Abridged Math Foundations for Signal Processing, Machine Learning, and Artificial Intelligence

A seven-part refresher

August 2026 (revised)

How to use this booklet

Machine learning is usually taught atop a full linear algebra course, yet it uses a surprisingly small slice of the subject — eigenvectors, a pillar of most linear algebra courses, barely appear. What machine learning actually leans on, constantly, is: vectors as feature representations, dot products as scores, norms as loss magnitudes, matrices as linear maps and datasets, spans/rank when reasoning about what a model can express, projections and least squares behind linear regression, and gradients of vector functions for gradient descent.

Part I covers exactly that slice, in seven lessons.

Machine learning also assumes a first course in *probability*, and unlike the eigenvector story, it uses this one heavily: Bayes' rule and conditional independence power the Bayesian-networks and HMM lectures, expectations define the utilities that MDPs and game-playing maximize, and every loss function is an expectation over data. **Part III** covers that probability core — probabilistic models and counting, conditioning and Bayes' rule, discrete and continuous random variables, expectation — in four lessons, following Bertsekas & Tsitsiklis, *Introduction to Probability* (2nd ed.), Chapters 1–3. Do it right after Part I.

Part II continues the linear algebra through the rest of Axler's book (Chapters 5–9): complex vector spaces, eigenvalues and diagonalization, the spectral theorem, unitary matrices and the discrete Fourier transform, the singular value decomposition, and determinants and trace. Its target is *Fourier analysis* — a subject whose linear-algebra backbone is precisely orthonormal expansions, eigen-analysis of linear systems, and the DFT as a unitary matrix. Part II is not needed for the machine-learning core and can be done at a relaxed pace.

Part IV continues the probability toward *statistical inference and statistical signal processing*, which assume a full first course in probability plus linear systems and Fourier transforms, and then build: random vectors and random processes, convergence and limit theorems, Markov and Gaussian processes, stationarity and autocorrelation, minimum-mean-square-error and linear estimation, Kalman filtering, hypothesis testing, and PCA. Part IV is the prerequisite half of that list, from Bertsekas & Tsitsiklis Chapters 4–8: derived distributions, covariance, conditional expectation as an estimator, random vectors and Gaussian vectors, limit theorems, the Bernoulli and Poisson processes, Markov chains, and least-mean-squares estimation. Part IV leans on Part III throughout and, at marked moments, on Part II: Lesson 19 reuses Lesson 6's projection geometry, Lesson 20 reuses Lesson 10's spectral theorem, and Lesson 23 reuses Lesson 9's eigenvalues.

Part V completes the undergraduate prerequisite set with the *signals-and-systems* core, following Oppenheim & Willsky (O&W), *Signals and Systems* (2nd ed.) — the linear-systems

layer that both Fourier analysis and statistical inference stand on: signals and system properties, LTI systems and convolution, Fourier series, the continuous-time and discrete-time Fourier transforms, frequency response and filtering, sampling and aliasing, and a working minimum of Laplace and z-transforms. Part V leans on Part II constantly (complex exponentials, the eigenfunction idea, the DFT): do it right after Part II. Parts IV and V are independent of each other; statistical inference wants both.

Applied signal processing — the DFT in practice, filter design and finite word lengths, multirate and sigma–delta conversion, random signals and spectrum estimation, Wiener/LMS/Kalman filtering, detection and digital communication, images, audio, and the electronics bench — is not in this booklet: it lives where it is exercised, as the theory sections of the **Course 3 labs** at diiv.io (embedded DSP on the STM32, Pi 5, and Jetson). Parts VI and VII below point there by lab number.

Part VI is the optional *rigor layer*: it explains, with the bare minimum of real analysis (Ross, with Rudin for several variables and Kreyszig for metric spaces), measure theory (Axler's MIRA), functional analysis (Bowers–Kalton, Kreyszig), complex analysis (Bak–Newman), distribution theory (Strichartz), and numerical linear algebra (Trefethen–Bau), *why* Part V's special integrals, derivatives, δ-manipulations, transform inversions, and functional derivatives are legal — when limits pass through integrals, what $\delta(t)$ actually is, why poles and ROCs behave, what space signals live in, and when to trust the arithmetic — and, along the way, proves the theorems of calculus itself — the mean value theorem, Taylor's theorem, both fundamental theorems, and their n-dimensional upgrades — before closing, in its last lesson, with the central limit theorem that Part IV states. Nothing in it is formally required by the earlier parts, but graduate Fourier analysis develops exactly its distribution theory, statistical inference's "almost surely" is its "almost everywhere," and it is deliberately built as the on-ramp to graduate-level DSP (estimation, adaptive filters, wavelets), whose native language — Hilbert spaces, operators, distributions, conditioning — it installs. Do it after Part V, at a relaxed pace, or interleave it with the courses themselves.

Part VII is the *optimization and information layer*: it adds the parts of Boyd–Vandenberghe's *Convex Optimization* and Polyanskiy–Wu's *Information Theory: From Coding to Learning* that recur in machine learning, signal processing, communications, and sensors. Convex optimization turns fitting, regularization, filter design, detector design, and resource allocation into problems with global certificates; information theory turns uncertainty, distinguishability, compression, channel capacity, and representation quality into computable quantities. Do it after Parts I, III, and V; Parts VI–VII make the signal-processing and rigor connections richer, but the first pass through Part VII is self-contained.

Each lesson has two interleaved layers:

- **Theory** in the style of Axler's *Linear Algebra Done Right* (LADR, 4th ed.) for Parts I–II, of Bertsekas–Tsitsiklis (B&T) for Parts III–IV, of Oppenheim–Willsky (O&W) for Part V, and of Ross, Axler (MIRA), Bak–Newman, Strichartz, Bowers–Kalton, Kreyszig, Rudin, and Trefethen–Bau for Part VI (one or two sources per lesson), and Boyd–Vandenberghe plus Polyanskiy–Wu for Part VII: precise definitions and theorems with complete proofs where feasible. Understanding *why* a fact is true is what makes it stick.

- **Concrete practice** in the style of Strang's *Introduction to Linear Algebra*: actual small examples — 2- and 3-dimensional vectors, dice, coins, three-state chains — computed by hand, with geometric pictures in words.

Exercises at the end of each lesson are tagged **[Proof]** or **[Hand]**. Do all of the [Hand] ones and as many [Proof] ones as time allows — the starred proofs are the most valuable.

What each subject needs from this booklet

Machine learning — linear models, SGD, features, neural networks, state-based search, Markov decision processes, game playing, constraint satisfaction, and graphical models (Bayesian networks and hidden Markov models) — needs Part I (the ML core) and Part III (the MDP, game-playing, and graphical-models material) — plus the Markov-chain lesson of Part IV if you want MDPs and HMMs to feel like review rather than news.

Statistical inference assumes all of Part III as a *floor* and then builds on precisely the material of Part IV; its linear-systems and Fourier prerequisite is Part V, whose own linear-algebra spine is Part II. *Fourier analysis* stands on Part V's first half, with Part II as its linear-algebra backbone. With all five parts, the undergraduate prerequisite set for all three subjects is complete.

Every lesson below carries a NEEDED FOR: tag. The booklet-wide map (*ML* = machine learning; *inference* = statistical inference):

Lessons	**Machine learning**	**Fourier / inference**
1–7 (Part I: linear algebra core)	**required**	assumed background for both
8–13 (Part II: eigen/Fourier/SVD)	not needed	**Fourier core**; inference (PCA, DFT)
14–17 (Part III: probability core)	**required**	**inference assumes it** (first-course level)
18 (derived dists, covariance)	not needed	**inference**
19 (conditional expectation, LMS)	not needed	**inference** (MMSE, Kalman lead-in)
20 (random vectors, Gaussians)	not needed	**inference** (PCA, Gaussian processes)
21 (limit theorems)	helpful (Monte Carlo)	**inference** (convergence, concentration)
22 (Bernoulli/Poisson processes)	not needed	**inference** (first random processes)
23 (Markov chains)	**helpful** (MDPs, HMMs)	**inference** (Markov processes)
24–30 (Part V: signals & systems)	not needed	**both** (linear-systems prerequisite)
31–43 (Part VI: the rigor layer)	not needed	optional for both; **grad DSP**
44–51 (Part VII: optimization & information)	**ML bridge**	**sensors, DSP, comms**

In one sentence: *for machine learning do Lessons 1–7 and 14–17 (and skim 23); Parts II and V are for Fourier analysis; Parts II–V together are the full statistical-inference load-out; the applied layer (DSP, images, communications, the bench) is the Course 3 labs; Part VI is the optional "why it all works" layer and the bridge to graduate DSP; Part VII is the convex-optimization and information-theory bridge for ML, sensors, DSP, and communications.*

Schedule

Part I — the linear algebra sprint:	
Day 1 (Sat 8/8)	Lesson 1: Vectors and linear combinations
Day 2 (Sun 8/9)	Lesson 2: Dot products, norms, orthogonality
Day 3 (Mon 8/10)	Lesson 3: Span, independence, basis, dimension
Day 4 (Tue 8/11)	Lesson 4: Matrices and linear maps
Day 5 (Wed 8/12)	Lesson 5: Null space, rank, and solving $Ax = b$
Day 6 (Thu 8/13)	Lesson 6: Projections and least squares
Day 7 (Fri 8/14)	Lesson 7: Gradients for machine learning
Day 9 (Sun 8/16)	Buffer: unfinished exercises; skim the self-check list
Part III — the probability core, one lesson per day, starting 8/17:	
Day 1	Lesson 14: Probability models, counting, conditioning, and Bayes' rule
Day 2	Lesson 15: Discrete random variables and expectation
Day 3	Lesson 16: Multiple random variables and conditioning
Day 4	Lesson 17: Continuous random variables and the normal
Part II — one lesson per day, after Part III:	
Day 1	Lesson 8: Complex vectors and complex exponentials
Day 2	Lesson 9: Eigenvalues, eigenvectors, diagonalization
Day 3	Lesson 10: Self-adjoint operators and the spectral theorem
Day 4	Lesson 11: Unitary matrices and the discrete Fourier transform
Day 5	Lesson 12: The singular value decomposition
Day 6	Lesson 13: Determinants and trace
Part V — one lesson per day, right after Part II:	
Day 1	Lesson 24: Signals and systems
Day 2	Lesson 25: LTI systems and convolution
Day 3	Lesson 26: Fourier series
Day 4	Lesson 27: The continuous-time Fourier transform
Day 5	Lesson 28: The discrete-time Fourier transform and frequency response
Day 6	Lesson 29: Sampling
Day 7	Lesson 30: Laplace and z-transforms, a working minimum
Part IV — one lesson per day, after Parts II & III:	
Day 1	Lesson 18: Derived distributions, covariance, correlation
Day 2	Lesson 19: Conditional expectation and least mean squares
Day 3	Lesson 20: Random vectors and Gaussian vectors
Day 4	Lesson 21: Limit theorems
Day 5	Lesson 22: The Bernoulli and Poisson processes
Day 6	Lesson 23: Markov chains

Part VI — optional, after Part V; relaxed pace, or interleaved with the courses:	
Day 1	Lesson 31: Limits, series, and when swapping is legal
Day 2	Lesson 32: Metric spaces, continuity, and compactness
Day 3	Lesson 33: Differentiation: the calculus theorems, done right
Day 4	Lesson 34: The Riemann integral and the fundamental theorem
Day 5	Lesson 35: Lebesgue integration, a.e., and the swap theorems
Day 6	Lesson 36: Banach/Hilbert spaces, L^2, orthonormal bases
Day 7	Lesson 37: Complex analysis I: analytic functions, power series, contour integrals
Day 8	Lesson 38: Complex analysis II: zeros, poles, residues, ROCs
Day 9	Lesson 39: The Fourier, Laplace, and z-transforms as operators, proved
Day 10	Lesson 40: Distributions: making δ honest
Day 11	Lesson 41: Dual spaces and functional derivatives
Day 12–13	Lesson 42: Numerical linear algebra: conditioning, stability, factorizations, iteration
Day 14	Lesson 43: Transforms of distributions, Chernoff, and the CLT proved
Part VII — after Parts I, III, and V; before ML/sensors/comms depth:	
Day 1	Lesson 44: Convex models: sets, functions, and problems
Day 2	Lesson 45: Regularized fitting and inverse problems
Day 3	Lesson 46: Duality, KKT conditions, and certificates
Day 4	Lesson 47: Gradient, Newton, and proximal algorithms
Day 5	Lesson 48: Entropy, KL divergence, and mutual information
Day 6	Lesson 49: The AEP and lossless compression
Day 7	Lesson 50: Testing, channels, and sensor information
Day 8	Lesson 51: Rate-distortion, compression, representation learning

Deeper reading. Pointers appear at the end of each lesson: LADR sections by number (e.g. 1A, 3C), Strang by chapter name (editions of Strang differ in page numbering), B&T by section number (e.g. B&T 1.3–1.5), and O&W by section number (e.g. O&W 4.3–4.4). Part VI points to Ross (§), Rudin (chapter), Kreyszig (§), MIRA (chapter/section), Bak–Newman (chapter), Strichartz (§), Bowers–Kalton (chapter), and Trefethen–Bau (lecture). Part VII points to Boyd–Vandenberghe by chapter and Polyanskiy–Wu by part/topic. Everything here is self-contained, though — the pointers are optional.

Notation. Following Axler, $\mathbb{F}$ denotes $\mathbb{R}$ or $\mathbb{C}$; in Part I you can always read $\mathbb{F} = \mathbb{R}$; Part II uses $\mathbb{F} = \mathbb{C}$ where it matters and says so. Vectors in $\mathbb{R}^n$ are written as columns; $x = (x_1, \dots, x_n)$ is shorthand for the column vector with entries $x_1, \dots, x_n$. In Parts III–IV, following B&T, $\mathbf{P}(A)$ is the probability of an event, random variables are capital letters $X, Y, \dots$, p_X is a probability mass function, f_X a density, $\mathbf{E}[X]$ an expectation, and $\mathrm{var}(X)$, $\mathrm{cov}(X, Y)$ variance and covariance. In Part V, following O&W, $x(t)$ is a continuous-time and $x[n]$ a discrete-time signal, δ and u are the unit impulse and unit step, h is an impulse response, and $X(j\omega)$, $X(e^{j\omega})$, $X(s)$, $X(z)$ are the CT Fourier, DT Fourier, Laplace, and z-transforms of x. Part V writes the imaginary unit as j, as every signals text does.

Contents

Part I: The Linear Algebra Core

1 Vectors and Linear Combinations

NEEDED FOR: Machine learning (core); assumed by everything later

1.1 Theory: the space $\mathbb{R}^n$

Definition 1.1 ($\mathbb{R}^n$)**.** For a positive integer n, $\mathbb{R}^n$ is the set of all lists of n real numbers:

$$\mathbb{R}^n = \{(x_1, \ldots, x_n) : x_k \in \mathbb{R} \text{ for } k = 1, \ldots, n\}.$$

Elements of $\mathbb{R}^n$ are called *vectors*; the number x_k is the kth *coordinate*.

Definition 1.2 (Addition and scalar multiplication)**.** For $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ and $\lambda \in \mathbb{R}$:

$$x + y = (x_1 + y_1, \ldots, x_n + y_n), \qquad \lambda x = (\lambda x_1, \ldots, \lambda x_n).$$

These two operations obey the familiar rules of arithmetic. Axler abstracts these rules into the definition of a *vector space*: a set V with an addition and scalar multiplication satisfying commutativity, associativity, existence of an additive identity 0 and additive inverses, a multiplicative identity ($1v = v$), and the distributive properties. $\mathbb{R}^n$ is the prototypical example, and for machine learning, $\mathbb{R}^n$ is essentially the only vector space you need. But two quick payoffs of the abstract view are worth internalizing, because their proofs teach the style of reasoning used everywhere below.

Proposition 1.3 (Unique additive identity)**.** *A vector space has exactly one additive identity.*

Proof. Suppose 0 and $0'$ are both additive identities, meaning $v + 0 = v$ and $v + 0' = v$ for every v. Then

$$0' = 0' + 0 = 0 + 0' = 0,$$

where the first equality holds because 0 is an additive identity, the second by commutativity of addition, and the third because $0'$ is an additive identity. Hence $0' = 0$. □

Proposition 1.4 (0 times a vector)**.** *In any vector space, $0v = 0$ for every vector v. (On the left, 0 is the number zero; on the right, the zero vector.)*

Proof. Using the distributive property with $0 = 0 + 0$ in $\mathbb{R}$:

$$0v = (0 + 0)v = 0v + 0v.$$

Adding the additive inverse $-(0v)$ to both sides gives $0 = 0v$. □

> **Why prove obvious things?** These proofs use only the listed axioms, never a picture. That discipline is exactly what you need later when statements stop being obvious — e.g., "the set of solutions of $Ax = 0$ is a subspace" or "a projection minimizes distance." The habit: every step cites a definition or a previously proved fact.

Definition 1.5 (Linear combination)**.** A *linear combination* of vectors $v_1, \ldots, v_m \in \mathbb{R}^n$ is a vector of the form

$$a_1 v_1 + \cdots + a_m v_m, \qquad a_1, \ldots, a_m \in \mathbb{R}.$$

1.2 Practice: vectors in 2D and 3D

A vector $v = (2, 1)$ in $\mathbb{R}^2$ is drawn as an arrow from the origin to the point $(2, 1)$. Addition is the parallelogram rule: place the tail of w at the head of v; the sum $v + w$ is the arrow from the origin to the final head. Scalar multiplication stretches ($\lambda > 1$), shrinks ($0 < \lambda < 1$), or flips ($\lambda < 0$) the arrow.

Example 1.6 (All combinations of one or two vectors)**.** Let $v = (1, 2)$ and $w = (3, 1)$ in $\mathbb{R}^2$.

- The set of all multiples $\{\lambda v : \lambda \in \mathbb{R}\}$ is the line through the origin and $(1, 2)$.
- The set of all combinations $\{av+bw\}$ is all of $\mathbb{R}^2$: given any target (t_1, t_2), we need $a+3b = t_1$ and $2a+b = t_2$; solving gives $b = (2t_1 - t_2)/5$ and $a = t_1 - 3b$, which always works. (Lesson 3 explains why it always works: v and w are linearly independent.)

In $\mathbb{R}^3$, the combinations of two (non-parallel) vectors fill a *plane* through the origin — not all of $\mathbb{R}^3$. Try to reach $(0, 0, 1)$ using only $u = (1, 0, 0)$ and $v = (0, 1, 0)$: every combination $au + bv = (a, b, 0)$ has third coordinate 0, so you cannot. Combinations of a third vector pointing out of that plane, e.g. $w = (0, 0, 1)$, fill all of $\mathbb{R}^3$.

Example 1.7 (Hand computation)**.** With $u = (1, 0, -1)$, $v = (2, 1, 0)$, $w = (0, 1, 2)$:

$$2u - v + 3w = (2, 0, -2) - (2, 1, 0) + (0, 3, 6) = (0, 2, 4).$$

Notice $(0, 2, 4) = 2w + 0u + 0v$ as well — the same vector can arise from different combinations. Lesson 3 turns this observation into the definition of linear *dependence*: indeed $2u - v + w = 0$ here (check it!), so these three vectors are dependent.

> **Machine-learning connection.** A feature extractor ϕ maps an input (an email, a sentence, a game state) to a feature vector $\phi(x) \in \mathbb{R}^n$. Adding evidence and scaling importance are literally vector addition and scalar multiplication. A linear model's weight vector $w \in \mathbb{R}^n$ lives in the same space.

1.3 Exercises

Exercise 1.1 (Hand)**.** Let $v = (2, -1)$ and $w = (1, 3)$. Compute $3v+2w$, $v-w$, and find scalars a, b with $av + bw = (0, 7)$.

Exercise 1.2 (Hand)**.** Which vectors in $\mathbb{R}^3$ are linear combinations of $u = (1, 1, 0)$ and $v = (0, 1, 1)$? Describe the set geometrically, and exhibit one specific vector in $\mathbb{R}^3$ that is *not* a combination of u and v, with justification.

Exercise 1.3 (Proof, $\star$)**.** Prove that in any vector space, $(-1)v = -v$ for every v; that is, $(-1)v$ is the additive inverse of v. (Use Proposition 1.4 and the distributive property. This is LADR 1.B territory.)

Exercise 1.4 (Proof)**.** Prove that the additive inverse of each vector is unique (mirror the proof of the unique-additive-identity proposition).

Deeper reading. LADR 1A ($\mathbb{R}^n$ and $\mathbb{C}^n$), 1B (vector space axioms); Strang, ch. "Introduction to Vectors" (vectors and linear combinations).

2 Dot Products, Norms, and Orthogonality

NEEDED FOR: Machine learning (core)

2.1 Theory: inner products

Definition 2.1 (Dot product)**.** For $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ in $\mathbb{R}^n$, the *dot product* (Axler: *Euclidean inner product*) is

$$x \cdot y = \langle x,\, y\rangle = x_1 y_1 + \cdots + x_n y_n.$$

The dot product on $\mathbb{R}^n$ satisfies, for all $x, y, z \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ (each is a one-line check from the definition):

- *positivity*: $\langle x,\, x\rangle \geq 0$, with equality if and only if $x = 0$;
- *symmetry*: $\langle x,\, y\rangle = \langle y,\, x\rangle$;
- *linearity in each slot*: $\langle x + z,\, y\rangle = \langle x,\, y\rangle + \langle z,\, y\rangle$ and $\langle \lambda x,\, y\rangle = \lambda\langle x,\, y\rangle$.

Every proof in this lesson uses only these three properties — never the coordinate formula. That is the Axler discipline, and it means the results hold for *any* inner product.

Definition 2.2 (Norm, orthogonality)**.** The *norm* of x is $\|x\| = \sqrt{\langle x,\, x\rangle} = \sqrt{x_1^2 + \cdots + x_n^2}$. Vectors x, y are *orthogonal* if $\langle x,\, y\rangle = 0$.

Theorem 2.3 (Pythagorean theorem)**.** *If $x, y \in \mathbb{R}^n$ are orthogonal, then $\|x+y\|^2 = \|x\|^2 + \|y\|^2$.*

Proof. Expanding with linearity and symmetry,

$$\|x + y\|^2 = \langle x + y,\, x + y\rangle = \langle x,\, x\rangle + 2\langle x,\, y\rangle + \langle y,\, y\rangle = \|x\|^2 + 2\langle x,\, y\rangle + \|y\|^2,$$

and orthogonality makes the middle term vanish. □

The next lemma is the engine of Lessons 2 and 6: any vector can be split into a piece along y and a piece orthogonal to y.

Lemma 2.4 (Orthogonal decomposition)**.** *Let $y \neq 0$. For any $x \in \mathbb{R}^n$, set*

$$c = \frac{\langle x,\, y\rangle}{\|y\|^2} \quad \text{and} \quad z = x - cy.$$

Then $\langle z,\, y\rangle = 0$ and $x = cy + z$.

Proof. The identity $x = cy + z$ holds by the definition of z. For orthogonality,

$$\langle z,\, y\rangle = \langle x - cy,\, y\rangle = \langle x,\, y\rangle - c\,\langle y,\, y\rangle = \langle x,\, y\rangle - \frac{\langle x,\, y\rangle}{\|y\|^2}\,\|y\|^2 = 0.$$ □

The vector cy is the *projection of x onto the line spanned by y*.

Theorem 2.5 (Cauchy–Schwarz inequality)**.** *For all $x, y \in \mathbb{R}^n$,*

$$|\langle x,\, y\rangle| \leq \|x\|\,\|y\|,$$

with equality if and only if one of x, y is a scalar multiple of the other.

Proof. If $y = 0$ both sides are 0. Otherwise decompose $x = cy + z$ as in Lemma 2.4. Since cy and z are orthogonal, the Pythagorean theorem gives

$$\|x\|^2 = \|cy\|^2 + \|z\|^2 = \frac{\langle x,\, y\rangle^2}{\|y\|^2} + \|z\|^2 \geq \frac{\langle x,\, y\rangle^2}{\|y\|^2},$$

because $\|z\|^2 \geq 0$. Multiplying by $\|y\|^2$ and taking square roots yields the inequality. Equality holds exactly when $\|z\| = 0$, i.e. $z = 0$, i.e. $x = cy$ is a multiple of y (and if x is a multiple of y, direct computation gives equality). □

Theorem 2.6 (Triangle inequality)**.** $\|x + y\| \leq \|x\| + \|y\|$ *for all* $x, y \in \mathbb{R}^n$.

Proof. Using Cauchy–Schwarz in the middle step,

$$\|x + y\|^2 = \|x\|^2 + 2\langle x,\, y\rangle + \|y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2. \qquad \square$$

By Cauchy–Schwarz, for nonzero x, y the ratio $\langle x,\, y\rangle/(\|x\|\|y\|)$ lies in $[-1, 1]$, so it is the cosine of a unique angle $\theta \in [0, \pi]$:

$$\cos\theta = \frac{\langle x,\, y\rangle}{\|x\|\,\|y\|}.$$

In $\mathbb{R}^2$ and $\mathbb{R}^3$ this matches the geometric angle between the arrows; in higher dimensions it is the *definition* of angle, and Cauchy–Schwarz is precisely what makes the definition legal.

2.2 Practice: 2D/3D computations

Example 2.7. Let $x = (3, 4)$ and $y = (4, -3)$. Then $\langle x,\, y\rangle = 12 - 12 = 0$: orthogonal. Norms: $\|x\| = \sqrt{9 + 16} = 5 = \|y\|$. Pythagoras check: $x + y = (7, 1)$, and $\|x + y\|^2 = 50 = 25 + 25$. ✓

Example 2.8 (Projection by hand)**.** Project $x = (2, 3)$ onto the line spanned by $y = (1, 1)$:

$$c = \frac{\langle x,\, y\rangle}{\|y\|^2} = \frac{2 + 3}{2} = \tfrac{5}{2}, \qquad cy = \left(\tfrac{5}{2}, \tfrac{5}{2}\right), \qquad z = x - cy = \left(-\tfrac{1}{2}, \tfrac{1}{2}\right).$$

Check: $\langle z,\, y\rangle = -\frac{1}{2} + \frac{1}{2} = 0$. ✓ The point $(\frac{5}{2}, \frac{5}{2})$ is the closest point to $(2, 3)$ on the line $y = x$ — Lesson 6 proves that projections are always the closest points.

Example 2.9 (Angles)**.** For $u = (1, 0, 1)$ and $v = (1, 1, 0)$: $\langle u,\, v\rangle = 1$, $\|u\| = \|v\| = \sqrt{2}$, so $\cos\theta = 1/2$ and $\theta = 60°$.

Machine-learning connection. A linear classifier predicts with the *score* $w \cdot \phi(x)$: positive score, predict $+1$; negative, predict -1. The decision boundary is the set $\{v : w \cdot v = 0\}$ — the hyperplane orthogonal to w. The *margin* of an example involves $\|w\|$, and cosine similarity between feature vectors is exactly $\cos\theta$ above. You will compute dot products constantly.

2.3 Exercises

Exercise 2.1 (Hand)**.** For $x = (1, 2, 2)$ and $y = (2, -2, 1)$: compute $\langle x,\, y\rangle$, $\|x\|$, $\|y\|$, and $\cos\theta$. Are x and y orthogonal?

Exercise 2.2 (Hand)**.** Project $x = (4, 1)$ onto $y = (2, 1)$; write $x = cy + z$ and verify $\langle z,\, y\rangle = 0$. Then compute $\|cy\|^2 + \|z\|^2$ and confirm it equals $\|x\|^2$.

Exercise 2.3 (Proof, $\star$)**.** (Parallelogram equality, LADR 6A) Prove that for all $x, y \in \mathbb{R}^n$,

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2,$$

using only the three inner-product properties. Interpret the identity for the parallelogram with sides x and y.

Exercise 2.4 (Proof)**.** Show that if $\langle x,\, v\rangle = 0$ for *every* $v \in \mathbb{R}^n$, then $x = 0$. (Hint: choose v wisely.)

Deeper reading. LADR 6A (inner products and norms — the proofs above follow Axler's); Strang, ch. "Introduction to Vectors" (lengths and dot products).

3 Span, Linear Independence, Basis, Dimension

NEEDED FOR: Machine learning (core)

3.1 Theory

Definition 3.1 (Span)**.** The *span* of $v_1, \ldots, v_m \in \mathbb{R}^n$ is the set of all their linear combinations:

$$\operatorname{span}(v_1, \ldots, v_m) = \{a_1 v_1 + \cdots + a_m v_m : a_1, \ldots, a_m \in \mathbb{R}\}.$$

The span of the empty list is defined to be $\{0\}$. If $\operatorname{span}(v_1, \ldots, v_m) = V$, we say the list *spans* V.

A *subspace* of $\mathbb{R}^n$ is a subset containing 0 that is closed under addition and scalar multiplication. Spans are exactly the subspaces you can describe by a finite list: $\operatorname{span}(v_1, \ldots, v_m)$ is the smallest subspace containing $v_1, \ldots, v_m$ (Exercise 3.4).

Definition 3.2 (Linear independence)**.** A list $v_1, \ldots, v_m$ is *linearly independent* if the only choice of scalars with

$$a_1 v_1 + \cdots + a_m v_m = 0$$

is $a_1 = \cdots = a_m = 0$. Otherwise the list is *linearly dependent*.

Equivalently (and this is the useful reading): independence means every vector in the span has *exactly one* representation as a combination. For if $a_1 v_1 + \cdots + a_m v_m = b_1 v_1 + \cdots + b_m v_m$, subtracting gives $\sum_k (a_k - b_k) v_k = 0$, so independence forces $a_k = b_k$ for all k.

Lemma 3.3 (Linear dependence lemma, LADR 2.19)**.** *If $v_1, \ldots, v_m$ is linearly dependent and $v_1 \neq 0$, then there exists an index $k \in \{2, \ldots, m\}$ such that $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$; moreover, removing v_k from the list leaves the span unchanged.*

Proof. By dependence there are scalars $a_1, \ldots, a_m$, not all zero, with $a_1 v_1 + \cdots + a_m v_m = 0$. Let k be the *largest* index with $a_k \neq 0$. If $k = 1$ we would have $a_1 v_1 = 0$ with $a_1 \neq 0$, hence $v_1 = 0$, contradicting our hypothesis; so $k \geq 2$. Solving for v_k:

$$v_k = -\frac{a_1}{a_k} v_1 - \cdots - \frac{a_{k-1}}{a_k} v_{k-1} \in \operatorname{span}(v_1, \ldots, v_{k-1}).$$

For the second claim: any combination using v_k can substitute this expression for v_k, producing the same vector as a combination of the others. Hence the span without v_k contains the span with it; the reverse containment is clear. □

Theorem 3.4 (Independent lists are no longer than spanning lists, LADR 2.22)**.** *In $\mathbb{R}^n$, if $u_1, \ldots, u_p$ is linearly independent and $w_1, \ldots, w_q$ spans $\mathbb{R}^n$ (or any subspace containing the u's), then $p \leq q$.*

Proof. We show that we can trade the w's for the u's one at a time, keeping a spanning list of length q at every step.

Step 1. The list $u_1, w_1, \ldots, w_q$ is linearly dependent (because u_1 is already in the span of the w's, it has a nontrivial representation, giving a dependence relation) and its first vector $u_1 \neq 0$ (independent lists cannot contain 0). By Lemma 3.3 we may remove some w_j — the lemma yields an index $k \geq 2$ whose vector lies in the span of its predecessors, and that vector must be a w, not a u, since the u's alone will always be independent — leaving a spanning list of length q containing u_1.

Step j. Suppose after $j - 1$ steps we have a spanning list of length q containing $u_1, \ldots, u_{j-1}$ (listed first) and $q - j + 1$ of the w's. Adjoin u_j right after $u_1, \ldots, u_{j-1}$: the resulting list of length $q + 1$ is dependent (u_j is in the span of a spanning list). By Lemma 3.3, some vector is in the span of its predecessors; it cannot be one of $u_1, \ldots, u_j$, since those are independent. So it is one of the remaining w's, and removing it keeps a spanning list of length q.

For this process to run through step p, there must be a w available to remove at each step; hence $q \geq p$. □

Definition 3.5 (Basis)**.** A *basis* of a subspace V is a list of vectors in V that is linearly independent and spans V.

Theorem 3.6 (Dimension is well defined, LADR 2.34)**.** *Any two bases of a subspace V have the same length. This common length is called the* dimension *of V, written* $\dim V$.

Proof. Let B_1 (length p) and B_2 (length q) be bases of V. Since B_1 is independent and B_2 spans V, Theorem 3.4 gives $p \leq q$. Swapping roles gives $q \leq p$. Hence $p = q$. □

Theorem 3.7 (Criterion for a basis)**.** *The list $v_1, \ldots, v_m$ is a basis of V if and only if every $v \in V$ can be written in* exactly one *way as $v = a_1 v_1 + \cdots + a_m v_m$.*

Proof. ($\Rightarrow$) Spanning gives at least one representation; independence gives at most one, by the uniqueness argument following the definition of independence. ($\Leftarrow$) Existence of representations for every v is spanning. Uniqueness applied to $v = 0$ says the only combination equal to 0 is the all-zeros one — independence. □

The *standard basis* of $\mathbb{R}^n$ is $e_1 = (1, 0, \ldots, 0), \ldots, e_n = (0, \ldots, 0, 1)$; hence $\dim \mathbb{R}^n = n$, and now "dimension" finally has a theorem-backed meaning matching the intuition: a line through 0 has dimension 1, a plane through 0 has dimension 2.

Why this matters. Theorem 3.7 is the licence to use *coordinates*: once you fix a basis, every vector *is* a unique list of numbers. The dimension theorem says "number of degrees of freedom" is intrinsic — it does not depend on which basis you happened to pick. These two facts are silently invoked every time machine learning says "represent the state as a vector in $\mathbb{R}^d$."

3.2 Practice: 2D/3D

Example 3.8 (Checking independence by hand)**.** Are $v_1 = (1, 2, 0)$, $v_2 = (0, 1, 1)$, $v_3 = (1, 4, 2)$ independent? Solve $av_1 + bv_2 + cv_3 = 0$:

$$\begin{aligned} a + c &= 0 \\ 2a + b + 4c &= 0 \implies a = -c, \quad b = -2c, \quad \text{2nd eq: } -2c - 2c + 4c = 0. \checkmark \\ b + 2c &= 0 \end{aligned}$$

The second equation is automatic, so c is free: taking $c = 1$ gives $-v_1 - 2v_2 + v_3 = 0$. *Dependent*, and indeed $v_3 = v_1 + 2v_2$: the third vector lies in the plane spanned by the first two, exactly as the linear dependence lemma predicts.

Example 3.9 (A basis of a plane)**.** The plane $P = \{(x_1, x_2, x_3) : x_1 + x_2 + x_3 = 0\}$ in $\mathbb{R}^3$ is a subspace (check closure!). The vectors $u = (1, -1, 0)$ and $v = (1, 0, -1)$ lie in P, are independent (neither is a multiple of the other), and span P: given (x_1, x_2, x_3) with $x_1 + x_2 + x_3 = 0$, write it as $-x_2 u - x_3 v$ — check: $-x_2(1, -1, 0) - x_3(1, 0, -1) = (-x_2 - x_3, x_2, x_3) = (x_1, x_2, x_3)$, using $x_1 = -x_2 - x_3$. So $\dim P = 2$: a plane, as geometry demands.

Machine-learning connection. The set of functions a linear model can represent is the span of its features: with features $\phi_1, \ldots, \phi_d$, the hypothesis class is $\{\sum_k w_k \phi_k\}$. Redundant (dependent) features do not enlarge the span — they add parameters without adding expressive power. "Add a feature to make the model more expressive" means "add a vector outside the current span."

3.3 Exercises

Exercise 3.1 (Hand)**.** Determine whether each list is independent; if dependent, exhibit a dependence relation. (a) $(2, 1), (4, 3)$ in $\mathbb{R}^2$. (b) $(1, 1, 1), (1, 2, 3), (0, 1, 2)$ in $\mathbb{R}^3$. (c) any list containing the zero vector.

Exercise 3.2 (Hand)**.** Find a basis for the line $\{(t, 2t, -t) : t \in \mathbb{R}\}$ and for the plane $x_1 - x_3 = 0$ in $\mathbb{R}^3$. State the dimension of each.

Exercise 3.3 (Proof, $\star$)**.** Prove that any list of $n + 1$ vectors in $\mathbb{R}^n$ is linearly dependent. (One line, given a theorem from this lesson.) Conclude that at most n features on a dataset of n-dimensional inputs can be "genuinely new" in the sense of enlarging the span at every step.

Exercise 3.4 (Proof)**.** Prove that $\operatorname{span}(v_1, \ldots, v_m)$ is a subspace, and that any subspace containing $v_1, \ldots, v_m$ contains $\operatorname{span}(v_1, \ldots, v_m)$.

Deeper reading. LADR 2A (span and linear independence), 2B (bases), 2C (dimension); Strang, ch. "Vector Spaces and Subspaces" (independence, basis and dimension).

4 Matrices and Linear Maps

NEEDED FOR: Machine learning (core)

4.1 Theory: linear maps first

Definition 4.1 (Linear map)**.** A function $T : \mathbb{R}^n \to \mathbb{R}^m$ is a *linear map* if for all $u, v \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$:

$$T(u+v) = Tu + Tv \qquad \text{and} \qquad T(\lambda v) = \lambda\, Tv.$$

Rotations about the origin, reflections through lines through the origin, scalings, shears, and projections onto subspaces are linear. Translation $v \mapsto v + b$ (for $b \neq 0$) is *not*: it moves 0.

Proposition 4.2. *If T is linear, then $T(0) = 0$.*

Proof. $T(0) = T(0+0) = T(0) + T(0)$; add $-T(0)$ to both sides. □

Theorem 4.3 (A linear map is determined by its values on a basis, LADR 3.4)**.** *Let $v_1, \dots, v_n$ be a basis of $\mathbb{R}^n$ and $w_1, \dots, w_n$ any vectors in $\mathbb{R}^m$. Then there exists a unique linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ with $Tv_k = w_k$ for each k.*

Proof. Existence. Every $x \in \mathbb{R}^n$ has a unique representation $x = a_1 v_1 + \cdots + a_n v_n$ (Theorem 3.7); define $Tx = a_1 w_1 + \cdots + a_n w_n$. This is well defined precisely because the representation is unique. Linearity: if $x = \sum a_k v_k$ and $y = \sum b_k v_k$, then $x + y = \sum (a_k + b_k) v_k$ is *the* representation of $x + y$, so $T(x+y) = \sum (a_k + b_k) w_k = Tx + Ty$; homogeneity is similar. Also $Tv_k = w_k$ by construction (take $a_k = 1$, others 0).

Uniqueness. If S is linear with $Sv_k = w_k$, then for $x = \sum a_k v_k$, linearity forces $Sx = \sum a_k S v_k = \sum a_k w_k = Tx$. □

> **The point.** A linear map on $\mathbb{R}^n$ is pinned down by n vectors' worth of information — its values on a basis. Choosing the standard basis, those values are the *columns of a matrix*. This theorem is why matrices and linear maps are interchangeable.

Definition 4.4 (Matrix of a map; matrix–vector product)**.** For linear $T : \mathbb{R}^n \to \mathbb{R}^m$, the *matrix of T* (with respect to standard bases) is the $m \times n$ array A whose kth column is $Te_k \in \mathbb{R}^m$. For $x = (x_1, \dots, x_n)$ we then have, by linearity,

$$Tx = T\Big(\sum_k x_k e_k\Big) = \sum_k x_k\, (Te_k),$$

which we *define* to be the product Ax. In words:

$$Ax = \text{the linear combination of the } \textit{columns} \text{ of } A \text{ with coefficients } x_1, \dots, x_n.$$

Entrywise, $(Ax)_i = \sum_k A_{i,k} x_k$: row i of A dotted with x. Both readings are used daily; the column reading is the more illuminating one.

