Real World Math

Figures current as of October 2026

Three courses in one volume: the calculus, linear algebra and optimization, and probability and statistics behind finance, energy, and artificial intelligence (AI), rebuilt for adults who want to use them. Every idea is attached to something real, from a battery's best duration to a portfolio's efficient frontier to the loss a language model minimizes, and every module ends with practice problems solved in full.

Part I — Calculus for the Real World. The algebra underneath, functions and graphs, exponentials and logarithms, rates of change, optimization, accumulation, change over time, approximation, and several variables, from a battery's best duration and a bond's duration to the option Greeks and the chain rule inside a neural network.

Part II — Linear Algebra and Optimization. Vectors, matrices, systems of equations, least squares, eigenvalues, portfolio risk, constrained optimization, linear programming, gradient methods, and the matrix arithmetic inside AI models.

Part III — Probability and Statistics. Probability, Bayes' rule, expected value, distributions and fat tails, the central limit theorem, estimation, hypothesis tests, regression, and how models are fit to data and judged.

How to use this volume. Each part is a complete course with everything it had on its own: a twenty-problem diagnostic, ten modules with Memorize This boxes, practice problems, and quizzes, and, at the end of the part, its Memorize This boxes gathered in one place, worked solutions to every problem, and an answer sheet. Module, section, figure, and problem numbers restart in each part, so a reference such as Module 4 or problem 3.8 points to the part it appears in. The parts build on one another in the order printed, but each also stands alone, and the diagnostic at the start of each part says where to begin. Page numbers run through the whole volume.

Contents

Part I — Calculus for the Real World **3**

Where to Start 5
Module 1 — The Algebra Underneath 7
Module 2 — Functions and Graphs 14
Module 3 — Exponentials and Logarithms 21
Module 4 — Rates of Change 28
Module 5 — The Rules of Differentiation 35
Module 6 — Optimization 43
Module 7 — Integrals as Accumulation 50
Module 8 — Differential Equations 57
Module 9 — Approximation 64
Module 10 — Several Variables 70
Everything to Memorize 77
Worked Solutions 84
Answer Sheet 92

Part II — Linear Algebra and Optimization **93**

Where to Start 95
Module 1 — Vectors 97
Module 2 — Matrices 105
Module 3 — Systems of Linear Equations 112
Module 4 — Least Squares 120
Module 5 — Eigenvalues and Eigenvectors 127
Module 6 — Covariance and Portfolio Risk 134
Module 7 — Optimization with Constraints 143
Module 8 — Linear Programming 152
Module 9 — Convexity and Gradient Methods 161
Module 10 — Linear Algebra Inside AI 169
Everything to Memorize 177
Worked Solutions 184
Answer Sheet 194

Part III — Probability and Statistics **195**

Where to Start 197
Module 1 — Probability and Its Rules 199
Module 2 — Conditional Probability and Bayes' Rule 205
Module 3 — Random Variables and Expected Value 212
Module 4 — Counting Events: Binomial and Poisson 219
Module 5 — Continuous Distributions and Fat Tails 225
Module 6 — Sums, Averages, and the Central Limit Theorem 232
Module 7 — Estimation and Confidence Intervals 238
Module 8 — Hypothesis Tests 246
Module 9 — Regression and Its Uncertainty 253
Module 10 — Fitting and Testing Models 261
Everything to Memorize 268
Worked Solutions 274
Answer Sheet 283

Calculus for the Real World

Figures current as of October 2026

The calculus behind finance, energy, and AI, rebuilt for adults who learned it once: the algebra underneath, rates of change, optimization, accumulation, change over time, approximation, and several variables. Every idea is attached to something real, from a battery's duration to a bond's price to a neural network's weights, and every module ends with practice problems solved in full at the back.

MODULE 1 — The Algebra Underneath

Goal: Rebuild the algebra every later module runs on.

Exponent rules, including negative and fractional powers; rearranging and solving equations, from break-even volumes to the quadratic formula; fractions and the difference quotient; units, orders of magnitude, and percentages — a battery's round trip, worked.

MODULE 2 — Functions and Graphs

Goal: Read and build the functions calculus works on.

Lines and slope as a rate; power functions, and why wind power rises with the cube of wind speed; composition and inverses; piecewise functions such as tiered tariffs; reading a chart honestly — truncated axes, log scales, and per-person figures.

MODULE 3 — Exponentials and Logarithms

Goal: Master the mathematics of growth.

Compounding and where e comes from; logarithms as the way to solve for time; doubling times and why the rule of 72 works; half-lives; log scales for learning curves and AI compute.

MODULE 4 — Rates of Change

Goal: Understand the derivative as a rate.

From average to instantaneous rate; limits only as far as needed; notation and units; marginal cost, the locational marginal price as the cost of the next megawatt-hour, and power as the derivative of energy.

MODULE 5 — The Rules of Differentiation

Goal: Differentiate fluently.

The power, product, quotient, exponential, and logarithm rules, then the chain rule at length: discount factors, the sigmoid, and backpropagation as the chain rule repeated. The heaviest drill in the course.

MODULE 6 — Optimization

Goal: Find the best value of anything written as a function.

Critical points, the second-derivative test, and endpoints; the profit-maximizing price; how long a battery should be; the square-root inventory formula; fitting a model by minimizing squared error.

MODULE 7 — Integrals as Accumulation

Goal: Add up a rate over time.

Riemann sums and the fundamental theorem; antiderivatives and substitution; energy and revenue from hourly data; the present value of a stream of cash flows and of perpetuities; consumer surplus.

MODULE 8 — Differential Equations

Goal: Model quantities whose change depends on their own size.

Growth and decay, cooling, a battery's state of charge with losses, and the S-curve of adoption; Euler's method, and gradient descent as Euler's method applied to a loss.

MODULE 9 — Approximation

Goal: Replace a hard function with an easy one, and know the error.

Tangent lines and Taylor polynomials; Newton's method for solving a yield or an internal rate of return; a bond's duration and convexity derived as the first and second derivatives of its price.

MODULE 10 — Several Variables

Goal: Measure how an output responds to each of its inputs.

Contour maps and partial derivatives as sensitivities; the option Greeks from the Black–Scholes price; a project model's sensitivities and the tornado chart; the gradient.

How to use this course. Start with the diagnostic on the next page; it says which modules to work and which to skim. Each module teaches through worked examples, then gives a Memorize This box of the formulas, definitions, and ideas to commit to memory, practice problems in three tiers, and a five-question quiz. Every symbol is defined in words where it first appears in a module, and again in its Memorize This box. Worked solutions to every practice problem are at the back, after all the Memorize This boxes gathered in one place; quiz answers are on the answer sheet. Work with pencil and paper, and a calculator with ln and e^x keys or a spreadsheet. The examples use illustrative numbers unless a source and date are given.

Where to Start

A twenty-problem diagnostic

Work these on paper before you start. A calculator is fine for problem 8; the rest need none. The four bands follow the order of the course. The answers are on the next page, with a guide to where to begin.

Band A — The algebra underneath (Module 1)

1. Simplify $x^3 \cdot x^5 / x^2$.

2. Evaluate $8^{2/3}$.

3. Solve $3(x - 4) = 2x + 5$.

4. Solve $x^2 - x - 12 = 0$.

5. The compound-growth formula $A = P(1 + r)^n$ says an amount P grows to A in n years at a yearly rate r. Solve it for r.

Band B — Functions, exponentials, and logarithms (Modules 2 and 3)

6. If $f(x) = 2x + 1$ and $g(x) = x^2$, find $f(g(3))$.

7. Find the slope of the line through the points (2, 5) and (6, 13).

8. Solve $e^{2x} = 7$.

9. Evaluate $\ln(e^5) + \log_{10} 1000$.

10. Money grows 6% a year, compounded yearly. About how many years does it take to double?

Band C — Derivatives (Modules 4 to 6)

11. Differentiate $f(x) = 5x^4 - 3x + 2$.

12. Differentiate $g(t) = e^{-0.5t}$.

13. Differentiate $h(x) = (3x + 1)^5$.

14. Differentiate $y = x \ln x$.

15. Find the value of x that minimizes $C(x) = x^2 - 8x + 20$.

Band D — Integrals and beyond (Modules 7 to 10)

16. Evaluate $\int_0^2 3x^2 \, dx$.

17. Find an antiderivative of e^{2x}.

18. A quantity y changes over time t, in years, at the rate $dy/dt = 0.03y$, starting from $y(0) = 100$. Write $y(t)$.

19. Use the tangent line to $\sqrt{x}$ at $x = 100$ to estimate $\sqrt{101}$.

20. If $f(x, y) = x^2 y + 3y$, find $\partial f / \partial x$.

Answers, and Where to Begin

#	Answer	Taught in
1	x^6	Module 1
2	4	Module 1
3	$x = 17$	Module 1
4	$x = 4$ or $x = -3$	Module 1
5	$r = (A/P)^{1/n} - 1$	Module 1
6	19	Module 2
7	2	Module 2
8	$x = \ln 7/2 \approx 0.973$	Module 3
9	8	Module 3
10	About 12 years by the rule of 72 (exactly 11.9)	Module 3
11	$f'(x) = 20x^3 - 3$	Module 5
12	$g'(t) = -0.5e^{-0.5t}$	Module 5
13	$h'(x) = 15(3x + 1)^4$	Module 5
14	$y' = \ln x + 1$	Module 5
15	$x = 4$	Module 6
16	8	Module 7
17	$e^{2x}/2$, plus any constant	Module 7
18	$y(t) = 100e^{0.03t}$	Module 8
19	10.05	Module 9
20	$2xy$	Module 10

Score each band out of five, then use the table. A low score on a band means working every module the band covers.

Band score	What to do
5 of 5	Skim the modules: read each Memorize This box and try the challenge problems.
3 or 4	Read the modules quickly and work the applied problems.
0 to 2	Work the modules through, worked examples and all.

If the algebra band goes badly, start at Module 1 whatever the other bands show: every later module leans on it. Most readers who last did calculus years ago do best starting at Module 1 and moving fast where it feels familiar.

Module 1 — The Algebra Underneath

Exponents, rearranging, solving, fractions, and units

Recommended Reading / Reference

Book: I. M. Gelfand and Alexander Shen, *Algebra* (1993) — a short, clear rebuild of school algebra by a great mathematician, written for students working on their own

Reference: Paul Dawkins, *Paul's Online Math Notes* (Lamar University) — free algebra review with worked examples and practice problems

1a. Exponents and Roots

Calculus is mostly algebra with one new idea added. When a calculus answer comes out wrong, the error is usually in the algebra: an exponent mishandled, a term cancelled that should not have been, an equation rearranged in the wrong order. This module rebuilds those skills on the quantities the rest of the course uses.

An exponent counts repeated multiplication: $x^3 = x \cdot x \cdot x$. For any numbers a and b and powers m and n, four rules follow from that definition, and every manipulation of powers in this course uses one of them.

$$a^m\,a^n = a^{m+n} \qquad \frac{a^m}{a^n} = a^{m-n}$$
$$(a^m)^n = a^{mn} \qquad (ab)^n = a^n\,b^n$$

Zero, negative, and fractional powers are defined so that the same rules keep working. Since $a^3/a^3 = 1$ and the division rule gives $a^{3-3} = a^0$, the zero power must equal 1. Since $a^0/a^n = a^{-n}$, a negative power is a reciprocal. And since $(a^{1/2})^2 = a^1$, the power one-half is the square root; in general, $a^{1/n}$ is the nth root.

$$a^0 = 1 \qquad a^{-n} = \frac{1}{a^n} \qquad a^{1/n} = \sqrt[n]{a} \qquad a^{m/n} = \left(\sqrt[n]{a}\right)^m$$

These are not curiosities. Calculus states its rules for powers of x, and rewriting $1/x^2$ as x^{-2} and $\sqrt{x}$ as $x^{1/2}$ lets one rule handle all of them. Finance writes the same way: the value today of a dollar paid in t years, at an interest rate r, is $1/(1+r)^t$, which is easier to work with as $(1+r)^{-t}$.

Worked example — A payment five years away

What is $1,000 paid in five years worth today, if money earns 4% a year? Divide by the growth factor, or multiply by the negative power:

$$1000\,(1.04)^{-5} = \frac{1000}{1.04^5} = \frac{1000}{1.21665} \approx 821.93$$

The payment is worth about $821.93 today. The number $1.04^{-5} \approx 0.822$ is the five-year discount factor at 4%; Module 7 adds up whole streams of payments discounted this way.

Three mistakes account for most exponent errors. A power does not distribute over a sum: $(a+b)^2 = a^2 + 2ab + b^2$, not $a^2 + b^2$. A root does not split a sum: $\sqrt{(9+16)} = 5$, while $\sqrt{9} + \sqrt{16} = 7$. And a negative exponent makes a reciprocal, not a negative number: $2^{-3} = 1/8$.

1b. Rearranging Equations

An equation is a balance: whatever is done to one side must be done to the other. To isolate a variable, undo the operations around it in the reverse of the order they were applied, the way a parcel is unwrapped from the outside in.

Energy E is power P multiplied by time t: $E = P \cdot t$. To find how long a battery runs, divide both sides by P to get $t = E/P$: a 100 MW battery holding 400 MWh runs for four hours at full power.

The compound-growth formula takes more unwrapping. A starting amount P, growing at rate r for n years, becomes

$$A = P\,(1 + r)^{n}$$

To find the growth rate, undo the operations in reverse: divide by P, take the nth root by raising both sides to the power $1/n$, and subtract 1.

$$\begin{aligned} \frac{A}{P} &= (1 + r)^{n} \\ \left(\frac{A}{P}\right)^{1/n} &= 1 + r \\ r &= \left(\frac{A}{P}\right)^{1/n} - 1 \end{aligned}$$

Worked example — A growth rate from two numbers

An investment grew from \$10,000 to \$16,000 in seven years. Its compound annual growth rate is $r = (16{,}000/10{,}000)^{1/7} - 1 = 1.6^{1/7} - 1$. A calculator gives $1.6^{1/7} \approx 1.0694$, so the rate is about 6.94% a year.

This is the calculation behind every compound annual growth rate quoted in the financial press. Dividing the 60% total gain by seven years gives about 8.6%, which overstates the rate, because it ignores that each year's growth builds on the growth before it.

1c. Solving Equations

A linear equation has its unknown to the first power only, and rearranging solves it. The commonest business example is break-even. A plant with fixed costs F sells each unit at price p and spends c to make it. Each unit sold contributes $p - c$ toward the fixed costs, so selling q units earns a profit of $(p - c)q - F$, which is zero when

$$(p - c)\,q - F = 0 \quad \Longrightarrow \quad q = \frac{F}{p - c}$$

Worked example — Break-even volume

A plant has fixed costs of \$2.4 million a year, sells its product at \$150, and spends \$90 on materials and labor for each unit. Each unit contributes \$60, so break-even is $2{,}400{,}000/60 = 40{,}000$ units a year. Every unit beyond that adds \$60 of profit.

Two linear equations in two unknowns are solved by substitution: solve one equation for one unknown and put the result into the other. A market settles where the quantity buyers want equals the quantity sellers offer.

Worked example — Where supply meets demand

Buyers want $q = 100 - 2p$ units at price p, and sellers offer $q = 3p - 20$. Setting the two equal, $100 - 2p = 3p - 20$, so $120 = 5p$ and $p = 24$. Either equation then gives $q = 52$.

A quadratic equation has a squared term: $ax^2 + bx + c = 0$, where the coefficients a, b, and c are numbers and a is not zero. Some factor at sight: $x^2 - 5x + 6 = (x - 2)(x - 3)$, so the solutions are 2 and 3. The quadratic formula solves any of them:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The expression under the root, $b^2 - 4ac$, decides how many solutions there are: two if it is positive, one if it is zero, and none among the real numbers if it is negative.

Worked example — Two prices, the same revenue

A store sells $q = 400 - 8p$ units a week at price p, so its weekly revenue is $R = p(400 - 8p)$. Which prices bring in \$3,000? Set $400p - 8p^2 = 3000$, divide every term by 8, and rearrange to $p^2 - 50p + 375 = 0$. The formula gives:

$$p = \frac{50 \pm \sqrt{2500 - 1500}}{2} = \frac{50 \pm 31.62}{2} \approx 9.19 \text{ or } 40.81$$

Both prices earn \$3,000 a week: the low one sells about 326 units, the high one about 74. Two solutions are usually two real choices. Module 6 finds the price between them that earns the most.

Most equations met in practice have no formula at all. The interest rate that makes a project's cash flows worth exactly its cost cannot be isolated by algebra. It is found by trying values and closing in, which Module 9 turns into a fast and reliable method.

1d. Fractions and the Difference Quotient

Fractions add over a common denominator: $a/b + c/d = (ad + bc)/(bd)$. They simplify by cancelling a factor that multiplies the whole top and the whole bottom, never a term joined by plus or minus. In $(x^2 - 9)/(x - 3)$, the top factors as $(x - 3)(x + 3)$, and cancelling the common factor leaves $x + 3$, true for every x except 3, where the original fraction divides by zero. In $(x + 3)/x$, nothing cancels: the x on top is a term, not a factor.

One fraction matters more than any other in this course. For a function f, the difference quotient measures how fast f changes between an input x and a nearby input $x + h$, where h is the size of the step:

$$\frac{f(x + h) - f(x)}{h}$$

Module 4 builds the derivative from it, and the algebra is always the same: expand, cancel what cancels, and divide by h. For $f(x) = x^2$:

$$\frac{(x + h)^2 - x^2}{h} = \frac{x^2 + 2xh + h^2 - x^2}{h} = \frac{2xh + h^2}{h} = 2x + h$$

As h shrinks toward zero, the quotient approaches $2x$, the slope of the curve $y = x^2$ at the point x. That is the derivative, reached with algebra alone.

Worked example — A difference quotient with fractions inside

For $f(x) = 1/x$, put the two fractions over a common denominator before dividing by h:

$$\frac{1}{h}\left(\frac{1}{x+h} - \frac{1}{x}\right) = \frac{x - (x+h)}{x(x+h)\,h} = -\frac{1}{x(x+h)}$$

As h shrinks, this approaches $-1/x^2$. The slope of $1/x$ is negative wherever it is defined: the curve always falls.

1e. Units, Scales, and Percentages

Units obey the same algebra as numbers. Power in megawatts times hours gives energy in megawatt-hours; a price in dollars per megawatt-hour times megawatt-hours gives dollars. Writing the units into a calculation and cancelling them like factors catches most errors before they happen. To convert, multiply by a fraction equal to one: 2.5 MWh × (1,000 kWh per MWh) = 2,500 kWh.

Worked example — A battery's round trip

A battery returns 85% of the energy put into it; the rest is lost as heat. To deliver 100 MWh it must take in $100/0.85 \approx 117.6$ MWh. If it charges overnight at $30/MWh and sells in the evening peak at $90/MWh, the round trip earns $100 \times 90 - 117.6 \times 30 \approx 5{,}471$ dollars.

The losses cost $529: the 17.6 MWh bought and never sold. Their cost rises with the price of charging power, so efficiency matters most when charging is expensive.

Very large and very small quantities are written in powers of ten, and multiplying or dividing them is exponent arithmetic: add the exponents to multiply, subtract them to divide. A gigawatt is 10^9 watts and a megawatt 10^6, so a gigawatt is a thousand megawatts. Suppose an AI training run needs 10^{25} arithmetic operations, and a cluster of chips performs 10^{18} operations a second in total. The run takes $10^{25}/10^{18} = 10^7$ seconds, about 116 days.

Percentages cause more confusion than any other arithmetic in finance. A percent change is measured against the starting value: (new − old)/old. A change between two percentages is measured in percentage points, and in finance in basis points, hundredths of a point. A bond yield that rises from 4.00% to 4.50% has risen 0.5 percentage points, or 50 basis points, a 12.5% increase in the yield itself.

Percent changes compound; they do not add. A price that falls 20% and then rises 20% ends at $0.8 \times 1.2 = 0.96$ of where it began, 4% down. Recovering a 20% fall takes a 25% rise, because $1/0.8 = 1.25$, and recovering a 50% fall takes a 100% rise.

1f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Multiplying and dividing powers	$a^m a^n = a^{m+n}$ $\quad \frac{a^m}{a^n} = a^{m-n}$ $\quad (a^m)^n = a^{mn}$ a is any base; m and n are any powers
Zero, negative, and fractional powers	$a^0 = 1$ $\quad a^{-n} = \frac{1}{a^n}$ $\quad a^{1/n} = \sqrt[n]{a}$ a is any nonzero base; $a^{1/n}$ is the nth root of a
Sums do not split	$(a+b)^2 = a^2 + 2ab + b^2$, and $\sqrt{(a+b)}$ is not $\sqrt{a} + \sqrt{b}$ a and b are any numbers
Quadratic formula	$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ Solves $ax^2 + bx + c = 0$, where a, b, and c are numbers and a is not zero
Growth rate from start to end	$r = \left(\frac{A}{P}\right)^{1/n} - 1$ P grows to A in n years; r is the compound annual growth rate
Break-even volume	$q = \frac{F}{p - c}$ F is the fixed cost, p the price, c the variable cost per unit, and q the units sold
Adding fractions	$a/b + c/d = (ad + bc)/(bd)$; cancel factors, never terms
Difference quotient	$\frac{f(x+h) - f(x)}{h}$ How fast the function f changes between the inputs x and $x + h$; h is the size of the step
Energy and power	Energy = power × time: 1 MW for 1 hour is 1 MWh, which is 1,000 kWh
Percentages	Percent change = (new − old)/old; 1 percentage point = 100 basis points; percent changes compound, so they do not add
The idea	Most calculus mistakes are algebra mistakes. Two answers to a quadratic are usually two real choices, like a low price and a high one that earn the same revenue.

1g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

1.1 Simplify $x^4 \cdot x^{-1} / x^2$.

1.2 Write each as a single power of x: $1/\sqrt{x}$, $x\sqrt{x}$, and $1/x^3$.

1.3 Evaluate without a calculator: $27^{2/3}$, $16^{-1/2}$, and $(1/8)^{-1/3}$.

1.4 Solve $5x - 7 = 2(x + 4)$.

1.5 Solve $x^2 + 2x - 15 = 0$ by factoring.

1.6 Simplify $(x^2 - 4)/(x^2 + 4x + 4)$.

1.7 Write $1/x - 1/(x + 1)$ as a single fraction.

Applied

1.8 A 50 MW battery holds 200 MWh. How long can it discharge at full power, and how long at 30 MW?

1.9 An investment grew from $25,000 to $40,000 in six years. What was its compound annual growth rate?

1.10 A power plant has fixed costs of $1.2 million a year. It sells power at $65/MWh, and its fuel and other variable costs are $25/MWh. How many megawatt-hours must it sell each year to break even?

1.11 Buyers want $q = 120 - 3p$ units at price p, and sellers offer $q = 2p - 10$. Find the market price and quantity.

1.12 A ten-year bond yield rises from 4.25% to 4.60%. How many basis points did it rise, and by what percent did the yield increase?

1.13 A battery with 88% round-trip efficiency must deliver 40 MWh. How much energy must it take in, and what does that energy cost at $28/MWh?

Challenge

1.14 Show that the difference quotient of $f(x) = x^3$ simplifies to $3x^2 + 3xh + h^2$. What does it approach as h shrinks toward zero?

1.15 A stock falls 30% and then rises 30%. What is its net change, and what rise would have recovered the fall exactly?

1.16 Solve $5{,}000 = 2{,}000(1.05)^t$ for t by trying values. (Module 3 shows how logarithms solve it in one step.)

Module 1 Quiz

Check Your Understanding

1. What is $8^{-2/3}$?

A) 4, the cube root of 8 squared
B) -4, since the power is negative
C) $-16/3$
D) $1/4$

2. Why does any nonzero number raised to the power zero equal 1?

A) It is an arbitrary convention, adopted so that tables of powers look tidy.
B) Because $a^n/a^n = 1$, and the rule for dividing powers says the same quotient equals $a^{n-n} = a^0$.
C) Because a negative power is a reciprocal, and 1 is its own reciprocal.
D) Because any number divided by zero is defined to be 1.

3. An investment grows from \$20,000 to \$30,000 in five years. Which expression gives its compound annual growth rate?

A) $0.5/5$
B) $1.5^5 - 1$
C) $1.5^{1/5} - 1$
D) $(30{,}000 - 20{,}000)/(5 \times 20{,}000)$

4. Which of these is true for every nonzero x?

A) $x^{-2} = 1/x^2$
B) $(x + 3)/x = 3$
C) $\sqrt{(x^2 + 9)} = x + 3$
D) $(x + 3)^2 = x^2 + 9$

5. A price falls 20% and then rises 20%. Where does it end?

A) Exactly where it started
B) 4% above where it started
C) 4% below where it started, because the rise applies to a smaller base: $0.8 \times 1.2 = 0.96$
D) About 2% below where it started, since the two moves nearly cancel

Module 2 — Functions and Graphs

Lines, powers, combinations, kinks, and honest charts

Recommended Reading / Reference

Book: Edward Tufte, *The Visual Display of Quantitative Information* (2nd ed., 2001) — the classic account of how charts inform and how they mislead

Reference: FRED, Federal Reserve Bank of St. Louis (fred.stlouisfed.org) — hundreds of thousands of economic data series, each chartable on a linear or a log axis

2a. What a Function Is

A function is a rule that turns each input into exactly one output. The notation $f(x)$ names the rule f and the input x. A battery's state of charge is a function of time, a store's revenue is a function of its price, and a bond's price is a function of its yield. Calculus studies how outputs change when inputs change, so every module from here on starts by writing the quantity of interest as a function.

The inputs a function accepts form its domain, and the outputs it can produce form its range. Real quantities limit both: a price cannot be negative, and a battery's charge runs from empty to full. Writing the domain down prevents a common error, which is accepting an answer the formula allows and the world does not, such as a negative price or a battery 110% full.

A graph draws a function, matching each input on the horizontal axis with its output on the vertical axis. Because each input has one output, no vertical line crosses the graph of a function more than once. Reading a graph means reading three things: where it is high or low, where it rises or falls, and how steeply.

2b. Lines and Slope

The simplest function is a line, $y = mx + b$. The intercept b is the output when the input is zero, and the slope m is how much the output changes for each one-unit change in the input. Between any two points on the line, (x_1, y_1) and (x_2, y_2), the slope is the change in output divided by the change in input. The Greek capital delta, Δ, is read “change in”:

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

Slope is a rate, and it carries units: dollars per unit, megawatt-hours per hour, degrees per kilometer. A line is the one function whose rate never changes. The derivative of Module 4 finds the slope of curves whose rate does.

Worked example — A cost line

A plant's cost of making q units a year is $C(q) = 2{,}400{,}000 + 90q$ dollars. The intercept, \$2.4 million, is the fixed cost paid even at zero output, and the slope, \$90 per unit, is the variable cost. Making 1,000 more units adds $90 \times 1{,}000 = 90{,}000$ dollars. Economists call the slope the marginal cost, and for a line it is the same at every level of output.

Given one point (x_1, y_1) and a slope m, the line through the point is $y - y_1 = m(x - x_1)$. This point-slope form is how the course writes tangent lines, which Module 9 uses to approximate curves.

Worked example — A household electricity bill

A utility charges a fixed $12 a month plus 16 cents a kilowatt-hour, so the bill is $B(u) = 12 + 0.16u$ for u kilowatt-hours. A household using 900 kWh pays $12 + 144 = 156$ dollars. Its average price is $156/900 \approx 17.3$ cents a kilowatt-hour, above the 16-cent slope, because the fixed charge is spread across the usage.

2c. Power Functions and Scaling

A power function multiplies a constant k by the input raised to a fixed power n: $y = kx^n$. The exponent sets the shape. With n above 1 the curve bends upward ever more steeply; with n between 0 and 1 it rises but flattens; with n negative it falls toward zero.

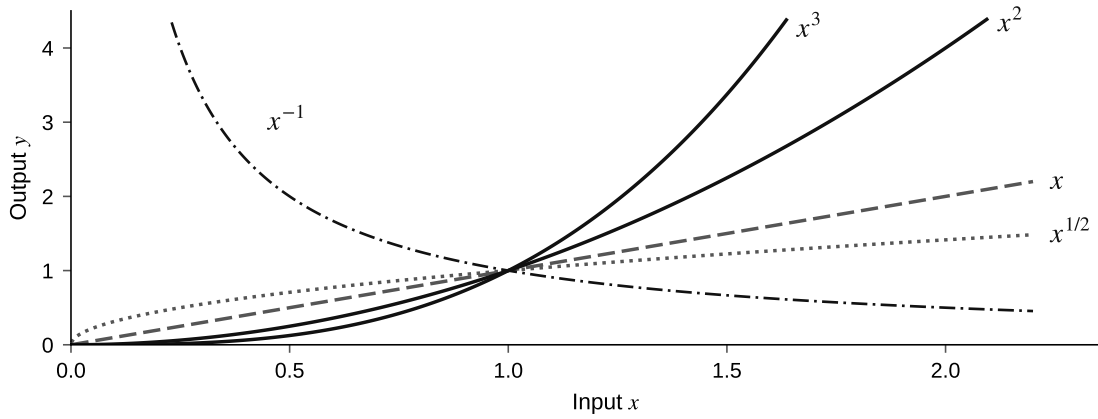


Figure 2.1. Power functions. Exponents above 1 bend upward, exponents between 0 and 1 flatten, and a negative exponent falls toward zero. All the positive powers pass through (1, 1).

The exponent also sets the scaling law, which is the practical point. Multiplying the input by a factor c multiplies the output by c^n, wherever the input starts. Double the side of a square and its area grows four times; double the side of a cube and its volume grows eight times.

Wind power is the most consequential power law in energy. The power carried by wind through a turbine's blades is

$$P = \tfrac{1}{2}\rho A v^3$$

where ρ is the density of air, about 1.225 kilograms per cubic meter at sea level, A is the area the blades sweep, and v is the wind speed. Power rises with the cube of the speed: wind 10% faster carries $1.1^3 \approx 1.33$ times the power, and wind twice as fast carries eight times as much. That is why developers pay for tall towers and windy ridges, and why a site's average wind speed matters more than almost any other number in a wind farm's finances.

Worked example — The power through a large rotor

A turbine's blades are 75 meters long, so they sweep $A = \pi \times 75^2 \approx 17{,}700$ square meters. At 10 meters a second, the wind's power through that circle is $0.5 \times 1.225 \times 17{,}700 \times 10^3 \approx 10.8$ million watts, or 10.8 MW. No turbine captures all of it. The physical ceiling, the Betz limit, is about 59%, and a good modern turbine might convert around 45% at this speed, or roughly 4.9 MW.

2d. Composition and Inverses

Functions chain. When one function's output becomes another's input, the result is a composition, written $f(g(x))$ and read from the inside out: apply g first, then f. Electricity demand on a summer afternoon depends on the temperature, $L = f(T)$, and the temperature depends on the hour, $T = g(h)$. Demand by the hour is the composition $f(g(h))$.

Compositions matter because the chain rule of Module 5 differentiates them, and most functions in finance and AI are compositions: a discount factor applied to a growth rate, an activation applied to a weighted sum. The chain rule's answer is the obvious one: how fast demand rises each hour is how fast it rises per degree, times how many degrees the temperature rises each hour.

An inverse function runs a function backward, taking the output and returning the input. If $f(c) = 1.8c + 32$ converts a Celsius temperature c to Fahrenheit, its inverse converts a Fahrenheit temperature F back: $f^{-1}(F) = (F - 32)/1.8$. The −1 marks an inverse, not a reciprocal. To find an inverse, solve $y = f(x)$ for x. Applying a function and then its inverse returns the starting value: $f^{-1}(f(x)) = x$.

Worked example — The price that hits a sales target

A store sells $q = 400 - 8p$ units a week at price p. The inverse gives the price that sells a target quantity: $p = (400 - q)/8$. To sell 240 units a week, the store charges $(400 - 240)/8 = 20$ dollars.

Only a function that never repeats an output has an inverse over its whole domain. The square x^2 gives 9 for both 3 and −3, so its inverse, the square root, is defined by choosing the positive input. Module 3 builds the most useful inverse in the course: the logarithm, which undoes the exponential and solves for time.

2e. Piecewise Functions

Many real rules change at a threshold. A piecewise function uses one formula on one stretch of inputs and another beyond it. Tiered electricity tariffs, income tax brackets, and loans with rate caps all work this way.

Worked example — A tiered tariff

A utility charges 12 cents a kilowatt-hour for the first 500 each month and 18 cents for every kilowatt-hour above that. Its monthly bill B, in dollars, for u kilowatt-hours is:

$$B(u) = \begin{cases} 0.12\,u, & u \le 500 \\ 60 + 0.18\,(u - 500), & u > 500 \end{cases}$$

A household using 800 kWh pays $60 + 0.18 \times 300 = 114$ dollars. The 60 is the cost of the whole first tier, 0.12×500, which makes the two pieces meet at 500: the bill has a kink there, not a jump. What jumps is the slope, the price of the next kilowatt-hour, from 12 cents to 18. Tax brackets work the same way, which is why a raise that crosses into a higher bracket never lowers take-home pay: only the dollars above the threshold pay the higher rate.

Two of the most important piecewise functions in finance and AI share a shape. A call option, the right to buy a stock at a strike price K, pays nothing at expiry if the stock ends below K and $S - K$ if it ends at S above it, so its payoff is $\max(S - K, 0)$. The rectified linear unit, or ReLU, the commonest activation in neural networks, is $\max(x, 0)$: zero for a negative input, the input itself for a positive one.

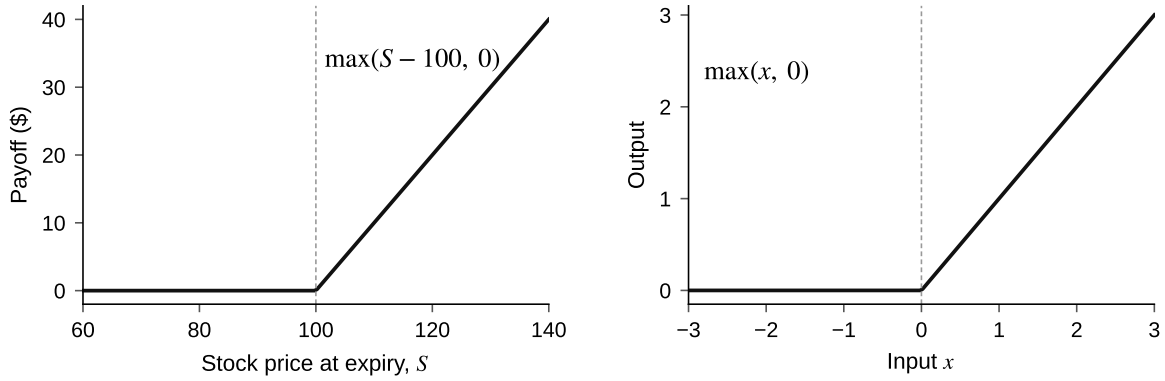


Figure 2.2. The same hockey stick twice: a call option's payoff at a strike of $100 (left) and the ReLU activation used in neural networks (right).

The kink is the point. A sum of many such hockey sticks, each bent at a different place, can approximate almost any curve. That is one reason networks built from ReLUs are so flexible, and why a portfolio of options can be shaped to almost any payoff.

2f. Reading a Chart Honestly

A chart is a function drawn for the eye, and its choices of axis can change what the eye concludes. Five checks catch most misleading charts.

Check	What it catches	What to do
Where the vertical axis starts	A truncated axis turns a 3% change into a towering one	Read the numbers, not the heights; bar charts should start at zero
Linear or log scale	On a linear axis steady percentage growth looks like a sudden explosion; on a log axis it is a straight line	Use a log axis for anything that compounds (Module 3)
Totals or per person	A large country's total dwarfs a small one's even when each person's share is smaller	Divide by population, output, or households before comparing
Nominal or real	Prices and incomes rise with inflation, so long charts in current dollars exaggerate growth	Adjust by a price index before comparing across decades
One axis or two	Two series on two scaled axes can be made to look as closely linked as the designer likes	Compare growth rates, or index both series to 100

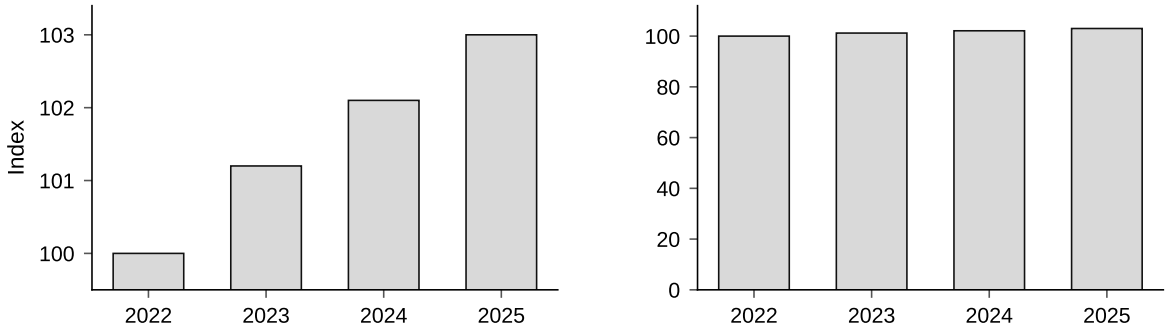


Figure 2.3. The same four years of an index rising 3% in total. Left, the axis starts at 99.5 and the rise looks like a surge; right, the axis starts at zero.

None of these choices is wrong in itself. A truncated axis shows small movements clearly, and a log axis is the right way to see a growth rate. The error is reading the picture without reading the axis.

2g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Slope	$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$ Δ means "change in"; (x_1, y_1) and (x_2, y_2) are two points on the line; slope has units of output per unit of input
Lines	$y = mx + b$; through the point (x_1, y_1) with slope m: $y - y_1 = m(x - x_1)$ m is the slope and b the intercept, the value of y when $x = 0$
Scaling a power function	Multiplying the input of $y = kx^n$ by a factor c multiplies the output by c^n k is a constant and n the power
Power in the wind	$P = \frac{1}{2}\rho A v^3$ P is power, ρ the density of air (about 1.225 kg per cubic meter), A the area the blades sweep, and v the wind speed; 10% more wind carries 33% more power
Composition	$f(g(x))$: apply g first, then f
Inverse	Solve $y = f(x)$ for x; $f^{-1}(f(x)) = x$; the −1 marks an inverse, not a reciprocal
Piecewise rules	A tier or bracket changes the slope at the threshold; only the units above it pay the new rate
Hockey sticks	Call option payoff $\max(S - K, 0)$; ReLU activation $\max(x, 0)$ S is the stock price at expiry and K the strike price
Chart checks	Axis start; linear or log; totals or per person; nominal or real; one axis or two
The idea	A function turns each input into exactly one output, and every calculus problem starts by writing its quantity as a function.

2h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

2.1 Find the equation of the line through (1, 7) and (4, 19).

2.2 If $f(x) = 3x^2 - 2$ and $g(x) = x + 1$, find $f(g(2))$ and $g(f(2))$.

2.3 Find the inverse of $f(x) = 5x - 4$.

2.4 For $y = 3x^4$, by what factor does y change when x doubles? When x triples?

2.5 A call option with a strike of 100 pays $\max(S - 100, 0)$, where S is the stock price at expiry. Evaluate the payoff at $S = 85$, 100, and 130.

2.6 What is the domain of $f(x) = 1/(x - 2)$? Of $g(x) = \sqrt{(x - 3)}$?

Applied

2.7 A utility charges $10 a month plus 14 cents a kilowatt-hour. Write the bill as a function of usage, and find the usage at which the bill is $136.

2.8 Under the tiered tariff of section 2e (12 cents for the first 500 kWh, 18 cents above), find the bill and the average price per kilowatt-hour for 350 kWh and for 1,200 kWh.

2.9 One site's wind averages 7 meters a second and another's 8. Using the cube law alone, how much more power does the windier site's wind carry?

2.10 A call option with a $50 strike cost $4. Write the buyer's profit at expiry as a function of the stock price S, and find the price at which the buyer breaks even.

2.11 Demand is $L = 2{,}000 + 60(T - 20)$ megawatts when the temperature is T degrees Celsius, for T above 20, and the temperature is $T = 20 + 1.5h$ at h hours after 9 a.m. Write demand as a function of h, and find it at 3 p.m.

2.12 An index rose from 412 to 420 over a year. A chart's vertical axis runs from 410 to 421. About what share of the chart's height does the rise fill, and what was the actual percent change?

Challenge

2.13 A battery cell makes heat in proportion to its volume and sheds it through its surface. For a cube-shaped cell of side s, write the ratio of surface area to volume as a function of s, and explain why larger cells are harder to cool.

2.14 Show that the inverse of $f(x) = (2x + 1)/(x - 3)$ is $f^{-1}(y) = (3y + 1)/(y - 2)$, and check it at $x = 4$.

2.15 A trader buys one call with a strike of 100 and sells two calls with a strike of 110. Write the payoff at expiry as a sum of hockey sticks, where S is the stock price at expiry, and evaluate it at $S = 95$, 105, 110, and 125.

2.16 Write $f(x) = \max(x, 0) - 2\max(x - 1, 0) + \max(x - 2, 0)$ piece by piece on $x < 0$, $0 \le x < 1$, $1 \le x < 2$, and $x \ge 2$, and describe its graph.

Module 2 Quiz

Check Your Understanding

1. Under a tariff with a fixed monthly charge plus a flat rate per kilowatt-hour, why is a household's average price per kilowatt-hour above the rate?

A) Because the rate rises with usage.
B) Because utilities add taxes to every bill that are charged as a percentage of the total amount due.
C) Because the average price includes the cost of the meter, which grows with the number of kilowatt-hours used.
D) Because the fixed charge is spread across every kilowatt-hour used.

2. Wind at one site is 20% faster than at another. Roughly how much more power does it carry?

A) 20%, since power is proportional to wind speed
B) About 73%
C) 44%, since power rises with the square of the speed
D) About 60%, because three times the 20% speed gain gives the power gain once the cube law is applied

3. In $f(g(x))$, which function is applied first?

A) g, the inner function
B) f
C) Both at once
D) Whichever is simpler to compute first

4. A tiered tariff charges 12 cents a kilowatt-hour up to 500 kWh and 18 cents above. What is the bill for 1,000 kWh?

A) $180, since the higher rate applies to all usage
B) $90
C) $120 at the lower rate
D) $150

5. Which chart choice can make steady percentage growth look like a sudden explosion?

A) A linear vertical axis on a series that compounds
B) A log scale
C) Per-person figures
D) A bar chart starting at zero

Module 3 — Exponentials and Logarithms

Compounding, e, logarithms, doubling and halving, and log scales

Recommended Reading / Reference

Book: Eli Maor, *e: The Story of a Number* (1994) — how compound interest, logarithms, and calculus led to the number e

Reference: Our World in Data (ourworldindata.org) — charts of technology costs, AI training compute, and other compounding series, most with a logarithmic axis

3a. Growth by a Constant Percentage

Linear growth adds the same amount each period; exponential growth multiplies by the same factor. An account earning 5% a year is multiplied by 1.05 each year. In general, a deposit P growing at a yearly rate r becomes, after t years, an amount

$$A = P(1 + r)^t$$

Early on the two kinds of growth look alike, but exponential growth pulls away, because each period's growth is earned on all the growth before it. At 5% a year, $1,000 earns $50 in its first year and about $126 in its twentieth. It grows to $1000 \times 1.05^{30} \approx 4{,}322$ dollars in thirty years, against $2,500 if it earned a flat $50 a year.

Interest is often credited more than once a year. At an annual rate r credited n times a year, each credit adds r/n, so after t years

$$A = P\left(1 + \frac{r}{n}\right)^{nt}$$

More frequent compounding earns slightly more, because interest starts earning interest sooner. Credited monthly, 6% a year becomes $(1 + 0.06/12)^{12} - 1 \approx 6.17\%$ over a full year, the effective annual rate a bank reports as the annual percentage yield.

Credit the interest ever more often, every day or every second, and the growth approaches a limit. At 100% interest credited n times a year, one year's growth factor is $(1 + 1/n)^n$: 2 for yearly credit, 2.5937 for ten credits, and 2.7169 for a thousand. The limit is the number e:

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = 2.71828\ldots$$

Growth compounded continuously at rate r multiplies a balance by e^{rt} after t years: $A = Pe^{rt}$. At 6%, continuous compounding gives $e^{0.06} - 1 \approx 6.18\%$ a year, barely more than monthly. The number e matters less for banking than for calculus: Module 5 shows that e^x is the one function that is its own rate of change, which makes it the natural language of growth and decay.

3b. Logarithms

A logarithm answers a question: to what power must a base be raised to give a number? Since $10^3 = 1{,}000$, the base-10 logarithm of 1,000 is 3, written $\log_{10} 1000 = 3$. The natural logarithm, ln, uses the base e. The logarithm is the inverse of the exponential, so each undoes the other:

$$e^{\ln x} = x \qquad \ln(e^x) = x \qquad \ln 1 = 0 \qquad \ln e = 1$$

For positive numbers a and b and any power k, three rules follow from the rules for exponents. They are why logarithms were invented: they turn multiplication into addition.

$$\ln(ab) = \ln a + \ln b \qquad \ln\frac{a}{b} = \ln a - \ln b \qquad \ln\left(a^k\right) = k \ln a$$

The third rule is the important one. It brings an unknown exponent down to where algebra can reach it, which is how to solve for time.

Worked example — How long to reach a target

At 7% a year, how long does $10,000 take to grow to $25,000? Write the equation, divide, take the natural log of both sides, and use the power rule:

$$\begin{aligned} 10{,}000\,(1.07)^t &= 25{,}000 \\ (1.07)^t &= 2.5 \\ t\,\ln 1.07 &= \ln 2.5 \\ t = \frac{\ln 2.5}{\ln 1.07} &= \frac{0.9163}{0.0677} \approx 13.5 \text{ years} \end{aligned}$$

The same steps solve any compound-growth problem for time: $t = \ln(A/P)/\ln(1 + r)$. A calculator's log key may use base 10 rather than e. Either works if the same base is used on top and bottom, since, for any base b, $\log_b x = \ln x / \ln b$.

3c. Doubling Times and the Rule of 72

Setting $A = 2P$ in the time formula gives the doubling time at a growth rate r:

$$t_2 = \frac{\ln 2}{\ln(1 + r)} \approx \frac{0.693}{r}$$

The approximation uses $\ln(1 + r) \approx r$, which holds for small rates and which Module 9 explains. In percentage terms, the doubling time is about 69.3 divided by the rate. The rule of 72 uses 72 instead, and not arbitrarily. Because $\ln(1 + r)$ is slightly less than r, the exact constant for yearly compounding is a little above 69.3, and near 8% it is almost exactly 72. Seventy-two also divides evenly by 2, 3, 4, 6, 8, 9, and 12.

Growth rate	Rule of 72 (years)	Exact, yearly compounding (years)
2%	36	35.0
4%	18	17.7
6%	12	11.9
8%	9	9.0
10%	7.2	7.3
12%	6	6.1

The rule of 72 against the exact doubling time, $\ln 2/\ln(1+r)$.

Doubling times turn growth rates into intuition. Prices rising 3% a year double in about 24 years, so a dollar loses half its purchasing power in that time. An economy growing 2% a year doubles its output in 35 years; one growing 7% doubles in about ten.

Worked example — How fast AI training compute doubles

Epoch AI, a research group that tracks the field, estimated in 2024 that the computing power used to train leading AI models had grown about 4 to 5 times a year since 2010. A quantity growing by a factor g a year doubles in $\ln 2/\ln g$ years. For $g = 4.5$, that is $0.693/1.504 \approx 0.46$ years, about five and a half months. By the same formula, transistor counts doubling every two years, the classic pace of Moore's law, grow by a factor of $2^{1/2} \approx 1.41$ a year.

3d. Decay and Half-Lives

Exponential decay is growth by a factor below 1. A quantity that loses a fixed fraction of itself each period shrinks toward zero without reaching it. Decay is usually described by its half-life T, the time for half of the quantity to go. Starting from an amount N_0, the amount $N(t)$ left after a time t is

$$N(t) = N_0 \left(\tfrac{1}{2}\right)^{t/T} = N_0\, e^{-\lambda t}, \qquad \lambda = \frac{\ln 2}{T}$$

The two forms are the same function, since $(1/2)^{t/T} = e^{-(\ln 2)t/T}$: the decay constant λ is $\ln 2$ divided by the half-life. Radioactive isotopes, drugs leaving the bloodstream, and discount factors all follow this curve.

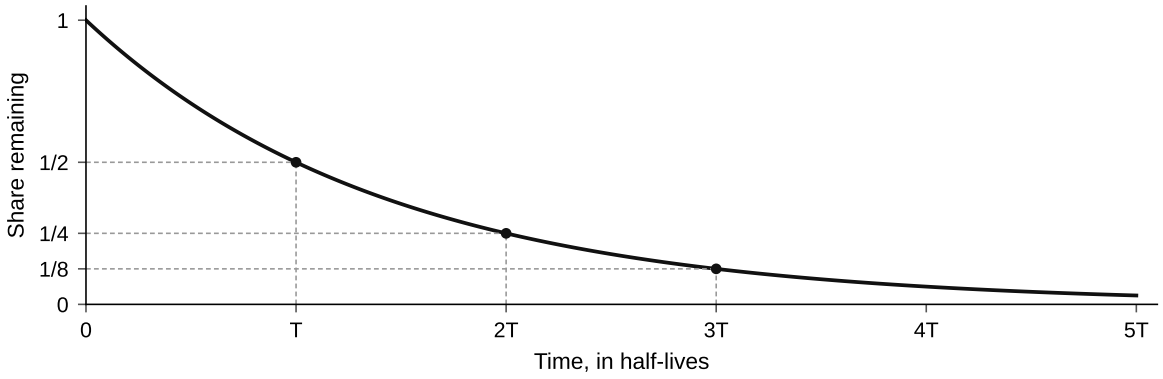


Figure 3.1. Exponential decay. Each half-life removes half of what is left: a half, then a quarter, then an eighth.

Worked example — Carbon dating

Carbon-14 has a half-life of about 5,730 years. A wood sample holds 30% of the carbon-14 it had when the tree died. Solve $0.3 = (1/2)^{t/5730}$ by taking logs: $t = 5730 \times \ln 0.3/\ln 0.5 \approx 5730 \times 1.737 \approx 9{,}950$ years.

Money decays too. Inflation at 3% a year shrinks a dollar's purchasing power by the factor 1.03^{-t}, a half-life of about 23 years. The discount factor $(1+r)^{-t}$ of Module 1 is the same curve, and it is why a payment due in 30 years is worth less than a quarter of its face value today at 5%: $1.05^{-30} \approx 0.23$.

3e. Log Scales and Learning Curves

On an ordinary axis, equal steps add equal amounts. On a logarithmic axis, equal steps multiply by equal factors, so 1, 10, 100, and 1,000 sit equally spaced. Because $\ln(Pe^{rt}) = \ln P + rt$, anything growing at a steady percentage rate plots as a straight line on a log axis, and the line's slope is the growth rate. A curve that bends upward on a log chart is accelerating; one that bends downward is slowing.

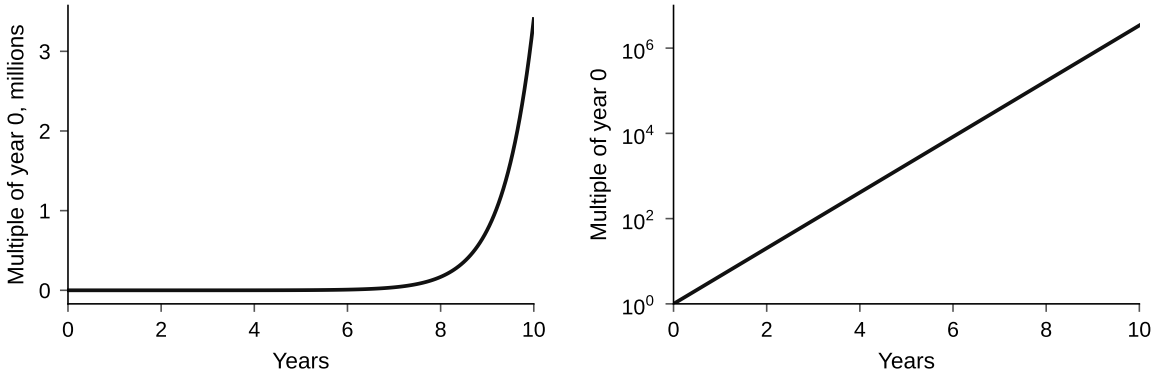


Figure 3.2. The same series, growing 4.5 times a year for ten years, on a linear axis (left) and a log axis (right). On the linear axis nothing seems to happen for seven years; on the log axis the growth is visibly steady.

Learning curves are the other place log scales earn their keep. For solar panels, batteries, and many manufactured goods, unit cost has fallen by a roughly constant percentage each time cumulative production doubled. That percentage is the learning rate: about 20% per doubling for solar modules over five decades, and about 18 to 20% for lithium-ion batteries. A learning curve is a power function of cumulative production Q, the total number of units ever made. If the cost was C_0 when cumulative production was Q_0, the cost C at any later Q is

$$C(Q) = C_0\left(\frac{Q}{Q_0}\right)^{-b}, \qquad 2^{-b} = 1 - \text{learning rate}$$

On axes that are logarithmic in both directions, a power function is a straight line with slope $-b$. A 20% learning rate means $2^{-b} = 0.8$, so $b = -\ln 0.8/\ln 2 \approx 0.32$.

Worked example — Ten doublings

A technology with a 20% learning rate goes through ten doublings of cumulative production, a 1,024-fold increase. Its cost falls to $0.8^{10} \approx 0.107$ of where it started, a decline of 89%. Solar module prices have fallen more than 99% since the mid-1970s, roughly what a 20% learning rate predicts over the twenty or so doublings of cumulative production since then.

Learning curves are regularities, not laws. They describe what happened while an industry scaled, and they can stall when raw-material costs bind or production stops growing. Read them like any log chart: as a record of a steady percentage rate, useful for exactly as long as the rate holds.

3f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Compound growth	$A = P(1+r)^t \qquad A = P\left(1+\frac{r}{n}\right)^{nt} \qquad A = Pe^{rt}$ P is the starting amount, A the amount after t years, r the annual rate, and n the number of times a year interest is credited
The number e	$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \approx 2.718$
Logs undo exponentials	$e^{\ln x} = x \qquad \ln(e^x) = x \qquad \ln 1 = 0 \qquad \ln e = 1$ ln is the natural logarithm, base e
Log rules	$\ln(ab) = \ln a + \ln b \qquad \ln\frac{a}{b} = \ln a - \ln b \qquad \ln(a^k) = k\ln a$ a and b are positive numbers; k is any power
Time to grow from P to A	$t = \frac{\ln(A/P)}{\ln(1+r)}$ t is in years when r is a yearly rate
Doubling time	$\ln 2/\ln(1+r)$, about 72 divided by the rate in percent; $\ln 2 \approx 0.693$
Half-life	$N = N_0(\frac{1}{2})^{t/T} = N_0e^{-\lambda t} \qquad \lambda = \frac{\ln 2}{T}$ N_0 is the starting amount, N the amount left after time t, T the half-life, and λ the decay constant
Log axes	Steady percentage growth is a straight line, and its slope is the growth rate
Learning curves	Cost falls by the learning rate with each doubling of cumulative production: about 20% for solar modules $C = C_0(Q/Q_0)^{-b}$: C_0 is the cost when cumulative production was Q_0, and 2^{-b} is one minus the learning rate
The idea	Exponential growth multiplies by a constant factor each period, so it always outruns linear growth: $1,000 at 5% becomes $4,322 in 30 years.

3g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

3.1 Evaluate $\ln 1$, $\ln(e^4)$, $\log_{10} 0.01$, and $e^{\ln 7}$.

3.2 Write $\ln 6 + \ln 5 - \ln 3$ as a single logarithm.

3.3 Solve $3^x = 81$ and $e^{0.5t} = 20$.

3.4 Use the rule of 72 to estimate doubling times at 3%, 6%, and 9% a year.

3.5 What is $5,000 worth after 10 years at 6% a year, compounded yearly? Monthly? Continuously?

3.6 A substance has a half-life of 8 days. What fraction remains after 24 days? After 30 days?

Applied

3.7 How long does money take to triple at 8% a year, compounded yearly?

3.8 A credit card charges 24% a year, compounded monthly. What is the effective annual rate?

3.9 Prices rise 4% a year. How many years until a dollar buys half as much? Use the rule of 72, then the exact formula.

3.10 If AI training compute grows 4 times a year, how many years until it has grown a thousandfold?

3.11 A battery cell costs $100 per kilowatt-hour today, and its industry's learning rate is 18%. If cumulative production doubles three more times, what does the learning curve predict?

3.12 A drug has a half-life of 6 hours. What decay constant describes it, and what fraction remains after 24 hours?

Challenge

3.13 Show that a continuous rate r and a yearly compounded rate R give the same growth when $R = e^{r} - 1$. What yearly rate matches a continuous 5%?

3.14 Using $\ln(1 + r) \approx r - r^{2}/2$, show that the doubling time is about $(0.693/r)(1 + r/2)$, and use it to explain why 72 works best near 8%.

3.15 On a chart with a log vertical axis, a series rises in a straight line from 10 to 1,000 over 8 years. What is its annual growth rate?

3.16 Solve $2^{x+1} = 3^{x}$ exactly, then give the decimal answer.

Module 3 Quiz

Check Your Understanding

1. At 9% a year, about how long does money take to double?

A) 9 years
B) 11 years
C) About 8 years
D) About 18 years, twice the rate

2. What is $\ln(e^3)$?

A) e^3
B) 3
C) About 20.1
D) It cannot be simplified without a calculator, since e is irrational

3. A sample has 25% of its original carbon-14 left. How old is it, given a half-life of 5,730 years?

A) 1,433 years
B) About 11,460 years
C) About 22,920 years, four half-lives
D) About 4,300 years, three-quarters of one half-life

4. On a chart with a logarithmic vertical axis, steady 7% annual growth appears as:

A) A flat line
B) A curve bending upward
C) A curve that flattens out over time, because each gridline is ten times the one below it
D) A straight line rising to the right

5. A technology's cost falls 20% with each doubling of cumulative production. After five doublings, what fraction of the original cost remains?

A) About a third, since each doubling multiplies the cost by 0.8: $0.8^5 \approx 0.33$
B) About 80%
C) About half
D) Nothing: five 20% cuts add up to 100%

Module 4 — Rates of Change

From average to instantaneous: the derivative as a rate

Recommended Reading / Reference

Book: Silvanus P. Thompson and Martin Gardner, *Calculus Made Easy* (1910; revised 1998) — the classic gentle introduction, built on the idea of a little bit of something

Reference: PJM Data Miner (dataminer2.pjm.com) — hourly and five-minute locational marginal prices across the PJM grid, free to download

4a. Average Rates of Change

Calculus is about rates. Its first question is how fast something changes, and the simplest answer is an average over an interval. For a function f and two inputs a and b, the average rate of change from a to b is the change in output divided by the change in input:

$$\frac{\Delta f}{\Delta x} = \frac{f(b) - f(a)}{b - a}$$

It is the slope of the straight line through the two points on the graph, a line called a secant, and its units are the output's units per unit of input.

A solar plant's meter records cumulative energy: the megawatt-hours produced since sunrise. The table shows a 100 MW plant on a clear day, with illustrative numbers. Dividing each hour's change in energy by one hour gives the plant's average power over that hour.

Hour	6 a.m.	7	8	9	10	11	Noon
Energy since 6 a.m. (MWh)	0	13.0	51.2	111.9	191.0	283.1	382.0
Average power over the next hour (MW)	13.0	38.2	60.7	79.1	92.1	98.9	—

A 100 MW solar plant on a clear day. Average power is the change in energy over each hour.

An average hides what happens inside its interval. From 9 to 10 a.m. the plant averaged 79.1 MW, but its output was climbing the whole time: at 9 a.m. it was producing less than that, and at 10 a.m. more. To find the power at 9 a.m. itself, shrink the interval.

4b. Limits, Only as Far as Needed

Measured over shorter and shorter intervals starting at 9 a.m., the plant's average power settles toward a single number:

Interval after 9 a.m.	1 hour	30 minutes	6 minutes	36 seconds	3.6 seconds
Average power (MW)	79.1	75.1	71.6	70.8	70.7

The number the averages approach, 70.7 MW, is the plant's power at 9 a.m. In symbols, write $E(t)$ for the cumulative energy in megawatt-hours at hour t, and h for the length of the interval in hours. The value the average approaches as h shrinks toward zero is its limit, written with $\lim$:

$$\lim_{h \to 0} \frac{E(9+h) - E(9)}{h} \approx 70.7 \text{ MW}$$

A limit is not found by setting h to zero, which would divide zero by zero. It is the value an expression approaches as h gets closer and closer to zero. Module 1 showed how algebra usually finds it: simplify the quotient until h can be set to zero safely.

Limits matter in this course because they define the derivative, and they also show where a derivative fails to exist. At a jump in a function there is no single value to approach. At a kink, like the ReLU's corner at zero, the slope approaches 1 from the right and 0 from the left, so the two sides disagree. Neural-network software simply assigns the ReLU a slope at zero, usually 0, and the choice almost never matters.

4c. The Derivative

The derivative of a function f at an input x is the limit of its average rate of change over a shrinking step h:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

It is the instantaneous rate of change of f at x. On the graph it is the slope of the tangent line, the straight line that touches the curve at that point and runs in the curve's direction there. Its units are the output's units per unit of input.

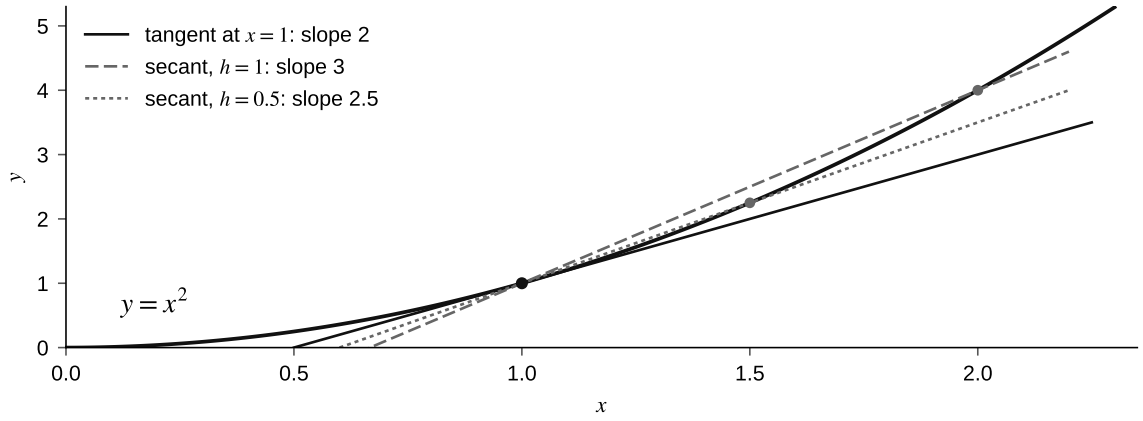


Figure 4.1. Secant lines through the point (1, 1) on $y = x^2$ tilt toward the tangent as the second point moves closer: slopes of 3 and 2.5 for steps h of 1 and 0.5, approaching the tangent's slope of 2.

Three notations for the derivative are in common use, and this course uses all of them. If $y = f(x)$, then

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

The second, Leibniz's notation, is read “dee y dee x.” It is not a fraction, but it behaves like one in the rules of Module 5, and it recalls where the derivative came from: a small change in y divided by the small change in x that caused it.

Module 1 found the derivative of x^2 this way: the difference quotient simplified to $2x + h$, which approaches $2x$. The same algebra handles any quadratic.

Worked example — The derivative of any quadratic

For $f(x) = a + bx + cx^2$, where a, b, and c are numbers, expand and cancel:

$$\frac{f(x+h) - f(x)}{h} = \frac{b\,h + c\,(2xh + h^2)}{h} = b + 2cx + ch \;\longrightarrow\; b + 2cx$$

So $f'(x) = b + 2cx$: the constant a contributes nothing, the linear term contributes its slope b, and the squared term contributes $2cx$. Module 5 turns patterns like this into rules, so the limit is rarely needed again.

Worked example — Revenue and price

A store sells $q = 400 - 8p$ units a week at price p dollars, so its weekly revenue is $R(p) = 400p - 8p^2$ dollars. By the quadratic result, $R'(p) = 400 - 16p$, in dollars of revenue per dollar of price. At $p = 20$, $R'(20) = 80$: a price one dollar higher would bring in about \$80 more a week. At $p = 30$, $R'(30) = -80$: raising the price would lose about \$80. At $p = 25$ the derivative is zero and revenue is at its peak, the price Module 6 finds directly.

4d. Rates in Practice

Most derivatives in business and engineering have names of their own. Marginal cost is the derivative of total cost with respect to output, the cost of one more unit. If $C(q)$ is the cost in dollars of producing q units, the marginal cost $C'(q)$ is in dollars per unit.

Worked example — Marginal and average cost

A plant's cost of producing q units is $C(q) = 5{,}000 + 20q + 0.01q^2$ dollars. Its marginal cost is $C'(q) = 20 + 0.02q$, which at $q = 1{,}000$ is \$40 a unit. Its average cost there is $C(1{,}000)/1{,}000 = 35{,}000/1{,}000 = 35$ dollars. Because the next unit costs more than the average, making it raises the average: marginal cost above average pulls the average up, and below average pulls it down.

Power markets price electricity at the margin. In an organized wholesale market such as PJM, the locational marginal price, or LMP, at a point on the grid is the cost of serving one more megawatt-hour of demand there, given every generator's offer and every transmission limit. It is a derivative: the rate at which the total cost of running the system changes as demand at that location rises. When transmission lines are congested, serving one more megawatt-hour in a crowded city can require a more expensive local plant, which is why prices differ by location.

Worked example — A marginal price

In a simplified system, the hourly cost of serving a demand of L megawatts is $C(L) = 15L + 0.002L^2$ dollars. The marginal cost of energy is $C'(L) = 15 + 0.004L$ dollars per megawatt-hour, which at a demand of 10,000 MW is \$55. Every megawatt-hour sold that hour clears at \$55, although the average cost of the energy is lower: $C(10{,}000)/10{,}000 = 35$ dollars.

Power is the derivative of energy. If $E(t)$ is the energy in megawatt-hours a battery holds at hour t, then dE/dt is in megawatt-hours per hour, which is megawatts: positive while charging, negative while

discharging. The solar plant's 70.7 MW at 9 a.m. was the derivative of its cumulative output. Module 7 runs the idea backward, adding up power over time to get energy.

A rate is often more useful relative to the size of the thing changing. A stock priced at S dollars that rises at dS/dt dollars a day has a relative rate of change of $(dS/dt)/S$ per day: a \$4 rise in a day is 5% for an \$80 stock and 1% for a \$400 one. Module 5 shows that this relative rate is the derivative of the logarithm.

4e. Reading Derivatives from Graphs

A derivative's sign and size can be read from a graph without algebra. Where f rises, f' is positive; where it falls, f' is negative; at the top of a peak or the bottom of a trough, where the tangent is flat, f' is zero. The steeper the graph, the larger the derivative in size.

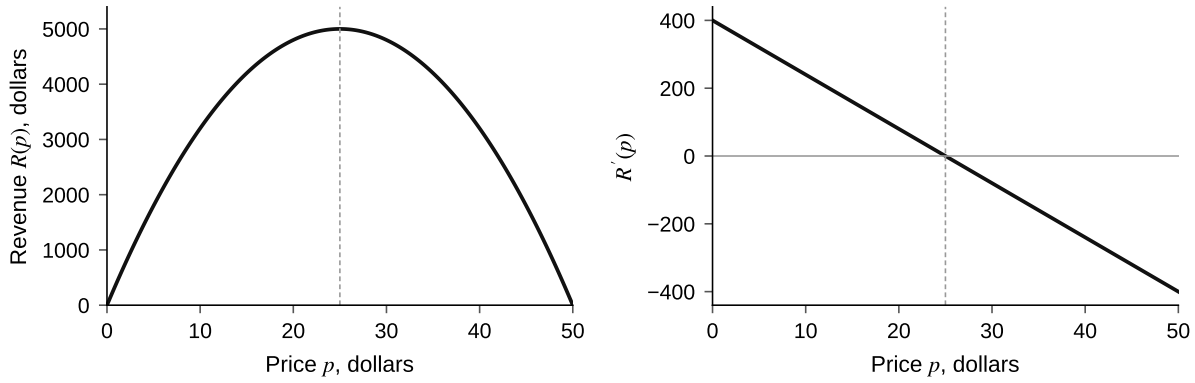


Figure 4.2. Weekly revenue (left) and its derivative (right). Revenue climbs while the derivative is positive, peaks where the derivative crosses zero at a price of \$25, and falls after.

The derivative is itself a function, and its own derivative, the second derivative $f''(x)$, measures how the slope is changing. When f'' is positive the slope is rising and the graph bends upward, like a cup; when it is negative the graph bends downward, like a cap. Revenue's second derivative is $R''(p) = -16$, negative everywhere, which is why its graph is a cap with a single peak. Module 6 uses the second derivative to tell peaks from troughs, and Module 9 uses it to measure a bond's convexity.

Data rarely arrives as a formula. To estimate a derivative from readings spaced h apart, use the readings on both sides of the point, a central difference:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}$$

For the solar plant at 9 a.m., with readings an hour on either side, $(191.0 - 51.2)/2 \approx 69.9$ MW, about 1% below the true 70.7 MW. The one-sided average from 9 to 10 a.m., 79.1 MW, was 12% too high. A central difference is centered on the point, so the errors from the two sides largely cancel.

4f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Average rate of change	$\frac{f(b) - f(a)}{b - a}$ Between the inputs a and b: the slope of the secant line, in output units per input unit
The derivative	$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$ The instantaneous rate of change at x and the slope of the tangent line; h is the step
Notation	$f'(x) = \frac{dy}{dx} = \frac{d}{dx} f(x)$ For $y = f(x)$; dy/dx is read “dee y dee x”
Quadratics	If $f(x) = a + bx + cx^2$, then $f'(x) = b + 2cx$ a, b, and c are numbers
Signs of the derivative	$f' > 0$: rising; $f' < 0$: falling; $f' = 0$: flat, as at a peak or a trough
Second derivative	$f'' > 0$: the graph bends upward, like a cup; $f'' < 0$: it bends downward, like a cap f'' is the derivative of f'
Marginal cost	$C'(q)$: the cost of one more unit; marginal above average pulls the average up $C(q)$ is the total cost of q units
Marginal price of power	LMP $= dC/dL$: the cost of serving one more MWh at a location C is the system's total cost and L the demand at that location
Power and energy	$P = dE/dt$: megawatt-hours per hour are megawatts $E(t)$ is stored or cumulative energy at hour t, and P is power
Central difference	$f'(x) \approx \frac{f(x + h) - f(x - h)}{2h}$ Estimates a derivative from readings spaced h apart
No derivative	At a jump or a kink, such as the ReLU's corner at zero
The idea	A derivative is a rate measured at a single point: how fast one quantity changes per unit change in another, found as the limit of averages over shorter and shorter intervals.

4g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

4.1 For $f(x) = x^2 + 3x$, find the average rate of change from $x = 1$ to $x = 4$.

4.2 Use the definition of the derivative to find $f'(x)$ for $f(x) = 5x - 2$.

4.3 Use the definition to find the derivative of $f(x) = 3x^2$.

4.4 For $f(x) = x^2$, compute the difference quotient at $x = 3$ for $h = 0.1$ and $h = 0.01$. What is $f'(3)$?

4.5 What are the units of dC/dq if C is a cost in dollars and q is output in tons? Of dE/dt if E is energy in megawatt-hours and t is time in hours?

4.6 A graph is rising and getting steeper. What are the signs of f' and f''?

Applied

4.7 A battery holds 380 MWh at 6 p.m. and 180 MWh at 8 p.m. What was its average power over those two hours?

4.8 A company's cost of producing q units is $C(q) = 8{,}000 + 30q + 0.02q^2$ dollars. Find the marginal cost at $q = 500$ and compare it with the average cost there.

4.9 Revenue at price p is $R(p) = p(600 - 12p)$ dollars. Find $R'(p)$, and say whether a small price rise from $p = 20$ raises or lowers revenue.

4.10 In a simplified power system, the hourly cost of serving a demand of L megawatts is $C(L) = 12L + 0.0025L^2$ dollars. Find the marginal cost of energy at 8,000 MW.

4.11 A plant's cumulative output reads 410 MWh at 1 p.m., 492 MWh at 2 p.m., and 566 MWh at 3 p.m. Estimate its power at 2 p.m. with a central difference.

4.12 A stock's price rises from \$80 to \$84 in a day. What is its rate of change, and what is its relative rate of change?

Challenge

4.13 Use the definition to show that the derivative of $\sqrt{x}$ is $1/(2\sqrt{x})$. (Multiply the top and bottom of the difference quotient by $\sqrt{(x+h)} + \sqrt{x}$.)

4.14 Show that the ReLU, $\max(x, 0)$, has no derivative at $x = 0$, by computing its difference quotient there for small positive and small negative h.

4.15 For $f(x) = x^3$ at $x = 2$ with $h = 0.1$, compute the one-sided difference $[f(2.1) - f(2)]/0.1$ and the central difference $[f(2.1) - f(1.9)]/0.2$, and compare them with the exact derivative, 12.

4.16 A tank holds $V(t) = 1{,}000 - 40t + 0.5t^2$ liters at t minutes, for t from 0 to 40. How fast is the volume changing at $t = 10$, and when does the tank stop draining?

Module 4 Quiz

Check Your Understanding

1. A battery's stored energy falls from 300 MWh to 200 MWh between 7 p.m. and 9 p.m. What is its average discharge power?

A) 100 MW
B) 250 MWh
C) 50 MW
D) 200 MW, because power is the energy that remains divided by the hours spent discharging

2. What does $f'(x)$ measure?

A) The value of f at x
B) The slope of the tangent line at x
C) The average of f over an interval around x
D) The total change in f from zero up to x, which is the area beneath the curve

3. A firm's cost of producing q units is $C(q) = 5{,}000 + 20q + 0.01q^2$ dollars. What is its marginal cost at $q = 1{,}000$?

A) \$20
B) \$35
C) \$40 per unit
D) \$35,000, because marginal cost is the total cost divided by the number of units

4. Why can't a derivative be found by setting $h = 0$ in the difference quotient?

A) Because the quotient becomes $0/0$, which has no value, so the derivative must be found as the limit the quotient approaches as h shrinks.
B) Because h must always be a whole number.
C) Because the derivative is defined only for linear functions.
D) Because setting $h = 0$ gives the average rate of change instead.

5. What is the derivative of the ReLU, $\max(x, 0)$, at $x = 0$?

A) It equals one, because the function rises to the right of zero.
B) It equals zero, because the function is flat to the left of zero.
C) It equals one-half, the average of the two slopes on either side.
D) It does not exist.

Module 5 — The Rules of Differentiation

Power, product, quotient, exponential, logarithm, and chain rules

Recommended Reading / Reference

Book: Gilbert Strang, *Calculus* (1991) — a complete text by an MIT mathematician with an applied bent, free from MIT OpenCourseWare

Reference: Andrej Karpathy, *micrograd* (github.com/karpathy/micrograd, 2020) — a tiny engine that computes derivatives for neural networks, showing backpropagation as the chain rule

5a. Powers, Sums, and Constants

Module 4 found derivatives from the definition: x^2 gives $2x$, x^3 gives $3x^2$, and $1/x$, which is x^{-1}, gives $-x^{-2}$. The pattern holds for every power. For any number n:

$$\frac{d}{dx}x^n = n\,x^{n-1}$$

The exponent comes down as a multiplier and drops by one. The rule covers negative and fractional powers, which is why Module 1 rewrote roots and reciprocals as powers: $\sqrt{x} = x^{1/2}$ has derivative $(1/2)x^{-1/2} = 1/(2\sqrt{x})$.

Three more rules let the power rule work on whole expressions. A constant has no rate of change, so its derivative is zero. A constant multiplier carries through: the derivative of $k \cdot f(x)$ is $k \cdot f'(x)$ for any number k. And a sum differentiates term by term: for functions f and g, the derivative of $f + g$ is $f' + g'$.

Worked example — Roots and reciprocals

Differentiate $y = 4x^3 - 2\sqrt{x} + 5/x$. Rewrite the root and the reciprocal as powers, then apply the power rule term by term:

$$\frac{dy}{dx} = \frac{d}{dx}\left(4x^3 - 2x^{1/2} + 5x^{-1}\right) = 12x^2 - x^{-1/2} - 5x^{-2} = 12x^2 - \frac{1}{\sqrt{x}} - \frac{5}{x^2}$$

5b. Exponentials and Logarithms

The exponential e^x has a property no other function shares: it is its own derivative.

$$\frac{d}{dx}e^x = e^x$$

At every point the curve's slope equals its height. The reason is the choice of e. The difference quotient of e^x is $e^x(e^h - 1)/h$, and e is exactly the base for which $(e^h - 1)/h$ approaches 1 as the step h shrinks: at $h = 0.001$ it is already 1.0005. For any constant k, the exponential e^{kx} grows at k times its own size, which Module 8 turns into the equation of growth and decay:

$$\frac{d}{dx}e^{kx} = k\,e^{kx}$$

Any other positive base a can be written as a power of e, since $a = e^{\ln a}$. So $a^x = e^{x \ln a}$, and its derivative is $a^x \ln a$. A balance growing 5% a year, 1.05^t after t years, grows at $\ln 1.05 \approx 0.0488$ times its size each year: 4.88% is the continuous rate equivalent to 5% compounded yearly.

The natural logarithm's derivative is simpler still. For positive x,

$$\frac{d}{dx} \ln x = \frac{1}{x}$$

A small change dx in x changes $\ln x$ by dx/x, the change as a fraction of x. That is why logarithms measure percentage changes, and why financial economists work with the logarithms of prices.

Worked example — A continuous discount factor

At a continuous rate of 5%, a dollar due in t years is worth $D(t) = e^{-0.05t}$ today. Its rate of change is $D'(t) = -0.05e^{-0.05t}$: the discount factor falls by 5% of its current value each year, which is what a constant 5% rate means.

5c. Products and Quotients

The derivative of a product is not the product of the derivatives. For functions f and g:

$$(fg)' = f'g + fg' \qquad \left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

The product rule has a plain meaning: when both factors change, the product changes by the first factor's change times the second factor, plus the first factor times the second's change. The quotient rule follows from it, as a challenge problem shows.

Worked example — Why a price rise can cut revenue

Revenue is price times quantity, $R(p) = p \cdot q(p)$, where the quantity sold, $q(p)$, depends on the price p. The product rule gives

$$R'(p) = q(p) + p\,q'(p)$$

A higher price earns more on each of the q units sold, the first term, but loses sales, the second term, since $q'(p)$ is negative. For $q = 400 - 8p$, $R'(p) = (400 - 8p) + p(-8) = 400 - 16p$, the result Module 4 reached from the definition.

Worked example — Average cost from marginal cost

Average cost is $A(q) = C(q)/q$, the total cost $C(q)$ divided by the q units made. The quotient rule gives $A'(q) = [C'(q) \cdot q - C(q)]/q^2 = [C'(q) - A(q)]/q$. Average cost falls while marginal cost is below it and rises once marginal cost is above it: Module 4's rule of thumb, now proved.

5d. The Chain Rule

Most functions met in practice are compositions, one function applied to another's output. If $y = f(u)$ and the inside function is $u = g(x)$, the chain rule says the rates multiply:

$$\frac{d}{dx} f(g(x)) = f'(g(x))\, g'(x) \qquad \text{or} \qquad \frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

Differentiate the outside function, leaving the inside alone, then multiply by the derivative of the inside. Module 2's summer demand is the model: if demand rises 60 MW per degree and the temperature rises 1.5 degrees an hour, demand rises $60 \times 1.5 = 90$ MW an hour.

Function	Outside, inside	Derivative
$(3x + 1)^5$	u^5, with $u = 3x + 1$	$5(3x + 1)^4 \cdot 3 = 15(3x + 1)^4$
e^{-rt}, as a function of t	e^u, with $u = -rt$	$e^{-rt} \cdot (-r) = -re^{-rt}$
$\ln(1 + x^2)$	$\ln u$, with $u = 1 + x^2$	$2x/(1 + x^2)$
$\sqrt{(1 + x^2)}$	$u^{1/2}$, with $u = 1 + x^2$	$(1/2)(1 + x^2)^{-1/2} \cdot 2x = x/\sqrt{(1 + x^2)}$

The chain rule at work. u names the inside function, and r is a constant rate.

Worked example — How a bond's price responds to its yield

A ten-year zero-coupon bond pays $1,000 at maturity. At a yield y, its price today is $P(y) = 1000(1 + y)^{-10}$. The outside function is the power u^{-10}, and the inside is $u = 1 + y$, whose derivative is 1:

$$P'(y) = -10{,}000\,(1 + y)^{-11} \qquad P'(0.04) = -\frac{10{,}000}{1.04^{11}} \approx -6{,}496$$

At a 4% yield the bond costs about $675.56, and its price falls about $6,496 per unit of yield. A unit of yield is 100 percentage points, so a rise of one point, 0.01, cuts the price by about $65, and a rise of one basis point by about 65 cents. Traders call that last figure the bond's DV01, its dollar value of a basis point; Module 9 turns the same derivative into duration.

5e. The Sigmoid and Backpropagation

Neural networks are built from long chains of simple functions, and they learn by the chain rule. The sigmoid σ, an activation that squashes any input z into a number between 0 and 1, is itself a composition:

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \left(1 + e^{-z}\right)^{-1}$$

By the chain rule, with the outside a power and the inside $1 + e^{-z}$, its derivative takes a convenient form:

$$\sigma'(z) = \frac{e^{-z}}{(1 + e^{-z})^2} = \sigma(z)(1 - \sigma(z))$$

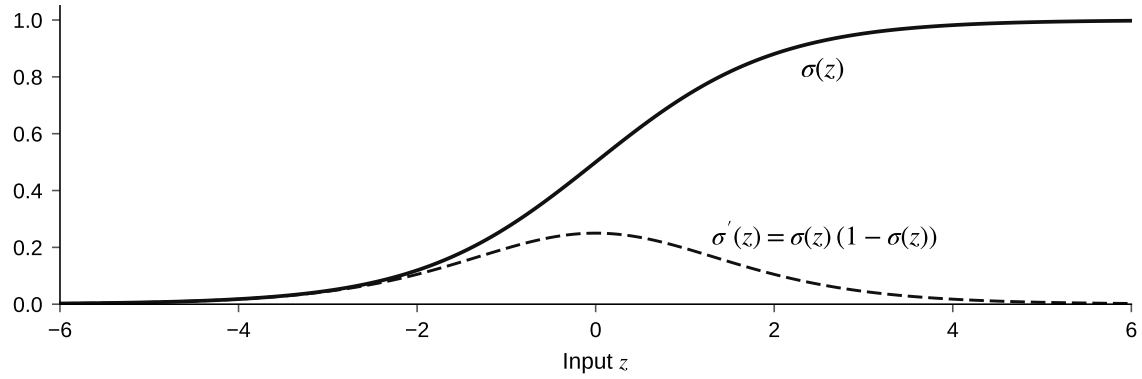


Figure 5.1. The sigmoid (solid) and its derivative (dashed). The derivative peaks at 0.25 where the input is zero and fades for large inputs of either sign, which is one reason networks built from sigmoids can learn slowly.

A single neuron takes an input x, multiplies it by a weight w, adds a bias b, and passes the sum $z = wx + b$ through the sigmoid to produce an output $a = \sigma(z)$. Training compares a with the target value y the neuron should produce, using a loss such as $L = (a - y)^2$, and adjusts w and b to make the loss smaller. That requires the derivative of the loss with respect to each weight, a chain of three links:

$$\frac{dL}{dw} = \frac{dL}{da} \cdot \frac{da}{dz} \cdot \frac{dz}{dw} = 2(a - y) \cdot \sigma(z)(1 - \sigma(z)) \cdot x$$

Worked example — One learning step

Take $x = 1.5$, $w = 0.8$, $b = -0.2$, and target $y = 1$. Then $z = 1.0$, $a = \sigma(1) \approx 0.731$, and the loss is $(0.731 - 1)^2 \approx 0.072$. The three links are $2(a - y) \approx -0.538$, $\sigma'(1) \approx 0.197$, and $x = 1.5$, so $dL/dw \approx -0.159$. The derivative is negative, so a larger weight lowers the loss. Moving w against the derivative by half its size, to $0.8 + 0.5 \times 0.159 \approx 0.879$, raises the output to 0.754 and cuts the loss to 0.061.

A network stacks thousands of such neurons in layers, so its loss is a composition many functions deep. Backpropagation is the chain rule applied from the loss backward, one layer at a time, reusing each layer's product of derivatives for the layer before it. That reuse makes the whole calculation cost only a small multiple of running the network forward once, which is what makes training networks with billions of weights possible. Module 8 shows that the step taken above, moving each weight against its derivative, is a numerical method for solving a differential equation.

5f. Logarithmic Derivatives and Elasticity

By the chain rule, the derivative of the logarithm of a positive function f is its relative rate of change:

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

When the input is time, this is the percentage growth rate. If an economy's output Y grows at 2% of its size per year, then $d(\ln Y)/dt = 0.02$, with t in years, which is why charts of growth plot the logarithm.

Economists measure how strongly buyers respond to price with the price elasticity of demand, written ε: the percentage change in the quantity q bought per percentage change in the price p.

$$\varepsilon = \frac{dq/q}{dp/p} = \frac{p}{q}\frac{dq}{dp}$$

Demand is inelastic when ε is between −1 and 0: quantity falls by a smaller percentage than price rises, so revenue rises with price. It is elastic when ε is below −1, and then revenue falls as price rises.

Worked example — Elasticity along a demand line

For $q = 400 - 8p$, the slope dq/dp is −8. At $p = 20$, $q = 240$ and $\varepsilon = (20/240)(-8) \approx -0.67$: inelastic, so a price rise raises revenue, as $R'(20) = 80$ showed. At $p = 30$, $q = 160$ and $\varepsilon = -1.5$: elastic, so revenue falls. At $p = 25$, $q = 200$ and $\varepsilon = -1$ exactly, where revenue peaks.

5g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Power rule	$\frac{d}{dx}x^n = n\,x^{n-1}$ n is any number, including negative and fractional powers
Constants and sums	$d(k)/dx = 0$; $(k \cdot f)' = k \cdot f'$; $(f + g)' = f' + g'$ k is a constant; f and g are functions of x
Exponentials	$\frac{d}{dx}e^x = e^x \qquad \frac{d}{dx}e^{kx} = k\,e^{kx} \qquad \frac{d}{dx}a^x = a^x \ln a$ k is a constant and a a positive base
Logarithm	$\frac{d}{dx}\ln x = \frac{1}{x}$ For positive x
Product and quotient	$(fg)' = f'g + fg' \qquad \left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$ f and g are functions of the same input
Chain rule	$\frac{d}{dx}f(g(x)) = f'(g(x))\,g'(x) \qquad \frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$ $u = g(x)$ is the inside function: differentiate the outside, then multiply by the inside's derivative
Sigmoid	$\sigma(z) = \frac{1}{1 + e^{-z}} \qquad \sigma'(z) = \sigma(z)(1 - \sigma(z))$ z is the neuron's input; the output lies between 0 and 1
Log derivative	$\frac{d}{dx}\ln f(x) = \frac{f'(x)}{f(x)}$ The relative, or percentage, rate of change of a positive function f
Elasticity	$\varepsilon = \frac{p}{q}\frac{dq}{dp}$ q is quantity and p price; revenue rises with price when ε is between −1 and 0
The idea	A handful of rules differentiate everything in this course. The chain rule, which multiplies the rates of nested functions, is the one used most, and backpropagation is the chain rule repeated.

