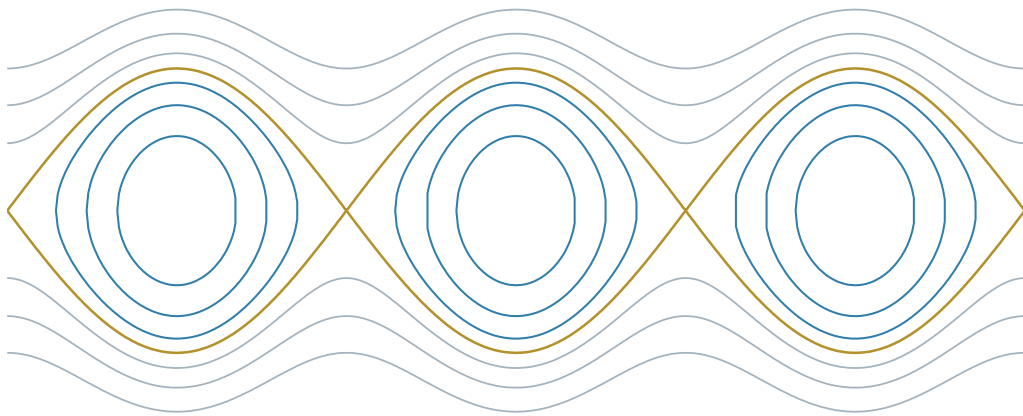


COURSE NOTES

Differential Equations

Steve Trettel



A first course in differential equations, told through the geometry of flows: what equations mean, when they can be solved, and how to see the behavior of solutions even when they can't.

Contents

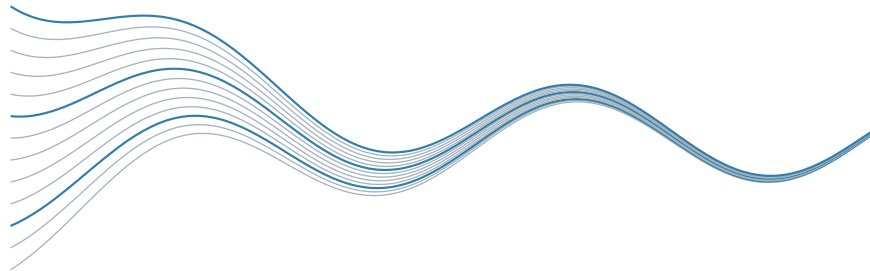
I	Local Rules, States, and Models	1
1	What is a Differential Equation	3
1.1	Local Rules	4
1.2	Slope Fields	6
1.3	Slope Fields with More Variables	10
2	Finding Solutions	19
2.1	Guessing and Checking Solutions	20
2.2	Antidifferentiation	22
2.3	Separating Variables	25
2.4	Undoing the Product Rule	31
2.5	Euler's Method	33
3	A Glimpse of the Theory	41
3.1	Existence and Uniqueness for Linear Equations	42
3.2	Existence and Uniqueness for Separable Equations	46
3.3	General Existence and Uniqueness	51
3.4	Understanding the Guarantees	52
3.5	Qualitative Solutions	53
4	Building and Exploring Models	71
4.1	Newton's Law of Cooling	72
4.2	Population Growth	74
4.2.1	From Exponential to Logistic Growth	74
4.2.2	Harvesting	82
4.3	Modeling with Systems	85
4.3.1	The Coffee and the Room	85
4.3.2	Interacting Populations	87

4.4	Forces and Second-Order Equations	97
4.4.1	From a restoring force to an equation	98
4.4.2	What motion does the model predict?	98
4.4.3	Revising the model: resistance to motion	100
4.5	Appendix: Modeling Earth's Temperature	101
4.5.1	A planet with one temperature	101
4.5.2	Adding an atmosphere	103
II	Linear Equations and Flows	115
5	Equations and Solution Spaces	117
5.1	Function Spaces and Solution Sets	118
5.1.1	Parameters for a Family of Functions	120
5.1.2	Linear Differential Equations	122
5.1.3	Affine Solution Sets	124
5.1.4	Nonlinear Solution Sets	125
5.2	A Common Formulation: First-Order Systems	127
5.2.1	Recording the Derivatives	128
5.2.2	Coupled Equations and the State	130
5.3	Existence, Uniqueness, and Initial Data	131
5.3.1	Initial Data as Parameters	132
6	Linear Evolution	139
6.1	Initial Data as Linear Coordinates	139
6.1.1	Higher-Order and Forced Equations	141
6.2	Bases of Solutions and Reconstruction	142
6.2.1	Solve a Few Times, Solve Them All	143
6.2.2	Choosing the Standard Initial States	145
6.3	Linear Evolution and Its Matrix	146
7	The Matrix Exponential	151
7.1	The Generator of Rotation	152
7.2	Repeated Small Steps	153
7.3	The Matrix Exponential	156
7.4	Matrix Flows We Can Compute	160
8	The Geometry of Linear Flows	165
9	Modes of Linear Systems	169

10 Forcing and Linear Response	175
10.1 The Homogeneous Family Behind a Forced Equation	175
10.2 Equilibrium Shifts in Linear Systems	175
 IIILow-Dimensional Nonlinear Dynamics	 177
11 Beyond Linear Flows	179
12 Local Nonlinear Dynamics	181
13 Cycles in the Plane	183
14 Bifurcations of Equilibria	185
15 Bifurcations of Oscillations	187
 A Calculus Review	 191
A.1 Differentiation Rules	191
A.2 Elementary Functions	193
A.2.1 Powers, reciprocals, and roots	193
A.2.2 Sine and cosine	193
A.2.3 Exponentials and logarithms	195
A.3 Integration and the Fundamental Theorem	197
A.3.1 Which integral are we writing?	197
A.3.2 Logarithms and variable endpoints	199
A.4 Substitution and Integration by Parts	199
A.5 Partial Fractions	201
 B Linear Algebra Review	 205
B.1 Vectors and Vector Spaces	205
B.2 Subspaces and Nonlinear Subsets	207
B.3 Bases and Dimension	209
B.4 Linear Maps, Kernels, and Images	211
B.5 Matrices and Coordinates	214
B.6 Linear Equations and Affine Spaces	216
B.7 Eigenvalues and Eigenvectors	218

PART I

Local Rules, States, and Models



What is a Differential Equation

This course is about differential equations. But what is a differential equation? We are all familiar with *algebraic* equations, where there is an unknown number (say x) satisfying some relationship involving the operations of algebra, like $x^2 + 4x = 5$. And *solving* an algebraic equation means to find the value(s) of x that actually realize this relationship.

A differential equation is a similar type of object, but for calculus. The unknown is no longer a *number*, but rather a *function* (say f), and the relationship can involve the operations of *calculus*, like $f'' + 4f' = 5f$. Solving a differential equation means finding the function(s) which actually have this relationship with one another.

There are many superficial similarities between differential equations, and the algebraic case you already know. Just like polynomials can come in many different *degrees* like $x + 2 = 5$ or $x^2 - 3x + 2 = 0$ or $x^5 - 3x^3 = 2$, differential equations come in different *orders*, where the order is defined as the highest derivative showing up in the equation. So $f'' = 4f$ is a second order equation, and $\frac{d^3}{dx^3}y = 4y + e^x$ is third order. And like algebraic equations, they can also come in *systems*: the two algebraic equations

$$\begin{aligned}x + y &= 3, \\x - y &= 1\end{aligned}$$

form a system for the two unknown numbers x and y . A system of differential equations asks the same thing of unknown functions. For example,

$$\begin{aligned}f' &= g, \\g' &= -f\end{aligned}$$

is a system for the two unknown functions f and g . The solution to this system is then a pair of functions: after thinking about the system a bit, can you see what a solution could be?

Like in algebra, checking a solution works is easy: if I give you $x = 1$ or $x = -5$ you can directly plug them into the equation and observe you get the same thing on both sides. And, if I give you $f = e^t$ or $f = e^{-5t}$, you can differentiate, plug into the differential equation, and see that each side gives you the same function.

In both cases the theory of *how to find these solutions* is more involved: even though you learn the rules of arithmetic in elementary school you don't learn to solve algebraic equations until high school. And similarly for differential equations: you learned the rules of differentiation in Calculus, but now it's finally our time to *put them to use!*

1.1 Local Rules

It's not really that much of an exaggeration to say that differential equations run the world. From economic models, across engineering and the sciences, to our best theories of fundamental physics, all are phrased in the language of differential equations. Perhaps much of this stems from the great mystery that our universe is so well described by mathematics. But there is also a practical, human reason that differential equations show up so often in our lives: it is usually much easier to describe what is happening *right now* than to describe the whole future at once.

The derivative is exactly this sort of local information. If $y(t)$ records some quantity which is changing over time, then $y'(t)$ tells us how it is changing near one particular instant. For a small step h ,

$$y(t + h) \approx y(t) + hy'(t).$$

Thus, an equation like

$$y' = F(t, y)$$

is a rule for how the function should behave nearby each point: if we know the current time t and the current value y , the right hand side tells us the current rate of change. The function $y(t)$ is the unknown history; $F(t, y)$ is the known local rule.

This is often exactly the kind of knowledge the world gives us. Imagine a hot cup of coffee sitting in a room. It may be difficult to say immediately what its temperature will be twenty minutes from now. But it is easy to say how its temperature is changing now: a cup hotter than the room cools down, a cup colder than the room warms up,

and a larger temperature difference produces a faster change. The simplest rule with these properties is

$$T' = -k(T - T_{\text{room}}), \quad k > 0.$$

Our simple observation about the rate of heating and cooling, translated into mathematics, gives a *differential equation* for the temperature T . And if we learn to solve such equations (giving us a function $T(t)$) we will have the answer to our original question - how cool will the coffee be in 20 minutes - without any more input from physics at all. Differential equations allow us to package together simple relational information, and solving them *integrates* all these local relationships into a global story.

The same idea works when several things are changing together. It is hard to predict the complete histories of a population of hares and a population of lynx. But some local relationships are easy to express: more hares mean more food for lynx, while more lynx mean more encounters in which hares are eaten. This seems like almost no information, yet thinking slowly through how to express this *precisely* (which we will do in the next chapter) allows one to write down a rather simple set of expressions for *the rate of change* H' of the number of hares, and L' for the number of Lynx, in terms of their current populations:

$$\begin{aligned} H' &= rH - bHL, \\ L' &= cHL - dL. \end{aligned}$$

Neither population can now be understood by itself: their two local rules are coupled.

Or consider a mass attached to a spring. The farther the spring is stretched, the harder it pulls back. A simplest precise version of this would say that the force it pulls with is proportional to the stretch distance x . and Newton's law turns force into acceleration:

$$mx'' = -kx.$$

None of these local rules tells us the complete history all at once. A *solution* is an overall function—or collection of functions—which obeys the rule at every point. It integrates all those nearby instructions into one coherent relationship. This is the reversal at the heart of differential equations: differentiation begins with a known history and extracts its local behavior; a differential equation begins with the local behavior and asks us to reconstruct the possible histories.

We can now collect this language precisely.

Definition 1.1 (Differential equations and solutions). A **differential equation** is an equation whose unknown is a function and which relates that function to one or more of its derivatives. Its **order** is the order of the highest derivative which appears.

A **solution** on an interval is a function which has the required derivatives and makes the equation true at every time in that interval. If several unknown functions must satisfy several differential equations together, the equations form a **system**, and a solution is the corresponding collection of functions.

1.2 Slope Fields

In algebra it is useful to have a *graph* to help conceptualize a problem (say for a quadratic equation, you can immediately see how many roots there'll be if you can draw a picture, even if you haven't exactly solved for them yet). And we might hope to find an analog for differential equations - something we can draw that will tell us how things behave, even if we haven't done all the work of producing an exact solution yet.

But what's the graph of a differential equation? For the moment, let's focus on the simplest general case: one unknown function $y(t)$ and one first-order equation

$$y' = F(t, y).$$

In algebra, we don't draw the number x but the relationship $y = f(x)$. Similarly here we can't draw the function $y(t)$ (since we don't know it yet!) but we will try to draw the relationship itself. This relationship is given to us by a function $F(t, y)$ which we can compute at any point (t_0, y_0) . Since this is equal to y' , the relationship tells us nothing about the *value*, but rather it answers the question

If a solution passes through (t_0, y_0) , what slope must it have there?

For example, consider

$$y' = t - y.$$

If a solution passes through $(1, -1)$, its slope there must be

$$y' = 1 - (-1) = 2.$$

Thus we do not know the solution, but we know one of its tangent lines! We can record that information by drawing a tiny line segment of slope 2 at the point $(1, -1)$.

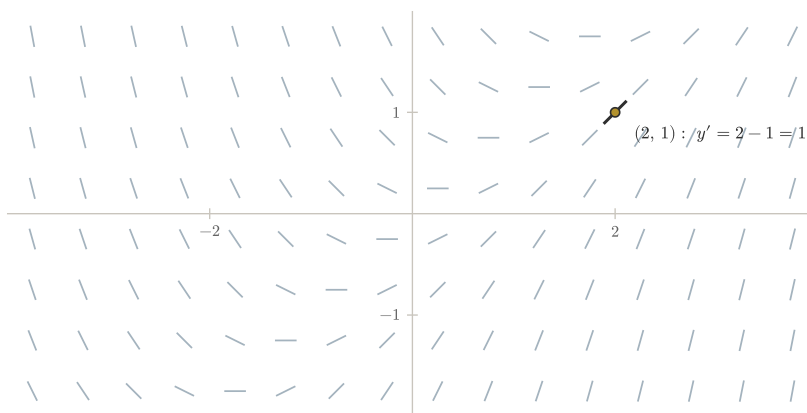


Figure 1.1 The slope field of $y' = t - y$. At each point the equation answers one small question: if a solution passed through here, what slope would it have? The field is that question asked everywhere — pure evaluation, no solving. In the live figure, move over the plane to ask the equation yourself, click to leave a slope behind, and try to draw a solution one slope at a time before revealing the field. (Draft caption — final wording at prose stage.)

Of course there is nothing special about this point. We can ask the same question at every other point in the plane and draw the answer there too.

The result is called a *slope field*. We have evaluated the rule everywhere, but we have not solved the equation. Different equations paint very different pictures. Some fields repeat their slopes along horizontal lines, others along vertical lines; some push curves away from the axis, while others pull them toward it. But every one of them is constructed and read in exactly the same way: evaluate the right hand side, then draw the resulting slope.

What does a *solution* look like in this picture? A function $y(t)$ is a solution precisely when its graph follows the field everywhere it goes. At each point $(t, y(t))$ on the graph, the curve's own slope $y'(t)$ must equal the slope $F(t, y(t))$ prescribed by the equation.

This is an everywhere requirement. A curve can follow the field perfectly for a long time and then make one wrong turn; at the point of that turn, it is not a solution. A function does not earn partial credit from the differential equation.

This pictorial description already has taught us something important: there is not just one solution to a differential equation. The slope field is filled by an entire family of curves, each following the same local rule. To choose one of them, we specify one piece of additional information: the value of the function at some time. Writing

$$y(t_0) = y_0$$

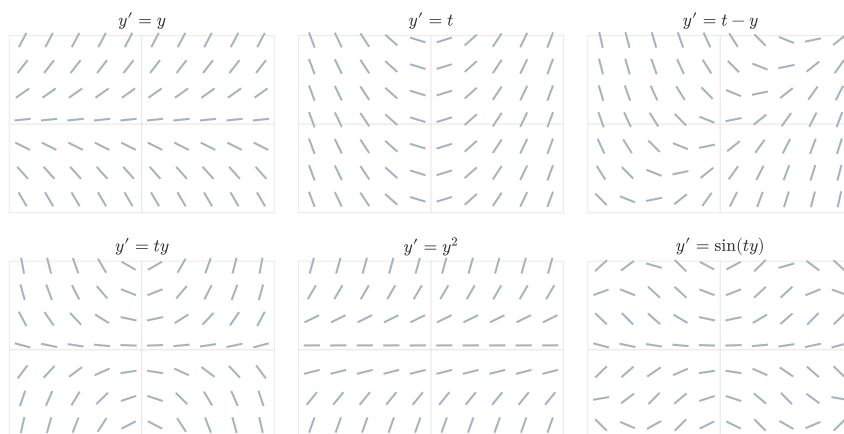


Figure 1.2 Six equations, six personalities. Every field is read the same way — flat ticks where $y' = 0$, steep ticks where the rule demands speed — but each equation paints its own portrait. (Draft caption — final wording at prose stage.)

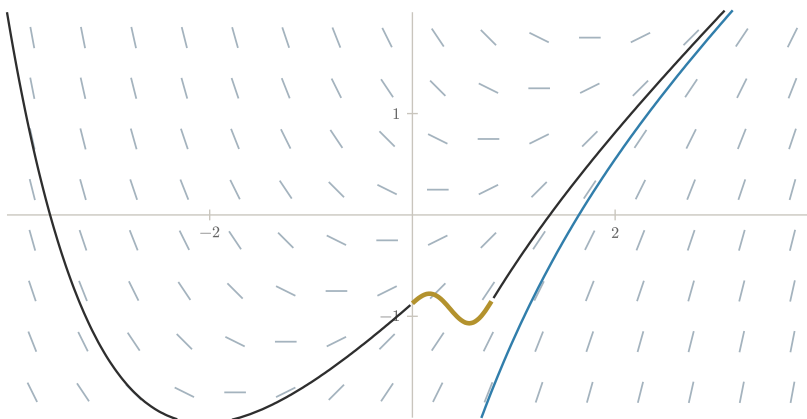


Figure 1.3 A solution and an impostor. The blue curve follows the field everywhere it goes. The other curve is a real solution on its dark stretches — but no single solution connects them, so the highlighted stretch is forced to disobey: watch it cut across the ticks. Being a solution is a pointwise obligation, and the failure has a location. In the live figure, click or drag anywhere to move the blue solution through that point; it remains tangent to the field everywhere. (Draft caption — final wording at prose stage.)

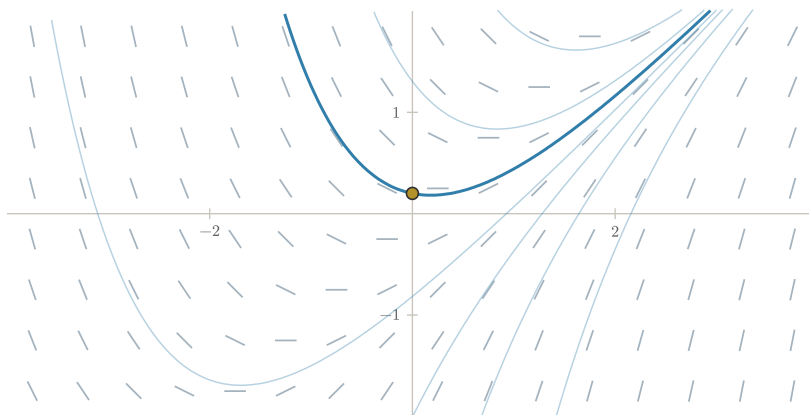


Figure 1.4 The gold point is an initial condition $y(t_0) = y_0$. In this figure, each starting point selects one member of the displayed solution family. Follow its curve in both time directions and compare its tangent with the field. In the live figure, click or drag to move the data. (Draft caption — final wording at prose stage.)

says that our solution must pass through the point (t_0, y_0) . This is called an *initial condition*, and the equation together with this condition is called an *initial value problem*:

$$\begin{aligned} y' &= F(t, y), \\ y(t_0) &= y_0. \end{aligned}$$

The differential equation describes the possible histories; the initial condition specifies a point the history must pass through.

Definition 1.2 (Initial conditions and initial value problems). An **initial condition** specifies the value of an unknown function at a chosen time, such as $y(t_0) = y_0$. An **initial value problem** consists of a differential equation together with initial conditions prescribed at one time. A solution of the initial value problem must satisfy both the equation and all of the initial conditions.

The slope field does tell us one thing about how solution curves can meet. At each point, the equation prescribes just one slope. If two solutions pass through the same point, they must have the same tangent there. They cannot meet with different tangents.

But could two curves meet with the same tangent and then separate? Does an initial condition always select exactly one solution, and how far can we follow it? We

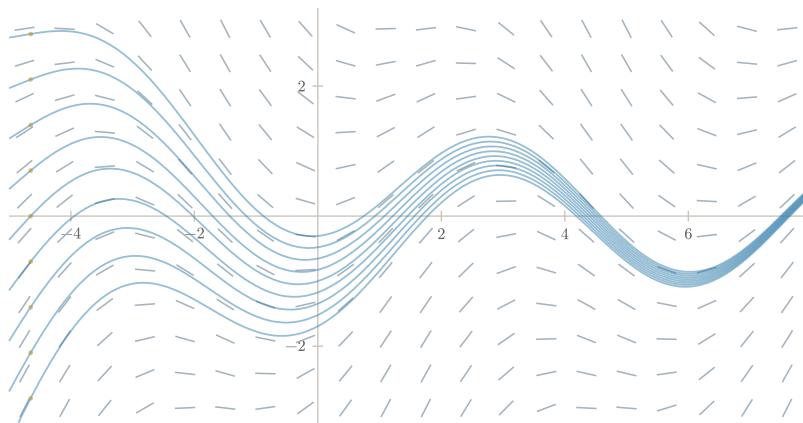


Figure 1.5 Build the flow. Click anywhere to add the complete solution through that point, or drag across the field to paint many solutions at once. Compare the blue histories with the slopes prescribed along them. Use the menu to change the local rule: the streams may gather, spread, or divide into horizontal bands, but every curve is still assembled from the slopes it meets. (Draft caption — final wording at prose stage.)

will investigate these questions in chapter 3. For now, we can explore the histories suggested by the slope field.

Even without a formula for a single solution, a slope field can tell us where solutions rise and fall, where they have horizontal tangents, and how different starting points may lead to different futures. This is genuine information about the unknown functions, extracted directly from the differential equation.

Try turning the local picture into a global one yourself. Each time you choose a point, the differential equation supplies the slopes which carry one solution through it. Add enough starting points and the slope field begins to fill with histories: the local rule becomes a flow.

1.3 Slope Fields with More Variables

So far our slope fields have described equations with one unknown function, $y(t)$. But we began this chapter with systems involving several unknown functions at once. What should a slope field look like for a system such as

$$\begin{cases} f' = F(t, f, g), \\ g' = G(t, f, g)? \end{cases}$$

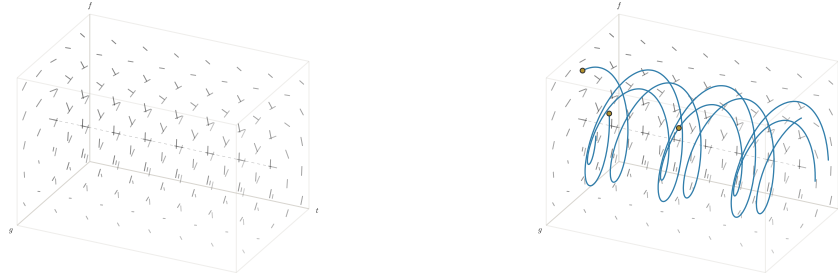


Figure 1.6 The slope field of the system $f' = g$, $g' = -f$, drawn in (t, f, g) -space. At each point the system prescribes one direction — one step of time, f' steps of f , g' steps of g — and the field is that answer drawn everywhere: the system itself, drawn. In the live figure, click near a tick to grow the integral curve through that point, drag to rotate, and switch to the damped sibling. (Draft caption — final wording at prose stage.)

A solution is now a *pair* of functions, $f(t)$ and $g(t)$. To graph them together, we need one axis for t , one for f , and one for g . Thus the graph of the solution is a curve in (t, f, g) -space:

$$\gamma(t) = (t, f(t), g(t)).$$

And what does the differential equation tell us at a possible point (t, f, g) ? It gives the two rates $F(t, f, g)$ and $G(t, f, g)$. If time moves forward by a small amount Δt , then f changes by approximately $F\Delta t$ and g changes by approximately $G\Delta t$. The total displacement is therefore $\Delta t(1, F, G)$, so the system assigns the direction

$$(1, F(t, f, g), G(t, f, g)).$$

This is exactly the slope-field idea again, only with one more component. We evaluate the equation at many possible points and draw a small segment in the direction it prescribes. The page has become a three-dimensional box, but the thought has not changed.

The figure uses the system

$$f' = g, \quad g' = -f,$$

which appeared near the beginning of the chapter. Its segment at (t, f, g) points in the direction $(1, g, -f)$. Clicking a segment asks for the solution through that point, and the curve grows by following one prescribed direction after another.

This also tells us exactly what it means for a pair of functions to solve a system. Along its graph,

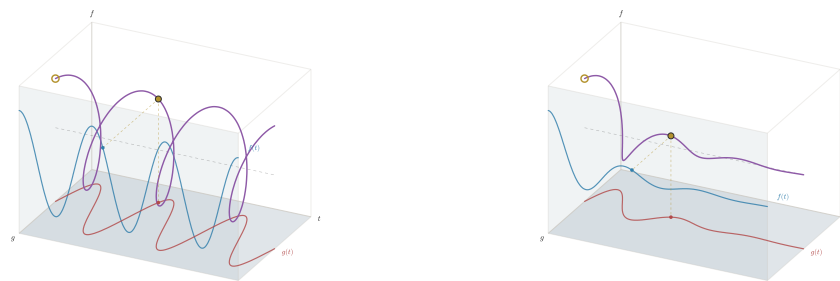


Figure 1.7 A solution of a system is a curve in space. The violet curve solves $f' = g$, $g' = -f$; its shadows on the wall and floor are the graphs of the two unknowns — a pair of waves, one lagging the other. The dashed line it winds around is the equilibrium solution. In the live figure, drag to rotate, drag the gold ring to move the initial condition, and switch to the damped sibling, whose curve corkscrews in toward the axis. (Draft caption — final wording at prose stage.)

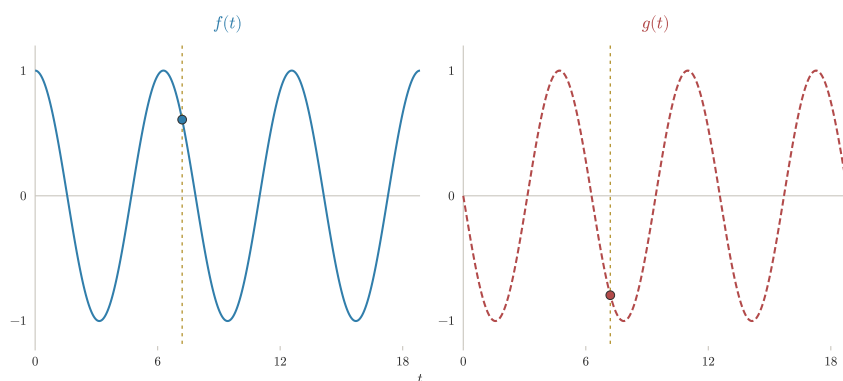


Figure 1.8 The two component graphs, with points marking the same time. Together they show one solution of $f' = g$, $g' = -f$, with $f(0) = 1$ and $g(0) = 0$.

$$\gamma'(t) = (1, f'(t), g'(t)) = (1, F(t, f(t), g(t)), G(t, f(t), g(t))),$$

so the curve is tangent to the field everywhere it goes. An initial condition must now specify both unknowns, say $f(t_0) = f_0$ and $g(t_0) = g_0$. That one point (t_0, f_0, g_0) again chooses one entire history.

The two shadows give us another way to draw the solution: put the graph of $f(t)$ beside the graph of $g(t)$. We will often display solutions this way, since ordinary graphs make it easy to read values and compare how the functions change.

These are two functions belonging to one solution of the system. At each time, we need both values to know what happens next. The equation $f' = g$ tells us that the slope of the first graph is the height of the second; $g' = -f$ tells us that the slope of the second is the negative of the height of the first.

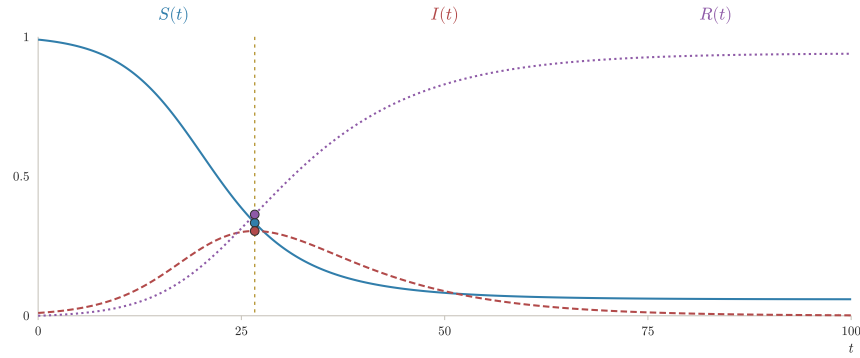


Figure 1.9 $S(t)$, $I(t)$, and $R(t)$ on shared axes, with a common time marker. The three curves show one constructed solution of the system.

We can therefore still read the differential equations in these pictures, but we can no longer draw the full slope field in either panel. A point (t, f) leaves out g , which we need to determine the slope. The three-dimensional picture held all that information at one point; here we have to look at the two graphs together.

This way of plotting solutions also works when there are more unknowns. For example, a simple model of an epidemic tracks three fractions of a population: the susceptible $S(t)$, the infectious $I(t)$, and the recovered $R(t)$. One such model is

$$\begin{aligned} S' &= -\beta SI, \\ I' &= \beta SI - \gamma I, \\ R' &= \gamma I, \end{aligned}$$

where $\beta, \gamma > 0$ describe the rates of infection and recovery. We will build this model from its assumptions in chapter 4. For now, look at what a solution consists of: three functions whose derivatives satisfy these three equations together.

The combined graph $(t, S(t), I(t), R(t))$ would use four coordinates. But we can easily draw the three functions against time. Here they share the same units, so we have placed them on the same axes. At the marked time, reading all three curves gives one snapshot of the epidemic.

Definition 1.3 (States and state spaces). The **state** at a particular time is the complete collection of current values which must be prescribed as initial data at that time. The set of all possible states is the **state space**. The individual values making up a state are its **state coordinates**.

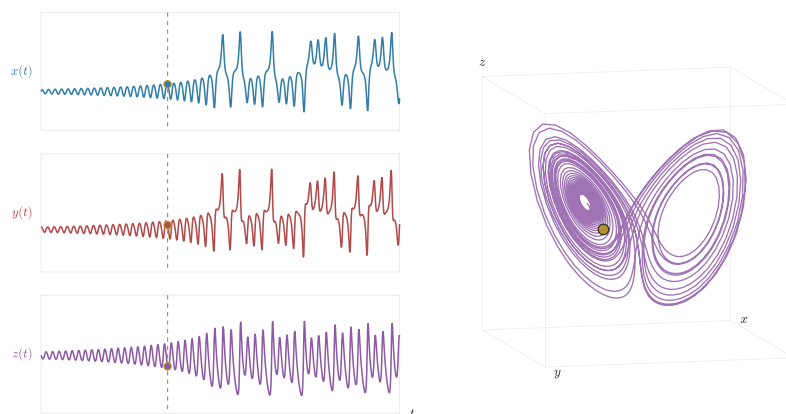


Figure 1.10 The marked values on the three time graphs give the coordinates of the gold point on the space curve, all at the same time.

For a famous example, here is a solution of the Lorenz system, drawn both as three functions of time and as the parametric curve $(x(t), y(t), z(t))$ in three-dimensional space.

We have learned several ways to picture a solution. But how do we find the functions whose graphs we are drawing? In the next chapter we will put our calculus techniques to work, beginning with the simplest possibility: guess a function and check whether it works.

PRACTICE • EXPLORATION • COMPUTATION

Problems

The first group checks the language and basic moves of the chapter. The explorations ask you to push those ideas somewhere the text has not already gone. The Python problems begin with supplied code: your job is to read it, change it, and explain the resulting picture.

CHECK YOUR UNDERSTANDING

1. Reading a local rule

A room is kept at 20°C . An object's temperature satisfies

$$T' = -k(T - 20), \quad k > 0.$$

- (a) Is the object warming or cooling when its temperature is 10°C ? What about 30°C ? At what temperature does the rule prescribe no change?
- (b) Compare the rates of change at 30°C and 40°C . What does the equation say about their relative sizes?
- (c) Two objects obey this same equation with the same k . One has just been placed in the room; the other has been there for an hour. If both are currently at 30°C , what does the rule say about their current rates of change? Does it need to know how they reached that temperature?

2. Obeying the equation at a point

Consider the differential equation $y' = y^2$.

- (a) Verify by substitution that $y(t) = \frac{1}{2-t}$ is a solution, on the interval $t < 2$.
- (b) The function $z(t) = \frac{1}{1+t^2}$ is smooth on the whole line, and its graph looks perfectly plausible. Show that it obeys the differential equation at exactly one point, and find that point. (In the language of figure 1.3: this curve is an impostor, and the failure has a location — in fact, every location but one.)

3. Checking the equation and the initial condition

Consider

$$y' = y, \quad y(0) = 2.$$

- (a) Check each proposed function against both requirements:

$$y(t) = e^t, \quad y(t) = 2e^t, \quad y(t) = 2 + t.$$

Which satisfy the differential equation? Which satisfy the initial condition? Which satisfy both?

- (b) Two proposed solution curves pass through the same point. One has slope 0 there and the other has slope 1. Could both solve the same equation $y' = F(t, y)$? Explain using the slope assigned at that point.
- (c) The curves $y = 0$ and $y = t^3$ meet at the origin. Find their slopes there. Does the observation that an equation assigns only one slope at each point rule out both curves being solutions? Explain what that observation does—and does not—tell us.

4. A solution you can spot

Suppose all you know about a differential equation is its shape, $y' = (y - 2t)g(t, y) + 2$, where g is some continuous function nobody has told you.

- (a) Show that $y(t) = 2t$ is a solution no matter what g is.
- (b) Explain in a sentence why no computation involving g was ever needed.

5. Two fields by hand

- (a) On the window $-2 \leq t \leq 2$, $-2 \leq y \leq 2$, sketch the slope field of $y' = -y$ by hand, evaluating at integer points. Then do the same for $y' = t$ on a second copy of the window.
- (b) Each of your two fields repeats itself, but in different directions: one is unchanged as you slide the picture horizontally, the other as you slide it vertically. Say which is which, and explain how you could have predicted this from the equations without drawing anything.
- (c) On each sketch, draw the solution curve passing through $(0, 1)$ by following your ticks.

6. Three rules, three signatures

Consider the three equations $y' = 1 - t$, $y' = t - 2y$, and $y' = t(1 - y)$. For each equation, find all points where the prescribed slope is zero, and sketch the resulting lines. Then decide whether its field repeats along vertical lines, horizontal lines, or neither. Explain how these signatures would let you recognize the equation from its slope field without solving it.

EXPLORATIONS

7. The shape of the field is the shape of the rule

- (a) A slope field has the property that its ticks are identical along every vertical line. What does this say about the right-hand side $F(t, y)$?
- (b) Another field has ticks identical along every *horizontal* line. What does this say about F ?
- (c) For an equation of the kind in the previous part — one whose rule involves only y , say $y' = f(y)$ — show that sliding a solution in time produces another solution: if $y(t)$ is a solution and c is any constant, then $z(t) = y(t - c)$ is also a solution. Explain how this fact and the horizontal repetition of the field are the same observation made twice.

8. One curve, two shadows

Consider the system $f' = g$, $g' = -f$.

- (a) Verify that $f(t) = \sin t$ and $g(t) = \cos t$ solve *both* equations.
- (b) Write the direction assigned by the system at $(t, f, g) = (0, 2, -3)$. Interpret the signs of its three components.
- (c) Draw the graphs of $f(t)$ and $g(t)$ in separate panels with the same time scale. Use $f' = g$ and $g' = -f$ to explain how the height of either graph determines the slope of the other at the same time.

PYTHON: PLOTTING LOCAL RULES

These problems use NumPy and Matplotlib. Begin with the supplied code and make one change at a time; a plot is part of your answer, but so is the sentence which explains what it shows.

9. Build a slope field

The code below draws the slope field of $y' = t - y$.

```
import numpy as np
import matplotlib.pyplot as plt

t = np.linspace(-3, 3, 21)
y = np.linspace(-3, 3, 21)
T, Y = np.meshgrid(t, y)
S = T - Y

length = np.sqrt(1 + S**2)
plt.quiver(T, Y, 1 / length, S / length,
           angles="xy", pivot="mid", color="steelblue")
plt.xlabel("t")
plt.ylabel("y")
plt.show()
```

- (a) Explain the jobs of `meshgrid`, `S`, and `length`. Why do the two components passed to `quiver` describe a segment of slope S ?
- (b) Change only the line defining `S` to draw the fields of $y' = -y$ and $y' = t$. For each field, identify the direction in which the pattern repeats.
- (c) Create a Boolean mask for points where $\text{abs}(S) < 0.15$ and plot those points on top of one field. What curve is your code detecting?

10. Plot the components of a solution

Use the following samples from the solution of $f' = g, g' = -f$.

```
t = np.linspace(0, 4 * np.pi, 500)
f = np.sin(t)
g = np.cos(t)
```

Create a figure with two panels: f against t and g against t . Give every axis a label and use the same time scale in both panels. Mark the two values at $t = \pi/4$. Explain why the two graphs together describe one solution of the system, and why either graph alone does not show its full slope field.

Finding Solutions

In the last chapter we learned what a differential equation *says*. A local rule assigns a slope at each possible time and state, and an initial condition specifies a point our solution must pass through. Drawing the slope field lets us begin to picture the possible histories. But how do we actually find one? How do we turn the local rule into a function we can evaluate, or a calculation we can carry out?

In one sense, that question is the subject of this entire book! But our first attempt will be deliberately modest. We will focus mostly on one unknown function and one first-order equation

$$y' = F(t, y).$$

We will take advantage of this simple situation to explore the different things one might want when asking for a solution—there’s no one-size-fits-all technique, nor should there be! Sometimes we want a *formula* for an answer, something we can recognize from calculus or derive using its techniques, then plug things into and get precise predictions. Other times we may know that whatever formula exists is too complicated to be useful, or that it’s essentially impossible to write down—in these cases we might prefer a numerical solution: an algorithm that follows the local rule to give us *good approximations* to the answer quickly, while avoiding thousands of pages of calculus and algebra. And other times we may find a simple plot reveals enough information that we don’t need a formula or rigorous numerics at all.

We will spend a little more time on the ‘formula’ case here, but not because it is necessarily more important. This chapter is also a chance to put some familiar calculus back to work: antidifferentiation, the chain rule, and the product rule. The point will not be to memorize a new collection of formulas. It will be to look at an equation and ask: is there a piece of calculus here which we already know how to undo?

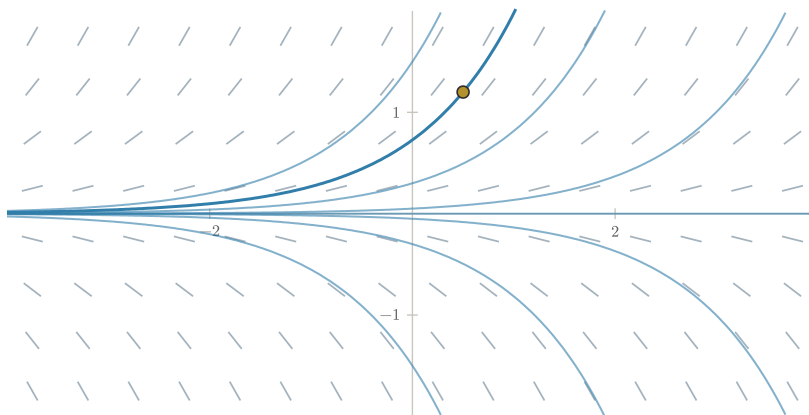


Figure 2.1 The family $y = Ce^t$ threading the slope field of $y' = y$. Every member satisfies the equation; the constant C is chosen by one piece of data. The $C = 0$ member lies exactly on the t -axis: the axis itself is a solution. In the live figure, drag the gold point — the readout answers with the formula.

2.1 Guessing and Checking Solutions

First things first: how do we confirm that we have a solution to a differential equation? If we are given a formula, we differentiate it and substitute into the equation to check that the two sides agree. If there is an initial condition, we check that too. We know how to differentiate elementary functions, so when someone proposes an elementary formula as a solution, we have a direct way to check their answer ourselves!

This opens up a possibility: *guess and check* could be a rigorous solution technique. We might recognize a function whose derivatives behave the way the equation asks, or try a few possibilities until something works. However we arrive at our guess, once we verify that it satisfies the equation and the initial condition, we have found a solution.

The simplest version of this idea is

$$y' = y.$$

We need a function which is equal to its own derivative, so e^t is a natural guess. Multiplying by any constant still works:

$$\frac{d}{dt}(Ce^t) = Ce^t.$$

Thus $y = Ce^t$ gives an entire family of solutions. One initial condition selects the value of C , and therefore selects one member of this family.

Recall the differential equation we placed beside its algebraic counterpart at the very beginning of the book:

$$f'' + 4f' = 5f.$$

At first this looks intimidating, since it asks for a sum of derivatives of f to somehow equal a multiple of f . But one (simple!) way this could happen is if each derivative were itself a multiple of f . Then we would just be adding multiples of the same function, and it would be no surprise to get another multiple of it. And we actually know functions whose derivatives are multiples of themselves—exponentials!

This suggests we might guess e^{rt} as a solution. Then $f' = re^{rt}$ and $f'' = r^2 e^{rt}$, so substitution gives

$$(r^2 + 4r)e^{rt} = 5e^{rt}.$$

Cancelling the exponential from both sides (since it is never zero, we can do this) reduces the differential equation to the algebraic equation

$$r^2 + 4r = 5.$$

Its solutions are $r = 1$ and $r = -5$, giving the two functions e^t and e^{-5t} . The resemblance between the two opening equations was not only superficial: the exponential *guess* turns the differential equation back into the algebraic one.

Remembering other relationships from calculus can extend our ability to guess (and check) even more equations. For the normalized spring equation

$$x'' = -x,$$

the functions $\cos t$ and $\sin t$ are natural candidates, since taking two derivatives negates each of them. And a direct check shows that these work—as does any sum or multiple of them, in fact! The spring oscillates just as we would expect.

Guessing and checking is a genuine solution technique, but not a systematic one: it works when we can recognize the right kind of function. So what else can we borrow from calculus? Not only remembered relationships, but the techniques themselves.

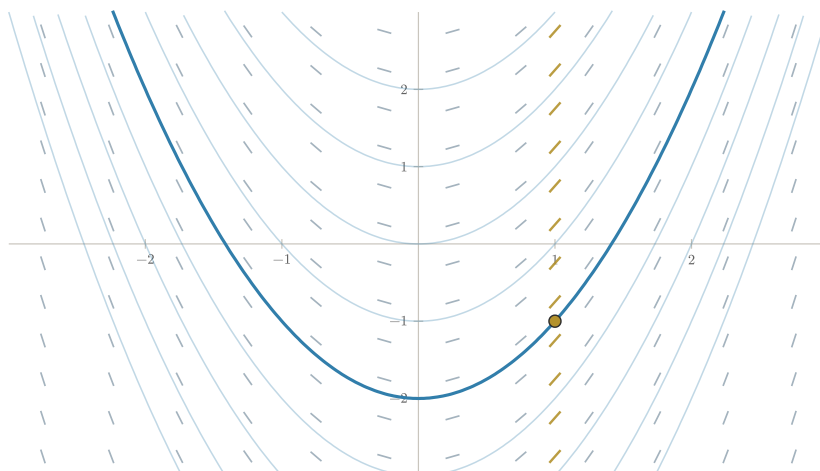


Figure 2.2 When $y' = g(t)$, the slope depends only on the horizontal coordinate. Every tick in one vertical column is therefore parallel, and every solution is a vertical translate of every other one. Choose a rule and drag the gold initial point; the highlighted column shows the one slope shared by every possible value of y at that time, while the readout turns the chosen initial condition into a value of C .

2.2 Antidifferentiation

So far we have borrowed facts from calculus to help us guess. But we can also borrow the techniques of calculus to produce solutions more systematically. In the easiest case, the differential equation simply tells us the derivative of the function we are looking for:

$$y' = g(t),$$

and the important thing here is that the right-hand side depends only on t , not on y . The slope does not care about the current *value* of the function at all; it cares only about the input. At a fixed time t , every point (t, y) therefore receives exactly the same slope, no matter how high or low the point lies. In the slope field, this shows up as entire vertical columns of identical ticks.

This is just the kind of problem calculus has always taught us to solve. We have some mystery function y , we know its derivative, and we want the function back. We undo the derivative: we antidifferentiate. If $G' = g$, then

$$y(t) = G(t) + C.$$

You have been solving differential equations this way since your first calculus course; we just did not usually call them differential equations.

The constant C is not an annoying ambiguity left behind by integration. It is the family of vertical translates we just saw in the slope field, written algebraically. The initial condition tells us which translate we want: it specifies one point the curve must pass through, and that fixes how high or low the curve sits.

For example, consider the initial value problem

$$\begin{cases} y' = 2t, \\ y(1) = 3. \end{cases}$$

Antidifferentiating gives $y = t^2 + C$, which is still a whole family of possibilities. Now the initial condition does its job. It says that our curve must pass through $(1, 3)$, so

$$3 = 1^2 + C.$$

Thus $C = 2$, and the particular history through our chosen initial point is

$$y(t) = t^2 + 2.$$

Sometimes even this tiny amount of differential-equation language is hidden by a little algebra. Suppose instead we are given

$$2y' + 4t = 6.$$

At first this is a relationship involving y' , rather than a formula for it. But we can solve for the derivative just as we would solve for any other unknown:

$$y' = 3 - 2t.$$

And now we are back in the situation we already understand. Integrating gives

$$y = 3t - t^2 + C.$$

There was no special differential-equation trick here. We used algebra to put the equation into a familiar form, and then used calculus to undo the derivative.

There is another way to write the same reasoning which will be more useful to us. The Fundamental Theorem of Calculus says that the total change from t_0 to t is the accumulated rate of change:

$$y(t) - y(t_0) = \int_{t_0}^t y'(s) \, ds.$$

Consequently, we obtain a result worth keeping close at hand.

Proposition 2.1 (*Accumulating a known rate*). Let g be continuous on an open interval I containing t_0 . Then the initial value problem

$$\begin{cases} y' = g(t), \\ y(t_0) = y_0 \end{cases}$$

has exactly one solution on I , given by

$$y(t) = y_0 + \int_{t_0}^t g(s) \, ds.$$

Any solution must satisfy this formula, by the Fundamental Theorem of Calculus calculation above. Conversely, differentiating the formula gives $y' = g(t)$, and setting $t = t_0$ gives $y(t_0) = y_0$. So the formula both produces a solution and leaves no room for a second one.

This formula says something wonderfully simple: start with the value you already have, then add up all the change that happens along the way. It is the local-to-global idea from the beginning of the book, now made into a calculation.

Does writing a definite integral really count as finding the solution? Consider

$$y' = e^{-t^2}, \quad y(0) = 0.$$

Our formula gives

$$y(t) = \int_0^t e^{-s^2} \, ds.$$

The integrand has no elementary antiderivative, so we cannot finish by replacing the integral with a familiar combination of functions. But the integral already defines a function! The Fundamental Theorem of Calculus tells us its derivative, and its value at $t = 0$ is zero, exactly as required.

This is an exact solution. If we want a decimal approximation to $y(1)$, we can approximate the integral with a Riemann sum. Those are two different accomplishments: the integral specifies the answer exactly, and the sum gives us a practical way to compute a value.

2.3 Separating Variables

We know what to do when the equation gives us $y' = g(t)$: integrate the known rate. But what if it gives us

$$y' = f(y)?$$

Equations whose rate depends only on the current value of y , with no explicit dependence on time, are called *autonomous*. We can still integrate both sides from the initial time t_0 to t :

$$y(t) - y_0 = \int_{t_0}^t f(y(s)) \, ds.$$

But now the function we are trying to find appears inside the integral! Can we rearrange the equation so that the integration becomes something we know how to do?

One possibility is to divide by $f(y)$, putting all the dependence on y on the left. But dividing by an expression involving our unknown requires some care.

Think back to $y' = y$. Dividing by y would give

$$\frac{y'}{y} = 1.$$

Yet $y(t) = 0$ solves the original equation: both sides are zero. It cannot satisfy the divided equation, because the quotient would be $0/0$! The original equation allows a solution that our rearrangement excludes.

We can check for this possibility before dividing. For $y' = f(y)$, suppose $f(a) = 0$. Then the constant function $y(t) = a$ satisfies the equation:

$$y'(t) = 0 = f(a).$$

The value stays fixed because the rule keeps prescribing zero change.

So we first look for zeros of f and record the corresponding constant solutions. Then, on an interval where $f(y(t)) \neq 0$, we can safely divide:

$$\frac{y'(t)}{f(y(t))} = 1.$$

Our calculation now applies on that interval; the constant solutions we found separately remain solutions of the original equation.