Definition 4.5 (Matrix multiplication)**.** For linear maps $S : \mathbb{R}^n \to \mathbb{R}^m$ and $T : \mathbb{R}^p \to \mathbb{R}^n$, the composition $S \circ T : \mathbb{R}^p \to \mathbb{R}^m$ is linear (check!). The product AB of the matrix A of S ($m \times n$)

and the matrix B of T ($n \times p$) is *defined* as the matrix of $S \circ T$; it is $m \times p$. Column k of AB is $(S \circ T)e_k = A(Be_k) = A\,(\text{column } k \text{ of } B)$, which yields the familiar formula

$$(AB)_{i,j} = \sum_{k=1}^{n} A_{i,k} B_{k,j}.$$

Because composition of functions is associative but not commutative, matrix multiplication is associative — $(AB)C = A(BC)$, no computation needed — and in general $AB \neq BA$.

Definition 4.6 (Transpose)**.** The *transpose* of an $m \times n$ matrix A is the $n \times m$ matrix A^{T} with $(A^{\mathsf{T}})_{i,j} = A_{j,i}$. Two facts used constantly: $(AB)^{\mathsf{T}} = B^{\mathsf{T}}A^{\mathsf{T}}$, and the dot product is a matrix product: $x \cdot y = x^{\mathsf{T}}y$. The defining property tying transpose to the dot product is

$$\langle Ax,\, y\rangle = \langle x,\, A^{\mathsf{T}}y\rangle \qquad \text{for all } x \in \mathbb{R}^n,\ y \in \mathbb{R}^m,$$

which follows by expanding both sides as $\sum_{i,k} A_{i,k}x_k y_i$.

4.2 Practice: 2D matrices you can see

Example 4.7 (Reading a matrix by its columns)**.**

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}: \qquad Ae_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad Ae_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

$e_1 \mapsto e_2$ and $e_2 \mapsto -e_1$: this is rotation by 90° counterclockwise. In general, rotation by angle θ has matrix $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ — its columns are where e_1, e_2 land, which is all Theorem 4.3 says you need to know.

Example 4.8 (Hand computation, both readings)**.**

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 1 \end{pmatrix}, \qquad x = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Column reading: $Ax = 2\begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix} - 1\begin{pmatrix} 2 \\ 4 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}$. Row reading: $(Ax)_1 = 1 \cdot 2 + 2 \cdot (-1) = 0$, $(Ax)_2 = 6 - 4 = 2$, $(Ax)_3 = 0 - 1 = -1$. Same answer, as it must be.

Example 4.9 (Composition order matters)**.** Let R = rotation by 90° (above) and $S = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ = projection onto the x_1-axis. Then

$$SR = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}, \qquad RS = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Track e_1 through each composition (rightmost map acts first): $SR\,e_1 = S(Re_1) = S(e_2) = 0$, while $RS\,e_1 = R(Se_1) = R(e_1) = e_2$. Rotate-then-project kills e_1; project-then-rotate sends it to e_2. Different maps, different matrices: $AB \neq BA$.

Machine-learning connection. A dataset of m examples with n features is an $m \times n$ matrix X whose *rows* are feature vectors. Then Xw computes all m scores $w \cdot \phi(x_i)$ in one matrix–vector product. A tiny neural network layer is $x \mapsto \sigma(Wx + b)$: a linear map, a shift, a nonlinearity. Reading Ax fluently in both the row picture and the column picture is the single highest-value matrix skill for the course.

4.3 Exercises

Exercise 4.1 (Hand)**.** Write down the 2×2 matrices of: (a) reflection through the line $x_2 = x_1$; (b) scaling by 3 in the x_1 direction only; (c) rotation by 180°. (Use the columns-are-images-of-e_k principle; no formulas to memorize.)

Exercise 4.2 (Hand)**.** With $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ (a *shear*) and $B = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, compute AB, BA, and $A^{\mathsf{T}}B^{\mathsf{T}}$. Check $(BA)^{\mathsf{T}} = A^{\mathsf{T}}B^{\mathsf{T}}$.

Exercise 4.3 (Proof, $\star$)**.** Prove that the composition of linear maps is linear, and that $\langle Ax,\, y\rangle = \langle x,\, A^{\mathsf{T}}y\rangle$ for all x, y by expanding both sides in coordinates.

Exercise 4.4 (Proof)**.** Using Theorem 4.3, prove: if two linear maps $S, T : \mathbb{R}^n \to \mathbb{R}^m$ agree on a basis, they are equal. Then explain in one sentence why an $m \times n$ matrix and a linear map $\mathbb{R}^n \to \mathbb{R}^m$ carry exactly the same information.

Deeper reading. LADR 3A (linear maps), 3C (matrices, matrix multiplication); Strang, chs. "Multiplying Matrices" / "Solving Linear Equations" (the idea of elimination, matrix operations).

5 Null Space, Rank, and Solving $Ax = b$

NEEDED FOR: Machine learning (core)

5.1 Theory: the fundamental theorem

Throughout, $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear with matrix A.

Definition 5.1 (Null space and range)**.**

$$\operatorname{null} T = \{x \in \mathbb{R}^n : Tx = 0\}, \qquad \operatorname{range} T = \{Tx : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m.$$

For the matrix A, $\operatorname{range} T$ is the span of the columns of A (by the column reading of Ax), also called the *column space*; $\operatorname{null} T$ is the solution set of $Ax = 0$. Both are subspaces (Exercise 5.4). The *rank* of A is $\operatorname{rank} A = \dim \operatorname{range} T$.

Proposition 5.2 (Injectivity ⇔ trivial null space, LADR 3.15)**.** *T is injective (one-to-one) if and only if* $\operatorname{null} T = \{0\}$.

Proof. ($\Rightarrow$) $T0 = 0$; if $Tx = 0$ too, injectivity gives $x = 0$. ($\Leftarrow$) If $Tu = Tv$, linearity gives $T(u - v) = 0$, so $u - v \in \operatorname{null} T = \{0\}$, i.e. $u = v$. □

This tiny proposition is used constantly: to check whether a linear map ever collapses two inputs together, you only have to look at what it sends to zero.

Theorem 5.3 (Fundamental theorem of linear maps / rank–nullity, LADR 3.21)**.** *For linear* $T : \mathbb{R}^n \to \mathbb{R}^m$*,*

$$\dim \operatorname{null} T + \dim \operatorname{range} T = n.$$

Proof. Let $u_1, \dots, u_k$ be a basis of $\operatorname{null} T$ (so $\dim \operatorname{null} T = k$). Extend it to a basis $u_1, \dots, u_k, v_1, \dots, v_r$ of $\mathbb{R}^n$ (add vectors outside the current span one at a time until spanning; independence is preserved at each step by the linear dependence lemma). Then $n = k + r$, and it suffices to show $Tv_1, \dots, Tv_r$ is a basis of $\operatorname{range} T$.

Spanning: any $x \in \mathbb{R}^n$ is $x = \sum a_i u_i + \sum b_j v_j$, so

$$Tx = \sum a_i \underbrace{Tu_i}_{=0} + \sum b_j Tv_j = \sum b_j Tv_j,$$

hence every vector of $\operatorname{range} T$ is a combination of the Tv_j.

Independence: suppose $\sum_j c_j Tv_j = 0$. By linearity $T\big(\sum_j c_j v_j\big) = 0$, so $\sum_j c_j v_j \in \operatorname{null} T$, hence $\sum_j c_j v_j = \sum_i d_i u_i$ for some scalars d_i. Rearranged, $\sum_j c_j v_j - \sum_i d_i u_i = 0$ is a linear relation on the basis of $\mathbb{R}^n$, so all c_j (and d_i) are zero. □

Corollary 5.4 (The square case is all-or-nothing)**.** *For* $T : \mathbb{R}^n \to \mathbb{R}^n$*: injective* $\Leftrightarrow$ *surjective* $\Leftrightarrow$ *invertible.*

Proof. By Theorem 5.3, $\dim \operatorname{null} T = 0 \iff \dim \operatorname{range} T = n \iff \operatorname{range} T = \mathbb{R}^n$ (a subspace of $\mathbb{R}^n$ of dimension n must be $\mathbb{R}^n$: take a basis of the subspace; it is an independent list of length n in $\mathbb{R}^n$, and extending it to a basis of $\mathbb{R}^n$ cannot add anything by Theorem 3.4). So injective $\iff$ surjective, and a bijective linear map has a linear inverse (Exercise 5.5). A square matrix A is *invertible* when such A^{-1} exists: $A^{-1}A = AA^{-1} = I$. □

What the fundamental theorem says. An $m \times n$ matrix consumes n input dimensions. Each dimension either survives into the output (contributing to rank) or is crushed to zero (contributing to the null space) — and nothing is lost or created: crushed + surviving $= n$. Consequences you can now read off instantly: a map to a smaller space ($m < n$) must crush something ($\dim \operatorname{null} T \geq n - m > 0$), so it cannot be injective; a map from a smaller space ($n < m$) has rank at most $n < m$, so it cannot be surjective.

Solving $Ax = b$**.** The complete picture, now provable in three lines:

1. $Ax = b$ has a solution $\iff$ $b \in \operatorname{range} T$ (that is what range means).
2. If x_p is one particular solution, the full solution set is $x_p + \operatorname{null} T = \{x_p + h : Ah = 0\}$: for if $Ax = b = Ax_p$ then $A(x - x_p) = 0$; conversely $A(x_p + h) = b + 0 = b$.
3. Hence the solution is unique iff $\operatorname{null} T = \{0\}$; for square A, Corollary 5.4 says a solution then exists for every b, namely $x = A^{-1}b$.

5.2 Practice: elimination by hand

Gaussian elimination is Strang's centerpiece: subtract multiples of rows to reach triangular form, then back-substitute. Row operations do not change the solution set (each is reversible and preserves the equations).

Example 5.5 (A unique solution)**.** Solve $Ax = b$:

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 4 \\ 1 & 1 & 0 \end{pmatrix} x = \begin{pmatrix} 3 \\ 8 \\ 1 \end{pmatrix}.$$

$R_2 \leftarrow R_2 - 2R_1$ and $R_3 \leftarrow R_3 - R_1$:

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix} x = \begin{pmatrix} 3 \\ 2 \\ -2 \end{pmatrix}; \qquad R_3 \leftarrow R_3 + R_2: \quad \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} x = \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix}.$$

Back-substitute: $x_3 = 0$, then $x_2 = 2 - 2x_3 = 2$, then $x_1 = 3 - 2x_2 - x_3 = -1$. Solution $x = (-1, 2, 0)$; three pivots, so rank 3, null space $\{0\}$, solution unique — the theory and the algorithm agree.

Example 5.6 (Null space by hand)**.** Find $\operatorname{null} A$ for $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix}$. The second row is twice the first, so $Ax = 0$ reduces to the single equation $x_1 + 2x_2 + 3x_3 = 0$: solve $x_1 = -2x_2 - 3x_3$ with x_2, x_3 free. Null space $= \operatorname{span}\big((-2, 1, 0), (-3, 0, 1)\big)$, dimension 2. Rank check: the columns of A are all multiples of $(1, 2)$, so rank $= 1$, and $1 + 2 = 3 = n$. ✓

> **Machine-learning connection.** "Does a weight vector exist that fits these constraints, and is it unique?" is exactly existence/uniqueness for a linear system. A nontrivial null space of the feature matrix means different weight vectors produce identical predictions on the training set — the model is over-parameterized on that data. Rank tells you the number of effective directions your data spans.

5.3 Exercises

Exercise 5.1 (Hand)**.** Solve by elimination, tracking pivots: $\left(\begin{smallmatrix} 2 & 1 \\ 4 & 3 \end{smallmatrix}\right) x = \left(\begin{smallmatrix} 5 \\ 11 \end{smallmatrix}\right)$, then $\left(\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right) x = \left(\begin{smallmatrix} 3 \\ 6 \end{smallmatrix}\right)$, then the same left side with right side $(3, 7)$. Describe all solutions in each case (unique / infinitely many / none) and reconcile with rank and the three-step picture above.

Exercise 5.2 (Hand)**.** For $A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix}$, find a basis of $\operatorname{null} A$ and a basis of the column space; verify rank–nullity.

Exercise 5.3 (Proof, $\star$)**.** Prove Proposition 5.2's consequence for systems: $Ax = b$ has *at most one* solution for every b if and only if $Ax = 0$ has only the zero solution.

Exercise 5.4 (Proof)**.** Prove that $\operatorname{null} T$ and $\operatorname{range} T$ are subspaces.

Exercise 5.5 (Proof)**.** Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is linear and bijective. Prove that its inverse function T^{-1} is linear. (Apply T to both sides of what you want to show.)

Deeper reading. LADR 3B (null spaces and ranges, fundamental theorem), 3D (invertibility); Strang, chs. "Solving Linear Equations" (elimination) and "Vector Spaces and Subspaces" (the null space, rank).

6 Orthogonal Projections and Least Squares

NEEDED FOR: Machine learning (core); reused in Lesson 19 for MMSE estimation

6.1 Theory

Definition 6.1 (Orthonormal list; orthogonal complement)**.** A list $e_1, \dots, e_k$ is *orthonormal* if $\langle e_i, e_j \rangle = 0$ for $i \neq j$ and $\|e_i\| = 1$ for all i. For a subspace $U \subseteq \mathbb{R}^n$, the *orthogonal complement* is

$$U^\perp = \{x \in \mathbb{R}^n : \langle x, u \rangle = 0 \text{ for all } u \in U\}.$$

Proposition 6.2 (Orthonormal lists are independent, LADR 6.25)**.** *Every orthonormal list is linearly independent, and if $v = a_1 e_1 + \cdots + a_k e_k$ then $a_j = \langle v, e_j \rangle$.*

Proof. Take the inner product of $v = \sum_i a_i e_i$ with e_j: by linearity, $\langle v, e_j \rangle = \sum_i a_i \langle e_i, e_j \rangle = a_j$, since all terms vanish except $i = j$, where $\langle e_j, e_j \rangle = 1$. For independence, apply this with $v = 0$: every coefficient equals $\langle 0, e_j \rangle = 0$. □

So orthonormal bases make coordinates *free*: no system-solving, just k dot products.

Theorem 6.3 (Orthogonal projection)**.** *Let $e_1, \dots, e_k$ be an orthonormal basis of a subspace $U \subseteq \mathbb{R}^n$ (one always exists: apply the Gram–Schmidt procedure below to any basis). Define $P_U : \mathbb{R}^n \to \mathbb{R}^n$ by*

$$P_U x = \langle x, e_1 \rangle \, e_1 + \cdots + \langle x, e_k \rangle \, e_k.$$

Then for every x: (a) $P_U x \in U$ and $x - P_U x \in U^\perp$; (b) $P_U x$ is the unique closest point of U to x:

$$\|x - P_U x\| < \|x - u\| \qquad \text{for every } u \in U \text{ with } u \neq P_U x.$$

Proof. (a) $P_U x \in U$ since it is a combination of the e_i. For the second part, for each j,

$$\langle x - P_U x, e_j \rangle = \langle x, e_j \rangle - \sum_i \langle x, e_i \rangle \langle e_i, e_j \rangle = \langle x, e_j \rangle - \langle x, e_j \rangle = 0,$$

and a vector orthogonal to each e_j is orthogonal to every combination of them, i.e. to all of U.

(b) Let $u \in U$. Write

$$x - u = \underbrace{(x - P_U x)}_{\in U^\perp} + \underbrace{(P_U x - u)}_{\in U}.$$

The two pieces are orthogonal, so by the Pythagorean theorem

$$\|x - u\|^2 = \|x - P_U x\|^2 + \|P_U x - u\|^2 \geq \|x - P_U x\|^2,$$

with equality iff $\|P_U x - u\| = 0$, i.e. $u = P_U x$. □

One picture to keep. Drop a perpendicular from x to the subspace U; its foot is $P_U x$. The proof of (b) is the Pythagorean theorem applied to the right triangle with vertices x, $P_U x$, and any other candidate u. Every least-squares fact below is this picture.

Gram–Schmidt in brief (LADR 6.32). Given an independent list $v_1, \dots, v_k$, define $e_1 = v_1/\|v_1\|$ and, inductively, subtract from v_j its projection onto the span of the previous e's, then normalize:

$$f_j = v_j - \sum_{i<j} \langle v_j,\, e_i \rangle\, e_i, \qquad e_j = f_j/\|f_j\|.$$

Each $f_j \neq 0$ (else v_j would be in the span of its predecessors, contradicting independence), each e_j is orthogonal to the previous ones by the same computation as in Theorem 6.3(a), and the spans agree at every stage.

Theorem 6.4 (Least squares / normal equations)**.** *Let A be $m \times n$ with linearly independent columns, and $b \in \mathbb{R}^m$. Then the minimization problem*

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|^2$$

has the unique solution given by the normal equations

$$A^{\mathsf{T}} A\, \hat{x} = A^{\mathsf{T}} b.$$

Proof. As x ranges over $\mathbb{R}^n$, the vector Ax ranges exactly over $U = \operatorname{range} A$ (the column space). By Theorem 6.3(b), the closest point of U to b is the unique vector $P_U b$; so minimizers are exactly the solutions of $A\hat{x} = P_U b$. Independence of the columns means A is injective on $\mathbb{R}^n$ (Proposition 5.2 applied to combinations of columns), so exactly one $\hat{x}$ satisfies this.

It remains to convert "$A\hat{x}$ is the projection of b" into equations. By Theorem 6.3(a), $A\hat{x} = P_U b$ iff the residual $b - A\hat{x}$ is orthogonal to U, i.e. orthogonal to every column of A, i.e.

$$A^{\mathsf{T}}(b - A\hat{x}) = 0 \iff A^{\mathsf{T}} A\, \hat{x} = A^{\mathsf{T}} b.$$

(For the converse direction of the iff: if $A^{\mathsf{T}}(b - A\hat{x}) = 0$, then $b - A\hat{x}$ is orthogonal to each column of A, hence to all of U, and adding $A\hat{x} \in U$ shows $A\hat{x}$ is the orthogonal projection of b, by the uniqueness of the decomposition $b = P_U b + (b - P_U b)$ into a U-part plus a $U^{\perp}$-part.)

Finally, $A^{\mathsf{T}} A$ is invertible when A has independent columns: if $A^{\mathsf{T}} A x = 0$ then $0 = \langle A^{\mathsf{T}} A x,\, x \rangle = \langle Ax,\, Ax \rangle = \|Ax\|^2$, so $Ax = 0$, so $x = 0$ by injectivity; now apply Corollary 5.4 to the square matrix $A^{\mathsf{T}} A$. □

6.2 Practice: fitting a line by hand

Example 6.5 (Least-squares line through three points)**.** Fit $y \approx c + dt$ to the data points $(t, y) = (0, 1), (1, 2), (2, 4)$. In matrix form, we want $Ax \approx b$ with unknowns $x = (c, d)$:

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}.$$

No exact solution exists (the three points are not collinear — check: slopes 1 and 2). Normal equations:

$$A^{\mathsf{T}}A = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}, \quad A^{\mathsf{T}}b = \begin{pmatrix} 7 \\ 10 \end{pmatrix} \implies \begin{cases} 3c + 3d = 7 \\ 3c + 5d = 10 \end{cases} \implies d = \tfrac{3}{2},\ c = \tfrac{7-3d}{3} = \tfrac{5}{6}.$$

Best line: $y = \frac{5}{6} + \frac{3}{2}t$. Residual: $b - A\hat{x} = (1, 2, 4) - (\frac{5}{6}, \frac{7}{3}, \frac{23}{6}) = (\frac{1}{6}, -\frac{1}{3}, \frac{1}{6})$. Sanity checks: the residual sums to zero (orthogonal to the all-ones first column) and $\langle \text{residual}, (0, 1, 2) \rangle = 0 - \frac{1}{3} + \frac{1}{3} = 0$ (orthogonal to the second column). The residual is orthogonal to the column space — the picture made numerical.

Machine-learning connection. Linear regression with the squared loss *is* this theorem: training loss $= \frac{1}{m}\|Xw - y\|^2$, and the closed-form minimizer solves $X^{\mathsf{T}}Xw = X^{\mathsf{T}}y$. Machine learning will usually minimize by gradient descent instead (Lesson 7) — but knowing the geometry (prediction = projection of the label vector onto the span of the features; residual $\perp$ features) is what makes regression make sense.

6.3 Exercises

Exercise 6.1 (Hand)**.** Run Gram–Schmidt on $v_1 = (1, 1, 0)$, $v_2 = (1, 0, 1)$ to produce an orthonormal basis e_1, e_2 of their span U. Then compute $P_U x$ for $x = (0, 0, 3)$ and verify $x - P_U x$ is orthogonal to both e_1 and e_2.

Exercise 6.2 (Hand)**.** Fit the least-squares line to $(0, 0), (1, 1), (2, 1), (3, 2)$ via the normal equations. Verify the residual is orthogonal to both columns of your A.

Exercise 6.3 (Proof, $\star$)**.** Prove that $\mathbb{R}^n = U \oplus U^{\perp}$ for every subspace U; that is, every x is uniquely a sum of a vector in U and a vector in $U^{\perp}$. (Existence: Theorem 6.3(a). Uniqueness: show $U \cap U^{\perp} = \{0\}$ — what is $\langle v, v \rangle$ for v in both?)

Exercise 6.4 (Proof)**.** Show $\langle P_U x, y \rangle = \langle x, P_U y \rangle$ for all x, y (projections are self-adjoint), and $P_U(P_U x) = P_U x$ (idempotent), directly from the formula in Theorem 6.3.

Deeper reading. LADR 6B (orthonormal bases, Gram–Schmidt), 6C (orthogonal complements, minimization — Axler solves this exact least-squares problem); Strang, ch. "Orthogonality" (projections; least squares approximations).

7 Gradients for Machine Learning

NEEDED FOR: Machine learning (core)

This lesson is the bridge from linear algebra to the opening machine-learning lectures. It is multivariable differentiation in the minimal dose: gradients of the two function shapes machine learning actually optimizes.

7.1 Theory

Definition 7.1 (Gradient)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$. The *partial derivative* $\frac{\partial f}{\partial x_k}(x)$ is the ordinary derivative of $t \mapsto f(x + te_k)$ at $t = 0$. If all partials exist and are continuous, the *gradient* is the vector

$$\nabla f(x) = \Big(\tfrac{\partial f}{\partial x_1}(x), \dots, \tfrac{\partial f}{\partial x_n}(x)\Big) \in \mathbb{R}^n,$$

and f admits the first-order approximation $f(x + h) \approx f(x) + \langle \nabla f(x),\, h\rangle$ for small h.

Proposition 7.2 (Steepest ascent)**.** *Suppose $\nabla f(x) \neq 0$. Among all unit vectors u, the directional rate of change $\langle \nabla f(x),\, u\rangle$ is maximized uniquely at $u = \nabla f(x)/\|\nabla f(x)\|$ and minimized uniquely at its negative.*

Proof. By Cauchy–Schwarz (Theorem 2.5), $|\langle \nabla f(x),\, u\rangle| \le \|\nabla f(x)\|\,\|u\| = \|\nabla f(x)\|$, with equality iff u is a scalar multiple of $\nabla f(x)$; the unit multiples are $\pm\nabla f(x)/\|\nabla f(x)\|$, giving the maximum with sign $+$ and the minimum with sign $-$. □

That is the entire justification for *gradient descent*: step against the gradient,

$$x^{(t+1)} = x^{(t)} - \eta\, \nabla f\big(x^{(t)}\big),$$

because $-\nabla f$ is locally the fastest way downhill. ($\eta > 0$ is the *step size* or *learning rate*.)

Proposition 7.3 (The two gradients to memorize)**.** *Fix $a \in \mathbb{R}^n$ and a symmetric $n \times n$ matrix M (meaning $M^{\mathsf{T}} = M$). Then*

$$\textit{(i)}\ \nabla_x \langle a,\, x\rangle = a; \qquad \textit{(ii)}\ \nabla_x\big(x^{\mathsf{T}} M x\big) = 2Mx.$$

Proof. (i) $\langle a,\, x\rangle = \sum_k a_k x_k$, so $\partial/\partial x_k$ picks out a_k; assembling the partials gives a.

(ii) Write $g(x) = x^{\mathsf{T}} M x = \sum_{i,j} M_{i,j} x_i x_j$. Differentiating with respect to x_k: the terms containing x_k are those with $i = k$ or $j = k$ (the term $i = j = k$ is $M_{k,k} x_k^2$, derivative $2M_{k,k}x_k$, consistent with counting it in both groups once each):

$$\frac{\partial g}{\partial x_k} = \sum_j M_{k,j} x_j + \sum_i M_{i,k} x_i = (Mx)_k + (M^{\mathsf{T}} x)_k = 2\,(Mx)_k,$$

using symmetry in the last step. Assembling over k gives $2Mx$. □

Corollary 7.4 (Gradient of the squared-error loss)**.** *For $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, the function $f(x) = \|Ax - b\|^2$ has*

$$\nabla f(x) = 2A^{\mathsf{T}}(Ax - b).$$

Consequently $\nabla f(\hat{x}) = 0 \iff A^{\mathsf{T}} A\hat{x} = A^{\mathsf{T}} b$: the normal equations of Lesson 6 are exactly the vanishing-gradient condition.

Proof. Expand using $\|v\|^2 = \langle v,\, v\rangle$, linearity, and $\langle Ax,\, b\rangle = \langle x,\, A^{\mathsf{T}} b\rangle$ (Lesson 4):

$$f(x) = \langle Ax - b,\, Ax - b\rangle = x^{\mathsf{T}}(A^{\mathsf{T}}A)x - 2\langle A^{\mathsf{T}} b,\, x\rangle + \|b\|^2.$$

The matrix $A^{\mathsf{T}}A$ is symmetric since $(A^{\mathsf{T}}A)^{\mathsf{T}} = A^{\mathsf{T}}A$. Now apply Proposition 7.3: the quadratic term contributes $2A^{\mathsf{T}}Ax$, the linear term $-2A^{\mathsf{T}}b$, the constant nothing. Sum: $2A^{\mathsf{T}}(Ax - b)$. □

Two roads, one summit. Lesson 6 found the least-squares solution by geometry (projection: residual $\perp$ columns). This lesson found it by calculus (set the gradient to zero). They agree — $A^{\mathsf{T}}(A\hat{x} - b) = 0$ either way — and machine learning takes a third road to the same place: iterate $w \leftarrow w - \eta \nabla f(w)$, which for this convex f converges to the same $\hat{x}$ (for suitably small η). Recognizing all three as one is the payoff of this booklet.

7.2 Practice: by hand

Example 7.5 (A full gradient computation)**.** $f(w_1, w_2) = (3w_1 + 2w_2 - 5)^2$. Chain rule per coordinate: with $r = 3w_1 + 2w_2 - 5$,

$$\frac{\partial f}{\partial w_1} = 2r \cdot 3, \qquad \frac{\partial f}{\partial w_2} = 2r \cdot 2, \qquad \nabla f = 2r\,(3, 2).$$

This matches Corollary 7.4 with $A = (3 \;\; 2)$ (one row), $b = 5$: $\nabla f = 2A^{\mathsf{T}}(Aw - b)$. Note the gradient is $2r$ times the feature vector $(3, 2)$ — *residual times features*, the shape of every gradient-descent update in machine learning's learning lectures.

Example 7.6 (Two steps of gradient descent by hand)**.** Minimize $f(w) = (w_1 - 1)^2 + 4(w_2)^2$, i.e. $w^{\mathsf{T}} M w - 2w_1 + 1$ with $M = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 4 \end{smallmatrix}\right)$; so $\nabla f(w) = (2(w_1 - 1),\, 8w_2)$. Start at $w^{(0)} = (0, 1)$ with $\eta = 0.1$:

$$\nabla f(w^{(0)}) = (-2, 8) \;\Rightarrow\; w^{(1)} = (0, 1) - 0.1\,(-2, 8) = (0.2, 0.2);$$

$$\nabla f(w^{(1)}) = (-1.6, 1.6) \;\Rightarrow\; w^{(2)} = (0.2, 0.2) - 0.1(-1.6, 1.6) = (0.36, 0.04).$$

Both coordinates march toward the minimizer $(1, 0)$ — the steep w_2 direction (curvature 4) converges faster at this η, and would oscillate if η were too large (check: the w_2 update is $w_2 \leftarrow (1 - 0.8)\, w_2$ here, and $\eta > 0.25$ makes $|1 - 8\eta| > 1$). Step-size sensitivity is a lived experience in machine learning homework; this two-line example is its essence.

Machine-learning connection. The learning lectures define the training loss as an average of per-example losses,

$$\text{TrainLoss}(w) = \tfrac{1}{m} \textstyle\sum_i \text{Loss}(x_i, y_i, w),$$

and minimize it by (stochastic) gradient descent. For the squared loss, $\nabla_w \text{Loss} = 2\,(w \cdot \phi(x) - y)\,\phi(x)$: residual times feature vector, exactly the pattern above. For the hinge loss on a misclassified example, $\nabla_w \text{Loss} = -y\,\phi(x)$. Every update you will implement has the form $w \leftarrow w - \eta\,(\text{scalar})\,\phi(x)$: a scalar–vector product plus a vector subtraction, both $O(n)$: one cheap pass over the features per example, which is why SGD scales to huge datasets.

7.3 Exercises

Exercise 7.1 (Hand)**.** Compute ∇f for: (a) $f(w) = \langle (1, -2, 3),\, w \rangle$; (b) $f(w) = \|w\|^2$; (c) $f(w) = (w \cdot \phi - y)^2$ for fixed $\phi \in \mathbb{R}^n$, $y \in \mathbb{R}$ (answer in terms of the residual).

Exercise 7.2 (Hand)**.** Run two iterations of gradient descent on $f(w) = (2w_1 + w_2 - 3)^2$ from $w^{(0)} = (0, 0)$ with $\eta = 0.1$, by hand. Does the loss decrease at each step? (Compute it.)

Exercise 7.3 (Proof, $\star$)**.** Prove Corollary 7.4 a second way, without Proposition 7.3: expand $f(x+h) = \|A(x+h)-b\|^2$ into $f(x) + \langle g,\, h\rangle + \|Ah\|^2$ for an explicit vector g, and identify $\nabla f(x) = g$ from the first-order approximation. (This perturbation technique is how gradients are computed cleanly in general.)

Exercise 7.4 (Proof)**.** Show that for symmetric M, the function $g(x) = x^{\mathsf{T}}Mx$ satisfies $g(x+h) - g(x) - \langle 2Mx,\, h\rangle = h^{\mathsf{T}}Mh$ *exactly* (no approximation), and explain how this re-proves Proposition 7.3(ii).

Deeper reading. Strang, ch. "Orthogonality" again (least squares) — and for the calculus, any multivariable-calculus text; introductory machine-learning problem sets exercise exactly these gradient patterns.

Part II: Eigenstructure, the Spectral Theorem, the DFT, and the SVD

8 Complex Vectors and Complex Exponentials

NEEDED FOR: Fourier analysis (core); statistical-inference background. Not needed for ML.

Fourier analysis lives over the complex numbers: the basic signals are the complex exponentials $e^{i\omega t}$, and the clean statements of orthogonality, eigenvalues, and the DFT all require $\mathbb{C}$. This lesson upgrades Lessons 1–2 from $\mathbb{R}$ to $\mathbb{C}$.

8.1 Theory

Recall $\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$ with $i^2 = -1$; the *conjugate* of $z = a + bi$ is $\bar{z} = a - bi$ and the *modulus* is $|z| = \sqrt{a^2 + b^2} = \sqrt{z\bar{z}}$. Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$ says $e^{i\theta}$ is the point on the unit circle at angle θ; thus $|e^{i\theta}| = 1$, $\overline{e^{i\theta}} = e^{-i\theta}$, and $e^{i\theta}e^{i\varphi} = e^{i(\theta+\varphi)}$ (angles add — multiplication by $e^{i\theta}$ is rotation by θ).

The space $\mathbb{C}^n$ is defined exactly like $\mathbb{R}^n$ with complex coordinates and complex scalars; everything in Lessons 1 and 3–5 (span, independence, basis, dimension, linear maps, rank–nullity) goes through verbatim with $\mathbb{F} = \mathbb{C}$ — the proofs never used anything about $\mathbb{R}$ beyond field arithmetic. The one definition that must change is the inner product.

Definition 8.1 (Inner product on $\mathbb{C}^n$)**.** For $x, y \in \mathbb{C}^n$,

$$\langle x,\, y\rangle = x_1\overline{y_1} + \cdots + x_n\overline{y_n}.$$

Why the conjugate? Because we need $\langle x,\, x\rangle = \sum_k x_k\overline{x_k} = \sum_k |x_k|^2$ to be a nonnegative real number, so that $\|x\| = \sqrt{\langle x,\, x\rangle}$ is a genuine length. Without the conjugate, $\langle x,\, x\rangle$ could be negative or non-real (try $x = (i)$ in $\mathbb{C}^1$: $i \cdot i = -1$).

The price is that symmetry becomes *conjugate symmetry* and linearity holds only in the first slot:

$$\langle y,\, x\rangle = \overline{\langle x,\, y\rangle}, \qquad \langle \lambda x,\, y\rangle = \lambda\langle x,\, y\rangle, \qquad \langle x,\, \lambda y\rangle = \bar{\lambda}\,\langle x,\, y\rangle.$$

(Each is a one-line check.) With these amendments, every proof of Lesson 2 survives unchanged in structure: positivity, the orthogonal decomposition lemma, Cauchy–Schwarz $|\langle x,\, y\rangle| \le \|x\|\|y\|$, the triangle inequality, and the Pythagorean theorem for orthogonal vectors ($\langle x,\, y\rangle = 0$) all hold in $\mathbb{C}^n$, and Lesson 6's orthonormal-basis and projection theory holds verbatim with $\langle \cdot,\, \cdot\rangle$ in this complex sense. This is LADR 6A's actual setting.

Lemma 8.2 (Roots of unity; the key to the DFT)**.** *Fix an integer $N \ge 1$ and let $\omega = e^{-2\pi i/N}$. Then $\omega^N = 1$, and for any integer m:*

$$\sum_{j=0}^{N-1} \omega^{jm} = \begin{cases} N & \text{if } N \mid m, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. $\omega^N = e^{-2\pi i} = 1$. If $N \mid m$, then $\omega^m = (\omega^N)^{m/N} = 1$ and every term of the sum is 1: total N. If $N \nmid m$, then $\omega^m \neq 1$ (since $\omega^m = e^{-2\pi i m/N} = 1$ exactly when m/N is an integer), so the finite geometric series gives

$$\sum_{j=0}^{N-1} (\omega^m)^j = \frac{(\omega^m)^N - 1}{\omega^m - 1} = \frac{(\omega^N)^m - 1}{\omega^m - 1} = \frac{1-1}{\omega^m - 1} = 0. \qquad \square$$

Geometrically: the numbers ω^{jm} for $j = 0, \ldots, N-1$ are equally spaced points on the unit circle (when $N \nmid m$), and equally spaced points on a circle sum to zero by symmetry. The lemma is that picture made algebra, and Lesson 11 will use it to prove that the Fourier basis is orthonormal.

8.2 Practice: by hand

Example 8.3 (Complex arithmetic warm-up)**.** $(1+i)^2 = 1 + 2i + i^2 = 2i$. Also $|1+i| = \sqrt{2}$ and $1 + i = \sqrt{2}\, e^{i\pi/4}$, so $(1+i)^2 = 2e^{i\pi/2} = 2i$: polar form turns powers into arithmetic on angles. Similarly $(1+i)^8 = 16\, e^{2\pi i} = 16$.

Example 8.4 (Fourth roots of unity)**.** For $N = 4$: $\omega = e^{-2\pi i/4} = -i$, and the powers $\omega^0, \omega^1, \omega^2, \omega^3 = 1, -i, -1, i$: the four compass points of the unit circle. Check the lemma for $m = 1$: $1 - i - 1 + i = 0$. ✓ For $m = 4$: $1 + 1 + 1 + 1 = 4$. ✓

Example 8.5 (A complex inner product)**.** $x = (1, i)$, $y = (i, 1)$ in $\mathbb{C}^2$: $\langle x, y\rangle = 1 \cdot \bar{i} + i \cdot \bar{1} = -i + i = 0$ — orthogonal. And $\|x\|^2 = |1|^2 + |i|^2 = 2$. Note $\langle y, x\rangle = \bar{0} = 0$ too, but in general the order matters up to conjugation.

Fourier connection. Fourier analysis's protagonist is $e^{2\pi i s t}$. Fourier series write a periodic signal as $f(t) = \sum_k c_k e^{2\pi i k t}$, with coefficients $c_k = \int_0^1 f(t)\, \overline{e^{2\pi i k t}}\, dt$ — read that as $\langle f, e_k\rangle$ for the inner product $\langle f, g\rangle = \int_0^1 f\bar{g}$ on functions. It is literally Proposition 6.2's formula $a_j = \langle v, e_j\rangle$, with integrals replacing sums: Fourier coefficients are coordinates in an orthonormal basis, and partial Fourier sums are orthogonal projections (Theorem 6.3) onto the span of finitely many exponentials — hence the best least-squares approximants. Carrying that dictionary into Fourier analysis makes the first three weeks feel familiar.

8.3 Exercises

Exercise 8.1 (Hand)**.** Write $z = -1 + i\sqrt{3}$ in polar form $re^{i\theta}$, then compute z^3 two ways (polar and direct expansion) and confirm they agree.

Exercise 8.2 (Hand)**.** List the sixth roots of unity $e^{-2\pi i j/6}$, $j = 0, \ldots, 5$, in the form $a + bi$. Verify by hand that they sum to 0, and that the sum of their squares is also 0 (which case of Lemma 8.2 is that?).

Exercise 8.3 (Proof, ⋆)**.** Prove the complex Pythagorean theorem and Cauchy–Schwarz: adapt the proofs of Theorem 2.3 and Theorem 2.5 to $\mathbb{C}^n$, flagging exactly where conjugate symmetry and conjugate-linearity in the second slot are used. (In the decomposition lemma, take $c = \langle x, y\rangle / \|y\|^2$ and verify $\langle z, y\rangle = 0$ still holds.)

Exercise 8.4 (Proof)**.** Show that for all $x, y \in \mathbb{C}^n$: $\langle x, y\rangle + \langle y, x\rangle = 2\,\mathrm{Re}\langle x, y\rangle$, and deduce $\|x + y\|^2 = \|x\|^2 + 2\,\mathrm{Re}\langle x, y\rangle + \|y\|^2$.

Deeper reading. LADR 1A ($\mathbb{C}^n$), 4 (polynomials over $\mathbb{C}$: the fundamental theorem of algebra, used next lesson), 6A (complex inner products); Strang, ch. "Complex Vectors and Matrices."

9 Eigenvalues, Eigenvectors, and Diagonalization

NEEDED FOR: Fourier analysis (core); statistical inference (Markov chains, PCA). Skimmable for the MDP side of ML.

9.1 Theory

Throughout, $T : V \to V$ is a linear map from a space to itself (an *operator*); V is $\mathbb{R}^n$ or $\mathbb{C}^n$.