5h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

5.1 Differentiate $x^5 - 4x^2 + 7$.

5.2 Differentiate $6\sqrt{x} + 2/x^3$.

5.3 Differentiate $3e^{2t} - 5\ln t$ with respect to t.

5.4 Differentiate $x^2\,e^x$.

5.5 Differentiate $x^3 \ln x$.

5.6 Differentiate $(2x + 1)/(x - 3)$.

5.7 Differentiate $(4x - 1)^6$.

5.8 Differentiate e^{-x^2}.

5.9 Differentiate $\ln(1 + 3x)$.

5.10 Differentiate $\sqrt{(5 + x^2)}$.

Applied

5.11 A 20-year zero-coupon bond pays \$1,000 at maturity and is priced at $P(y) = 1000(1 + y)^{-20}$ at a yield y. Find $P'(y)$ at a 5% yield, and estimate the price change if the yield rises by 0.1 percentage point.

5.12 A deposit is worth $A(t) = 10{,}000e^{0.04t}$ dollars after t years. How fast is it growing at $t = 10$?

5.13 Demand is $q = 1{,}200 - 15p$ units at price p. Find the price elasticity at $p = 30$ and at $p = 50$. At each price, does a small price rise raise revenue?

5.14 A firm's total cost of q units is $C(q) = 2{,}000 + 10q + 0.05q^2$ dollars. Write its average cost $A(q) = C(q)/q$, find $A'(q)$, and find the output at which average cost is lowest.

5.15 For the neuron of section 5e, with $x = 1.5$, $w = 0.8$, $b = -0.2$, and target $y = 1$, find the derivative of the loss with respect to the bias b.

5.16 In a simple model, a battery's capacity after n full cycles is $Q(n) = Q_0\,e^{-0.0001n}$, where Q_0 is its capacity when new. At what rate does capacity fall per cycle, as a share of current capacity? What share is left after 3,000 cycles?

Challenge

5.17 Derive the quotient rule from the product and chain rules, by writing $f/g = f \cdot g^{-1}$.

5.18 Show that the derivative of a^x is $a^x \ln a$, by writing $a^x = e^{x\ln a}$.

5.19 Differentiate x^x for positive x. (Write $x^x = e^{x\ln x}$.)

5.20 A two-layer network computes a hidden value $h = \sigma(w_1 x)$ from an input x, then an output $a = \sigma(w_2 h)$, where w_1 and w_2 are weights. Use the chain rule to write da/dw_1.

Module 5 Quiz

Check Your Understanding

1. What is the derivative of $e^{-0.05t}$ with respect to t?

A) $e^{-0.05t}$
B) $0.05e^{-0.05t}$
C) $-e^{-0.05t}$
D) $-0.05e^{-0.05t}$

2. By the product rule, what is the derivative of $x^2 e^x$?

A) $2xe^x$
B) $2xe^x + x^2e^x$
C) x^2e^x
D) $2x + e^x$, the sum of the two separate derivatives, since products differentiate term by term

3. Why is backpropagation described as the chain rule applied repeatedly?

A) Because a network is a long chain of nested functions.
B) Because it adjusts each weight in proportion to how often that weight has been used during training.
C) Because the loss is a sum over training examples, and the derivative of a sum is the sum of the derivatives.
D) Because the network's outputs are fed back into its inputs, so each layer's output becomes the next layer's input in reverse.

4. Demand has an elasticity of −0.5 at the current price. A small price increase will:

A) Cut revenue
B) Leave revenue unchanged
C) Raise revenue
D) Raise both quantity and revenue

5. What is the derivative of the sigmoid, $\sigma(z) = 1/(1 + e^{-z})$?

A) $\sigma(z)^2$
B) $\sigma(z)(1 - \sigma(z))$, the output times one minus the output
C) e^{-z}
D) $-1/(1 + e^{-z})^2$, by the power rule alone

Module 6 — Optimization

Finding the best price, size, order, and fit

Recommended Reading / Reference

Book: Avinash Dixit, *Optimization in Economic Theory* (2nd ed., 1990) — a short classic on how economists set up and solve best-choice problems

Reference: Lazard, *Levelized Cost of Energy+* (published each June, lazard.com) — cost benchmarks for generation and storage, the inputs to sizing problems like this module's

6a. Critical Points

Every optimization problem in this course has the same shape: a quantity to make as large or as small as possible, such as profit, cost, or error, that depends on a choice, such as a price, a size, or a parameter. Write the quantity as a function of the choice, and calculus finds the best choice.

At the top of a smooth peak or the bottom of a smooth trough, the tangent line is flat, so the derivative is zero. An input c where $f'(c) = 0$ is called a critical point of the function f. Most optimization comes down to finding critical points and checking which is best, with two cautions. A best value can sit at the edge of what is allowed, such as a price that cannot be negative or a battery that cannot outgrow its site, so the endpoints of the allowed range must be checked too. And a function with a kink, like the ReLU of Module 2, can peak where the derivative does not exist.

Worked example — The revenue-maximizing price

Module 4's store earns weekly revenue $R(p) = 400p - 8p^2$ dollars at price p. Setting $R'(p) = 400 - 16p$ to zero gives $p = 25$. The allowed prices run from 0 to 50, above which nothing sells, and revenue is zero at both ends, so the peak is at $25, where the store sells 200 units and takes in $5,000 a week.

6b. Peaks and Troughs

A critical point can be a peak, a trough, or neither: x^3 is flat at zero but keeps rising. The second derivative tells which. At a critical point c:

$$f''(c) < 0 \;\Rightarrow\; \text{a peak (local maximum)} \qquad f''(c) > 0 \;\Rightarrow\; \text{a trough (local minimum)}$$

A cap-shaped graph, bending downward, has its peak at the critical point; a cup-shaped graph has its trough there. When $f''(c) = 0$ the test says nothing, and the sign of f' on either side decides: rising then falling is a peak, falling then rising a trough.

“Local” matters. A local maximum is the highest point nearby, not necessarily the highest anywhere. When a function has a single critical point and its second derivative keeps one sign across the whole allowed range, as the functions in this module do, the local answer is also the overall one.

Worked example — Marginal revenue equals marginal cost

A firm selling q units earns revenue $R(q) = 60q - 0.02q^2$ and pays costs $C(q) = 5{,}000 + 20q + 0.01q^2$, both in dollars. Its profit, written π by economists (here not the circle constant), is $\pi(q) = R(q) - C(q) = 40q - 0.03q^2 - 5{,}000$. Setting $\pi'(q) = 40 - 0.06q$ to zero gives $q \approx 667$ units, and $\pi''(q) = -0.06$ is negative, so this is a peak. Profit there is about $8,333.

Setting $\pi'(q) = R'(q) - C'(q)$ to zero is the same as setting marginal revenue equal to marginal cost: $R'(q) = C'(q)$. Below that output, the next unit brings in more than it costs; above it, less. Here $R'(q) = 60 - 0.04q$ and $C'(q) = 20 + 0.02q$ meet at $q \approx 667$.

6c. Pricing for Profit

A firm that chooses its price rather than its output faces the same calculation. If each unit costs c dollars to make and buyers take $q(p)$ units at price p, profit is the margin per unit times the units sold:

$$\pi(p) = (p - c)\, q(p)$$

Worked example — A price above the revenue peak

The store's demand is $q = 400 - 8p$ units a week, and each unit costs it $10. Profit is $\pi(p) = (p - 10)(400 - 8p)$. By the product rule, $\pi'(p) = (400 - 8p) - 8(p - 10) = 480 - 16p$, which is zero at $p = 30$. The store sells 160 units and makes $20 \times 160 = 3{,}200$ dollars a week. At the revenue-maximizing price of $25 it would sell 200 units but make only $15 \times 200 = 3{,}000$.

The profit-maximizing price sits above the revenue-maximizing one because the extra sales a lower price wins each cost money to make. Elasticity, from Module 5, gives a rule that holds at any profit-maximizing price: the markup over cost, as a share of the price, equals one over the size of the elasticity ε, written $|\varepsilon|$, the elasticity with its sign dropped.

$$\frac{p - c}{p} = \frac{1}{|\varepsilon|}$$

At $p = 30$, $q = 160$ and $\varepsilon = (30/160)(-8) = -1.5$, and the markup is $(30 - 10)/30 = 2/3$, which is $1/1.5$. Firms whose customers have few alternatives, and so respond weakly to price, charge large markups; firms in close competition charge small ones.

6d. How Long Should a Battery Be?

A battery has two sizes. Its power, in megawatts, sets how fast it can charge and discharge; its energy, in megawatt-hours, sets how long it can run. Energy divided by power is its duration in hours. Longer durations earn more, because the battery can serve more of each day's high-priced hours, but each added hour earns less than the one before: the first hour captures the day's sharpest price spike, and later hours capture flatter ones. Energy capacity, meanwhile, costs about the same for every added hour.

A simple model captures this. With illustrative numbers, suppose each kilowatt of power earns $R(h) = 150(1 - e^{-h/3})$ dollars a year when the battery holds h hours of energy, and energy capacity costs 12 dollars per kilowatt-hour each year. The net value per kilowatt, $N(h)$, is greatest where its derivative is zero:

$$N(h) = 150\left(1 - e^{-h/3}\right) - 12h$$
$$N'(h) = 50\,e^{-h/3} - 12 = 0$$
$$h = 3\ln\frac{50}{12} \approx 4.3 \text{ hours}$$

The second derivative, $N''(h) = -(50/3)e^{-h/3}$, is negative for every h, so 4.3 hours is the best duration in this model. The derivative states the economics: keep adding hours until the last hour's extra revenue, $50e^{-h/3}$ dollars, falls to its cost, $12.

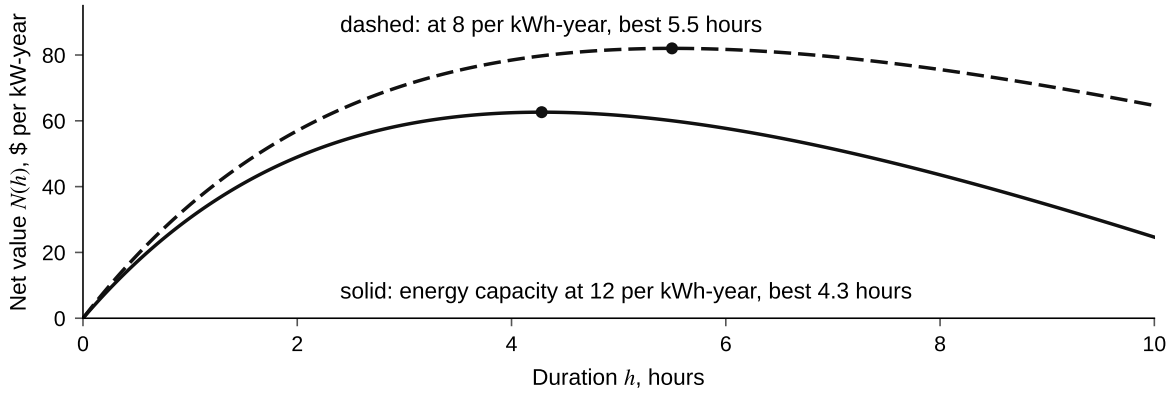


Figure 6.1. Net value per kilowatt against duration, with illustrative numbers. When energy capacity costs $12 per kilowatt-hour each year (solid) the best duration is 4.3 hours; at $8 (dashed) it is 5.5 hours.

Worked example — Cheaper storage, longer batteries

If energy capacity falls to $8 per kilowatt-hour each year, the best duration becomes $3\ln(50/8) \approx 5.5$ hours. Falling costs for energy capacity push the best duration longer: each added hour still earns less than the one before, but it now costs less too.

6e. The Square-Root Inventory Formula

A business that sells D units a year from stock must decide how many to order at a time. Each order has a fixed cost S, for paperwork, shipping, and handling, and each unit held in stock costs H a year, for storage, insurance, and the money tied up. Ordering Q units at a time means D/Q orders a year and, since stock runs down from Q to zero between deliveries, an average of $Q/2$ units on hand. The yearly cost of ordering and holding, $T(Q)$, is

$$T(Q) = \frac{D}{Q}S + \frac{Q}{2}H$$

Setting the derivative, $T'(Q) = -DS/Q^2 + H/2$, to zero gives the economic order quantity:

$$Q^* = \sqrt{\frac{2DS}{H}}$$

Worked example — Stocking solar connectors

A distributor sells 12,000 boxes of solar connectors a year. Each order costs \$150 to place, and each box costs \$4 a year to hold. The best order is $Q^* = \sqrt{(2 \times 12{,}000 \times 150/4)} = \sqrt{900{,}000} \approx 949$ boxes, about 12.6 orders a year, for a yearly cost of about \$3,795. At that quantity the ordering and holding costs are equal, about \$1,897 each, which is true of every economic order quantity.

The square root makes the answer forgiving. Ordering 25% more than Q^* raises the yearly cost by only 2.5%, because near its minimum the cost curve is nearly flat. Errors in the inputs are damped too: if the holding cost is really twice the estimate, the best order falls only by a factor of $\sqrt{2}$, about 29%.

6f. Fitting a Model by Minimizing Squared Error

Fitting a model to data is an optimization problem: choose the model's parameters to make its errors as small as possible. Take n data points, the pairs (x_i, y_i) for i from 1 to n, and a model $y = wx$, a line through the origin with slope w. The model's error at point i is $y_i - wx_i$. Squaring each error makes every term positive and penalizes large misses most. The Greek capital sigma, Σ, means “add up over all the points,” and the total squared error $E(w)$ is

$$E(w) = \sum_{i=1}^{n} (y_i - w\,x_i)^2$$

By the chain rule, $E'(w) = -2\,\Sigma\, x_i(y_i - wx_i)$. Setting it to zero and solving for w gives the least-squares slope:

$$w = \frac{\sum x_i\, y_i}{\sum x_i^2}$$

Worked example — Demand per degree

On four summer afternoons, electricity demand ran y megawatts above its mild-weather level when the temperature was x degrees above 20°C: the points were (2, 130), (4, 230), (6, 370), and (8, 480). Then $\Sigma\, x_i\, y_i = 260 + 920 + 2{,}220 + 3{,}840 = 7{,}240$ and $\Sigma\, x_i^2 = 4 + 16 + 36 + 64 = 120$, so $w = 7{,}240/120 \approx 60.3$ megawatts per degree.

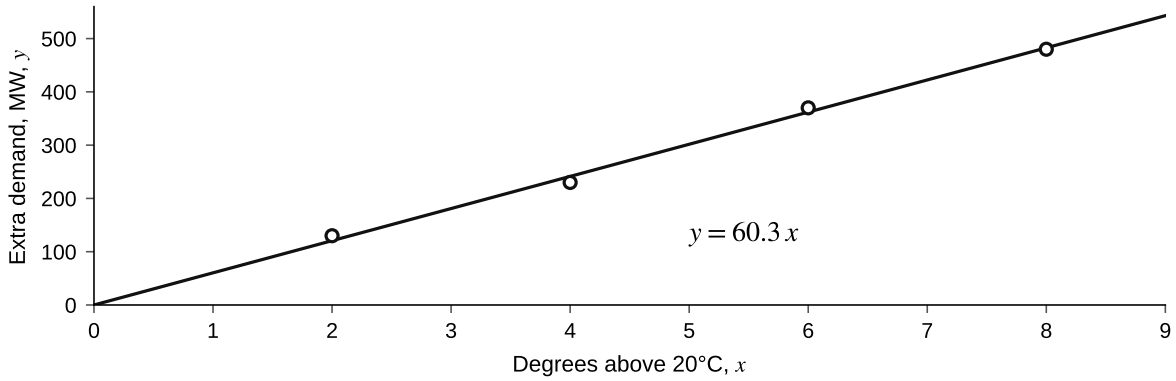


Figure 6.2. The least-squares line through the origin for four afternoons of demand data. The slope, 60.3 megawatts per degree, minimizes the total squared vertical distance from the points to the line.

The second derivative, $E''(w) = 2\Sigma x_i^2$, is positive, so this is a minimum. The same idea, with more parameters, fits every regression line and trains every AI model: choose the parameters that minimize a loss. With two or more parameters, each partial derivative must be zero, which Module 10 handles. With billions of parameters, solving for the zero by algebra is hopeless, and models instead step downhill, as Module 8 shows.

6g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Critical point	An input c where $f'(c) = 0$; check the endpoints of the allowed range and any kinks too f is the quantity being optimized
Second-derivative test	$f''(c) < 0:$ peak $\qquad f''(c) > 0:$ trough At a critical point c; when $f''(c) = 0$, check the sign of f' on either side
Profit-maximizing output	Marginal revenue equals marginal cost: $R'(q) = C'(q)$ R is revenue, C cost, and q output
Markup rule	$\frac{p - c}{p} = \frac{1}{\lvert\varepsilon\rvert}$ p is the price, c the unit cost, and $\lvert\varepsilon\rvert$ the size of the elasticity of demand
Battery duration	Add hours until the last hour's extra revenue falls to its cost Duration is energy capacity divided by power capacity
Economic order quantity	$Q^* = \sqrt{\frac{2DS}{H}}$ D units a year, S dollars per order, H dollars to hold a unit for a year; ordering and holding costs are equal at Q^*
Least squares through the origin	$w = \frac{\sum x_i y_i}{\sum x_i^2}$ Fits $y = wx$ to the points (x_i, y_i); Σ means add up over all the points
Recipe	Write the quantity as a function of the choice, state the allowed range, set the derivative to zero, check the second derivative and the endpoints, and interpret the answer
The idea	The best value of a smooth function sits where its derivative is zero or at an edge of what is allowed, and the second derivative tells a peak from a trough.

6h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

6.1 Find the critical point of $f(x) = x^2 - 10x + 3$ and say whether it is a peak or a trough.

6.2 Find and classify the critical points of $f(x) = x^3 - 12x$.

6.3 Find the largest value of $f(x) = -2x^2 + 8x + 1$.

6.4 Find the smallest value of $g(x) = x + 9/x$ for positive x.

6.5 Find the largest and smallest values of $f(x) = x^2 - 4x$ on the interval from $x = 0$ to $x = 5$.

Applied

6.6 Demand is $q = 1{,}000 - 20p$ units at price p, and each unit costs $15 to make. Find the profit-maximizing price, the quantity sold, and the profit.

6.7 A firm's revenue is $R(q) = 80q - 0.04q^2$ and its cost $C(q) = 4{,}000 + 20q + 0.01q^2$, in dollars, for q units. Find the output where marginal revenue equals marginal cost, and the profit there.

6.8 In the battery model of section 6d, with revenue $R(h) = 150(1 - e^{-h/3})$ dollars per kilowatt each year for h hours of duration, find the best duration if energy capacity costs $18 per kilowatt-hour each year.

6.9 A distributor sells 4,800 units a year. Each order costs $60 to place, and holding a unit costs $2.50 a year. Find the economic order quantity and the number of orders a year.

6.10 Fit the line $y = wx$ through the origin to the points (1, 2.1), (2, 3.9), and (3, 6.2).

6.11 A rectangular storage yard is built against a building, so fencing is needed on only three sides. There are 600 meters of fence. With x the length of each side running out from the building, what dimensions give the largest area?

Challenge

6.12 Show that at the economic order quantity the yearly ordering cost equals the yearly holding cost.

6.13 With demand $q = 400 - 8p$ and a unit cost of c dollars, show that the profit-maximizing price is $p = 25 + c/2$. What happens to the price when the unit cost rises by $2?

6.14 Derive the least-squares slope by differentiating $E(w) = \Sigma(y_i - wx_i)^2$, and show that the second derivative is positive.

6.15 In the battery model, what is the best duration if energy capacity costs only $6 per kilowatt-hour each year?

6.16 For the solar-connector distributor ($D = 12{,}000$, $S = 150$, $H = 4$), compute the yearly cost at Q^* and at 1.25 times Q^*. Why is the difference so small?

Module 6 Quiz

Check Your Understanding

1. At a critical point of f, what is true of f'?

A) It is zero
B) It is at its largest
C) It equals the second derivative
D) It equals the function's value

2. If $f'(c) = 0$ and $f''(c) < 0$, then at c the function has:

A) A minimum
B) A zero
C) An inflection point, where the bend reverses
D) A local maximum

3. Why does a firm with production costs set a higher price than the one that maximizes revenue?

A) Because higher prices always raise profit.
B) Because demand becomes more elastic as costs rise, so the firm must raise prices to keep revenue constant.
C) Because the extra sales a lower price wins each cost money to make.
D) Because the firm wants to sell as many units as possible, and a higher price attracts more buyers who value quality.

4. At the economic order quantity, how do the yearly ordering and holding costs compare?

A) Holding costs are twice the ordering costs.
B) They are equal.
C) Ordering costs are zero, because orders are placed only when stock runs out.
D) Holding costs are half the ordering costs, because average inventory is half the order size.

5. To fit $y = wx$ to data by least squares, what is w chosen to minimize?

A) The sum of the errors
B) The number of points off the line
C) The largest single error, so no point is badly fitted
D) The sum of the squared differences between each y_i and wx_i

Module 7 — Integrals as Accumulation

Adding up a rate over time: energy, money, and value

Recommended Reading / Reference

Book: Steven Strogatz, *Infinite Powers* (2019) — a lively history of calculus, at its best on how adding up slices became integration

Reference: U.S. Energy Information Administration, Hourly Electric Grid Monitor (eia.gov/electricity/gridmonitor) — hourly demand and generation by region, the kind of data this module integrates

7a. Accumulation and Area

Module 4 went from an amount to its rate, from the energy a solar plant had produced to its power. Integration goes the other way, from a rate to an amount. When a rate is constant, the amount is the rate times the time: 100 MW for 4 hours is 400 MWh. When the rate varies, cut the time into short slices, treat the rate as constant within each slice, and add up rate times time over all of them.

The solar plant's power readings, taken each hour from 6 a.m. to 6 p.m., were 0, 25.9, 50.0, 70.7, 86.6, 96.6, 100.0, 96.6, 86.6, 70.7, 50.0, 25.9, and 0 MW. Treating each hour's power as its starting reading and adding up gives 759.6 MWh for the day. On a graph, each hour is a thin rectangle under the power curve, and the energy is the area under the curve.

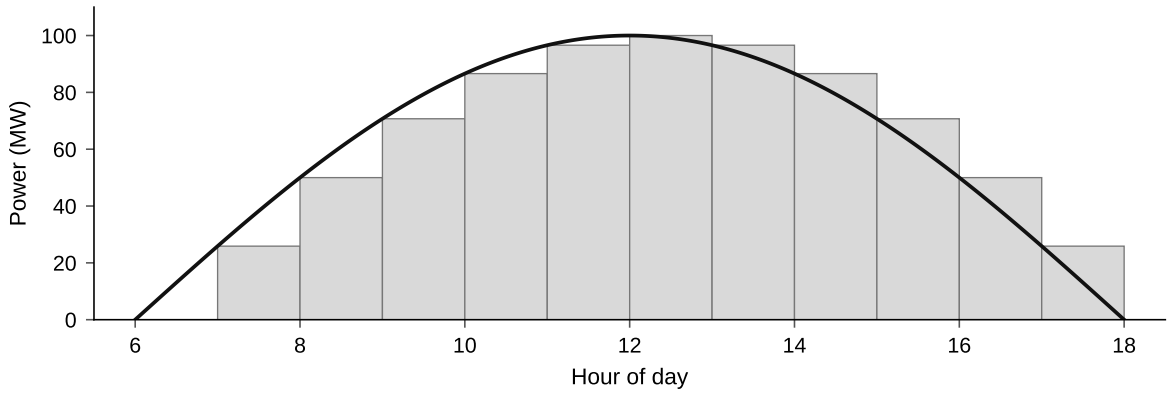


Figure 7.1. Hourly rectangles under the solar plant's power curve. Each rectangle is one hour's power times one hour, an estimate of that hour's energy. Together they give 759.6 MWh; the area under the curve is 763.9 MWh.

In the limit of ever-thinner slices, the sum becomes the definite integral. To accumulate a rate $f(x)$ from $x = a$ to $x = b$, cut the interval into n slices of width Δx, take a point x_i in each slice i, and add up:

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\,\Delta x$$

The integral sign is a stretched S, for sum, and dx recalls the slices' vanishing width. The numbers a and b are the limits of integration. Area below the horizontal axis counts as negative, so an integral measures net accumulation: energy put into a battery and energy taken out offset each other.

7b. The Fundamental Theorem of Calculus

Computing an integral as a limit of sums is laborious. The fundamental theorem of calculus, found independently by Isaac Newton in the 1660s and Gottfried Leibniz in the 1670s, replaces it with a subtraction. If F is any function whose derivative is f, called an antiderivative of f, then

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

The reason is the link between energy and power from Module 4. Power is the derivative of cumulative energy, so adding up power over an interval must give the change in cumulative energy across it. Accumulating a rate gives the change in whatever the rate is the rate of.

Antiderivatives come from reading the derivative rules backward. Each is unique only up to an added constant, written C, since a constant's derivative is zero; in a definite integral the constant cancels.

Function	Antiderivative	Condition
x^n	$x^{n+1}/(n+1) + C$	Any power n except −1
$1/x$	$\ln x + C$	For positive x
e^{kx}	$e^{kx}/k + C$	Any constant k except 0
$k \cdot f(x) + g(x)$	$k \cdot F(x) + G(x) + C$	F and G are antiderivatives of f and g

Each row reverses a derivative rule from Module 5. C is any constant.

Worked example — A tapering discharge

A battery's output falls steadily as it empties: $P(t) = 100 - 10t$ megawatts at t hours, for four hours. An antiderivative is $100t - 5t^2$, so the energy delivered is

$$\int_0^4 (100 - 10t)\,dt = \Big[\,100t - 5t^2\,\Big]_0^4 = 400 - 80 = 320 \text{ MWh}$$

The bracket notation $[F(t)]_0^4$ means $F(4) - F(0)$. Reading the chain rule backward gives substitution, the integration counterpart of the chain rule: an integral of the form $\int f(g(x))\,g'(x)\,dx$ equals $F(g(x)) + C$. For example, $\int 2x(1 + x^2)^3\,dx = (1 + x^2)^4/4 + C$, since differentiating the answer by the chain rule returns the function inside the integral.

7c. Integrating Real Data

Meters and markets report numbers, not formulas, so most integrals in practice are computed from data. The trapezoid rule joins neighboring readings with straight lines and adds up the areas of the trapezoids beneath them. For readings $f_0, f_1, \ldots, f_n$ spaced h apart:

$$\int_a^b f(x)\,dx \approx h\left(\tfrac{1}{2}f_0 + f_1 + f_2 + \cdots + f_{n-1} + \tfrac{1}{2}f_n\right)$$

For the solar plant's hourly readings the rule also gives 759.6 MWh, 0.6% below the true 763.9 MWh. The rule runs low on any curve that bends downward, because each straight chord passes beneath the curve. Readings every 15 minutes would cut the error about sixteenfold, since the error shrinks with the square of the spacing.

Revenue is an integral too. A battery discharging at power $P(t)$ into a market paying a price of $\lambda(t)$ dollars per megawatt-hour, where λ is the Greek letter lambda, earns $\int P(t)\,\lambda(t)\,dt$ dollars. With hourly prices, the integral is a sum of each hour's energy times that hour's price.

Worked example — An evening's arbitrage

A battery discharges 100 MW for four hours in the evening peak, when prices are \$85, \$120, \$140, and \$95 per megawatt-hour. It earns $100 \times (85 + 120 + 140 + 95) = 44{,}000$ dollars. It charged with 470 MWh overnight at \$25, a cost of \$11,750, so the day's margin is \$32,250.

7d. The Present Value of a Stream

Module 3's discount factor values a single payment: at a continuous rate r, a dollar due in t years is worth e^{-rt} today. A project that pays out continuously, at a rate of $c(t)$ dollars a year at time t, is a stream of tiny payments, and its present value PV is their sum over its life of T years:

$$PV = \int_0^T c(t)\, e^{-rt}\, dt$$

For a constant stream c, the antiderivative of e^{-rt} is $-e^{-rt}/r$, so

$$PV = c\,\frac{1 - e^{-rT}}{r} \longrightarrow \frac{c}{r} \quad \text{as } T \to \infty$$

Worked example — A storage project's cash flow

A battery project nets \$2 million a year for 15 years. At a continuous discount rate of 8%, it is worth $2 \times (1 - e^{-1.2})/0.08 \approx 17.5$ million dollars today, against \$30 million of cash undiscounted. The last five years' \$10 million are worth only about \$3.7 million today.

A stream that never ends, a perpetuity, is worth c/r: \$1 a year forever at 5% is worth \$20. If the stream grows at a rate g below r, so that $c(t) = c \cdot e^{gt}$, its value is

$$\int_0^\infty c\, e^{gt}\, e^{-rt}\, dt = \frac{c}{r - g}$$

This growing-perpetuity formula, in its yearly form named for the economist Myron Gordon, is how analysts value a stock from its dividends. Its sensitivity is the lesson. A stock paying \$3 a year, with dividends growing 4% a year, is worth $3/(0.09 - 0.04) = 60$ dollars at a 9% discount rate and \$75 at 8%: a quarter more for one point of discount rate, because the gap $r - g$ is what matters.

7e. Consumer Surplus

Integrals also measure value that never shows up as cash. A demand curve, read sideways, gives the most a buyer would pay for each unit. For the store's demand, $q = 400 - 8p$, solving for price gives $p = 50 - q/8$: the first unit is worth nearly \$50 to someone, and the 200th is worth \$25. At a market price p^*, every buyer who values a unit above p^* gains the difference. Their total gain, the consumer surplus CS, is the area between the demand curve $D(q)$, the price a buyer would pay for the qth unit, and the price line, out to the quantity sold, q^*:

$$CS = \int_0^{q^*} \left(D(q) - p^*\right) dq$$

Worked example — The store's customers

At a price of \$25 the store sells 200 units, and $CS = \int_0^{200} (50 - q/8 - 25)\, dq = [25q - q^2/16]_0^{200} = 5{,}000 - 2{,}500 = 2{,}500$ dollars a week: the store's customers value what they buy at \$2,500 more than they pay for it.

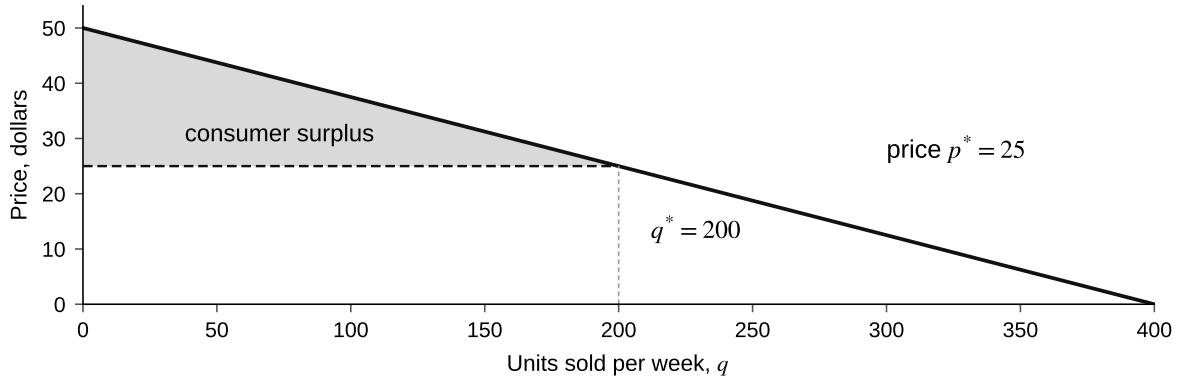


Figure 7.2. Consumer surplus is the shaded area between the demand curve and the price line, out to the 200 units sold: \$2,500 a week.

Producer surplus is the matching area between the price line and the supply curve. Together they measure what a market's trades are worth to the people making them, which is how economists value the gains from trade and the losses from taxes and price caps.

7f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Definite integral	$\int_a^b f(x)\,dx$ The net accumulation of the rate f from a to b: the area under its graph, counted negative below the axis
Fundamental theorem	$\int_a^b f(x)\,dx = F(b) - F(a)$ F is any antiderivative of f: a function with $F' = f$
Antiderivatives	$\int x^n dx = \frac{x^{n+1}}{n+1} + C \qquad \int \frac{dx}{x} = \ln x + C \qquad \int e^{kx} dx = \frac{e^{kx}}{k} + C$ n is not −1, x is positive in the logarithm, k is not 0, and C is any constant
Substitution	$\int f(g(x))\,g'(x)\,dx = F(g(x)) + C$ The chain rule read backward; F is an antiderivative of f
Energy and revenue	Energy $= \int P(t)\,dt$; revenue $= \int P(t)\,\lambda(t)\,dt$ P is power in MW, t time in hours, and λ the price in dollars per MWh
Trapezoid rule	$h\left(\frac{1}{2}f_0 + f_1 + \cdots + f_{n-1} + \frac{1}{2}f_n\right)$ For readings f_0 to f_n spaced h apart; the error shrinks with the square of h
Present value of a stream	$PV = \int_0^T c(t)\,e^{-rt}\,dt \qquad \text{constant } c: \ \frac{c\,(1 - e^{-rT})}{r}$ $c(t)$ dollars a year for T years at a continuous rate r
Perpetuities	level: $\frac{c}{r} \qquad$ growing: $\frac{c}{r - g}$ g is the growth rate, which must be below r
Consumer surplus	$\int_0^{q^*} (D(q) - p^*)\,dq$ $D(q)$ is what a buyer would pay for the qth unit; p^* is the price and q^* the quantity sold
The idea	An integral adds up a rate over an interval, and the fundamental theorem says the total is the change in any function whose derivative is that rate.

7g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

7.1 Evaluate $\int_0^3 (2x + 1)\,dx$.

7.2 Evaluate $\int_1^4 \sqrt{x}\,dx$.

7.3 Evaluate $\int_0^1 e^{2x}\,dx$.

7.4 Evaluate $\int_1^e (1/x)\,dx$.

7.5 Find an antiderivative of $6x^2(1 + x^3)^4$.

7.6 Evaluate $\int_0^2 x\, e^{x^2}\, dx$.

Applied

7.7 A battery discharges at $P(t) = 120 - 15t$ megawatts for t from 0 to 4 hours. How much energy does it deliver?

7.8 A plant's power readings are 0, 40, 75, 90, 80, 50, and 0 MW at hours 0 through 6. Estimate the energy it delivered with the trapezoid rule.

7.9 A battery discharges 50 MW for four hours at prices of \$70, \$110, \$150, and \$90 per megawatt-hour, after charging with 230 MWh overnight at \$25. What does the day's arbitrage earn?

7.10 Find the present value of \$1 million a year received continuously for 20 years, at a continuous rate of 6%.

7.11 A stock pays a dividend of \$2.40 a year that grows 3% a year. What is it worth under the growing-perpetuity formula at a discount rate of 8%? At 7%?

7.12 Buyers would pay $D(q) = 80 - 0.5q$ dollars for the qth unit. At a price of \$30, what is the consumer surplus?

Challenge

7.13 Show that the present value of a constant stream c for T years at a continuous rate r is $c(1 - e^{-rT})/r$, and that it approaches c/r as T grows.

7.14 Show that $\int_0^\infty c\, e^{gt}\, e^{-rt}\, dt = c/(r - g)$ when g is below r. What happens when g is at least r?

7.15 Integration by parts, $\int u\, v'\, dt = uv - \int u'\, v\, dt$, reverses the product rule. Use it with $u = t$ and $v = -e^{-rt}/r$ to find $\int_0^\infty t\, e^{-rt}\, dt$, and divide by $\int_0^\infty e^{-rt}\, dt = 1/r$ to find the average time, weighted by present value, at which a perpetuity's cash arrives. What is it at 5%?

7.16 The trapezoid rule underestimated the solar plant's energy. Explain why it underestimates the area under any curve that bends downward, and why it would overestimate the area under one that bends upward.

Module 7 Quiz

Check Your Understanding

1. A battery discharges at $P(t) = 100 - 10t$ megawatts for four hours. How much energy does it deliver?

A) 60 MW

B) 80 MW

C) 320 MWh

D) 400 MWh, since it starts at 100 MW and runs four hours

2. What does the fundamental theorem of calculus say?

A) That integrating a rate gives the change in any function whose rate it is.

B) That every function has a derivative.

C) That the area under a curve equals the curve's highest value times the width of the interval.

D) That derivatives and integrals both measure slopes, one at a point and the other averaged over an interval.

3. Why is the present value of a perpetuity finite even though its payments never end?

A) Because payments stop after a century.

B) Because discounting shrinks distant payments so fast that their sum converges, to c/r.

C) Because inflation eventually makes the payments worthless.

D) Because the payments themselves fall over time, so the total of an endless stream is limited by the first payment.

4. Adding up hourly readings of a power plant's output with the trapezoid rule estimates its:

A) Energy

B) Peak power

C) Average price

D) Round-trip efficiency

5. What does consumer surplus measure?

A) The seller's profit.

B) The amount by which demand exceeds supply at the current price.

C) The total amount buyers spend, which is the area under the demand curve up to the quantity sold.

D) The value buyers receive beyond what they pay.

Module 8 — Differential Equations

Quantities whose change depends on their size: growth, cooling, charge, and adoption

Recommended Reading / Reference

Book: Steven Strogatz, *Nonlinear Dynamics and Chaos* (2nd ed., 2015) — the most readable serious text on differential equations, rich in applications

Reference: International Energy Agency, *Global EV Outlook* (published each year, iea.org) — electric-car sales shares by country, adoption curves in real data

8a. What a Differential Equation Says

Modules 4 to 7 started from a function and studied its rate of change. Often the order is reversed: what is known is a rule for the rate, and the function must be found from it. A differential equation states such a rule, usually giving the rate of a quantity in terms of its current value. For a quantity y that changes over time t:

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

where f says how fast y is changing when it has the value y at time t. A solution is a whole function $y(t)$ that obeys the rule at every moment. It is pinned down by a starting value, the initial condition $y(0)$, which this module also writes y_0.

Differential equations are how science and engineering state cause and effect. A balance earning interest grows faster the larger it is; a hot object cools faster the hotter it is; a technology spreads faster the more people already use it, until few are left to convert. Each statement is a rule for a rate, and each solution is a path through time.

8b. Growth and Decay

The simplest differential equation says the rate is proportional to the quantity, with a constant k:

$$\frac{dy}{dt} = k\,y \qquad \Longrightarrow \qquad y(t) = y_0\,e^{kt}$$

Module 5 showed that e^{kt} grows at k times its own size, so $y_0\,e^{kt}$ obeys the rule, and at $t = 0$ it equals y_0. When k is positive the solution is the continuous compounding of Module 3; when k is negative it is decay, with a half-life of $\ln 2/|k|$, where $|k|$ is k without its sign. Any quantity that grows or shrinks at a steady percentage rate obeys this equation: a deposit, a population, a radioactive sample, the computing power behind AI.

8c. Cooling, and Settling Toward Equilibrium

A hot object cools at a rate proportional to how much hotter it is than its surroundings. If T is its temperature, T_{env} the temperature around it, and k a constant set by how readily it sheds heat, Newton's law of cooling says

$$\frac{dT}{dt} = -k\,(T - T_{\text{env}}) \qquad \Longrightarrow \qquad T(t) = T_{\text{env}} + (T_0 - T_{\text{env}})\,e^{-kt}$$

where T_0 is the starting temperature. The gap to the surroundings decays exponentially, and $1/k$, the time constant, is the time for the gap to shrink to about 37% of its starting size, a factor of e. The same equation describes anything that settles toward an equilibrium at a rate proportional to its distance from it: a building's indoor temperature after the heat fails, a price adjusting toward its long-run level, a battery cell after a hard discharge.

Worked example — A battery module after a hard discharge

A battery module is at 45°C when a discharge ends, in a 25°C room, and it sheds heat with $k = 0.5$ per hour, a time constant of two hours. Its temperature is $T(t) = 25 + 20e^{-0.5t}$ degrees at t hours. It reaches 30°C when $20e^{-0.5t} = 5$, that is, when $t = 2\ln 4 \approx 2.8$ hours.

8d. A Battery's State of Charge

A battery's stored energy E, in megawatt-hours, changes at the rate power flows in or out, and losses make the bookkeeping lopsided. Charging at a power P_c stores only a share η_c of it, the charging efficiency, where η is the Greek letter eta. Discharging at a power P_d draws more than it delivers, the delivered power divided by the discharging efficiency η_d. And even idle, a battery loses a small share λ of its stored energy each hour, to self-discharge and to the systems that cool and control it:

$$\frac{dE}{dt} = \begin{cases} \eta_c\,P_c, & \text{charging} \\ -P_d/\eta_d, & \text{discharging} \\ -\lambda\,E, & \text{idle} \end{cases}$$

The first two lines integrate directly, as in Module 7: charging at a steady 50 MW for two hours with $\eta_c = 0.95$ stores $0.95 \times 100 = 95$ MWh. The product $\eta_c\,\eta_d$ is the round-trip efficiency; with 95% each way it is $0.95 \times 0.95 \approx 0.90$, so a tenth of the energy bought is lost. The idle line is the decay equation of section 8b, with the solution $E(t) = E_0\,e^{-\lambda t}$, where E_0 is the energy stored when the battery goes idle.

Worked example — A full battery left overnight

A 400 MWh battery sits full for 12 hours, and with illustrative numbers its idle losses take 0.2% of its stored energy each hour, so $\lambda = 0.002$. By morning it holds $400e^{-0.024} \approx 390.5$ MWh. The 9.5 MWh lost is a cost of holding charge, and one reason operators time charging close to the hours they expect to discharge.

8e. The S-Curve of Adoption

New technologies spread slowly, then quickly, then slowly again, and the logistic equation captures that shape. If N is the number of users, K the most there can be, and r a growth constant, then

$$\frac{dN}{dt} = r\,N\left(1 - \frac{N}{K}\right)$$

When N is small, the bracket is close to 1 and growth is exponential: each user recruits others. As N approaches K, the bracket shrinks toward zero and growth stalls, because few potential users remain. The solution is the S-curve, centered on the time t_0 at which N reaches half of K:

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

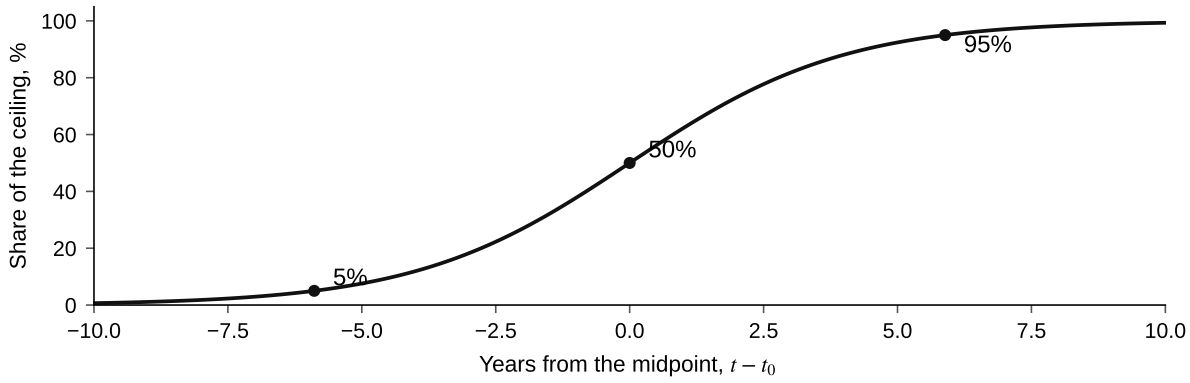


Figure 8.1. The logistic S-curve with $r = 0.5$ a year. Growth is fastest at the midpoint, and the climb from 5% to 50% takes as long as the climb from 50% to 95%: about 5.9 years each.

Growth is fastest at the midpoint, when $N = K/2$, and the curve is symmetric about it: the climb from 5% of K to 50% takes as long as the climb from 50% to 95%, $\ln 19 / r$ in each case.

Worked example — Electric cars in Norway

Battery-electric cars took 82.4% of Norway's new-car sales in 2023 and 88.9% in 2024, according to the Norwegian Road Federation. For a share s following a logistic curve whose ceiling is 100%, the quantity $\ln(s/(1 - s))$ rises in a straight line at the rate r. It went from $\ln(0.824/0.176) \approx 1.54$ to $\ln(0.889/0.111) \approx 2.08$ in a year, so $r \approx 0.54$ a year. Running the line back, the curve's midpoint, 50%, sits about $1.54/0.54 \approx 2.9$ years before 2023, around 2020.

The logistic curve is a model of a pattern, not a law. Real adoption depends on prices, policy, and what the alternatives cost, and the ceiling K is often the very thing in doubt: whether electric cars, heat pumps, or rooftop solar saturate at all of a market or well short of it is what forecasters argue about. Fitting an S-curve to the early, exponential stretch of data is notoriously unreliable, because the bend has not happened yet.

8f. Euler's Method, and How Machines Learn

Most differential equations have no formula for their solution. Euler's method, published by Leonhard Euler in the 1760s, solves them numerically in small steps. From the current value y_n at time t_n, step forward by a short time h, assuming the rate holds steady over the step:

$$y_{n+1} = y_n + h\, f(t_n,\, y_n)$$

Smaller steps give a more accurate path, at the cost of more steps. Yearly compounding is Euler's method applied to continuous growth: for $dy/dt = 0.05y$ with yearly steps, each step multiplies by 1.05, so \$100 grows to $100 \times 1.05^{10} \approx 162.89$ dollars in ten years, against the exact $100e^{0.5} \approx 164.87$.

Euler's method is also how neural networks learn. Picture a weight w sliding downhill on its loss $L(w)$ at a speed equal to the slope, $dw/dt = -L'(w)$: it moves against the derivative until it settles at a minimum. Euler's method with a step size α, called the learning rate, turns that slide into gradient descent, the rule used to train modern AI models:

$$w_{n+1} = w_n - \alpha\, L'(w_n)$$

Module 5's learning step, which moved a weight against its derivative by half the derivative's size, was one such step with $\alpha = 0.5$. The step size matters, and the stability of Euler's method explains why.

Worked example — A learning rate too large

For the loss $L(w) = (w - 3)^2$, whose minimum is at $w = 3$, the derivative is $L'(w) = 2(w - 3)$, so each step multiplies the distance from 3 by $1 - 2\alpha$. Starting from $w = 0$:

Learning rate	Step 1	Step 2	Step 3	Step 4	Step 5	Behavior
$\alpha = 0.1$	0.60	1.08	1.46	1.77	2.02	Converges slowly
$\alpha = 0.4$	2.40	2.88	2.98	3.00	3.00	Converges fast
$\alpha = 0.9$	5.40	1.08	4.54	1.77	3.98	Overshoots, converges
$\alpha = 1.1$	6.60	−1.32	8.18	−3.22	10.47	Diverges

Gradient descent on $L(w) = (w - 3)^2$ from $w = 0$. The distance to the minimum shrinks whenever $1 - 2\alpha$ lies between −1 and 1, that is, for α between 0 and 1.

Too small a learning rate wastes steps; too large a one overshoots the minimum and can diverge, the instability every Euler method suffers when its steps are too long for the dynamics. Practitioners tune learning rates for exactly this reason, and often shrink them as training proceeds.

8g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Differential equation	A rule for the rate of change of a quantity in terms of its current value: $dy/dt = f(t, y)$ The solution $y(t)$ is fixed by its starting value, $y_0 = y(0)$
Growth and decay	$\frac{dy}{dt} = ky \quad \Longrightarrow \quad y = y_0 e^{kt}$ k is the continuous rate: positive for growth, negative for decay
Settling toward equilibrium	$\frac{dT}{dt} = -k\,(T - T_{\text{env}}) \implies T = T_{\text{env}} + (T_0 - T_{\text{env}})\,e^{-kt}$ T_{env} is the surroundings and T_0 the start; after one time constant, $1/k$, the gap is down to 37%
Battery state of charge	Charging: $dE/dt = \eta_c P_c$; discharging: $dE/dt = -P_d/\eta_d$; idle: $dE/dt = -\lambda E$ η_c and η_d are the efficiencies, P_c and P_d the powers, and λ the idle loss rate; the round trip is $\eta_c \eta_d$
Logistic growth	$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \quad \Longrightarrow \quad N = \frac{K}{1 + e^{-r(t - t_0)}}$ K is the ceiling and t_0 the midpoint; growth is fastest at $K/2$, and 5% to 50% takes $\ln 19/r$
Euler's method	$y_{n+1} = y_n + h\,f(t_n, y_n)$ h is the step; smaller steps are more accurate
Gradient descent	$w_{n+1} = w_n - \alpha\,L'(w_n)$ L is the loss, w a weight, and α the learning rate; too large a rate overshoots and can diverge
The idea	A differential equation says how a quantity changes given where it is now; solving it, by formula or by many small Euler steps, gives its whole path through time.

8h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

8.1 Verify that $y = 5e^{3t}$ solves $dy/dt = 3y$ with $y(0) = 5$.

8.2 Solve $dy/dt = -0.2y$ with $y(0) = 50$, and find $y(5)$.

8.3 A quantity obeying $dy/dt = ky$ doubles every 10 years. Find k.

8.4 Solve $dT/dt = -0.5(T - 20)$ with $T(0) = 80$, and find $T(2)$.

8.5 In the logistic equation with $K = 1{,}000$ and $r = 0.4$, at what N is growth fastest, and what is dN/dt there?

8.6 Take two Euler steps of size $h = 0.5$ for $dy/dt = y$ with $y(0) = 1$, and compare the result with the exact $y(1) = e$.

Applied

8.7 A battery module at 45°C cools toward a 25°C room with $k = 0.5$ per hour. How long until it reaches 35°C?

8.8 A battery charges at 50 MW for 2 hours with a charging efficiency of 95%, then discharges at 50 MW with a discharging efficiency of 95%. How much energy does it store, how long can it discharge, and how much does it deliver?

8.9 A full 400 MWh battery sits idle for 12 hours, losing 0.2% of its stored energy each hour. How much energy remains?

8.10 An adoption share follows $s(t) = 1/(1 + e^{-0.5(t - t_0)})$, with t in years. How long does it take to go from 10% to 90%?

8.11 Run gradient descent on $L(w) = (w - 3)^2$ from $w = 0$ with a learning rate of 0.25 for three steps.

8.12 A deposit grows continuously at 6% a year while \$1,000 a year is added continuously, so $dA/dt = 0.06A + 1{,}000$, with $A(0) = 0$. Verify that $A(t) = (1{,}000/0.06)(e^{0.06t} - 1)$ solves it, and find $A(20)$.

Challenge

8.13 Solve $dT/dt = -k(T - T_{\text{env}})$ by writing $u = T - T_{\text{env}}$, where the surroundings' temperature T_{env} is constant.

8.14 Show that $N(t) = K/(1 + e^{-r(t - t_0)})$ satisfies $dN/dt = rN(1 - N/K)$.

8.15 For gradient descent on $L(w) = (w - 3)^2$, show that each step multiplies the error $w - 3$ by $1 - 2\alpha$, and find the learning rates α for which the method converges.

8.16 Show that n Euler steps of size $h = t/n$ for $dy/dt = ry$ give $y_0(1 + rt/n)^n$, and say what happens as n grows.

Module 8 Quiz

Check Your Understanding

1. What function solves $dy/dt = ky$ with $y(0) = y_0$?

A) $y = ky$
B) $y = y_0 + kt$
C) $y = y_0\, e^{kt}$
D) $y = e^{k}$

2. In the logistic model $dN/dt = rN(1 - N/K)$, when is growth fastest?

A) At the start
B) When $N = K$
C) When N is small, since growth then compounds fastest
D) When N is half of K, midway up the S-curve, where the curve is steepest

3. What is Euler's method?

A) It solves equations by factoring.
B) It finds the exact solution of any differential equation.
C) It steps a solution forward by adding the step size times the current rate of change.
D) It finds where a function is largest by setting its derivative to zero and checking the second derivative.

4. What happens if gradient descent on $L(w) = (w - 3)^2$ uses a learning rate α above 1?

A) The steps converge faster than with a small rate.
B) The steps diverge.
C) The steps stop exactly at the minimum after one move.
D) Nothing changes, since the learning rate only affects how long each step takes to compute.

5. A battery module at 45°C cools in a 25°C room. Under Newton's law of cooling, its temperature:

A) Approaches 25°C exponentially
B) Falls linearly
C) Overshoots below 25°C and then rebounds
D) Stays at 45°C until the room warms, since heat flows only from cold to hot

Module 9 — Approximation

Tangent lines, Taylor polynomials, Newton's method, and bond risk

Recommended Reading / Reference

Book: Frank J. Fabozzi, *Bond Markets, Analysis, and Strategies* (10th ed., 2021) — the standard text on bond pricing, duration, and convexity

Reference: U.S. Department of the Treasury, Daily Treasury Par Yield Curve Rates (home.treasury.gov) — the yields that price bonds like this module's examples

9a. Linear Approximation

Near any point where it is smooth, a function looks like its tangent line. That is the most used fact in applied mathematics, because lines are easy. For a function f near an input a, a small step Δx changes the output by about the slope times the step:

$$f(a + \Delta x) \approx f(a) + f'(a)\,\Delta x$$

Module 5's bond example was this approximation at work: the derivative of price with respect to yield, times a small change in yield, estimated the change in price. So is every sensitivity in finance and engineering quoted as so much per unit.

Worked example — A square root without a calculator

To estimate $\sqrt{101}$, use $f(x) = \sqrt{x}$ near $a = 100$, where $f(100) = 10$ and $f'(100) = 1/(2\sqrt{100}) = 0.05$. With $\Delta x = 1$, $\sqrt{101} \approx 10 + 0.05 = 10.05$. The true value is 10.0499: the tangent line is excellent this close.

The approximation $\ln(1 + x) \approx x$ for a small number x, which follows because $\ln 1 = 0$ and the slope of $\ln$ at 1 is 1, explains two facts from earlier modules. It is why a small percentage change and the change in a logarithm are nearly the same number, and why the doubling time at a growth rate r, $\ln 2/\ln(1 + r)$, is close to $0.693/r$.

9b. Taylor Polynomials

A tangent line matches a function's value and slope at a point. A Taylor polynomial, named for the English mathematician Brook Taylor, also matches its curvature and higher derivatives, and so tracks the function farther away. The quadratic version adds a term with the second derivative:

$$f(a + \Delta x) \approx f(a) + f'(a)\,\Delta x + \tfrac{1}{2} f''(a)\,(\Delta x)^2$$

The full pattern continues with the third and higher derivatives, each divided by a factorial: $k!$, read "k factorial," is the product $1 \times 2 \times \ldots \times k$, so $3! = 6$. Near zero, the most useful functions have simple Taylor polynomials, written here with x as the step from zero and n any number:

$$e^x \approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$$

$$\ln(1 + x) \approx x - \frac{x^2}{2} + \frac{x^3}{3}$$

$$(1 + x)^n \approx 1 + n\,x + \frac{n(n-1)}{2}\,x^2$$

Worked example — Four terms of e

With $x = 0.1$, the series gives $1 + 0.1 + 0.005 + 0.000167 \approx 1.105167$, against the true $e^{0.1} = 1.105171$: four terms, six correct digits. The first term left out, $x^4/24 \approx 0.000004$, is a good estimate of the error, which is the usual rule for these series when the step is small.

Taylor's idea also explains why gradient descent works. One step moves a weight w by $-\alpha L'(w)$, where L is the loss and α the learning rate, and the linear approximation says the loss changes by about the slope times the step: $L'(w) \times (-\alpha L'(w)) = -\alpha L'(w)^2$. That is a decrease whenever the slope is not zero and α is small enough for the approximation to hold.

9c. Newton's Method

Many equations can only be solved numerically: the yield that prices a bond, the discount rate at which a project breaks even. Isaac Newton's method solves an equation $g(x) = 0$ by repeatedly replacing the curve with its tangent line and jumping to where the line crosses zero. From a guess x_n, the next guess is

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$$

Near a solution, Newton's method roughly doubles the number of correct digits at every step. To find $\sqrt{2}$, solve $g(x) = x^2 - 2 = 0$, with $g'(x) = 2x$. From $x_0 = 1.5$, the steps give 1.41667, then 1.414216, then 1.4142136, correct to every digit shown.

Worked example — The internal rate of return

A project costs $100 today and pays $30 at the end of each of the next five years. Its internal rate of return, or IRR, is the discount rate r at which the payments are worth exactly the cost: the solution of $g(r) = -100 + \Sigma\, 30/(1 + r)^t = 0$, with the sum over the years t from 1 to 5. At $r = 10\%$, g is 13.72 and its derivative is −290.5, so the first Newton step is $0.10 + 13.72/290.5 \approx 14.72\%$. The next steps give 15.233% and 15.238%: the project returns about 15.2% a year.

Spreadsheet functions that compute a bond's yield or a project's IRR run Newton's method or a close relative behind the scenes.

9d. Duration and Convexity

A bond is a stream of fixed payments, c_t dollars at t years from now, and its price P at a yield y is their present value, each payment discounted at y:

$$P(y) = \sum_{t=1}^{n} \frac{c_t}{(1+y)^t}$$

Here n is the year of the last payment. Applying the quadratic Taylor polynomial to the price, and dividing by the price so that the change is a percentage, gives the relationship every bond trader uses, where Δy is the change in yield:

$$\frac{\Delta P}{P} \approx -D\,\Delta y + \tfrac{1}{2} C\,(\Delta y)^2, \qquad D = -\frac{P'(y)}{P}, \qquad C = \frac{P''(y)}{P}$$

The first-order sensitivity D is the modified duration, and the second-order coefficient C is the convexity. Duration has a second meaning that gives it its name. The Macaulay duration, the present-value-weighted average time at which the bond's payments arrive, equals $D(1+y)$. A zero-coupon bond, with its single payment at maturity, has a Macaulay duration equal to its maturity; a coupon bond's is shorter, because some of its value arrives early.

Worked example — A ten-year bond

A ten-year bond paying a 4% coupon once a year is priced at its \$1,000 face value when its yield is 4%. Its Macaulay duration is 8.44 years, its modified duration $D = 8.44/1.04 \approx 8.11$, and its convexity $C \approx 80.7$. If its yield rises one percentage point, so $\Delta y = 0.01$:

$$\frac{\Delta P}{P} \approx -8.11 \times 0.01 + \tfrac{1}{2} \times 80.7 \times (0.01)^2 = -8.11\% + 0.40\% = -7.71\%$$

The exact fall, from repricing the bond at 5%, is 7.72%. Duration alone overstated the loss by 0.4 points, and convexity recovered almost all of it. Convexity works in the holder's favor both ways: when the yield falls a point, the price rises 8.53%, more than duration's 8.11%. The dollar value of a basis point, the DV01, is $D \times P \times 0.0001$, about 81 cents per \$1,000 of this bond.

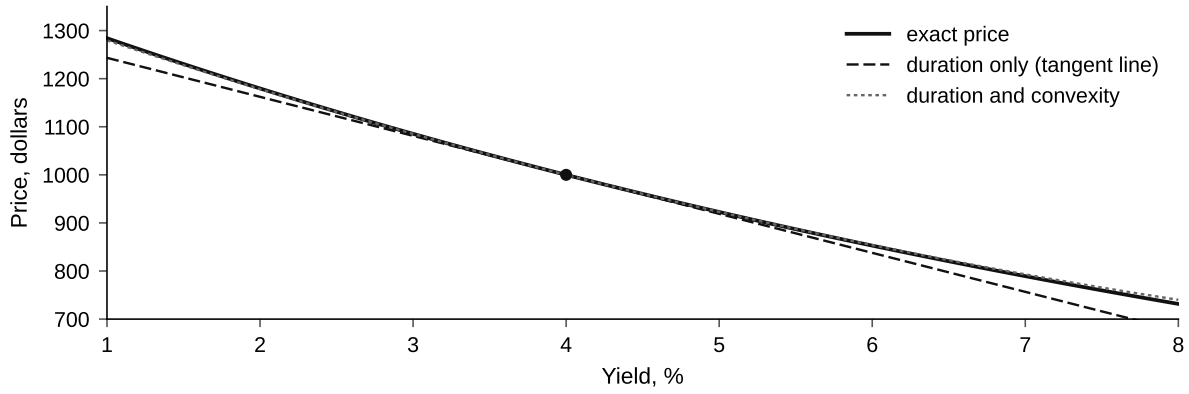


Figure 9.1. The ten-year 4% bond's price against its yield. The tangent line at 4% is duration's estimate; adding convexity bends the estimate to follow the curve, which lies above the tangent on both sides.

9e. Knowing the Error

Every approximation in this module has the same weakness: it is built from information at one point, so it degrades with distance. Three warnings follow. First, the error grows with the step, roughly as the first term left out: duration alone is fine for a basis point and misleading for two percentage points. Second, approximations fail at kinks. A call option near its strike at expiry has a slope of 0 on one side and 1 on the other, and no single tangent line describes both. Third, an approximation fitted to calm conditions says nothing about a jump, such as a default or a market halt.

The working rule is to use the simplest approximation whose first neglected term is too small to matter, and to check it against an exact calculation whenever the step is large.

9f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Linear approximation	$f(a+\Delta x) \approx f(a) + f'(a)\,\Delta x$ Near the input a, for a small step Δx: the tangent line
Quadratic approximation	$f(a+\Delta x) \approx f(a) + f'(a)\,\Delta x + \frac{1}{2} f''(a)\,(\Delta x)^2$ Adds the curvature; f'' is the second derivative
Series near zero	$e^x \approx 1 + x + \dfrac{x^2}{2} \qquad \ln(1+x) \approx x - \dfrac{x^2}{2} \qquad (1+x)^n \approx 1 + nx$ x is small and n is any number; the first term left out estimates the error
Factorial	$k! = 1 \times 2 \times \ldots \times k$; for example, $3! = 6$ and $4! = 24$
Newton's method	$x_{n+1} = x_n - \dfrac{g(x_n)}{g'(x_n)}$ Solves $g(x) = 0$ from a guess x_0; near a solution, correct digits roughly double each step
Bond price change	$\dfrac{\Delta P}{P} \approx -D\,\Delta y + \frac{1}{2} C\,(\Delta y)^2$ P is the price, Δy the change in yield, D the modified duration, and C the convexity
Duration and convexity	$D = -\dfrac{P'(y)}{P} \qquad C = \dfrac{P''(y)}{P} \qquad D_{\text{Mac}} = D\,(1+y)$ D_{Mac}, the Macaulay duration, is the present-value-weighted average time of the payments; for a zero-coupon bond it equals the maturity
DV01	$D \times P \times 0.0001$: the price change for a one-basis-point move in yield
The idea	Near a point, a smooth function looks like its tangent line and, a little farther out, like its Taylor polynomial. A bond's duration is the first derivative of its price and its convexity the second.

9g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

9.1 Estimate $\sqrt{49.5}$ with a linear approximation.

9.2 Estimate $e^{0.05}$ with the first three terms of its series.

9.3 Estimate $\ln 1.03$ with one term of its series, then with two.

9.4 Estimate 1.02^{10} with $(1+x)^n \approx 1 + nx$, then add the quadratic term.

9.5 Take two steps of Newton's method for $x^2 - 5 = 0$, starting from $x_0 = 2$.

9.6 A bond priced at \$1,000 has a modified duration of 6. Ignoring convexity, estimate its price change for a 0.5-point rise in yield, and find its DV01.

Applied

9.7 A zero-coupon bond maturing in 8 years yields 4%. Find its Macaulay and modified durations, and estimate its percentage price change for a 0.25-point rise in yield.

9.8 A bond with modified duration 7.5 and convexity 70 sees its yield rise 2 percentage points. Estimate its percentage price change with duration alone, then with convexity.

9.9 A project costs \$1,000 and pays \$1,210 in two years, so its IRR solves $g(r) = -1{,}000 + 1{,}210/(1+r)^2 = 0$. Take two Newton steps from $r = 5\%$, then find the exact answer.

9.10 In a simple model, a battery's capacity after n cycles is $Q_0\, e^{-0.0001 n}$, where Q_0 is its capacity when new. Use $e^x \approx 1 + x$ to estimate the share of capacity left after 500 cycles, and compare with the exact value.

9.11 Estimate the discount factor 1.05^{-10} with the quadratic approximation of $(1+x)^n$, and compare with the exact value. Why is the error larger than in drill problem 9.4?

9.12 A five-year bond with a 4% annual coupon trades at \$950. At a 4% yield its price would be \$1,000, with $P'(0.04) \approx -4{,}451.8$. Take one Newton step from 4% toward its yield.

Challenge

9.13 Use the linear approximation to show that one gradient-descent step, from w to $w - \alpha L'(w)$, lowers the loss by about $\alpha L'(w)^2$ when the learning rate α is small.

9.14 Derive the series for e^x up to the x^3 term from the fact that every derivative of e^x is e^x.

9.15 For a zero-coupon bond with face value F maturing in T years, $P(y) = F(1+y)^{-T}$. Show that its modified duration is $T/(1+y)$ and its convexity $T(T+1)/(1+y)^2$.

9.16 At expiry, a call option with strike K is worth $\max(S - K, 0)$, where S is the stock price. Explain why a linear approximation of its value around $S = K$ fails.

Module 9 Quiz

Check Your Understanding

1. What is the linear approximation of $f(a + \Delta x)$?

A) $f'(a)\Delta x$
B) $f(a) + f'(a)\Delta x$
C) $f(a)$
D) $f(a)\Delta x$, the value times the step

2. What does a bond's modified duration measure?

A) Its price sensitivity to yield
B) The years until the bond matures
C) The bond's coupon divided by its price
D) How much the duration itself changes as yields move, which traders call the bond's convexity

3. Starting from $x_0 = 1.5$, Newton's method for $x^2 - 2 = 0$ gives a first step of:

A) 1.5
B) 0.75, half the start
C) 1.4167
D) 1.4142, exactly

4. What is the quadratic Taylor approximation of e^x near zero?

A) $1 + x$
B) $x + x^2$
C) $1 + x^2$
D) $1 + x + x^2/2$

5. Why does adding convexity improve a duration estimate for a large change in yield?

A) Because yields cannot fall
B) Because convexity adds the coupon income
C) Because duration is only accurate for zero-coupon bonds, and convexity corrects for the coupons
D) Because the price-yield curve bends, and convexity, the second derivative, measures the bend that the tangent line misses

Module 10 — Several Variables

Partial derivatives, sensitivities, the option Greeks, and the gradient

Recommended Reading / Reference

Book: John C. Hull, *Options, Futures, and Other Derivatives* (11th ed., 2021) — the standard reference on option pricing and the Greeks

Reference: Cboe Global Markets (cboe.com) — delayed option quotes showing each contract's delta, gamma, theta, and vega

10a. Functions of Several Inputs

Most quantities that matter depend on more than one input. A battery project's value depends on its revenue, its cost, its life, and the discount rate. An option's price depends on the stock price, the time to expiry, interest rates, and how volatile the stock is. A neural network's loss depends on every one of its weights. A function of two inputs, $f(x, y)$, gives one output for each pair of inputs x and y.

A function of two inputs can be drawn as a surface over a flat map, or more usefully as a contour map: curves joining the points where the output takes the same value, like the contour lines on a hiking map. Where contours crowd together the surface is steep; where they spread apart it is nearly flat.

10b. Partial Derivatives

A partial derivative measures how the output changes when one input moves and every other input is held fixed. For $f(x, y)$, the partial derivative with respect to x, over a small step h, is written with a curly ∂, read “partial”:

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x + h,\ y) - f(x,\ y)}{h}$$

To compute one, treat every other input as a constant and use the rules of Module 5. For $f(x, y) = x^2 y + 3y$, holding y fixed gives $\partial f/\partial x = 2xy$, and holding x fixed gives $\partial f/\partial y = x^2 + 3$. Second partial derivatives work the same way, and for the smooth functions in this course the order does not matter: differentiating by x and then y gives the same result as y and then x.

The partial derivatives combine into a linear approximation, the several-input version of Module 9's tangent line. Small moves Δx and Δy in the inputs change the output by about the sum of each sensitivity times its move:

$$\Delta f \approx \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y$$

This is how risk managers attribute a portfolio's daily profit or loss to its sources: so much from the stock market, so much from interest rates, so much from currencies, each a sensitivity times a move.

Worked example — Wind speed and air density

The power in the wind, $P = (1/2)\rho A v^3$, depends on the air density ρ, the area A the blades sweep, and the wind speed v. Its partial derivatives are $\partial P/\partial v = (3/2)\rho A v^2$ and $\partial P/\partial \rho = (1/2)Av^3$. As shares of P, a 1% rise in wind speed raises power by about 3%, and a 1% rise in air density, as on a cold day, raises it by about 1%. For any power function, the percentage sensitivity to each input, its elasticity, is that input's exponent.

10c. How Sensitive Is a Project?

Every financial model is a function of several inputs, and its partial derivatives say which inputs matter. Module 7 valued a battery project netting c million dollars a year for T years, discounted at a continuous rate r. If it costs K million dollars to build, its net present value, or NPV, is

$$\mathrm{NPV}(c,\ r,\ T,\ K) = c\,\frac{1 - e^{-rT}}{r} - K$$

Worked example — A battery project's sensitivities

With $c = 2$, $T = 15$ years, $r = 8\%$, and $K = 14$, the NPV is \$3.47 million. The partial derivative with respect to revenue is $(1 - e^{-rT})/r \approx 8.74$: each extra \$1 million a year adds \$8.74 million of value. With respect to cost it is −1. With respect to the discount rate it is about −105, so one percentage point on the rate, 0.01, moves the NPV by about \$1.05 million. If revenue rises to \$2.2 million and the rate to 9%, the linear approximation gives $3.47 + 8.74 \times 0.2 - 105 \times 0.01 \approx 4.16$, against an exact \$4.11 million.

Analysts show such sensitivities in a tornado chart: each input is moved across a plausible range with the others held fixed, and the resulting swings in value are drawn as bars, widest at the top.

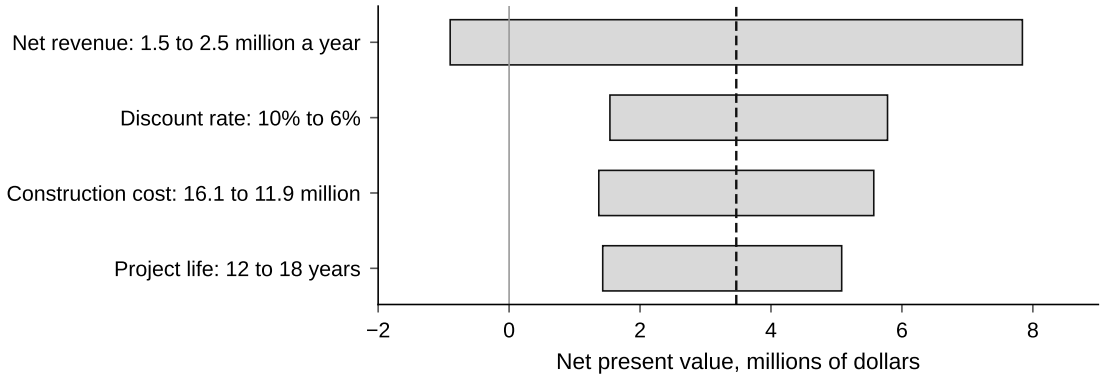


Figure 10.1. A tornado chart for the battery project, with illustrative numbers. Each bar shows the range of value as one input moves across its range, with the others at their base values; the dashed line is the base case, \$3.47 million.

The chart ranks the questions worth researching. The revenue forecast dominates: a range of \$1.5 million to \$2.5 million a year swings the project from a loss to its best case, while construction cost, the discount rate, and the project's life each matter about half as much. Effort spent pinning down revenue is worth more than effort spent on the rest.

10d. The Option Greeks

Fischer Black and Myron Scholes published a formula for the value of a stock option in 1973, and Robert Merton extended it the same year; Scholes and Merton received the Nobel prize in economics for it in 1997, Black having died in 1995. For a call option, the right to buy a stock now priced at S for a strike price K at a time T years from now, the formula is

$$C = S\,N(d_1) - K\,e^{-rT}N(d_2), \qquad d_1 = \frac{\ln(S/K) + \left(r + \tfrac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}, \qquad d_2 = d_1 - \sigma\sqrt{T}$$

where C is the call's value, r the continuous interest rate, and σ the stock's volatility, the standard deviation of its yearly return. N is the cumulative normal distribution: $N(d)$ is the probability that a standard bell-curve variable falls below d, available in any spreadsheet as NORM.S.DIST(d, TRUE). Deriving the formula needs probability beyond this course; using it needs only the formula and a spreadsheet.

Traders manage options through the formula's partial derivatives, called the Greeks after the letters that name most of them:

Greek	Partial derivative	Meaning
Delta	$\partial C/\partial S = N(d_1)$	Change in the option's price per $1 move in the stock
Gamma, Γ	$\partial^2 C/\partial S^2$	How fast delta itself changes: the option's convexity
Vega	$\partial C/\partial \sigma$	Change in price per unit of volatility
Theta	Rate of change with time	Time decay: the value lost as expiry approaches
Rho	$\partial C/\partial r$	Change in price per unit of interest rate

The option Greeks. Vega is not a Greek letter, but the name stuck.

Worked example — An at-the-money call

Take a stock at $100, a call with a strike of $100 expiring in six months ($T = 0.5$), an interest rate of 4%, and volatility of 25%. Then $d_1 \approx 0.20$ and $d_2 \approx 0.02$, and the call is worth about $8.01. Its delta is $N(0.20) \approx 0.58$: the option gains about 58 cents for each dollar the stock rises. Its gamma is about 0.022; its vega about 27.6 per unit of volatility, or 28 cents per volatility point; and its theta about −$8.91 a year, or 2.4 cents a day.

Delta and gamma play the roles that duration and convexity played in Module 9. For a stock move of ΔS dollars, the option's price changes by about

$$\Delta C \approx N(d_1)\,\Delta S + \tfrac{1}{2}\Gamma\,(\Delta S)^2$$

For a $3 rise, $0.58 \times 3 + 0.5 \times 0.022 \times 9 \approx 1.84$, against an exact $1.837. A trader who sells this call can offset its delta by holding 0.58 shares of the stock per option, so that small moves in the stock leave the combined position unchanged. Gamma says how quickly that hedge goes stale and must be adjusted.

10e. The Gradient

The partial derivatives of a function, collected into a list, form its gradient, written ∇f and read “grad f”:

$$\nabla f(x,\,y) = \left(\frac{\partial f}{\partial x},\,\frac{\partial f}{\partial y}\right)$$

The gradient is a direction on the contour map. It points the way the function rises fastest, its length is that fastest rate of rise, and it crosses the contour lines at right angles. Walking against the gradient is the fastest way down.

That is gradient descent in many dimensions, the training method of Modules 5 and 8 with every weight at once. For a loss L that depends on a list of weights w, each step moves every weight against its own partial derivative, scaled by the learning rate α:

$$w_{\text{new}} = w - \alpha \, \nabla L(w)$$

Backpropagation, the chain rule run backward through a network, computes all the partial derivatives in one pass, which is how models with billions of weights are trained.

Worked example — Descent down a narrow valley

Take $L(w_1, w_2) = (w_1 - 1)^2 + 4(w_2 + 2)^2$, a valley whose lowest point is at $(1, -2)$ and which is four times steeper in the w_2 direction. Its gradient is $(2(w_1 - 1), 8(w_2 + 2))$. From $(-1, 0)$ with $\alpha = 0.2$, the first step moves to $(-0.2, -3.2)$, then $(0.28, -1.28)$, then $(0.57, -2.43)$: steady progress along the gentle direction, but a zigzag across the steep one.

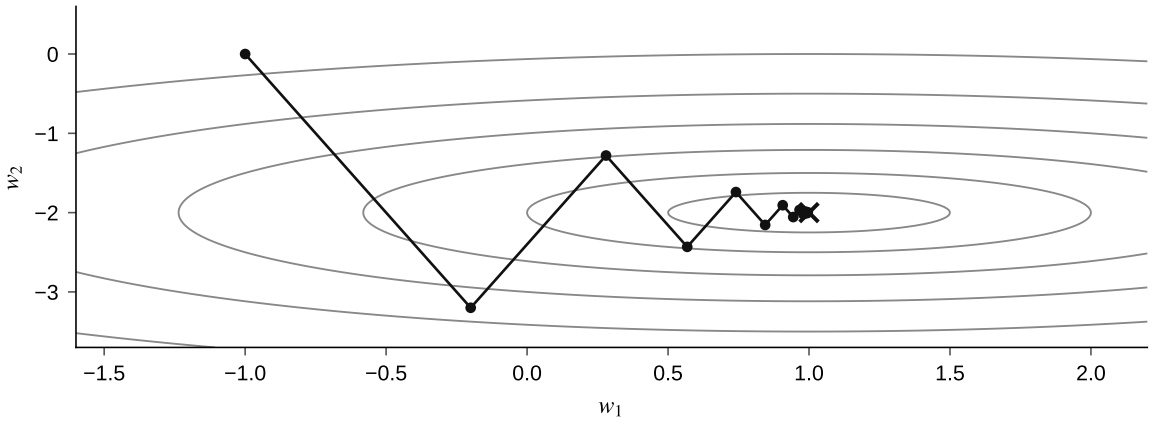


Figure 10.2. Gradient descent on a narrow valley, from (−1, 0) with a learning rate of 0.2. The contours are lines of equal loss; the path zigzags across the steep direction while it closes in along the gentle one, reaching the minimum, marked ×, at (1, −2).

The zigzag is the several-input version of Module 8's instability: the learning rate must be small enough for the steepest direction, which makes progress slow in the gentle ones. Much of the craft of training large models, in methods with names like momentum and Adam, goes into adapting the step to each direction.

Setting every partial derivative to zero finds the critical points of a function of several inputs, as $f'(x) = 0$ did for one input. Fitting a line with both a slope and an intercept is the standard example, and it is the first challenge problem below.

10f. Memorize This

Commit these to memory. The practice problems assume them, and so does much of the mathematics you will meet after this course.

Partial derivative	$\frac{\partial f}{\partial x}$ The rate of change of f as x moves with every other input held fixed: treat the other inputs as constants
Linear approximation	$\Delta f \approx \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y$ For small moves Δx and Δy in the inputs x and y
Elasticity	$\frac{x}{f}\frac{\partial f}{\partial x}$ The percentage change in f per 1% change in x; for a power function, the input's exponent
Black–Scholes call	$C = S\,N(d_1) - K\,e^{-rT}N(d_2)$ S is the stock price, K the strike, T the years to expiry, r the interest rate, σ the volatility, and N the cumulative normal; d_1 and d_2 are given in section 10d
The Greeks	Delta $= \partial C/\partial S = N(d_1)$; gamma $= \partial^2 C/\partial S^2$; vega $= \partial C/\partial \sigma$; theta is time decay; rho $= \partial C/\partial r$ C is the option's value
Option price change	$\Delta C \approx N(d_1)\,\Delta S + \frac{1}{2}\Gamma\,(\Delta S)^2$ ΔS is the stock's move and Γ the gamma: the options version of duration and convexity
Gradient	$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$ Points the way f rises fastest, at right angles to the contour lines
Gradient descent	$w_{\text{new}} = w - \alpha\,\nabla L(w)$ Every weight in the list w moves against its own partial derivative of the loss L; α is the learning rate
Critical point	Every partial derivative equals zero
The idea	With several inputs, the derivative splits into one partial derivative per input. Together they form the gradient, which gives every sensitivity at once and points the way uphill.

10g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

10.1 For $f(x, y) = 3x^2y - 2y^3$, find $\partial f/\partial x$ and $\partial f/\partial y$.

10.2 For $g(x, y) = x\,e^{xy}$, find $\partial g/\partial y$.

10.3 For $f(x, y) = x^2 + xy + y^2$, find the gradient at the point (1, 2).

10.4 Find the critical point of $f(x, y) = x^2 + y^2 - 4x + 6y$.

10.5 For $f(x, y) = x^3y^2$, check that the two mixed second partial derivatives are equal.

10.6 For $f(x, y) = x^2y$, estimate $f(2.1, 2.95)$ with the linear approximation from the point (2, 3).

Applied

10.7 Using the elasticities of the wind-power formula, estimate the percentage change in the wind's power if the wind speed rises 2% and the air density falls 1%.

10.8 Using the sensitivities of section 10c, estimate the project's NPV if construction cost rises by $1 million and the discount rate falls to 7%.

10.9 A call option has a delta of 0.56 and a gamma of 0.022. Estimate its price change if the stock rises $3, and if it falls $3.

10.10 A trader holds 100 call options, each on one share, with a delta of 0.56. How many shares must the trader sell short to make the position insensitive to small moves in the stock?

10.11 An option's vega is 27.6 per unit of volatility. Estimate its price change if volatility rises from 25% to 28%.

10.12 Run two steps of gradient descent on $L(w_1, w_2) = (w_1 - 1)^2 + 4(w_2 + 2)^2$ from (0, 0) with a learning rate of 0.1.

Challenge

10.13 Fit the line $y = a + bx$ to the points (1, 3), (2, 5), and (3, 6) by setting both partial derivatives of $E(a, b) = \Sigma(y_i - a - bx_i)^2$ to zero, where a is the intercept, b the slope, and Σ adds over the three points (x_i, y_i).

10.14 The dot product of two directions (p, q) and (u, v) is $pu + qv$, and it is zero exactly when they are at right angles. Use the linear approximation to show that the gradient is at right angles to a contour line.

10.15 For $P = kx^a y^b$, where k, a, and b are constants, show that the elasticity of P with respect to x is a and with respect to y is b.

10.16 For $L(w_1, w_2) = (w_1 - 1)^2 + 4(w_2 + 2)^2$, show that each gradient-descent step multiplies the error in w_1 by $1 - 2\alpha$ and the error in w_2 by $1 - 8\alpha$. For which learning rates does descent converge, and why does the steep direction set the limit?

Module 10 Quiz

Check Your Understanding

1. For $f(x, y) = x^2 y$, what is $\partial f/\partial x$?

A) $2xy + x^2$
B) $2xy$
C) $x^2 y$, unchanged
D) $2x$, treating y as zero

2. What does an option's delta measure?

A) The change in the option's price per $1 change in the stock
B) The option's price
C) The time until expiry
D) How fast the option loses value each day as expiry approaches, which traders call its time decay

3. Which way does the gradient of a function point?

A) It points downhill.
B) It points along the contour lines, where the function stays constant.
C) It points in the direction of steepest increase.
D) It points toward the function's minimum, which is why it is used directly to find the lowest point.

4. A project's NPV, in millions of dollars, has a partial derivative of −105 with respect to the discount rate. What does that mean?

A) NPV is negative.
B) Raising the discount rate by one percentage point cuts NPV by $105 million.
C) The project loses $105 for every dollar invested, so it should be rejected regardless of the rate.
D) A one-point rise in the rate cuts NPV by about $1.05 million.

5. How does gradient descent train a model with many weights?

A) Each weight moves against its own partial derivative.
B) It tries every combination of weights and keeps the one with the lowest loss.
C) It sets each partial derivative to zero and solves the equations exactly by algebra.
D) It adjusts only the weight with the largest partial derivative, leaving the others unchanged until later.

Everything to Memorize

The Memorize This boxes from every module, in one place

Module 1: The Algebra Underneath

Multiplying and dividing powers	$a^m a^n = a^{m+n}$ $\quad \frac{a^m}{a^n} = a^{m-n}$ $\quad (a^m)^n = a^{mn}$ a is any base; m and n are any powers
Zero, negative, and fractional powers	$a^0 = 1$ $\quad a^{-n} = \frac{1}{a^n}$ $\quad a^{1/n} = \sqrt[n]{a}$ a is any nonzero base; $a^{1/n}$ is the nth root of a
Sums do not split	$(a+b)^2 = a^2 + 2ab + b^2$, and $\sqrt{(a+b)}$ is not $\sqrt{a} + \sqrt{b}$ a and b are any numbers
Quadratic formula	$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ Solves $ax^2 + bx + c = 0$, where a, b, and c are numbers and a is not zero
Growth rate from start to end	$r = \left(\frac{A}{P}\right)^{1/n} - 1$ P grows to A in n years; r is the compound annual growth rate
Break-even volume	$q = \frac{F}{p - c}$ F is the fixed cost, p the price, c the variable cost per unit, and q the units sold
Adding fractions	$a/b + c/d = (ad + bc)/(bd)$; cancel factors, never terms
Difference quotient	$\frac{f(x+h) - f(x)}{h}$ How fast the function f changes between the inputs x and $x + h$; h is the size of the step
Energy and power	Energy = power × time: 1 MW for 1 hour is 1 MWh, which is 1,000 kWh
Percentages	Percent change = (new − old)/old; 1 percentage point = 100 basis points; percent changes compound, so they do not add
The idea	Most calculus mistakes are algebra mistakes. Two answers to a quadratic are usually two real choices, like a low price and a high one that earn the same revenue.

Module 2: Functions and Graphs

Slope	$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$ Δ means "change in"; (x_1, y_1) and (x_2, y_2) are two points on the line; slope has units of output per unit of input
Lines	$y = mx + b$; through the point (x_1, y_1) with slope m: $y - y_1 = m(x - x_1)$ m is the slope and b the intercept, the value of y when $x = 0$
Scaling a power function	Multiplying the input of $y = kx^n$ by a factor c multiplies the output by c^n k is a constant and n the power

Power in the wind	$P = \frac{1}{2}\rho A v^3$ P is power, ρ the density of air (about 1.225 kg per cubic meter), A the area the blades sweep, and v the wind speed; 10% more wind carries 33% more power
Composition	$f(g(x))$: apply g first, then f
Inverse	Solve $y = f(x)$ for x; $f^{-1}(f(x)) = x$; the −1 marks an inverse, not a reciprocal
Piecewise rules	A tier or bracket changes the slope at the threshold; only the units above it pay the new rate
Hockey sticks	Call option payoff $\max(S - K, 0)$; ReLU activation $\max(x, 0)$ S is the stock price at expiry and K the strike price
Chart checks	Axis start; linear or log; totals or per person; nominal or real; one axis or two
The idea	A function turns each input into exactly one output, and every calculus problem starts by writing its quantity as a function.

Module 3: Exponentials and Logarithms

Compound growth	$A = P(1 + r)^t \qquad A = P\left(1 + \frac{r}{n}\right)^{nt} \qquad A = Pe^{rt}$ P is the starting amount, A the amount after t years, r the annual rate, and n the number of times a year interest is credited
The number e	$e = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n \approx 2.718$
Logs undo exponentials	$e^{\ln x} = x \qquad \ln(e^x) = x \qquad \ln 1 = 0 \qquad \ln e = 1$ ln is the natural logarithm, base e
Log rules	$\ln(ab) = \ln a + \ln b \qquad \ln\frac{a}{b} = \ln a - \ln b \qquad \ln(a^k) = k \ln a$ a and b are positive numbers; k is any power
Time to grow from P to A	$t = \frac{\ln(A/P)}{\ln(1 + r)}$ t is in years when r is a yearly rate
Doubling time	$\ln 2/\ln(1 + r)$, about 72 divided by the rate in percent; $\ln 2 \approx 0.693$
Half-life	$N = N_0(\frac{1}{2})^{t/T} = N_0 e^{-\lambda t} \qquad \lambda = \frac{\ln 2}{T}$ N_0 is the starting amount, N the amount left after time t, T the half-life, and λ the decay constant
Log axes	Steady percentage growth is a straight line, and its slope is the growth rate
Learning curves	Cost falls by the learning rate with each doubling of cumulative production: about 20% for solar modules $C = C_0(Q/Q_0)^{-b}$: C_0 is the cost when cumulative production was Q_0, and 2^{-b} is one minus the learning rate
The idea	Exponential growth multiplies by a constant factor each period, so it always outruns linear growth: $1,000 at 5% becomes $4,322 in 30 years.