Integrating both sides, we obtain

$$\int \frac{y'(t)}{f(y(t))} dt = \int 1 dt.$$

The left side is a familiar substitution problem! The denominator contains $y(t)$, and the numerator contains its derivative. Set

$$u = y(t), \quad du = y'(t) dt.$$

The substitution turns the equation into

$$\int \frac{du}{f(u)} = t + C.$$

Now the integrand is a known function of the integration variable u . We can look for its antiderivative using our calculus techniques, even though we do not yet know the function $y(t)$ which supplied the substitution.

The initial condition determines the constant. If $y(t_0) = y_0$, we can write the resulting relationship as

$$t - t_0 = \int_{y_0}^{y(t)} \frac{du}{f(u)}.$$

Notice what the integral gives us: the elapsed time expressed in terms of the value of y reached. To recover y as a function of t , we still need to solve this relationship for y . Sometimes that is easy; sometimes the integral relationship is the most useful way to leave the answer.

This is the idea of *separation of variables*: arrange the equation so that each side can be integrated with respect to just one variable. The substitution we worked through is often written more briefly as

$$\frac{dy}{f(y)} = dt, \quad \int \frac{dy}{f(y)} = \int dt.$$

Here the dy packages the substitution $dy = y'(t) dt$; it is the same calculus calculation in a shorter notation.

And the same move works for a larger class of equations. Suppose

$$y' = g(t)h(y).$$

After checking the constant solutions where $h(y) = 0$, we can divide on intervals where $h(y) \neq 0$ to obtain

$$\frac{y'(t)}{h(y(t))} = g(t).$$

Integrating and making the same substitution on the left gives

$$\int \frac{dy}{h(y)} = \int g(t) dt.$$

Now the right side also asks us to integrate a function, but the two variables have been separated: each side is an antidifferentiation problem in one variable.

Definition 2.2 (Separable equations). A first-order equation is **separable** on a region if it can be written

$$y' = g(t)h(y).$$

On any part of the region where $h(y) \neq 0$, **separation of variables** rewrites the equation as

$$\frac{dy}{h(y)} = g(t) dt$$

and integrates both sides. Any values where $h(y) = 0$ must be checked separately before dividing.

Let us first work through an example where we *can* finish by solving for y . Consider

$$\begin{cases} y' = ty, \\ y(0) = 2. \end{cases}$$

Before dividing by y , we notice that $y = 0$ is a constant solution. Our required initial value is 2, so any solution with this initial value must remain positive on a sufficiently short interval around $t = 0$, by continuity. On that interval we can divide by y and separate:

$$\frac{dy}{y} = t dt.$$

Integrating both sides gives

$$\ln |y| = \frac{t^2}{2} + C.$$

This time the last relationship is easy to invert. Exponentiating and absorbing the possible sign into the arbitrary constant gives

$$y = Ae^{t^2/2}.$$

Allowing $A = 0$ even puts the constant solution back into the family. Finally, the initial condition says $2 = Ae^0$, so $A = 2$ and

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

Now check what we have found:

$$\frac{d}{dt} \left(2e^{t^2/2} \right) = 2te^{t^2/2} = ty, \quad y(0) = 2.$$

We began the calculation on a short interval where division was allowed, but the resulting function satisfies the equation and initial condition on the whole real line.

Let us try another equation where the integration leaves us with a little more algebra:

$$y' = \frac{t}{1+y}, \quad y(0) = 1.$$

Separating and integrating gives

$$(1+y) dy = t dt, \quad y + \frac{y^2}{2} = \frac{t^2}{2} + C.$$

The initial condition says $1 + 1/2 = C$, so $C = 3/2$. Multiplying by 2 and rearranging, we obtain

$$y^2 + 2y = t^2 + 3.$$

How do we extract y from this relationship? Complete the square:

$$(y+1)^2 = t^2 + 4, \quad y = -1 \pm \sqrt{t^2 + 4}.$$

At $t = 0$, the positive branch gives $y = 1$ and the negative branch gives $y = -3$. Our initial condition selects the positive branch, so

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

Differentiating verifies the equation:

$$y'(t) = \frac{t}{\sqrt{t^2 + 4}} = \frac{t}{1 + y(t)}.$$

The denominator $1 + y(t) = \sqrt{t^2 + 4}$ never vanishes, and $y(0) = 1$, so this formula solves the initial value problem on the whole real line.

But separation does not always leave a relationship which is useful to invert. For example, suppose

$$y' = \frac{t}{1 + y^2}.$$

Moving the y -dependent factor to the left gives

$$(1 + y^2) dy = t dt.$$

After integrating, we obtain

$$y + \frac{y^3}{3} = \frac{t^2}{2} + C.$$

This relation specifies the solution even though we have not isolated y . We could use the cubic formula to do so, but the result would be much harder to read and would tell us very little. An equation relating t and y in this way is called an *implicit solution*, and it is every bit as exact as an explicit formula $y = f(t)$. If we want a numerical value at a particular time, we can solve the resulting cubic numerically.

Definition 2.3 (Implicit and explicit solutions). An **explicit solution** gives the unknown directly as $y = f(t)$. An **implicit solution** gives a relation

$$G(t, y) = C$$

whose differentiable branches $y(t)$ satisfy the differential equation. The second form is no less exact merely because y has not been isolated.

Before leaving separation, let us return to cooling. Newton's law said that the rate of change of temperature is proportional to the difference between the object's temperature and its surroundings, with $k > 0$:

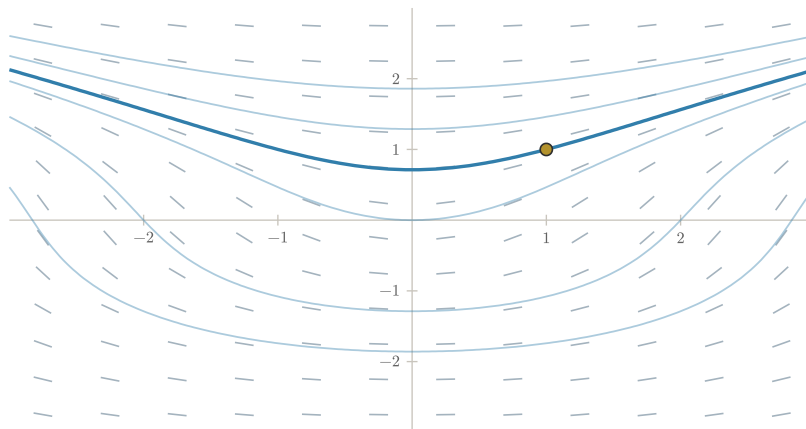


Figure 2.3 An implicit solution is still a curve we can draw. Each blue curve consists of the points satisfying $y + y^3/3 = t^2/2 + C$, even though we never isolate y . Drag the gold point to choose an initial condition; the readout determines C and draws the one implicit solution through that point.

$$T' = -k(T - T_{\text{room}}).$$

The repeated expression $T - T_{\text{room}}$ is the clue. Let $u = T - T_{\text{room}}$ be that temperature difference. Since the room temperature is constant, $u' = T'$, and the equation becomes

$$u' = -ku.$$

We can find the constant solution $u = 0$ first, and then separate the nonzero solutions:

$$\frac{du}{u} = -k dt.$$

Thus $\ln |u| = -kt + C$. If $T(0) = T_0$, then $u(0) = T_0 - T_{\text{room}}$, so $u = (T_0 - T_{\text{room}})e^{-kt}$. Returning to temperature gives

$$T(t) = T_{\text{room}} + (T_0 - T_{\text{room}})e^{-kt}.$$

Now the qualitative story we told from the local rule is visible in a single formula. The exponential factor tends toward zero, so a hot object cools toward the room tem-

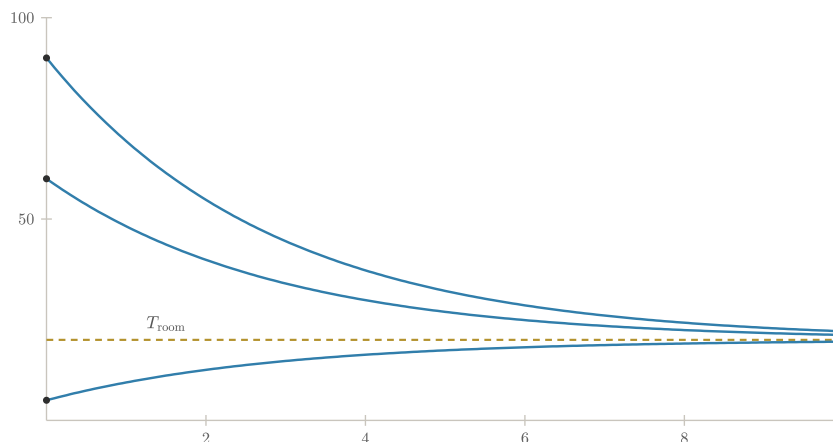


Figure 2.4 Cooling, solved. Three starting temperatures ride the same formula $T(t) = T_{\text{room}} + (T_0 - T_{\text{room}})e^{-kt}$ — hot mugs cooling down and a cold drink warming up are one calculation. In the live figure, drag the room-temperature line itself: every solution chases the new room temperature.

perature while a cold object warms toward it. The same equation, and the same formula, contains both stories.

2.4 Undoing the Product Rule

Separation was really the chain rule used in reverse. Can we play the same game with another familiar rule? Sometimes an equation contains the pieces of a product rule, but the product they came from has been broken apart. Consider

$$(1 + t^2)y' + 2ty = 1.$$

The two terms on the left look as though they came from differentiating a product. Indeed,

$$\frac{d}{dt}((1 + t^2)y) = (1 + t^2)y' + 2ty.$$

So we can group the entire left-hand side into one derivative:

$$\frac{d}{dt}((1 + t^2)y) = 1.$$

Now antidifferentiation gives

$$(1 + t^2)y = t + C, \quad \text{and hence} \quad y = \frac{t + C}{1 + t^2}.$$

Nothing new had to be invented here. We simply recognized the remains of a product rule and put the product back together.

Often the needed product is not visible until we create it. For example, look at

$$y' + 2ty = 2t.$$

The left side is not currently the derivative of a product. Suppose we multiply the whole equation by some function $m(t)$:

$$my' + 2tmy = 2tm.$$

If the left side is to equal $(my)' = my' + m'y$, then we need the coefficient of y to satisfy

$$m' = 2tm.$$

We recognize one function with exactly that derivative relationship: $m(t) = e^{t^2}$. Multiplying by it turns the equation into

$$e^{t^2}y' + 2te^{t^2}y = 2te^{t^2},$$

or, after putting the product back together,

$$\frac{d}{dt}(e^{t^2}y) = 2te^{t^2}.$$

Both sides can now be antiderivated:

$$e^{t^2}y = e^{t^2} + C, \quad \text{so} \quad y = 1 + Ce^{-t^2}.$$

The multiplying function e^{t^2} is called an *integrating factor*. But the name is not the point; the point is the question which produced it: what could we multiply by so that the product rule gathers several terms into one derivative?

We managed to invent the right multiplier for this equation, but was that good luck? In the next chapter we will make this construction systematic for a whole class of equations, and use it to prove that their initial value problems have exactly one solution.

First, though, we need another way to find answers. Our calculus techniques depend on recognizing a useful structure in the equation. What can we do when we do not find one? The slope field still gives us a direction at every point. Perhaps we can turn the act of following those directions into a calculation.

2.5 Euler's Method

Let us try this on the initial value problem

$$\begin{cases} y' = y(1 - y) + 0.1t, \\ y(0) = 0.1. \end{cases}$$

This looks suspiciously like the sort of equation we should be able to solve for homework. But the $0.1t$ prevents us from separating the variables, and there is no product rule on the left waiting to be put back together. None of the calculus techniques we have developed gives us a formula.

Still, the differential equation answers one small question perfectly. At the initial point $(0, 0.1)$, it tells us the slope:

$$F(0, 0.1) = 0.1(1 - 0.1) + 0.1(0) = 0.09.$$

What can we do with one slope? The tangent-line approximation says that for a short time h , the change in y should be approximately

$$\text{change in } y \approx h(0.09).$$

If we take $h = 1$, this predicts a new value

$$y_1 = 0.1 + 1(0.09) = 0.19.$$

Of course, there is no reason to keep trusting the old slope once we arrive at this new point. So we do something much more sensible: ask the differential equation again! It gives us the slope at $(1, 0.19)$, we follow that slope for another short step, and then we ask again. The equation supplies the local rule; our computation repeatedly checks in with it as the state changes.

In general, if our current approximation is (t_n, y_n) , this repeated conversation becomes Euler's method.

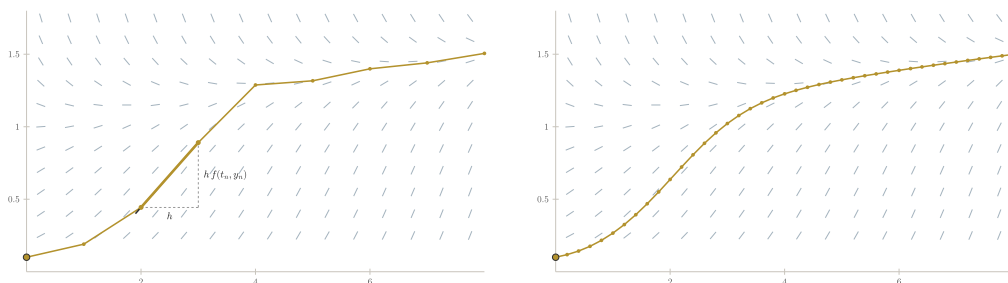


Figure 2.5 Euler's method turns local slopes into a polygonal history. Each segment borrows the slope at its left endpoint: h units across and $hF(t_n, y_n)$ units up. Smaller steps ask the equation for new information more often and bend through the field more faithfully. In the live figure, click a step to inspect it, change h , play the construction, or drag the initial point.

Definition 2.4 (Euler's method). For the initial value problem $y' = F(t, y)$, $y(t_0) = y_0$, choose a step size h . Starting from (t_0, y_0) , **Euler's method** produces successive approximations by

$$\begin{aligned} t_{n+1} &= t_n + h, \\ y_{n+1} &= y_n + hF(t_n, y_n). \end{aligned}$$

There is nothing more hidden in the method: it is the familiar tangent-line approximation, restarted over and over. Continuing our $h = 1$ calculation by hand gives

n	t_n	y_n	$F(t_n, y_n)$	y_{n+1}
0	0	0.100	0.090	0.190
1	1	0.190	0.254	0.444
2	2	0.444	0.447	0.891

There is something important hiding in this little table. The values y_n are not points sampled from a solution we secretly know. They are the approximations we have produced so far, and each new slope is evaluated at one of those approximate points. The computation builds the history and uses the history it has built to decide what to do next.

The step size h answers a practical question: how long are we willing to trust one tangent line? With $h = 1$, we travel fairly far before checking the rule again. With $h = 0.2$, we ask five times as often. This costs five times as many slope calculations, but the smaller-step polygons in the figure appear to settle toward a common curve. That is good evidence that we are finding the solution—though it is not yet a proof.

The method is quite easy to translate into code. The following Python loop says exactly what we just said: ask for a slope, use it to update the value, update the time, and repeat.

```
def euler(F, t0, y0, h, steps):
    t = t0
    y = y0
    points = [(t, y)]

    for _ in range(steps):
        y = y + h * F(t, y)
        t = t + h
        points.append((t, y))

    return points

def F(t, y):
    return y * (1 - y) + 0.1 * t

points = euler(F, 0, 0.1, 0.5, 16)
```

The list `points` is the polygon drawn in the figure. There is no hidden solver inside the code. This tiny loop *is* the solver.

Earlier, a Riemann sum gave us a way to approximate the accumulated change when the rate was a known function of time. Euler's method has extended that idea to a rate which depends on the unknown value as well. To see this, add up the first n updates:

$$y_n = y_0 + \sum_{k=0}^{n-1} hF(t_k, y_k).$$

If $F(t, y) = g(t)$ does not depend on y , this is an ordinary left Riemann sum for $\int g(t) dt$: all the rates can be known before the sum begins. In general, each rate $F(t_k, y_k)$ depends on the value the sum has reached so far. Euler's method is a Riemann sum which learns what to add next from its own running total.

How much should we trust the result? We can test the method on equations we already know how to solve, and we can repeat a computation with smaller and smaller values of h to see whether the answers agree. Both are useful evidence. Neither, by

itself, guarantees that an arbitrary computation is accurate. Building that guarantee requires a mathematical theory of error and stability, which we will develop when we study how to trust a computed trajectory. For now, Euler's method gives us the central idea on which much more powerful numerical solvers are built: follow the local rule for a short time, then ask again.

We can now find solutions in several ways. A remembered function may give us a successful guess; a calculus rule may reconstruct an exact answer; a sequence of short steps may give us a useful numerical approximation. But these methods also leave us with questions.

If we cannot find a formula, is there nevertheless a solution to find? If we have checked one, could there be another with the same initial condition? And how far in time does the solution continue? Euler's update lets us ask for another step, but that alone does not tell us whether an exact solution exists at the time we are trying to reach.

These are questions about the differential equation itself, whether or not we can solve it explicitly. In the next chapter we will begin answering them.

Further Reading

- Thomas W. Judson's open textbook, *The ODE Project*, Sections 1.2–1.4, develops separable equations, direction fields and phase lines, and numerical approximation through examples and activities.
- Jiří Lebl's *Notes on Diffy Qs*, Chapter 1 gives a concise second pass through first-order equations, including implicit solutions, autonomous equations, and Euler's method.
- MIT OpenCourseWare's 18.03SC Unit I offers another route through the same analytic, geometric, and numerical viewpoints, with free notes, videos, practice problems, and solutions.

PRACTICE • EXPLORATION • COMPUTATION

Problems

The routine problems ask you to guess and check, use calculus to find solutions, and step numerically with Euler's method. In the Python problems, begin with the supplied structure and make the computation your own.

CHECK YOUR UNDERSTANDING

1. Exponentials turn calculus into algebra

For each equation, try $y = e^{rt}$, determine every value of r which works, and verify the resulting functions by substitution.

(a) $y'' - y' - 6y = 0$.

(b) $y''' + y'' - 4y' - 4y = 0$.

(c) The three exponential solutions you found for the third-order equation all satisfy $y(0) = 1$. Calculate their values of $y'(0)$ and $y''(0)$. Does specifying only $y(0)$ distinguish these solutions?

2. A spring with a different clock

Consider $x'' = -9x$.

(a) Motivate a trigonometric guess, then verify that $x(t) = A \cos(3t) + B \sin(3t)$ solves the equation for every A and B .

(b) Choose A and B so that $x(0) = 2$ and $x'(0) = -6$.

(c) What is the period of this motion? Check your answer directly from the formula rather than quoting a memorized spring formula.

3. Accumulating a known rate

Solve each initial value problem exactly. Give an integral-defined answer when an elementary antiderivative is not available.

(a) $y' = 3t^2 - 4t + 1, y(1) = 2$.

(b) $2y' + 6t = 4, y(-1) = 0$.

(c) $y' = e^{-t^2}, y(0) = 3$. Then explain what the Fundamental Theorem of Calculus lets you know about your answer even though you cannot simplify its integral.

4. Separate carefully

For each equation, find every constant solution before dividing, then find the remaining solutions and impose the initial condition.

(a) $y' = t(1 - y), y(0) = 3$.

(b) $y' = y(2 - y), y(0) = 1$.

(c) In the previous part, identify the step which would have discarded two solutions if you had performed it too early.

5. An implicit solution is still a solution

Separate $y' = t/(1 + y^4)$ with $y(0) = 1$. Leave your answer as an implicit relation. Differentiate that relation with respect to t to verify that its branches satisfy the original differential equation.

6. Put the product back together

Solve each equation by recognizing or creating a product derivative.

- (a) $(1 + t)y' + y = t^2$ on an interval which does not contain $t = -1$.
- (b) $y' + 3t^2y = 3t^2$, using the integrating factor e^{t^3} .
- (c) Check the family you found in the previous part by substitution. Which member is the constant solution?

7. Euler by hand

Apply Euler's method to $y' = t - y$, $y(0) = 1$.

- (a) With $h = 0.5$, build a table through $t = 2$ containing t_n , y_n , the sampled slope, and y_{n+1} .
- (b) The exact solution is $y = t - 1 + 2e^{-t}$. Compare your approximation at $t = 2$ with the exact value.
- (c) Without recomputing, point to the entries of your table which show that Euler's method used its own approximate history to choose later slopes.

EXPLORATIONS

8. One equation, two methods

Consider

$$y' + 2ty = 2t.$$

- (a) Rewrite the equation in separable form. First find any constant solutions. Then separate variables on intervals where the factor you divide by is nonzero, and solve.
- (b) Solve the same equation by multiplying by e^{t^2} and recognizing a product derivative.
- (c) Show that your two answers describe the same family, even if their constants look different. Which solution needed separate attention during separation? How does it appear in the product-rule answer?
- (d) Use either answer to find a solution satisfying $y(0) = 3$, and check it directly.

9. Choosing a branch

Consider the equation

$$y' = \frac{t}{1 + y}$$

with two different initial conditions:

$$y(2) = 0 \quad \text{and} \quad y(2) = -2.$$

(a) Separate and integrate for each initial condition. Show that both lead to the same implicit relation

$$(y + 1)^2 = t^2 - 3.$$

(b) Solve this relation for y . Which branch satisfies each initial condition? Differentiate both formulas to check the differential equation.

(c) For each answer, find the largest open interval containing $t = 2$ on which the formula satisfies the original equation. What happens to y as t approaches the left endpoint? Why can't that endpoint be included, even though the formula has a finite value there?

(d) The formulas also give real values when $t < -\sqrt{3}$. Explain why those values do not extend either solution through the intervening gap.

PYTHON: FROM A LOOP TO EVIDENCE

10. Write Euler's method once

Complete the two missing update lines in this function.

```
def euler(F, t0, y0, h, steps):
    t = t0
    y = y0
    ts = [t]
    ys = [y]

    for _ in range(steps):
        # update y using the slope at the old point
        # update t
        ts.append(t)
        ys.append(y)

    return ts, ys
```

(a) Test your function on $y' = y$, $y(0) = 1$ through $t = 2$, using step sizes $h = 0.5, 0.25$, and 0.125 . Plot all three polygons with e^t on the same axes.

(b) For each step size, compute the absolute error at $t = 2$. Put the results in a small table. What happens to the error when h is halved?

(c) Change the right-hand side to $F(t, y) = y(1 - y) + 0.1t$. Which lines of your solver change, and which do not? Explain why that separation is useful.

11. Plot an implicit family

The separated equation in figure 2.3 produced the family $y + y^3/3 = t^2/2 + C$.

Use `numpy.linspace` to make grids of t and y , evaluate

$$G = Y + Y^{**3} / 3 - T^{**2} / 2$$

on the grid, and use `plt.contour(T, Y, G, levels=[...])` to draw at least five values of C . Mark one initial point, compute its value of C , and highlight the corresponding contour. Explain why a contour plot can draw the exact family without ever solving explicitly for y .

A Glimpse of the Theory

There is an old joke that a mathematician will prove a problem has a solution, then stop and leave finding it to the engineers. There is some truth to this! Knowing that a solution exists can be the beginning of a mathematical investigation, even if we never write it down.

Suppose we are studying a differential equation and want to say, “Take the solution $y(t)$, and now let us see what we can learn about it.” Perhaps we want to know whether it stays positive, whether it can pass another solution, or whether it must keep increasing. We might be able to answer these questions directly from the equation. But are we entitled to begin with “take the solution”? Is there a function there to investigate? And does calling it *the* solution make sense?

We might not even know the exact equation. A model could contain parameters we have not measured precisely, or a function whose general properties we understand without knowing its formula. Can we make statements which hold for every equation consistent with that information?

These are questions for a *general theory* of differential equations. We want results whose hypotheses describe what we need to know about an equation, and whose conclusions tell us something about its solutions. Such results let us reason about functions we have not found, and about whole classes of equations at once.

We will begin by establishing *existence*: that there is a solution to investigate. We will also ask about *uniqueness*: whether two solutions satisfying the same equation and initial condition must agree. For one important class of equations, the calculus from the last chapter is already enough to prove both.

3.1 Existence and Uniqueness for Linear Equations

We have already proved our first existence-and-uniqueness result, although we were thinking about antidifferentiation when we did it. Suppose f is continuous on an open interval I containing t_0 . Then

$$y' = f(t), \quad y(t_0) = y_0$$

has the solution

$$y(t) = y_0 + \int_{t_0}^t f(s) \, ds.$$

The Fundamental Theorem of Calculus verifies that its derivative is $f(t)$, and setting $t = t_0$ gives the required initial value. We have produced a solution, so we have proved *existence*.

But we also proved *uniqueness*. Suppose someone proposes another solution. Whatever function they give us, the Fundamental Theorem of Calculus requires it to satisfy

$$y(t) - y_0 = \int_{t_0}^t y'(s) \, ds = \int_{t_0}^t f(s) \, ds.$$

That forces exactly the formula above. There is no room for a different answer!

Notice how little we needed to know about f . We never chose a particular function, and we never evaluated an integral. Continuity was enough to make the construction work. Moreover, it works throughout I : for every t in that interval, we can integrate along the segment joining t_0 to t .

Now let us change the equation a little:

$$y' = f(t)y, \quad y(t_0) = y_0.$$

The rate now depends on the unknown value as well as on time, but separation gives us a way forward. On an interval where $y \neq 0$,

$$\frac{dy}{y} = f(t) \, dt.$$

To keep the notation manageable, write

$$A(t) = \int_{t_0}^t f(s) ds.$$

Thus $A' = f$ and $A(t_0) = 0$. Integrating the separated equation gives

$$\ln |y| = A(t) + C,$$

and exponentiating produces a family of the form $y = Ce^{A(t)}$, with a renamed constant. Allowing $C = 0$ includes the constant solution we excluded when dividing. The initial condition selects $C = y_0$, suggesting

$$y(t) = y_0 e^{A(t)}.$$

Let us check it:

$$y'(t) = y_0 e^{A(t)} A'(t) = f(t)y(t), \quad y(t_0) = y_0 e^0 = y_0.$$

The formula works for every initial value, including zero, and it is defined throughout I . We have proved existence again.

Have we also proved uniqueness? Our separation calculation divided by y , so we need to be careful about what happens at zero. It is easy to verify that the zero function solves the equation. What we need to rule out is some other function which satisfies the same initial condition but behaves differently afterward.

The formula we found suggests how to proceed. If $y = Ce^{A(t)}$, then multiplying by $e^{-A(t)}$ leaves a constant. Could the equation itself force that product to be constant?

Take any solution of $y' = f(t)y$. By the product rule,

$$\begin{aligned} \frac{d}{dt} (e^{-A(t)} y(t)) &= e^{-A(t)} y'(t) - A'(t) e^{-A(t)} y(t) \\ &= e^{-A(t)} (y'(t) - f(t)y(t)) \\ &= 0. \end{aligned}$$

A function with derivative zero on an interval is constant, by the Mean Value Theorem. At the initial time, this constant is

$$e^{-A(t_0)} y(t_0) = y_0.$$

Therefore every solution with the prescribed initial value must satisfy

$$e^{-A(t)}y(t) = y_0,$$

which forces $y(t) = y_0 e^{A(t)}$.

This argument never divided by y , so it works even when $y_0 = 0$. We have proved uniqueness for every initial value.

The factor $e^{-A(t)}$ is an integrating factor, just like the one we constructed in the last chapter. Here it turned the equation into the assertion that a product has derivative zero. Let us see what happens if we change the equation once more:

$$y' = f(t)y + g(t).$$

An equation of this form is called a **first-order linear equation**. We now have a prescribed contribution $g(t)$ to the rate, in addition to the term proportional to y . The same product-rule calculation gives

$$\frac{d}{dt} \left(e^{-A(t)}y(t) \right) = e^{-A(t)}(y'(t) - f(t)y(t)) = e^{-A(t)}g(t).$$

The product is no longer constant, but its derivative is a known function of time. We are back to the first problem we solved: accumulating a prescribed rate.

This gives us a theorem for the entire class.

Theorem 3.1 (Existence and uniqueness for first-order linear equations). Let f and g be continuous on an open interval I , and let $t_0 \in I$. Then, for every initial value y_0 , the initial value problem

$$y' = f(t)y + g(t), \quad y(t_0) = y_0$$

has exactly one solution on all of I .

Proof. Keep

$$A(t) = \int_{t_0}^t f(s) ds.$$

The product derivative we just found tells us to integrate:

$$e^{-A(t)}y(t) - e^{-A(t_0)}y(t_0) = \int_{t_0}^t e^{-A(s)}g(s) ds.$$

Since $A(t_0) = 0$ and $y(t_0) = y_0$, we obtain the candidate

$$y(t) = e^{A(t)} \left(\gamma_0 + \int_{t_0}^t e^{-A(s)} g(s) ds \right).$$

Every part of this expression is defined for every $t \in I$. Continuity of f gives us A , and continuity of $e^{-A}g$ gives us the second integral. To verify the equation, differentiate using the product rule and the Fundamental Theorem of Calculus:

$$\begin{aligned} y'(t) &= f(t)e^{A(t)} \left(\gamma_0 + \int_{t_0}^t e^{-A(s)} g(s) ds \right) + e^{A(t)} e^{-A(t)} g(t) \\ &= f(t)y(t) + g(t). \end{aligned}$$

At $t = t_0$, the integral vanishes and the exponential equals 1, so $y(t_0) = \gamma_0$. This proves existence on all of I .

For uniqueness, suppose y_1 and y_2 both solve the initial value problem. We want to show that they agree, so let us study their difference:

$$u = y_2 - y_1.$$

Subtracting the equations gives

$$u' = (f(t)y_2 + g(t)) - (f(t)y_1 + g(t)) = f(t)u.$$

Their initial values agree, so $u(t_0) = 0$. But we have already proved that $u' = f(t)u$ with this initial value has exactly one solution: $u = 0$. Consequently $y_1 = y_2$, proving uniqueness. ■

This is a big deal. Whenever we have a first-order linear equation with continuous coefficients and an initial value, we are entitled to say “take the solution” and begin investigating it. The theorem guarantees that there is a function there, that the initial value selects it uniquely, and that it exists throughout the coefficient interval. If the coefficients are continuous for all real time, the solution exists for all real time too.

The result applies even when we have not specified the coefficients completely. We might know their signs, bounds on their values, or how they depend on a parameter. As long as they are continuous, existence and uniqueness are settled. Those other properties can then become the information we use to investigate the solution.

And uniqueness tells us what a successful guess has accomplished. If we find a function satisfying the equation and initial condition, it must agree with the solution constructed in the proof. As we vary γ_0 , the formula gives every solution on I , because every such solution has a value at t_0 .

Of course, the formula may be an inconvenient way to compute. It contains an integral inside an exponential, and then another integral involving that exponential. Euler's method gives us a different way to approximate values. But the formula has already done its theoretical work: we know that the solution exists and is unique without evaluating either integral.

We have another calculus technique available. Separation of variables also produced answers in the last chapter, including for nonlinear equations. Can we use it to establish a general result too? And will its construction work throughout the interval where the equation is defined?

3.2 Existence and Uniqueness for Separable Equations

In the last chapter, separation turned an equation of the form

$$y' = g(t)h(y)$$

into two integration problems. But the resulting relationship did not always give us an explicit formula for y . Sometimes there was more algebra to do, and sometimes we left the answer implicit. Can we still use that relationship to prove that a solution exists and is unique? And how far in time can we follow it?

Suppose g is continuous on an open interval of times I , and h is continuous on an open interval of possible values of y . Choose an initial time t_0 and an initial value y_0 in these intervals. For now, assume $h(y_0) \neq 0$, so we do not begin at one of the equilibrium solutions that division would discard.

Recall what separation tells us to try:

$$\int_{y_0}^{y(t)} \frac{du}{h(u)} = \int_{t_0}^t g(s) ds.$$

Each side integrates a known function; the unknown value $y(t)$ appears as the upper limit on the left. Let us name the two functions defined by these integrals:

$$H(y) = \int_{y_0}^y \frac{du}{h(u)}, \quad G(t) = \int_{t_0}^t g(s) ds.$$

The relationship becomes $H(y(t)) = G(t)$. If we can invert H , we can recover the unknown function:

$$y(t) = H^{-1}(G(t)).$$

Does H have an inverse? Since b is continuous and $b(y_0) \neq 0$, there is a small open interval J around y_0 on which b never vanishes. It cannot change sign there without passing through zero. The Fundamental Theorem of Calculus gives

$$H'(y) = \frac{1}{b(y)},$$

so H' also has a constant, nonzero sign on J . Thus H is strictly monotone and has an inverse on its range. The inverse-function rule from calculus tells us that this inverse is differentiable.

We do not have to find a formula for the inverse to know that it exists! Monotonicity has established that each value in the range of H comes from exactly one value of y .

There is one more thing to check: does $G(t)$ belong to that range? At the initial time,

$$G(t_0) = 0 = H(y_0).$$

Since y_0 is inside J , the range of H contains an open interval around zero. Continuity of G therefore guarantees that $G(t)$ stays in that range for all t sufficiently close to t_0 . Our proposed solution is defined, at least nearby.

Now differentiate it. Using the derivative of an inverse function and the Fundamental Theorem of Calculus,

$$\begin{aligned} y'(t) &= (H^{-1})'(G(t))G'(t) \\ &= \frac{G'(t)}{H'(H^{-1}(G(t)))} \\ &= \frac{g(t)}{1/b(y(t))} \\ &= g(t)b(y(t)). \end{aligned}$$

Also $y(t_0) = H^{-1}(0) = y_0$. This proves existence.

For uniqueness, suppose we have any solution with the same initial condition. By continuity, its values remain in J for a short interval around t_0 . On that interval, the chain rule gives

$$\frac{d}{dt}H(y(t)) = \frac{y'(t)}{b(y(t))} = g(t).$$

Integrating from t_0 to t , and using $H(y_0) = 0$, forces $H(y(t)) = G(t)$. Applying the inverse forces exactly the function we constructed. There is only one solution nearby.

Notice what we have accomplished without evaluating either integral or finding a formula for an inverse. Calculus tells us that these functions exist, and their composition gives the solution. Continuity of g and b , together with $b(y_0) \neq 0$, was enough for the whole argument.

But we wanted to know how far the solution goes. We chose J small only to make sure division was allowed; nothing requires us to stop there. Now take J to be the largest open interval containing y_0 , within the allowed values of y , on which b never vanishes. Define H on all of this interval. It is still strictly monotone, so its inverse works on the whole range $H(J)$.

Our formula gives a solution whenever

$$G(t) \in H(J).$$

Starting at t_0 , follow this condition in both time directions. Let I_* be the largest open interval containing t_0 , inside I , on which it holds. Then

$$y(t) = H^{-1}(G(t)), \quad t \in I_*.$$

This is how far the inverse works. The same differentiation verifies the solution throughout I_* , and the same separation argument forces this answer for any solution through our initial point whose values stay in J . We may not be able to calculate the endpoints explicitly, but the range condition describes the interval using known functions.

The requirement of one interval matters. If G leaves the range of H and returns later, those later values do not continue our solution through the gap.

Have we found the whole solution interval? Division leaves a question unanswered: could our solution reach an equilibrium and continue beyond it? And what if the initial value itself is an equilibrium? Our argument has not settled either question.

If we assume a little more about b , we can finish the analysis. Suppose b' exists and is continuous. Then an equilibrium initial value gives exactly one solution: the constant solution. A solution starting away from an equilibrium cannot reach it or approach it at a finite time inside the given time interval. The interval found by our con-

struction is its maximal solution interval. We prove these remaining claims in problem 12 and problem 13.

Theorem 3.2 (Existence and uniqueness for separable equations). Suppose g is continuous on an open interval of times and h has a continuous derivative on an open interval of values of y . Choose an initial time t_0 and an initial value y_0 in these intervals. Then

$$y' = g(t)h(y), \quad y(t_0) = y_0$$

has a unique solution on a largest open interval containing t_0 .

If $h(y_0) = 0$, the solution is $y(t) = y_0$ throughout the given time interval. If $h(y_0) \neq 0$, it is the solution $H^{-1}(G(t))$ on the interval I_* constructed above.

Here *largest*, or *maximal*, means that we continue the solution as far as possible in both directions, keeping t and $y(t)$ in the intervals where our assumptions hold. There is no extension to a larger time interval that still solves the equation with those restrictions.

Let us see the interval construction at work in a concrete example.

Example 3.3 (A solution with a finite lifespan). Consider the initial value problem

$$\begin{cases} y' = y^2, \\ y(0) = 1. \end{cases}$$

Here $g(t) = 1$ and $h(y) = y^2$. The only zero of h is 0, so the largest interval containing our initial value 1 on which we can divide by h is $J = (0, \infty)$. The two integrals from our construction are

$$H(y) = \int_1^y \frac{du}{u^2} = 1 - \frac{1}{y}, \quad G(t) = \int_0^t 1 \, ds = t.$$

As y runs from 0 to ∞ , $H(y)$ increases from $-\infty$ toward 1, never reaching 1. Thus its range is $(-\infty, 1)$. The inverse works precisely when $G(t) = t < 1$! Solving $H(y) = G(t)$ gives

$$1 - \frac{1}{y} = t, \quad y(t) = \frac{1}{1-t}, \quad t \in (-\infty, 1).$$

We can check the formula directly:

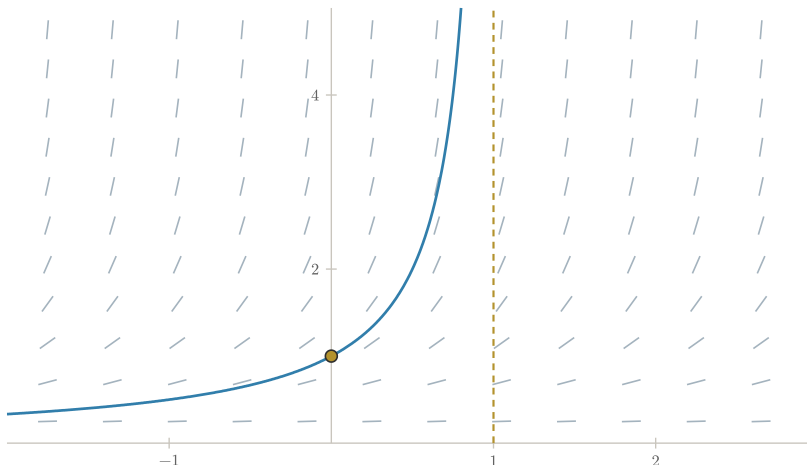


Figure 3.1 A formula reveals a lifespan. The solution of $y' = y^2, y(0) = y_0$ is $y = y_0/(1 - y_0 t)$. For $y_0 > 0$, its maximal interval through $t = 0$ is $(-\infty, 1/y_0)$; the dashed line marks the finite right endpoint. The slope field is smooth at every point of the plane; the finite lifespan does not come from a singularity in the slope rule. Drag the starting point: bigger starts die sooner. Beyond the asymptote the formula still has values, but the solution through $(0, y_0)$ does not.

$$y'(t) = \frac{1}{(1-t)^2} = y(t)^2, \quad y(0) = 1.$$

As t approaches 1 from below, $y(t)$ grows without bound. The interval $(-\infty, 1)$ is maximal: no finite value at $t = 1$ could continue this history. The algebraic expression still has values for $t > 1$, but those values do not extend the solution through the pole. Our range condition has found both the solution and its whole interval.

For linear equations, our construction worked throughout the coefficient interval. For separable nonlinear equations, we first worked locally, then followed the inverse as far as it would go. Even when the equation is defined for every real time and every real value of y , as in $y' = y^2$, the solution may exist on a smaller time interval.

The extra hypothesis also prepares our next step. For a separable equation the local rule is $F(t, y) = g(t)b(y)$. Holding t fixed and differentiating with respect to y gives

$$\frac{\partial F}{\partial y}(t, y) = g(t)b'(y).$$

Our stronger assumptions make both F and this derivative continuous. The next theorem uses these same conditions to guarantee existence and uniqueness even when separation and the product rule give us no formula.

3.3 General Existence and Uniqueness

Our proofs so far used the special form of the equation to construct a solution. For a general equation $y' = F(t, y)$, we may have neither a product rule to undo nor variables we can separate. But the conditions we just encountered still give us an existence-and-uniqueness theorem.

We will write F_y for the partial derivative $\partial F / \partial y$, computed by holding t fixed and differentiating with respect to y . For a linear equation, $F(t, y) = f(t)y + g(t)$, this is simply $F_y(t, y) = f(t)$. For a separable equation, we just found $F_y(t, y) = g(t)h'(y)$. In both cases, our assumptions made F and F_y continuous. The following theorem applies whenever these two continuity conditions hold near the initial point, regardless of whether the equation has either special form.

Theorem 3.4 (Local existence and uniqueness). Suppose that $F(t, y)$ and its partial derivative $\frac{\partial F}{\partial y}(t, y)$ are continuous in a neighborhood of (t_0, y_0) . Then the initial value problem

$$\begin{cases} y' = F(t, y), \\ y(t_0) = y_0 \end{cases}$$

has exactly one solution on some open interval containing t_0 .

The hypotheses are also a particularly understandable sufficient test, not the most general version of the theorem. Continuity of F keeps the slope rule from jumping abruptly and is enough to guarantee a local solution; continuity of $\partial F / \partial y$ controls how quickly the rule can change as the state y changes and guarantees uniqueness.

We will leave the proof of general existence out. But we can sketch why uniqueness holds using the linear theory we have already developed.

Proof sketch of uniqueness. The uniqueness argument has the same shape as the linear case. Suppose u and v are two solutions with the same initial value, and consider their difference $w = u - v$. Then

$$w' = F(t, u) - F(t, v).$$

Wherever $u \neq v$, write this as

$$w' = a(t)w, \quad a(t) = \frac{F(t, u(t)) - F(t, v(t))}{u(t) - v(t)}.$$

This quotient measures how much the prescribed rate changes when we change the value of the solution. When the two values coincide, we define $a(t) = F_y(t, u(t))$.

The mean value theorem and continuity of F_y ensure that this fills in the quotient continuously near the initial time.

So the difference satisfies a first-order linear equation! Since $w(t_0) = 0$, our linear uniqueness result forces $w = 0$: the two solutions agree nearby. Wherever the theorem's hypotheses continue to hold, repeating the argument extends their agreement throughout their common interval. If the interval of agreement had an endpoint inside the solutions' common domain, continuity would make them agree there too, and local uniqueness would extend their agreement past it.

3.4 Understanding the Guarantees

The phrase “some interval” is important: this is a *local* theorem. It promises that a unique history begins at our chosen point, but not that the history continues forever. We have already seen this with $y' = y^2$: the rule is smooth everywhere, but the solution through $y(0) = 1$ blows up at $t = 1$.

The hypotheses also deserve a careful reading. They are *sufficient*: when they hold, we get the promised conclusion. If one fails, the theorem is silent. We have to investigate the equation itself to find out whether existence or uniqueness actually fails.

Example 3.5 (A rate with a sudden switch). Consider the initial value problem

$$y' = \begin{cases} 0, & t < 0, \\ 1, & t \geq 0, \end{cases} \quad y(0) = 0.$$

The prescribed rate jumps at $t = 0$, so F is not continuous there. Can any differentiable function obey this rule on an open interval containing zero?

Before zero, its derivative would have to be zero, so the function would be constant. After zero, its derivative would have to be one, so it would have the form $t + C$. Continuity at zero and the initial condition force both constants to be zero. The only possible candidate is

$$y(t) = \begin{cases} 0, & t < 0, \\ t, & t \geq 0. \end{cases}$$

But this function has a corner at zero: its left derivative is zero and its right derivative is one. It is not differentiable there! Thus no solution exists on an open interval containing the initial time. A solution must satisfy the differential equation at every time in its interval, including the time when the rate switches.

What if F is continuous, but the derivative hypothesis fails? A local solution exists, but it need not be unique.

Example 3.6 (One point, two solutions). Consider the initial value problem

$$\begin{cases} y' = 3y^{2/3}, \\ y(0) = 0. \end{cases}$$

The rule $F(y) = 3y^{2/3}$ is continuous at $y = 0$, but its derivative $F'(y) = 2y^{-1/3}$ is not. And indeed, both

$$y(t) = 0 \quad \text{and} \quad y(t) = t^3$$

pass through the origin and satisfy the differential equation. The same initial point therefore leads to more than one possible history.

3.5 Qualitative Solutions

We began this chapter wanting to say “take the solution” and investigate it, even without knowing its formula. We now have conditions which let us do that. What can we learn?

In Chapter 1, we noticed that the differential equation prescribes just one slope at each point. Two solution curves therefore cannot meet with different tangents. But that leaves a possibility: could they meet with the same tangent and then separate? They might even cross while sharing a tangent at the crossing.

Uniqueness rules out all of these possibilities. Two solutions which differ anywhere on their common interval cannot even *touch*.

Why? Suppose two solutions meet at time t_* , with

$$u(t_*) = v(t_*) = a.$$

We can treat the meeting point as an initial condition. Both functions solve

$$y' = F(t, y), \quad y(t_*) = a.$$

Uniqueness forces them to agree near t_* , and the continuation argument from the preceding section extends that agreement throughout their common interval. This works in both time directions. The word “initial” does not require us to look only forward!

Corollary 3.7 (Solutions do not touch). Suppose F and F_y are continuous near every point of two solution curves of $y' = F(t, y)$. If the solutions agree at one time, they agree throughout the interval where both are defined.

This gives us a way to compare solutions without finding either one. Suppose

$$u(t_0) < v(t_0).$$

Could their order ever reverse? The difference $v - u$ is continuous, so to change from positive to negative it would have to pass through zero. But then the two solutions would meet, forcing them to agree at t_0 as well. That contradicts how they started. In fact, the same argument rules out even a moment of equality.

Thus

$$u(t_0) < v(t_0) \implies u(t) < v(t)$$

throughout their common interval. We call this **preservation of order**.

The solutions might both rise, both fall, or change direction. They might move closer together or farther apart. Whatever they do, the one which starts above stays above while both exist. The equation may depend explicitly on time; nothing in the argument required the special form $y' = f(y)$.

Now a single solution we can recognize becomes useful information about all the others. Its graph acts as a barrier: solutions beginning above it stay above it, and solutions beginning below it stay below it. We do not need formulas for those other solutions to know a region they can never enter.

Some of the easiest solutions to recognize are the ones which do not move. We already looked for these in Chapter 2, before dividing by a factor that might be zero. Now they can help us understand other solutions too.

For the constant function $y(t) = a$, the derivative is zero. Thus it solves $y' = F(t, y)$ precisely when

$$F(t, a) = 0$$

at every time in the interval we are considering. Checking this at just one time is not enough: the equation must keep prescribing zero slope along the entire horizontal line.

Definition 3.8 (Equilibrium solutions). A constant solution $y(t) = a$ of a differential equation is called an **equilibrium solution**, and its value a is called an **equilibrium**.

An equilibrium gives us a solution curve without any integration. But uniqueness makes that horizontal line much more useful: it becomes a barrier for every other solution.

Corollary 3.9 (Equilibria are barriers). Under the uniqueness hypotheses above, a solution which begins above an equilibrium remains strictly above it throughout their common interval of existence. A solution which begins below remains strictly below.

The proof is already contained in preservation of order: compare the solution with the constant solution $y(t) = a$. It cannot even touch the equilibrium unless it agrees with that equilibrium throughout their common interval.

Consequently, if we find two equilibria $a < b$, any solution starting between them stays between them while it exists. We may have no formula for that solution, but we already know a region it cannot leave!

Nothing here requires the equation to be autonomous. The rate may depend on time; what matters is that the constant functions really are solutions and that uniqueness applies.

Barriers tell us where a solution can go. What can we say about how it moves there? We already know a useful test from calculus: look at the sign of its derivative. The differential equation gives us that sign without requiring us to find the solution first!

Where $F(t, y) > 0$, the slope field points upward, so a solution is increasing while it remains in that region. Where $F(t, y) < 0$, it is decreasing. If we can use solution barriers to keep a history in a region where the sign is fixed, we know its direction of motion throughout its interval of existence.

Let us return to the equation whose slope field we drew in Chapter 1:

$$y' = t - y.$$

Below the diagonal $y = t$, we have $t - y > 0$, so solutions rise. Above it, we have $t - y < 0$, so solutions fall. Along the diagonal itself, the assigned slope is zero.

Does that make the diagonal a solution barrier? Check it! The function $y(t) = t$ has derivative 1, but the equation asks for $t - t = 0$. It is not a solution. A solution passing through this diagonal has a horizontal tangent there; it does not follow the diagonal. Uniqueness therefore gives us no reason to forbid crossing this line. Finding where F vanishes tells us where solutions have horizontal tangents, but those points need not themselves form a solution curve.