Definition 9.1 (Eigenvalue, eigenvector)**.** A scalar $\lambda \in \mathbb{F}$ is an *eigenvalue* of T if there exists $v \neq 0$ with

$$Tv = \lambda v;$$

any such v is an *eigenvector* for λ. Equivalently (Proposition 5.2): λ is an eigenvalue iff $T - \lambda I$ is not injective; the eigenvectors for λ, together with 0, form the subspace $\operatorname{null}(T - \lambda I)$, the *eigenspace*.

An eigenvector is a direction the operator does not turn — only stretches by the factor λ. On such directions, applying T repeatedly is trivial: $T^k v = \lambda^k v$.

Theorem 9.2 (Independence of eigenvectors, LADR 5.11)**.** *Eigenvectors corresponding to distinct eigenvalues are linearly independent.*

Proof. Let $v_1, \ldots, v_m$ be eigenvectors for distinct eigenvalues $\lambda_1, \ldots, \lambda_m$, and suppose for contradiction the list is dependent. Since each $v_i \neq 0$, the linear dependence lemma (Lemma 3.3) gives a smallest index k with $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$, and $v_1, \ldots, v_{k-1}$ is independent by minimality of k. Write

$$v_k = a_1 v_1 + \cdots + a_{k-1} v_{k-1}.$$

Apply T to both sides, and separately multiply both sides by λ_k:

$$\lambda_k v_k = a_1 \lambda_1 v_1 + \cdots + a_{k-1} \lambda_{k-1} v_{k-1}, \qquad \lambda_k v_k = a_1 \lambda_k v_1 + \cdots + a_{k-1} \lambda_k v_{k-1}.$$

Subtracting, $0 = \sum_{i<k} a_i(\lambda_i - \lambda_k) v_i$. Independence of $v_1, \ldots, v_{k-1}$ forces $a_i(\lambda_i - \lambda_k) = 0$ for each i; since the eigenvalues are distinct, every $a_i = 0$, making $v_k = 0$ — contradicting that v_k is an eigenvector. □

Corollary 9.3. *An operator on an n-dimensional space has at most n distinct eigenvalues.*

Proof. Eigenvectors for distinct eigenvalues form an independent list, and independent lists in an n-dimensional space have length at most n (Theorem 3.4). □

Do eigenvalues exist at all? Over $\mathbb{R}$, not always: rotation of $\mathbb{R}^2$ by $90°$ turns every direction, so it has no real eigenvalue. Over $\mathbb{C}$, always — and Axler's proof is a gem, using no determinants:

Theorem 9.4 (Existence of eigenvalues over $\mathbb{C}$, LADR 5.19)**.** *Every operator on a nonzero finite-dimensional complex vector space has an eigenvalue.*

Proof. Let T be an operator on $\mathbb{C}^n$, $n \geq 1$, and pick any $v \neq 0$. The $n+1$ vectors

$$v,\, Tv,\, T^2v, \ldots, T^nv$$

must be linearly dependent (Theorem 3.4 again: $n+1$ vectors in an n-dimensional space). So there are scalars $a_0, \ldots, a_n$, not all zero, with $a_0v + a_1Tv + \cdots + a_nT^nv = 0$. Let m be the largest index with $a_m \neq 0$; then $m \geq 1$, because $m = 0$ would give $a_0v = 0$ with $a_0 \neq 0$ and $v \neq 0$.

Now the polynomial $p(z) = a_0 + a_1z + \cdots + a_mz^m$ has degree $m \geq 1$, so by the fundamental theorem of algebra it factors completely over $\mathbb{C}$:

$$p(z) = a_m(z - \lambda_1)(z - \lambda_2)\cdots(z - \lambda_m)$$

for some $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$. Because powers of T commute with each other, the same factorization holds with T substituted for z:

$$0 = a_0v + a_1Tv + \cdots + a_mT^mv = a_m(T - \lambda_1 I)(T - \lambda_2 I)\cdots(T - \lambda_m I)\, v.$$

The right side is a composition of operators applied to the nonzero vector v, producing 0. If every factor $T - \lambda_j I$ were injective, the composition would be injective and could not send $v \neq 0$ to 0. Hence some $T - \lambda_j I$ is not injective: λ_j is an eigenvalue. □

Why this proof is worth knowing. It explains *where eigenvalues come from*: dependence among $v, Tv, T^2v, \ldots$ produces a polynomial identity in T, and eigenvalues are the roots. The fundamental theorem of algebra — every complex polynomial factors into linear pieces — is the sole reason $\mathbb{C}$ guarantees eigenvalues while $\mathbb{R}$ does not (over $\mathbb{R}$, $z^2 + 1$ refuses to factor, and that irreducible quadratic *is* the 90° rotation's obstruction).

Definition 9.5 (Diagonalizable)**.** An operator T on V ($\dim V = n$) is *diagonalizable* if V has a basis $v_1, \ldots, v_n$ of eigenvectors of T. In matrix terms: $A = PDP^{-1}$, where D is diagonal with the eigenvalues $\lambda_1, \ldots, \lambda_n$ on its diagonal and P is the invertible matrix whose columns are the corresponding eigenvectors.

The matrix identity is just bookkeeping for "A acts diagonally on the eigenbasis": $AP = PD$ says column-by-column that $Av_k = \lambda_k v_k$ (compute both sides' kth columns: AP has kth column Av_k, and PD has kth column $\lambda_k v_k$). By Theorem 9.2, an operator on an n-dimensional space with n *distinct* eigenvalues is automatically diagonalizable. Repeated eigenvalues may or may not ruin it: I has one eigenvalue with an n-dimensional eigenspace (diagonalizable — it is diagonal), while the shear $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ has the single eigenvalue 1 with only a one-dimensional eigenspace (not diagonalizable: no basis of eigenvectors exists).

Proposition 9.6 (Powers via diagonalization)**.** *If $A = PDP^{-1}$ then $A^k = PD^kP^{-1}$, and D^k is diagonal with entries λ_j^k.*

Proof. $A^2 = PDP^{-1}PDP^{-1} = PD^2P^{-1}$; induction gives the rest. Diagonal matrices multiply entrywise on the diagonal. □

Understanding a diagonalizable operator's long-run behavior thus reduces to scalar arithmetic: components along eigenvectors with $|\lambda| < 1$ decay, $|\lambda| > 1$ grow, $|\lambda| = 1$ persist.

9.2 Practice: by hand

Example 9.7 (Full eigen-computation, 2×2)**.** $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Find λ making $A - \lambda I$ singular. A 2×2 matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is singular exactly when $ad - bc = 0$ (its columns are then parallel — Lesson 13 names this quantity the determinant). Here:

$$(2-\lambda)^2 - 1 = 0 \iff 2-\lambda = \pm 1 \iff \lambda = 1 \text{ or } 3.$$

Eigenvectors: for $\lambda = 3$, solve $(A - 3I)v = 0$: $\left(\begin{smallmatrix} -1 & 1 \\ 1 & -1 \end{smallmatrix}\right) v = 0 \Rightarrow v = (1,1)$. For $\lambda = 1$: $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) v = 0 \Rightarrow v = (1,-1)$. Two distinct eigenvalues $\Rightarrow$ diagonalizable, and note $(1,1) \perp (1,-1)$ — not an accident: A is symmetric, and Lesson 10 proves symmetric matrices always get orthogonal eigenvectors.

Geometrically, A stretches the diagonal direction $(1,1)$ by 3 and the anti-diagonal $(1,-1)$ by 1: an anisotropic stretch along tilted axes.

Example 9.8 (Powers put to work)**.** With A as above, compute $A^{10}(1,0)$ without ten multiplications: decompose $(1,0) = \frac{1}{2}(1,1) + \frac{1}{2}(1,-1)$, then

$$A^{10}(1,0) = \tfrac{1}{2}\,3^{10}(1,1) + \tfrac{1}{2}\,1^{10}(1,-1) = \tfrac{1}{2}\big(3^{10}+1,\ 3^{10}-1\big) = (29525,\, 29524).$$

Eigen-coordinates turn matrix powers into scalar powers — the essential trick behind solving recurrences, Markov chains, and linear ODEs.

> **Fourier connection.** A linear time-invariant (LTI) system — any circuit or filter Fourier analysis studies — has the complex exponentials as its *eigenfunctions*: feed in $e^{2\pi i s t}$ and out comes $H(s)\,e^{2\pi i s t}$, the same signal scaled by a complex number. The scale factor $H(s)$ (the frequency response / transfer function) is precisely an eigenvalue, indexed by frequency. "Diagonalize the system in the Fourier basis" is the slogan under all of Fourier analysis: convolution in time becomes multiplication (by eigenvalues) in frequency. Lesson 11 proves the finite-dimensional version of this exactly.

9.3 Exercises

Exercise 9.1 (Hand)**.** Find all eigenvalues and eigenvectors of (a) $\left(\begin{smallmatrix} 3 & 0 \\ 0 & -2 \end{smallmatrix}\right)$, (b) $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ (reflection through $x_2 = x_1$ — interpret the answer geometrically), (c) $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ over $\mathbb{R}$ and then over $\mathbb{C}$.

Exercise 9.2 (Hand)**.** Diagonalize $A = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$ and use Proposition 9.6 to find a closed form for $A^k(1,0)$.

Exercise 9.3 (Proof, $\star$)**.** Prove that if T is invertible with eigenvalue λ (necessarily $\lambda \neq 0$ — why?), then λ^{-1} is an eigenvalue of T^{-1}, with the same eigenvectors. Also show: λ^k is an eigenvalue of T^k for every $k \geq 1$.

Exercise 9.4 (Proof)**.** Show the shear $S = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ is not diagonalizable: find all eigenvalues, show the eigenspace is one-dimensional, and conclude no basis of eigenvectors exists.

Deeper reading. LADR 5A (invariant subspaces, eigenvalues), 5B (existence via minimal polynomials — a refinement of Theorem 9.4), 5D (diagonalizable operators); Strang, ch. "Eigenvalues and Eigenvectors."

10 Self-Adjoint Operators and the Spectral Theorem

NEEDED FOR: Fourier analysis (core); statistical inference (covariance matrices, PCA — reused in Lesson 20)

10.1 Theory

Work in $\mathbb{F}^n$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) with the inner product of Lessons 2 and 8. Recall from Lesson 4 the defining property of the transpose; over $\mathbb{C}$ its correct analogue is the *conjugate transpose* (*adjoint*) $A^* = \overline{A^{\mathsf{T}}}$, which satisfies $\langle Ax,\, y\rangle = \langle x,\, A^*y\rangle$ for all x, y (same expansion as before, tracking conjugates). Over $\mathbb{R}$, $A^* = A^{\mathsf{T}}$.

Definition 10.1 (Self-adjoint)**.** A is *self-adjoint* (*Hermitian*; over $\mathbb{R}$: *symmetric*) if $A^* = A$; equivalently $\langle Ax,\, y\rangle = \langle x,\, Ay\rangle$ for all x, y.

Proposition 10.2 (Real eigenvalues, LADR 7.12)**.** *Every eigenvalue of a self-adjoint matrix is real.*

Proof. Let $Av = \lambda v$ with $v \neq 0$. Then

$$\lambda\|v\|^2 = \langle \lambda v,\, v\rangle = \langle Av,\, v\rangle = \langle v,\, Av\rangle = \langle v,\, \lambda v\rangle = \bar{\lambda}\,\|v\|^2.$$

Since $\|v\|^2 \neq 0$, $\lambda = \bar{\lambda}$: real. □

Proposition 10.3 (Orthogonal eigenspaces)**.** *If A is self-adjoint, eigenvectors for distinct eigenvalues are orthogonal.*

Proof. Let $Av = \lambda v$, $Aw = \mu w$, $\lambda \neq \mu$ (both real by Proposition 10.2). Then

$$\lambda\langle v,\, w\rangle = \langle Av,\, w\rangle = \langle v,\, Aw\rangle = \bar{\mu}\langle v,\, w\rangle = \mu\langle v,\, w\rangle,$$

so $(\lambda - \mu)\langle v,\, w\rangle = 0$, forcing $\langle v,\, w\rangle = 0$. □

Theorem 10.4 (Spectral theorem, LADR 7.29/7.31)**.** *Let A be a self-adjoint $n \times n$ matrix (real symmetric, or complex Hermitian). Then $\mathbb{F}^n$ has an orthonormal basis of eigenvectors of A, and all eigenvalues are real. Equivalently, $A = Q\Lambda Q^*$ with Q having orthonormal columns ($Q^*Q = I$) and Λ real diagonal.*

Proof. Induct on n. For $n = 1$ any unit vector works. For $n > 1$: A has an eigenvalue λ_1. (Over $\mathbb{C}$ this is Theorem 9.4. Over $\mathbb{R}$: regard the real symmetric A as a complex matrix; it is Hermitian, so it has a complex eigenvalue λ_1, real by Proposition 10.2; then $A - \lambda_1 I$ is a real matrix that is singular over $\mathbb{C}$, hence has rank $< n$, hence is singular over $\mathbb{R}$ — rank is the same whether we allow real or complex coefficients, since elimination never leaves the field the entries live in — so a real eigenvector exists.)

Pick a unit eigenvector e_1 for λ_1, and let $U = \{x : \langle x,\, e_1\rangle = 0\}$ be its orthogonal complement, of dimension $n - 1$ (Exercise 6.3). The key point is that A maps U into U: for $u \in U$,

$$\langle Au,\, e_1\rangle = \langle u,\, Ae_1\rangle = \langle u,\, \lambda_1 e_1\rangle = \lambda_1\langle u,\, e_1\rangle = 0,$$

so $Au \in U$. The restriction of A to U is an operator on an $(n-1)$-dimensional inner product space, still self-adjoint (the identity $\langle Ax,\, y\rangle = \langle x,\, Ay\rangle$ holds in particular for $x, y \in U$). By

the induction hypothesis, U has an orthonormal basis $e_2, \ldots, e_n$ of eigenvectors of A. Then $e_1, e_2, \ldots, e_n$ is orthonormal (each later e_j lies in $U \perp e_1$) and consists of eigenvectors.

For the matrix form: let Q have columns $e_1, \ldots, e_n$. Orthonormality of columns says exactly $Q^*Q = I$, and $Ae_k = \lambda_k e_k$ says $AQ = Q\Lambda$; since Q is square with $Q^*Q = I$, $Q^{-1} = Q^*$ and $A = Q\Lambda Q^*$. □

What the spectral theorem buys you. Self-adjoint matrices are exactly the ones that are a *pure stretch along some perpendicular axes*: rotate into the right orthonormal frame (Q^*), scale each axis by a real number (Λ), rotate back (Q). No shearing, no hidden rotation. Because the eigenbasis is orthonormal, coordinates in it are free ($a_j = \langle x, e_j \rangle$, Proposition 6.2) — so analyzing A's action on any vector costs n dot products. This is the finite-dimensional model of what Fourier analysis does for differential operators.

A useful companion definition: a self-adjoint A is *positive semidefinite* (PSD) if $\langle Ax, x \rangle \geq 0$ for all x. By the spectral theorem this holds iff all eigenvalues are ≥ 0 (expand x in the eigenbasis: $\langle Ax, x \rangle = \sum_j \lambda_j |\langle x, e_j \rangle|^2$). The matrices A^*A from least squares (Lesson 6) are always PSD: $\langle A^*Ax, x \rangle = \|Ax\|^2 \geq 0$ — this observation powers the SVD in Lesson 12 (and, in Lesson 20, the fact that covariance matrices are PSD).

10.2 Practice: by hand

Example 10.5 (Spectral decomposition, fully worked)**.** From Example 9.7, $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ has eigenpairs $\lambda = 3$, $v = (1, 1)$ and $\lambda = 1$, $v = (1, -1)$, already orthogonal as Proposition 10.3 promised. Normalize:

$$Q = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \qquad A = Q\Lambda Q^{\mathsf{T}}.$$

Verify one entry: $(Q\Lambda Q^{\mathsf{T}})_{11} = \frac{1}{2}(3 \cdot 1 \cdot 1 + 1 \cdot 1 \cdot 1) = 2$. ✓ Also read off positivity: eigenvalues $3, 1 > 0$, so $\langle Ax, x \rangle > 0$ for all $x \neq 0$ — A is *positive definite*.

Example 10.6 (Quadratic forms become sums of squares)**.** For the same A, with new coordinates $y = Q^{\mathsf{T}}x$ (the eigen-coordinates),

$$x^{\mathsf{T}}Ax = x^{\mathsf{T}}Q\Lambda Q^{\mathsf{T}}x = y^{\mathsf{T}}\Lambda y = 3y_1^2 + y_2^2.$$

The level sets $x^{\mathsf{T}}Ax = 1$ are ellipses with axes along $(1, 1)$ and $(1, -1)$ and semi-axis lengths $1/\sqrt{3}$ and 1. Diagonalization untangles every quadratic form into independent squared terms.

Fourier connection. Symmetric structure is why real signals have the Fourier symmetries Fourier analysis drills ($\hat{f}(-s) = \overline{\hat{f}(s)}$, etc.), and self-adjointness of the underlying operators is what makes Fourier eigen-expansions come with orthogonality — which in turn yields Parseval/Rayleigh identities: energy in time equals energy in frequency, i.e., $\|x\|$ is preserved when you switch to an orthonormal (eigen)basis. The next lesson makes the norm-preservation statement exact for the DFT.

10.3 Exercises

Exercise 10.1 (Hand)**.** Find the spectral decomposition $Q\Lambda Q^{\mathsf{T}}$ of $\left(\begin{smallmatrix} 0 & 2 \\ 2 & 3 \end{smallmatrix}\right)$, and write the quadratic form $q(x) = 4x_1x_2 + 3x_2^2$ as a combination of squares in the eigen-coordinates. Is the matrix PSD?

Exercise 10.2 (Hand)**.** The Hermitian matrix $H = \left(\begin{smallmatrix} 1 & i \\ -i & 1 \end{smallmatrix}\right)$ (note $H^* = H$): find its (real!) eigenvalues and a pair of orthogonal eigenvectors in $\mathbb{C}^2$; verify orthogonality with the complex inner product.

Exercise 10.3 (Proof, $\star$)**.** Prove: a self-adjoint matrix with all eigenvalues in $\{0, 1\}$ is exactly an orthogonal projection P_U (as in Theorem 6.3) onto the span U of its $\lambda = 1$ eigenvectors. (Use the spectral theorem to write $A = \sum_j \lambda_j e_j e_j^*$, i.e. $Ax = \sum_j \lambda_j \langle x,\, e_j\rangle e_j$, and compare with the projection formula.)

Exercise 10.4 (Proof)**.** Show that for any (not necessarily self-adjoint) real A, the matrix $S = \frac{1}{2}(A + A^{\mathsf{T}})$ is symmetric and $x^{\mathsf{T}}Ax = x^{\mathsf{T}}Sx$ for all x — so quadratic forms lose nothing by assuming symmetry (as Proposition 7.3 did).

Deeper reading. LADR 7A (self-adjoint and normal operators), 7B (spectral theorem — Axler also proves the normal case over $\mathbb{C}$, the full characterization of orthonormally diagonalizable operators), 7C (positive operators); Strang, ch. "Symmetric Matrices and Positive Definiteness."

11 Unitary Matrices and the Discrete Fourier Transform

NEEDED FOR: Fourier analysis (core); statistical-inference background (transform-domain reasoning)

11.1 Theory

Definition 11.1 (Unitary / orthogonal matrix)**.** A square matrix Q over $\mathbb{C}$ is *unitary* if $Q^*Q = I$ (over $\mathbb{R}$, with $Q^{\mathsf{T}}Q = I$: *orthogonal*).

Theorem 11.2 (Three faces of unitarity, LADR 7.51)**.** *For a square matrix Q, the following are equivalent:*

(a) $Q^*Q = I$ (columns are orthonormal);

(b) $\langle Qx,\, Qy\rangle = \langle x,\, y\rangle$ for all x, y (inner products preserved);

(c) $\|Qx\| = \|x\|$ for all x (lengths preserved).

Moreover a unitary Q is invertible with $Q^{-1} = Q^*$, which is again unitary.

Proof. (a) $\Rightarrow$ (b): $\langle Qx,\, Qy\rangle = \langle x,\, Q^*Qy\rangle = \langle x,\, y\rangle$, using the adjoint's defining property. (b) $\Rightarrow$ (c): take $y = x$ and square roots. (c) $\Rightarrow$ (a): first note entry (j, k) of Q^*Q is $\langle Qe_k,\, Qe_j\rangle$. From (c) and the expansion $\|Q(x+y)\|^2 = \|Qx\|^2 + 2\,\mathrm{Re}\langle Qx,\, Qy\rangle + \|Qy\|^2$ (Exercise 8.4), preserving all norms forces $\mathrm{Re}\langle Qx,\, Qy\rangle = \mathrm{Re}\langle x,\, y\rangle$ for all x, y; replacing x by ix recovers the imaginary parts likewise (over $\mathbb{R}$, the real-part identity is already everything). So (b) holds, and applying

it to standard basis vectors gives $\langle Qe_k,\, Qe_j\rangle = \langle e_k,\, e_j\rangle = \delta_{jk}$: the columns are orthonormal, which is (a).

Finally, $Q^*Q = I$ with Q square makes Q injective (if $Qx = 0$ then $x = Q^*Qx = 0$), hence invertible by Corollary 5.4, and multiplying $Q^*Q = I$ by Q^{-1} on the right gives $Q^* = Q^{-1}$; then $(Q^*)^*Q^* = QQ^{-1} = I$ shows Q^* is unitary. □

Unitary matrices are the rigid motions of $\mathbb{C}^n$ fixing the origin: they preserve all lengths and angles. A change of basis between two orthonormal bases is always unitary, and conversely.

Definition 11.3 (The DFT matrix)**.** Fix $N \geq 1$ and $\omega = e^{-2\pi i/N}$. The *discrete Fourier transform* of $x \in \mathbb{C}^N$ (indexing coordinates from 0) is the vector $\hat{x} = Wx$, where W is the $N \times N$ matrix with entries

$$W_{k,n} = \omega^{kn}, \qquad \text{i.e.} \qquad \hat{x}_k = \sum_{n=0}^{N-1} x_n\, e^{-2\pi ikn/N}.$$

The number $\hat{x}_k$ measures the correlation of the signal x with the frequency-k oscillation.

Theorem 11.4 (The DFT is (essentially) unitary)**.** *$W^*W = NI$. Hence $F = W/\sqrt{N}$ is unitary, the DFT is invertible with*

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} \hat{x}_k\, e^{+2\pi ikn/N} \qquad (\textit{the inverse DFT}),$$

and Parseval's identity holds: $\sum_n |x_n|^2 = \frac{1}{N}\sum_k |\hat{x}_k|^2$.

Proof. Entry (m,n) of W^*W is $\sum_k \overline{W_{k,m}} W_{k,n} = \sum_{k=0}^{N-1} \overline{\omega^{km}}\,\omega^{kn} = \sum_{k=0}^{N-1} \omega^{k(n-m)}$, which by the roots-of-unity lemma (Lemma 8.2) equals N if $m = n$ and 0 otherwise — here $|n-m| < N$, so $N \mid (n-m)$ occurs only for $m = n$. Thus $W^*W = NI$.

Consequently $F^*F = I$: F is unitary, so $W^{-1} = \frac{1}{N}W^*$, whose entry (n,k) is $\frac{1}{N}\overline{\omega^{kn}} = \frac{1}{N}e^{+2\pi ikn/N}$ — the displayed inversion formula. Parseval is Theorem 11.2(c) for F: $\|Fx\| = \|x\|$, i.e. $\frac{1}{N}\|Wx\|^2 = \|x\|^2$. □

Definition 11.5 (Circular convolution)**.** For $c, x \in \mathbb{C}^N$, with indices taken mod N,

$$(c * x)_n = \sum_{j=0}^{N-1} c_j\, x_{n-j \bmod N}.$$

Equivalently $c * x = Cx$ where C is the *circulant* matrix whose column j is c cyclically shifted down by j.

Theorem 11.6 (Convolution theorem)**.** *$\widehat{c * x}_k = \hat{c}_k\hat{x}_k$ for every k. Equivalently: every circulant matrix is diagonalized by the DFT, $C = W^{-1}\,\mathrm{diag}(\hat{c})\,W$, with eigenvalues $\hat{c}_0, \ldots, \hat{c}_{N-1}$.*

Proof. Compute, substituting $m = n - j \bmod N$ (as n runs over residues mod N with j fixed, so does m, and $\omega^{kn} = \omega^{k(j+m)}$ since $\omega^{kN} = 1$ makes exponents matter only mod N):

$$\widehat{c * x}_k = \sum_n \sum_j c_j x_{n-j}\, \omega^{kn} = \sum_j c_j \omega^{kj} \sum_m x_m \omega^{km} = \hat{c}_k\hat{x}_k.$$

For the matrix form: the identity says $W(Cx) = \mathrm{diag}(\hat{c})(Wx)$ for all x, i.e. $WC = \mathrm{diag}(\hat{c})W$, i.e. $C = W^{-1}\,\mathrm{diag}(\hat{c})W$. Reading it as a diagonalization: the columns of W^{-1} (the sampled complex exponentials) are eigenvectors of *every* circulant, with eigenvalues the DFT of the first column. □

The punchline of Part II. Shift-invariant linear operations (circulants) are precisely the ones diagonalized by the Fourier basis, and their eigenvalues are the transform of their impulse response c. Filtering = multiply in frequency. This is Lesson 9's "diagonalization turns matrix powers into scalars" plus Lesson 10's "orthonormal eigenbases make coordinates free," specialized to the one basis that never changes: the complex exponentials. A Fourier-analysis course spends weeks developing the continuous version.

11.2 Practice: by hand

Example 11.7 (The 4-point DFT, fully)**.** $N = 4$, $\omega = -i$ (Lesson 8). The matrix and a transform:

$$W = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix}, \qquad x = (1,0,1,0) \;\Rightarrow\; \hat{x} = Wx = (2,0,2,0).$$

Sense check: x alternates with period 2 — energy at frequency 0 (its mean is not zero) and frequency 2 (the Nyquist alternation), none at frequencies 1, 3. Parseval: $\sum |x_n|^2 = 2$ and $\frac{1}{4}\sum |\hat{x}_k|^2 = \frac{1}{4}(4+4) = 2$. ✓

Example 11.8 (Convolution theorem, by hand)**.** Take the averaging filter $c = \frac{1}{2}(1,1,0,0)$ and $x = (1,0,1,0)$ with $N = 4$. Directly: $(c * x)_n = \frac{1}{2}(x_n + x_{n-1}) = \frac{1}{2}(1,1,1,1)$. Via frequency: $\hat{c} = \frac{1}{2}(2,\ 1-i,\ 0,\ 1+i)$, and $\hat{c} \odot \hat{x} = (2,0,0,0)$ (entrywise), whose inverse DFT is $\frac{1}{4}\cdot 2 \cdot (1,1,1,1) = \frac{1}{2}(1,1,1,1)$. ✓ The filter's zero at frequency 2 ($\hat{c}_2 = 0$) is why the alternating component of x is annihilated: averaging adjacent samples kills the fastest oscillation.

Fourier connection. This lesson is the last third of Fourier analysis in miniature: the DFT section of the course defines exactly this matrix, proves exactly this inversion and convolution theorem, and then studies the FFT — an algorithm computing Wx in $O(N \log N)$ instead of the obvious $O(N^2)$, arguably the most important algorithm in engineering. Everything earlier in the course (Fourier series, the continuous transform, sampling) has the same skeleton with integrals in place of sums; if you own this lesson, the Fourier course's DFT weeks are review.

11.3 Exercises

Exercise 11.1 (Hand)**.** Write out the 3-point DFT matrix using $\omega = e^{-2\pi i/3} = -\frac{1}{2} - \frac{\sqrt{3}}{2}i$, and compute $\hat{x}$ for $x = (1,1,1)$ and for $x = (1, \omega^{-1}, \omega^{-2})$. Explain both answers in one sentence each.

Exercise 11.2 (Hand)**.** For $N = 4$, write the circulant matrix C for $c = (0,1,0,0)$ (the cyclic shift). Verify directly that $v = (1, i^{-1}, i^{-2}, i^{-3}) = (1, -i, -1, i)$ satisfies $Cv = \lambda v$, and identify λ as an entry of $\hat{c}$.

Exercise 11.3 (Proof, $\star$)**.** Prove that the product of two unitary matrices is unitary and that every eigenvalue λ of a unitary matrix has $|\lambda| = 1$. (For the latter: apply Theorem 11.2(c) to an eigenvector.) Conclude the DFT eigenvalue fact: $|\hat{c}_k| \le \sum_j |c_j|$ for a circulant's eigenvalues, directly from the definition of $\hat{c}$ and the triangle inequality.

Exercise 11.4 (Proof)**.** Show that the set of circulant $N \times N$ matrices is closed under products and that any two circulants commute. (Two lines with Theorem 11.6: simultaneous diagonalization.)

Deeper reading. LADR 7D (isometries and unitary operators); Strang, ch. "Complex Vectors and Matrices" (the Fast Fourier Transform); Osgood, *Lectures on the Fourier Transform and Its Applications*, chapter on the DFT.

12 The Singular Value Decomposition

NEEDED FOR: Fourier-adjacent; statistical inference (PCA). Not needed for ML.

The spectral theorem needs a self-adjoint square matrix. The SVD is its extension to *every* matrix, square or not — LADR Chapter 7's finale, and the workhorse behind `lstsq`, `matrix_rank`, PCA, and compression.

12.1 Theory

Theorem 12.1 (Singular value decomposition, LADR 7.70)**.** *Let A be a real $m \times n$ matrix of rank r. Then there exist orthonormal vectors $v_1, \dots, v_r \in \mathbb{R}^n$, orthonormal vectors $u_1, \dots, u_r \in \mathbb{R}^m$, and numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ (the* singular values*) such that*

$$A = \sigma_1 u_1 v_1^{\mathsf{T}} + \cdots + \sigma_r u_r v_r^{\mathsf{T}}, \quad \textit{equivalently} \quad Av_i = \sigma_i u_i \ (i \leq r), \quad Av = 0 \textit{ for } v \perp v_1, \dots, v_r.$$

(The same holds over $\mathbb{C}$ with $*$ in place of T.)

Proof. The matrix $A^{\mathsf{T}}A$ is symmetric and PSD (Lesson 10: $\langle A^{\mathsf{T}}Ax, x\rangle = \|Ax\|^2 \geq 0$). By the spectral theorem it has an orthonormal eigenbasis $v_1, \dots, v_n$ of $\mathbb{R}^n$ with real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$; write $\lambda_i = \sigma_i^2$ with $\sigma_i \geq 0$, and let r' be the number of nonzero σ_i.

Key computation. For any i, j:

$$\langle Av_i,\, Av_j\rangle = \langle A^{\mathsf{T}}Av_i,\, v_j\rangle = \sigma_i^2\langle v_i,\, v_j\rangle = \begin{cases} \sigma_i^2 & i = j \\ 0 & i \neq j. \end{cases}$$

So the vectors $Av_1, \dots, Av_n$ are mutually orthogonal with $\|Av_i\| = \sigma_i$; in particular $Av_i = 0$ exactly when $\sigma_i = 0$. For $i \leq r'$ define $u_i = Av_i/\sigma_i$: unit vectors, orthonormal by the computation above.

The formula. Both sides of the claimed identity are linear maps; by Theorem 4.3 it suffices to check them on the basis $v_1, \dots, v_n$. For $j \leq r'$: the right side gives $\sum_i \sigma_i u_i (v_i^{\mathsf{T}} v_j) = \sigma_j u_j = Av_j$. For $j > r'$: the right side gives $0 = Av_j$. They agree.

Rank. range A is spanned by the Av_j, i.e. by $\sigma_1 u_1, \dots, \sigma_{r'} u_{r'}$, an orthogonal (hence independent) list — so $\operatorname{rank} A = r'$, forcing $r' = r$. □

In matrix form, padding the v's and u's to full orthonormal bases: $A = U\Sigma V^{\mathsf{T}}$ with U $(m \times m)$ and V $(n \times n)$ orthogonal and Σ $(m \times n)$ diagonal with the σ_i's. Read as a pipeline, every matrix is: rotate the input frame (V^{T}), stretch each axis by σ_i and embed into the output space (Σ), rotate the output frame (U). Rotation–stretch–rotation, no exceptions.

Corollary 12.2 (What the SVD reveals at a glance)**.** *With notation as above:*

(a) $\operatorname{rank} A$ = *number of nonzero singular values (this is what* `matrix_rank` *counts, with a tolerance);*

(b) $\|Av\| \le \sigma_1 \|v\|$ *for all* v*, with equality at* $v = v_1$*:* σ_1 *is the operator's maximum stretch factor;*

(c) the null space of A *is* $\operatorname{span}(v_{r+1}, \dots, v_n)$ *and the range is* $\operatorname{span}(u_1, \dots, u_r)$ *— the SVD hands you orthonormal bases for both of Lesson 5's subspaces.*

Proof. (a) was shown in the proof. (c): $Av = 0$ iff all eigen-coordinates of v along $v_1, \dots, v_r$ vanish (by the orthogonality of the Av_i and $\|Av_i\| = \sigma_i > 0$); the range statement was the last line of the proof. (b): expand $v = \sum_i c_i v_i$; then $Av = \sum_{i \le r} c_i \sigma_i u_i$, and by orthonormality of the u_i (Pythagoras, repeatedly):

$$\|Av\|^2 = \sum_{i \le r} \sigma_i^2 c_i^2 \le \sigma_1^2 \sum_i c_i^2 = \sigma_1^2 \|v\|^2,$$

with equality when all weight sits on c_1. □

Dropping all but the k largest terms of $\sum_i \sigma_i u_i v_i^{\mathsf{T}}$ gives the rank-k matrix A_k closest to A (Eckart–Young theorem; statement worth knowing, proof omitted) — the principle behind PCA and lossy compression: keep the directions with the biggest stretch, discard the rest.

12.2 Practice: by hand

Example 12.3 (A full 2×2 SVD)**.** $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ (the shear that refused to diagonalize in Lesson 9 — but nothing refuses an SVD). Compute

$$A^{\mathsf{T}} A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}: \quad (1-\lambda)(2-\lambda) - 1 = 0 \iff \lambda^2 - 3\lambda + 1 = 0 \iff \lambda = \tfrac{3 \pm \sqrt{5}}{2}.$$

So $\sigma_1 = \sqrt{(3+\sqrt{5})/2} \approx 1.618$ and $\sigma_2 = \sqrt{(3-\sqrt{5})/2} \approx 0.618$ (the golden ratio and its reciprocal!). Eigenvector for λ_1: solve $(1-\lambda_1)a + b = 0$, giving $v_1 \propto (1, \lambda_1 - 1)$; then $u_1 = Av_1/\sigma_1$. The shear stretches one tilted direction by φ and compresses an orthogonal one by $1/\varphi$ — invisible to eigen-analysis, plain in the SVD.

Fourier connection. SVD is Fourier-adjacent rather than Fourier-core: it appears when Fourier methods meet data — e.g., analyzing which frequency patterns dominate a family of signals (PCA of spectrograms), or characterizing the maximum gain σ_1 of a non-shift-invariant system where the convolution theorem does not apply. It also closes a Part I loop: `lstsq` (Lesson 6) is implemented via SVD because Corollary 12.2 exposes range and null space in orthonormal coordinates, making the projection of b trivial even when columns are dependent. (In statistical inference it returns as PCA: Lesson 20 shows the covariance matrix's eigendecomposition *is* the SVD story for data.)

12.3 Exercises

Exercise 12.1 (Hand)**.** Compute the SVD ingredients of $A = \left(\begin{smallmatrix} 3 & 0 \\ 0 & -2 \end{smallmatrix}\right)$: singular values, v_i's, u_i's. (Careful: singular values are positive; where did the sign of -2 go?) Compare with the eigenvalues from Exercise 9.1(a).

Exercise 12.2 (Hand)**.** For $A = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right)$ (rank 1), find σ_1, u_1, v_1 and write $A = \sigma_1 u_1 v_1^{\mathsf{T}}$ exactly.

Exercise 12.3 (Proof, $\star$)**.** Prove that for symmetric PSD A, the singular values equal the eigenvalues and one may take $u_i = v_i$. Then show by example (Exercise 12.1!) that for symmetric A with a negative eigenvalue, singular values are the absolute values of eigenvalues.

Exercise 12.4 (Proof)**.** Using Corollary 12.2(b) applied to A and to A^{-1} (square invertible case), show $\|Ax\|/\|x\|$ ranges over $[\sigma_n, \sigma_1]$, and interpret $\kappa = \sigma_1/\sigma_n$ as the worst-case ratio by which A distorts relative lengths. (κ is the *condition number*; `np.linalg.cond` computes it, and it governs how much rounding error `solve` can amplify.)

Deeper reading. LADR 7E–7F (SVD and its consequences — Axler's proof is the one given here), 7C (positive operators and square roots); Strang, ch. "The Singular Value Decomposition."

13 Determinants and Trace

NEEDED FOR: Fourier analysis (stretch theorem); general background. Not needed for ML.

Axler famously postpones determinants to his final chapter because nothing earlier needed them — and indeed nothing earlier *here* needed them. But they earn their place: as the volume-scaling factor of a linear map (which is exactly how they enter multivariable integrals and multidimensional Fourier transforms) and as a compact certificate of invertibility.

13.1 Theory

Definition 13.1 (Determinant, characterized)**.** The determinant is the unique function $\det : (n \times n \text{ matrices}) \to \mathbb{F}$ that, viewed as a function of the n columns, is

(i) *multilinear*: linear in each column separately, the others held fixed;

(ii) *alternating*: $\det = 0$ whenever two columns are equal;

(iii) *normalized*: $\det I = 1$.

Existence and uniqueness are LADR 9B–9C; uniqueness forces the explicit permutation formula

$$\det A = \sum_{\pi} \operatorname{sign}(\pi)\, A_{\pi(1),1} \cdots A_{\pi(n),n},$$

summed over the $n!$ permutations π of $\{1, \dots, n\}$. For $n = 2, 3$ this yields the familiar

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc, \qquad \det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = aei + bfg + cdh - ceg - bdi - afh.$$

Two consequences of (i)–(ii), each one line: swapping two columns flips the sign of det (expand det with both columns replaced by their sum, and cancel using (ii)); and adding a multiple of one column to another leaves det unchanged (multilinearity splits it into det A plus a determinant with a repeated column). The same statements hold for rows, since $\det A = \det A^{\mathsf{T}}$ (visible from the permutation formula: transposing reindexes the sum by inverse permutations, which have the same signs).

Theorem 13.2 (Multiplicativity)**.** $\det(AB) = \det(A)\det(B)$.

Proof. Fix A and consider $f(B) = \det(AB)$ as a function of the columns of B. Column j of AB is Ab_j (Lesson 4), so f is multilinear in the columns of B (because $b_j \mapsto Ab_j$ is linear and det is multilinear) and alternating (equal columns $b_i = b_j$ give equal columns $Ab_i = Ab_j$). A short argument from uniqueness — any multilinear alternating function is a scalar multiple of det, the scalar being its value at I — gives $f(B) = f(I)\det(B) = \det(A)\det(B)$. □

Theorem 13.3 (Determinant as invertibility certificate and eigenvalue product)**.** *(a) A is invertible $\iff$ $\det A \neq 0$; in that case $\det(A^{-1}) = 1/\det A$.*

(b) If A (over $\mathbb{C}$) has eigenvalues $\lambda_1, \ldots, \lambda_n$ listed with multiplicity, then $\det A = \lambda_1 \cdots \lambda_n$; consequently λ is an eigenvalue of A iff $\det(A - \lambda I) = 0$ — the characteristic polynomial route used informally in Example 9.7.