Module 4: Rates of Change

Average rate of change	$\frac{f(b) - f(a)}{b - a}$ Between the inputs a and b: the slope of the secant line, in output units per input unit
The derivative	$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$ The instantaneous rate of change at x and the slope of the tangent line; h is the step
Notation	$f'(x) = \frac{dy}{dx} = \frac{d}{dx} f(x)$ For $y = f(x)$; dy/dx is read “dee y dee x”
Quadratics	If $f(x) = a + bx + cx^2$, then $f'(x) = b + 2cx$ a, b, and c are numbers
Signs of the derivative	$f' > 0$: rising; $f' < 0$: falling; $f' = 0$: flat, as at a peak or a trough
Second derivative	$f'' > 0$: the graph bends upward, like a cup; $f'' < 0$: it bends downward, like a cap f'' is the derivative of f'
Marginal cost	$C'(q)$: the cost of one more unit; marginal above average pulls the average up $C(q)$ is the total cost of q units
Marginal price of power	LMP $= dC/dL$: the cost of serving one more MWh at a location C is the system's total cost and L the demand at that location
Power and energy	$P = dE/dt$: megawatt-hours per hour are megawatts $E(t)$ is stored or cumulative energy at hour t, and P is power
Central difference	$f'(x) \approx \frac{f(x + h) - f(x - h)}{2h}$ Estimates a derivative from readings spaced h apart
No derivative	At a jump or a kink, such as the ReLU's corner at zero
The idea	A derivative is a rate measured at a single point: how fast one quantity changes per unit change in another, found as the limit of averages over shorter and shorter intervals.

Module 5: The Rules of Differentiation

Power rule	$\frac{d}{dx} x^n = n\, x^{n-1}$ n is any number, including negative and fractional powers
Constants and sums	$d(k)/dx = 0$; $(k \cdot f)' = k \cdot f'$; $(f + g)' = f' + g'$ k is a constant; f and g are functions of x
Exponentials	$\frac{d}{dx} e^x = e^x$ $\quad$ $\frac{d}{dx} e^{kx} = k\, e^{kx}$ $\quad$ $\frac{d}{dx} a^x = a^x \ln a$ k is a constant and a a positive base
Logarithm	$\frac{d}{dx} \ln x = \frac{1}{x}$ For positive x

Product and quotient	$(fg)' = f'g + fg' \qquad \left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$ f and g are functions of the same input
Chain rule	$\frac{d}{dx}f(g(x)) = f'(g(x))\,g'(x) \qquad \frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$ $u = g(x)$ is the inside function: differentiate the outside, then multiply by the inside's derivative
Sigmoid	$\sigma(z) = \frac{1}{1 + e^{-z}} \qquad \sigma'(z) = \sigma(z)(1 - \sigma(z))$ z is the neuron's input; the output lies between 0 and 1
Log derivative	$\frac{d}{dx}\ln f(x) = \frac{f'(x)}{f(x)}$ The relative, or percentage, rate of change of a positive function f
Elasticity	$\varepsilon = \frac{p}{q}\frac{dq}{dp}$ q is quantity and p price; revenue rises with price when ε is between −1 and 0
The idea	A handful of rules differentiate everything in this course. The chain rule, which multiplies the rates of nested functions, is the one used most, and backpropagation is the chain rule repeated.

Module 6: Optimization

Critical point	An input c where $f'(c) = 0$; check the endpoints of the allowed range and any kinks too f is the quantity being optimized
Second-derivative test	$f''(c) < 0:$ peak $\qquad f''(c) > 0:$ trough At a critical point c; when $f''(c) = 0$, check the sign of f' on either side
Profit-maximizing output	Marginal revenue equals marginal cost: $R'(q) = C'(q)$ R is revenue, C cost, and q output
Markup rule	$\frac{p - c}{p} = \frac{1}{\lvert\varepsilon\rvert}$ p is the price, c the unit cost, and $\lvert\varepsilon\rvert$ the size of the elasticity of demand
Battery duration	Add hours until the last hour's extra revenue falls to its cost Duration is energy capacity divided by power capacity
Economic order quantity	$Q^* = \sqrt{\frac{2DS}{H}}$ D units a year, S dollars per order, H dollars to hold a unit for a year; ordering and holding costs are equal at Q^*
Least squares through the origin	$w = \frac{\sum x_i y_i}{\sum x_i^2}$ Fits $y = wx$ to the points (x_i, y_i); Σ means add up over all the points
Recipe	Write the quantity as a function of the choice, state the allowed range, set the derivative to zero, check the second derivative and the endpoints, and interpret the answer
The idea	The best value of a smooth function sits where its derivative is zero or at an edge of what is allowed, and the second derivative tells a peak from a trough.

Module 7: Integrals as Accumulation

Definite integral	$\int_a^b f(x)\,dx$ The net accumulation of the rate f from a to b: the area under its graph, counted negative below the axis
Fundamental theorem	$\int_a^b f(x)\,dx = F(b) - F(a)$ F is any antiderivative of f: a function with $F' = f$
Antiderivatives	$\int x^n dx = \frac{x^{n+1}}{n+1} + C \qquad \int \frac{dx}{x} = \ln x + C \qquad \int e^{kx} dx = \frac{e^{kx}}{k} + C$ n is not −1, x is positive in the logarithm, k is not 0, and C is any constant
Substitution	$\int f(g(x))\,g'(x)\,dx = F(g(x)) + C$ The chain rule read backward; F is an antiderivative of f
Energy and revenue	Energy $= \int P(t)\,dt$; revenue $= \int P(t)\,\lambda(t)\,dt$ P is power in MW, t time in hours, and λ the price in dollars per MWh
Trapezoid rule	$h\left(\frac{1}{2}f_0 + f_1 + \cdots + f_{n-1} + \frac{1}{2}f_n\right)$ For readings f_0 to f_n spaced h apart; the error shrinks with the square of h
Present value of a stream	$PV = \int_0^T c(t)\,e^{-rt}\,dt \qquad \text{constant } c: \ \frac{c\,(1 - e^{-rT})}{r}$ $c(t)$ dollars a year for T years at a continuous rate r
Perpetuities	$\text{level: } \frac{c}{r} \qquad \text{growing: } \frac{c}{r - g}$ g is the growth rate, which must be below r
Consumer surplus	$\int_0^{q^*} (D(q) - p^*)\,dq$ $D(q)$ is what a buyer would pay for the qth unit; p^* is the price and q^* the quantity sold
The idea	An integral adds up a rate over an interval, and the fundamental theorem says the total is the change in any function whose derivative is that rate.

Module 8: Differential Equations

Differential equation	A rule for the rate of change of a quantity in terms of its current value: $dy/dt = f(t, y)$ The solution $y(t)$ is fixed by its starting value, $y_0 = y(0)$
Growth and decay	$\frac{dy}{dt} = ky \quad \Longrightarrow \quad y = y_0 e^{kt}$ k is the continuous rate: positive for growth, negative for decay
Settling toward equilibrium	$\frac{dT}{dt} = -k\,(T - T_{\text{env}}) \implies T = T_{\text{env}} + (T_0 - T_{\text{env}})\,e^{-kt}$ T_{env} is the surroundings and T_0 the start; after one time constant, $1/k$, the gap is down to 37%
Battery state of charge	Charging: $dE/dt = \eta_c P_c$; discharging: $dE/dt = -P_d/\eta_d$; idle: $dE/dt = -\lambda E$ η_c and η_d are the efficiencies, P_c and P_d the powers, and λ the idle loss rate; the round trip is $\eta_c\,\eta_d$

Logistic growth	$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \quad \Longrightarrow \quad N = \frac{K}{1 + e^{-r(t - t_0)}}$ K is the ceiling and t_0 the midpoint; growth is fastest at $K/2$, and 5% to 50% takes $\ln 19 / r$
Euler's method	$y_{n+1} = y_n + h\, f(t_n, y_n)$ h is the step; smaller steps are more accurate
Gradient descent	$w_{n+1} = w_n - \alpha\, L'(w_n)$ L is the loss, w a weight, and α the learning rate; too large a rate overshoots and can diverge
The idea	A differential equation says how a quantity changes given where it is now; solving it, by formula or by many small Euler steps, gives its whole path through time.

Module 9: Approximation

Linear approximation	$f(a + \Delta x) \approx f(a) + f'(a)\, \Delta x$ Near the input a, for a small step Δx: the tangent line
Quadratic approximation	$f(a + \Delta x) \approx f(a) + f'(a)\, \Delta x + \frac{1}{2} f''(a)\, (\Delta x)^2$ Adds the curvature; f'' is the second derivative
Series near zero	$e^x \approx 1 + x + \frac{x^2}{2} \qquad \ln(1 + x) \approx x - \frac{x^2}{2} \qquad (1 + x)^n \approx 1 + nx$ x is small and n is any number; the first term left out estimates the error
Factorial	$k! = 1 \times 2 \times \ldots \times k$; for example, $3! = 6$ and $4! = 24$
Newton's method	$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$ Solves $g(x) = 0$ from a guess x_0; near a solution, correct digits roughly double each step
Bond price change	$\frac{\Delta P}{P} \approx -D\, \Delta y + \frac{1}{2} C\, (\Delta y)^2$ P is the price, Δy the change in yield, D the modified duration, and C the convexity
Duration and convexity	$D = -\frac{P'(y)}{P} \qquad C = \frac{P''(y)}{P} \qquad D_{\text{Mac}} = D\,(1 + y)$ D_{Mac}, the Macaulay duration, is the present-value-weighted average time of the payments; for a zero-coupon bond it equals the maturity
DV01	$D \times P \times 0.0001$: the price change for a one-basis-point move in yield
The idea	Near a point, a smooth function looks like its tangent line and, a little farther out, like its Taylor polynomial. A bond's duration is the first derivative of its price and its convexity the second.

Module 10: Several Variables

Partial derivative	$\frac{\partial f}{\partial x}$ The rate of change of f as x moves with every other input held fixed: treat the other inputs as constants
Linear approximation	$\Delta f \approx \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y$ For small moves Δx and Δy in the inputs x and y
Elasticity	$\frac{x}{f}\frac{\partial f}{\partial x}$ The percentage change in f per 1% change in x; for a power function, the input's exponent
Black–Scholes call	$C = S\,N(d_1) - K\,e^{-rT}N(d_2)$ S is the stock price, K the strike, T the years to expiry, r the interest rate, σ the volatility, and N the cumulative normal; d_1 and d_2 are given in section 10d
The Greeks	Delta $= \partial C/\partial S = N(d_1)$; gamma $= \partial^2 C/\partial S^2$; vega $= \partial C/\partial \sigma$; theta is time decay; rho $= \partial C/\partial r$ C is the option's value
Option price change	$\Delta C \approx N(d_1)\,\Delta S + \frac{1}{2}\Gamma\,(\Delta S)^2$ ΔS is the stock's move and Γ the gamma: the options version of duration and convexity
Gradient	$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$ Points the way f rises fastest, at right angles to the contour lines
Gradient descent	$w_{\text{new}} = w - \alpha\,\nabla L(w)$ Every weight in the list w moves against its own partial derivative of the loss L; α is the learning rate
Critical point	Every partial derivative equals zero
The idea	With several inputs, the derivative splits into one partial derivative per input. Together they form the gradient, which gives every sensitivity at once and points the way uphill.

Worked Solutions

Every practice problem, solved step by step

Module 1: The Algebra Underneath

1.1 Add exponents to multiply and subtract them to divide: $x^{4-1-2} = x^1 = x$.

1.2 $1/\sqrt{x} = x^{-1/2}$; $x\sqrt{x} = x^1 \cdot x^{1/2} = x^{3/2}$; $1/x^3 = x^{-3}$.

1.3 The cube root of 27 is 3, so $27^{2/3} = 3^2 = 9$. The square root of 16 is 4, so $16^{-1/2} = 1/4$. The cube root of $1/8$ is $1/2$, and the negative power turns it over: $(1/8)^{-1/3} = 2$.

1.4 Expand the right side: $5x - 7 = 2x + 8$. Subtract $2x$ from both sides and add 7: $3x = 15$, so $x = 5$.

1.5 Two numbers that multiply to −15 and add to 2 are 5 and −3, so $(x + 5)(x - 3) = 0$, and $x = -5$ or $x = 3$.

1.6 Factor the top and bottom: $(x - 2)(x + 2)/(x + 2)^2$. Cancel one factor of $x + 2$ to get $(x - 2)/(x + 2)$, for every x except −2.

1.7 Over the common denominator $x(x + 1)$: $[(x + 1) - x]/[x(x + 1)] = 1/[x(x + 1)]$.

1.8 From $t = E/P$: $200/50 = 4$ hours at full power, and $200/30 \approx 6.7$ hours, about 6 hours 40 minutes, at 30 MW.

1.9 $r = (40{,}000/25{,}000)^{1/6} - 1 = 1.6^{1/6} - 1$. A calculator gives $1.6^{1/6} \approx 1.0815$, so the rate is about 8.15% a year. Dividing the 60% total gain by six would overstate it as 10%.

1.10 Each MWh contributes $65 - 25 = 40$ dollars, so $q = F/(p - c) = 1{,}200{,}000/40 = 30{,}000$ MWh a year.

1.11 Set the two equal: $120 - 3p = 2p - 10$, so $130 = 5p$ and $p = 26$. Then $q = 120 - 78 = 42$, and the supply equation agrees: $52 - 10 = 42$.

1.12 The change is 0.35 percentage points, or 35 basis points. As a share of the starting yield it is $0.35/4.25 \approx 0.082$, an 8.2% increase.

1.13 It must take in $40/0.88 \approx 45.45$ MWh, which costs $45.45 \times 28 \approx 1{,}273$ dollars. The 5.45 MWh lost on the round trip account for about $153 of that.

1.14 Expand $(x + h)^3 = x^3 + 3x^2h + 3xh^2 + h^3$, then cancel and divide:

$$\frac{(x + h)^3 - x^3}{h} = \frac{3x^2h + 3xh^2 + h^3}{h} = 3x^2 + 3xh + h^2$$

As h shrinks, the last two terms vanish and the quotient approaches $3x^2$, the slope of $y = x^3$ at x.

1.15 It ends at $0.7 \times 1.3 = 0.91$ of its starting price, a 9% loss. Full recovery needs a rise of $1/0.7 - 1 \approx 0.429$, about 42.9%.

1.16 Divide by 2,000: $1.05^t = 2.5$. Trying values, $1.05^{18} \approx 2.41$ and $1.05^{19} \approx 2.53$, so t lies between 18 and 19.

Since 2.5 sits about three-quarters of the way from 2.41 to 2.53, $t \approx 18.75$. The exact answer, $\ln 2.5/\ln 1.05 \approx 18.78$, comes from Module 3.

Module 2: Functions and Graphs

2.1 The slope is $(19-7)/(4-1) = 12/3 = 4$. Then $7 = 4(1) + b$ gives $b = 3$, so the line is $y = 4x + 3$.

2.2 $g(2) = 3$, so $f(g(2)) = f(3) = 27 - 2 = 25$. $f(2) = 10$, so $g(f(2)) = 11$. The order matters.

2.3 Solve $y = 5x - 4$ for x: $x = (y+4)/5$, so $f^{-1}(y) = (y+4)/5$. Check: $f(2) = 6$ and $f^{-1}(6) = 2$.

2.4 The constant 3 does not matter: doubling multiplies y by $2^4 = 16$, and tripling by $3^4 = 81$.

2.5 0, 0, and 30. Below the strike the option pays nothing; above it, the excess.

2.6 f accepts every x except 2, where it would divide by zero. g needs $x - 3 \geq 0$, so its domain is $x \geq 3$.

2.7 For u kilowatt-hours, $B(u) = 10 + 0.14u$ dollars. Setting $10 + 0.14u = 136$ gives $0.14u = 126$, so $u = 900$ kWh.

2.8 At 350 kWh: $0.12 \times 350 = 42$ dollars, an average of 12 cents. At 1,200 kWh: $60 + 0.18 \times 700 = 186$ dollars, an average of $186/1{,}200 = 15.5$ cents. The average rises toward 18 cents as usage grows, but never reaches it.

2.9 $(8/7)^3 \approx 1.49$: about 49% more power from a 14% faster wind. Real comparisons need the whole distribution of wind speeds, not just the average, but the cube law sets the scale.

2.10 Profit is $\max(S - 50, 0) - 4$. Above the strike, $S - 54 = 0$ at $S = 54$: the stock must end at $54 for the payoff to repay the $4 cost.

2.11 Substitute: $L = 2{,}000 + 60(1.5h) = 2{,}000 + 90h$. At 3 p.m., $h = 6$, so $L = 2{,}540$ MW. The slope, 90 MW per hour, is 60 MW per degree times 1.5 degrees per hour, the chain rule in miniature.

2.12 The rise covers $(420-412)/(421-410) = 8/11 \approx 73\%$ of the chart's height, but the index rose only $8/412 \approx 1.9\%$.

2.13 Surface area is $6s^2$ and volume s^3, so the ratio is $6s^2/s^3 = 6/s$. It falls as s grows: doubling the side multiplies the heat by 8 but the cooling surface only by 4. That is the square-cube law, and the reason large cells need active cooling.

2.14 Multiply out $y(x-3) = 2x + 1$, so $yx - 3y = 2x + 1$. Collect the x terms: $x(y-2) = 3y + 1$, so $x = (3y+1)/(y-2)$. Check: $f(4) = 9/1 = 9$, and $f^{-1}(9) = 28/7 = 4$.

2.15 Payoff $= \max(S - 100, 0) - 2\max(S - 110, 0)$. At 95 it is 0; at 105, 5; at 110, 10; at 125, $25 - 30 = -5$. The position profits from a moderate rise and loses on a large one.

2.16 Below 0 every term is 0. On $0 \leq x < 1$, $f(x) = x$. On $1 \leq x < 2$, $f(x) = x - 2(x-1) = 2 - x$. From 2 on, $f(x) = 2 - x + (x-2) = 0$. The graph is a tent rising from 0 to a peak of 1 at $x = 1$ and back to 0 at 2: three ReLUs build a shape none of them has alone.

Module 3: Exponentials and Logarithms

3.1 0, 4, −2, and 7. The logarithm and the exponential undo each other, and $0.01 = 10^{-2}$.

3.2 $\ln(6 \times 5/3) = \ln 10$.

3.3 $81 = 3^4$, so $x = 4$. Taking logs, $0.5t = \ln 20$, so $t = 2\ln 20 \approx 5.99$.

3.4 About 24, 12, and 8 years. The exact figures for yearly compounding are 23.4, 11.9, and 8.0.

3.5 Yearly: $5{,}000 \times 1.06^{10} \approx 8{,}954.24$. Monthly: $5{,}000 \times 1.005^{120} \approx 9{,}096.98$. Continuously: $5{,}000e^{0.6} \approx 9{,}110.59$. Compounding more often adds a little, never a lot.

3.6 24 days is three half-lives, leaving $(1/2)^3 = 1/8$. After 30 days, $(1/2)^{30/8} = 2^{-3.75} \approx 0.074$, about 7.4%.

3.7 $t = \ln 3 / \ln 1.08 \approx 1.0986 / 0.0770 \approx 14.3$ years.

3.8 Each month adds 2%: $1.02^{12} - 1 \approx 0.268$, an effective rate of about 26.8%.

3.9 The rule gives $72/4 = 18$ years; exactly, $\ln 2 / \ln 1.04 \approx 17.7$ years.

3.10 Solve $4^t = 1{,}000$: $t = \ln 1000 / \ln 4 \approx 6.908 / 1.386 \approx 5.0$ years.

3.11 Each doubling multiplies cost by 0.82: $100 \times 0.82^3 \approx 55.1$, so about $55 per kilowatt-hour.

3.12 $\lambda = \ln 2 / 6 \approx 0.116$ per hour. Twenty-four hours is four half-lives, leaving $(1/2)^4 = 1/16$, or 6.25%.

3.13 One year of continuous growth multiplies by e^r, and one year at R multiplies by $1 + R$. They agree when $1 + R = e^r$, so $R = e^r - 1$. For 5%, $R = e^{0.05} - 1 \approx 5.13\%$.

3.14 $\ln(1 + r) \approx r(1 - r/2)$, so $t_2 = \ln 2 / \ln(1 + r) \approx (0.693/r)/(1 - r/2)$. For small r, dividing by $1 - r/2$ is close to multiplying by $1 + r/2$. In percent the constant is $69.3(1 + r/2)$, which at $r = 0.08$ is $69.3 \times 1.04 \approx 72.1$.

3.15 It grew a hundredfold in 8 years, so its yearly factor is $100^{1/8} = 10^{0.25} \approx 1.778$, a growth rate of about 78% a year.

3.16 Take logs: $(x + 1)\ln 2 = x \ln 3$, so $x(\ln 3 - \ln 2) = \ln 2$ and $x = \ln 2 / \ln(3/2) \approx 0.6931 / 0.4055 \approx 1.71$.

Module 4: Rates of Change

4.1 $f(4) = 28$ and $f(1) = 4$, so the average rate is $(28 - 4)/(4 - 1) = 8$.

4.2 $[5(x + h) - 2 - (5x - 2)]/h = 5h/h = 5$, so $f'(x) = 5$: a line's derivative is its slope.

4.3 $[3(x + h)^2 - 3x^2]/h = (6xh + 3h^2)/h = 6x + 3h$, which approaches $6x$ as h shrinks.

4.4 $(3.1^2 - 9)/0.1 = 0.61/0.1 = 6.1$ and $(3.01^2 - 9)/0.01 = 0.0601/0.01 = 6.01$. The quotients approach $f'(3) = 6$.

4.5 Dollars per ton, and megawatt-hours per hour, which are megawatts.

4.6 Rising means f' is positive; getting steeper means the slope is increasing, so f'' is positive too.

4.7 $(180 - 380)/(8 - 6) = -100$ MWh per hour: it discharged at an average of 100 MW.

4.8 By the quadratic result, $C'(q) = 30 + 0.04q$, which is $50 at $q = 500$. Average cost is $C(500)/500 = 28{,}000/500 = 56$ dollars. Marginal cost is below average, so making more units still pulls the average down.

4.9 $R(p) = 600p - 12p^2$, so $R'(p) = 600 - 24p$. At $p = 20$, $R'(20) = 120$: a one-dollar rise adds about $120, so revenue rises.

4.10 $C'(L) = 12 + 0.005L$, which at $L = 8{,}000$ is $12 + 40 = 52$ dollars per megawatt-hour.

4.11 $(566 - 410)/2 = 78$ MW.

4.12 Its rate of change is $4 a day. Its relative rate is $4/80 = 0.05$, or 5% a day.

4.13 The top becomes $(x + h) - x = h$:

$$\frac{\sqrt{x+h} - \sqrt{x}}{h} \cdot \frac{\sqrt{x+h} + \sqrt{x}}{\sqrt{x+h} + \sqrt{x}} = \frac{h}{h\left(\sqrt{x+h} + \sqrt{x}\right)} = \frac{1}{\sqrt{x+h} + \sqrt{x}}$$

As h shrinks, this approaches $1/(2\sqrt{x})$.

4.14 For positive h, $(h-0)/h = 1$. For negative h, $(0-0)/h = 0$. The quotient approaches 1 from one side and 0 from the other, so there is no single limit and no derivative.

4.15 One-sided: $(9.261 - 8)/0.1 = 12.61$, off by 0.61. Central: $(9.261 - 6.859)/0.2 = 12.01$, off by only 0.01.

4.16 By the quadratic result, $V'(t) = -40 + t$ liters per minute. At $t = 10$ it is −30: draining at 30 liters a minute. The rate reaches zero at $t = 40$ minutes.

Module 5: The Rules of Differentiation

5.1 $5x^4 - 8x$. The constant contributes nothing.

5.2 Rewrite as $6x^{1/2} + 2x^{-3}$. The derivative is $3x^{-1/2} - 6x^{-4} = 3/\sqrt{x} - 6/x^4$.

5.3 $6e^{2t} - 5/t$.

5.4 By the product rule, $2x \cdot e^x + x^2 \cdot e^x = xe^x(2 + x)$.

5.5 By the product rule, $3x^2 \ln x + x^3 \cdot (1/x) = x^2(3 \ln x + 1)$.

5.6 By the quotient rule, $[2(x-3) - (2x+1) \cdot 1]/(x-3)^2 = -7/(x-3)^2$.

5.7 By the chain rule, $6(4x-1)^5 \cdot 4 = 24(4x-1)^5$.

5.8 The inside is $-x^2$, with derivative $-2x$, so the result is $-2xe^{-x^2}$.

5.9 $(1/(1+3x)) \cdot 3 = 3/(1+3x)$.

5.10 $(1/2)(5+x^2)^{-1/2} \cdot 2x = x/\sqrt{(5+x^2)}$.

5.11 $P'(y) = -20{,}000(1+y)^{-21}$, so $P'(0.05) = -20{,}000/1.05^{21} \approx -7{,}179$. A rise of 0.001 changes the price by about $-7{,}179 \times 0.001 \approx -7.18$ dollars, from about $376.89. The exact fall is $7.11; the estimate is close because the step is small.

5.12 $A'(t) = 400e^{0.04t}$, so $A'(10) = 400e^{0.4} \approx 596.73$ dollars a year: 4% of the balance at that time, $14,918.

5.13 $dq/dp = -15$. At $p = 30$, $q = 750$ and $\varepsilon = (30/750)(-15) = -0.6$: inelastic, so revenue rises with price. At $p = 50$, $q = 450$ and $\varepsilon = (50/450)(-15) \approx -1.67$: elastic, so revenue falls.

5.14 $A(q) = 2{,}000/q + 10 + 0.05q$, so $A'(q) = -2{,}000/q^2 + 0.05$. Setting it to zero gives $q^2 = 40{,}000$, so $q = 200$. There marginal cost, $C'(200) = 10 + 0.1 \times 200 = 30$, equals average cost, $A(200) = 10 + 10 + 10 = 30$, as the quotient rule predicts.

5.15 The last link is $dz/db = 1$ instead of x, so $dL/db = 2(a-y) \cdot \sigma'(z) \cdot 1 \approx -0.538 \times 0.197 \approx -0.106$.

5.16 The log derivative is $Q'(n)/Q(n) = -0.0001$: capacity falls by 0.01% of its current value each cycle. After 3,000 cycles, $e^{-0.3} \approx 0.741$, so about 74% remains.

5.17 The chain rule gives $(g^{-1})' = -g^{-2}g'$. Then $(f \cdot g^{-1})' = f'g^{-1} - fg^{-2}g' = (f'g - fg')/g^2$.

5.18 The exponent $x \ln a$ has derivative $\ln a$, a constant, so the chain rule gives $e^{x \ln a} \cdot \ln a = a^x \ln a$.

5.19 By the product rule, the exponent $x \ln x$ has derivative $\ln x + 1$. So the derivative is $e^{x \ln x}(\ln x + 1) = x^x(\ln x + 1)$.

5.20 Work from the output back: $da/dw_1 = \sigma'(w_2 h) \cdot w_2 \cdot \sigma'(w_1 x) \cdot x$. Each layer contributes its own local derivative, and their product is the signal backpropagation carries to the first weight.

Module 6: Optimization

6.1 $f'(x) = 2x - 10 = 0$ at $x = 5$, and $f''(x) = 2$ is positive, so it is a trough, with $f(5) = -22$.

6.2 $f'(x) = 3x^2 - 12 = 0$ at $x = 2$ and $x = -2$. $f''(x) = 6x$ is positive at 2, a trough with $f(2) = -16$, and negative at −2, a peak with $f(-2) = 16$.

6.3 $f'(x) = -4x + 8 = 0$ at $x = 2$, and f'' is negative, so the largest value is $f(2) = -8 + 16 + 1 = 9$.

6.4 $g'(x) = 1 - 9/x^2 = 0$ at $x = 3$, and $g''(x) = 18/x^3$ is positive there, so the smallest value is $g(3) = 6$.

6.5 The critical point is $x = 2$, where $f(2) = -4$. At the endpoints, $f(0) = 0$ and $f(5) = 5$. The smallest value is −4, at the critical point; the largest is 5, at an endpoint.

6.6 Profit is $\pi(p) = (p - 15)(1{,}000 - 20p)$, so $\pi'(p) = (1{,}000 - 20p) - 20(p - 15) = 1{,}300 - 40p$. It is zero at $p = 32.50$, where 350 units sell and profit is $17.50 \times 350 = 6{,}125$ dollars.

6.7 $R'(q) = 80 - 0.08q$ and $C'(q) = 20 + 0.02q$ are equal when $60 = 0.10q$, at $q = 600$. Revenue is $33,600 and cost $19,600, so profit is $14,000.

6.8 $N'(h) = 50e^{-h/3} - 18 = 0$ gives $h = 3\ln(50/18) \approx 3.1$ hours. Dearer energy capacity makes the best battery shorter.

6.9 $Q^* = \sqrt{(2 \times 4{,}800 \times 60/2.5)} = \sqrt{230{,}400} = 480$ units, so $4{,}800/480 = 10$ orders a year.

6.10 $\Sigma x_i y_i = 2.1 + 7.8 + 18.6 = 28.5$ and $\Sigma x_i^2 = 1 + 4 + 9 = 14$, so $w = 28.5/14 \approx 2.04$.

6.11 The side parallel to the building is $600 - 2x$, so the area is $A(x) = x(600 - 2x)$. $A'(x) = 600 - 4x = 0$ at $x = 150$: the yard is 150 by 300 meters, 45,000 square meters.

6.12 Ordering cost is $(D/Q)S$ and holding cost $(Q/2)H$. They are equal when $Q^2 = 2DS/H$, which is exactly the condition that defines Q^*.

6.13 $\pi(p) = (p - c)(400 - 8p)$, so $\pi'(p) = 400 - 8p - 8(p - c) = 400 - 16p + 8c$, which is zero at $p = 25 + c/2$. A $2 rise in cost raises the price by $1: the store passes on half of it.

6.14 By the chain rule, $E'(w) = -2\,\Sigma\, x_i(y_i - wx_i)$. Setting it to zero gives $\Sigma\, x_i\, y_i = w\,\Sigma\, x_i^2$, so $w = \Sigma\, x_i\, y_i / \Sigma\, x_i^2$. Then $E''(w) = 2\,\Sigma\, x_i^2$, positive whenever any x_i is not zero.

6.15 $h = 3\ln(50/6) \approx 6.4$ hours, half again as long as at $12.

6.16 At $Q^* \approx 949$ the cost is about $3,795; at about 1,186 boxes it is about $1{,}518 + 2{,}372 = 3{,}890$ dollars, only 2.5% more. In general the cost at k times Q^* is $(1/k + k)/2$ times the minimum, which stays close to 1 for k near 1: the cost curve is flat at its bottom.

Module 7: Integrals as Accumulation

7.1 An antiderivative is $x^2 + x$, so the integral is $(9 + 3) - 0 = 12$.

7.2 $\sqrt{x} = x^{1/2}$ has antiderivative $(2/3)x^{3/2}$, so the integral is $(2/3)(8 - 1) = 14/3 \approx 4.67$.

7.3 $[e^{2x}/2]_0^{\,1} = (e^2 - 1)/2 \approx 3.19$.

7.4 $\ln e - \ln 1 = 1 - 0 = 1$.

7.5 Substitute $u = 1 + x^3$, whose derivative is $3x^2$: the function is $2u^4$ times u', so an antiderivative is $(2/5)(1 + x^3)^5$. Check: differentiating gives $2(1 + x^3)^4 \cdot 3x^2$.

7.6 With $u = x^2$, whose derivative is $2x$, the integral is $(1/2)\int_0^4 e^u\,du = (e^4 - 1)/2 \approx 26.8$.

7.7 $[120t - 7.5t^2]_0^{\,4} = 480 - 120 = 360$ MWh.

7.8 With $h = 1$: $0/2 + 40 + 75 + 90 + 80 + 50 + 0/2 = 335$ MWh.

7.9 Revenue is $50 \times (70 + 110 + 150 + 90) = 21{,}000$ dollars; charging cost $230 \times 25 = 5{,}750$. The margin is $15,250.

7.10 $(1 - e^{-1.2})/0.06 \approx 11.65$, so about $11.65 million, against $20 million undiscounted.

7.11 $2.40/(0.08 - 0.03) = 48$ dollars; at 7%, $2.40/0.04 = 60$ dollars, a 25% jump for one point of discount rate.

7.12 The quantity sold solves $80 - 0.5q = 30$, so $q^* = 100$. Then $CS = \int_0^{100} (50 - 0.5q)\,dq = [50q - 0.25q^2]_0^{100} = 2{,}500$ dollars.

7.13 $\int_0^T c\,e^{-rt}\,dt = c[-e^{-rt}/r]_0^T = c(1 - e^{-rT})/r$. As T grows, e^{-rT} shrinks to zero, leaving c/r.

7.14 The integrand is $c\,e^{-(r-g)t}$, a decaying exponential when $r - g$ is positive, and its integral from 0 to infinity is $c/(r - g)$. When $g \geq r$ the payments' present values do not shrink, and the total grows without limit: the formula breaks down, as valuations do when growth is assumed to outrun discounting forever.

7.15 $\int_0^\infty t\,e^{-rt}\,dt = [-te^{-rt}/r]_0^\infty + (1/r)\int_0^\infty e^{-rt}\,dt = 0 + 1/r^2$. Dividing by $1/r$ gives $1/r$: 20 years at 5%. Module 9 calls this kind of weighted average a duration.

7.16 A curve that bends downward lies above the straight chord between any two of its points, so each trapezoid misses a sliver above it. A curve that bends upward lies below its chords, so each trapezoid includes a sliver too much.

Module 8: Differential Equations

8.1 $dy/dt = 15e^{3t} = 3 \times 5e^{3t} = 3y$, and $y(0) = 5e^0 = 5$.

8.2 $y(t) = 50e^{-0.2t}$, so $y(5) = 50e^{-1} \approx 18.39$.

8.3 $e^{10k} = 2$, so $k = \ln 2/10 \approx 0.0693$, a continuous rate of about 6.9% a year.

8.4 $T(t) = 20 + 60e^{-0.5t}$, so $T(2) = 20 + 60e^{-1} \approx 42.1$.

8.5 At $N = K/2 = 500$, where $dN/dt = 0.4 \times 500 \times (1 - 0.5) = 100$ per unit of time.

8.6 Each step multiplies by $1 + 0.5 = 1.5$: $y \approx 1.5$, then 2.25. The exact value is about 2.718; smaller steps would close the gap.

8.7 The gap of 20 degrees must halve: $20e^{-0.5t} = 10$, so $t = 2 \ln 2 \approx 1.4$ hours.

8.8 It stores $0.95 \times 50 \times 2 = 95$ MWh. Delivering 50 MW drains $50/0.95 \approx 52.6$ MWh an hour, so it runs $95/52.6 \approx 1.8$ hours and delivers about $50 \times 1.8 \approx 90$ MWh: a round trip of about 90%.

8.9 $400e^{-0.002 \times 12} = 400e^{-0.024} \approx 390.5$ MWh.

8.10 $s = 0.1$ when $e^{-0.5(t - t_0)} = 9$, so $t - t_0 = -2 \ln 9 \approx -4.4$; by symmetry $s = 0.9$ at $+4.4$. The climb takes about 8.8 years.

8.11 Each step is $w - 0.25 \times 2(w - 3) = w - 0.5(w - 3)$, halving the distance to 3: 1.5, then 2.25, then 2.625.

8.12 The derivative is $(1{,}000/0.06) \times 0.06e^{0.06t} = 1{,}000e^{0.06t}$, and $0.06A + 1{,}000 = 1{,}000(e^{0.06t} - 1) + 1{,}000 = 1{,}000e^{0.06t}$, so the two sides agree; also $A(0) = 0$. Then $A(20) = 16{,}666.67 \times (e^{1.2} - 1) \approx 38{,}669$ dollars.

8.13 Since T_{env} is constant, $du/dt = dT/dt = -ku$, so $u = u_0\,e^{-kt}$. Adding T_{env} back gives $T(t) = T_{\text{env}} + (T_0 - T_{\text{env}})e^{-kt}$.

8.14 Write $u = e^{-r(t-t_0)}$, so $N = K/(1+u)$ and $du/dt = -ru$. By the chain rule, $dN/dt = -K(1+u)^{-2}(-ru) = rKu/(1+u)^2$. Also $rN(1-N/K) = r(K/(1+u)) \times (u/(1+u)) = rKu/(1+u)^2$, the same.

8.15 $w_{n+1} - 3 = w_n - 3 - 2\alpha(w_n - 3) = (1-2\alpha)(w_n - 3)$. The error shrinks to zero when $|1-2\alpha| < 1$, which is $0 < \alpha < 1$.

8.16 Each step multiplies by $1 + rh = 1 + rt/n$, and n steps multiply by $(1 + rt/n)^n$. As n grows, this approaches e^{rt}, Module 3's limit: compounding ever more often is Euler's method with ever smaller steps, converging to the exact solution.

Module 9: Approximation

9.1 Near $a = 49$, $f(49) = 7$ and $f'(49) = 1/14$, so $\sqrt{49.5} \approx 7 + 0.5/14 \approx 7.0357$. The true value is 7.0356.

9.2 $1 + 0.05 + 0.00125 = 1.05125$; the true value is 1.05127.

9.3 One term gives 0.03; two give $0.03 - 0.00045 = 0.02955$. The true value is 0.029559.

9.4 $1 + 10 \times 0.02 = 1.2$. The quadratic term adds $(10 \times 9/2) \times 0.02^2 = 0.018$, giving 1.218. The true value is 1.21899.

9.5 $x_1 = 2 - (4-5)/4 = 2.25$; $x_2 = 2.25 - (5.0625 - 5)/4.5 \approx 2.236111$. The true root, $\sqrt{5}$, is 2.236068.

9.6 $-6 \times 0.005 = -3\%$, a fall of about $30. The DV01 is $6 \times 1{,}000 \times 0.0001 = 0.60$, so 60 cents per basis point.

9.7 The Macaulay duration equals the maturity, 8 years, so $D = 8/1.04 \approx 7.69$, and the price changes by about $-7.69 \times 0.0025 \approx -1.92\%$.

9.8 Duration alone: $-7.5 \times 0.02 = -15\%$. Convexity adds $0.5 \times 70 \times 0.02^2 = 1.4\%$, for an estimate of −13.6%. For a move this large, the convexity term matters.

9.9 $g'(r) = -2{,}420/(1+r)^3$. At 5%, $g \approx 97.51$ and $g' \approx -2{,}090.5$, so $r_1 \approx 0.05 + 0.0466 = 9.66\%$. A second step gives about 9.998%. Exactly, $(1+r)^2 = 1.21$, so $r = 10\%$.

9.10 $e^{-0.05} \approx 1 - 0.05 = 0.95$, so about 95%; exactly, $e^{-0.05} \approx 0.9512$.

9.11 With $n = -10$ and $x = 0.05$: $1 - 0.5 + (-10)(-11)/2 \times 0.0025 = 0.6375$, against the exact 0.6139. The step $nx = -0.5$ is not small, so the terms left out are not negligible.

9.12 Solve $P(y) - 950 = 0$: $y_1 = 0.04 - (1{,}000 - 950)/(-4{,}451.8) \approx 0.04 + 0.0112 = 5.12\%$. Further steps converge to about 5.16%.

9.13 $L(w - \alpha L'(w)) \approx L(w) + L'(w)(-\alpha L'(w)) = L(w) - \alpha L'(w)^2$. The change is negative whenever $L'(w)$ is not zero.

9.14 At zero, e^x and all its derivatives equal 1. The Taylor polynomial is $1 + 1 \cdot x + (1/2!)x^2 + (1/3!)x^3 = 1 + x + x^2/2 + x^3/6$.

9.15 $P'(y) = -TF(1+y)^{-T-1}$, so $-P'/P = T/(1+y)$. $P''(y) = T(T+1)F(1+y)^{-T-2}$, so $P''/P = T(T+1)/(1+y)^2$.

9.16 The value has slope 0 below the strike and 1 above it, and no derivative at the strike itself. Any tangent line built on one side misprices the other side completely. Option traders therefore track how the slope itself changes, the gamma of Module 10.

Module 10: Several Variables

10.1 $\partial f/\partial x = 6xy$ and $\partial f/\partial y = 3x^2 - 6y^2$.

10.2 With x held constant, the chain rule gives $x \cdot x\,e^{xy} = x^2\,e^{xy}$.

10.3 $\nabla f = (2x + y, x + 2y)$, which at (1, 2) is (4, 5).

10.4 $\partial f/\partial x = 2x - 4$ and $\partial f/\partial y = 2y + 6$ are both zero at (2, −3).

10.5 $\partial f/\partial x = 3x^2y^2$, whose derivative with respect to y is $6x^2y$. $\partial f/\partial y = 2x^3y$, whose derivative with respect to x is also $6x^2y$.

10.6 $f(2, 3) = 12$, $\partial f/\partial x = 2xy = 12$, and $\partial f/\partial y = x^2 = 4$. The estimate is $12 + 12 \times 0.1 + 4 \times (-0.05) = 13.0$; exactly, $f(2.1, 2.95) \approx 13.01$.

10.7 Power's elasticity is 3 with respect to speed and 1 with respect to density, so the change is about $3 \times 2\% - 1 \times 1\% = 5\%$.

10.8 $3.47 - 1 \times 1 + (-105) \times (-0.01) \approx 3.52$ million dollars. The exact value, $2(1 - e^{-1.05})/0.07 - 15$, is about $3.57 million.

10.9 Rise: $0.56 \times 3 + 0.5 \times 0.022 \times 9 \approx 1.78$. Fall: $-1.68 + 0.099 \approx -1.58$. Gamma helps the holder both ways.

10.10 The options gain about $100 \times 0.56 = 56$ dollars per dollar the stock rises, so selling 56 shares short offsets them.

10.11 The change in volatility is 0.03, so the price rises by about $27.6 \times 0.03 \approx 0.83$ dollars.

10.12 The gradient at (0, 0) is (−2, 16), so the first step reaches $(0.2, -1.6)$. There the gradient is (−1.6, 3.2), so the second step reaches $(0.36, -1.92)$.

10.13 Setting $\partial E/\partial a$ and $\partial E/\partial b$ to zero gives $3a + 6b = 14$ and $6a + 14b = 31$, using $\Sigma x_i = 6$, $\Sigma y_i = 14$, $\Sigma x_i^2 = 14$, and $\Sigma x_i y_i = 31$. Solving, $b = 1.5$ and $a = 5/3 \approx 1.67$.

10.14 Along a contour f does not change, so for a small step $(\Delta x, \Delta y)$ along it, $\Delta f \approx (\partial f/\partial x)\Delta x + (\partial f/\partial y)\Delta y = 0$. That sum is the dot product of the gradient with the step, so the two are at right angles.

10.15 $\partial P/\partial x = akx^{a-1}y^b$, so $(x/P)(\partial P/\partial x) = akx^a\,y^b/(kx^a\,y^b) = a$. By the same steps, the elasticity with respect to y is b.

10.16 The steps are $w_1 - 1 \to (1 - 2\alpha)(w_1 - 1)$ and $w_2 + 2 \to (1 - 8\alpha)(w_2 + 2)$. Both errors shrink only if $|1 - 2\alpha| < 1$ and $|1 - 8\alpha| < 1$, which together require $0 < \alpha < 0.25$. The steep direction, with curvature 8, fails first, so it caps the rate for every direction.

Answer Sheet

Answers to every module's Check Your Understanding quiz.

Module	Answers
Module 1 Quiz	1. D 2. B 3. C 4. A 5. C
Module 2 Quiz	1. D 2. B 3. A 4. D 5. A
Module 3 Quiz	1. C 2. B 3. B 4. D 5. A
Module 4 Quiz	1. C 2. B 3. C 4. A 5. D
Module 5 Quiz	1. D 2. B 3. A 4. C 5. B
Module 6 Quiz	1. A 2. D 3. C 4. B 5. D
Module 7 Quiz	1. C 2. A 3. B 4. A 5. D
Module 8 Quiz	1. C 2. D 3. C 4. B 5. A
Module 9 Quiz	1. B 2. A 3. C 4. D 5. D
Module 10 Quiz	1. B 2. A 3. C 4. D 5. A

Linear Algebra and Optimization

Figures current as of October 2026

The mathematics of many numbers at once, and of the best choice under limits, for adults who want to use it: vectors and matrices, systems of equations, least squares, eigenvalues, portfolio risk, constrained optimization, linear programming, gradient methods, and the arithmetic inside artificial intelligence (AI) models. Every idea is attached to something real, from a battery's dispatch schedule to a portfolio's risk to a neural network's layers, and every module ends with practice problems solved in full at the back.

MODULE 1 — Vectors

Goal: Handle many numbers as one object.

Vectors as lists and as arrows; adding, scaling, and linear combinations; the dot product as returns and revenue; length and distance; angles and cosine similarity, the arithmetic of AI search; projection.

MODULE 2 — Matrices

Goal: Use a matrix as a machine that transforms vectors.

The matrix–vector product as many dot products at once, from a fund of funds to a neural-network layer; multiplication as one step after another; transpose, identity, and inverse; checking shapes the way units are checked.

MODULE 3 — Systems of Linear Equations

Goal: Solve many equations in many unknowns, and know when they cannot be solved.

Elimination; one solution, none, or infinitely many, with the determinant as the quick test; replicating an option from a stock and a bond; the economy's input–output table; how power injected at one point splits across a meshed grid.

MODULE 4 — Least Squares

Goal: Fit a model to more data than it has parameters.

The best fit as a projection, and the normal equations; demand against temperature and weekday; a stock's beta and a factor model; overfitting, and the ridge penalty that controls it.

MODULE 5 — Eigenvalues and Eigenvectors

Goal: Find the directions a matrix only stretches.

Markov chains settling to a steady state, from customer churn to credit ratings to web search; growth rates of linked systems; principal components, and the three shapes that explain almost every move in the yield curve.

MODULE 6 — Covariance and Portfolio Risk

Goal: Measure how things move together, and what that does to risk.

Variance, covariance, and correlation from data; the covariance matrix and a portfolio's variance; why diversification works and where it fails; risk contributions and risk parity.

MODULE 7 — Optimization with Constraints

Goal: Find the best choice under limits, and put a price on each limit.

Lagrange multipliers as shadow prices; economic dispatch, where the multiplier is the price of power; the minimum-variance portfolio and the efficient frontier; inequality constraints.

MODULE 8 — Linear Programming

Goal: Solve the optimization problems that run power markets and supply chains.

Feasible regions, corners, and the simplex method in outline; duality; scheduling a battery against a day of prices; dispatch with a transmission limit, where one price splits into locational prices and congestion is a shadow price.

MODULE 9 — Convexity and Gradient Methods

Goal: Understand how large models are actually optimized.

Convex and nonconvex problems; the gradient and the matrix of second derivatives; why descent zigzags, explained by eigenvalues; momentum, stochastic gradient descent, and Adam.

MODULE 10 — Linear Algebra Inside AI

Goal: See the matrix arithmetic inside a modern AI model.

A network as matrix products and simple nonlinear steps; attention worked on a tiny example; counting the arithmetic, and why it runs on graphics processors; low rank and the singular value decomposition, from compression to fine-tuning.

How to use this course. Start with the diagnostic on the next page; it says which modules to work and which to skim. Each module teaches through worked examples, then shows the spreadsheet functions that do the same work at real size. It ends with a Memorize This box of the formulas, definitions, and ideas to commit to memory, practice problems in three tiers, and a five-question quiz. Vectors and matrices are set in bold, and every symbol is defined in words where it first appears in a module and again in its Memorize This box. Worked solutions to every practice problem are at the back, after all the Memorize This boxes gathered in one place; quiz answers are on the answer sheet. Work the small examples by hand and the large ones in a spreadsheet. The examples use illustrative numbers unless a source and date are given.

Where to Start

A twenty-problem diagnostic

Work these on paper before you start; a calculator is fine throughout. The four bands follow the order of the course, except the calculus band, which covers ideas the course uses rather than teaches. The answers are on the next page, with a guide to where to begin.

Band A — Algebra and equations (Modules 1 and 3)

1. Solve the pair of equations $x + y = 10$ and $x - y = 4$.

2. Solve the pair of equations $2x + 3y = 12$ and $x - y = 1$.

3. Expand $(a + b)^2$.

4. Evaluate $\sqrt{(3^2 + 4^2)}$.

5. Evaluate $\Sigma_{i=1}{}^{4}\, i^2$, the sum of i^2 as i runs from 1 to 4.

Band B — Vectors and matrices (Modules 1 and 2)

6. Find the dot product of (1, 2, 3) and (4, 0, −1): multiply matching entries and add.

7. Find the length of the vector (2, −1, 2).

8. Multiply the matrix with rows (1, 2) and (3, 4) by the vector (1, 1).

9. Find the inverse of the matrix with rows (2, 0) and (0, 4).

10. Are the vectors (1, 2) and (−2, 1) perpendicular?

Band C — Calculus the course uses (Modules 7 and 9)

11. Differentiate $f(x) = 3x^2 - 5x$.

12. Find the value of x that minimizes $x^2 - 6x + 10$.

13. Differentiate e^{-2t} with respect to t.

14. For $f(x, y) = x^2 + 3xy$, find the partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$.

15. Find the gradient of $x^2 + y^2$, the vector of its two partial derivatives, at the point (1, 2).

Band D — Averages and spread (Modules 4 and 6)

16. Find the average of 4, 8, 6, and 2.

17. Find the variance of 2, 4, and 6: the average of their squared distances from their mean.

18. A portfolio holds 70% stocks that returned 10% and 30% bonds that returned 4%. What did it return?

19. Two series always move in opposite directions, in exact proportion. What is their correlation?

20. A stock's yearly returns have a standard deviation of 20%. What is their variance?

Answers, and Where to Begin

#	Answer	Where
1	$x = 7$ and $y = 3$	Module 3
2	$x = 3$ and $y = 2$	Module 3
3	$a^2 + 2ab + b^2$	Module 1
4	5	Module 1
5	30	Module 1
6	1	Module 1
7	3	Module 1
8	(3, 7)	Module 2
9	The matrix with rows (0.5, 0) and (0, 0.25)	Module 2
10	Yes: their dot product is 0	Module 1
11	$6x - 5$	Used from Module 7
12	$x = 3$	Used from Module 7
13	$-2e^{-2t}$	Used from Module 7
14	$2x + 3y$ and $3x$	Used from Module 7
15	(2, 4)	Module 7
16	5	Module 4
17	$8/3 \approx 2.67$	Module 6
18	8.2%	Module 6
19	−1	Module 6
20	0.04, which is 0.2 squared	Module 6

Score each band out of five, then use the table. A low score on a band means working every module the band covers.

Band score	What to do
5 of 5	Skim the modules the band covers: read each Memorize This box and try the challenge problems.
3 or 4	Read those modules quickly and work the applied problems.
0 to 2	Work those modules through, worked examples and all.

The calculus band is the one this course does not teach from scratch. If it goes badly, review derivatives and partial derivatives in any introductory calculus text before Module 7; Modules 1 to 6 need no calculus, apart from challenge problems 1.15 and 4.13. If the first two bands go badly, work Modules 1 to 3 slowly, since everything after them is built from vectors, matrices, and equations.

Module 1 — Vectors

Lists of numbers, the dot product, length, angle, and projection

Recommended Reading / Reference

Book: Gilbert Strang, *Introduction to Linear Algebra* (6th ed., 2023) — the standard applied text, by the mathematician whose recorded lectures on the subject are among the most watched anywhere

Reference: Grant Sanderson, *Essence of Linear Algebra* (3Blue1Brown, 2016) — free animated videos that show vectors and matrices as geometry

1a. What a Vector Is

Linear algebra is the mathematics of many numbers handled together. Its basic object is the vector, an ordered list of numbers. A portfolio split among three assets is a vector of three weights. A day of hourly electricity prices is a vector of 24 prices. Order matters: the third entry of a price vector is always the price for the hour from 2 to 3 a.m.

This course writes vectors in bold lowercase letters, such as $\mathbf{w}$ or $\mathbf{p}$. The numbers inside a vector are its components, written in ordinary italic with a subscript that gives their position: w_1 is the first component of $\mathbf{w}$, and w_i is the ith. A vector with n components has dimension n. Written out, a vector is a column of numbers. To save space, the text often writes it as a row in parentheses, so the portfolio with half its money in stocks, 30% in bonds, and 20% in gold is

$$\mathbf{w} = \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} = (0.5,\ 0.3,\ 0.2)$$

A vector with two components can be drawn as an arrow from the origin to the point with those coordinates: (3, 1) points three units right and one up. Arrows give the intuition, and the arithmetic works the same way in 24 dimensions or in 3,000, where no one can draw. The zero vector, $\mathbf{0}$, has every component equal to zero.

Artificial intelligence (AI) systems handle words the same way. A text-embedding model turns a passage of text into a vector, chosen during training so that passages with similar meanings get similar vectors. OpenAI's text-embedding-3-small model, for example, turns any passage into a vector of 1,536 numbers. Section 1e shows how a search engine compares them.

1b. Adding and Scaling

Vectors with the same number of components add component by component. Multiplying a vector by an ordinary number, called a scalar, multiplies every component by it. For vectors $\mathbf{a}$ and $\mathbf{b}$ with n components each and a scalar c:

$$\mathbf{a} + \mathbf{b} = (a_1 + b_1,\ a_2 + b_2,\ \ldots,\ a_n + b_n)$$

$$c\,\mathbf{a} = (c\,a_1,\ c\,a_2,\ \ldots,\ c\,a_n)$$

Drawn as arrows, addition places the second arrow at the tip of the first, and the sum runs from the start of one to the end of the other. Scaling stretches an arrow or shrinks it, and a negative scalar also turns it around.

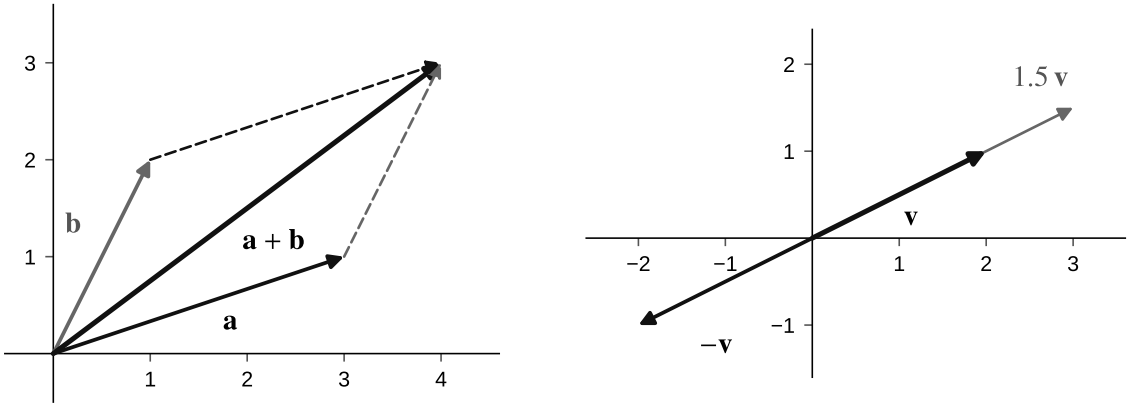


Figure 1.1. Left: adding $\mathbf{a} = (3, 1)$ and $\mathbf{b} = (1, 2)$ tip to tail gives $\mathbf{a} + \mathbf{b} = (4, 3)$, the diagonal of the parallelogram the two arrows make. Right: scaling $\mathbf{v} = (2, 1)$ by 1.5 stretches it, and scaling by −1 turns it around.

Adding scaled vectors gives a linear combination, such as $c_1\mathbf{v}_1 + c_2\mathbf{v}_2$ for vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ and scalars c_1 and c_2. It is the most important construction in this course. A fund of funds holds a linear combination of its funds' holdings. A grid's net load is demand minus its wind and solar output, which is a linear combination too. A neural network's layer computes thousands of linear combinations at once.

Worked example — Blending two funds

Fund A holds stocks, bonds, and gold in the proportions $\mathbf{a} = (0.8, 0.2, 0)$, and fund B in the proportions $\mathbf{b} = (0.2, 0.6, 0.2)$. An investor puts 60% of her money in A and 40% in B, so her holdings are the linear combination

$$\begin{aligned} 0.6\,\mathbf{a} + 0.4\,\mathbf{b} &= (0.48,\ 0.12,\ 0) + (0.08,\ 0.24,\ 0.08) \\ &= (0.56,\ 0.36,\ 0.08) \end{aligned}$$

She holds 56% stocks, 36% bonds, and 8% gold. The proportions still add to 1, as they must when each fund is fully invested and the blend's own weights add to 1.

1c. The Dot Product

The dot product of two vectors with the same number of components multiplies matching components and adds the results. The answer is a single number, not a vector:

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=1}^{n} a_i\, b_i$$

The Greek capital sigma, Σ, means “add up,” here over the components as i runs from 1 to n. The dot product is the workhorse of linear algebra, because so many quantities are sums of matching products.

A portfolio's return is one. With weights $\mathbf{w} = (0.5, 0.3, 0.2)$ in stocks, bonds, and gold, and a year's returns $\mathbf{r} = (0.10, 0.04, -0.02)$ on those three assets, the portfolio returned

$$\begin{aligned} \mathbf{w} \cdot \mathbf{r} &= 0.5\,(0.10) + 0.3\,(0.04) + 0.2\,(-0.02) \\ &= 0.050 + 0.012 - 0.004 = 0.058 \end{aligned}$$

or 5.8%. Each term is one asset's contribution: stocks added 5.0 percentage points, bonds 1.2, and gold took away 0.4.

Worked example — A battery's day in one number

Split a day into six four-hour blocks. A battery rated at 25 megawatts (MW) of power and 100 megawatt-hours (MWh) of energy charges 60 MWh in the first block and 50 MWh in the second. It discharges 100 MWh in the fifth block, from 4 to 8 p.m. Its schedule, in MWh sold, is $\mathbf{s} = (-60, -50, 0, 0, 100, 0)$, negative when it buys. The blocks' average prices, in dollars per MWh, are $\mathbf{p} = (28, 24, 35, 42, 95, 60)$, and the day's revenue is their dot product:

$$\begin{aligned}\mathbf{p} \cdot \mathbf{s} &= 28\,(-60) + 24\,(-50) + 95\,(100) \\ &= -1{,}680 - 1{,}200 + 9{,}500 = 6{,}620\end{aligned}$$

The battery earned $6,620. It bought 110 MWh to sell 100, a round trip of about 91%, and the dot product counts both sides of the trade at once.

The dot product is symmetric: $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$. It also distributes over addition: $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$ for a third vector $\mathbf{c}$ of the same size. The second rule is why a portfolio's return splits cleanly into one contribution per asset. An average is a dot product too: the average of n numbers is their dot product with a vector whose every component is $1/n$.

1d. Length and Distance

The length of a vector $\mathbf{x}$, also called its norm and written $\|\mathbf{x}\|$, is the square root of the vector's dot product with itself:

$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

In two dimensions this is Pythagoras' theorem: the arrow (3, 4) has length $\sqrt{(9 + 16)} = 5$. The formula keeps working in any number of dimensions. The distance between two vectors $\mathbf{a}$ and $\mathbf{b}$ is the length of their difference, $\|\mathbf{a} - \mathbf{b}\|$.

Worked example — Which past day looks most like today?

A trader keeps each day's prices as a six-block vector, in dollars per MWh. Today's is $\mathbf{t} = (30, 25, 40, 45, 90, 55)$. Day A's was $\mathbf{d}_A = (28, 24, 35, 42, 95, 60)$, and day B's was $\mathbf{d}_B = (40, 38, 45, 48, 70, 52)$. Today differs from day A by (2, 1, 5, 3, −5, −5) and from day B by (−10, −13, −5, −3, 20, 3), so the distances are

$$\|\mathbf{t} - \mathbf{d}_A\| = \sqrt{4 + 1 + 25 + 9 + 25 + 25} = \sqrt{89} \approx 9.4$$

$$\|\mathbf{t} - \mathbf{d}_B\| = \sqrt{100 + 169 + 25 + 9 + 400 + 9} = \sqrt{712} \approx 26.7$$

Day A is much closer. Forecasters use exactly this idea: find the past days whose vectors lie nearest today's, and see what came next on those days. With real data the vectors hold 24 hourly prices, or 288 five-minute ones, and the arithmetic is the same.

Dividing a nonzero vector by its length gives a unit vector, of length 1, pointing the same way: $\mathbf{x}/\|\mathbf{x}\|$. This separates a vector's direction from its size. A price profile and the same profile at one and a half times the level point in exactly the same direction; only their lengths differ.

1e. Angles and Cosine Similarity

The dot product also measures the angle between two vectors. For nonzero vectors $\mathbf{a}$ and $\mathbf{b}$ separated by an angle θ:

$$\cos\theta = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|}$$

The ratio is called the cosine similarity. It is 1 when the vectors point the same way, 0 when they are at right angles, and −1 when they point in opposite directions. Two vectors whose dot product is zero are perpendicular, also called orthogonal. The cosine ignores length entirely, so it compares shapes: today's price profile has a cosine similarity of 0.998 with day A's and 0.978 with day B's.

Worked example — Searching by meaning

Suppose an embedding model, far smaller than a real one, maps three words to vectors of three components: “battery” to $\mathbf{b} = (0.9, 0.1, 0.3)$, “storage” to $\mathbf{s} = (0.8, 0.2, 0.4)$, and “tariff” to $\mathbf{f} = (0.1, 0.9, 0.2)$. Then $\mathbf{b}\cdot\mathbf{s} = 0.86$, $\mathbf{b}\cdot\mathbf{f} = 0.24$, and the three lengths are about 0.954, 0.917, and 0.927. With θ_{bs} the angle between “battery” and “storage” and θ_{bf} the angle between “battery” and “tariff”:

$$\cos\theta_{bs} = \frac{0.86}{0.954 \times 0.917} \approx 0.98$$

$$\cos\theta_{bf} = \frac{0.24}{0.954 \times 0.927} \approx 0.27$$

“Battery” and “storage” point nearly the same way, about 10 degrees apart; “battery” and “tariff” are about 74 degrees apart. A search engine built on embeddings ranks passages by exactly this number. It turns the query into a vector and returns the passages whose vectors have the highest cosine similarity to it. Real embeddings have hundreds or thousands of components, but the arithmetic is the same.

Cosine similarity returns in Module 6 under another name. When two series of returns are first measured from their own averages, the cosine of the angle between them is their correlation.

1f. Projection

Often one vector must be split into a part that runs along another vector and a part perpendicular to it. The projection of $\mathbf{b}$ onto $\mathbf{a}$ is the multiple of $\mathbf{a}$ closest to $\mathbf{b}$, and the residual $\mathbf{e}$ is what is left over:

$$\text{proj}_{\mathbf{a}}\,\mathbf{b} = \frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}\cdot\mathbf{a}}\,\mathbf{a} \qquad \mathbf{e} = \mathbf{b} - \text{proj}_{\mathbf{a}}\,\mathbf{b}$$

The residual is perpendicular to $\mathbf{a}$: $\mathbf{a}\cdot\mathbf{e} = 0$. Among all the multiples of $\mathbf{a}$, the projection is the one at the shortest distance from $\mathbf{b}$, as Figure 1.2 shows. That makes projection the geometry of every best fit in this course.

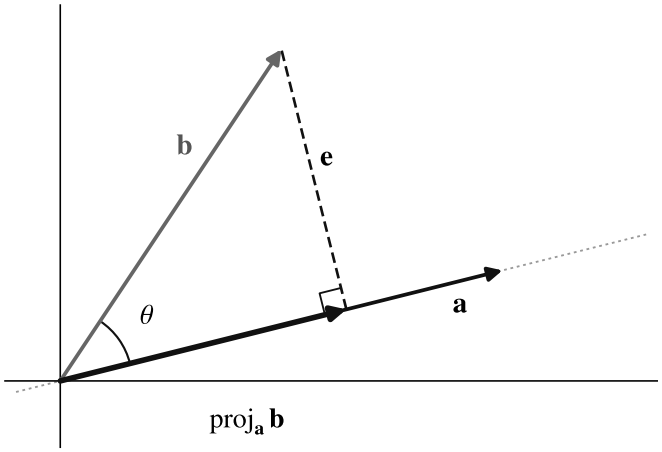


Figure 1.2. Projecting $\mathbf{b} = (2, 3)$ onto $\mathbf{a} = (4, 1)$. The projection, about (2.59, 0.65), is the multiple of $\mathbf{a}$ nearest $\mathbf{b}$, and the residual $\mathbf{e}$ meets the line through $\mathbf{a}$ at a right angle. The angle θ between $\mathbf{a}$ and $\mathbf{b}$ is about 42 degrees.

Worked example — How much of a stock's move is the market?

Over four months, a market index returned $\mathbf{m} = (2, -1, 3, -2)$ and a stock returned $\mathbf{y} = (3, -2, 4, -1)$, in percent. Then $\mathbf{m} \cdot \mathbf{y} = 6 + 2 + 12 + 2 = 22$ and $\mathbf{m} \cdot \mathbf{m} = 4 + 1 + 9 + 4 = 18$, so the projection's coefficient is $22/18 \approx 1.22$:

$$\text{proj}_{\mathbf{m}}\,\mathbf{y} \approx 1.22\,\mathbf{m} \approx (2.44,\ -1.22,\ 3.67,\ -2.44)$$

$$\mathbf{e} = \mathbf{y} - \text{proj}_{\mathbf{m}}\,\mathbf{y} \approx (0.56,\ -0.78,\ 0.33,\ 1.44)$$

The coefficient, 1.22, is a simple version of the stock's beta: the stock tends to move about 1.22% for each 1% the market moves. The residual is the part of the stock's moves that the market does not explain. Its dot product with $\mathbf{m}$ is exactly zero before rounding, as it must be. Module 4 adds an intercept and explains why a real estimate needs years of data, but the idea is this projection.

In a Spreadsheet

Put each vector in a column, say $\mathbf{a}$ in cells A2 to A7 and $\mathbf{b}$ in B2 to B7. Four formulas then do this module's arithmetic for vectors of any size:

To compute	Formula
Dot product $\mathbf{a} \cdot \mathbf{b}$	=SUMPRODUCT(A2:A7, B2:B7)
Length $\|\|\mathbf{a}\|\|$	=SQRT(SUMSQ(A2:A7))
Cosine similarity	=SUMPRODUCT(A2:A7, B2:B7)/SQRT(SUMSQ(A2:A7)*SUMSQ(B2:B7))
Projection coefficient, $(\mathbf{a} \cdot \mathbf{b})/(\mathbf{a} \cdot \mathbf{a})$	=SUMPRODUCT(A2:A7, B2:B7)/SUMSQ(A2:A7)

1g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Vector	An ordered list of numbers, set in bold: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ x_i is the ith component, and n, the number of components, is the dimension
Adding and scaling	$\mathbf{a} + \mathbf{b} = (a_1 + b_1, \ldots, a_n + b_n) \qquad c\,\mathbf{a} = (c\,a_1, \ldots, c\,a_n)$ c is a scalar, an ordinary number; a sum of scaled vectors, such as $c_1\mathbf{v}_1 + c_2\mathbf{v}_2$, is a linear combination
Dot product	$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i\, b_i$ Multiply matching components and add; a portfolio's return is $\mathbf{w} \cdot \mathbf{r}$, its weights $\mathbf{w}$ times its assets' returns $\mathbf{r}$
Length and distance	$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} \qquad \text{distance} = \|\mathbf{a} - \mathbf{b}\|$ $\|\mathbf{x}\|$ is the norm; $\mathbf{x}/\|\mathbf{x}\|$ is the unit vector pointing the same way
Angle	$\cos\theta = \dfrac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|}$ θ is the angle between nonzero vectors $\mathbf{a}$ and $\mathbf{b}$; the ratio is the cosine similarity, from −1 to 1
Perpendicular	$\mathbf{a} \cdot \mathbf{b} = 0$: the vectors are at right angles, also called orthogonal
Projection	$\text{proj}_{\mathbf{a}}\,\mathbf{b} = \dfrac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{a} \cdot \mathbf{a}}\,\mathbf{a}$ The multiple of $\mathbf{a}$ closest to $\mathbf{b}$; the residual, $\mathbf{b}$ minus its projection, is perpendicular to $\mathbf{a}$
The idea	A vector packs many numbers into one object. One operation, the dot product, computes returns, revenues, lengths, angles, and projections.

1h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

1.1 Compute (2, −1, 3) + 2(1, 4, 0).

1.2 Compute the dot product of (1, 3, −2) and (4, −1, 5).

1.3 Find the length of (1, 2, 2) and the unit vector in its direction.

1.4 Find the distance between the points (1, 5) and (4, 1).

1.5 Find the cosine of the angle between (1, 0) and (1, 1), and the angle itself.

1.6 Find the number c that makes (2, c) perpendicular to (3, −6).

1.7 Project (3, 4) onto (1, 0), and find the residual.

Applied

1.8 A portfolio holds four funds with weights (0.4, 0.3, 0.2, 0.1), and the funds returned 12%, 5%, −3%, and 8%. What did the portfolio return, and which fund contributed most?

1.9 A battery's six-block schedule, in MWh sold, is $\mathbf{s} = (-50, -60, 0, 0, 60, 40)$, and the blocks' prices are $\mathbf{p} = (30, 22, 38, 45, 110, 70)$ dollars per MWh. What did it earn, and what was its round trip?

1.10 Three trades in a day were 200 MWh at \$40, 500 MWh at \$44, and 300 MWh at \$50. Find the volume-weighted average price (VWAP), the average price with each trade weighted by its share of the volume.

1.11 Today's six-block demand profile, in gigawatts, is (22, 20, 28, 31, 35, 27). Day A's was (21, 19, 27, 32, 37, 28) and day B's (24, 22, 26, 27, 30, 26). Which day is nearer to today?

1.12 A search engine represents each document by how often it uses three words: battery, grid, and tariff. Document 1 is (3, 2, 0), document 2 is (2, 2, 1), and the query is (1, 1, 0). Rank the documents by cosine similarity to the query.

1.13 Over four months a market returned (1, −2, 2, 1) and a fund returned (2, −3, 3, 0), in percent. Find the projection coefficient of the fund's returns onto the market's, and check that the residual is perpendicular to the market.

Challenge

1.14 Show that $\|\mathbf{a} + \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + 2\mathbf{a} \cdot \mathbf{b} + \|\mathbf{b}\|^2$. What does it become when $\mathbf{a}$ and $\mathbf{b}$ are perpendicular?

1.15 For a nonzero vector $\mathbf{a}$, the squared distance from $\mathbf{b}$ to a multiple $c\mathbf{a}$ is $D(c) = \|\mathbf{b} - c\mathbf{a}\|^2$. Find the c that minimizes it, and show that the residual is then perpendicular to $\mathbf{a}$.

1.16 Use the previous problem to show that $(\mathbf{a} \cdot \mathbf{b})^2 \le \|\mathbf{a}\|^2 \|\mathbf{b}\|^2$ for nonzero vectors, so the cosine similarity always lies between −1 and 1.

1.17 Many embedding models scale every vector to length 1 before storing it. Show that cosine similarity then equals the plain dot product, and explain why that saves work when one query is compared with a million stored passages.

Module 1 Quiz

Check Your Understanding

1. What is the dot product of (1, 2, 3) and (2, 0, −1)?

A) (2, 0, −3), a vector of the matching products
B) −1
C) 1
D) About 8.4, the product of their lengths

2. What is the length of the vector (2, 3, 6)?

A) 7, the square root of 4 + 9 + 36
B) 11
C) 49
D) About 3.3, the square root of 2 + 3 + 6

3. Two nonzero vectors have a dot product of zero. What does that tell you?

A) One of them must be the zero vector
B) They point in opposite directions
C) They are perpendicular
D) Wherever one has a nonzero component, the other has a zero, so the two vectors never share a nonzero position

4. Why is the projection of one vector onto another the best fit of the first by a multiple of the second?

A) Because the projection always has the same length as the vector being fitted.
B) Because the projection points in the direction of the vector being fitted.
C) Because the residual is zero whenever the two vectors are not parallel.
D) Because the leftover part is perpendicular to the direction, so no other multiple can come closer to the vector being fitted.

5. A trader compares today's six-block price vector $\mathbf{t}$ with a past day's $\mathbf{d}$. What does the distance $\mathbf{t - d}$ measure?

A) It measures how different the two days' prices were.
B) It measures the average price difference between the two days, block by block, in dollars.
C) It measures the angle between the two price profiles, ignoring their levels entirely.
D) It measures how much a battery would have earned by trading between the two days.

Module 2 — Matrices

Tables of numbers that transform vectors

Recommended Reading / Reference

Book: Stephen Boyd and Lieven Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares* (2018) — built around applications, and free online from its authors

Reference: Microsoft, Excel function reference for MMULT, MINVERSE, and TRANSPOSE (support.microsoft.com) — how a spreadsheet multiplies and inverts blocks of cells

2a. What a Matrix Is

A matrix is a rectangular table of numbers. One with m rows and n columns is an $m \times n$ matrix, read "m by n." This course writes matrices in bold capital letters, such as $\mathbf{A}$, and the entry in row i and column j as a_{ij}. A vector with n components is a matrix with one column, an $n \times 1$ matrix.

Tables like this are everywhere. A week of hourly power prices is a 7 × 24 matrix, one row per day. A fund family's holdings form a matrix with a row for each asset and a column for each fund. A grayscale photograph is a matrix of pixel brightnesses. The weights of one layer of a neural network form a matrix, and in a large model a single such matrix can hold tens of millions of numbers.

A matrix can be read by rows or by columns, and each row and each column is a vector. The holdings matrix $\mathbf{H}$ below has a row for each asset, stocks, bonds, and gold, and a column for each of the two funds that Module 1 blended. Column j is fund j's mix of assets; row i is how much of asset i each fund holds:

$$\mathbf{H} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.6 \\ 0 & 0.2 \end{bmatrix}$$

2b. Matrix Times Vector

A matrix times a vector gives a vector. For an $m \times n$ matrix $\mathbf{A}$ and a vector $\mathbf{x}$ with n components, the product $\mathbf{Ax}$ has m components, and component i is the dot product of row i of $\mathbf{A}$ with $\mathbf{x}$:

$$(\mathbf{Ax})_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$$

The same product has a second reading, which is often the more useful one. With $\mathbf{a}_j$ standing for column j of $\mathbf{A}$, the product is a linear combination of the columns, and the components of $\mathbf{x}$ are the weights:

$$\mathbf{Ax} = x_1\,\mathbf{a}_1 + x_2\,\mathbf{a}_2 + \cdots + x_n\,\mathbf{a}_n$$

Rows give the arithmetic; columns give the meaning.

Worked example — A fund of funds

An investor holds 60% of her money in the first fund and 40% in the second, so her allocation is $\mathbf{x} = (0.6, 0.4)$. Her exposure to each asset is $\mathbf{Hx}$:

$$\mathbf{Hx} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.6 \\ 0 & 0.2 \end{bmatrix} \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix} = \begin{bmatrix} 0.8\,(0.6) + 0.2\,(0.4) \\ 0.2\,(0.6) + 0.6\,(0.4) \\ 0\,(0.6) + 0.2\,(0.4) \end{bmatrix} = \begin{bmatrix} 0.56 \\ 0.36 \\ 0.08 \end{bmatrix}$$

This is Module 1's blend again, now as one matrix product. The row reading computes each asset's share. The column reading says the portfolio is 0.6 of the first fund's column plus 0.4 of the second's.

A square matrix can also be pictured as a transformation of the plane, or of space. Since $\mathbf{Ax}$ is a combination of columns, the matrix sends (1, 0) to its first column and (0, 1) to its second, and every other vector goes wherever the same combination of columns takes it. Straight lines stay straight and the origin stays put, which is what makes the transformation linear.

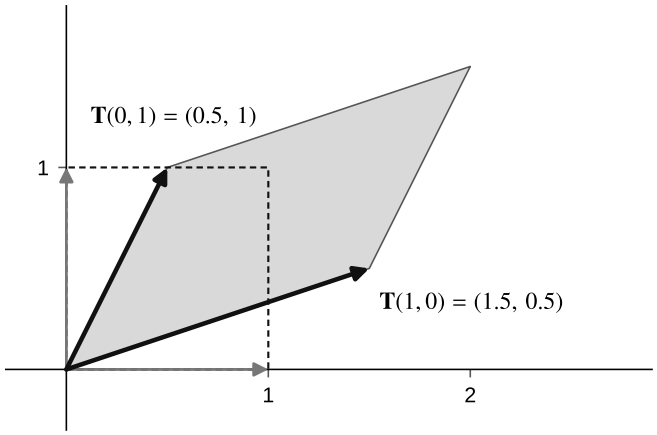


Figure 2.1. The matrix $\mathbf{T}$ with columns (1.5, 0.5) and (0.5, 1) sends (1, 0) to its first column and (0, 1) to its second. The dashed unit square becomes the shaded parallelogram, whose area, 1.25, Module 3 calls the determinant.

Worked example — One layer of a neural network

A layer of a neural network takes an input vector $\mathbf{x}$, multiplies it by a weight matrix $\mathbf{W}$, adds a bias vector $\mathbf{b}$, and passes each component of the result through a simple function. The simplest, long the most common, is the rectified linear unit, or ReLU, which keeps positive numbers and replaces negative ones with zero. With two inputs and three outputs:

$$\mathbf{Wx} + \mathbf{b} = \begin{bmatrix} 0.5 & -1 \\ 1 & 0.5 \\ -0.5 & 0.25 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.1 \\ -0.2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 2.5 \\ -0.75 \end{bmatrix} + \begin{bmatrix} 0.1 \\ -0.2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 2.3 \\ -0.75 \end{bmatrix}$$

The ReLU turns (0.1, 2.3, −0.75) into (0.1, 2.3, 0), the layer's output. Training a network means adjusting the entries of $\mathbf{W}$ and $\mathbf{b}$, the network's parameters, of which this layer has $6 + 3 = 9$. A large model's layers do essentially this, with matrices thousands of rows and columns wide.

2c. Matrix Times Matrix

Two matrices can be multiplied when the first has as many columns as the second has rows. If $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$, their product $\mathbf{AB}$ is $m \times p$. Its entry in row i and column j is the dot product of row i of $\mathbf{A}$ with column j of $\mathbf{B}$, summing over the shared index k:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}\,b_{kj}$$

Equivalently, each column of $\mathbf{AB}$ is $\mathbf{A}$ times the matching column of $\mathbf{B}$. Shapes work like units: the inner sizes must match, and they cancel, while the outer sizes give the answer's shape. A 3 × 2 matrix times a 2 × 4 matrix is 3 × 4. A mismatch is the matrix version of adding dollars to megawatts, and checking shapes catches errors the same way.

Worked example — From orders to materials

A plant makes three products. The matrix $\mathbf{M}$ gives the kilograms of steel and of copper, its two rows, used in one unit of each product, its three columns. Customers place two kinds of standing order, and $\mathbf{Q}$ lists the units of each product, its rows, in each kind of order, its columns:

$$\mathbf{M} = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 4 & 1 \end{bmatrix} \qquad \mathbf{Q} = \begin{bmatrix} 10 & 0 \\ 5 & 20 \\ 0 & 8 \end{bmatrix} \qquad \mathbf{MQ} = \begin{bmatrix} 25 & 44 \\ 20 & 88 \end{bmatrix}$$

The product, a 2 × 2 matrix, gives the materials each kind of order uses. The first takes 25 kilograms of steel and 20 of copper, and the second 44 of steel and 88 of copper. The products, the shared dimension, have cancelled out.

Multiplying matrices means doing one transformation after another: $\mathbf{ABx}$ applies $\mathbf{B}$ to $\mathbf{x}$ first, then $\mathbf{A}$. Order matters, and in general $\mathbf{AB}$ and $\mathbf{BA}$ differ: stretching the plane sideways and then turning it a quarter turn is not the same as turning first and then stretching. Grouping does not matter: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ for any three matrices whose shapes fit.

That last fact explains something basic about neural networks. Two layers that only multiply by matrices, with weight matrices $\mathbf{W}_1$ and then $\mathbf{W}_2$, compute $\mathbf{W}_2(\mathbf{W}_1\mathbf{x})$. That is the same as one layer with the single matrix $\mathbf{W}_2\mathbf{W}_1$, so stacking them adds nothing. The simple nonlinear step between layers, such as the ReLU, is what lets a deep network represent more than any single matrix could.

Multiplying two $n \times n$ matrices takes n^3 multiplications: there are n^2 entries, each a dot product of length n. Doubling the size multiplies the work by eight. That cubic growth is why AI computing is measured in trillions of operations a second, and why it runs on chips built to multiply matrices.

2d. Transpose and Identity

The transpose of a matrix, written $\mathbf{A}^T$, swaps its rows and columns: the entry in row i and column j of $\mathbf{A}^T$ is a_{ji}, so an $m \times n$ matrix has an $n \times m$ transpose. A column vector's transpose is a row. That lets a dot product of vectors $\mathbf{x}$ and $\mathbf{y}$ be written as a matrix product, a $1 \times n$ row times an $n \times 1$ column:

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y} \qquad (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

The transpose of a product reverses the order, the way taking off socks and shoes reverses putting them on. A matrix equal to its own transpose is symmetric. Covariance matrices, in Module 6, are always symmetric.

The identity matrix $\mathbf{I}$ has ones on its diagonal, from top left to bottom right, and zeros everywhere else. It leaves every vector unchanged, $\mathbf{Ix} = \mathbf{x}$, and it plays the role among matrices that 1 plays among numbers: $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$.

2e. The Inverse

The inverse of a square matrix $\mathbf{A}$, written $\mathbf{A}^{-1}$, undoes it: $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. If $\mathbf{A}$ sends $\mathbf{x}$ to $\mathbf{y}$, then $\mathbf{A}^{-1}$ sends $\mathbf{y}$ back to $\mathbf{x}$. For a 2 × 2 matrix with entries a, b, c, and d, there is a formula:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

The number $ad - bc$ is the matrix's determinant. When it is zero there is no inverse: the matrix squashes the plane onto a line, and nothing can undo that. Module 3 takes up what this means for solving equations.

Worked example — Which mix of two funds gives 60/40?

Two balanced funds hold stocks and bonds in the proportions (0.8, 0.2) and (0.2, 0.8), the columns of the matrix $\mathbf{F}$. What allocation $\mathbf{x}$ across the two funds gives a portfolio of 60% stocks and 40% bonds, the target $\mathbf{t} = (0.6, 0.4)$? The allocation must satisfy $\mathbf{Fx} = \mathbf{t}$, so $\mathbf{x} = \mathbf{F}^{-1}\mathbf{t}$. The determinant is $0.64 - 0.04 = 0.6$, and

$$\begin{aligned} \mathbf{x} = \mathbf{F}^{-1}\mathbf{t} &= \frac{1}{0.6}\begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.8 \end{bmatrix}\begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix} \\ &= \frac{1}{0.6}\begin{bmatrix} 0.40 \\ 0.20 \end{bmatrix} = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix} \end{aligned}$$

Two-thirds in the first fund and one-third in the second gives exactly 60/40. In practice nobody computes an inverse to solve a problem like this: elimination, in Module 3, is faster and more accurate. The inverse is still the right way to think about undoing a transformation.

In a Spreadsheet

A spreadsheet treats a matrix as a block of cells. Modern spreadsheets spill a matrix answer into the cells below and to the right of the formula; older ones need the output block selected first. With a matrix in A2:C3 and another in E2:F4:

To compute	Formula
Product $\mathbf{AB}$	=MMULT(A2:C3, E2:F4)
Transpose $\mathbf{A}^T$	=TRANSPOSE(A2:C3)
Inverse $\mathbf{A}^{-1}$ of a square block	=MINVERSE(A2:B3)
Determinant of a square block	=MDETERM(A2:B3)
Identity matrix, 3 × 3	=MUNIT(3)

2f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Matrix	An $m \times n$ table of numbers, set in bold capitals, such as $\mathbf{A}$ m rows and n columns; a_{ij} is the entry in row i and column j
Matrix times vector	$(\mathbf{Ax})_i = \sum_{j=1}^{n} a_{ij} x_j \qquad \mathbf{Ax} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$ Each component is a row of $\mathbf{A}$ dotted with $\mathbf{x}$; the whole is a combination of the columns $\mathbf{a}_j$
Matrix times matrix	$(\mathbf{AB})_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$ Row i of $\mathbf{A}$ dotted with column j of $\mathbf{B}$; an $m \times n$ matrix times an $n \times p$ matrix is $m \times p$
Order and grouping	$\mathbf{AB}$ and $\mathbf{BA}$ usually differ; $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ $\mathbf{ABx}$ applies $\mathbf{B}$ first, then $\mathbf{A}$
Transpose	$(\mathbf{A}^T)_{ij} = a_{ji} \qquad (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \qquad \mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y}$ Rows and columns swap; a matrix equal to its transpose is symmetric
Identity	$\mathbf{Ix} = \mathbf{x}$ and $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$ Ones on the diagonal and zeros elsewhere
Inverse of a 2 × 2 matrix	$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ $ad - bc$ is the determinant; when it is zero there is no inverse
Linear layers	$\mathbf{W}_2(\mathbf{W}_1\mathbf{x}) = (\mathbf{W}_2\mathbf{W}_1)\mathbf{x}$ $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices; two purely linear layers equal one, so networks put a nonlinear step between them, such as the ReLU, which turns each input z into $\max(z, 0)$
Cost	Multiplying two $n \times n$ matrices takes n^3 multiplications: doubling the size multiplies the work by eight
The idea	A matrix is a machine that takes a vector and returns a combination of its own columns. Multiplying matrices chains such machines, in order.

2g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

2.1 Compute the product of the matrix with rows (1, 2), (3, 4), and (5, 6) and the vector (1, −1).

2.2 For $\mathbf{A}$ with rows (1, 2) and (0, 1) and $\mathbf{B}$ with rows (1, 0) and (3, 1), compute $\mathbf{AB}$ and $\mathbf{BA}$.

2.3 $\mathbf{A}$ is 2 × 3, $\mathbf{B}$ is 3 × 4, and $\mathbf{C}$ is 4 × 2. Which of $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{BC}$, $\mathbf{CA}$, and $\mathbf{ABC}$ are defined, and what shapes are they?

2.4 Write the transpose of the matrix with rows (1, 2, 3) and (4, 5, 6), and compute the dot product of (1, 2) and (3, 4) as a row times a column.

2.5 Find the inverse of the matrix with rows (4, 7) and (2, 6), and check it.

2.6 Does the matrix with rows (2, 4) and (1, 2) have an inverse?

Applied

2.7 Three funds hold stocks, bonds, and cash in the proportions (0.6, 0.4, 0), (0.3, 0.6, 0.1), and (0.9, 0, 0.1). An investor puts half her money in the first, 30% in the second, and 20% in the third. What is her mix of stocks, bonds, and cash?

2.8 A layer has weights $\mathbf{W}$ with rows (1, −1), (2, 0), and (−1, 1) and bias $\mathbf{b} = (0, -1, 0.5)$. Find its output for the input (3, 1) after the ReLU, and count its parameters.