There is another nearby line which does solve the equation:

$$y(t) = t - 1.$$

Its derivative is 1, and substituting into the right-hand side gives $t - (t - 1) = 1$. Thus this line really is a barrier. A solution which starts below it stays below it, and one which starts above stays above, throughout their common interval. The linear theorem tells us that these solutions exist on the whole real line.

Now combine the two kinds of information. If $y(t_0) < t_0 - 1$, preservation of order gives

$$y(t) < t - 1$$

at every time. This keeps the solution below the diagonal $y = t$, where the slopes are positive. So the solution is increasing throughout its entire history. We have learned this by checking one particularly simple solution and reading the signs of the slope field, without finding a formula for the solution we are investigating.

Let us try these ideas on some equations where the calculations are particularly simple. For an autonomous equation $y' = f(y)$, the rule does not change with time. Such equations occur frequently in scientific models, and we will meet them throughout the book. Here, their simple form makes it easy to find solution barriers and read the signs of the slopes.

For $y' = f(y)$, finding an equilibrium means solving the algebraic equation

$$f(a) = 0.$$

There is no time variable to check: once $f(a) = 0$, the constant function $y(t) = a$ satisfies the differential equation at every time. Draw these equilibrium solutions as horizontal lines in the slope field. Under our uniqueness hypotheses, every other solution must stay on its own side of each line.

What happens between them? Suppose $a < b$ are two neighboring equilibria. There are no zeros of f between a and b , and continuity prevents f from changing sign without passing through zero. Thus either

$$f(y) > 0 \quad \text{throughout } a < y < b,$$

or $f(y) < 0$ throughout that interval. We can determine which by checking the sign at just one value between the roots.

Now our two general observations work together. A solution starting between a and b stays between them, and the sign tells us whether it keeps increasing or keeps decreasing. We can begin sketching its history without doing any integration!

The same sign reasoning works above the largest equilibrium or below the smallest, when those exist. There may be no second equilibrium to bound the motion, but the solution still cannot change direction while it stays in an interval where f has no zeros. We must keep its lifespan in mind, though: moving in one direction does not guarantee that a solution exists forever, as $y' = y^2$ already showed us.

Let us try this on

$$y' = 1 - y^2.$$

The equilibria occur where $1 - y^2 = 0$, so begin by drawing the horizontal solution lines $y = -1$ and $y = 1$.

Now choose an initial value between -1 and 1 . The solution cannot touch either equilibrium, so its entire history stays in this strip. And since $1 - y^2 > 0$ there, it keeps increasing. Try sketching a curve with these properties: it must follow the upward slopes while staying strictly below the horizontal line $y = 1$.

But there is still something to justify. Must this solution actually approach 1? Could it flatten out toward some smaller value instead? And do we know that its history continues for arbitrarily large times?

Our sketch suggests answers, but now it's up to us to try to pull them out of the mathematics.

The separable result settles the lifespan question. A solution increasing between -1 and 1 cannot grow without bound. Nor can its history end at a value inside the strip, where our inverse construction still works, or approach the equilibrium at 1 in finite time. So it continues for every future time.

Since the solution increases and stays below 1 , it must approach a limit L , with $-1 < y_0 \leq L \leq 1$. Could L be smaller than 1 ? Then $1 - L^2 > 0$, so as the solution approaches L , its rate approaches a positive number. Eventually it would gain at least some fixed amount per unit time, making it impossible to settle at L . Thus

$$\lim_{t \rightarrow \infty} y(t) = 1.$$

The same reasoning works between neighboring equilibria of any $y' = f(y)$ with continuously differentiable f : the solution approaches the equilibrium in its direction of motion. In problem 14, you will make this argument precise for general f .

What happens outside this strip? Above $y = 1$, the slopes are negative, so solutions decrease. Uniqueness keeps them above the equilibrium, and the same continuation and limit argument shows that they approach 1 . Thus solutions on both sides pile up toward the horizontal solution $y = 1$.

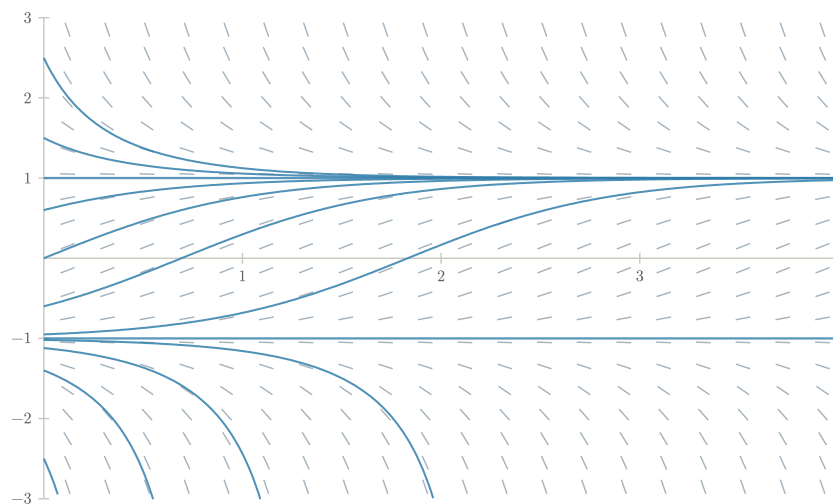


Figure 3.2 For $y' = 1 - y^2$, draw the two equilibrium histories and compare the slopes in the three regions they separate. Solutions between -1 and 1 rise toward 1 , and those above 1 fall toward it. Below -1 , solutions fall without bound and leave the visible window before their finite lifespan ends. In the live figure, click the slope field to release a state and follow its solution.

Below $y = -1$, the slopes are also negative, but now decreasing carries solutions away from the equilibrium. There are no further equilibria below them. Could a solution nevertheless approach some finite lower value? Our previous argument rules this out: at any such value, the rate would still be negative, keeping the solution moving downward. These solutions therefore decrease without bound.

In fact, they reach $-\infty$ in finite time: the solution ends by becoming unbounded, just as our earlier $y' = y^2$ solution did by growing toward $+\infty$. You will prove this by separation in problem 11.

Solutions on both sides of 1 approach it as time increases; we call this equilibrium **attracting**. Near -1 , solutions on either side move away; we call it **repelling**.

Must an equilibrium attract from both sides or repel from both sides? Compare

$$y' = -y \quad \text{and} \quad y' = y^2.$$

Both have their only equilibrium at zero. For the first equation, slopes are positive below zero and negative above it. Solutions on both sides move toward the equilibrium.

For $y' = y^2$, however, every slope away from zero is positive. Solutions starting below zero increase toward it, while those starting above zero increase away from it.

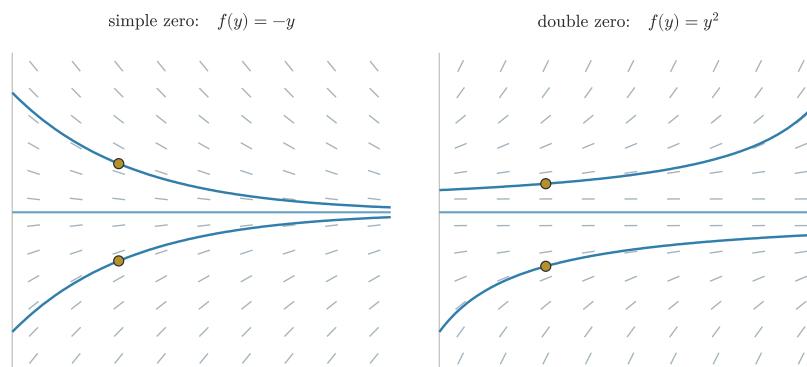


Figure 3.3 Across a simple zero, the sign of f changes, so the slopes tilt in opposite directions on the two sides of the equilibrium wall. Across this double zero, the sign stays positive, so the slopes tilt upward on both sides. Gold states flow toward the simple zero from both sides, but toward the double zero from below and away from it above.

The equilibrium attracts from one side and repels from the other. We call this behavior **semistable**.

The difference is visible in the algebra. In $f(y) = -y$, the factor y occurs once, so its sign changes across zero. In $f(y) = y^2$, the factor is squared, so the sign stays positive on both sides.

More generally, a polynomial has a **simple zero** at a when the factor $y - a$ occurs exactly once, and a **double zero** when it occurs exactly twice. Near a simple zero, the sign changes across the equilibrium; near a double zero, it does not. Checking those signs tells us how solutions move on each side.

We can recognize the same behavior in Newton's law of cooling, $T' = -k(T - T_{\text{room}})$, with $k > 0$. Above room temperature the rate is negative; below it the rate is positive. The equilibrium $T = T_{\text{room}}$ is attracting, and uniqueness prevents any other solution from reaching or crossing it. Thus hot objects cool toward room temperature and cold objects warm toward it.

An equation can also have more than one attracting equilibrium. For $y' = y - y^3 = y(1 - y)(1 + y)$, the equilibria are -1 , 0 , and 1 . Check the signs on either side of each: -1 and 1 are attracting, while 0 is repelling. Every positive initial value gives a solution approaching 1 , every negative initial value gives one approaching -1 , and an initial value of zero gives the constant solution. The initial condition therefore selects between two attracting outcomes, with the equilibrium at zero separating the initial values that lead to each.

In each of our autonomous examples, a nonconstant solution keeps moving in one direction. This holds even when there are no equilibria to trap it between. Could a solution of $y' = f(y)$ ever turn around?

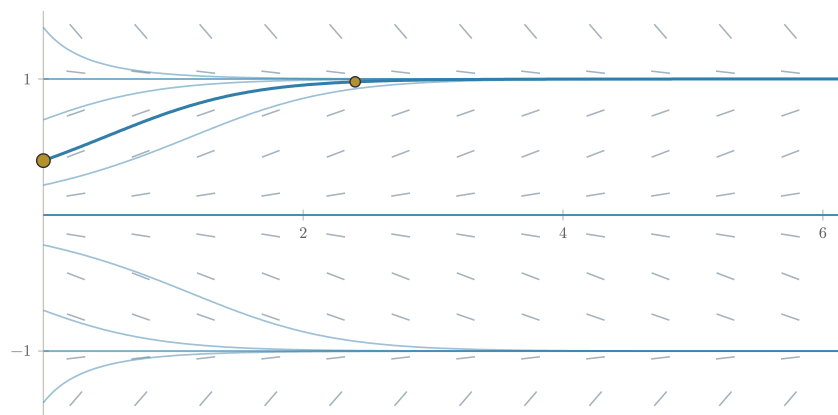


Figure 3.4 One rule, two possible destinations. For $y' = y - y^3$, the attracting equilibria at -1 and 1 collect the histories on their side of the repelling equilibrium at 0 . In the live figure, drag the gold initial state up and down. Its blue history changes immediately, and the moving gold point follows that history toward its equilibrium.

To change direction, its derivative would have to pass through zero. But if $y'(t_*) = 0$, then

$$f(y(t_*)) = 0.$$

The solution has reached an equilibrium! Uniqueness then forces it to agree with that constant solution throughout its interval. Thus a nonconstant solution can never have zero derivative. Since $y' = f(y)$ is continuous, its sign stays fixed.

Proposition 3.10 (Nonoscillation). Suppose f is continuously differentiable. Every nonconstant solution of $y' = f(y)$ is strictly increasing or strictly decreasing throughout its interval of existence. In particular, there are no nonconstant periodic solutions.

So if we want to model something that oscillates, like a spring, an equation of the form $y' = f(y)$ won't do. For the spring we used $x'' = -x$: allowing a second derivative makes oscillation possible.

For the autonomous equations we have just studied, the slope field repeats the same information across every horizontal line. The slope at a given height depends only on that height. If we want, we can compress this picture onto a single copy of the y -axis.

Mark the equilibria on this line. Between them, draw arrows pointing upward where $f(y) > 0$ and downward where $f(y) < 0$. This picture is called a **phase line**.

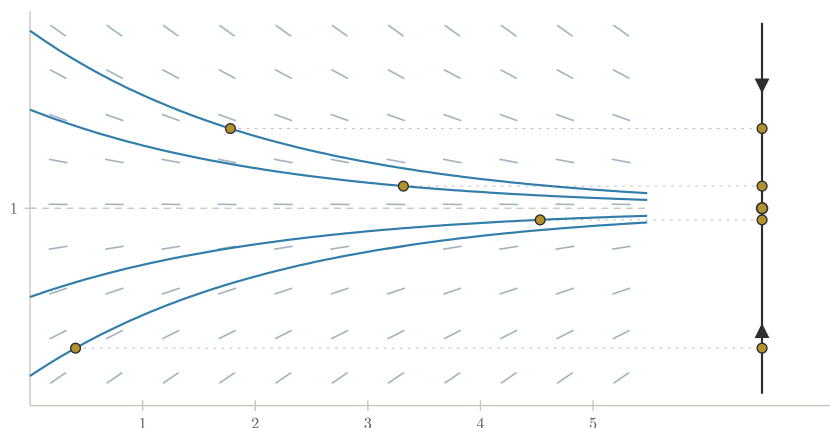


Figure 3.5 Each moving point on a cooling solution casts a horizontal shadow onto the phase line, keeping only its current temperature. Watch the point and its shadow move together toward the room-temperature value. The arrows on the line record the direction of that motion.

It records the equilibrium values and directions of motion that we already used to understand the solution curves.

Watch a point move along one of the cooling curves. Its horizontal shadow on the phase line keeps only its temperature. As the original point moves forward in time, its shadow moves toward room temperature. We can follow the same history as a curve in the slope field or as a moving point on the line.

The phase line is useful when we want to see these directions and equilibria at a glance. But the arrows alone do not tell us how long the motion takes. For $y' = 1 - y^2$, a downward arrow below -1 records decreasing motion; the separation calculation tells us that the solution becomes unbounded in finite time.

Compare the phase lines below. Can you identify which equilibria attract, which repel, and which attract from only one side? Check the arrows against the signs of each equation. The final example, $y' = 1$, has no equilibria at all: every state moves upward.

We began by asking whether we could say “take the solution” without knowing how to find it. Existence and uniqueness gave us conditions under which that makes sense. Then we put the solution to work: comparing it with other solutions, finding regions it cannot leave, and investigating where it goes as time passes. These arguments apply to whole classes of equations, even when the formulas change or disappear. In the next chapter, we will build models of our own, turning assumptions about how something works into a differential equation and investigating what it predicts.

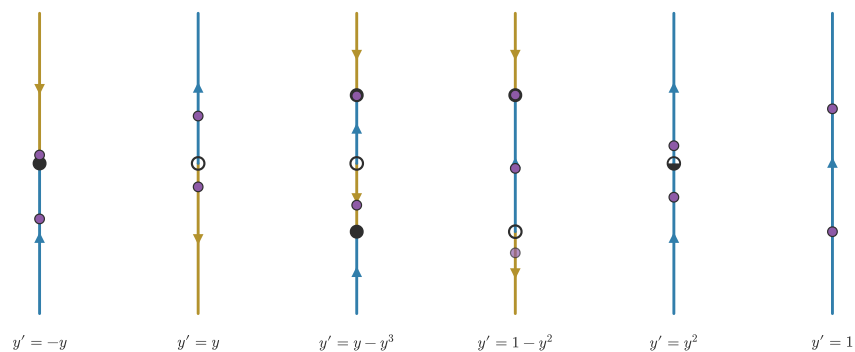


Figure 3.6 Six equations, each summarized on a phase line. The arrows point upward where the rate is positive and downward where it is negative. Filled dots mark attracting equilibria, open dots mark repelling equilibria, and the half-filled dot records attraction from below and repulsion from above. The violet points move along the lines as the states change with time.

Further Reading

- Thomas W. Judson's open textbook, *The ODE Project*, Sections 1.2–1.4, develops separable equations, direction fields and phase lines, and numerical approximation through examples and activities.
- Jiří Lebl's *Notes on Diffy Qs*, Chapter 1 gives a concise second pass through first-order equations, including implicit solutions, autonomous equations, and Euler's method.
- MIT OpenCourseWare's 18.03SC Unit I offers another route through the same analytic, geometric, and numerical viewpoints, with free notes, videos, practice problems, and solutions.

PRACTICE • EXPLORATION • COMPUTATION

Problems

CHECK YOUR UNDERSTANDING

1. What interval does the linear theorem guarantee?

Without finding solution formulas, use theorem 3.1 to answer the following.

(a) Consider

$$y' = (\sin t)y + e^{-t^2}, \quad y(0) = 2.$$

On what largest open interval containing the initial time does the theorem guarantee a unique solution? Identify the coefficient functions and check its hypotheses.

(b) Answer the same question for

$$y' = \frac{y}{1-t} + \sin t, \quad y(0) = 2.$$

How does your answer change if the initial condition is instead $y(2) = 2$?

(c) The general local existence-and-uniqueness theorem also applies near each of these initial points. What does the linear theorem tell us about the solution's interval that the local theorem alone does not?

2. Reading the theorem

For each initial value problem, decide whether theorem 3.4 guarantees a unique solution near the initial point. If not, say exactly which hypothesis fails.

(a) $y' = 4 + y^3$, with $y(0) = 1$.

(b) $y' = \sqrt{y}$, with $y(1) = 0$; then the same equation with $y(1) = 1$.

(c) $y' = \frac{t}{y-2}$, with $y(0) = 2$.

(d) $y' = |y|$, with $y(0) = 0$.

3. The fence

Suppose $F(t, y)$ and $\partial F / \partial y$ are continuous everywhere, and suppose that $F(t, 1) = 0$ for every t .

(a) Show that the constant function $y(t) = 1$ is a solution of $y' = F(t, y)$.

(b) Now let $y(t)$ be any solution with $y(0) = 0$. Show that $y(t) < 1$ throughout its interval of existence: the line $y = 1$ is a fence no other solution can touch.

4. A solution we know tells us about solutions we don't

In Chapter 1, we checked that $y(t) = 2t$ solves

$$y' = (y - 2t)g(t, y) + 2$$

without needing to know the function g . Now suppose g and its partial derivative g_y are continuous everywhere.

(a) Verify the known solution again. Check that the right-hand side of the differential equation satisfies the hypotheses of the general existence-and-uniqueness theorem.

(b) Let u and v be solutions with

$$u(0) = -1, \quad v(0) = 1.$$

Without finding either solution, show that

$$u(t) < 2t < v(t)$$

throughout the interval where both are defined. Explain why this holds for negative as well as positive times.

(c) Suppose $g(t, y) = -3$. Calculate $u'(0)$ and $v'(0)$. The lower solution is initially rising and the upper solution is initially falling. Could they eventually meet? Explain how your answer fits with part (b).

5. Read the phase line

Consider $y' = y(y - 2)(4 - y)$.

- (a) Find every equilibrium and draw the phase line.
- (b) Classify each equilibrium as attracting, repelling, or semistable.
- (c) Describe the long-term fate of solutions beginning at -1 , 1 , 3 , and 5 without solving the equation.
- (d) Explain why none of the nonconstant solutions can oscillate.

6. A periodic solution?

Suppose y solves $y' = y$ and satisfies $y(0) = y(1)$.

- (a) Use the family $y = Ce^t$ to determine every possibility.
- (b) Give a phase-line explanation of the same conclusion which does not begin from a formula.

EXPLORATIONS

7. Sufficient, but not necessary

In problem 2, we saw that theorem 3.4 is silent about $y' = |y|$, $y(0) = 0$.

Silence is not a verdict: the theorem's hypotheses are sufficient, not necessary. Let us find out what actually happens.

- (a) Suppose a solution is positive at some time. On any interval where $y > 0$ the equation reads $y' = y$. What are the solutions of $y' = y$, and can any of them ever equal zero? If such a positive interval had a finite endpoint inside the solution's domain, what would continuity require there?
- (b) Make the corresponding argument on intervals where $y < 0$, using $y' = -y$. Explain why a largest negative interval cannot have a finite endpoint inside the solution's domain.
- (c) Conclude that the only solution with $y(0) = 0$ is $y \equiv 0$: uniqueness holds here even though the theorem declined to promise it. Compare with theorem 3.6, where the same hypothesis failed and uniqueness really was lost.

8. Waiting before leaving

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

- (a) Find the nonzero solutions separately in the regions $y > 0$ and $y < 0$.
 (b) Verify that for every $a \geq 0$, the piecewise function

$$y_a(t) = \begin{cases} 0, & t \leq a, \\ (t - a)^2, & t > a \end{cases}$$

solves the differential equation and initial condition.

- (c) Where does the hypothesis of theorem 3.4 fail? Explain how the family above turns that failed guarantee into visible behavior.

9. Design a phase line

- (a) Construct a polynomial $f(y)$ for which $y' = f(y)$ has exactly three equilibria: attracting at $y = -2$ and $y = 3$, and repelling at $y = 0$. Draw the phase line to check your design.
 (b) Can an autonomous equation $y' = f(y)$ with continuously differentiable f have exactly three simple equilibria and have all three attracting? Give a sign argument.
 (c) Construct a continuously differentiable example with exactly three equilibria in which all three are attracting from at least one side. State which equilibria are semistable.

10. Confident nonsense past a blowup

Euler's update does not know that the solution of $y' = y^2$, $y(0) = 1$ becomes infinite at $t = 1$.

- (a) Use $h = 0.25$ to compute y_1 through y_6 . What finite value does the algorithm report at $t = 1.5$?
 (b) Compare that table with theorem 3.3. Why is continuing the recursion algebraically valid but mathematically meaningless as an approximation to the original history?
 (c) State a practical warning this example gives about trusting numerical output.

11. Decreasing without bound in finite time

Consider

$$y' = 1 - y^2, \quad y(t_0) = y_0 < -1.$$

The signs and equilibrium barriers tell us that the solution keeps decreasing below -1 . We will use separation to determine how much time passes before it becomes unbounded.

- (a) Separate variables and show that, while the solution exists,

$$t - t_0 = \int_{y_0}^{y(t)} \frac{du}{1 - u^2} = \int_{y(t)}^{y_0} \frac{du}{u^2 - 1}.$$

Explain why this elapsed time is positive when $t > t_0$, even though $y(t) < y_0$.

- (b) Use partial fractions to evaluate

$$\int_{-\infty}^{y_0} \frac{du}{u^2 - 1} = \frac{1}{2} \log \left(\frac{y_0 - 1}{y_0 + 1} \right).$$

Check that the result is finite and positive for every $y_0 < -1$.

(c) Use the range of the separated integral and the maximality conclusion of theorem 3.2 to show that the right endpoint of the maximal solution interval is

$$T = t_0 + \frac{1}{2} \log \left(\frac{y_0 - 1}{y_0 + 1} \right),$$

and that $y(t) \rightarrow -\infty$ as $t \rightarrow T$ from below. Explain why no solution can continue this history through T .

GUIDED PROOFS

12. Why equilibria cannot be reached or left

Suppose g is continuous on an open time interval I , and b has a continuous derivative on an open interval of allowed values of y . Consider

$$y' = g(t)b(y).$$

We will prove the equilibrium claims made in the separable section using calculus. Do not invoke the general existence-and-uniqueness theorem or its consequences about solution barriers: this exercise supplies an earlier proof for separable equations.

(a) Let J be a largest open interval of allowed values on which b never vanishes. Suppose a is a finite endpoint of J that is still inside the allowed value interval. Explain why $b(a) = 0$.

Use continuity of b' to bound $|b'|$ by some $M > 0$ near a . Apply the Mean Value Theorem to show that

$$|b(y)| \leq M|y - a|$$

for y sufficiently close to a .

(b) Fix $y_* \in J$ and define

$$H(y) = \int_{y_*}^y \frac{du}{b(u)}.$$

Use the preceding bound to compare $1/|b(y)|$ with $1/(M|y - a|)$. Show that integrating $1/|b(y)|$ toward a from inside J gives a divergent improper integral. Why does the constant sign of b on J imply that $|H(y)| \rightarrow \infty$ as $y \rightarrow a$ from that side?

(c) Suppose a solution takes values in J on a time interval containing s . By the chain rule and the Fundamental Theorem of Calculus,

$$H(y(t)) - H(y(s)) = \int_s^t g(r) dr.$$

If b is a finite endpoint of this time interval lying inside I , explain why the right side has a finite limit as $t \rightarrow b$. Use the previous part to rule out $y(t) \rightarrow a$ for any endpoint a of J inside the allowed value interval. Treat both left and right time endpoints. This also rules out approaching such an equilibrium at a finite endpoint where the solution has not yet been defined.

(d) Let a solution be defined on an open interval $K \subseteq I$, and suppose $b(y(s)) \neq 0$ at some $s \in K$. Take J to be the largest interval containing $y(s)$ on which b never vanishes. Show that $y(t)$ stays in J for every $t \in K$.

Hint: Take the largest time interval around s on which the solution stays in J . If it ended at a time b inside K , continuity would put $y(b)$ at an endpoint of J inside the allowed value interval. Apply the previous part. Thus a solution starting away from equilibrium cannot reach any equilibrium at a finite time in its domain.

(e) Now suppose $b(a) = 0$ and $y(t_0) = a$. Verify that $y(t) = a$ is a solution on all of I , and prove that every solution with this initial value is constant throughout its interval.

Hint: If $y(s) \neq y(t_0)$ at another time, the Mean Value Theorem gives an intermediate time r at which $y'(r) \neq 0$. The differential equation then implies $b(y(r)) \neq 0$. Apply the preceding part starting at r : could this same solution have the equilibrium value a at t_0 ? Your argument should work whether $s < t_0$ or $s > t_0$.

13. Following the inverse as far as it works

Keep the assumptions of problem 12, and suppose $b(y_0) \neq 0$. Recall the construction from the text: J is the largest open interval containing y_0 among the allowed values on which b never vanishes, and

$$H(y) = \int_{y_0}^y \frac{du}{b(u)}, \quad G(t) = \int_{t_0}^t g(s) ds.$$

Let I_* be the largest open interval containing t_0 within I on which $G(t) \in H(J)$. The text proved directly that $y(t) = H^{-1}(G(t))$ is a solution on I_* . Use that construction and problem 12 to prove that this interval is maximal.

(a) Suppose b is a finite endpoint of I_* inside I , and that $y(t) \rightarrow \ell \in J$ as $t \rightarrow b$ from within I_* . Show that $G(b) = H(\ell)$. Since $H(J)$ is open, continuity of G then puts $G(t)$ inside $H(J)$ for times on both sides of b . Explain why this would extend the inverse formula beyond b , contradicting the choice of I_* .

(b) Suppose there were a solution extending this history through a finite endpoint b of I_* inside I . Continuity would give a finite limit $\ell = y(b)$ inside the allowed value interval. Explain why ℓ must lie either in J or at an endpoint of J .

Rule out the first possibility using the preceding part and the second using problem 12. Notice that you only need a limit under the assumption that an extension exists; you do not need to prove that every solution has a limit at every endpoint.

(c) Explain why an endpoint of the prescribed time interval, escape from the allowed value interval, or unbounded growth cannot permit a continuation that keeps both time and the solution values in their prescribed intervals. Treat both time directions and conclude that I_* is the maximal solution interval.

Finally, use the separation identity and the fact that solutions with $b(y_0) \neq 0$ remain in J to show that any solution with the same initial data agrees with $H^{-1}(G(t))$ and is a restriction of this maximal solution. Together with equilibrium uniqueness from the previous exercise, this completes the claims of theorem 3.2.

14. Between neighboring equilibria

Suppose f is continuously differentiable on an open interval containing $[a, b]$, where $a < b$, $f(a) = f(b) = 0$, and f has no zeros between a and b . Let $y(t)$ be the maximal solution of

$$y' = f(y), \quad y(t_0) = y_0, \quad a < y_0 < b.$$

We will prove that the solution exists for every future time and approaches one of the two equilibria. You may use the separable result theorem 3.2, including its conclusions about maximality and the impossibility of reaching an equilibrium in finite time.

(a) Use preservation of order to show that $a < y(t) < b$ throughout the solution's interval of existence. Why does continuity of f then imply that the solution is either strictly increasing throughout this interval or strictly decreasing throughout it?

(b) Suppose the maximal interval has a finite right endpoint B . Explain why monotonicity and boundedness give a limit $L \in [a, b]$ as $t \rightarrow B$ from below. If $a < L < b$, use the inverse construction from the separable section to show that the solution extends through B . If $L = a$ or $L = b$, use the equilibrium nonarrival conclusion. Deduce that the solution exists for all $t \geq t_0$.

(c) Show that $y(t)$ has a limit $L \in [a, b]$ as $t \rightarrow \infty$. Suppose $f(L) \neq 0$. Use continuity to find a constant $c > 0$ and a time T such that, for all $t \geq T$, either $y'(t) \geq c$ or $y'(t) \leq -c$. Integrate this inequality from T to t to contradict boundedness. Conclude that $f(L) = 0$.

(d) Use the direction of motion and the initial value to show that

$$\lim_{t \rightarrow \infty} y(t) = \begin{cases} b, & \text{if } f > 0 \text{ on } (a, b), \\ a, & \text{if } f < 0 \text{ on } (a, b). \end{cases}$$

Explain why the proof does not require evaluating a separated integral or finding a formula for $y(t)$.

PYTHON: COMPARING SOLUTION HISTORIES

15. Many initial values, one phase line

Use the Euler function from Chapter 2 to compute solutions of $y' = y - y^3$ from initial values $-2, -1.2, -0.5, 0, 0.5, 1.2, 2$.

Use $h = 0.01$ through $t = 8$. Plot the histories on one set of axes and mark the three equilibrium lines. Then make a second plot showing only the final computed value against the initial value. Explain how the two plots reveal the same two basins separated by the repelling equilibrium.

ADVANCED EXPLORATIONS

16. Dependence on initial conditions

Uniqueness tells us what happens when two initial values agree exactly. What if they are only close? For linear equations, we can calculate exactly how their difference affects the solutions.

Let f and g be continuous on an open interval I containing t_0 . Suppose y and $\tilde{y}$ solve the same equation

$$y' = f(t)y + g(t),$$

with initial values $y(t_0) = y_0$ and $\tilde{y}(t_0) = \tilde{y}_0$. The linear theorem gives both solutions throughout I .

(a) Set $w = \tilde{y} - y$. Derive a differential equation and initial condition for w . Explain why the prescribed contribution $g(t)$ disappears when we compare the two solutions.

(b) Use an integrating factor to prove that

$$\tilde{y}(t) - y(t) = (\tilde{y}_0 - y_0)e^{\int_{t_0}^t f(s) ds}.$$

Hint: Set $A(t) = \int_{t_0}^t f(s) ds$ and differentiate $e^{-A(t)}w(t)$. Use the initial value to determine the resulting constant. This should also work when $w(t_0) = 0$.

Recover uniqueness and preservation of strict order from this formula. Why can the difference never change sign?

(c) Fix $T > 0$ with $[t_0, t_0 + T] \subset I$, and suppose $|f(t)| \leq M$ on this interval, where $M \geq 0$. Show that

$$|\tilde{y}(t) - y(t)| \leq |\tilde{y}_0 - y_0|e^{MT} \quad \text{for every } t \in [t_0, t_0 + T].$$

Hint: Bound $\int_{t_0}^t f(s) ds$ using $f(s) \leq |f(s)| \leq M$ and $t - t_0 \leq T$.

Given a tolerance $\varepsilon > 0$, show that requiring

$$|\tilde{y}_0 - y_0| < \varepsilon e^{-MT}$$

keeps the solutions within ε of each other throughout this time interval. Explain how this gives **continuous dependence on the initial value**: on any fixed finite interval, we can make the solutions as close as we want by choosing their initial values sufficiently close.

(d) Apply the difference formula to $y' = -y$ and to $y' = y$, with initial values prescribed at $t_0 = 0$. In each case, describe what happens to a nonzero initial difference as time increases.

For which equation does every sufficiently small initial difference stay small for all future time? For which does every nonzero initial difference eventually exceed any fixed tolerance? Explain why both equations still have the continuous-dependence property from the preceding part. Pay attention to the distinction between fixing a finite observation interval and asking for closeness for all future time.

Building and Exploring Models

In the last two chapters, we learned how to find solutions to differential equations and how to understand their behavior without finding a formula. But where do the equations come from? If we begin with a question about the world, how do we decide which equation to write?

This is the task of *mathematical modeling*. We begin by identifying the quantities that seem relevant to our question. Then we decide which changes we need to follow and which quantities we can reasonably treat as constant. Using these ingredients and their derivatives, we try to turn an ordinary description of what is happening into a precise mathematical relationship.

There is an art to these choices. The same description may suggest several equations, and a quantity we treat as constant for one question may need to vary for another. Our aim is to build a model simple enough that we can understand it and make predictions with it, but detailed enough to capture the behavior that matters. Once we have an equation, mathematics tells us what it implies. We can then compare those consequences with observations and use what we learn to revise our choices.

We will begin by returning to Newton's law of cooling, this time slowing down to examine how we build the model. Then we will investigate population growth, letting measurements guide us from a simple model to a more complicated one, before adding harvesting. Elsewhere in the chapter, models of interacting populations, epidemics, and moving springs will ask us to follow more quantities and relate them in new ways. The chapter closes with a climate appendix, where we apply the same modeling approach to Earth's temperature. Ecology, climate, and mechanics will keep returning throughout the book as we develop the mathematics their models lead us to.

4.1 Newton's Law of Cooling

Suppose we leave a hot cup of coffee on the table and want to know how warm it will be twenty minutes from now. What quantities might matter? The coffee's current temperature matters, of course. But so does the temperature of its surroundings: the same coffee would cool differently in a warm kitchen and a cold garage.

Let us begin with those two quantities: the temperature of the coffee and the temperature of the room. Which should we treat as variables, and which as constants?

Over twenty minutes, the coffee's temperature may change substantially, while the room's temperature changes very little. So a reasonable first choice is to let the coffee's temperature be a function $T(t)$, while treating the room's temperature as a constant T_{room} . We are also simplifying by describing the coffee with just one temperature, ignoring differences between the liquid near the surface and the liquid deeper in the cup.

These are choices about our model. Over a single second, we might reasonably treat both temperatures as nearly constant. Over a whole day, the room may warm and cool enough that we need to follow its temperature too. We choose what to hold fixed by considering the time scale and the question we want to answer.

Now we have our ingredients. What can we say about how they are related? A cup hotter than the room cools down, a cup colder than the room warms up, and a larger temperature difference produces a faster change. The quantity

$$T - T_{\text{room}}$$

therefore seems useful: its sign tells us which way the temperature should change, and its magnitude tells us how far the coffee is from matching its surroundings.

But "a larger difference produces a faster change" does not tell us exactly how much faster. We have to make a choice. One simple possibility is *proportionality*: doubling the temperature difference doubles the rate of change. With the sign chosen so that hot coffee cools, this gives

$$T' = -k(T - T_{\text{room}}), \quad k > 0.$$

Why do we need the proportionality constant k ? Even before considering any details of heat transfer, the units tell us that something is missing without it. The temperature difference is measured in degrees, while T' is measured in degrees *per unit time*. Suppose we change our time unit from seconds to hours. At the same physical instant, the temperature difference has the same numerical value, but the rate expressed in degrees per hour is 3600 times the rate expressed in degrees per second! The numerical value of k must change by the same factor.

Thus k has units of inverse time. Its value also describes how quickly this particular object responds to its surroundings: a larger k gives a faster change for the same temperature difference. Our verbal observations have not determined that value; we would need measurements or further physical information. Treating k as constant throughout the cooling is another assumption of our simple model.

Let us check the equation against the observations that motivated it. When $T > T_{\text{room}}$, the right side is negative, so the coffee cools. When $T < T_{\text{room}}$, it is positive, so the coffee warms. And when the temperatures agree, the rate is zero. The equation says what we intended!

We have recovered **Newton's law of cooling**, which we already solved in chapter 2:

$$T(t) = T_{\text{room}} + (T(0) - T_{\text{room}})e^{-kt}.$$

Once we have reason to trust this model—for example, because we have seen it agree with measurements in many situations—we can put it to work. But our formula still contains k . The cooling law itself does not tell us its value: that depends on the thermal properties of the object and its surroundings, as well as our choice of time unit.

Fortunately, we do not need to derive k from those physical details. We can *measure* our way to it! Suppose we know the constant room temperature and measure the hot coffee's temperature at time 0. Our formula then has only one undetermined constant left. Measuring the coffee's temperature once more, at a later time, gives an equation we can solve for k . With that value in hand, the formula predicts the temperature at times we have not measured.

This is a very useful feature of modeling. We began by wanting an entire function—the coffee's future temperature—but the model reduced what we needed to learn to finitely many numbers. A few measurements determine those numbers, and the equation supplies the rest of the history. In the problems, you will use two measurements of a cooling cup of coffee to predict a third.

Try this with the measurements in Figure figure 4.1. Here we have more than two observations, so we can also see how well one choice of k describes the whole cooling history.

Of course, determining the constant assumes that we have chosen a suitable model. Let us now turn to population growth, where measurements will help us both choose the constants and investigate whether the equation itself needs to change.

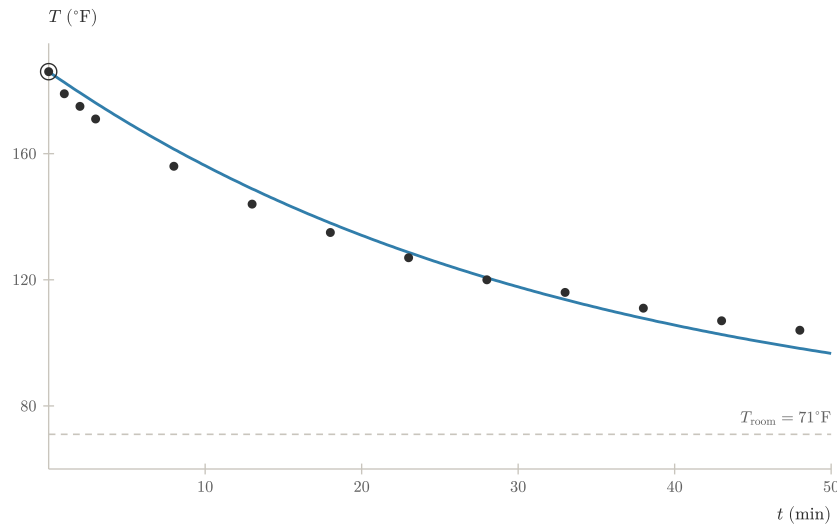


Figure 4.1 The dots are reported coffee-temperature measurements from *Active Prelude to Calculus*. The initial temperature is 186°F and the room temperature is 71°F ; keep both fixed and adjust k to fit the curve. With $k = 0.030 \text{ min}^{-1}$, the model follows the main cooling trend, although it does not pass through every measurement.

4.2 Population Growth

How can we predict the size of a growing population? Let us begin with a population living where food and space are plentiful, and try to describe its growth with a single quantity: the amount of population present, $N(t)$.

4.2.1 From Exponential to Logistic Growth

What determines how quickly this quantity changes? New individuals are born, while others die. For a first model, suppose each individual makes approximately the same average contribution to the population's net growth. Then twice as many individuals should produce twice as much net growth over the same short interval. As with cooling, a simple way to express this assumption is proportionality:

$$N' = rN.$$

The coefficient r measures the net growth per unit of population per unit time.

Over time, the conditions in which a population lives can change, and its growth rate can change with them. In the United States, the annual birth rate is estimated to have been about 50 births per 1,000 people around 1800; in 2024 it was 10.7. Over those centuries, family life, economic conditions, and expectations about having chil-

dren changed enormously as the country shifted from a predominantly rural society toward an urban one. Births are only one contribution to population growth—deaths and migration matter too—but even this one contribution is clearly something we cannot treat as constant forever. (Historical estimates; 2024 measurements.)

But we may not need to describe centuries of change. Over a shorter window—perhaps a few hours for a yeast culture, or a decade for a human population—the conditions controlling growth may change relatively little while the population itself changes appreciably. This gives us a reason to try treating r as constant in our first simplified model. As with the coffee and its surroundings, we are choosing to follow one quantity while holding another approximately fixed.

The units work just as they did for cooling. Since N' has units of population per time and N has units of population, r must have units of inverse time. Its numerical value depends both on our time unit and on how quickly this particular population grows.

We already know the solution. If $N(t_0) = N_0$, then

$$N(t) = N_0 e^{r(t-t_0)}.$$

For $r > 0$, our assumption predicts exponential growth. Once we measure the initial population, the only undetermined constant is r . We can therefore try what we did with coffee: use measurements to choose the coefficient, and see how well the resulting curve describes the population.

In 1913, Tor Carlson measured the amount of yeast growing in brewer's wort at hourly intervals. Let us imagine watching his experiment as it happens. We have seen the first six measurements, taken an hour apart, and want to use them to build a prediction.

The first measurement gives $N(1) = 9.6$, so our exponential model becomes

$$N(t) = 9.6 e^{r(t-1)}.$$

Just as with the coffee, the initial measurement fixes one constant, leaving a single parameter for us to determine. Here it is r , the growth coefficient. Changing r changes how quickly the curve rises, while every choice passes through the same measured starting point.

This looks promising! One constant describes the rapid growth quite well. We have done more than draw a curve through measurements: we have chosen a differential equation, used the measurements to determine its remaining parameter, and obtained a function we can evaluate at future times.

But those future values are predictions. To find out whether they are useful, we have to let the experiment continue.

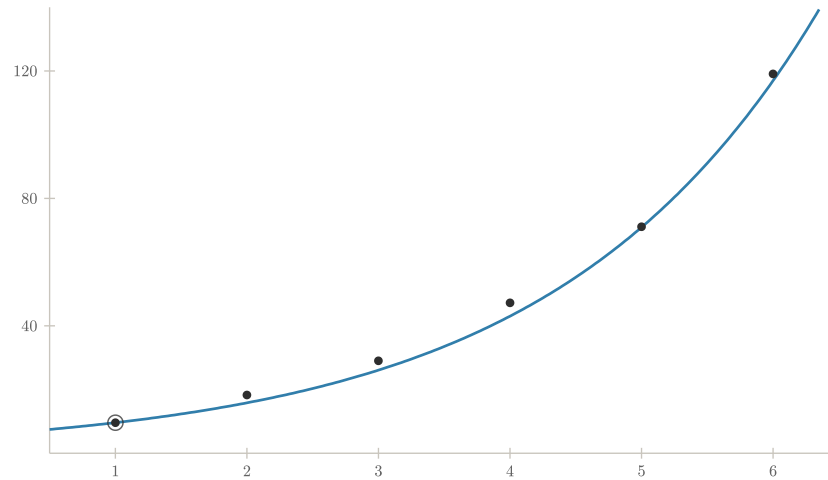


Figure 4.2 The dots show Carlson's first six measurements. Adjust r to fit the exponential curve to them, keeping $N(1) = 9.6$ fixed. With $r = 0.50 \text{ h}^{-1}$, the model predicts 117.0 at hour 6; Carlson measured 119.1.

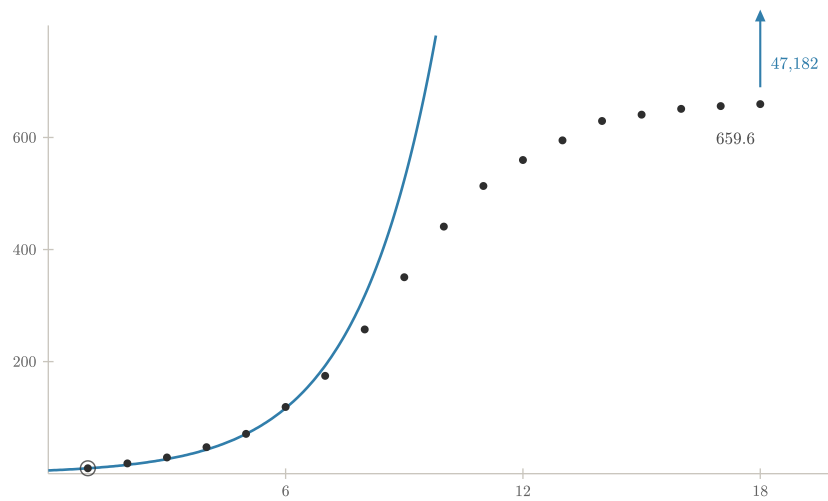


Figure 4.3 Now reveal the remaining measurements. The amount of yeast levels off near 660, while our exponential curve continues upward. Try changing r : a value that matches the later measurements misses the early growth, and a value that matches the early growth misses the later measurements.

At hour 18, our model predicts

$$N(18) = 9.6e^{0.50(18-1)} \approx 47,182,$$

while Carlson measured 659.6. The prediction is about seventy-two times too large!

The same model that worked well over a few hours has failed over a longer interval. And adjusting its one constant cannot fix the problem. We need to reconsider the assumption that each unit of population keeps contributing the same net growth per unit time.

Can we use the measurements to check our assumption? We said that each unit of population contributes the same net growth per unit time. To investigate that contribution, divide the total growth rate by the amount of population present:

$$\frac{N'}{N}.$$

This is called the *per-capita growth rate*. Our exponential model says that it equals the constant r . But what do the measurements say?

We do not have measurements of N' directly. We have the amount of yeast at particular times. Between two consecutive measurements, we can estimate the derivative using a difference quotient:

$$N' \approx \frac{N_{n+1} - N_n}{\Delta t}.$$

To estimate the growth per unit of population, we also need the amount of yeast present during that interval. It changes from N_n to N_{n+1} , so let us use their average. This gives

$$\text{estimated per-capita growth rate} = \frac{N_{n+1} - N_n}{\left(\frac{N_{n+1} + N_n}{2}\right) \Delta t}.$$

We can compute this number for each pair of consecutive measurements, then plot it against the average amount of yeast present.

The data suggest a way forward! The per-capita rate is not constant, but its decrease looks approximately linear. We could try many curves through these points; a line is a simple next choice.

Let r denote the rate when the population is very small, and let K denote the population at which the line reaches zero. The line with these two intercepts is

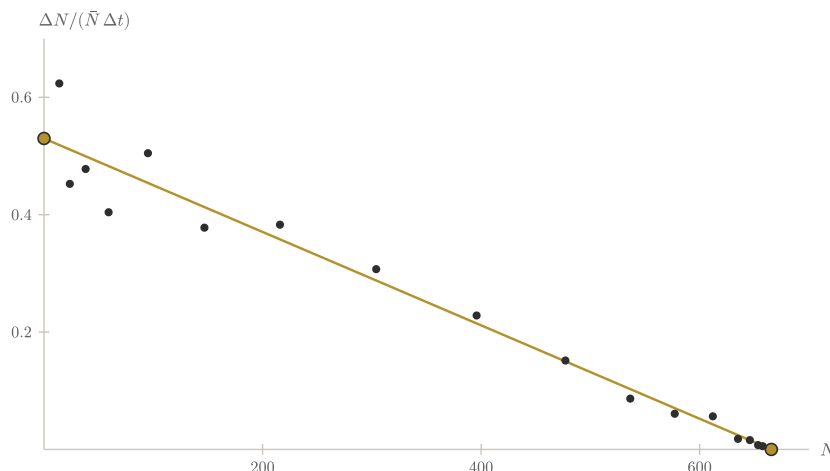


Figure 4.4 Each dot compares the average amount of yeast during an interval with its estimated per-capita growth rate. If our exponential model were accurate, the dots would lie near one horizontal level. Instead, the rate decreases as the amount of yeast increases. Try fitting a straight line to this relationship.

$$\frac{N'}{N} = r \left(1 - \frac{N}{K} \right).$$

Multiplying by N , we obtain the **logistic equation**:

$$N' = rN \left(1 - \frac{N}{K} \right).$$

Notice what we have changed. The growth per unit of population now *varies* as the population grows. We still have constant parameters, r and K , but they describe a changing rate. The measurements have helped us replace one assumption with another. The parameter r still has units of inverse time, while K has the same units as N .

We have not established that the decrease must be linear. We have chosen a simple relationship suggested by the data. Now we need to investigate what this new equation predicts—and whether those predictions describe the experiment.

Before solving anything, let us see what the equation tells us:

$$N' = rN \left(1 - \frac{N}{K} \right).$$

When N is much smaller than K , the factor $1 - N/K$ is nearly 1, so

$$N' \approx rN.$$

Our revised model therefore retains the early exponential growth that worked so well! The correction becomes important as the population grows. At $N = K$, it makes the growth rate zero; above K , it makes the rate negative.

We can summarize these observations with a phase line:

$$0 \longrightarrow K \longleftarrow .$$

Both $N = 0$ and $N = K$ are equilibrium solutions. A population between them grows, but uniqueness prevents its solution from crossing K . By the reasoning from chapter 3, it approaches K as time increases. A population starting above K decreases toward the same equilibrium. This is why we call K the **carrying capacity** of the model: it is the population level toward which every positive starting population tends.

So we already know something useful. Our new equation predicts growth that eventually levels off, just as Carlson's measurements do. But does it level off at the right height, and does it get there at the right pace?