Proof. (a) If A is invertible, $1 = \det I = \det(AA^{-1}) = \det A \det A^{-1}$ by Theorem 13.2, so $\det A \neq 0$ and the reciprocal formula holds. If A is not invertible, its columns are dependent (Lesson 5), so some column is a combination of the others (dependence lemma); subtracting that combination from it — which does not change det — produces a zero column, and multilinearity with the zero column gives $\det A = 0$.

(b) Over $\mathbb{C}$, every matrix is similar to an upper-triangular one: $A = STS^{-1}$ with T upper triangular whose diagonal lists the eigenvalues with multiplicity (LADR 5C, from Theorem 9.4 by induction; statement used here without the inductive details). By multiplicativity, $\det A = \det S \det T \det S^{-1} = \det T$, and the determinant of a triangular matrix is the product of its diagonal entries — expand by the permutation formula: only the identity permutation avoids all the zero entries below the diagonal. Hence $\det A = \lambda_1 \cdots \lambda_n$. Applying this to $A - \lambda I$, whose eigenvalues are $\lambda_i - \lambda$: it is singular iff some $\lambda_i = \lambda$ iff its determinant $\prod_i (\lambda_i - \lambda)$ vanishes. □

Volume. $|\det A|$ is the factor by which the map $x \mapsto Ax$ scales n-dimensional volume: the unit square (cube) maps to the parallelogram (parallelepiped) spanned by the columns of A, whose volume is $|\det A|$. Properties (i)–(iii) are exactly the axioms signed volume must satisfy (scale one edge, the volume scales; degenerate flat shapes have volume zero; the unit cube has volume 1), which is the honest reason the determinant is defined this way. The sign records orientation: negative when the map flips handedness.

Definition 13.4 (Trace)**.** $\operatorname{tr} A = \sum_i A_{i,i}$, the sum of diagonal entries.

Proposition 13.5. *(a) $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ for any A ($m \times n$) and B ($n \times m$). (b) Over $\mathbb{C}$, $\operatorname{tr} A = \lambda_1 + \cdots + \lambda_n$ (eigenvalues with multiplicity).*

Proof. (a) Both sides equal $\sum_i \sum_j A_{i,j} B_{j,i}$ — expand each definition and swap the order of summation. (b) With $A = STS^{-1}$ as above: $\operatorname{tr} A = \operatorname{tr}\big((ST)S^{-1}\big) = \operatorname{tr}\big(S^{-1}(ST)\big) = \operatorname{tr} T = \sum_i \lambda_i$, using (a) with the grouping shown. □

Trace and determinant are the two coefficients of the characteristic polynomial you can read off instantly; for 2×2: $\det(A - \lambda I) = \lambda^2 - (\operatorname{tr} A)\lambda + \det A$, so eigenvalues of a 2×2 are $\frac{\operatorname{tr} A \pm \sqrt{(\operatorname{tr} A)^2 - 4 \det A}}{2}$ — worth memorizing.

13.2 Practice: by hand

Example 13.6. $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$: $\operatorname{tr} A = 4$, $\det A = 3$; eigenvalues solve $\lambda^2 - 4\lambda + 3 = 0$: $\lambda = 1, 3$. Matches Example 9.7, with product $3 = \det$ and sum $4 = \operatorname{tr}$. ✓

Example 13.7 (Volume scaling)**.** The map $A = \left(\begin{smallmatrix} 2 & 1 \\ 0 & 3 \end{smallmatrix}\right)$ sends the unit square to the parallelogram spanned by $(2, 0)$ and $(1, 3)$: base 2, height 3, area $6 = \det A$. The shear part (the 1) moved the top edge sideways without changing the area — exactly the "add a multiple of one column to another" invariance.

Fourier connection. Determinants enter Fourier analysis exactly once, but at a memorable moment: the higher-dimensional Fourier transform. For $f : \mathbb{R}^n \to \mathbb{C}$ and an invertible A, the transform of the stretched signal $f(Ax)$ is $\frac{1}{|\det A|} \hat{f}(A^{-\mathsf{T}} \xi)$ — the general stretch theorem. The $|\det A|$ is the volume-scaling factor from the change of variables in the integral, and $A^{-\mathsf{T}}$ is why frequency space transforms "contragradiently" (images stretched in space compress in frequency, along dual axes). When Osgood derives this for 2-D images and CT scans, you will know exactly where every symbol comes from.

13.3 Exercises

Exercise 13.1 (Hand)**.** Compute by row reduction, tracking swaps and column operations' effects: $\det \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 1 & 0 & 3 \end{pmatrix}$. Cross-check with the 3×3 formula.

Exercise 13.2 (Hand)**.** Using trace and det only, find the eigenvalues of $\left(\begin{smallmatrix} 5 & 4 \\ 1 & 2 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$ (the latter over $\mathbb{C}$; interpret).

Exercise 13.3 (Proof, $\star$)**.** From the three defining properties alone (no permutation formula), derive $\det \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = ad - bc$: expand $\det\big((a, c), (b, d)\big)$ by multilinearity into four determinants of standard-basis columns, and evaluate each using alternation and normalization.

Exercise 13.4 (Proof)**.** Prove that similar matrices ($B = S^{-1}AS$) have the same determinant, trace, and eigenvalues — so all three are properties of the underlying linear map, independent of basis.

Deeper reading. LADR 9A (bilinear and quadratic forms), 9B (alternating multilinear forms), 9C (determinants — the characterization above); for the skipped Chapter 8 (generalized eigenvectors, i.e. what replaces diagonalization when it fails, and the trace's operator-theoretic role in 8D): read after the Fourier work if the shear example left you curious — no Fourier topic requires it. Strang, ch. "Determinants."

Part III: The Probability Core

Machine learning assumes a first course in probability, and it collects on the debt: Bayes' rule and conditional independence power the Bayesian-network and HMM lectures, expectations define the utilities that MDPs and expectimax maximize, and every training loss is an expectation over data. This part is the minimum working dose, following Bertsekas & Tsitsiklis (B&T), *Introduction to Probability* (2nd ed.), Chapters 1–3, in the same style as Part I. Statistical inference assumes all of this as a floor, so nothing here is wasted on the longer road either.

14 Probability Models, Counting, Conditioning, and Bayes' Rule

NEEDED FOR: Machine learning (core: Bayesian networks, HMMs); statistical inference (assumed)

14.1 Theory

Definition 14.1 (Probability model)**.** A *probabilistic model* consists of a *sample space* Ω — the set of all possible outcomes of an experiment, exactly one of which occurs — and a *probability law* $\mathbf{P}$ assigning to each event $A \subseteq \Omega$ a number $\mathbf{P}(A)$, satisfying the axioms:

(i) *nonnegativity*: $\mathbf{P}(A) \geq 0$ for every event A;

(ii) *additivity*: if $A_1, A_2, \ldots$ are disjoint events, then $\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots$;

(iii) *normalization*: $\mathbf{P}(\Omega) = 1$.

Everything else in Parts III–IV is a consequence of these three axioms — the same game as Lesson 1's vector-space axioms, and worth playing for the same reason.

Proposition 14.2 (First consequences)**.** *For events A, B:*

(a) $\mathbf{P}(\varnothing) = 0$ *and* $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$*;*

(b) if $A \subseteq B$ *then* $\mathbf{P}(A) \leq \mathbf{P}(B)$ *(monotonicity);*

(c) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$*, hence* $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$ *(the union bound).*

Proof. (a) Ω and $\varnothing$ are disjoint with union Ω, so $1 = \mathbf{P}(\Omega) = \mathbf{P}(\Omega) + \mathbf{P}(\varnothing)$, giving $\mathbf{P}(\varnothing) = 0$; A and A^c are disjoint with union Ω, so $\mathbf{P}(A) + \mathbf{P}(A^c) = 1$.

(b) $B = A \cup (B \cap A^c)$, a disjoint union, so $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \cap A^c) \geq \mathbf{P}(A)$ by nonnegativity.

(c) Write $A \cup B$ as the disjoint union $A \cup (B \cap A^c)$ and B as the disjoint union $(A \cap B) \cup (B \cap A^c)$. Then

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B \cap A^c) = \mathbf{P}(A) + \big(\mathbf{P}(B) - \mathbf{P}(A \cap B)\big),$$

and the inequality follows since $\mathbf{P}(A \cap B) \geq 0$. □

When Ω is finite with equally likely outcomes, $\mathbf{P}(A) = |A|/|\Omega|$: probability is counting (the *discrete uniform law*). A second situation is just as common (B&T §1.6): every outcome in A has the same already-known probability p, so that $\mathbf{P}(A) = p \cdot |A|$ — exactly how the binomial probabilities of Lesson 15 will be computed. Either way, computing the probability means *counting* the elements of A, so we pause to assemble the counting toolkit, following B&T §1.6.

Theorem 14.3 (The Counting Principle)**.** *Consider a process consisting of r stages. Suppose there are n_1 possible results at the first stage, and, for every possible result of the first $i-1$ stages, exactly n_i possible results at the ith stage. Then the total number of possible results of the r-stage process is $n_1 n_2 \cdots n_r$.*

Proof. Picture a tree: the root branches into n_1 nodes, each of those into n_2, and so on for r levels. Each result of the process is exactly one root-to-leaf path, so the count is the number of leaves, which by induction on r is $n_1 \cdots n_r$: the r-level tree hangs n_r leaves below each of the $n_1 \cdots n_{r-1}$ leaves of the $(r-1)$-level tree. (Note the hypothesis: the *number* n_i must not depend on the earlier results, though the available choices themselves may.) □

Two selection regimes cover most counting problems: *sampling with replacement* (the same item may be chosen again — successive coin flips, the digits of a phone number) and *sampling without replacement* (each item at most once — cards dealt from a deck). The Counting Principle disposes of the *ordered* versions of both:

Proposition 14.4 (Ordered selections)**.** *From n distinct objects:*

(a) the number of ordered sequences of k choices made with *replacement is n^k;*

*(b) the number of ordered sequences of $k \le n$ distinct objects (k-*permutations*; sampling* without *replacement) is $n(n-1)\cdots(n-k+1) = \dfrac{n!}{(n-k)!}$;*

(c) in particular ($k = n$), the number of permutations *(orderings) of n objects is $n!$.*

Proof. All three are the Counting Principle. (a) n choices at each of k stages. (b) n choices at the first stage and — whatever was chosen before — exactly $n-i+1$ remaining choices at the ith. (c) is (b) with $k = n$, using the convention $0! = 1$. □

Order often does not matter: a poker hand is a *set* of five cards, however it was dealt. The trick is to count ordered selections anyway, then divide by the number of orderings that yield the same unordered selection — legal precisely because that number ($k!$ below) is the same for every unordered selection.

Proposition 14.5 (Combinations)**.** *The number of k-element subsets (*combinations*) of an n-element set is the* binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}.$$

Proof. Each k-element subset arises from exactly $k!$ of the $n!/(n-k)!$ ordered k-permutations — one per ordering of its elements. So (number of subsets) $\cdot\, k! = n!/(n-k)!$. □

Proposition 14.6 (Three identities, by counting)**.** *For integers $0 \le k \le n$:*

(a) $\displaystyle\binom{n}{k} = \binom{n}{n-k}$*;*

(b) $\displaystyle\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$ *(Pascal's rule; $1 \le k \le n-1$);*

(c) $\displaystyle\sum_{k=0}^{n}\binom{n}{k}=2^n$.

Proof. Each identity is two counts of the same set — no algebra required. (a) Choosing which k objects to take is the same choice as choosing which $n-k$ to leave behind. (b) Split the k-element subsets of $\{1,\dots,n\}$ by whether they contain element 1: those that do are counted by $\binom{n-1}{k-1}$ (choose the rest from the other $n-1$), those that don't by $\binom{n-1}{k}$. (c) The left side counts the subsets of an n-element set, grouped by size. So does 2^n: build a subset by one binary decision per element — in or out — which is the Counting Principle with n stages of 2 choices; equivalently, subsets correspond one-to-one to bit-vectors of length n. □

Proposition 14.7 (Partitions: the multinomial coefficient)**.** *Let $n_1+\cdots+n_r=n$ with all $n_i\ge 0$. The number of ways to partition an n-element set into r labeled groups, the ith of size n_i, is*

$$\binom{n}{n_1,\dots,n_r}=\frac{n!}{n_1!\,n_2!\cdots n_r!}.$$

Proof. Form the groups one at a time: $\binom{n}{n_1}$ choices for the first group, then $\binom{n-n_1}{n_2}$ for the second, and so on. By the Counting Principle the count is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r}=\frac{n!}{n_1!\,(n-n_1)!}\cdot\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!}\cdots=\frac{n!}{n_1!\cdots n_r!},$$

the intermediate factorials cancelling telescopically. ($r=2$ recovers Proposition 14.5: a combination is a partition into "chosen" and "left behind.") □

With counting in hand, we return to the axioms' most important derived notion: reasoning from partial information.

Definition 14.8 (Conditional probability)**.** For events A,B with $\mathbf{P}(B)>0$, the *conditional probability of A given B* is

$$\mathbf{P}(A\mid B)=\frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}.$$

Conditioning is re-normalization: the outcomes in B become the new sample space, with probabilities scaled up by $1/\mathbf{P}(B)$ so they again sum to 1. One checks directly that $\mathbf{P}(\cdot\mid B)$ satisfies the three axioms — so *every* theorem about probability laws applies verbatim to conditional laws. This innocuous remark is used constantly.

Theorem 14.9 (Multiplication rule, total probability, Bayes' rule)**.** *Let $A_1,\dots,A_n$ be disjoint events whose union is Ω (a* partition*), each with $\mathbf{P}(A_i)>0$, and let B be an event with $\mathbf{P}(B)>0$.*

(a) Multiplication rule: $\mathbf{P}(A\cap B)=\mathbf{P}(B)\,\mathbf{P}(A\mid B)=\mathbf{P}(A)\,\mathbf{P}(B\mid A)$ *(whenever the conditioning events have positive probability); more generally*

$$\mathbf{P}(A_1\cap\cdots\cap A_n)=\mathbf{P}(A_1)\,\mathbf{P}(A_2\mid A_1)\cdots\mathbf{P}(A_n\mid A_1\cap\cdots\cap A_{n-1}).$$

(b) Total probability theorem: $\displaystyle\mathbf{P}(B)=\sum_{i=1}^{n}\mathbf{P}(A_i)\,\mathbf{P}(B\mid A_i)$.

(c) Bayes' rule:

$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i)\,\mathbf{P}(B \mid A_i)}{\sum_{j=1}^{n} \mathbf{P}(A_j)\,\mathbf{P}(B \mid A_j)}.$$

Proof. (a) is the definition of conditional probability rearranged; the general form follows by induction, multiplying one more conditional probability at each step (each factor's numerator cancels the previous factor's intersection).

(b) The sets $B \cap A_1, \ldots, B \cap A_n$ are disjoint (the A_i are) and their union is B (the A_i cover Ω). By additivity and then the multiplication rule,

$$\mathbf{P}(B) = \sum_i \mathbf{P}(B \cap A_i) = \sum_i \mathbf{P}(A_i)\,\mathbf{P}(B \mid A_i).$$

(c) By definition and then (a) for the numerator and (b) for the denominator:

$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A_i)\,\mathbf{P}(B \mid A_i)}{\sum_j \mathbf{P}(A_j)\,\mathbf{P}(B \mid A_j)}.$$ □

Bayes' rule is *inference*: the A_i are competing hypotheses ("causes"), B is the observed evidence ("effect"), and the rule converts the easy direction $\mathbf{P}(\text{effect} \mid \text{cause})$ — usually given by a model — into the wanted direction $\mathbf{P}(\text{cause} \mid \text{effect})$. The $\mathbf{P}(A_i)$ are *priors*, the $\mathbf{P}(A_i \mid B)$ *posteriors*.

Definition 14.10 (Independence)**.** Events A and B are *independent* if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B)$ (equivalently, when $\mathbf{P}(B) > 0$: $\mathbf{P}(A \mid B) = \mathbf{P}(A)$ — learning B tells you nothing about A). Events A and B are *conditionally independent given* C if $\mathbf{P}(A \cap B \mid C) = \mathbf{P}(A \mid C)\,\mathbf{P}(B \mid C)$.

Two warnings, both of which machine learning's graphical-models lectures weaponize: independence does *not* survive conditioning (two independent coin flips become dependent given "the two flips differ"), and conditional independence does not imply independence (nor conversely). Independence is a modeling *assumption* about the world, not something you can usually read off the numbers.

14.2 Practice: by hand

Example 14.11 (Counting first: phone numbers and poker hands)**.** (After B&T §1.6.) A local phone number is a 7-digit sequence whose first digit is neither 0 nor 1: by the Counting Principle there are $8 \cdot 10^6$ of them. A poker hand is 5 cards from 52: dealt *in order* there are $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 = 52!/47!$ possible deals, and each hand accounts for 5! of them, so there are

$$\binom{52}{5} = \frac{52!}{5!\,47!} = 2{,}598{,}960$$

hands. If all hands are equally likely, the probability that a hand is all spades is $\binom{13}{5}/\binom{52}{5} = 1287/2{,}598{,}960 \approx 4.95 \times 10^{-4}$ — probability by counting, twice over.

Example 14.12 (Anagrams and coin flips)**.** How many distinct letter sequences can be formed by rearranging SYSTEMS? The seven positions are partitioned into labeled groups: which three

positions get an S, and which single positions get the Y, T, E, M. By Proposition 14.7 the count is $\binom{7}{3,1,1,1,1} = 7!/3! = 840$. And the probability of exactly two heads in four fair flips is

$$\frac{\binom{4}{2}}{2^4} = \frac{6}{16} = \frac{3}{8}:$$

the 2^4 outcome sequences are equally likely, and choosing *which* two flips come up heads is a combination. These are the binomial probabilities of Lesson 15, arriving a lesson early.

Example 14.13 (Two rolls of a fair die)**.** $\Omega = \{(i,j) : 1 \le i,j \le 6\}$, all 36 outcomes equally likely. Let $A =$ "the sum is 8" $= \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$, so $\mathbf{P}(A) = 5/36$. Let $B =$ "the first roll is even" (18 outcomes, $\mathbf{P}(B) = 1/2$). Then $A \cap B = \{(2,6),(4,4),(6,2)\}$, so

$$\mathbf{P}(A \mid B) = \frac{3/36}{18/36} = \frac{1}{6} > \frac{5}{36} = \mathbf{P}(A):$$

evidence that the first roll is even makes a sum of 8 slightly *more* likely — A and B are dependent.

Example 14.14 (The false-positive puzzle — Bayes' rule you must own)**.** A test for a condition affecting 1 in 1000 people is 99% accurate in both directions: $\mathbf{P}(+ \mid \text{sick}) = 0.99$ and $\mathbf{P}(- \mid \text{healthy}) = 0.99$. You test positive. The probability you are sick is *not* 0.99; by Bayes' rule with the partition $\{\text{sick}, \text{healthy}\}$:

$$\mathbf{P}(\text{sick} \mid +) = \frac{0.001 \times 0.99}{0.001 \times 0.99 + 0.999 \times 0.01} = \frac{0.00099}{0.00099 + 0.00999} \approx 0.09.$$

About 9%: the false positives from the huge healthy population swamp the true positives from the tiny sick one. The prior matters as much as the test.

Example 14.15 (Conditional independence, concretely)**.** Two coins: one fair, one with $\mathbf{P}(\text{heads}) = 0.9$. Pick one at random and flip it twice. Given *which* coin was picked, the flips are independent. Unconditionally they are not: a first head is evidence for the biased coin, which makes a second head more likely. Compute: $\mathbf{P}(H_1) = \frac{1}{2}(0.5) + \frac{1}{2}(0.9) = 0.7$, while

$$\mathbf{P}(H_2 \mid H_1) = \frac{\mathbf{P}(H_1 \cap H_2)}{\mathbf{P}(H_1)} = \frac{\frac{1}{2}(0.25) + \frac{1}{2}(0.81)}{0.7} = \frac{0.53}{0.7} \approx 0.757 > 0.7.$$

This two-coin experiment is a two-node Bayesian network, and the calculation you just did is exactly machine learning's "inference by enumeration."

Machine-learning connection. A Bayesian network *is* the multiplication rule with a conditional-independence assumption: the joint distribution factors as $\mathbf{P}(X_1,\dots,X_n) = \prod_i \mathbf{P}(X_i \mid \text{Parents}(X_i))$. Inference in the network — filtering in an HMM, diagnosing from symptoms — is Bayes' rule plus total probability, applied systematically. Example 14.14 *is* a two-variable Bayesian network. When machine learning says "condition on the evidence and renormalize," it is invoking the fact that $\mathbf{P}(\cdot \mid B)$ is a probability law.

14.3 Exercises

Exercise 14.1 (Hand)**.** Three fair coin flips. List the sample space. Let $A =$ "at least two heads" and $B =$ "the first flip is heads." Compute $\mathbf{P}(A)$ and $\mathbf{P}(A \mid B)$, and check whether A and B are independent.

Exercise 14.2 (Hand)**.** A spam filter knows: 30% of email is spam, the word "free" appears in 40% of spam and 5% of non-spam. Apply Bayes' rule to find $\mathbf{P}(\text{spam} \mid \text{"free"})$. Then recompute with a prior of 80% spam and note how the posterior moves.

Exercise 14.3 (Proof, $\star$)**.** Prove the general union bound $\mathbf{P}\big(\bigcup_{i=1}^{n} A_i\big) \le \sum_{i=1}^{n} \mathbf{P}(A_i)$ by induction from Proposition 14.2(c). (This one-line tool underlies half of the analyses in machine learning theory.)

Exercise 14.4 (Proof)**.** Verify that $\mathbf{P}(\cdot \mid B)$ satisfies the three probability axioms. Then prove the "chain rule with a spectator": $\mathbf{P}(A \cap C \mid B) = \mathbf{P}(C \mid B)\,\mathbf{P}(A \mid B \cap C)$ whenever the conditioning events have positive probability.

Exercise 14.5 (Proof)**.** Show that if A and B are independent, so are A and B^c, and so are A^c and B^c.

Exercise 14.6 (Hand)**.** A byte is a string of 8 bits. How many bytes are there? How many contain exactly three 1s? If a byte is chosen uniformly at random, what is the probability that it contains exactly three 1s? Exactly zero? Exactly eight?

Exercise 14.7 (Hand)**.** (a) How many distinct letter sequences can be formed by rearranging FILTER? By rearranging SIGNALS? (b) A committee of 4 is to be chosen from 10 people. How many committees are there? How many contain one designated person, and how many do not? Check that your three committee counts agree with Pascal's rule, Proposition 14.6(b).

Exercise 14.8 (Proof)**.** (The club identity.) From n people, form a club consisting of one leader plus any set (possibly empty) of additional members. Count the clubs in two ways — leader first, then members; and size-k club first, then its leader — and conclude that

$$\sum_{k=1}^{n} k \binom{n}{k} = n\,2^{\,n-1}.$$

Exercise 14.9 (Proof, $\star$)**.** (Stars and bars.) Show that the number of ways to place k indistinguishable balls into n labeled bins — equivalently, the number of *unordered* samples of size k drawn *with* replacement from n objects — is $\binom{n+k-1}{k}$. (Encode a placement as a binary string: a 0 for each ball and a 1 for each of the $n-1$ walls between adjacent bins; show that placements correspond one-to-one to strings with k 0s and $n-1$ 1s. Note why the divide-by-duplicates trick behind Proposition 14.5 does *not* apply directly here: the number of ordered placements per unordered placement is not constant.)

Deeper reading. B&T 1.1–1.2 (models and axioms), 1.3–1.4 (conditioning, total probability, Bayes), 1.5 (independence), 1.6 (counting, developed above); the machine learning "Bayesian networks" module notes re-derive Theorem 14.9 in network notation.

15 Discrete Random Variables and Expectation

NEEDED FOR: Machine learning (core: utilities, expectimax, losses); statistical inference (assumed)

15.1 Theory

Definition 15.1 (Random variable, PMF)**.** A *random variable* is a function $X : \Omega \to \mathbb{R}$ — a number determined by the outcome. It is *discrete* if its range is finite or countable. The *probability mass function* (PMF) of a discrete X is

$$p_X(x) = \mathbf{P}(X = x) \qquad \big(\text{shorthand for } \mathbf{P}(\{\omega \in \Omega : X(\omega) = x\})\big),$$

and satisfies $p_X(x) \geq 0$ and $\sum_x p_X(x) = 1$ (additivity over the disjoint events $\{X = x\}$).

The four discrete distributions to know cold:

- *Bernoulli(p)*: $X \in \{0, 1\}$ with $p_X(1) = p$. Models one yes/no trial.
- *Binomial(n, p)*: X = number of successes in n independent Bernoulli(p) trials; $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, \ldots, n$.
- *Geometric(p)*: X = number of trials up to and including the first success; $p_X(k) = (1-p)^{k-1} p$ for $k = 1, 2, \ldots$.
- *Poisson(λ)*: $p_X(k) = e^{-\lambda} \lambda^k / k!$ for $k = 0, 1, 2, \ldots$; the $n \to \infty$, $np \to \lambda$ limit of the binomial (Lesson 22 uses this).

Definition 15.2 (Expectation, variance)**.** The *expectation* (mean) of a discrete random variable X is

$$\mathbf{E}[X] = \sum_x x \, p_X(x)$$

(assuming the sum converges absolutely), and its *variance* is

$$\operatorname{var}(X) = \mathbf{E}\big[(X - \mathbf{E}[X])^2\big] \geq 0, \qquad \text{with } \textit{standard deviation } \sigma_X = \sqrt{\operatorname{var}(X)}.$$

Theorem 15.3 (Expected value rule / LOTUS)**.** *For any function $g : \mathbb{R} \to \mathbb{R}$,*

$$\mathbf{E}[g(X)] = \sum_x g(x) \, p_X(x).$$

Proof. Let $Y = g(X)$; Y is discrete with $p_Y(y) = \sum_{x:\, g(x)=y} p_X(x)$ (additivity over the disjoint events $\{X = x\}$ composing $\{Y = y\}$). Then

$$\mathbf{E}[Y] = \sum_y y \, p_Y(y) = \sum_y \sum_{x:\, g(x)=y} y \, p_X(x) = \sum_y \sum_{x:\, g(x)=y} g(x) \, p_X(x) = \sum_x g(x) \, p_X(x),$$

since every x appears in exactly one inner sum. □

The name LOTUS — "law of the unconscious statistician" — records that people apply it without noticing it needs proof: the naive move (average $g(x)$ against p_X, without ever computing $p_{g(X)}$) is correct.

Theorem 15.4 (Linearity of expectation)**.** *For random variables X, Y on the same experiment and constants a, b:*

$$\mathbf{E}[aX + b] = a\,\mathbf{E}[X] + b, \qquad \mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y].$$

The second identity requires no *independence.*

Proof. The first is LOTUS with $g(x) = ax+b$: $\sum_x (ax+b)p_X(x) = a\sum_x xp_X(x) + b\sum_x p_X(x) = a\mathbf{E}[X] + b$. For the second, work on the underlying sample space (for simplicity, countable Ω, writing $p(\omega) = \mathbf{P}(\{\omega\})$):

$$\mathbf{E}[X + Y] = \sum_\omega \big(X(\omega) + Y(\omega)\big)p(\omega) = \sum_\omega X(\omega)p(\omega) + \sum_\omega Y(\omega)p(\omega) = \mathbf{E}[X] + \mathbf{E}[Y]. \qquad \square$$

Proposition 15.5 (Variance identities)**.** $\operatorname{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$*, and* $\operatorname{var}(aX + b) = a^2 \operatorname{var}(X)$*.*

Proof. Write $\mu = \mathbf{E}[X]$. Expanding the square and using linearity,

$$\operatorname{var}(X) = \mathbf{E}[X^2 - 2\mu X + \mu^2] = \mathbf{E}[X^2] - 2\mu\mathbf{E}[X] + \mu^2 = \mathbf{E}[X^2] - \mu^2.$$

For the second: $aX+b$ has mean $a\mu+b$, so $\operatorname{var}(aX+b) = \mathbf{E}[(aX+b-a\mu-b)^2] = \mathbf{E}[a^2(X-\mu)^2] = a^2 \operatorname{var}(X)$ — the shift b disappears, as it should: shifting a distribution does not change its spread. $\square$

Example 15.6 (The catalog, computed)**.** *Bernoulli(p):* $\mathbf{E}[X] = p$, and $\mathbf{E}[X^2] = p$ (since $X^2 = X$), so $\operatorname{var}(X) = p - p^2 = p(1-p)$ — maximized at $p = \frac{1}{2}$, the most unpredictable coin.

Binomial(n, p): write $X = X_1 + \cdots + X_n$ as a sum of Bernoulli indicators. By linearity (no independence needed!): $\mathbf{E}[X] = np$. For the variance, independence *is* needed (Lesson 16): $\operatorname{var}(X) = np(1-p)$.

Geometric(p): $\mathbf{E}[X] = 1/p$. Slick derivation via the total expectation theorem (next lesson) or the derivative trick: for $|q| < 1$, differentiate $\sum_{k\geq 0} q^k = \frac{1}{1-q}$ to get $\sum_{k\geq 1} kq^{k-1} = \frac{1}{(1-q)^2}$, so with $q = 1-p$: $\mathbf{E}[X] = p\sum_{k\geq 1} kq^{k-1} = p \cdot \frac{1}{p^2} = \frac{1}{p}$.

Poisson(λ): $\mathbf{E}[X] = \sum_{k\geq 1} k\, e^{-\lambda}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k\geq 1}\frac{\lambda^{k-1}}{(k-1)!} = \lambda$; a similar computation gives $\operatorname{var}(X) = \lambda$ too.

15.2 Practice: by hand

Example 15.7 (Expected utility, the machine learning shape)**.** A game: pay 1 to play; roll a die; win 4 if you roll a six, else nothing. Your net gain G has $p_G(3) = \frac{1}{6}$, $p_G(-1) = \frac{5}{6}$, so

$$\mathbf{E}[G] = 3 \cdot \tfrac{1}{6} + (-1) \cdot \tfrac{5}{6} = -\tfrac{1}{3}.$$

Play 100 times and linearity says your expected total is -33.3: no cleverness about dependence between rounds required. Decisions in machine learning (expectimax, MDP policies) are rankings of actions by exactly this kind of number.

Example 15.8 (Indicators make hard counts easy)**.** n people throw their hats in a pile and each takes one back uniformly at random. How many get their own hat, on average? Let X_i be

the indicator that person i succeeds: $\mathbf{E}[X_i] = \mathbf{P}(X_i = 1) = 1/n$. The total is $X = \sum_i X_i$, so by linearity

$$\mathbf{E}[X] = \sum_{i=1}^{n} \frac{1}{n} = 1,$$

for *every* n — even though the X_i are dependent (if $n-1$ people got their own hats, so did the last). Linearity through dependence is the single most-used trick in average-case analysis.

Machine-learning connection. An MDP policy is evaluated by its *expected utility*; the value of a chance node in expectimax is $\sum_s \mathbf{P}(s)\, V(s)$ — an expectation; the training loss $\frac{1}{m}\sum_i \text{Loss}_i$ is the expectation of the loss under the empirical distribution of the data, and generalization asks how it relates to the expectation under the true distribution. When machine learning writes $V_\pi(s) = \sum_{s'} T(s,a,s')[\text{Reward} + \gamma V_\pi(s')]$, read it as: value = expected (reward + discounted future value). Expectation is machine learning's unit of account.

15.3 Exercises

Exercise 15.1 (Hand)**.** Let X be the number of heads in 3 fair flips. Tabulate p_X, compute $\mathbf{E}[X]$ and $\text{var}(X)$ from the table, and confirm both against the Binomial$(3, \frac{1}{2})$ formulas.

Exercise 15.2 (Hand)**.** A quiz has 5 multiple-choice questions, 4 options each; you guess uniformly and independently. Let X = number correct. What are $\mathbf{E}[X]$ and $\text{var}(X)$? What is $\mathbf{P}(X \geq 1)$? (Complement trick.)

Exercise 15.3 (Proof, $\star$)**.** Prove that $\mathbf{E}[X]$ minimizes the mean squared error: for any constant c, $\mathbf{E}[(X-c)^2] = \text{var}(X) + (c - \mathbf{E}[X])^2$, so the unique minimizing c is $\mathbf{E}[X]$. (Expand; linearity does the rest. Lesson 19 upgrades this from constants to functions of an observation — it becomes MMSE estimation.)

Exercise 15.4 (Proof)**.** Derive $\mathbf{E}[X] = 1/p$ for the geometric distribution using the memoryless recursion: condition on the first trial to get $\mathbf{E}[X] = 1 + (1-p)\mathbf{E}[X]$, and justify the conditioning step. (This is the total expectation theorem of Lesson 16 in action; compare with the derivative trick of Example 15.6.)

Deeper reading. B&T 2.1–2.4 (PMFs, expectation, variance; the distribution catalog); B&T 2.4 has the geometric-series variance computations spelled out.

16 Multiple Random Variables and Conditioning

NEEDED FOR: Machine learning (core: naive Bayes, HMMs, MDP value computations); statistical inference (assumed)

16.1 Theory

Definition 16.1 (Joint, marginal, conditional PMFs)**.** For discrete random variables X, Y on the same experiment, the *joint PMF* is $p_{X,Y}(x,y) = \mathbf{P}(X = x,\, Y = y)$. The *marginals* recover

each variable alone:

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y)$$

(additivity over the partition by the other variable's value). For $p_Y(y) > 0$, the *conditional PMF* is

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)},$$

a genuine PMF in x for each fixed y. X and Y are *independent* if $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$ for all x, y.

All of Lesson 14 transfers: the multiplication rule reads $p_{X,Y}(x,y) = p_Y(y)\,p_{X|Y}(x \mid y)$, and Bayes' rule swaps which variable is conditioned on. Everything extends to three or more variables by adding coordinates.

Definition 16.2 (Conditional expectation)**.** $\mathbf{E}[X \mid Y = y] = \sum_x x\,p_{X|Y}(x \mid y)$: the mean of X in the re-normalized world where $Y = y$ is known.

Theorem 16.3 (Total expectation theorem)**.**

$$\mathbf{E}[X] = \sum_y p_Y(y)\,\mathbf{E}[X \mid Y = y].$$

Proof. Expand the right side using the definitions and the multiplication rule:

$$\sum_y p_Y(y) \sum_x x\,p_{X|Y}(x \mid y) = \sum_y \sum_x x\,p_{X,Y}(x,y) = \sum_x x \sum_y p_{X,Y}(x,y) = \sum_x x\,p_X(x) = \mathbf{E}[X],$$

where the swap of summation order is legal by absolute convergence. □

Read it as *divide and conquer for averages*: average within each scenario y, then average the scenario averages, weighted by how likely each scenario is. This theorem is the engine of every MDP value calculation and of Lesson 19.

Theorem 16.4 (Expectations and products under independence)**.** *If X and Y are independent, then*

$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y] \qquad \textit{and} \qquad \operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y).$$

Proof. For the product, using LOTUS on the pair (whose proof is identical to Theorem 15.3's) and factoring the joint PMF:

$$\mathbf{E}[XY] = \sum_{x,y} xy\,p_X(x)p_Y(y) = \Big(\sum_x x p_X(x)\Big)\Big(\sum_y y p_Y(y)\Big) = \mathbf{E}[X]\,\mathbf{E}[Y].$$

For the variance, let $\tilde{X} = X - \mathbf{E}[X]$ and $\tilde{Y} = Y - \mathbf{E}[Y]$ (still independent, with mean zero). Then

$$\operatorname{var}(X+Y) = \mathbf{E}[(\tilde{X}+\tilde{Y})^2] = \mathbf{E}[\tilde{X}^2] + 2\,\mathbf{E}[\tilde{X}\tilde{Y}] + \mathbf{E}[\tilde{Y}^2] = \operatorname{var}(X) + 2\,\mathbf{E}[\tilde{X}]\mathbf{E}[\tilde{Y}] + \operatorname{var}(Y),$$

and the cross term is $2 \cdot 0 \cdot 0 = 0$. □

Contrast with Theorem 15.4: means *always* add; variances add only under independence (Lesson 18 measures the failure by *covariance*). This completes the binomial variance claimed in Example 15.6: a sum of n independent Bernoulli(p)'s has variance $np(1-p)$.

16.2 Practice: by hand

Example 16.5 (A full joint-PMF workout)**.** Let (X, Y) have joint PMF given by the table (rows $x = 0, 1$; columns $y = 0, 1, 2$):

$p_{X,Y}$	$y=0$	$y=1$	$y=2$
$x=0$	0.1	0.2	0.1
$x=1$	0.2	0.1	0.3

Marginals: $p_X(0) = 0.4$, $p_X(1) = 0.6$; $p_Y = (0.3, 0.3, 0.4)$. Conditional: $p_{X|Y}(1 \mid 2) = 0.3/0.4 = 0.75$. Independence fails: $p_{X,Y}(1, 2) = 0.3$ but $p_X(1)p_Y(2) = 0.24$. Conditional expectation: $\mathbf{E}[X \mid Y = 2] = 0.75$, while $\mathbf{E}[X] = 0.6$ — observing $Y = 2$ raises the estimate of X. Total expectation check: $0.3 \cdot \frac{0.2}{0.3} + 0.3 \cdot \frac{0.1}{0.3} + 0.4 \cdot 0.75 = 0.2 + 0.1 + 0.3 = 0.6 = \mathbf{E}[X]$. ✓

Example 16.6 (Naive Bayes, by hand)**.** Classify an email as spam ($C = 1$) or not ($C = 0$) from two binary features: F_1 = contains "free," F_2 = contains "meeting." Model: $\mathbf{P}(C = 1) = 0.3$, and *given the class*, features are conditionally independent with

$$\mathbf{P}(F_1 = 1 \mid C = 1) = 0.4, \quad \mathbf{P}(F_1 = 1 \mid C = 0) = 0.05,$$
$$\mathbf{P}(F_2 = 1 \mid C = 1) = 0.02, \quad \mathbf{P}(F_2 = 1 \mid C = 0) = 0.3.$$

Observe $F_1 = 1, F_2 = 0$. By Bayes' rule with the factored likelihood:

$$\mathbf{P}(C = 1 \mid F_1 = 1, F_2 = 0) = \frac{0.3 \times 0.4 \times 0.98}{0.3 \times 0.4 \times 0.98 + 0.7 \times 0.05 \times 0.7} = \frac{0.1176}{0.1176 + 0.0245} \approx 0.83.$$

"Free" without "meeting" pushes the posterior from 30% to 83% spam. The conditional-independence assumption is what let us multiply the per-feature likelihoods — with d features it reduces 2^d joint parameters to $2d$.

> **Machine-learning connection.** Example 16.6 *is* a Bayesian network ($C \to F_1$, $C \to F_2$), and the computation is machine learning's inference-by-enumeration algorithm. An HMM adds time: hidden states form a chain (Lesson 23), each emitting an observation, and filtering alternates the multiplication rule (predict) with Bayes' rule (update). The total expectation theorem is also how Bellman equations average over next states in MDPs. Master the two examples above and the graphical-models unit becomes bookkeeping.

16.3 Exercises

Exercise 16.1 (Hand)**.** Roll two fair dice; let X = the minimum, Y = the maximum. Tabulate the joint PMF for $X, Y \in \{1, 2, 3\}$ restricted to rolls of a 3-sided die (i.e. two fair 3-sided dice), find both marginals, and compute $\mathbf{E}[X \mid Y = 3]$.