2.9 Three plants generate over four hours, in MWh: a nuclear plant (100, 100, 100, 100), a gas plant (0, 50, 80, 20), and a peaking plant (0, 0, 30, 0). Their fuel costs are \$10, \$40, and \$120 per MWh. Find the cost of each hour as a row vector times a matrix.

2.10 For the plant in section 2c, with materials matrix $\mathbf{M}$, a customer orders 4, 2, and 1 units of the three products. How much steel and copper does the order take?

2.11 Let $\mathbf{S}$ stretch the plane sideways by 2, so its rows are (2, 0) and (0, 1), and let $\mathbf{R}$ turn it a quarter turn counterclockwise, with rows (0, −1) and (1, 0). Apply $\mathbf{RS}$ and $\mathbf{SR}$ to (1, 0).

2.12 Two linear layers have weights $\mathbf{W}_1$, with rows (1, 2), (0, 1), and (3, −1), and $\mathbf{W}_2$, with rows (1, 0, 2) and (−1, 1, 0). Compute $\mathbf{W}_2\mathbf{W}_1$, and check that it gives the same output as the two layers for the input (1, 1).

Challenge

2.13 Show that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$, using the formula for an entry of a product.

2.14 Show that if $\mathbf{A}$ and $\mathbf{B}$ have inverses, the inverse of $\mathbf{AB}$ is $\mathbf{B}^{-1}\mathbf{A}^{-1}$.

2.15 A model multiplies two 4,096 × 4,096 matrices. How many multiplications and additions does it take, and how long would that take on a chip that does 10^15 operations a second?

2.16 A rotation of the plane by an angle θ has rows (cos θ, −sin θ) and (sin θ, cos θ). Show that it does not change lengths, and that its transpose is its inverse.

Module 2 Quiz

Check Your Understanding

1. If A is 3 × 2 and B is 2 × 4, what is the shape of AB?

A) 2 × 2
B) 2 × 3
C) 4 × 3, because the order of the outer sizes reverses
D) 3 × 4, because the matching inner sizes cancel and the outer sizes remain

2. Read by columns, what is the product Ax?

A) The matrix's rows, scaled
B) A combination of the matrix's columns
C) A single number
D) Every entry times every component, added up

3. Why do neural networks put a nonlinear step, such as the ReLU, between layers?

A) To speed up training
B) To keep every weight positive
C) Two linear layers equal one
D) To keep every output between zero and one

4. What is the inverse of the matrix with rows (2, 0) and (0, 5)?

A) The matrix with rows (−2, 0) and (0, −5), the negatives of its entries
B) It has no inverse
C) Rows (0.5, 0) and (0, 0.2)
D) The matrix with rows (5, 0) and (0, 2), with its diagonal entries swapped

5. Why does multiplying two square matrices take eight times as many multiplications when their size doubles?

A) Because the matrices hold eight times as many entries.
B) Because each entry has to be computed twice.
C) Because the computer must keep a second copy of each matrix while it works, and both copies double.
D) Because there are four times as many entries to compute, and each entry is a dot product that is twice as long.

Module 3 — Systems of Linear Equations

Elimination, when a solution exists, and three systems that run markets

Recommended Reading / Reference

Book: Wassily Leontief, *Input-Output Economics* (2nd ed., 1986) — the Nobel-winning economist's own account of the tables in section 3e

Reference: U.S. Bureau of Economic Analysis, Input-Output Accounts (bea.gov) — official tables of what each U.S. industry buys from every other

3a. Equations as a Matrix

A system of linear equations asks for numbers that satisfy several linear equations at once. Any such system can be written as one matrix equation, $\mathbf{Ax} = \mathbf{b}$, where the matrix $\mathbf{A}$ holds the coefficients, the vector $\mathbf{x}$ the unknowns, and the vector $\mathbf{b}$ the right-hand sides. The pair $x + y = 4$ and $x - y = 2$ is

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

There are two ways to picture it. Read by rows, each equation is a line in the plane, and the solution is the point where the lines cross, here (3, 1). Read by columns, the question is which combination of the columns (1, 1) and (1, −1) makes the vector (4, 2): three of the first plus one of the second.

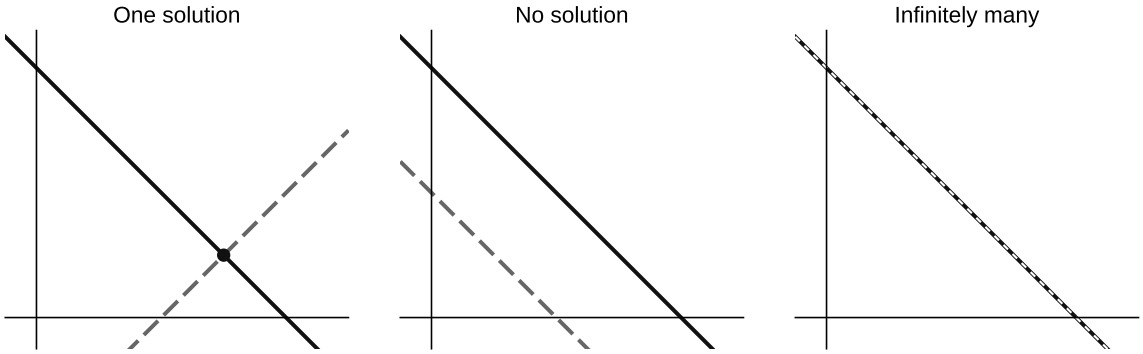


Figure 3.1. The row picture of two equations in two unknowns. Left: $x + y = 4$ and $x - y = 2$ cross once, at (3, 1). Middle: $x + y = 4$ and $x + y = 2$ are parallel and never meet. Right: $x + y = 4$ and $2x + 2y = 8$ are the same line, drawn twice.

3b. Elimination

Elimination solves any system by subtracting multiples of one equation from another until the unknowns peel off one at a time. Each step replaces the system with an easier one that has the same solutions. Take three equations in three unknowns:

$$\begin{aligned} x + y + z &= 6 \\ 2x + y - z &= 1 \\ x - y + 2z &= 5 \end{aligned}$$

Subtract twice the first equation from the second, and the first from the third, to remove x from both:

$$\begin{aligned} x + y + z &= 6 \\ -y - 3z &= -11 \\ -2y + z &= -1 \end{aligned}$$

Then subtract twice the new second equation from the third, to remove y:

$$\begin{aligned} x + y + z &= 6 \\ -y - 3z &= -11 \\ 7z &= 21 \end{aligned}$$

The last equation gives $z = 3$. Substituting back up the triangle, the second gives $y = 2$ and the first $x = 1$. That second phase is called back substitution, and the numbers left on the diagonal at the end, here 1, −1, and 7, are the pivots.

Software runs the same steps with one refinement: before each step it swaps rows so that the pivot is as large as possible, which keeps rounding errors small. Solving n equations this way takes about $n^3/3$ multiplications. A system of 1,000 equations takes about 333 million, a fraction of a second on a laptop.

3c. One Solution, None, or Infinitely Many

Elimination always ends in one of three ways. If it produces an impossible equation, such as $0 = 5$, there is no solution: the equations contradict each other, like two parallel lines. Otherwise there is exactly one solution if every unknown gets a pivot, and infinitely many if some unknown gets none, as when, in a square system, an equation drops out as $0 = 0$: the equations repeat each other, like one line drawn twice. The number of pivots is the matrix's rank.

For a square matrix, one number gives the answer in advance. The determinant, written $\det\mathbf{A}$, is $ad - bc$ for a 2 × 2 matrix with entries a, b, c, and d, and it has longer formulas for larger matrices. It is the factor by which the matrix scales areas, or volumes in three dimensions, and its sign says whether the matrix flips the plane over. It also equals the product of the pivots, with the sign flipped once for each row swap: for the system above, $1 \times (-1) \times 7 = -7$.

When the determinant is not zero, the system has exactly one solution, whatever the right-hand sides. When it is zero, the matrix squashes space flat, and the system has no solution or infinitely many.

Nearly parallel equations are a warning even when the determinant is not exactly zero. The pair $x + y = 2$ and $x + 1.01y = 2.01$ has the solution (1, 1). Change the second right-hand side to 2.02, a shift of half a percent, and the solution jumps to (0, 2). The two lines are nearly parallel, so a tiny move in one slides their crossing point a long way. Such a system is called ill-conditioned. The determinant's size alone is a poor gauge: multiplying both equations by 10 multiplies it by 100 but leaves the solution just as sensitive.

Finance meets ill-conditioning whenever two instruments are nearly the same. Hedging one bond with another whose payments almost match it takes enormous offsetting positions, and small errors in either price swing the hedge wildly.

3d. Replicating an Option

The price of an option, a contract that pays off only in some outcomes, can be found by solving a system of equations. The idea is to build a portfolio of simpler things, a stock and a safe bond, whose payoff matches the option's in every outcome. Two investments that pay the same in every outcome must cost the same today. Otherwise a trader could buy the cheaper, sell the dearer, and keep the difference at no risk. That is the principle of no arbitrage.

Worked example — A one-period call option

A stock trades at \$100, and in a year it will be worth either \$120 or \$90. A safe bond costs \$1 today and pays \$1.04 in a year, a 4% rate. A call option with a strike price of \$100 pays the stock's excess over \$100: \$20 if the stock rises, nothing if it falls. Find the number of shares, Δ (the Greek capital delta), and the number of bonds, B, that together pay the same as the call in both outcomes:

$$120\,\Delta + 1.04\,B = 20$$
$$90\,\Delta + 1.04\,B = 0$$

Subtracting the second equation from the first removes B: $30\Delta = 20$, so $\Delta = 2/3$ of a share. The second equation then gives $1.04B = -60$, so $B \approx -57.69$, which means borrowing \$57.69 at 4%. The replicating portfolio costs $100 \times 2/3 - 57.69 \approx 8.97$ dollars today, so the call must cost \$8.97 too.

The chance that the stock rises never entered the calculation. Whether investors think the stock will probably rise or probably fall, the call is worth \$8.97, because it can be built from the stock and the bond. Repeated over many short periods, this two-equation system becomes the binomial model that banks use to price options. In the limit of very short periods, it gives the Black–Scholes formula.

3e. The Economy's Input–Output Table

Industries buy from each other. Making electricity takes fuel, equipment, and services, and making equipment takes electricity. To deliver goods to final users, such as households, governments, and buyers abroad, each industry must also produce what the other industries use along the way. Wassily Leontief, who received the Nobel prize in economics for the method in 1973, wrote this as a linear system.

Let $\mathbf{x}$ be the vector of each industry's total output, in dollars, and $\mathbf{d}$ the final demand for each industry's goods. The matrix $\mathbf{A}$ holds the input coefficients: a_{ij} is the dollars of industry i's output used to make one dollar of industry j's. Total output must cover both what industries use and what final users buy, and $\mathbf{I}$ is the identity matrix:

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{d} \quad \Longrightarrow \quad (\mathbf{I} - \mathbf{A})\,\mathbf{x} = \mathbf{d}$$

$$\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{d}$$

Worked example — Energy and manufacturing

In a two-industry economy, each dollar of energy takes 20 cents of energy and 40 cents of manufactured goods to produce. Each dollar of manufactured goods takes 30 cents of energy and 10 cents of manufactured goods. Final users want \$100 of energy and \$200 of manufactured goods:

$$\mathbf{A} = \begin{bmatrix} 0.2 & 0.3 \\ 0.4 & 0.1 \end{bmatrix} \qquad \mathbf{d} = \begin{bmatrix} 100 \\ 200 \end{bmatrix} \qquad (\mathbf{I} - \mathbf{A})^{-1} = \begin{bmatrix} 1.5 & 0.5 \\ 0.67 & 1.33 \end{bmatrix}$$

So $\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{d}$ = (250, 333): the economy must produce $250 of energy and $333 of manufactured goods to deliver $100 and $200 to final users. The matrix $(\mathbf{I} - \mathbf{A})^{-1}$ is the table of total requirements. Its first column says that each dollar of energy delivered to a final user takes $1.50 of energy output and 67 cents of manufacturing output, counting every round of inputs to inputs.

Economists use these tables to trace how a shock spreads through the industries that supply each other. A surge in data-center construction, for example, raises demand for electrical equipment, which raises demand for copper, steel, and the power to make them.

3f. How Power Divides Across a Grid

Electricity on a meshed grid, where there is more than one path between two points, does not follow contracts. It follows physics, splitting across every available path in inverse proportion to each path's impedance, its opposition to the flow. Grid operators compute the split by solving a linear system. They use a simplified form of the full equations that engineers call direct-current (DC) power flow, after its resemblance to a direct-current circuit.

In that model each bus, a connection point on the grid, has a voltage angle, and the angle at bus i is written θ_i. The flow on a line from bus i to bus j is $(\theta_i - \theta_j)/X_{ij}$, where X_{ij} is the line's reactance, the main part of its impedance on a high-voltage line. At every bus, the power injected must equal the power flowing out on its lines. One bus serves as the reference, with its angle set to zero.

Worked example — Three buses in a triangle

A generator at bus 1 injects 100 megawatts (MW), a load at bus 3 takes 100 MW, and bus 2 has neither. Three identical lines join every pair of buses, each with a reactance of 1. With bus 3 as the reference, so $\theta_3 = 0$, the power balances at buses 1 and 2 are:

$$(\theta_1 - \theta_2) + (\theta_1 - \theta_3) = 100$$
$$(\theta_2 - \theta_1) + (\theta_2 - \theta_3) = 0$$

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 100 \\ 0 \end{bmatrix}$$

The solution is $\theta_1 \approx 66.7$ and $\theta_2 \approx 33.3$. The direct line from bus 1 to bus 3 carries 66.7 MW, and the indirect path through bus 2 carries the other 33.3 MW, because that path has twice the reactance. The fractions, 2/3 on the direct line and 1/3 on each line of the indirect path, are the shift factors for a transfer from bus 1 to bus 3.

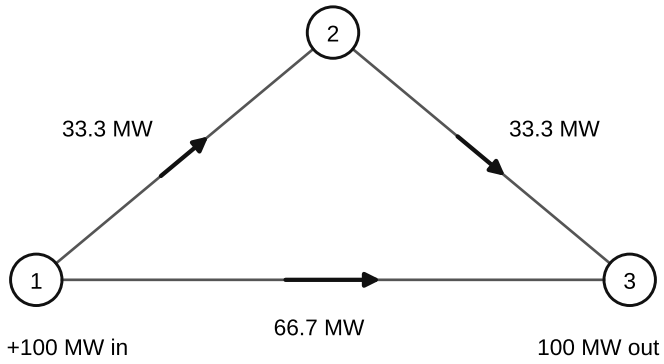


Figure 3.2. Power injected at bus 1 and taken out at bus 3 splits two to one between the direct line and the path through bus 2, which has twice the reactance.

Shift factors are how grid operators decide whether a transaction will overload a line. PJM, which runs the grid for all or parts of 13 states and the District of Columbia and is named for Pennsylvania, New Jersey, and Maryland, where it began, computes them for every line it watches. If the direct line in the example could carry only 50 MW, no contract could deliver 100 MW from bus 1 to bus 3, because two-thirds of any such transfer lands on that line. Some of the load would have to come from a generator nearer to it, usually a more expensive one. Module 8 shows how that cost becomes a price.

In a Spreadsheet

To solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ in a spreadsheet, put the coefficients in A2:C4 and the right-hand sides in E2:E4. For large systems, statistical software solves by elimination directly, which is faster and more accurate than forming the inverse.

To compute	Formula
The solution $\mathbf{x}$	=MMULT(MINVERSE(A2:C4), E2:E4)
The determinant	=MDETERM(A2:C4)
Total requirements $(\mathbf{I} - \mathbf{A})^{-1}$ for a 3 × 3 table	=MINVERSE(MUNIT(3) - A2:C4)

3g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Matrix form	$\mathbf{Ax} = \mathbf{b}$ $\mathbf{A}$ holds the coefficients, $\mathbf{x}$ the unknowns, and $\mathbf{b}$ the right-hand sides
Elimination	Subtract multiples of equations to make a triangle, then solve from the bottom up by back substitution The numbers left on the diagonal are the pivots; n equations take about $n^3/3$ multiplications
Three outcomes	None if an equation becomes $0 = c$ with c not zero; otherwise one solution if every unknown has a pivot, and infinitely many if some unknown has none The number of pivots is the rank
Determinant	$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$ Its size is the factor by which the matrix scales area; it is the product of the pivots, with the sign flipped for each row swap; not zero means exactly one solution
Ill-conditioning	Nearly parallel equations: a tiny change in the data moves the solution a long way
No arbitrage	An option is worth what the portfolio replicating its payoff costs In one period, solve for the shares Δ and bonds B that match the payoff in each outcome
Input–output	$\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{d}$ $\mathbf{x}$ is total output, $\mathbf{d}$ final demand, and a_{ij} the input from industry i per dollar of industry j's output
Power flow	Flow splits across parallel paths in inverse proportion to their reactance; the fractions are shift factors In the DC model the flow from bus i to bus j is $(\theta_i - \theta_j)/X_{ij}$, with θ_i the voltage angles and X_{ij} the line's reactance
The idea	Elimination solves any linear system or shows why it cannot be solved. The same arithmetic prices options, traces an economy's supply chains, and routes power.

3h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

3.1 Solve $2x + y = 7$ and $x - y = 2$ by elimination.

3.2 Solve by elimination: $x + 2y - z = 3$, $2x + 5y + z = 14$, and $x - y + 3z = 5$.

3.3 Find the determinant of the matrix with rows (3, 1) and (4, 2). How many solutions does a system with this matrix have?

3.4 How many solutions do $x + 2y = 4$ and $2x + 4y = 8$ have? And $x + 2y = 4$ and $2x + 4y = 9$?

3.5 Find the number c that makes the matrix with rows (2, c) and (3, 6) have no inverse.

3.6 Find the determinant of the coefficient matrix in problem 3.2 from its pivots.

Applied

3.7 A stock trades at \$50 and will be worth \$60 or \$40 in a year. A bond costs \$1 and pays \$1.05. Price a call option with a \$50 strike by replication.

3.8 In the setting of section 3d, price a put option with a strike of \$100, which pays the stock's shortfall below \$100: nothing if the stock rises to \$120, \$10 if it falls to \$90.

3.9 Two industries, energy and manufacturing, have input coefficients with rows (0.1, 0.4) and (0.3, 0.2), and final demand is \$50 of energy and \$150 of manufactured goods. Find the total outputs.

3.10 In the three-bus triangle of section 3f, the direct line from bus 1 to bus 3 is replaced with one of reactance 2, while the other two lines keep reactance 1. How does a 100 MW transfer from bus 1 to bus 3 now divide?

3.11 Two bonds pay \$100 and \$100, and \$100 and \$101, in two scenarios. A liability pays \$100 and \$102. How many of each bond replicate it? What if the liability paid \$100 and \$103 instead?

3.12 Two ores contain 2% and 0.5% copper. How many tonnes of each make 100 tonnes of a 1.1% blend?

Challenge

3.13 For the system $ax + by = e$ and $cx + dy = f$, use elimination to show that $x = (ed - bf)/(ad - bc)$ and $y = (af - ec)/(ad - bc)$ when $ad - bc$ is not zero.

3.14 In section 3d, let $q = (1.04 \times 100 - 90)/(120 - 90)$. Show that discounting the call's payoff, weighted by q and $1 - q$, gives the same price as replication.

3.15 Check put–call parity in the one-period model: the call price minus the put price should equal the stock price minus the strike's value today, $100 - 100/1.04$.

3.16 Two parallel paths between the same two buses have reactances X_1 and X_2. Show that a transfer of F megawatts sends $F\,X_2/(X_1 + X_2)$ along the first path.

Module 3 Quiz

Check Your Understanding

1. Elimination on two equations in two unknowns turns the second equation into 0 = 0. What does that mean?

A) The system has no solution
B) Infinitely many solutions
C) Exactly one solution remains
D) The determinant is negative

2. What does the determinant of a 2 × 2 matrix measure?

A) How much it scales areas
B) Its largest entry
C) Its rank
D) The sum of its diagonal entries

3. A one-period call option can be priced without knowing the probability that the stock rises. Why?

A) Because options are always fairly priced
B) Because a portfolio of the stock and a bond can match the option's payoff in every outcome, so the option must cost what that portfolio costs
C) Because the probability is always one-half
D) Because the bond's interest rate already includes the stock's expected return

4. On a meshed grid, how does power sent from one bus to another divide between two paths?

A) Equally, whatever the lines
B) By contract
C) All on the line with the most capacity
D) Inversely to impedance

5. In an input–output table, what does the total-requirements matrix show?

A) It shows the total output needed per dollar of final demand.
B) It shows how much final demand each industry can supply from one dollar of its own total output.
C) It shows only the direct inputs one industry buys from another, before any further rounds of inputs.
D) It shows which industries are profitable, since output above final demand is counted as each industry's profit.

Module 4 — Least Squares

Fitting a model to more data than it has parameters

Recommended Reading / Reference

Book: Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning* (2nd ed., 2009) — least squares, ridge regression, and overfitting, from three Stanford statisticians, free online

Reference: Kenneth R. French, Data Library (Tuck School of Business, Dartmouth) — free monthly and daily returns for the market, size, and value factors in section 4e

4a. More Data Than Unknowns

Module 3 solved systems with as many equations as unknowns. Data usually supplies far more equations than unknowns. Five afternoons of temperature and electricity demand, for example, give five equations for a straight line's two numbers, its intercept and its slope. No line passes through all five points, so the system has no exact solution, and the question becomes which line comes closest.

Write the model as $\mathbf{A}\mathbf{x} \approx \mathbf{b}$. The vector $\mathbf{b}$ holds the n observations, the matrix $\mathbf{A}$ has one row per observation and one column per input to the model, and the vector $\mathbf{x}$ holds the model's coefficients. For a line $y = c + mT$, with intercept c, slope m, and temperature T, each row of $\mathbf{A}$ is (1, T) for one afternoon, $\mathbf{b}$ lists the observed values of y, and $\mathbf{x} = (c, m)$. The residuals, $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$, are the misses: observed minus fitted.

Least squares chooses the coefficients that make the sum of the squared residuals, $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2$, as small as possible. Squaring treats misses above and below the line alike and punishes large misses most. It also leads to a formula.

4b. The Best Fit Is a Projection

As $\mathbf{x}$ varies, $\mathbf{A}\mathbf{x}$ ranges over every combination of the columns of $\mathbf{A}$. The best fit is the combination closest to $\mathbf{b}$: the projection of $\mathbf{b}$ onto the space the columns span, just as Module 1 projected one vector onto another. The residual of a projection is perpendicular to what was projected onto, so here it is perpendicular to every column of $\mathbf{A}$: $\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0}$. Rearranged, that condition gives the normal equations:

$$\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

The matrix $\mathbf{A}^T\mathbf{A}$ is square, with one row and one column per coefficient, however many observations there are. Solving this small system gives the least-squares coefficients. Normal is an old word for perpendicular.

4c. Fitting a Line

Worked example — Demand against temperature

On five summer afternoons, the temperature T, in degrees Celsius, and the electricity demand y, in gigawatts (GW), were:

Temperature T (°C)	22	25	28	31	34
Demand y (GW)	30.6	33.9	36.0	38.7	41.8

With a row (1, T) for each afternoon, the normal equations are

$$\begin{bmatrix} 5 & 140 \\ 140 & 4{,}010 \end{bmatrix} \begin{bmatrix} c \\ m \end{bmatrix} = \begin{bmatrix} 181.0 \\ 5{,}149.6 \end{bmatrix}$$

Here 5 is the number of afternoons, 140 the sum of the temperatures, 4,010 the sum of their squares, 181.0 the sum of the demands, and 5,149.6 the sum of each temperature times its demand. Solving gives $c \approx 10.81$ and $m \approx 0.907$: demand rises about 0.91 GW for each degree. At 36 degrees, hotter than any afternoon in the data, the line predicts $10.81 + 0.907 \times 36 \approx 43.5$ GW.

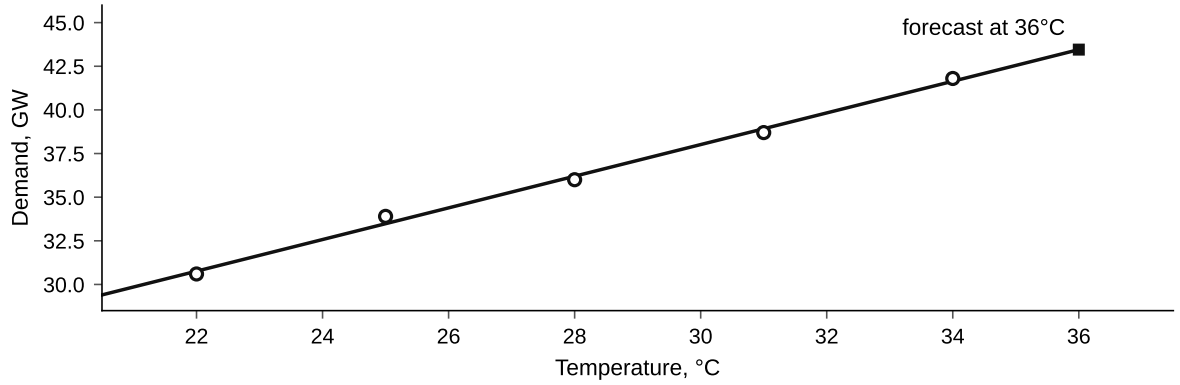


Figure 4.1. The least-squares line through five afternoons of demand. The residuals, the vertical gaps between the points and the line, are at most 0.42 GW, too small to see at this scale; the line makes the sum of their squares as small as possible.

For a line, the solution has a memorable form. With T_i and y_i the ith afternoon's temperature and demand, and bars marking averages:

$$m = \frac{\sum_i (T_i - \bar{T})(y_i - \bar{y})}{\sum_i (T_i - \bar{T})^2} \qquad c = \bar{y} - m\,\bar{T}$$

Here the average temperature is 28 degrees and the average demand 36.2 GW, so the slope is $81.6/90 \approx 0.907$. The second formula says the line passes through the point of averages. The residuals are −0.16, 0.42, −0.20, −0.22, and 0.16 GW; with an intercept in the model, they always add to zero.

How well does the line fit? R^2, the share of the variation in demand the line explains, is one minus the sum of the squared residuals divided by the sum of the squared deviations of demand from its average. Here that is $1 - 0.316/74.3 \approx 0.996$, a very close fit.

4d. Several Inputs at Once

Least squares handles any number of inputs, each adding a column to $\mathbf{A}$. Demand depends on the day of the week as well as the weather. A column that is 1 on weekdays and 0 at weekends, called a dummy variable, captures that.

Worked example — Temperature and weekdays

Eight days of data are below. With a row (1, T, w) for each day, where w is 1 on a weekday and 0 at the weekend, the normal equations form a 3 × 3 system, which a spreadsheet's LINEST function solves directly:

Temperature T (°C)	Weekday w	Demand (GW)
24	1	33.0
27	1	35.5
30	1	38.6
33	1	41.0
26	0	32.1
29	0	35.0
32	0	37.6
35	1	43.0

$$y \approx 8.47 + 0.911\,T + 2.59\,w$$

Demand rises about 0.91 GW per degree, as before, and weekdays add about 2.6 GW for offices and factories. Each coefficient measures one input's effect with the others held fixed, which is the point of fitting them together. Here $R^2 \approx 0.999$.

4e. A Stock's Beta and Factor Models

The same tool measures how a stock moves with the market. Fitting a stock's monthly return against the market's, each measured above the return on a safe Treasury bill, gives

$$r_s - r_f = \alpha + \beta\,(r_m - r_f) + \varepsilon$$

where r_s, r_m, and r_f are the returns on the stock, the market, and the bill. The slope β, beta, is the stock's sensitivity to the market. The intercept α, alpha, is the average return the market does not explain, and ε, epsilon, is each month's residual. The capital asset pricing model, for which William Sharpe shared the 1990 Nobel prize in economics, holds that a stock's expected excess return is its beta times the market's, so alpha should average zero.

Worked example — Six months of returns

Over six months, the market's excess returns were 2.0, −1.5, 3.0, 0.5, −2.5, and 1.5 percent, and a stock's were 3.4, −2.6, 2.9, 1.9, −3.1, and 0.9. Their averages are 0.50% and 0.57%. The formula for a line's slope gives $\beta = 27.75/22.5 \approx 1.23$, and then $\alpha = 0.57 - 1.23 \times 0.50 \approx -0.05$ percent a month.

The stock moved about 1.23% for each 1% move in the market, and the market explained 88% of the variation in its returns over these six months. Six months is far too few to trust either number. Even with years of real data, a single stock's beta is uncertain by a tenth or more, and the market often explains less than half of a stock's movement.

Eugene Fama and Kenneth French added two more inputs in 1993: the return of small companies' stocks minus large ones', and the return of cheap stocks, those with a high book value relative to their market value, minus expensive ones'. Their three-factor model is the same least-squares fit with three inputs instead of one. Investors use such models to separate a manager's skill, which shows up as alpha, from exposure to factors that anyone can buy cheaply.

4f. Overfitting and the Ridge Penalty

Adding inputs never worsens the fit to the data in hand, because least squares can use extra freedom only to shrink the residuals. It does not always improve predictions. A model with as many coefficients as observations can pass through every point, fitting the noise along with the signal, and then predict new data badly. This is overfitting.

Worked example — Six points, two models

Six points scatter around the line $y = 1 + 0.5x$. A straight line, with two coefficients, fits them with an average squared residual of 0.145. A polynomial of degree five, using powers of x up to the fifth and six coefficients, passes through all six exactly. On five new points from the same source, the line's average squared error is 0.056 and the polynomial's is 0.441, eight times worse.

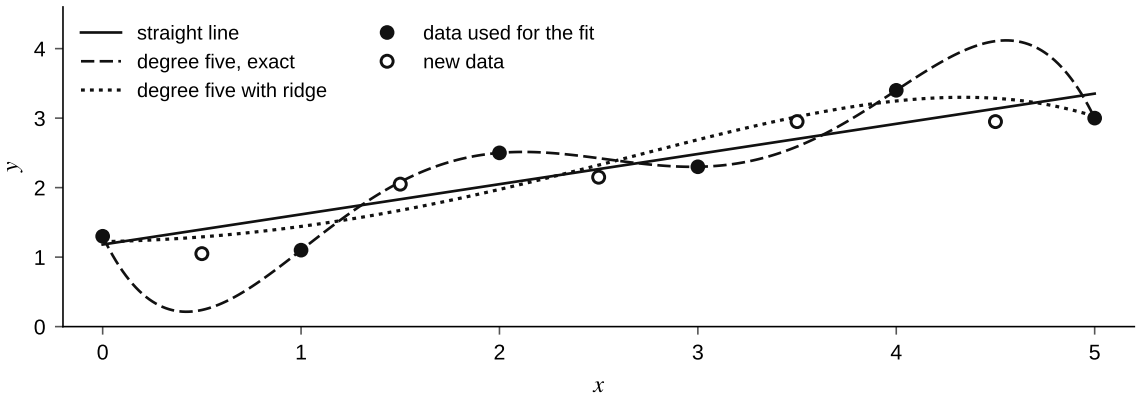


Figure 4.2. The degree-five polynomial passes through all six points used for the fit but swings between them, missing the new points. The ridge penalty tames the swings.

The ridge penalty is the standard remedy. It adds the sum of the squared coefficients, weighted by a number λ (lambda) that the modeler chooses, to the quantity being minimized. That changes the normal equations only slightly, adding λ times the identity matrix $\mathbf{I}$:

$$\text{minimize} \;\; \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 + \lambda \, \|\mathbf{x}\|^2$$

$$(\mathbf{A}^T\mathbf{A} + \lambda \, \mathbf{I}) \, \mathbf{x} = \mathbf{A}^T\mathbf{b}$$

Large coefficients now cost something, so the fit uses them only where the data support them strongly. With $\lambda = 0.5$, leaving the intercept unpenalized as is usual, by putting a zero in place of the identity matrix's first diagonal entry, the degree-five polynomial's coefficients shrink and its curve flattens. Its error on the six points rises to 0.096, but on the new points it falls to 0.070, close to the line's. The modeler chooses λ by holding some data back and picking the value that predicts the held-back data best.

Training an AI model faces the same trade-off on a vast scale. A model with billions of parameters could memorize much of its training data, and a penalty on the size of its weights, called weight decay in machine learning, is one of the standard defenses.

In a Spreadsheet

Spreadsheets fit lines and planes directly. With the inputs in columns and the observations of y in another:

To compute	Formula
Slope and intercept of a line	=SLOPE(B2:B6, A2:A6) and =INTERCEPT(B2:B6, A2:A6)
Several inputs at once	=LINEST(D2:D9, B2:C9): the coefficients come back in reverse column order, intercept last
R^2 for a line	=RSQ(B2:B6, A2:A6)
A prediction from the fitted line	=TREND(B2:B6, A2:A6, 36)

4g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Least squares	Choose $\mathbf{x}$ to minimize $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2$, the sum of the squared residuals $\mathbf{b}$ holds the observations, $\mathbf{A}$ has a row per observation and a column per input, and $\mathbf{x}$ holds the coefficients
Normal equations	$\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\mathbf{b}$ The residual is perpendicular to every column of $\mathbf{A}$: the fit is a projection
Slope and intercept of a line	$m = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad c = \bar{y} - m\,\bar{x}$ For $y = c + mx$ fitted to points (x_i, y_i); bars mark averages, and the line passes through the point of averages
R squared	$R^2 = 1 - \frac{\sum_i r_i^2}{\sum_i (y_i - \bar{y})^2}$ The share of the variation in y that the model explains; r_i are the residuals
Dummy variable	A column of 1s and 0s that measures the effect of a category, such as weekdays
Beta and alpha	$r_s - r_f = \alpha + \beta\,(r_m - r_f) + \varepsilon$ r_s, r_m, and r_f are the returns on a stock, the market, and a Treasury bill; β is the slope, α the unexplained average return, and ε the residual
Overfitting	More coefficients always fit past data at least as well but can predict new data worse: judge a model on data held back from the fit
Ridge penalty	$(\mathbf{A}^T\mathbf{A} + \lambda\,\mathbf{I})\,\mathbf{x} = \mathbf{A}^T\mathbf{b}$ λ weights a penalty on the squared size of the coefficients; a larger λ shrinks them toward zero
The idea	With more data than unknowns, the best fit is a projection, and its residual is perpendicular to everything the model can do. Fit on some data and judge on the rest.

4h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

4.1 Write $\mathbf{A}$ and $\mathbf{b}$ for fitting $y = c + mx$ to the points (1, 2), (2, 3), and (3, 5).

4.2 Form and solve the normal equations for problem 4.1.

4.3 Find the residuals of the fit in problem 4.2. Check that they add to zero and are perpendicular to the column of x values.

4.4 Use the formula with averages to fit a line to the points (2, 1), (4, 5), and (6, 6).

4.5 Find R^2 for the fit in problem 4.4.

4.6 A model fits 30 days of demand with a constant, temperature, and a weekday dummy. What are the shapes of $\mathbf{A}$, $\mathbf{A}^T\mathbf{A}$, and $\mathbf{x}$?

Applied

4.7 Using the fitted model $y \approx 8.47 + 0.911T + 2.59w$ from section 4d, predict demand on a 31°C weekday and on a 31°C Sunday.

4.8 A battery's capacity, measured after 0, 500, 1,000, and 1,500 full cycles, was 100%, 97.6%, 95.3%, and 92.4% of its capacity when new. Fit a line and estimate the capacity after 2,500 cycles.

4.9 Over five months, a market's excess returns were 1, −2, 3, 0, and −2 percent, and a fund's were 2, −1, 4, 1, and −1. Estimate the fund's beta and alpha.

4.10 A fund's three-factor regression gives a market beta of 1.0, a loading of 0.6 on small-minus-large, 0.4 on cheap-minus-expensive, and an alpha of zero. What does that suggest about the manager?

4.11 For a line through the origin, $y \approx wx$, the ridge solution is $w = \Sigma xy/(\Sigma x^2 + \lambda)$, with the sums over the data points. Compute w for x = (1, 2, 3) and y = (2, 4, 7), with $\lambda = 0$ and with $\lambda = 6$.

4.12 As inputs are added to a model, its error on the training data falls from 2.0 to 0.1, while its error on held-back data falls to 0.8 and then rises to 1.5. Which model should be used, and why?

Challenge

4.13 Derive the normal equations for a line $y = c + mx$ by setting the partial derivatives of $E(c, m) = \Sigma(y_i - c - mx_i)^2$ to zero, with the sums over the data points (x_i, y_i).

4.14 Show that when a model includes an intercept, its residuals add to zero and its fitted line passes through the point of averages.

4.15 Show that the least-squares slope of a line through the origin is Module 1's projection coefficient.

4.16 What happens to the ridge solution as λ shrinks toward zero, and as it grows very large?

Module 4 Quiz

Check Your Understanding

1. What does least squares minimize?

A) The largest error
B) The sum of the errors
C) The sum of the squared differences between the data and the fit
D) The number of data points that do not lie exactly on the fitted line, counted one by one

2. Why is the least-squares residual perpendicular to every column of the data matrix?

A) Because the data are noisy
B) Because the columns have length one
C) Because the residual is always zero for the first column, which holds the ones of the intercept
D) Because the best fit is the projection of the data onto the space the columns span, and a projection leaves a perpendicular remainder

3. A stock's beta against the market is 1.3. What does that mean?

A) It returns 1.3% a month
B) It moves about 1.3% per 1% market move
C) It is 30% safer than the market as a whole
D) It has beaten the market by 1.3 percentage points a year after adjusting for risk, which is called alpha

4. What does the ridge penalty do to a model's coefficients?

A) Shrinks them toward zero
B) Makes them larger
C) Makes them positive
D) Sets the largest one to zero

5. A polynomial of degree five fits six data points exactly. What is the likely problem?

A) Too few parameters
B) A singular matrix
C) Overfitting
D) Negative residuals

Module 5 — Eigenvalues and Eigenvectors

Directions a matrix only stretches, and what they say about the long run

Recommended Reading / Reference

Book: Amy N. Langville and Carl D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (2006) — eigenvectors and Markov chains at work in web search

Reference: U.S. Department of the Treasury, Daily Treasury Par Yield Curve Rates (home.treasury.gov) — free daily yields at every maturity, for computing principal components like those of section 5e from real data

5a. Directions a Matrix Only Stretches

Most vectors change direction when a matrix multiplies them. A few special directions do not: the matrix only stretches them, shrinks them, or flips them around. A nonzero vector $\mathbf{v}$ with

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

is an eigenvector of the square matrix $\mathbf{A}$, and the number λ (lambda) is its eigenvalue, the factor by which the matrix stretches that direction. The word comes from German, in which eigen means “own”: these are the matrix's own directions.

Worked example — A matrix and its two directions

The matrix $\mathbf{A}$ with rows (2, 1) and (1, 2) sends (1, 1) to (3, 3), which is 3 times (1, 1). It sends (1, −1) to (1, −1) itself. So (1, 1) is an eigenvector with eigenvalue 3, and (1, −1) is one with eigenvalue 1. Any other vector changes direction: (1, 0) goes to (2, 1).

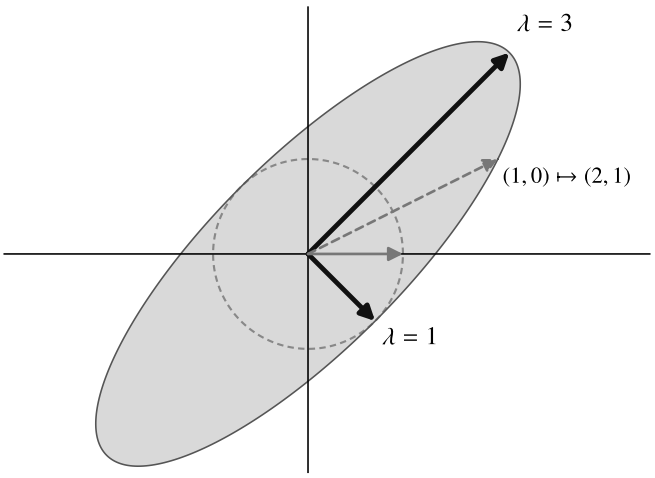


Figure 5.1. The matrix $\mathbf{A}$ with rows (2, 1) and (1, 2) turns the dashed unit circle into the shaded ellipse. Along (1, 1) it stretches by 3, along (1, −1) by 1, and every other direction turns, as (1, 0) does.

Every vector is a combination of the two eigenvectors, so the matrix's whole action can be read from them. It triples everything along the direction (1, 1) and leaves the perpendicular direction alone, which is why the circle of unit vectors becomes an ellipse whose long axis points along (1, 1).

5b. Finding Eigenvalues

Rewriting $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ as $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$, where $\mathbf{I}$ is the identity matrix, shows that the matrix $\mathbf{A} - \lambda\mathbf{I}$ sends a nonzero vector to zero. So it has no inverse, and its determinant must be zero. For a 2 × 2 matrix with rows (a, b) and (c, d), that condition is a quadratic equation in λ:

$$\begin{aligned}\det(\mathbf{A} - \lambda\mathbf{I}) &= (a - \lambda)(d - \lambda) - bc \\ &= \lambda^2 - (a + d)\,\lambda + (ad - bc) = 0\end{aligned}$$

The sum $a + d$ of the diagonal entries is called the trace. The two eigenvalues add up to the trace and multiply to the determinant. For the matrix above, the trace is 4 and the determinant is 3, so $\lambda^2 - 4\lambda + 3 = 0$, which gives 3 and 1. Each eigenvalue's eigenvector then comes from solving the system for $\mathbf{v}$ by elimination.

Larger matrices have more eigenvalues, one per row when repeats are counted, and some can be complex numbers, which describe rotation. Software finds them all. Symmetric matrices, equal to their own transposes, are the well-behaved case that statistics and finance mostly meet: their eigenvalues are real numbers, and their eigenvectors are perpendicular to each other.

5c. What Repeated Multiplication Does

Eigenvalues matter most when a matrix is applied again and again, as when a system steps forward in time by $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$, where $\mathbf{x}_k$ is the state after k steps. Write the starting vector as a combination of eigenvectors. Each step multiplies each eigenvector's part by its own eigenvalue, so after k steps the part along an eigenvector with eigenvalue λ has been multiplied by λ^k. The eigenvalue largest in size soon dominates. It sets the long-run growth rate, and its eigenvector sets the long-run shape.

Worked example — A population by age

A fish stock has juveniles and adults. Each year, every adult produces 1.5 surviving juveniles, half of the juveniles survive to become adults, and 80% of adults survive another year. With j_k juveniles and a_k adults in year k:

$$\begin{bmatrix} j_{k+1} \\ a_{k+1} \end{bmatrix} = \begin{bmatrix} 0 & 1.5 \\ 0.5 & 0.8 \end{bmatrix} \begin{bmatrix} j_k \\ a_k \end{bmatrix}$$

The eigenvalues solve $\lambda^2 - 0.8\lambda - 0.75 = 0$: about 1.354 and −0.554. Starting from 100 juveniles and no adults, the total swings at first, through 50, 115, 129.5, and 189.85. It then settles into growth of 35.4% a year, with juveniles making up 52.6% of the stock, the proportions of the leading eigenvector. The negative eigenvalue's part, which flips sign each year, fades as 0.554^k. Demographers project human populations with the same arithmetic, in matrices with a row for every age.

5d. Markov Chains and Steady States

A Markov chain tracks shares that move between states each period with fixed probabilities. Its transition matrix $\mathbf{P}$ has an entry p_{ij} for the probability of moving from state j to state i, so each column adds to 1, and the shares evolve by $\mathbf{x}_{k+1} = \mathbf{P}\mathbf{x}_k$. Because each column adds to 1, $\mathbf{P}$ always has an eigenvalue of exactly 1. That eigenvalue's eigenvector, scaled so its shares add to 1, is the steady state: the mix that one more step leaves unchanged.

Worked example — Customers switching suppliers

Each year, 10% of an electricity retailer's customers switch to its rival, and 5% of the rival's customers switch the other way. Starting from an even split:

$$\mathbf{P} = \begin{bmatrix} 0.90 & 0.05 \\ 0.10 & 0.95 \end{bmatrix} \qquad \mathbf{P}\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.475 \\ 0.525 \end{bmatrix}$$

The steady state balances the flows. When the first retailer's share is s, it loses $0.10s$ a year and gains $0.05(1 - s)$, and the two match at $s = 1/3$. The retailer is headed for a third of the market, whatever the starting split. The other eigenvalue, $0.90 + 0.95 - 1 = 0.85$, sets the pace: the gap to the steady state shrinks by 15% a year, halving about every 4.3 years.

Rating agencies publish transition matrices like this for corporate credit ratings, usually transposed, with a row rather than a column for each starting rating, and banks multiply them to get default probabilities over many years.

Worked example — Default over several years

Suppose, with illustrative numbers, that each year an investment-grade borrower stays investment grade with probability 0.968, is downgraded to speculative grade with probability 0.03, and defaults with probability 0.002. A speculative-grade borrower is upgraded with probability 0.03, stays put with probability 0.93, and defaults with probability 0.04. Default is permanent. Multiplying the matrix by itself gives the two-year probabilities:

$$\mathbf{P} = \begin{bmatrix} 0.968 & 0.03 & 0 \\ 0.03 & 0.93 & 0 \\ 0.002 & 0.04 & 1 \end{bmatrix} \qquad \mathbf{P}^2 \approx \begin{bmatrix} 0.938 & 0.057 & 0 \\ 0.057 & 0.866 & 0 \\ 0.005 & 0.077 & 1 \end{bmatrix}$$

An investment-grade borrower's chance of default within two years is 0.51%, more than twice the one-year 0.2%, because some borrowers are downgraded first and then default from the lower grade. Higher powers give 2.0% over five years and 5.9% over ten.

Google's original ranking of web pages, PageRank, is also the steady state of a Markov chain. Picture a surfer who follows a random link from the current page 85% of the time and otherwise jumps to a page chosen at random; Sergey Brin and Larry Page proposed that 85% figure in 1998. A page's rank is the share of time the surfer spends on it in the long run.

Worked example — Ranking three pages

Page A links to B and C, B links to C, and C links to A. The surfer's transition matrix mixes the link matrix, with weight 0.85, and a jump to any of the three pages, with weight 0.15. Its steady state gives ranks of 0.388 for A, 0.215 for B, and 0.397 for C. Page C ranks first, because two pages link to it. Page A ranks a close second: only one page links to it, but that page is C, the highest-ranked, which sends A all of its traffic.

5e. Principal Components

Symmetric matrices have a special place in data analysis, because a covariance matrix is symmetric. A covariance matrix records how much each of a set of quantities varies and how much each pair varies together; Module 6 builds one from data. Its eigenvectors, called principal components, are the patterns in

which the quantities tend to move together, and each eigenvalue is the variance, the amount of movement, that its pattern accounts for. Principal component analysis (PCA) ranks the patterns from largest to smallest.

Worked example — The shapes of the yield curve

Take the daily changes in Treasury yields at maturities of 2, 5, 10, and 30 years, with illustrative volatilities of 6 to 7 basis points a day and correlations between 0.70 and 0.95. The eigenvectors of their 4 × 4 covariance matrix, and the share of the total variance each explains, are:

Component	2-year	5-year	10-year	30-year	Share
Level	0.42	0.55	0.54	0.48	90.1%
Slope	0.72	0.23	−0.26	−0.61	7.4%
Curvature	0.50	−0.47	−0.43	0.59	1.9%
Fourth	0.25	−0.65	0.68	−0.24	0.6%

Principal components of daily yield changes, from an illustrative covariance matrix. Each row is an eigenvector; the share is its eigenvalue divided by the sum of all four.

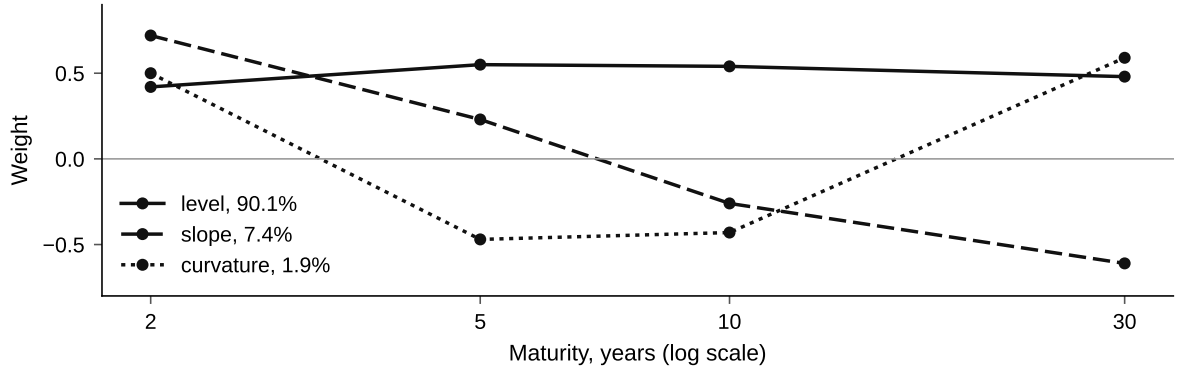


Figure 5.2. The first three principal components of yield changes across maturities: a parallel shift, a tilt, and a bend.

The first component moves every maturity the same way, a parallel shift, and explains 90% of the variance. The second tilts the curve, raising short yields while lowering long ones, and explains 7%. The third bends it, moving the two ends against the middle, and explains about 2%. Robert Litterman and José Scheinkman found the same three patterns in real Treasury returns in 1991, named them level, steepness, and curvature, and showed that together they explain almost all of the variation. Bond managers measure and hedge their exposure to each of the three.

PCA works the same way on any set of correlated data: stock returns, power prices at different hubs, or the hourly shape of demand. It finds the few patterns that carry most of the movement, which is also the idea behind compressing data, taken up again in Module 10.

In a Spreadsheet

Spreadsheets have no eigenvalue function, but three workarounds cover the common cases, with the trace in B1 and the determinant in B2 for the first:

To compute	Formula
Eigenvalues of a 2 × 2 matrix	=(B1+SQRT(B1^2-4*B2))/2 and =(B1-SQRT(B1^2-4*B2))/2
Steady state of a Markov chain	Repeat =MMULT(P, x) on the shares until they stop changing
Largest eigenvalue and its eigenvector	Repeat =MMULT(A, x), dividing by the vector's length each time: the power method
All eigenvalues and principal components	Statistical software, or the Python function numpy.linalg.eigh for a symmetric matrix

5f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Eigenvector and eigenvalue	$\mathbf{A}\mathbf{v} = \lambda\,\mathbf{v}$ $\mathbf{v}$ is a nonzero vector that the square matrix $\mathbf{A}$ only stretches, by the factor λ
Finding them	$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ For a 2 × 2 matrix with rows (a, b) and (c, d), $\lambda^2 - (a + d)\lambda + (ad - bc) = 0$; the eigenvalues add to the trace, $a + d$, and multiply to the determinant
Symmetric matrices	Real eigenvalues and perpendicular eigenvectors
Repeated multiplication	$\mathbf{A}^k\mathbf{v} = \lambda^k\mathbf{v}$ In $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$, the eigenvalue largest in size sets the long-run growth rate and its eigenvector the long-run shape
Markov chain	$\mathbf{x}_{k+1} = \mathbf{P}\,\mathbf{x}_k$ p_{ij} is the probability of moving from state j to state i, so columns add to 1; the steady state is the eigenvector with eigenvalue 1, and the second-largest eigenvalue in size sets how fast the chain approaches it
Several periods	$\mathbf{P}^k$ gives the probabilities of moving between states over k periods
PageRank	A page's share of time, in the long run, of a random surfer who follows links 85% of the time
Principal components	The eigenvectors of a covariance matrix: patterns of joint movement, each explaining its eigenvalue's share of the variance; for yields, level, slope, and curvature
The idea	A matrix's eigenvectors are the directions it only stretches. Applied over and over, a matrix is ruled by its largest eigenvalue, which is why eigenvalues predict growth, steady states, and the main patterns in data.

5g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

5.1 Check that (1, 2) is an eigenvector of the matrix with rows (3, 1) and (2, 4), and find its eigenvalue.

5.2 Find the eigenvalues of the matrix with rows (4, 1) and (2, 3).

5.3 Find an eigenvector for each eigenvalue in problem 5.2.

5.4 Find the eigenvalues of the matrix with rows (3, 0) and (0, −2), and of the matrix with rows (2, 5) and (0, 4).

5.5 Check that the eigenvalues in problem 5.2 add to the trace and multiply to the determinant.

5.6 For $\mathbf{A}$ with rows (2, 1) and (1, 2), compute $\mathbf{A}^3$ times (1, 1) and times (1, 0), using eigenvectors.

Applied

5.7 Each year 20% of one retailer's customers switch to its rival, and 10% of the rival's switch back. Find the first retailer's steady-state share and how fast the market approaches it.

5.8 Under the section 5d matrix, a retailer starts with half the market. What is its share after three years?

5.9 Page A links to B, B links to A and C, and C links to A. With the surfer following links 85% of the time, find the ranks. (Since the ranks add to 1, each page gets $0.15/3 = 0.05$ from random jumps, plus 0.85 times the rank flowing in along links.)

5.10 Using the two-year matrix in section 5d, what is the chance that an investment-grade borrower is speculative grade two years later?

5.11 A population has juveniles and adults. Each adult produces 2 surviving juveniles a year, 40% of juveniles survive to adulthood, and 50% of adults survive. Find the long-run growth rate.

5.12 A covariance matrix of yield changes has eigenvalues 158.8, 13.1, 3.4, and 1.0. What share of the variance does each component explain, and how many components are needed to explain 97%?

Challenge

5.13 Show that if $\mathbf{A}$ is symmetric, eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ with different eigenvalues λ_1 and λ_2 are perpendicular.

5.14 Show that a matrix $\mathbf{P}$ whose columns each add to 1 has an eigenvalue of 1.

5.15 In the power method, a vector is multiplied by a matrix over and over. Show that the part along the second eigenvector shrinks, relative to the part along the first, by the ratio of their eigenvalues each step. How fast is that for the fish stock of section 5c?

5.16 For a 2 × 2 matrix, use the quadratic $\lambda^2 - (a+d)\lambda + (ad - bc) = 0$ to show that the eigenvalues add to the trace and multiply to the determinant.

Module 5 Quiz

Check Your Understanding

1. If $\mathbf{A}\mathbf{v} = 3\mathbf{v}$ for a nonzero vector $\mathbf{v}$, what is 3?

A) The determinant
B) An eigenvalue
C) The vector's length
D) The matrix's trace

2. In a Markov chain of market shares, what is the steady state?

A) The starting shares
B) The largest entry of the transition matrix
C) The eigenvector with eigenvalue 1, scaled so its shares add to 1: the mix that one more period leaves unchanged
D) Equal shares

3. What sets how fast a two-state Markov chain approaches its steady state?

A) The size of the second eigenvalue
B) The first eigenvalue
C) The starting shares
D) The steady-state shares, since a chain heading for a lopsided split always takes longer to get there

4. In a principal component analysis of daily yield-curve changes, what does the first component usually look like?

A) A steepening, with short yields rising and long yields falling
B) A hump
C) Random noise across maturities, with no pattern from one maturity to the next
D) A parallel shift

5. What is a web page's PageRank, in the terms of this module?

A) Its share in the chain's steady state
B) The number of other pages that link to it
C) The number of links on the page that point to other pages
D) The eigenvalue of the link matrix that belongs to the page, which measures how much traffic the page sends to others

Module 6 — Covariance and Portfolio Risk

How things move together, and what that does to risk

Recommended Reading / Reference

Book: Edwin J. Elton, Martin J. Gruber, Stephen J. Brown, and William N. Goetzmann, *Modern Portfolio Theory and Investment Analysis* (9th ed., 2014) — the standard text on covariance, diversification, and portfolio risk

Reference: Aswath Damodaran, Historical Returns on Stocks, Bonds and Bills (New York University, Stern School of Business) — free yearly U.S. returns since 1928 for stocks, bonds, bills, real estate, and gold, enough to build this module's covariance matrices from real data

6a. Variance and Standard Deviation

An investment's average return says nothing about how far any one month can stray from it, and a portfolio's risk depends on more than the risks of its parts. This module measures how much quantities vary and how they vary together. It then gathers both into a matrix that gives the risk of any portfolio in one line.

Take six months of returns, in percent, for a stock index and a bond fund. The numbers are illustrative:

Month	1	2	3	4	5	6	Average
Stock return (%)	6.0	−4.0	3.0	−2.0	5.0	−2.0	1.0
Bond return (%)	0.3	2.1	0.9	−0.6	0.0	1.5	0.7
Stock deviation	5.0	−5.0	2.0	−3.0	4.0	−3.0	
Bond deviation	−0.4	1.4	0.2	−1.3	−0.7	0.8	
Product of the two	−2.0	−7.0	0.4	3.9	−2.8	−2.4	

A deviation is a month's return minus the average return. The products of the deviations are used in section 6b.

The variance measures spread: roughly, the average squared distance from the average. With n observations, x_i the ith of them, and a bar over a letter marking its average, the sample variance is

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where σ^2 is read sigma squared. Dividing by $n - 1$ rather than n corrects for measuring the deviations from the data's own average, which sits slightly closer to the data than the true average does. With dozens of observations the difference hardly matters. Dividing by n instead gives the population variance, the plain average of the squared deviations.

The standard deviation, σ, is the square root of the variance, which puts it back in the units of the data. The standard deviation of an investment's returns is called its volatility.

Worked example — Six months of stocks and bonds

The stock deviations in the table are 5, −5, 2, −3, 4, and −3, and their squares add to 88. The variance is $88/5 = 17.6$, in percent squared, and the standard deviation is $\sqrt{17.6} \approx 4.20\%$ a month. For the bonds, the squared deviations add to 4.98, the variance is 0.996, and the standard deviation is about 1.00% a month.

Volatility is usually quoted per year. If months are independent of each other, variances add over time, so a year's variance is 12 times a month's, and volatility grows with the square root of time:

$$\sigma_{\text{annual}} = \sqrt{12}\,\sigma_{\text{monthly}}$$

The stocks' annual volatility is about $4.20\% \times 3.46 \approx 14.5\%$, and the bonds' about 3.5%. Daily volatilities scale up by the square root of about 252, the number of trading days in a year.

6b. Covariance and Correlation

Covariance measures how two series move together. It averages the products of their deviations, with x_i and y_i the two series' values in period i:

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

When the two tend to be above their averages at the same times, and below them at the same times, most products are positive and so is the covariance. When one tends to be high while the other is low, the covariance is negative. For the stocks and bonds, the products in the last row of the table add to −9.9, so their covariance is $-9.9/5 = -1.98$, in percent squared.

A covariance's size depends on the units and on the two volatilities, so it is hard to read on its own. Dividing it by the two standard deviations, σ_x and σ_y, gives the correlation, ρ (rho), a pure number between −1 and 1:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}$$

Here $\rho = -1.98/(4.20 \times 1.00) \approx -0.47$. The bonds tended to rise in the months the stocks fell, but far from perfectly.

Correlation is also a cosine. Put the stock deviations into a vector $\mathbf{d}_x$ and the bond deviations into $\mathbf{d}_y$. The two factors of $n - 1$ cancel, and the correlation is the cosine of the angle between the two vectors, the measure of similarity from Module 1:

$$\rho_{xy} = \frac{\mathbf{d}_x \cdot \mathbf{d}_y}{\|\mathbf{d}_x\|\,\|\mathbf{d}_y\|} = \frac{-9.9}{\sqrt{88}\,\sqrt{4.98}} \approx -0.47$$

A correlation of 1 means the deviation vectors point the same way, so one series' deviations are a positive multiple of the other's. A correlation of −1 means they point in opposite directions. A correlation of 0 means they are perpendicular: there is no straight-line relationship between the two series. Here the angle between them is about 118 degrees.

Correlation captures only straight-line co-movement, and it changes over time. For most of the two decades before 2022, U.S. stocks and high-quality bonds tended to move in opposite directions, so bonds cushioned losses on stocks. In 2022, as inflation and interest rates rose, both fell. The Standard & Poor's 500 index of large U.S. companies (S&P 500) returned −18.1% including dividends, and the Bloomberg U.S. Aggregate index of investment-grade bonds lost 13.0%.

A correlation estimated from the past is a forecast, and it can fail when it matters most.

6c. The Covariance Matrix

With more than two assets, the variances and covariances fill a square table, the covariance matrix. It is written $\mathbf{\Sigma}$, a bold capital sigma, which is a matrix and not the sign for a sum. The entry in row i and column j is the covariance of assets i and j, written σ_{ij}. The variances sit on the diagonal, and the matrix is symmetric, equal to its own transpose.

The matrix comes from data in a single product. Put each asset's deviations from its average into a column of a matrix $\mathbf{X}$, with one row per period. Each entry of $\mathbf{X}^T\mathbf{X}$ is then the dot product of two columns, a sum of products of deviations, so

$$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

For the six months of stocks and bonds, $\mathbf{X}$ is 6 × 2, and

$$\mathbf{\Sigma} = \frac{1}{5}\begin{bmatrix} 88 & -9.9 \\ -9.9 & 4.98 \end{bmatrix} = \begin{bmatrix} 17.6 & -1.98 \\ -1.98 & 0.996 \end{bmatrix}$$

Risk managers often build the matrix the other way round, from volatilities and correlations estimated or chosen separately. Since a correlation is a covariance divided by two volatilities, each covariance is a correlation times two volatilities: $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$. Here ρ_{ij} is the correlation of assets i and j, and σ_i and σ_j are their volatilities.

Worked example — Stocks, bonds, and gold

Suppose stocks have an annual volatility of 16%, bonds 6%, and gold 15%. The correlations are 0.1 between stocks and bonds, 0.1 between stocks and gold, and 0.3 between bonds and gold. These illustrative figures are used again in this module and the next. In decimals, the covariance of stocks and bonds is $0.1 \times 0.16 \times 0.06 = 0.00096$, and the whole matrix is

$$\mathbf{\Sigma} = \begin{bmatrix} 0.0256 & 0.00096 & 0.0024 \\ 0.00096 & 0.0036 & 0.0027 \\ 0.0024 & 0.0027 & 0.0225 \end{bmatrix}$$

A covariance matrix has one more property, which follows from the next section: no portfolio can have a negative variance. Its eigenvalues, the variances of Module 5's principal components, are therefore never negative. A matrix assembled by hand from correlations that cannot all hold at once breaks this rule, and problem 6.16 gives an example.

6d. A Portfolio's Variance

A portfolio is described by its weights, $\mathbf{w}$, the shares of its money in each asset, which add to 1. Its return is the dot product $\mathbf{w} \cdot \mathbf{r}$, where $\mathbf{r}$ holds the assets' returns, as in Module 1. Its variance depends on every variance and every covariance at once:

$$\sigma_p^2 = \mathbf{w}^T\mathbf{\Sigma}\,\mathbf{w}$$

Here $\sigma_p{}^2$ is the portfolio's variance. The product $\mathbf{\Sigma w}$ is a vector, and its dot product with $\mathbf{w}$ is a single number. For two assets, with weights w_1 and w_2, volatilities σ_1 and σ_2, and correlation ρ, multiplying out gives the formula behind every discussion of diversification:

$$\sigma_p^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2\,w_1 w_2\,\rho\,\sigma_1\sigma_2$$

Worked example — A 60/40 portfolio

Put 60% in stocks with a volatility of 16% and 40% in bonds with a volatility of 6%, with a correlation of 0.1 between them. The three terms of the variance are $0.36 \times 0.0256 = 0.009216$, $0.16 \times 0.0036 = 0.000576$, and $2 \times 0.6 \times 0.4 \times 0.1 \times 0.16 \times 0.06 \approx 0.000461$. They add to about 0.01025, and the volatility is its square root, 10.13%. The weighted average of the two volatilities is 12%.

The portfolio is less risky than the average of its parts because the two assets do not move in lockstep. With a correlation of 0 the volatility would be 9.90%, and with −0.3 it would be 9.17%. With 0.5 it would be 11.00%. Only a correlation of exactly 1 gives the full 12%; short of that, combining assets always brings risk below their weighted average.

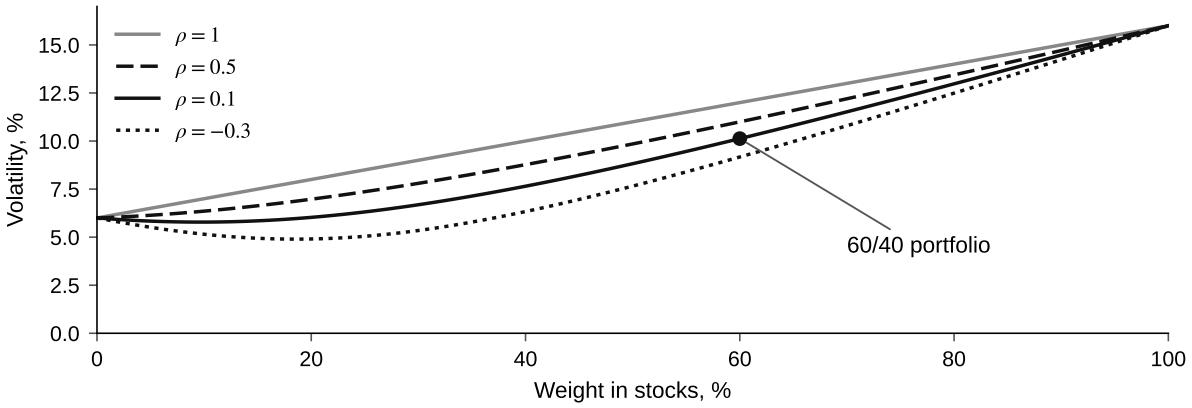


Figure 6.1. The volatility of a mix of stocks, with a volatility of 16%, and bonds, with 6%, as the weight in stocks rises. Only with a correlation of 1 is the mix's risk the weighted average of the two; the lower the correlation, the deeper the curve dips.

Worked example — Three assets

With the matrix of section 6c and weights of 50% stocks, 30% bonds, and 20% gold, $\mathbf{\Sigma w}$ = (0.01357, 0.00210, 0.00651). Its dot product with $\mathbf{w}$ is $0.5 \times 0.01357 + 0.3 \times 0.00210 + 0.2 \times 0.00651 \approx 0.00872$, so the portfolio's volatility is $\sqrt{0.00872} \approx 9.34\%$. The weighted average of the three volatilities is 12.8%.

6e. Why Diversification Works, and Where It Stops

Take n stocks, each with volatility σ, every pair with correlation ρ, and put the fraction $1/n$ of the money in each. The covariance matrix holds n variances, each σ^2, and $n(n-1)$ covariances, each $\rho\sigma^2$, and every product of two weights is $1/n^2$. So

$$\sigma_p^2 = \frac{\sigma^2}{n} + \left(1 - \frac{1}{n}\right)\rho\,\sigma^2$$

As n grows, the first term, the stocks' own risk, shrinks toward zero. The second, the risk they share, approaches $\rho\sigma^2$. The portfolio's volatility falls toward a floor of $\sigma\sqrt{\rho}$ and never reaches zero.

Number of stocks	1	2	5	10	30	100	Unlimited
Volatility	30.0%	24.2%	19.9%	18.2%	17.1%	16.6%	16.4%

Equally weighted stocks, each with a volatility of 30% and every pair with a correlation of 0.3.

Most of the benefit arrives with the first ten to thirty holdings. Beyond that, adding stocks barely helps, because what remains is the risk they share. That shared part is market risk, which Module 4 measured with beta, and finance calls it systematic risk. The part that diversifies away is called specific, or idiosyncratic, risk.

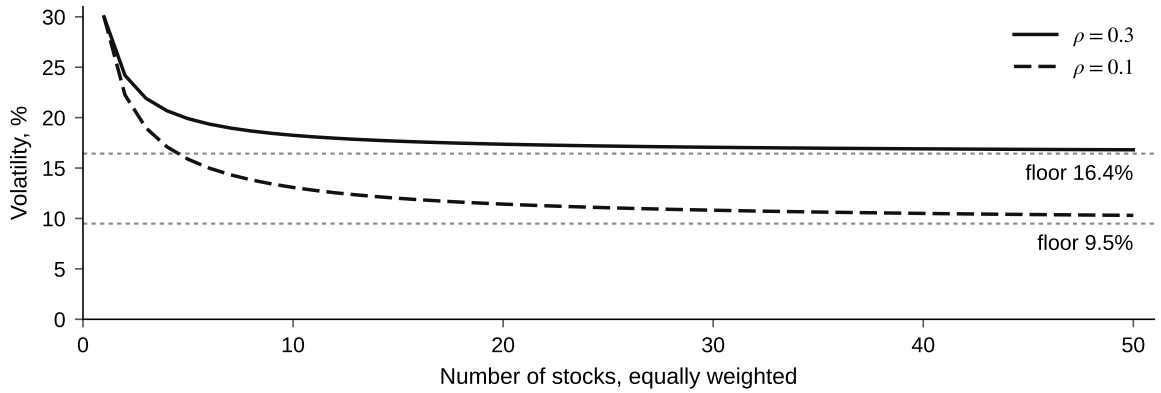


Figure 6.2. Portfolio volatility as stocks are added, each with a volatility of 30%. The stocks' own risk diversifies away quickly; the risk they share, set by their correlation, sets a floor.

Where diversification stops matters most in a crisis. Correlations among risky assets tend to rise in a sell-off, as in 2008, when stocks around the world, lower-rated corporate bonds, and commodities all fell steeply together while U.S. Treasury bonds rose. A higher ρ raises the floor just when protection is wanted. The fall of both stocks and bonds in 2022 taught the same lesson about the two assets most portfolios hold.

6f. Risk Contributions and Risk Parity

How much of a portfolio's risk does each asset bring? Asset i contributes its weight, w_i, times the ith entry of $\mathbf{\Sigma w}$, written $(\mathbf{\Sigma w})_i$. Added up over the assets, these contributions are the dot product of $\mathbf{w}$ with $\mathbf{\Sigma w}$, which is the portfolio's whole variance. So each asset's share of the risk is

$$\text{share}_i = \frac{w_i\,(\mathbf{\Sigma w})_i}{\mathbf{w}^T\mathbf{\Sigma}\,\mathbf{w}}$$

and the shares add to 100%.

Worked example — Where a 60/40 portfolio's risk comes from

For the 60/40 portfolio of section 6d, $\mathbf{\Sigma w}$ = (0.01574, 0.00202). Stocks contribute $0.6 \times 0.01574 \approx 0.00945$ and bonds $0.4 \times 0.00202 \approx 0.00081$, out of a total variance of 0.01025. Stocks are 60% of the money but 92% of the risk.

The three-asset portfolio of section 6d is similar. Stocks are half the money and 78% of the risk, bonds 30% of the money and 7% of the risk, and gold 20% of the money and 15% of the risk.

Risk parity turns this around: it chooses the weights so that every asset contributes the same share of the risk. For two assets, whatever their correlation, that means weights inversely proportional to volatility, so that

$w_1\sigma_1 = w_2\sigma_2$; problem 6.15 shows why. For stocks at 16% and bonds at 6%, the weights are 27.3% and 72.7%, and each asset brings half the risk.

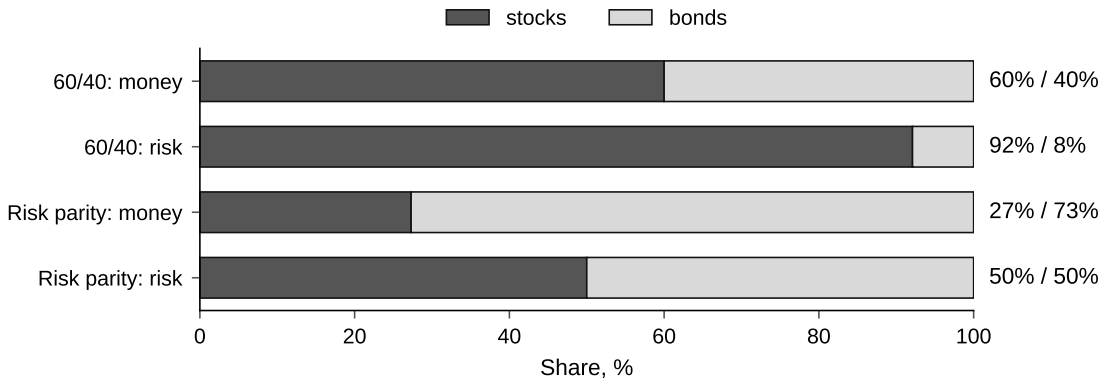


Figure 6.3. Shares of money and of risk. In a 60/40 portfolio, stocks bring 92% of the risk; risk parity holds more bonds so that each asset brings half.

The result is a low-risk portfolio, with a volatility of only 6.47%, so funds that follow the approach usually borrow to scale it up: about 1.55 times leverage brings the volatility to 10%. Bridgewater Associates launched one of the first such funds, All Weather, in 1996, and Edward Qian of PanAgora Asset Management coined the name risk parity in 2005.

The approach needs only the covariance matrix, not forecasts of returns, which are much harder to make. Its weak points are the borrowing it relies on and its large bond holdings, which suffer when bonds fall along with stocks, as they did in 2022.

In a Spreadsheet

Spreadsheets do this module's arithmetic at any size. Suppose one asset's monthly returns are in A2:A61 and another's in B2:B61:

To compute	Formula
Average	=AVERAGE(A2:A61)
Variance and standard deviation	=VAR.S(A2:A61) and =STDEV.S(A2:A61); VAR.P and STDEV.P divide by n instead of $n - 1$
Annual volatility from monthly returns	=STDEV.S(A2:A61)*SQRT(12)
Covariance and correlation	=COVARIANCE.S(A2:A61, B2:B61) and =CORREL(A2:A61, B2:B61)
Covariance matrix, with each asset's deviations in a column of D2:F61	=MMULT(TRANSPOSE(D2:F61), D2:F61)/(ROWS(D2:F61)-1)
Portfolio variance, with the covariance matrix in H2:J4 and the weights in L2:L4	=MMULT(TRANSPOSE(L2:L4), MMULT(H2:J4, L2:L4))
Each asset's share of the risk	=L2:L4*MMULT(H2:J4, L2:L4), divided by the portfolio variance

6g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Variance and standard deviation	$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ σ^2 is the variance of the observations $x_1, \ldots, x_n$, and the bar marks their average; the standard deviation σ is its square root, and for returns it is called volatility
Annualizing	$\sigma_{\text{annual}} = \sqrt{12}\,\sigma_{\text{monthly}}$ Variances add over independent periods, so volatility grows with the square root of time
Covariance	$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ σ_{xy} is the covariance of two series x_i and y_i; it is positive when they tend to be above their averages at the same times
Correlation	$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ σ_x and σ_y are the standard deviations; ρ lies between −1 and 1 and is the cosine of the angle between the two series' deviation vectors
Covariance matrix	$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$ $\mathbf{X}$ holds each asset's deviations in a column; the entries are $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, with the variances on the diagonal, and the matrix is symmetric
Portfolio variance	$\sigma_p^2 = \mathbf{w}^T\mathbf{\Sigma}\,\mathbf{w}$ $\mathbf{w}$ holds the weights; for two assets with weights w_1 and w_2, volatilities σ_1 and σ_2, and correlation ρ, ${\sigma_p}^2 = {w_1}^2{\sigma_1}^2 + {w_2}^2{\sigma_2}^2 + 2w_1w_2\rho\sigma_1\sigma_2$
Diversification limit	$\sigma_p^2 = \frac{\sigma^2}{n} + \left(1 - \frac{1}{n}\right)\rho\,\sigma^2$ For n equally weighted assets, each with volatility σ and every pair with correlation ρ; the volatility falls toward $\sigma\sqrt{\rho}$, the risk they share
Risk contribution	$\text{share}_i = \frac{w_i\,(\mathbf{\Sigma}\mathbf{w})_i}{\mathbf{w}^T\mathbf{\Sigma}\,\mathbf{w}}$ w_i is asset i's weight and $(\mathbf{\Sigma}\mathbf{w})_i$ the ith entry of $\mathbf{\Sigma}\mathbf{w}$; the shares add to 100%, and risk parity makes them equal, which for two assets means $w_1\sigma_1 = w_2\sigma_2$
The idea	Risk adds up through the covariance matrix, not through the weights alone. Diversification removes the risk that assets do not share, never the risk they do, and correlations measured in calm times can rise in a crisis.

6h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

6.1 Find the average, the sample variance, and the standard deviation of 2, 4, 4, and 6.

6.2 An asset's monthly volatility is 4%. Assuming independent months, what is its annual volatility?

6.3 Two series have deviations from their averages of (1, 0, −1) and (0, 1, −1). Find their correlation.

6.4 Two assets have volatilities of 20% and 10% and a correlation of −0.2. Write their covariance matrix.

6.5 Find the volatility of a portfolio split equally between the two assets of problem 6.4.

6.6 A covariance matrix has rows (0.09, 0.012) and (0.012, 0.04). Find the two volatilities and the correlation.

Applied

6.7 A portfolio holds 60% stocks, with a volatility of 18%, and 40% bonds, with a volatility of 7%, and their correlation is 0.3. Find the portfolio's volatility and compare it with the weighted average of the two volatilities.

6.8 An equally weighted portfolio holds 20 stocks, each with a volatility of 35%, every pair with a correlation of 0.25. Find its volatility, and the floor that adding more such stocks approaches.

6.9 A portfolio is half stocks, with a volatility of 16%, and half bonds, with 6%, and the two are uncorrelated. What share of the risk comes from stocks?

6.10 Find two-asset risk-parity weights for stocks with a volatility of 18% and bonds with 7%, uncorrelated. What is the portfolio's volatility, and what leverage brings it to 10%?

6.11 Over five months, two funds returned 2, −1, 3, −2, and 0 percent, and 1, −1, 2, −1, and −1 percent. Find their correlation.

6.12 Using the three-asset matrix of section 6c, find the volatility of a portfolio with 40% stocks, 40% bonds, and 20% gold.

Challenge

6.13 Multiply out $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ for two assets to get the two-asset formula.

6.14 Derive the formula of section 6e for n equally weighted assets.

6.15 Show that two assets contribute equally to risk exactly when $w_1\sigma_1 = w_2\sigma_2$, whatever their correlation.

6.16 Three assets have the same volatility. Can each have a correlation of −0.9 with each of the others? Test an equally weighted portfolio.

Module 6 Quiz

Check Your Understanding

1. What does a correlation of zero between two assets' returns mean?

A) Both average returns are zero
B) They move in exactly opposite directions
C) Holding both gives no diversification
D) Their deviations from average have no straight-line relationship: the deviation vectors are perpendicular

2. An asset's monthly volatility is 3%. If months are independent, what is its annual volatility?

A) 3%
B) About 10.4%
C) 36%, twelve times 3%
D) About 6%, since over a year most monthly gains and losses cancel, leaving about double the monthly figure

3. Stocks have a volatility of 16% and bonds 6%, with a correlation of 0.1. Roughly what share of a 60/40 portfolio's risk comes from stocks?

A) 60%
B) About 8%
C) About 92%
D) About 80%, the stocks' weight times volatility as a share of the total

4. Why does adding more stocks to an equally weighted portfolio eventually stop reducing its risk?

A) Each new stock is more volatile than the ones already held.
B) Trading costs grow with every stock that is added.
C) The weights become too small to affect the total risk.
D) The risk the stocks share through their correlations remains; only each stock's own risk diversifies away.

5. What does a risk-parity portfolio make equal across its assets?

A) Risk contributions
B) The money in each asset
C) Expected returns, so that no asset is favored
D) Each asset's expected return per unit of volatility, so every holding offers the same reward for its risk

Module 7 — Optimization with Constraints

The best choice under limits, and the price of each limit

Recommended Reading / Reference

Book: Allen J. Wood, Bruce F. Wollenberg, and Gerald B. Sheblé, *Power Generation, Operation, and Control* (3rd ed., 2013) — the engineers' standard text on economic dispatch, which section 7c works in miniature

Reference: Harry Markowitz, “Portfolio Selection,” *The Journal of Finance* 7, no. 1 (1952) — the fifteen-page paper that introduced the efficient frontier of section 7d

7a. Optimizing Under a Constraint

Many decisions are optimizations with limits attached. A grid operator meets demand at the lowest cost but must produce exactly what customers use. An investor wants the least risk but needs a target return, with weights that add to one. This module solves such problems with Lagrange multipliers, named for the 18th-century mathematician Joseph-Louis Lagrange, and finds that the method also puts a price on every limit.

The method needs one idea from calculus. For a function of two variables, $f(x, y)$, the partial derivative $\partial f/\partial x$ is the rate at which f changes as x changes with y held fixed, and $\partial f/\partial y$ is defined the same way. The gradient, ∇f, read del f, is the vector of the partial derivatives:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

The gradient points in the direction in which f rises fastest, and it is perpendicular to the contours of f, the curves along which f is constant. For $f(x, y) = x^2 + y^2$, the gradient is $(2x, 2y)$, so at the point (1, 2) it is (2, 4).

Worked example — The closest point on a line

Find the point on the line $x + y = 2$ closest to the origin. That means minimizing the squared distance, $f(x, y) = x^2 + y^2$, subject to the constraint $g(x, y) = 2$, where $g(x, y) = x + y$.

Moving along the line, f falls as long as the line cuts across the contours of f toward smaller values. At the minimum, the line just touches a contour without crossing it. There the line and the contour are parallel, so the perpendiculars to them, the gradients ∇f and ∇g, are parallel too:

$$\nabla f = \lambda \nabla g$$

The number λ (lambda) that scales one gradient into the other is the Lagrange multiplier. Here $\nabla f = (2x, 2y)$ and $\nabla g = (1, 1)$, so $2x = \lambda$ and $2y = \lambda$. Hence $x = y$, and the constraint gives $x = y = 1$, with $\lambda = 2$. The closest point is (1, 1), at a squared distance of 2.

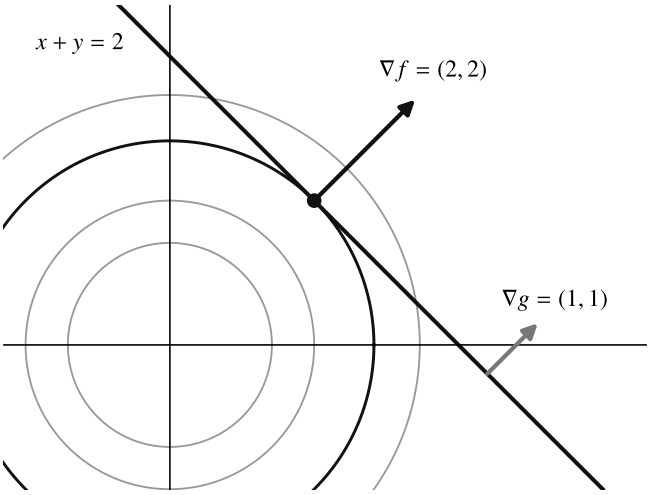


Figure 7.1. Contours of the squared distance $x^2 + y^2$ are circles. The constraint line just touches the circle through (1, 1), where the gradient of f points the same way as the gradient of the constraint, which is perpendicular to the line everywhere.

The same equations come from a recipe that works for any number of variables and constraints. Form the Lagrangian,

$$L(x, y, \lambda) = f(x, y) - \lambda\,(g(x, y) - c)$$

where c is the constraint's level, 2 here. Then set its partial derivatives with respect to x, y, and λ to zero. The first two give $\nabla f = \lambda \nabla g$, and the third gives back the constraint. Each additional constraint brings its own multiplier and its own term.

7b. The Multiplier as a Shadow Price

The multiplier is more than a device for solving equations. Let $f^*(c)$ be the best value of f that can be achieved when the constraint's level is c. Then the multiplier is the rate at which that best value changes as the limit moves:

$$\frac{df^*}{dc} = \lambda$$

Economists call the multiplier a shadow price, because it prices something, here the limit, that has no market price of its own. In the example, the line $x + y = c$ has its closest point at $(c/2, c/2)$, so $f^*(c) = c^2/2$. Its derivative is c, which is 2 at $c = 2$: exactly λ.

Moving the line out to $x + y = 2.1$ therefore raises the smallest squared distance by about $\lambda \times 0.1 = 0.2$. The exact change is $2.205 - 2 = 0.205$.

Shadow prices answer the questions managers actually ask. What is one more megawatt of transmission capacity worth? What does a tighter risk limit cost in expected return? How much would another hour of a scarce machine's time add to profit? Each answer is a Lagrange multiplier.

7c. Economic Dispatch

Every few minutes, the operator of a power grid decides how much each generator should produce to meet demand at the lowest total cost. This is economic dispatch, and it is a Lagrange problem.

Worked example — Three generators and 1,000 megawatts

Three generators must together supply 1,000 megawatts (MW). Producing P_i MW for an hour costs generator i the amount $C_i(P_i)$, in dollars:

$$C_1 = 20P_1 + 0.02P_1^2$$
$$C_2 = 22P_2 + 0.01P_2^2$$
$$C_3 = 30P_3 + 0.025P_3^2$$

Each generator's marginal cost is the derivative of its cost: the cost of one more megawatt-hour (MWh), the energy of one megawatt for an hour. The marginal costs are $C_1'(P_1) = 20 + 0.04P_1$, $C_2'(P_2) = 22 + 0.02P_2$, and $C_3'(P_3) = 30 + 0.05P_3$, in dollars per MWh.

Minimizing the total cost subject to $P_1 + P_2 + P_3 = 1{,}000$, the Lagrange condition sets every marginal cost equal to the multiplier:

$$20 + 0.04P_1 = 22 + 0.02P_2 = 30 + 0.05P_3 = \lambda$$

So $P_1 = 25(\lambda - 20)$, $P_2 = 50(\lambda - 22)$, and $P_3 = 20(\lambda - 30)$. These must add to 1,000, so $95\lambda - 2{,}200 = 1{,}000$, which gives $\lambda \approx 33.68$. The generators produce about 342, 584, and 74 MW, and the hour's total cost is $27,795.

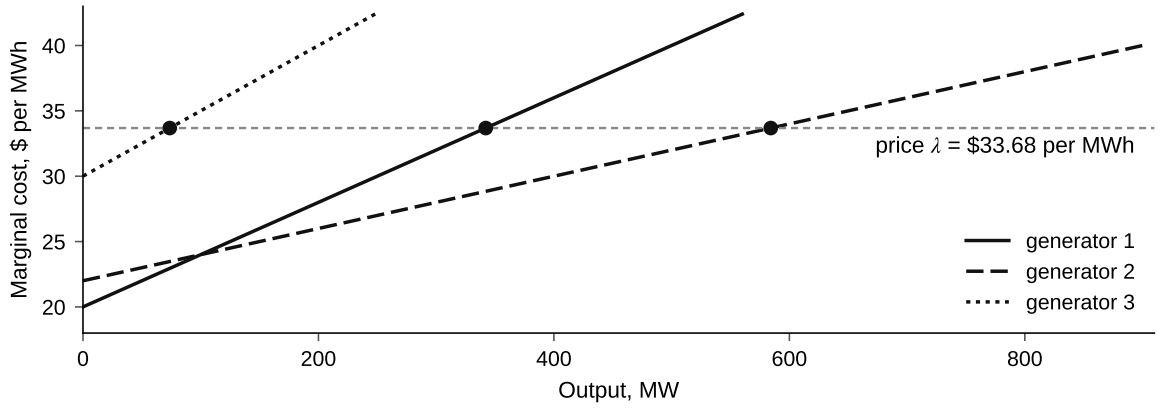


Figure 7.2. Economic dispatch. Each generator runs up its marginal-cost line until its marginal cost reaches the common price, $33.68 per MWh, at which their outputs add to the demand of 1,000 MW.