The calculus from chapter 2 lets us make this comparison precisely. Away from the equilibria, we can separate variables:

$$\frac{dN}{N(1 - N/K)} = r dt.$$

The partial-fraction identity

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

lets us integrate the left side. For an initial population $N(t_0) = N_0$ with $0 < N_0 < K$, solving the resulting equation for N gives

$$N(t) = \frac{K}{1 + \left(\frac{K - N_0}{N_0}\right) e^{-r(t - t_0)}}.$$

Now we can plot this solution against the measurements. As before, the first measurement fixes the initial condition. The two parameters left to choose are r and K , and the per-capita plot has already suggested values for both.

This is worth celebrating! We began with a simple assumption, watched its prediction fail, and used the measurements to suggest a revision. Solving the revised equation then produced a curve that follows the whole experiment.

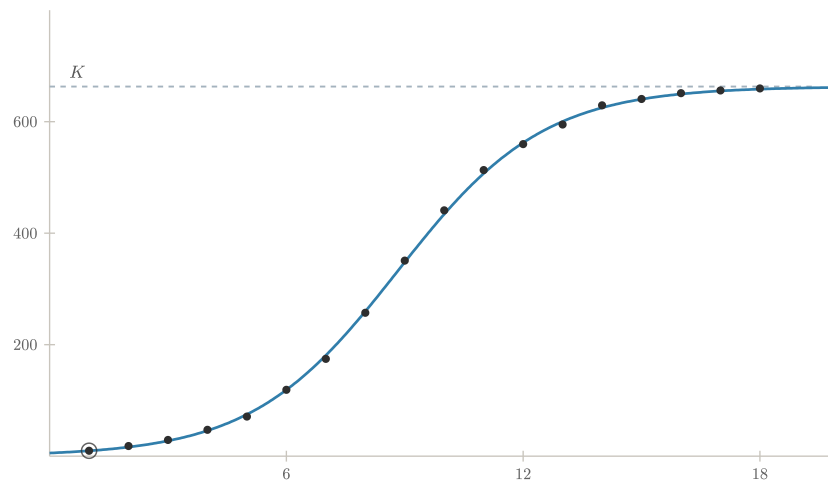


Figure 4.5 The logistic solution starts at Carlson’s first measurement, $N(1) = 9.6$. With $r = 0.54 \text{ h}^{-1}$ and $K = 663$, it follows the measurements through both the early growth and the later plateau. Adjust r and K yourself. The approximate values suggested by the per-capita plot, $r \approx 0.53 \text{ h}^{-1}$ and $K \approx 666$, also give a close fit.

The logistic model is only a little more complicated than the exponential model: we added one parameter and let the growth per unit of population decrease linearly. Yet that change captures both the early rapid growth and the later slowdown. We have found a model that is still simple enough to understand, but now describes the important behavior over a much longer interval.

We built the logistic model by examining one yeast culture. How much of its success carries over to other populations?

In the 1930s, Georgy Gause measured populations of the single-celled organism *Paramecium* growing in laboratory cultures. His measurements fluctuate more than Carlson’s yeast record, but we can recognize the same broad pattern: rapid growth at first, followed by a slowdown toward a persistent population level.

Try fitting the logistic model to these measurements. We need different values of r and K , but the equation itself stays the same. It does not reproduce every fluctuation. It does capture the larger pattern around which those fluctuations occur.

Now make a much larger leap. Instead of a laboratory culture observed over days, consider the population of the United States, recorded by the census every ten years. These measurements span more than two centuries, during which the country’s territory, economy, medicine, and ways of life changed enormously. Can the same simple equation describe anything useful here?

The census comparison is striking. A logistic curve captures much of the population’s long sweep, despite the enormous complexity of the history behind those

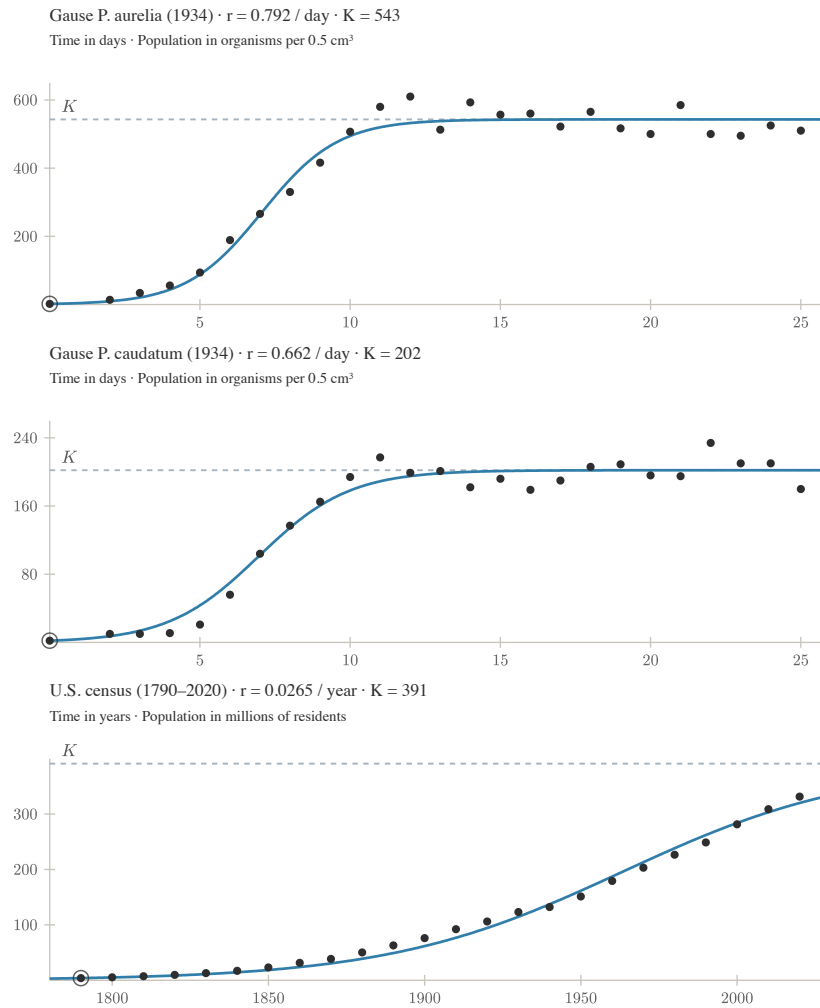


Figure 4.6 Choose among Carlson’s yeast, Gause’s two *Paramecium* cultures, and the U.S. census from 1790 through 2020. For each record, the first measurement fixes the initial population; adjust r and K to fit the remaining observations. Notice the different population and time units. How much of each record can you describe with just these two parameters?

measurements. This is the kind of success we are looking for in a simple model: a small amount of mathematics organizes a great deal of observed behavior.

But interpreting its parameters requires judgment. In the laboratory cultures, the measurements show a recognizable plateau. The census record has not settled at a population ceiling, so its fitted K is an extrapolation beyond what we have observed.

And remember our earlier discussion of changing birth rates. The logistic model allows growth per unit of population to change, but it describes that change using only the population itself. A successful fit does not tell us that crowding explains changes in family size, or that immigration and technology can be ignored when predicting the future.

We can learn something from the broad pattern without claiming to have explained every cause. Whether that is enough depends on the question we want the model to answer.

4.2.2 Harvesting

So far, we have asked how a population grows. Now suppose we want to use that population: we have a trout pond, and we would like to catch fish. How much can we harvest without eventually emptying the pond?

Let $P(t)$ denote the fish population. Suppose its natural growth is described by the logistic model,

$$P' = rP \left(1 - \frac{P}{K} \right).$$

Fishing introduces another process that changes P . To include it, we need to decide how the harvesting works. Let us suppose fish are removed at a constant rate of H fish per year, regardless of how many fish are currently in the pond. This contributes a negative term:

$$P' = rP \left(1 - \frac{P}{K} \right) - H.$$

For a concrete example, suppose the pond has a carrying capacity of 200 fish and a growth coefficient of 1 per year. Including harvesting gives

$$P' = P \left(1 - \frac{P}{200} \right) - H,$$

where time is measured in years and H is the number of fish removed per year.

The population stays constant when natural growth exactly replaces the fish we catch:

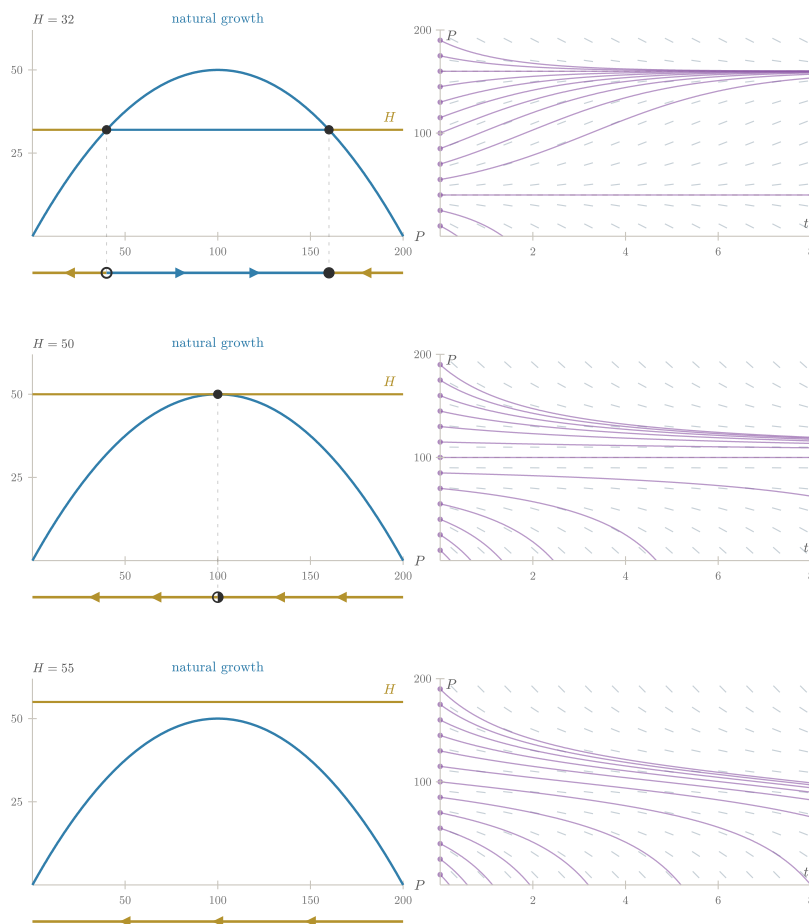


Figure 4.7 Compare the pond's natural growth rate with the constant harvesting rate. Their intersections give the equilibria, and the signs of their difference determine the phase-line arrows and solution histories. Increase H and watch how the possible futures change.

$$P \left(1 - \frac{P}{200} \right) = H.$$

We can find these equilibria graphically. Plot the natural growth rate against the population: it is a parabola. The constant harvesting rate is a horizontal line, and their intersections are the populations where growth and harvesting balance.

Where the parabola lies above the line, the population grows; where it lies below, the population declines. So this one picture also tells us how to draw the phase-line arrows!

For $H = 32$ fish per year, the two rates balance at

$$P = 40 \quad \text{and} \quad P = 160.$$

If the population falls below 40, its natural growth cannot replace the fish we remove, and it keeps declining. Between 40 and 160, growth exceeds harvesting, so the population increases. Above 160, harvesting again exceeds natural growth, and the population decreases. Thus 40 is an unstable threshold, while 160 is an attracting equilibrium.

The same harvest can therefore have very different consequences depending on how many fish we begin with!

What happens if we increase the quota? The horizontal line rises, and the two intersections approach one another. The growth curve reaches its maximum at $P = 100$, where natural growth adds 50 fish per year.

At $H = 50$, the two equilibria meet. Our equation becomes

$$P' = -\frac{(P - 100)^2}{200}.$$

A population above 100 decreases toward the equilibrium, but one below 100 keeps declining. If we raise the harvest above 50, the pond cannot replace the catch at *any* population: there are no positive equilibria.

The model has helped us identify a maximum sustainable harvesting rate, but it has also shown us why operating at that maximum is precarious. At $H = 50$, even a small drop below the equilibrium leads to continued decline. Errors in our population estimate or growth model, or an unusually poor year, could matter enormously. A harvesting policy needs room for that uncertainty.

There is another limitation we can read directly from the equation. At $P = 0$, it still says $P' = -H$: our constant quota keeps removing fish from an empty pond! We must stop interpreting the solution as a population when it reaches zero.

We could instead model harvesting with a term such as $-qP$, removing fish at a rate proportional to the population. That rate would automatically approach zero as the pond empties. But it would describe a different harvesting policy and give us a different equation to investigate.

Once again, choosing the term is part of choosing the model. One constant subtraction created a population threshold and changed which long-term outcomes were possible.

4.3 Modeling with Systems

So far, we have built models around a single changing quantity: the temperature of an object or the size of a population. But sometimes predicting one quantity requires us to follow another. An object changes the temperature of the room around it, which in turn affects how the object cools. Predators change the population of their prey, while the available prey affects how the predators grow.

In these situations, we need several unknown functions and equations describing how they change together. We are naturally led to a **system of differential equations**. Let us return to temperature and population to see how this happens.

4.3.1 The Coffee and the Room

When we modeled a cooling cup of coffee, we treated the room's temperature as constant. But the hot object also warms its surroundings. What if that change is large enough to matter?

Let $T(t)$ be the object's temperature and $R(t)$ the room's temperature. We will still describe each with a single temperature, but now both are unknown functions of time. Our rule for the object stays the same:

$$T' = -k(T - R), \quad k > 0.$$

We need a new rule for R' . For this first model, suppose the room's temperature changes only because of its interaction with the object. When the object is hotter than the room, the room warms; when the object is colder, the room cools. A larger temperature difference should produce a faster change. Just as before, let us choose proportionality as a simple way to make that description precise:

$$R' = c(T - R), \quad c > 0.$$

The new constant c describes how strongly the room responds. There is no reason for it to equal k : the object and the room need not change temperature equally quickly. In fact, for a cup of coffee in a large room, we would expect c to be much smaller than k : the coffee can cool substantially while the room barely warms.

Putting the two rules together gives our system:

$$\begin{aligned} T' &= -k(T - R), \\ R' &= c(T - R). \end{aligned}$$

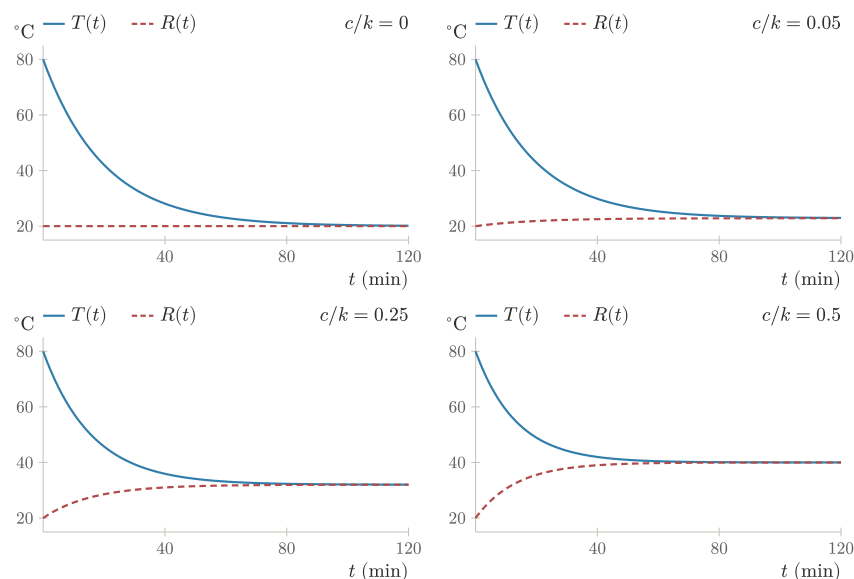


Figure 4.8 The object begins at 80°C and the room at 20°C . The solid curve follows the object and the dashed curve follows the room. Begin with $c = 0$, then increase c/k to make the room respond more strongly. Compare $c/k = 0, 0.05, 0.25$, and 0.5 , keeping $k = 0.05 \text{ min}^{-1}$. These are illustrative numerical solutions, not fitted temperature measurements.

If we set $c = 0$, then $R' = 0$: the room stays at its initial temperature, and we recover our original cooling model exactly!

For $c > 0$, the equations must be studied together. To know how quickly the object cools, we need the room's current temperature. But that temperature is itself changing in response to the object! A solution now consists of *two functions*, $T(t)$ and $R(t)$, satisfying their equations together. To begin a prediction, we need both initial temperatures.

We have not yet developed systematic methods for finding formulas for solutions of systems like this. Learning how to solve and understand systems will be one of the main topics of the rest of the book. But we can already explore them numerically. Euler's method asks us to compute both rates from the current temperatures, then use those rates to advance both temperatures by one short time step. We must compute both rates *before* updating either temperature, so they refer to the same moment.

The two temperatures approach agreement. When c is small compared with k , they meet near the room's initial temperature. Increasing c/k makes the room warm more, and the two temperatures approach a higher common value. Try changing k while keeping c/k fixed: the histories unfold at a different speed, but approach the same temperature.

Allowing one more quantity to change gave us a different prediction and two equations to study together. Let us now see how interactions between populations lead to the same mathematical need.

4.3.2 Interacting Populations

Suppose we want to predict how a population of hares will change. Earlier, we tried to do this using only the number of hares. But if lynx are hunting them, that number alone is no longer enough. The same hare population might grow when there are few lynx and decline when there are many. And the lynx population depends on how much food the hares provide. As with the coffee and the room, predicting either quantity requires us to follow both.

Let $H(t)$ and $L(t)$ be the hare and lynx populations. Before putting them together, let us consider how each might change on its own.

In the absence of lynx, suppose the hares have plenty of food and space. Our earlier assumption of a constant average contribution to net growth gives

$$H' = aH, \quad a > 0.$$

We recognize exponential growth. This is our first approximation to what the hares would do without predators.

What should happen to the lynx without hares to eat? We expect their population to decline, but that does not mean every lynx dies at the same instant. For a simple model, suppose that over each short interval, approximately the same fraction of the surviving lynx dies. Twice as many remaining lynx then means twice as many deaths per unit time. This gives

$$L' = -cL, \quad c > 0,$$

and hence exponential decline, $L(t) = L(0)e^{-ct}$. A constant death rate per individual is a modeling choice; it does not describe every detail of starvation. Like our other continuous population models, it also stops being a useful description when individual animals matter.

Now let the two populations interact.

What about encounters? If the populations mix uniformly, doubling the number of hares should double the opportunities for any lynx to find one, and doubling the number of lynx should double the total opportunities again. The simplest expression with both properties is the product HL . Let bHL be the rate at which encounters re-

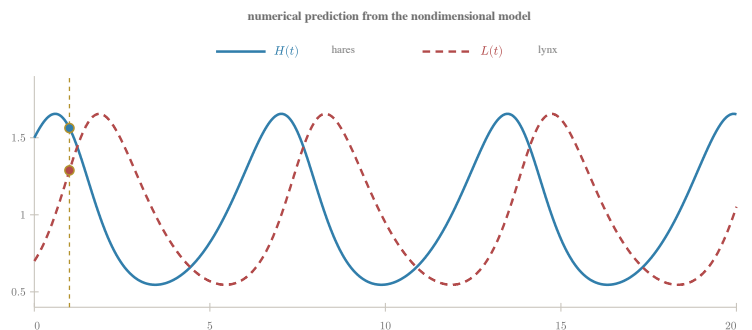


Figure 4.9 For the illustrative family $H' = H(1 - L)$, $L' = \gamma L(H - 1)$, follow the solid hare history and the dashed lynx history. Both populations repeatedly rise and fall! Change the relative predator response γ and compare the timing of their peaks.

move hares, and let dHL be the rate at which food from those encounters contributes to lynx growth. Then

$$\begin{aligned} H' &= aH - bHL, \\ L' &= -cL + dHL. \end{aligned}$$

This is the **Lotka–Volterra predator–prey model**. The coefficients b and d need not be equal: one successful encounter removes one hare, but its contribution to predator reproduction depends on how food is converted into new lynx.

The limiting cases check the bookkeeping. If $L = 0$, then $H' = aH$ and hares grow exponentially. If $H = 0$, then $L' = -cL$ and lynx decline exponentially. If either population is zero, the interaction term vanishes. The signs of all four terms match the verbal story.

We have built a system because the populations affect one another. A solution consists of both histories, $H(t)$ and $L(t)$, and a prediction needs both starting populations.

We do not need to fit a particular ecosystem in order to explore one consequence of the model. Choose simple population and time units, take $a = b = 1$ and $c = d = \gamma$. This gives the illustrative family

$$\begin{aligned} H' &= H(1 - L), \\ L' &= \gamma L(H - 1), \end{aligned}$$

Here γ controls how quickly the predator population responds relative to the prey population. Let us compute both histories and see what the interaction produces.

The populations oscillate! When hares are abundant, the lynx have enough food for their population to grow. More lynx then put more pressure on the hares, whose population declines. With less food available, the lynx begin to decline too. Fewer predators give the hares a chance to recover, and the cycle begins again. In the computed histories, the hare peak comes before the lynx peak: food becomes abundant first, and the predator population responds afterward.

Remember the conclusion from chapter 3: a nonconstant solution of a single smooth autonomous equation $N' = f(N)$ must be increasing or decreasing. It cannot keep rising and falling periodically. But knowing H alone no longer determines H' : the same number of hares can be growing or declining, depending on the lynx population. By following two interacting quantities, we have opened up a new kind of behavior.

We can explore these oscillations numerically now. Explaining their structure mathematically will be part of our later study of systems.

This is a prediction worth taking back to evidence. The next figure compares the Hudson Bay hare–lynx fur record, a laboratory Paramecium–Didinium experiment, and a prey–predatory-mite experiment. Each population has been divided by its own observed maximum, so we are comparing relative timing, not fitting one scaled model to all three records.

The resemblance is striking. Yet fur returns depend on trapping as well as abundance, and normalization reveals timing while hiding scale. The evidence supports a qualitative mechanism; it does not establish one four-parameter model for every predator–prey system.

In Chapter 3, we sometimes compressed a solution picture onto a phase line by keeping the current value and leaving out time. We can do something similar here, but the current state consists of two populations.

Return to our illustrative model

$$H' = H(1 - L), \quad L' = \gamma L(H - 1).$$

At any time t , the two heights on the solution graphs give us one point:

$$(H(t), L(t)).$$

Put the hare population on the horizontal axis and the lynx population on the vertical axis. As time passes, this point moves through the plane, tracing a parametric curve. We call this plane the **phase plane** and the curve a **trajectory**.

The differential equations tell us how the point moves. Its horizontal velocity is H' , and its vertical velocity is L' . For example, at $(H, L) = (1.5, 0.7)$,

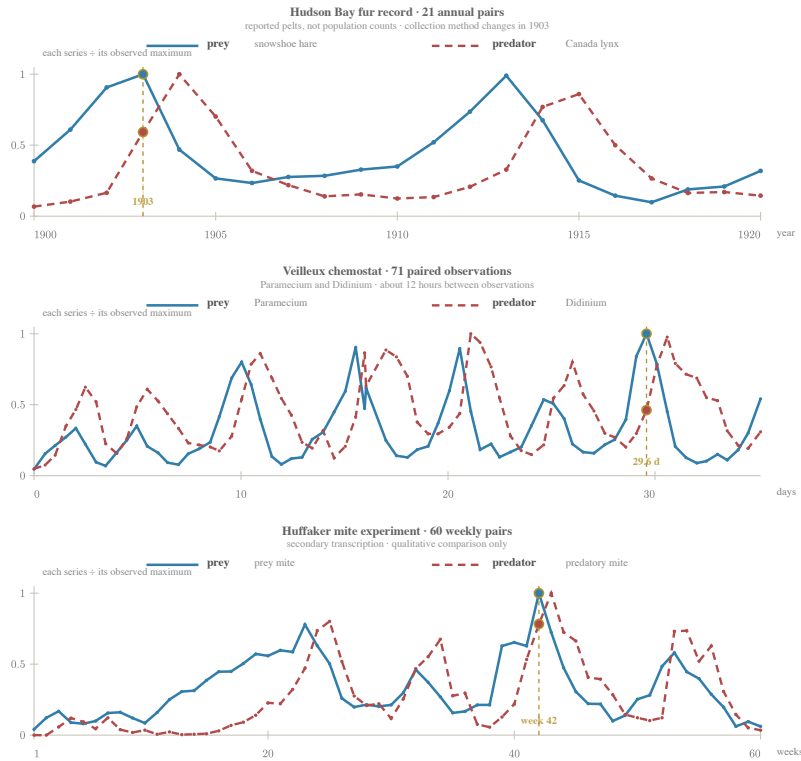


Figure 4.10 Does the prey peak first outside our computed example? Compare the Hudson Bay hare–lynx record, a *Paramecium*–*Didinium* chemostat, and Huffaker’s mite experiment. Each series is divided by its own observed maximum, so we are comparing timing rather than abundance or fitted parameters. The Hudson Bay values are pelts rather than population counts, and the collection method changed in 1903. The Huffaker series is a secondary transcription included only for qualitative comparison.

$$H' = 1.5(1 - 0.7) = 0.45, \quad L' = \gamma(0.7)(1.5 - 1) = 0.35\gamma.$$

Both populations are increasing, so the point moves upward and to the right. We can repeat this calculation at other points to draw a field of arrows.

Our equations are **autonomous**: their rates depend on the current populations, without explicitly depending on time. This is what lets us assign one fixed arrow to each point in the plane. Whenever the populations have those values, the equations prescribe the same velocity.

The arrows and trajectories together form a **phase portrait**. Here the numerical trajectory appears to make a closed loop: the populations return together to their earlier values as the cycle repeats. We have observed this in a computation; we have not yet proved that the trajectory closes.

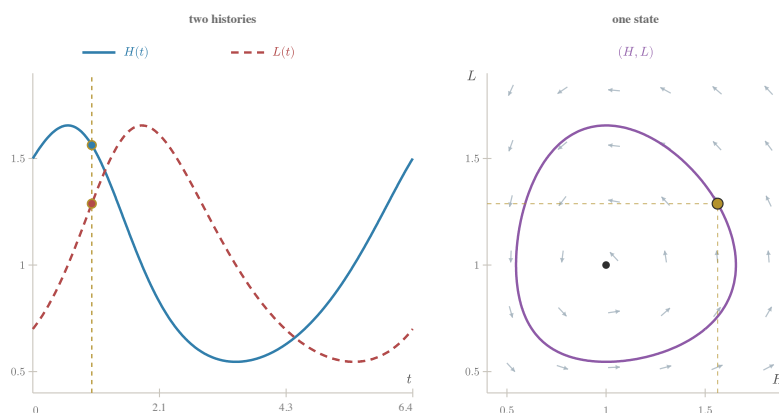


Figure 4.11 The histories on the left and the trajectory on the right show the same numerical solution. The horizontal coordinate of the moving point is the current hare population; its vertical coordinate is the current lynx population. Follow the point through the population cycle, or drag either picture to inspect a particular instant. Changing γ changes both views.

This curve is not generally the graph of lynx population as a function of hare population. The same number of hares can occur with different numbers of lynx, at different stages of the cycle. Nor does the curve alone tell us when the populations reach a particular point. The time histories retain that information; the phase-plane picture makes their motion together easier to see.

Limited food for the prey

We can now see how each assumption entered the equations: hare reproduction contributed aH , predator decline contributed $-cL$, and encounters contributed the two HL terms. That also tells us where to revise the model. Let us try two extensions, beginning with the hares' own food supply.

Without lynx, our model says the hares grow exponentially forever. But hares need grass and other plants to eat, and the available land cannot support unlimited growth. We have already built a simple model for this kind of limitation! Replace the exponential growth term with logistic growth:

$$\begin{aligned} H' &= aH \left(1 - \frac{H}{K}\right) - bHL, \\ L' &= -cL + dHL. \end{aligned}$$

Here K describes the carrying capacity of the hare population in the absence of lynx. The terms describing encounters stay the same. We have changed one assumption about how the hares grow on their own.

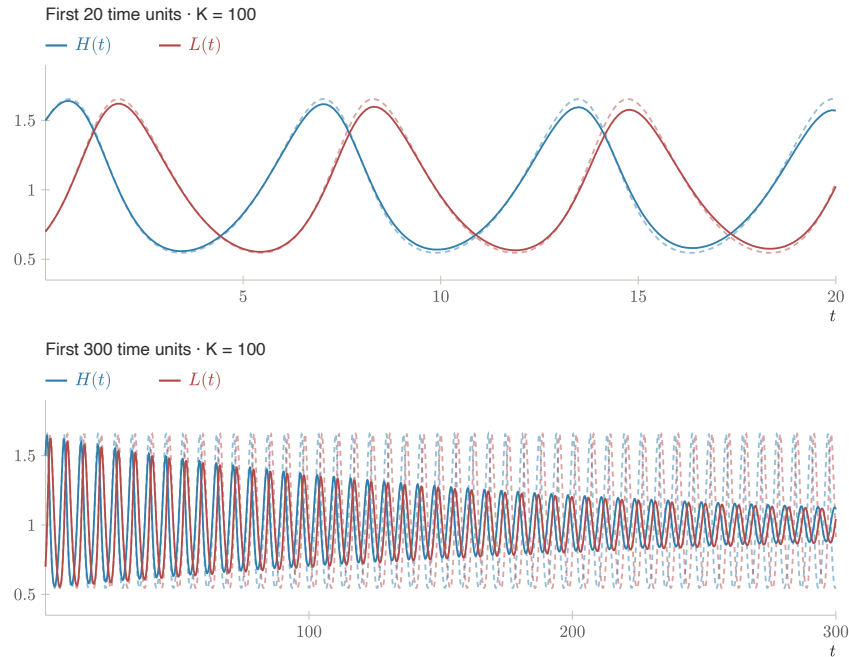


Figure 4.12 Compare the original model with the logistic-prey revision, both starting at $H(0) = 1.5$ and $L(0) = 0.7$, with $\gamma = 1$. Both populations share the same axes: blue for hares and red for lynx. The dashed curves show the original model; the solid curves include limited food for the hares. Start with $K = 100$ and the first 20 time units, then reduce K or extend the view to 300. These are illustrative model comparisons, not fits to new measurements.

How much difference does this make? If H stays much smaller than K , the factor $1 - H/K$ stays close to 1, so the correction is small. We might expect our original model to give much the same prediction. Let us compare them using the same initial populations and the same coefficients. Using the scaled populations from our earlier example gives

$$\begin{aligned} H' &= H \left(1 - \frac{H}{K} \right) - HL, \\ L' &= \gamma L (H - 1). \end{aligned}$$

With a large carrying capacity, the hares are far from using up all the grass available to them. Over the first few cycles, the original model does a pretty good job! The extra term makes little difference to the behavior we see. For smaller carrying capacities, the peaks in the revised model visibly shrink as time passes. The cycles no longer repeat in the same way.

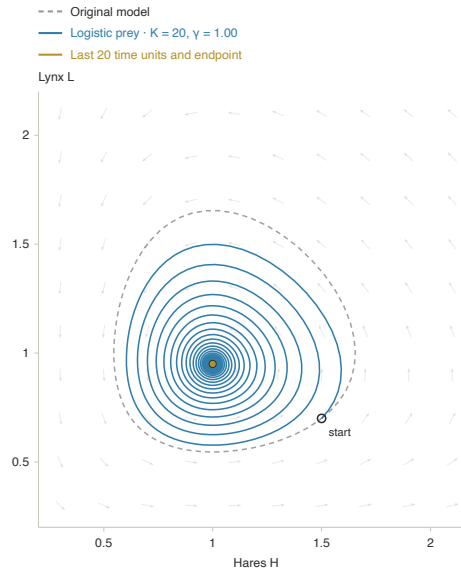


Figure 4.13 Follow the revised model over many cycles, starting from the same populations as above. The dashed loop is the original model; the blue trajectory includes limited food, and the arrows show the revised rates. Gold highlights the last 20 time units and the endpoint. Begin with $K = 20$, $\gamma = 1$, and 300 time units. Adjust K , the predator rate γ , or the elapsed time to see how the trajectory develops.

Let us follow that change in the phase plane. A repeating population cycle would trace the same closed loop each time. What happens to the trajectory when the successive peaks get smaller?

The trajectory winds inward: each trip around the center corresponds to a smaller swing in the populations. The computation suggests that both populations eventually settle at constant values. Try increasing K . The trajectory stays close to the original loop for more turns before the inward drift becomes noticeable.

We did not need to account for grass to explain why predators and prey can oscillate. But including it changes our prediction about whether those oscillations persist. If food is plentiful and we want to follow just a few cycles, the simpler model may be enough. If we want to predict what happens over many cycles, even a small limitation on the hares' food supply can matter. How far ahead we want to predict is part of choosing the model!

Hunting interacting populations

For our second extension, return to the original predator–prey model and add hunting. The trout pond taught us that a harvesting term is a modeling decision, and that removing a fixed fraction of the population is the policy which shuts itself off as the population disappears. So suppose hunters remove hares at a per-capita rate e and lynx at a per-capita rate f . We will consider $0 \leq e < 1$ and $f \geq 0$, so the hares can still grow when lynx are absent. Taking $\gamma = 1$, the two new terms are $-eH$ and $-fL$:

$$\begin{aligned} H' &= H(1 - L) - eH = H(1 - e - L), \\ L' &= L(H - 1) - fL = L(H - 1 - f). \end{aligned}$$

Before computing anything, guess. Hunting hares should mean fewer hares, and hunting lynx should mean fewer lynx. Now look for the balance point, the state at which neither population changes. When both populations are positive, the hare rate vanishes only when $L = 1 - e$, and the lynx rate vanishes only when $H = 1 + f$. So the balance sits at

$$(H, L) = (1 + f, 1 - e).$$

Read this carefully, because it contradicts the guess. The hare balance depends only on how hard we hunt *lynx*, and the lynx balance depends only on how hard we hunt *hares*. Hunting hares leaves the hare balance exactly where it was and lowers the lynx balance instead.

The figure shows why the guess was wrong. Hunting hares does remove hares, but every hare we take is one the lynx do not eat. The hares can only hold their own once the lynx have declined to $1 - e$, so it is the lynx that have the lower balance. Hunting lynx has the mirror effect: a hunted lynx population needs more food to hold steady, so the lynx are balanced only when the hares reach $1 + f$, and it is the hares that have the higher balance. The cycles persist throughout, with each population spending part of every cycle above its balance and part below. What the hunting rates control is the level the cycles surround.

Nothing about the individual terms announced this. The term $-eH$ is as simple as a term can be, and on the hares alone its effect would be just what we guessed. Its effect on the coupled system became visible only when we balanced both rates at once. This is essentially the observation that first led Volterra to the model: when fishing in the Adriatic fell during the First World War, the share of predatory fish in the catch rose. The model captures that direction of change. Easing the harvest raises the predator balance and lowers the prey balance.

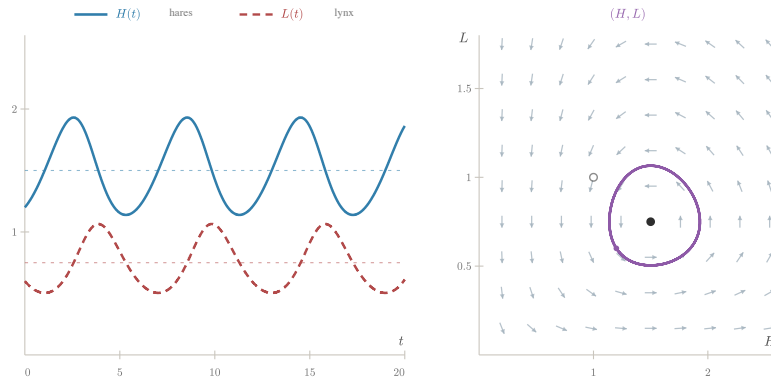


Figure 4.14 Hunt both populations at fixed per-capita rates e and f . The histories still rise and fall, but now about the dashed balance levels $H = 1 + f$ and $L = 1 - e$, and the trajectory loops around the new balance point rather than the hollow unhunted balance at $(1, 1)$. Raise the hare hunting rate and watch which dashed level moves.

Learning rates from an outbreak

An epidemic gives us another reason to follow several populations. The number of infectious people matters, but it is not enough: the same number can spread an illness very differently depending on how many people are still able to catch it. Let $S(t)$ count these *susceptible* people, $I(t)$ those who are *infectious* and can transmit the illness, and $R(t)$ those who have *recovered* from it. Here we use numbers of people rather than the fractions plotted in Chapter 1. For a first model, suppose recovered people cannot catch the illness again during this outbreak, and neglect births, deaths, and people entering or leaving the population we are following.

How quickly do new infections occur? We can reuse our reasoning about predator–prey encounters. If people mix uniformly, doubling the susceptible population should double an infectious person’s opportunities to pass on the illness. Doubling the infectious population should double the total opportunities again. This suggests an infection rate bSI , where the positive constant b accounts for how often people meet and how readily those encounters transmit the illness.

For recovery, suppose approximately the same fraction of the infectious population recovers over each short interval of a given length. Twice as many infectious people then means twice as many recoveries per unit time, giving a recovery rate γI for another positive constant γ .

Now follow where people go. Each new infection removes one person from S and adds one to I . Each recovery removes one from I and adds one to R . So the

susceptible population only loses people, the infectious population both gains and loses them, and the recovered population only gains:

$$\begin{aligned}S' &= -bSI, \\I' &= bSI - \gamma I, \\R' &= \gamma I.\end{aligned}$$

These are the **SIR equations**. The same bSI appears with opposite signs in the first two equations because an infection transfers a person between the groups; the recovery terms work the same way. We have three equations for three unknown histories.

As with the cooling constant, our assumptions have given us a model without determining its constants. We can learn about b and γ by choosing values, computing the resulting histories, and comparing them with measurements.

In October 1967, a common-cold outbreak on Tristan da Cunha was recorded through daily reports of illness and recovery. A published table of 21 daily observations gives us two curves to compare with our model. We use the number currently ill as an approximation to I , and the number recovered from this outbreak as R . This assumes that the period with symptoms approximates the period when someone can transmit the illness.

There is another unknown here: how many people were susceptible at the start? We must estimate $S(0)$ as well, keeping the observed $I(0) = 1$ and $R(0) = 0$. Take $t = 0$ at the first observation and measure time in days. For each choice of the three unknown values, a numerical solution gives us both histories to compare with the data.

The model captures the rise and fall of illness along with the accumulating recoveries. Try changing one parameter at a time. Increasing b makes new infections occur faster at the same values of S and I , while increasing γ moves people out of the infectious group faster. Changing $S(0)$ changes how many people can catch the illness in the first place. Each choice affects both curves, so matching the illness history alone is not enough: the recovery measurements give us another way to check it.

The equations can also suggest a more direct way to estimate a constant. In the exercises, we use $R' = \gamma I$ and the area under the illness curve to obtain $\gamma \approx 0.28$ per day. This is close to the value found by fitting both curves, though the two procedures need not give exactly the same answer.

There is still room to improve the model. The fit misses the sharp illness peak and predicts a slower decline. Its estimate of about forty initially susceptible people is also much smaller than the island's population: the fit does not tell us whether the others

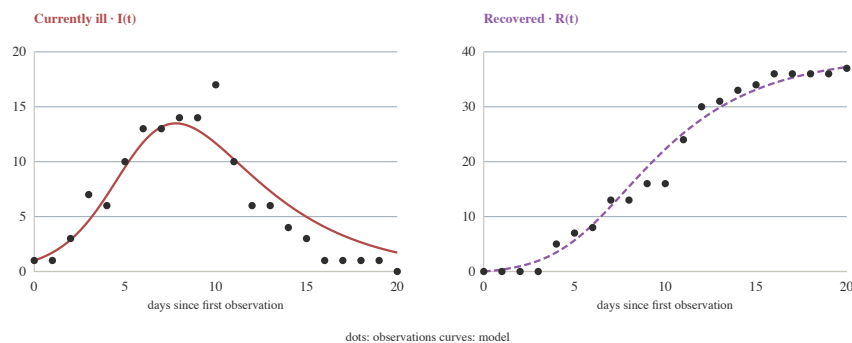


Figure 4.15 Dots show the recorded illness and recovery counts; curves show numerical solutions of the SIR equations. Adjust the transmission rate b , recovery rate γ , and initial susceptible count $S(0)$ to compare both histories. The fitted values $b \approx 0.0204 \text{ person}^{-1} \text{ day}^{-1}$, $\gamma \approx 0.267 \text{ day}^{-1}$, and $S(0) \approx 40.3$ were chosen by numerically minimizing the sum of squared discrepancies in both curves.

were immune or whether uniform mixing was a poor assumption. These are estimates within our model, not direct measurements of its mechanisms.

We could revise those assumptions too. An **SEIR model** adds an exposed group between susceptible and infectious, allowing time between catching an illness and being able to transmit it. The published comparison considers this extension using the same data.

From the coffee and the room to interacting populations, the need for systems has come from asking what else we must keep track of. Once those quantities affect one another, their equations belong together. We can already compute approximate solutions and compare them with observations; learning how to solve and understand these systems will be one of the main tasks of the rest of the book.

4.4 Forces and Second-Order Equations

Pull a mass attached to a horizontal spring away from its resting position and let go. It moves back and forth. How could we build a model of that motion?

Let $x(t)$ be the displacement from the resting position, positive to the right and negative to the left. The mass m and the spring's stiffness will matter too; we will treat both as constant. For a first model, suppose the mass slides without friction and ignore air resistance. We want to predict $x(t)$ from how we start the motion.

4.4.1 From a restoring force to an equation

The spring pulls the mass back toward its resting position. When $x > 0$, the force points left; when $x < 0$, it points right. A greater displacement also produces a stronger pull. As with cooling, we need to choose a precise relationship. Let us try proportionality: doubling the displacement doubles the force. Then

$$F_{\text{spring}} = -kx, \quad k > 0.$$

The constant k measures the stiffness, and the minus sign makes the force oppose the displacement. This is **Hooke's law**, a useful approximation for many springs over a limited range of stretching and compression.

What does this force tell us about x ? Recall that x' is velocity: it says how quickly the position changes, including the direction of motion. Its derivative x'' is acceleration, the rate at which velocity changes. Newton's second law says that total force equals mass times acceleration. Since the spring supplies the only horizontal force in our model,

$$mx'' = -kx.$$

Our description has led us to a **second-order differential equation**! The force gives us a rule for the second derivative. In particular, a mass to the right of its resting position has negative acceleration. It might still be moving right, but the spring is slowing that motion and will pull it back. The same position can occur on the outward trip and on the return trip, with different velocities.

This is why knowing the initial position alone cannot determine the motion. We must also specify how we launch the mass:

$$x(0) = x_0, \quad x'(0) = v_0.$$

These are our two initial conditions. As with interacting populations, we have found that predicting one quantity requires more information than its current value.

4.4.2 What motion does the model predict?

Dividing by m gives $x'' = -(k/m)x$. We have seen this kind of equation in chapter 2: taking two derivatives of a sine or cosine gives a negative multiple of the original function. In particular,

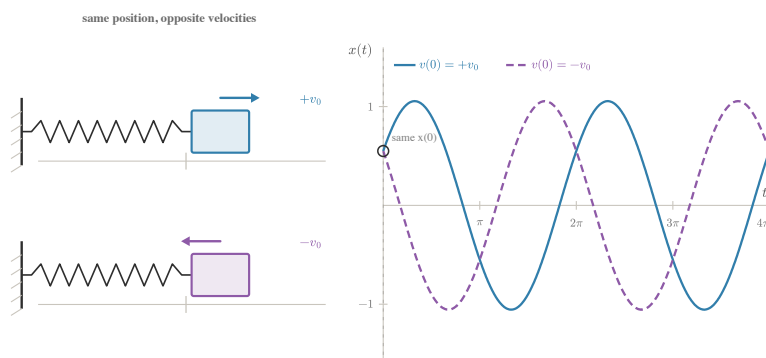


Figure 4.16 For $x'' = -x$, both masses begin at the same position, but one has velocity $+v_0$ and the other $-v_0$. Choose the common position and velocity magnitude, then play the motions. Their histories leave the same point with opposite slopes and follow different paths.

$$\frac{d^2}{dt^2} \cos(\omega t) = -\omega^2 \cos(\omega t),$$

so choosing $\omega = \sqrt{k/m}$ gives the multiplier we need. The sine works the same way, and we can check that

$$x(t) = x_0 \cos(\omega t) + \frac{v_0}{\omega} \sin(\omega t), \quad \omega = \sqrt{\frac{k}{m}},$$

satisfies both the equation and our initial conditions: at $t = 0$, its value is x_0 and its derivative is v_0 .

The model predicts oscillation. For a nonzero motion, the time to complete one cycle is

$$\mathcal{T} = \frac{2\pi}{\omega} = 2\pi \sqrt{\frac{m}{k}}.$$

A heavier mass oscillates more slowly; a stiffer spring makes it oscillate more quickly. We can also work backward: measuring the period for a known mass lets us determine k , much as temperature measurements let us find the cooling constant. You will do this in the exercises.

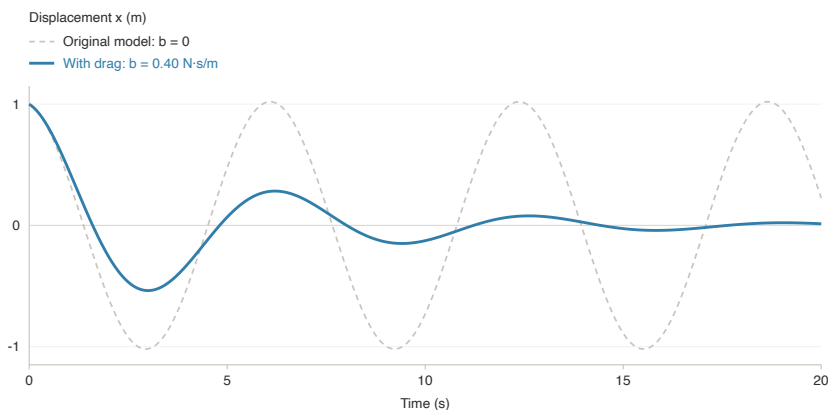


Figure 4.17 Compare the models with and without drag, starting from the same position and velocity. Here $m = 1 \text{ kg}$, $k = 1 \text{ N m}^{-1}$, $x(0) = 1 \text{ m}$, and $x'(0) = -0.2 \text{ m s}^{-1}$. Initially, the solid curve uses $b = 0.4 \text{ N s m}^{-1}$; the dashed curve uses $b = 0$. Vary b from 0 to 1 and compare the motions. These are computed solutions for illustrative parameters.

4.4.3 Revising the model: resistance to motion

Our formula predicts that the oscillations continue forever, with the same amplitude. A spring we release on a table usually settles down. What did we leave out? We deliberately ignored resistance to motion. Let us revisit that choice.

A resisting force should point opposite the velocity: left when the mass moves right, and right when it moves left. For a simple model, suppose its magnitude is proportional to speed. With a new positive constant b , this gives

$$F_{\text{drag}} = -bx'.$$

Notice the difference between the two forces. The spring responds to *position*; drag responds to *velocity*. They can point in the same direction or in opposite directions, depending on how the mass is moving. Adding their contributions to the total force gives

$$mx'' = -kx - bx', \quad \text{or} \quad mx'' + bx' + kx = 0.$$

We have changed the model by adding one term. Setting $b = 0$ recovers our first equation. For $b > 0$, we do not yet have a systematic solution method, but we can compute motions and see what the new term does.

With these positive choices of b , the successive swings shrink and the mass approaches its resting position. The added force has captured something our first model

missed. Whether this particular resistance law describes a real spring well, and which value of b to use, would require measurements.

How can we compute these motions with the first-order methods we already know? Let $v = x'$ be velocity. Then the second-order equation becomes

$$\begin{aligned}x' &= v, \\v' &= -\frac{k}{m}x - \frac{b}{m}v.\end{aligned}$$

The first equation says how velocity changes position; the second says how the forces change velocity. Starting from (x_0, v_0) , Euler's method can advance both together, just as it advanced the coffee and room temperatures. Conversely, substituting $v = x'$ into the second equation recovers our original second-order equation.

Position and velocity together form the **state** of this model. We can describe its motion with one second-order equation or two first-order equations. So the two extensions we have met in this chapter are connected: learning to work with systems will also help us understand equations with higher derivatives.

4.5 Appendix: Modeling Earth's Temperature

Sunlight warms Earth, while radiation carries energy away into space. Can we turn this description into a model of Earth's temperature? Let us try the same process we used for coffee and populations: choose what to hold constant, build an equation, and see what its predictions tell us.