Exercise 16.2 (Hand)**.** Extend Example 16.6: observe $F_1 = 1, F_2 = 1$. Compute the posterior. Explain in one sentence why the two features pull in opposite directions.

Exercise 16.3 (Proof, $\star$)**.** Prove the total expectation theorem in its conditional form: $\mathbf{E}[X \mid A] = \sum_y p_{Y|A}(y)\, \mathbf{E}[X \mid Y = y, A]$ for an event A with $\mathbf{P}(A) > 0$, by applying Theorem 16.3 to the conditional law $\mathbf{P}(\cdot \mid A)$ (recall: it satisfies the axioms, so every theorem applies to it).

Exercise 16.4 (Proof)**.** Show that if X and Y are independent, then so are $g(X)$ and $h(Y)$ for any functions g, h; conclude $\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)]\, \mathbf{E}[h(Y)]$.

Deeper reading. B&T 2.5–2.7 (joint PMFs, conditioning, independence); B&T 4.3 revisits conditional expectation as a random variable, which is Lesson 19's starting point.

17 Continuous Random Variables and the Normal

NEEDED FOR: Machine learning (light: Gaussian noise, continuous features); statistical inference (assumed, heavily)

17.1 Theory

Definition 17.1 (PDF, CDF)**.** A random variable X is *continuous* if there is a nonnegative function f_X (the *probability density function*, PDF) with

$$\mathbf{P}(a \le X \le b) = \int_a^b f_X(x)\,dx \quad \text{for all } a \le b, \qquad \int_{-\infty}^{\infty} f_X(x)\,dx = 1.$$

The *cumulative distribution function* (CDF) of any random variable is $F_X(x) = \mathbf{P}(X \le x)$; for continuous X, $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$ and $f_X = F_X'$ wherever the derivative exists.

Two immediate consequences. First, $\mathbf{P}(X = x) = \int_x^x f_X = 0$ for every single point: individual values have probability zero, and only *intervals* carry probability — $f_X(x)$ is probability *per unit length* ($\mathbf{P}(x \le X \le x + \delta) \approx f_X(x)\,\delta$ for small δ), not a probability; it can exceed 1. Second, the CDF is the universal translator: it is defined for discrete and continuous variables alike, and differentiating (continuous) or differencing (discrete) recovers the local description.

Expectations are integrals now,

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\,dx, \qquad \mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\,dx,$$

and *every* algebraic theorem of Lessons 15–16 (linearity, LOTUS, variance identities, total expectation, independence factoring) holds verbatim with sums replaced by integrals and PMFs by PDFs — the proofs transpose symbol-for-symbol. Joint PDFs $f_{X,Y}$, marginals ($\int f_{X,Y}\,dy$), and conditionals ($f_{X|Y} = f_{X,Y}/f_Y$) likewise mirror the discrete case.

The continuous catalog:

- *Uniform(a,b)*: $f_X(x) = \frac{1}{b-a}$ on $[a,b]$; $\mathbf{E}[X] = \frac{a+b}{2}$, $\mathrm{var}(X) = \frac{(b-a)^2}{12}$.
- *Exponential(λ)*: $f_X(x) = \lambda e^{-\lambda x}$ for $x \ge 0$; $\mathbf{E}[X] = 1/\lambda$, $\mathrm{var}(X) = 1/\lambda^2$; CDF $F_X(x) = 1 - e^{-\lambda x}$. The continuous geometric: memoryless (Lesson 22 proves it and builds the Poisson process on it).
- *Normal (Gaussian) $\mathcal{N}(\mu, \sigma^2)$*:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-\mu)^2/2\sigma^2}, \qquad \mathbf{E}[X] = \mu, \quad \mathrm{var}(X) = \sigma^2.$$

Proposition 17.2 (Standardization)**.** *If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a \neq 0$, then $aX + b \sim \mathcal{N}(a\mu + b,\, a^2\sigma^2)$. In particular $Z = (X - \mu)/\sigma \sim \mathcal{N}(0,1)$ (the* standard normal*), so every normal probability reduces to the standard CDF $\Phi(z) = \mathbf{P}(Z \le z)$:*

$$\mathbf{P}(X \le x) = \Phi\Big(\frac{x - \mu}{\sigma}\Big).$$

Proof. Let $Y = aX + b$ with $a > 0$ (the case $a < 0$ is symmetric). CDF first, then differentiate — the standard recipe for derived distributions (Lesson 18 makes it a method):

$$F_Y(y) = \mathbf{P}(aX + b \le y) = \mathbf{P}\Big(X \le \frac{y-b}{a}\Big) = F_X\Big(\frac{y-b}{a}\Big),$$

so by the chain rule

$$f_Y(y) = \frac{1}{a} f_X\Big(\frac{y-b}{a}\Big) = \frac{1}{\sqrt{2\pi}\,a\sigma} \exp\Big(-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}\Big),$$

which is the $\mathcal{N}(a\mu + b, a^2\sigma^2)$ density. Taking $a = 1/\sigma$, $b = -\mu/\sigma$ gives $Z \sim \mathcal{N}(0,1)$, and the displayed identity is the CDF computation just done. □

Why this family rules: sums of many small independent effects are approximately normal (the central limit theorem, Lesson 21 — this is a theorem, not folklore), and the family is closed under the linear operations of Part I (Lesson 20 does the vector version). Useful numbers: $\mathbf{P}(|X - \mu| \le \sigma) \approx 0.68$, $\le 2\sigma \approx 0.95$, $\le 3\sigma \approx 0.997$.

17.2 Practice: by hand

Example 17.3 (Working a normal probability)**.** Scores are $\mathcal{N}(500, 100^2)$. What fraction exceeds 650?

$$\mathbf{P}(X > 650) = 1 - \Phi\Big(\frac{650 - 500}{100}\Big) = 1 - \Phi(1.5) \approx 1 - 0.9332 = 0.067.$$

No new integral was done — standardization plus one table/`scipy` lookup.

Example 17.4 (Exponential tail)**.** Requests arrive with exponential($\lambda = 2$ per minute) gaps. Probability a gap exceeds 1 minute: $\mathbf{P}(X > 1) = e^{-2} \approx 0.135$. Given a gap has already lasted 1 minute, the chance it lasts past 2 minutes is $\mathbf{P}(X > 2 \mid X > 1) = e^{-4}/e^{-2} = e^{-2}$ — the same 0.135: memorylessness, proved in general in Lesson 22.

Example 17.5 (Continuous Bayes' rule)**.** A signal Θ is $\mathcal{N}(0,1)$; you observe $Y = \Theta + N$ where the noise N is $\mathcal{N}(0,1)$, independent. Then $f_{Y|\Theta}(y \mid \theta)$ is the $\mathcal{N}(\theta, 1)$ density, and Bayes' rule for densities gives

$$f_{\Theta|Y}(\theta \mid y) = \frac{f_\Theta(\theta)\, f_{Y|\Theta}(y \mid \theta)}{\int f_\Theta(t)\, f_{Y|\Theta}(y \mid t)\, dt} \propto e^{-\theta^2/2} e^{-(y-\theta)^2/2} \propto e^{-(\theta - y/2)^2},$$

after completing the square in θ. So $\Theta \mid Y = y \sim \mathcal{N}(y/2, \frac{1}{2})$: the posterior mean $y/2$ splits the difference between the prior mean 0 and the data y, weighted by their precisions. This tiny computation is the seed of Kalman filtering (statistical inference) and of every "Gaussian prior × Gaussian likelihood" argument.

Machine-learning / statistical-inference connection. Machine learning keeps continuous probability in a supporting role: Gaussian noise in continuous-state models and particle filtering, densities for continuous features. Statistical inference promotes it to the lead: signals are continuous random variables, noise is Gaussian, and Example 17.5 — estimate a Gaussian signal from a noisy observation — is statistical inference's central estimation problem, solved there in full generality (Lessons 19–20 build the machinery).

17.3 Exercises

Exercise 17.1 (Hand)**.** $X \sim \text{Uniform}(0, 2)$. Compute $\mathbf{E}[X]$, $\text{var}(X)$, and $\mathbf{E}[X^3]$ (LOTUS). Then find the CDF of $Y = X^2$ and differentiate to get f_Y. (CDF-then-differentiate — the derived-distribution recipe of Lesson 18.)

Exercise 17.2 (Hand)**.** $X \sim \mathcal{N}(10, 4)$. Using Φ: compute $\mathbf{P}(X \leq 13)$, $\mathbf{P}(|X - 10| > 4)$, and the value c with $\mathbf{P}(X \leq c) = 0.99$ (use $\Phi^{-1}(0.99) \approx 2.33$). Careful: $\sigma = 2$, not 4.

Exercise 17.3 (Proof, $\star$)**.** Prove that the normal density integrates to 1: let $I = \int_{-\infty}^{\infty} e^{-x^2/2} dx$, compute I^2 as a double integral, switch to polar coordinates, and conclude $I = \sqrt{2\pi}$. (One of the great tricks; Fourier analysis meets it again when computing the Fourier transform of a Gaussian.)

Exercise 17.4 (Proof)**.** Show the exponential is the *only* memoryless continuous positive distribution: if $\mathbf{P}(X > s + t) = \mathbf{P}(X > s)\mathbf{P}(X > t)$ for all $s, t \geq 0$, then the function $g(t) = \ln \mathbf{P}(X > t)$ is additive, and (assuming right-continuity) $g(t) = -\lambda t$ for some $\lambda \geq 0$, i.e. $\mathbf{P}(X > t) = e^{-\lambda t}$.

Deeper reading. B&T 3.1–3.3 (PDFs, CDFs, the normal), 3.4–3.6 (joint densities, conditioning, continuous Bayes); the polar-coordinates trick is B&T's Section 3.3 sidebar.

Part IV: Random Vectors, Limit Theorems, and Stochastic Processes

Statistical inference opens where Part III stops and moves fast: random vectors, convergence and limit theorems, random processes (IID, Markov, Gaussian), stationarity and autocorrelation, MMSE and linear estimation, Kalman filtering, hypothesis testing, PCA. This part is the on-ramp: B&T Chapters 4–8, with the linear-algebra connections (Lessons 6, 10, 13) made explicit, because at this level probability and linear algebra stop being separate subjects. *None of this part is required for machine learning* except as noted for Lesson 23 (Markov chains, which make the MDP and HMM material feel like review).

18 Derived Distributions, Covariance, and Correlation

NEEDED FOR: Statistical inference. Not needed for ML.

18.1 Theory

Theorem 18.1 (Derived distributions: the CDF method and the monotone formula)**.** *Let X be continuous with PDF f_X and let $Y = g(X)$.*

(a) Always*: $F_Y(y) = \mathbf{P}(g(X) \le y)$, computed as an integral of f_X over $\{x : g(x) \le y\}$; differentiate to get f_Y.*

(b) If g is strictly monotone and differentiable with inverse h (so $X = h(Y)$), then

$$f_Y(y) = f_X\big(h(y)\big)\,\big|h'(y)\big|.$$

Proof. (a) is the definition of the CDF plus the definition of a PDF. For (b) with g increasing:

$$F_Y(y) = \mathbf{P}(g(X) \le y) = \mathbf{P}(X \le h(y)) = F_X(h(y)),$$

and the chain rule gives $f_Y(y) = f_X(h(y))\,h'(y)$, with $h' > 0$. For g decreasing the event flips to $\{X \ge h(y)\}$, so $F_Y(y) = 1 - F_X(h(y))$ and $f_Y(y) = -f_X(h(y))h'(y)$ with $h' < 0$; both cases are covered by $|h'(y)|$. □

The factor $|h'(y)|$ is one-dimensional volume scaling — the same role $|\det A|$ played in Lesson 13's change of variables, and the preview of the multivariate formula (Jacobian determinant) statistical inference uses. Proposition 17.2 was this theorem applied to $g(x) = ax + b$.

Definition 18.2 (Covariance, correlation)**.** For random variables X, Y with means μ_X, μ_Y:

$$\operatorname{cov}(X, Y) = \mathbf{E}\big[(X - \mu_X)(Y - \mu_Y)\big] = \mathbf{E}[XY] - \mu_X\mu_Y,$$

(the second form by expanding and linearity), and the *correlation coefficient* is

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} \qquad (\sigma_X, \sigma_Y > 0).$$

X, Y are *uncorrelated* if $\operatorname{cov}(X, Y) = 0$.

Theorem 18.3 (Covariance is an inner product; $|\rho| \le 1$)**.** *On the space of zero-mean random variables (with finite variance), define $\langle X, Y\rangle = \mathbf{E}[XY]$. This satisfies the three inner-product properties of Lesson 2 (positivity — with $\langle X, X\rangle = 0$ forcing $X = 0$ with probability 1 — symmetry, and linearity in each slot). Consequently, by the Cauchy–Schwarz proof of Theorem 2.5, transcribed verbatim:*

$$\big|\mathbf{E}[XY]\big| \le \sqrt{\mathbf{E}[X^2]}\sqrt{\mathbf{E}[Y^2]},$$

i.e. $|\operatorname{cov}(X,Y)| \le \sigma_X\sigma_Y$, i.e. $-1 \le \rho \le 1$, with $|\rho| = 1$ exactly when Y is (with probability 1) a scalar multiple of X.

Proof. Positivity: $\mathbf{E}[X^2] \ge 0$ since $X^2 \ge 0$; and $\mathbf{E}[X^2] = 0$ forces $\mathbf{P}(|X| > \epsilon) = 0$ for every $\epsilon > 0$ (else the expectation would exceed $\epsilon^2\mathbf{P}(|X| > \epsilon) > 0$), so $X = 0$ with probability 1. Symmetry is $XY = YX$; linearity in each slot is linearity of expectation (Theorem 15.4). With the axioms verified, Lesson 2's proof of Cauchy–Schwarz applies word for word — it used nothing but the axioms. Applying it to the centered variables $X - \mu_X$, $Y - \mu_Y$ gives the covariance statement, and the equality case (one variable a multiple of the other) translates to $Y - \mu_Y = c(X - \mu_X)$: a perfect linear relationship. □

This theorem is the hinge of Part IV: *random variables form an inner product space*, in which uncorrelated = orthogonal, standard deviation = norm, and $\rho = \cos\theta$. Lesson 19 will cash this in: best estimation becomes orthogonal projection.

Proposition 18.4 (Variance of a sum, in general)**.**

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y),$$

and independent ⇒ uncorrelated (by Theorem 16.4), but not conversely.

Proof. Expand $\mathbf{E}[(\tilde{X} + \tilde{Y})^2]$ with $\tilde{X}, \tilde{Y}$ centered, as in Theorem 16.4; the cross term is now $2\operatorname{cov}(X, Y)$, kept rather than killed. For the non-converse, one example suffices: let X be uniform on $\{-1, 0, 1\}$ and $Y = X^2$. Then $\mathbf{E}[XY] = \mathbf{E}[X^3] = 0 = \mathbf{E}[X]\mathbf{E}[Y]$: uncorrelated. But $\mathbf{P}(X = 0, Y = 0) = \frac{1}{3} \ne \frac{1}{3} \cdot \frac{1}{3}$: dependent — Y is a deterministic function of X! Covariance detects only *linear* relationships. □

18.2 Practice: by hand

Example 18.5 (Derived distribution, start to finish)**.** $X \sim \text{Uniform}(0, 1)$; find the PDF of $Y = -\ln X$. The inverse of $g(x) = -\ln x$ is $h(y) = e^{-y}$, with $|h'(y)| = e^{-y}$, so for $y > 0$:

$$f_Y(y) = f_X(e^{-y}) \cdot e^{-y} = 1 \cdot e^{-y} = e^{-y}:$$

exponential(1). This is how computers generate exponential random variables from uniform ones (*inverse transform sampling*).

Example 18.6 (Covariance from a joint table)**.** For the joint PMF of Example 16.5: $\mathbf{E}[X] = 0.6$, $\mathbf{E}[Y] = 0.3 \cdot 0 + 0.3 \cdot 1 + 0.4 \cdot 2 = 1.1$, and

$$\mathbf{E}[XY] = \sum_{x,y} xy\, p_{X,Y}(x, y) = 1 \cdot 1 \cdot 0.1 + 1 \cdot 2 \cdot 0.3 = 0.7,$$

so $\operatorname{cov}(X, Y) = 0.7 - 0.6 \times 1.1 = 0.04 > 0$: weakly positively associated, matching Lesson 16's observation that seeing $Y = 2$ raised the estimate of X. With $\operatorname{var}(X) = 0.24$ and $\operatorname{var}(Y) = 0.69$: $\rho = 0.04/\sqrt{0.24 \times 0.69} \approx 0.098$.

Statistical-inference connection. Replace "two random variables X, Y" by "one random signal sampled at two times X_t, X_s" and covariance becomes the *autocorrelation function* $R_X(t,s) = \mathbf{E}[X_t X_s]$ — statistical inference's central object, whose Fourier transform (for stationary processes) is the power spectral density. Every property proved here ($|R| \le \sigma^2$ via Cauchy–Schwarz, bilinearity) is quoted there in the first two weeks.

18.3 Exercises

Exercise 18.1 (Hand)**.** $X \sim \text{Uniform}(0,1)$. Find the PDF of (a) $Y = X^2$ (CDF method — careful, g is monotone here, so check both routes agree) and (b) $Y = 3X+1$ (monotone formula).

Exercise 18.2 (Hand)**.** Two fair dice, $S =$ sum, $D =$ difference (first minus second). Compute $\text{cov}(S, D)$. (Slick route: $S = X + Y$, $D = X - Y$; expand by bilinearity of covariance.) Are S and D independent? (Check a corner case, e.g. $S = 12$ vs $D = 5$.)

Exercise 18.3 (Proof, $\star$)**.** Prove the bilinearity of covariance,

$$\text{cov}\Big(\sum_i a_i X_i,\ \sum_j b_j Y_j\Big) = \sum_i \sum_j a_i b_j \,\text{cov}(X_i, Y_j),$$

and deduce the general variance-of-a-sum formula

$$\text{var}\Big(\sum_i X_i\Big) = \sum_i \text{var}(X_i) + 2\sum_{i<j} \text{cov}(X_i, X_j).$$

Exercise 18.4 (Proof)**.** Let $Y = aX + b$ with $a \neq 0$. Show $\rho(X,Y) = \text{sign}(a)$, confirming $|\rho| = 1$ iff the relationship is exactly linear, with the sign recording direction.

Deeper reading. B&T 4.1 (derived distributions), 4.2 (covariance and correlation); B&T 4.4–4.5 (transforms and random sums; transforms are developed in Lesson 43, where they prove the central limit theorem).

19 Conditional Expectation and Least Mean Squares

NEEDED FOR: Statistical inference (core: MMSE, linear estimation, Kalman lead-in). Not needed for ML.

19.1 Theory

Lesson 16 defined the *number* $\mathbf{E}[X \mid Y = y]$. The upgrade that powers estimation theory is to see it as a *random variable*: define $\mathbf{E}[X \mid Y]$ to be the random variable that takes the value $\mathbf{E}[X \mid Y = y]$ when $Y = y$ — a function of Y.

Theorem 19.1 (Law of iterated expectations)**.** $\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \mathbf{E}[X]$.

Proof. $\mathbf{E}[X \mid Y]$ is a function of Y, say $g(Y)$ with $g(y) = \mathbf{E}[X \mid Y = y]$. By LOTUS and then the total expectation theorem (Theorem 16.3):

$$\mathbf{E}[g(Y)] = \sum_y g(y)\, p_Y(y) = \sum_y \mathbf{E}[X \mid Y = y]\, p_Y(y) = \mathbf{E}[X].$$

(Continuous case: same lines with integrals.) □

Theorem 19.2 (Conditional expectation is the least-mean-squares estimator)**.** *Among* all *functions g, the mean squared error $\mathbf{E}\big[(X-g(Y))^2\big]$ is minimized by $g(Y)=\mathbf{E}[X \mid Y]$.*

Proof. Fix y and work in the conditional world given $Y=y$ (a legitimate probability law, Lesson 14). There, $g(y)$ is just a constant c, and Exercise 15.3 proved $\mathbf{E}[(X-c)^2 \mid Y=y]$ is minimized uniquely by $c=\mathbf{E}[X \mid Y=y]$. So for every y,

$$\mathbf{E}\big[(X-g(Y))^2 \mid Y=y\big] \;\geq\; \mathbf{E}\big[(X-\mathbf{E}[X \mid Y=y])^2 \mid Y=y\big],$$

and averaging both sides over y (total expectation, Theorem 16.3) gives $\mathbf{E}[(X-g(Y))^2] \geq \mathbf{E}[(X-\mathbf{E}[X \mid Y])^2]$. □

Theorem 19.3 (Orthogonality principle)**.** *Let $\tilde{X}=X-\mathbf{E}[X \mid Y]$ be the estimation error. Then for* every *function h,*

$$\mathbf{E}\big[\tilde{X}\,h(Y)\big]=0:$$

the error is uncorrelated with every function of the data. In the inner product of Theorem 18.3, the error is orthogonal *to the space of estimators, so $\mathbf{E}[X \mid Y]$ is the orthogonal projection of X onto the space of functions of Y.*

Proof. Condition on Y and iterate (Theorem 19.1):

$$\mathbf{E}[\tilde{X}h(Y)]=\mathbf{E}\Big[\mathbf{E}\big[\tilde{X}h(Y) \mid Y\big]\Big]=\mathbf{E}\Big[h(Y)\,\mathbf{E}\big[\tilde{X} \mid Y\big]\Big],$$

where $h(Y)$ slides out of the inner conditional expectation because, given Y, it is a constant. And $\mathbf{E}[\tilde{X} \mid Y]=\mathbf{E}[X \mid Y]-\mathbf{E}[X \mid Y]=0$. □

This is Lesson 6, re-lived: Theorem 6.3 said the projection is the closest point and its residual is orthogonal to the subspace; Theorems 19.2 and 19.3 say it again with $\|\cdot\|^2=\mathbf{E}[(\cdot)^2]$. Same geometry, new inner product.

Linear least mean squares. $\mathbf{E}[X \mid Y]$ can be a complicated nonlinear function, and computing it needs the full joint distribution. Statistical inference's workhorse compromise: project onto the smaller subspace of *linear* (affine) estimators $\hat{X}=aY+b$ — computable from means, variances, and one covariance alone.

Theorem 19.4 (Linear LMS)**.** *The minimizers of $\mathbf{E}[(X-aY-b)^2]$ are*

$$a=\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(Y)}, \qquad b=\mu_X-a\mu_Y,$$

so the best linear estimator and its error variance are

$$\hat{X}_L=\mu_X+\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(Y)}\,(Y-\mu_Y), \qquad \mathbf{E}[(X-\hat{X}_L)^2]=(1-\rho^2)\,\sigma_X^2.$$

Proof. For fixed a, the best constant shift is $b=\mathbf{E}[X-aY]=\mu_X-a\mu_Y$ (Exercise 15.3 again), leaving $\mathbf{E}[(\tilde{X}-a\tilde{Y})^2]$ with centered variables. Expand:

$$J(a)=\sigma_X^2-2a\operatorname{cov}(X,Y)+a^2\sigma_Y^2,$$

a parabola in a minimized at $a=\operatorname{cov}(X,Y)/\sigma_Y^2$ (set $J'(a)=0$; $J''=2\sigma_Y^2>0$). Substituting back:

$$J(a^*)=\sigma_X^2-\frac{\operatorname{cov}(X,Y)^2}{\sigma_Y^2}=\sigma_X^2\Big(1-\frac{\operatorname{cov}(X,Y)^2}{\sigma_X^2\sigma_Y^2}\Big)=(1-\rho^2)\sigma_X^2.$$ □

Read the error formula as a scoreboard for ρ: correlation is exactly the fraction of standard deviation that linear prediction can remove (ρ^2 is the "variance explained"). And notice: Theorem 19.4 *is* Lesson 6's least-squares projection onto the two-"vector" span $\{1, Y\}$, with the normal equations $\mathbf{E}[(X - aY - b) \cdot 1] = 0$, $\mathbf{E}[(X - aY - b) \cdot Y] = 0$ — residual orthogonal to each spanning vector.

19.2 Practice: by hand

Example 19.5 (Nonlinear vs. linear estimator)**.** Let $Y \sim \text{Uniform}(-1, 1)$ and $X = Y^2$ (no noise!). The best estimator is $\mathbf{E}[X \mid Y] = Y^2$: zero error. The best *linear* estimator: $\text{cov}(X, Y) = \mathbf{E}[Y^3] - \mathbf{E}[Y^2]\mathbf{E}[Y] = 0$, so $a = 0$ and $\hat{X}_L = \mu_X = \mathbf{E}[Y^2] = \frac{1}{3}$ — a constant, with error variance $\text{var}(Y^2) = \mathbf{E}[Y^4] - \frac{1}{9} = \frac{1}{5} - \frac{1}{9} = \frac{4}{45}$. Linear estimation is blind to this perfectly predictable X because the dependence is purely nonlinear ($\rho = 0$). Moral: LLMS is powerful exactly when correlation captures the relationship — which Lesson 20 shows is always the case for Gaussians.

Example 19.6 (Signal plus noise, the statistical inference primal example)**.** X (signal) has mean 0, variance σ_X^2; N (noise) has mean 0, variance σ_N^2, uncorrelated with X; observe $Y = X + N$. Then $\text{cov}(X, Y) = \sigma_X^2$ and $\text{var}(Y) = \sigma_X^2 + \sigma_N^2$, so

$$\hat{X}_L = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2} Y, \qquad \text{error variance} = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}.$$

The estimator shrinks the observation by the *signal-to-noise ratio*: trust Y when the signal dominates, shrink toward the prior mean 0 when the noise does. With $\sigma_X^2 = \sigma_N^2 = 1$ this is $\hat{X} = Y/2$ — exactly the posterior mean of Example 17.5, where everything was Gaussian and the linear estimator was fully optimal. Kalman filtering (statistical inference) is this formula applied recursively as observations stream in.

Statistical-inference connection. Statistical inference's estimation unit is these two theorems at scale: MMSE = conditional expectation; LLMS = projection using covariances; the orthogonality principle is the certificate for both, and for the Wiener and Kalman filters built from them. The habit to carry in: *when someone says "best estimate," ask "projection onto what subspace, in which inner product?"*

19.3 Exercises

Exercise 19.1 (Hand)**.** For the joint table of Example 16.5: compute $\mathbf{E}[X \mid Y]$ (a value for each y), verify the law of iterated expectations numerically, and find the LLMS estimator $\hat{X}_L = aY + b$ from the moments computed in Lesson 18. Compare its MSE with that of $\mathbf{E}[X \mid Y]$.

Exercise 19.2 (Hand)**.** In Example 19.6 with $\sigma_X = 2, \sigma_N = 1$: compute $\hat{X}_L$ as a function of Y and the error variance. What fraction of the signal's variance does the observation remove?

Exercise 19.3 (Proof, $\star$)**.** (Law of total variance.) Define $\text{var}(X \mid Y)$ as the variance computed in the conditional world, a function of Y. Prove

$$\text{var}(X) = \mathbf{E}\big[\,\text{var}(X \mid Y)\big] + \text{var}\,\big(\mathbf{E}[X \mid Y]\big),$$

by applying $\text{var} = \mathbf{E}[\square^2] - (\mathbf{E}[\square])^2$ twice and the law of iterated expectations. Interpret: total uncertainty = average leftover uncertainty + uncertainty explained by the data.

Exercise 19.4 (Proof)**.** Derive Theorem 19.4 from the orthogonality principle instead of calculus: require $\mathbf{E}[(X - aY - b) \cdot 1] = 0$ and $\mathbf{E}[(X - aY - b) \cdot Y] = 0$ and solve the two linear equations. (These are normal equations, Lesson 6 — with $\mathbf{E}[\cdot\,\cdot]$ as the inner product.)

Deeper reading. B&T 4.3 (conditional expectation and variance revisited — iterated expectations, total variance); B&T 8.3–8.4 (Bayesian least mean squares and *linear* LMS estimation — the two theorems above in their inference setting); then statistical inference's linear-estimation notes will read as review.

20 Random Vectors and Gaussian Vectors

NEEDED FOR: Statistical inference (core: PCA, Gaussian processes, Kalman). Uses Lessons 10 & 13. Not needed for ML.

20.1 Theory

Definition 20.1 (Random vector, mean, covariance matrix)**.** A *random vector* $X = (X_1, \dots, X_n)$ is a vector of random variables on one experiment. Its *mean vector* is $\mu = \mathbf{E}[X] = (\mathbf{E}[X_1], \dots, \mathbf{E}[X_n])$, and its *covariance matrix* is the $n \times n$ matrix

$$\Sigma = \mathbf{E}\big[(X - \mu)(X - \mu)^{\mathsf{T}}\big], \qquad \Sigma_{ij} = \operatorname{cov}(X_i, X_j),$$

with variances on the diagonal.

Proposition 20.2 (Covariance matrices transform like quadratic forms)**.** *For any (deterministic) $m \times n$ matrix A and $b \in \mathbb{R}^m$, the random vector $AX + b$ has mean $A\mu + b$ and covariance matrix $A\Sigma A^{\mathsf{T}}$. In particular, for a single linear functional $w \in \mathbb{R}^n$:*

$$\operatorname{var}(w^{\mathsf{T}} X) = w^{\mathsf{T}} \Sigma\, w.$$

Proof. The mean is linearity of expectation, coordinate by coordinate. For the covariance, let $\tilde{X} = X - \mu$; then $AX + b$ centered is $A\tilde{X}$, and

$$\mathbf{E}\big[(A\tilde{X})(A\tilde{X})^{\mathsf{T}}\big] = \mathbf{E}\big[A\tilde{X}\tilde{X}^{\mathsf{T}}A^{\mathsf{T}}\big] = A\,\mathbf{E}[\tilde{X}\tilde{X}^{\mathsf{T}}]\,A^{\mathsf{T}} = A\Sigma A^{\mathsf{T}},$$

pulling the constant matrices through the (entrywise) expectation. The scalar case is w^{T} as a $1 \times n$ matrix. □

Theorem 20.3 (Covariance matrices are exactly the PSD matrices)**.** *Every covariance matrix is symmetric positive semidefinite. Conversely, every symmetric PSD matrix is the covariance matrix of some random vector.*

Proof. Symmetry: $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$. PSD: for any $w \in \mathbb{R}^n$, $w^{\mathsf{T}}\Sigma w = \operatorname{var}(w^{\mathsf{T}}X) \geq 0$ by Proposition 20.2. Conversely, given symmetric PSD Σ, the spectral theorem (Theorem 10.4) gives $\Sigma = Q\Lambda Q^{\mathsf{T}}$ with $\Lambda = \operatorname{diag}(\lambda_i)$, $\lambda_i \geq 0$; set $\Sigma^{1/2} = Q\Lambda^{1/2}Q^{\mathsf{T}}$ (a symmetric matrix with $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$). Take Z with independent, zero-mean, unit-variance coordinates — so $\operatorname{cov}(Z) = I$ — and set $X = \Sigma^{1/2}Z$: by Proposition 20.2, $\operatorname{cov}(X) = \Sigma^{1/2} I\, \Sigma^{1/2} = \Sigma$. □

Lesson 10's abstract PSD theory just became concrete: *every* covariance matrix has an orthonormal eigenbasis with nonnegative eigenvalues, the eigenvectors are the *principal components* (directions of uncorrelated variation), and the eigenvalues are the variances along them. Changing to the eigenbasis ($Y = Q^{\mathsf{T}}\tilde{X}$, which has covariance $Q^{\mathsf{T}}\Sigma Q = \Lambda$, diagonal) *decorrelates* the vector — that transformation is PCA, and the top-k eigenvalue truncation is Lesson 12's Eckart–Young story applied to data.

Definition 20.4 (Gaussian random vector)**.** X is a *Gaussian (normal) random vector*, $X \sim \mathcal{N}(\mu, \Sigma)$, if it can be written $X = \mu + BZ$ where Z has independent $\mathcal{N}(0,1)$ coordinates and B is any matrix with $BB^{\mathsf{T}} = \Sigma$. Equivalently (the equivalence is proved with the multivariate form of Lesson 43's transform uniqueness theorem; we adopt the constructive definition): every linear combination $w^{\mathsf{T}}X$ is a (one-dimensional) normal random variable. For invertible Σ, the density is

$$f_X(x) = \frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}} \exp\Big(-\tfrac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\Big).$$

The $\sqrt{\det\Sigma}$ is Lesson 13's volume factor (the map B stretches the unit Gaussian ball into the ellipsoid of the distribution), and the level sets of f_X are the ellipsoids $x^{\mathsf{T}}\Sigma^{-1}x = \text{const}$ whose axes are — again — the eigenvectors of Σ (Lesson 10's quadratic-form picture).

Theorem 20.5 (The Gaussian miracles)**.** *Let $X \sim \mathcal{N}(\mu, \Sigma)$.*

(a) Linearity: $AX + b \sim \mathcal{N}(A\mu + b,\, A\Sigma A^{\mathsf{T}})$ — *Gaussians stay Gaussian under every linear map.*

(b) Uncorrelated $\Rightarrow$ independent: *if two coordinates (or blocks) of a jointly Gaussian vector are uncorrelated, they are independent.*

Proof. (a) Write $X = \mu + BZ$ with $BB^{\mathsf{T}} = \Sigma$; then $AX + b = (A\mu + b) + (AB)Z$ with $(AB)(AB)^{\mathsf{T}} = A\Sigma A^{\mathsf{T}}$ — literally the definition, plus Proposition 20.2 for the parameters.

(b) In the density case: if Σ is block-diagonal $\left(\begin{smallmatrix}\Sigma_1 & 0\\ 0 & \Sigma_2\end{smallmatrix}\right)$, then Σ^{-1} is block-diagonal too, $\det\Sigma = \det\Sigma_1 \det\Sigma_2$ (Lesson 13 on the block-triangular form), and the exponent splits as a sum of a term in x_1 and a term in x_2 — so f_X *factors* into the two marginal Gaussian densities, which is independence. □

Part (b) is a miracle worth staring at: Proposition 18.4 warned that uncorrelated is far weaker than independent (recall X and X^2) — *except* in the jointly Gaussian world, where covariance captures all dependence. This is also why Lesson 19's *linear* LMS estimator is fully optimal for Gaussians (Example 19.6): no nonlinear function can see structure that covariances don't already report.

20.2 Practice: by hand

Example 20.6 (A 2-D Gaussian, fully worked)**.** Let $\Sigma = \begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}$, $\mu = 0$ — Example 9.7's matrix as a covariance. Eigen-decomposition (known): eigenvalues $3, 1$ with orthonormal eigenvectors $\frac{1}{\sqrt{2}}(1,1)$, $\frac{1}{\sqrt{2}}(1,-1)$. Read off:

- The density's ellipses have long axis along $(1,1)$ (variance 3) and short axis along $(1,-1)$ (variance 1): the cloud is a tilted ellipse.

- PCA: the first principal component is $(1,1)/\sqrt{2}$, "explaining" $3/(3+1) = 75\%$ of the total variance $\operatorname{tr}\Sigma = 4$ (trace = sum of eigenvalues, Lesson 13).
- Whitening: $Y = \Lambda^{-1/2}Q^{\mathsf{T}}X$ has covariance I (check with Proposition 20.2) — rotate to the eigenbasis, divide each axis by its standard deviation.
- $\operatorname{var}(X_1 - X_2) = w^{\mathsf{T}}\Sigma w$ with $w = (1,-1)$: $2+2-2\cdot 1 = 2$ — the quadratic form doing what Lesson 10 promised.

Statistical-inference connection. A *Gaussian process* is this lesson with infinitely many coordinates: any finite set of samples $(X_{t_1}, \ldots, X_{t_n})$ is a Gaussian vector, so the whole process is specified by a mean function and a covariance (autocorrelation) function. Kalman filtering is Theorem 20.5(a) plus Lesson 19's estimator, iterated. PCA of data is Theorem 20.3's eigen-story with Σ estimated from samples. When statistical inference says "jointly Gaussian, hence independent given uncorrelated," that is miracle (b).

20.3 Exercises

Exercise 20.1 (Hand)**.** Let X_1, X_2 be uncorrelated with variances $1, 4$, and $Y_1 = X_1 + X_2$, $Y_2 = X_1 - X_2$. Write the matrix A with $Y = AX$, compute $\operatorname{cov}(Y) = A\Sigma A^{\mathsf{T}}$, and read off $\operatorname{var}(Y_1)$, $\operatorname{var}(Y_2)$, $\operatorname{cov}(Y_1, Y_2)$. When would Y_1, Y_2 be uncorrelated?

Exercise 20.2 (Hand)**.** For $\Sigma = \left(\begin{smallmatrix} 5 & 2 \\ 2 & 2 \end{smallmatrix}\right)$: find eigenvalues and orthonormal eigenvectors, the fraction of variance along the first principal component, and the whitening transform. Sketch (in words) the density's ellipses.

Exercise 20.3 (Proof, $\star$)**.** Prove that $\Sigma^{1/2} = Q\Lambda^{1/2}Q^{\mathsf{T}}$ is the *unique* symmetric PSD square root of a symmetric PSD Σ. (Uniqueness: if $S^2 = \Sigma$ with S symmetric PSD, show S and Σ have the same eigenvectors — diagonalize S and square. LADR 7C proves this as the "positive operators have unique positive square roots" theorem.)

Exercise 20.4 (Proof)**.** Using Theorem 20.5(a), show that if $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ invertible, the whitened vector $\Lambda^{-1/2}Q^{\mathsf{T}}(X - \mu)$ is $\mathcal{N}(0, I)$ — i.e. its coordinates are independent standard normals. (This is why "every Gaussian is a linearly transformed standard one" can be run in both directions.)

Deeper reading. B&T 4.2 (correlation matrices, in the covariance section); the bivariate normal supplement on the B&T website; LADR 7C (positive operators, square roots) for the linear-algebra half; statistical inference's random-vectors notes then formalize the characteristic-function view.

21 Limit Theorems

NEEDED FOR: Statistical inference (core: convergence, concentration); machine learning (background: why Monte Carlo works)

21.1 Theory

Throughout, $X_1, X_2, \ldots$ are independent identically distributed (i.i.d.) with mean μ and variance $\sigma^2 < \infty$, and

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

is the *sample mean*. By linearity and independence: $\mathbf{E}[M_n] = \mu$ and $\mathrm{var}(M_n) = \sigma^2/n$ — averaging keeps the target and shrinks the spread. The limit theorems sharpen "shrinks" into guarantees.

Theorem 21.1 (Markov and Chebyshev inequalities)**.** *(a) If $X \geq 0$, then for any $a > 0$:* $\mathbf{P}(X \geq a) \leq \dfrac{\mathbf{E}[X]}{a}$.

(b) For any random variable Y with mean μ and variance σ^2, and any $\epsilon > 0$:

$$\mathbf{P}\big(|Y - \mu| \geq \epsilon\big) \leq \frac{\sigma^2}{\epsilon^2}.$$

Proof. (a) Let I be the indicator of $\{X \geq a\}$. Then $aI \leq X$ always: if the event occurs, $aI = a \leq X$; if not, $aI = 0 \leq X$. Taking expectations (monotonicity of $\mathbf{E}$, which follows from nonnegativity of $\mathbf{E}$ applied to $X - aI$): $a\mathbf{P}(X \geq a) = \mathbf{E}[aI] \leq \mathbf{E}[X]$.