Every generator runs at the same marginal cost. If one ran at a lower marginal cost than another, shifting a megawatt to it would cut the total cost, so the cheapest plan equalizes them. Power engineers call this the equal incremental cost rule.

The multiplier, $33.68 per MWh, is the system's marginal cost: serving 1,001 MW instead of 1,000 costs $33.69 more. In a wholesale electricity market, generators offer their output at prices that reflect their marginal costs. The market then pays every dispatched generator the same clearing price, which is this multiplier. The price of power is a Lagrange multiplier.

Real dispatch adds each generator's minimum and maximum output, the cost of starting units up, and the limits of the transmission network, which Module 8 takes up. This solution also assumes that all three generators run. Below 650 MW of demand, the formulas would give generator 3 a negative output, and section 7e shows what happens instead.

7d. The Minimum-Variance Portfolio and the Efficient Frontier

Harry Markowitz framed investing as the same kind of problem in 1952: choose the weights that give the least variance for a required expected return. For that work he shared the 1990 Nobel prize in economics with Merton Miller and William Sharpe.

The ingredients are the weights $\mathbf{w}$, the covariance matrix $\mathbf{\Sigma}$ of Module 6, and the vector $\boldsymbol{\mu}$ (mu) of expected returns, each asset's average return looking forward. With $\mathbf{1}$ a vector of ones, so that $\mathbf{1}^T\mathbf{w}$ is the sum of the weights, and m the target return, the problem is:

$$\begin{aligned} &\text{minimize} && \mathbf{w}^T\mathbf{\Sigma}\,\mathbf{w} \\ &\text{subject to} && \boldsymbol{\mu}^T\mathbf{w} = m \ \text{ and } \ \mathbf{1}^T\mathbf{w} = 1 \end{aligned}$$

Start with the return target dropped, keeping only the weights adding to 1. The answer is the minimum-variance portfolio, the least risky mix of the assets. The gradient of $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$, the vector of its partial derivatives with respect to the weights, is $2\mathbf{\Sigma}\mathbf{w}$, and the gradient of the sum of the weights is $\mathbf{1}$.

So the Lagrange condition is $2\mathbf{\Sigma}\mathbf{w} = \lambda\mathbf{1}$, which makes $\mathbf{w}$ a multiple of $\mathbf{\Sigma}^{-1}\mathbf{1}$, using the inverse from Module 2. Scaling it so the weights add to 1 gives

$$\mathbf{w}_{\min} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}$$

For two assets with volatilities σ_1 and σ_2 and correlation ρ, the weight of the first is $w_1 = (\sigma_2{}^2 - \rho\sigma_1\sigma_2)/(\sigma_1{}^2 + \sigma_2{}^2 - 2\rho\sigma_1\sigma_2)$, and the second gets $w_2 = 1 - w_1$. For stocks at 16% and bonds at 6%, with a correlation of 0.1, it puts 9.7% in stocks, for a volatility of 5.78%. That is below the bonds' own 6%: a little of the riskier asset lowers the risk, because its movements partly offset the bonds'.

Worked example — Stocks, bonds, and gold

Use the covariance matrix of section 6c, with assumed expected returns of 7% for stocks, 4% for bonds, and 5% for gold. Solving $\mathbf{\Sigma}\mathbf{v} = \mathbf{1}$ for the vector $\mathbf{v} = \mathbf{\Sigma}^{-1}\mathbf{1}$ gives (28.3, 262.8, 9.9), whose entries add to 301.0. Dividing by that sum, the minimum-variance portfolio is 9.4% stocks, 87.3% bonds, and 3.3% gold, with a volatility of 5.76% and an expected return of 4.31%.

Now restore the return target and solve again for each target m. The least volatility at each expected return traces a curve. Its upper half, from the minimum-variance portfolio upward, is the efficient frontier: the portfolios with the highest expected return for their risk. Every other portfolio is dominated, since some frontier portfolio offers the same return with less risk.

The table gives the frontier for the three assets with every weight kept between 0 and 1, which rules out borrowing and short sales, as most funds must. Those limits are inequalities, the subject of section 7e, and a spreadsheet's Solver finds each row in a moment.

Expected return	Stocks	Bonds	Gold	Volatility
4.31%: minimum variance	9.4%	87.3%	3.3%	5.76%
5.0%	29.9%	59.8%	10.2%	6.81%
5.5%	44.9%	39.8%	15.3%	8.52%
5.8%	53.9%	27.8%	18.4%	9.75%
6.0%	59.9%	19.7%	20.4%	10.62%
6.5%	75.0%	0.0%	25.0%	12.93%
7.0%	100.0%	0.0%	0.0%	16.00%

The efficient frontier for stocks, bonds, and gold with no short sales, from the covariance matrix of section 6c and assumed expected returns of 7%, 4%, and 5%. Weights may not add to exactly 100% because of rounding.

The familiar 60/40 portfolio, with an expected return of 5.8% and a volatility of 10.13%, lies inside the frontier. The frontier portfolio with the same return, 54% stocks, 28% bonds, and 18% gold, has a volatility of 9.75%. Equal weights do worse: a third in each asset expects 5.33% with a volatility of 8.42%, while the frontier delivers that return with 7.89%.

Gold earns its place even though it is far riskier than bonds, because it adds expected return with only a low correlation to stocks.

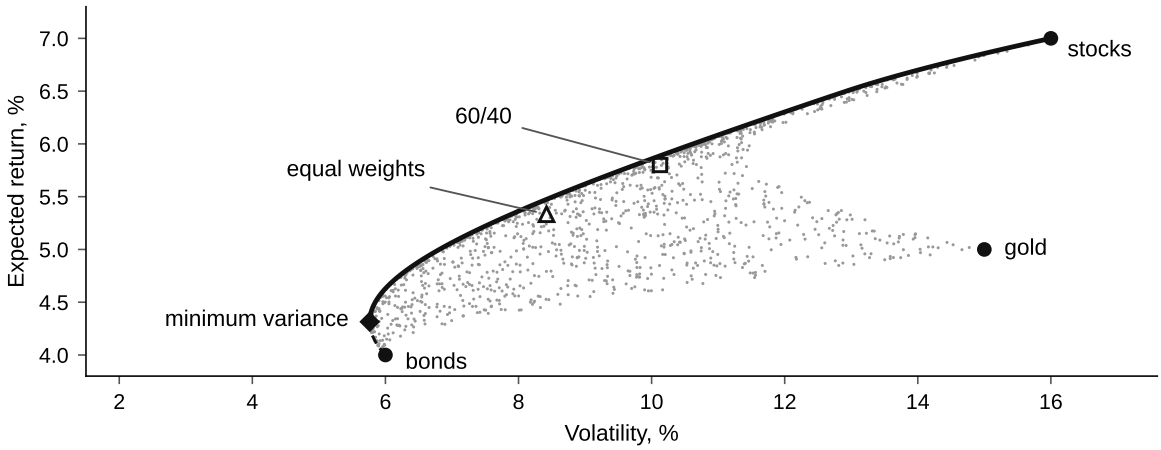


Figure 7.3. Gray dots are 1,500 random mixes of stocks, bonds, and gold with no short sales. The solid curve is the efficient frontier; the dashed curve below it holds the least risky mixes for returns below the minimum-variance portfolio's, which no one should choose.

The frontier is only as good as its inputs. Expected returns are hard to estimate, and the optimizer leans hardest on whichever assets look best, so it magnifies errors in the estimates. Raising gold's assumed return from 5% to 6% moves the frontier portfolio at 5.8% from 54% stocks and 18% gold to 41% stocks and 29% gold.

Practitioners therefore cap the weights, or blend the estimated returns toward neutral values, much as the ridge penalty of Module 4 shrinks coefficients. Others rely on the covariance matrix alone, as minimum-variance and risk-parity portfolios do.

7e. Inequality Constraints

Many limits are inequalities. A generator cannot produce more than its capacity or less than zero, and a fund that cannot sell short must keep every weight at zero or above. At the optimum, each inequality either binds, holding exactly, or has room to spare. A limit with room to spare does not shape the solution, so its multiplier is zero.

These conditions, added to the Lagrange conditions, are the Karush–Kuhn–Tucker (KKT) conditions. William Karush wrote them down in a 1939 master's thesis, and Harold Kuhn and Albert Tucker published them in 1951. Their key condition is complementary slackness:

$$\lambda_j \, s_j = 0 \qquad \lambda_j \geq 0, \quad s_j \geq 0$$

Here s_j is the slack, the room left under limit j, and λ_j is that limit's multiplier. At least one of the two is zero: either the limit binds, with no slack, or its shadow price is zero. When a cost is being minimized, a limit's multiplier is never negative, since loosening a limit can never raise the lowest achievable cost.

Worked example — A capacity limit

Suppose generator 2 of section 7c can produce at most 500 MW. The unlimited plan asked it for 584 MW, so the limit binds. Generator 2 runs at 500 MW, and generators 1 and 3 share the other 500 MW at a common marginal cost: $25(\lambda - 20) + 20(\lambda - 30) = 500$, so $\lambda \approx 35.56$. They produce about 389 and 111 MW, and the hour costs $27,944, which is $150 more than without the limit.

Generator 2's marginal cost at 500 MW is $22 + 0.02 \times 500 = 32$ dollars per MWh, below the system price of $35.56. The gap, $3.56 per MWh, is the capacity limit's shadow price: each extra megawatt of capacity at generator 2 would save about $3.56 an hour. With a limit of 600 MW instead, the limit would have room to spare, and its shadow price would be zero.

Lower limits work the same way. With only 600 MW of demand, equal marginal costs for all three generators would require a negative output from generator 3. Instead its output stays at zero and its lower limit binds. Generators 1 and 2 meet the demand at a price of $29.33, below the $30 marginal cost of generator 3's first megawatt.

The no-short-sales frontier of section 7d shows the same pattern. Above an expected return of about 6.5%, the solution without limits would sell bonds short, borrowing them to buy more stocks and gold. The limit of zero on the bond weight then binds, and the frontier bends away from the one without limits. Below that return the limit has room to spare, and the two frontiers coincide.

In a Spreadsheet

Spreadsheets solve constrained problems with Excel's Solver add-in. With the covariance matrix in B2:D4, a column of three 1s in F2:F4, and the weights in H2:H4:

To compute	Formula
Minimum-variance weights	=MMULT(MINVERSE(B2:D4), F2:F4)/SUM(MMULT(MINVERSE(B2:D4), F2:F4))
Portfolio volatility	=SQRT(MMULT(TRANSPOSE(H2:H4), MMULT(B2:D4, H2:H4)))
A frontier portfolio	Solver: minimize the volatility cell by changing H2:H4, with the weights adding to 1, the expected return equal to the target, and each weight at least 0, using the GRG Nonlinear method (generalized reduced gradient)
Economic dispatch	Goal Seek, under What-If Analysis: set the total-output cell to the demand by changing the cell that holds λ
Each limit's shadow price	Solver's Sensitivity Report lists the Lagrange multiplier of every constraint

7f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Partial derivative and gradient	$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$ $\partial f/\partial x$ is the rate of change of f as x changes with y held fixed; the gradient ∇f points where f rises fastest, perpendicular to its contours
Lagrange condition	$\nabla f = \lambda \nabla g$ At the best point of f on the constraint $g(x, y) = c$, the two gradients are parallel; λ is the Lagrange multiplier
Lagrangian	$L = f - \lambda (g - c)$ Set its partial derivatives with respect to every variable, and to λ, equal to zero
Shadow price	$\frac{df^*}{dc} = \lambda$ $f^*(c)$ is the best value achievable at constraint level c; the multiplier is the marginal value of the limit
Economic dispatch	$C_1'(P_1) = C_2'(P_2) = \cdots = \lambda$ Generator i produces P_i at cost $C_i(P_i)$; every generator between its limits runs at the same marginal cost, λ, which is the price of power
Minimum-variance portfolio	$\mathbf{w}_{\min} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}$ $\mathbf{\Sigma}$ is the covariance matrix and $\mathbf{1}$ a vector of ones; for two assets with volatilities σ_1 and σ_2 and correlation ρ, $w_1 = (\sigma_2{}^2 - \rho\sigma_1\sigma_2)/(\sigma_1{}^2 + \sigma_2{}^2 - 2\rho\sigma_1\sigma_2)$
Efficient frontier	The portfolios with the highest expected return for their risk; any portfolio inside it is dominated by a frontier portfolio with the same return and less risk
Complementary slackness	$\lambda_j s_j = 0$ s_j is the room left under limit j and λ_j its multiplier: either the limit binds or its shadow price is zero; with the Lagrange conditions, these are the Karush–Kuhn–Tucker (KKT) conditions
The idea	At a constrained optimum the gradients line up, and each multiplier prices its limit. The price of power and the cost of a risk limit are both Lagrange multipliers.

7g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

7.1 Find the gradient of $f(x, y) = x^2 + 3xy$ at the point (1, 2).

7.2 Minimize $x^2 + y^2$ subject to $x + 2y = 5$.

7.3 Maximize xy subject to $x + y = 10$.

7.4 Use the multiplier from problem 7.3 to estimate the maximum of xy when the constraint is $x + y = 10.2$, and compare with the exact answer.

7.5 Two uncorrelated assets have volatilities of 20% and 10%. Find the minimum-variance weights and volatility.

7.6 Two generators have marginal costs of $20 + 0.05P_1$ and $30 + 0.02P_2$ dollars per MWh. If the price is $40 per MWh, how much does each produce?

Applied

7.7 For the generators of problem 7.6, find the price and each output when demand is 600 MW.

7.8 In problem 7.7, demand rises by 1 MW. How is the extra megawatt shared, and how much does the price rise?

7.9 In the dispatch of section 7c, suppose generator 1 can produce at most 300 MW. Find the new price, the outputs, and the limit's shadow price.

7.10 Find the minimum-variance portfolio of the two assets in problem 6.4, with volatilities of 20% and 10% and a correlation of −0.2.

7.11 A lab splits a budget of $17 million between two research programs. With x and y the millions spent on each, the value created is $10\sqrt{x} + 6\sqrt{y}$ million dollars. Find the best split and the value of one more million of budget.

7.12 A 50/50 mix of the stocks and bonds of section 7d has an expected return of 5.5%. Find its volatility, and compare it with the frontier portfolio that has the same return.

Challenge

7.13 Derive the two-asset minimum-variance weight by writing the portfolio variance as a function of w_1 alone, with $w_2 = 1 - w_1$, and setting its derivative to zero.

7.14 Starting from the Lagrangian $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} - \lambda(\mathbf{1}^T\mathbf{w} - 1)$, show that the minimum-variance portfolio is $\boldsymbol{\Sigma}^{-1}\mathbf{1}$ divided by $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$.

7.15 Generators with costs $C_i = a_iP_i + b_iP_i^2$, where a_i and b_i are positive constants, all run and together meet demand D. Show that the price is $\lambda = (D + \Sigma a_i/(2b_i))/\Sigma(1/(2b_i))$, with the sums taken over the generators, and check it against section 7c.

7.16 In the dispatch of section 7c, generator 1 is limited to 400 MW. Does the limit bind, and what is its shadow price?

Module 7 Quiz

Check Your Understanding

1. In a constrained optimization, what does a constraint's Lagrange multiplier measure?

A) The best value itself
B) How much the best value changes per unit change in the limit
C) The slope of the constraint
D) The fraction of the limit that is used up at the optimum, a number that always lies between zero and one

2. In economic dispatch, every generator running between its limits produces at the same marginal cost. What is that common marginal cost?

A) The average cost of all the power produced in the hour
B) The fuel price
C) The system's price of power
D) The marginal cost of the cheapest generator, which always runs at full output before any other unit starts

3. What does complementary slackness say about an inequality constraint at the optimum?

A) It always binds
B) Its multiplier is negative
C) Either the limit binds, or its multiplier is zero: a limit with room to spare has a shadow price of zero
D) Its multiplier equals the room left under the limit at the optimum

4. Which portfolio sits at the far left tip of the efficient frontier, where risk is lowest?

A) Equal weights
B) The minimum-variance portfolio
C) The least volatile asset
D) The risk-parity portfolio of the assets

5. What is the efficient frontier?

A) The best return for each level of risk
B) Every portfolio that holds all of the assets in equal proportions, so that no single position dominates
C) The single asset with the highest expected return, whatever its risk
D) The portfolios whose assets have the lowest average volatility

Module 8 — Linear Programming

Linear objectives, linear limits, and the prices hidden in the solution

Recommended Reading / Reference

Book: Daniel S. Kirschen and Goran Strbac, *Fundamentals of Power System Economics* (2nd ed., 2019) — how power markets dispatch generators and set prices by location, the subject of sections 8e and 8f

Reference: PJM Interconnection, Data Miner 2 (dataminer2.pjm.com) — free hourly locational marginal prices for thousands of points on the grid, each split into its energy, congestion, and loss components

8a. Linear Objectives, Linear Limits

Module 7 optimized smooth functions under a few limits. Many of the largest optimization problems in business are simpler in one way and harder in another. The objective and every limit are linear, but there are thousands or millions of variables and limits, most of them inequalities. These are linear programs, an old usage in which a program is a plan.

George Dantzig devised the simplex method for solving them in 1947, while planning for the U.S. Air Force. Leonid Kantorovich had formulated similar problems in the Soviet Union in 1939, and he and Tjalling Koopmans shared the 1975 Nobel prize in economics for the theory of allocating resources. Oil refiners were among the first businesses to use linear programming, in the early 1950s, and grid operators now solve dispatch problems of this kind every few minutes.

Worked example — Training or serving

An artificial intelligence (AI) data center has 100 pods, blocks of servers packed with graphics processing units (GPUs), and 150 megawatts (MW) of power. A pod used for training models draws 2 MW and earns $40 thousand a day. A pod used for answering users' requests, called inference, draws 1 MW and earns $30 thousand a day. With x pods training and y serving, the operator wants to

$$\begin{aligned} \text{maximize} \quad & 40x + 30y \\ \text{subject to} \quad & x + y \le 100 \\ & 2x + y \le 150 \\ & x \ge 0, \;\; y \ge 0 \end{aligned}$$

The first limit counts pods and the second megawatts, and profit is in thousands of dollars a day. Every term is a number times a variable, so the problem is linear.

In general, a linear program chooses a vector of variables $\mathbf{x}$ to maximize $\mathbf{c}^T\mathbf{x}$, where $\mathbf{c}$ holds the payoff per unit of each variable, subject to $\mathbf{A}\mathbf{x} \le \mathbf{b}$ and $\mathbf{x} \ge \mathbf{0}$. Each row of the matrix $\mathbf{A}$ is one limit, and the vector $\mathbf{b}$ holds the amounts available. An inequality between vectors holds entry by entry.

8b. The Feasible Region and Its Corners

The points that satisfy every limit form the feasible region. Each linear inequality keeps the points on one side of a line, so the region is a polygon, and in more dimensions a many-sided solid. For the data center, its corners are (0, 0), (75, 0), (50, 50), and (0, 100).

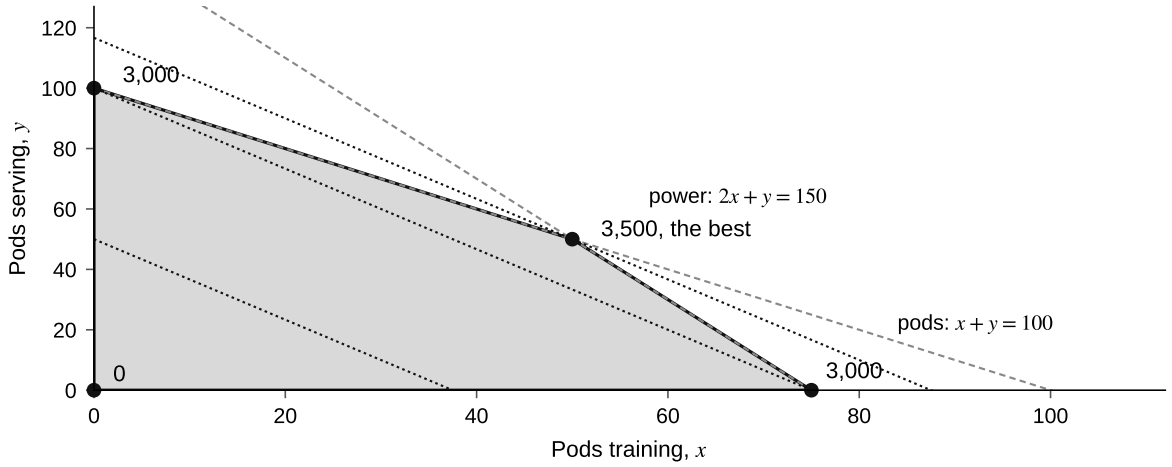


Figure 8.1. The feasible region, shaded, with the profit at each corner in thousands of dollars a day. The dotted lines are lines of equal profit, at 1,500, 3,000, and 3,500; the highest one that still touches the region meets it at the corner (50, 50).

Lines along which profit is constant, such as $40x + 30y = 3{,}000$, are parallel. Slid toward higher profit, the last point of the region they touch is a corner, here (50, 50), with a profit of $3.5 million a day. Half the pods train and half serve. Both limits bind: every pod and every megawatt is in use.

Corner (x, y)	Pods used	Megawatts used	Profit ($ thousand a day)
(0, 0)	0	0	0
(75, 0)	75	150	3,000
(50, 50)	100	150	3,500
(0, 100)	100	100	3,000

This is the central fact of linear programming. If a linear program has a best solution, one of the corners is best. A linear objective rises steadily in one direction and never curves back, so it keeps improving until the region runs out, and the region runs out at a corner. When the lines of equal profit run parallel to an edge, every point on that edge ties.

8c. The Simplex Method

Checking every corner works for two variables but not for two thousand, since the number of corners explodes. The simplex method walks instead. It starts at a corner, moves along an edge to a neighboring corner with a better objective, and repeats until no neighbor is better. Because the feasible region has no dents, a corner with no better neighbor is the best corner of all.

Each inequality first gets a slack variable, the unused amount of its resource, which turns it into an equation: $x + y + s_1 = 100$ and $2x + y + s_2 = 150$, with s_1 and s_2 zero or above. At each corner, two of the four variables are zero, and each move from corner to corner is an elimination step of the kind in Module 3.

Worked example — The walk

Start at (0, 0), where the slacks are 100 pods and 150 MW. Training earns more per pod, so increase x. Power runs out first, at $x = 75$, and the walk reaches (75, 0), with a profit of 3,000. Written in terms of the two variables now at zero, the profit is $3{,}000 + 10y - 20s_2$.

Increasing y still adds 10 a unit, because each serving pod brought in displaces half a training pod: $30 - 40/2 = 10$. The pods run out at $y = 50$, and the walk reaches (50, 50). There the profit becomes $3{,}500 - 20s_1 - 10s_2$. Raising either slack, which means leaving a pod or a megawatt idle, can only lower profit, so the walk stops. The two numbers in that last line, 20 and 10, reappear in section 8d as prices.

Real solvers handle millions of variables, with refinements that skip most of the corners. The main alternative, interior-point methods, cuts through the inside of the region instead of walking its edges; it grew from an algorithm Narendra Karmarkar published in 1984.

8d. Duality and Shadow Prices

Every linear program has a twin, its dual, which prices the resources instead of choosing the activities. Give each pod a value u and each megawatt a value v, in thousands of dollars a day. A sound valuation must be at least as high as the profit each activity earns from the resources it uses. A training pod uses one pod and 2 MW, so $u + 2v \geq 40$; a serving pod uses one pod and 1 MW, so $u + v \geq 30$. The dual asks for the cheapest such valuation of everything the data center has:

$$\begin{aligned} \text{minimize} \quad & 100u + 150v \\ \text{subject to} \quad & u + 2v \geq 40 \\ & u + v \geq 30 \\ & u \geq 0, \;\; v \geq 0 \end{aligned}$$

Its solution is $u = 20$ and $v = 10$, and its minimum is $100 \times 20 + 150 \times 10 = 3{,}500$, exactly the best profit. That equality is the strong duality theorem: whenever a linear program has a best solution, so does its dual, and their best values are equal.

The dual solution holds the shadow prices of Module 7. One more pod is worth $20 thousand a day, and one more megawatt $10 thousand a day. With 151 MW, the best plan becomes 51 training pods and 49 serving, and the profit rises by exactly 10. Complementary slackness holds here too: a resource with room to spare would have a shadow price of zero.

In general, the dual has one variable for each limit of the original, called the primal, and uses the transpose of the same matrix, with a vector $\mathbf{p}$ of prices:

$$\begin{aligned} \text{primal:} \quad & \text{maximize } \mathbf{c}^T\mathbf{x} \;\; \text{subject to } \mathbf{A}\mathbf{x} \leq \mathbf{b}, \;\; \mathbf{x} \geq \mathbf{0} \\ \text{dual:} \quad & \text{minimize } \mathbf{b}^T\mathbf{p} \;\; \text{subject to } \mathbf{A}^T\mathbf{p} \geq \mathbf{c}, \;\; \mathbf{p} \geq \mathbf{0} \end{aligned}$$

Shadow prices tell a manager which limit to relax first and what relaxing it is worth. They hold only over a range: here, power keeps its $10 value as long as the best plan still uses every pod and every megawatt.

8e. Scheduling a Battery

A battery earns money by charging when power is cheap and discharging when it is dear. Some energy is lost on the round trip: with a round-trip efficiency η (eta), each megawatt-hour (MWh) bought returns η MWh. A charge and discharge therefore pays only if the selling price times η exceeds the buying price. At 85% efficiency, power bought at $40 per MWh must sell above $40/0.85 \approx \$47$.

A whole day is a linear program. For each hour t, let p_t be the price, c_t the power used to charge, d_t the power discharged, and s_t the energy stored at the end of the hour. The battery maximizes its net revenue,

$$\text{maximize} \quad \sum_{t=1}^{24} p_t \, (d_t - c_t)$$

subject to the storage balance $s_t = s_{t-1} + \eta c_t - d_t$ in every hour, the power limits $0 \le c_t \le P$ and $0 \le d_t \le P$, and the energy limit $0 \le s_t \le E$. Here P is the battery's power rating and E its energy capacity. Every term is linear.

Worked example — A day of prices

A battery of 10 MW and 40 MWh, a four-hour battery, has 85% round-trip efficiency and starts and ends the day empty. The prices follow a summer pattern: high twenties overnight, a small morning peak, a dip around midday, and a steep evening peak to $110 per MWh. Solving the linear program, with 72 variables, gives the schedule in Figure 8.2.

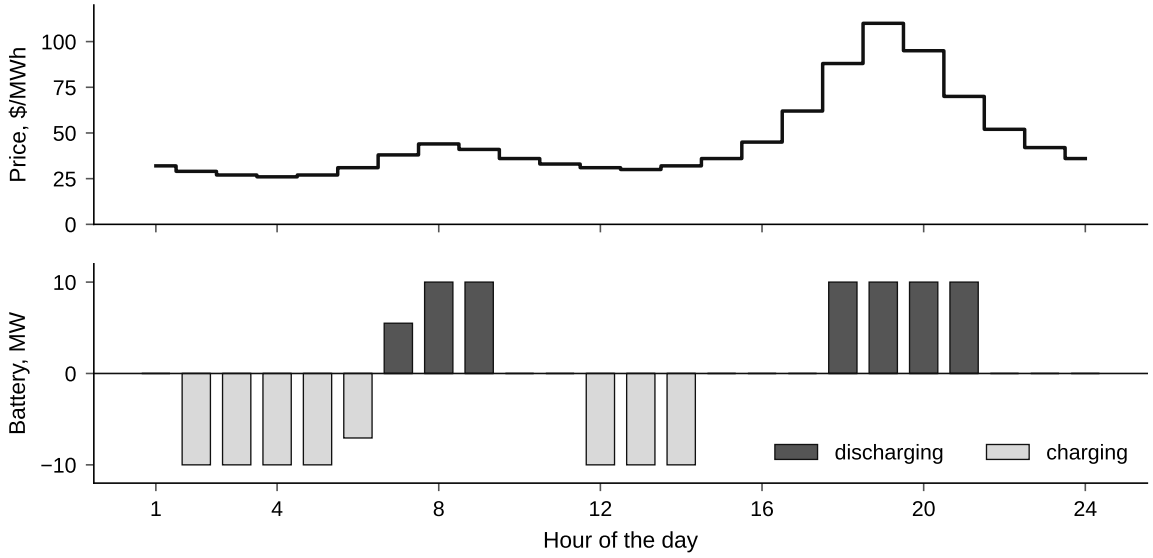


Figure 8.2. The best schedule for a 10 MW, 40 MWh battery over one day of prices. It charges overnight, sells into the morning peak, recharges at midday, and discharges into the evening peak.

The solver finds two cycles, not one. It charges overnight, sells part of that energy into the morning peak, recharges at midday, and discharges fully into the evening peak. Over the day it buys 77.1 MWh for $2,239 and sells 65.5 MWh for $4,689, a profit of $2,450. The best single cycle earns $2,328, so the morning cycle, easy to miss by eye, adds about $122.

The limits have shadow prices, as always. Re-solving with one more MWh of energy capacity raises the day's profit by about $26, and with one more MW of power by about $111, so per unit, a megawatt of extra power is worth about four times a megawatt-hour of extra storage. Many battery warranties also cap the energy discharged per day; a cap of one full cycle, 40 MWh, would cost this battery about $122 a day.

8f. Transmission Limits and Locational Prices

Module 3 found that power sent from bus 1 to bus 3 of a three-bus triangle, with identical lines, splits two to one: two-thirds on the direct line and one-third through bus 2. Now put a cheap generator at bus 1, offering up to 200 MW at $20 per MWh, and a dearer one at bus 2, offering up to 200 MW at $40. A load of 100 MW sits at bus 3, and the direct line from bus 1 to bus 3 can carry at most 50 MW.

Without that limit, the cheap generator would supply all 100 MW for $2,000 an hour, but 66.7 MW of it would cross the direct line. With g_1 the output of generator 1 and g_2 that of generator 2, the flow on the direct line is $(2/3)g_1 + (1/3)g_2$, and the dispatch is a linear program:

$$\begin{aligned} \text{minimize} \quad & 20g_1 + 40g_2 \\ \text{subject to} \quad & g_1 + g_2 = 100 \\ & \tfrac{2}{3}g_1 + \tfrac{1}{3}g_2 \le 50 \\ & 0 \le g_1 \le 200, \;\; 0 \le g_2 \le 200 \end{aligned}$$

The solution splits the load evenly, 50 MW from each generator, for $3,000 an hour. The direct line carries exactly its 50 MW limit, the line from bus 2 to bus 3 carries the other 50 MW, and the line between buses 1 and 2 carries nothing.

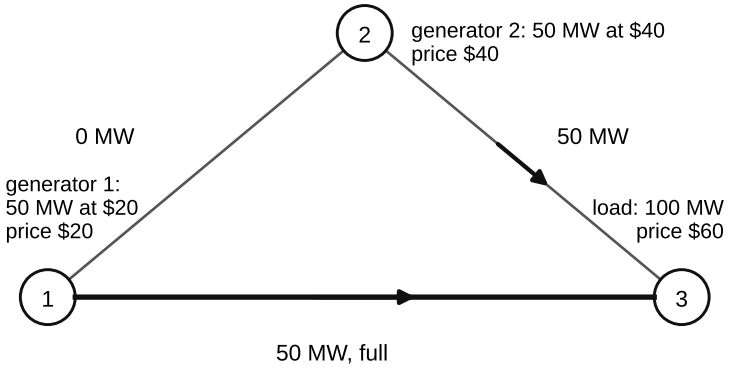


Figure 8.3. Dispatch with the direct line from bus 1 to bus 3 at its 50 MW limit. Each generator supplies half the load, and the price differs at every bus.

The price of power is now different at each bus. A locational marginal price (LMP) is the cost of serving one more megawatt at a bus, at the lowest cost the grid's limits allow. At bus 1 it is $20, since generator 1 can serve a local load without sending power over any line, and at bus 2 it is $40 for the same reason.

At bus 3 it is $60, more than either generator charges. One more MW at bus 3 cannot come from generator 1 alone, because two-thirds of it would land on the full line. It takes 2 MW more from generator 2 and 1 MW less from generator 1. That changes the direct line's flow by $(2/3)(-1) + (1/3)(2) = 0$, at a cost of $2 \times 40 - 20 = 60$ dollars.

The line's limit has a shadow price too. With 51 MW of room on the direct line, generator 1 could produce 3 MW more and generator 2 3 MW less, saving $3 \times (40 - 20) = 60$ dollars an hour. The line's congestion price is $60 per MW of capacity per hour.

Every bus's price splits the same way. With bus 3 as the reference, each bus's price is the reference price minus the congestion price times the bus's shift factor on the full line: $60 - 60 \times 2/3 = 20$ at bus 1 and

$60 - 60 \times 1/3 = 40$ at bus 2. PJM publishes each of its locational prices split into an energy component, a congestion component, and a loss component, which this lossless model leaves out.

Buyers at bus 3 pay $60 for 100 MW, $6,000 an hour, while the two generators receive $1,000 and $2,000. The $3,000 left over is congestion rent, equal to the line's $60 congestion price times its 50 MW of flow. Market operators pay it out to holders of financial transmission rights (FTRs), contracts that pay the difference in price between two points and so let a buyer at bus 3 hedge the cost of power bought at bus 1.

In a Spreadsheet

Spreadsheets solve linear programs exactly, with Solver's Simplex LP method, LP standing for linear programming. Suppose the variables are in B2:C2, their payoffs per unit in B3:C3, each limit's coefficients in a row of B5:C6, and the amounts available in E5:E6:

To compute	Formula
Objective	=SUMPRODUCT(B2:C2, B3:C3)
Amount of each resource used	=SUMPRODUCT($B$2:$C$2, B5:C5), filled down
Solve	Solver: maximize the objective by changing B2:C2, with each amount used at most its limit in E5:E6, Make Unconstrained Variables Non-Negative checked, and Simplex LP as the method
Shadow prices	Solver's Sensitivity Report: the Shadow Price column, with the range over which each price holds
A day of battery operation	One row per hour, with charge, discharge, and stored energy as variables, and each hour's storage balance as an equality constraint

8g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Linear program	$\text{maximize } \mathbf{c}^T\mathbf{x} \text{ subject to } \mathbf{A}\mathbf{x} \le \mathbf{b},\ \mathbf{x} \ge \mathbf{0}$ $\mathbf{x}$ holds the decision variables, $\mathbf{c}$ the payoff per unit of each, each row of $\mathbf{A}$ one limit, and $\mathbf{b}$ the amounts available
Corner principle	If a linear program has a best solution, a corner of the feasible region is best; the simplex method walks from corner to better corner
Slack variable	The unused amount of a limit: $x + y \le 100$ becomes $x + y + s = 100$, with the slack s zero or above
Duality	$\text{minimize } \mathbf{b}^T\mathbf{p} \text{ subject to } \mathbf{A}^T\mathbf{p} \ge \mathbf{c},\ \mathbf{p} \ge \mathbf{0}$ The dual has one price in $\mathbf{p}$ for each limit; at the optimum its value equals the primal's, and its solution is the limits' shadow prices
Battery arbitrage	A charge and discharge pays only if $\eta p_{\text{sell}} > p_{\text{buy}}$ η is the round-trip efficiency, p_{buy} the price paid to charge, and p_{sell} the price received on discharge
Locational marginal price	The cost of serving one more MW at a bus within the grid's limits; with a congested line, prices differ by bus and can exceed every generator's offer
Price components	$\text{LMP}_k = \lambda - \mu\,\varphi_k$ λ is the price at the reference bus, μ (mu) the congestion price of the full line, and φ_k (phi) the shift factor of bus k on that line
Congestion rent	Buyers' payments minus sellers' receipts: the congestion price times the flow on the full line, paid out to holders of financial transmission rights
The idea	A linear program's best plan sits at a corner, and its dual prices every limit. The price of power at each point on a grid is one of those dual prices.

8h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

8.1 Maximize $3x + 2y$ subject to $x + y \le 4$, $x + 3y \le 6$, $x \ge 0$, and $y \ge 0$ by checking the corners.

8.2 Write the dual of problem 8.1 and check that its best value is also 12.

8.3 Add slack variables to the data center problem and give their values at the corners (75, 0) and (50, 50).

8.4 In problem 8.1, which limits bind at the best corner, and what are their shadow prices?

8.5 A battery with 90% round-trip efficiency can buy power at $30 per MWh and sell later at $34. Does the trade pay?

8.6 In the three-bus example, generator 1 produces 80 MW and generator 2 produces 20 MW for the load at bus 3. What flows on the direct line from bus 1 to bus 3?

Applied

8.7 The data center's power rises from 150 MW to 160 MW. Find the new best plan and profit, and compare the gain with the shadow price.

8.8 Over what range of power does the data center's shadow price of $10 thousand per MW hold?

8.9 A battery of 1 MW and 1 MWh with 80% round-trip efficiency faces prices of $20, $50, $30, and $80 in four hours, starting and ending empty. Find its best schedule and profit.

8.10 In the three-bus example, the direct line's limit is raised to 60 MW. Find the dispatch, the cost, and the prices.

8.11 In the three-bus example, a load of 20 MW is added at bus 1. How do the dispatch and the cost change?

8.12 A financial transmission right of 10 MW from bus 1 to bus 3 pays the difference in price between the two buses. What does it pay in an hour of the three-bus example?

Challenge

8.13 Show that for the data center, any feasible production (x, y) and any feasible dual prices (u, v) satisfy $40x + 30y \le 100u + 150v$.

8.14 Start the simplex method at the corner (0, 100) of the data center problem and take one step. What does the final profit line show?

8.15 Show that in the three-bus example the congestion rent equals the congestion price times the line's flow, using each bus's price $\lambda - \mu\varphi_k$.

8.16 Explain why, when prices are positive, the best battery schedule never charges and discharges in the same hour. What can change when prices are negative?

Module 8 Quiz

Check Your Understanding

1. Why does a linear program's best solution lie at a corner of the feasible region?

A) Corners are the only points that satisfy every one of the limits.

B) The simplex method can only visit corners, so the points between them are never even considered by the solver.

C) A linear objective is largest wherever two limits cross at a right angle, and each corner is such a crossing.

D) A linear objective improves steadily in one direction until the region ends, at a corner.

2. In the data center example, the shadow price of power is $10 thousand per MW a day. What does that mean?

A) Power costs $10 thousand per MW a day

B) One more MW would add about $10 thousand a day to the best profit

C) Each MW in use lowers the day's profit by $10 thousand, so idle power is better

D) The operator should be willing to pay $10 thousand per MW a day for any amount of extra power, however much it buys

3. What does strong duality say about a linear program and its dual?

A) The dual always has fewer variables than the original problem.

B) The dual's solution matches the original's, variable by variable.

C) When both have best solutions, their best values are equal, so the shadow prices account for every dollar of the best profit.

D) The dual is always easier to solve, which is why solvers use it.

4. A battery with 80% round-trip efficiency charges at $40 per MWh. What selling price just breaks even?

A) About $50

B) About $48: $40 plus 20% to cover the loss

C) $40, the price it paid, since the stored energy keeps its value

D) About $32, since only 80% of the energy comes back out

5. In the three-bus example, why is the price at bus 3 higher than either generator's offer?

A) Transmission losses

B) A shortage of generating capacity

C) A price cap

D) Redispatch around the full line

Module 9 — Convexity and Gradient Methods

When a minimum is the minimum, and how large models find it

Recommended Reading / Reference

Book: Stephen Boyd and Lieven Vandenberghe, *Convex Optimization* (2004) — the standard text on convex problems, free online from its authors

Reference: Sebastian Ruder, “An Overview of Gradient Descent Optimization Algorithms” (2016) — a free survey of momentum, Adam, and their relatives, the methods of sections 9e and 9f

9a. Convex and Nonconvex Problems

The methods of Modules 7 and 8 find points where nothing nearby is better. Whether such a point is the best anywhere depends on the shape of the problem. In a convex problem it always is, and that property separates the problems that can be solved reliably at enormous scale from those that can only be attacked.

A function is convex if the straight line between any two points on its graph never dips below the graph. In symbols, for any points x and y and any fraction t from 0 to 1,

$$f\big(t\,x + (1-t)\,y\big) \le t\,f(x) + (1-t)\,f(y)$$

A bowl is convex; a landscape of hills and valleys is not. For a function of one variable with a second derivative, convexity means $f''(x) \ge 0$ everywhere: the slope never decreases. A set of points is convex if the segment between any two of its points stays inside it, as the feasible region of a linear program does.

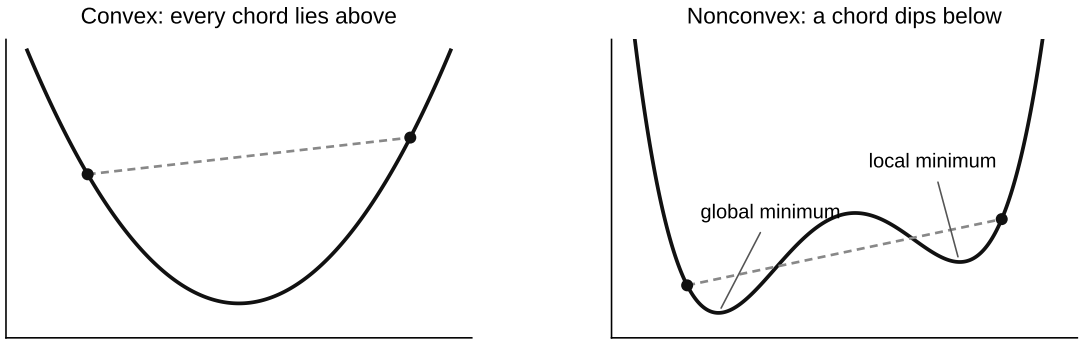


Figure 9.1. A convex function, left, has one valley, and the chord between any two points lies above it. The nonconvex function, right, has two valleys; descending from the wrong side ends in the higher one.

The payoff is a theorem: for a convex function minimized over a convex set, every local minimum is a global minimum. A method that only ever goes downhill cannot get stuck in the wrong valley. Linear programs, least squares, the minimum-variance portfolio, and economic dispatch with rising marginal costs are all convex.

Training a neural network is not. Its loss, the measure of its errors on the training data, has countless dips and flat stretches, and section 9f returns to what that means.

9b. The Gradient and the Hessian

For a function of many variables, the gradient ∇f collects the first partial derivatives, as in Module 7. The Hessian, $\mathbf{H}$, collects the second ones: its entry in row i and column j is $\partial^2 f/\partial x_i \partial x_j$, the rate at which the ith partial derivative changes as the jth variable moves. For smooth functions the order of differentiation does not matter, so the Hessian is symmetric.

Near a point $\mathbf{x}$, a function looks like a bowl described by these two objects. For a small step $\mathbf{h}$,

$$f(\mathbf{x}+\mathbf{h}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{h} + \tfrac{1}{2}\mathbf{h}^T \mathbf{H}\, \mathbf{h}$$

The last term is the curvature. The eigenvalues of the symmetric matrix $\mathbf{H}$, from Module 5, are the curvatures along its eigenvectors. If every eigenvalue is zero or positive everywhere, the function curves up in every direction and is convex. A negative eigenvalue marks a direction in which it curves down.

Worked example — A long, narrow bowl

For $f(x, y) = x^2 + 10y^2$, the gradient is $(2x, 20y)$, and the Hessian is the diagonal matrix with entries 2 and 20. Both eigenvalues are positive, so the function is convex, with its minimum at the origin. Its contours are ellipses about 3.2 times as long as they are wide, since moving along x raises f more slowly than moving along y.

The ratio of the largest eigenvalue to the smallest is the condition number, κ (kappa). Here $\kappa = 20/2 = 10$. It measures how stretched the bowl is, and section 9d shows that it sets how hard the function is to minimize.

9c. Gradient Descent

Solving $\nabla f = \mathbf{0}$ directly is out of reach for most large problems. Gradient descent instead takes repeated small steps downhill, against the gradient:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\, \nabla f(\mathbf{x}_k)$$

Here $\mathbf{x}_k$ is the position after k steps, and the step size α (alpha) is called the learning rate in machine learning.

Worked example — Halving the distance

For $f(x) = (x-3)^2$, the derivative is $2(x-3)$. With $\alpha = 0.25$, each step is $x_{k+1} = x_k - 0.5(x_k - 3)$, which halves the distance to the minimum at 3. From $x_0 = 0$, the steps go to 1.5, 2.25, 2.625, and 2.8125.

The step size matters. Here each step multiplies the distance to the minimum by $1 - 2\alpha$. A small α creeps; $\alpha = 0.5$ jumps to the minimum in one step. Any α above 1 makes the factor larger than 1 in size, so each step overshoots by more than the last and the iteration flies apart.

Least squares, from Module 4, can be solved the same way. With $\mathbf{A}$ holding the data, one row per observation, and $\mathbf{b}$ the observed values, the gradient of $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ is $2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$, which takes only two matrix–vector products. A problem far too large for the normal equations can still be solved by descending.

9d. Why Descent Zigzags

On a stretched bowl, the gradient points toward the steep walls rather than along the valley floor, so descent bounces from wall to wall. The eigenvalues say exactly what that costs. For a quadratic bowl with Hessian $\mathbf{H}$ and its minimum at the origin, the gradient is $\mathbf{Hx}$, so one step multiplies the position by $\mathbf{I} - \alpha\mathbf{H}$, where $\mathbf{I}$ is the identity matrix. Along an eigenvector with eigenvalue λ, the step multiplies by $1 - \alpha\lambda$.

Every such factor must be smaller than 1 in size, which requires $\alpha < 2/\lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue. The steep direction caps the step size, and the shallow direction, with the smallest eigenvalue $\lambda_{\min}$, then crawls. The best compromise, $\alpha = 2/(\lambda_{\max} + \lambda_{\min})$, shrinks the error each step by the factor

$$r = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa - 1}{\kappa + 1}$$

Worked example — The narrow bowl again

For $x^2 + 10y^2$, the best step is $2/22 \approx 0.091$, and the factor is $9/11 \approx 0.82$. Cutting the distance to the minimum a thousandfold takes about 35 steps. The path zigzags, because the factor along y is −0.82: each step overshoots the valley floor and lands on the opposite wall.

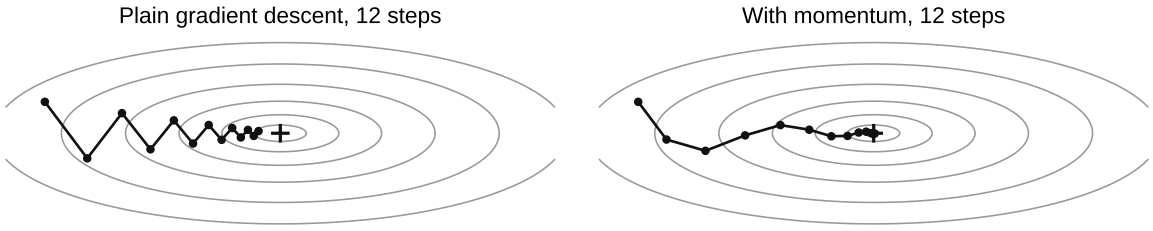


Figure 9.2. Descent on the narrow bowl $x^2 + 10y^2$ from the same start: plain steps of 0.09, and momentum with steps of 0.06 and $\beta = 0.5$. Plain descent bounces between the walls and creeps along the floor; momentum damps the bouncing and reaches the minimum, marked +, within 12 steps.

Real problems can be far worse. With a condition number of 10,000, the factor is 0.9998, and a thousandfold reduction takes about 34,500 steps. This is Module 3's ill-conditioning in another form: a stretched problem is both sensitive and slow.

Two standard remedies change the problem's shape. Rescaling the inputs so each varies over a similar range, called normalization, rounds the bowl. Newton's method uses the Hessian itself, stepping by $-\mathbf{H}^{-1}\nabla f$, which reaches the minimum of a quadratic bowl in a single step. Computing and inverting the Hessian is out of the question for a model with billions of parameters, so large models rely on the cheaper ideas that follow.

9e. Momentum

Momentum, proposed by Boris Polyak in 1964 as the heavy-ball method, lets each step carry part of the previous one. With a velocity $\mathbf{v}$ and a momentum factor β (beta) between 0 and 1,

$$\mathbf{v}_{k+1} = \beta\,\mathbf{v}_k - \alpha\,\nabla f(\mathbf{x}_k)$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_{k+1}$$

Across the valley the gradient flips sign from step to step, so those parts of the velocity cancel. Along the valley it keeps pointing the same way, so the velocity builds, like a ball gathering speed downhill. With its two settings tuned to the eigenvalues, momentum shrinks the error each step by a factor of about $(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$.

For $\kappa = 10$, that is 0.52 a step instead of 0.82, so in the long run a thousandfold reduction takes about 11 steps instead of 35. For $\kappa = 10{,}000$, it takes about 345 steps instead of 34,500. These are long-run rates, and the first steps can take longer. Training large models usually uses $\beta = 0.9$.

9f. Stochastic Gradients and Adam

A model's loss is usually an average over its training examples. Its exact gradient needs a pass through all of them, which for a large model means trillions of words. Stochastic gradient descent (SGD) instead estimates the gradient from a small random batch of examples and steps at once.

Each step is noisy but cheap. The average of B examples' gradients has noise with a standard deviation $1/\sqrt{B}$ times that of a single example's, the same arithmetic that diversifies a portfolio of uncorrelated assets in Module 6. One pass through the data is called an epoch.

Worked example — Same work, more progress

Fit a line to 1,000 points. A full gradient step costs 1,000 example gradients. In the same work, SGD with batches of 10 takes 100 steps and reaches a loss within about 15% of the best possible, while the full-gradient method is still more than five times the minimum. SGD then jitters around the minimum rather than settling, which is why training gradually shrinks the learning rate.

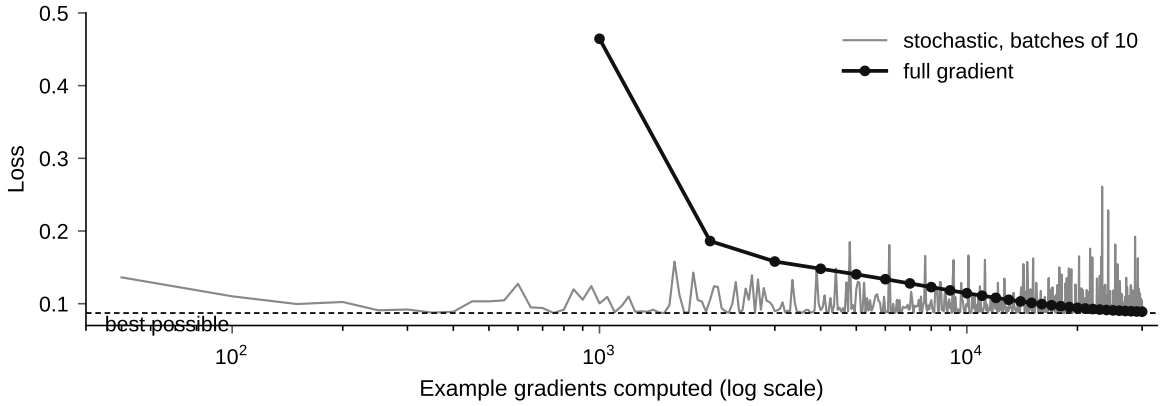


Figure 9.3. Fitting a line to 1,000 points. For the same amount of computation, stochastic steps get close to the best fit long before full-gradient steps do, but then hover around it.

Adam, introduced by Diederik Kingma and Jimmy Ba in 2014, adds momentum and gives every parameter its own step size. For each parameter w, with gradient g_k at step k, it keeps a running average m of the gradient and a running average v of its square:

$$m_k = \beta_1 m_{k-1} + (1-\beta_1)\, g_k$$
$$v_k = \beta_2 v_{k-1} + (1-\beta_2)\, g_k^2$$
$$w_k = w_{k-1} - \alpha \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \varepsilon}$$

The hats mark the two averages after a correction for starting at zero: dividing by $1-{\beta_1}^k$ and by $1-{\beta_2}^k$. The usual settings are $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and ε (epsilon) is a tiny number, such as 10^{-8}, that prevents division by zero.

Dividing by the square root of the average squared gradient makes each step about α in size, whatever the gradient's scale. A parameter whose gradients are a thousand times smaller than another's moves just as fast, which matters in a network whose layers see gradients of very different sizes. A variant called AdamW, which applies the weight penalty of Module 4 separately, is the usual choice for training large language models.

Worked example — Adam's first two steps

Take one parameter with $\alpha = 0.001$ and gradients of 4 and then 2. After step 1, $m = 0.4$ and $v = 0.016$, which the corrections turn into 4 and 16, so the step is $0.001 \times 4/\sqrt{16} = 0.001$. After step 2, $m = 0.56$ and $v \approx 0.0200$, corrected to about 2.95 and 10.0, so the step is about $0.001 \times 2.95/3.16 \approx 0.00093$. Halving the gradient barely changed the step.

Neural-network losses are not convex, so none of the guarantees of section 9a apply. In practice SGD and Adam train them well anyway. Researchers' explanations include the noise of small batches, which helps steps escape flat regions, and the finding that in very many dimensions most points where the gradient vanishes are saddles, curving down in some direction, rather than poor local minima.

In a Spreadsheet

Gradient methods fit a spreadsheet one step per row. With the current point in B2, the learning rate in F1, and the momentum factor in F2:

To compute	Formula
A gradient step for $(x-3)^2$	=B2-$F$1*2*(B2-3), filled down
Momentum, with the velocity in C2	=$F$2*C2-$F$1*2*(B2-3) for the next velocity in C3, and =B2+C3 for the next point
Gradient of least squares	=2*MMULT(TRANSPOSE(A), MMULT(A, x)-b), with A, x, and b as named ranges
Eigenvalues of a 2 × 2 Hessian	With its trace in H1 and determinant in H2: =(H1+SQRT(H1^2-4*H2))/2 and =(H1-SQRT(H1^2-4*H2))/2
Large models	Libraries such as PyTorch, whose torch.optim.SGD and torch.optim.AdamW carry out this module's methods

9g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Convex function	$f(t\,x + (1-t)\,y) \le t\,f(x) + (1-t)\,f(y)$ For any points x and y and any t from 0 to 1, the chord never dips below the graph; in one variable, $f''(x) \ge 0$
Why convexity matters	For a convex function on a convex set, every local minimum is global; least squares, linear programs, and minimum-variance portfolios are convex, and neural-network training is not
Hessian	$f(\mathbf{x}+\mathbf{h}) \approx f(\mathbf{x}) + \nabla f \cdot \mathbf{h} + \frac{1}{2}\mathbf{h}^T\mathbf{H}\,\mathbf{h}$ $\mathbf{H}$ holds the second partial derivatives and $\mathbf{h}$ is a small step; eigenvalues of $\mathbf{H}$ all zero or positive everywhere mean convex
Gradient descent	$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\,\nabla f(\mathbf{x}_k)$ α is the step size, or learning rate; it must be below $2/\lambda_{\max}$, with $\lambda_{\max}$ the largest eigenvalue of the Hessian
Condition number	$\kappa = \dfrac{\lambda_{\max}}{\lambda_{\min}}, \qquad r = \dfrac{\kappa - 1}{\kappa + 1}$ The ratio of the Hessian's largest and smallest eigenvalues; r is the best factor by which plain descent shrinks the error each step
Momentum	$\mathbf{v}_{k+1} = \beta\,\mathbf{v}_k - \alpha\,\nabla f(\mathbf{x}_k), \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_{k+1}$ $\mathbf{v}$ is the velocity and β the momentum factor, often 0.9; well tuned, the factor per step improves to about $(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$
Stochastic gradient	A gradient estimated from a random batch of B examples; its noise shrinks like $1/\sqrt{B}$, and one pass through the data is an epoch
Adam	$w_k = w_{k-1} - \alpha\,\dfrac{\hat{m}_k}{\sqrt{\hat{v}_k} + \varepsilon}$ m and v are running averages of a parameter w's gradient and squared gradient, with hats marking their corrected values; each step is about α in size whatever the gradient's scale
The idea	Convex problems can be solved reliably because going downhill always leads to the bottom. Large models are not convex, but they are trained by the same downhill steps, made cheap by sampling and steady by momentum and Adam.

9h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

9.1 Is $f(x) = x^2 - 4x + 7$ convex? Find its minimum.

9.2 Find the Hessian of $f(x, y) = 3x^2 + 2xy + y^2$ and decide whether f is convex.

9.3 Take one gradient step on $f(x) = (x-5)^2$ from $x = 1$ with $\alpha = 0.1$.

9.4 For $f(x) = 4x^2$, which step sizes make gradient descent converge, and which reaches the minimum in one step?

9.5 A Hessian has eigenvalues 1 and 50. Find the condition number and the best factors per step for plain descent and for momentum.

9.6 A data set has 50,000 examples, and training uses batches of 100. How many steps make one epoch?

Applied

9.7 Run gradient descent on $f(x, y) = x^2 + 10y^2$ from (10, 1) with $\alpha = 0.05$. Where is it after two steps, and about how many steps does x need to fall below 0.01?

9.8 Fit $y = wx$ to the points (1, 2), (2, 3), and (3, 7) by gradient descent on the sum of squared errors, from $w = 0$ with $\alpha = 0.02$. Take two steps and compare with the exact answer.

9.9 For the loss $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, the Hessian is $2\mathbf{A}^T\mathbf{A}$. If the eigenvalues of $\mathbf{A}^T\mathbf{A}$ are 1 and 100, what learning rates are safe, and what is the condition number?

9.10 A parameter's gradient on Adam's first step is −0.002, with $\alpha = 0.001$. How far does Adam move it, and how far would plain gradient descent with the same α move it?

9.11 Batches of 100 give gradient noise with a standard deviation of 0.2. What batch size halves it, and what does that cost?

9.12 Apply one step of Newton's method, $x - f'(x)/f''(x)$, to $f(x) = x^2 - 4x + 7$ from $x = 10$.

Challenge

9.13 For $f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$, with $\mathbf{H}$ symmetric and its eigenvalues positive, show that gradient descent multiplies the part of $\mathbf{x}$ along each eigenvector by $1 - \alpha\lambda$.

9.14 Show that the best fixed step for eigenvalues $\lambda_{\min}$ and $\lambda_{\max}$ is $2/(\lambda_{\min} + \lambda_{\max})$, with error factor $(\kappa - 1)/(\kappa + 1)$.

9.15 Show that a local minimum of a convex function is a global minimum.

9.16 Find and classify the points where $f(x) = x^4 - 2x^2$ has zero slope. What does gradient descent find from $x = 0.1$ and from $x = -0.1$?

Module 9 Quiz

Check Your Understanding

1. Gradient descent diverges, each step overshooting more than the last. What is the usual fix?

A) More iterations
B) A smaller step size
C) A larger step size, to get there sooner
D) Starting from a random point

2. Why does gradient descent zigzag in a long, narrow valley?

A) The gradient points toward the steep walls rather than along the valley floor, so each step overshoots across the valley
B) The learning rate is too small
C) The function is not convex
D) Rounding errors pile up

3. Which of these problems is convex?

A) Training a deep neural network
B) Finding the lowest point of a hilly landscape
C) Least squares
D) Choosing which of five plants to build

4. What does Adam do differently from plain gradient descent?

A) It computes the exact Hessian at every step.
B) It replaces every gradient with the exact gradient of the whole data set, computed afresh at every single step of training.
C) It adds a penalty on the size of the weights.
D) It scales each parameter's step by running averages of its gradients and their squares.

5. What does momentum do in gradient descent?

A) Doubles the step size
B) Removes all noise
C) It replaces the gradient with a random direction at each step, which lets the method escape from saddle points and local minima entirely
D) Carries part of each step into the next

Module 10 — Linear Algebra Inside AI

The matrix arithmetic of a modern AI model, from one layer to the data center

Recommended Reading / Reference

Book: Simon J. D. Prince, *Understanding Deep Learning* (2023) — a clear modern text on neural networks and transformers, free online from its author

Reference: Ashish Vaswani and others, “Attention Is All You Need” (2017) — the paper that introduced the transformer, the design of section 10c and of today's large language models

10a. A Network Is a Chain of Matrix Products

Module 2 met a single layer of a neural network: multiply the input vector by a weight matrix, add a bias vector, and pass each component through a simple nonlinear step. A network chains such layers, each one's output feeding the next. With two hidden layers and one output:

$$\begin{aligned}\mathbf{h}_1 &= \phi(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)\\ \mathbf{h}_2 &= \phi(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2)\\ y &= \mathbf{w}_3 \cdot \mathbf{h}_2 + b_3\end{aligned}$$

Here $\mathbf{x}$ is the input, each $\mathbf{W}$ a weight matrix, each $\mathbf{b}$ a bias vector, and the vectors $\mathbf{h}$ the outputs of the hidden layers. The final layer has a weight vector $\mathbf{w}_3$ and a single bias b_3, and its output y is the network's prediction.

The step φ (phi) is often the rectified linear unit (ReLU), the simplest choice, which keeps positive numbers and replaces negative ones with zero: $\varphi(z) = \max(0, z)$, applied to each component. Without it the layers would collapse into a single matrix, as Module 2 showed. With it, each layer bends and recombines what the layer before it found. A network with enough hidden units can approximate any reasonable function as closely as wanted, a result proved in 1989. Large language models mostly use smooth variants of the ReLU, but the idea is the same.

Worked example — A network with 13 parameters

Take the input $\mathbf{x} = (1, 2)$ and a hidden layer of three units, with weight rows (1, −1), (0.5, 0.5), and (−1, 1) and biases (0.5, 0, −0.5). Then $\mathbf{W}_1\mathbf{x} + \mathbf{b}_1 = (-0.5, 1.5, 0.5)$, and the ReLU turns it into $\mathbf{h} = (0, 1.5, 0.5)$. With output weights (2, 1, −1) and an output bias of 0.5, the prediction is $0 + 1.5 - 0.5 + 0.5 = 1.5$.

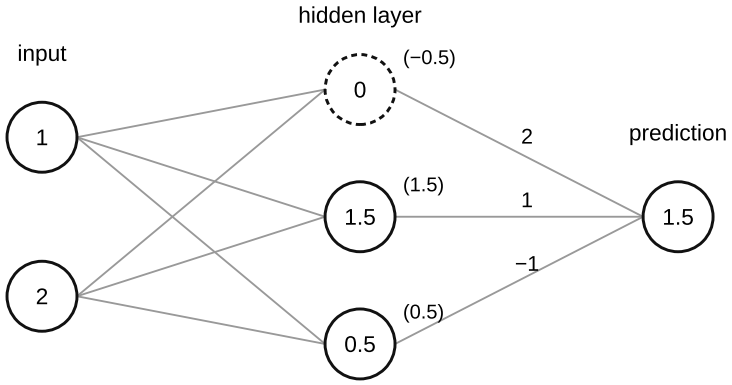


Figure 10.1. The 13-parameter network on the input (1, 2). Numbers in parentheses are each hidden unit's value before the ReLU; the first is negative, so that unit, dashed, outputs 0. The output weights label the lines on the right.

The numbers that training adjusts are the parameters. Here the hidden layer has $3 \times 2 = 6$ weights and 3 biases, and the output layer 3 weights and 1 bias, 13 in all. In general a layer from n inputs to m outputs has $mn + m$ parameters. Large language models have tens or hundreds of billions.

Networks process inputs in batches. Stacking a batch of inputs as the rows of a matrix $\mathbf{X}$ turns each layer's many matrix–vector products into one matrix–matrix product, $\mathbf{X}\mathbf{W}^T$, which is the operation that graphics processors are built to do fast.

10b. Learning by Backpropagation

Training adjusts every parameter to reduce a loss, such as the squared error $(y - t)^2$ between the prediction y and the target t. Gradient descent, from Module 9, needs the gradient of the loss with respect to every parameter. Backpropagation computes all of them in one backward pass through the network, applying the chain rule of calculus layer by layer. David Rumelhart, Geoffrey Hinton, and Ronald Williams made it famous in 1986.

Each step backward is linear algebra again. For a layer $\mathbf{z} = \mathbf{W}\mathbf{x}$, let $\boldsymbol{\delta}$ (delta) be the gradient of the loss with respect to $\mathbf{z}$. The gradient with respect to the weights is the outer product $\boldsymbol{\delta}\mathbf{x}^T$, a matrix with one entry per weight. The gradient passed back to the layer's input is $\mathbf{W}^T\boldsymbol{\delta}$: the transpose of Module 2 carries the error backward. A ReLU passes the gradient through where its input was positive and blocks it elsewhere.

Worked example — One step of training

In the 13-parameter network, let the target be $t = 1$. The prediction of 1.5 gives a loss of $0.5^2 = 0.25$, and the loss's gradient with respect to the prediction is $2 \times 0.5 = 1$. The output weights' gradient is that 1 times $\mathbf{h}$, or (0, 1.5, 0.5). Passed back through the output weights, the error is (2, 1, −1). The ReLU had switched off the first hidden unit, so its share is blocked, leaving $\boldsymbol{\delta} = (0, 1, -1)$.

The hidden weights' gradient is $\boldsymbol{\delta}\mathbf{x}^T$, with rows (0, 0), (1, 2), and (−1, −2). One gradient step with $\alpha = 0.02$ changes every parameter a little, and the prediction falls from 1.5 to about 1.19, cutting the loss from 0.25 to 0.037. Training a large model repeats exactly this, for billions of parameters and trillions of examples.

10c. Attention

Language models work on tokens, words or pieces of words, each represented by an embedding vector like those of Module 1. The transformer design, introduced in 2017, lets every token gather information from the others through attention. Each token's vector is multiplied by three learned matrices to give a query, a key, and a value. A token's query is compared with every token's key by a dot product, and the results become weights for averaging the values.

Stack the queries, keys, and values of all the tokens as the rows of matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, with d components in each query and key. Attention is then

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$$

The softmax turns each row of scores into weights that are positive and add to 1: a score z_j becomes $e^{z_j}/\Sigma_k e^{z_k}$, where the sum runs over the row and $e \approx 2.718$ is the base of natural logarithms. Dividing by $\sqrt{d}$ keeps the scores from growing with the number of components, which would push the softmax toward putting all the weight on one token.

Worked example — Three tokens

Let the queries and the keys be the same, with rows (1, 0), (0, 1), and (1, 1), so $d = 2$, and let the values have rows (2, 0), (0, 2), and (1, 1). The scores, $\mathbf{Q}\mathbf{K}^T/\sqrt{2}$, and the weight matrix $\mathbf{A}$ that the softmax makes of them, row by row, are:

$$\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{2}} \approx \begin{bmatrix} 0.71 & 0 & 0.71 \\ 0 & 0.71 & 0.71 \\ 0.71 & 0.71 & 1.41 \end{bmatrix}$$

$$\mathbf{A} \approx \begin{bmatrix} 0.40 & 0.20 & 0.40 \\ 0.20 & 0.40 & 0.40 \\ 0.25 & 0.25 & 0.50 \end{bmatrix}$$

Each row of $\mathbf{A}$ says how much one token attends to each token. The first token attends equally to itself and to the third, whose key resembles its query, and half as much to the second. The output, $\mathbf{A}\mathbf{V}$, has rows (1.20, 0.80), (0.80, 1.20), and (1.00, 1.00): each token's new vector blends the values of the tokens it attended to.

A real model runs dozens of these attention heads side by side, each with its own matrices, in each of dozens of layers, and follows each attention step with a small two-layer network applied to every token. When it generates text, each token may attend only to earlier ones. The score matrix has a row and a column for every token, so its cost grows with the square of the text's length, which is why long documents are expensive.

10d. Counting the Arithmetic

A matrix–vector product with an $m \times n$ matrix takes mn multiplications and about as many additions, roughly $2mn$ operations on decimal numbers, called floating-point operations. In a forward pass, each parameter of a model takes part in about one multiplication and one addition per token, so a model with N parameters needs about $2N$ operations per token.

Training also needs the backward pass, which costs about twice the forward pass, so training takes about $6N$ operations per token. For D training tokens, the total is

$$\text{training operations} \approx 6\,N\,D$$

Worked example — A 70-billion-parameter model

Meta reported that its Llama 3 models were trained on more than 15 trillion tokens. For the version with 70 billion parameters, $6ND \approx 6 \times (70 \times 10^9) \times (15 \times 10^{12}) \approx 6.3 \times 10^{24}$ operations. Each token it generates afterward takes about $2N = 1.4 \times 10^{11}$ operations.

Graphics processing units (GPUs) suit this arithmetic. A matrix product is millions of independent multiply-and-add steps, and a GPU has thousands of arithmetic units working at once, with special circuits for multiplying small blocks of matrices. Nvidia's H100 is rated at about 989 trillion operations a second on 16-bit numbers, drawing up to 700 watts.

At that peak rate, 6.3×10^{24} operations would take about 1.8 million GPU-hours. Meta's model card reports 6.4 million H100 GPU-hours for the 70-billion-parameter model, so its chips averaged about 28% of their peak. That shortfall is normal once memory traffic, communication between chips, and interruptions are counted. At 700 watts, 6.4 million GPU-hours is about 4.5 gigawatt-hours (GWh) of electricity for the chips alone, before servers and cooling.

Memory is the other limit. Stored as 16-bit numbers, 2 bytes each, 70 billion parameters take 140 gigabytes (GB), more than one H100's 80 GB, so the model must be split across several chips.

10e. Low Rank and the Singular Value Decomposition

Every matrix, square or not, can be written as a rotation, a stretch along perpendicular directions, and a second rotation, where either rotation may include a reflection. This is the singular value decomposition (SVD):

$$\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$$

The columns of $\mathbf{U}$ and of $\mathbf{V}$ are perpendicular unit vectors, and $\mathbf{\Sigma}$, here a diagonal matrix rather than a covariance matrix, holds the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$. Their squares are the eigenvalues of $\mathbf{A}^T\mathbf{A}$, and the columns of $\mathbf{V}$ are its eigenvectors. For a table of data with its averages removed, those columns are the principal components of Module 5.

Written out, the decomposition is a sum of simple pieces, each a column times a row: $\mathbf{A} = \sigma_1\mathbf{u}_1{\mathbf{v}_1}^T + \sigma_2\mathbf{u}_2{\mathbf{v}_2}^T + \ldots$, where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith columns of $\mathbf{U}$ and $\mathbf{V}$. Keeping only the first k pieces gives the best approximation of $\mathbf{A}$ by a matrix of rank k, one built from k such pieces. Carl Eckart and Gale Young proved that in 1936.

Worked example — A rank-1 approximation

The matrix with rows (4, 0) and (3, −5) has singular values $\sqrt{40} \approx 6.32$ and $\sqrt{10} \approx 3.16$, with $\mathbf{u}_1 = (1, 2)/\sqrt{5}$ and $\mathbf{v}_1 = (1, -1)/\sqrt{2}$. Its best rank-1 approximation, $\sigma_1\mathbf{u}_1{\mathbf{v}_1}^T$, has rows (2, −2) and (4, −4). What is left over, with rows (2, 2) and (−1, −1), is exactly the second piece, whose size is σ_2.

Low rank saves storage and arithmetic. A 1,000 × 1,000 matrix holds a million numbers, but its rank-50 approximation needs only $50 \times (1{,}000 + 1{,}000 + 1) = 100{,}050$, about a tenth. When the singular values fall

off quickly, as they do for photographs and for many matrices learned from data, the approximation loses little.

Fine-tuning a large model shows the idea at work. Low-rank adaptation (LoRA), introduced by Edward Hu and colleagues in 2021, leaves each weight matrix $\mathbf{W}$ frozen and learns a low-rank change, $\mathbf{BC}$, where $\mathbf{B}$ has r columns and $\mathbf{C}$ has r rows for a small rank r. For a 4,096 × 4,096 matrix and $r = 8$, that is $8 \times (4{,}096 + 4{,}096) = 65{,}536$ numbers to train instead of 16.8 million, about 0.4%.

Its authors reported cutting the trainable parameters of a 175-billion-parameter model by a factor of 10,000. From embeddings to attention to fine-tuning, the arithmetic of a modern AI model is the arithmetic of this course, at enormous scale: dot products, matrix products, gradients, and eigenvalues and singular values.

In a Spreadsheet

A spreadsheet can run a small network or a single attention head. With a weight matrix W, an input column x, and a bias column b as named ranges:

To compute	Formula
One layer with a ReLU	=LET(z, MMULT(W, x)+b, IF(z>0, z, 0))
Softmax of the scores in A1:A3	=EXP(A1:A3)/SUM(EXP(A1:A3))
Attention scores	=MMULT(Q, TRANSPOSE(K))/SQRT(2), then a softmax along each row
Operations to train a model	=6*N*D, with N parameters and D training tokens
Singular value decomposition	Not built in; the Python function numpy.linalg.svd returns U, the singular values, and the transpose of V

10f. Memorize This

Commit these to memory. The practice problems assume them, and so does much of the mathematics you will meet after this course.

Layer	$\mathbf{h} = \phi(\mathbf{W}\mathbf{x} + \mathbf{b}), \qquad \phi(z) = \max(0, z)$ $\mathbf{W}$ is the weight matrix, $\mathbf{b}$ the bias vector, and φ the ReLU step applied to each component; a layer from n inputs to m outputs has $mn + m$ parameters
Backpropagation	$\dfrac{\partial L}{\partial \mathbf{W}} = \boldsymbol{\delta}\,\mathbf{x}^T, \qquad$ passed back: $\mathbf{W}^T\boldsymbol{\delta}$ For a layer $\mathbf{z} = \mathbf{W}\mathbf{x}$ with loss L, $\boldsymbol{\delta}$ is the gradient of L with respect to $\mathbf{z}$
Softmax	$\text{softmax}(\mathbf{z})_j = \dfrac{e^{z_j}}{\sum_k e^{z_k}}$ Turns a list of scores z_j into positive weights that add to 1
Attention	$\text{softmax}\left(\dfrac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$ $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ hold each token's query, key, and value as rows, and d is the length of a query; the cost grows with the square of the number of tokens
Arithmetic	$2N$ per token, $\qquad 6ND$ to train N is the number of parameters and D the number of training tokens; a matrix–vector product with an $m \times n$ matrix takes about $2mn$ operations
Singular value decomposition	$\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T = \sum_i \sigma_i\,\mathbf{u}_i\mathbf{v}_i^T$ $\mathbf{U}$ and $\mathbf{V}$ have perpendicular unit columns $\mathbf{u}_i$ and $\mathbf{v}_i$, and the diagonal $\mathbf{\Sigma}$ holds the singular values σ_i, largest first
Best low-rank approximation	Keep the first k terms of the sum: no matrix of rank k comes closer
Low-rank adaptation	$\mathbf{W} + \mathbf{B}\mathbf{C}$ $\mathbf{W}$ stays frozen while $\mathbf{B}$, of size $d \times r$, and $\mathbf{C}$, of size $r \times k$, are trained: $r(d + k)$ numbers instead of dk
The idea	A modern AI model is matrix products, a simple nonlinear step, and gradients, repeated at enormous scale. Counting the arithmetic predicts its cost, and low rank cuts it.

10g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

10.1 Compute $\varphi(\mathbf{W}\mathbf{x} + \mathbf{b})$ with the ReLU, for $\mathbf{W}$ with rows (2, −1) and (1, 1), $\mathbf{x} = (1, 3)$, and $\mathbf{b} = (0, -2)$.

10.2 How many parameters does a network with 784 inputs, one hidden layer of 128 units, and 10 outputs have?

10.3 Find the softmax of the scores (2, 1, 0).

10.4 Show that adding the same number c to every score leaves the softmax unchanged.

10.5 About how many operations does multiplying a 4,096 × 4,096 matrix by a vector take?

10.6 How much memory does a model with 7 billion parameters need, stored as 16-bit numbers?

Applied

10.7 A 7-billion-parameter model is trained on 2 trillion tokens. Estimate the operations per generated token and the total training operations.

10.8 How many H100 GPU-hours would the training in problem 10.7 take, if the chips ran at 40% of their peak of 989 trillion operations a second?

10.9 A query (1, 0) meets keys (2, 0), (0, 2), and (1, 1), with $d = 2$. Find the attention weights.

10.10 For n tokens with queries and keys of length d, computing the scores takes about $2n^2d$ operations. How does the cost change when the text grows from 4,000 to 8,000 tokens?

10.11 Low-rank adaptation is applied to an 8,192 × 8,192 weight matrix with $r = 16$. How many numbers are trained, and what fraction of the full matrix is that?

10.12 A 2,000 × 500 matrix is stored as its rank-20 approximation. How many numbers does that take, and what fraction of the original?

Challenge

10.13 Show that two layers with biases but no nonlinear step collapse into one.

10.14 For a layer $\mathbf{z} = \mathbf{W}\mathbf{x}$ and a loss L, show that the gradient with respect to $\mathbf{W}$ is $\boldsymbol{\delta}\mathbf{x}^T$, with $\boldsymbol{\delta}$ the gradient with respect to $\mathbf{z}$.

10.15 Using $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, show that the squares of the singular values are the eigenvalues of $\mathbf{A}^T\mathbf{A}$.

10.16 For the matrix with rows (4, 0) and (3, −5), check that the leftover after the rank-1 approximation is the second piece, using $\mathbf{u}_2 = (2, -1)/\sqrt{5}$ and $\mathbf{v}_2 = (1, 1)/\sqrt{2}$.

Module 10 Quiz

Check Your Understanding

1. What does backpropagation compute?

A) The network's output for a new input, one layer at a time
B) The gradient of the loss with respect to every weight
C) The best learning rate
D) Which training examples the model finds hardest, so that they can be shown to it more often during training

2. In attention, what does the softmax step produce?

A) Weights that add to 1
B) The keys and queries for every token in the sequence
C) A count of how often each word appears
D) The final answer, chosen as the single most likely next word from the whole vocabulary of the model

3. A model has 70 billion parameters. Roughly how much arithmetic does it take to process one token?

A) About 70 billion operations
B) About 6.3 trillion operations
C) About 140 billion operations: a multiplication and an addition for every parameter
D) About 4.9 quadrillion operations

4. What does a rank-1 approximation keep from a matrix's singular value decomposition?

A) All singular values but the largest
B) The top singular term
C) The diagonal
D) The first row and column

5. What does low-rank adaptation (LoRA) train?

A) A low-rank update
B) Every weight of the model
C) Only the final layer's biases
D) The embeddings of its training data

Everything to Memorize

The Memorize This boxes from every module, in one place

Module 1: Vectors

Vector	An ordered list of numbers, set in bold: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ x_i is the ith component, and n, the number of components, is the dimension
Adding and scaling	$\mathbf{a} + \mathbf{b} = (a_1 + b_1, \ldots, a_n + b_n) \qquad c\,\mathbf{a} = (c\,a_1, \ldots, c\,a_n)$ c is a scalar, an ordinary number; a sum of scaled vectors, such as $c_1\mathbf{v}_1 + c_2\mathbf{v}_2$, is a linear combination
Dot product	$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i\, b_i$ Multiply matching components and add; a portfolio's return is $\mathbf{w} \cdot \mathbf{r}$, its weights $\mathbf{w}$ times its assets' returns $\mathbf{r}$
Length and distance	$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} \qquad \text{distance} = \|\mathbf{a} - \mathbf{b}\|$ $\|\mathbf{x}\|$ is the norm; $\mathbf{x}/\|\mathbf{x}\|$ is the unit vector pointing the same way
Angle	$\cos\theta = \dfrac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|}$ θ is the angle between nonzero vectors $\mathbf{a}$ and $\mathbf{b}$; the ratio is the cosine similarity, from −1 to 1
Perpendicular	$\mathbf{a} \cdot \mathbf{b} = 0$: the vectors are at right angles, also called orthogonal
Projection	$\text{proj}_{\mathbf{a}}\,\mathbf{b} = \dfrac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{a} \cdot \mathbf{a}}\,\mathbf{a}$ The multiple of $\mathbf{a}$ closest to $\mathbf{b}$; the residual, $\mathbf{b}$ minus its projection, is perpendicular to $\mathbf{a}$
The idea	A vector packs many numbers into one object. One operation, the dot product, computes returns, revenues, lengths, angles, and projections.

Module 2: Matrices

Matrix	An $m \times n$ table of numbers, set in bold capitals, such as $\mathbf{A}$ m rows and n columns; a_{ij} is the entry in row i and column j
Matrix times vector	$(\mathbf{Ax})_i = \sum_{j=1}^{n} a_{ij}\, x_j \qquad \mathbf{Ax} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$ Each component is a row of $\mathbf{A}$ dotted with $\mathbf{x}$; the whole is a combination of the columns $\mathbf{a}_j$
Matrix times matrix	$(\mathbf{AB})_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}$ Row i of $\mathbf{A}$ dotted with column j of $\mathbf{B}$; an $m \times n$ matrix times an $n \times p$ matrix is $m \times p$
Order and grouping	$\mathbf{AB}$ and $\mathbf{BA}$ usually differ; $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ $\mathbf{ABx}$ applies $\mathbf{B}$ first, then $\mathbf{A}$
Transpose	$(\mathbf{A}^T)_{ij} = a_{ji} \qquad (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \qquad \mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y}$ Rows and columns swap; a matrix equal to its transpose is symmetric
Identity	$\mathbf{Ix} = \mathbf{x}$ and $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$ Ones on the diagonal and zeros elsewhere

Inverse of a 2 × 2 matrix	$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ $ad - bc$ is the determinant; when it is zero there is no inverse
Linear layers	$\mathbf{W}_2(\mathbf{W}_1\mathbf{x}) = (\mathbf{W}_2\mathbf{W}_1)\mathbf{x}$ $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices; two purely linear layers equal one, so networks put a nonlinear step between them, such as the ReLU, which turns each input z into $\max(z, 0)$
Cost	Multiplying two $n \times n$ matrices takes n^3 multiplications: doubling the size multiplies the work by eight
The idea	A matrix is a machine that takes a vector and returns a combination of its own columns. Multiplying matrices chains such machines, in order.

Module 3: Systems of Linear Equations

Matrix form	$\mathbf{A}\mathbf{x} = \mathbf{b}$ $\mathbf{A}$ holds the coefficients, $\mathbf{x}$ the unknowns, and $\mathbf{b}$ the right-hand sides
Elimination	Subtract multiples of equations to make a triangle, then solve from the bottom up by back substitution The numbers left on the diagonal are the pivots; n equations take about $n^3/3$ multiplications
Three outcomes	None if an equation becomes $0 = c$ with c not zero; otherwise one solution if every unknown has a pivot, and infinitely many if some unknown has none The number of pivots is the rank
Determinant	$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$ Its size is the factor by which the matrix scales area; it is the product of the pivots, with the sign flipped for each row swap; not zero means exactly one solution
Ill-conditioning	Nearly parallel equations: a tiny change in the data moves the solution a long way
No arbitrage	An option is worth what the portfolio replicating its payoff costs In one period, solve for the shares Δ and bonds B that match the payoff in each outcome
Input–output	$\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{d}$ $\mathbf{x}$ is total output, $\mathbf{d}$ final demand, and a_{ij} the input from industry i per dollar of industry j's output
Power flow	Flow splits across parallel paths in inverse proportion to their reactance; the fractions are shift factors In the DC model the flow from bus i to bus j is $(\theta_i - \theta_j)/X_{ij}$, with θ_i the voltage angles and X_{ij} the line's reactance
The idea	Elimination solves any linear system or shows why it cannot be solved. The same arithmetic prices options, traces an economy's supply chains, and routes power.

Module 4: Least Squares

Least squares	Choose $\mathbf{x}$ to minimize $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2$, the sum of the squared residuals $\mathbf{b}$ holds the observations, $\mathbf{A}$ has a row per observation and a column per input, and $\mathbf{x}$ holds the coefficients
Normal equations	$\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\mathbf{b}$ The residual is perpendicular to every column of $\mathbf{A}$: the fit is a projection
Slope and intercept of a line	$m = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad c = \bar{y} - m\,\bar{x}$ For $y = c + mx$ fitted to points (x_i, y_i); bars mark averages, and the line passes through the point of averages
R squared	$R^2 = 1 - \frac{\sum_i r_i^2}{\sum_i (y_i - \bar{y})^2}$ The share of the variation in y that the model explains; r_i are the residuals
Dummy variable	A column of 1s and 0s that measures the effect of a category, such as weekdays
Beta and alpha	$r_s - r_f = \alpha + \beta\,(r_m - r_f) + \varepsilon$ r_s, r_m, and r_f are the returns on a stock, the market, and a Treasury bill; β is the slope, α the unexplained average return, and ε the residual
Overfitting	More coefficients always fit past data at least as well but can predict new data worse: judge a model on data held back from the fit
Ridge penalty	$(\mathbf{A}^T\mathbf{A} + \lambda\,\mathbf{I})\,\mathbf{x} = \mathbf{A}^T\mathbf{b}$ λ weights a penalty on the squared size of the coefficients; a larger λ shrinks them toward zero
The idea	With more data than unknowns, the best fit is a projection, and its residual is perpendicular to everything the model can do. Fit on some data and judge on the rest.

Module 5: Eigenvalues and Eigenvectors

Eigenvector and eigenvalue	$\mathbf{A}\mathbf{v} = \lambda\,\mathbf{v}$ $\mathbf{v}$ is a nonzero vector that the square matrix $\mathbf{A}$ only stretches, by the factor λ
Finding them	$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ For a 2 × 2 matrix with rows (a, b) and (c, d), $\lambda^2 - (a + d)\lambda + (ad - bc) = 0$; the eigenvalues add to the trace, $a + d$, and multiply to the determinant
Symmetric matrices	Real eigenvalues and perpendicular eigenvectors
Repeated multiplication	$\mathbf{A}^k\mathbf{v} = \lambda^k\mathbf{v}$ In $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$, the eigenvalue largest in size sets the long-run growth rate and its eigenvector the long-run shape
Markov chain	$\mathbf{x}_{k+1} = \mathbf{P}\,\mathbf{x}_k$ p_{ij} is the probability of moving from state j to state i, so columns add to 1; the steady state is the eigenvector with eigenvalue 1, and the second-largest eigenvalue in size sets how fast the chain approaches it
Several periods	$\mathbf{P}^k$ gives the probabilities of moving between states over k periods
PageRank	A page's share of time, in the long run, of a random surfer who follows links 85% of the time

Principal components	The eigenvectors of a covariance matrix: patterns of joint movement, each explaining its eigenvalue's share of the variance; for yields, level, slope, and curvature
The idea	A matrix's eigenvectors are the directions it only stretches. Applied over and over, a matrix is ruled by its largest eigenvalue, which is why eigenvalues predict growth, steady states, and the main patterns in data.