4.5.1 A planet with one temperature

Our ingredients are the surface temperature, the sunlight Earth absorbs, and the radiation it emits. Let $T(t)$ be the mean surface temperature, setting aside differences between places. Average the absorbed sunlight over the whole planet and call its rate Q . We will treat Q as constant, ignoring daily and seasonal changes. For our first attempt, suppose all radiation emitted by the surface escapes directly to space.

How should these quantities determine T' ? Absorbing energy warms the planet; emitting energy cools it. Just as with growth and harvesting, we need a positive contribution and a negative one. The temperature should rise when more energy arrives than leaves, and fall when more leaves than arrives.

To make this precise, suppose a fixed amount of energy is needed to raise the temperature by one degree. Call that amount C , measured per square meter of surface.

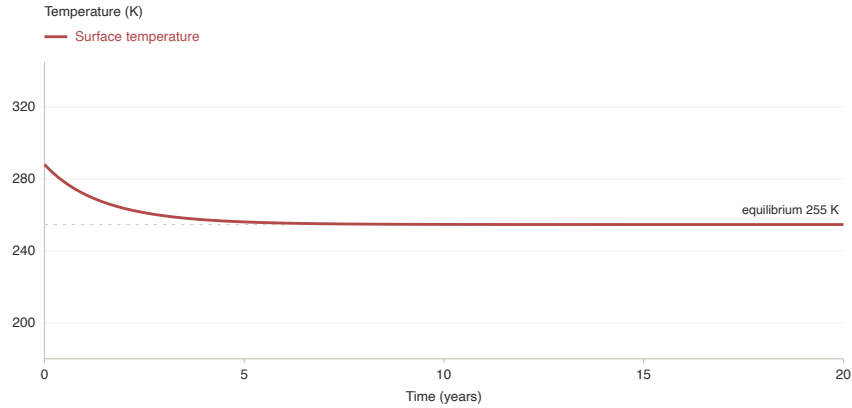


Figure 4.18 Start at 288 K and watch the model cool. Try a colder initial temperature and a larger heat capacity. The default $C = 2 \times 10^8 \text{ J m}^{-2}\text{K}^{-1}$ is illustrative; the curves show the model's predictions, not measurements of Earth's temperature history.

This constant is the model's *heat capacity*: a larger C means the same energy gain produces a smaller temperature change. Then CT' is the rate at which energy accumulates per square meter, so

$$CT' = Q - \text{rate of energy emitted per square meter.}$$

We need a rule for the outgoing radiation. Here physics supplies one: an ideal emitter, called a *blackbody*, radiates at rate σT^4 per square meter. This is the Stefan–Boltzmann law, with $\sigma \approx 5.67 \times 10^{-8} \text{ W m}^{-2}\text{K}^{-4}$. It uses absolute temperature, so we must measure T in kelvin. Choosing to approximate the surface's radiation by this law gives

$$CT' = Q - \sigma T^4.$$

We now have a model! The temperature varies, while Q , C , and σ are constant. The absorbed sunlight is approximately $Q = 238 \text{ W m}^{-2}$: about 238 joules arrive each second per square meter, after reflection and averaging over the planet. This value comes from observations of Earth's energy budget. With a choice of C and a starting temperature, we can compute a solution.

At a constant temperature, $T' = 0$, so $Q = \sigma T^4$: incoming sunlight and outgoing radiation balance. Solving gives

$$T = \left(\frac{Q}{\sigma} \right)^{1/4} \approx 255 \text{ K.}$$

Above this temperature, radiation removes more energy than sunlight supplies, so the planet cools; below it, the planet warms. That explains where the plotted solutions are heading. Changing C changes how quickly they get there, but not the temperature where the two rates balance.

The prediction is about -18°C (0°F). Earth's average surface temperature is closer to 15°C (59°F), or 288 K. Our model is too cold, and adjusting C cannot fix it! What did we leave out?

4.5.2 Adding an atmosphere

We let radiation escape straight from the surface to space. But Earth's atmosphere absorbs some of that radiation and emits radiation of its own, including radiation back toward the surface. To include this interaction, we need to follow the atmosphere's temperature too, just as we followed both the coffee and the room.

Call the surface temperature $T_s(t)$ and the atmospheric temperature $T_a(t)$. We will describe the atmosphere as one layer with one temperature, ignoring differences with height.

Sunlight is concentrated at shorter wavelengths than the radiation emitted by Earth's surface, and the atmosphere absorbs them differently. For a simple model, suppose sunlight passes through the atmosphere and warms the surface, while all infrared radiation from the surface is absorbed by the atmosphere. Keep the same absorbed sunlight rate Q as before.

Absorbing radiation warms the atmosphere. The atmosphere also emits radiation of its own: downward toward the surface and upward toward space. Let us apply the same blackbody law to each side of our layer, giving an emission rate σT_a^4 in each direction. For now, these exchanges of radiation are the only ways our model gains and loses heat.

We can now write the surface's equation. It gains energy from sunlight at rate Q and from the atmosphere at rate σT_a^4 , and loses energy by emitting radiation at rate σT_s^4 . Choosing a constant surface heat capacity C_s gives

$$C_s T_s' = Q + \sigma T_a^4 - \sigma T_s^4.$$

The atmosphere gains energy from the surface at rate σT_s^4 . Its upward and downward emissions each remove energy at rate σT_a^4 . With its own constant heat capacity C_a , its equation is

$$C_a T_a' = \sigma T_s^4 - 2\sigma T_a^4.$$

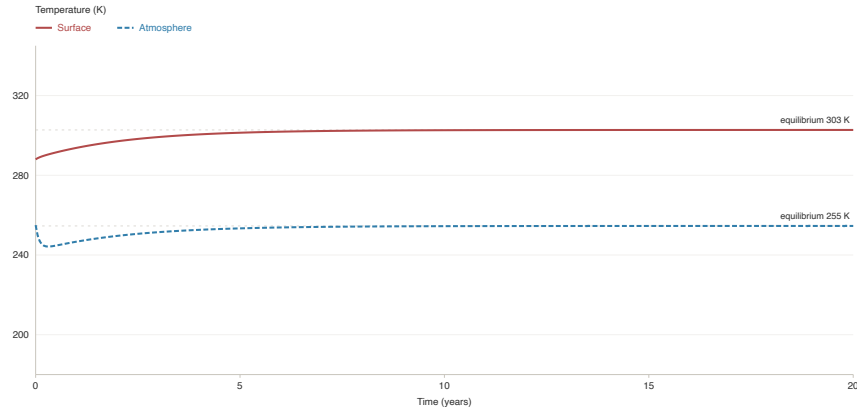


Figure 4.19 Begin with the surface at 288 K and the atmosphere at 255 K. Change either initial temperature or the surface heat capacity and follow both temperatures. The illustrative defaults are $C_s = 2 \times 10^8$ and $C_a = 2 \times 10^7 \text{ J m}^{-2}\text{K}^{-1}$; these values have not been fitted to an observed warming history.

The factor of two counts the two directions. Each temperature now affects the other's rate of change, giving us a system. Choosing both initial temperatures lets us explore it numerically as we did with the coffee and the room.

The temperatures settle at different values. Setting both derivatives to zero tells us that incoming and outgoing radiation balance for each:

$$Q + \sigma T_a^4 = \sigma T_s^4, \quad \sigma T_s^4 = 2\sigma T_a^4.$$

Substituting the second equation into the first gives $\sigma T_a^4 = Q$, and then $\sigma T_s^4 = 2Q$. Thus

$$T_a = \left(\frac{Q}{\sigma}\right)^{1/4} \approx 255 \text{ K}, \quad T_s = \left(\frac{2Q}{\sigma}\right)^{1/4} \approx 303 \text{ K}.$$

Adding the atmosphere has made the surface warmer: it now receives radiation from the atmosphere as well as sunlight. Unlike the coffee and room, these temperatures need not become equal, because energy keeps entering from the Sun and leaving for space. This simple greenhouse model shows how an atmosphere can maintain a warmer surface.

But 303 K is about 30°C (86°F), so our revised prediction is too warm. One assumption to revisit is that the atmosphere absorbs *all* radiation from the surface. In the exercises, you will let some pass through and see how the prediction changes.

We have repeated the same process: a choice of quantities and simple rules gave us an equation; its prediction led us to revise those choices. Adding a second temperature

captured an important interaction, and the new prediction suggests where to look next.

Further Reading

- Raymond Pearl’s 1927 survey, “The Growth of Populations”, reprints Carlson’s yeast measurements and places the logistic model in its early historical setting.
- Georgy Gause’s 1934 book *The Struggle for Existence* reports the *Paramecium* cultures used in the logistic data laboratory.
- The U.S. Census Bureau’s historical population tables provide the decennial counts used there from 1790 through 2020.
- Thomas W. Judson’s open *ODE Project* develops population, harvesting, predator–prey, epidemic, and mechanical models through activities and computation.
- Brian Rose’s open, notebook-based *Climate Laboratory Book* builds energy-balance and greenhouse models in more physical detail than we used here.
- Courchamp, Clutton-Brock, and Grenfell’s review “Inverse Density Dependence and the Allee Effect” explains the biological mechanisms behind low-population decline and the strong-Allee model.

PRACTICE · EXPLORATION · COMPUTATION

Problems

Modeling begins before an equation is solved. In every problem, name the state, say what each term means, and check signs or units before trusting a calculation. The explorations ask you to make new modeling choices; more than one answer may be defensible when the assumptions are stated clearly.

CHECK YOUR UNDERSTANDING

1. Read the assumptions from the equation

For each model, identify the state variables and parameters, describe the mechanism represented by every term, and give one limiting case which checks your interpretation.

- (a) $N' = 0.4N(1 - N/800)$.
- (b) $V' = 12 - 0.03V$.
- (c) $H' = aH - bHL, L' = -cL + dHL$.
- (d) $mx'' = -kx - bx'$.

2. Predicting the coffee's temperature

A cup of coffee sits in a room held at 20°C. Its temperature is initially 80°C, and ten minutes later it is 50°C. Assume that Newton's law of cooling describes its temperature throughout this interval and afterward.

- (a) Use the two measurements to determine k , taking time in minutes. State its units and write the resulting temperature function with no undetermined constants.
- (b) Predict the coffee's temperature twenty minutes after the first measurement.

3. Exponential first, logistic later

A bacterial culture begins with mass $N(0) = 12$ mg and initially has a per-capita growth rate of 0.7 h^{-1} .

- (a) Write the exponential model and predict the mass after six hours.
- (b) Suppose the environment has carrying capacity $K = 500$ mg. Write the logistic revision and use its exact solution to predict the mass after six hours.
- (c) At what population does the logistic model's *total* growth rate reach its maximum? At what population has its per-capita rate fallen to half of the low-density value? Explain why these are the same population in this model.

4. Harvesting a pond

For $P' = P(1 - P/200) - H$, answer the following without first using the quadratic formula.

- (a) Complete the square in the natural-growth term to show that its maximum is 50 at $P = 100$.
- (b) For $H = 42$, find both equilibrium populations and classify them from the sign of P' .
- (c) A manager wants an attracting equilibrium of at least 140 fish. Determine the largest constant quota the model permits. Then give one reason to choose a smaller quota in practice.

5. A tank as a balance law

A tank initially holds 300 L. Water enters at 18 L/min, and a drain removes 4% of the current volume per minute.

- (a) Derive the differential equation and check the units of both terms.
- (b) Find and classify the equilibrium volume without solving the equation.

(c) Solve the initial value problem. Use the formula to answer when the volume is within 5 L of equilibrium.

(d) The physical tank holds only 400 L. Does the mathematical solution ever overflow it? Explain how this feasibility check changes if the initial volume is 500 L.

6. What changes a climate equilibrium?

The absorbed sunlight in the climate appendix can be written as $Q = \frac{S_0}{4}(1 - \alpha)$, where S_0 is the incoming solar intensity and α is the fraction reflected. The factor $1/4$ averages sunlight intercepted by Earth's disk over its spherical surface.

Suppose absorbed sunlight increases by a constant amount F_0 . The equation becomes $CT' = \frac{S_0}{4}(1 - \alpha) - \sigma T^4 + F_0$.

(a) Derive the equilibrium temperature and show directly which parameters can move it.

(b) Without solving for $T(t)$, explain what doubling C changes and what it cannot change.

(c) Compute the new equilibrium when $S_0 = 1361 \text{ W m}^{-2}$, $\alpha = 0.30$, and $F_0 = 4 \text{ W m}^{-2}$. Compare it with the unforced equilibrium. Use $\sigma = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$.

7. Two histories, one moving point

Consider the chapter's predator-prey model with $\gamma = 1$:

$$H' = H(1 - L), \quad L' = L(H - 1).$$

(a) Calculate (H', L') at each state

$$(2, \frac{1}{2}), \quad (2, 2), \quad (\frac{1}{2}, 2), \quad (\frac{1}{2}, \frac{1}{2}).$$

Draw the four points in the phase plane and attach an arrow showing the velocity at each. Interpret each arrow in terms of growing or declining populations.

(b) In the chapter's figure showing time histories beside the phase-plane trajectory, locate a maximum of the hare population and the next maximum of the lynx population. Mark the corresponding points on a sketch of the trajectory. At which point is the motion vertical, and at which is it horizontal? Explain using the derivatives.

(c) At the states $(2, \frac{1}{2})$ and $(2, 2)$, the hare population is the same. Can an equation $H' = f(H)$ describe the hare's rate of change at both states? Explain what information is missing when we keep track of hares alone.

8. Transfers must cancel

Here we use fractions of the population rather than counts. If the constant total population is N , divide each count by N . The count-based infection term bSI then becomes a fraction-based term with coefficient $\beta = bN$; the recovery coefficient γ stays the same. Below, S, I, R denote the fractions.

$$S' = -\beta SI, \quad I' = \beta SI - \gamma I, \quad R' = \gamma I.$$

A quantity computed from a solution is **conserved** if it remains constant along that solution.

- (a) Differentiate $S + I + R$ and explain why every transfer cancels. Deduce that $S + I + R = 1$ for all times of existence if the initial fractions sum to 1.
- (b) Explain why knowing $S(t)$ and $I(t)$ determines $R(t)$. Write the two equations we could solve first, and the formula recovering R afterward.
- (c) At the state $(S, I, R) = (0.80, 0.15, 0.05)$ with $\beta = 0.6$ and $\gamma = 0.2$, compute (S', I', R') and interpret every sign.
- (d) Perform one simultaneous Euler step with $h = 0.1$. Verify that the three new fractions still sum to 1.
- (e) Add the three Euler update formulas at an arbitrary state. Show that their sum is unchanged by every simultaneous step, apart from rounding errors in a computer calculation.
- (f) A student updates S , then uses that *new* value of S while updating I . Explain why this is not the simultaneous Euler update and why exact cancellation may be lost.

9. Spring data determine parameters

A 0.50 kg mass on an undamped spring completes one oscillation every 1.2 s.

- (a) Use $\mathcal{T} = 2\pi\sqrt{m/k}$ to determine the spring stiffness k , including its units.
- (b) The mass begins 4 cm to the right of equilibrium with velocity -0.10 m/s. Find the constants in $x(t) = A \cos(\omega t) + B \sin(\omega t)$.
- (c) Find the first time at which the mass passes through equilibrium. Check that the sign of its velocity then agrees with the initial motion.
- (d) Let $v = x'$. Write the spring equation as a first-order system for (x, v) , using the parameters you found, and give its initial state. Explain why specifying the initial position alone would not determine the motion.

EXPLORATIONS

10. A threshold built one factor at a time

Suppose a population has trouble finding mates when its numbers are small. We want a model in which populations below 75 decline toward extinction, populations between 75 and 600 grow, and populations above 600 decline toward 600.

- (a) Draw the desired phase line. Where must the rate vanish?
- (b) Start with the logistic rate $rN(1 - N/600)$, where $r > 0$. On which interval does it have the wrong sign?
- (c) Find a dimensionless factor, linear in N , which is negative below 75, zero at 75, and positive above 75. Multiply the logistic rate by your factor and check all the required signs.
- (d) Describe the futures of populations starting at 40, 200, and 800. Explain why they cannot cross the equilibrium populations.
- (e) If time is measured in years, what units must r have?

11. Constant harvest or proportional harvest?

Replace the fixed quota in the trout model by removal proportional to the population: $P' = P(1 - P/200) - qP$.

- (a) Find and classify the equilibria as q varies.
- (b) Compare the behavior near $P = 0$ with constant harvesting. Which policy prevents the model from demanding removal from an empty pond?
- (c) Compare the constant quota $H = 32$ with a proportional policy which removes the same number of fish when $P = 160$. Choose q and compare their predictions when $P = 40$. What assumption about policy does each term encode?

12. The vanishing snowball

A spherical snowball melts at a rate proportional to its surface area. Its radius is 10 cm initially and 8 cm five minutes later. Assume its density remains constant.

- (a) Start with volume $V = \frac{4}{3}\pi r^3$ and surface area $A = 4\pi r^2$. Translate the melting assumption into an equation for V' , then use the chain rule to obtain an equation for r' .
- (b) Determine when the snowball disappears. Why is the radius linear in time even though both volume and area are nonlinear functions of radius?
- (c) Name one physical change near the end of melting which could make the model fail.

13. An oven that is still warming

An object is placed in an oven just as the oven is switched on. Both initially have temperature 20°C . Suppose the oven's temperature is prescribed by

$$A(t) = 200 - 180e^{-t/10}, \quad t \geq 0,$$

where time is measured in minutes and temperature in degrees Celsius. Model the object by a single temperature $T(t)$, and assume it obeys Newton's law with $k = 1/20 \text{ min}^{-1}$. Assume the object does not appreciably affect the oven's temperature.

- (a) Write the initial value problem for T . What does the model predict for $T'(0)$? Explain why that does not mean the object will remain at its initial temperature.
- (b) In Chapter 2 we subtracted the constant surrounding temperature. Try the corresponding substitution $u = T - A(t)$ here. Derive

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

What does the extra term represent?

- (c) Solve for $T(t)$ using an integrating factor. Check both the differential equation and the initial condition.
- (d) Show that $20 < T(t) < A(t) < 200$ for every $t > 0$, and find the limiting temperature of the object. Sketch the two temperature histories together.
- (e) The temperature difference $A(t) - T(t)$ starts at zero and eventually returns toward zero. When is it largest? Use the differential equation to explain why this is also when the object warms most rapidly.

14. A partially transparent greenhouse

Modify the climate appendix's greenhouse model by assuming the atmosphere absorbs a fraction ε of the surface's infrared radiation and transmits the rest directly to space, where $0 \leq \varepsilon \leq 1$. The layer emits $\varepsilon\sigma T_a^4$ upward and the same amount downward. Sunlight still passes through the atmosphere, supplying energy to the surface at rate Q per unit area.

Let $C_s, C_a > 0$ be the constant heat capacities per unit area of the surface and atmosphere.

(a) Write differential equations for T_s and T_a . Account for each incoming and outgoing contribution, and check that $\varepsilon = 1$ recovers the chapter's equations.

(b) Add the equations to find

$$\frac{d}{dt}(C_s T_s + C_a T_a).$$

Identify the two ways radiation escapes to space. Explain why energy exchanged between the surface and atmosphere disappears from this combined balance.

(c) Suppose $0 < \varepsilon \leq 1$. Set both derivatives equal to zero and solve for the equilibrium surface temperature in terms of Q, σ, ε . Do the heat capacities affect this temperature?

(d) Now set $\varepsilon = 0$ in the original differential equations. What happens to the atmospheric temperature? Find the surface equilibrium and compare it with the limit of your formula from part (c) as $\varepsilon \rightarrow 0$. Why does the equilibrium calculation no longer determine an atmospheric temperature?

(e) Find the value of ε which makes the surface equilibrium 288 K when the bare-planet equilibrium is 255 K. Explain what matching this one number does—and does not—establish about the model.

15. Vaccination as a transfer

Suppose susceptible individuals in an SIR model are vaccinated at a per-capita rate ν and move directly to the removed compartment.

(a) Modify all three equations. Make the new transfer appear once with a minus sign and once with a plus sign.

(b) Show that $S + I + R$ remains conserved.

(c) At what instantaneous rate does vaccination reduce the susceptible population when $S = 0.7$ and $\nu = 0.03$ per day? Explain why a constant term -0.03 would represent a different program.

PYTHON: MODELS AS EXPERIMENTS

16. Recover a growth law from Carlson's data

The arrays below contain the eighteen hourly yeast measurements used in the chapter.

```
import numpy as np
import matplotlib.pyplot as plt
```

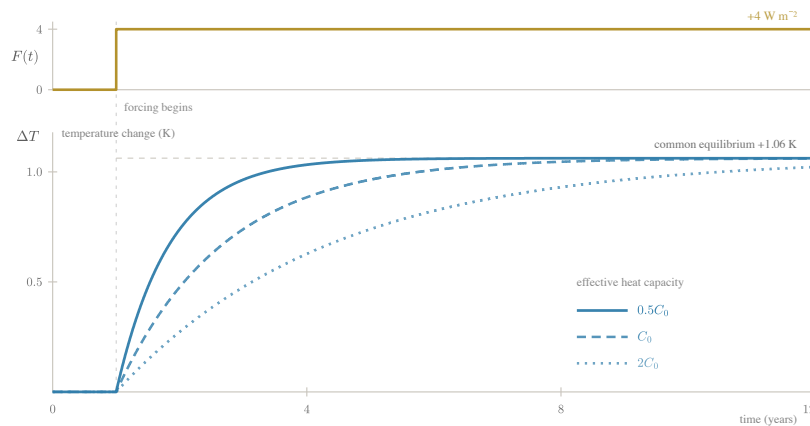


Figure 4.20 In this illustrative experiment, the absorbed sunlight increases by 4 W m^{-2} at year 1. The three curves use $0.5C_0$, C_0 , and $2C_0$, with $C_0 = 2 \times 10^8 \text{ J m}^{-2}\text{K}^{-1}$. They approach the same new equilibrium, about 1.06 K warmer, at different speeds. The following calculations let you explore other prescribed inputs.

```
t = np.arange(1, 19, dtype=float)
N = np.array([
    9.6, 18.3, 29.0, 47.2, 71.1, 119.1, 174.6, 257.3, 350.7,
    441.0, 513.3, 559.7, 594.8, 629.4, 640.8, 651.1, 655.9, 659.6
])
```

- Compute the midpoint population and midpoint estimate of per-capita growth for every consecutive pair. Plot the estimated rate against the midpoint population. Label the axes with units.
- Use `np.polyfit` to fit a line to this rate plot. Read estimates of r and K from the line's two intercepts.
- Plot the data together with the logistic solution using your estimates. Describe one feature the curve captures and one feature the two-parameter model leaves unexplained.

17. Force a climate model

Suppose the absorbed sunlight in the climate appendix changes by a prescribed amount $F(t)$, while reflection and the radiation law retain their chosen values. These are hypothetical experiments with the model; the inputs below are not measured solar histories.

Adapt your Euler function from Chapter 2 to the model $T' = (Q - \sigma T^4 + F(t))/C$.

Use $Q = 238 \text{ W m}^{-2}$ and start from the corresponding unforced equilibrium. Measure time in years, take $C = 8$ and $40 \text{ W yr m}^{-2} \text{ K}^{-1}$, use $h = 0.02$ year, and simulate fifty years. The right-hand side then returns kelvins per year.

- Write three forcing functions: a step from 0 to 4 W m^{-2} at year 1, a ramp increasing by 0.1 W m^{-2} per year, and a zero-mean sinusoid of amplitude 4 W m^{-2} and period ten years.

- (b) Plot the response to each forcing for two values of C . Keep every other parameter and the time grid fixed. Which conclusions are robust when C changes?
- (c) For the periodic forcing, measure the time difference between one forcing peak and the following temperature peak. How does this lag change with C ?

18. Does one recovery rate describe the outbreak?

These are the daily illness and recovery counts from the Tristan da Cunha record used in the chapter. Time is measured in days from the first observation. As in the text, use the number currently ill as an approximation to the infectious population I .

$I = [1, 1, 3, 7, 6, 10, 13, 13, 14, 14, 17, 10, 6, 6, 4, 3, 1, 1, 1, 1, 0]$

$R = [0, 0, 0, 0, 5, 7, 8, 13, 13, 16, 16, 24, 30, 31, 33, 34, 36, 36, 36, 36, 37]$

- (a) Starting from $R' = \gamma I$, show that a constant recovery rate must satisfy

$$\gamma = \frac{R(b) - R(a)}{\int_a^b I(t) dt}$$

on any interval with a nonzero denominator. Check the units.

- (b) Estimate the integral over $[0, 20]$ by adding trapezoid areas between consecutive observations. Use it to estimate γ and compare your answer with the value obtained by fitting both curves in the chapter.
- (c) Repeat the calculation separately for $[0, 10]$ and $[10, 20]$. If the observations followed the model exactly and we knew the integrals exactly, how would these estimates compare? How closely do your estimates agree?
- (d) Use your whole-record estimate $\hat{\gamma}$ to calculate

$$\hat{R}(n) = R(0) + \hat{\gamma} \sum_{j=0}^{n-1} \frac{I(j) + I(j+1)}{2}, \quad n = 1, \dots, 20.$$

Plot these reconstructed recovery counts alongside the observations. Why must the final counts agree? Do the intermediate counts agree?

- (e) Discuss what this comparison tells us about using one constant recovery rate. Consider the discrete observations, the integral approximation, and the assumption that being ill coincides with being infectious. Does disagreement by itself establish that the biological recovery rate changed?

19. Simulate transfers and audit conservation

Write a function returning the three SIR rates and use simultaneous vector Euler updates

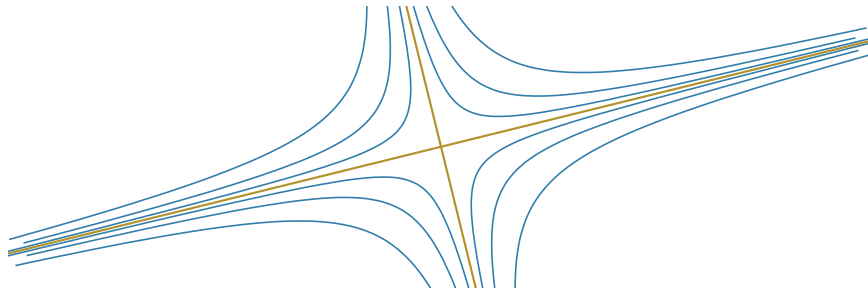
```
state = state + h * rates(state)
```

to simulate an epidemic from $(S, I, R) = (0.99, 0.01, 0)$. Use $\beta = 0.6 \text{ day}^{-1}$, $\gamma = 0.2 \text{ day}^{-1}$, $b = 0.05 \text{ day}$, and run for sixty days.

- (a) Plot all three compartments against time and plot $S + I + R$ on a second axis. Report the largest deviation of the total from 1.
- (b) Add the vaccination transfer from the previous exploration. Compare the peak infected fraction and the time of that peak for $\nu = 0, 0.01, 0.03 \text{ day}^{-1}$.
- (c) Deliberately implement the sequential-update mistake from the routine problem. Plot the total again. Explain how a conservation law became a test of your code rather than merely a fact about the equations.

PART II

Linear Equations and Flows



Equations and Solution Spaces

In the first part of the book, many of our solutions arrived with constants attached. Antidifferentiating $y' = 2t$ gave the whole family $y = t^2 + C$. Guessing and checking gave $y = Ce^t$ for $y' = y$, and $x = A \cos t + B \sin t$ for the spring $x'' = -x$. Even the system $f' = g$, $g' = -f$, which the very first page of the book asked you to solve just by thinking about it, has a whole family of solutions:

$$f(t) = A \cos t + B \sin t, \quad g(t) = -A \sin t + B \cos t.$$

Often, our next move was to use an initial condition to pick out the constants and study that one solution. Sometimes the formulas were already written in terms of initial data: the logistic solution in terms of the initial population N_0 , and the spring's motion in terms of its initial position and velocity. We could see the whole family, but the problem at hand often asked us to select one member.

Let's not do that this time. Before we impose any initial condition, an equation already has a whole collection of solutions: for the spring, one for every choice of A and B . What does that collection look like, taken all at once?

This asks us to study a set whose members are entire functions rather than numbers. One thing we would like to know is how many numbers it takes to single out one of its members. The formulas suggest an answer, one constant for $y' = y$ and two for the spring, but many equations have no formula at all, and we would like an answer that doesn't depend on finding one. We would also like to know what kind of set it is. If we already know a few of its members, can we build others from them?

That last question is one we can simply try.

5.1 Function Spaces and Solution Sets

Start with the spring. We already know two of its solutions, $\cos t$ and $\sin t$. Suppose u and v are *any* two solutions of $x'' = -x$, and add them. Differentiation respects sums, so

$$(u + v)'' = u'' + v'' = -u - v = -(u + v).$$

The sum is another solution! The same calculation works for a multiple cu , whose second derivative is $cu'' = -cu$. So starting from just $\cos t$ and $\sin t$, scaling and adding produces every function $A \cos t + B \sin t$ in the family we wrote down. We can build new spring solutions out of old ones as freely as we like.

Now try the same thing with $y' = y^2$, whose solutions taught us in chapter 3 that a perfectly smooth rule can still produce finite-time blowup. If u and v are two solutions, then

$$(u + v)' = u' + v' = u^2 + v^2.$$

But for $u + v$ to be another solution, the equation would require

$$(u + v)' = (u + v)^2 = u^2 + 2uv + v^2.$$

The extra term $2uv$ vanishes only where one of the two solutions does. Differentiation respects addition, but squaring does not: the sum of two solutions follows the wrong rule. Scaling fails the same way, since $(cu)' = cu^2$ while the equation demands $(cu)^2 = c^2u^2$.

Look at what we just did. We added two whole functions, multiplied a function by a number, and asked whether the result still solved the equation. That is the arithmetic of vectors. A list of three numbers counts as a vector because we can add lists and multiply them by scalars, and linear algebra applies the same idea to matrices, polynomials, and functions. It may have been a while since you thought of a function as a vector, but that is exactly the viewpoint we need for studying a whole collection of solutions at once.

A solution is always a solution *on an interval*. On an interval I , we need a function with enough derivatives to make every expression in the equation meaningful, satisfying the equation at every time in I . The interval matters. The function $y(t) = 1/(1-t)$ solves $y' = y^2$ on $(-\infty, 1)$, but cannot provide a solution through $t = 1$. A formula and the interval on which it satisfies the equation belong together.

Fix an open interval I . If f and g are real-valued functions on I , their sum and a scalar multiple are defined by

$$(f + g)(t) = f(t) + g(t), \quad (cf)(t) = cf(t) \quad \text{for every } t \in I.$$

These are **pointwise operations**: perform the arithmetic separately at each input to obtain a new function. The familiar algebraic rules hold because they hold at every input. For instance, $f + g = g + f$ because $f(t) + g(t) = g(t) + f(t)$ for every t . The zero vector is the function which is zero everywhere, and the additive inverse of f is the function $-f$. Thus real-valued functions on a fixed interval form a vector space.

For differential equations, we need functions we can differentiate. Write $C^k(I)$ for the space of real-valued functions on I with continuous derivatives through order k . In particular, $C^1(I)$ consists of continuously differentiable functions, while $C^2(I)$ requires two continuous derivatives. These are vector spaces too: adding or scaling functions preserves the required differentiability. We use $C(I)$ for the space of continuous functions. The linear algebra appendix reviews these spaces alongside the familiar spaces of numerical vectors.

These spaces are enormous! Even $C^k(I)$ for a fixed k contains all polynomials. The functions $1, t, t^2, \dots$ are linearly independent: a finite linear combination that vanishes throughout an interval is a zero polynomial, so all its coefficients vanish. Consequently, $C^k(I)$ contains arbitrarily large linearly independent collections. It is **infinite-dimensional**; no finite list of functions can form a basis for the whole space.

We will not need to find a basis for such a space. Its role is to hold the functions among which we are looking for solutions. For example, the solutions of $y' = y$ on I form the subset

$$\mathcal{S} = \{y \in C^1(I) : y'(t) = y(t) \text{ for every } t \in I\}.$$

This notation separates the two requirements. Membership in $C^1(I)$ makes differentiation available; the condition after the colon asks that the derivative agree with the function. We call $\mathcal{S}$ the **solution set**, and $C^1(I)$ its **ambient space**. For a second-order equation we can work in $C^2(I)$ instead. In each case, the equation selects a subset of a much larger space of functions.

Systems fit into the same picture. A solution of the system $f' = g, g' = -f$ from the introduction is a *pair* of functions, and neither is an answer on its own: the derivative of the first must equal the second, and the derivative of the second must equal the negative of the first. We can regard the pair as a single **vector-valued function**,

$$X(t) = \begin{pmatrix} f(t) \\ g(t) \end{pmatrix}.$$

At each time, $X(t)$ is a vector in $\mathbb{R}^2$. The whole function X records how that vector changes with time. Its derivative is computed component by component, so the system can be written

$$X'(t) = \begin{pmatrix} f'(t) \\ g'(t) \end{pmatrix} = \begin{pmatrix} g(t) \\ -f(t) \end{pmatrix}.$$

This is still the same pair of equations. Collecting the functions into a vector lets us speak of *one solution* while keeping track of all its components.

More generally, a first-order system for N unknown functions, with the equations solved for their derivatives, has the form

$$\begin{aligned} x_1' &= F_1(t, x_1, \dots, x_N), \\ &\vdots \\ x_N' &= F_N(t, x_1, \dots, x_N). \end{aligned}$$

Each rate may depend on every current value. Collecting the unknowns into $X = (x_1, \dots, x_N)$ and the known rules into $F = (F_1, \dots, F_N)$ gives the compact expression $X' = F(t, X)$. The functions F_i may contain products, squares, or other nonlinear expressions, as they did in our population models. They may also depend explicitly on time. Vector notation accommodates all of these possibilities.

For systems, use $C^k(I, \mathbb{R}^N)$, the space of vector-valued functions whose components belong to $C^k(I)$. Addition, scaling, and differentiation act component by component. Notice that N counts the components of the value $X(t)$ at one time; it is not the dimension of this function space. Even one component can be any of infinitely many linearly independent functions.

5.1.1 Parameters for a Family of Functions

How large is the subset selected by an equation? We can begin with the families we already know. On $\mathbb{R}$, every solution of $y' = y$ has the form $y_C(t) = Ce^t$. The assignment

$$\mathbb{R} \longrightarrow C^1(\mathbb{R}), \quad C \longmapsto y_C$$

takes a number and returns an entire function. Changing C moves through the solution set. Every solution occurs, and two different values of C give different solutions: just evaluate at $t = 0$ to recover $y_C(0) = C$. Thus one real number gives a coordinate for each member of this family.

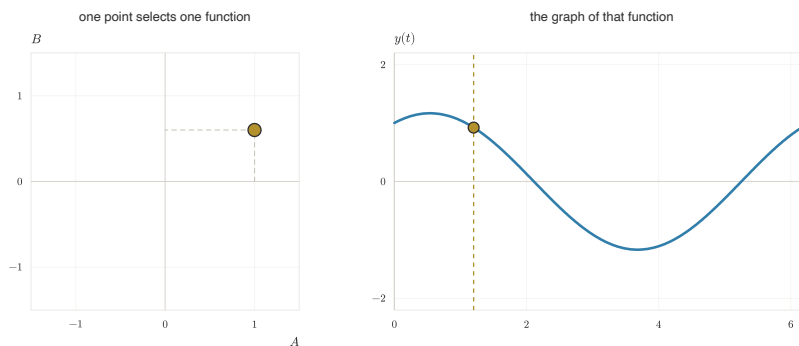


Figure 5.1 Choose a point (A, B) on the left to select the entire function $y(t) = A \cos t + B \sin t$ on the right. Moving the time marker along that graph leaves (A, B) fixed; moving (A, B) chooses a different function. The left panel gives coordinates for functions, while the right panel shows the values of one function over time.

Geometrically, these functions form a line through zero in function space. They are precisely the scalar multiples of the nonzero vector e^t . This line looks quite different from the collection of exponential graphs we drew earlier. Each entire graph represents just one point of the line! Moving along the graph means varying t within one function. Moving along the line in function space means varying C to choose a different function.

The spring gives us a family with two parameters:

$$y_{A,B}(t) = A \cos t + B \sin t.$$

Its parameters can be recovered from $y_{A,B}(0) = A$ and $y'_{A,B}(0) = B$, so different pairs give different functions. These functions form the plane spanned by the vectors $\cos t$ and $\sin t$ in $C^2(\mathbb{R})$. Their independence is easy to check: if $A \cos t + B \sin t$ is the zero function, its value and derivative at zero force $A = B = 0$. Choosing a point (A, B) in the parameter plane therefore chooses one point in this plane of functions.

This is how a differential equation can leave us with finitely many choices even though we began in an infinite-dimensional space. We have already verified that this whole plane consists of spring solutions. Proving that it contains *every* spring solution is a further claim; the existence-and-uniqueness theory will help us settle claims of this kind without guessing what other functions might work.

5.1.2 Linear Differential Equations

A subset of a vector space is a **vector subspace** if it contains zero and is closed under linear combinations. Thus, for a solution set to be a vector space, the zero function must solve the equation, and whenever u and v solve it, $au + bv$ must solve it for all real numbers a, b . The exponential line and the plane of spring solutions have this property. Can we recognize it from the equations themselves?

Differentiation already respects the arithmetic we need:

$$(au + bv)' = au' + bv'.$$

In the language of linear algebra, differentiation is a **linear map** from $C^1(I)$ to $C(I)$. Its input is a function, and its output is another function. More generally, a map L between vector spaces is linear when $L(au + bv) = aL(u) + bL(v)$. We often call a map acting on functions an **operator**.

Consider the operator

$$L : C^2(I) \longrightarrow C(I), \quad L[y] = y'' + y.$$

It takes a function, differentiates it twice, and adds the original function. Both operations respect linear combinations, so

$$L[au + bv] = (au + bv)'' + (au + bv) = a(u'' + u) + b(v'' + v) = aL[u] + bL[v].$$

Our spring equation $y'' = -y$ is exactly the equation $L[y] = 0$. Its solutions are the functions the operator sends to zero: the **kernel** of L . Recall why a kernel is a subspace. It contains zero, and if $L[u] = L[v] = 0$, then $L[au + bv] = a \cdot 0 + b \cdot 0 = 0$. We have proved that the entire solution set is a vector space without first finding all its members!

This reasoning works for much more than the spring. Multiplication by a fixed function of time is also linear: for any known coefficient $p(t)$, $p(t)(au(t) + bv(t)) = ap(t)u(t) + bp(t)v(t)$. Consequently, an operator of the form

$$L[y] = a_n(t)y^{(n)} + \dots + a_1(t)y' + a_0(t)y$$

is linear. With continuous coefficients, it maps $C^n(I)$ to $C(I)$. The coefficients can change with time; what matters is that they are fixed functions, independent of the unknown y .

Definition 5.1 (Linear differential equations). A differential equation is **linear** if it can be written

$$L[u] = g,$$

where g is a known function and L is a **linear differential operator**: a linear map built from differentiating the unknown and multiplying by known functions of t . We call g the **forcing**. The equation is **homogeneous** when $g = 0$ and **forced** otherwise. An equation that cannot be written in this form is **nonlinear**.

The most familiar case is a scalar equation of order n ,

$$a_n(t)y^{(n)} + \dots + a_1(t)y' + a_0(t)y = g(t),$$

whose operator is the one we just built. For example, $y' = y$ is the homogeneous linear equation $y' - y = 0$, and $y' = ty$ is homogeneous and linear too: the coefficient t varies, but the dependence on the unknown is still linear. Squaring y or y' would destroy the linearity of the corresponding operator.

For a system, consider

$$x' = 2y, \quad y' = -3x.$$

Its operator sends the pair $X = (x, y)$ to $L[X] = (x' - 2y, y' + 3x)$. This is a linear map from $C^1(I, \mathbb{R}^2)$ to $C(I, \mathbb{R}^2)$, and once again the solution set is its kernel. To give this space some concrete members, let $\omega = \sqrt{6}$. Direct differentiation verifies the family

$$\begin{aligned} x(t) &= A \cos(\omega t) + \frac{2B}{\omega} \sin(\omega t), \\ y(t) &= B \cos(\omega t) - \frac{3A}{\omega} \sin(\omega t). \end{aligned}$$

Here $x(0) = A$ and $y(0) = B$. In checking the second equation, the identity $\omega^2 = 6$ makes the coefficients agree. Adding two of these vector-valued solutions adds their parameters; multiplying a solution by a scalar multiplies both parameters by that scalar.

The same kernel argument applies to all these equations, regardless of their order or number of components.

Theorem 5.2 (Superposition). The solutions of a homogeneous linear differential equation or system on a fixed interval form a vector space. In particular, if u and v are solutions, then $au + bv$ is a solution for every pair of real numbers a, b .

The proof is the kernel calculation above. The word *linear* describes this arithmetic of functions, not the shapes of their graphs. Our solutions can grow exponentially or oscillate while still belonging to a vector space.

5.1.3 Affine Solution Sets

What about a linear equation with a nonzero right-hand side? We have been solving these since antidifferentiation. For example, $y' = 2t$ has the full solution family

$$y(t) = t^2 + C.$$

This is a line in function space too, but it does not pass through zero: the zero function does not have derivative $2t$. Adding two solutions gives $2t^2 + C_1 + C_2$, whose derivative is $4t$, so the sum leaves the solution set. Yet subtracting two solutions gives the constant $C_1 - C_2$. The direction of the line is the space of constant functions, which solves the associated homogeneous equation $y' = 0$.

Recall that an **affine space** inside a vector space is a translate $p + V = \{p + v : v \in V\}$ of a vector subspace V . The point p locates the translate, while V describes its directions. Our family $t^2 + C$ is the translate of the constant functions by the particular function t^2 . The constants of integration were describing an affine space all along.

The same relationship holds for every solvable linear equation. Suppose $L[u] = g$, and we have found one particular solution u_p . If u is any other solution, linearity gives

$$L[u - u_p] = L[u] - L[u_p] = g - g = 0.$$

Thus $u - u_p$ belongs to $\ker L$. Conversely, adding any homogeneous solution $b \in \ker L$ to u_p gives $L[u_p + b] = g + 0 = g$. These two calculations establish the whole solution set:

Theorem 5.3 (The solution set of a linear equation). If L is a linear differential operator and u_p solves $L[u] = g$ on an interval I , then all solutions on I form the affine space

$$u_p + \ker L.$$

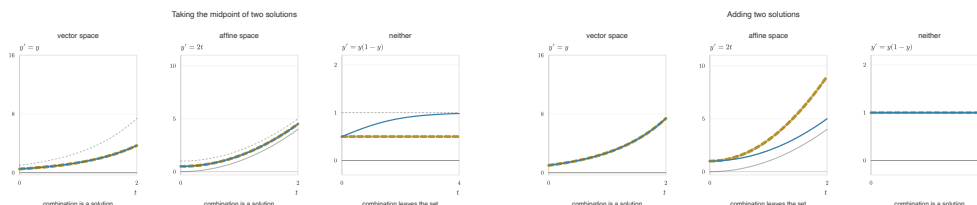


Figure 5.2 Add two solutions, or take their midpoint, and compare the result with the solution starting at the same value. The gold dashed curve is the combination; the blue curve is the solution. For $y' = y$ they agree for both operations. For the forced equation $y' = 2t$ the sum fails, but the midpoint works, since it is still a particular solution plus a homogeneous one. For the logistic equation, the constant solutions $u = 0$ and $v = 1$ have the midpoint $1/2$, which is not a solution. Try other initial values too: a combination that happens to work for one pair need not work for every pair.

In words, every solution is one particular solution plus a solution of the homogeneous equation.

When $g = 0$, we can choose $u_p = 0$ and recover the vector space $\ker L$. When g is not the zero function, zero is not a solution, so the affine solution set is not a vector subspace. In either case, this description assumes that a particular solution exists; it does not yet prove existence for arbitrary equations or initial data.

5.1.4 Nonlinear Solution Sets

For a linear equation, we now know which arithmetic preserves solutions: linear combinations when the equation is homogeneous, and a particular solution plus a homogeneous one when it is forced. The calculation for $y' = y^2$ showed that addition and scaling need not preserve solutions of a nonlinear equation. The logistic equation $y' = y(1 - y)$ has the same difficulty. For general nonlinear equations, we have no superposition principle: knowing several solutions does not, by itself, give a way to construct the solution from another initial state.

For a second-order comparison, place the spring $x'' = -x$ beside

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

Here $x(t) = t$ and $x(t) = -t$ are two solutions: each has second derivative zero and first derivative with square one. Their sum is the zero function, whose derivatives both vanish, so substituting it would require $0 = 1$. Doubling fails too: $x = 2t$ has $x'' = 0$, but $1 - (x')^2 = -3$. A nonlinear dependence on a derivative breaks the arithmetic just as a nonlinear dependence on the function does.

Yet this equation has plenty of solutions. Here is a whole family of them:

$$x_{a,b}(t) = a + \log(\cosh t + b \sinh t).$$

We do not need to discover this formula to make use of it; we can check it. Write $b(t) = \cosh t + b \sinh t$, so that $b'' = b$. Wherever $b > 0$, differentiating the proposed solution gives

$$x'_{a,b} = \frac{b'}{b}, \quad x''_{a,b} = \frac{b''}{b} - \left(\frac{b'}{b}\right)^2 = 1 - (x'_{a,b})^2.$$

The formula works. And since $b(0) = 1$ and $b'(0) = b$, its parameters are exactly the initial data:

$$x_{a,b}(0) = a, \quad x'_{a,b}(0) = b.$$

We can choose initial position and initial velocity independently, just as we could for the spring. But they enter this family differently. Changing a translates the graph vertically, while changing b changes the function inside the logarithm. In fact, adding any constant to a solution preserves this equation, since its derivatives do not change. So we can obtain some new solutions from a known one. What we lack is the freedom to combine solutions by arbitrary addition and scaling. We still have two independent initial-data parameters, even though they do not enter as the coefficients of two fixed solutions.

For any real a, b , the formula is defined on an interval around zero, because $b(0) = 1$. If $|b| \leq 1$, it is defined for all real t :

$$b(t) = \frac{1+b}{2}e^t + \frac{1-b}{2}e^{-t} > 0.$$

In particular, $b = 1$ gives $x = a + t$, and $b = -1$ gives $x = a - t$, recovering our earlier solutions. For $|b| > 1$, b vanishes at a finite time on one side of zero; we use the interval containing zero on which it is positive. The parameters can affect how long a solution exists, as well as the shape of its graph.

The spring family and the logarithmic family each use two numbers, the initial position and velocity, to pick out one member. Their different arithmetic does not change that count. So far, though, every count like this has come from a formula. How many numbers pick out a solution when there is no formula to read them from? Chapter 3 answered this for a single first-order equation $y' = F(t, y)$: under the theorem's hypotheses, one initial value determines a unique local solution, even when we

cannot find a formula. We would like the same kind of answer for every equation we have met.

5.2 A Common Formulation: First-Order Systems

We already found a connection between higher-order equations and systems when modeling the spring. Its equation $x'' = -x$ tells us the acceleration from the position, but position alone does not determine the motion. Two masses can pass through the same position with different velocities. To describe how either one moves, we need to keep track of both quantities.

Give velocity its own name, $v = x'$. We can now ask for the first derivative of each quantity we are tracking. The derivative of position is velocity, and the derivative of velocity is the acceleration supplied by the equation:

$$\begin{aligned}x' &= v, \\v' &= -x.\end{aligned}$$

We have replaced one second-order equation with two first-order equations. The new variable records information the original problem already needed. The initial conditions $x(t_0) = x_0$ and $x'(t_0) = v_0$ become the single vector condition $X(t_0) = (x_0, v_0)$ for $X = (x, v)$.

Nothing about this construction required the spring equation to be linear. Apply it to our nonlinear example $x'' = 1 - (x')^2$. With the same choice $v = x'$, we obtain

$$\begin{aligned}x' &= v, \\v' &= 1 - v^2.\end{aligned}$$

The second equation now involves only v , so we can solve it by separation and then integrate $x' = v$. The solutions $x = t$ and $x = -t$ correspond to the system solutions $(x, v) = (t, 1)$ and $(x, v) = (-t, -1)$. In the system, we can see why those motions work: the velocities 1 and -1 make $v' = 0$ and remain constant.