(b) Apply (a) to the nonnegative variable $(Y - \mu)^2$ with threshold ϵ^2:

$$\mathbf{P}(|Y - \mu| \geq \epsilon) = \mathbf{P}\big((Y - \mu)^2 \geq \epsilon^2\big) \leq \frac{\mathbf{E}[(Y - \mu)^2]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}. \qquad \square$$

Theorem 21.2 (Weak law of large numbers)**.** *For every $\epsilon > 0$,*

$$\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big) \leq \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0.$$

We say M_n converges in probability *to μ.*

Proof. Chebyshev applied to M_n, whose mean is μ and variance is σ^2/n. $\square$

Three remarks. (1) The bound is quantitative: to pin μ within ϵ with 95% confidence, $n = 20\sigma^2/\epsilon^2$ samples suffice — crude but honest, and the pattern (accuracy costs $1/\epsilon^2$) is right. (2) Chebyshev is distribution-free and correspondingly loose; the Chernoff bound (Lesson 43, built from transforms) trades generality for exponentially better bounds. (3) The *strong* law (B&T 5.5) upgrades the conclusion to $\mathbf{P}(M_n \to \mu) = 1$; the distinction between the two convergence modes is a statistical-inference lecture of its own.

Theorem 21.3 (Central limit theorem)**.** *Let $Z_n = \dfrac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$ (the sum, standardized to mean 0 and variance 1). Then for every z,*

$$\mathbf{P}(Z_n \leq z) \xrightarrow[n\to\infty]{} \Phi(z),$$

the standard normal CDF — regardless of the distribution of the X_i.

The proof needs one more tool — the moment generating function, under which a sum of independent variables becomes a product — and is given in full in Lesson 43: the transform of Z_n converges to the Gaussian's $e^{s^2/2}$, and a continuity theorem converts that into convergence of CDFs. Here the statement and its use are what matter. The practical reading: for large n,

$$X_1 + \cdots + X_n \;\approx\; \mathcal{N}(n\mu,\, n\sigma^2), \qquad M_n \;\approx\; \mathcal{N}(\mu,\, \sigma^2/n),$$

which converts "how unlikely is this deviation?" into a Φ lookup (Lesson 17). The CLT is why the Gaussian owns Lesson 20 and statistical inference: sums of many small independent contributions — thermal noise being the canonical example — are Gaussian by theorem.

21.2 Practice: by hand

Example 21.4 (Polling)**.** Poll n voters; each answer is Bernoulli(p), and M_n estimates p. Since $\sigma^2 = p(1-p) \le \frac{1}{4}$: Chebyshev guarantees $\mathbf{P}(|M_n - p| \ge 0.03) \le \frac{1/4}{n(0.03)^2}$, which is $\le 5\%$ once $n \ge 5556$. The CLT sharpens this: $\mathbf{P}(|M_n - p| \ge 0.03) \approx 2\big(1 - \Phi(0.03\sqrt{n}/\sqrt{p(1-p)}\,)\big) \le 5\%$ once $0.03\sqrt{4n} \ge 1.96$, i.e. $n \ge 1068$ — the famous "$\pm 3\%$ needs about a thousand people," independent of the population size.

Example 21.5 (CLT for a die)**.** Sum 100 fair-die rolls. $\mu = 3.5$, $\sigma^2 = \frac{35}{12} \approx 2.92$ per roll, so the sum is approximately $\mathcal{N}(350, 292)$ and

$$\mathbf{P}(\text{sum} > 380) \approx 1 - \Phi\Big(\tfrac{380-350}{\sqrt{292}}\Big) = 1 - \Phi(1.76) \approx 0.04.$$

An exact computation would need a 100-fold convolution; the CLT needs two moments.

Statistical-inference / machine-learning connection. Statistical inference devotes a unit to exactly this ladder — Markov → Chebyshev → Chernoff, then modes of convergence, then the CLT — and uses it to reason about detection error rates and generalization in learning. For machine learning the payoff is quieter but constant: every Monte Carlo estimate (particle filters, MCTS rollouts, sampled gradients in SGD) is a sample mean whose $\sigma/\sqrt{n}$ error bar and approximate normality come from these theorems.

21.3 Exercises

Exercise 21.1 (Hand)**.** A server handles requests with mean processing time 2 ms and standard deviation 1 ms (distribution unknown). Bound $\mathbf{P}$(total time for 400 requests $>$ 900 ms) two ways: Chebyshev, then CLT approximation. Compare.

Exercise 21.2 (Hand)**.** How many fair-coin flips guarantee, via Chebyshev, that the observed fraction of heads is within 0.01 of $\frac{1}{2}$ with probability at least 0.99? What does the CLT suggest instead?

Exercise 21.3 (Proof, $\star$)**.** Prove the "one-sided Chebyshev" warm-up: for any Y with mean 0 and variance σ^2 and any $a > 0$, $\mathbf{P}(Y \ge a) \le \frac{\sigma^2}{\sigma^2 + a^2}$. (Hint: for any $t \ge 0$, $\{Y \ge a\} \subseteq \{(Y+t)^2 \ge (a+t)^2\}$; apply Markov and minimize over t.)

Exercise 21.4 (Proof)**.** Show that convergence in probability to μ does *not* require $\mathrm{var}(M_n) \to 0$ in general: construct Y_n with $Y_n \to 0$ in probability but $\mathrm{var}(Y_n) \to \infty$. (Take $Y_n = n$ with probability $1/n^1$... adjust the exponent until it works, and say why.)

Deeper reading. B&T 5.1–5.4 (inequalities, WLLN, convergence in probability, CLT), 5.5 (strong law); B&T 4.4 (transforms) and Problem 5.12 are the proof, worked out in Lesson 43.

22 The Bernoulli and Poisson Processes

NEEDED FOR: Statistical inference (first random processes; independent increments). Not needed for ML.

22.1 Theory

A *random process* is a collection of random variables indexed by time. The two canonical arrival processes — coin flips in discrete time, and its continuous-time limit — are where statistical inference starts building intuition for the general theory.

Definition 22.1 (Bernoulli process)**.** A *Bernoulli process* with rate p is a sequence $X_1, X_2, \ldots$ of i.i.d. Bernoulli(p) random variables: independent trials, each a success ("arrival") with probability p.

Everything about it follows from Part III: the number of arrivals in n slots is Binomial(n, p); the time to the first arrival is Geometric(p); and the process is *memoryless* — whatever has happened through time n, the future $X_{n+1}, X_{n+2}, \ldots$ is a fresh Bernoulli process, by independence. Consequently the *interarrival times* (gaps between successive arrivals) are i.i.d. Geometric(p): after each arrival the process restarts.

Definition 22.2 (Poisson process)**.** A *Poisson process* with rate $\lambda > 0$ counts arrivals in continuous time: with $N(t)$ = number of arrivals in $(0, t]$, the process satisfies

(i) (*Poisson marginals*) $N(t) \sim \text{Poisson}(\lambda t)$: $\mathbf{P}(N(t) = k) = e^{-\lambda t}(\lambda t)^k/k!$;

(ii) (*independent, stationary increments*) arrival counts in disjoint intervals are independent, and the count in an interval depends only on its length.

Theorem 22.3 (The Poisson process is the limit of Bernoulli processes)**.** *Chop $(0, t]$ into n slots of width t/n and run a Bernoulli process with per-slot probability $p_n = \lambda t/n$. Then the number of successes $S_n \sim$ Binomial(n, p_n) satisfies, for each fixed k,*

$$\mathbf{P}(S_n = k) \xrightarrow[n\to\infty]{} e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

Proof. Write $a = \lambda t$. Then

$$\mathbf{P}(S_n = k) = \binom{n}{k}\left(\frac{a}{n}\right)^k\left(1 - \frac{a}{n}\right)^{n-k} = \frac{a^k}{k!} \cdot \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\to 1} \cdot \underbrace{\left(1 - \frac{a}{n}\right)^n}_{\to e^{-a}} \cdot \underbrace{\left(1 - \frac{a}{n}\right)^{-k}}_{\to 1},$$

using $(1 - a/n)^n \to e^{-a}$ and the fact that the first fraction is a product of k factors each tending to 1. □

Theorem 22.4 (Interarrival times are exponential)**.** *In a Poisson process of rate λ, the time T_1 to the first arrival is exponential(λ), and successive interarrival times are i.i.d. exponential(λ).*

Proof. $\mathbf{P}(T_1 > t) = \mathbf{P}(N(t) = 0) = e^{-\lambda t}$: exactly the exponential tail, so $T_1 \sim$ exponential(λ). For the later gaps: by independent and stationary increments, the process after any fixed time s is a fresh Poisson process independent of the past (this is the memorylessness of Lesson 17's exponential, now at the process level; the full argument that this also holds after the *random* time of an arrival is B&T 6.2), so each gap has the same exponential distribution, independently. □

Two standard consequences, both provable in a few lines from the definitions (Exercises): *merging* two independent Poisson processes of rates λ_1, λ_2 yields a Poisson process of rate $\lambda_1+\lambda_2$, and *splitting* a rate-λ process by an independent coin (p / $1-p$) yields two *independent* Poisson processes of rates $p\lambda$, $(1-p)\lambda$.

22.2 Practice: by hand

Example 22.5 (A packet link)**.** Packets arrive as a Poisson process at $\lambda = 3$ per second.

$$\mathbf{P}(\text{exactly 2 in one second}) = e^{-3}\frac{3^2}{2!} \approx 0.224, \qquad \mathbf{P}(\text{none in 2 seconds}) = e^{-6} \approx 0.0025.$$

Expected arrivals in a minute: $\lambda t = 180$; and by Lesson 21's CLT (a Poisson is a sum of many thin-slice Bernoullis), the count over a minute is approximately $\mathcal{N}(180, 180)$ — Poisson variance equals its mean.

Example 22.6 (Merging and the memoryless race)**.** Emails arrive at rate 2/hour, texts at rate 4/hour, independent Poisson. Interruptions of either kind: Poisson at rate 6/hour, so the wait for the next interruption is exponential(6) — mean 10 minutes. Given an interruption occurs, it is a text with probability 4/6: in a merged process, each arrival came from stream i with probability $\lambda_i / \sum_j \lambda_j$, independent of the past.

Statistical-inference connection. These are your first *random processes*, and they carry the vocabulary statistical inference runs on: independent increments, stationarity (increment distributions depend only on interval length — the first example of a stationarity property), memorylessness, and the interplay between the counting view $N(t)$ and the interarrival view $T_1, T_2, \ldots$. Shot noise in electronics is modeled directly as a Poisson process; the Wiener process (Brownian motion) that statistical inference introduces later is the same independent-stationary-increments idea with Gaussian increments in place of Poisson.

22.3 Exercises

Exercise 22.1 (Hand)**.** A Bernoulli process has $p = 0.1$ per slot. Compute: the probability of no arrival in 10 slots; the expected time to the third arrival (sum of three geometrics); the probability the first arrival is in slot 5 given none in the first 3 (memorylessness makes this instant).

Exercise 22.2 (Hand)**.** Buses (rate 4/hour) and taxis (rate 8/hour) pass a corner as independent Poisson processes. You wait for either. Find: the mean wait; the probability the first three vehicles are all taxis; the probability exactly 2 buses pass in 30 minutes.

Exercise 22.3 (Proof, $\star$)**.** Prove the merging theorem for marginals: if $N_1(t), N_2(t)$ are independent Poisson($\lambda_1 t$), Poisson($\lambda_2 t$), then $N_1(t) + N_2(t)$ is Poisson($(\lambda_1 + \lambda_2)t$). (Convolve the PMFs and recognize the binomial theorem inside the sum.)

Exercise 22.4 (Proof)**.** For a Poisson process, show $\operatorname{cov}\big(N(s), N(t)\big) = \lambda \min(s,t)$ for $s \le t$ by writing $N(t) = N(s) + \big(N(t) - N(s)\big)$ and using independent increments. (This "min" covariance reappears verbatim for the Wiener process in statistical inference.)

Deeper reading. B&T 6.1 (Bernoulli process), 6.2 (Poisson process — including the random-incidence paradox, a classic statistical-inference exam topic worth reading now).

23 Markov Chains

NEEDED FOR: Statistical inference (core: Markov processes); machine learning (helpful: the substrate of MDPs and HMMs). Uses Lesson 9.

23.1 Theory

Definition 23.1 (Discrete-time Markov chain)**.** A *Markov chain* on states $\{1, \dots, m\}$ is a sequence $X_0, X_1, \dots$ such that the next state depends on the past only through the present (the *Markov property*):

$$\mathbf{P}(X_{n+1} = j \mid X_n = i,\, X_{n-1}, \dots, X_0) = \mathbf{P}(X_{n+1} = j \mid X_n = i) = P_{ij},$$

where the *transition probabilities* P_{ij} form the $m \times m$ *transition matrix* P: each row is a PMF, $P_{ij} \ge 0$, $\sum_j P_{ij} = 1$.

Theorem 23.2 (Chapman–Kolmogorov: n-step transitions are matrix powers)**.** *Let $r_{ij}(n) = \mathbf{P}(X_n = j \mid X_0 = i)$. Then $r_{ij}(n) = (P^n)_{ij}$; and if the row vector $\pi^{(0)}$ is the PMF of X_0, the PMF of X_n is $\pi^{(0)} P^n$.*

Proof. Induct on n; the case $n = 1$ is the definition. For the step, condition on the state at time n (total probability, Theorem 14.9(b)) and use the Markov property:

$$r_{ij}(n+1) = \sum_{k=1}^{m} \mathbf{P}(X_n = k \mid X_0 = i)\, \mathbf{P}(X_{n+1} = j \mid X_n = k) = \sum_k r_{ik}(n) P_{kj},$$

which is exactly the (i,j) entry of $P^n P = P^{n+1}$. The distribution statement is the same computation with the initial PMF weighting row i. □

Probability question → Lesson 4 matrix multiplication; long-run behavior → Lesson 9 eigenvalues. That is the whole design.

Definition 23.3 (Stationary distribution)**.** A row vector π with $\pi_j \ge 0$, $\sum_j \pi_j = 1$ is a *stationary distribution* if

$$\pi P = \pi \qquad \text{(the } \textit{balance equations } \pi_j = \textstyle\sum_k \pi_k P_{kj}\text{, plus normalization).}$$

Started from π, the chain keeps distribution π forever (Theorem 23.2).

$\pi P = \pi$ says: π^{T} is an eigenvector of P^{T} with eigenvalue 1. Eigenvalue 1 always exists (P has row sums 1, so $P\mathbf{1} = \mathbf{1}$; P and P^{T} have the same eigenvalues by Lesson 13, $\det(P - \lambda I) = \det(P^{\mathsf{T}} - \lambda I)$). The substantive question is convergence:

Theorem 23.4 (Steady-state convergence — statement)**.** *Suppose the chain is* irreducible *(every state can reach every state) and* aperiodic *(no forced cycling; e.g. some state has $P_{ii} > 0$). Then there is a unique stationary π, with all $\pi_j > 0$, and for* every *starting state i:*

$$r_{ij}(n) \xrightarrow[n\to\infty]{} \pi_j.$$

Moreover π_j is the long-run fraction of time spent in state j.

B&T 7.3 proves this; the eigen-story explains it (for the diagonalizable case): the other eigenvalues of P have $|\lambda| < 1$ under these hypotheses, so in the eigen-expansion of $\pi^{(0)}P^n$ every component except the $\lambda = 1$ one dies geometrically — Lesson 9's "$|\lambda| < 1$ decays" as a theorem about forgetting the initial condition. The convergence *rate* is set by the second-largest $|\lambda|$ (the "spectral gap"), the number statistical inference and MCMC practitioners both watch.

23.2 Practice: by hand

Example 23.5 (A two-state weather chain, end to end)**.** Sunny (1) or rainy (2): $P = \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}$.

Balance: $\pi_1 = 0.9\pi_1 + 0.5\pi_2$ gives $0.1\pi_1 = 0.5\pi_2$, so $\pi = (\frac{5}{6}, \frac{1}{6})$: sunny five days in six, whatever today is. Two-step check from rain: $r_{21}(2) = (P^2)_{21} = 0.5\cdot 0.9 + 0.5\cdot 0.5 = 0.70$, already climbing toward $\frac{5}{6} \approx 0.83$. Eigenvalues: $\operatorname{tr} P = 1.4$, $\det P = 0.4$, so $\lambda = 1,\ 0.4$ (Lesson 13's trace/det shortcut) — errors shrink by a factor 0.4 per step, and after n steps the deviation from steady state is proportional to 0.4^n.

Example 23.6 (Absorption — a gambler)**.** States $0, 1, 2, 3$: you hold \$$k$, bet \$1 at fair odds each round, stop at \$0 (ruin) or \$3 (goal). States 0 and 3 are *absorbing* ($P_{00} = P_{33} = 1$). Let a_k = probability of reaching 3 from k. Conditioning on one step (total probability):

$$a_k = \tfrac{1}{2}a_{k-1} + \tfrac{1}{2}a_{k+1}, \qquad a_0 = 0,\ a_3 = 1,$$

so a is linear in k: $a_k = k/3$. Fair game, fair chances — proportional to your stake. (B&T 7.4 develops absorption probabilities and expected absorption times in general; the same first-step conditioning gives both.)

Machine-learning / statistical-inference connection. A Markov chain is the "environment half" of machine learning's MDPs: add actions that choose which row of transition probabilities applies, plus rewards, and value iteration is first-step conditioning (the gambler calculation) iterated to a fixed point. An HMM is a Markov chain observed through noise, and its filtering updates ride on Theorem 23.2. In statistical inference, Markov chains are the first structured random process — and the eigen-analysis of P (stationarity, spectral gap, reversibility) is a recurring exam theme. This is the one Part IV lesson worth doing *before* the machine-learning work if time permits.

23.3 Exercises

Exercise 23.1 (Hand)**.** A machine is up (1) or down (2): $P_{12} = 0.05$, $P_{21} = 0.6$. Find the stationary distribution, the long-run fraction of downtime, and both eigenvalues of P. Roughly how many steps until the initial condition is forgotten (deviation shrunk by $\sim 0.4^n$-style factor — compute the actual base)?

Exercise 23.2 (Hand)**.** For the gambler with unfair odds (win probability 0.4): re-derive the first-step equations and solve for a_1, a_2 with $a_0 = 0, a_3 = 1$. (The general solution is a geometric progression in $(0.6/0.4)^k$; or just solve the 2×2 linear system — Lesson 5.)

Exercise 23.3 (Proof, $\star$)**.** Prove that if π satisfies the *detailed balance* equations $\pi_i P_{ij} = \pi_j P_{ji}$ for all i, j, then π is stationary. (Sum over i.) Verify detailed balance holds for Example 23.5 and exhibit a 3-state chain where stationarity holds but detailed balance fails (hint: probability flowing around a cycle $1 \to 2 \to 3 \to 1$).

Exercise 23.4 (Proof)**.** Show that if P and Q are transition matrices, so is PQ; conclude P^n is a transition matrix for every n (rows of $r_{ij}(n)$ are PMFs, as they must be).

Deeper reading. B&T 7.1–7.3 (Markov chains, classification, steady state), 7.4 (absorption — the MDP-adjacent part), 7.5 (continuous time, inference-relevant); MDP treatments in machine learning re-derive first-step analysis as the Bellman equation.

Part V: The Signals-and-Systems Core

Both Fourier analysis and statistical inference stand on *linear systems and Fourier transforms*. This part is that course in minimum lessons, following Oppenheim & Willsky (O&W), *Signals and Systems* (2nd ed.), Chapters 1–7, 9–10, in the same style. It leans on Part II throughout: complex exponentials (Lesson 8), eigenvalues and the eigenfunction idea (Lesson 9), and the DFT and convolution theorem (Lesson 11) are the finite-dimensional shadows of everything below. *Nothing in this part is needed for machine learning.* With Part V done, the undergraduate prerequisite set for machine learning, Fourier analysis, and statistical inference is complete.

Following every signals text, Part V writes the imaginary unit as j.

24 Signals and Systems

NEEDED FOR: Fourier analysis and statistical inference (prerequisite, linear-systems level). Not needed for ML.

24.1 Theory

Definition 24.1 (Signals)**.** A *continuous-time (CT) signal* is a function $x(t)$, $t \in \mathbb{R}$; a *discrete-time (DT) signal* is a sequence $x[n]$, $n \in \mathbb{Z}$ — i.e. a vector with infinitely many coordinates, which is why Part I's language keeps applying. The *energy* of a signal is $E = \int_{-\infty}^{\infty} |x(t)|^2 dt$ (CT) or $\sum_n |x[n]|^2$ (DT) — a squared norm, Lesson 2 — and its *average power* is the per-unit-time limit of the same quantity.

Definition 24.2 (Transformations of time; periodicity)**.** From x we form $x(t - t_0)$ (*shift right* by t_0), $x(-t)$ (*reversal*), and $x(at)$ (*scaling*: compression for $|a| > 1$, stretch for $|a| < 1$). A signal is *periodic* with period $T > 0$ (resp. integer $N > 0$) if $x(t + T) = x(t)$ for all t (resp. $x[n + N] = x[n]$ for all n). Any signal splits uniquely into even plus odd parts: $x = \underbrace{\frac{1}{2}(x(t) + x(-t))}_{\text{even}} + \underbrace{\frac{1}{2}(x(t) - x(-t))}_{\text{odd}}$.

Proposition 24.3 (CT vs. DT complex exponentials — the crucial asymmetry)**.** *(a) $e^{j\omega_0 t}$ is periodic for every $\omega_0 \neq 0$, with fundamental period $2\pi/|\omega_0|$; distinct frequencies give distinct signals.*

(b) $e^{j\omega_0 n}$ satisfies $e^{j(\omega_0+2\pi)n} = e^{j\omega_0 n}$ for all n: DT frequencies are identical mod 2π, so only $\omega_0 \in [-\pi, \pi)$ are genuinely different, and $\omega_0 = \pm\pi$ is the fastest possible DT oscillation ($e^{j\pi n} = (-1)^n$). Moreover $e^{j\omega_0 n}$ is periodic in n iff $\omega_0/2\pi$ is rational.

Proof. (a) $e^{j\omega_0(t+T)} = e^{j\omega_0 t}e^{j\omega_0 T} = e^{j\omega_0 t}$ iff $\omega_0 T \in 2\pi\mathbb{Z}$; the smallest positive T is $2\pi/|\omega_0|$. (b) $e^{j2\pi n} = 1$ for every integer n (Lesson 8), giving the mod-2π identity. Periodicity: $e^{j\omega_0 N} = 1$ iff $\omega_0 N = 2\pi k$ iff $\omega_0/2\pi = k/N$ is rational. □

Part (b) is the deep reason sampling (Lesson 29) can destroy information: a DT sequence cannot tell frequencies apart mod 2π. It is also why the DTFT of Lesson 28 will be periodic.

Definition 24.4 (Unit impulse and unit step)**.** In DT: $\delta[n] = 1$ if $n = 0$, else 0; and $u[n] = 1$ for $n \geq 0$, else 0; note $u[n] = \sum_{k=0}^{\infty} \delta[n-k]$ and $\delta[n] = u[n] - u[n-1]$. In CT: $u(t) = 1$ for $t > 0$, else 0, and the *unit impulse* $\delta(t)$ is the idealized limit of a pulse of width Δ, height $1/\Delta$, as $\Delta \to 0$; it is defined *operationally* by the *sifting property*

$$\int_{-\infty}^{\infty} x(t)\,\delta(t-t_0)\,dt = x(t_0)$$

for every signal x continuous at t_0. (Making δ a genuine object is distribution theory; O&W 1.4 and 2.5 give the honest operational account, and Fourier analysis develops the rigorous one — every formal δ-manipulation below can be certified by running it against a test function.)

Definition 24.5 (System properties)**.** A *system* maps input signals to output signals, $y = T\{x\}$. It is

- *memoryless* if y at time t depends only on x at time t;
- *causal* if y at time t depends only on x at times $\leq t$;
- *(BIBO) stable* if every bounded input produces a bounded output;
- *time-invariant* if shifting the input shifts the output: $x(t-t_0) \mapsto y(t-t_0)$ for all t_0;
- *linear* if $T\{ax_1 + bx_2\} = aT\{x_1\} + bT\{x_2\}$ — Lesson 4's definition, verbatim, on the vector space of signals.

Proposition 24.6. *A linear system maps the zero signal to the zero signal.*

Proof. $T\{0\} = T\{0 \cdot x\} = 0 \cdot T\{x\} = 0$ — the same one-liner as Proposition 1.4 and Lesson 4, because it is the same fact. □

24.2 Practice: by hand

Example 24.7 (Classifying systems)**.**

system	memoryless	causal	stable	TI	linear
$y(t) = 2x(t) + 3$	✓	✓	✓	✓	×
$y[n] = x[n] - x[n-1]$	×	✓	✓	✓	✓
$y(t) = x(2t)$	×	×	✓	×	✓
$y[n] = n\,x[n]$	✓	✓	×	×	✓
$y(t) = x^2(t)$	✓	✓	✓	✓	×

Spot checks worth doing: $y(t) = 2x(t) + 3$ fails linearity because zero in gives $3 \neq 0$ out. $y(t) = x(2t)$ is not causal ($y(-1) = x(-2)$ is fine, but $y(1) = x(2)$ looks ahead) and not time-invariant: shifting the input by t_0 produces $x(2t - t_0)$, but shifting the output produces $x(2(t-t_0)) = x(2t - 2t_0)$ — different. $y[n] = nx[n]$ is unstable: the bounded input $x[n] = 1$ gives the unbounded output n.

Example 24.8 (Even/odd decomposition)**.** $x[n] = u[n]$: even part $\frac{1}{2}(u[n] + u[-n]) = \frac{1}{2} + \frac{1}{2}\delta[n]$, odd part $\frac{1}{2}(u[n] - u[-n]) = \frac{1}{2}\,\mathrm{sign}[n]$. Check they sum to $u[n]$ at $n = -1, 0, 1$: $0, 1, 1$. ✓

Fourier / statistical-inference connection. "Linear plus time-invariant" is the exact class of systems Fourier methods diagonalize — Lesson 9's LTI-connection box, about to become Lessons 25–28. The properties table is not busywork: statistical inference constantly asks which manipulations of a random process preserve stationarity, and the answer is "the time-invariant ones"; Fourier analysis's transform rules each mirror one row of it.

24.3 Exercises

Exercise 24.1 (Hand)**.** Sketch (or tabulate) $x(t) = u(t+1) - u(t-2)$, then $x(2t)$, $x(t/2)$, and $x(1-t)$. For the last one: which operations, in which order?

Exercise 24.2 (Hand)**.** Classify (all five properties, with one-line justifications): (a) $y[n] = x[-n]$; (b) $y(t) = \int_{-\infty}^{t} x(\tau)\, d\tau$; (c) $y[n] = \cos(x[n])$.

Exercise 24.3 (Proof, $\star$)**.** Prove that the even/odd decomposition is unique: if $x = e_1 + o_1 = e_2 + o_2$ with e_i even and o_i odd, then $e_1 = e_2$ and $o_1 = o_2$. Then show energy splits: $\sum_n |x[n]|^2 = \sum_n |e[n]|^2 + \sum_n |o[n]|^2$ (the cross terms form an odd signal — even and odd signals are *orthogonal*, Lesson 2's Pythagoras again).

Exercise 24.4 (Proof)**.** Determine the fundamental period of $x[n] = e^{j(2\pi/3)n} + e^{j(3\pi/4)n}$ (least common period of the two parts), and prove that a sum of two periodic DT signals is always periodic — then explain why the analogous CT claim is *false* ($\cos t + \cos \pi t$).

Deeper reading. O&W 1.1–1.4 (signals, transformations, exponentials, impulses), 1.5–1.6 (systems and properties).

25 LTI Systems and Convolution

NEEDED FOR: Fourier analysis and statistical inference (prerequisite). Not needed for ML.

25.1 Theory

The sifting property writes any DT signal as a combination of shifted impulses:

$$x[n] = \sum_{k=-\infty}^{\infty} x[k]\, \delta[n-k].$$

The shifted impulses are a "basis" for signal space, and the coefficients are the signal's own values. Lesson 4 taught what happens next: a linear map is determined by what it does to a basis.

Theorem 25.1 (LTI systems are convolutions)**.** *Let T be a linear time-invariant DT system and let $h = T\{\delta\}$ be its* impulse response*. Then for every input,*

$$y[n] = (x * h)[n] = \sum_{k=-\infty}^{\infty} x[k]\, h[n-k].$$

Conversely, every such convolution is LTI. The same holds in CT with $y(t) = (x * h)(t) = \int_{-\infty}^{\infty} x(\tau)\, h(t-\tau)\, d\tau$ and $h = T\{\delta\}$.

Proof. (DT; the CT statement is its limit, O&W 2.2.) By time invariance, $T\{\delta[\cdot - k]\} = h[\cdot - k]$. By linearity applied to the sifting representation (for inputs of finite duration this is a finite sum; the infinite case needs a continuity assumption that all systems in practice satisfy):

$$y[n] = T\Big\{ \sum_k x[k]\, \delta[\cdot - k] \Big\}[n] = \sum_k x[k]\, h[n-k].$$

Conversely, the formula is linear in x (sums split), and replacing x by $x[\cdot - n_0]$ and substituting $k' = k - n_0$ shifts y by n_0: time-invariant. □

This is Theorem 4.3 for signal space: *one* signal, h, pins down the entire system, exactly as a matrix's columns pin down a linear map. Indeed, writing $y[n] = \sum_k H_{n,k} x[k]$ with $H_{n,k} = h[n-k]$: an LTI system is an (infinite) matrix that is constant along diagonals (*Toeplitz*) — and Lesson 11's circulant matrices were the finite, wrap-around version of exactly this.

Proposition 25.2 (Algebra of convolution)**.** *$x*h = h*x$ (commutative); $x*(h_1*h_2) = (x*h_1)*h_2$ (associative — cascaded LTI systems combine into one, in either order); $x*(h_1+h_2) = x*h_1 + x*h_2$ (distributive — parallel systems add); $x*\delta = x$ (identity).*

Proof. Commutativity: substitute $m = n-k$ in the defining sum, $\sum_k x[k]h[n-k] = \sum_m x[n-m]h[m]$. Identity: sifting. Distributivity: split the sum. Associativity: write out the double sum for both sides and reindex — both equal $\sum_k \sum_m x[k]\, h_1[m]\, h_2[n-k-m]$. □

Theorem 25.3 (Causality and stability, read off h)**.** *An LTI system is causal iff $h[n] = 0$ for all $n < 0$; it is BIBO stable iff h is absolutely summable, $\sum_k |h[k]| < \infty$ (CT: $h(t) = 0$ for $t < 0$; $\int |h| < \infty$).*

Proof. Causality. Write $y[n] = \sum_k h[k]x[n-k]$ (commuted form). If $h[k] = 0$ for $k < 0$, the sum involves only $x[n], x[n-1], \ldots$: causal. If $h[k_0] \neq 0$ for some $k_0 < 0$, the input δ produces output h, which is nonzero at time $k_0 < 0$ although the input was zero before time 0: not causal.

Stability, sufficiency. If $|x[n]| \le B$ for all n, the triangle inequality gives

$$|y[n]| \le \sum_k |h[k]|\, |x[n-k]| \le B \sum_k |h[k]| < \infty.$$

Necessity. Suppose $\sum_k |h[k]| = \infty$. Feed in the bounded input $x[n] = \operatorname{sign}\big(h[-n]\big)$ (magnitude ≤ 1). Then $y[0] = \sum_k h[k]x[-k] = \sum_k h[k] \operatorname{sign}(h[k]) = \sum_k |h[k]| = \infty$: a bounded input with an unbounded output. □

Difference and differential equations. Most implemented systems are described by linear constant-coefficient equations. The simplest, run as a recursion, already shows the shape of everything: $y[n] - ay[n-1] = x[n]$ with initial rest gives, feeding in δ: $h[0] = 1$, $h[1] = a$, $h[2] = a^2, \ldots$, i.e.

$$h[n] = a^n u[n],$$

which is causal always and stable iff $|a| < 1$ (geometric series) — Lesson 9's "$|\lambda| < 1$ decays" as a system-theory fact. Lesson 30 gives the systematic transform treatment.

25.2 Practice: by hand

Example 25.4 (A convolution table)**.** $x = (1, 2, 1)$ at $n = 0, 1, 2$ and $h = (1, 1)$ at $n = 0, 1$. Slide-and-sum:

$$y[0] = 1, \quad y[1] = 2 + 1 = 3, \quad y[2] = 1 + 2 = 3, \quad y[3] = 1; \qquad y = (1, 3, 3, 1).$$

Sanity checks: length $3+2-1 = 4$; $\sum y = (\sum x)(\sum h) = 4 \cdot 2$ ✓ (that identity is the convolution property at frequency zero, Lesson 27). Commuted order gives the same table transposed. ✓

Example 25.5 (Pulse $*$ pulse = triangle)**.** Let $x(t) = h(t) = 1$ on $[0, 1]$, else 0. Then $y(t) = \int x(\tau)h(t-\tau)d\tau$ is the overlap length of $[0, 1]$ and $[t-1, t]$: $y(t) = t$ on $[0, 1]$, $2 - t$ on $[1, 2]$, 0 elsewhere — a triangle. Convolution smooths: the output is continuous although both inputs jump. (Convolve again with the pulse and the result is piecewise-parabolic; each convolution adds a degree of smoothness — the picture behind Fourier analysis's "smoother signal ⇔ faster-decaying transform.")

Example 25.6 (First-order smoother)**.** $y[n] = 0.5\,y[n-1] + x[n]$, input $x = u$: $y[0] = 1$, $y[1] = 1.5$, $y[2] = 1.75$, $y[3] = 1.875$, climbing to $\sum_k (0.5)^k = 2$. The step response is the running convolution $u * h$, and its limit is $\sum h[k]$ — the "DC gain."

Fourier / statistical-inference connection. Convolution is Fourier analysis's second-most-important operation (after the transform itself), and the pair "convolution in time ↔ multiplication in frequency" (Lessons 27–28, finite version already proved as Theorem 11.6) is Fourier analysis's engine. In statistical inference, filtered random processes are convolutions $y = h * x$ with x random; the autocorrelation of the output involves h convolved with itself, and stability (Theorem 25.3) is what keeps output variances finite.

25.3 Exercises

Exercise 25.1 (Hand)**.** Compute $(1, 1, 1) * (1, -1)$ and $(1, 2) * (1, 2)$ by table. Check the lengths and the sum-product identity of Example 25.4.

Exercise 25.2 (Hand)**.** Compute $u * h$ for $h[n] = (0.5)^n u[n]$ in closed form (finite geometric series), and identify the DC gain. Then find the impulse response of the *cascade* of that system with $y[n] = x[n] - x[n-1]$ (a differencer), using associativity — what does the cascade do to a constant input, and why is that obvious from h?

Exercise 25.3 (Proof, $\star$)**.** Prove that the cascade of two causal LTI systems is causal and of two stable LTI systems is stable, directly from Theorem 25.3 and the convolution formula ($h = h_1 * h_2$; bound $\sum |h|$ by $\big(\sum |h_1|\big)\big(\sum |h_2|\big)$).

Exercise 25.4 (Proof)**.** Show that an LTI system is memoryless iff $h[n] = c\,\delta[n]$ for some constant c — so *every* nontrivial LTI system has memory, and "matrix diagonal" is the right mental image.

Deeper reading. O&W 2.1–2.3 (convolution sum and integral, properties), 2.4 (difference and differential equations), 2.5 (singularity functions).

26 Fourier Series

NEEDED FOR: Fourier analysis (its opening act) and statistical inference (prerequisite). Not needed for ML.

26.1 Theory

Proposition 26.1 (Complex exponentials are eigenfunctions of LTI systems)**.** *Let h be the impulse response of a CT LTI system, and let $s \in \mathbb{C}$ be such that $H(s) = \int_{-\infty}^{\infty} h(\tau)e^{-s\tau}d\tau$ converges. Then the input e^{st} produces the output $H(s)\,e^{st}$. In DT, z^n produces $H(z)\,z^n$ with $H(z) = \sum_k h[k]z^{-k}$.*

Proof.

$$y(t) = \int h(\tau)\, e^{s(t-\tau)}d\tau = e^{st}\int h(\tau)e^{-s\tau}d\tau = H(s)\, e^{st}. \qquad \square$$

Eigenvector in, eigenvalue out (Lesson 9) — so if we can write signals as combinations of exponentials, LTI systems act diagonally. Fourier series do exactly that for periodic signals.

Lemma 26.2 (Orthonormality of harmonics)**.** *Fix $T > 0$ and $\omega_0 = 2\pi/T$. Then*

$$\frac{1}{T}\int_T e^{jk\omega_0 t}\,\overline{e^{jl\omega_0 t}}\,dt = \frac{1}{T}\int_T e^{j(k-l)\omega_0 t}\,dt = \begin{cases} 1 & k = l, \\ 0 & k \neq l, \end{cases}$$

the integral over any interval of length T.

Proof. For $k = l$ the integrand is 1. For $m = k - l \neq 0$: $\int_0^T e^{jm\omega_0 t}dt = \frac{e^{jm\omega_0 T}-1}{jm\omega_0} = \frac{e^{j2\pi m}-1}{jm\omega_0} = 0$. $\square$

This is Lemma 8.2 with the sum over roots of unity replaced by an integral over the circle: the harmonics $\{e^{jk\omega_0 t}\}$ are an orthonormal list in the inner product $\langle f,\, g\rangle = \frac{1}{T}\int_T f\bar{g}$.

Theorem 26.3 (Fourier series)**.** *Let x be periodic with period T and representable as $x(t) = \sum_{k=-\infty}^{\infty} a_k e^{jk\omega_0 t}$ (see the convergence remark below). Then the coefficients are forced:*

$$a_k = \frac{1}{T}\int_T x(t)\, e^{-jk\omega_0 t}\,dt,$$

and Parseval's relation holds: $\dfrac{1}{T}\displaystyle\int_T |x(t)|^2 dt = \sum_k |a_k|^2.$

Proof. Multiply the expansion by $\overline{e^{jl\omega_0 t}} = e^{-jl\omega_0 t}$, integrate over one period, and swap sum and integral (legal under the convergence assumptions):

$$\frac{1}{T}\int_T x(t)e^{-jl\omega_0 t}dt = \sum_k a_k \cdot \frac{1}{T}\int_T e^{j(k-l)\omega_0 t}dt = a_l$$

by Lemma 26.2 — exactly Proposition 6.2's "coordinates in an orthonormal basis are inner products," $a_l = \langle x,\, e_l\rangle$. Parseval is the Pythagorean theorem in the same inner product: expand $\langle x,\, x\rangle = \sum_k \sum_l a_k \bar{a}_l \langle e_k,\, e_l\rangle = \sum_k |a_k|^2$. $\square$

Convergence, honestly. For x with $\int_T |x|^2 < \infty$, the partial sums converge to x in energy (they are orthogonal projections onto $\text{span}\{e^{jk\omega_0 t} : |k| \le N\}$ — Theorem 6.3 verbatim — hence the best least-squares approximants at every N). Pointwise convergence needs the Dirichlet conditions (O&W 3.4); at a jump the series converges to the midpoint, and near a jump the partial sums overshoot by $\approx 9\%$ no matter how many terms are kept — the *Gibbs phenomenon*.

Two properties used constantly (both one-line from the coefficient formula): a time shift multiplies coefficients by a phase, $x(t - t_0) \leftrightarrow a_k e^{-jk\omega_0 t_0}$; and for real x, $a_{-k} = \overline{a_k}$ (conjugate symmetry — take conjugates inside the integral).