Module 6: Covariance and Portfolio Risk

Variance and standard deviation	$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ σ^2 is the variance of the observations $x_1, \ldots, x_n$, and the bar marks their average; the standard deviation σ is its square root, and for returns it is called volatility
Annualizing	$\sigma_{\text{annual}} = \sqrt{12}\,\sigma_{\text{monthly}}$ Variances add over independent periods, so volatility grows with the square root of time
Covariance	$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ σ_{xy} is the covariance of two series x_i and y_i; it is positive when they tend to be above their averages at the same times
Correlation	$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ σ_x and σ_y are the standard deviations; ρ lies between −1 and 1 and is the cosine of the angle between the two series' deviation vectors
Covariance matrix	$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$ $\mathbf{X}$ holds each asset's deviations in a column; the entries are $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, with the variances on the diagonal, and the matrix is symmetric
Portfolio variance	$\sigma_p^2 = \mathbf{w}^T\mathbf{\Sigma}\,\mathbf{w}$ $\mathbf{w}$ holds the weights; for two assets with weights w_1 and w_2, volatilities σ_1 and σ_2, and correlation ρ, ${\sigma_p}^2 = {w_1}^2{\sigma_1}^2 + {w_2}^2{\sigma_2}^2 + 2w_1w_2\rho\sigma_1\sigma_2$
Diversification limit	$\sigma_p^2 = \frac{\sigma^2}{n} + \left(1 - \frac{1}{n}\right)\rho\,\sigma^2$ For n equally weighted assets, each with volatility σ and every pair with correlation ρ; the volatility falls toward $\sigma\sqrt{\rho}$, the risk they share
Risk contribution	$\text{share}_i = \frac{w_i\,(\mathbf{\Sigma}\mathbf{w})_i}{\mathbf{w}^T\mathbf{\Sigma}\,\mathbf{w}}$ w_i is asset i's weight and $(\mathbf{\Sigma}\mathbf{w})_i$ the ith entry of $\mathbf{\Sigma}\mathbf{w}$; the shares add to 100%, and risk parity makes them equal, which for two assets means $w_1\sigma_1 = w_2\sigma_2$
The idea	Risk adds up through the covariance matrix, not through the weights alone. Diversification removes the risk that assets do not share, never the risk they do, and correlations measured in calm times can rise in a crisis.

Module 7: Optimization with Constraints

Partial derivative and gradient	$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$ $\partial f / \partial x$ is the rate of change of f as x changes with y held fixed; the gradient ∇f points where f rises fastest, perpendicular to its contours
Lagrange condition	$\nabla f = \lambda \nabla g$ At the best point of f on the constraint $g(x, y) = c$, the two gradients are parallel; λ is the Lagrange multiplier
Lagrangian	$L = f - \lambda (g - c)$ Set its partial derivatives with respect to every variable, and to λ, equal to zero
Shadow price	$\frac{df^*}{dc} = \lambda$ $f^*(c)$ is the best value achievable at constraint level c; the multiplier is the marginal value of the limit
Economic dispatch	$C_1'(P_1) = C_2'(P_2) = \cdots = \lambda$ Generator i produces P_i at cost $C_i(P_i)$; every generator between its limits runs at the same marginal cost, λ, which is the price of power
Minimum-variance portfolio	$\mathbf{w}_{\min} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}$ $\mathbf{\Sigma}$ is the covariance matrix and $\mathbf{1}$ a vector of ones; for two assets with volatilities σ_1 and σ_2 and correlation ρ, $w_1 = (\sigma_2{}^2 - \rho\sigma_1\sigma_2)/(\sigma_1{}^2 + \sigma_2{}^2 - 2\rho\sigma_1\sigma_2)$
Efficient frontier	The portfolios with the highest expected return for their risk; any portfolio inside it is dominated by a frontier portfolio with the same return and less risk
Complementary slackness	$\lambda_j s_j = 0$ s_j is the room left under limit j and λ_j its multiplier: either the limit binds or its shadow price is zero; with the Lagrange conditions, these are the Karush–Kuhn–Tucker (KKT) conditions
The idea	At a constrained optimum the gradients line up, and each multiplier prices its limit. The price of power and the cost of a risk limit are both Lagrange multipliers.

Module 8: Linear Programming

Linear program	$\text{maximize } \mathbf{c}^T\mathbf{x} \text{ subject to } \mathbf{A}\mathbf{x} \le \mathbf{b},\ \mathbf{x} \ge \mathbf{0}$ $\mathbf{x}$ holds the decision variables, $\mathbf{c}$ the payoff per unit of each, each row of $\mathbf{A}$ one limit, and $\mathbf{b}$ the amounts available
Corner principle	If a linear program has a best solution, a corner of the feasible region is best; the simplex method walks from corner to better corner
Slack variable	The unused amount of a limit: $x + y \le 100$ becomes $x + y + s = 100$, with the slack s zero or above
Duality	$\text{minimize } \mathbf{b}^T\mathbf{p} \text{ subject to } \mathbf{A}^T\mathbf{p} \ge \mathbf{c},\ \mathbf{p} \ge \mathbf{0}$ The dual has one price in $\mathbf{p}$ for each limit; at the optimum its value equals the primal's, and its solution is the limits' shadow prices
Battery arbitrage	A charge and discharge pays only if $\eta p_{\text{sell}} > p_{\text{buy}}$ η is the round-trip efficiency, p_{buy} the price paid to charge, and p_{sell} the price received on discharge

Locational marginal price	The cost of serving one more MW at a bus within the grid's limits; with a congested line, prices differ by bus and can exceed every generator's offer
Price components	$\mathrm{LMP}_k = \lambda - \mu\,\varphi_k$ λ is the price at the reference bus, μ (mu) the congestion price of the full line, and φ_k (phi) the shift factor of bus k on that line
Congestion rent	Buyers' payments minus sellers' receipts: the congestion price times the flow on the full line, paid out to holders of financial transmission rights
The idea	A linear program's best plan sits at a corner, and its dual prices every limit. The price of power at each point on a grid is one of those dual prices.

Module 9: Convexity and Gradient Methods

Convex function	$f\big(t\,x + (1-t)\,y\big) \le t\,f(x) + (1-t)\,f(y)$ For any points x and y and any t from 0 to 1, the chord never dips below the graph; in one variable, $f''(x) \ge 0$
Why convexity matters	For a convex function on a convex set, every local minimum is global; least squares, linear programs, and minimum-variance portfolios are convex, and neural-network training is not
Hessian	$f(\mathbf{x}+\mathbf{h}) \approx f(\mathbf{x}) + \nabla f \cdot \mathbf{h} + \frac{1}{2}\mathbf{h}^T\mathbf{H}\,\mathbf{h}$ $\mathbf{H}$ holds the second partial derivatives and $\mathbf{h}$ is a small step; eigenvalues of $\mathbf{H}$ all zero or positive everywhere mean convex
Gradient descent	$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\,\nabla f(\mathbf{x}_k)$ α is the step size, or learning rate; it must be below $2/\lambda_{\max}$, with $\lambda_{\max}$ the largest eigenvalue of the Hessian
Condition number	$\kappa = \dfrac{\lambda_{\max}}{\lambda_{\min}}, \qquad r = \dfrac{\kappa - 1}{\kappa + 1}$ The ratio of the Hessian's largest and smallest eigenvalues; r is the best factor by which plain descent shrinks the error each step
Momentum	$\mathbf{v}_{k+1} = \beta\,\mathbf{v}_k - \alpha\,\nabla f(\mathbf{x}_k), \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_{k+1}$ $\mathbf{v}$ is the velocity and β the momentum factor, often 0.9; well tuned, the factor per step improves to about $(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$
Stochastic gradient	A gradient estimated from a random batch of B examples; its noise shrinks like $1/\sqrt{B}$, and one pass through the data is an epoch
Adam	$w_k = w_{k-1} - \alpha\,\dfrac{\hat{m}_k}{\sqrt{\hat{v}_k} + \varepsilon}$ m and v are running averages of a parameter w's gradient and squared gradient, with hats marking their corrected values; each step is about α in size whatever the gradient's scale
The idea	Convex problems can be solved reliably because going downhill always leads to the bottom. Large models are not convex, but they are trained by the same downhill steps, made cheap by sampling and steady by momentum and Adam.

Module 10: Linear Algebra Inside AI

Layer	$\mathbf{h} = \phi(\mathbf{W}\mathbf{x} + \mathbf{b}), \qquad \phi(z) = \max(0, z)$ $\mathbf{W}$ is the weight matrix, $\mathbf{b}$ the bias vector, and φ the ReLU step applied to each component; a layer from n inputs to m outputs has $mn + m$ parameters
Backpropagation	$\frac{\partial L}{\partial \mathbf{W}} = \boldsymbol{\delta}\,\mathbf{x}^T, \qquad$ passed back: $\mathbf{W}^T\boldsymbol{\delta}$ For a layer $\mathbf{z} = \mathbf{W}\mathbf{x}$ with loss L, $\boldsymbol{\delta}$ is the gradient of L with respect to $\mathbf{z}$
Softmax	$\text{softmax}(\mathbf{z})_j = \frac{e^{z_j}}{\sum_k e^{z_k}}$ Turns a list of scores z_j into positive weights that add to 1
Attention	$\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$ $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ hold each token's query, key, and value as rows, and d is the length of a query; the cost grows with the square of the number of tokens
Arithmetic	$2N$ per token, $\qquad 6ND$ to train N is the number of parameters and D the number of training tokens; a matrix–vector product with an $m \times n$ matrix takes about $2mn$ operations
Singular value decomposition	$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ $\mathbf{U}$ and $\mathbf{V}$ have perpendicular unit columns $\mathbf{u}_i$ and $\mathbf{v}_i$, and the diagonal $\boldsymbol{\Sigma}$ holds the singular values σ_i, largest first
Best low-rank approximation	Keep the first k terms of the sum: no matrix of rank k comes closer
Low-rank adaptation	$\mathbf{W} + \mathbf{B}\mathbf{C}$ $\mathbf{W}$ stays frozen while $\mathbf{B}$, of size $d \times r$, and $\mathbf{C}$, of size $r \times k$, are trained: $r(d + k)$ numbers instead of dk
The idea	A modern AI model is matrix products, a simple nonlinear step, and gradients, repeated at enormous scale. Counting the arithmetic predicts its cost, and low rank cuts it.

Worked Solutions

Every practice problem, solved step by step

Module 1: Vectors

1.1 Double the second vector to get (2, 8, 0), then add component by component: (4, 7, 3).

1.2 $4 - 3 - 10 = -9$.

1.3 The length is $\sqrt{(1 + 4 + 4)} = 3$, so the unit vector is (1/3, 2/3, 2/3).

1.4 Their difference is (−3, 4), whose length is $\sqrt{(9 + 16)} = 5$.

1.5 The dot product is 1 and the lengths are 1 and $\sqrt{2}$, so $\cos\theta = 1/\sqrt{2} \approx 0.707$, and θ is 45 degrees.

1.6 The dot product $6 - 6c$ must be zero, so $c = 1$.

1.7 The coefficient is $3/1 = 3$, so the projection is (3, 0), and the residual is (0, 4), which is perpendicular to (1, 0).

1.8 The dot product is $4.8 + 1.5 - 0.6 + 0.8 = 6.5$, so the portfolio returned 6.5%. The first fund contributed most, 4.8 percentage points.

1.9 $\mathbf{p} \cdot \mathbf{s} = -1{,}500 - 1{,}320 + 6{,}600 + 2{,}800 = 6{,}580$ dollars. It bought 110 MWh and sold 100, a round trip of about 91%.

1.10 The volume shares are (0.2, 0.5, 0.3), and their dot product with the prices (40, 44, 50) is $8 + 22 + 15 = 45$. The VWAP is $45.00 per MWh.

1.11 Today minus day A is (1, 1, 1, −1, −2, −1), of length $\sqrt{9} = 3$. Today minus day B is (−2, −2, 2, 4, 5, 1), of length $\sqrt{54} \approx 7.3$. Day A is nearer.

1.12 For document 1, $5/(\sqrt{2} \times \sqrt{13}) \approx 0.981$. For document 2, $4/(\sqrt{2} \times 3) \approx 0.943$. Document 1 ranks first.

1.13 The coefficient is $14/10 = 1.4$. The residual is (2, −3, 3, 0) − 1.4(1, −2, 2, 1) = (0.6, −0.2, 0.2, −1.4), and its dot product with (1, −2, 2, 1) is $0.6 + 0.4 + 0.4 - 1.4 = 0$.

1.14 Use the definition of length and distribute the dot product:

$$\|\mathbf{a} + \mathbf{b}\|^2 = (\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) = \mathbf{a} \cdot \mathbf{a} + 2\,\mathbf{a} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{b}$$

When the vectors are perpendicular the middle term is zero, leaving $\|\mathbf{a} + \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$: Pythagoras' theorem in any number of dimensions.

1.15 Expanding as in the previous problem, $D(c) = \mathbf{b} \cdot \mathbf{b} - 2c\,\mathbf{a} \cdot \mathbf{b} + c^2\,\mathbf{a} \cdot \mathbf{a}$. Its derivative, $-2\mathbf{a} \cdot \mathbf{b} + 2c\,\mathbf{a} \cdot \mathbf{a}$, is zero at $c = \mathbf{a} \cdot \mathbf{b}/\mathbf{a} \cdot \mathbf{a}$, the projection's coefficient. Then $\mathbf{a} \cdot (\mathbf{b} - c\mathbf{a}) = \mathbf{a} \cdot \mathbf{b} - c\,\mathbf{a} \cdot \mathbf{a} = 0$.

1.16 The smallest squared distance cannot be negative. Substituting the best c into $D(c)$ gives $\mathbf{b} \cdot \mathbf{b} - (\mathbf{a} \cdot \mathbf{b})^2/\mathbf{a} \cdot \mathbf{a} \geq 0$, and multiplying by $\mathbf{a} \cdot \mathbf{a}$ gives the inequality. Dividing both sides by $\|\mathbf{a}\|^2\|\mathbf{b}\|^2$ shows that the squared cosine is at most 1. This is the Cauchy–Schwarz inequality.

1.17 With both lengths equal to 1, the cosine formula's denominator is 1, so the cosine is the dot product itself. The lengths are computed once, when each passage is stored, instead of for every query, so each comparison costs one dot product and nothing more.

Module 2: Matrices

2.1 Each row dotted with (1, −1) gives $1-2$, $3-4$, and $5-6$, so the product is (−1, −1, −1).

2.2 $\mathbf{AB}$ has rows (7, 2) and (3, 1); $\mathbf{BA}$ has rows (1, 2) and (3, 7). The two orders give different matrices.

2.3 $\mathbf{AB}$ is 2 × 4, $\mathbf{BC}$ is 3 × 2, $\mathbf{CA}$ is 4 × 3, and $\mathbf{ABC}$ is 2 × 2. $\mathbf{BA}$ is not defined: a 3 × 4 matrix cannot multiply a 2 × 3 one, because 4 does not match 2.

2.4 The transpose has rows (1, 4), (2, 5), and (3, 6). The row (1, 2) times the column (3, 4) is $3+8=11$.

2.5 The determinant is $24-14=10$, so the inverse is one-tenth of the matrix with rows (6, −7) and (−2, 4): rows (0.6, −0.7) and (−0.2, 0.4). Multiplying the two matrices gives the identity.

2.6 No. Its determinant is $4-4=0$: its second column is twice its first, so it squashes the plane onto a line.

2.7 With the funds as the columns of a matrix and the allocation (0.5, 0.3, 0.2), the product is $0.5(0.6,0.4,0)+0.3(0.3,0.6,0.1)+0.2(0.9,0,0.1)$ = (0.57, 0.38, 0.05): 57% stocks, 38% bonds, and 5% cash.

2.8 $\mathbf{Wx}$ = (2, 6, −2); adding the bias gives (2, 5, −1.5); the ReLU gives (2, 5, 0). The layer has $6+3=9$ parameters.

2.9 With the plants' outputs as the rows of $\mathbf{G}$ and the costs $\mathbf{c}=(10,40,120)$, the hourly costs are $\mathbf{c}^T\mathbf{G}$ = ($1,000, $3,000, $7,800, $1,800). The third hour costs most, because the peaking plant runs.

2.10 $\mathbf{M}$ times (4, 2, 1) is $(2\times4+1\times2+3\times1, 0+4\times2+1\times1)$ = (13, 9): 13 kilograms of steel and 9 of copper.

2.11 $\mathbf{RS}$ stretches first: (1, 0) becomes (2, 0), then turns to (0, 2). $\mathbf{SR}$ turns first: (1, 0) becomes (0, 1), which the stretch leaves at (0, 1). The orders give different results.

2.12 $\mathbf{W}_2\mathbf{W}_1$ has rows (7, 0) and (−1, −1). The two layers give $\mathbf{W}_1(1,1)$ = (3, 1, 2), then $\mathbf{W}_2$ = (7, −2). The single matrix gives (7, −2) too.

2.13 Entry (i,j) of $(\mathbf{AB})^T$ is entry (j,i) of $\mathbf{AB}$, which is $\Sigma_k\, a_{jk}\, b_{ki}$. Entry (i,j) of $\mathbf{B}^T\mathbf{A}^T$ is the sum over k of entry (i,k) of $\mathbf{B}^T$ times entry (k,j) of $\mathbf{A}^T$, which is $\Sigma_k\, b_{ki}\, a_{jk}$. The two sums are the same.

2.14 Multiply: $(\mathbf{AB})(\mathbf{B}^{-1}\mathbf{A}^{-1})=\mathbf{A}(\mathbf{B}\mathbf{B}^{-1})\mathbf{A}^{-1}=\mathbf{A}\mathbf{I}\mathbf{A}^{-1}=\mathbf{I}$. To undo two steps, undo the last one first.

2.15 There are $4{,}096^3\approx6.9\times10^{10}$ multiplications and nearly as many additions, about 1.4×10^{11} operations in all. At 10^{15} a second that is about 0.14 milliseconds, so such a chip can do several thousand of these products a second.

2.16 For $\mathbf{x}=(x,y)$, the rotated vector is $(x\cos\theta-y\sin\theta, x\sin\theta+y\cos\theta)$. Squaring and adding its components, the cross terms cancel and $\cos^2\theta+\sin^2\theta=1$ leaves x^2+y^2. Multiplying the transpose by the matrix gives diagonal entries $\cos^2\theta+\sin^2\theta=1$ and off-diagonal entries $-\cos\theta\sin\theta+\sin\theta\cos\theta=0$: the identity.

Module 3: Systems of Linear Equations

3.1 Adding the equations removes y: $3x=9$, so $x=3$, and then $y=1$.

3.2 Subtract twice the first equation from the second and the first from the third: $y+3z=8$ and $-3y+4z=2$. Add three times the new second equation to the third: $13z=26$.

So $z=2$, then $y=8-6=2$, and $x=3-4+2=1$.

3.3 $6 - 4 = 2$, which is not zero, so there is exactly one solution for any right-hand side. The matrix doubles areas.

3.4 The first pair is one line written twice: infinitely many solutions. In the second, subtracting twice the first equation from the second gives $0 = 1$: no solution.

3.5 Its determinant, $12 - 3c$, must be zero, so $c = 4$.

3.6 The pivots were 1, 1, and 13, with no row swaps, so the determinant is 13.

3.7 The call pays \$10 or nothing. Subtracting $40\Delta + 1.05B = 0$ from $60\Delta + 1.05B = 10$ gives $20\Delta = 10$, so $\Delta = 0.5$. Then $1.05B = -20$, so $B \approx -19.05$.

The call is worth $50 \times 0.5 - 19.05 \approx 5.95$ dollars.

3.8 Subtracting $90\Delta + 1.04B = 10$ from $120\Delta + 1.04B = 0$ gives $30\Delta = -10$, so $\Delta = -1/3$: the replicating portfolio sells a third of a share short. Then $1.04B = 40$, so $B \approx 38.46$.

The put is worth $-33.33 + 38.46 \approx 5.13$ dollars.

3.9 $\mathbf{I} - \mathbf{A}$ has rows (0.9, −0.4) and (−0.3, 0.8), with determinant $0.72 - 0.12 = 0.6$. Its inverse has rows (1.33, 0.67) and (0.5, 1.5).

Total output is (166.7, 250): \$166.7 of energy and \$250 of manufactured goods.

3.10 The balance at bus 1 becomes $(\theta_1 - \theta_2) + \theta_1/2 = 100$, and at bus 2 it is still $2\theta_2 - \theta_1 = 0$. The solution is $\theta_1 = 100$ and $\theta_2 = 50$.

The direct line carries $100/2 = 50$ MW and the path through bus 2 the other 50 MW. Both paths now have a reactance of 2, so they split the transfer evenly.

3.11 Solve $100a + 100b = 100$ and $100a + 101b = 102$: $b = 2$ and $a = -1$. For \$103, $b = 3$ and $a = -2$.

A change of one dollar in one payment moves both positions by a whole bond: nearly identical bonds make an ill-conditioned hedge.

3.12 With x and y the tonnes of each ore, $x + y = 100$ and $0.02x + 0.005y = 1.1$. Substituting $y = 100 - x$ gives $0.015x = 0.6$, so 40 tonnes of the richer ore and 60 of the poorer.

3.13 Multiply the first equation by d and the second by b, then subtract: $(ad - bc)x = ed - bf$. Multiply the first by c and the second by a, then subtract: $(bc - ad)y = ec - af$.

Dividing by the determinant gives both formulas. This is Cramer's rule for two equations.

3.14 $q = 14/30 \approx 0.467$. The weighted payoff is $0.467 \times 20 + 0.533 \times 0 \approx 9.33$, and discounting at 4% gives $9.33/1.04 \approx 8.97$, the replication price.

The number q is the risk-neutral probability: the probability that would make the stock earn exactly the bond's 4% on average.

3.15 The call is worth \$8.974 and the put, from problem 3.8, \$5.128. Their difference is \$3.846, and $100 - 100/1.04 \approx 3.846$. A call bought and a put sold always pay $S - 100$, where S is the stock's final price, so the pair must cost what the stock costs, less the strike's discounted value.

3.16 Both paths see the same angle difference, $\Delta\theta$, so their flows are $\Delta\theta/X_1$ and $\Delta\theta/X_2$. They must add to F, so $\Delta\theta = F X_1 X_2/(X_1 + X_2)$.

The first path's flow is then $\Delta\theta/X_1 = F X_2/(X_1 + X_2)$: each path's share is proportional to the other path's reactance.

Module 4: Least Squares

4.1 $\mathbf{A}$ has rows (1, 1), (1, 2), and (1, 3), and $\mathbf{b}$ = (2, 3, 5).

4.2 $\mathbf{A}^T\mathbf{A}$ has rows (3, 6) and (6, 14), and $\mathbf{A}^T\mathbf{b}$ = (10, 23). The determinant is $42 - 36 = 6$, so $c = (14 \times 10 - 6 \times 23)/6 \approx 0.333$ and $m = (3 \times 23 - 6 \times 10)/6 = 1.5$. The line is $y \approx 0.333 + 1.5x$.

4.3 The fitted values are 1.833, 3.333, and 4.833, so the residuals are 0.167, −0.333, and 0.167. They add to zero, and their dot product with (1, 2, 3) is $0.167 - 0.667 + 0.5 = 0$.

4.4 The averages are 4 and 4. The deviations are (−2, 0, 2) and (−3, 1, 2), so the slope is $(6 + 0 + 4)/(4 + 0 + 4) = 1.25$, and the intercept is $4 - 1.25 \times 4 = -1$.

4.5 The fitted values are 1.5, 4, and 6.5, so the residuals are −0.5, 1, and −0.5, with squares adding to 1.5. The squared deviations of y from its average add to $9 + 1 + 4 = 14$. So $R^2 = 1 - 1.5/14 \approx 0.89$.

4.6 $\mathbf{A}$ is 30 × 3, $\mathbf{A}^T\mathbf{A}$ is 3 × 3, and $\mathbf{x}$ has three components.

4.7 Weekday: $8.47 + 0.911 \times 31 + 2.59 \approx 39.3$ GW. Sunday: $8.47 + 0.911 \times 31 \approx 36.7$ GW.

4.8 The averages are 750 cycles and 96.325%. The sum of the products of the deviations is −6,275, and the sum of the squared cycle deviations is 1,250,000, so the slope is about −0.00502% per cycle, or −0.50% per 100 cycles. The intercept is $96.325 + 0.00502 \times 750 \approx 100.09$.

At 2,500 cycles the line gives about 87.5%. That is an extrapolation beyond the data, and battery fade often speeds up late in life.

4.9 The averages are 0% and 1%. The products of the deviations add to $1 + 4 + 9 + 0 + 4 = 18$, as do the squared market deviations, so $\beta = 1.0$, and $\alpha = 1 - 1.0 \times 0 = 1$ percent a month.

An alpha of 1% a month would be remarkable if it lasted, but five months prove nothing.

4.10 The fund's returns are explained by its tilt toward small and cheap stocks, exposures that low-cost index funds provide. With an alpha of zero, there is no evidence of skill beyond those tilts.

4.11 $\Sigma xy = 2 + 8 + 21 = 31$ and $\Sigma x^2 = 14$. With $\lambda = 0$, $w = 31/14 \approx 2.21$, the ordinary least-squares slope. With $\lambda = 6$, $w = 31/20 = 1.55$: the penalty shrinks the coefficient toward zero.

4.12 The model with the lowest error on held-back data, 0.8. The models after it fit the training data better only by fitting its noise, which is overfitting.

4.13 The partial derivative with respect to c is $-2\Sigma(y_i - c - mx_i)$ and with respect to m is $-2\Sigma x_i(y_i - c - mx_i)$. Setting both to zero gives $nc + m\Sigma x_i = \Sigma y_i$ and $c\Sigma x_i + m\Sigma {x_i}^2 = \Sigma x_i y_i$, where n is the number of points.

These are the rows of $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ for a matrix $\mathbf{A}$ with rows (1, x_i).

4.14 The first column of $\mathbf{A}$ is all ones, and the residual is perpendicular to it, so the residuals add to zero. Dividing the first normal equation, $nc + m\Sigma x_i = \Sigma y_i$, by n gives $c + m$ times the average of x equals the average of y.

4.15 With a single column $\mathbf{a}$ of x values, the normal equations are $(\mathbf{a} \cdot \mathbf{a})w = \mathbf{a} \cdot \mathbf{b}$, so $w = \mathbf{a} \cdot \mathbf{b}/\mathbf{a} \cdot \mathbf{a}$, the coefficient of the projection of $\mathbf{b}$ onto $\mathbf{a}$.

4.16 As λ shrinks toward zero, the ridge equations become the normal equations, and the solution approaches ordinary least squares. As λ grows, the penalty dominates and the penalized coefficients shrink toward zero. With the intercept unpenalized, the fit approaches a flat line at the average of y.

Module 5: Eigenvalues and Eigenvectors

5.1 The product is $(3 + 2, 2 + 8)$ = (5, 10), which is 5 times (1, 2). The eigenvalue is 5.

5.2 The trace is 7 and the determinant $12 - 2 = 10$, so $\lambda^2 - 7\lambda + 10 = 0$, which gives 5 and 2.

5.3 For 5, $\mathbf{A} - 5\mathbf{I}$ has rows (−1, 1) and (2, −2), so (1, 1) works. For 2, $\mathbf{A} - 2\mathbf{I}$ has rows (2, 1) and (2, 1), so (1, −2) works.

5.4 For a diagonal or triangular matrix, the eigenvalues are the diagonal entries: 3 and −2, then 2 and 4.

5.5 $5 + 2 = 7$, the trace, and $5 \times 2 = 10$, the determinant.

5.6 (1, 1) has eigenvalue 3, so $\mathbf{A}^3$ times it is 27 times (1, 1) = (27, 27). (1, 0) is half of (1, 1) plus half of (1, −1), so $\mathbf{A}^3$ times it is half of (27, 27) plus half of (1, −1) = (14, 13).

5.7 Balance: $0.2s = 0.1(1 - s)$, so $s = 1/3$. The second eigenvalue is $0.8 + 0.9 - 1 = 0.7$, so the gap to the steady state shrinks by 30% a year, halving in about 1.9 years.

5.8 The gap to the steady state, $1/2 - 1/3 = 1/6$, shrinks by the factor 0.85 each year, so after three years the share is $1/3 + (1/6) \times 0.85^3 \approx 0.436$.

5.9 With r_A, r_B, and r_C the ranks of the three pages, the ranks satisfy $r_A = 0.05 + 0.85(r_B/2 + r_C)$, $r_B = 0.05 + 0.85r_A$, and $r_C = 0.05 + 0.85r_B/2$. Solving this linear system by elimination gives $r_A \approx 0.397$, $r_B \approx 0.388$, and $r_C \approx 0.215$.

5.10 About 0.057, or 5.7%: the entry of $\mathbf{P}^2$ in the speculative-grade row and the investment-grade column.

5.11 The matrix has rows (0, 2) and (0.4, 0.5), so $\lambda^2 - 0.5\lambda - 0.8 = 0$, and the larger root is $(0.5 + \sqrt{3.45})/2 \approx 1.179$. The population grows about 17.9% a year in the long run.

5.12 The total is 176.3, so the shares are about 90.1%, 7.4%, 1.9%, and 0.6%. The first two components together explain 97.5%.

5.13 $\lambda_1(\mathbf{v}_1 \cdot \mathbf{v}_2) = (\mathbf{A}\mathbf{v}_1) \cdot \mathbf{v}_2 = \mathbf{v}_1 \cdot (\mathbf{A}\mathbf{v}_2) = \lambda_2(\mathbf{v}_1 \cdot \mathbf{v}_2)$; the middle step uses the symmetry of $\mathbf{A}$, which equals its transpose. Since $\lambda_1 \neq \lambda_2$, the dot product $\mathbf{v}_1 \cdot \mathbf{v}_2$ must be zero.

5.14 Let $\mathbf{u}$ be the vector of all ones. Each column adding to 1 means $\mathbf{u}^T\mathbf{P} = \mathbf{u}^T$, so 1 is an eigenvalue of $\mathbf{P}^T$. A matrix and its transpose have the same determinant, so $\det(\mathbf{P} - \mathbf{I}) = \det(\mathbf{P}^T - \mathbf{I}) = 0$, and 1 is an eigenvalue of $\mathbf{P}$ too.

5.15 With λ_1 and λ_2 the first and second eigenvalues, after k steps the two parts have been multiplied by ${\lambda_1}^k$ and ${\lambda_2}^k$, so their ratio has been multiplied by $(\lambda_2/\lambda_1)^k$. For the fish stock the ratio is $0.554/1.354 \approx 0.41$ in size, so the second part shrinks to about 41% of its relative size each year.

5.16 If the roots are λ_1 and λ_2, the quadratic equals $(\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2$. Matching coefficients, $\lambda_1 + \lambda_2 = a + d$ and $\lambda_1\lambda_2 = ad - bc$.

Module 6: Covariance and Portfolio Risk

6.1 The average is 4, and the deviations are −2, 0, 0, and 2, whose squares add to 8. The sample variance is $8/3 \approx 2.67$, and the standard deviation is $\sqrt{2.67} \approx 1.63$.

6.2 $4\% \times \sqrt{12} \approx 4\% \times 3.46 \approx 13.9\%$.

6.3 The dot product of the deviation vectors is $0 + 0 + 1 = 1$, and each has length $\sqrt{2}$. The correlation is the cosine of the angle between them, $1/(\sqrt{2} \times \sqrt{2}) = 0.5$.

6.4 The variances are $0.2^2 = 0.04$ and $0.1^2 = 0.01$, and the covariance is $-0.2 \times 0.2 \times 0.1 = -0.004$. So $\mathbf{\Sigma}$ has rows (0.04, −0.004) and (−0.004, 0.01).

6.5 The variance is $0.25 \times 0.04 + 0.25 \times 0.01 + 2 \times 0.25 \times (-0.004) = 0.0105$, so the volatility is $\sqrt{0.0105} \approx 10.2\%$, below the 15% average of the two volatilities.

6.6 The volatilities are $\sqrt{0.09} = 30\%$ and $\sqrt{0.04} = 20\%$. The correlation is $0.012/(0.3 \times 0.2) = 0.2$.

6.7 The three terms of the variance are $0.36 \times 0.0324 = 0.011664$, $0.16 \times 0.0049 = 0.000784$, and $2 \times 0.6 \times 0.4 \times 0.3 \times 0.18 \times 0.07 \approx 0.001814$. They add to about 0.01426, a volatility of about 11.9%. The weighted average is $0.6 \times 18\% + 0.4 \times 7\% = 13.6\%$.

6.8 The variance is $0.1225/20 + (19/20) \times 0.25 \times 0.1225 \approx 0.0352$, a volatility of about 18.8%. The floor is $35\% \times \sqrt{0.25} = 17.5\%$.

6.9 With no covariance, stocks contribute $0.5 \times 0.5 \times 0.0256 = 0.0064$ and bonds $0.5 \times 0.5 \times 0.0036 = 0.0009$. Stocks bring $0.0064/0.0073 \approx 88\%$ of the risk with half the money.

6.10 The weights are proportional to $1/18$ and $1/7$, which gives 28% stocks and 72% bonds. The variance is $0.28^2 \times 0.0324 + 0.72^2 \times 0.0049 \approx 0.00508$, a volatility of about 7.13%. Leverage of $10/7.13 \approx 1.40$ times brings it to 10%.

6.11 The averages are 0.4 and 0. The deviations are (1.6, −1.4, 2.6, −2.4, −0.4) and (1, −1, 2, −1, −1). Their dot product is 11, and their squared lengths are 17.2 and 8. The correlation is $11/\sqrt{(17.2 \times 8)} \approx 0.94$.

6.12 $\mathbf{\Sigma w}$ = (0.011104, 0.002364, 0.00654), and its dot product with the weights is $0.4 \times 0.011104 + 0.4 \times 0.002364 + 0.2 \times 0.00654 \approx 0.006695$. The volatility is $\sqrt{0.006695} \approx 8.18\%$.

6.13 $\mathbf{\Sigma}$ has rows (σ_1^2, $\rho\sigma_1\sigma_2$) and ($\rho\sigma_1\sigma_2$, σ_2^2), so $\mathbf{\Sigma w}$ = ($w_1\sigma_1^2 + w_2\rho\sigma_1\sigma_2$, $w_1\rho\sigma_1\sigma_2 + w_2\sigma_2^2$). Its dot product with $\mathbf{w}$ is $w_1^2\sigma_1^2 + 2w_1w_2\rho\sigma_1\sigma_2 + w_2^2\sigma_2^2$.

6.14 Each weight is $1/n$, so each term of $\mathbf{w}^T\mathbf{\Sigma w}$ is $1/n^2$ times an entry of $\mathbf{\Sigma}$. The n diagonal entries give $n\sigma^2/n^2 = \sigma^2/n$. The $n(n-1)$ others give $n(n-1)\rho\sigma^2/n^2 = (1 - 1/n)\rho\sigma^2$.

6.15 Asset 1 contributes $w_1^2\sigma_1^2 + w_1w_2\rho\sigma_1\sigma_2$, and asset 2 contributes $w_2^2\sigma_2^2 + w_1w_2\rho\sigma_1\sigma_2$. The shared term is the same in both, so the contributions are equal exactly when $w_1^2\sigma_1^2 = w_2^2\sigma_2^2$. For positive weights, that means $w_1\sigma_1 = w_2\sigma_2$.

6.16 With $n = 3$ and $\rho = -0.9$, the formula of section 6e gives $\sigma^2/3 + (2/3)(-0.9)\sigma^2 \approx -0.27\sigma^2$, where σ is the common volatility. A variance cannot be negative, so the answer is no.

If the first asset moves strongly against the second, and the third strongly against the first, then the third must move with the second. The formula shows that a correlation shared by all three pairs can be no lower than −0.5.

Module 7: Optimization with Constraints

7.1 The partial derivatives are $2x + 3y$ and $3x$, so the gradient at (1, 2) is (8, 3).

7.2 The Lagrange condition gives $2x = \lambda$ and $2y = 2\lambda$, so $y = 2x$. The constraint gives $x + 4x = 5$, so $x = 1$, $y = 2$, and $\lambda = 2$. The minimum is $1 + 4 = 5$.

7.3 The gradient of xy is (y, x) and that of $x + y$ is (1, 1), so $y = \lambda$ and $x = \lambda$. Then $x = y = 5$, $\lambda = 5$, and the maximum is 25.

7.4 The estimate is $25 + \lambda \times 0.2 = 25 + 5 \times 0.2 = 26$. Exactly, $x = y = 5.1$, and the maximum is $5.1^2 = 26.01$.

7.5 With $\rho = 0$, $w_1 = 0.01/(0.04 + 0.01) = 0.2$, so the weights are 20% and 80%. The variance is $0.2^2 \times 0.04 + 0.8^2 \times 0.01 = 0.008$, a volatility of about 8.94%.

7.6 Each produces until its marginal cost equals \$40: $P_1 = (40 - 20)/0.05 = 400$ MW and $P_2 = (40 - 30)/0.02 = 500$ MW.

7.7 From the marginal costs, $P_1 = 20(\lambda - 20)$ and $P_2 = 50(\lambda - 30)$. They must add to 600, so $70\lambda - 1{,}900 = 600$, and $\lambda \approx 35.71$. Then $P_1 \approx 314.3$ MW and $P_2 \approx 285.7$ MW.

7.8 Each extra dollar of price adds 20 MW from generator 1 and 50 MW from generator 2, so the price rises by $1/70 \approx 0.014$ dollars per MWh. Generator 1 supplies $20/70 \approx 29\%$ of the extra megawatt and generator 2 the other 71%: the generator with the flatter marginal-cost line takes the larger share.

7.9 Generator 1 runs at 300 MW, and generators 2 and 3 share the other 700: $50(\lambda - 22) + 20(\lambda - 30) = 700$, so $\lambda = 2{,}400/70 \approx 34.29$. Then $P_2 \approx 614.3$ MW and $P_3 \approx 85.7$ MW.

Generator 1's marginal cost at 300 MW is $20 + 0.04 \times 300 = 32$, so the shadow price of its limit is about $34.29 - 32 = 2.29$ dollars per MWh.

7.10 Here $\rho\sigma_1\sigma_2 = -0.004$, so $w_1 = (0.01 + 0.004)/(0.04 + 0.01 + 0.008) \approx 0.241$. The weights are about 24.1% and 75.9%, and the volatility is about 8.14%.

7.11 The Lagrange condition is $5/\sqrt{x} = \lambda$ and $3/\sqrt{y} = \lambda$, so $\sqrt{x}/\sqrt{y} = 5/3$ and $x/y = 25/9$. With $x + y = 17$, $x = 12.5$ and $y = 4.5$.

Then $\lambda = 5/\sqrt{12.5} \approx 1.41$: one more million of budget adds about \$1.41 million of value.

7.12 The variance is $0.25 \times 0.0256 + 0.25 \times 0.0036 + 2 \times 0.25 \times 0.00096 = 0.00778$, a volatility of about 8.82%. The frontier portfolio with 5.5% expected return has a volatility of 8.52%, so adding gold cuts the risk by 0.3 percentage points with no loss of return.

7.13 The variance is $w_1{}^2\sigma_1{}^2 + (1 - w_1)^2\sigma_2{}^2 + 2w_1(1 - w_1)\rho\sigma_1\sigma_2$. Its derivative with respect to w_1 is $2w_1\sigma_1{}^2 - 2(1 - w_1)\sigma_2{}^2 + 2(1 - 2w_1)\rho\sigma_1\sigma_2$.

Setting it to zero gives $w_1(\sigma_1{}^2 + \sigma_2{}^2 - 2\rho\sigma_1\sigma_2) = \sigma_2{}^2 - \rho\sigma_1\sigma_2$, the formula of section 7d.

7.14 Setting the gradient with respect to $\mathbf{w}$ to zero gives $2\boldsymbol{\Sigma}\mathbf{w} - \lambda\mathbf{1} = \mathbf{0}$, so $\mathbf{w} = (\lambda/2)\boldsymbol{\Sigma}^{-1}\mathbf{1}$. The constraint $\mathbf{1}^T\mathbf{w} = 1$ then requires $\lambda/2 = 1/(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1})$.

7.15 Each marginal cost, $a_i + 2b_iP_i$, equals λ, so $P_i = (\lambda - a_i)/(2b_i)$. Adding over the generators, $D = \lambda\Sigma(1/(2b_i)) - \Sigma a_i/(2b_i)$, which rearranges to the formula.

In section 7c, the values of $1/(2b_i)$ are 25, 50, and 20, adding to 95, and the values of $a_i/(2b_i)$ are 500, 1,100, and 600, adding to 2,200. So $\lambda = (1{,}000 + 2{,}200)/95 \approx 33.68$, as before.

7.16 Without the limit, generator 1 produces about 342 MW, which leaves 58 MW of room under 400. The limit has slack, so by complementary slackness its multiplier is zero. The dispatch, the price, and the cost are all unchanged.

Module 8: Linear Programming

8.1 The corners are (0, 0), (4, 0), (3, 1), and (0, 2), where the two limits cross at (3, 1). The objective there is 0, 12, 11, and 4, so the best is 12 at (4, 0).

8.2 With prices u and v for the two limits: minimize $4u + 6v$ subject to $u + v \geq 3$, $u + 3v \geq 2$, $u \geq 0$, and $v \geq 0$. The second limit of the primal has room to spare at (4, 0), so $v = 0$, and then $u = 3$, giving $4 \times 3 = 12$.

8.3 The equations are $x + y + s_1 = 100$ and $2x + y + s_2 = 150$. At (75, 0), $s_1 = 25$ idle pods and $s_2 = 0$. At (50, 50), both slacks are zero.

8.4 At (4, 0), the limit $x + y \le 4$ binds, with shadow price 3, and so does $y \ge 0$. The limit $x + 3y \le 6$ has 2 units of room, so its shadow price is zero.

8.5 Yes, barely. Each MWh bought returns 0.9 MWh, worth $0.9 \times 34 = 30.60$ dollars, against a cost of \$30. The trade earns 60 cents per MWh bought.

8.6 $(2/3) \times 80 + (1/3) \times 20 = 53.3 + 6.7 = 60$ MW, which exceeds the line's 50 MW limit.

8.7 Both limits still bind: $x + y = 100$ and $2x + y = 160$ give $x = 60$ and $y = 40$. Profit is $40 \times 60 + 30 \times 40 = 3{,}600$, up 100, which is 10 MW times the shadow price of 10.

8.8 The best corner is where $x + y = 100$ meets $2x + y = P$, at $x = P - 100$ and $y = 200 - P$, for P megawatts of power. Both must be zero or above, so the price holds from 100 MW to 200 MW.

Above 200 MW every pod trains and power has room to spare, so its shadow price drops to zero.

8.9 Charge fully in hour 1 for \$20, storing 0.8 MWh. In hour 2, sell 0.6 MWh for \$30 and keep 0.2 MWh. Charge fully again in hour 3 for \$30, which adds 0.8 MWh and fills the battery. Sell the full 1 MWh in hour 4 for \$80. The profit is $30 + 80 - 20 - 30 = 60$ dollars.

Selling all of the first charge in hour 2 would earn \$54, and holding all of it for hour 4 only \$52.50. Keeping back 0.2 MWh fills exactly the room that hour 3's charge, limited to 1 MW, cannot.

8.10 The flow on the direct line is $(2/3)g_1 + (1/3)(100 - g_1) = 33.3 + g_1/3$, which reaches 60 MW at $g_1 = 80$. So generator 1 produces 80 MW and generator 2 produces 20 MW, at a cost of $1{,}600 + 800 = 2{,}400$ dollars an hour.

Both generators are still running below capacity and the line is still full, so the prices stay at \$20, \$40, and \$60.

8.11 Generator 1 serves it locally, rising to 70 MW, with no change in any line's flow. The cost rises by $20 \times 20 = 400$ dollars an hour, matching the price of \$20 at bus 1.

8.12 $(60 - 20) \times 10 = 400$ dollars. A buyer at bus 3 that has contracted for 10 MW from bus 1 pays that much extra in congestion, so the right offsets it.

8.13 Since x and y are zero or above and $u + 2v \ge 40$ and $u + v \ge 30$, the profit is at most $(u + 2v)x + (u + v)y = u(x + y) + v(2x + y)$. The limits make that at most $100u + 150v$.

So every feasible plan earns at most any feasible valuation, which is why the two best values can meet but never cross.

8.14 At (0, 100) the pod limit binds, so $y = 100 - x - s_1$, and the profit is $40x + 30(100 - x - s_1) = 3{,}000 + 10x - 30s_1$. Increasing x pays, and the power limit, $s_2 = 50 - x + s_1$, stops it at $x = 50$.

At (50, 50), the profit is $3{,}500 - 20s_1 - 10s_2$. Both slacks lower it, so the corner is best, and the coefficients are the shadow prices, 20 and 10.

8.15 With net withdrawal w_k at bus k, the rent is $\Sigma_k (\lambda - \mu\varphi_k)w_k$. Withdrawals sum to zero over the grid, so the λ part vanishes, leaving $-\mu\Sigma_k \varphi_k w_k$.

The flow on the line is the sum of each bus's net injection, $-w_k$, times its shift factor, so the rent is μ times the flow: $60 \times 50 = 3{,}000$ dollars.

8.16 Cutting both by matching amounts, charging δ less and discharging $\eta\delta$ less, leaves the stored energy unchanged and saves $p_t\delta - p_t\eta\delta = (1 - \eta)p_t\delta$, which is positive when the price is positive.

When the price is negative, the same change costs money, so a battery may charge and discharge at once to waste energy on purpose and be paid for taking it.

Module 9: Convexity and Gradient Methods

9.1 Yes: $f''(x) = 2$, which is positive everywhere. Setting $f'(x) = 2x - 4$ to zero gives $x = 2$, where $f = 3$.

9.2 The Hessian has rows (6, 2) and (2, 2). Its trace is 8 and its determinant 8, so its eigenvalues solve $\lambda^2 - 8\lambda + 8 = 0$, giving about 6.83 and 1.17. Both are positive, so f is convex.

9.3 The gradient at 1 is $2(1 - 5) = -8$, so the step goes to $1 - 0.1 \times (-8) = 1.8$.

9.4 The second derivative is 8, so each step multiplies x by $1 - 8\alpha$. Convergence needs $\alpha < 2/8 = 0.25$, and $\alpha = 1/8$ reaches the minimum in one step.

9.5 $\kappa = 50$. Plain descent: $49/51 \approx 0.961$. Momentum: $(\sqrt{50} - 1)/(\sqrt{50} + 1) \approx 6.07/8.07 \approx 0.752$.

9.6 $50{,}000/100 = 500$ steps.

9.7 Each step multiplies x by $1 - 0.05 \times 2 = 0.9$ and y by $1 - 0.05 \times 20 = 0$. After one step the point is (9, 0), and after two (8.1, 0).

The x part shrinks by 0.9 a step, so it needs $10 \times 0.9^k < 0.01$, or about 66 steps.

9.8 The loss is $\Sigma(wx_i - y_i)^2$, with gradient $2(14w - 29)$, since $\Sigma x_i^2 = 14$ and $\Sigma x_i y_i = 29$. Step 1: $w = 0 + 0.02 \times 58 = 1.16$. Step 2: the gradient is $2(16.24 - 29) = -25.52$, so $w \approx 1.67$.

The exact least-squares answer is $29/14 \approx 2.07$. Each step shrinks the gap by the factor $1 - 0.02 \times 28 = 0.44$.

9.9 The Hessian's eigenvalues are 2 and 200, so α must be below $2/200 = 0.01$. The condition number is $200/2 = 100$.

9.10 On the first step the corrected averages are the gradient and its square, so the step is $\alpha \times g/|g|$: Adam moves the parameter up by 0.001. Plain descent moves it by $0.001 \times 0.002 = 0.000002$, five hundred times less.

9.11 Noise shrinks like $1/\sqrt{B}$, so halving it takes four times the batch: 400 examples. Each step then costs four times as much computation.

9.12 $f'(10) = 16$ and $f''(10) = 2$, so the step goes to $10 - 16/2 = 2$, the minimum, in one step, as it always does for a quadratic.

9.13 The gradient is $\mathbf{Hx}$, so a step gives $(\mathbf{I} - \alpha\mathbf{H})\mathbf{x}$. Write $\mathbf{x}$ as a combination of eigenvectors. Each eigenvector $\mathbf{u}$ with eigenvalue λ is sent to $\mathbf{u} - \alpha\lambda\mathbf{u} = (1 - \alpha\lambda)\mathbf{u}$, so its part is multiplied by $1 - \alpha\lambda$.

9.14 The slowest part shrinks by the larger of $|1 - \alpha\lambda_{\min}|$ and $|1 - \alpha\lambda_{\max}|$. Raising α lowers the first and raises the second, so the best α makes them equal: $1 - \alpha\lambda_{\min} = \alpha\lambda_{\max} - 1$, or $\alpha = 2/(\lambda_{\min} + \lambda_{\max})$.

The factor is then $1 - 2\lambda_{\min}/(\lambda_{\min} + \lambda_{\max}) = (\lambda_{\max} - \lambda_{\min})/(\lambda_{\max} + \lambda_{\min})$, which is $(\kappa - 1)/(\kappa + 1)$.

9.15 Suppose x^* is a local minimum but some point y has $f(y) < f(x^*)$. By convexity, every point $tx^* + (1 - t)y$ on the segment between them has f at most $tf(x^*) + (1 - t)f(y)$, which is below $f(x^*)$ for any t less than 1.

Points with t just below 1 lie arbitrarily close to x^*, so x^* would not be a local minimum after all.

9.16 $f'(x) = 4x^3 - 4x = 0$ at $x = 0$ and $x = \pm 1$. Since $f''(x) = 12x^2 - 4$, $x = 0$ is a local maximum, with $f'' = -4$, and $x = \pm 1$ are minima, with $f'' = 8$ and $f = -1$.

From 0.1, descent rolls to 1; from −0.1, it rolls to −1. Here both valleys are equally deep, but in general a nonconvex function's result depends on the starting point.

Module 10: Linear Algebra Inside AI

10.1 $\mathbf{Wx} = (2 - 3, 1 + 3) = (-1, 4)$, and adding $\mathbf{b}$ gives (−1, 2). The ReLU turns that into (0, 2).

10.2 $784 \times 128 + 128 = 100{,}480$ for the hidden layer and $128 \times 10 + 10 = 1{,}290$ for the output layer, 101,770 in all.

10.3 $e^2 \approx 7.389$, $e^1 \approx 2.718$, and $e^0 = 1$, which add to about 11.107. The weights are about 0.665, 0.245, and 0.090.

10.4 Each term becomes $e^{z_j + c} = e^c e^{z_j}$, so the factor e^c appears in every term of the numerator and the denominator, and cancels.

10.5 $2 \times 4{,}096^2 \approx 33.6$ million.

10.6 Each parameter takes 2 bytes, so $7 \times 10^9 \times 2 = 14 \times 10^9$ bytes, 14 GB.

10.7 About $2N = 1.4 \times 10^{10}$ operations per token, and $6ND = 6 \times (7 \times 10^9) \times (2 \times 10^{12}) = 8.4 \times 10^{22}$ to train.

10.8 The effective rate is $0.4 \times 989 \times 10^{12} \approx 3.96 \times 10^{14}$ operations a second. Then $8.4 \times 10^{22} / (3.96 \times 10^{14}) \approx 2.1 \times 10^8$ seconds, about 59,000 GPU-hours.

10.9 The dot products are 2, 0, and 1, which become 1.414, 0, and 0.707 after dividing by $\sqrt{2}$. Their exponentials are about 4.113, 1, and 2.028, adding to 7.141, so the weights are about 0.576, 0.140, and 0.284.

10.10 It grows with the square of n, so doubling the length multiplies the cost by four.

10.11 $16 \times (8{,}192 + 8{,}192) = 262{,}144$ numbers, against $8{,}192^2 \approx 67.1$ million, about 0.39%.

10.12 $20 \times (2{,}000 + 500 + 1) = 50{,}020$ numbers, against 1,000,000, about 5%.

10.13 $\mathbf{W}_2(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2 = (\mathbf{W}_2\mathbf{W}_1)\mathbf{x} + (\mathbf{W}_2\mathbf{b}_1 + \mathbf{b}_2)$, a single layer with weight matrix $\mathbf{W}_2\mathbf{W}_1$ and bias $\mathbf{W}_2\mathbf{b}_1 + \mathbf{b}_2$.

10.14 The weight W_{ij} affects only $z_i = \Sigma_j W_{ij} x_j$, and changes it at the rate x_j. By the chain rule its gradient is $\delta_i x_j$, the entry in row i and column j of the outer product $\boldsymbol{\delta}\mathbf{x}^T$.

10.15 $\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$, since the columns of $\mathbf{U}$ are perpendicular unit vectors. Multiplying by $\mathbf{V}$ gives $\mathbf{A}^T\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Sigma}^2$: each column of $\mathbf{V}$ is an eigenvector, with eigenvalue $\sigma_i{}^2$.

10.16 $\sigma_2 \mathbf{u}_2 \mathbf{v}_2{}^T = \sqrt{10} \times (2, -1)^T (1, 1) / (\sqrt{5} \times \sqrt{2})$, and $\sqrt{5} \times \sqrt{2} = \sqrt{10}$, so it is the matrix with rows (2, 2) and (−1, −1). That is exactly the original minus the rank-1 approximation.

Answer Sheet

Answers to every module's Check Your Understanding quiz.

Module	Answers
Module 1 Quiz	1. B 2. A 3. C 4. D 5. A
Module 2 Quiz	1. D 2. B 3. C 4. C 5. D
Module 3 Quiz	1. B 2. A 3. B 4. D 5. A
Module 4 Quiz	1. C 2. D 3. B 4. A 5. C
Module 5 Quiz	1. B 2. C 3. A 4. D 5. A
Module 6 Quiz	1. D 2. B 3. C 4. D 5. A
Module 7 Quiz	1. B 2. C 3. C 4. B 5. A
Module 8 Quiz	1. D 2. B 3. C 4. A 5. D
Module 9 Quiz	1. B 2. A 3. C 4. D 5. D
Module 10 Quiz	1. B 2. A 3. C 4. B 5. A

Probability and Statistics

Figures current as of October 2026

The mathematics of uncertainty and of learning from data, for adults who want to use it: the rules of probability, Bayes' rule, expected value and risk, the distributions behind counts, prices, and failures, the law of averages, estimation, testing, regression, and how a model is fit to data and judged. Every idea is attached to something real, from a power grid's reliability to a portfolio's worst days to an artificial intelligence (AI) model's training, and every module ends with practice problems solved in full at the back.

MODULE 1 — Probability and Its Rules

Goal: Put numbers on uncertainty and combine them correctly.

Outcomes, events, and the rules for not, or, and and; independence and its failures; counting arrangements; systems in series and in parallel, and why the chance that something somewhere goes wrong grows so fast.

MODULE 2 — Conditional Probability and Bayes' Rule

Goal: Update a probability when evidence arrives.

Conditional probability; Bayes' rule and base rates, from a medical test to fraud alerts; odds and likelihood ratios; how a spam filter learns from words; weighing a claim of skill.

MODULE 3 — Random Variables and Expected Value

Goal: Value a gamble, a contract, or a decision.

Random variables and their distributions; expected value and its rules; variance and standard deviation; pricing insurance, the value of better information, and why the average is not the whole story.

MODULE 4 — Counting Events: Binomial and Poisson

Goal: Model how many times something happens.

The binomial distribution for successes in repeated trials, from loan defaults to experiments on websites; the Poisson distribution for rare events, from outages to insurance claims; and what happens when events are not independent.

MODULE 5 — Continuous Distributions and Fat Tails

Goal: Model quantities that vary smoothly, and their extremes.

Densities and probabilities as areas; the normal distribution and z-scores; the lognormal for prices; the exponential for waiting times and failures; fat tails, Value at Risk, and expected shortfall.

MODULE 6 — Sums, Averages, and the Central Limit Theorem

Goal: See why averages are predictable when single outcomes are not.

Means and variances of sums; the law of large numbers; the central limit theorem and the standard error; insurance pools; Monte Carlo simulation of a battery's year; and why shared risks do not average away.

MODULE 7 — Estimation and Confidence Intervals

Goal: Estimate an unknown from a sample, and say how sure you are.

Samples and estimators; standard errors; confidence intervals for averages and proportions, from polls to government surveys; small samples and the t distribution; the bootstrap.

MODULE 8 — Hypothesis Tests

Goal: Decide whether a difference is real.

Null hypotheses, test statistics, and p-values, and what they do not mean; the two kinds of error and statistical power; experiments on websites; multiple comparisons, overfitted backtests, and the replication crisis.

MODULE 9 — Regression and Its Uncertainty

Goal: Fit a relationship and know how far to trust it.

Least squares from the statistician's side; standard errors, t-statistics, and intervals for coefficients and forecasts; several inputs at once; checking residuals; and why correlation is not causation.

MODULE 10 — Fitting and Testing Models

Goal: See how models, AI models included, are fit to data and judged.

Likelihood and maximum likelihood; logistic regression for probabilities; cross-entropy, the loss that language models minimize; training, validation, and test data; accuracy, precision, recall, and calibration.

How to use this course. Start with the diagnostic on the next page; it says which modules to work and which to skim. Each module teaches through worked examples, then shows the spreadsheet functions that do the same work at real size. It ends with a Memorize This box of the formulas, definitions, and ideas to commit to memory, practice problems in three tiers, and a five-question quiz. Every symbol is defined in words where it first appears in a module and again in its Memorize This box. Worked solutions to every practice problem are at the back, after all the Memorize This boxes gathered in one place; quiz answers are on the answer sheet. Work the small examples by hand and the large ones in a spreadsheet. The examples use illustrative numbers unless a source and date are given.

Where to Start

A twenty-problem diagnostic

Work these on paper before you start; a calculator is fine throughout. The four bands follow the order of the course, except the last, which covers algebra the course uses rather than teaches. The answers are on the next page, with a guide to where to begin.

Band A — Fractions, percentages, and powers (Modules 1 to 4)

1. Multiply 0.2 by 0.15.

2. Evaluate $1 - 0.9^3$.

3. Write 3/8 as a percentage.

4. Evaluate $5 \times 4 \times 3 \times 2 \times 1$.

5. A price rises 10% and then falls 10%. What is the overall change?

Band B — Chance and counting (Modules 1 and 2)

6. A fair coin is tossed three times. What is the chance of three heads?

7. A fair die is rolled. What is the chance of an even number or a 5?

8. In how many ways can 2 people be chosen from 5, when order does not matter?

9. Two independent events have probabilities 0.5 and 0.2. What is the chance that both happen?

10. An event has probability 0.03. What is the probability that it does not happen?

Band C — Averages and spread (Modules 3, 6, and 7)

11. Find the average of 3, 7, 8, 10, and 12.

12. Find the median, the middle value, of 2, 9, 4, 7, and 100.

13. Evaluate the weighted average $0.3 \times 10 + 0.7 \times 20$.

14. The numbers 4 and 8 have an average of 6. What is the average of their squared distances from 6?

15. Evaluate $\sqrt{(0.25 \times 0.75 / 100)}$.

Band D — Algebra, logarithms, and graphs (Modules 2, 3, 5, 9, and 10)

16. Solve $3x - 7 = 11$.

17. Evaluate $\ln(e^3)$, where ln is the natural logarithm.

18. Evaluate $\Sigma_{i=1}^{3} 2i$, the sum of $2i$ as i runs from 1 to 3.

19. On the line $y = 2 + 0.5x$, what is y when $x = 10$, and how much does y rise for each unit of x?

20. Find the value of x that makes $10x - x^2$ as large as possible.

Answers, and Where to Begin

#	Answer	Where
1	0.03	Module 1
2	0.271	Module 1
3	37.5%	Module 1
4	120	Module 1
5	A fall of 1%: $1.1 \times 0.9 = 0.99$	Module 3
6	1/8	Module 1
7	4/6, or 2/3	Module 1
8	10	Module 1
9	0.1	Module 1
10	0.97	Module 1
11	8	Module 3
12	7	Module 7
13	17	Module 3
14	4	Module 3
15	About 0.0433	Module 7
16	$x = 6$	Used from Module 5
17	3	Used from Module 2
18	12	Used from Module 3
19	7, and 0.5	Used from Module 9
20	$x = 5$	Used from Module 10

Score each band out of five, then use the table. A low score on a band means working every module the band covers.

Band score	What to do
5 of 5	Skim the modules the band covers: read each Memorize This box and try the challenge problems.
3 or 4	Read those modules quickly and work the applied problems.
0 to 2	Work those modules through, worked examples and all.

The last band is the one this course does not teach from scratch. If it goes badly, review linear equations, logarithms, and graphs of lines in any algebra text; Modules 1 to 4 need very little of it. If the first two bands go badly, work Modules 1 and 2 slowly, since everything after them is built on their rules.

Module 1 — Probability and Its Rules

Outcomes, events, and the rules for combining chances

Recommended Reading / Reference

Book: Joseph K. Blitzstein and Jessica Hwang, *Introduction to Probability* (2nd ed., 2019) — the text of Harvard's introductory probability course, free online with its recorded lectures

Reference: North American Electric Reliability Corporation (NERC), Long-Term Reliability Assessment (annual) — the grid's probabilistic checks on whether supply will cover demand, region by region

1a. Outcomes, Events, and Probabilities

A probability is a number from 0 to 1 that measures how likely something is: an impossible event has probability 0, and a certain one probability 1. It has two readings, and this course uses both. One is a long-run frequency: a 30% chance of rain means it rains on about 30% of days like this one. The other is a degree of belief, such as a 20% chance that a new drug wins approval, where no long run exists.

The vocabulary is small. An experiment, such as rolling a die, has outcomes, and the set of all of them is the sample space. An event is a set of outcomes, such as rolling an even number. When the outcomes are equally likely, the probability of an event is its number of outcomes divided by the total:

$$P(A) = \frac{\text{number of outcomes in } A}{\text{total number of outcomes}}$$

Here $P(A)$ is the probability of an event A. Rolling two dice has 36 equally likely outcomes, and six of them sum to 7, so the chance of a 7 is $6/36 = 1/6$. Most real probabilities, such as the chance that a power plant trips or a borrower defaults, come instead from data: the share of past cases in which it happened.

1b. Not, Or, And

Three rules combine probabilities. The complement of an event A, written A^c and read not A, happens whenever A does not, so

$$P(A^c) = 1 - P(A)$$

The event that A or B happens, or both, is written $A \cup B$, and the event that both happen is written $A \cap B$. Adding $P(A)$ and $P(B)$ counts the outcomes in both twice, so the addition rule subtracts them once:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Worked example — Solar panels and electric cars

In a utility's survey, 40% of customers have solar panels, 25% have an electric vehicle (EV), and 15% have both. The share with at least one is $0.40 + 0.25 - 0.15 = 0.50$, half of all customers, and the share with neither is $1 - 0.50 = 0.50$.

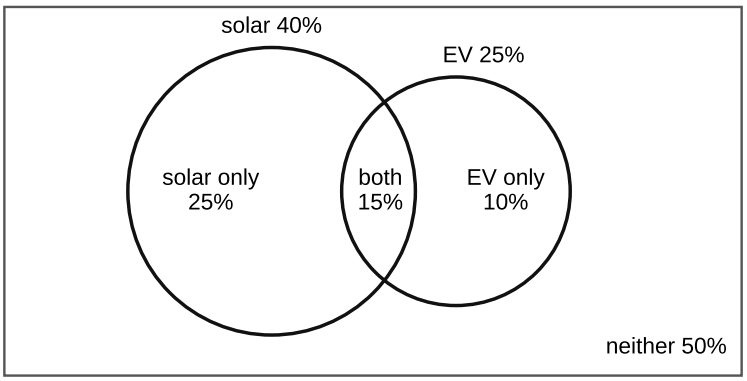


Figure 1.1. The survey's customers. Adding the two circles counts the 15% with both twice, so the addition rule subtracts it once.

Events that cannot happen together, such as a die showing 2 and showing 5, are mutually exclusive. Their overlap is empty, so their probabilities simply add.

1c. Independence and the Multiplication Rule

Two events are independent when knowing that one happened does not change the chance of the other. Then the chance that both happen is the product:

$$P(A \cap B) = P(A)\,P(B)$$

Solar panels and electric cars are not independent in the survey. If they were, $0.40 \times 0.25 = 10\%$ of customers would have both, not 15%: owning one makes owning the other more likely.

Independence is what makes redundancy work. If each of two transformers at a substation fails on a given day with probability 0.001, independently, both fail together with probability 0.001^2, one in a million. The same arithmetic, run the other way, explains why large systems have problems constantly. The chance that none of n independent components fails is the product of their chances of working, so

$$P(\text{at least one fails}) = 1 - (1 - p)^n$$

where p is each component's chance of failing. A data center with 1,000 servers, each failing on a given day with probability 0.001, has a $1 - 0.999^{1{,}000} \approx 63\%$ chance of at least one failure that day.

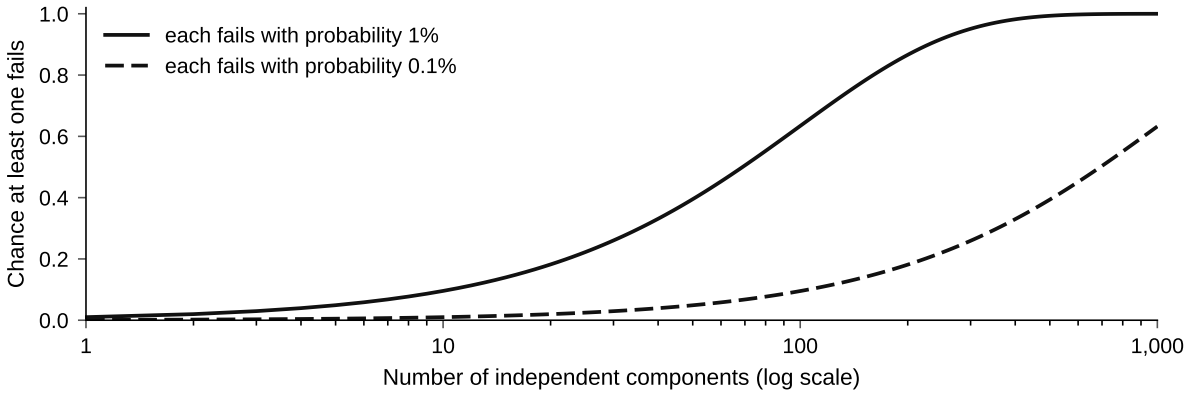


Figure 1.2. The chance that at least one of many independent components fails. Small chances accumulate: with a 1% chance each, 70 components give a better-than-even chance of a failure.

Independence is an assumption, and it fails exactly when it matters most. In February 2021, a winter storm in Texas froze gas wells, pipelines, and power plants at the same time. Failures that a model of independent outages would treat as nearly impossible happened together. A shared cause behind several failures is called a common-mode failure, and problem 1.16 measures how much it matters.

1d. Counting

When outcomes are equally likely, probability becomes counting. If one choice has m options and a second has n, the pair has mn. Arranging n different items in order can be done in $n!$ ways, read n factorial: $n \times (n-1) \times \ldots \times 1$. Choosing k items from n, when order does not matter, can be done in

$$\binom{n}{k} = C(n,k) = \frac{n!}{k!\,(n-k)!}$$

ways, read n choose k. A battery operator choosing 4 charging hours out of 24 has $C(24,4) = 10{,}626$ possible plans. The Powerball lottery draws 5 of 69 white balls and 1 of 26 red ones, so there are $C(69,5) \times 26 = 292{,}201{,}338$ possible tickets, and a ticket's chance of the jackpot is 1 in about 292 million.

Worked example — Shared birthdays

In a group of 23 people, what is the chance that two share a birthday? Count the opposite. Ignoring leap years, the chance that all 23 birthdays differ is $(365 \times 364 \times \ldots \times 343)/365^{23} \approx 0.493$, so the chance of a shared birthday is about 51%. Most people guess far lower, but 23 people make $C(23,2) = 253$ pairs, and each pair is another chance.

The same arithmetic tells engineers how often two different files will receive the same digital fingerprint from a hashing function, and why the fingerprints need vastly more possible values than there are files.

1e. Systems in Series and in Parallel

Reliability engineering applies these rules to whole systems. Components in series must all work, like the links of a chain. Components in parallel back each other up, so the system fails only if every one of them fails:

$$\begin{aligned} \text{series:} \quad & P(\text{works}) = p_1\, p_2 \cdots p_n \\ \text{parallel:} \quad & P(\text{works}) = 1 - (1-p_1)(1-p_2)\cdots(1-p_n) \end{aligned}$$

Here p_i is the chance that component i works, and the failures are assumed independent. Five components in series, each working 99.9% of the time, give $0.999^5 \approx 99.5\%$. Two transformers in parallel, each available 99% of the time, give $1 - 0.01^2 = 99.99\%$.

Grid planners turn these rules into standards. The N−1 criterion requires the high-voltage transmission grid to keep serving customers after the loss of any single element, such as a line, a transformer, or a generator; the local lines that feed homes are not held to it. For supply overall, the most common target in North America is a loss-of-load expectation of one day in ten years: on average, supply should fall short on no more than 0.1 days a year.

In a Spreadsheet

Spreadsheets have a function for each rule. With a probability in B2 and a count in B3:

To compute	Formula
Complement	=1-B2
At least one of n independent events	=1-(1-B2)^B3
Choosing and arranging	=COMBIN(24, 4) for 24 choose 4, =PERMUT(10, 3) for ordered choices, and =FACT(5) for $5!$
A random event with probability B2	=RAND()<B2, which is TRUE with probability B2 each time the sheet recalculates
Series and parallel, with probabilities in B2:B6	=PRODUCT(B2:B6) and =1-PRODUCT(1-B2:B6); the second is an array formula, so in Google Sheets write =ARRAYFORMULA(1-PRODUCT(1-B2:B6)), and in Excel before 2021 confirm it with Ctrl+Shift+Enter

1f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Probability	A number from 0 to 1; with equally likely outcomes, the number of outcomes in the event divided by the total
Complement	$P(A^c) = 1 - P(A)$ A^c is the event that A does not happen
Addition rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $A \cup B$ means A or B or both, and $A \cap B$ means both; for mutually exclusive events the last term is zero
Independence	$P(A \cap B) = P(A)\,P(B)$ Holds when knowing that one event happened does not change the chance of the other
At least one	$1 - (1 - p)^n$ The chance that at least one of n independent events, each with probability p, happens
Counting	$C(n, k) = \dfrac{n!}{k!\,(n-k)!}$ Ways to choose k of n items when order does not matter; $n! = n \times (n-1) \times \ldots \times 1$ is the number of orders of n items
Series and parallel	$p_1 p_2 \cdots p_n$ and $1 - (1 - p_1) \cdots (1 - p_n)$ The chance that a system works when every component must work, and when any one suffices; p_i is the chance that component i works
Common-mode failure	A shared cause, such as a storm, that makes failures happen together and breaks the independence redundancy relies on
The idea	Probabilities combine by a few rules: subtract from 1 for not, add and remove the overlap for or, and multiply for and when events are independent. Many small chances add up fast, and independence is an assumption to check.

1g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

1.1 A card is drawn from a standard deck of 52. What is the chance that it is a heart or a king?

1.2 What is the chance of at least one six in four rolls of a die?

1.3 In how many ways can 3 projects be chosen from 10? In how many ways can first, second, and third place be awarded among 10?

1.4 Independent events have $P(A) = 0.3$ and $P(B) = 0.5$. Find the chances of both, of at least one, and of neither.

1.5 Two backup generators each start with probability 0.95, independently. What is the chance that at least one starts?

1.6 In the survey of section 1b, check directly whether having solar panels and having an EV are independent.

Applied

1.7 A wind farm has 50 turbines, each out of service on a given day with probability 0.02, independently. What is the chance that all are running? That at least one is out?

1.8 A portfolio holds 20 bonds, each with a 1% chance of default this year, independently. What is the chance of at least one default? What would make the true chance different?

1.9 A data pipeline has four stages in series, each working 99.5% of the time. How reliable is it? How reliable is it if every stage is duplicated in parallel?

1.10 A lottery asks players to choose 6 numbers from 49. What is the chance that one ticket matches all six?

1.11 A town is served by three lines, each out of service with probability 0.01, independently. The town keeps power if at least two lines are up. What is the chance it keeps power?

1.12 Two independent spam checks each miss 10% of spam. A message is caught if either check catches it. What share of spam gets through?

Challenge

1.13 Show the addition rule by splitting $A \cup B$ into pieces that do not overlap.

1.14 Show that if A and B are independent, then so are A and B^c.

1.15 Check the birthday calculation: why is the chance that 23 birthdays all differ equal to $(365 \times 364 \times \dots \times 343)/365^{23}$?

1.16 On stormy days, 2% of all days, each of two generators fails with probability 0.2, independently; on other days each fails with probability 0.005. Find the chance that both fail on a random day, and compare it with the answer that treats the two as independent.

Module 1 Quiz

Check Your Understanding

1. Each of 100 independent components fails with probability 1%. Why is the chance that at least one fails far above 1%?

A) Every component is one more chance of failure, and the chances accumulate.

B) Components wear each other out, so failures become more likely as their number grows.

C) The components share a power supply, which makes them fail together.

D) With 100 components a failure is certain every day, since 100 times 1% equals 100%.

2. What does it mean for two events to be independent?

A) They cannot both happen at once, because each excludes the other.

B) Knowing one happened doesn't change the other's chance.

C) One event causes the other, so they always occur together.

D) Their probabilities add up to exactly 1, so one or the other must happen.

3. In a survey, 40% of customers have solar, 25% have an EV, and 15% have both. What share have at least one?

A) 65%

B) About 55%

C) 50%, adding the two and subtracting the overlap

D) 40%, the larger of the two shares, since owners of both are already counted in it

4. Why do grid planners worry about common-mode failures?

A) They are the most frequent kind of equipment fault.

B) They affect only the oldest equipment on a system.

C) They make components fail at random, on separate days.

D) One cause, such as a storm or a freeze, can knock out several components at once, so failures stop being independent and the chance of losing them together soars.

5. In how many ways can 3 projects be chosen from 8, when order does not matter?

A) 336

B) 8 cubed, or 512

C) 56

D) 40,320, the number of ways to arrange all eight projects in order from first to last

Module 2 — Conditional Probability and Bayes' Rule

How evidence changes a probability

Recommended Reading / Reference

Book: Sharon Bertsch McGrayne, *The Theory That Would Not Die* (2011) — the history of Bayes' rule, from code-breaking to insurance to the search for lost submarines

Reference: Paul Graham, “A Plan for Spam” (2002) — the free essay that popularized Bayesian spam filters, the subject of section 2d

2a. Conditional Probability

New information changes probabilities. The probability of A given that B happened, written $P(A|B)$, restricts attention to the cases where B happened and asks how often A happened among them:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Worked example — Late payments and default

A lender's records cover 1,000 loans. 100 had a late payment in their first year, and 20 of those later defaulted. Of the 900 without a late payment, 18 defaulted.

	Defaulted	Did not default	Total
Late payment	20	80	100
No late payment	18	882	900
Total	38	962	1,000

The overall default rate is $38/1{,}000 = 3.8\%$. Given a late payment it is $20/100 = 20\%$, and given none, $18/900 = 2\%$. Turned around, the conditional probability is a different number: of the loans that defaulted, $20/38 \approx 53\%$ had a late payment.

Rearranging the definition gives the general multiplication rule, $P(A \cap B) = P(B)P(A|B)$, which holds whether or not the events are independent. Independence is the special case $P(A|B) = P(A)$.

2b. Bayes' Rule

The probability usually known is the chance of the evidence given the cause. The one wanted is the chance of the cause given the evidence. Bayes' rule, named for the 18th-century English minister Thomas Bayes, turns one into the other:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

The bottom of the fraction, the overall chance of the evidence, comes from adding up its chance under each possibility, the law of total probability:

$$P(B) = P(B \mid A)\,P(A) + P(B \mid A^c)\,P(A^c)$$

$P(A)$ is called the prior, the probability before the evidence, and $P(A|B)$ the posterior, the probability after it.

Worked example — A medical test

A disease affects 1% of people. A test detects it 90% of the time, its sensitivity, and wrongly flags 5% of healthy people, its false-positive rate. Someone tests positive. The chance of a positive is $0.90 \times 0.01 + 0.05 \times 0.99 = 0.0585$, so the chance of disease given a positive is $0.009/0.0585 \approx 15\%$.

Counting people makes the result plain. Of 10,000 people tested, 100 have the disease, and 90 of them test positive. Of the 9,900 who are healthy, 5% test positive, 495 people. So 585 people test positive, and only 90 of them, about 15%, are sick.

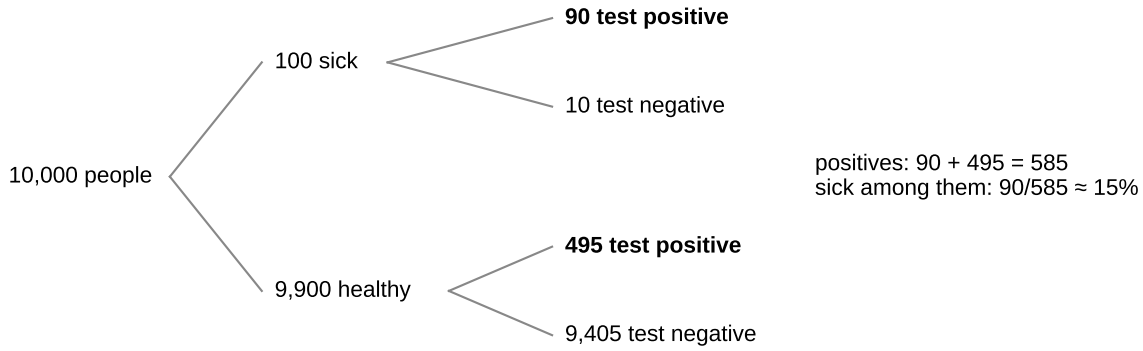


Figure 2.1. The medical test in counts. Because healthy people vastly outnumber sick ones, the healthy people wrongly flagged, 495, outnumber the sick people correctly flagged, 90.

Ignoring how rare the condition is, its base rate, is the most common mistake in reasoning with probabilities. A rare condition and an imperfect test produce mostly false alarms, however accurate the test sounds. Figure 2.2 shows how the answer depends on the base rate.

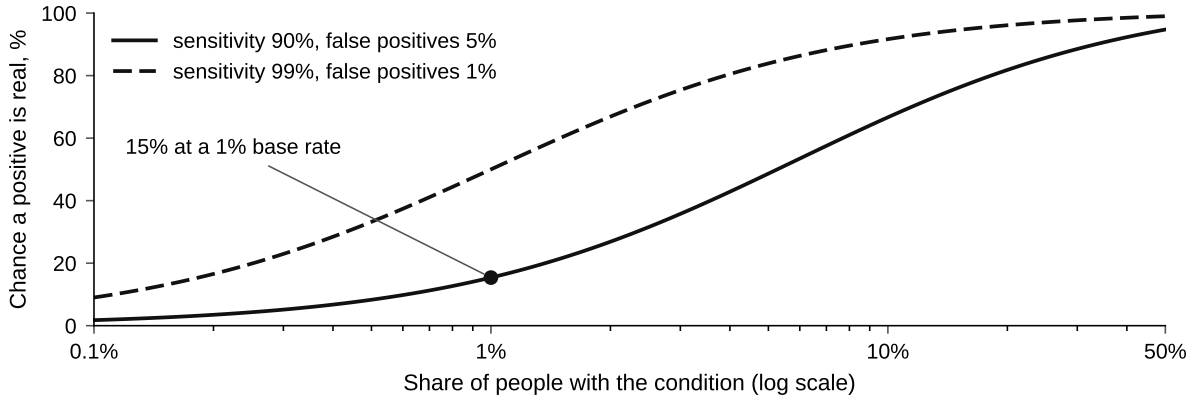


Figure 2.2. The chance that a positive test is real depends heavily on how common the condition is. The better test gains most, 35 to 40 points, when 1% to 5% of people have the condition; when the condition is very rare, even its positives are mostly false.

2c. Odds and Likelihood Ratios

Bayes' rule is simplest in odds. The odds of an event are its probability divided by the probability that it does not happen, $p/(1-p)$, so a probability of 20% is odds of 1 to 4. The strength of a piece of evidence is its likelihood ratio: how much more likely the evidence is if the event is true than if it is false. Then

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}$$

For the test, a positive has a likelihood ratio of $0.90/0.05 = 18$. The prior odds are 1 to 99, so the posterior odds are 18 to 99, a probability of $18/117 \approx 15\%$, as before. A second, independent positive multiplies the odds by 18 again, to 324 to 99, a probability of about 77%.

Worked example — Fraud alerts

One card transaction in 1,000 is fraudulent. A detector flags 95% of frauds and 1% of legitimate transactions. A flag's likelihood ratio is $0.95/0.01 = 95$, and the prior odds are 1 to 999, so the posterior odds are 95 to 999: only about 8.7% of flags are fraud.

The other 91% of flags fall on honest customers, which is one reason a flag usually prompts a quick check with the customer rather than a closed account.

2d. How a Spam Filter Learns

In 2002 the programmer Paul Graham showed that a filter built on Bayes' rule could catch nearly all spam by learning from words. The filter counts, in mail already sorted, how often each word appears in spam and in legitimate mail. Each word in a new message then contributes its likelihood ratio.

Treating the words as independent given the kind of message, which they are not, gives the naive Bayes filter: the posterior odds are the prior odds times the likelihood ratio of every word in the message. The simplification is crude, but it works well in practice.

Worked example — Two words

Half of incoming mail is spam. The word “winner” appears in 20% of spam and 0.5% of legitimate mail, a likelihood ratio of 40. The word “meeting” appears in 1% of spam and 10% of legitimate mail, a ratio of 0.1. A message containing both has odds of $1 \times 40 \times 0.1 = 4$ to 1 of being spam, an 80% chance.

Multiplying many small numbers loses precision, so filters add logarithms instead. The log of the posterior odds is the log of the prior odds plus each word's log likelihood ratio, a weighted sum of which words appear. That is the form of logistic regression, the classifier of Module 10.

2e. Weighing a Claim of Skill

Bayes' rule also weighs a track record. Suppose 10% of trading strategies are skilled and win 55% of their trades, while the rest win 50%, by luck. A strategy wins 60 of 100 trades. The likelihood ratio of that record is $(0.55^{60} \times 0.45^{40})/0.5^{100} \approx 4.5$, and the prior odds are 1 to 9, so the posterior odds are $4.5/9 = 0.5$. The chance that the strategy is skilled is only about a third.

A record that looks impressive is weak evidence when lucky records are common and skill is rare. The same reasoning applies to fund managers, to a new drug's first small trial, and to a model that beats a benchmark once.

Reversing a conditional probability while ignoring the base rate has a name in courtrooms, the prosecutor's fallacy. The chance of evidence like this if the accused is innocent is not the chance that the accused is innocent given the evidence. Problem 2.12 shows the same error in a face-recognition system.

In a Spreadsheet

With the prior probability in B2, the sensitivity in B3, and the false-positive rate in B4:

To compute	Formula
Posterior by Bayes' rule	=B3*B2/(B3*B2+B4*(1-B2))
Odds from a probability, and back	=B2/(1-B2), and =B5/(1+B5) for odds in B5
Likelihood ratio of a positive	=B3/B4
Naive Bayes, with each word's log likelihood ratio in C2:C20	=EXP(LN(B2/(1-B2))+SUM(C2:C20)) for the posterior odds
A two-way table from raw records	=COUNTIFS(A:A, "late", B:B, "default") for each cell

2f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Conditional probability	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ The chance of A among the cases where B happened; $P(A\|B)$ is usually not $P(B\|A)$
Total probability	$P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$ The overall chance of B, adding its chance when A happens and when it does not
Bayes' rule	$P(A \mid B) = \dfrac{P(B \mid A)\,P(A)}{P(B)}$ Turns the chance of the evidence given the cause into the chance of the cause given the evidence; $P(A)$ is the prior and $P(A\|B)$ the posterior
Odds	$\text{odds} = \dfrac{p}{1-p}, \qquad p = \dfrac{\text{odds}}{1+\text{odds}}$ A probability p of 20% is odds of 1 to 4
Bayes in odds	$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}$ The likelihood ratio of evidence B is $P(B\|A)/P(B\|A^c)$; for pieces of evidence that are independent once it is known whether A happened, the ratios multiply
Base rates	A rare condition and an imperfect test give mostly false alarms: at 1% prevalence, 90% sensitivity, and 5% false positives, only about 15% of positives are real
Naive Bayes	The posterior odds are the prior odds times each feature's likelihood ratio, treating the features as independent; in logs, a weighted sum
The idea	Evidence updates a probability by its likelihood ratio, but the starting point matters: when the prior is small, even strong evidence can leave the posterior modest.

2g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

2.1 From the loan table of section 2a, find the chance that a loan that defaulted had a late payment.

2.2 Events have $P(A) = 0.3$, $P(B) = 0.4$, and $P(A \cap B) = 0.12$. Find $P(A|B)$ and decide whether A and B are independent.

2.3 Convert a probability of 0.2 to odds, and odds of 3 to 1 to a probability.

2.4 A test has 99% sensitivity and a 1% false-positive rate, for a condition that 0.1% of people have. What is the chance that a positive is real?

2.5 Repeat problem 2.4 with odds and the likelihood ratio.

2.6 Rain falls on 30% of summer days. Demand peaks above a utility's alert level on 10% of rainy days and 40% of dry ones. What share of days reach the alert level?

Applied

2.7 In problem 2.6, given that demand reached the alert level, what is the chance that it rained?

2.8 2% of a utility's transformers are at risk of failure. A monitor flags 80% of those and 10% of healthy ones. What is the chance that a flagged transformer is at risk? What if a second, independent monitor also flags it?

2.9 40% of a company's mail is spam. “Invoice” appears in 5% of spam and 2% of legitimate mail; “unsubscribe” in 30% of spam and 6% of legitimate mail. What is the chance that a message with both words is spam, under naive Bayes?

2.10 A forecaster puts the chance of a recession next year at 20%. Then a warning sign appears that, in illustrative history, preceded 80% of recessions and appeared in 15% of other years. What is the updated chance?

2.11 A coin is either fair or biased toward heads, landing heads 70% of the time, with prior chances of 50% each. It lands heads 8 times in 10 tosses. What is the chance that it is biased?

2.12 A face-recognition system wrongly matches 1 face in 100,000. It scans 2 million faces at a stadium for one person who is there, and it will flag that person. About how many flags will there be, and what is the chance that a given flag is the right person?

Challenge

2.13 Derive Bayes' rule from the definition of conditional probability.

2.14 Show that Bayes' rule in odds form follows from writing Bayes' rule for A and for A^c given the same evidence B.

2.15 For a rare condition with prevalence p, show that the chance a positive is real is about p times the likelihood ratio, when that product is small.

2.16 Show that in logs, a naive Bayes filter adds a fixed weight for each word present, and explain why that makes it a linear model.

Module 2 Quiz

Check Your Understanding

1. A test catches 90% of cases of a disease and gives false alarms 5% of the time. 1% of people have it. Roughly what share of positives are real?

A) About 15%
B) 90%
C) About 50%, since the test is right far more often than it is wrong
D) 95%, which is one minus the false-alarm rate

2. What does Bayes' rule do?

A) Adds probabilities
B) Updates a probability in light of new evidence
C) Counts outcomes
D) It guarantees that a probability, once estimated from enough data, never needs to be revised again when new information arrives

3. Prior odds are 1 to 99, and a positive test has a likelihood ratio of 18. What are the posterior odds?

A) 18 to 1
B) 1 to 1,782
C) 99 to 18
D) 18 to 99, the prior odds times the likelihood ratio, a probability of about 15%

4. Why do most fraud alerts turn out to be false alarms?

A) The detectors are built to flag every transaction that looks unusual in any way at all.
B) Fraudsters deliberately trigger alerts on honest accounts to hide their own activity.
C) Fraud is rare, so false alarms outnumber real cases.
D) Banks set their systems to raise alarms often, so that customers look at their statements more.