More generally, consider a second-order equation

$$x'' = G(t, x, x').$$

The same substitution produces $x' = v$, $v' = G(t, x, v)$. We should check that these descriptions really have the same solutions. If x solves the scalar equation on an

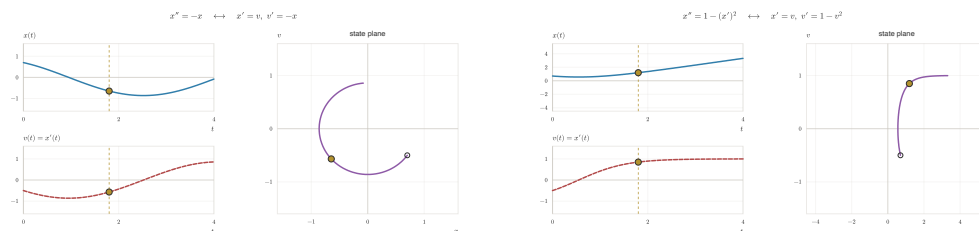


Figure 5.3 Follow one solution in its two descriptions. The graphs of $x(t)$ and $v(t) = x'(t)$ share a time marker; their values give the gold point $(x(t), v(t))$ in the state plane. Move time to trace the same solution, or change either initial value to choose another. Switching equations shows the same correspondence for $x'' = -x$ and $x'' = 1 - (x')^2$. Here a point in the state plane records a solution at one instant; the entire state curve records its motion over the displayed interval.

interval I , define $v = x'$. Then $x' = v$ by definition, and $v' = x'' = G(t, x, v)$ by the original equation. Thus (x, x') solves the system on I .

Conversely, suppose a pair (x, v) solves the system. Its first equation forces $v = x'$. Differentiating gives $x'' = v' = G(t, x, v) = G(t, x, x')$, so its first component solves the original scalar equation. These constructions undo each other. We have a one-to-one correspondence between the two solution sets, including the solutions with any specified initial position and velocity. No existence or uniqueness theorem was needed to establish this correspondence; it follows directly from the equations.

5.2.1 Recording the Derivatives

What happens for a third-order equation? Recording y and y' is no longer enough to write down their rates: the derivative of y' is y'' , which the equation has not yet determined from those two values. So record y'' too. Its derivative is y''' , and that is what the third-order equation supplies. The construction stops when we have recorded all derivatives below the highest one.

For an equation of order $n \geq 2$ written as

$$y^{(n)} = G(t, y, y', \dots, y^{(n-1)}),$$

introduce the coordinates

$$x_1 = y, \quad x_2 = y', \quad \dots, \quad x_n = y^{(n-1)}.$$

Differentiating each coordinate gives the first-order system

$$\begin{aligned}
x_1' &= x_2, \\
x_2' &= x_3, \\
&\vdots \\
x_{n-1}' &= x_n, \\
x_n' &= G(t, x_1, \dots, x_n).
\end{aligned}$$

The first $n - 1$ equations record the meaning of our new coordinates. The last carries the original differential equation. Thus the state is $X = (y, y', \dots, y^{(n-1)})$, and prescribing its value at t_0 specifies exactly the usual n initial values for the scalar equation.

The equivalence argument works just as it did for order two. A scalar solution produces a system solution by collecting its derivatives. Starting with a system solution, the first equation gives $x_2 = x_1'$, the next gives $x_3 = x_1''$, and continuing gives $x_n = x_1^{(n-1)}$. The last equation then says that $y = x_1$ solves the original equation. In particular, the coordinates of a system solution cannot be chosen as unrelated functions: the derivative relations tie them together.

For linear equations, this conversion also preserves linearity. If $a_n(t)$ is nonzero, divide $a_n y^{(n)} + \dots + a_1 y' + a_0 y = g$ by a_n to obtain the last system equation

$$x_n' = -\frac{a_0}{a_n}x_1 - \frac{a_1}{a_n}x_2 - \dots - \frac{a_{n-1}}{a_n}x_n + \frac{g}{a_n}.$$

Together with the derivative chain, this is a linear system, homogeneous when $g = 0$. The correspondence $y \mapsto (y, y', \dots, y^{(n-1)})$ itself respects addition and scaling. Our two descriptions therefore agree about the linear or affine structure of their solution sets too.

Every first-order linear system has this shape. Collecting the coefficients into a matrix, it reads

$$X' = A(t)X + g(t),$$

where $A(t)$ is an $N \times N$ matrix of known functions and $g(t)$ is a known vector-valued function. In the language of theorem 5.1, its operator is $L[X] = X' - A(t)X$ and its forcing is g . The converted scalar equation is one example; the system $x' = 2y$, $y' = -3x$ from the previous section is another, with

$$A = \begin{pmatrix} 0 & 2 \\ -3 & 0 \end{pmatrix}.$$

This is the form the next chapter starts from.

5.2.2 Coupled Equations and the State

The same construction applies when several equations involve higher derivatives. Suppose $q = (q_1, \dots, q_m)$ is a vector of unknown functions satisfying

$$q'' = G(t, q, q').$$

Record the vector $v = q'$ along with q . We obtain $q' = v$ and $v' = G(t, q, v)$, a first-order system with $2m$ scalar coordinates. Each component contributes its value and its first derivative to the state. Equations with different orders work the same way: for each unknown, record its derivatives up to one below the order of the equation that determines its highest derivative. If those orders are $n_1, \dots, n_m$, the resulting state has $N = n_1 + \dots + n_m$ coordinates. This assumes the original system has been solved for those highest derivatives in terms of time and the recorded quantities.

Equations solved for their highest derivatives in this way are said to be in **normal form**, and the conversion applies to them. A scalar linear equation can be put in normal form wherever its leading coefficient $a_n(t)$ is nonzero; at a zero of a_n we cannot divide.

Within this scope, all the equations now have the same shape:

$$X' = F(t, X), \quad X(t_0) = X_0, \quad X \in \mathbb{R}^N.$$

Here X is exactly the state of theorem 1.3: the current values and derivatives we must record as initial data. The differential equation supplies their rates of change. The starting time t_0 is also part of the initial value problem: when F depends explicitly on time, the same state prescribed at different times can have different rates.

Could we reduce everything to one scalar first-order equation instead? Even the spring shows what would be lost. At a fixed time and position, we may choose different velocities, while a scalar equation $x' = f(t, x)$ assigns just one velocity. Recording a vector allows us to retain both initial choices. We have reduced the order by keeping more coordinates; we have not removed the information in those coordinates.

This leaves us free to use whichever description is convenient. The scalar equation $x'' = -x$ makes the relationship between acceleration and position immediate, while its system makes the full state explicit. A theorem about first-order systems can now apply to both descriptions. The next question is whether specifying X_0 actually determines a solution: does a history begin at every initial state, and can more than one history begin there?

5.3 Existence, Uniqueness, and Initial Data

For one first-order equation, we already have a theorem that answers this question. Continuity of the rate function and its derivative with respect to the unknown gives local existence and uniqueness. For a system, the corresponding hypotheses ask us to check how each component of the rate depends on each component of the state.

Write $F = (F_1, \dots, F_N)$ and $X = (x_1, \dots, x_N)$. The partial derivative $\partial F_i / \partial x_j$ measures how the i th rate changes as we vary the j th state coordinate, holding time and the other coordinates fixed. There are N^2 such derivatives. The theorem has the same form in every dimension.

Theorem 5.4 (Local existence and uniqueness for systems). Suppose $F(t, X)$ and all partial derivatives $\partial F_i / \partial x_j$ are continuous on an open set $D \subseteq \mathbb{R} \times \mathbb{R}^N$. For every $(t_0, X_0) \in D$, the initial value problem

$$X' = F(t, X), \quad X(t_0) = X_0$$

has a unique solution on a sufficiently small open interval containing t_0 , with $(t, X(t)) \in D$.

We will use the system theorem without proving it. Our scalar theorem is its $N = 1$ case; the conversion of higher-order equations tells us how to apply the same result far beyond that case. The hypotheses are sufficient conditions, not a claim that every equation failing this test must fail to have a unique solution.

For example, our nonlinear second-order equation became

$$x' = v, \quad v' = 1 - v^2.$$

Here $F_1(x, v) = v$ and $F_2(x, v) = 1 - v^2$, with state derivatives

$$\frac{\partial F_1}{\partial x} = 0, \quad \frac{\partial F_1}{\partial v} = 1, \quad \frac{\partial F_2}{\partial x} = 0, \quad \frac{\partial F_2}{\partial v} = -2v.$$

Everything is continuous everywhere. Thus any choice of $x(t_0)$ and $v(t_0)$ gives a unique local system solution. By the correspondence we proved, any choice of $x(t_0)$ and $x'(t_0)$ gives a unique local solution of the second-order equation. We have answered the question for every initial position and velocity without solving the equation for any of them.

The same reasoning applies to an order- n equation in normal form. If G and its partial derivatives with respect to $y, y', \dots, y^{(n-1)}$ are continuous near the prescribed

data, then the converted system satisfies the theorem's hypotheses. Each derivative-chain equation has a coordinate as its right-hand side; only the last equation requires us to check G . The n prescribed initial values therefore determine a unique local scalar solution. Coupled higher-order equations are handled by the same check on their combined state.

Uniqueness also settles a question we left open about the spring. We verified that $x(t) = A \cos t + B \sin t$ solves $x'' = -x$. Given any solution on an interval containing zero, take $A = x(0)$ and $B = x'(0)$. The displayed formula and the given solution then have the same initial state, so uniqueness forces them to agree near zero. They agree throughout their common interval: if agreement could stop at some time inside that interval, their states there would agree by continuity, and local uniqueness at that time would extend the agreement. This is the same continuation argument we used for scalar first-order equations. There are no missing spring solutions! The plane of functions we found is the whole solution set on $\mathbb{R}$, and its restrictions give the solutions on smaller intervals containing zero.

Existence is still a local guarantee. Even when the hypotheses hold everywhere, a solution need not exist for all time. We already know this from $y' = y^2$: the solution with $y(0) = a \neq 0$ is $y(t) = a/(1 - at)$ on its interval through zero, and it blows up at $t = 1/a$. Different initial values can therefore give different intervals of existence. The theorem guarantees a history near the initial time, not that we can follow it indefinitely. In fact, only the zero solution of $y' = y^2$ exists on all of $\mathbb{R}$.

Uniqueness matters just as much. The functions $y = 0$ and $y = t^3$ from Chapter 3 solve $y' = 3y^{2/3}$ with the same value at zero. For that equation, an initial value cannot distinguish the two histories. When we use initial data to identify a solution, we are using a theorem about the equation, not merely counting the constants in a formula.

5.3.1 Initial Data as Parameters

How can finitely many numbers specify an entire function? A general function can take values independently at many different times. A solution of a differential equation has much less freedom: its values must fit together according to the equation. Existence and uniqueness tell us just how much freedom remains.

Fix an initial time t_0 . Under the hypotheses of our theorem, each admissible initial state ξ determines a unique local solution X_ξ , with $X_\xi(t_0) = \xi$. Existence supplies a solution near t_0 . Uniqueness means that any two solutions with those data agree throughout their common interval containing t_0 . Going in the other direction is easy: given a solution defined near t_0 , evaluate it there to recover the state that selected it.

The initial state therefore works as a set of coordinates for local solutions. For an equation whose state has N coordinates, N numbers recorded at a single time deter-

mine the solution near that time. This is remarkable information to have without a solution formula: we know how many initial values specify a solution before we have found a single one.

This answers the question we asked of our formulas. The count comes from the equation, through its state, not from how many constants happen to appear in a formula. We could write the solutions of $y' = y$ as $(A + B)e^t$, but only the sum $A + B$ matters. The state of $y' = y$ has one coordinate, and one number is what it takes.

The local qualification matters. Different initial states can give solutions on different intervals, as the blowup of $y' = y^2$ showed. Restricting a solution to a smaller interval containing t_0 also preserves its initial data. The data determine how the solution behaves where it is defined; they do not prescribe which interval we write it on.

For a higher-order scalar equation, remember what these coordinates mean. We first regard its solution as the vector-valued solution $X = (y, y', \dots, y^{(n-1)})$. Evaluation then records

$$(y(t_0), y'(t_0), \dots, y^{(n-1)}(t_0)).$$

So the order- n scalar equation has a locally n -parameter solution family under the hypotheses above. The system formulation explains why derivative values count as parameters in exactly the same way as the component values of a first-order system.

Return now to the nonlinear family we checked earlier:

$$x_{a,b}(t) = a + \log(\cosh t + b \sinh t), \quad x'' = 1 - (x')^2.$$

We verified that each function solves the equation wherever the logarithm's argument is positive. Have we found all the solutions near zero, or could there be others that our formula missed? Take any solution x on an interval containing zero and record $a = x(0)$, $b = x'(0)$. The function $x_{a,b}$ exists near zero and has exactly those initial data. Uniqueness forces $x = x_{a,b}$ near zero, and the continuation argument extends their agreement throughout their common interval. There are no missing local solutions. We did not need a systematic method for solving the equation to establish that our verified family was complete!

We began this chapter by asking two questions about the collection of all solutions of an equation. What kind of set is it? Homogeneous linear equations give vector spaces, and fixed forcing gives translates of those spaces whenever a particular solution exists. Nonlinear equations generally have neither structure, so the same reconstruction by linear algebra is not guaranteed. How many initial values specify a local solution? As many as there are coordinates in the state, for every equation our theorem covers.

The spring and $x'' = 1 - (x')^2$ show that these answers are separate facts. Both need two numbers. For the spring, those numbers select the combination $x(0) \cos t + x'(0) \sin t$; for the other equation, they enter the logarithmic formula in a different way. Knowing how many numbers we need does not tell us what arithmetic the solutions allow.

For homogeneous linear equations, these two kinds of information fit together remarkably well. Evaluating at t_0 respects addition and scaling, and those operations keep us inside the solution set. In the next chapter we put the two together: the initial state becomes a set of linear coordinates for whole solutions, which will tell us the dimension of the solution space, give us bases of solutions, and describe evolution in time by a matrix.

PRACTICE · EXPLORATION · COMPUTATION

Problems

The objects in this chapter live in several different spaces. A state is a list of numbers, a solution is a function, and a solution set is a collection of functions. Keep track of which object you are working with; many of the calculations below become more useful once we know what they say about the whole collection.

CHECK YOUR UNDERSTANDING

1. A point, a graph, and a function

Consider the family $x_{A,B}(t) = A \cos t + B \sin t$ of solutions to $x'' = -x$, and the corresponding system state $X = (x, x')$.

- (a) Choose $(A, B) = (1, 0)$. Write the resulting scalar function $x(t)$ and vector-valued function $X(t)$. Which of these objects is a point in $C^2(\mathbb{R})$? Which is a point in $C^1(\mathbb{R}, \mathbb{R}^2)$?
- (b) At $t = \pi/2$, record three points: the parameter point (A, B) , the graph point $(t, x(t))$, and the state point $(x(t), x'(t))$. Explain what information each point records.
- (c) Hold (A, B) fixed and vary t . Which of those points moves? Then hold $t = \pi/2$ fixed and vary A . Are you moving along one solution or selecting different solutions? Use figure 5.1 to check your interpretation.

2. How many independent parameters?

- (a) The formulas $y_C(t) = Ce^t$ and $z_{A,B}(t) = (A + B)e^t$ describe families in $C^1(\mathbb{R})$. Do they describe the same set of functions? Give two different pairs (A, B) that describe the same function, and explain how many independent parameters the set has.
- (b) Show that if $A \cos t + B \sin t = C \cos t + D \sin t$ for every t , then $A = C$ and $B = D$. Which two measurements of a function recover its parameters?
- (c) For $|a| < 1$, consider $y_a(t) = a/(1-at)$ on $(-1, 1)$. Recover a from the function and show that different parameters give different functions. Does having one independent parameter imply that this family is a vector subspace? Test a specific pair or a scalar multiple.
- (d) In $y(t) = Ce^{rt}$, explain the different roles of C and r when we are describing solutions of the fixed equation $y' = 2y$.

3. Recognizing linear structure

For each equation in parts (a)–(d), decide whether it is linear as written. If it is, identify its linear differential operator and forcing, and decide whether it is homogeneous. State what this tells you about its nonempty solution set on a common interval. If it is nonlinear, identify the term that prevents the linear-operator argument from applying.

- (a) $y'' + t^2 y' - \sin(t)y = 0$.
- (b) $y' + 3y = e^{-t}$.
- (c) $y'' + y^2 = 0$.
- (d) $x' = tx - 2y, y' = x + (1 + t^2)y$.
- (e) For $y' = y(1 - y)$, use the constant solutions zero and one to prove that the solution set on $\mathbb{R}$ is not affine. Why would merely finding the zero solution be insufficient to decide its structure?

4. Arithmetic in an affine solution set

Let L be a linear differential operator, and suppose u and v solve $L[y] = g$ on the same interval. Assume g is not the zero function.

- (a) Calculate $L[u - v]$ and $L[u + v]$. Which equation does each function solve? Explain why the failure of addition here does not mean L is nonlinear.
- (b) For real constants a, b , show that $au + bv$ solves $L[y] = g$ if and only if $a + b = 1$. What changes when $g = 0$?
- (c) Suppose u_p is one solution of $L[y] = g$. Prove that every solution is $u_p + h$ for some $h \in \ker L$, and that every such function is a solution.
- (d) Apply the description to $y' = 2t$. Identify a particular solution and the homogeneous solution space. Choose two distinct solutions and calculate their difference and their midpoint.

5. Write the state, then recover the equation

In each conversion, name every state coordinate and specify how the initial data translate. Do not solve the equations.

(a) Convert $x'' = 1 - (x')^2$, with $x(0) = a$ and $x'(0) = b$, to a first-order system. Starting instead from a solution of your system, show directly that its first component solves the scalar equation.

(b) Convert

$$y''' + a_2(t)y'' + a_1(t)y' + a_0(t)y = 0$$

to a first-order system. Write its initial state in terms of $y(0)$, $y'(0)$, and $y''(0)$.

(c) Convert the coupled equations

$$x'' = -x + y, \quad y'' = x - y$$

to a first-order system. How many scalar coordinates does its state have? List the initial data, and explain how a system solution recovers both original equations.

(d) For the equation $a_2(t)y'' + a_1(t)y' + a_0(t)y = g(t)$, what condition on a_2 lets you carry out the usual conversion throughout an interval? Explain which step can fail at a zero of a_2 .

6. What does the theorem tell us?

Apply theorem 5.4. Check the right-hand side and its state partial derivatives; do not infer uniqueness merely from how many initial values have been supplied.

(a) For $x' = 2y$, $y' = -3x$, check the four state partial derivatives. Does the theorem apply at every initial state? State precisely its conclusion about the interval of existence.

(b) Repeat for $x' = v$, $v' = 1 - v^2$. Translate the conclusion into a statement about initial value problems for the scalar second-order equation.

(c) Regard $y' = 3y^{2/3}$ as a one-component system, where $y^{2/3} = (\sqrt[3]{y})^2$. Compare the initial values $y(0) = 1$ and $y(0) = 0$. Where does the theorem apply? At zero, verify two distinct solutions with the same initial value. Distinguish the failed hypothesis from the separate calculation establishing nonuniqueness.

7. Have we found every solution?

For the system $x' = 2y$, $y' = -3x$, let $\omega = \sqrt{6}$ and consider the proposed family

$$\begin{aligned} x_{A,B}(t) &= A \cos(\omega t) + \frac{2B}{\omega} \sin(\omega t), \\ y_{A,B}(t) &= B \cos(\omega t) - \frac{3A}{\omega} \sin(\omega t). \end{aligned}$$

(a) Verify both differential equations. Recover A and B by evaluating the solution at zero.

(b) Let (u, v) be any solution on an interval containing zero. Choose A, B from its initial state and use uniqueness to prove that it agrees with the displayed solution near zero. Explain why agreement extends throughout the common interval.

(c) Write the solution with initial state $(x(0), y(0)) = (2, -1)$. Which part of your argument shows that the formula works, and which shows that there are no other solutions with those data?

8. The interval changes the solution set

For $y' = y^2$, the solution starting at $y(0) = a$ is

$$y_a(t) = \frac{a}{1 - at}$$

on its interval of existence containing zero.

- (a) For which real values of a is this solution defined throughout $(-1, 1)$? Include the end-point values of your parameter range if they work. How does the answer change if the solution must exist on an open interval containing the closed interval $[-1, 1]$?
- (b) Answer the first question for $(-2, 2)$, then for $\mathbb{R}$. Explain why a one-parameter local family need not remain a one-parameter family when we require solutions for all time.

EXPLORATIONS

9. Can solution graphs cross?

Consider the two spring solutions $u(t) = \sin t$ and $v(t) = -\sin t$.

- (a) Verify that both solve $x'' = -x$. Find their graph intersections and compare their derivatives at those times.
- (b) Write the corresponding state histories $U = (u, u')$ and $V = (v, v')$. At a time when the scalar graphs cross, do the system solutions have the same state? Explain why uniqueness permits the graph crossings.
- (c) Could two distinct spring solutions have both the same value and the same derivative at the same time? Give an argument using the system theorem rather than the explicit solution formula.
- (d) Show that U and V trace the same unit circle in the state plane, but never occupy the same point at the same time. Explain the distinction between a geometric curve and a function of time. Why does tracing the same curve not contradict uniqueness?

10. How long does the nonlinear family exist?

Return to

$$x_{a,b}(t) = a + \log h_b(t), \quad h_b(t) = \cosh t + b \sinh t,$$

which solves $x'' = 1 - (x')^2$ wherever $h_b > 0$. We want the largest open interval containing zero on which this formula gives a solution.

- (a) Use the exponential definitions of $\cosh$ and $\sinh$ to write $h_b(t)$ as a linear combination of e^t and e^{-t} . Show that $h_b(t) > 0$ for every real t when $|b| \leq 1$. What are the solutions when $b = 1$ and $b = -1$?
- (b) When $b > 1$ or $b < -1$, find the time t_* at which $h_b(t_*) = 0$. Determine on which side of t_* the function is positive, and give the requested interval of existence in each case.

- (c) What happens to $x_{a,b}(t)$ as t approaches t_* from inside that interval? Explain why the solution cannot extend through t_* as a real-valued solution, even though the system's right-hand side and all its state partial derivatives are continuous everywhere.
- (d) Which initial states (a, b) give solutions for every real t ? Which give solutions for every $t \geq 0$? Sketch both sets in the initial-data plane. Explain why changing a does not change the interval of existence.

Linear Evolution

In chapter 5, initial data became coordinates for whole solutions. We could specify a function by finitely many numbers because the differential equation and uniqueness determined everything else. For a nonlinear equation, those coordinates pick out solutions, but the solutions they pick out cannot be combined. For a homogeneous linear equation, the solutions form a vector space. What happens when we put these two facts together?

The coordinates become linear coordinates: adding initial states adds the corresponding solutions, and scaling an initial state scales its entire solution. This means that a basis of states can give us a basis of functions. Once we have solved from those few initial states, every other initial value problem can be answered by linear algebra.

We will develop this for systems whose coefficients may depend on time. The solutions need not have convenient formulas. The point of the structure is that we can use it even when finding individual solutions is difficult.

6.1 Initial Data as Linear Coordinates

Consider an N -component homogeneous linear system

$$X' = A(t)X,$$

where $A(t)$ is an $N \times N$ matrix with continuous entries on an open interval I . Fix $t_0 \in I$. From the last chapter, we know that every initial state gives a unique local solution. We also know that linear combinations of solutions remain solutions wherever they are all defined.

For linear equations, we can strengthen the existence conclusion: the unique solution extends throughout the interval on which the coefficients are continuous. The same strengthening holds with a continuous forcing term. The improvement con-

cerns how long the solution exists: the local guarantee from chapter 5 becomes a guarantee on all of I .

Theorem 6.1 (Global existence and uniqueness for linear systems). Suppose the entries of $A(t)$ and the components of $g(t)$ are continuous on an open interval I , and let $t_0 \in I$. For every $X_0 \in \mathbb{R}^N$, the initial value problem

$$X' = A(t)X + g(t), \quad X(t_0) = X_0$$

has exactly one solution on all of I .

We use this theorem without proof. The word *global* means throughout I : a solution cannot blow up at a finite time inside the interval where its coefficients and forcing remain continuous. If those functions are continuous on all of $\mathbb{R}$, every solution exists for all real time. This is the system version of the stronger linear existence theorem from chapter 3. The constant-coefficient construction in the next chapter will give an explicit special case; the theorem here also covers coefficients that vary with time.

Return to the homogeneous equation $X' = A(t)X$, and let

$$\mathcal{S}(I) = \{X \in C^1(I, \mathbb{R}^N) : X' = A(t)X\}.$$

This is a vector subspace of the function space $C^1(I, \mathbb{R}^N)$. Evaluation at t_0 defines a map

$$E_{t_0} : \mathcal{S}(I) \longrightarrow \mathbb{R}^N, \quad E_{t_0}(X) = X(t_0).$$

The previous chapter used this map to recover the parameters of a solution. Now both its domain and codomain are vector spaces. Does it respect their arithmetic? For any solutions U, V and real numbers a, b ,

$$E_{t_0}(aU + bV) = (aU + bV)(t_0) = aU(t_0) + bV(t_0) = aE_{t_0}(U) + bE_{t_0}(V).$$

Yes! Evaluation is linear. Existence and uniqueness give the rest.

Theorem 6.2 (Initial data as linear coordinates). Let $A(t)$ have continuous entries on an open interval I , and fix $t_0 \in I$. Evaluation at t_0 is a linear isomorphism from the solution space of $X' = A(t)X$ onto $\mathbb{R}^N$. Consequently,

$$\dim \mathcal{S}(I) = N.$$

Proof. We have just checked linearity. For every $X_0 \in \mathbb{R}^N$, existence supplies a solution $X \in \mathcal{S}(I)$ with $E_{t_0}(X) = X_0$, so evaluation is surjective. If two solutions have the same image, they have the same initial state; uniqueness says they are the same solution on I . Thus evaluation is also injective.

A bijective linear map is a linear isomorphism. Isomorphic vector spaces have the same dimension, so $\dim \mathcal{S}(I) = N$. ■

We have determined the dimension without finding a single solution formula. The surrounding function space is infinite-dimensional, but the equation selects a subspace with exactly as many dimensions as there are state coordinates. And we know a concrete isomorphism: to turn a solution into a vector of numbers, evaluate it at t_0 .

Suppose U and V have initial states U_0 and V_0 at t_0 . Superposition says that $aU + bV$ is a solution, and its initial state is $aU_0 + bV_0$. Uniqueness therefore identifies it as the solution from that initial state. The same coefficients combine the initial states and the entire solutions.

6.1.1 Higher-Order and Forced Equations

For a homogeneous scalar equation

$$a_n(t)y^{(n)} + a_{n-1}(t)y^{(n-1)} + \dots + a_1(t)y' + a_0(t)y = 0,$$

suppose all coefficients are continuous on I and a_n never vanishes there. The conversion from chapter 5 gives an n -component homogeneous linear system with continuous coefficients. The correspondence $y \mapsto (y, y', \dots, y^{(n-1)})$ is itself linear, so the scalar solution space is n -dimensional. Its initial-data isomorphism records

$$y \mapsto (y(t_0), y'(t_0), \dots, y^{(n-1)}(t_0)).$$

For a scalar equation, evaluating only $y(t_0)$ would lose information. The derivative values are the remaining state coordinates.

With a fixed continuous forcing, we obtain an affine space instead. For $X' = A(t)X + g(t)$, the existence theorem supplies a particular solution X_p on I . Chapter 5 then identifies the full solution set as

$$X_p + \mathcal{S}(I).$$

Its affine dimension is N , the dimension of its space of directions $\mathcal{S}(I)$. If we choose $X_p(t_0) = 0$ and let U be the homogeneous solution with $U(t_0) = X_0$, then $X_p + U$ solves the forced equation with initial state X_0 . The same reasoning gives affine

dimension n for the scalar order- n forced equation under the coefficient assumptions above. These statements concern all solutions of the equation; prescribing the full initial data selects just one of them. Methods for finding particular forced solutions will come later.

6.2 Bases of Solutions and Reconstruction

An isomorphism transports a basis from one vector space to another. Choose a basis $b_1, \dots, b_N$ of $\mathbb{R}^N$, and let U_j be the solution with $U_j(t_0) = b_j$. These functions form a basis of $\mathcal{S}(I)$. To express an arbitrary solution in that basis, we only need to express its initial state in the basis $b_1, \dots, b_N$.

Theorem 6.3 (A basis of states gives a basis of solutions). Let $U_1, \dots, U_N$ solve the same homogeneous linear system $X' = A(t)X$ on an open interval I , with A continuous and $t_0 \in I$. They form a basis of the solution space if and only if $U_1(t_0), \dots, U_N(t_0)$ form a basis of $\mathbb{R}^N$.

For such a basis, the unique solution from $X(t_0) = X_0$ is

$$X(t) = c_1 U_1(t) + \dots + c_N U_N(t),$$

where the coefficients are determined by

$$X_0 = c_1 U_1(t_0) + \dots + c_N U_N(t_0).$$

Proof. Evaluation is a linear isomorphism, so it preserves linear independence and spanning in both directions. For the reconstruction formula, superposition says that $c_1 U_1 + \dots + c_N U_N$ is a solution. The equation for the coefficients makes its initial state X_0 , so uniqueness identifies it as the solution from X_0 . The coefficients are unique because the chosen initial states form a basis. ■

The independence assertion is worth interpreting. We can test whether whole functions are independent by examining their states at a single time. For arbitrary functions, that would be false: many different functions pass through the same value. Here uniqueness is built into the isomorphism. A linear combination of solutions that has zero initial state must be the zero solution everywhere on I .

Return to the oscillator from chapter 5,

$$x' = 2y, \quad y' = -3x.$$

With $\omega = \sqrt{6}$, we already know the two solutions

$$U(t) = \begin{pmatrix} \cos(\omega t) \\ \frac{3}{\omega} \sin(\omega t) \end{pmatrix}, \quad V(t) = \begin{pmatrix} \frac{2}{\omega} \sin(\omega t) \\ \cos(\omega t) \end{pmatrix}.$$

Their initial states are $U(0) = (1, 0)$ and $V(0) = (0, 1)$, so the basis theorem immediately identifies them as a basis of solutions. For initial state (a, b) , the entire solution is $aU + bV$. For example, $(2, -1) = 2(1, 0) - (0, 1)$ gives

$$X(t) = 2U(t) - V(t) = \begin{pmatrix} 2 \cos(\omega t) - \frac{2}{\omega} \sin(\omega t) \\ -\frac{6}{\omega} \sin(\omega t) - \cos(\omega t) \end{pmatrix}.$$

There was no new integration to perform. We used two solved functions and a calculation in $\mathbb{R}^2$. The theorem tells us that this procedure reaches every solution, and that each one has unique coefficients. Nothing in its proof required sine and cosine formulas.

6.2.1 Solve a Few Times, Solve Them All

We can use the same structure when solutions must be approximated numerically. Consider the synthetic time-dependent system

$$\begin{aligned} x' &= -\frac{1}{4}x + \left(2 + \frac{1}{4} \sin t\right)y, \\ y' &= -\left(2 + \frac{1}{4} \cos t\right)x - \frac{1}{4}y. \end{aligned}$$

Without the sine and cosine perturbations, its rate is $-\frac{1}{4}(x, y) + (2y, -2x)$: an inward radial component plus a clockwise turning component. The perturbations make the turning coefficients vary differently with time. This is a mathematical example chosen to explore reconstruction with time-dependent coefficients. All coefficients are continuous on $\mathbb{R}$, so its solution space is two-dimensional and every solution exists for all real time.

Let U, V now denote this system's solutions from $U(0) = (2, 1)$ and $V(0) = (1, 2)$. These initial vectors are independent, so the solutions form a basis. We can approximate the functions and save their values on a common time grid.

To reconstruct the solution from $(5, 4)$, solve

$$a(2, 1) + b(1, 2) = (5, 4).$$

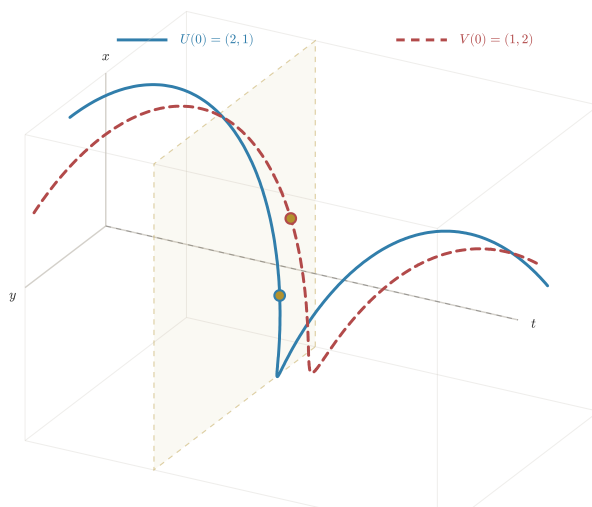


Figure 6.1 The curves approximate $t \mapsto (t, U(t))$ and $t \mapsto (t, V(t))$ in (t, x, y) -space. Rotate the picture to inspect their shape, then drag the common time marker to compare their states at the same time. Keeping the time coordinate visible lets us read each curve as a function of time.

This gives $2a + b = 5$, $a + 2b = 4$, hence $a = 2$, $b = 1$. The exact solution is $2U + V$. Combining the stored approximations with those same coefficients gives an approximation to it.

For comparison, numerical approximations at $t = 2$ give

$$U(2) \approx (-1.082775, 0.853062), \quad V(2) \approx (-1.378345, 0.066657).$$

The table compares reconstruction with separate numerical integrations from three initial states. Values were combined before rounding.

$X(0)$	(a, b)	$aU(2) + bV(2)$	direct solve at $t = 2$
$(5, 4)$	$(2, 1)$	$(-3.543895, 1.772781)$	$(-3.543895, 1.772781)$
$(1, -1)$	$(1, -1)$	$(0.295570, 0.786404)$	$(0.295570, 0.786404)$
$(0, 3)$	$(-1, 2)$	$(-1.673915, -0.719747)$	$(-1.673915, -0.719747)$

Try this with Euler's method: compute from $(2, 1)$ and $(1, 2)$ using the same grid, then combine the stored values to approximate a third solution. Why should this agree with a direct Euler run? Write the system as $X' = A(t)X$. Its update is itself linear:

$$X_{n+1} = (I + hA(t_n))X_n.$$

If $X_n = aU_n + bV_n$, then

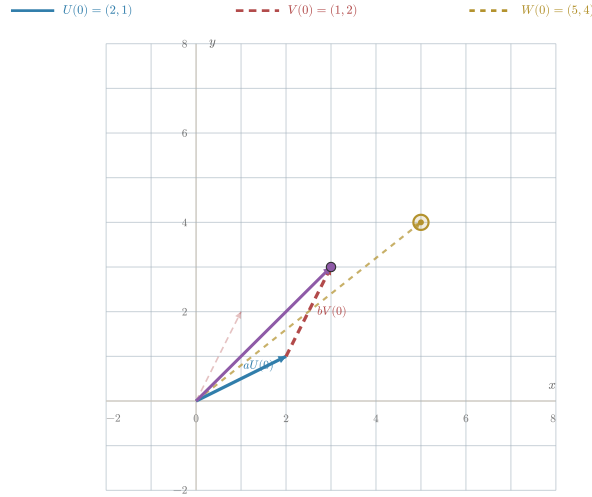


Figure 6.2 The blue and red vectors are the known initial states $U(0)$ and $V(0)$; the gold vector is the target $(5, 4)$. Adjust a, b until $aU(0) + bV(0)$ reaches the target. Those coefficients also combine the entire solutions.

$$X_{n+1} = (I + hA(t_n))(aU_n + bV_n) = aU_{n+1} + bV_{n+1}.$$

The equality holds initially, so induction gives it at every step. Apart from floating-point roundoff, combining Euler approximations on the same grid gives exactly the same discrete answer as another Euler run. Its accuracy against the true solution is a separate question, still governed by the step size and numerical error. Agreement between two ways of computing the same approximation does not eliminate that error.

6.2.2 Choosing the Standard Initial States

Our numerical basis began at $(2, 1)$ and $(1, 2)$, so each new problem required a small linear system to find its coefficients. In the exact oscillator example, we had already made a more convenient choice: the standard initial vectors. For a general N -component system, let R_j be the solution with

$$R_j(t_0) = e_j, \quad j = 1, \dots, N.$$

We call these the **standard solutions** at t_0 . Since the e_j form a basis, so do the R_j . If $X_0 = (\xi_1, \dots, \xi_N)$, its coordinates in the standard basis are already visible. The solution from X_0 is therefore

$$X(t) = \xi_1 R_1(t) + \cdots + \xi_N R_N(t).$$

We can compute these standard solutions exactly when formulas are available, or approximate them numerically as above. Either way, the same collection serves every initial state at t_0 . How can we package the collection so that one operation performs the reconstruction?

6.3 Linear Evolution and Its Matrix

Fix a time t . The reconstruction formula expresses $X(t)$ as a linear combination of the vectors $R_1(t), \dots, R_N(t)$, with the entries of X_0 as coefficients. This is exactly how matrix multiplication works: put those vectors into columns,

$$\Phi_{t_0}(t) = (R_1(t) \cdots R_N(t)).$$

Multiplying by an initial state carries out the reconstruction:

$$\Phi_{t_0}(t)X_0 = \xi_1 R_1(t) + \cdots + \xi_N R_N(t) = X(t).$$

We call $\Phi_{t_0}(t)$ the **flow matrix** from t_0 to t . For each fixed t , it represents the linear map taking an initial state at t_0 to the state of its solution at t . At the initial time, each column is its original basis vector, so $\Phi_{t_0}(t_0) = I$. At every other time its columns still form a basis: E_t is an isomorphism just as E_{t_0} is. Thus the flow matrix is invertible. Evolving to a later time does not lose the information needed to recover the initial state.

The starting time matters. For our time-dependent system, numerical integration gives

$$\Phi_0(1) \approx \begin{pmatrix} -0.462693 & 0.632462 \\ -0.661972 & -0.406010 \end{pmatrix}, \quad \Phi_1(2) \approx \begin{pmatrix} -0.433324 & 0.699270 \\ -0.627664 & -0.386833 \end{pmatrix}.$$

Both maps advance the state by one unit of time, but the coefficient matrix changes during those two intervals. Equal elapsed times need not produce equal evolution maps. This is why we retain t_0 in the notation.

We have built the flow matrix out of solutions. Can we describe it by a differential equation of its own? Differentiate column by column:

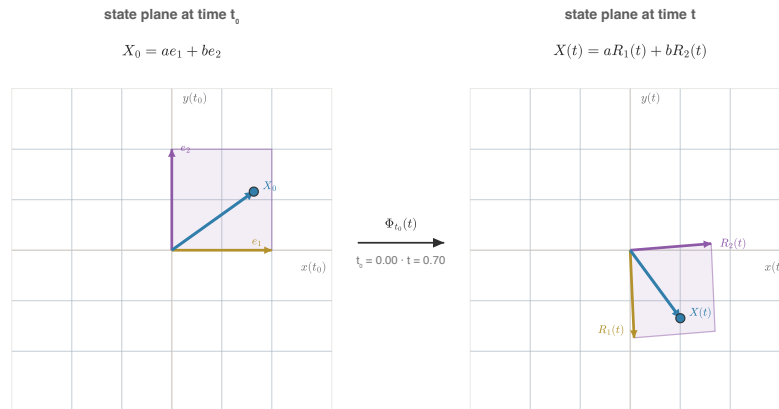


Figure 6.3 For the time-dependent system above, the left panel shows $X_0 = ae_1 + be_2$. The flow map sends the basis vectors to $R_1(t)$ and $R_2(t)$, so the right panel shows $X(t) = aR_1(t) + bR_2(t)$. Drag X_0 , or change the starting and elapsed times, to compare the two linear combinations.

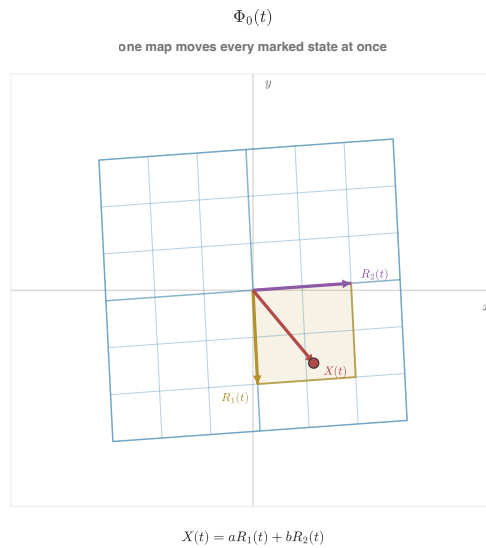


Figure 6.4 The map $\Phi_0(t)$ acts simultaneously on a grid of initial states, the standard basis, the unit square, and one selected state. The coefficients a, b remain fixed while the basis vectors and selected state move. Each frame displays one linear map at one time.

$$\begin{aligned}
\Phi'_{t_0}(t) &= (R'_1(t) \cdots R'_N(t)) \\
&= (A(t)R_1(t) \cdots A(t)R_N(t)) \\
&= A(t)\Phi_{t_0}(t).
\end{aligned}$$

Thus all the standard initial value problems fit into one matrix initial value problem:

$$\Phi'_{t_0}(t) = A(t)\Phi_{t_0}(t), \quad \Phi_{t_0}(t_0) = I.$$

A solution of this matrix equation supplies the solution from *every* initial state by $X(t) = \Phi_{t_0}(t)X_0$. Numerically, the Euler updates for all the columns can likewise be collected into

$$\widehat{\Phi}_{n+1} = (I + hA(t_n))\widehat{\Phi}_n, \quad \widehat{\Phi}_0 = I,$$

where $t_n = t_0 + nb$. This computes an approximation to the flow matrix one step at a time.

There is still a differential equation hidden in the construction: we obtained the columns by solving from the standard initial states. The original equation gives us $A(t)$, which assigns the rate of change at an instant. We want $\Phi_{t_0}(t)$, which carries out the resulting evolution over an interval. In the next chapter we specialize to a constant matrix A and ask how to construct that evolution directly from the matrix itself.

PRACTICE · EXPLORATION · COMPUTATION

Problems

CHECK YOUR UNDERSTANDING

1. Determine the dimension

For each homogeneous linear problem below, assume continuous coefficients and equations solved for their highest derivatives. State how many independent pieces of initial data are needed and hence the dimension of the solution space. How many solutions form a basis?

List the initial data at $t = 0$ for the solutions corresponding to the standard basis of the state space.

- (a) $y''' + a_2(t)y'' + a_1(t)y' + a_0(t)y = 0$.
- (b) A homogeneous first-order system in four unknown functions.
- (c) Two coupled second-order equations for $x(t)$ and $y(t)$. First rewrite the state in first-order form before counting.

PYTHON: REUSING COMPUTED SOLUTIONS

2. Combine computed solutions

Use the Euler function from Chapter 2 to compute solutions of $y' = ty$ from initial values 2, -1, and 5 on the same time grid.

- (a) Form the array $2Y_2 - Y_{-1}$ from the first two computed solutions. Which initial value does this combination have? Compare it point by point with the separately computed solution from that initial value.
- (b) Plot the difference between the combined and directly computed arrays. Is the difference exactly zero, close to zero, or visibly large? Explain why Euler's update itself preserves superposition for this linear equation.
- (c) Repeat the experiment with the logistic equation $y' = y(1 - y/10)$. Plot the combined approximation and the directly computed approximation together, then plot their difference. At what step does the mismatch first appear?

The Matrix Exponential

Chapter 6 left us with a matrix initial value problem. For the system $X' = A(t)X$, the flow matrix beginning at time t_0 satisfies

$$\Phi'_{t_0}(t) = A(t)\Phi_{t_0}(t), \quad \Phi_{t_0}(t_0) = I.$$

We can construct this matrix by solving once from each standard initial state, either exactly or numerically. That procedure computes the columns of Φ_{t_0} separately. It does not yet tell us how to construct the whole matrix directly from the coefficient matrix $A(t)$.

We now restrict attention to a constant system $X' = AX$. In this case the starting time can be removed by resetting the clock. If $X(t_0) = X_0$, define $Y(\tau) = X(t_0 + \tau)$. The chain rule gives

$$Y'(\tau) = X'(t_0 + \tau) = AX(t_0 + \tau) = AY(\tau), \quad Y(0) = X_0.$$

Thus the evolution depends only on the elapsed time $\tau = t - t_0$: $\Phi_{t_0}(t) = \Phi_0(t - t_0)$. We may set the initial time equal to zero, construct $\Phi_0(t)$, and replace t by $t - t_0$ when the original clock is needed.

The problem is therefore precise: given a constant $N \times N$ matrix A , construct the matrix-valued solution of

$$\Phi'(t) = A\Phi(t), \quad \Phi(0) = I$$

directly from A .

7.1 The Generator of Rotation

The solutions of the normalized spring $x'' = -x$ are already known. With state $X = (x, v)$, where $v = x'$, its equation becomes the constant system

$$X' = \begin{pmatrix} x' \\ v' \end{pmatrix} = \begin{pmatrix} v \\ -x \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix} = JX, \quad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

At a state $X = (x, v)$, the matrix J assigns the velocity $JX = (v, -x)$. This vector is perpendicular to X , since $(x, v) \cdot (v, -x) = xv - vx = 0$, and it has the same length because $\|JX\|^2 = v^2 + x^2 = \|X\|^2$. Thus the differential equation prescribes clockwise motion tangent to the circle through the current state, with unit angular speed.

We also know the solution from an arbitrary initial state $X_0 = (x_0, v_0)$. The position and velocity are

$$x(t) = x_0 \cos t + v_0 \sin t, \quad v(t) = -x_0 \sin t + v_0 \cos t.$$

Putting these two equations into a single matrix equation gives

$$X(t) = R(t)X_0, \quad R(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

The matrix J describes the velocity at one instant, while $R(t)$ advances every initial state through a time interval of length t . These matrices are connected by the same equation that governed the flow matrix in Chapter 6. Indeed,

$$R'(t) = \begin{pmatrix} -\sin t & \cos t \\ -\cos t & -\sin t \end{pmatrix},$$

$$JR(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} = \begin{pmatrix} -\sin t & \cos t \\ -\cos t & -\sin t \end{pmatrix}.$$

Therefore $R'(t) = JR(t)$, and direct evaluation gives $R(0) = I$. In particular,

$$R'(0) = JR(0) = J.$$

Here we obtained $R(t)$ from a solution formula that was already available. If the sine and cosine formulas were not known, how could we recover $R(t)$ from the matrix J alone?

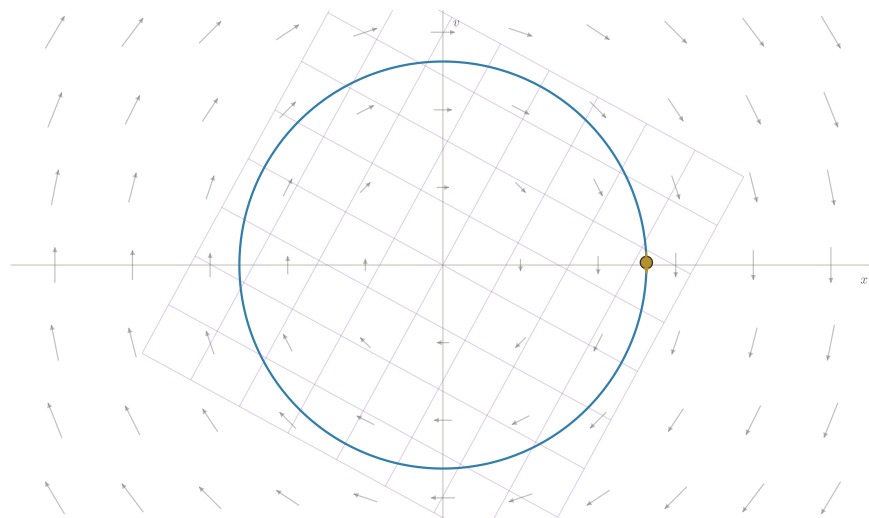


Figure 7.1 For $X' = JX$, every velocity arrow is perpendicular to the position vector and grows in proportion to its length. The state therefore travels around a circle while the grid rotates rigidly with it. In the live figure, drag the gold state and watch the entries of the rotation matrix $R(t)$ change as the motion continues.