Theorem 26.4 (Filtering a periodic signal)**.** *If $x = \sum_k a_k e^{jk\omega_0 t}$ drives an LTI system with frequency response $H(j\omega) = \int h(\tau) e^{-j\omega\tau} d\tau$, the output is*

$$y(t) = \sum_k a_k \, H(jk\omega_0) \, e^{jk\omega_0 t} :$$

each harmonic passes through independently, scaled by the eigenvalue at its frequency.

Proof. Linearity plus Proposition 26.1 at $s = jk\omega_0$, term by term. □

DT Fourier series = the DFT. A DT signal with period N has only N distinct harmonics ($e^{j(2\pi/N)kn}$, $k = 0, \dots, N-1$, by Proposition 24.3b), so the DT Fourier series is a finite sum — and its analysis/synthesis pair is exactly Lesson 11's DFT and inverse DFT (up to the $1/N$ placement). Everything proved there (unitarity, Parseval, the convolution theorem for circulants) *is* the DT Fourier series theory.

26.2 Practice: by hand

Example 26.5 (The square wave, the one to memorize)**.** Let x have period T, with $x(t) = 1$ for $|t| < T_1$ and 0 elsewhere in the period. Then $a_0 = 2T_1/T$, and for $k \neq 0$:

$$a_k = \frac{1}{T} \int_{-T_1}^{T_1} e^{-jk\omega_0 t} dt = \frac{e^{jk\omega_0 T_1} - e^{-jk\omega_0 T_1}}{jk\omega_0 T} = \frac{2\sin(k\omega_0 T_1)}{k\omega_0 T} = \frac{\sin(k\omega_0 T_1)}{k\pi}.$$

For the half-duty case $T = 4T_1$ ($\omega_0 T_1 = \pi/2$): $a_k = \frac{\sin(k\pi/2)}{k\pi}$ — $\frac{1}{2}$, $\frac{1}{\pi}$, 0, $-\frac{1}{3\pi}$, 0, $\frac{1}{5\pi}$, … for $k = 0, 1, 2, 3, 4, 5$. The coefficients decay like $1/k$: slow, because the signal jumps. (Smooth signals decay faster — differentiate the shift-and-phase property to see each derivative buy one more power of $1/k$.)

Example 26.6 (Filtering the square wave)**.** Send the $T = 4T_1$ square wave through an ideal lowpass filter keeping only $|k| \le 1$: the output is $\frac{1}{2} + \frac{2}{\pi}\cos(\omega_0 t)$ — a DC level plus one smooth cosine; all edges gone. Keep $|k| \le 5$ and the corners begin to return, with Gibbs ears at the jumps. Filtering is coefficient surgery: Theorem 26.4 with $H = 1$ in the passband and 0 outside.

Fourier connection. A Fourier-analysis course opens with precisely this lesson — Osgood's first chapters are Fourier series, coefficients-as-inner-products, and Parseval — before taking $T \to \infty$ to reach the Fourier transform (as Lesson 27 does next). Arriving with Lemma 26.2 and Theorem 26.3 already owned turns those weeks into review, exactly as the Lesson 8 box promised.

26.3 Exercises

Exercise 26.1 (Hand)**.** Compute the Fourier series coefficients of $x(t) = \sin^2(\omega_0 t)$ directly (no integrals: rewrite with the double-angle identity and read off the a_k). Which k are nonzero?

Exercise 26.2 (Hand)**.** From Example 26.5 with $T = 4T_1$, evaluate the series at $t = 0$ to prove $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots = \frac{\pi}{4}$ (Leibniz), and apply Parseval to prove $\sum_{k \text{ odd}} \frac{1}{k^2} = \frac{\pi^2}{8}$.

Exercise 26.3 (Proof, $\star$)**.** Prove the differentiation property: if x has coefficients a_k and x' is also nice-enough periodic, then x' has coefficients $jk\omega_0 a_k$ (integrate by parts over one period; the boundary terms cancel by periodicity). Conclude: m derivatives of smoothness force $|a_k| = O(1/k^m)$ decay.

Exercise 26.4 (Proof)**.** Prove the DT orthogonality relation $\frac{1}{N}\sum_{n=0}^{N-1} e^{j(2\pi/N)(k-l)n} = \delta_{kl}$ (for $k, l \in \{0, \ldots, N-1\}$) from Lemma 8.2, and write out the resulting DT Fourier series pair. Compare with Lesson 11's DFT to fix where the $1/N$ sits in each convention.

Deeper reading. O&W 3.2 (eigenfunctions), 3.3–3.5 (CT Fourier series, convergence, properties), 3.6–3.7 (DT Fourier series), 3.8–3.9 (LTI systems and filtering).

27 The Continuous-Time Fourier Transform

NEEDED FOR: Fourier analysis (its core object) and statistical inference (prerequisite). Not needed for ML.

27.1 Theory

Aperiodic signals have no harmonics to expand in — but an aperiodic signal is a periodic one whose period went to infinity. Carrying Theorem 26.3 through that limit (O&W 4.1 does it carefully: $Ta_k \to X(j\omega)$ at $\omega = k\omega_0$, and the Fourier sum becomes a Riemann integral) yields the transform pair that Fourier analysis lives on:

Definition 27.1 (CT Fourier transform)**.**

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t}\, dt \qquad \Longleftrightarrow \qquad x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(j\omega)\, e^{j\omega t}\, d\omega.$$

$X(j\omega)$ is the signal's density of content at frequency ω — the continuous analogue of "coefficient = inner product with a harmonic."

Example 27.2 (The pairs to know cold)**.** Each is a direct integral:

$$\begin{aligned}
e^{-at}u(t)\ (a>0) &\longleftrightarrow \frac{1}{a+j\omega} && \text{(one-sided decay)}\\
x(t)=1,\ |t|<T_1 &\longleftrightarrow \frac{2\sin(\omega T_1)}{\omega} && \text{(rectangle} \leftrightarrow \text{sinc)}\\
\delta(t) &\longleftrightarrow 1 && \text{(all frequencies equally)}\\
1 \longleftrightarrow 2\pi\,\delta(\omega), \qquad e^{j\omega_0 t} &\longleftrightarrow 2\pi\,\delta(\omega-\omega_0) && \text{(tones are frequency impulses)}\\
\cos\omega_0 t &\longleftrightarrow \pi\big[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)\big].
\end{aligned}$$

The first: $\int_0^\infty e^{-(a+j\omega)t}dt = \frac{1}{a+j\omega}$. The second: $\int_{-T_1}^{T_1} e^{-j\omega t}dt = \frac{2\sin(\omega T_1)}{\omega}$ — the same computation as the square wave's a_k (Example 26.5), now with a continuous frequency variable. The δ pairs are certified by the sifting property inside the inversion integral. Also worth knowing: a Gaussian transforms to a Gaussian ($e^{-t^2/2} \leftrightarrow \sqrt{2\pi}\, e^{-\omega^2/2}$ — Exercise 17.3's integral, completed in the exponent), the unique fixed-point shape.

Theorem 27.3 (The property toolbox)**.** *Let $x \leftrightarrow X$ and $y \leftrightarrow Y$. Then:*

(a) Linearity: $ax + by \leftrightarrow aX + bY$.

(b) Time shift: $x(t - t_0) \leftrightarrow e^{-j\omega t_0}X(j\omega)$ — *magnitude untouched, phase tilted.*

(c) Scaling: $x(at) \leftrightarrow \frac{1}{|a|}X\big(j\frac{\omega}{a}\big)$ *for $a \neq 0$ — compress time, stretch frequency, and vice versa.*

(d) Duality: *if $x(t) \leftrightarrow X(j\omega)$ then $X(jt) \leftrightarrow 2\pi\, x(-\omega)$.*

(e) Differentiation: $x'(t) \leftrightarrow j\omega\, X(j\omega)$.

(f) Parseval: $\displaystyle\int |x(t)|^2 dt = \frac{1}{2\pi}\int |X(j\omega)|^2 d\omega$.

Proof. (a) integrals are linear. (b) substitute $\tau = t - t_0$: $\int x(t-t_0)e^{-j\omega t}dt = e^{-j\omega t_0}\int x(\tau)e^{-j\omega\tau}d\tau$. (c) substitute $\tau = at$; for $a > 0$, $\int x(at)e^{-j\omega t}dt = \frac{1}{a}\int x(\tau)e^{-j(\omega/a)\tau}d\tau$; for $a < 0$ the limits flip, producing the $|a|$ — this is Lesson 13's stretch theorem in one dimension, with $1/|a|$ as the volume factor and ω/a as the contragradient frequency variable. (d) the inversion formula *is* a forward transform up to sign and 2π; write it at $-\omega$ and rename variables. (e) differentiate the inversion formula under the integral: each $e^{j\omega t}$ contributes a factor $j\omega$. (f) is Theorem 11.2(c) in integral form: the transform is (up to $\sqrt{2\pi}$) a unitary map, and lengths are preserved; the direct proof expands $|x|^2 = x\bar{x}$ via the inversion formula and swaps integrals. □

Theorem 27.4 (The convolution property — the punchline)**.**

$$y = h * x \qquad \Longleftrightarrow \qquad Y(j\omega) = H(j\omega)\, X(j\omega),$$

and dually, multiplication in time is $\frac{1}{2\pi}\times$ convolution in frequency: $x(t)p(t) \leftrightarrow \frac{1}{2\pi}(X * P)(j\omega)$.

Proof. Transform the convolution and swap the integrals:

$$Y(j\omega) = \int\Big(\int x(\tau)h(t-\tau)d\tau\Big)e^{-j\omega t}dt = \int x(\tau)e^{-j\omega\tau}\Big(\int h(t-\tau)e^{-j\omega(t-\tau)}dt\Big)d\tau = X(j\omega)H(j\omega),$$

using the shift substitution inside the inner integral. The dual statement follows by duality (or the mirror-image computation). □

This is Theorem 11.6 — "circulants are diagonalized by the DFT" — with integrals in place of sums: LTI systems are diagonal in the Fourier basis, with $H(j\omega)$ (the transform of the impulse response) as the eigenvalue at frequency ω. Solving systems, designing filters, and inverting channels all reduce to multiplication and division of transforms.

27.2 Practice: by hand

Example 27.5 (Two-sided decay)**.** $x(t) = e^{-a|t|}$, $a > 0$: split at 0 and use the first pair twice (once reversed),

$$X(j\omega) = \frac{1}{a + j\omega} + \frac{1}{a - j\omega} = \frac{2a}{a^2 + \omega^2},$$

real and even, as it must be for a real even signal. As $a \to 0$ the time signal flattens toward 1 and the transform sharpens toward $2\pi\delta(\omega)$ — scaling property intuition made visible.

Example 27.6 (RC lowpass filter)**.** The circuit equation $RC\,y'(t) + y(t) = x(t)$ has impulse response $h(t) = \frac{1}{RC}e^{-t/RC}u(t)$. Transforming with properties (e) and (a): $(1 + j\omega RC)Y = X$, so

$$H(j\omega) = \frac{1}{1 + j\omega RC}, \qquad |H(j\omega)| = \frac{1}{\sqrt{1 + (\omega RC)^2}}:$$

gain ≈ 1 for $\omega \ll 1/RC$, rolling off like $1/\omega$ above the *cutoff* $\omega_c = 1/RC$ (where $|H| = 1/\sqrt{2}$, the "-3 dB point"). One first-order system, read three ways: differential equation, impulse response, frequency response.

Fourier / statistical-inference connection. Lessons 26–27 are the first half of Fourier analysis in miniature: series, transform, properties, convolution. What Fourier analysis adds is depth (distributions to make $\delta(\omega)$ rigorous, the sampling and DFT chapters — Lessons 29 and 11 here — and applications from diffraction to CT scans). Statistical inference's stake: the power spectral density of a stationary process is the Fourier transform of its autocorrelation (Lesson 18), and filtered noise obeys $S_y(\omega) = |H(j\omega)|^2 S_x(\omega)$ — Theorem 27.4 applied to second moments. That formula is the reason statistical inference lists this material as a prerequisite.

27.3 Exercises

Exercise 27.1 (Hand)**.** Using pairs and properties only (no new integrals): find the transforms of (a) $e^{-3(t-2)}u(t-2)$; (b) $te^{-at}u(t)$ (differentiate $\frac{1}{a+j\omega}$ with respect to ω — which property is that?); (c) $\sin\omega_0 t$.

Exercise 27.2 (Hand)**.** For the RC filter of Example 27.6 with $RC = 10^{-3}$: what are the cutoff in Hz, $|H|$ at 50 Hz and at 10 kHz, and the output when $x(t) = \cos(2\pi \cdot 100\,t) + \cos(2\pi \cdot 10^4\,t)$? (Use Theorem 26.4: evaluate H at each tone.)

Exercise 27.3 (Proof, $\star$)**.** Prove the modulation property $x(t)\cos(\omega_0 t) \leftrightarrow \frac{1}{2}X(j(\omega - \omega_0)) + \frac{1}{2}X(j(\omega + \omega_0))$ from linearity and the $e^{j\omega_0 t}$ shift-in-frequency property (prove that one first, by the mirror of Theorem 27.3b). This two-line result is AM radio (O&W Ch. 8), and Lesson 29 reuses it as the sampling theorem's engine.

Exercise 27.4 (Proof)**.** Derive the transform of the triangle $\big(1 - |t|/T\big)$ on $|t| < T$ by writing it as $\frac{1}{T}(\text{rect} * \text{rect})$ and applying Theorem 27.4: the answer is a sinc^2. Confirm positivity of the transform and explain it (autocorrelation of a real signal — Lesson 18's Cauchy–Schwarz guarantees the peak is at 0).

Deeper reading. O&W 4.1–4.3 (the transform and its properties), 4.4–4.5 (convolution and multiplication properties), 4.7 (LCCDE systems); Osgood, *Lectures on the Fourier Transform and Its Applications*, chapters 2–4 tell the same story with more depth and better jokes.

28 The Discrete-Time Fourier Transform and Frequency Response

NEEDED FOR: Fourier analysis (DFT background) and statistical inference (prerequisite; DSP fluency). Not needed for ML.

28.1 Theory

Definition 28.1 (DTFT)**.** For a DT signal $x[n]$,

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]\, e^{-j\omega n} \qquad \Longleftrightarrow \qquad x[n] = \frac{1}{2\pi}\int_{2\pi} X(e^{j\omega})\, e^{j\omega n}\, d\omega.$$

$X(e^{j\omega})$ is 2π-periodic — immediately from Proposition 24.3(b) — which is why one period ($\omega \in [-\pi, \pi]$) carries everything and the inversion integral runs over one period only. The notation $X(e^{j\omega})$ advertises the periodicity: the transform is a function on the unit circle, foreshadowing Lesson 30's z-plane.

Example 28.2 (The two DT pairs to know cold)**.** *Geometric decay:* for $|a| < 1$,

$$a^n u[n] \;\longleftrightarrow\; \sum_{n\geq 0} (ae^{-j\omega})^n = \frac{1}{1-ae^{-j\omega}},$$

a geometric series, convergent because $|ae^{-j\omega}| = |a| < 1$.

Rectangular window: $x[n] = 1$ for $|n| \leq N_1$:

$$X(e^{j\omega}) = \sum_{n=-N_1}^{N_1} e^{-j\omega n} = \frac{\sin\left(\omega(N_1 + \frac{1}{2})\right)}{\sin(\omega/2)},$$

by the same finite geometric sum as Lemma 8.2 — the *periodic sinc* (Dirichlet kernel), the DT sibling of rect $\leftrightarrow$ sinc.

All of Theorem 27.3 transfers (shift $\to$ phase, Parseval with $\frac{1}{2\pi}\int_{2\pi}$, etc.), and so does the punchline, by the same swap-the-sums proof as Theorem 27.4:

Theorem 28.3 (DT convolution property)**.** *$y = x * h \iff Y(e^{j\omega}) = X(e^{j\omega})H(e^{j\omega})$, where $H(e^{j\omega}) = \sum_k h[k]e^{-j\omega k}$ is the system's* frequency response *— the eigenvalue function of Proposition 26.1 evaluated on the unit circle $z = e^{j\omega}$.*

Frequency response from a difference equation. For $\sum_k a_k y[n-k] = \sum_k b_k x[n-k]$, transform both sides with the shift property and divide:

$$H(e^{j\omega}) = \frac{\sum_k b_k e^{-j\omega k}}{\sum_k a_k e^{-j\omega k}}$$

— a ratio of polynomials in $e^{-j\omega}$, computable by inspection. The first-order smoother $y[n] - ay[n-1] = x[n]$ gives $H = 1/(1 - ae^{-j\omega})$, consistent with $h[n] = a^n u[n]$ and Example 28.2: for $0 < a < 1$, $|H|$ is largest at $\omega = 0$ and smallest at $\omega = \pi$ — a lowpass; flip the sign of a and the roles swap — a highpass.

The DFT is the sampled DTFT. For a length-N signal, Lesson 11's DFT values are exactly

$$\hat{x}_k = X(e^{j\omega})\Big|_{\omega=2\pi k/N} :$$

N equally spaced samples of the DTFT around the circle. That is why `np.fft.fft` computes "the spectrum": it evaluates this lesson's object on a grid, in $O(N \log N)$.

Why ideal filters are ideals. The perfect lowpass $H(e^{j\omega}) = 1$ for $|\omega| < \omega_c$, else 0, inverts to $h[n] = \frac{\sin(\omega_c n)}{\pi n}$ — nonzero for all $n < 0$ (noncausal) and decaying only like $1/|n|$, so $\sum |h[n]| = \infty$ (unstable by Theorem 25.3). Realizable filters can only approximate the ideal, trading passband flatness, rolloff sharpness, and delay — the design space O&W Chapter 6 maps and DSP courses inhabit.

28.2 Practice: by hand

Example 28.4 (Reading a two-point filter)**.** $y[n] = \frac{1}{2}(x[n] + x[n-1])$: $H(e^{j\omega}) = \frac{1}{2}(1 + e^{-j\omega}) = e^{-j\omega/2}\cos(\omega/2)$. So $|H(1)| = 1$ at DC and $H = 0$ at $\omega = \pi$: the two-point averager passes constants and annihilates $(-1)^n$ — Example 11.8's observation ("averaging kills the Nyquist alternation"), re-derived in one line from the frequency response. The differencer $y[n] = x[n] - x[n-1]$ has $H = 1 - e^{-j\omega} = 2je^{-j\omega/2}\sin(\omega/2)$: zero at DC, maximal at π — the complementary highpass.

Example 28.5 (First-order filter numbers)**.** $a = 0.9$: $|H(e^{j0})| = \frac{1}{1-0.9} = 10$ and $|H(e^{j\pi})| = \frac{1}{1+0.9} \approx 0.53$ — a 19:1 preference for slow signals. The half-power frequency solves $|1 - 0.9e^{-j\omega}|^2 = 2(0.1)^2$, giving $\omega_c \approx 0.105$: narrow, matching the long memory $h[n] = (0.9)^n$ (time constant $\approx 1/(1-a) = 10$ samples). Bandwidth and memory are reciprocal — the DT face of the RC filter's $\omega_c = 1/RC$.

> **Fourier / statistical-inference connection.** Fourier analysis's DFT chapter is this lesson plus Lesson 11 plus sampling; arriving with "DFT = sampled DTFT" in hand collapses it. For statistical inference (and any DSP work), the frequency response is the everyday tool: white noise into H comes out with spectrum $|H(e^{j\omega})|^2$, and the AR(1) process $x[n] = a\,x[n-1] + w[n]$ is exactly white noise through this lesson's first-order filter — compute $|H|^2$ for $a = 0.95$ and you have its power spectral density.

28.3 Exercises

Exercise 28.1 (Hand)**.** Compute the DTFT of $x = \delta[n] - \delta[n-2]$, factor out the phase, and locate its zeros in $[-\pi, \pi]$. What tones does this filter annihilate?

Exercise 28.2 (Hand)**.** For the three-point averager $y[n] = \frac{1}{3}(x[n] + x[n-1] + x[n-2])$: find $H(e^{j\omega})$ via the geometric sum (Example 28.2's window, shifted), locate its zeros, and evaluate $|H|$ at $\omega = 0,\ 2\pi/3,\ \pi$.

Exercise 28.3 (Proof, $\star$)**.** Prove Parseval for the DTFT, $\sum_n |x[n]|^2 = \frac{1}{2\pi}\int_{2\pi} |X(e^{j\omega})|^2 d\omega$, by inserting the definition of X, swapping sum and integral, and using $\frac{1}{2\pi}\int_{2\pi} e^{j\omega(m-n)}d\omega = \delta_{mn}$ (Lemma 26.2's DT twin — prove that too).

Exercise 28.4 (Proof)**.** Show that if $x[n]$ is real, $X(e^{-j\omega}) = \overline{X(e^{j\omega})}$, and deduce that $|H|$ of any real-coefficient filter is an even function of ω — so plotting $[0, \pi]$ suffices, which is why every DSP plot you will ever see stops at π (or at the "Nyquist frequency" in Hz).

Deeper reading. O&W 5.1–5.4 (DTFT, properties, convolution), 5.8 (LCCDE frequency responses), 6.1–6.4 (magnitude–phase, ideal vs. nonideal filters), 6.6 (first-order DT systems). For the dedicated discrete-time treatment, Oppenheim–Schafer–Buck, *Discrete-Time Signal Processing* (2nd ed.), Chs. 2–5, is the natural expansion of this lesson and the Course 3 filter and multirate labs (6.1–6.3, 5.4).

29 Sampling

NEEDED FOR: Fourier analysis (a signature topic) and statistical inference (prerequisite; all data is sampled). Not needed for ML.

29.1 Theory

Sampling bridges the CT world of Lessons 26–27 and the DT world of Lesson 28 — and explains when the bridge loses nothing.

Lemma 29.1 (Spectrum of the impulse train)**.** *Let $p(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT)$ and $\omega_s = 2\pi/T$. Then*

$$P(j\omega) = \frac{2\pi}{T} \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_s) :$$

an impulse train in time is an impulse train in frequency.

Proof. p is periodic with period T; its Fourier series coefficients are $a_k = \frac{1}{T}\int_{-T/2}^{T/2} \delta(t) e^{-jk\omega_0 t} dt = \frac{1}{T}$ for every k (sifting). So $p = \frac{1}{T}\sum_k e^{jk\omega_s t}$, and transforming term by term with the pair $e^{j\omega_0 t} \leftrightarrow 2\pi\delta(\omega - \omega_0)$ (Example 27.2) gives the claim. □

Theorem 29.2 (Spectrum replication)**.** *Let $x_p(t) = x(t)\,p(t)$ (the impulse-sampled signal, carrying exactly the sample values $x(nT)$). Then*

$$X_p(j\omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X\big(j(\omega - k\omega_s)\big) :$$

sampling in time replicates the spectrum at every multiple of ω_s, scaled by $1/T$.

Proof. Multiplication in time is $\frac{1}{2\pi}\times$ convolution in frequency (Theorem 27.4), and convolving X with the frequency impulse train of Lemma 29.1 places a copy of X at each $k\omega_s$ (sifting, once per impulse):

$$X_p = \frac{1}{2\pi}\, X * P = \frac{1}{2\pi} \cdot \frac{2\pi}{T} \sum_k X\big(j(\omega - k\omega_s)\big). \qquad \square$$

Theorem 29.3 (Sampling theorem (Nyquist–Shannon))**.** *Suppose x is* band-limited*: $X(j\omega) = 0$ for $|\omega| > \omega_M$. If*

$$\omega_s > 2\omega_M$$

then the replicas in Theorem 29.2 do not overlap, and x is exactly recoverable from its samples: pass x_p through an ideal lowpass filter of gain T and cutoff $\omega_c \in (\omega_M,\ \omega_s - \omega_M)$, which selects the $k = 0$ copy. Equivalently, in the time domain,

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT)\, \frac{\sin\big(\pi(t - nT)/T\big)}{\pi(t - nT)/T}$$

*(*sinc interpolation*: each sample carries a sinc, and the sincs vanish at every other sample instant). The critical rate $2\omega_M$ is the* Nyquist rate*. If $\omega_s < 2\omega_M$, the replicas overlap and add* — aliasing — *and reconstruction is impossible in general.*

Proof. Non-overlap: copy k occupies $[k\omega_s - \omega_M,\ k\omega_s + \omega_M]$; consecutive copies are disjoint iff $\omega_s - \omega_M > \omega_M$. The ideal lowpass with gain T then returns exactly $X(j\omega)$, hence x. For the interpolation formula: the filter's impulse response is $T \cdot \frac{\sin(\omega_c t)}{\pi t}$ (the inverse transform of a rect — Example 27.2 read backwards through duality), and the filter output is $x_p * h = \sum_n x(nT)\, h(t - nT)$ (convolution with an impulse train of weighted impulses; sifting once per term); with $\omega_c = \omega_s/2$ this is the displayed sum. □

Aliasing, concretely. Sample $\cos(\omega_0 t)$ with $\omega_s/2 < \omega_0 < \omega_s$: the replica arithmetic puts spectral lines at $\pm(\omega_s - \omega_0)$ inside the base band, so the reconstruction is $\cos\big((\omega_s - \omega_0)t\big)$ — a *lower* frequency impersonating the original. Wagon wheels spinning backwards in films, and the requirement of an analog *anti-aliasing* lowpass filter in front of every real sampler, are both this computation.

DT processing of CT signals. The practical payoff (O&W 7.4): sample above Nyquist, apply a DT filter $H_d(e^{j\omega})$, reconstruct. Within the band, the overall CT system is LTI with frequency response $H_d(e^{j\omega T})$ — so software can implement any (band-limited) analog filter, which is why signal processing moved into code.

29.2 Practice: by hand

Example 29.4 (Alias arithmetic)**.** $x(t) = \cos(2\pi \cdot 60\, t)$ sampled at $f_s = 100\,\text{Hz}$ (below Nyquist $= 120\,\text{Hz}$): the reconstructed tone appears at $100 - 60 = 40\,\text{Hz}$. At $f_s = 150\,\text{Hz}$ (above Nyquist): $60\,\text{Hz}$ survives exactly. Checking with samples: at $f_s = 100$, $x(n/100) = \cos(2\pi \cdot 0.6\, n) = \cos(2\pi \cdot 0.6n - 2\pi n) = \cos(2\pi \cdot (-0.4)n) = \cos(2\pi \cdot 0.4\, n)$ — the samples of a $40\,\text{Hz}$ tone, indistinguishable (Proposition 24.3(b) again: DT frequencies live mod 2π).

Example 29.5 (Budgeting a sampler)**.** Speech is intelligible below $\sim$4 kHz: telephony samples at 8 kHz. Hearing tops out near 20 kHz: CDs sample at 44.1 kHz — just above Nyquist, leaving a narrow $[20, 22.05]$ kHz guard band for the anti-aliasing filter to roll off in, which is why audiophile debates about that filter exist.

Fourier / statistical-inference connection. The sampling theorem is a Fourier-analysis centerpiece — Osgood devotes a chapter to it and its interplay with the DFT — and this

proof (replicate, then window) is the one taught there. For statistical inference, sampling is the standing assumption: every "random process" met in data is samples of something, sampled band-limited noise stays faithfully represented, and undersampled noise *folds* its out-of-band power into the base band (aliased noise), a standard exam trap. And the whole story is Proposition 24.3's asymmetry cashing out: DT cannot resolve frequencies mod ω_s, so you must not put content there.

29.3 Exercises

Exercise 29.1 (Hand)**.** For each, give the Nyquist rate and what appears at the output if sampled at $\omega_s = 2\pi \cdot 1000$: (a) $\cos(2\pi \cdot 300\, t)$; (b) $\cos(2\pi \cdot 700\, t)$; (c) their sum. (For (c): can any post-processing undo the damage? Why not?)

Exercise 29.2 (Hand)**.** A signal has spectrum supported on $[-2\pi \cdot 5\,\text{kHz},\ 2\pi \cdot 5\,\text{kHz}]$. It is sampled at 8 kHz *after* an anti-aliasing filter with cutoff 4 kHz. What band survives end-to-end? What would the spectrum near 3.5 kHz look like *without* the filter (which input frequencies fold there)?

Exercise 29.3 (Proof, $\star$)**.** Prove the samples-orthogonality behind the interpolation formula: the shifted sincs $\phi_n(t) = \operatorname{sinc}\big((t-nT)/T\big)$ satisfy $\phi_n(mT) = \delta_{mn}$ (immediate) *and* are orthogonal in L^2: $\int \phi_n \phi_m\, dt = T\,\delta_{mn}$ (transform each sinc to a rect via Example 27.2/duality, and integrate the product in the frequency domain via Parseval — phases from the shifts give Lemma 26.2's integral). So band-limited reconstruction is one more orthonormal expansion, and Theorem 6.3's geometry applies verbatim.

Exercise 29.4 (Proof)**.** Suppose x is band-limited to ω_M and you sample at exactly the Nyquist rate $\omega_s = 2\omega_M$. Exhibit a nonzero band-limited signal all of whose samples vanish (consider $\sin(\omega_M t)$), concluding that the strict inequality in Theorem 29.3 matters.

Deeper reading. O&W 7.1–7.3 (sampling theorem, reconstruction, aliasing), 7.4 (DT processing of CT signals); the modulation view of sampling is O&W 8.5–8.6 (pulse-amplitude modulation).

30 Laplace and z-Transforms, a Working Minimum

NEEDED FOR: Fourier analysis (background) and statistical inference (prerequisite; system stability fluency). Not needed for ML.

30.1 Theory

The Fourier transform evaluates a system's eigenvalue function $H(s)$ (Proposition 26.1) only on the axis $s = j\omega$. The Laplace transform keeps the whole complex plane — which is what lets it handle unstable systems, transients, and initial conditions.

Definition 30.1 (Laplace transform, ROC)**.** $X(s) = \int_{-\infty}^{\infty} x(t)e^{-st}dt$ for those $s \in \mathbb{C}$ where the integral converges absolutely — the *region of convergence* (ROC), always a vertical strip determined by $\operatorname{Re} s$. On $s = j\omega$ (when the ROC contains it), $X(j\omega)$ is the Fourier transform.

Example 30.2 (Same formula, different signals)**.**

$$e^{-at}u(t) \;\longleftrightarrow\; \frac{1}{s+a}, \quad \operatorname{Re} s > -a; \qquad -e^{-at}u(-t) \;\longleftrightarrow\; \frac{1}{s+a}, \quad \operatorname{Re} s < -a.$$

(Both are the same integral computed over the half-line where the signal lives.) Identical algebra, different ROCs, completely different signals — one right-sided, one left-sided. *The ROC is part of the transform*; quoting $X(s)$ without it determines nothing.

Proposition 30.3 (ROC facts for rational transforms)**.** *Let $X(s)$ be rational with poles $p_1, \dots, p_r$ (roots of the denominator — Lesson 13's territory). Then:*

(a) the ROC is a vertical strip containing no pole, bounded by poles (or extending to $\pm\infty$);

(b) x right-sided (e.g. causal) $\iff$ ROC is the half-plane to the right of the rightmost pole;

(c) the system with impulse response x is BIBO stable $\iff$ the ROC contains the $j\omega$-axis;

(d) hence: causal and stable $\iff$ all poles satisfy $\operatorname{Re} p_i < 0$ *(left half-plane).*

Proof sketch. (a) absolute convergence depends only on $\operatorname{Re} s$, and it fails at a pole. (b) for a right-sided x, convergence at s_0 implies convergence for all $\operatorname{Re} s > \operatorname{Re} s_0$ (the factor $e^{-(s-s_0)t}$ only helps as $t \to +\infty$); Example 30.2 is the template. (c) stability is $\int |h| < \infty$ (Theorem 25.3), which is exactly absolute convergence of the transform at $s = j\omega$. (d) combine (b) and (c): the right-of-rightmost-pole half-plane contains the axis iff every pole is strictly to its left. The instance $h = e^{-at}u(t)$: pole at $-a$, causal always, stable iff $a > 0$ — matching the direct computation $\int_0^\infty e^{-at}dt < \infty$. □

Inversion by partial fractions. Rational transforms invert by table lookup after a partial-fraction expansion (Lesson 5's linear algebra: matching coefficients is solving a linear system). Worked once, causal case:

$$H(s) = \frac{1}{(s+1)(s+2)} = \frac{1}{s+1} - \frac{1}{s+2} \quad\Longrightarrow\quad h(t) = \big(e^{-t} - e^{-2t}\big)u(t),$$

with ROC $\operatorname{Re} s > -1$; poles at $-1, -2$ in the left half-plane: stable, as h's decay confirms. Differential equations transform to algebra the same way as in Lesson 28: $\sum_k a_k y^{(k)} = \sum_k b_k x^{(k)}$ gives $H(s) = \big(\sum b_k s^k\big)/\big(\sum a_k s^k\big)$, by the differentiation property $x' \leftrightarrow sX$ (same proof as Theorem 27.3(e)).

Definition 30.4 (z-transform)**.** The DT twin: $X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}$ for $z \in \mathbb{C}$ in the ROC — always an annulus $r_1 < |z| < r_2$. On the unit circle $z = e^{j\omega}$ (when the ROC contains it), this is the DTFT — which is why Lesson 28 wrote $X(e^{j\omega})$.

The dictionary translates line by line ($e^{st} \to z^n$; vertical strips $\to$ annuli; left half-plane $\to$ inside the unit circle):

$$a^n u[n] \;\longleftrightarrow\; \frac{1}{1-az^{-1}}, \;\; |z| > |a|; \qquad \text{causal + stable} \iff \text{all poles satisfy } |p_i| < 1.$$

The first is Example 28.2's geometric series with $e^{j\omega}$ promoted to z; the second is Proposition 30.3(d) with the geometry renamed — and it finally unifies every "$|a| < 1$" in this booklet:

Lesson 9's decaying eigenvalues, Lesson 23's spectral gap, Lesson 25's stable recursion, Lesson 28's lowpass smoother. They are all "poles inside the unit circle."

Difference equations give rational $H(z) = \big(\sum_k b_k z^{-k}\big)/\big(\sum_k a_k z^{-k}\big)$ by the shift property $x[n-1] \leftrightarrow z^{-1}X(z)$, and $|H(e^{j\omega})|$ can be read geometrically from the pole–zero plot: magnitude response = (product of distances from $e^{j\omega}$ to zeros) / (product of distances to poles) — peaks near poles, notches near zeros (Exercise 28.1 built exactly such a notch).

30.2 Practice: by hand

Example 30.5 (A second-order system, end to end)**.** $H(z) = \dfrac{1}{(1-0.9z^{-1})(1+0.5z^{-1})}$, causal. Poles at 0.9 and -0.5: both inside the unit circle — stable. Partial fractions:

$$H(z) = \frac{9/14}{1-0.9z^{-1}} + \frac{5/14}{1+0.5z^{-1}} \implies h[n] = \Big(\tfrac{9}{14}(0.9)^n + \tfrac{5}{14}(-0.5)^n\Big)u[n]\,:$$

a slow decay (pole near the circle) plus a fast alternating one. Frequency response by geometry: the pole at 0.9 sits near $z = 1$, so $|H|$ peaks at $\omega = 0$; the pole at -0.5 mildly boosts $\omega = \pi$. The long-run behavior is owned by the pole of largest modulus — Lesson 9's dominant eigenvalue, verbatim.

Fourier / statistical-inference connection. Fourier analysis mostly stays on the $j\omega$-axis, but its treatment of one-sided signals and of the Laplace–Fourier relationship assumes this fluency; statistical inference uses pole locations as the standing language for stability of estimators and filters (a Kalman filter's error dynamics are stable because certain poles are inside the unit circle). If you can go from a difference equation to $H(z)$ to poles to a stability verdict to $h[n]$ without notes, this lesson has done its job.

30.3 Exercises

Exercise 30.1 (Hand)**.** Invert $X(s) = \dfrac{s+3}{(s+1)(s+2)}$ (causal): partial fractions, $x(t)$, stability verdict. Then re-answer for the ROC $\operatorname{Re} s < -2$ — which signal is it now, and is the Fourier transform defined?

Exercise 30.2 (Hand)**.** For $y[n] = 0.8y[n-1] + x[n] - x[n-1]$: find $H(z)$, its pole and zero, sketch $|H(e^{j\omega})|$ from the geometry (zero at $z = 1$!), and classify the filter (lowpass / highpass / notch). Compute $h[n]$ by partial fractions or long division.

Exercise 30.3 (Proof, $\star$)**.** Prove Proposition 30.3(c) in the DT case: the system with impulse response h is BIBO stable iff the ROC of $H(z)$ contains the unit circle. (Both directions are Theorem 25.3 plus the definition of absolute convergence at $|z| = 1$.)

Exercise 30.4 (Proof)**.** Prove the final-value flavor of Lesson 9 in transform language: if $H(z)$ is causal with all poles strictly inside the unit circle, then $h[n] \to 0$ geometrically, at rate $\max_i |p_i|$. (Partial fractions: each term is $c_i\, p_i^n u[n]$, possibly times a polynomial in n for repeated poles — bound it.)

Deeper reading. O&W 9.1–9.3, 9.5, 9.7 (Laplace, ROC, inversion, properties, LTI analysis), 10.1–10.5, 10.7 (the z-transform mirror); 9.4/10.4 (geometric evaluation from pole–zero plots) rewards the extra hour.

Part VI: The Analysis Behind the Transforms

Part V computed freely: it differentiated a square wave, integrated δ's, swapped $\int$ with $\sum$ and $\int$ with $\int$, evaluated transforms by table and partial fractions, and promised each time that the move "can be certified." This part is the certificate. It answers, with the bare minimum of real analysis, measure theory, functional analysis, complex analysis, distribution theory, and numerical linear algebra, the six questions Part V kept deferring:

1. When may a limit pass through a sum or an integral? (Lesson 31 — and why the Gibbs overshoot is a theorem, not a numerical artifact.)
2. What integral makes "energy," "almost everywhere," and every double-integral swap honest? (Lesson 35.)
3. What space do signals live in, and in what sense does a Fourier series actually converge? (Lesson 36.)
4. Why do poles, ROCs, partial fractions, and inversion integrals work? (Lesson 38.)
5. What *is* $\delta(t)$, and why may we differentiate discontinuous signals and Fourier-transform constants and impulse trains? (Lesson 40.)
6. What is a derivative with respect to a *function*, and where do matched filters and regularized deconvolution come from? (Lesson 41.)

Lesson 42 adds the compute layer — conditioning, backward stability, the factorizations inside `solve`, `lstsq`, `eigh`, `svd`, and `fft`, and the Krylov iterations that replace them at scale; Lesson 43 adds the probability payoff — the transforms of Part IV's distributions (moment generating and characteristic functions), the Chernoff bound, and the proof of the central limit theorem that Lesson 21 deferred. Sources, one per lesson, all self-contained here at working depth: Ross, *Elementary Analysis* (2nd ed.); Axler, *Measure, Integration & Real Analysis* (MIRA); Bak–Newman, *Complex Analysis* (3rd ed.); Strichartz, *A Guide to Distribution Theory and Fourier Transforms*; Bowers–Kalton, *An Introductory Course in Functional Analysis*; Trefethen–Bau, *Numerical Linear Algebra*; Bertsekas–Tsitsiklis §4.4 and Problem 5.12 for Lesson 43. The ground rule throughout is *bare minimum, honestly connected*: a proof appears when it teaches a mechanism you will reuse; a theorem is stated with a pointer when its proof is a course of its own. *Nothing here is needed for machine learning*, and Fourier analysis and statistical inference can formally be survived without it — but graduate Fourier analysis develops exactly Lesson 40, statistical inference's "almost surely" is exactly Lesson 35's "almost everywhere," and graduate DSP (estimation, adaptive filters, wavelets, multirate) speaks the language of the Course 3 DSP labs (Modules 5–6 and 9) as its native tongue. Part VI keeps Part V's convention j for the imaginary unit, even inside complex analysis.