5. A naive Bayes spam filter treats words as independent given the kind of message. What does it multiply together?

A) Every word's count
B) The probability of each word appearing in any email, ignoring whether the email is spam or not
C) The lengths of the emails
D) The prior odds and each word's likelihood ratio

Module 3 — Random Variables and Expected Value

Valuing uncertain outcomes, and the limits of the average

Recommended Reading / Reference

Book: Peter L. Bernstein, *Against the Gods: The Remarkable Story of Risk* (1996) — how expected value, insurance, and the measurement of risk were invented

Reference: Social Security Administration, Actuarial Life Table (ssa.gov) — the free table of the chance of dying within a year at each age, the raw material of life insurance pricing

3a. Random Variables and Their Distributions

A random variable is a number set by chance: tomorrow's power price, the number of defaults in a loan pool, the payout of an insurance policy. A capital letter such as X stands for the variable, and lowercase letters such as x for the values it can take. Its distribution lists the possible values and their probabilities, which add to 1.

Worked example — A battery's evening

A battery operator's profit tomorrow depends on the evening price spike. With no spike, probability 0.6, it earns $200. With a moderate spike, probability 0.3, it earns $800. With an extreme spike, probability 0.1, it earns $3,000.

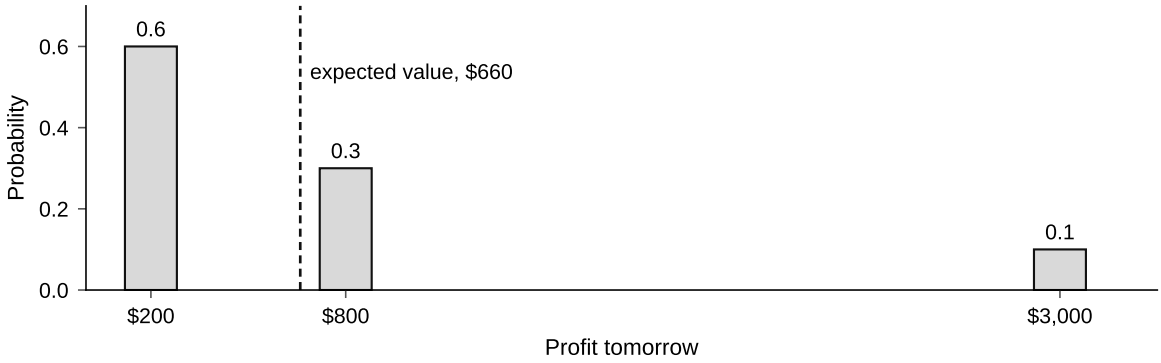


Figure 3.1. The distribution of the battery's profit. The expected value sits well above the most likely outcome, pulled up by the rare extreme spike.

The cumulative distribution gives the chance that the variable is at most a given value: here, the chance of earning at most $800 is $0.6 + 0.3 = 0.9$.

3b. Expected Value

The expected value of a random variable, written $E[X]$ and also called its mean, μ (mu), is the probability-weighted average of its values:

$$E[X] = \sum_i x_i \, p_i$$

Here x_i are the values, p_i their probabilities, and Σ_i, a capital sigma, means the sum over all the values. The expected value is the long-run average: over many evenings like tomorrow, the profit would average this much. For the battery, $E[X] = 0.6 \times 200 + 0.3 \times 800 + 0.1 \times 3{,}000 = 660$ dollars. The extreme spike, one evening in ten, supplies $300/660 \approx 45\%$ of the expected profit, which is why storage owners watch the tails of the price distribution as closely as its middle.

Expected values obey two rules that make them easy to work with. Scaling and shifting pass straight through: $E[aX + b] = aE[X] + b$ for any numbers a and b. And the expected value of a sum is the sum of the expected values, $E[X + Y] = E[X] + E[Y]$, whether or not the variables are independent.

Worked example — Expected defaults

A pool holds 20 bonds, each with a 1% chance of default. Count each bond as 1 if it defaults and 0 if not, a variable whose expected value is 0.01. The number of defaults is the sum of the 20 counts, so its expected value is $20 \times 0.01 = 0.2$, however the defaults are related.

3c. Variance and Standard Deviation

Two investments with the same expected value can differ enormously in risk. The variance measures the spread around the mean, and the standard deviation, its square root, puts it back in the original units:

$$\operatorname{Var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2, \qquad \sigma = \sqrt{\operatorname{Var}(X)}$$

Here σ (sigma) is the standard deviation. For the battery, the squared deviations from 660 are $(-460)^2$, 140^2, and $2{,}340^2$. Weighting them by their probabilities gives a variance of 680,400, so $\sigma \approx 825$ dollars, more than the expected profit itself.

Scaling a variable by a scales its variance by a^2. For a sum, the variances add, plus a term for how the two variables move together, their covariance:

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y)$$

The covariance is the average product of the two variables' deviations from their means. It is zero for independent variables, whose variances then simply add. Two independent batteries like the one above earn $1,320 in expectation, with a standard deviation of $\sqrt{2} \times 825 \approx 1{,}167$ dollars: twice the expected profit, but only about 1.4 times the risk.

3d. Pricing Insurance

Insurance is the oldest business built on expected values. A one-year life insurance policy of $100,000, for someone with a 0.4% chance of dying during the year, has an expected payout of $0.004 \times 100{,}000 = 400$ dollars, its actuarially fair price. The insurer charges more, say $520, to cover its costs, the capital it must hold, and a profit. Actuaries take the probabilities from life tables, which list the chance of dying within a year at each age.

Why buy a policy that costs more than its expected payout? Because the loss it covers would be ruinous, and each dollar matters more when money is scarce. Economists describe this with a utility function, which turns wealth into satisfaction and rises ever more slowly as wealth grows. A buyer who weighs outcomes by utility rather than dollars will pay something above the expected loss to remove a large risk.

The insurer can afford to take the risk because it pools many independent policies, whose total is far more predictable than any one of them, as Module 6 shows.

3e. The Value of Information

Expected values also price information. A data center must decide today whether to lock in tomorrow's power at $60 per megawatt-hour (MWh) or to buy at tomorrow's spot price. The spot price will be $40 with probability 0.7 or $120 with probability 0.3, an expected $64. Without more information, locking in is cheaper on average: $60 against an expected $64.

Tomorrow's spot price	Probability	Best choice with a perfect forecast	Cost per MWh
$40	0.7	Buy at spot	$40
$120	0.3	Lock in	$60
Expected cost			$46

With a perfect forecast, the buyer would buy at spot when the price will be $40 and lock in when it will be $120, for an expected cost of $0.7 \times 40 + 0.3 \times 60 = 46$ dollars. The difference, $60 - 46 = 14$ dollars per MWh, is the expected value of perfect information: the most the buyer should pay for a perfect forecast. A real forecast, being imperfect, is worth less.

3f. When the Average Misleads

Expected value is the right guide for decisions repeated many times on a small scale. It can mislead when one outcome can wipe out the decision-maker, or when gains and losses compound.

Worked example — Average versus typical

An investment gains 50% or loses 40% each year, with equal chances. Its expected return is $0.5 \times 50 + 0.5 \times (-40) = 5$ percent a year, so expected wealth grows. But a typical sequence has about as many gains as losses, and a gain followed by a loss multiplies wealth by $1.5 \times 0.6 = 0.9$. After ten years, expected wealth is $1.05^{10} \approx 1.63$ times the start, while the typical investor, the median one with half doing better and half worse, has $0.9^{5} \approx 0.59$.

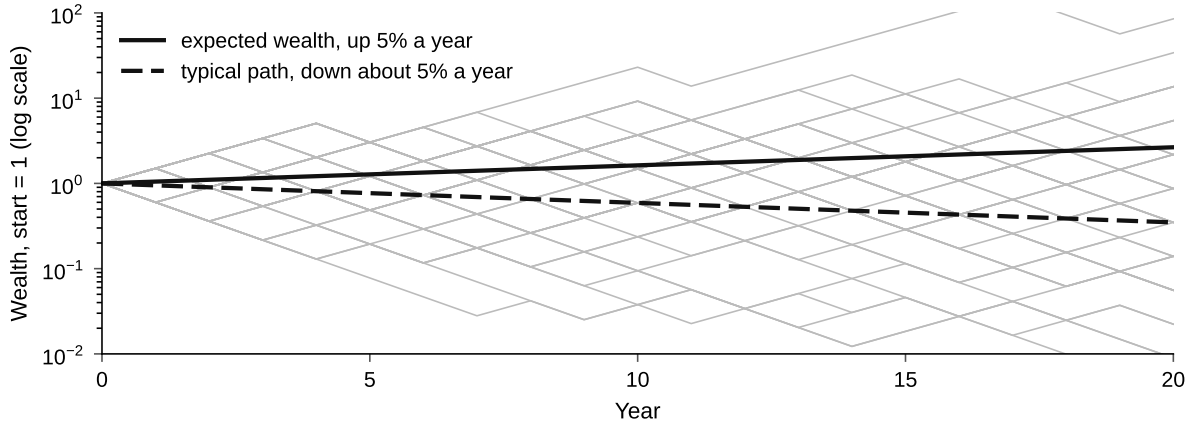


Figure 3.2. Forty simulated paths of the investment that gains 50% or loses 40% each year, in gray. Expected wealth rises because a few paths soar; the typical path shrinks.

The average is pulled up by a few extremely lucky paths. What matters for a single investor over time is the growth rate of the typical path, the probability-weighted average of the natural logarithm, ln, of each year's growth factor: $0.5 \ln 1.5 + 0.5 \ln 0.6 \approx -0.053$, a loss of about 5% a year.

Betting a fixed fraction of wealth each time turns this into a rule. For a bet at even odds won with probability p, the fraction of wealth that maximizes long-run growth is $2p - 1$, the Kelly criterion, published by John Kelly in 1956. With a 55% chance of winning, it says to bet 10% of wealth each time. Betting about twice that fraction, or more, makes the typical path shrink, even though every bet has a positive expected value.

In a Spreadsheet

With the values in A2:A4 and their probabilities in B2:B4:

To compute	Formula
Expected value	=SUMPRODUCT(A2:A4, B2:B4)
Variance, with the expected value in E1	=SUMPRODUCT((A2:A4-E1)^2, B2:B4), and its square root for the standard deviation
A random draw from the distribution	=LOOKUP(RAND(), {0, 0.6, 0.9}, {200, 800, 3000}), with the cumulative probabilities as thresholds
Growth rate of the typical path	=0.5*LN(1.5)+0.5*LN(0.6)
Kelly fraction for an even-odds bet won with probability B6	=2*B6-1

3g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Random variable	A number set by chance; its distribution lists the possible values x_i and their probabilities p_i, which add to 1
Expected value	$E[X] = \mu = \sum_i x_i\, p_i$ The probability-weighted average, and the long-run average over many repetitions; μ (mu) is the mean
Linearity	$E[aX + b] = aE[X] + b, \qquad E[X + Y] = E[X] + E[Y]$ a and b are constants; the second rule holds even when X and Y are dependent
Variance	$\mathrm{Var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2$ The standard deviation σ is its square root, in the units of X; scaling X by a scales the variance by a^2
Variance of a sum	$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$ The covariance, the average product of the deviations from the means, is zero for independent variables, whose variances simply add
Value of information	The expected outcome when deciding after the uncertainty is resolved, minus the expected outcome when deciding before it: the most a perfect forecast is worth
Typical growth	$g = \sum_i p_i \ln(1 + r_i)$ The long-run growth rate of a single compounding path, where r_i are the possible returns, p_i their probabilities, and ln the natural logarithm; it is below the expected return and can be negative when that is positive
Kelly criterion	For a bet at even odds won with probability p, betting the fraction $2p - 1$ of wealth each time maximizes long-run growth
The idea	The expected value is the long-run average and adds up simply, but risk lives in the spread and the tails. For a single compounding path, the typical outcome can fall short of the average, and even shrink.

3h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

3.1 Find the expected value, variance, and standard deviation of one roll of a fair die.

3.2 A bet pays \$10 with probability 0.3 and loses \$5 with probability 0.7. What is its expected value?

3.3 A random variable has mean 4 and variance 9. Find the mean and variance of $2X + 3$.

3.4 A variable takes the values 0, 1, and 2 with probabilities 0.5, 0.3, and 0.2. Find its mean and variance.

3.5 50 loans each default with probability 2%. What is the expected number of defaults?

3.6 Independent variables X and Y have standard deviations 3 and 4. What is the standard deviation of $X + Y$?

Applied

3.7 A $300,000 house has a 0.2% chance a year of being destroyed by fire. What is the actuarially fair premium for full cover, and what would an insurer charge with a 30% loading for costs and profit?

3.8 A generator can sell its output in the day-ahead market at $55 per MWh, or wait for the real-time market, where prices will be $40, $60, or $150 with probabilities 0.5, 0.4, and 0.1. Compare the expected values and the risk.

3.9 A well costs $10 million to drill and has a 20% chance of finding oil worth $60 million. A seismic survey would reveal in advance whether the oil is there. What is the expected value of drilling now, and the most the survey is worth?

3.10 An asset returns +30% or −20% each year, with equal chances. Find its expected return and the growth rate of its typical path.

3.11 A bet at even odds is won with probability 0.55. What fraction of wealth does the Kelly criterion say to bet, and what is the long-run growth per bet?

3.12 Two solar farms' daily outputs have standard deviations of 10 and 20 MWh and a correlation of −0.5. What is the standard deviation of their total?

Challenge

3.13 Show that $E[X + Y] = E[X] + E[Y]$ for discrete random variables, without assuming independence.

3.14 Show that $E[(X - \mu)^2] = E[X^2] - \mu^2$.

3.15 In the St. Petersburg game, a coin is tossed until the first head; if that takes k tosses, the game pays 2^k dollars. Show that the expected payout is infinite. What is the game worth to someone who values wealth by its logarithm?

3.16 Derive the Kelly fraction by choosing the fraction f that maximizes $g(f) = p\ln(1 + f) + (1 - p)\ln(1 - f)$.

Module 3 Quiz

Check Your Understanding

1. Why might a person sensibly pay more for insurance than the expected value of the loss?

A) Over many years, a policyholder expects to get back more than the premiums paid.
B) A large loss would hurt far more than the premium does, because each dollar matters more when money is scarce.
C) Expected values apply only to games of chance, not to houses.
D) The premium is tax deductible, which makes it cheaper than it looks.

2. A bet pays $10 with probability 0.3 and loses $5 with probability 0.7. What is its expected value?

A) −$0.50
B) $5
C) $2.50, the average of the win and the loss, weighted equally
D) +$1.50 a bet

3. Which rule holds for any two random variables, dependent or not?

A) Expected values add
B) Variances add
C) Means multiply
D) Spreads add

4. What is the value of perfect information?

A) The forecast's accuracy rate
B) Gain from knowing first
C) The expected value of the outcome
D) The price charged for a forecast

5. An investment gains 50% or loses 40% each year, with equal chances. What happens to a typical investor over many years?

A) Grows 5% a year
B) Doubles every 14 years, since the expected return is 5% a year
C) Stays even
D) Shrinks, about 5% a year

Module 4 — Counting Events: Binomial and Poisson

How many times something happens, and what independence hides

Recommended Reading / Reference

Book: Michael Lewis, *The Big Short* (2010) — the story of the mortgage bonds whose defaults turned out to be anything but independent

Reference: U.S. Energy Information Administration (EIA), Annual Electric Power Industry Report, Form EIA-861 — free reliability data, including how often the average customer loses power

4a. Repeated Trials: The Binomial Distribution

Many questions count successes in repeated trials: how many of 100 loans default, how many of 10,000 visitors buy, how many of 1,000 test questions a model gets right. When there are n independent trials, each a success with the same probability p, the number of successes X has the binomial distribution:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n$$

Each particular sequence with k successes and $n - k$ failures has probability $p^k(1-p)^{n-k}$, and there are $C(n, k)$, read n choose k, such sequences. The mean and variance are

$$E[X] = np, \qquad \mathrm{Var}(X) = np(1-p)$$

Worked example — A pool of loans

A pool holds 100 loans, each defaulting this year with probability 2%, independently. The expected number of defaults is 2, with a standard deviation of $\sqrt{(100 \times 0.02 \times 0.98)} = 1.4$. The chance of no defaults is $0.98^{100} \approx 0.133$, and the chance of exactly two is $C(100, 2) \times 0.02^2 \times 0.98^{98} \approx 0.273$.

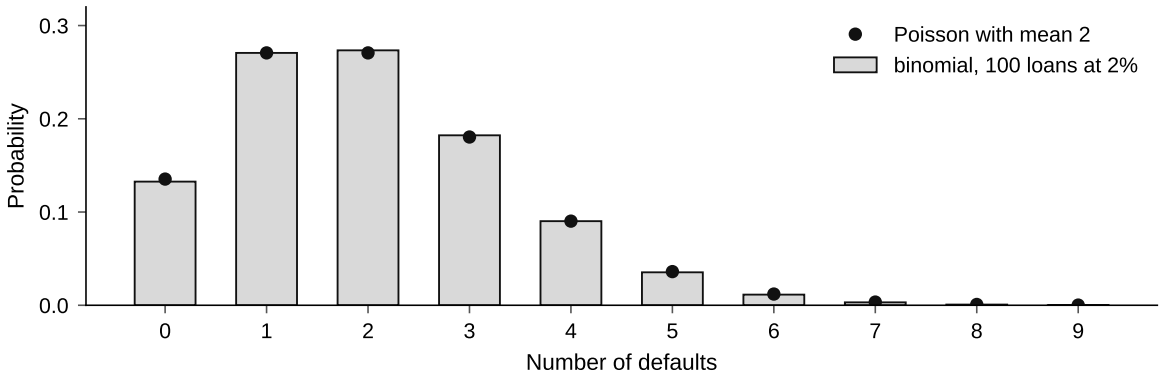


Figure 4.1. The number of defaults among 100 independent loans, each with a 2% chance. The Poisson distribution with the same mean, the dots, is almost identical.

The same arithmetic shows how much a measured success rate wobbles. A model with a true accuracy of 80%, tested on 1,000 questions, will score about 800 correct, with a standard deviation of $\sqrt{(1{,}000 \times 0.8 \times 0.2)} \approx 12.6$. Its measured accuracy will typically land within about 1.3 percentage points of 80%, so two models whose scores differ by a point or two may not differ at all.

4b. Rare Events: The Poisson Distribution

Some events happen at random moments with no fixed number of trials: power outages, insurance claims, server crashes, customers arriving. If they occur independently at a steady average rate, the number in a period has the Poisson distribution, named for the French mathematician Siméon Denis Poisson:

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}, \qquad k = 0, 1, 2, \ldots$$

Here λ (lambda) is the average number of events in the period, and $e \approx 2.718$. The mean and the variance both equal λ, so the standard deviation is $\sqrt{\lambda}$.

Worked example — Outages

A utility's customers average 1.3 interruptions a year, an illustrative figure. For one customer, the chance of a year with none is $e^{-1.3} \approx 0.27$. The chance of three or more is $1 - e^{-1.3}(1 + 1.3 + 1.3^2/2) \approx 0.14$: one customer in seven has a bad year even when nothing about the grid has changed.

The Poisson is also the limit of the binomial when trials are many and each success is rare, with $\lambda = np$. For the loan pool, it gives the chance of no defaults as $e^{-2} \approx 0.135$, close to the binomial's 0.133. A famous early use, by Ladislaus Bortkiewicz in 1898, counted the soldiers in Prussian army corps killed by horse kicks each year and found that the Poisson fitted them well.

4c. Claims, Reserves, and Capacity

Because the Poisson's standard deviation is $\sqrt{\lambda}$, larger books of business are steadier in proportion: a count averaging 100 varies by about 10%, one averaging 10,000 by about 1%. Insurers set their reserves, and grid and data center operators their spare capacity, from the upper tail of such counts.

Worked example — An insurer's year

An insurer expects 50 large claims a year. The standard deviation is $\sqrt{50} \approx 7.1$. The Poisson distribution puts the chance of more than 65 claims at about 1.7%, so reserves sized for 65 claims would fall short about one year in 60.

4d. When Events Are Not Independent

The binomial and the Poisson both assume that events are independent. Defaults are not: they cluster in recessions, when one shared cause, a weak economy, raises every borrower's risk at once. A simple way to model this is a mixture, in which the default probability itself depends on the state of the economy.

Worked example — A good year and a bad year

Suppose that in a normal year, 80% of years, each loan defaults with probability 1%, and in a recession, 20% of years, with probability 6%. The average default rate is $0.8 \times 1 + 0.2 \times 6 = 2$ percent, the same as before. But the chance of 6 or more defaults in the pool of 100 is now about 11%, against 1.5% if every loan defaulted independently at 2%.

The difference grows further out. Eleven or more defaults have a chance of about 1 in 180,000 under independence, but about 1 in 130 under the mixture, a risk more than a thousand times larger.

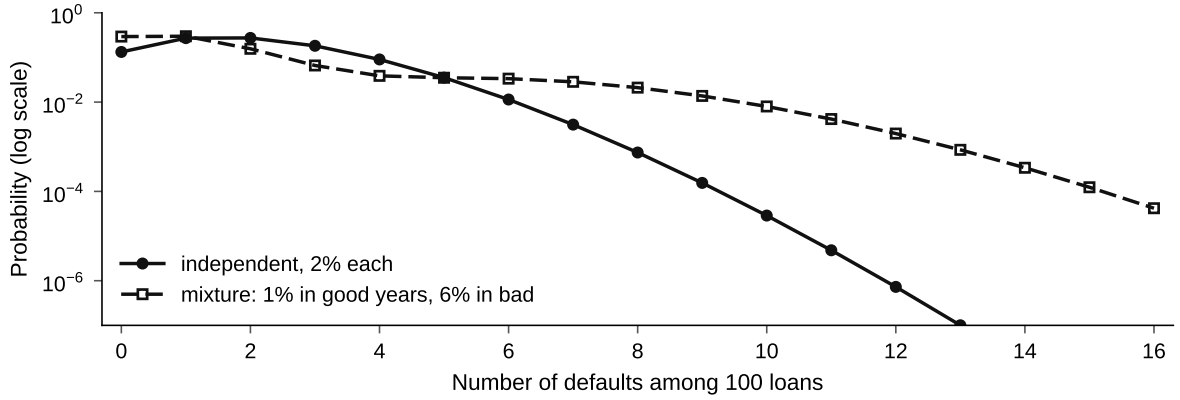


Figure 4.2. Two models with the same 2% average default rate. When defaults share a cause, the chance of many defaults at once is larger by orders of magnitude.

That is the arithmetic of the mortgage bonds that failed in 2007 and 2008. Bonds built from pools of mortgages were divided into slices, and the senior slices, which lost money only if defaults across the pool ran very high, were rated as safe as government debt. The ratings leaned on models in which defaults were only weakly linked. When house prices fell across the country at once, defaults rose together, and slices that were supposed to be nearly riskless lost heavily.

4e. Experiments on Websites

Online businesses test a change by showing two versions of a page to randomly chosen visitors, called an A/B test, and comparing how many in each group buy or sign up. Each visitor is a trial, so the number of conversions is binomial.

Worked example — Is 5.3% better than 5.0%?

With 10,000 visitors and a true conversion rate of 5%, conversions average 500, with a standard deviation of $\sqrt{(10{,}000 \times 0.05 \times 0.95)} \approx 21.8$. The measured rate varies by about 0.22 percentage points, and the difference between two such groups by about 0.31 points.

A measured difference of 0.3 points, 5.3% against 5.0%, is therefore about one standard deviation, the size of gap that chance produces often. Module 8 turns this comparison into a formal test.

In a Spreadsheet

Spreadsheets have both distributions built in, with n trials, success probability p, and Poisson mean λ:

To compute	Formula
Binomial chance of exactly k	=BINOM.DIST(k, n, p, FALSE); TRUE instead gives the chance of k or fewer
Poisson chance of exactly k	=POISSON.DIST(k, λ, FALSE); TRUE instead gives the chance of k or fewer
The count a binomial stays at or below 99% of the time	=BINOM.INV(n, p, 0.99)
A random binomial count, for simulation	=BINOM.INV(n, p, RAND())
Chance of more than 65 Poisson events with mean 50	=1-POISSON.DIST(65, 50, TRUE)

4f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Binomial distribution	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ The chance of k successes in n independent trials, each a success with probability p
Binomial mean and variance	$E[X] = np, \qquad \mathrm{Var}(X) = np(1-p)$ A success rate measured from n trials varies by about $\sqrt{(p(1-p)/n)}$
Poisson distribution	$P(X = k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ Counts of independent events at an average rate of λ per period; the mean and the variance are both λ, and $e \approx 2.718$
Poisson approximation	For many trials, each with a small chance of success, the binomial is close to the Poisson with $\lambda = np$
Relative steadiness	A Poisson count's standard deviation is $\sqrt{\lambda}$, so larger books vary less in proportion: about 10% at $\lambda = 100$, 1% at $\lambda = 10{,}000$
Dependence	When events share a cause, the average can stay the same while the chance of many events at once grows enormously
The idea	Counts of independent events follow the binomial or the Poisson, and they grow steadier in proportion as they grow larger. Dependence, not the average rate, decides how bad the worst years are.

4g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

4.1 A fair coin is tossed 5 times. What is the chance of exactly 3 heads?

4.2 For 10 trials with success probability 0.1, find the chances of no successes and of exactly one.

4.3 Find the mean and standard deviation of a binomial count with $n = 400$ and $p = 0.25$.

4.4 For a Poisson count with $\lambda = 2$, find the chances of 0, of exactly 2, and of at most 2.

4.5 What are the mean and standard deviation of a Poisson count with $\lambda = 9$?

4.6 A server crashes 3 times a month on average, at random. What is the chance of a month with no crash?

Applied

4.7 A wind farm has 40 turbines, each out of service on 3% of days, independently. Find the expected number out, and the chance that more than 3 are out on a given day.

4.8 Two models, each tested on its own random sample of 1,000 questions, score 812 and 798. Is the gap larger than chance would produce?

4.9 An insurer expects 50 large claims a year, a Poisson count. What is the chance of more than 65, and what does the normal shortcut, $(65.5 - 50)/\sqrt{50}$, suggest?

4.10 500 transactions each have a 0.2% chance of being fraudulent. Compare the binomial and Poisson chances of at least one fraud.

4.11 Each loan in a pool of 100 defaults with probability 1% in good years, 75% of years, and 5% in bad years, the rest. What is the average default rate, and how does the chance of 6 or more defaults compare with independent defaults at that rate?

4.12 An A/B test shows each version to 20,000 visitors at a conversion rate near 4%. How much does each measured rate vary by chance?

Challenge

4.13 Show that the binomial probabilities add to 1.

4.14 Show that a binomial count has mean np, using linearity.

4.15 Show that a binomial count has variance $np(1 - p)$.

4.16 Show that the binomial chance of k successes approaches the Poisson as n grows with $np = \lambda$ fixed.

Module 4 Quiz

Check Your Understanding

1. 100 loans each default with probability 2%, independently. What is the expected number of defaults?

A) 20
B) 0.2 defaults
C) 2 defaults
D) About 1.4, the standard deviation of the count

2. What does the Poisson distribution model?

A) The heights of people in a large population
B) Counts of rare events
C) Waiting times between events
D) Whether a single event happens or not, with no way of counting repeats within a period

3. Why were losses on mortgage bonds in 2008 so much larger than models of independent defaults predicted?

A) Defaults rose together when house prices fell.
B) Borrowers defaulted at random, one at a time.
C) The bonds were too small to spread the risk across enough loans.
D) Borrowers repaid their loans early, so the bonds earned less interest than expected.

4. For a Poisson count with mean 9, what is the standard deviation?

A) 9
B) 81
C) 4.5
D) 3, the square root

5. Two models score 812 and 798 on separate random samples of 1,000 test questions. What is the best reading?

A) A clear win
B) Too big for chance
C) Within ordinary noise
D) Proof the first is better

Module 5 — Continuous Distributions and Fat Tails

Quantities that vary smoothly, and how extreme they get

Recommended Reading / Reference

Book: Benoit Mandelbrot and Richard L. Hudson, *The (Mis)Behavior of Markets* (2004) — the case that market prices have far fatter tails than the bell curve allows

Reference: Federal Reserve Bank of St. Louis, FRED (fred.stlouisfed.org) — free daily data on stock indexes, interest rates, and energy prices, for measuring real tails

5a. Densities and Areas

Prices, temperatures, and waiting times can take any value in a range, not just whole numbers. For such a continuous random variable X, the chance of any one exact value is zero, and probabilities belong to intervals instead. They come from a density, a curve $f(x)$ whose area over an interval, from a to b, is the chance of landing in it:

$$P(a \le X \le b) = \text{area under } f(x) \text{ from } a \text{ to } b = \int_a^b f(x)\,dx$$

The integral sign, $\int$, is calculus notation for that area, and the total area under a density is 1. The mean and variance are defined as for discrete variables, with areas in place of sums. The simplest density is flat: a delay equally likely to be anywhere from 0 to 10 minutes has a density of 0.1 on that range, so the chance of waiting under 3 minutes is $3 \times 0.1 = 0.3$.

The cumulative distribution, $F(x) = P(X \le x)$, is the area to the left of x. Turned around, it gives percentiles: the 99th percentile is the value with 99% of the area below it.

5b. The Normal Distribution

The normal distribution, the bell curve, is described by its mean μ (mu) and standard deviation σ (sigma), with density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)}$$

where $e \approx 2.718$ and $\pi \approx 3.1416$. Nobody computes with the formula by hand. What matters is its shape and a few landmarks: about 68% of values lie within one standard deviation of the mean, 95% within two, and 99.7% within three. Module 6 explains why the normal turns up so often: sums of many small independent effects are close to normal.

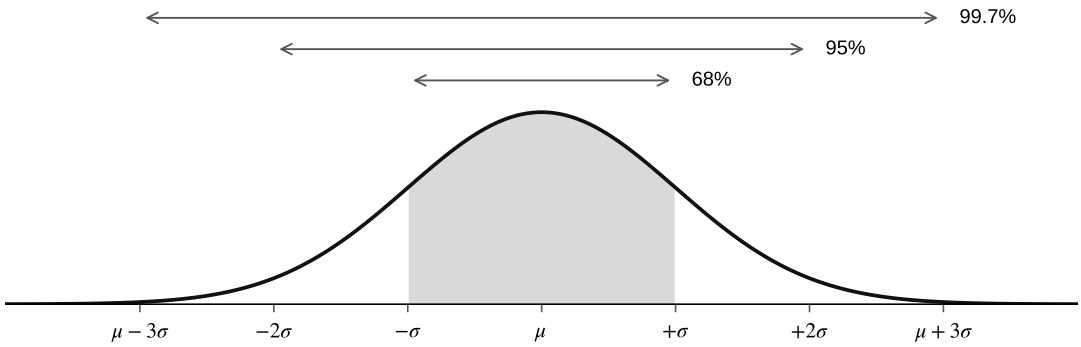


Figure 5.1. The normal distribution. About 68% of the area, shaded, lies within one standard deviation of the mean, 95% within two, and 99.7% within three.

Any normal variable can be measured in standard deviations from its mean, its z-score, $z = (x - \mu)/\sigma$. The standard normal distribution, with mean 0 and standard deviation 1, then gives every probability.

Worked example — A power plant's month

A plant's monthly output is normal with a mean of 500 gigawatt-hours (GWh) and a standard deviation of 30 GWh. A contract requires at least 450 GWh. The z-score of 450 is $(450 - 500)/30 \approx -1.67$, and the standard normal puts about 4.8% of its area below −1.67, so the plant falls short about one month in 21.

5c. The Lognormal Distribution: Prices

Stock prices cannot fall below zero, and they move in percentages rather than in fixed amounts. A standard model says that the logarithm of the price ratio over a period, the log return, is normal. The price itself is then lognormal: never negative, and skewed to the right, with a long tail of large gains. The Black–Scholes formula for option prices, published in 1973, rests on this model. Electricity prices are an exception to the rule of no negative prices: they can turn negative when supply swamps demand.

If the log return has mean m and standard deviation s, the median price ratio is e^{m}, while the mean price ratio is higher, $e^{m + s^2/2}$, because the tail of large gains pulls the mean up.

Worked example — A stock in a year

A stock costs $100. Its log return over a year has mean 0.06 and standard deviation 0.20. The median price in a year is $100 e^{0.06} \approx 106.18$ dollars and the mean $100 e^{0.08} \approx 108.33$. The chance of ending below $80 is the chance that the log return falls below $\ln 0.8 \approx -0.223$, a z-score of $(-0.223 - 0.06)/0.20 \approx -1.42$, about 7.8%.

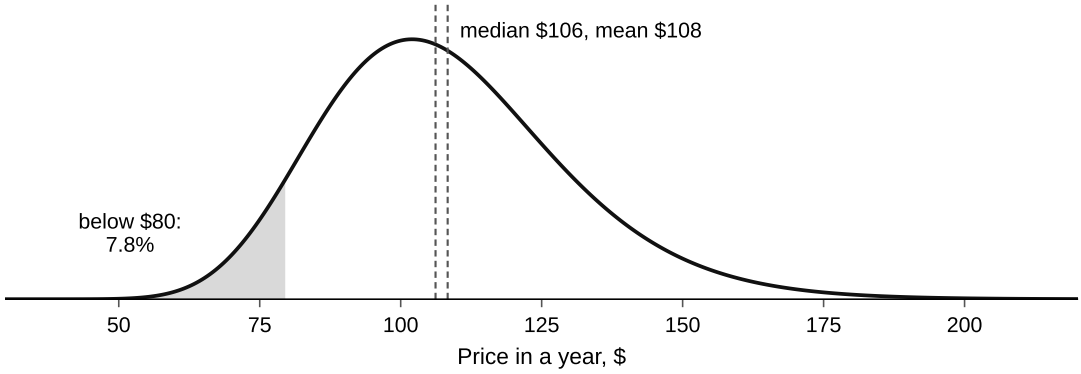


Figure 5.2. The lognormal distribution of a $100 stock's price a year later. The long right tail pulls the mean above the median; the shaded area is the 7.8% chance of ending below $80.

5d. The Exponential Distribution: Waiting and Failure

If events arrive at random at an average rate of λ (lambda) per unit of time, as in the Poisson model, the waiting time T until the next one has the exponential distribution. The chance of waiting longer than a time t, and the expected wait $E[T]$, are

$$P(T > t) = e^{-\lambda t}, \qquad E[T] = \frac{1}{\lambda}$$

Its striking property is that it has no memory. However long you have already waited, the remaining wait has the same distribution as at the start. For equipment, that means a constant failure rate: a unit that has survived 20 years is as good as new.

Worked example — Transformer lifetimes

If a transformer fails at a constant rate of 0.02 a year, its expected lifetime is $1/0.02 = 50$ years, and the chance it survives 10 years is $e^{-0.2} \approx 0.82$. In a fleet of 1,000, about 20 fail each year.

Real equipment is rarely so simple. Failure rates are often high early, when manufacturing defects show up, low in midlife, and rising again as parts wear out, a pattern engineers call the bathtub curve. The Weibull distribution, whose failure rate can steadily rise or steadily fall, is the usual model for each stage; the whole bathtub takes a combination of them.

5e. Fat Tails

The normal distribution's tails thin out extremely fast: a move of 4 standard deviations or more, in either direction, has a chance of about 1 in 16,000. On daily market data, that would be about one day every 63 years of trading. Real markets produce such days several times a decade or more, and occasionally far worse.

On October 19, 1987, the Standard & Poor's 500 index of large U.S. stocks (S&P 500) fell 20.5% in a day. With a typical daily standard deviation of about 1%, that was a move of some 20 standard deviations. The normal distribution gives it a chance smaller than 1 in 10^{90}, so it should never have happened in the life of the universe.

Distributions with more weight in the tails than the normal are called fat-tailed. A common model is Student's t distribution, whose tails get fatter as its parameter, the degrees of freedom, gets smaller. Scaled to the same standard deviation, a t distribution with 3 degrees of freedom produces moves beyond 4 standard deviations about once every 160 days, a hundred times as often as the normal.

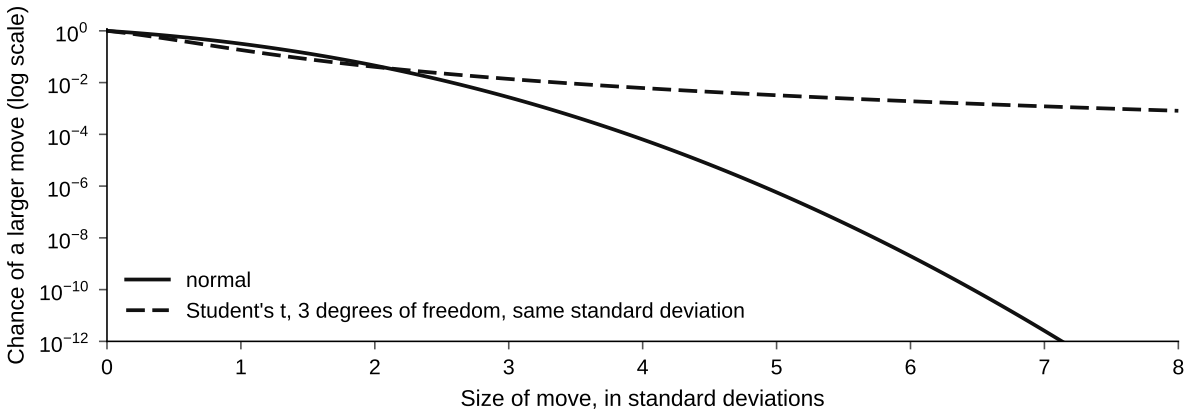


Figure 5.3. The chance of a move larger than a given number of standard deviations. The normal's tail collapses; the fat-tailed distribution's declines slowly, so extreme moves remain plausible.

Fat tails appear in power prices, which can jump a hundredfold in a shortage, in insurance losses from catastrophes, and in the sizes of cities and fortunes. Wherever they appear, a model built on the normal will understate how bad the worst cases are.

5f. Value at Risk and Expected Shortfall

Banks and funds summarize their tail risk in two numbers. The Value at Risk (VaR) at 99% is the loss exceeded on only 1% of days. The expected shortfall is the average loss on those worst 1% of days, so it also measures how bad the bad days are.

Worked example — A $10 million portfolio

A portfolio of $10 million has daily returns with a standard deviation of 1%. If they were normal, the 99% one-day VaR would be $2.33 \times 0.01 \times 10{,}000{,}000 \approx 233{,}000$ dollars, since 2.33 standard deviations cuts off the lowest 1% of the normal. The expected shortfall, the average loss beyond that point, would be about 2.67 standard deviations, or $267,000.

With fat tails, both are larger, and the expected shortfall more so, since it averages the extreme losses rather than marking where they begin. The Basel Committee's market-risk rules for banks, finalized in 2019, replaced 99% VaR with expected shortfall at 97.5%, which for a normal distribution is about the same size, 2.34 standard deviations, but grows when the tails are fat.

In a Spreadsheet

Spreadsheets have every distribution of this module, with mean in B1 and standard deviation in B2:

To compute	Formula
Normal chance below x	=NORM.DIST(x, B1, B2, TRUE), or =NORM.S.DIST(z, TRUE) for a z-score
Percentile of a normal	=NORM.INV(0.99, B1, B2); =NORM.S.INV(0.99) gives 2.326
Lognormal chance below x, for a log with mean m and standard deviation s	=LOGNORM.DIST(x, m, s, TRUE)
Exponential chance of failure by time t	=EXPON.DIST(t, λ, TRUE)
Student's t chance below x	=T.DIST(x, degrees_of_freedom, TRUE)
99% one-day VaR of a normal portfolio worth V	=-NORM.INV(0.01, B1, B2)*V

5g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Density	$P(a \le X \le b) = \int_a^b f(x)\,dx$ For a continuous variable, probability is area under the density $f(x)$, the total area is 1, and any single value has probability 0
Normal landmarks	About 68% of values within one standard deviation σ of the mean μ, 95% within two, and 99.7% within three
z-score	$z = \frac{x - \mu}{\sigma}$ The distance from the mean in standard deviations; the standard normal turns it into a probability
Lognormal	$\text{median} = e^m, \qquad \text{mean} = e^{m+s^2/2}$ For a price ratio whose log is normal with mean m and standard deviation s; never negative, skewed right
Exponential	$P(T > t) = e^{-\lambda t}, \qquad E[T] = 1/\lambda$ The wait T for random events at rate λ; memoryless, meaning a constant failure rate
Fat tails	More probability far from the mean than the normal allows; a 4-standard-deviation move is a once-in-63-years day under the normal, but routine in markets
VaR and expected shortfall	The 99% VaR is the loss exceeded 1% of the time; the expected shortfall is the average loss in that worst 1%; for a normal, 2.33 and 2.67 standard deviations
The idea	Continuous probabilities are areas. The normal is the right model for sums of many small effects and the wrong one for the extremes of markets and catastrophes, where tails are fat.

5h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

5.1 A delay is equally likely to be anywhere from 0 to 10 minutes. What is the chance it lasts between 2 and 5 minutes, and what is its mean?

5.2 A normal variable has mean 100 and standard deviation 15. Find the chance of a value above 130, and the chance of a value between 85 and 115.

5.3 Find the z-score of 72 when the mean is 60 and the standard deviation 8.

5.4 Find the 99th percentile of a normal with mean 50 and standard deviation 10.

5.5 Customers arrive at random at 0.5 an hour. What is the chance of waiting more than 3 hours for the next, and the mean wait?

5.6 A price ratio's log is normal with mean 0 and standard deviation 0.5. Find its median and mean.

Applied

5.7 In the worked example of section 5b, how often would the plant fall short of 440 GWh?

5.8 For the stock of section 5c, what is the chance it ends the year above $130?

5.9 A $50 million portfolio has normal daily returns with a standard deviation of 1.2%. Find its 99% one-day VaR and expected shortfall.

5.10 Transformers fail at a constant rate of 0.02 a year. What is the chance that one survives 30 years? How many failures a year should a fleet of 1,000 expect?

5.11 Under the normal, how often is a daily move larger than 4 standard deviations? How often under the t distribution of section 5e, scaled to the same standard deviation?

5.12 The S&P 500 fell 20.5% on October 19, 1987. If daily moves had a normal distribution with a standard deviation of 1%, how unusual would that be?

Challenge

5.13 Show that the exponential distribution is memoryless: the chance of surviving another t years, given survival for s years, does not depend on s.

5.14 Show that the z-score $Z = (X - \mu)/\sigma$ of any variable with mean μ and standard deviation σ has mean 0 and standard deviation 1.

5.15 For a lognormal price ratio with $s = 1$, how much larger is the mean than the median? Why?

5.16 Explain why the expected shortfall at any level is never smaller than the VaR at the same level.

Module 5 Quiz

Check Your Understanding

1. Why does a continuous variable have zero chance of taking any one exact value?

A) Continuous variables are always measured with error.
B) Probability is area under the density, and a single point has no width.
C) Its density is too small to measure at any one point.
D) Exact values occur, but only once in every few thousand observations, so their probability rounds down to zero.

2. Under a normal distribution, about what share of values lie within two standard deviations of the mean?

A) 68%
B) 99.7%
C) About 50%
D) About 95%, leaving about 2.5% in each tail

3. Why are asset prices often modeled as lognormal rather than normal?

A) They cannot fall below zero
B) They move by fixed amounts
C) Because returns are always positive in the long run, so a price must keep rising
D) Lognormal prices are symmetric, so gains and losses are equally likely

4. What does the expected shortfall at 99% measure?

A) The largest loss that has ever happened in the whole history of the portfolio's returns
B) The loss that is exceeded on exactly 1% of days, and no more than that
C) The average loss in the worst 1% of cases
D) The total amount the portfolio is expected to lose over a year

5. A transformer fails at a constant rate of 0.02 a year. What is its expected lifetime?

A) 50 years
B) 0.02 years
C) About 35 years, the time by which half have failed
D) 25 years, half of 50, since units fail midway through life on average

Module 6 — Sums, Averages, and the Central Limit Theorem

Why averages are predictable when single outcomes are not

Recommended Reading / Reference

Book: Leonard Mlodinow, *The Drunkard's Walk: How Randomness Rules Our Lives* (2008) — the law of averages and its misreadings, from casinos to careers

Reference: Nicholas Metropolis and Stanislaw Ulam, “The Monte Carlo Method,” *Journal of the American Statistical Association* 44 (1949) — the short paper that named the simulation method of section 6d

6a. Sums of Random Variables

Many quantities that matter are sums of many uncertain pieces: an insurer's claims, a utility's demand, a battery's revenue over a year. If $X_1, \ldots, X_n$ are independent and each has mean μ and standard deviation σ, the rules of Module 3 give the sum and the average:

$$\begin{aligned} \text{sum:} &\quad \text{mean } n\mu, \quad \text{standard deviation } \sigma\sqrt{n} \\ \text{average:} &\quad \text{mean } \mu, \quad \text{standard deviation } \sigma/\sqrt{n} \end{aligned}$$

The sum's spread grows, but only like the square root of n, so relative to its size it shrinks. The average's spread shrinks toward zero.

Worked example — An insurance pool

Each policy in a pool has an expected claim of \$500 a year but a standard deviation of \$5,000: most policies pay nothing and a few pay a great deal. For one policy, the outcome is a lottery. For 10,000 independent policies, total claims average \$5 million with a standard deviation of $5{,}000 \times \sqrt{10{,}000} = 500{,}000$ dollars, so the cost per policy is \$500, give or take \$50.

6b. The Law of Large Numbers

The law of large numbers says that the average of many independent draws from the same distribution gets closer and closer to the mean. Casinos, insurers, and lotteries are built on it: no single bet or policy is predictable, but the average over millions is.

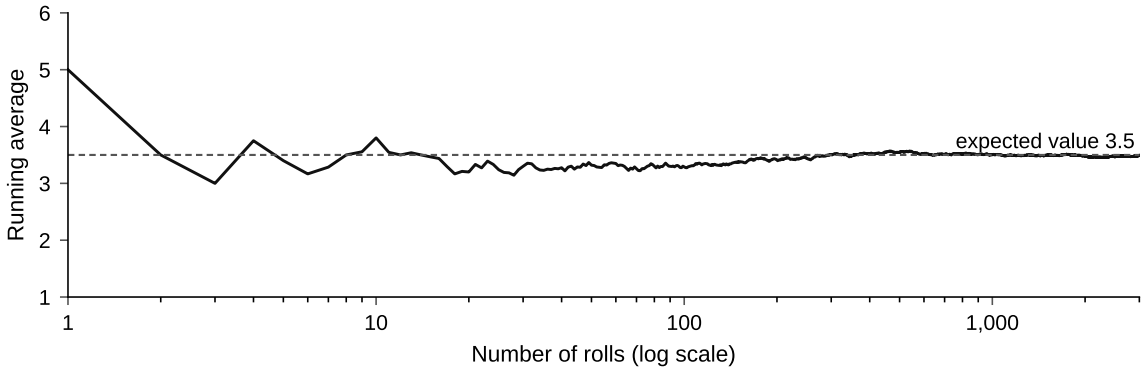


Figure 6.1. The running average of 3,000 rolls of a die. It wanders at first and then settles toward 3.5, with its typical error shrinking like one over the square root of the number of rolls.

The law is often misread. It does not say that outcomes balance out in the short run: after five reds at roulette, black is no more likely than before. The average converges because the early run is swamped by the many draws that follow, not because chance corrects it. Believing otherwise is the gambler's fallacy.

6c. The Central Limit Theorem

The central limit theorem says more: the sum or average of many independent draws is approximately normal, whatever the shape of the distribution being drawn from, as long as its standard deviation is finite. The sum is close to normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

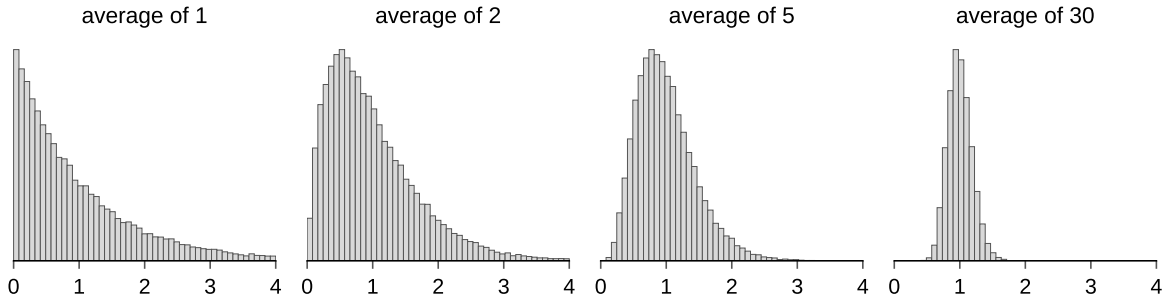


Figure 6.2. Averages of draws from a lopsided distribution, the exponential. One draw is heavily skewed; the average of 30 is already close to a symmetric bell centered on the mean of 1.

This is why the normal appears everywhere. Measurement errors, a utility's total demand, and the error in a poll are each the sum of many small independent effects. The standard deviation of an average, $\sigma/\sqrt{n}$, is called its standard error, and Modules 7 to 9 are built on it.

Worked example — The insurance pool's bad year

For the pool of 10,000 policies, total claims are close to normal with mean \$5 million and standard deviation \$500,000. Claims above \$6 million are 2 standard deviations above the mean, a chance of about 2.3%. An insurer can hold reserves for such a year, though it could never reserve sensibly for one policy.

The theorem needs independence and a finite standard deviation, and it describes the middle of the distribution better than its far tails. When draws share a common cause, or when single draws can be enormous, as with catastrophe losses, the sum's tails stay fatter than the normal's.

6d. Monte Carlo Simulation

When a quantity depends on many uncertain inputs in a complicated way, simulate it. Draw random values for the inputs, compute the result, and repeat thousands of times; the results form the quantity's distribution. Stanislaw Ulam and John von Neumann developed the method at Los Alamos in the 1940s, and it was named for the Monte Carlo casino.

Worked example — A battery's year

Take the battery of Module 3, whose daily profit is $200, $800, or $3,000 with probabilities 0.6, 0.3, and 0.1, and treat the 365 days as independent. A simulation of 10,000 years gives an average annual profit of about $241,000, with a standard deviation of about $15,700.

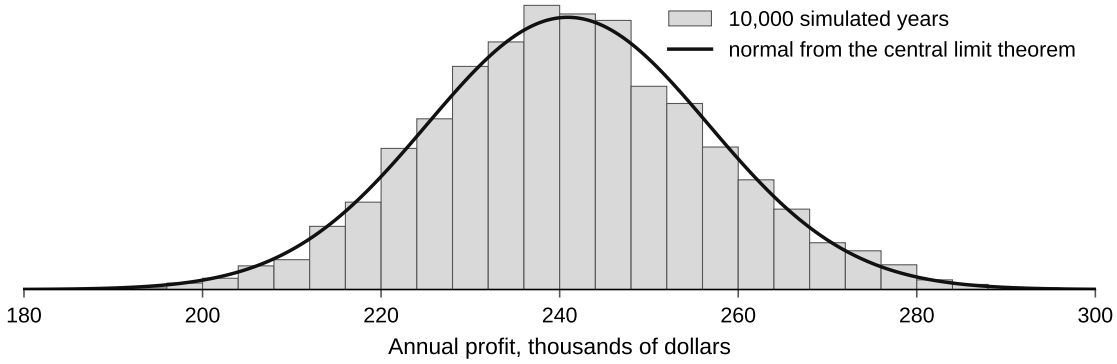


Figure 6.3. Simulated annual profit of the battery, with the normal curve that the central limit theorem predicts. The two agree closely, but the simulation's left tail is slightly thinner.

The central limit theorem predicts a mean of $365 \times 660 = 240{,}900$ dollars and a standard deviation of $825 \times \sqrt{365} \approx 15{,}760$ dollars, matching the simulation. It puts the chance of a year below $210,000 at about 2.5%; the simulation finds 2.1%, because the rare spikes make each day's profit lopsided, and the sum of 365 lopsided days is still slightly lopsided.

A simulation's own error shrinks like one over the square root of the number of runs, so quadrupling the runs halves it. The method's real weakness is its inputs: here, treating days as independent ignores heat waves, which bring spikes in runs, so the true spread of annual profit is wider.

6e. When Averaging Fails

Averaging removes the risk that draws do not share, never the risk they do. If each pair of draws has correlation ρ (rho), the variance of the average of n of them is

$$\mathrm{Var}(\text{average}) = \frac{\sigma^2}{n} + \left(1 - \frac{1}{n}\right)\rho\,\sigma^2$$

which approaches $\rho\sigma^2$ rather than zero as n grows. Twenty wind farms in one region, each with output varying by 30 megawatts (MW) and a correlation of 0.6 because the same weather drives them, have a total that varies by about 472 MW. Independent farms would vary by only 134 MW. Diversification works when the pieces are truly separate: wind farms far apart, policies in different regions, assets with different drivers.

In a Spreadsheet

Spreadsheets can run a simulation directly. With a mean in B1 and a standard deviation in B2:

To compute	Formula
A random normal draw	=NORM.INV(RAND(), B1, B2)
Mean and standard deviation of a sum of n independent draws	=n*B1 and =SQRT(n)*B2
Standard error of the average of the data in A2:A101	=STDEV.S(A2:A101)/SQRT(COUNT(A2:A101))
Repeat a whole simulation many times	What-If Analysis, Data Table: point a one-column table at the result cell, and each row recalculates every RAND
Share of simulated results below a threshold	=COUNTIF(C2:C10001, "<210000")/10000

6f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Sums and averages	sum: $n\mu,\ \sigma\sqrt{n}$ average: $\mu,\ \sigma/\sqrt{n}$ Mean and standard deviation for n independent draws, each with mean μ and standard deviation σ
Standard error	$\mathrm{SE} = \dfrac{\sigma}{\sqrt{n}}$ The standard deviation of an average; halving it takes four times the data
Law of large numbers	The average of many independent draws approaches the mean; it does not make outcomes balance out in the short run
Central limit theorem	The sum or average of many independent draws is approximately normal, whatever their distribution, if its standard deviation is finite
Monte Carlo	Simulate the inputs many times and read off the distribution of the result; its error shrinks like one over the square root of the number of runs
Correlated averages	$\dfrac{\sigma^2}{n} + \left(1 - \dfrac{1}{n}\right)\rho\,\sigma^2 \ \to\ \rho\,\sigma^2$ The variance of an average of n draws with pairwise correlation ρ: shared risk does not average away
The idea	Independent pieces cancel each other's noise, so totals are steady in proportion and averages are nearly normal. Shared causes and huge single outcomes are where that comfort ends.

6g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

6.1 25 independent draws each have mean 10 and standard deviation 4. Find the mean and standard deviation of their sum and of their average.

6.2 How much more data does it take to halve a standard error? To cut it to a tenth?

6.3 What are the mean and standard error of the average of 100 rolls of a die?

6.4 Each policy has an expected claim of $100 and a standard deviation of $1,000. Compare the relative spread of total claims for 400 policies and for 40,000.

6.5 A simulation of 1,000 runs estimates an average with a standard error of $2,000. How many runs would bring it to $500?

6.6 Use the central limit theorem to estimate the chance of 60 or more heads in 100 tosses of a fair coin.

Applied

6.7 For the battery of section 6d, use the central limit theorem to find the chance of a year below $210,000.

6.8 A call center receives 2 calls a minute on average, at random. Use the normal approximation to estimate the chance of more than 140 calls in an hour.

6.9 50 stocks each have an annual standard deviation of 30%, and every pair has a correlation of 0.3. What is the standard deviation of an equally weighted portfolio?

6.10 A poll of 1,000 voters finds 52% support. Find the standard error of that share, and the range within about two standard errors.

6.11 Twenty wind farms in one region each have output with a standard deviation of 30 MW, and every pair has a correlation of 0.6. Find the standard deviation of their total, and compare it with independent farms.

6.12 A project's cost is equally likely to be anywhere from $80 million to $120 million, and its revenue is normal with mean $130 million and standard deviation $20 million, independently. Estimate the chance of a loss.

Challenge

6.13 Show that the average of n independent draws with standard deviation σ has variance σ^2/n.

6.14 Derive the formula of section 6e for the variance of an average of n correlated draws.

6.15 Chebyshev's inequality says any variable is at least k standard deviations from its mean with probability at most $1/k^2$. Use it to show that the average of independent draws converges to the mean.

6.16 An old trick makes a nearly normal random number by adding 12 independent draws that are each equally likely to be anywhere from 0 to 1, and subtracting 6. Why does it work?

Module 6 Quiz

Check Your Understanding

1. Why do an insurer's total claims become more predictable as it writes more independent policies?

A) Each policy becomes less risky
B) The spread of the total grows only like the square root of the number of policies, while the total grows in proportion to it
C) Large insurers pay smaller claims
D) Larger insurers pick only the safest customers, so most of their policies never pay out at all

2. To halve the standard error of an average, how many more observations are needed?

A) Twice as many
B) Eight times as many
C) Four times as many
D) It cannot be halved, because the standard error depends only on the spread of the data, not on the number of observations

3. What does the central limit theorem say?

A) All data are normal
B) The average of many observations equals the mean exactly, whatever their distribution
C) Small samples are normal
D) Sums of many independent variables are approximately normal

4. What is a Monte Carlo simulation?

A) A model of casino profits
B) Repeated random draws
C) An exact formula for any risk
D) A method that runs the model once, with every input set to its average value

5. Why does adding more wind farms in one region smooth total output less than independence would suggest?

A) Their outputs are correlated, because the same weather drives them all.
B) Wind farms interfere with each other's turbines.
C) Larger fleets are switched off more often for maintenance.
D) Each new farm adds less capacity than the one before, because the best sites are always built first.

Module 7 — Estimation and Confidence Intervals

How to estimate an unknown from a sample, and say how sure you are

Recommended Reading / Reference

Book: David Spiegelhalter, *The Art of Statistics* (2019) — learning from data through real cases, with a clear, formula-light account of estimates, intervals, and the bootstrap

Reference: U.S. Census Bureau, American Community Survey (census.gov/programs-surveys/acs) — the bureau's largest household survey, which publishes every estimate with its margin of error at 90% confidence

7a. Samples and Estimators

Almost nothing worth knowing can be measured in full. A pollster cannot ask every voter, a utility cannot meter every appliance, and a battery maker cannot test every cell to failure. Statistics estimates an unknown quantity, called a parameter, from a sample, and then says how far off the estimate is likely to be.

A rule that turns a sample into an estimate is an estimator. For n values $x_1, \ldots, x_n$, the sample average $\bar{x}$ (x-bar), their sum divided by n, estimates the population mean μ. The share $\hat{p}$ (p-hat) of a sample who say yes estimates the true share p. The median, the middle value once the data are sorted, estimates the middle of the population.

The sample standard deviation s measures the spread of the data around their average:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

Dividing by $n - 1$ rather than n corrects a slight understatement. The data sit closer to their own average than to the true mean, so their squared distances from $\bar{x}$ run a little small, and the smaller divisor makes up for it.

An estimate can be wrong in two ways. Random error comes from which sample happened to be drawn, and it shrinks as the sample grows. Bias is a systematic error, from a sample that does not represent the population or a question that pushes answers one way, and more data does not shrink it at all.

Worked example — A poll of two million

In 1936 the *Literary Digest* magazine mailed some 10 million ballots and got more than 2 million back. It predicted that Alf Landon would beat Franklin Roosevelt, with 57% of the vote between them. Roosevelt won 61% of the popular vote and 46 of the 48 states. George Gallup, polling a far smaller sample chosen to resemble the electorate, called the winner correctly.

The magazine's lists, drawn largely from telephone directories, car registrations, and its own subscribers, overrepresented the better-off in the Depression, and the minority who mailed ballots back leaned further toward Landon. A sample of 2 million has a random error of a few hundredths of a percentage point, yet the poll missed by about 20 points. More data from the wrong source only makes a wrong answer look precise.

7b. Standard Errors

Another sample would give a different estimate. The standard deviation of an estimate across all the samples that might have been drawn is its standard error (SE), the yardstick for random error. Module 6 gave it for an average, and the binomial of Module 4 gives it for a share:

$$\mathrm{SE}(\bar{x}) = \frac{s}{\sqrt{n}}, \qquad \mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

Both put the sample's own spread in place of the unknown true one, which is accurate enough once n is more than a few dozen. Both shrink like $1/\sqrt{n}$, so halving an error takes four times the data.

Worked example — Household electricity use

A utility reads the smart meters of 400 randomly chosen homes and finds an average of 29.4 kilowatt-hours (kWh) a day, with a standard deviation of 8 kWh. The standard error of the average is $8/\sqrt{400} = 0.4$ kWh, the typical distance between an average of 400 homes and the true average of all of them.

7c. Confidence Intervals

By the central limit theorem, an estimate from a large sample is close to normal, centered on the true value, and the normal puts 95% of its area within 1.96 standard deviations of the center. So the interval

$$\text{estimate} \pm 1.96 \times \mathrm{SE}$$

catches the true value in 95% of samples. It is called a 95% confidence interval, and 1.96 times the standard error is its margin of error. For 90% confidence the multiplier is 1.645, and for 99% it is 2.576.

Worked example — Household electricity use, continued

The 95% confidence interval for the average use of all the utility's homes is $29.4 \pm 1.96 \times 0.4$, from 28.6 to 30.2 kWh a day.

The confidence belongs to the method, not to any one interval. The true average is a fixed number, either inside 28.6 to 30.2 or not. What is known is that the procedure that produced the interval succeeds 19 times in 20, as Figure 7.1 shows.

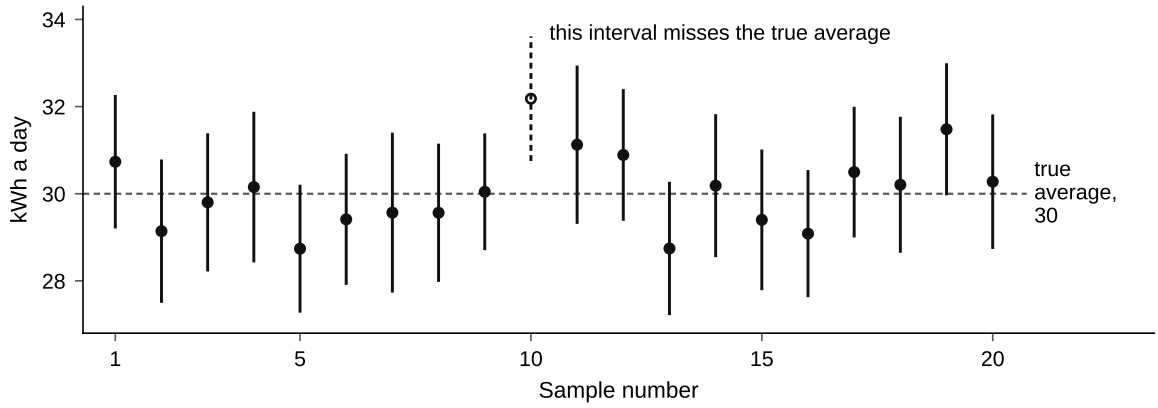


Figure 7.1. Twenty samples of 100 homes, each with its 95% confidence interval, from homes whose true average use is 30 kWh a day. Nineteen intervals catch the true value; one, dashed, misses it.

Polls

For a poll, the margin of error is largest when support is near 50%, where $\hat{p}(1-\hat{p})$ reaches its maximum of 0.25. Then

$$\text{margin of error} = 1.96\sqrt{\frac{0.25}{n}} \approx \frac{1}{\sqrt{n}}$$

so a poll of 1,000 people has a margin of about 3 percentage points, the figure printed under most news polls. Turned around, the margin fixes the sample size. A margin of m, written as a decimal, needs $n = (1.96/m)^2 \times 0.25$ people: 1,068 for 3 points, and 2,401 for 2.

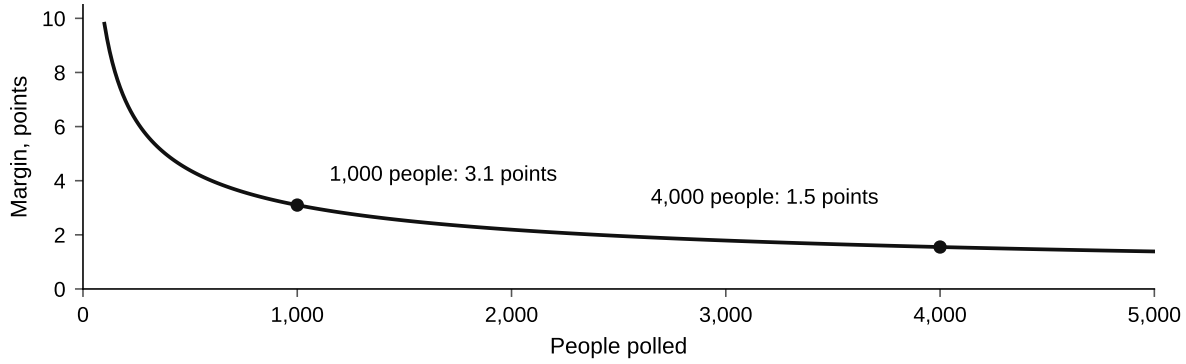


Figure 7.2. The 95% margin of error of a poll, for support near 50%. It falls like one over the square root of the sample size, so halving the 3-point margin of a 1,000-person poll takes 4,000 people.

Two cautions apply. The margin covers only random sampling error, not bias from who answers or how a question is worded. And the gap between two candidates in the same poll has a margin nearly twice as large, since a voter lost by one is often a voter gained by the other. For shares $\hat{p}_1$ and $\hat{p}_2$ of the same n people, the gap's standard error is

$$\text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - (\hat{p}_1 - \hat{p}_2)^2}{n}}$$

Government Surveys

Official statistics come from samples too, and the best agencies publish their margins. The American Community Survey reports every estimate with a margin of error at 90% confidence. The Bureau of Labor Statistics (BLS) estimates the monthly change in U.S. payroll jobs from a survey of employers, with a 90% margin of error that in recent reports has been roughly 120,000 jobs.

So a reported gain of 50,000 jobs is consistent with anything from a loss of 70,000 to a gain of 170,000. Month-to-month headlines about jobs often treat differences smaller than the margin as news.

7d. Small Samples and the t Distribution

With a small sample, s is itself a rough estimate of the true spread, and the interval must widen to allow for that. In 1908 William Gosset, a chemist at the Guinness brewery in Dublin, worked out by how much, publishing under the name Student. The interval for an average uses Student's t distribution, met in Module 5 as a model of fat tails, with $n - 1$ degrees of freedom:

$$\bar{x} \pm t_{n-1} \frac{s}{\sqrt{n}}$$

Here t_{n-1} is the multiplier that leaves 2.5% in each tail. It is 2.78 for 5 observations, 2.26 for 10, 2.06 for 25, and 1.98 for 100, approaching 1.96 as the sample grows.

Worked example — Six battery cells

A battery maker cycles six cells 500 times and measures their loss of capacity: 2.1, 2.5, 1.8, 2.9, 2.4, and 2.2 percent. The average is 2.32 percent and $s \approx 0.376$, so the standard error is $0.376/\sqrt{6} \approx 0.154$. With 5 degrees of freedom the multiplier is 2.571, and the 95% interval runs from 1.92 to 2.71 percent.

Using 1.96 would give 2.02 to 2.62, an interval too narrow to deserve its 95%. The t interval assumes the measurements themselves are roughly normal, which matters with six values and fades as the sample grows.

7e. The Bootstrap

Some estimates have no handy formula for their standard error: a median, a ratio, the 99th percentile of losses. The bootstrap, introduced by the statistician Bradley Efron in 1979, gets one by brute force. Treat the sample as a stand-in for the population, draw new samples of the same size from it, and recompute the estimate on each. The spread of the recomputed estimates measures the estimate's uncertainty.

The new samples are drawn with replacement: each value can be drawn again after it has been drawn, so a resample repeats some values and leaves out others. The middle 95% of the resampled estimates, from their 2.5th to their 97.5th percentile, is a 95% interval.

Worked example — How long outages last

A utility's crews restored power after 19 outages, in hours: 0.4, 0.5, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, 2.1, 2.5, 3.0, 3.8, 4.9, 6.5, and 11.2. The median is 1.4 hours, while a few long repairs pull the average up to 2.46. In 10,000 resamples, the 2.5th and 97.5th percentiles of the medians are 1.0 and 2.5 hours, so the 95% interval for the median is 1.0 to 2.5 hours.

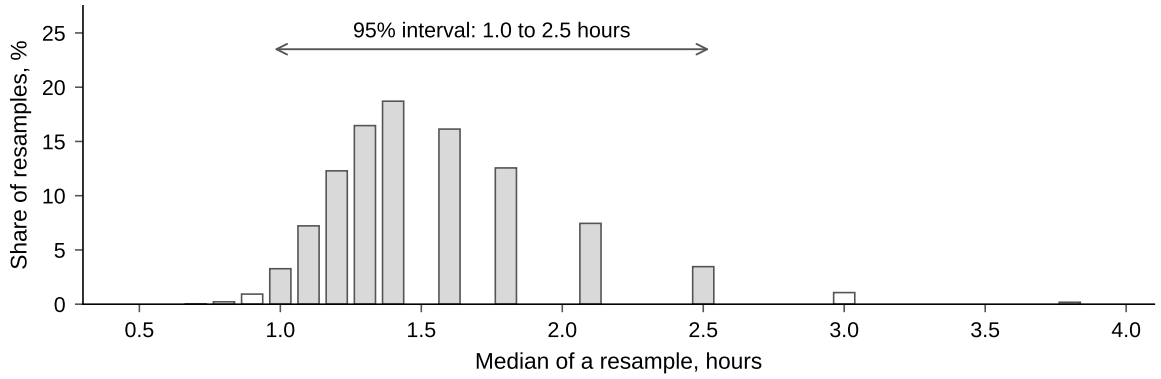


Figure 7.3. The medians of 10,000 resamples of the 19 outage times. A median can only be one of the data values, so the bars sit at those values. The shaded bars run from the 2.5th to the 97.5th percentile.

The bootstrap needs the sample to represent the population, and it struggles with very small samples and with extremes such as the maximum, which no resample can exceed. Within those limits it frees estimation from formulas, and statisticians use it for almost any estimate that a computer can recompute.

In a Spreadsheet

Spreadsheets compute every interval in this module. For data in A2:A101:

To compute	Formula
Average, sample standard deviation, and count	=AVERAGE(A2:A101), =STDEV.S(A2:A101), and =COUNT(A2:A101)
95% margin of error for the average, large sample	=CONFIDENCE.NORM(0.05, STDEV.S(A2:A101), COUNT(A2:A101))
95% margin of error for the average, small sample	=CONFIDENCE.T(0.05, STDEV.S(A2:A101), COUNT(A2:A101))
The t multiplier for n observations	=T.INV.2T(0.05, n-1); for six observations it gives 2.571
Standard error of a share $\hat{p}$ of n people	=SQRT(p*(1-p)/n), with the share in place of p
One bootstrap resample of 19 values in A2:A20	=INDEX(A$2:A$20, RANDBETWEEN(1, 19)), copied down 19 rows; take its MEDIAN, and repeat with a Data Table

7f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Two kinds of error	Random error comes from which sample was drawn and shrinks like $1/\sqrt{n}$; bias is systematic, and more data does not remove it
Sample standard deviation	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ The spread of data $x_1, \ldots, x_n$ around their average $\bar{x}$; dividing by $n - 1$ corrects for measuring from $\bar{x}$
Standard errors	$\mathrm{SE}(\bar{x}) = \frac{s}{\sqrt{n}}, \qquad \mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ For an average $\bar{x}$ of n values with standard deviation s, and for a share $\hat{p}$ of n people
Confidence interval	$\text{estimate} \pm z \times \mathrm{SE}$ The multiplier z is 1.645 for 90% confidence, 1.96 for 95%, and 2.576 for 99%; that share of intervals made this way contains the true value
Poll margin	$1.96\sqrt{0.25/n} \approx 1/\sqrt{n}$ The 95% margin of error for a share near 50% from n people: about 3 points for 1,000; the gap between two candidates has nearly twice the margin
Small samples	$\bar{x} \pm t_{n-1}\, s/\sqrt{n}$ Student's t multiplier t_{n-1}, with $n - 1$ degrees of freedom: 2.57 for 6 values, 2.26 for 10, near 1.96 beyond 100
Bootstrap	Resample the data with replacement many times, recompute the estimate each time, and take the middle 95% of the results as the interval
The idea	An estimate is worth only as much as its error bar. Random error shrinks with more data; bias does not, and no margin of error includes it.

7g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

7.1 Find the average, median, and sample standard deviation of 4, 7, 9, 10, and 15.

7.2 A sample of 100 values has a standard deviation of 12. Find the standard error of its average. What would it be for 400 values with the same spread?

7.3 A poll of 600 voters finds 45% support. Find the standard error and the 95% confidence interval.

7.4 An average of 50 has a standard error of 1. Give its 90%, 95%, and 99% confidence intervals.

7.5 How many people must a poll survey for a 95% margin of error of 3 points on a share near 50%? For 2 points?

7.6 Find the 95% t multipliers for samples of 5 and of 25. How much wider than with 1.96 is each interval?

Applied

7.7 A survey of 2,000 people in the labor force finds 5.5% unemployed. Find the 95% margin of error.

7.8 A turbine's efficiency is measured five times: 41.2, 40.8, 41.9, 41.5, and 40.6 percent. Find a 95% confidence interval for its true efficiency.

7.9 A county's median household income is published as $68,000 with a margin of error of $4,000 at 90% confidence. Find the standard error and the 95% margin of error.

7.10 A poll of 1,000 voters has candidate A at 48% and candidate B at 46%. Is A clearly ahead?

7.11 A monthly jobs report shows a gain of 50,000 jobs, with a 90% margin of error of 120,000. What true changes are consistent with it? Does it show that jobs grew?

7.12 In the outage data of section 7e, suppose the longest outage had lasted 30 hours instead of 11.2. How would the average and the median change?

Challenge

7.13 Show that $p(1-p)$ is largest at $p = 1/2$, and that the 95% margin of error for a share near 50% is about $1/\sqrt{n}$.

7.14 Twenty independent studies each report a 95% confidence interval. What is the chance that all twenty contain their true values? How many should be expected to miss?

7.15 Compare the *Literary Digest* poll's standard error with its actual error. What does the comparison show?

7.16 Show that dividing the squared distances from $\bar{x}$ by n would underestimate the variance σ^2 on average.

Module 7 Quiz

Check Your Understanding

1. What does it mean that a poll's 95% confidence interval runs from 41% to 49%?

A) There is a 95% chance that the true share is exactly 45%.
B) The poll is 95% accurate.
C) Support will be between 41% and 49% on election day.
D) The interval came from a method that catches the true share in 95% of the samples it could have drawn.

2. A poll of 1,000 people finds support near 50%. About how large is its 95% margin of error?

A) About 10 points
B) About 1.5 points
C) About 3 points
D) Under 1 point

3. Why was the *Literary Digest*'s 1936 poll of more than 2 million people so far wrong?

A) Its sample was too small for a national election.
B) Its lists and its replies leaned toward Landon, and a larger sample cannot remove that kind of bias.
C) Many voters changed their minds in the weeks after it was taken.
D) Its margin of error was too wide to call the race.

4. Why does a 95% interval from only six measurements use a multiplier of 2.57 instead of 1.96?

A) Small samples always overstate the average, and the larger multiplier corrects that bias
B) The normal multiplier of 1.96 applies only to shares, not averages
C) The spread is itself estimated
D) A t distribution is used whenever the data are not normal

5. What does the bootstrap do?

A) Resamples the data to gauge uncertainty
B) Collects a larger sample
C) Removes outliers before computing the average
D) Replaces the data with a fitted normal distribution

Module 8 — Hypothesis Tests

Deciding whether a difference is real, and how tests mislead

Recommended Reading / Reference

Book: Stuart Ritchie, *Science Fictions* (2020) — how flexible analyses, selective publishing, and hype filled journals with findings that do not replicate, and what reformers are doing about it

Reference: American Statistical Association (ASA), “The ASA Statement on p-Values: Context, Process, and Purpose,” by Ronald Wasserstein and Nicole Lazar, *The American Statistician* 70 (2016) — six principles for using and reading p-values

8a. The Logic of a Test

A trading signal called the market's direction correctly on 61 of 100 days. Is it skilled, or was it lucky? A hypothesis test answers by asking how surprising the result would be if there were nothing to find.

The test starts from a null hypothesis, the claim that there is no effect: here, that the signal is right with probability 0.5, like a coin toss. It then measures how far the data fall from what the null predicts, in standard errors. This test statistic is

$$z = \frac{\text{estimate} - \text{null value}}{\text{SE}}$$

Under the null, the number of correct calls is binomial with mean 50 and standard deviation $\sqrt{(100 \times 0.5 \times 0.5)} = 5$, so $z = (61 - 50)/5 = 2.2$.

The p-value is the chance, if the null hypothesis were true, of a result at least as far from it as the one observed. Here it is the chance of a z-score beyond 2.2 in either direction, about 0.028, the two shaded tails of Figure 8.1. By a convention that goes back to the statistician Ronald Fisher in the 1920s, a p-value below 0.05 is called statistically significant.

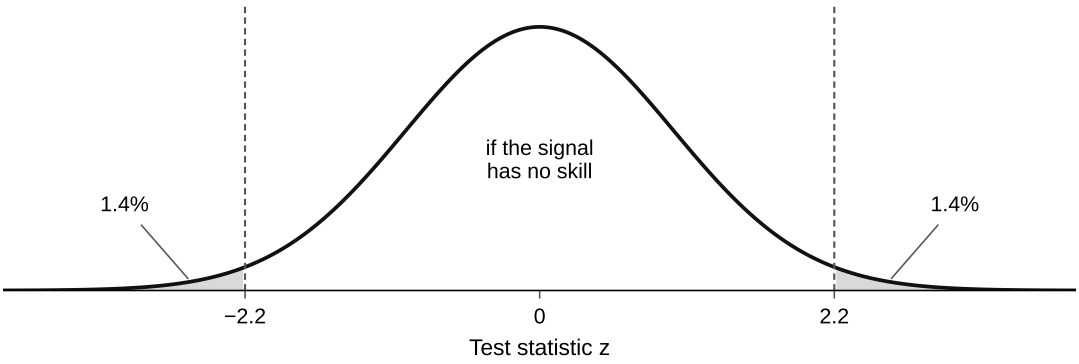


Figure 8.1. The p-value of the trading signal: the chance, if it had no skill, of a z-score at least 2.2 from zero. Each shaded tail holds 1.4%, so the p-value is about 0.028.

The test is two-sided: a signal right on only 39 days would be just as surprising, and just as useful, traded in reverse. An exact calculation with the binomial distribution of Module 4 gives 0.035 rather than 0.028; either way, a coin would land this far from 50, in one direction or the other, only about one time in 30. That is evidence against pure luck, though, as section 8e shows, not yet proof of skill.

8b. What a p-Value Does Not Mean

The p-value is the most misread number in science. It describes the data on the assumption that the null hypothesis is true. It is not the chance that the null is true, and a p-value of 0.028 does not mean a 97.2% chance that the signal is skilled.

In 2016 the American Statistical Association (ASA) took the unusual step of issuing principles for p-values. Among them: p-values do not measure the probability that a hypothesis is true; conclusions should not rest only on whether a p-value passes a threshold; and a p-value does not measure the size of an effect or its importance. A tiny, useless effect is significant in a large enough sample, and a large, valuable one can miss in a small one.

The chance that a significant result is real comes from Bayes' rule, in Module 2. It needs a prior, how often the ideas being tested are true, and the test's power, its chance of detecting an effect that is really there.

Worked example — How many discoveries are real?

Suppose 10% of the ideas a field tests are true, tests use the 5% level, and they detect a true effect 80% of the time. Of 1,000 ideas, 100 are true and 80 of those test significant; of the 900 false ones, 45 test significant too. So 45 of the 125 significant results, 36%, are false, far more than the 5% the threshold seems to promise.

With power of only 35%, common in small studies, the false share rises to 56%. The medical researcher John Ioannidis made this argument in a widely read 2005 paper, “Why Most Published Research Findings Are False.”

8c. Experiments on Websites

Module 4 compared two versions of a web page, each shown to 10,000 visitors, converting 5.0% and 5.3%. A test asks whether that gap could be chance. Under the null hypothesis that both versions convert at the same rate, the best estimate of the rate pools the two groups, and the test statistic is

$$z = \frac{\hat{p}_B - \hat{p}_A}{\sqrt{\bar{p}\,(1-\bar{p})\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}$$

where $\hat{p}_A$ and $\hat{p}_B$ are the two conversion rates, n_A and n_B the numbers of visitors, and $\bar{p}$ the pooled rate, all conversions divided by all visitors.

Worked example — Is 5.3% better than 5.0%?

The pooled rate is 5.15%, so the denominator is $\sqrt{(0.0515 \times 0.9485 \times 2/10{,}000)} \approx 0.0031$, and $z = 0.003/0.0031 \approx 0.96$. The two-sided p-value is 0.34: a gap this size turns up about a third of the time between two identical pages. The test has found nothing either way.

Had the new page converted 5.6%, z would be 1.89 and the p-value 0.058, just above the line. A p-value of 0.058 is barely different from one of 0.045, yet only one is called significant. A confidence interval is the better report: here the lift is 0.6 points, with a 95% interval from about 0 to 1.2 points, perhaps nothing and perhaps a gain worth a great deal.

A 5% test and a 95% interval nearly always agree: the test rejects “no difference” when the interval excludes zero. Online tests also invite a subtle error. Checking a running test every day and stopping as soon as its

p-value dips below 0.05 gives chance many tries. With ten looks at two identical pages, about 19% of tests end in a false win, not 5%. The remedy is to fix the sample size in advance, or to use methods built for repeated looks.

8d. Two Errors and Power

A test can go wrong in two ways. A Type I error, a false alarm, rejects a null hypothesis that is true; the significance level, written α (alpha), caps its chance, usually at 5%. A Type II error, a miss, fails to reject a null hypothesis that is false. Its chance is written β (beta), and the power of the test, its chance of detecting a real effect, is $1 - \beta$.

Power depends on the size of the effect, the noise in the data, and the sample size. For a test of two rates at the 5% level, a sample of n per group gives 80% power to detect true rates p_A and p_B when

$$n = \frac{(1.96 + 0.84)^2 \left[p_A(1 - p_A) + p_B(1 - p_B)\right]}{(p_B - p_A)^2}$$

Here 1.96 cuts off 2.5% in each tail, and 0.84 is the z-score that 80% of the normal lies below.

Worked example — Designing a test

To detect a lift from 5.0% to 5.5% takes $7.84 \times (0.0475 + 0.0520)/0.005^2 \approx 31{,}000$ visitors per version. The 10,000-visitor test of section 8c had only about a 35% chance of detecting such a lift.

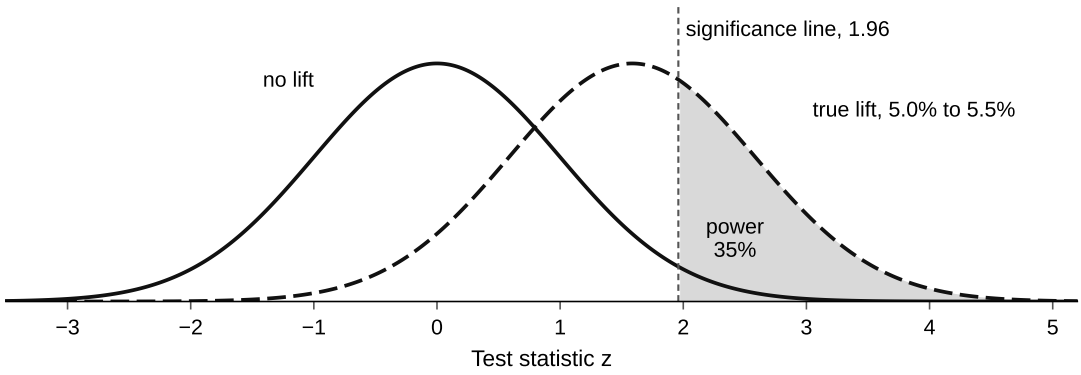


Figure 8.2. Why the 10,000-visitor test is weak. With no lift, z centers on 0 (solid curve); with a true lift from 5.0% to 5.5%, it centers near 1.6 (dashed), and only the shaded 35% of that curve clears the significance line. The solid curve's thin tail beyond the line is the 2.5% chance of a false alarm in that direction.

Weak tests do double damage. They miss most real effects, and when they do find one, its size tends to be exaggerated, since only the luckiest overestimates clear the bar.

8e. Many Tests at Once

Run enough tests and something will come up significant. Twenty independent tests of effects that do not exist, each at the 5% level, give a $1 - 0.95^{20} \approx 0.64$ chance of at least one false alarm. The simplest guard, the Bonferroni correction, divides the threshold by the number of tests, here to $0.05/20 = 0.0025$.

Worked example — The best of 100 backtests

An analyst tests 100 trading strategies on five years of past prices, and none has any real edge. A strategy's Sharpe ratio is its average annual return above a safe rate, divided by the standard deviation of that return. Estimated from five years, each Sharpe ratio has a standard error of about $1/\sqrt{5} \approx 0.45$. The best of 100 is typically about 2.5 standard errors above zero, a Sharpe ratio near 1.1, which would pass for a good fund's record.

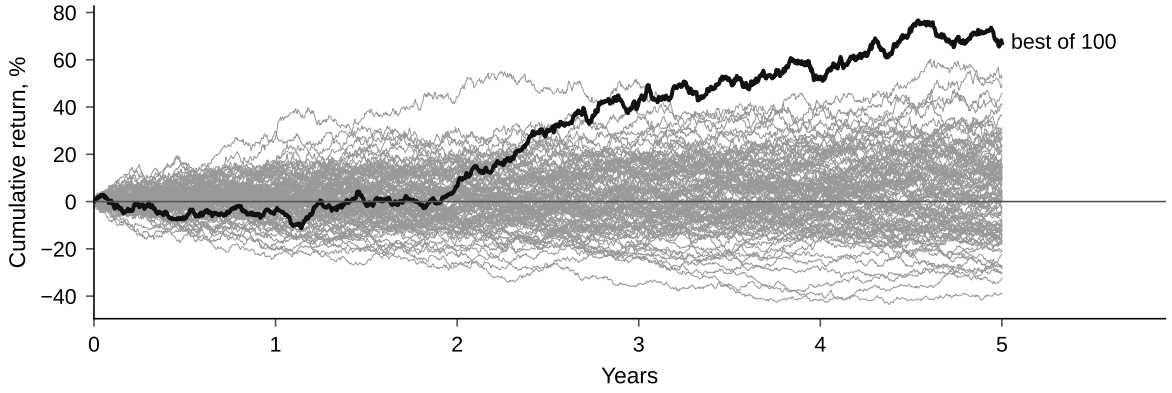


Figure 8.3. A hundred simulated trading strategies with no edge at all. The best, the heavy line, gained about 67% in five years, a Sharpe ratio of 1.1, by selection alone.

Backtests that try many variations and report the best are overfitted: they fit the noise of the past. The defense is to count the trials, demand evidence in proportion, and judge the winner on data it has never seen, the subject of Module 10.

Science has the same problem at scale. In a project published in 2015, the Open Science Collaboration repeated 100 published psychology experiments. 97% of the originals had reported significant results, but only 36% of the repeats found them. Flexible analysis, in which researchers try several measures, subgroups, or exclusions and report what works, together with journals' taste for positive results, had filled the literature with false alarms.

The reforms follow from this module: register the hypothesis and the analysis before collecting data, report every test that was run, publish results whether or not they are significant, and design studies with enough power to find what they seek.