7.2 Repeated Small Steps

Before trying to recover the rotation matrix from J , consider the scalar equation with the same constant-coefficient form:

$$y' = ay, \quad y(0) = y_0.$$

Euler's method with step size h replaces the continuous evolution by the repeated update

$$y_{k+1} = y_k + h a y_k = (1 + h a) y_k.$$

To travel from time 0 to time t in n equal steps, take $h = t/n$. Repeating the update n times gives

$$y_n = \left(1 + \frac{at}{n}\right)^n y_0.$$

Calculus identifies the limit of these growth factors:

$$e^{at} = \lim_{n \rightarrow \infty} \left(1 + \frac{at}{n}\right)^n.$$

Now apply exactly the same numerical idea to $X' = AX$. One Euler step sends the current state X_k to

$$X_{k+1} = X_k + hAX_k = (I + hA)X_k.$$

Because A is constant, every step uses the same matrix. After n steps of size t/n , the computed state is therefore

$$X_n = \left(I + \frac{t}{n}A\right)^n X_0.$$

For the spring generator J , we can see precisely what these repeated updates do. The preceding section showed that JX is perpendicular to X and has the same length. Thus the new radius $X + hJX$ is the hypotenuse of a right triangle whose perpendicular legs are X and hJX . The Pythagorean theorem gives

$$\begin{aligned}\|X + hJX\|^2 &= \|X\|^2 + b^2\|JX\|^2 \\ &= (1 + b^2)\|X\|^2.\end{aligned}$$

One step therefore multiplies every length by $\sqrt{1 + b^2}$. Its signed clockwise turning angle is $\arctan b$: the ratio of the perpendicular leg to the original radius is b . In terms of the rotation matrix from the previous section, the entire Euler update is

$$I + hJ = \sqrt{1 + b^2} R(\arctan b).$$

Repeating this same scaling and rotation n times, with $b = t/n$, yields

$$\left(I + \frac{t}{n}J\right)^n = \left(1 + \frac{t^2}{n^2}\right)^{n/2} R\left(n \arctan \frac{t}{n}\right).$$

Both finite-step errors are visible here. The first factor makes the state spiral outward, while the angle after n steps is not quite t . But both errors disappear as the step size shrinks. For the radial factor,

$$0 \leq \frac{n}{2} \log \left(1 + \frac{t^2}{n^2}\right) \leq \frac{t^2}{2n} \longrightarrow 0,$$

so $(1 + t^2/n^2)^{n/2} \rightarrow 1$. When $t \neq 0$, the accumulated angle satisfies

$$n \arctan \frac{t}{n} = t \frac{\arctan(t/n)}{t/n} \longrightarrow t,$$

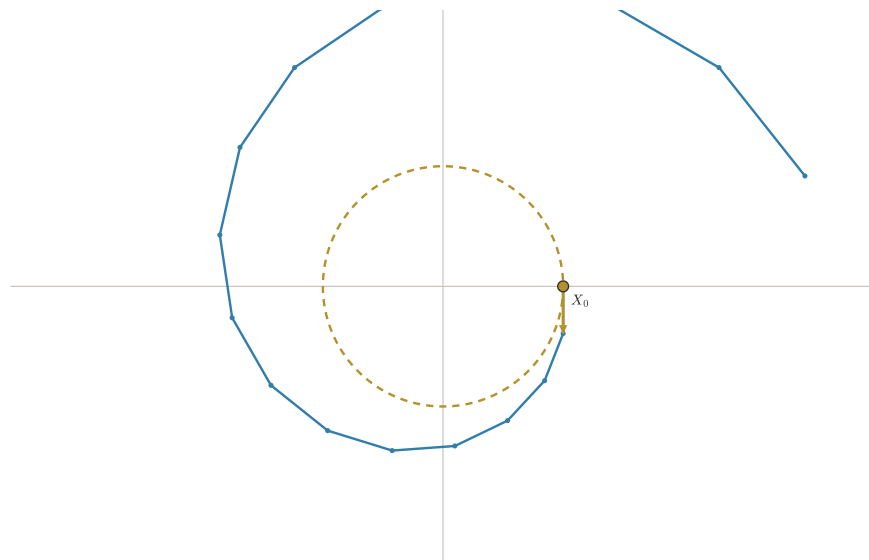


Figure 7.2 Each Euler step multiplies the state by $I + bJ$. This turns the state through $\arctan b$, but it also stretches lengths by $\sqrt{1 + b^2}$, so the polygon spirals outside the true circle. In the live figure, increase the number of steps per loop: the accumulated stretch shrinks and $(I + \frac{t}{n}J)^n$ approaches the rotation $R(t)$.

and the case $t = 0$ is immediate. Consequently,

$$\left(I + \frac{t}{n}J\right)^n \longrightarrow R(t).$$

Starting only from the local rule J , Euler's small linear turns recover the exact rotation in the limit.

The scalar and matrix calculations now have the same shape:

$$e^{at} = \lim_{n \rightarrow \infty} \left(1 + \frac{at}{n}\right)^n, \quad R(t) = \lim_{n \rightarrow \infty} \left(I + \frac{t}{n}J\right)^n.$$

For a general constant matrix A , this resemblance suggests both a name and a possible limiting formula for the exact flow:

$$e^{tA} \stackrel{?}{=} \lim_{n \rightarrow \infty} \left(I + \frac{t}{n}A\right)^n.$$

The question mark matters. We have proved this matrix limit only for the special generator J , and an Euler approximation does not by itself define the exact solution. The exact equation $\Phi' = A\Phi$, $\Phi(0) = I$ must tell us what the matrix e^{tA} is.

7.3 The Matrix Exponential

The Euler product suggested what to call the flow, but the exact equation must construct it. We return to

$$\Phi' = A\Phi, \quad \Phi(0) = I.$$

For the scalar equation $y' = ay$, the initial condition $y(0) = 1$ characterizes e^{at} . The matrix equation has precisely the same form: the scalar 1 has become the identity matrix I , and multiplication by the number a has become left multiplication by the matrix A . If a matrix-valued function deserves to be called e^{tA} , it is the solution of this initial value problem.

There is nothing mysterious about differentiating such a function. The set $M_N(\mathbb{R})$ of real $N \times N$ matrices is a vector space with N^2 coordinates, one for each entry. A matrix-valued function is a curve in this space, and its derivative is computed entry by entry. Moreover, the map $B \mapsto AB$ is linear on $M_N(\mathbb{R})$. Thus the equation above is an ordinary linear system in N^2 scalar coordinates. Viewed column by column, it is also exactly the collection of N standard-history problems from Chapter 6.

How might we construct its solution? In Calculus II, power series gave a way to solve differential equations when no formula was yet available. Try the same idea here by seeking a candidate of the form

$$\Phi(t) = C_0 + C_1 t + C_2 t^2 + C_3 t^3 + \dots,$$

where each coefficient C_k is a constant matrix. We are not assuming that an unknown solution must be analytic; we are asking whether a power series can produce one. Formally differentiating the candidate and multiplying it by A give

$$\begin{aligned} \Phi'(t) &= C_1 + 2C_2 t + 3C_3 t^2 + \dots, \\ A\Phi(t) &= AC_0 + AC_1 t + AC_2 t^2 + \dots. \end{aligned}$$

The initial condition requires $C_0 = I$. Matching equal powers of t in the differential equation then requires

$$C_1 = AC_0 = A, \quad 2C_2 = AC_1 = A^2, \quad 3C_3 = AC_2 = \frac{A^3}{2!}.$$

In general, $(k+1)C_{k+1} = AC_k$, so the pattern continues by induction:

$$C_k = \frac{A^k}{k!}.$$

We have therefore arrived at one possible solution,

$$I + tA + \frac{t^2 A^2}{2!} + \frac{t^3 A^3}{3!} + \dots$$

So far this is only a formal candidate. Before differentiating it as though it were a function, we must know that its partial sums actually converge. For a matrix $B = (b_{ij})$, define the **row-sum norm**

$$\|B\| = \max_i \sum_j |b_{ij}|.$$

This is also commonly called the infinity norm, though we will use the more descriptive name here.

This norm interacts especially well with matrix multiplication. If $B = (b_{ij})$ and $C = (c_{ij})$, then

$$\begin{aligned} \|BC\| &= \max_i \sum_j \left| \sum_\ell b_{i\ell} c_{\ell j} \right| \\ &\leq \max_i \sum_\ell |b_{i\ell}| \sum_j |c_{\ell j}| \\ &\leq \left(\max_i \sum_\ell |b_{i\ell}| \right) \left(\max_\ell \sum_j |c_{\ell j}| \right) \\ &= \|B\| \|C\|. \end{aligned}$$

Applying this inequality repeatedly gives $\|A^k\| \leq \|A\|^k$. Consequently,

$$\left\| \frac{t^k A^k}{k!} \right\| \leq \frac{|t|^k \|A\|^k}{k!}.$$

The scalar series formed from the quantities on the right is

$$\sum_{k=0}^{\infty} \frac{|t|^k \|A\|^k}{k!} = e^{|t| \|A\|},$$

which converges for every real t . The comparison therefore proves that our matrix series converges for every t as well. Since $M_N(\mathbb{R})$ has only finitely many coordinates, convergence in this norm is equivalent to convergence of every matrix entry.

We may now make the promised definition.

Definition 7.1 (Matrix exponential). For a real square matrix A , its **matrix exponential** is the matrix-valued function

$$e^{tA} := \sum_{k=0}^{\infty} \frac{t^k A^k}{k!}.$$

The notation does not mean that we exponentiate the entries of A separately. It names the matrix built from the powers $I, A, A^2, A^3, \dots$ by this convergent series.

Each entry of the matrix exponential is now a scalar power series with infinite radius of convergence. The power-series theorem from calculus therefore permits term-by-term differentiation. Reindexing the resulting series gives

$$\begin{aligned} \frac{d}{dt} e^{tA} &= \sum_{k=1}^{\infty} \frac{k t^{k-1} A^k}{k!} \\ &= A \sum_{k=1}^{\infty} \frac{t^{k-1} A^{k-1}}{(k-1)!} \\ &= A e^{tA}. \end{aligned}$$

At $t = 0$, every term except the first vanishes, so $e^{0A} = I$. The series really does solve the matrix initial value problem that produced it. Uniqueness now identifies it with the flow matrix.

Theorem 7.2 (Flow of a constant linear system). For a constant real matrix A , the unique solution of

$$X' = AX, \quad X(t_0) = X_0$$

is

$$X(t) = e^{(t-t_0)A} X_0.$$

Proof. Set $X(t) = e^{(t-t_0)A} X_0$. The calculation above gives

$$X'(t) = A e^{(t-t_0)A} X_0 = AX(t),$$

while $X(t_0) = e^{0A} X_0 = X_0$. Thus this function satisfies both the differential equation and the initial condition. Uniqueness says it is the solution. ■

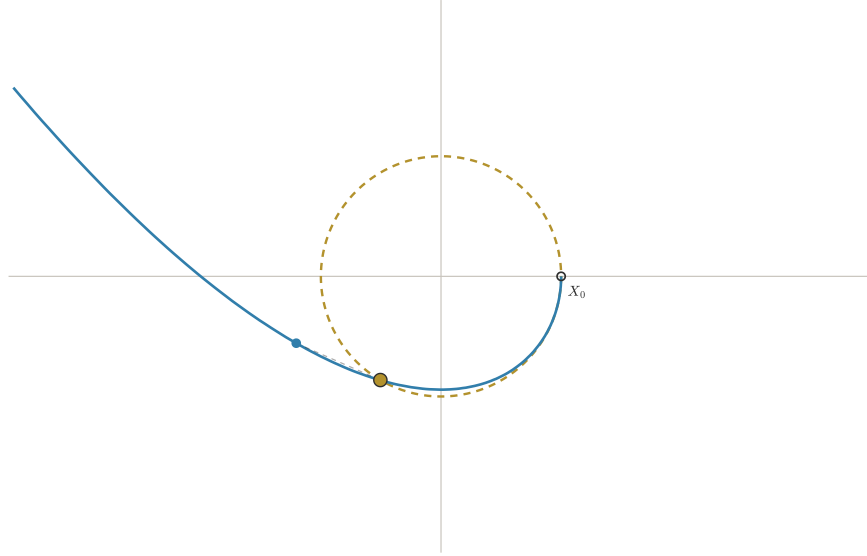


Figure 7.3 In the live figure, choose how many terms of the exponential series to keep. The resulting polynomial path follows the true circle for longer as more terms are added, then eventually peels away. Drag the gold state around the circle to compare its true position with the position predicted by the chosen truncation. In the limit, the series produces the flow we denote by e^{tJ} .

The spring shows what this construction has accomplished. Since $J^2 = -I$, the powers repeat in groups of four: $J^3 = -J$ and $J^4 = I$. Because the exponential series converges absolutely, we may collect its even and odd powers to obtain

$$\begin{aligned}
 e^{tJ} &= I + tJ + \frac{t^2 J^2}{2!} + \frac{t^3 J^3}{3!} + \cdots \\
 &= I \left(1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \cdots \right) + J \left(t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots \right) \\
 &= I \cos t + J \sin t \\
 &= R(t).
 \end{aligned}$$

We began with the instantaneous rule J and, without assuming the trigonometric solution, constructed the entire rotation flow from its powers.

The matrix exponential also obeys the familiar law for adding elapsed times. Fix s and consider, as functions of t ,

$$F(t) = e^{(t+s)A}, \quad G(t) = e^{tA} e^{sA}.$$

Both satisfy $Z' = AZ$, and both equal e^{sA} when $t = 0$. Uniqueness gives

$$e^{(t+s)A} = e^{tA}e^{sA}.$$

Taking $s = -t$ then gives

$$e^{tA}e^{-tA} = e^{0A} = I, \quad (e^{tA})^{-1} = e^{-tA}.$$

Thus every constant linear system produces an invertible flow for every elapsed time. The remaining difficulty is computational: when the powers of A do not repeat as transparently as the powers of J , how can we recognize the motion encoded by their series?

7.4 Matrix Flows We Can Compute

The exponential series gives an exact flow for every constant matrix, but an infinite series is not always the most revealing form of the answer. For some matrices, the pattern in the powers turns the series into a familiar formula and makes the resulting motion visible.

Begin with a scalar multiple of the identity, $A = \lambda I$. Since $(\lambda I)^k = \lambda^k I$, every term of the matrix series is a scalar multiple of the identity:

$$\begin{aligned} e^{tA} &= \sum_{k=0}^{\infty} \frac{t^k (\lambda I)^k}{k!} \\ &= \left(\sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \right) I \\ &= e^{\lambda t} I. \end{aligned}$$

The local rule $AX = \lambda X$ points along the line through X , outward when $\lambda > 0$ and inward when $\lambda < 0$. Its flow multiplies every state by the same factor $e^{\lambda t}$, so every state remains on its original ray.

The coordinates can also grow at different rates. If

$$D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$

then every power of D is still diagonal:

$$D^k = \begin{pmatrix} \lambda_1^k & 0 \\ 0 & \lambda_2^k \end{pmatrix}.$$

Substitution into the series gives

$$e^{tD} = \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix},$$

and therefore

$$e^{tD} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} e^{\lambda_1 t} x_0 \\ e^{\lambda_2 t} y_0 \end{pmatrix}.$$

This formula may look like entrywise exponentiation, but it works only because the diagonal matrix does not mix the coordinates. The two scalar equations $x' = \lambda_1 x$ and $y' = \lambda_2 y$ evolve independently. In particular, $D = \text{diag}(1, -1)$ expands the first coordinate while contracting the second as time moves forward.

Some matrices simplify for a different reason: their higher powers vanish. Consider

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Here $N^2 = 0$, so every term of the exponential series after the linear term vanishes:

$$e^{tN} = I + tN = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

Thus the flow sends

$$(x_0, y_0) \mapsto (x_0 + ty_0, y_0).$$

The second coordinate remains fixed while the first changes at the constant rate y_0 . Every point on the x -axis is fixed, and horizontal layers slide past one another faster when they lie farther from that axis. This is a **shear**.

The first and third of these motions can occur simultaneously. Suppose

$$A = aI + \omega J.$$

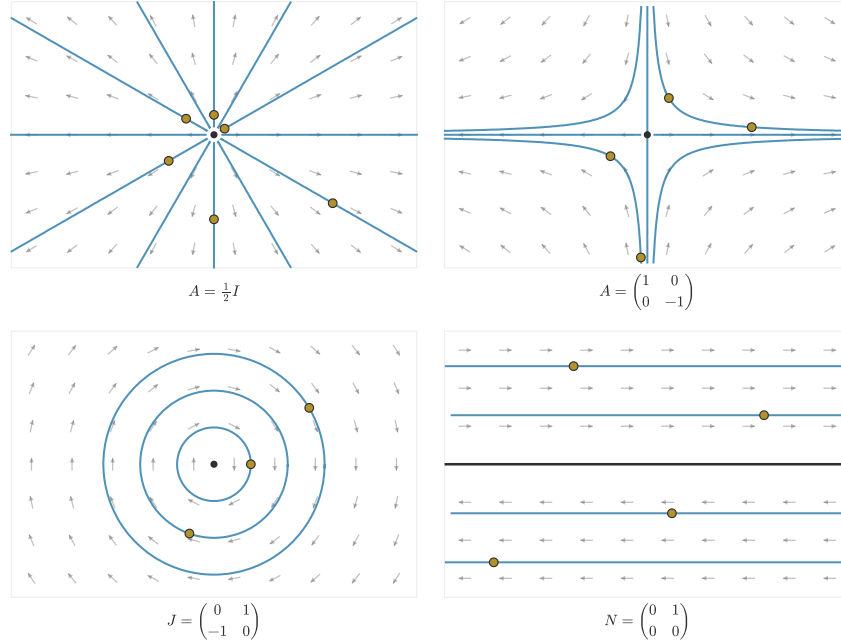


Figure 7.4 These four matrices have simple patterns in their powers but produce quite different motion. Uniform scaling pushes every ray outward, diagonal scaling gives the two axes different rates, J rotates the plane, and the nilpotent matrix slides horizontal layers past one another while leaving the x -axis fixed. The gold states move at the speeds prescribed by their respective vector fields.

At each state X , the term aX points radially while ωJX points tangentially. This suggests a flow that scales by e^{at} while rotating through the angle ωt . To check that the two motions really combine in this way, set

$$F(t) = e^{at}(I \cos(\omega t) + J \sin(\omega t)).$$

Using $J^2 = -I$, differentiation gives

$$\begin{aligned} F'(t) &= e^{at}(aI \cos(\omega t) + aJ \sin(\omega t) \\ &\quad - \omega I \sin(\omega t) + \omega J \cos(\omega t)), \end{aligned}$$

while multiplication by A gives

$$\begin{aligned} AF(t) &= e^{at}(aI + \omega J)(I \cos(\omega t) + J \sin(\omega t)) \\ &= e^{at}(aI \cos(\omega t) + aJ \sin(\omega t) + \omega J \cos(\omega t) - \omega I \sin(\omega t)). \end{aligned}$$

The two expressions agree, and $F(0) = I$. Uniqueness therefore proves

$$e^{t(aI+\omega J)} = e^{at}(I \cos(\omega t) + J \sin(\omega t)).$$

When $a > 0$, states spiral outward; when $a < 0$, they spiral inward; and when $a = 0$, we recover the circular motion of the normalized spring. The magnitude of ω sets the angular speed, while its sign sets the direction.

We can also write the formula as

$$e^{t(aI+\omega J)} = e^{taI} e^{t\omega J}.$$

This factorization works because aI and ωJ commute. It is not a general law: if two matrices B and C do not commute, then e^{B+C} need not equal $e^B e^C$.

These examples succeeded because the powers of each matrix followed a pattern we could recognize. The damped spring brings us to a matrix for which that pattern is not immediately visible. If the spring constant is k and the damping coefficient is b , its state $X = (x, v)$ satisfies

$$X' = \begin{pmatrix} 0 & 1 \\ -k & -b \end{pmatrix} X.$$

The theorem from the preceding section gives the exact solution

$$X(t) = e^{tA} X_0, \quad A = \begin{pmatrix} 0 & 1 \\ -k & -b \end{pmatrix}.$$

But the powers $A, A^2, A^3, \dots$ do not reveal the motion as readily as the powers of I, D, N , or J . The exponential notation has constructed the flow without yet making it easy to calculate or interpret.

What makes the earlier matrices simple is that they act predictably along particular directions. Can we find directions adapted to a more complicated matrix, so that its action becomes simple there? That question leads from the matrix exponential to the geometry of linear systems.

The Geometry of Linear Flows

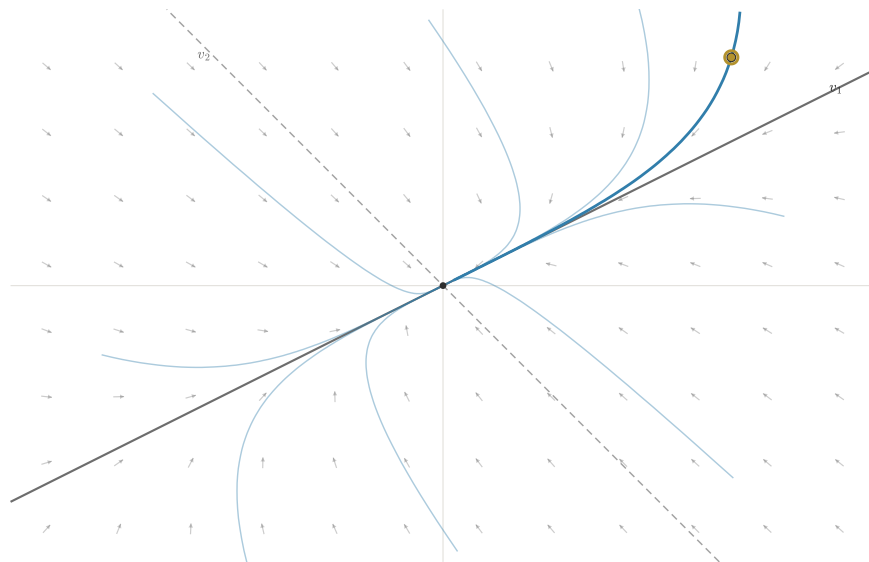


Figure 8.1 In the live figure, drag the gold state onto either invariant line. Its trajectory remains on that line and decays at the corresponding exponential rate. Away from the lines, trajectories curve toward the origin and eventually become tangent to the slower direction. The moving point makes that slow final approach visible.

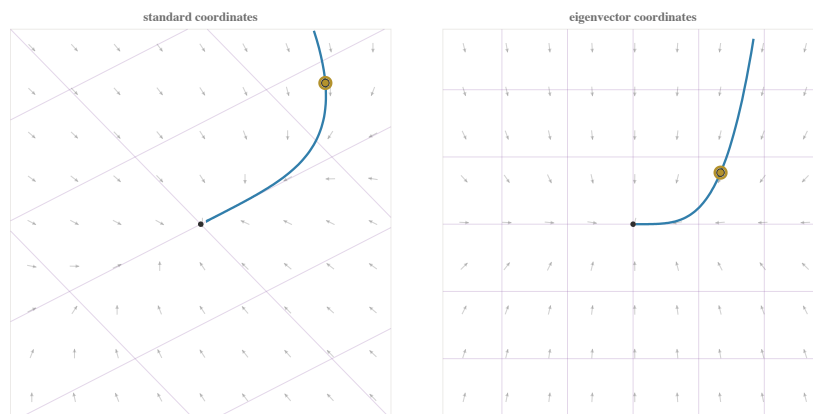


Figure 8.2 The two panels show the same flow in different coordinates. In standard coordinates the invariant grid is slanted; in eigenvector coordinates it is square, and each coordinate decays independently at its own rate. In the live figure, drag the gold state in either panel and follow its synchronized copy in the other.

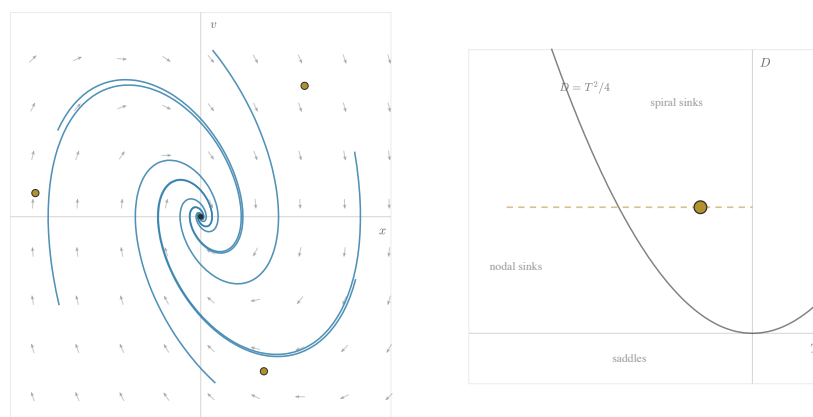


Figure 8.3 For $x'' + bx' + 2x = 0$, the state matrix has trace $-b$ and determinant 2. In the live figure, drag the gold point along this horizontal path in the trace–determinant plane. As b increases, the phase portrait changes from a center to a spiral sink, reaches a critically damped improper node on the parabola, and then becomes an overdamped node. The spiral sinks lie above the parabola.

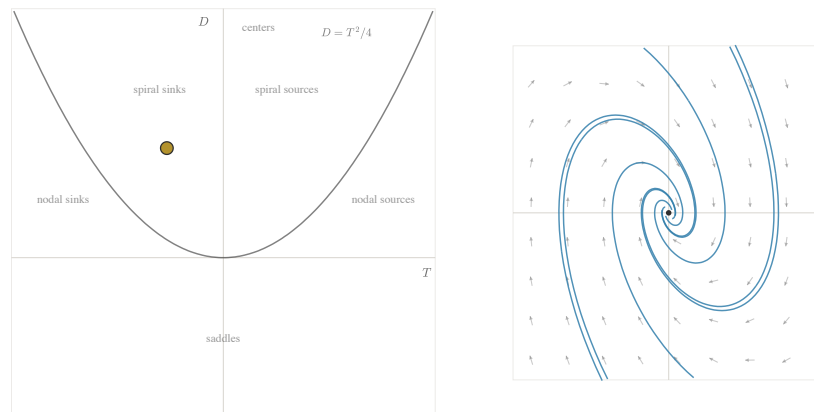


Figure 8.4 In the live figure, choose a point in the trace–determinant plane and compare it with the phase portrait beside it. Saddles lie below the horizontal axis, nodes between that axis and the parabola, and spirals above the parabola. Centers occur where the trace is zero and the determinant is positive. The live controls snap to the axes and the parabola so that the boundary cases can be examined directly.

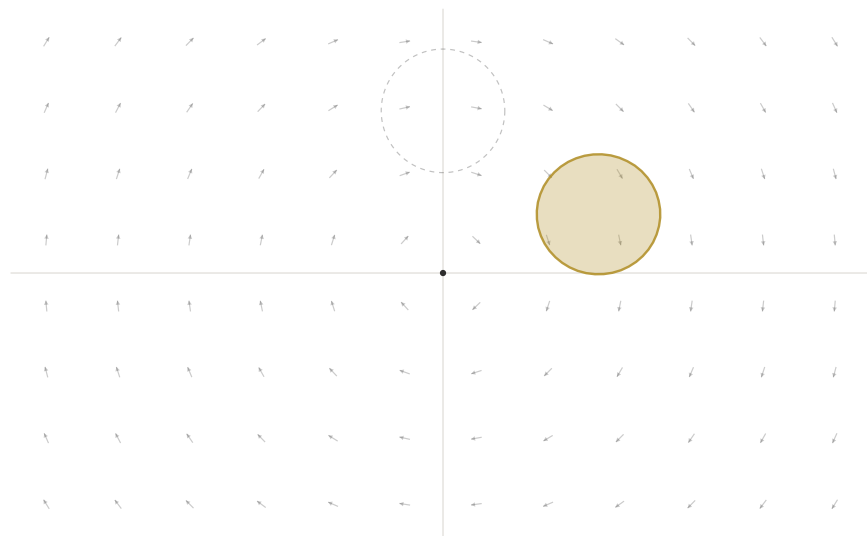


Figure 8.5 In the live figure, switch among the three flows and follow the same small region of initial states. A center carries it around without shrinking, a spiral sink draws it toward the origin, and the trace-zero saddle stretches it into a thin filament without changing its area. The area readout tracks $\det e^{tA} = e^{t \operatorname{tr} A}$ throughout the motion.

Modes of Linear Systems

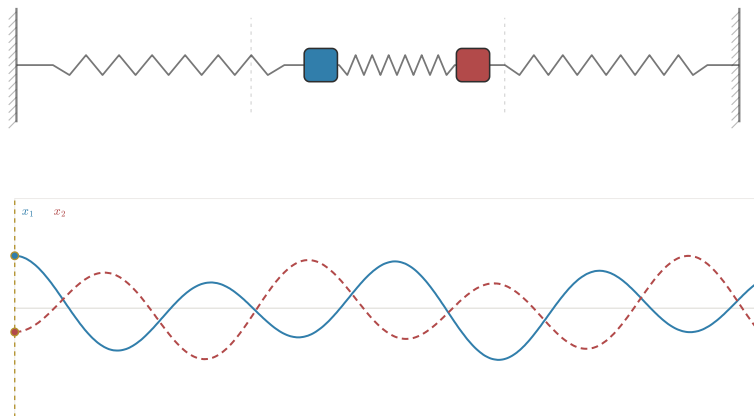


Figure 9.1 In the live figure, drag the two carts to choose a starting displacement, then let go. Most choices mix two frequencies and change shape as they move. If the carts begin with equal displacements, however, they retain that shape and oscillate with $\omega = 1$. If they begin with equal and opposite displacements, they retain the second shape and oscillate with $\omega = \sqrt{3}$.

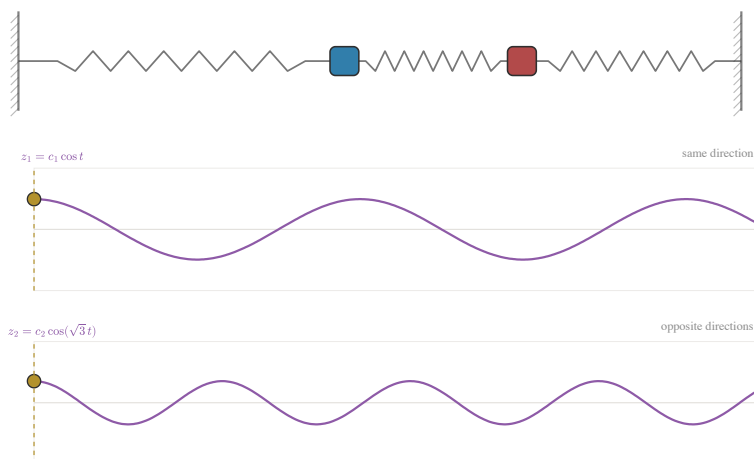


Figure 9.2 The motion shown above is assembled from $c_1 \cos(t)$ times the in-phase shape and $c_2 \cos(\sqrt{3}t)$ times the opposite shape. In the live figure, drag the two amplitude controls. Setting either coefficient to zero isolates one mode; combining them reconstructs the coupled motion of the carts.

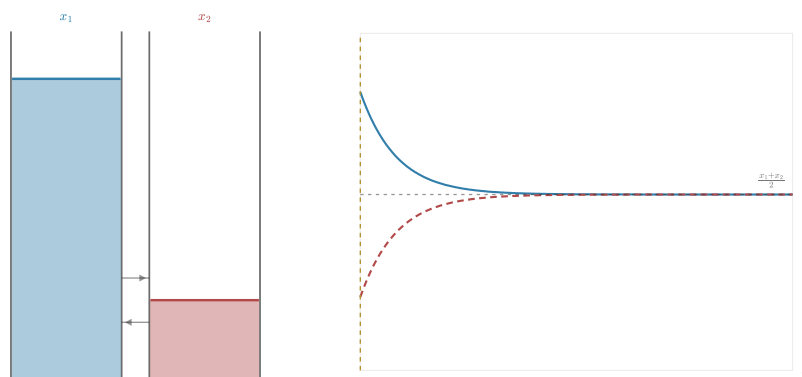


Figure 9.3 In the live figure, drag either brine level. The mean level remains fixed while the difference between the tanks decays exponentially. These two patterns—a constant sum and a decaying difference—reconstruct every approach to equal levels, even though nothing oscillates.



Figure 9.4 In the live figure, move through the modes of twelve coupled masses on a ring. Higher modes vary more rapidly around the ring and have higher frequencies; paired modes share a frequency, while the uniform mode has no restoring force. Switch from $q'' = -Lq$ to $u' = -Lu$ to make the same spatial pattern decay instead of oscillate.

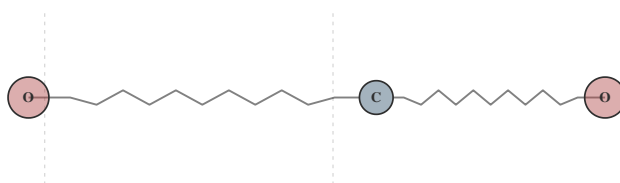


Figure 9.5 The three-mass model is calibrated to the two stretching frequencies of carbon dioxide. In the live figure, replace ^{12}C with ^{13}C without changing the fitted bond constants. The predicted antisymmetric frequency shifts to 2282 cm^{-1} , compared with the observed 2283.5 cm^{-1} . The symmetric frequency does not shift because the carbon atom remains still in that mode.

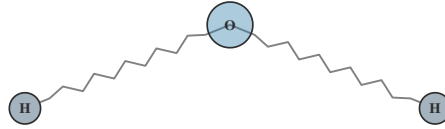


Figure 9.6 The six-coordinate calculation uses water's bent geometry and published force constants in $Kv = \omega^2 Mv$. Three zero eigenvalues correspond to rigid motions of the molecule, leaving the three vibrational modes shown. After fitting H_2O , keep the same force constants and change only the isotope masses: the resulting D_2O and HDO frequencies agree with the observed values to about one percent.

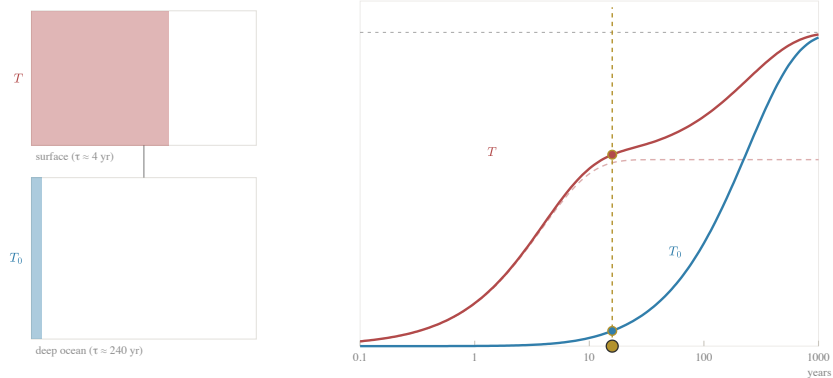


Figure 9.7 In the live figure, drag the marker along the logarithmic time axis after a step of forcing. The two-box ocean–atmosphere model responds on two time scales: the surface adjusts over about four years while the deep ocean changes little, then both boxes continue warming together over roughly two and a half centuries. The surface completes about 60% of its eventual warming during the fast part.

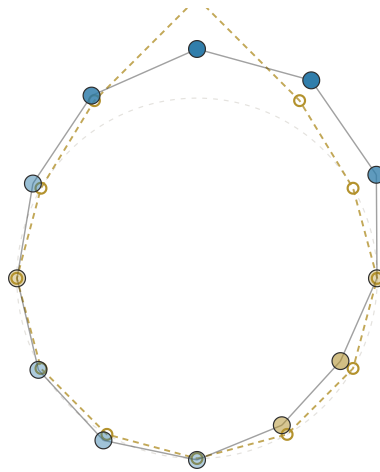


Figure 9.8 In the live figure, drag a mass to create the gold target shape, then choose how many of the lowest-frequency modes to retain. A few modes recover the broad, smooth part of the shape. Reproducing the sharp corner requires the higher-frequency modes as well.

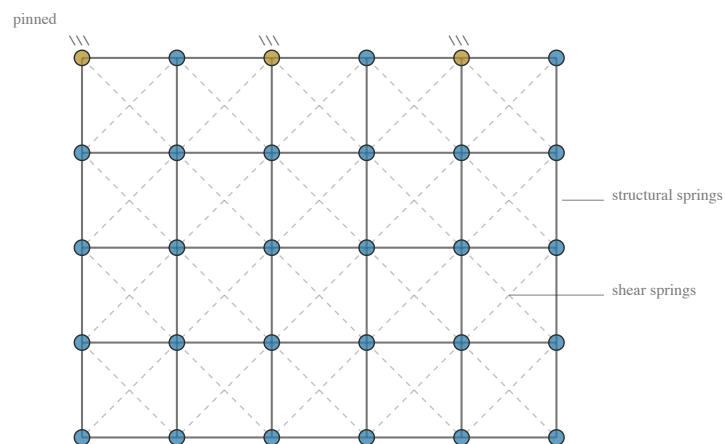


Figure 9.9 This curtain contains a few hundred point masses joined by structural and shear constraints. Gravity and a light breeze move those masses in real time; switch views to see the network beneath the rendered surface. Models of this kind are fast enough for interactive animation, but this example is not fitted to the behavior of a particular fabric. Drag the figure to orbit around it.

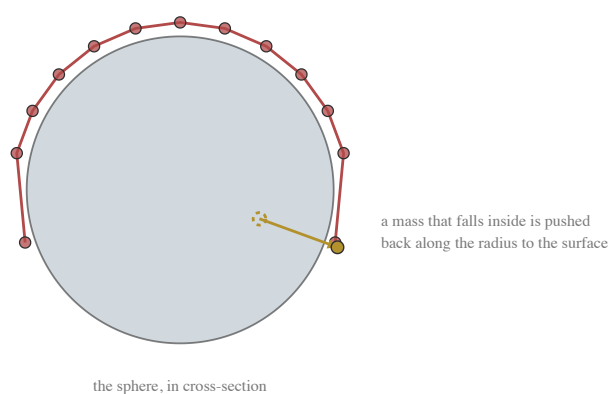


Figure 9.10 Drop the square sheet onto the sphere. Any mass which enters the sphere is pushed back to its surface and loses most of its speed on contact. Masses which are far apart on the sheet are also kept from passing through one another, allowing the folds to stack as the cloth settles. Replay the drop, switch to the mass-and-spring view, or drag to orbit around the final drape.

Forcing and Linear Response

10.1 The Homogeneous Family Behind a Forced Equation

- Open with Newton cooling supplied by a constant heater. The input shifts the equilibrium, while every solution differs from that equilibrium by a homogeneous cooling history.
- Compare the result with $A\mathbf{x} = \mathbf{b}$: one particular solution plus a solution of $A\mathbf{x} = 0$. For differential equations, prove that two solutions with the same forcing differ by a homogeneous solution, and conclude that every solution is one particular solution plus a homogeneous solution.
- Describe the forced solution family as an affine translate of the homogeneous solution space. Keep the language of null spaces, bases, and translates from ordinary linear algebra without introducing a differential operator or its kernel.
- Use this affine description as the starting point for constructing responses to time-dependent forcing later in the chapter.

10.2 Equilibrium Shifts in Linear Systems

- Build the greenhouse-air and rock-bed model $U' = KU + b$ from heat exchange. Find its equilibrium U_* and shift to the deviation $V = U - U_*$, which satisfies the homogeneous system $V' = KV$.
- Use the already-developed homogeneous flow to write $U(t) = U_* + \Phi(t)(U_0 - U_*)$. Interpret this as an equilibrium plus a freely evolving deviation.

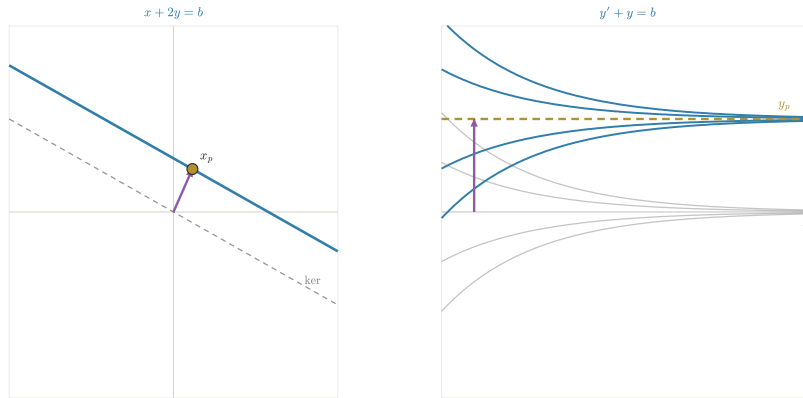


Figure 10.1 In the live figure, change b in either panel. The solutions of $x + 2y = b$ form a line obtained by translating the null space of $x + 2y = 0$. The solutions of $y' + y = b$ form a family obtained by translating the homogeneous solutions of $y' + y = 0$. Changing b translates both solution sets while their homogeneous parts remain fixed.

- Keep the application focused on the relationship between forced and free response. Do not repeat the construction of the flow matrix or the earlier development of eigenvalues, eigenvectors, and modes.
- Retain connected mixing tanks as a possible substantial problem: derive the forced system, find its equilibrium, subtract it, and reuse its homogeneous flow.
- Use the equilibrium shift as a simple entry into the chapter's larger question: how does a linear system transform a changing external input into an observable response?

PART III

Low-Dimensional Nonlinear Dynamics

CHAPTER 11

Beyond Linear Flows

CHAPTER 12

Local Nonlinear Dynamics

CHAPTER 13

Cycles in the Plane

CHAPTER 14

Bifurcations of Equilibria

CHAPTER 15

Bifurcations of Oscillations

Appendices

Calculus Review

Much of our work with differential equations asks us to recognize calculus we already know. Checking a proposed solution means differentiating it; finding one often means recognizing a chain rule or a product rule waiting to be undone. This appendix reviews those rules and the functions we most often use them on.

We will move briskly. Rather than collect a long table of derivatives, we will recover the familiar formulas from a few algebraic and geometric facts. The integration techniques will then follow by reversing those derivative calculations. Short questions along the way give you a chance to try the ideas; answers are collected at the end.

A.1 Differentiation Rules

The derivative records the first-order change of a function:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}.$$

When this limit exists, a small input change h produces the change $f'(x)h$, with an error whose ratio to h tends to zero. This is the local linear approximation behind tangent lines and Euler's method.

The limit definition immediately gives the derivative of a constant as zero and the derivative of x as 1. It also gives **linearity**:

$$(af + bg)' = af' + bg'$$

for constant scalars a, b . The derivative of a product requires more care. Both factors change, and we can separate those changes by writing

$$\begin{aligned} f(x+h)g(x+h) - f(x)g(x) &= (f(x+h) - f(x))g(x+h) \\ &\quad + f(x)(g(x+h) - g(x)). \end{aligned}$$

Divide by h and take the limit. Differentiability implies continuity, so $g(x+h) \rightarrow g(x)$, giving the **product rule**:

$$(fg)' = f'g + fg'.$$

Composition combines changes in a different way. An input change first passes through g , then through f . Their local scaling factors multiply, giving the **chain rule**:

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

The derivative of f is evaluated at the intermediate value $g(x)$, not at the original input x .

This also tells us how to differentiate an inverse. Suppose $g = f^{-1}$, so $f(g(x)) = x$. Differentiating gives

$$f'(g(x))g'(x) = 1, \quad g'(x) = \frac{1}{f'(g(x))}.$$

The inverse-function theorem justifies the differentiability used here: if f is continuously differentiable near a and $f'(a) \neq 0$, then f has a differentiable local inverse near $f(a)$, with

$$(f^{-1})'(f(a)) = \frac{1}{f'(a)}.$$

Geometrically, exchanging input and output exchanges the roles of rise and run. The nonzero derivative matters: the inverse of x^3 has a vertical tangent at zero rather than a finite derivative there.

Try this 1. Differentiate $f(x)f(x)g(x)$ using only the product rule, and then differentiate $f(g(x))g(x)$.

Try this 2. A continuously differentiable, invertible function satisfies $f(1) = 2$ and $f'(1) = 4$. What is $(f^{-1})'(2)$? Why does the formula ask for $f'(1)$ rather than $f'(2)$?

A.2 Elementary Functions

A.2.1 Powers, reciprocals, and roots

The product rule produces every positive-integer power derivative. The case $n = 1$ is already known. If $(x^n)' = nx^{n-1}$, then

$$(x^{n+1})' = (x^n x)' = nx^{n-1}x + x^n = (n+1)x^n.$$

Induction therefore proves the rule for all positive integers n . Linearity now differentiates any polynomial.

We can find other derivatives by differentiating identities. Let $r(x) = 1/x$ for $x \neq 0$. Since $xr(x) = 1$, the product rule gives

$$r(x) + xr'(x) = 0, \quad r'(x) = -\frac{1}{x^2}.$$

The chain rule consequently gives $(1/g)' = -g'/g^2$ wherever $g \neq 0$. Applying the product rule to $f/g = f(1/g)$ yields the **quotient rule**:

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}.$$

For the square root, set $s(x) = \sqrt{x}$ on $x > 0$. It is the inverse of squaring on the positive half-line, so the inverse rule ensures it is differentiable. Differentiate $s(x)s(x) = x$:

$$2s(x)s'(x) = 1, \quad s'(x) = \frac{1}{2\sqrt{x}}.$$

These identities give formulas on their stated domains. Neither dividing by x at zero nor dividing by $\sqrt{x}$ at zero is allowed. With the chain rule, however, we immediately obtain such calculations as

$$\frac{d}{dt}\sqrt{t^2 + 4} = \frac{2t}{2\sqrt{t^2 + 4}} = \frac{t}{\sqrt{t^2 + 4}},$$

which is valid for every real t .

A.2.2 Sine and cosine

The derivative formulas for sine and cosine come from the geometry of a small angle. Measure angles in radians. For $0 < h < \pi/2$, draw the unit circle with center O , the

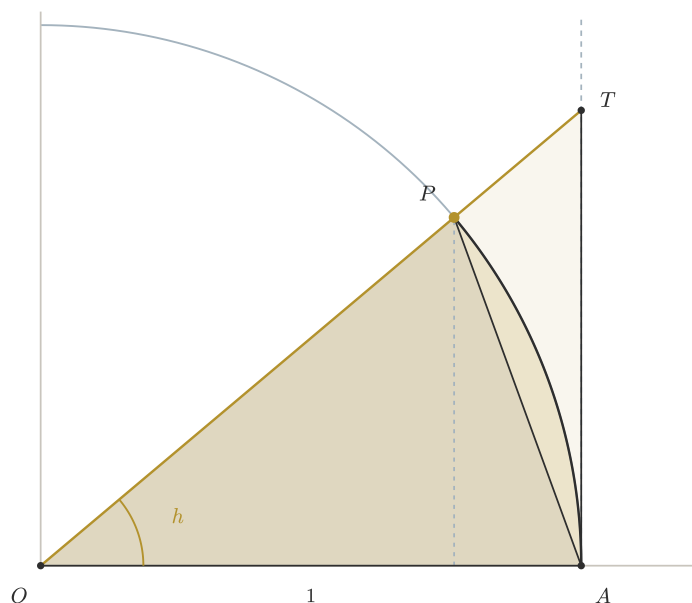


Figure A.1 The inscribed triangle lies inside the sector, and the sector inside the tangent triangle. Their common base OA has length one; the triangle heights are $\sin b$ and $\tan b$. Vary the angle to compare the three areas as b decreases. The area inequalities, rather than the appearance of the picture, establish the limit.

point $A = (1, 0)$, and $P = (\cos b, \sin b)$. Extend the ray OP to meet the tangent line $x = 1$ at $T = (1, \tan b)$.

The triangle OAP lies inside the circular sector OAP , which lies inside the triangle OAT . Comparing their areas gives

$$\frac{1}{2} \sin b < \frac{1}{2} b < \frac{1}{2} \tan b.$$

The sector has area $b/2$ because its angle is the fraction $b/(2\pi)$ of a full circle of area π . This is where radian measure enters.

Rearranging the inequalities gives

$$\cos b < \frac{\sin b}{b} < 1.$$

As $b \rightarrow 0^+$, the point P approaches A , so $\cos b \rightarrow 1$. The squeeze theorem gives the limit 1; because $\sin b/b$ is even, the same limit holds from the left. Using the circle identity $\sin^2 b + \cos^2 b = 1$ then gives

$$\frac{\cos b - 1}{b} = -\frac{\sin b}{b} \frac{\sin b}{1 + \cos b} \rightarrow 0.$$

Now the angle-addition identity supplies the derivative:

$$\begin{aligned} \frac{\sin(x+b) - \sin x}{b} &= \sin x \frac{\cos b - 1}{b} + \cos x \frac{\sin b}{b} \\ &\rightarrow \cos x. \end{aligned}$$

The cosine addition identity similarly gives $-\sin x$. Thus

$$(\sin x)' = \cos x, \quad (\cos x)' = -\sin x.$$

The quotient rule gives $(\tan x)' = 1/\cos^2 x$ wherever $\cos x \neq 0$. On $(-\pi/2, \pi/2)$, tangent has the inverse arctan, so the inverse derivative rule gives

$$(\arctan x)' = \cos^2(\arctan x) = \frac{1}{1+x^2}.$$

Here the last equality uses $1 + \tan^2 \theta = 1/\cos^2 \theta$.

A.2.3 Exponentials and logarithms

Begin with a familiar exponential a^x , where $a > 0$. We take the ordinary exponent laws and differentiability of these functions as known from calculus. The addition law separates the difference quotient:

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

Since $a^0 = 1$, the limit of the last factor is exactly the slope of a^x at zero. So the derivative of an exponential is the exponential itself, multiplied by its slope at zero.

Which base makes that extra factor equal to 1? That base is e : the number $e \approx 2.71828$ is characterized by

$$\lim_{b \rightarrow 0} \frac{e^b - 1}{b} = 1.$$

We take the existence of this base as known from calculus. The difference-quotient calculation above now gives

$$\frac{d}{dx}e^x = e^x.$$

The **natural exponential** e^x has both value 1 and slope 1 at zero, and its derivative equals its value everywhere.

The **natural logarithm**, written $\log x$ or $\ln x$, is the inverse of this exponential. Its domain is $x > 0$, and

$$e^{\log x} = x, \quad \log(e^t) = t.$$

The inverse rule now does the work:

$$(\log x)' = \frac{1}{e^{\log x}} = \frac{1}{x}.$$

The exponential addition law also gives $\log(ab) = \log a + \log b$ for positive a, b . For another positive base a , write $a = e^{\log a}$ and use the chain rule:

$$a^x = e^{x \log a}, \quad \frac{d}{dx}a^x = (\log a)a^x.$$

Likewise, for a real exponent α and $x > 0$,

$$x^\alpha = e^{\alpha \log x}, \quad \frac{d}{dx}x^\alpha = e^{\alpha \log x} \frac{\alpha}{x} = \alpha x^{\alpha-1}.$$

This extends the power rule beyond the integer case proved earlier. For integer powers we also have the larger domains already allowed by their algebraic definitions.

One form of the logarithm derivative is particularly useful when integrating. For $x < 0$, the chain rule gives $(\log(-x))' = (-1)/(-x) = 1/x$. Together with the positive case, this says

$$\frac{d}{dx} \log |x| = \frac{1}{x} \quad (x \neq 0), \quad \frac{d}{dt} \log |g(t)| = \frac{g'(t)}{g(t)} \quad (g(t) \neq 0).$$

Try this 3. Differentiate $e^{-t} \sin(2t)$ and $(\log t)/t$, stating where each formula is valid.

Try this 4. Differentiate $\log |1 - t^2|$. On which intervals is this function differentiable? Does a negative value of $1 - t^2$ prevent us from using the formula?

A.3 Integration and the Fundamental Theorem

An integral adds up a rate over an interval. For continuous f , the definite integral $\int_a^b f(s) ds$ is the limit of sums of values of f times small interval widths. Contributions below zero carry a negative sign. Reversing the bounds reverses the sign, and an interval of zero length gives zero.

Fix a and let the other endpoint vary:

$$F_a(t) = \int_a^t f(s) ds.$$

We have defined a function of t . The letter s is a dummy variable which runs through the integral; renaming it does not change the answer. The subscript on F_a records the base point, which stays fixed.

Theorem A.1 (Fundamental Theorem of Calculus). Let f be continuous on an interval I and let $a \in I$. Then $F_a(t) = \int_a^t f(s) ds$ satisfies $F'_a(t) = f(t)$ at interior points of I . If $G' = f$ on I , then for $a, b \in I$,

$$\int_a^b f(s) ds = G(b) - G(a).$$

The first statement comes from

$$\frac{F_a(t+h) - F_a(t)}{h} = \frac{1}{h} \int_t^{t+h} f(s) ds.$$

This is the average value of f over a shrinking interval, so continuity makes its limit $f(t)$. For the second statement, $G - F_a$ has derivative zero and hence is constant by the mean value theorem. Evaluating that constant at a gives $G(t) - F_a(t) = G(a)$.

A.3.1 Which integral are we writing?

An **antiderivative** of f is a function G with $G' = f$. Any two antiderivatives differ by a constant on an interval, since their difference has derivative zero. The **indefinite integral** notation records the whole family:

$$\int f(t) dt = G(t) + C.$$

There are no bounds because no value or base point has yet been selected. By contrast, a definite integral with fixed bounds gives a number, while a variable bound gives a particular function. For the same integrand,

$$\int 2t \, dt = t^2 + C, \quad \int_0^2 2s \, ds = 4, \quad \int_0^t 2s \, ds = t^2.$$

The last expression selects the antiderivative which is zero at $t = 0$. There is no missing $+C$ in a definite integral: its bounds have already fixed its value. To describe every antiderivative, we add a constant *outside* the definite integral:

$$y(t) = C + \int_0^t f(s) \, ds.$$

This will be our default form when zero belongs to the interval under consideration. In this form $C = y(0)$. There is nothing mandatory about zero, however. We can choose any base point a in the interval and write

$$y(t) = C + \int_a^t f(s) \, ds, \quad C = y(a).$$

Changing the base point changes the constant needed to describe the same function. An initial condition $y(t_0) = y_0$ suggests an especially convenient choice:

$$y(t) = y_0 + \int_{t_0}^t f(s) \, ds.$$

If we keep another base point a , then $C = y_0 - \int_a^{t_0} f(s) \, ds$. If instead we start from an arbitrary antiderivative G , the formula $y = G + C$ requires $C = y_0 - G(t_0)$. The symbol C does not have a fixed numerical meaning independent of how we write the antiderivative.

For example, the problem $y' = 2t$, $y(1) = 3$ from Chapter 2 can be written in all three ways:

$$y(t) = t^2 + 2 = 2 + \int_0^t 2s \, ds = 3 + \int_1^t 2s \, ds.$$

The constants 2 and 3 differ because the integrals have different base points. Both expressions describe the same function.

An integral also defines an explicit function when no elementary antiderivative is available. For instance,

$$H(t) = \int_0^t e^{-s^2} ds$$

is a particular function with $H(0) = 0$ and $H'(t) = e^{-t^2}$. The integral specifies its value at every t ; an elementary formula is not required. Numerical integration approximates those values, while the Fundamental Theorem gives its derivative exactly.

A.3.2 Logarithms and variable endpoints

We defined the logarithm as an inverse function and proved that its derivative is $1/x$. The Fundamental Theorem now identifies its integral representation:

$$\int_1^x \frac{du}{u} = \log x - \log 1 = \log x, \quad x > 0.$$

Thus the inverse of the exponential also measures accumulated area under $1/x$. Notice the base point 1. We could not replace it by zero here: the integrand is undefined there and the integral diverges. Base points must respect the interval on which the calculation makes sense.

Finally, combine the Fundamental Theorem with the chain rule. If $u(t)$ and $v(t)$ are differentiable and their values lie in the interval of continuity of f , then

$$\frac{d}{dt} \int_{u(t)}^{v(t)} f(s) ds = f(v(t))v'(t) - f(u(t))u'(t).$$

One way to see both terms is to write the integral as $F_a(v(t)) - F_a(u(t))$. No antiderivative formula for f is needed.

Try this 5. Solve $y' = \cos t$, $y(\pi/2) = 3$. Write the answer as an elementary antiderivative plus a constant, as a constant plus an integral from 0 to t , and as an integral based at the initial time.

Try this 6. Differentiate $\int_t^{t^2} e^{-s^2} ds$ without evaluating the integral.

A.4 Substitution and Integration by Parts

The chain rule says that if $F' = f$, then $(F(g(t)))' = f(g(t))g'(t)$. Read backward, it gives

$$\int f(g(t))g'(t) dt = F(g(t)) + C.$$

The substitution notation $u = g(t)$, $du = g'(t) dt$ records this same calculation. Its purpose is to recognize an outer function together with the derivative of its input. For example,

$$\int \frac{2t}{1+t^2} dt = \int \frac{du}{u} = \log u + C = \log(1+t^2) + C, \quad u = 1+t^2.$$

We do not need an absolute value in the final expression because $1+t^2 > 0$. A substitution in a definite integral also changes its bounds:

$$\int_a^b f(g(t))g'(t) dt = \int_{g(a)}^{g(b)} f(u) du.$$

For this form of substitution, f continuous and g continuously differentiable suffice; g need not be one-to-one. The identity follows by evaluating $F(g(b)) - F(g(a))$ on both sides. Reversals in direction are already accounted for by the signed derivative and bounds.

For a product, $(uv)' = u'v + uv'$. Integrating and rearranging gives **integration by parts**:

$$\int u(t)v'(t) dt = u(t)v(t) - \int u'(t)v(t) dt.$$

Indefinite integrals here represent antiderivative families; one arbitrary constant in the final answer accounts for the freedom on both sides. With fixed bounds the precise evaluation is

$$\int_a^b uv' dt = u(b)v(b) - u(a)v(a) - \int_a^b u'v dt.$$

The useful choice is one for which differentiating u simplifies it, while integrating v' is possible. For $\int te^t dt$, choose $u = t$ and $v' = e^t$. Then

$$\int te^t dt = te^t - \int e^t dt = (t-1)e^t + C.$$

Sometimes both techniques are needed in one calculation. Assignment 2 led to

$$\int 2t^3 e^{t^2} dt.$$

The exponential suggests $u = t^2$, but we must transform the entire integrand: $2t^3 dt = t^2(2t dt) = u du$. Thus

$$\int 2t^3 e^{t^2} dt = \int u e^u du = (u - 1)e^u + C = (t^2 - 1)e^{t^2} + C.$$

The substitution exposes the product which integration by parts can simplify. Differentiating the final answer checks both steps at once.

Try this 7. Use integration by parts to find $\int t \log t dt$ on $t > 0$.

Try this 8. Evaluate $\int_0^1 t/(1+t^2) dt$ by substitution, changing the bounds when you change variables.

A.5 Partial Fractions

Sometimes the calculus becomes easy after an algebraic rearrangement. For the logistic equation in Chapter 4 we needed to integrate

$$\frac{1}{x(1 - x/K)}, \quad K > 0.$$

Its denominator has two distinct linear factors. Try expressing it as a sum with one factor in each denominator:

$$\frac{1}{x(1 - x/K)} = \frac{A}{x} + \frac{B}{K - x}.$$

Multiplying by $x(K - x)$ turns the rational identity into a polynomial identity:

$$K = A(K - x) + Bx = AK + (B - A)x.$$

Matching the constant and linear coefficients gives $A = B = 1$. We have therefore reduced the integral to

$$\int \frac{dx}{x(1 - x/K)} = \log |x| - \log |K - x| + C.$$

The minus sign in the second logarithm comes from the derivative of $K - x$. Differentiation verifies the decomposition and its integral. The answer applies on each interval avoiding 0 and K ; constants on different intervals need not agree.

The general method follows the factorization of the denominator. First use polynomial division if the numerator's degree is at least the denominator's. For the remaining proper fraction, each linear factor $x - a$ contributes a term $A/(x - a)$. A repeated factor $(x - a)^m$ requires all powers through m :

$$\frac{A_1}{x-a} + \frac{A_2}{(x-a)^2} + \cdots + \frac{A_m}{(x-a)^m}.$$

A quadratic factor $q(x)$ with no real root requires a linear numerator $(Bx + C)/q(x)$; if it is repeated, include such a numerator over each power of q . Multiply by the common denominator and match coefficients to find the unknown constants.

Linear-factor terms integrate using powers and logarithms. For a single irreducible quadratic, split its numerator into a multiple of q' and a constant. The first part integrates to a logarithm; completing the square in the remaining part leads to arctangent. The basic identity is

$$\int \frac{dx}{(x-b)^2 + a^2} = \frac{1}{a} \arctan \frac{x-b}{a} + C, \quad a > 0,$$

which follows by substitution and the arctangent derivative. More elaborate repeated-quadratic integrals need further reductions; the linear factors and single quadratics cover the calculations we need here.

Partial fractions does not introduce a new differentiation rule. It rewrites a rational function until the derivatives we already understand become visible.

Try this 9. Decompose and integrate $1/[x(x+1)]$. State the intervals on which your antiderivative formulas apply.

Try this 10. Find an antiderivative of $(2x+3)/(x^2+2x+5)$ by separating a derivative-of-the-denominator term and then completing the square.

Answers to the Review Questions

1. The derivatives are $2ff'g + f^2g'$ and $f'(g(x))g'(x)g(x) + f(g(x))g'(x)$.
2. $(f^{-1})'(2) = 1/4$. The inverse takes input 2 back to 1, so we use the derivative of f at that recovered input.
3. The first derivative is $e^{-t}(2\cos(2t) - \sin(2t))$, valid for every real t . The second is $(1 - \log t)/t^2$, valid on $t > 0$.
4. The derivative is $-2t/(1-t^2)$ on $(-\infty, -1)$, $(-1, 1)$, and $(1, \infty)$. The absolute value allows negative values of $1-t^2$; only its zeros are excluded.
5. All three expressions are

$$y(t) = \sin t + 2 = 2 + \int_0^t \cos s \, ds = 3 + \int_{\pi/2}^t \cos s \, ds.$$

6. The derivative is $2te^{-t^4} - e^{-t^2}$. Both endpoint derivatives are needed.

7. Choose $u = \log t$ and $v' = t$, so $u' = 1/t$ and $v = t^2/2$. Then

$$\int t \log t \, dt = \frac{t^2}{2} \log t - \int \frac{t}{2} \, dt = \frac{t^2}{2} \log t - \frac{t^2}{4} + C.$$

8. Set $u = 1 + t^2$, so $du = 2t \, dt$. The new bounds are 1 and 2, giving $\frac{1}{2} \int_1^2 du/u = \frac{1}{2} \log 2$.

9. $1/[x(x+1)] = 1/x - 1/(x+1)$, so an antiderivative is $\log |x| - \log |x+1| + C$. The intervals are $(-\infty, -1)$, $(-1, 0)$, and $(0, \infty)$.

10. The numerator is $(2x+2)+1$ and the denominator is $(x+1)^2+4$. Thus an antiderivative, valid on all of $\mathbb{R}$, is

$$\log(x^2 + 2x + 5) + \frac{1}{2} \arctan \frac{x+1}{2} + C.$$

Further Reading

OpenStax's freely available *Calculus* gives additional worked practice: Volume 1, Chapter 3 reviews differentiation, and Volume 1, Chapter 5 develops integration and the Fundamental Theorem. Volume 2, Chapter 3 extends the integration techniques, including integration by parts and partial fractions.

Linear Algebra Review

You have seen linear algebra before, but perhaps it has been a while since you last found a basis or computed a kernel. This appendix is a chance to recover those ideas and put them to work in a setting where the vectors may be functions. We will move back and forth between the familiar arithmetic of lists of numbers and the arithmetic of polynomials and other functions. The point is to recognize the same mathematics in both places.

Work through the calculations with a pencil, and pause at the short review questions before reading their answers at the end. We use real numbers as scalars until the final section, where complex eigenvalues will give us a reason to allow complex coordinates too.

B.1 Vectors and Vector Spaces

What can we do with vectors? In $\mathbb{R}^2$, we can add them and multiply them by numbers. If $u = (1, 2)$ and $v = (3, -1)$, then

$$2u - v = 2(1, 2) - (3, -1) = (-1, 5).$$

Each coordinate follows ordinary arithmetic. Nothing about this calculation depends on drawing arrows, and that is useful: we would like the same arithmetic to work for objects we cannot conveniently draw as arrows.

For instance, take the functions $p(t) = 1 + 2t$ and $q(t) = 3 - t$. We can make exactly the same calculation:

$$(2p - q)(t) = 2(1 + 2t) - (3 - t) = -1 + 5t.$$

The constant coefficients combine just as the first coordinates did, and the coefficients of t combine just as the second coordinates did. The functions are different objects from their coefficient pairs, but their arithmetic has the same structure.

More generally, if f and g are real-valued functions on the same interval I , define

$$(f + g)(t) = f(t) + g(t), \quad (cf)(t) = cf(t) \quad \text{for every } t \in I.$$

These operations are called **pointwise** addition and scalar multiplication: we perform the numerical operation separately at each input. The result is another whole function. In particular, the zero vector in this setting is the function which is zero at every input, not merely a function whose graph passes through the origin.

Matrices admit the same kind of arithmetic. Two real 2×2 matrices can be added entry by entry, and each entry can be multiplied by a scalar. They therefore give us another collection of objects with vector arithmetic. Matrix multiplication is an additional operation; it is not needed to make matrices into vectors.

What must this arithmetic satisfy? We need sums and scalar multiples to stay inside the collection, and we need the familiar algebraic rules to continue working.

Definition B.1 (Real vector space). A **real vector space** is a set V with an addition operation and an operation of multiplication by real scalars. For $u, v, w \in V$ and $a, b \in \mathbb{R}$, the results $u + v$ and au belong to V , and the following rules hold:

$$\begin{aligned} u + v &= v + u, & (u + v) + w &= u + (v + w), \\ a(u + v) &= au + av, & (a + b)u &= au + bu, \\ a(bu) &= (ab)u, & 1u &= u. \end{aligned}$$

There is a zero vector $0 \in V$ with $u + 0 = u$, and every vector u has an additive inverse $-u$ with $u + (-u) = 0$.

For functions with pointwise operations, these rules come from the rules for real numbers. For example, at every t ,

$$(a(f + g))(t) = a(f(t) + g(t)) = af(t) + ag(t) = (af + ag)(t).$$

Since the two functions agree at every input, they are the same function. This proves the distributive rule for functions.

We will often restrict the functions we allow. Write P_n for the space of real polynomials of degree *at most* n , including the zero polynomial. The phrase "at most" matters: two quadratic polynomials can add to a constant or to zero, so polynomials

of degree exactly two would not form a vector space. We also use $C(I)$ for continuous real-valued functions on I , and $C^k(I)$ for functions with continuous derivatives through order k . Adding or scaling such functions preserves the required continuity and differentiability, so these are vector spaces too.

These function spaces can be much larger than $\mathbb{R}^n$. No finite list of polynomials can span all polynomials: if the largest degree in the list is m , every linear combination still has degree at most m , and misses t^{m+1} . We call a vector space **infinite-dimensional** when it has no finite basis. We will review basis and dimension shortly; for now, the point is that no fixed finite list of coefficients describes every function in a general function space. Our calculations will take place in smaller finite-dimensional spaces inside it.

Try this 1. For $f(t) = 1 - t$ and $g(t) = t + t^2$, compute $3f - 2g$. What is the additive inverse of f ? Is f the zero function because $f(1) = 0$?

Try this 2. Explain why real 2×3 matrices form a vector space under entrywise operations. Does their rectangular shape cause a problem for any vector-space rule?

B.2 Subspaces and Nonlinear Subsets

Once we have a vector space, we can ask whether a smaller collection inside it still supports the same arithmetic. Consider the line

$$U = \{(x, y) \in \mathbb{R}^2 : y = 2x\}.$$

Its vectors have the form $(a, 2a)$. Adding two gives $(a, 2a) + (b, 2b) = (a+b, 2(a+b))$, which stays on the line. Scaling gives $c(a, 2a) = (ca, 2ca)$, which stays there too. The zero vector lies on the line, so we have a vector space inside $\mathbb{R}^2$.

Definition B.2 (Subspace). A **subspace** of a vector space V is a subset $U \subseteq V$ which is itself a vector space using the addition and scalar multiplication from V .

We do not need to check every vector-space rule again. Commutativity and distributivity already hold in V , so they hold for members of U . What could fail is that a required result leaves the smaller set.

Proposition B.3 (Subspace test). A subset U of a real vector space V is a subspace if and only if it contains 0 and, for every $u, v \in U$ and $a, b \in \mathbb{R}$, it contains $au + bv$.

The condition includes sums, scalar multiples, and additive inverses. It is often convenient to check addition and scalar multiplication separately. To show that a set fails the test, however, just one counterexample is enough.

Move our line upward by one unit:

$$H = \{(x, y) \in \mathbb{R}^2 : y = 2x + 1\}.$$

This set does not contain $(0, 0)$, so it cannot be a subspace. Adding its point $(0, 1)$ to itself gives $(0, 2)$, which is not on the line. Yet the line still has a simple relationship to U : every member is $(0, 1) + u$ for some $u \in U$. It is a translated subspace, an **affine space**. We will return to that structure in section B.6.

Now consider a different subset:

$$Q = \{(x, y) \in \mathbb{R}^2 : y = x^2\}.$$

This parabola does contain zero. Does that make it a subspace? The point $(1, 1)$ belongs to Q , but its double $(2, 2)$ does not: the parabola requires the second coordinate to be $2^2 = 4$. Knowing one point has not given us its scalar multiples. Addition fails for the same pair of points.

The parabola is not a translated line either. Any translated subspace contains the midpoint of each pair of its points: translating the points back to the subspace, averaging, and translating forward stays in the set. But the midpoint of $(0, 0)$ and $(1, 1)$ is $(1/2, 1/2)$, and $1/2 \neq (1/2)^2$. Thus Q is neither a vector subspace nor an affine space.

We can see all three possibilities among functions. In $C(\mathbb{R})$, the functions satisfying $f(0) = 0$ form a subspace, since

$$(af + bg)(0) = af(0) + bg(0) = 0.$$

The functions satisfying $f(0) = 1$ form a translate of that subspace: subtract the constant function 1 and the value at zero becomes zero. They do not form a subspace themselves.

For a nonlinear counterpart, consider the following family inside P_1 :

$$\mathcal{Q} = \{f_a(t) = a + a^2t : a \in \mathbb{R}\}.$$

Its coefficient pairs lie on exactly the parabola we just studied. The zero function belongs to this family, as does $1 + t$, but $2 + 2t$ does not: its constant coefficient would force $a = 2$, while its coefficient of t would then need to be 4. The midpoint function

$(1 + t)/2$ fails for the same reason. These functions live in a vector space, but this particular family is neither a subspace nor an affine space.

This is one difficulty we should expect when studying nonlinear solution sets. Even if we know several members, their linear combinations need not belong to the set. A curve may have a simple parameterization, as our parabola does, without allowing the arithmetic of a vector space. Linear algebra alone no longer reconstructs the family from a basis of solutions.

We should test the set rather than judge only the appearance of its defining equation. The nonlinear-looking condition $y^2 = 0$ still describes the subspace $y = 0$. Nonlinear equations do not automatically have solution sets of one particular shape.

Try this 3. Is the set of even continuous functions a subspace of $C(\mathbb{R})$? Recall that even means $f(-t) = f(t)$ for every t . What about the set of functions with $f(0) \geq 0$?

Try this 4. The circle $x^2 + y^2 = 1$ is a subset of $\mathbb{R}^2$. Give one reason it is not a subspace and a midpoint counterexample showing that it is not an affine space either.

B.3 Bases and Dimension

Suppose we know a few vectors and want to generate others from them. With $u = (1, 1)$ and $v = (1, -1)$, can we construct $(3, 1)$? Writing

$$au + bv = (a + b, a - b) = (3, 1)$$

gives $a + b = 3$ and $a - b = 1$. Adding the equations gives $a = 2$, and then $b = 1$. Thus $(3, 1) = 2u + v$.

An expression $a_1v_1 + \dots + a_kv_k$ is a **linear combination** of $v_1, \dots, v_k$. The **span** of these vectors is the set of all their linear combinations, with arbitrary real coefficients. Their span is always a subspace: adding or scaling combinations just changes the coefficients. It is the smallest subspace containing the given vectors.

Our two vectors span all of $\mathbb{R}^2$. Indeed, the equations $a + b = x$ and $a - b = y$ have the solution

$$a = \frac{x + y}{2}, \quad b = \frac{x - y}{2}$$

for every (x, y) . The representation is also unique. To understand that second feature, suppose the same vector has two representations. Subtracting them produces a linear combination of u and v equal to zero. The question of uniqueness is therefore a question about combinations producing zero.

Definition B.4 (Linear independence). The vectors $v_1, \dots, v_k$ are **linearly independent** if

$$a_1 v_1 + \dots + a_k v_k = 0$$

forces $a_1 = \dots = a_k = 0$. Otherwise they are **linearly dependent**.

For u and v above, a zero combination gives $a + b = 0$ and $a - b = 0$, so both coefficients vanish. They are independent. If we added the third vector $w = (2, 0)$ to our list, independence would fail because $u + v - w = 0$. The third vector contributes no new direction to the span.

Definition B.5 (Basis and dimension). A **basis** of a vector space is a linearly independent collection which spans the space. Every vector then has a unique expression as a linear combination of the basis vectors. In a finite-dimensional space, every basis has the same number of vectors; that number is the **dimension**.

Spanning supplies existence of the representation; independence supplies uniqueness. These are different jobs, and both are necessary.

For example, $1, t, t^2$ form a basis of P_2 . They span it because every member has the form $a + bt + ct^2$. To check independence, suppose

$$a + bt + ct^2 = 0 \quad \text{for every } t \in \mathbb{R}.$$

Evaluating at 0 gives $a = 0$. Differentiating the identity and evaluating at 0 gives $b = 0$, and differentiating twice gives $2c = 0$. Thus all three coefficients vanish, so $\dim P_2 = 3$.

The words "for every t " are essential. Independence concerns the functions, not their values at one chosen input. All three functions t, t^2, t^3 vanish at zero, yet they are independent as functions. One evaluation discards most of the information in a function.

In the ordered basis $\mathcal{B} = (1, t, t^2)$, the polynomial $p(t) = 2 + 3t + t^2$ has coordinate vector

$$[p]_{\mathcal{B}} = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}.$$

Now use $\mathcal{C} = (1, 1 + t, (1 + t)^2)$. To find the new coordinates, write

$$a + b(1 + t) + c(1 + t)^2 = (a + b + c) + (b + 2c)t + ct^2.$$

Matching coefficients with $2 + 3t + t^2$ gives $c = 1$, $b = 1$, and $a = 0$. Hence

$$[p]_{\mathcal{C}} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

The same coefficient matching works uniquely for every polynomial in P_2 , so $\mathcal{C}$ is another basis. The polynomial did not change; the numbers describing it changed because we changed the basis.

A few standard finite-dimensional facts are useful to remember. In a space of dimension n , an independent list has at most n vectors, and a spanning list has at least n . An independent list of exactly n vectors is automatically a basis, as is a spanning list of exactly n vectors. An independent list can be extended to a basis; a spanning list can have redundant vectors removed until it becomes a basis. These facts formalize the idea that dimension counts independent directions, not the number of entries used to write a particular object.

Try this 5. Do $1 + t$, $1 - t$, and t^2 form a basis of P_2 ? If so, find the coordinates of $2 + 3t + t^2$ in this ordered basis.

Try this 6. Let U be the subspace of $\mathbb{R}^3$ satisfying $x + y + z = 0$. Find a basis and its dimension. Why does writing its vectors with three entries not make its dimension three?

B.4 Linear Maps, Kernels, and Images

We have been combining vectors inside a space. What happens when we apply an operation to them? A matrix sends an input vector to an output vector; differentiation sends an input function to another function. Both respect the arithmetic we have been reviewing:

$$A(au + bv) = aAu + bAv, \quad (af + bg)' = af' + bg'.$$

Definition B.6 (Linear map). A map $T : V \rightarrow W$ between real vector spaces is **linear** if

$$T(au + bv) = aT(u) + bT(v)$$

for every $u, v \in V$ and $a, b \in \mathbb{R}$.

A linear map necessarily sends zero to zero. A map which adds a fixed nonzero vector, such as $x \mapsto x + 1$ on $\mathbb{R}$, fails this test. Sending zero to zero is necessary, but not sufficient: $x \mapsto x^2$ does that and still fails linearity.

The domain and codomain are part of a map's specification. Consider

$$D : P_2 \longrightarrow P_1, \quad Dp = p'.$$

Differentiation sends a polynomial of degree at most two to one of degree at most one. If $p = a + bt + ct^2$, then $Dp = b + 2ct$. Which inputs disappear, and which outputs can occur?

Definition B.7 (Kernel and image). For a linear map $T : V \rightarrow W$, its **kernel** and **image** are

$$\ker T = \{v \in V : T(v) = 0\}, \quad \text{im } T = \{T(v) : v \in V\}.$$

The kernel is a subspace of V , and the image is a subspace of W .

For the kernel, the subspace test follows from $T(au + bv) = aT(u) + bT(v) = 0$ whenever u, v lie in the kernel. For the image, a combination of outputs $aT(u) + bT(v)$ is the output $T(au + bv)$. Both contain zero.

In our differentiation example, $Dp = 0$ forces $b = c = 0$, so

$$\ker D = \text{span}\{1\}.$$

Every polynomial $r + st$ in P_1 is the derivative of $rt + (s/2)t^2$. Therefore $\text{im } D = P_1$. Differentiation loses the constant coefficient, but it reaches every allowed output.

A map is **injective** when distinct inputs have distinct outputs, and **surjective** when every member of its codomain occurs as an output. For a linear map, injectivity is equivalent to having kernel $\{0\}$. Indeed, $T(u) = T(v)$ is equivalent to $T(u - v) = 0$. A nonzero kernel vector is exactly an ambiguity in recovering the input from its output.

Thus $D : P_2 \rightarrow P_1$ is surjective but not injective. If we instead declare the codomain to be P_2 , the same differentiation rule is no longer surjective: no derivative of a member of P_2 equals t^2 . If we restrict the domain to the polynomials in P_2 with zero constant coefficient, then differentiation becomes both injective and surjective onto P_1 . The spaces named on either side of the arrow matter!

A linear map which is both injective and surjective is a **linear isomorphism**. Its inverse is linear too: recovering inputs from a combination of outputs gives the corresponding combination of inputs. The coordinate map $p \mapsto [p]_{\mathcal{B}}$ is an isomorphism from P_2 to $\mathbb{R}^3$. This gives a precise meaning to saying that their linear algebra is the same.

Evaluation is another useful example:

$$E_0 : C(\mathbb{R}) \longrightarrow \mathbb{R}, \quad E_0(f) = f(0).$$

This map is linear because $(af + bg)(0) = af(0) + bg(0)$. Its kernel is the subspace of functions vanishing at zero. It is surjective because constant functions attain any requested output. It is far from injective: many different functions have the same value at zero.

For finite-dimensional domains, there is a useful accounting rule.

Theorem B.8 (Rank–nullity). If $T : V \rightarrow W$ is linear and V is finite-dimensional, then

$$\dim V = \dim \ker T + \dim \operatorname{im} T.$$

The two dimensions on the right are called the **nullity** and **rank**.

Here is why. Choose a basis $k_1, \dots, k_r$ of the kernel and extend it to a basis $k_1, \dots, k_r, v_1, \dots, v_s$ of V . Applying T to a linear combination erases the k_i terms, so $T(v_1), \dots, T(v_s)$ span the image. They are also independent. If $\sum_j a_j T(v_j) = 0$, then $\sum_j a_j v_j$ lies in the kernel and can be written as a combination of the k_i . Independence of the extended basis forces every a_j to be zero. Thus the image has dimension s , and $\dim V = r + s$ as claimed.

For polynomial differentiation, the count is $3 = 1 + 2$: one independent constant is lost and two independent output coefficients remain. We will use this numerical dimension formula only in finite-dimensional spaces.

Try this 7. For $T : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ given by $T(x, y, z) = (x + y, y + z)$, find a basis of the kernel and determine the image. Check rank–nullity. Is T injective? Surjective?

Try this 8. Let $J : P_1 \rightarrow P_2$ send q to $Jq(t) = \int_0^t q(s) ds$. Find its kernel and image. Explain why it is an isomorphism onto the subspace of polynomials with zero constant coefficient, but not onto all of P_2 .

B.5 Matrices and Coordinates

How much of a linear map must we know to determine all of it? Suppose $v_1, \dots, v_n$ is a basis of its domain. Every input has a unique expression $v = c_1 v_1 + \dots + c_n v_n$, and linearity forces

$$T(v) = c_1 T(v_1) + \dots + c_n T(v_n).$$

We only need the images of the basis vectors. Once those are known, every other output is a linear combination with the same coefficients.

Choose an ordered basis $\mathcal{B} = (v_1, \dots, v_n)$ of the domain and $\mathcal{C} = (w_1, \dots, w_m)$ of the codomain. Place the coordinates of $T(v_j)$ into column j of a matrix M . The formula above becomes

$$[T(v)]_{\mathcal{C}} = M[v]_{\mathcal{B}}.$$

This is why matrix multiplication takes a linear combination of columns: the columns describe what happens to basis vectors, and the input coordinates tell us how much of each image to use. A map from an n -dimensional space to an m -dimensional space has an $m \times n$ matrix in the chosen bases.

For $D : P_2 \rightarrow P_1$, use the bases $(1, t, t^2)$ and $(1, t)$. The basis images are

$$D(1) = 0, \quad D(t) = 1, \quad D(t^2) = 2t.$$

Their coordinate columns give

$$M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Multiplying confirms the meaning:

$$M \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} b \\ 2c \end{pmatrix}.$$

The output column records the polynomial $b + 2ct$, the derivative of $a + bt + ct^2$. The zero first column also makes the lost constant visible.

Composition explains matrix multiplication. If $T : U \rightarrow V$ has matrix A and $S : V \rightarrow W$ has matrix B , using the same basis of V for the intermediate coordinates, then

$$[S(T(u))] = B(A[u]) = (BA)[u].$$

The rightmost map acts first. For example, differentiation from P_1 to P_0 has matrix $N = \begin{pmatrix} 0 & 1 \end{pmatrix}$. Differentiating twice from P_2 to P_0 therefore has matrix

$$NM = \begin{pmatrix} 0 & 0 & 2 \end{pmatrix},$$

which sends (a, b, c) to $2c$. Order matters, and in this example the reverse matrix product is not even defined because the dimensions do not match. Even for square matrices, the products AB and BA need not agree.

An isomorphism between finite-dimensional spaces has an invertible square matrix. Its inverse matrix records the inverse map. In particular, changing basis is invertible: we have changed how we describe a vector, not discarded any information about it.

Return to the two bases $\mathcal{B} = (1, t, t^2)$ and $\mathcal{C} = (1, 1+t, (1+t)^2)$ of P_2 . Put the $\mathcal{B}$ -coordinates of the $\mathcal{C}$ basis vectors into columns:

$$P = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then $[p]_{\mathcal{B}} = P[p]_{\mathcal{C}}$. Solving the triangular system gives

$$P^{-1} = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Applying P^{-1} to $(2, 3, 1)$ recovers $(0, 1, 1)$, as in our earlier calculation. For a map $T : V \rightarrow V$ with matrix A in the old basis, the matrix in the new basis is

$$A_{\text{new}} = P^{-1}AP.$$

Read this from right to left: convert the input to old coordinates, apply the old matrix, then convert the output to new coordinates. This explains the formula without memorizing where the inverse goes.

Try this 9. Find the matrix of $J : P_1 \rightarrow P_2$ from Try this 8 in the standard polynomial bases. Multiply the matrices to check that $D \circ J$ is the identity on P_1 . What does $J \circ D$ do to a polynomial in P_2 ?

Try this 10. In $\mathbb{R}^2$, let $\mathcal{C} = ((1, 1), (1, -1))$. Write the matrix P which converts $\mathcal{C}$ -coordinates to standard coordinates. Find the $\mathcal{C}$ -coordinates of $(4, 2)$, and check by reconstructing the vector.

B.6 Linear Equations and Affine Spaces

We now have the language to describe an entire solution set, not just calculate one answer. Begin with

$$\begin{aligned}x + 2y &= 3, \\2x + 4y &= 6.\end{aligned}$$

Subtracting twice the first equation from the second leaves $0 = 0$. There is only one independent constraint. Taking $y = s$ gives

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 - 2s \\ s \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} + s \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \quad s \in \mathbb{R}.$$

One particular solution locates the line, while a kernel vector gives its direction. For

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix},$$

the kernel is $\text{span}\{(-2, 1)\}$. Moving in this direction changes the input without changing its output under A .

Why must this description work for every consistent linear equation? Let $T : V \rightarrow W$ be linear, and suppose v_p is one solution of $T(v) = b$. If v is any other solution, then

$$T(v - v_p) = T(v) - T(v_p) = b - b = 0.$$

Thus $v - v_p \in \ker T$. Conversely, if $b \in \ker T$, then $T(v_p + b) = b + 0 = b$. We have proved both directions of the statement:

Theorem B.9 (Solutions of a linear equation). For a linear map $T : V \rightarrow W$, the equation $T(v) = b$ has a solution exactly when $b \in \text{im } T$. If v_p is one solution, the full solution set is

$$\{v : T(v) = b\} = v_p + \ker T = \{v_p + b : b \in \ker T\}.$$

The notation $T^{-1}(\{b\})$ is also used for this set, the *preimage* of b . It does not assert that T has an inverse map. Our matrix A does not: each attainable output has a whole line of inputs.

An equation can also be inconsistent. If we change the second right-hand side in the matrix system to 7, subtraction leaves $0 = 1$. Equivalently, $(3, 7)$ is not in the image of A . There is no particular solution from which to start a translation.

Definition B.10 (Affine subspace). An **affine subspace** of a vector space V is a nonempty set of the form $v_p + U$, where U is a vector subspace. We also call this an affine space when the surrounding vector space is understood. If U is finite-dimensional, the affine space has dimension $\dim U$.

Every vector subspace is also an affine space: choose $v_p = 0$. An affine space contains zero exactly when it is a vector subspace. Indeed, if $0 = v_p + u$ for some $u \in U$, then $v_p = -u \in U$, and translation by v_p leaves U unchanged. When distinguishing solution sets, we can therefore say "a vector space," "affine but not a vector space," or "neither." The empty set counts as neither under our definition.

What arithmetic survives in a translated space? Let $v = v_p + u$ and $w = v_p + z$. For any real a ,

$$av + (1 - a)w = v_p + au + (1 - a)z$$

still belongs to $v_p + U$. More generally, linear combinations whose coefficients sum to 1 stay inside the affine space. These are **affine combinations**. Arbitrary sums and scalar multiples can fail, but midpoints and entire lines through pairs of points remain available. Differences of points lie in the direction subspace U .

The distinction is just as useful for functions. Prescribe a polynomial derivative:

$$Dp = 2 + 6t, \quad p \in P_2.$$

We can find one answer by antidifferentiating: $p_p(t) = 2t + 3t^2$. Since $\ker D$ consists of constants, every answer is

$$p(t) = 2t + 3t^2 + c, \quad c \in \mathbb{R}.$$

The familiar constant of integration is the kernel freedom! This is a one-dimensional affine space inside the three-dimensional vector space P_2 . It is not a vector

subspace, since the zero polynomial does not have the prescribed derivative. Subtracting any two answers gives a constant, while averaging two answers gives another answer.

Compare this with the family $a + a^2t$ from section B.2. Its failure to contain the midpoint of two of its members already rules out an affine description. There is no fixed subspace whose translate gives that parabola of functions. The ambient function space still lets us add and subtract functions, but the constraint need not respect those operations. That is the structural question we will ask of differential equations: which operations on known solutions produce new solutions of the same equation?

Try this 11. Find all (x, y, z) with $x + y + z = 1$. Write the answer as a particular vector plus a subspace, using your basis from Try this 6. Find its affine dimension. Does adding two solutions preserve the equation? Does taking their midpoint?

Try this 12. Describe all $p \in P_2$ with $p' = 1 + 2t$. Then impose $p(0) = 3$. Describe how the solution set changes. Is the final singleton a vector subspace, an affine space, or neither?

B.7 Eigenvalues and Eigenvectors

A linear map becomes a matrix once we choose a basis, and the preceding sections showed that we can choose different bases. Could a good choice make the map particularly simple? The best possible outcome would be for each basis vector to map to a multiple of itself. Then applying the map would scale each coordinate separately, without mixing them.

Consider

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

The standard basis vectors do not have that property: $A(1, 0) = (2, 1)$ and $A(0, 1) = (1, 2)$. But try the vectors $v_1 = (1, 1)$ and $v_2 = (1, -1)$:

$$Av_1 = (3, 3) = 3v_1, \quad Av_2 = (1, -1) = v_2.$$

Along one line the map triples vectors; along the other it leaves them unchanged. These vectors form a basis, so their two simple behaviors determine the action of A on every vector.

Definition B.11 (Eigenvalues and eigenvectors). Let $T : V \rightarrow V$ be linear. A nonzero vector v is an **eigenvector** with **eigenvalue** λ if

$$T(v) = \lambda v.$$

The **eigenspace** for an eigenvalue λ is $\ker(T - \lambda I)$, where I is the identity map.

Zero is excluded from being an eigenvector because $T(0) = \lambda 0$ for every λ ; it cannot identify a special scaling factor. Zero does belong to every eigenspace, together with the eigenvectors for that eigenvalue. The eigenvalue itself can be zero: a nonzero vector in the kernel is an eigenvector with eigenvalue zero.

How do we find these vectors without guessing them? For a square matrix, rewrite $Av = \lambda v$ as

$$(A - \lambda I)v = 0.$$

We need this equation to have a nonzero solution. A square matrix has a nonzero kernel exactly when it is singular, which is equivalent to its determinant being zero. Thus the possible eigenvalues satisfy

$$\det(A - \lambda I) = 0.$$

Recall that the determinant of $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $ad - bc$. For our matrix, the eigenvalue equation is

$$(2 - \lambda)^2 - 1 = (\lambda - 1)(\lambda - 3) = 0.$$

Once $\lambda = 3$ is known, solving $(A - 3I)v = 0$ gives $y = x$; the eigenspace is $\text{span}\{(1, 1)\}$. For $\lambda = 1$, the equations give $y = -x$, so the eigenspace is $\text{span}\{(1, -1)\}$. Finding eigenvalues supplies the scalars; finding the corresponding kernels supplies the vectors.

Place these eigenvectors into columns:

$$P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

The two eigenvector equations together say $AP = P\Lambda$. Because the columns of P are a basis, P is invertible, and we obtain

$$P^{-1}AP = \Lambda, \quad A = P\Lambda P^{-1}.$$

This is **diagonalization**: writing a linear map in a basis of eigenvectors makes its matrix diagonal. Conversely, any basis which makes the matrix diagonal must consist of eigenvectors, since each column then has only its own coordinate possibly nonzero. Thus a matrix is diagonalizable exactly when it has a basis of eigenvectors.

Eigenvectors for distinct eigenvalues are linearly independent. In particular, an $n \times n$ matrix with n distinct eigenvalues over the chosen scalar field has an eigenbasis. Repeated eigenvalues require us to check the dimensions of their eigenspaces; repetition alone does not decide whether diagonalization is possible.

Why is diagonalization useful? When we multiply powers, adjacent factors $P^{-1}P$ cancel, giving

$$A^k = P\Lambda^k P^{-1}, \quad \Lambda^k = \begin{pmatrix} 3^k & 0 \\ 0 & 1 \end{pmatrix} \quad (k = 0, 1, 2, \dots).$$

In eigenvector coordinates, repeated matrix multiplication is ordinary scalar exponentiation. For example, $(3, 1) = 2v_1 + v_2$, so $A^k(3, 1) = 2 \cdot 3^k v_1 + v_2$ without multiplying out A^k .

This is the algebra we will need when we study the matrix exponential in the main text: an eigenbasis makes all powers of a matrix act coordinate by coordinate.

We cannot assume such a basis exists. For

$$B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

the only eigenvalue is 1, but $(B - I)(x, y) = (y, 0)$, so its eigenspace is only the line $y = 0$. There are not two independent eigenvectors, and B cannot be diagonalized. By contrast, the identity matrix also has only eigenvalue 1, but every nonzero vector is an eigenvector, so it already is diagonal.

The scalar field matters too. The real rotation matrix

$$R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

turns each nonzero real vector through a right angle, so it has no real eigenvector. Its characteristic equation is $\lambda^2 + 1 = 0$. If we allow complex coordinates, with $i^2 = -1$, the eigenvalues are i and $-i$, with eigenvectors $(1, -i)$ and $(1, i)$ respectively. They form a basis of $\mathbb{C}^2$, so R is diagonalizable over the complex numbers even though it is not diagonalizable over the reals. This is why complex numbers will appear when we study real oscillations.

Finally, an eigenvector can be a function. Differentiation on the space of smooth real-valued functions on $\mathbb{R}$ sends $e^{\lambda t}$ to $\lambda e^{\lambda t}$ for every real λ . The familiar exponential

is therefore an eigenvector of differentiation. This does not assert that these functions form a basis of that infinite-dimensional space; the eigenvector equation alone is enough to explain why exponentials are natural functions to try in linear differential equations.

Try this 13. Find the eigenvalues and eigenspaces of $A = \begin{pmatrix} 4 & 1 \\ 0 & 2 \end{pmatrix}$. Write $A = P\Lambda P^{-1}$ using an eigenbasis, and use the basis to compute $A^k(3, 2)$ for nonnegative integers k .

Try this 14. Regard differentiation as a map $D : P_2 \rightarrow P_2$. Show that its only eigenvalue is zero and find the eigenspace. Is this map diagonalizable? Explain why the exponential eigenfunctions from the preceding paragraph do not contradict your answer.

Answers to the Review Questions

1. $3f - 2g = 3 - 5t - 2t^2$, and $-f = -1 + t$. The function f is not zero: for example, $f(0) = 1$. Being the zero function requires vanishing at every input.

2. Addition and scalar multiplication stay among 2×3 matrices and satisfy the vector-space rules entry by entry. The zero matrix and entrywise negatives supply zero and additive inverses. Rectangular shape causes no difficulty; multiplication of two such matrices is not required.

3. Even functions contain the zero function, and $(af + bg)(-t) = af(t) + bg(t) = (af + bg)(t)$, so they form a subspace. The condition $f(0) \geq 0$ fails closure under negative scaling: the constant function 1 belongs, while its negative does not.

4. The circle does not contain zero. Its points $(1, 0)$ and $(-1, 0)$ have midpoint $(0, 0)$ outside the circle, so it is not affine either.

5. In the identity $a(1+t) + b(1-t) + ct^2 = 0$, coefficient matching gives $a+b=0$, $a-b=0$, and $c=0$, hence all coefficients vanish. Three independent vectors in the three-dimensional space P_2 form a basis. For the requested polynomial, $a+b=2$, $a-b=3$, and $c=1$, so the coordinates are $(5/2, -1/2, 1)$.

6. Write $(x, y, z) = (x, y, -x - y) = x(1, 0, -1) + y(0, 1, -1)$. The two displayed vectors span U and are independent by looking at their first two entries. Thus $\dim U = 2$. Three entries record the vector, but the constraint leaves only two independent choices.

7. Kernel vectors satisfy $x = -y$ and $z = -y$, so $\ker T = \text{span}\{(-1, 1, -1)\}$. Every (r, s) is the image of $(r, 0, s)$, hence the image is $\mathbb{R}^2$. The dimension count is $3 = 1 + 2$. The map is surjective and not injective.

8. $J(a + bt) = at + (b/2)t^2$. Its kernel is $\{0\}$ and its image is $\text{span}\{t, t^2\}$. It is injective and maps onto this subspace, but cannot produce a polynomial with nonzero constant term.

9. The matrix of J is

$$K = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

With M the matrix of D from section B.5,

$$MK = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad KM = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus $D \circ J$ is the identity on P_1 , while $(J \circ D)p = p - p(0)$ removes the constant coefficient.

10. $P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. Solving $a + b = 4$, $a - b = 2$ gives coordinates $(3, 1)$. Indeed, $3(1, 1) + (1, -1) = (4, 2)$.

11. The full set is

$$(0, 0, 1) + a(1, 0, -1) + b(0, 1, -1), \quad a, b \in \mathbb{R}.$$

It is affine of dimension two and does not contain zero. Adding two solutions gives coordinates summing to 2, so the sum fails the equation. Their midpoint has coordinates summing to 1 and succeeds.

12. The derivative condition gives $p = c + t + t^2$, an affine line. The value at zero forces $c = 3$, leaving the singleton $\{3 + t + t^2\}$. This is a zero-dimensional affine space, a translate of the zero subspace. It is not a vector subspace because it does not contain the zero polynomial.

13. The characteristic equation is $(4 - \lambda)(2 - \lambda) = 0$. The eigenspaces for 4 and 2 are respectively $\text{span}\{(1, 0)\}$ and $\text{span}\{(1, -2)\}$. Thus we can take

$$P = \begin{pmatrix} 1 & 1 \\ 0 & -2 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}.$$

Since $(3, 2) = 4(1, 0) - (1, -2)$, the requested vector is

$$A^k(3, 2) = 4 \cdot 4^k(1, 0) - 2^k(1, -2) = (4 \cdot 4^k - 2^k, 2 \cdot 2^k).$$

14. If $p' = \lambda p$ with $\lambda \neq 0$ and p a nonzero polynomial, the right side has the same degree as p , whereas the derivative has lower degree (or is zero for a constant). This is impossible. For $\lambda = 0$ the eigenspace consists of constants and has dimension one, so it cannot supply a basis of P_2 . The map is not diagonalizable. For $\lambda \neq 0$, $e^{\lambda t}$ does not belong to P_2 ; changing the domain changes the available eigenvectors.

Further Reading

Dan Margalit and Joseph Rabinoff's *Interactive Linear Algebra* is a free online textbook with interactive figures for exploring the geometry behind the algebra.

Jim Hefferon's *Linear Algebra* is a freely available first-course text with a separate answer book. Its treatments of systems, vector spaces, and maps provide additional worked practice for this review.

Sheldon Axler's *Linear Algebra Done Right*, fourth edition, is freely available from the author's site. Chapters 1–3 develop vector spaces, bases, and linear maps in greater depth; Chapter 5 treats eigenvalues and eigenvectors.