In a Spreadsheet

Spreadsheets compute p-values directly. With a test statistic in B1:

To compute	Formula
Two-sided p-value for a z-statistic	=2*(1-NORM.S.DIST(ABS(B1), TRUE))
Two-sided p-value for a t-statistic with d degrees of freedom	=T.DIST.2T(ABS(B1), d)
p-value comparing the averages of two columns of data	=T.TEST(A2:A31, C2:C31, 2, 3): two-tailed, allowing different spreads
z for two conversion rates in D1 and D2, from visitors in E1 and E2	=(D2-D1)/SQRT(D3*(1-D3)*(1/E1+1/E2)), with the pooled rate =(D1*E1+D2*E2)/(E1+E2) in D3
Chance of at least one false alarm in k tests at 5%	=1-0.95^k
Visitors per group for 80% power, true rates a and b	=(1.96+0.8416)^2*(a*(1-a)+b*(1-b))/(b-a)^2

8f. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Test statistic	$z = \dfrac{\text{estimate} - \text{null value}}{\text{SE}}$ How far the data fall from what the null hypothesis, the claim of no effect, predicts, in standard errors (SE)
p-value	The chance, if the null hypothesis were true, of a result at least as far from it as the one observed; below 0.05 is called significant by convention
What it is not	Not the chance that the null hypothesis is true, not the size or importance of an effect, and not the chance that a result will replicate
Two errors	Type I: a false alarm, rejecting a true null, with chance α, the significance level. Type II: a miss, with chance β. Power is $1 - \beta$
Two rates	$z = \dfrac{\hat{p}_B - \hat{p}_A}{\sqrt{\bar{p}(1-\bar{p})(1/n_A + 1/n_B)}}$ Rates $\hat{p}_A$ and $\hat{p}_B$ from n_A and n_B visitors; $\bar{p}$ is the pooled rate, all conversions over all visitors
Sample size for 80% power	$n = \dfrac{7.84\,[p_A(1-p_A) + p_B(1-p_B)]}{(p_B - p_A)^2}$ Per group, at the 5% level, to detect true rates p_A and p_B; 7.84 is $(1.96 + 0.84)^2$
Many tests	$1 - 0.95^k$ The chance of at least one false alarm in k independent tests of effects that do not exist; the Bonferroni correction tests each at $0.05/k$
The idea	A test asks whether chance alone could plausibly explain the data. Its answer is only as honest as the count of tests behind it, and it says nothing by itself about how large or important an effect is.

8g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

8.1 A test statistic is $z = 2.5$. Find the two-sided p-value. Is the result significant at the 5% level? At 1%?

8.2 Which z-statistics are significant, two-sided, at the 10%, 5%, and 1% levels?

8.3 A coin lands heads 58 times in 100 tosses. Is there evidence that it is unfair?

8.4 A smoke alarm tests the null hypothesis “there is no fire.” Describe its Type I and Type II errors. What does a more sensitive alarm trade?

8.5 Twelve independent tests of effects that do not exist are run at the 5% level. Find the chance of at least one significant result, and the Bonferroni threshold.

8.6 A study has 80% power at the 5% level. If the effect it seeks is real, what is the chance it misses it? If the effect is not real, what is the chance of a false alarm?

Applied

8.7 An A/B test shows each version of a page to 20,000 visitors. Version A converts 4.0% and version B 4.4%. Test whether B is better.

8.8 How many visitors per version would a test need for 80% power to detect a lift from 10% to 11%?

8.9 A trading signal is right on 55 of 100 days. Test it. Then test a signal that is right on 550 of 1,000 days.

8.10 A fund manager beats the market in 8 of 10 years. If each year were a coin toss, what is the chance of 8 or more wins? Among 1,000 managers with no skill, how many would do it?

8.11 In a field where 20% of tested hypotheses are true, studies have 50% power and use the 5% level. What share of significant results are real?

8.12 An experiment's 95% confidence interval for a lift is −0.1 to 1.3 percentage points. Would a two-sided test at 5% reject “no lift”? What does the interval say that the test does not?

Challenge

8.13 Show that a two-sided test at the 5% level rejects a null value v exactly when v lies outside the 95% confidence interval, the estimate plus or minus 1.96 standard errors.

8.14 The best of 100 worthless strategies, tested on five years, shows a Sharpe ratio near 1.1. How many years of data would bring the expected best down to 0.5?

8.15 Explain why checking an A/B test every day, and stopping as soon as its p-value falls below 0.05, makes false wins more likely.

8.16 Derive the sample-size formula of section 8d.

Module 8 Quiz

Check Your Understanding

1. A test of a new drug gives a p-value of 0.03. What does that mean?

A) There is a 3% chance the drug has no effect.

B) The drug's effect is small.

C) There is a 97% chance the drug works, so the result will almost certainly hold up when the trial is repeated with new patients.

D) If the drug had no effect, results this extreme would occur about 3% of the time.

2. What is the power of a test?

A) Its chance of a false alarm when nothing is there

B) Its chance of detecting a real effect

C) The size of the effect

D) One minus the p-value of the test result

3. A team tests 20 website changes that in fact do nothing, each at the 5% level. What should it expect?

A) All 20 tests will correctly find no significant effect.

B) Exactly one of the 20 changes will test significant.

C) About one false alarm on average, with a 64% chance of at least one, so some change will probably test significant.

D) Significant results can appear only if the samples are too small.

4. An analyst backtests 100 strategies that have no real edge and reports the best. Why does its record look good?

A) Markets change over time

B) Trading costs were left out

C) Old prices contain recording errors

D) Selection favors luck

5. What is a Type II error?

A) Missing a real effect: failing to reject a false null hypothesis

B) A false alarm

C) Rejecting a null hypothesis that is true

D) Using a test that does not suit the data, such as a normal test on a sample too small for it

Module 9 — Regression and Its Uncertainty

Fitting a relationship, and knowing how far to trust it

Recommended Reading / Reference

Book: Joshua Angrist and Jörn-Steffen Pischke, *Mastering 'Metrics: The Path from Cause to Effect* (2014) — how economists use regression and natural experiments to separate cause from correlation, told through real studies

Reference: U.S. Energy Information Administration (EIA), Hourly Electric Grid Monitor (eia.gov/electricity/gridmonitor) — hourly electricity demand for each of the regional operators that balance the U.S. grid, ready to regress on temperatures from weather records

9a. Fitting a Line

Regression describes how one quantity moves with another. Table 9.1 shows ten summer days for a utility: each day's high temperature, in degrees Fahrenheit (°F), and its peak electricity demand, in gigawatts (GW). Hot days bring high demand, because air conditioners work harder.

High, °F	78	80	83	86	88	89	91	93	95	97
Peak, GW	21.1	22.0	22.6	24.8	25.1	24.5	25.9	27.9	26.9	29.2

Table 9.1. Ten summer days: the high temperature and the peak demand.

The model is a straight line plus an error, $y = a + bx + e$. Here x is the input, the temperature; y is the outcome, the demand; a is the intercept, the line's height at $x = 0$; b is the slope, the change in y for each unit of x; and e is the error, everything else that moves demand.

Least squares chooses a and b to make the sum of the squared residuals, the vertical distances from the points to the line, as small as possible. For data $(x_1, y_1), \ldots, (x_n, y_n)$ with averages $\bar{x}$ and $\bar{y}$, the answer is

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad a = \bar{y} - b\,\bar{x}$$

Worked example — Demand and temperature

The temperatures average $\bar{x} = 88$ and the demands $\bar{y} = 25.0$. The squared temperature deviations sum to 358, and the products of the deviations to 143.2. So $b = 143.2/358 = 0.40$ GW per °F, and $a = 25.0 - 0.40 \times 88 = -10.2$. The fitted line is $\hat{y} = -10.2 + 0.40x$, where $\hat{y}$ (y-hat) is the fitted demand: each extra degree adds about 0.4 GW, or 400 megawatts, to the peak.

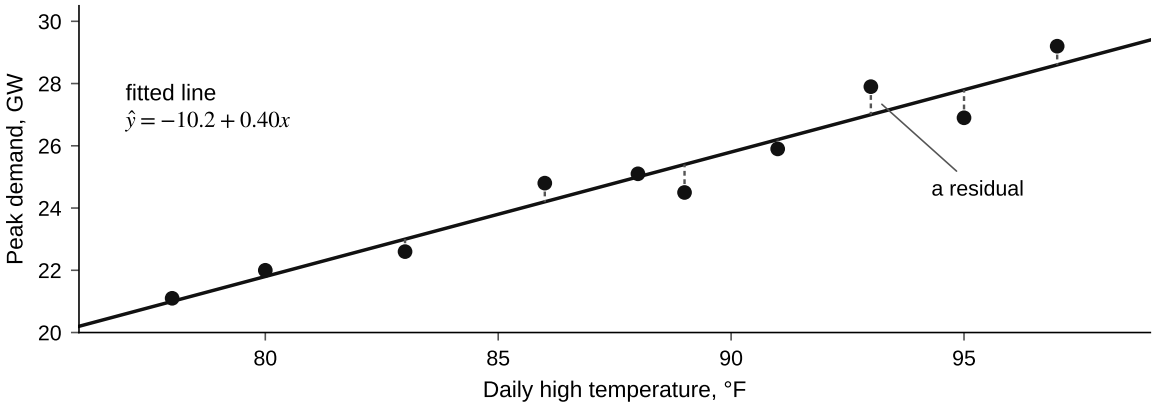


Figure 9.1. The ten days of Table 9.1 with the least-squares line. The dashed segments are the residuals, whose squares the line makes as small as possible in total.

The intercept, −10.2 GW at 0°F, is not a real prediction: no data come near freezing, and nothing says the line holds there. It only positions the line.

The strength of a straight-line relationship is measured by the correlation r, which runs from −1 to 1. Its square, R^2, is the share of the variation in y that the line accounts for. Here $r = 0.97$ and $R^2 = 0.94$: temperature accounts for 94% of the day-to-day variation in peak demand. The slope and the correlation are linked by $b = r\,s_y/s_x$, where s_x and s_y are the standard deviations of the inputs and the outcomes.

The word regression comes from Francis Galton, who found in 1886 that the children of very tall parents were tall, but on average less extremely so. Their heights “regressed” toward the average. The same pull appears wherever two measurements are imperfectly correlated: this year's top performers, in sport or business, tend to do less well next year.

9b. How Sure Is the Slope?

Ten other summer days would give a somewhat different line. The noise around the line is measured by the standard deviation of the residuals $e_i = y_i - \hat{y}_i$, with $n - 2$ in the divisor because two quantities, a and b, were estimated:

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

The slope's standard error, and the t-statistic for the null hypothesis that the true slope is zero, follow:

$$\mathrm{SE}(b) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \qquad t = \frac{b}{\mathrm{SE}(b)}$$

An interval for the true slope is b plus or minus t_{n-2} standard errors, where t_{n-2} is Student's multiplier of Module 7 with $n - 2$ degrees of freedom.

Worked example — How sure is 0.40?

The residuals are 0.1, 0.2, −0.4, 0.6, 0.1, −0.9, −0.3, 0.9, −0.9, and 0.6 GW, and their squares sum to 3.46. So $s = \sqrt{(3.46/8)} \approx 0.66$ GW, and the slope's standard error is $0.66/\sqrt{358} \approx 0.0349$. The t-statistic is $0.40/0.0349 \approx 11.5$, far beyond any threshold. With the multiplier 2.306 for 8 degrees of freedom, the 95% interval for the slope is 0.32 to 0.48 GW per °F.

Regression software reports these columns for every coefficient: the estimate, its standard error, the t-statistic, and the p-value. The standard errors can be trusted only if the errors are independent and similar in size from point to point; section 9e shows how to check.

9c. Forecasts and Their Uncertainty

A forecast is the line's value at a new input x_0, namely $\hat{y}_0 = a + bx_0$. It is uncertain for two reasons: the line itself is estimated, and a single day has its own error e. The standard error of a forecast for a single new day is

$$s_{\text{new}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

The 1 stands for the day's own noise, and it dominates. The other two terms are the line's uncertainty, which grows with the distance of x_0 from $\bar{x}$.

Worked example — A 92-degree day

The forecast is $-10.2 + 0.40 \times 92 = 26.6$ GW. Its standard error is $0.66 \times \sqrt{(1 + 0.1 + 16/358)} \approx 0.71$, so the 95% prediction interval is $26.6 \pm 2.306 \times 0.71$, about 25.0 to 28.2 GW. A utility arranging reserves for that day would plan for as much as 28.2 GW, not 26.6.

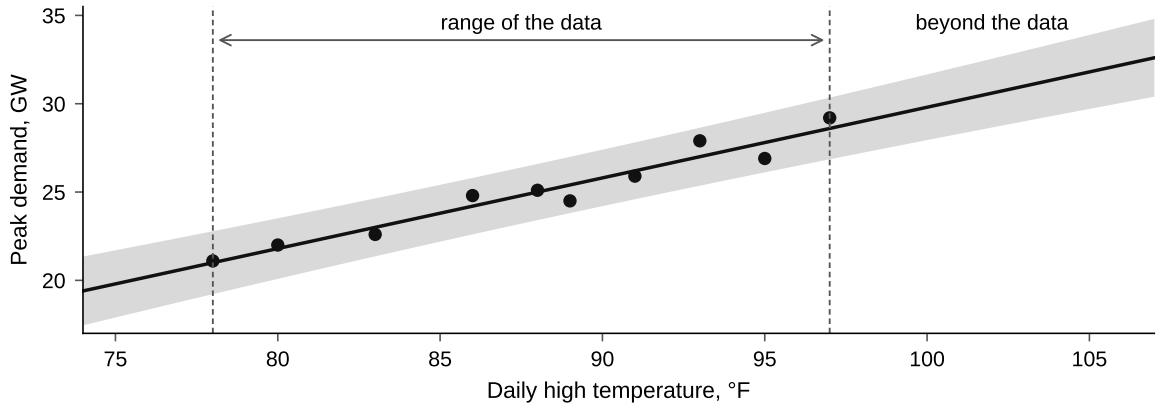


Figure 9.2. The fitted line with its 95% prediction interval for a single day, shaded. The interval widens away from the average temperature of 88°F; beyond the hottest day in the data, it rests on the assumption that the line still holds.

At 105°F, hotter than any day in the data, the formula gives 31.8 GW, with a 95% prediction interval of plus or minus 2.1 GW. But the formula assumes the line continues, and nothing in the data says so: in extreme heat, air conditioners may run flat out while demand levels off, or failing equipment may change everything. Extrapolation, forecasting beyond the range of the data, rests on assumptions the data cannot check.

9d. Several Inputs at Once

Demand depends on more than temperature: offices and factories add load on weekdays. Multiple regression fits several inputs $x_1, \ldots, x_k$ at once,

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

choosing the coefficients by least squares as before. Each coefficient b_j is the change in y for a unit change in x_j with the other inputs held fixed. An input can be an indicator, equal to 1 or 0, such as 1 on weekdays and 0 at weekends.

Worked example — A summer of data

A fit to 92 summer days, with the temperature as x_1 and a weekday indicator as x_2, gives $\hat{y} = -10.4 + 0.40 x_1 + 1.48 x_2$. The weekday coefficient has a standard error of 0.15: at the same temperature, a weekday adds about 1.5 GW to the peak.

Logarithms turn slopes into percentages. If the logarithm of demand is regressed on the logarithm of price, the slope is an elasticity: a slope of −0.25 means that a 10% rise in price cuts demand by about 2.5%.

Leaving out an input that matters changes the meaning of the others when it moves with them. If hot days fell mostly at weekends, a regression on temperature alone would blend the effect of heat with the effect of weekends. This omitted-variable bias is the statistical form of the problem in section 9f. And when two inputs move together, such as temperature and humidity, the data cannot cleanly separate their effects, so both coefficients get large standard errors even when the fit as a whole is good.

9e. Checking the Residuals

Every regression rests on assumptions: the relationship is a straight line, the errors are independent, and their spread is the same everywhere. The residuals are the evidence. Plotted against the fitted values, they should look like random scatter around zero.

A curve in the residuals means the relationship bends. A fan, spreading as the fitted value grows, means the noise grows with the level. A run of residuals of the same sign over time means the errors are linked from one day to the next, and the standard errors are too small. A single point far from the rest can drag the line toward itself.

In 1973 the statistician Francis Anscombe published four small data sets with the same averages, the same standard deviations, the same correlation of 0.82, and the same least-squares line, $\hat{y} = 3 + 0.5x$. Plotted, they could hardly be more different.

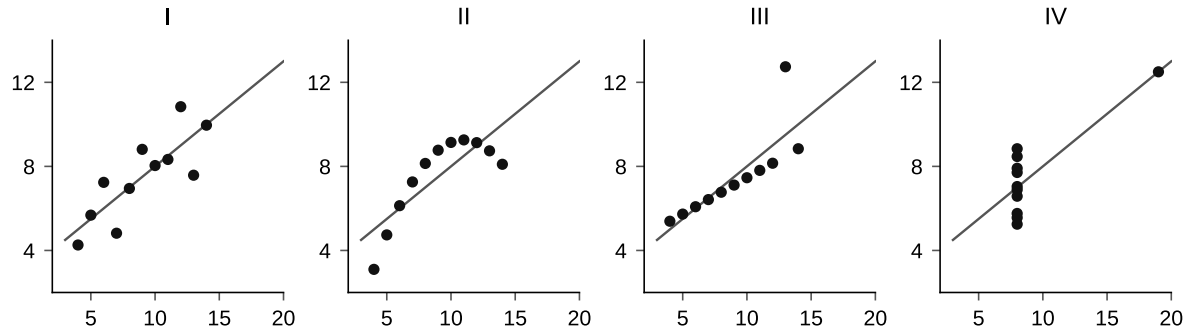


Figure 9.3. Anscombe's quartet: four data sets with the same summary statistics and the same fitted line. Only the first suits a straight line; the second is a curve, the third has an outlier, and in the fourth one point alone sets the slope.

The lesson is the oldest in data analysis: look at the data, and at the residuals, before trusting a summary.

9f. Correlation Is Not Causation

A regression coefficient measures association. Whether it measures a cause depends on how the data arose. Three things produce correlation without cause: a confounder that drives both quantities, causation that runs the other way, and selection of which cases are seen.

Worked example — Smart thermostats

Suppose homes with smart thermostats use 8% less electricity than homes without them. Their owners differ in other ways: they may be wealthier, live in newer and better-insulated homes, or care more about energy. The 8% mixes the device's effect with all of these. A utility that gives thermostats to a randomly chosen half of its customers, and compares them with the other half, measures the device alone.

Confounding can even reverse a comparison. In 1973, 44% of the men who applied for graduate study at the University of California, Berkeley, were admitted, against 35% of the women. Yet department by department the gap vanished: about as many departments favored women as favored men, and in four of the six largest, women were admitted at higher rates. Women had applied mostly to the departments that admitted few applicants of either sex, so the overall comparison misled.

Randomized experiments, like the website tests of Module 8, settle causation because chance alone decides who is treated. When an experiment is impossible, economists look for natural experiments, in which something close to chance decided. Angrist and Pischke, this module's reading, call their main tools the “furious five”: randomized trials, regression, instrumental variables, regression discontinuity designs, and differences-in-differences.

In a Spreadsheet

Spreadsheets fit and judge lines directly. With inputs in A2:A11 and outcomes in B2:B11:

To compute	Formula
Slope and intercept	=SLOPE(B2:B11, A2:A11) and =INTERCEPT(B2:B11, A2:A11)
Correlation r and R^2	=CORREL(A2:A11, B2:B11) and =RSQ(B2:B11, A2:A11)
Standard deviation of the residuals, s	=STEYX(B2:B11, A2:A11)
Forecast at a new input x_0	=FORECAST.LINEAR(x0, B2:B11, A2:A11)
Coefficients with their standard errors	=LINEST(B2:B11, A2:A11, TRUE, TRUE), filling 5 rows by 2 columns; for several inputs, give it a block of input columns
The t multiplier for n points	=T.INV.2T(0.05, n-2)

9g. Memorize This

Commit these to memory. The practice problems, and the modules after this one, assume them.

Least squares	$b = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \qquad a = \bar{y} - b\bar{x}$ The line $\hat{y} = a + bx$ that minimizes the sum of squared residuals, for data (x_i, y_i) with averages $\bar{x}$ and $\bar{y}$
Slope and correlation	$b = r\dfrac{s_y}{s_x}$ r is the correlation and s_x, s_y are the standard deviations; $R^2 = r^2$ is the share of the variation in y that the line accounts for
Residual spread	$s = \sqrt{\dfrac{\sum e_i^2}{n-2}}$ The standard deviation of the residuals $e_i = y_i - \hat{y}_i$, the vertical distances from the n points to the line
Slope's uncertainty	$\mathrm{SE}(b) = \dfrac{s}{\sqrt{\sum(x_i - \bar{x})^2}}, \qquad t = \dfrac{b}{\mathrm{SE}(b)}$ The interval for the true slope is b plus or minus t_{n-2} standard errors (SE), with $n - 2$ degrees of freedom
Forecast for one new case	$s\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$ The standard error of a forecast at input x_0; it grows away from $\bar{x}$, and beyond the data the line itself is an assumption
Several inputs	Each coefficient b_j is the change in y per unit of x_j with the other inputs held fixed; leaving out an input that matters biases the others whenever it moves with them
Correlation and cause	Confounders, reverse causation, and selection all produce correlation without cause; randomized experiments, or natural ones, separate them
The idea	A fitted line is an estimate like any other, with standard errors for its slope and wider ones for its forecasts. Check the residuals, stay within the data, and do not read cause into a correlation.

9h. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

9.1 On the line $y = 2 + 0.5x$, find y at $x = 10$ and at $x = 12$. What does the slope mean?

9.2 Data have $\bar{x} = 5$, $\bar{y} = 12$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 40$, and $\Sigma(x_i - \bar{x})^2 = 20$. Find the least-squares line.

9.3 Two quantities have a correlation of 0.6. What share of the variation in one does a line on the other account for?

9.4 From 30 points, a slope is 1.5 with a standard error of 0.5. Find its t-statistic and a 95% interval.

9.5 The line $\hat{y} = 2 + 2x$ is fit to data that include the point (6, 15). Find the point's fitted value and residual.

9.6 The correlation of two quantities is 0.8, $s_x = 4$, and $s_y = 10$. Find the slope of the least-squares line of y on x.

Applied

9.7 With the fitted line of section 9a, forecast the peak on a 90°F day and on a 95°F day. How much do the extra 5 degrees add?

9.8 Using $s = 0.66$ and the data of section 9a, find the 95% prediction interval for a single 88°F day.

9.9 In the regression of section 9d, what does the coefficient 1.48 on the weekday indicator mean? Is it clearly different from zero?

9.10 A regression of home prices on floor area finds a slope of $150 per square foot, with a standard error of $20, from 200 sales. Give a 95% interval for the slope.

9.11 A regression of the logarithm of gasoline use on the logarithm of its price has a slope of −0.25. If the price rises 20%, by about how much does use change?

9.12 Cities with more police officers per person tend to have more crime. Give two reasons this does not show that police cause crime.

Challenge

9.13 Show that the least-squares line passes through the point $(\bar{x}, \bar{y})$.

9.14 Show that the residuals of a least-squares line sum to zero.

9.15 A school's test scores in two years have a correlation of 0.5 with the same average and spread. A school scored 2 standard deviations above average this year. What should be expected next year, and why?

9.16 The correlation is $r = \Sigma(x_i - \bar{x})(y_i - \bar{y})/((n-1)s_x s_y)$. Show that the least-squares slope equals $r\, s_y/s_x$.

Module 9 Quiz

Check Your Understanding

1. In the line $\hat{y} = -10.2 + 0.40x$, fitted to peak demand in GW and temperature in °F, what does 0.40 mean?

A) Demand at 0°F
B) The correlation
C) Gigawatts added per degree
D) The share of demand explained by heat

2. What does a residual measure?

A) The slope's error
B) The fitted value
C) The share of variance explained
D) How far a point lies above or below the line

3. From 10 points, a slope is 0.40 with a standard error of 0.05. What can be concluded?

A) The slope is certainly between 0.35 and 0.45, with no chance of lying outside that range.
B) The slope is clearly not zero; its t-statistic is 8.
C) Ten points are too few to say anything about the slope.
D) The line explains 40% of the variation in demand.

4. Why is a forecast at 105°F less trustworthy than one at 88°F, when the data run from 78°F to 97°F?

A) Because it extrapolates beyond the data
B) Because the residuals are always larger on hot days, as the spread of demand grows with temperature
C) Because the slope changes sign above 100°F
D) Because R^2, the share of the variation that the line accounts for, is lower at high temperatures

5. Homes with smart thermostats use 8% less electricity than homes without them. Why does that not show that the thermostats cause the savings?

A) The difference is too small to matter.
B) The owners may differ in other ways, such as income or home size, that also affect use.
C) Correlations between two quantities are always due to chance.
D) A difference of 8% is too large to be due to chance, so the thermostats must have caused it.

Module 10 — Fitting and Testing Models

How models, AI models included, are fit to data and judged

Recommended Reading / Reference

Book: Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning* (2nd ed., 2021) — the standard first text on fitting and judging predictive models, free online from its authors

Reference: UCI Machine Learning Repository (archive.ics.uci.edu) — hundreds of free data sets, from credit card defaults to medical records, for practicing the methods of this module

10a. Likelihood

Every model in this course has had parameters: a probability, a mean, a slope. A general way to fit them is maximum likelihood. The likelihood of a parameter value is the probability of the observed data if that value were true, and the maximum likelihood estimate is the value that makes the observed data most probable.

Worked example — Seven renewals in ten

Seven of ten customers renew a contract. If each renews independently with probability p, the chance of the observed sequence of renewals is $L(p) = p^7(1-p)^3$. At $p = 0.5$ it is about 0.00098; at $p = 0.7$ it is about 0.00222, 2.3 times as large, and no value of p does better. The maximum likelihood estimate is 0.7, the observed share.

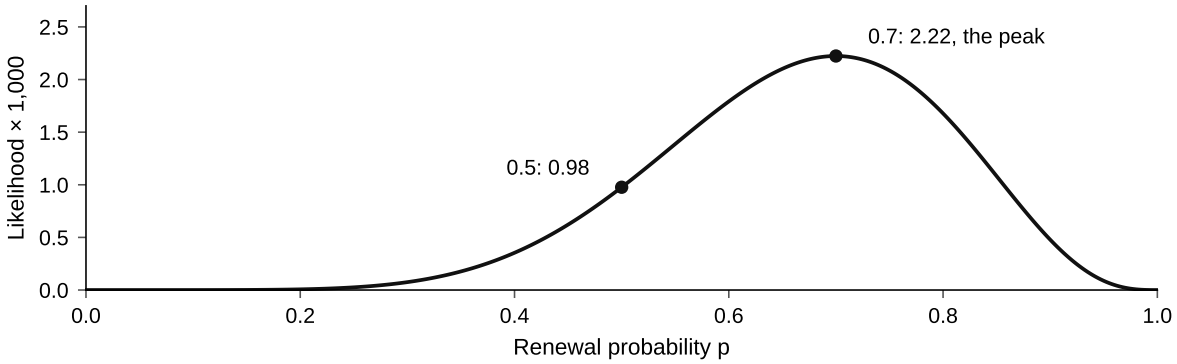


Figure 10.1. The likelihood of seven renewals in ten, for each possible renewal probability. It peaks at 0.7, the maximum likelihood estimate.

Likelihoods are products of many small probabilities, so in practice one maximizes their logarithm, the log-likelihood, which turns products into sums and peaks in the same place. Here it is $7 \ln p + 3 \ln(1-p)$. Calculus gives the slope of this curve as $7/p - 3/(1-p)$, which is zero, so that the curve is flat at its peak, when $p = 0.7$.

Maximum likelihood ties much of this course together. For counts, it estimates the Poisson mean by the average count. For a line with normal errors, it gives exactly the least-squares line of Module 9, since the most probable line is the one with the smallest squared errors. For models with millions of parameters, no one solves for the peak by algebra: computers climb toward it step by step.

10b. Logistic Regression

Many questions have yes-or-no answers: will a loan default, will a customer leave, is an email spam? A straight line for the probability would wander below 0 and above 1. Logistic regression instead makes the log-odds a straight line in the input x, with intercept a and slope b:

$$\ln\frac{p}{1-p} = a + bx, \qquad \text{or equivalently} \qquad p = \frac{1}{1+e^{-(a+bx)}}$$

Here p is the probability of a yes, and $p/(1-p)$ is its odds, as in Module 2. The probability rises along an S-shaped curve from 0 to 1, and each unit of x multiplies the odds by e^b.

Worked example — Loans and debt

A lender's logistic regression of default on x, a borrower's debt payments as a percentage of income, gives $a = -6$ and $b = 0.08$. At 30% the log-odds are $-6 + 2.4 = -3.6$, a default probability of $1/(1+e^{3.6}) \approx 2.7\%$. At 50% they are −2, a probability of 11.9%. Each extra 10 points of debt multiplies the odds of default by $e^{0.8} \approx 2.2$.

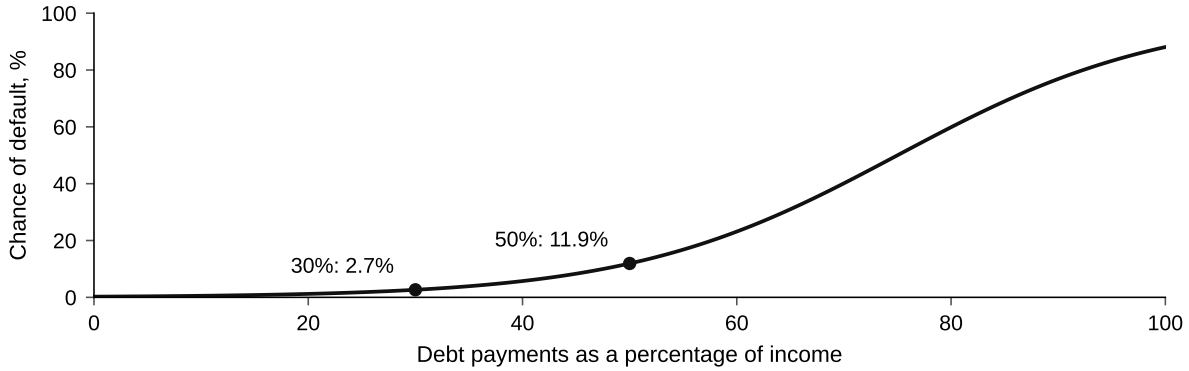


Figure 10.2. The lender's logistic regression. The chance of default stays low for modest debts and then climbs steeply, but it can never pass 100%.

Logistic regression is fit by maximum likelihood, and with several inputs it adds their effects on the log-odds. That is the form of the spam filter of Module 2, whose log-odds were a prior plus one weight for each word. A neural network classifier ends the same way: its last step turns weighted sums into probabilities.

10c. Cross-Entropy: How AI Models Are Scored

To fit a classifier, maximize the likelihood of the true answers. Equivalently, minimize the average of $-\ln q$, where q is the probability the model gave to the answer that turned out to be right. Over N examples, with q_i the probability given to the right answer of example i, this average is the cross-entropy loss:

$$\text{loss} = -\frac{1}{N}\sum_{i=1}^{N} \ln q_i$$

A model that gives the right answer probability 1 has loss 0. One that gives it 0.5 pays 0.69, and one that gives it 0.01 pays 4.6. Confident mistakes are punished hardest.

Large language models are trained this way. At each position in a text, the model gives a probability to every possible next token, a word or a piece of one, and its loss is $-\ln q$ for the token that actually comes next.

Training adjusts billions of parameters to lower the average loss over trillions of tokens: Meta's Llama 3 models, released in 2024, were trained on more than 15 trillion.

The loss is often reported as perplexity, e raised to the power of the average loss. A perplexity of 4 means the model is, on average, as uncertain as if it were choosing among 4 equally likely tokens.

Worked example — Scoring a prediction

A model gives the word that actually comes next a probability of 0.25. Its loss on that word is $-\ln 0.25 \approx 1.39$. Had it given the word 0.9, the loss would be 0.11; had it given 0.01, the loss would be 4.61. A model whose average loss is 1.39 has a perplexity of $e^{1.39} \approx 4$.

10d. Training, Validation, and Test Data

A model's error on the data it was fit to is an optimistic guide to its error on new data. Give a model enough flexibility and it will fit the noise along with the signal, the overfitting of the backtests in Module 8.

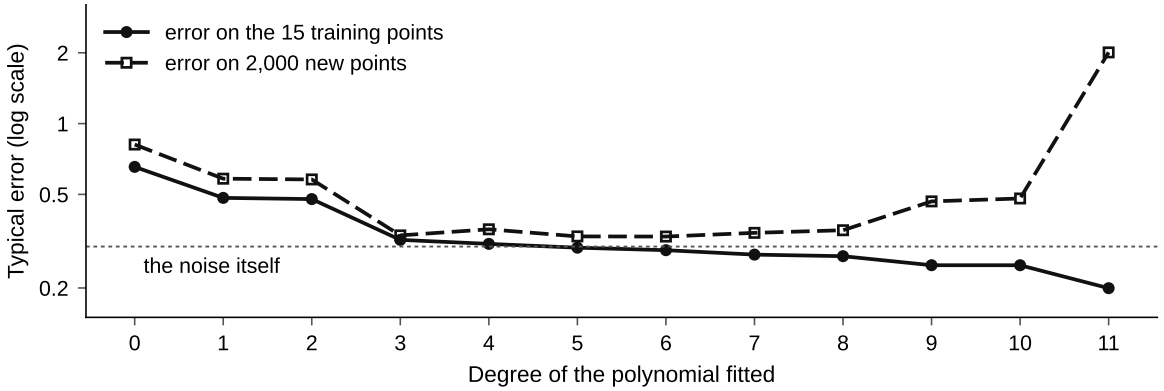


Figure 10.3. Polynomials of rising degree fitted to 15 noisy points. The training error keeps falling, below even the noise in the data; the error on new points is lowest at a moderate degree and then climbs as the curve chases the noise.

Figure 10.3 fits polynomials, curves with more and more adjustable bends, to 15 noisy points. The error on the training points falls steadily. The error on new points falls at first, as the curve captures the real pattern, and then rises as the curve starts to fit the noise.

The remedy is to split the data three ways. The training set fits the model. The validation set chooses among models and settings, such as the polynomial's degree. The test set, held back until the very end and used once, gives an honest estimate of performance on new data. Any decision made by looking at the test set uses it up, because the decision fits the test set's noise.

When data are scarce, cross-validation reuses them: split the data into, say, five parts, fit on four and validate on the fifth, rotate through all five, and average. For data in time order, such as prices or demand, the validation data must come after the training data, or the model will be judged with knowledge of the future.

10e. Judging a Classifier

A classifier that answers yes or no can be wrong in two ways, like the tests of Module 8. Three measures summarize its record. Accuracy is the share of all cases it gets right. Precision is the share of its yes answers that are right. Recall is the share of the actual yes cases that it catches.

Worked example — A fraud detector

Of 10,000 card transactions, 100 are fraud. A model flags 300, of which 80 are fraud. Its recall is $80/100 = 80\%$, and its precision is $80/300 \approx 27\%$. Its accuracy is $(80 + 9{,}680)/10{,}000 = 97.6\%$, since it correctly passes 9,680 of the 9,900 honest transactions. A model that never flags anything scores 99% accuracy and catches no fraud at all.

When the thing sought is rare, accuracy says little, as Bayes' rule warned in Module 2. Most classifiers output a probability, and a threshold turns it into yes or no. Lowering the threshold catches more fraud, raising recall, but flags more honest customers, lowering precision. The right threshold depends on what each kind of mistake costs.

A model's probabilities should also mean what they say. A model is well calibrated if the events it gives 70% happen about 70% of the time, and weather forecasts are judged this way. OpenAI's 2023 technical report on GPT-4 showed that the model, after its initial training to predict text, was well calibrated on multiple-choice questions, and that the further training that made it a chat assistant hurt its calibration significantly.

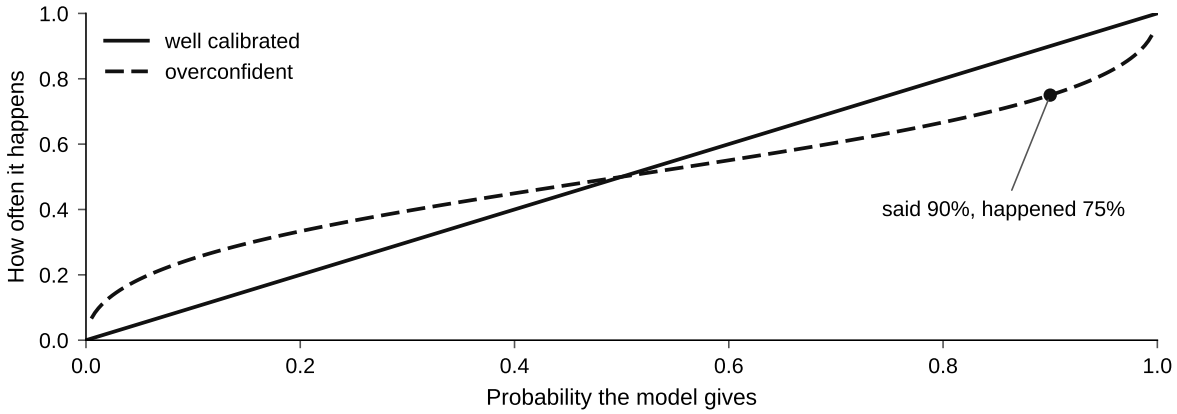


Figure 10.4. Calibration. A well-calibrated model's probabilities match how often events happen; an overconfident one's 90% comes true only 75% of the time, and its 10% a quarter of the time.

Calibration matters whenever a probability feeds a decision: a 90% that comes true only 75% of the time will be bet on too heavily. To check it, group a model's predictions by the probability given, and compare each group's probability with how often its events happened.

In a Spreadsheet

Spreadsheets can do the arithmetic of this module. With counts and probabilities in cells:

To compute	Formula
Log-likelihood of k successes in n trials at probability p	=k*LN(p)+(n-k)*LN(1-p)
Probability from log-odds z	=1/(1+EXP(-z))
Cross-entropy, for the probabilities in A2:A101 given to the right answers	=-SUMPRODUCT(LN(A2:A101))/COUNT(A2:A101)
Perplexity from an average loss in B1	=EXP(B1)
Precision and recall, with true positives in C1, false positives in C2, and false negatives in C3	=C1/(C1+C2) and =C1/(C1+C3)
Fit a logistic regression	Solver: maximize the summed log-likelihood over the cells that hold a and b

10f. Memorize This

Commit these to memory. The practice problems assume them, and so does every model you will meet after this course.

Maximum likelihood	Choose the parameter values that make the observed data most probable; for a probability, the estimate is the observed share
Logistic regression	$\ln\frac{p}{1-p} = a + bx, \qquad p = \frac{1}{1+e^{-(a+bx)}}$ The log-odds of a yes, with probability p, is a line in the input x; each unit of x multiplies the odds by e^b
Cross-entropy	$\text{loss} = -\frac{1}{N}\sum_{i=1}^{N} \ln q_i$ The average, over N examples, of minus the log of the probability q_i given to the right answer; minimizing it maximizes the likelihood
Perplexity	e raised to the average cross-entropy loss: the number of equally likely choices that would carry the same uncertainty
Three data sets	Training data fit the model, validation data choose among models, and test data, used once at the end, measure performance on new data
Classifier measures	Accuracy: the share of all cases right. Precision: the share of yes answers that are right. Recall: the share of actual yes cases caught
Calibration	Events given a probability of 70% should happen about 70% of the time; check by grouping predictions and comparing with outcomes
The idea	Fit a model by making the data probable, judge it on data it has never seen, and score its probabilities honestly. Every modern AI model is fit and judged this way.

10g. Practice Problems

Drill problems build fluency, applied problems put the module to work on real quantities, and challenge problems stretch it. Worked solutions to every problem are at the back of the course.

Drill

10.1 Eighteen of 60 customers cancel their service. If each cancels independently with probability p, what is the maximum likelihood estimate of p?

10.2 Convert log-odds of 0, 2, and −3 into probabilities.

10.3 An event has probability 0.2. Find its odds and its log-odds.

10.4 A classifier gives the right answer probabilities of 0.5, 0.9, and 0.01 on three examples. Find its loss on each, and the average.

10.5 A language model's average loss is 2.0. What is its perplexity?

10.6 A model makes 50 true positives, 10 false positives, 25 false negatives, and 915 true negatives. Find its accuracy, precision, and recall.

Applied

10.7 For the lender of section 10b, find the chance of default at 40% debt-to-income, and the factor by which the odds change between 30% and 60%.

10.8 A feeder's outages over six months number 2, 0, 3, 1, 1, and 5. If outages are Poisson with mean λ a month, what is the maximum likelihood estimate of λ? Under it, what is the chance of a month with no outage?

10.9 Compare the fraud detector of section 10e with a model that flags every transaction. Find both models' precision, recall, and accuracy.

10.10 Lowering the fraud detector's threshold makes it flag 1,000 transactions, catching 95 of the 100 frauds. Find its new precision and recall. When would the change be worth it?

10.11 A forecaster gave a 30% chance of rain on 200 days, and it rained on 46 of them. Is the forecaster well calibrated at 30%?

10.12 A model's error rate is 2% on its training data and 9% on its validation data. What is likely happening, and what could be done?

Challenge

10.13 Show that the maximum likelihood estimate of p from k successes in n independent trials is k/n.

10.14 Show that for a line with normal errors of fixed standard deviation σ, maximizing the likelihood is the same as least squares.

10.15 Show that the probability $p = 1/(1 + e^{-z})$ has log-odds z.

10.16 A team tries 50 versions of a model and reports the one that scores best on the test set. Why is that score too optimistic, and by roughly how much, if all 50 are in fact equally good and each score has a standard error of 1 point?

Module 10 Quiz

Check Your Understanding

1. What is the maximum likelihood estimate of a parameter?

A) It is the value most likely to be true, given the prior and the data together.
B) It is the value that makes the sum of the residuals exactly equal to zero.
C) It is the average of all the values that the parameter could possibly take.
D) It is the value that makes the observed data most probable.

2. A language model gives the word that actually comes next a probability of 0.25. What is its cross-entropy loss on that word?

A) 0.25
B) About 4, its perplexity
C) About 1.39
D) About 0.75, the chance of a miss

3. Why is a test set kept apart until the very end?

A) Any choice made by looking at it would fit its noise, so it would no longer give an honest measure of performance on new data.
B) Test data are usually of lower quality than training data.
C) Models train faster when they see less data.
D) It is needed to compute the model's gradients.

4. A fraud model flags 300 of 10,000 transactions and catches 80 of the 100 frauds. What is its precision?

A) It is 80%, since it catches 80 of the 100 frauds.
B) It is 97.6%, the share of all the transactions that it labels correctly.
C) It is 3%, the share of all transactions that it flags as fraud.
D) It is about 27%, since 80 of its 300 flags are frauds.

5. What does it mean for a model's probabilities to be well calibrated?

A) It is right most of the time
B) Its probabilities are close to 0 or 1, so it is always confident, which makes it more useful than a model that hedges between the two
C) Events it gives 70% happen about 70% of the time
D) Its probabilities add up to 100%

Everything to Memorize

The Memorize This boxes from every module, in one place

Module 1: Probability and Its Rules

Probability	A number from 0 to 1; with equally likely outcomes, the number of outcomes in the event divided by the total
Complement	$P(A^c) = 1 - P(A)$ A^c is the event that A does not happen
Addition rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $A \cup B$ means A or B or both, and $A \cap B$ means both; for mutually exclusive events the last term is zero
Independence	$P(A \cap B) = P(A)\,P(B)$ Holds when knowing that one event happened does not change the chance of the other
At least one	$1 - (1 - p)^n$ The chance that at least one of n independent events, each with probability p, happens
Counting	$C(n, k) = \dfrac{n!}{k!\,(n-k)!}$ Ways to choose k of n items when order does not matter; $n! = n \times (n-1) \times \ldots \times 1$ is the number of orders of n items
Series and parallel	$p_1 p_2 \cdots p_n$ and $1 - (1 - p_1) \cdots (1 - p_n)$ The chance that a system works when every component must work, and when any one suffices; p_i is the chance that component i works
Common-mode failure	A shared cause, such as a storm, that makes failures happen together and breaks the independence redundancy relies on
The idea	Probabilities combine by a few rules: subtract from 1 for not, add and remove the overlap for or, and multiply for and when events are independent. Many small chances add up fast, and independence is an assumption to check.

Module 2: Conditional Probability and Bayes' Rule

Conditional probability	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ The chance of A among the cases where B happened; $P(A\|B)$ is usually not $P(B\|A)$
Total probability	$P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$ The overall chance of B, adding its chance when A happens and when it does not
Bayes' rule	$P(A \mid B) = \dfrac{P(B \mid A)\,P(A)}{P(B)}$ Turns the chance of the evidence given the cause into the chance of the cause given the evidence; $P(A)$ is the prior and $P(A\|B)$ the posterior
Odds	$\text{odds} = \dfrac{p}{1-p}, \qquad p = \dfrac{\text{odds}}{1 + \text{odds}}$ A probability p of 20% is odds of 1 to 4
Bayes in odds	$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}$ The likelihood ratio of evidence B is $P(B\|A)/P(B\|A^c)$; for pieces of evidence that are independent once it is known whether A happened, the ratios multiply

Base rates	A rare condition and an imperfect test give mostly false alarms: at 1% prevalence, 90% sensitivity, and 5% false positives, only about 15% of positives are real
Naive Bayes	The posterior odds are the prior odds times each feature's likelihood ratio, treating the features as independent; in logs, a weighted sum
The idea	Evidence updates a probability by its likelihood ratio, but the starting point matters: when the prior is small, even strong evidence can leave the posterior modest.

Module 3: Random Variables and Expected Value

Random variable	A number set by chance; its distribution lists the possible values x_i and their probabilities p_i, which add to 1
Expected value	$E[X] = \mu = \sum_i x_i\, p_i$ The probability-weighted average, and the long-run average over many repetitions; μ (mu) is the mean
Linearity	$E[aX + b] = aE[X] + b, \qquad E[X + Y] = E[X] + E[Y]$ a and b are constants; the second rule holds even when X and Y are dependent
Variance	$\mathrm{Var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2$ The standard deviation σ is its square root, in the units of X; scaling X by a scales the variance by a^2
Variance of a sum	$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$ The covariance, the average product of the deviations from the means, is zero for independent variables, whose variances simply add
Value of information	The expected outcome when deciding after the uncertainty is resolved, minus the expected outcome when deciding before it: the most a perfect forecast is worth
Typical growth	$g = \sum_i p_i \ln(1 + r_i)$ The long-run growth rate of a single compounding path, where r_i are the possible returns, p_i their probabilities, and ln the natural logarithm; it is below the expected return and can be negative when that is positive
Kelly criterion	For a bet at even odds won with probability p, betting the fraction $2p - 1$ of wealth each time maximizes long-run growth
The idea	The expected value is the long-run average and adds up simply, but risk lives in the spread and the tails. For a single compounding path, the typical outcome can fall short of the average, and even shrink.

Module 4: Counting Events: Binomial and Poisson

Binomial distribution	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ The chance of k successes in n independent trials, each a success with probability p
Binomial mean and variance	$E[X] = np, \qquad \mathrm{Var}(X) = np(1-p)$ A success rate measured from n trials varies by about $\sqrt{(p(1-p)/n)}$
Poisson distribution	$P(X = k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ Counts of independent events at an average rate of λ per period; the mean and the variance are both λ, and $e \approx 2.718$
Poisson approximation	For many trials, each with a small chance of success, the binomial is close to the Poisson with $\lambda = np$
Relative steadiness	A Poisson count's standard deviation is $\sqrt{\lambda}$, so larger books vary less in proportion: about 10% at $\lambda = 100$, 1% at $\lambda = 10{,}000$
Dependence	When events share a cause, the average can stay the same while the chance of many events at once grows enormously
The idea	Counts of independent events follow the binomial or the Poisson, and they grow steadier in proportion as they grow larger. Dependence, not the average rate, decides how bad the worst years are.

Module 5: Continuous Distributions and Fat Tails

Density	$P(a \le X \le b) = \int_a^b f(x)\,dx$ For a continuous variable, probability is area under the density $f(x)$, the total area is 1, and any single value has probability 0
Normal landmarks	About 68% of values within one standard deviation σ of the mean μ, 95% within two, and 99.7% within three
z-score	$z = \dfrac{x - \mu}{\sigma}$ The distance from the mean in standard deviations; the standard normal turns it into a probability
Lognormal	$\text{median} = e^m, \qquad \text{mean} = e^{m + s^2/2}$ For a price ratio whose log is normal with mean m and standard deviation s; never negative, skewed right
Exponential	$P(T > t) = e^{-\lambda t}, \qquad E[T] = 1/\lambda$ The wait T for random events at rate λ; memoryless, meaning a constant failure rate
Fat tails	More probability far from the mean than the normal allows; a 4-standard-deviation move is a once-in-63-years day under the normal, but routine in markets
VaR and expected shortfall	The 99% VaR is the loss exceeded 1% of the time; the expected shortfall is the average loss in that worst 1%; for a normal, 2.33 and 2.67 standard deviations
The idea	Continuous probabilities are areas. The normal is the right model for sums of many small effects and the wrong one for the extremes of markets and catastrophes, where tails are fat.

Module 6: Sums, Averages, and the Central Limit Theorem

Sums and averages	$\text{sum: } n\mu,\ \sigma\sqrt{n} \qquad \text{average: } \mu,\ \sigma/\sqrt{n}$ Mean and standard deviation for n independent draws, each with mean μ and standard deviation σ
Standard error	$\text{SE} = \dfrac{\sigma}{\sqrt{n}}$ The standard deviation of an average; halving it takes four times the data
Law of large numbers	The average of many independent draws approaches the mean; it does not make outcomes balance out in the short run
Central limit theorem	The sum or average of many independent draws is approximately normal, whatever their distribution, if its standard deviation is finite
Monte Carlo	Simulate the inputs many times and read off the distribution of the result; its error shrinks like one over the square root of the number of runs
Correlated averages	$\dfrac{\sigma^2}{n} + \left(1 - \dfrac{1}{n}\right)\rho\,\sigma^2 \ \to\ \rho\,\sigma^2$ The variance of an average of n draws with pairwise correlation ρ: shared risk does not average away
The idea	Independent pieces cancel each other's noise, so totals are steady in proportion and averages are nearly normal. Shared causes and huge single outcomes are where that comfort ends.

Module 7: Estimation and Confidence Intervals

Two kinds of error	Random error comes from which sample was drawn and shrinks like $1/\sqrt{n}$; bias is systematic, and more data does not remove it
Sample standard deviation	$s = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ The spread of data $x_1, \ldots, x_n$ around their average $\bar{x}$; dividing by $n-1$ corrects for measuring from $\bar{x}$
Standard errors	$\text{SE}(\bar{x}) = \dfrac{s}{\sqrt{n}}, \qquad \text{SE}(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ For an average $\bar{x}$ of n values with standard deviation s, and for a share $\hat{p}$ of n people
Confidence interval	$\text{estimate} \pm z \times \text{SE}$ The multiplier z is 1.645 for 90% confidence, 1.96 for 95%, and 2.576 for 99%; that share of intervals made this way contains the true value
Poll margin	$1.96\sqrt{0.25/n} \approx 1/\sqrt{n}$ The 95% margin of error for a share near 50% from n people: about 3 points for 1,000; the gap between two candidates has nearly twice the margin
Small samples	$\bar{x} \pm t_{n-1}\, s/\sqrt{n}$ Student's t multiplier t_{n-1}, with $n-1$ degrees of freedom: 2.57 for 6 values, 2.26 for 10, near 1.96 beyond 100
Bootstrap	Resample the data with replacement many times, recompute the estimate each time, and take the middle 95% of the results as the interval

The idea	An estimate is worth only as much as its error bar. Random error shrinks with more data; bias does not, and no margin of error includes it.

Module 8: Hypothesis Tests

Test statistic	$z = \dfrac{\text{estimate} - \text{null value}}{\text{SE}}$ How far the data fall from what the null hypothesis, the claim of no effect, predicts, in standard errors (SE)
p-value	The chance, if the null hypothesis were true, of a result at least as far from it as the one observed; below 0.05 is called significant by convention
What it is not	Not the chance that the null hypothesis is true, not the size or importance of an effect, and not the chance that a result will replicate
Two errors	Type I: a false alarm, rejecting a true null, with chance α, the significance level. Type II: a miss, with chance β. Power is $1 - \beta$
Two rates	$z = \dfrac{\hat{p}_B - \hat{p}_A}{\sqrt{\bar{p}(1-\bar{p})(1/n_A + 1/n_B)}}$ Rates $\hat{p}_A$ and $\hat{p}_B$ from n_A and n_B visitors; $\bar{p}$ is the pooled rate, all conversions over all visitors
Sample size for 80% power	$n = \dfrac{7.84\,[p_A(1-p_A) + p_B(1-p_B)]}{(p_B - p_A)^2}$ Per group, at the 5% level, to detect true rates p_A and p_B; 7.84 is $(1.96 + 0.84)^2$
Many tests	$1 - 0.95^k$ The chance of at least one false alarm in k independent tests of effects that do not exist; the Bonferroni correction tests each at $0.05/k$
The idea	A test asks whether chance alone could plausibly explain the data. Its answer is only as honest as the count of tests behind it, and it says nothing by itself about how large or important an effect is.

Module 9: Regression and Its Uncertainty

Least squares	$b = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \qquad a = \bar{y} - b\bar{x}$ The line $\hat{y} = a + bx$ that minimizes the sum of squared residuals, for data (x_i, y_i) with averages $\bar{x}$ and $\bar{y}$
Slope and correlation	$b = r\dfrac{s_y}{s_x}$ r is the correlation and s_x, s_y are the standard deviations; $R^2 = r^2$ is the share of the variation in y that the line accounts for
Residual spread	$s = \sqrt{\dfrac{\sum e_i^2}{n-2}}$ The standard deviation of the residuals $e_i = y_i - \hat{y}_i$, the vertical distances from the n points to the line
Slope's uncertainty	$\text{SE}(b) = \dfrac{s}{\sqrt{\sum(x_i - \bar{x})^2}}, \qquad t = \dfrac{b}{\text{SE}(b)}$ The interval for the true slope is b plus or minus t_{n-2} standard errors (SE), with $n - 2$ degrees of freedom

Forecast for one new case	$s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$ The standard error of a forecast at input x_0; it grows away from $\bar{x}$, and beyond the data the line itself is an assumption
Several inputs	Each coefficient b_j is the change in y per unit of x_j with the other inputs held fixed; leaving out an input that matters biases the others whenever it moves with them
Correlation and cause	Confounders, reverse causation, and selection all produce correlation without cause; randomized experiments, or natural ones, separate them
The idea	A fitted line is an estimate like any other, with standard errors for its slope and wider ones for its forecasts. Check the residuals, stay within the data, and do not read cause into a correlation.

Module 10: Fitting and Testing Models

Maximum likelihood	Choose the parameter values that make the observed data most probable; for a probability, the estimate is the observed share
Logistic regression	$\ln\frac{p}{1-p}=a+bx, \qquad p=\frac{1}{1+e^{-(a+bx)}}$ The log-odds of a yes, with probability p, is a line in the input x; each unit of x multiplies the odds by e^b
Cross-entropy	$\text{loss}=-\frac{1}{N}\sum_{i=1}^{N}\ln q_i$ The average, over N examples, of minus the log of the probability q_i given to the right answer; minimizing it maximizes the likelihood
Perplexity	e raised to the average cross-entropy loss: the number of equally likely choices that would carry the same uncertainty
Three data sets	Training data fit the model, validation data choose among models, and test data, used once at the end, measure performance on new data
Classifier measures	Accuracy: the share of all cases right. Precision: the share of yes answers that are right. Recall: the share of actual yes cases caught
Calibration	Events given a probability of 70% should happen about 70% of the time; check by grouping predictions and comparing with outcomes
The idea	Fit a model by making the data probable, judge it on data it has never seen, and score its probabilities honestly. Every modern AI model is fit and judged this way.

Worked Solutions

Every practice problem, solved step by step

Module 1: Probability and Its Rules

1.1 There are 13 hearts and 4 kings, and one card is both, so the chance is $(13 + 4 - 1)/52 = 16/52 \approx 0.308$.

1.2 The chance of no six in four rolls is $(5/6)^4 \approx 0.482$, so the chance of at least one is about 0.518, slightly better than even.

1.3 $C(10, 3) = 120$ choices when order does not matter, and $10 \times 9 \times 8 = 720$ when it does.

1.4 Both: $0.3 \times 0.5 = 0.15$. At least one: $0.3 + 0.5 - 0.15 = 0.65$. Neither: $1 - 0.65 = 0.35$, which is also 0.7×0.5.

1.5 Both fail to start with probability $0.05^2 = 0.0025$, so at least one starts with probability 0.9975.

1.6 Independence would require the chance of both to be $0.40 \times 0.25 = 0.10$. The survey found 0.15, so they are not independent: the two tend to go together.

1.7 All running: $0.98^{50} \approx 0.364$. At least one out: about 0.636. On nearly two days in three, a wind farm of this size has at least one turbine down.

1.8 $1 - 0.99^{20} \approx 0.182$. If defaults tend to come together, as they do in recessions, the chance of at least one default is lower than this, but the chance of several at once is much higher.

1.9 In series: $0.995^4 \approx 0.980$. Duplicating a stage makes it fail only if both copies fail, so it works with probability $1 - 0.005^2 = 0.999975$, and the pipeline with $0.999975^4 \approx 0.9999$.

1.10 There are $C(49, 6) = 13{,}983{,}816$ possible choices, so the chance is 1 in about 14 million.

1.11 All three up: $0.99^3 \approx 0.970299$. Exactly one out: three choices of which, each with probability 0.01×0.99^2, for 0.029403 in all. Together: about 0.9997.

1.12 Spam gets through only if both checks miss it: $0.1 \times 0.1 = 0.01$, or 1%.

1.13 $A \cup B$ is made of three separate pieces: A without B, B without A, and $A \cap B$. $P(A)$ covers the first and third pieces and $P(B)$ the second and third, so $P(A) + P(B)$ counts the third twice. Subtracting $P(A \cap B)$ once leaves each piece counted once.

1.14 $P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c)$.

1.15 There are 365^{23} equally likely ways to assign birthdays. For all to differ, the first person has 365 choices, the second 364, and so on to the 23rd, with 343. Dividing gives the chance, about 0.493.

1.16 Both fail with probability $0.02 \times 0.2^2 + 0.98 \times 0.005^2 \approx 0.000825$. Each generator alone fails on $0.02 \times 0.2 + 0.98 \times 0.005 = 0.0089$ of days.

Treating them as independent gives $0.0089^2 \approx 0.000079$, about ten times too low. The storms, a common cause, make failures arrive together.

Module 2: Conditional Probability and Bayes' Rule

2.1 Of 38 defaults, 20 had a late payment: $20/38 \approx 0.526$.

2.2 $P(A|B) = 0.12/0.4 = 0.3$, which equals $P(A)$, so they are independent. Equivalently, $0.3 \times 0.4 = 0.12$.

2.3 0.2 is odds of $0.2/0.8 = 0.25$, or 1 to 4. Odds of 3 to 1 are a probability of $3/4 = 0.75$.

2.4 $0.99 \times 0.001 = 0.00099$ of people are true positives and $0.01 \times 0.999 = 0.00999$ false ones, so the chance is $0.00099/0.01098 \approx 9\%$.

2.5 The likelihood ratio is $0.99/0.01 = 99$, and the prior odds are 1 to 999, so the posterior odds are 99 to 999, a probability of $99/1{,}098 \approx 9\%$.

2.6 By total probability: $0.3 \times 0.1 + 0.7 \times 0.4 = 0.31$.

2.7 The chance of rain given an alert is $0.03/0.31 \approx 0.097$: about 10%, down from the 30% base rate, because rain makes peaks less likely.

2.8 The likelihood ratio of a flag is $0.8/0.1 = 8$, and the prior odds are 2 to 98. One flag gives odds of 16 to 98, a probability of about 14%.

A second flag multiplies the odds by 8 again, to 128 to 98, about 57%.

2.9 The prior odds are $0.4/0.6 \approx 0.667$. The likelihood ratios are 2.5 and 5, so the posterior odds are $0.667 \times 2.5 \times 5 \approx 8.33$, a probability of about 89%.

2.10 The likelihood ratio is $0.80/0.15 \approx 5.33$, and the prior odds are 1 to 4, so the posterior odds are about 1.33, a probability of about 57%.

2.11 The likelihood ratio is $(0.7^8 \times 0.3^2)/0.5^{10} \approx 0.00519/0.000977 \approx 5.31$. With prior odds of 1, the posterior odds are 5.31, a probability of about 84%.

2.12 Expect about $2{,}000{,}000/100{,}000 = 20$ false matches plus the one true match, about 21 flags. The chance that a given flag is the right person is about $1/21$, under 5%, though each false match had only a 1-in-100,000 chance.

2.13 By definition, $P(A \cap B) = P(A|B)P(B)$ and also $P(A \cap B) = P(B|A)P(A)$. Setting the two equal and dividing by $P(B)$ gives $P(A|B) = P(B|A)P(A)/P(B)$.

2.14 $P(A|B) = P(B|A)P(A)/P(B)$ and $P(A^c|B) = P(B|A^c)P(A^c)/P(B)$. Dividing the first by the second cancels $P(B)$ and gives the posterior odds as the prior odds $P(A)/P(A^c)$ times the likelihood ratio $P(B|A)/P(B|A^c)$.

2.15 The prior odds $p/(1-p)$ are close to p when p is small, so the posterior odds are about p times the likelihood ratio. When the odds are small, the probability, the odds divided by one plus the odds, is close to the odds themselves.

2.16 The log posterior odds are the log prior odds plus the sum of $\ln$ of each present word's likelihood ratio. Writing $x_j = 1$ if word j appears and 0 otherwise, that is a constant plus $\Sigma_j w_j x_j$, with weight w_j equal to the word's log likelihood ratio: a weighted sum of the inputs, which is what linear means.

Module 3: Random Variables and Expected Value

3.1 $E[X] = (1 + 2 + \ldots + 6)/6 = 3.5$. $E[X^2] = (1 + 4 + 9 + 16 + 25 + 36)/6 = 91/6$, so the variance is $91/6 - 3.5^2 = 35/12 \approx 2.92$, and the standard deviation is about 1.71.

3.2 $0.3 \times 10 + 0.7 \times (-5) = 3 - 3.5 = -0.5$: the bet loses 50 cents on average.

3.3 The mean is $2 \times 4 + 3 = 11$, and the variance is $2^2 \times 9 = 36$: adding a constant does not change the spread.

3.4 The mean is $0.3 + 0.4 = 0.7$. $E[X^2] = 0.3 + 0.8 = 1.1$, so the variance is $1.1 - 0.49 = 0.61$.

3.5 By linearity, $50 \times 0.02 = 1$, whether or not the defaults are related.

3.6 The variances add: $9 + 16 = 25$, so the standard deviation is 5, not 7.

3.7 The expected loss is $0.002 \times 300{,}000 = 600$ dollars a year. With a 30% loading, the premium is $600 \times 1.3 = 780$ dollars.

3.8 Waiting has an expected price of $20 + 24 + 15 = 59$ dollars, above $55. Its standard deviation is $\sqrt{(0.5 \times 19^2 + 0.4 \times 1^2 + 0.1 \times 91^2)} \approx 31.8$ dollars.

Waiting pays $4 more on average, but with a large spread; the choice depends on how much the owner can stand a bad day.

3.9 Drilling now: $0.2 \times 60 - 10 = 2$ million dollars. With perfect information, drill only when there is oil: $0.2 \times (60 - 10) = 10$ million. The survey is worth up to $10 - 2 = 8$ million dollars.

3.10 Expected return: $0.5 \times 30 + 0.5 \times (-20) = 5$ percent. Typical growth: $\sqrt{(1.3 \times 0.8)} = \sqrt{1.04} \approx 1.0198$, about 2% a year, well below 5%.

3.11 The fraction is $2 \times 0.55 - 1 = 0.1$. The growth per bet is $0.55 \ln 1.1 + 0.45 \ln 0.9 \approx 0.0050$, about 0.5% a bet.

3.12 The covariance is $-0.5 \times 10 \times 20 = -100$, so the variance of the total is $100 + 400 - 200 = 300$, and the standard deviation is about 17.3 MWh, well below the 30 MWh that adding the two standard deviations would give, and below the larger farm's 20 MWh alone.

3.13 $E[X + Y] = \Sigma_{x,y}(x + y)P(x, y)$, summing over all pairs of values, where $P(x, y)$ is the chance that $X = x$ and $Y = y$. Split it into $\Sigma_{x,y} xP(x, y) + \Sigma_{x,y} yP(x, y)$. Summing $P(x, y)$ over y gives $P(x)$, so the first sum is $E[X]$; likewise the second is $E[Y]$. Independence was never used.

3.14 Expand: $E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2$, using linearity and $E[X] = \mu$.

3.15 The chance of first head on toss k is 2^{-k}, so each possible payout contributes $2^{-k} \times 2^k = 1$ dollar, and the sum of infinitely many 1s is infinite.

The expected logarithm of the payout is $\Sigma_k 2^{-k} k \ln 2 = 2 \ln 2$, so the certain amount with the same expected logarithm is 4 dollars, since $\ln 4 = 2 \ln 2$. Daniel Bernoulli proposed this resolution in 1738.

3.16 The derivative is $p/(1 + f) - (1 - p)/(1 - f)$. Setting it to zero gives $p(1 - f) = (1 - p)(1 + f)$, so $f = 2p - 1$.

Module 4: Counting Events: Binomial and Poisson

4.1 $C(5, 3) \times 0.5^5 = 10/32 = 0.3125$.

4.2 $0.9^{10} \approx 0.349$ and $10 \times 0.1 \times 0.9^9 \approx 0.387$.

4.3 The mean is $400 \times 0.25 = 100$, and the standard deviation is $\sqrt{(400 \times 0.25 \times 0.75)} = \sqrt{75} \approx 8.66$.

4.4 $e^{-2} \approx 0.135$; $e^{-2} \times 2^2/2 \approx 0.271$; and at most 2 is $e^{-2}(1 + 2 + 2) \approx 0.677$.

4.5 The mean is 9, and the standard deviation is $\sqrt{9} = 3$.

4.6 With $\lambda = 3$, the chance is $e^{-3} \approx 0.050$, about one month in 20.

4.7 The expected number is $40 \times 0.03 = 1.2$. From the binomial, the chance of more than 3 is about 0.031, roughly one day a month.

4.8 Each score has a standard deviation of about $\sqrt{(1{,}000 \times 0.8 \times 0.2)} \approx 12.6$, so the difference of two independent scores has one of about $\sqrt{2} \times 12.6 \approx 17.9$. The gap of 14 is under one standard deviation, well within ordinary noise. On a shared set of questions the two scores would move together, and comparing only the questions on which the models disagree can detect a smaller gap.

4.9 The Poisson gives about 0.017. The shortcut gives $15.5/7.07 \approx 2.19$ standard deviations, a chance of about 0.014 under the normal curve of Module 5, a little too low because the Poisson's upper tail is longer. The extra 0.5 counts the whole step from 65 to 66, since claims come in whole numbers.

4.10 Binomial: $1 - 0.998^{500} \approx 0.6325$. Poisson, with $\lambda = 500 \times 0.002 = 1$: $1 - e^{-1} \approx 0.6321$. They differ by less than 0.0004.

4.11 The average is $0.75 \times 1 + 0.25 \times 5 = 2$ percent. Under independence at 2%, the chance of 6 or more is about 1.5%. Under the mixture it is about 9.6%: 0.75 times a tiny chance in good years, plus 0.25 times about 38% in bad years.

4.12 The count's standard deviation is $\sqrt{(20{,}000 \times 0.04 \times 0.96)} \approx 27.7$, so the rate's is about $27.7/20{,}000 \approx 0.14$ percentage points.

4.13 By the binomial theorem, $\Sigma_k C(n,k) p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1$.

4.14 Write the count as the sum of n indicators, each 1 if its trial succeeds and 0 if not. Each has expected value p, so by linearity the sum has expected value np, with no need for the formula.

4.15 Each indicator I has variance $E[I^2] - p^2 = p - p^2 = p(1-p)$, since $I^2 = I$. The indicators are independent, so the variances add: $np(1-p)$.

4.16 Write $C(n,k)p^k(1-p)^{n-k} = [n(n-1)\ldots(n-k+1)/n^k] \times (\lambda^k/k!) \times (1-\lambda/n)^n \times (1-\lambda/n)^{-k}$.

As n grows, the first bracket and the last factor approach 1, and $(1-\lambda/n)^n$ approaches $e^{-\lambda}$, leaving $e^{-\lambda}\lambda^k/k!$.

Module 5: Continuous Distributions and Fat Tails

5.1 The density is 0.1 on the range, so the chance is $(5-2) \times 0.1 = 0.3$. The mean is the midpoint, 5 minutes.

5.2 130 is $z = 2$, so the chance above it is about 2.3%. 85 to 115 is within one standard deviation, about 68%.

5.3 $(72 - 60)/8 = 1.5$.

5.4 $50 + 2.326 \times 10 \approx 73.3$.

5.5 $e^{-0.5 \times 3} = e^{-1.5} \approx 0.223$, and the mean wait is $1/0.5 = 2$ hours.

5.6 The median is $e^0 = 1$, and the mean is $e^{0.5^2/2} = e^{0.125} \approx 1.13$.

5.7 $z = (440 - 500)/30 = -2$, so the chance is about 2.3%, roughly one month in 44.

5.8 The log return must exceed $\ln 1.3 \approx 0.262$, a z-score of $(0.262 - 0.06)/0.20 \approx 1.01$, so the chance is about 15.6%.

5.9 VaR: $2.33 \times 0.012 \times 50{,}000{,}000 \approx 1.40$ million dollars. Expected shortfall: $2.67 \times 0.012 \times 50{,}000{,}000 \approx 1.60$ million.

5.10 $e^{-0.02\times 30} = e^{-0.6} \approx 0.55$. The fleet expects $1{,}000 \times 0.02 = 20$ failures a year, a Poisson count with a standard deviation of about 4.5.

5.11 Normal: about 1 day in 15,800, or once every 63 years of 252 trading days. The t distribution with 3 degrees of freedom: about 0.6% of days, roughly 1.6 days a year.

5.12 It is a move of 20.5 standard deviations. The normal puts the chance of a fall that large at about 10^{-93}, so the normal model is simply wrong about the tail.

5.13 $P(T > s + t | T > s) = P(T > s + t)/P(T > s) = e^{-\lambda(s+t)}/e^{-\lambda s} = e^{-\lambda t}$, the same as the chance of surviving t years from new.

5.14 By linearity, the mean is $(\mu - \mu)/\sigma = 0$. Dividing by σ divides the standard deviation by σ, and subtracting μ does not change it, so the standard deviation is $\sigma/\sigma = 1$.

5.15 The ratio is $e^{m+1/2}/e^{m} = e^{0.5} \approx 1.65$: the mean is 65% higher. Large gains are possible and losses are capped at the whole price, so a long right tail pulls the mean up.

5.16 The expected shortfall is the average of the losses beyond the VaR. Every one of those losses is at least the VaR, so their average is too.

Module 6: Sums, Averages, and the Central Limit Theorem

6.1 Sum: mean 250, standard deviation $4 \times \sqrt{25} = 20$. Average: mean 10, standard deviation $4/\sqrt{25} = 0.8$.

6.2 Because it shrinks like $1/\sqrt{n}$, halving it takes four times the data, and cutting it to a tenth takes a hundred times.

6.3 The mean is 3.5. One roll has a standard deviation of about 1.71, so the average of 100 has $1.71/10 \approx 0.17$.

6.4 400 policies: mean \$40,000, standard deviation $1{,}000 \times 20 = 20{,}000$, or 50% of the mean. 40,000 policies: mean \$4 million, standard deviation $1{,}000 \times 200 = 200{,}000$, or 5%.

6.5 The error must fall by a factor of 4, which takes $4^2 = 16$ times the runs: 16,000.

6.6 The count has mean 50 and standard deviation $\sqrt{(100 \times 0.5 \times 0.5)} = 5$. Using 59.5 as the cutoff, $z = (59.5 - 50)/5 = 1.9$, a chance of about 2.9%.

6.7 The mean is \$240,900 and the standard deviation about \$15,760, so $z = (210{,}000 - 240{,}900)/15{,}760 \approx -1.96$, a chance of about 2.5%.

6.8 The hourly count is Poisson with mean 120 and standard deviation $\sqrt{120} \approx 10.95$. With 140.5 as the cutoff, $z \approx 1.87$, a chance of about 3.1%.

6.9 $\sqrt{(0.09/50 + (1 - 1/50) \times 0.3 \times 0.09)} = \sqrt{(0.0018 + 0.02646)} \approx 0.168$, about 16.8%, close to the floor of $0.30 \times \sqrt{0.3} \approx 0.164$, or 16.4%.

6.10 $\sqrt{(0.52 \times 0.48/1{,}000)} \approx 0.0158$, or 1.6 points. Two standard errors either side gives roughly 49% to 55%.

6.11 The variance is $20 \times 900 + 20 \times 19 \times 0.6 \times 900 = 223{,}200$, so the standard deviation is about 472 MW. Independent farms would give $\sqrt{(20 \times 900)} \approx 134$ MW.

6.12 The cost has variance $40^2/12 \approx 133$, so the profit has mean \$30 million and standard deviation $\sqrt{(133 + 400)} \approx 23.1$ million. Treating the profit as roughly normal, a loss is $30/23.1 \approx 1.3$ standard deviations below the mean, a chance of about 10%.

6.13 The variance of the sum is $n\sigma^2$, because independent variances add. Dividing a variable by n divides its variance by n^2, so the average has variance $n\sigma^2/n^2 = \sigma^2/n$.

6.14 The variance of the sum is $n\sigma^2$ plus a covariance $\rho\sigma^2$ for each of the $n(n-1)$ ordered pairs. Dividing by n^2 gives $\sigma^2/n + (n-1)\rho\sigma^2/n$, which is the formula.

6.15 The average has standard deviation $\sigma/\sqrt{n}$. For any fixed distance d, the chance of missing the mean by more than d is at most $\sigma^2/(nd^2)$, taking $k = d\sqrt{n}/\sigma$. That bound goes to zero as n grows.

6.16 Each draw has mean 1/2 and variance 1/12. The sum of 12 has mean 6 and variance 1, so subtracting 6 gives mean 0 and standard deviation 1, and by the central limit theorem the sum is close to normal.

Module 7: Estimation and Confidence Intervals

7.1 The average is $45/5 = 9$, and the median, the middle value, is also 9. The squared distances from 9 are 25, 4, 0, 1, and 36, which sum to 66, so $s = \sqrt{(66/4)} = \sqrt{16.5} \approx 4.06$.

7.2 $12/\sqrt{100} = 1.2$, and $12/\sqrt{400} = 0.6$: half as large, for four times the data.

7.3 $\sqrt{(0.45 \times 0.55/600)} \approx 0.0203$, or 2.0 points. The interval is $45 \pm 1.96 \times 2.03$, about 41.0% to 49.0%.

7.4 48.4 to 51.6, then 48.0 to 52.0, then 47.4 to 52.6, using multipliers of 1.645, 1.96, and 2.576. More confidence costs a wider interval.

7.5 $(1.96/0.03)^2 \times 0.25 \approx 1{,}067.1$, so 1,068 people. For 2 points, $(1.96/0.02)^2 \times 0.25 = 2{,}401$.

7.6 With 4 degrees of freedom the multiplier is 2.776, about 42% wider than 1.96. With 24 it is 2.064, about 5% wider.

7.7 $\sqrt{(0.055 \times 0.945/2{,}000)} \approx 0.0051$, so the margin is $1.96 \times 0.51 \approx 1.0$ point, giving an interval of about 4.5% to 6.5%.

7.8 The average is 41.2 and $s \approx 0.524$, so the standard error is $0.524/\sqrt{5} \approx 0.235$. With 4 degrees of freedom the multiplier is 2.776, giving 41.2 ± 0.65, or 40.55% to 41.85%.

7.9 The standard error is $4{,}000/1.645 \approx 2{,}430$ dollars. The 95% margin is $1.96 \times 2{,}430 \approx 4{,}770$ dollars.

7.10 The standard error of the gap is $\sqrt{((0.48 + 0.46 - 0.02^2)/1{,}000)} \approx 0.0307$, so its 95% margin is about 6.0 points. A 2-point lead is well within it, and the poll cannot say who is ahead.

7.11 Anything from a loss of 70,000 to a gain of 170,000, at 90% confidence. The range includes losses, so the report alone does not show that jobs grew.

7.12 The total rises by 18.8 hours, from 46.7 to 65.5, so the average rises from 2.46 to about 3.45 hours. The median stays 1.4, since the middle value has not moved. One extreme value can move an average a long way and a median not at all.

7.13 $p(1-p) = 1/4 - (p - 1/2)^2$, which is largest when $p = 1/2$, where it equals 1/4. The margin is then $1.96 \times \sqrt{(1/(4n))} = 0.98/\sqrt{n}$, about $1/\sqrt{n}$.

7.14 $0.95^{20} \approx 0.36$, so it is more likely than not that at least one misses. The expected number of misses is $20 \times 0.05 = 1$.

7.15 With at least 2 million replies, the standard error of a share near 57% was at most $\sqrt{(0.571 \times 0.429/2{,}000{,}000)} \approx 0.00035$, under 0.04 points. Landon actually won about 37.5% of the votes cast for the two of them, so the poll missed by about 20 points, more than 500 standard errors. The error was bias, which no sample size fixes.

7.16 Squared distances measured from the true mean μ and from $\bar{x}$ are linked by

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu)^2$$

On average, the first sum on the right is $n\sigma^2$, and $(\bar{x}-\mu)^2$ is the variance of the average, σ^2/n. So the left side averages $n\sigma^2-\sigma^2=(n-1)\sigma^2$. Dividing by n gives $(n-1)\sigma^2/n$, slightly too small; dividing by $n-1$ gives exactly σ^2.

Module 8: Hypothesis Tests

8.1 The chance of a z-score beyond 2.5 in either direction is $2\times 0.0062\approx 0.012$. That is significant at 5% but not at 1%.

8.2 Those beyond 1.645, 1.96, and 2.576 respectively, in either direction.

8.3 Under the null, heads has mean 50 and standard deviation 5, so $z=(58-50)/5=1.6$, and the two-sided p-value is about 0.11. A fair coin does this often enough that there is no real evidence of bias.

8.4 A Type I error is an alarm with no fire; a Type II error is silence during a fire. A more sensitive alarm makes fewer misses at the cost of more false alarms.

8.5 $1-0.95^{12}\approx 0.46$, and $0.05/12\approx 0.0042$.

8.6 20%, the Type II error rate β; and 5%, the significance level α.

8.7 The pooled rate is 4.2%, so the denominator is $\sqrt{(0.042\times 0.958\times 2/20{,}000)}\approx 0.00201$, and $z=0.004/0.00201\approx 1.99$. The two-sided p-value is about 0.046, just significant at 5%. The 95% interval for the lift runs from almost nothing to 0.8 points.

8.8 $7.84\times(0.09+0.0979)/0.01^2\approx 14{,}700$ visitors per version.

8.9 For 100 days, $z=(55-50)/5=1.0$, a p-value of 0.32: no evidence. For 1,000 days, the null's standard deviation is $\sqrt{(1{,}000\times 0.25)}\approx 15.8$, so $z=50/15.8\approx 3.16$, a p-value of about 0.0016. The same 55% rate is strong evidence over ten times the data.

8.10 Using the binomial counts $C(10,8)=45$, $C(10,9)=10$, and $C(10,10)=1$, the chance is $(45+10+1)/1{,}024\approx 0.055$. About 55 of 1,000 managers would compile such a record by luck alone.

8.11 Of 1,000 hypotheses, 200 are true and 100 of them test significant; 40 of the 800 false ones also do. So $100/140\approx 71\%$ of significant results are real.

8.12 No, since zero lies inside the interval. The interval also shows that the data allow anything from a slight loss to a gain of 1.3 points, so the experiment is inconclusive, which is not the same as showing no effect.

8.13 The test statistic is the distance from v to the estimate, divided by the standard error. The test rejects when that ratio exceeds 1.96 in size, that is, when v is more than 1.96 standard errors from the estimate. That is exactly when v lies outside the interval.

8.14 On average, the best of 100 is about 2.5 standard errors above zero, and the standard error is $1/\sqrt{T}$ for T years. Setting $2.5/\sqrt{T}=0.5$ gives $\sqrt{T}=5$, or about 25 years.

8.15 Each look is another chance for noise to cross the line, and the test stops at its luckiest moment rather than at a planned size. With ten looks, the chance of a false win is about 19%. Fixing the sample size in advance, or using a stricter threshold at each look, restores the 5%.

8.16 With n visitors per group, the standard error of the gap in rates, σ_g, is

$$\sigma_g = \sqrt{\frac{p_A(1-p_A) + p_B(1-p_B)}{n}}$$

If the true gap is $\delta = p_B - p_A$, the statistic z is roughly normal with mean δ/σ_g and standard deviation 1. It exceeds 1.96 with probability 80% when $\delta/\sigma_g - 1.96 = 0.84$, that is, when $\sigma_g = \delta/2.80$. Squaring both sides and solving for n gives the formula.

Module 9: Regression and Its Uncertainty

9.1 $y = 7$ and $y = 8$. The slope means that y rises by 0.5 for each unit of x.

9.2 $b = 40/20 = 2$ and $a = 12 - 2 \times 5 = 2$, so the line is $\hat{y} = 2 + 2x$.

9.3 $R^2 = 0.6^2 = 0.36$, or 36%.

9.4 $t = 1.5/0.5 = 3$. With 28 degrees of freedom the multiplier is 2.048, so the interval is 1.5 ± 1.02, from 0.48 to 2.52.

9.5 The fitted value is $2 + 2 \times 6 = 14$, and the residual is $15 - 14 = 1$.

9.6 $b = 0.8 \times 10/4 = 2$.

9.7 $-10.2 + 0.40 \times 90 = 25.8$ GW and $-10.2 + 0.40 \times 95 = 27.8$ GW, so the extra 5 degrees add 2.0 GW.

9.8 At $\bar{x} = 88$, the standard error is $0.66 \times \sqrt{(1 + 0.1)} \approx 0.69$. The interval is $25.0 \pm 2.306 \times 0.69$, about 23.4 to 26.6 GW.

9.9 At the same temperature, a weekday's peak is about 1.48 GW higher than a weekend day's. Its t-statistic is $1.48/0.15 \approx 10$, so it is far from zero.

9.10 With 198 degrees of freedom the multiplier is about 1.97, so the interval is 150 ± 39, about \$111 to \$189 per square foot.

9.11 By about $-0.25 \times 20 = -5$ percent. More exactly, use is multiplied by $1.2^{-0.25} \approx 0.955$, a fall of 4.5%.

9.12 Reverse causation: cities with more crime hire more police. Confounding: large, dense cities tend to have both more crime and more police. Showing an effect needs police numbers that changed for reasons unrelated to crime.

9.13 At $x = \bar{x}$, the line gives $a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}$.

9.14 $\Sigma e_i = \Sigma(y_i - a - bx_i) = n\bar{y} - na - nb\bar{x} = n(\bar{y} - a - b\bar{x})$, which is zero because $a = \bar{y} - b\bar{x}$.

9.15 The line in standard units has slope $r = 0.5$, so the forecast is $0.5 \times 2 = 1$ standard deviation above average. Part of this year's high score was luck, which does not repeat. This is regression to the mean, the effect Galton found in heights.

9.16 The sample variance gives $\Sigma(x_i - \bar{x})^2 = (n-1)s_x^2$, and the definition of r gives $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = (n-1) r\, s_x\, s_y$. Dividing the second by the first gives $b = r\, s_y/s_x$.

Module 10: Fitting and Testing Models

10.1 The observed share, $18/60 = 0.3$.

10.2 $1/(1 + e^0) = 0.5$, $1/(1 + e^{-2}) \approx 0.881$, and $1/(1 + e^3) \approx 0.047$.

10.3 The odds are $0.2/0.8 = 0.25$, and the log-odds are $\ln 0.25 \approx -1.39$.

10.4 $-\ln 0.5 \approx 0.69$, $-\ln 0.9 \approx 0.11$, and $-\ln 0.01 \approx 4.61$. The average is about 1.80, most of it from the one confident mistake.

10.5 $e^2 \approx 7.4$: on average, as uncertain as choosing among about seven equally likely tokens.

10.6 Accuracy $(50 + 915)/1{,}000 = 96.5\%$, precision $50/60 \approx 83\%$, and recall $50/75 \approx 67\%$.

10.7 The log-odds are $-6 + 3.2 = -2.8$, a probability of $1/(1 + e^{2.8}) \approx 5.7\%$. From 30% to 60%, the odds are multiplied by $e^{0.08 \times 30} = e^{2.4} \approx 11$.

10.8 The average count, $12/6 = 2$ outages a month. The chance of a month with none is $e^{-2} \approx 0.135$.

10.9 Flagging everything gives precision $100/10{,}000 = 1\%$, recall 100%, and accuracy 1%. The detector has precision 27%, recall 80%, and accuracy 97.6%.

10.10 Precision $95/1{,}000 = 9.5\%$ and recall 95%. The 700 extra flags catch 15 more frauds, about 47 extra checks per extra fraud caught, 46 of them of honest transactions, so it pays when a missed fraud costs more than about 47 checks.

10.11 It rained on 23% of those days. If the true chance were 30%, the share in 200 days would have a standard error of $\sqrt{(0.3 \times 0.7/200)} \approx 0.032$, so 23% is about 2.2 standard errors below, a two-sided p-value of about 0.03. There is some evidence that this forecaster's 30% is too high.

10.12 The model is overfitting: it has learned noise in the training data. Simplify it, add training data, or stop training earlier, choosing by the validation error; then measure the final choice once on the test set.

10.13 The log-likelihood is $k \ln p + (n - k) \ln(1 - p)$. Its slope, by calculus, is $k/p - (n - k)/(1 - p)$, which is zero when $k(1 - p) = (n - k)p$, that is, when $p = k/n$.

10.14 Each point's density is proportional to $e^{-e_i^2/(2\sigma^2)}$, where e_i is its residual. The product over all points is proportional to $e^{-\Sigma e_i^2/(2\sigma^2)}$, which is largest when Σe_i^2 is smallest.

10.15 $1 - p = e^{-z}/(1 + e^{-z})$, so $p/(1 - p) = 1/e^{-z} = e^z$, and $\ln(p/(1 - p)) = z$.

10.16 Choosing the best of 50 on the test set selects for luck, like the best of 100 backtests in Module 8. The best of 50 equally good scores is typically about 2.2 standard errors above their true level, so the reported score overstates by about 2 points. Only a fresh test set measures the chosen model honestly.

Answer Sheet

Answers to every module's Check Your Understanding quiz.

Module	Answers
Module 1 Quiz	1. A 2. B 3. C 4. D 5. C
Module 2 Quiz	1. A 2. B 3. D 4. C 5. D
Module 3 Quiz	1. B 2. A 3. A 4. B 5. D
Module 4 Quiz	1. C 2. B 3. A 4. D 5. C
Module 5 Quiz	1. B 2. D 3. A 4. C 5. A
Module 6 Quiz	1. B 2. C 3. D 4. B 5. A
Module 7 Quiz	1. D 2. C 3. B 4. C 5. A
Module 8 Quiz	1. D 2. B 3. C 4. D 5. A
Module 9 Quiz	1. C 2. D 3. B 4. A 5. B
Module 10 Quiz	1. D 2. C 3. A 4. D 5. C