31 Limits Done Right: Sequences, Series, and When Swapping Is Legal

NEEDED FOR: Optional rigor layer for Fourier analysis and statistical inference; foundation for graduate DSP. Not needed for ML.

31.1 Theory

Everything in analysis rests on one axiom that distinguishes $\mathbb{R}$ from the rationals.

Definition 31.1 (Completeness axiom)**.** Every nonempty set $S \subseteq \mathbb{R}$ that is bounded above has a *least* upper bound $\sup S \in \mathbb{R}$.

Definition 31.2 (Limit of a sequence)**.** $s_n \to s$ means: for every $\varepsilon > 0$ there is N such that $|s_n - s| < \varepsilon$ for all $n > N$. A sequence is *Cauchy* if for every $\varepsilon > 0$ there is N with $|s_n - s_m| < \varepsilon$ for all $n, m > N$.

Theorem 31.3 (Monotone and Cauchy convergence)**.** *(a) Every bounded monotone sequence converges. (b) A sequence of reals converges iff it is Cauchy.*

Proof. (a) Say $s_1 \le s_2 \le \cdots \le B$. Let $s = \sup_n s_n$ (completeness). Given $\varepsilon > 0$, $s - \varepsilon$ is not an upper bound, so some $s_N > s - \varepsilon$; by monotonicity $s - \varepsilon < s_n \le s$ for all $n > N$. (b) Convergent $\Rightarrow$ Cauchy is the triangle inequality. Conversely a Cauchy sequence is bounded, so $t_n = \sup_{m \ge n} s_m$ is a bounded decreasing sequence; by (a) it converges, say to s, and the Cauchy condition squeezes $s_n \to s$ as well (Ross §10 fills in the two-line estimate). □

Part (b) is the tool that proves convergence *without knowing the limit* — the only way to handle series whose sums have no closed form.

Definition 31.4 (Series)**.** $\sum_{k=1}^{\infty} a_k = \lim_{n\to\infty} \sum_{k=1}^{n} a_k$ when the limit of partial sums exists. The series *converges absolutely* if $\sum |a_k| < \infty$.

Proposition 31.5 (The three workhorses)**.** *(a) (Geometric) $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for $|r| < 1$, and diverges for $|r| \ge 1$.*

(b) (Comparison) If $|a_k| \le b_k$ and $\sum b_k < \infty$, then $\sum a_k$ converges, absolutely.

(c) Absolute convergence implies convergence.

Proof. (a) $(1 - r)\sum_{k=0}^{n} r^k = 1 - r^{n+1}$ and $r^{n+1} \to 0$ iff $|r| < 1$. (c) Partial sums of $\sum a_k$ are Cauchy because $\left|\sum_{k=m}^{n} a_k\right| \le \sum_{k=m}^{n} |a_k|$, and the right side is the Cauchy tail of a convergent series; apply Theorem 31.3(b). (b) The same estimate, with b_k as majorant, plus (c). □

Theorem 31.6 (Order matters, unless it doesn't)**.** *If $\sum a_k$ converges absolutely, every rearrangement converges to the same sum. If it converges only conditionally (like $1 - \frac{1}{2} + \frac{1}{3} - \cdots = \ln 2$), then for* any *target $t \in \mathbb{R} \cup \{\pm\infty\}$ some rearrangement converges to t (Riemann). (Proof: Ross §22 / Rudin; the mechanism — greedily overshoot with positive terms, undershoot with negative — is worth one minute of thought.)*

This is why Part V's hypotheses are what they are. Theorem 25.3 demanded $\sum_k |h[k]| < \infty$, and Proposition 25.2 freely reindexed the double sum $\sum_k \sum_m x[k]h_1[m]h_2[n-k-m]$: absolute convergence is precisely the license to reorder and regroup infinite sums as if they were finite. Without it, Theorem 31.6 says the "sum" is not even well defined.

From numbers to functions. A signal-processing limit is almost never a limit of numbers; it is a limit of *functions* (partial Fourier sums, truncated impulse responses, mollified pulses). Two inequivalent meanings:

Definition 31.7 (Pointwise and uniform convergence)**.** $f_n \to f$ *pointwise* on S if $f_n(t) \to f(t)$ for each fixed $t \in S$; *uniformly* if

$$\|f_n - f\|_\infty = \sup_{t \in S} |f_n(t) - f(t)| \longrightarrow 0,$$

i.e. one N works for every t simultaneously. (Note the sup-norm: uniform convergence is convergence in a norm on function space — Lesson 36 makes that view official.)

Example 31.8. $f_n(t) = t^n$ on $[0,1]$ converges pointwise to $f(t) = 0$ for $t < 1$, $f(1) = 1$ — but not uniformly: $\sup_t |f_n(t) - f(t)| = 1$ for every n (take t just below 1). The limit of these continuous functions is discontinuous.

Theorem 31.9 (Uniform limits preserve continuity)**.** *If each f_n is continuous on S and $f_n \to f$ uniformly, then f is continuous.*

Proof. The $\varepsilon/3$ argument. Fix t_0 and $\varepsilon > 0$. Choose n with $\|f_n - f\|_\infty < \varepsilon/3$, then $\eta > 0$ with $|f_n(t) - f_n(t_0)| < \varepsilon/3$ for $|t - t_0| < \eta$ (continuity of f_n). Then

$$|f(t) - f(t_0)| \le \underbrace{|f(t) - f_n(t)|}_{<\varepsilon/3} + \underbrace{|f_n(t) - f_n(t_0)|}_{<\varepsilon/3} + \underbrace{|f_n(t_0) - f(t_0)|}_{<\varepsilon/3} < \varepsilon. \qquad \square$$

Corollary 31.10 (Gibbs is inevitable)**.** *The partial Fourier sums of the square wave are continuous (finite sums of cosines); the square wave is not. Therefore the convergence* cannot *be uniform: by Theorem 31.9,* $\sup_t |S_N(t) - x(t)| \not\to 0$*. The overshoot of Lesson 26's partial sums, stalled near 9% of the jump, is this corollary made visible (its exact value,* $\frac{1}{\pi}\int_0^\pi \frac{\sin u}{u}\,du - \frac{1}{2} \approx 0.0895$*, is a one-line numerical integral). No amount of terms, and no amount of floating-point care, removes it.*

Theorem 31.11 (Swapping limit and integral: the uniform case)**.** *If $f_n \to f$ uniformly on a* bounded *interval $[a,b]$ and each f_n is integrable, then f is integrable and $\int_a^b f_n \to \int_a^b f$.*

Proof. $\left|\int_a^b f_n - \int_a^b f\right| \le \int_a^b |f_n - f| \le (b-a)\,\|f_n - f\|_\infty \to 0$. (Integrability of f: Ross §25.) $\square$

Example 31.12 (Both hypotheses are load-bearing)**.** *Bounded interval fails:* $f_n = \mathbf{1}_{[n,\,n+1]}$ (a bump marching to $+\infty$) converges to 0 pointwise, and even "locally uniformly," yet $\int_{\mathbb{R}} f_n = 1 \not\to 0$. *Uniformity fails:* $f_n = n \cdot \mathbf{1}_{[0,\,1/n]}$ converges to 0 pointwise on $[0,1]$ but $\int_0^1 f_n = 1$. Mass escaping to infinity or into a spike is exactly what the hypotheses forbid — keep both pictures; Lesson 35 tames them with domination instead of uniformity, and Lesson 40 will make the second one *converge after all*, to δ, in a weaker sense.

Theorem 31.13 (Weierstrass M-test)**.** *If $\|g_k\|_\infty \le M_k$ on S and $\sum_k M_k < \infty$, then $\sum_k g_k$ converges uniformly on S (and absolutely at each point).*

Proof. The partial sums are uniformly Cauchy: $\|\sum_{k=m}^n g_k\|_\infty \le \sum_{k=m}^n M_k$, a tail of a convergent series. Apply Theorem 31.3(b) at each t, and note the resulting pointwise limit is approached at the uniform rate $\sum_{k>n} M_k$. $\square$

Corollary 31.14 (Absolutely summable Fourier series behave perfectly)**.** *If $\sum_k |a_k| < \infty$, then $x(t) = \sum_k a_k e^{jk\omega_0 t}$ converges uniformly, x is continuous, and term-by-term integration against anything bounded is legal (Theorem 31.11) — in particular the coefficient formula $a_l = \frac{1}{T}\int_T x(t)e^{-jl\omega_0 t}dt$ of Theorem 26.3 is an honest computation, not a formal one: swap $\int_T \sum_k$ to $\sum_k \int_T$ and invoke Lemma 26.2.*

By the smoothness–decay law (Exercise 26.3), $\sum |a_k| < \infty$ holds whenever x is continuously differentiable ($|a_k| = O(1/k^2)$); the square wave, with its $|a_k| \sim 1/k$, sits precisely on the wrong side of this test — which is Gibbs again, from the coefficient side.

Theorem 31.15 (Term-by-term differentiation)**.** *If $\sum_k g_k$ converges at one point and the* differentiated *series $\sum_k g_k'$ converges uniformly on an interval, then $\sum g_k$ converges uniformly there, its sum is differentiable, and $\big(\sum g_k\big)' = \sum g_k'$. (Proof: Ross §26.)*

Note where the hypothesis sits: on the *derivatives*. $\sum \sin(kt)/k^2$ converges uniformly (M-test, $M_k = 1/k^2$) but its differentiated series $\sum \cos(kt)/k$ does not converge absolutely — one uniform-convergence hypothesis per derivative you take. That is why each degree of smoothness of x costs one power of decay of a_k, and it is the honest content of "differentiate the Fourier series of the triangle wave to get the square wave." (For genuinely discontinuous signals, differentiation must wait for Lesson 40, where it becomes *always* legal — at the price of the answer being a distribution.)

31.2 Practice: by hand

Example 31.16 (The four-example zoo, worth memorizing)**.** On the stated sets, with $f = \lim f_n$ pointwise:

f_n	set	uniform?	moral
t^n	$[0,1]$	$\times$	uniform limit would be continuous
$\frac{\sin(nt)}{\sqrt{n}}$	$\mathbb{R}$	✓	$f_n'(0) = \sqrt{n} \to \infty$: derivatives need their own test
$n\,\mathbf{1}_{[0,1/n]}$	$[0,1]$	$\times$	spike: $\int f_n = 1 \neq 0 = \int f$
$\mathbf{1}_{[n,n+1]}$	$\mathbb{R}$	$\times$	escape to ∞: same integral failure

The second row is the sharpest: $\|f_n - 0\|_\infty = 1/\sqrt{n} \to 0$, uniform convergence at its best, yet the derivatives diverge — convergence of f_n never controls f_n', exactly as Theorem 31.15's hypothesis warns.

Example 31.17 (A stability margin from a geometric tail)**.** For $h[n] = a^n u[n]$, $|a| < 1$, the truncation error of the impulse response is $\sum_{k>N} |a|^k = \frac{|a|^{N+1}}{1-|a|}$ (geometric tail). So an FIR truncation of this IIR filter is within ε of the true system — in the operator sense of Lesson 36 — once $N > \log\big(\varepsilon(1-|a|)\big)/\log|a|$. For $a = 0.9$, $\varepsilon = 10^{-6}$: $N \geq 153$. This one-line tail estimate is the entire theory behind "truncate the impulse response" in filter design.

Fourier / statistical-inference / grad DSP connection. Fourier analysis states its convergence theorems in exactly this vocabulary (pointwise vs. uniform vs. the L^2 sense of Lesson 36), and Gibbs (Corollary 31.10) is a lecture there. Statistical inference's limit theorems (Lesson 21) juggle three modes of convergence of random variables; the pointwise/uniform/norm trichotomy here is the deterministic warm-up — and Theorem 31.3 is

what "the empirical average converges" silently uses. In graduate DSP, every adaptive algorithm's convergence proof is a sequence argument, and every filter-truncation error bound is a tail estimate like the one above.

31.3 Exercises

Exercise 31.1 (Hand)**.** For each family decide pointwise limit and uniformity, with a one-line sup computation: (a) $f_n(t) = \frac{t}{1+nt^2}$ on $\mathbb{R}$; (b) $f_n(t) = \frac{nt}{1+nt}$ on $[0,1]$; (c) $f_n(t) = t^n(1-t)$ on $[0,1]$. (Answers: (a) uniform, sup $= \frac{1}{2\sqrt{n}}$; (b) not uniform — limit jumps at 0; (c) uniform — maximize to get sup $\sim \frac{1}{en}$.)

Exercise 31.2 (Hand)**.** How many terms of $\sum_k 1/k^2$ guarantee the partial sum is within 10^{-3} of $\pi^2/6$? (Bound the tail by $\int_N^\infty t^{-2}dt$.) Compare with the alternating series $\sum(-1)^{k+1}/k^2$, where the error is at most the first omitted term — conditional-style cancellation converges faster, but Theorem 31.6 says you must not reorder it carelessly.

Exercise 31.3 (Proof, $\star$)**.** Prove Theorem 31.13 in full (uniformly Cauchy $\Rightarrow$ uniformly convergent included), then use it to show that $x(t) = \sum_{k=1}^{\infty} 2^{-k}\cos(3^k t)$ is continuous on all of $\mathbb{R}$. (This is Weierstrass's famous example; it is differentiable *nowhere* — continuity from Theorem 31.9 costs nothing, differentiability is a different currency, Theorem 31.15.)

Exercise 31.4 (Proof)**.** Show that if $f_n \to f$ uniformly on $\mathbb{R}$ and each f_n is bounded, then f is bounded — and conclude (contrapositive) that $\frac{1}{t}\mathbf{1}_{(0,1)}$ is not a uniform limit of bounded functions. Then verify that the convergence $\frac{t}{1+nt^2} \to 0$ *is* uniform although each function is bounded by a different constant.

Deeper reading. Ross §§4, 7–12 (completeness, sequences), 14–15 (series), 23–26 (sequences and series of functions, uniform convergence, term-by-term theorems).

32 Metric Spaces, Continuity, and Compactness

NEEDED FOR: Optional rigor layer: the vocabulary (open, closed, compact, complete) used by every later analysis lesson and by convergence statements in optimization. Not needed for ML.

32.1 Theory

Lesson 31 did analysis on the real line. But the objects this book cares about — signal vectors in $\mathbb{R}^k$, functions on an interval, sequences in ℓ^2 — do not live on a line, and the theorems we want ("the minimum is attained," "the iteration converges," "the limit of continuous things is continuous") should not have to be re-proved for each new home. The right abstraction carries over everything from Lesson 31 at the cost of one definition: all we ever used about $|a-b|$ was that it is a *distance*.

Definition 32.1 (Metric space)**.** A *metric* on a set S is a function $d(x,y)$ defined for pairs $x, y \in S$ with (D1) $d(x,x) = 0$ and $d(x,y) > 0$ for $x \neq y$; (D2) $d(x,y) = d(y,x)$; (D3) the triangle inequality $d(x,z) \leq d(x,y) + d(y,z)$. The pair (S,d) is a *metric space* — the pair, because one set can carry many useful metrics (Ross §13).

Example 32.2 (The working zoo)**.** (a) $\mathbb{R}$ with $d(a,b) = |a-b|$. (b) $\mathbb{R}^k$ with the Euclidean metric $d(x,y) = \|x-y\|_2$ — D3 is the triangle inequality of Lesson 2, i.e. Cauchy–Schwarz. (c) $\mathbb{R}^k$ again with $d_\infty(x,y) = \max_j |x_j - y_j|$: same set, different metric, and the two are related by $d_\infty \le d_2 \le \sqrt{k}\, d_\infty$, so they have the *same* convergent sequences. (d) $C[a,b]$, the continuous functions on $[a,b]$, with $d(f,g) = \|f-g\|_\infty = \max_t |f(t)-g(t)|$: convergence in this metric *is* uniform convergence — Lesson 31's definition, now recognized as ordinary metric convergence in a bigger space (Kreyszig §1.1–1.2). (e) Sequence spaces like ℓ^∞ and ℓ^p (Kreyszig §1.2), waiting for Lesson 36.

Definition 32.3 (Convergence, Cauchy, completeness)**.** In (S,d): $s_n \to s$ means $d(s_n,s) \to 0$; (s_n) is *Cauchy* if for every $\varepsilon > 0$ there is N with $d(s_m,s_n) < \varepsilon$ for $m,n > N$. The space is *complete* if every Cauchy sequence converges to a point of the space (Ross §13.2). Completeness is the property Theorem 31.3(b) gave $\mathbb{R}$; it is what lets you prove convergence without producing the limit first.

Theorem 32.4 ($\mathbb{R}^k$ is complete; Bolzano–Weierstrass)**.** *(a) A sequence in $\mathbb{R}^k$ converges (is Cauchy) iff each of its k coordinate sequences converges (is Cauchy) in $\mathbb{R}$; hence $\mathbb{R}^k$ is complete. (b) Every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.*

Proof. (a) The coordinatewise squeeze $|x_j - y_j| \le d_2(x,y) \le \sqrt{k}\max_j |x_j - y_j|$ transfers both properties between $\mathbb{R}^k$ and the coordinates; then apply Theorem 31.3(b) coordinatewise (Ross 13.3–13.4). (b) Extract a subsequence making coordinate 1 converge (Bolzano–Weierstrass on $\mathbb{R}$, itself from Theorem 31.3(a) via a monotone subsequence), then a further subsequence for coordinate 2, and so on k times; later extractions do not disturb earlier limits (Ross 13.5). □

Not every metric space is complete: $\mathbb{Q}$ is not (a rational sequence can chase $\sqrt{2}$), and $C[a,b]$ under the *integral* metric $\int |f-g|$ is not — a sequence of ramps steepening into a step is Cauchy but has no continuous limit. That failure is precisely why Lesson 35 has to build L^1 and L^2: they are the *completions* of these too-small spaces (Kreyszig §1.6).

The topological vocabulary. Four words, used on every later page (Ross §13.6–13.9): $s_0 \in E$ is *interior* to E if some ball $\{s : d(s,s_0) < r\} \subseteq E$; E is *open* if every point is interior ($E = E^\circ$); E is *closed* if its complement is open; the *closure* E^- is the smallest closed set containing E, and the *boundary* is $E^- \setminus E^\circ$. Arbitrary unions and finite intersections of open sets are open — and the finiteness matters: $\bigcap_n (-\frac{1}{n}, \frac{1}{n}) = \{0\}$ is not open.

Proposition 32.5 (Closed = stable under limits)**.** *E is closed iff it contains the limit of every convergent sequence of its own points; and E^- is exactly the set of limits of sequences from E (Ross 13.9).*

This is the version you actually use: constraint sets like $\{x : \|Ax-b\| \le \varepsilon\}$ are closed because they are preimages of closed sets under continuous maps, so limits of feasible points stay feasible — the silent step in every "take a minimizing sequence" argument of Part VII.

Definition 32.6 (Compactness)**.** An *open cover* of E is a family of open sets whose union contains E; E is *compact* if every open cover has a finite subcover (Ross 13.11). In $\mathbb{R}^k$ this abstract property has a checkable characterization:

Theorem 32.7 (Heine–Borel)**.** *$E \subseteq \mathbb{R}^k$ is compact iff it is closed and bounded.*

Proof sketch. Compact ⇒ bounded: cover E by the nested open boxes $\{x : \max_j |x_j| < m\}$, $m \in \mathbb{N}$; a finite subcover means one box already contains E. Compact ⇒ closed: around any $x_0 \notin E$ the open sets $V_m = \{x : d(x, x_0) > \frac{1}{m}\}$ cover E; a finite subcover keeps a whole ball around x_0 out of E. For the converse, first show a closed bounded box (k-cell) is compact by bisection: if some cover had no finite subcover, halving the box k-fold repeatedly yields nested sub-boxes with the same defect and diameters → 0; their intersection point (nonempty by Theorem 32.4 — Ross 13.10) lies in one open set of the cover, which then swallows all small enough sub-boxes, contradiction. A closed subset of a compact set is compact, which finishes it (Ross 13.12–13.13). The bisection is Theorem 31.3's "zoom in on the bad half" argument, promoted to k dimensions — and the same mechanism reappears in Goursat's proof of Cauchy's theorem (Lesson 38). □

Continuity, in one uniform language. (Ross §21.) For $f : (S, d) \to (S^*, d^*)$, the following are equivalent: (i) for each s_0 and $\varepsilon > 0$ there is $\delta > 0$ with $d(s, s_0) < \delta \Rightarrow d^*(f(s), f(s_0)) < \varepsilon$; (ii) $s_n \to s$ implies $f(s_n) \to f(s)$; (iii) preimages of open sets are open (Ross 21.1–21.3). Version (ii) is the workhorse for computations; version (iii) is why solution sets and constraint sets inherit open/closedness for free.

Theorem 32.8 (Continuity meets compactness)**.** *Let $f : S \to S^*$ be continuous. (a) If $E \subseteq S$ is compact, then $f(E)$ is compact. (b) (Extreme Value Theorem) A continuous real-valued f on a nonempty compact E attains a maximum and a minimum on E. (c) On a compact E, continuity upgrades to* uniform *continuity: one δ serves every point at once (Ross 21.4, 19.2).*

Proof. (a) Pull back an open cover of $f(E)$ through (iii), take the finite subcover, push forward. (b) By (a) and Heine–Borel, $f(E) \subseteq \mathbb{R}$ is closed and bounded, so $\sup f(E)$ is finite and, being a limit of points of $f(E)$ (Proposition 32.5), belongs to $f(E)$. (c) is the covering argument of Ross §19 (or: a failure of uniformity yields two sequences at distance → 0 with f-values staying ε apart; compactness extracts convergent subsequences and contradicts (ii)). □

Part (b) is the most-used theorem in this book that Part I never stated: "the least-squares minimum exists," "the maximum-likelihood estimate exists on the constraint set," "some unit vector maximizes $\|Av\|$" (which is how the SVD's σ_1 exists, Lesson 12) are all instances — continuous objective, closed and bounded feasible set, done. When a feasible set is unbounded (most of Part VII), coercivity restores compactness by trapping minimizing sequences in a ball.

Connectedness, briefly. E is *disconnected* if it splits into two nonempty pieces separated by open sets, *connected* otherwise; continuous images of connected sets are connected, and the connected subsets of $\mathbb{R}$ are exactly the intervals (Ross §22). Hence the Intermediate Value Theorem: a continuous real function on an interval takes every value between any two of its values (Ross 22.2–22.3). This is the theorem behind "a root exists between a sign change" — bisection root-finding, and the existence half of many spec statements ("the filter's response crosses −3 dB somewhere").

32.2 Practice: by hand

Example 32.9 (Interior, closure, boundary drill)**.** In $\mathbb{R}$ with the usual metric, let $E = (0, 1] \cup \{2\}$. Then $E^\circ = (0, 1)$ (2 has no ball inside E); $E^- = [0, 1] \cup \{2\}$ (add the limit 0; Proposition 32.5);

boundary $= \{0, 1, 2\}$. E is neither open (1 is not interior) nor closed (0 is a missing limit). Compact? Not closed, so no. $[0, 1] \cup \{2\}$ *is* compact: closed and bounded. ✓

Example 32.10 (Same set, three metrics)**.** On $\mathbb{R}$, compare $d_1(x, y) = |x - y|$, $d_2(x, y) = \min\{|x - y|, 1\}$, and $d_3(x, y) = |x - y|^2$. d_2 is a metric (check D3 by cases) with exactly the same convergent sequences as d_1 — boundedness of the metric costs nothing locally. d_3 fails D3: $d_3(0, 2) = 4 > 1 + 1 = d_3(0, 1) + d_3(1, 2)$. Squaring a metric can break the triangle inequality; taking $\sqrt{\ }$ never does (Exercise below).

Where you will use this. Lesson 35: "L^2 is complete" is the whole point of building it. Lesson 36: norms induce metrics, and every statement there (ℓ^2 complete, projections exist) is a metric-space statement. Lesson 38: open sets are the domains of holomorphy; Goursat's bisection is Heine–Borel's. Part VII: existence of minimizers (Theorem 32.8(b) plus coercivity), closedness of constraint sets, and the completeness that makes gradient descent's limit exist before you know what it is. In numerical work, Theorem 32.4(a) is the license to test convergence of a vector iteration coordinate by coordinate.

32.3 Exercises

Exercise 32.1 (Hand)**.** For each candidate metric on $\mathbb{R}$, verify D1–D3 or exhibit a failure: (a) $d(x, y) = \sqrt{|x - y|}$; (b) $d(x, y) = (x-y)^2$; (c) $d(x, y) = |x-2y|$; (d) the discrete metric $d(x, y) = 1$ for $x \neq y$, 0 for $x = y$. For the ones that are metrics, describe what a ball $\{y : d(x, y) < r\}$ looks like.

Exercise 32.2 (Hand)**.** For $E_1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\} \cup \{(1, 0)\}$, $E_2 = \{(x, y) : xy = 1\}$, and $E_3 = \{(x, y) : x^2 + y^2 \leq 1,\ x \geq 0\}$: find interior, closure, and boundary; decide open/closed/neither; decide compact or not (Heine–Borel), and for the non-compact ones name the failing hypothesis.

Exercise 32.3 (Proof)**.** Prove Proposition 32.5: E is closed iff every convergent sequence of points of E has its limit in E. Then use it to show $\{x \in \mathbb{R}^k : \|Ax - b\|_2 \leq \varepsilon\}$ is closed for any matrix A.

Exercise 32.4 (Proof, $\star$)**.** Prove Theorem 32.8(a) and (b) in full from the open-cover definition and the preimage characterization of continuity. Then deduce this Part-I fact: $\sigma_1 = \max_{\|v\|=1} \|Av\|$ is attained — name the compact set and the continuous function.

Deeper reading. Ross §13 (metric spaces, completeness, Heine–Borel), §§17–20 (continuity on $\mathbb{R}$), §§21–22 (continuity and connectedness in metric spaces); Kreyszig §§1.1–1.6 (the metric-space zoo, completion — the bridge to Lesson 36).

33 Differentiation: The Calculus Theorems, Done Right

NEEDED FOR: Optional rigor layer: every rule of calculus you already use, with its proof; the honest foundation under gradients (Lesson 7) and every optimization step in Part VII. Not needed for ML.

33.1 Theory

Part I used derivatives as a machine; this lesson proves the machine. Everything below is Ross ch. 5, compressed to what the rest of the book leans on.

Definition 33.1 (Derivative)**.** For f defined on an open interval containing a,

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a},$$

when the limit exists; equivalently $f(a+h) = f(a) + f'(a)h + r(h)$ with $r(h)/h \to 0$: the best *linear* approximation at a, with a remainder that dies faster than h. Keep the second form in view — it is the one that survives to $\mathbb{R}^n$.

Theorem 33.2 (Differentiable ⇒ continuous)**.** *If $f'(a)$ exists, f is continuous at a: as $x \to a$, $f(x) - f(a) = \frac{f(x)-f(a)}{x-a} \cdot (x-a) \to f'(a) \cdot 0 = 0$ (Ross 28.2). The converse fails — $|x|$ at 0, and Lesson 31's Weierstrass series* everywhere.

Theorem 33.3 (The algebra of derivatives)**.** *If f, g are differentiable at a (Ross 28.3): (i) $(cf)' = cf'$ and $(f+g)' = f' + g'$; (ii)* product rule $(fg)'(a) = f(a)g'(a) + f'(a)g(a)$*; (iii)* quotient rule $(f/g)'(a) = \dfrac{g(a)f'(a) - f(a)g'(a)}{g(a)^2}$ *when $g(a) \neq 0$.*

Proof. (ii) Add and subtract the mixed term:

$$\frac{f(x)g(x) - f(a)g(a)}{x-a} = f(x)\,\frac{g(x)-g(a)}{x-a} + g(a)\,\frac{f(x)-f(a)}{x-a} \longrightarrow f(a)g'(a) + g(a)f'(a),$$

using continuity of f at a (Theorem 33.2) on the first factor. (iii) Prove $(1/g)'(a) = -g'(a)/g(a)^2$ the same way, then apply (ii). □

Theorem 33.4 (Chain rule)**.** *If f is differentiable at a and g at $f(a)$, then $(g \circ f)'(a) = g'(f(a)) \cdot f'(a)$.*

Proof idea. The tempting two-factor telescoping $\frac{g(f(x))-g(f(a))}{f(x)-f(a)} \cdot \frac{f(x)-f(a)}{x-a}$ divides by zero whenever $f(x) = f(a)$ near a. The repair (Ross 28.4): replace the first factor by the function $h(y) = \frac{g(y)-g(f(a))}{y-f(a)}$ for $y \neq f(a)$, $h(f(a)) = g'(f(a))$ — continuous at $f(a)$ by differentiability of g — so that $g(f(x)) - g(f(a)) = h(f(x))\,(f(x) - f(a))$ holds *identically*, zeros included, and the limit is legal. □

The mean value package. Three statements, each one line from the last (Ross §29), all resting on Theorem 32.8(b):

Theorem 33.5 (Interior extrema; Rolle; Mean Value Theorem)**.** *(a) If f on an open interval attains a max or min at x_0 and $f'(x_0)$ exists, then $f'(x_0) = 0$: if $f'(x_0) > 0$, points just right of x_0 would beat it; if $f'(x_0) < 0$, points just left would (Ross 29.1). (b) (Rolle) If f is continuous on $[a,b]$, differentiable on (a,b), and $f(a) = f(b)$, then $f'(x) = 0$ for some $x \in (a,b)$: the max and min of f on the compact $[a,b]$ exist; if both sit at endpoints f is constant, otherwise an interior extremum triggers (a) (Ross 29.2). (c) (MVT) If f is continuous on $[a,b]$, differentiable on (a,b), then some $x \in (a,b)$ has*

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

— subtract the chord $L(t)$ from f and apply Rolle to $g = f - L$ (Ross 29.3).

Part (a) is why "set the gradient to zero" finds candidates and only candidates — the entire first-order logic of Part VII in one line. The MVT's consequences are the facts calculus treats as obvious (Ross 29.4–29.7): $f' \equiv 0$ on an interval forces f constant; $f' = g'$ forces $f = g + c$ (*the* uniqueness statement behind antiderivatives, used the moment Lesson 34 proves the Fundamental Theorem); $f' > 0$ forces strictly increasing. None of these is available from the definition alone — you need a point where the derivative *sees* the secant.

Theorem 33.6 (Generalized MVT; L'Hospital)**.** *(a) For f, g continuous on $[a, b]$, differentiable on (a, b), some $x \in (a, b)$ has $f'(x)[g(b) - g(a)] = g'(x)[f(b) - f(a)]$ (apply Rolle to the right cross-difference; Ross 30.1). (b) If $\lim_{x\to s} f'(x)/g'(x) = L$ exists and either $f, g \to 0$ or $|g| \to \infty$ at s (any of a, $a^{\pm}$, $\pm\infty$), then $\lim_{x\to s} f(x)/g(x) = L$ (Ross 30.2).*

L'Hospital is (a) plus bookkeeping — and its hypotheses earn their keep: the *derivative* ratio must converge, and $g' \neq 0$ nearby. (Use it twice on $\sin(x)/x$-type limits and you go in a circle: that limit *is* the derivative of sin at 0.)

Theorem 33.7 (Taylor's theorem with Lagrange remainder)**.** *If $f^{(n)}$ exists on $(a, b) \ni c$, then for each $x \in (a, b)$ there is y between c and x with*

$$f(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(c)}{k!}(x - c)^k + R_n(x), \qquad R_n(x) = \frac{f^{(n)}(y)}{n!}\,(x - c)^n.$$

Proof idea. Define M by demanding equality at x with M in place of $f^{(n)}(y)$; the difference $g(t) = f(t) - \sum_{k<n} \frac{f^{(k)}(c)}{k!}(t - c)^k - \frac{M(t-c)^n}{n!}$ has $g(c) = g'(c) = \cdots = g^{(n-1)}(c) = 0$ and $g(x) = 0$, so n successive applications of Rolle's theorem walk a zero of each derivative toward c until $g^{(n)}(y) = f^{(n)}(y) - M = 0$ (Ross 31.3, Wolfe's proof). $n = 1$ *is* the MVT. □

This is the workhorse behind every "≈" in the book: the remainder is not folklore but $\frac{f^{(n)}(y)}{n!}(x - c)^n$, so bounding one derivative on an interval bounds the approximation error — the $O(\|h\|^2)$ in Lesson 7's gradient expansions, the step-size conditions of Part VII, and the linearization error every extended Kalman filter budgets for. For power series $f(x) = \sum a_k x^k$ inside the radius of convergence, term-by-term differentiation (Theorem 31.15) gives $a_k = f^{(k)}(0)/k!$ — Taylor coefficients and power-series coefficients are the same thing (Ross 31.1).

33.2 From one variable to n: the same theorems, upgraded

Everything above survives in $\mathbb{R}^n$ once "$f'(a)$ is a number" is replaced by the right idea — already present in Definition 33.1's second form. This subsection is the bridge from Ross to Rudin (ch. 9); Part I's gradient facts are the $m = 1$ row of it.

Definition 33.8 (Total derivative (Rudin 9.11))**.** $f : \mathbb{R}^n \supseteq E \to \mathbb{R}^m$ is *differentiable* at x if a linear map $A : \mathbb{R}^n \to \mathbb{R}^m$ satisfies

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

and then $f'(x) = A$ — unique (Rudin 9.12), a *matrix*, not a number. Its entries are the partial derivatives: $[f'(x)]_{ij} = \partial f_i/\partial x_j$, the *Jacobian* (Rudin 9.17). For $m = 1$ the $1 \times n$ derivative is the transpose of the gradient ∇f, and $f'(x)h = \langle \nabla f(x), h \rangle$ — Lesson 7's directional derivative, with steepest ascent along ∇f (Rudin 9.18–19 discussion).

Three one-variable habits need care in $\mathbb{R}^n$ (Rudin 9.16, 9.21): existence of all partials does *not* imply differentiability (partials only probe axis directions); but *continuous* partials do — $f \in \mathcal{C}^1$ iff all $\partial f_i/\partial x_j$ exist and are continuous, and $\mathcal{C}^1$ implies differentiable. This is the hypothesis under which Part I freely wrote Jacobians.

Theorem 33.9 (Chain rule and the MVT bound in $\mathbb{R}^n$)**.** *(a) If f is differentiable at x_0 and g at $f(x_0)$, then $(g \circ f)'(x_0) = g'(f(x_0))\, f'(x_0)$ — a* product of matrices *(Rudin 9.15). Backpropagation is this equation read right to left. (b) There is no exact n-dimensional MVT for vector-valued f, but the useful consequence survives: if E is convex and $\|f'(x)\| \le M$ on E, then $\|f(b) - f(a)\| \le M\|b - a\|$ (restrict f to the segment $\gamma(t) = (1-t)a + tb$, apply the one-variable MVT to $t \mapsto f(\gamma(t))$; Rudin 9.19). In particular $f' \equiv 0$ on a convex open set forces f constant.*

Part (b) is the Lipschitz estimate quoted throughout Part VII: a bounded Hessian makes ∇f Lipschitz, which is exactly the "L-smoothness" constant that sets gradient-descent step sizes.

Theorem 33.10 (Contraction principle; inverse and implicit function theorems)**.** *(a) (Banach fixed point, Rudin 9.23) If X is a* complete *metric space (Lesson 32) and $\varphi : X \to X$ satisfies $d(\varphi(x), \varphi(y)) \le c\, d(x, y)$ with $c < 1$, then φ has exactly one fixed point, and the iteration $x_{k+1} = \varphi(x_k)$ converges to it from any start, geometrically: $d(x_{k+1}, x_k) \le c^k\, d(x_1, x_0)$. (b) (Inverse function theorem, Rudin 9.24) If $f \in \mathcal{C}^1$ near a and the matrix $f'(a)$ is invertible, then f is a bijection between a neighborhood of a and a neighborhood of $f(a)$, with $\mathcal{C}^1$ inverse and $g'(y) = [f'(g(y))]^{-1}$. The proof solves $f(x) = y$ by running (a) on $\varphi(x) = x + f'(a)^{-1}(y - f(x))$ — Newton's stance: invert the linearization, iterate. (c) (Implicit function theorem, Rudin 9.28) If $F(x, y) = 0$, $F \in \mathcal{C}^1$, and the block $\partial F/\partial y$ is invertible at a solution, then y is locally a $\mathcal{C}^1$ function of x, with $y'(x) = -(\partial F/\partial y)^{-1}\, \partial F/\partial x$.*

Statement (a) is the abstract engine: it is why fixed-point iterations, Newton steps near a good root, and the value-iteration loops of Lesson 23-style problems converge — completeness supplies the limit, contraction supplies the rate. Statement (c) is the differentiation-under-constraints fact that KKT analysis (Lesson 46) and every sensitivity computation quietly invoke. n-dimensional *integration* needs no new theory here: Lesson 35 integrates on any measure space, and its Fubini–Tonelli theorem is the statement that $\int_{\mathbb{R}^2}$ may be computed one variable at a time.

33.3 Practice: by hand

Example 33.11 (Every rule at once)**.** $f(x) = \dfrac{e^x \sin x}{1 + x^2}$. Product rule inside, quotient rule outside:

$$f'(x) = \frac{(1 + x^2)\, e^x(\sin x + \cos x) - 2x\, e^x \sin x}{(1 + x^2)^2}.$$

At $x = 0$: $f'(0) = \frac{1 \cdot 1 \cdot (0+1) - 0}{1} = 1$. Sanity check by Taylor: $e^x \sin x = x + x^2 + O(x^3)$ and $(1 + x^2)^{-1} = 1 + O(x^2)$, so $f(x) = x + O(x^2)$: slope 1. ✓

Example 33.12 (Taylor remainder as an error budget)**.** How well does $1 + \frac{x}{2}$ approximate $\sqrt{1 + x}$ on $|x| \le 0.1$? Here $f(x) = (1 + x)^{1/2}$, $f''(y) = -\frac{1}{4}(1 + y)^{-3/2}$, so by Theorem 33.7 with $n = 2$, $c = 0$:

$$|R_2(x)| = \Big|\frac{f''(y)}{2}\Big| x^2 \le \frac{(0.9)^{-3/2}}{8}(0.1)^2 \approx 1.5 \times 10^{-3}.$$

One derivative bound → one guaranteed error bar — the pattern for every linearization in the labs (small-signal models, EKF Jacobians, $\ln(1+x) \approx x$ steps in decibel arithmetic).

ML / optimization connection. Lesson 7 computed ∇f for least squares by pattern matching; this lesson is the license: $\mathcal{C}^1$ implies differentiable, chain rule as matrix product (backprop), interior extrema have zero gradient, smoothness constants from Theorem 33.9(b). Part VII's convergence proofs (Lesson 47) are Taylor's theorem with the remainder tracked, and Newton's method is the inverse function theorem run numerically. The contraction principle is the one theorem shared by all of them.

33.4 Exercises

Exercise 33.1 (Hand)**.** Differentiate, naming the rule at each step, and evaluate at the given point: (a) $x^2 \sin(1/x)$ at general $x \neq 0$; then decide from the *definition* whether $f(x) = x^2 \sin(1/x)$, $f(0) = 0$, is differentiable at 0, and whether f' is continuous there. (b) $\dfrac{\ln(1+x^2)}{x}$ at $x = 1$. (c) $g(t) = f(\gamma(t))$ for $f(x_1, x_2) = x_1^2 x_2$, $\gamma(t) = (\cos t,\ t^2)$, via the $\mathbb{R}^n$ chain rule (Jacobian times velocity), at $t = 0$.

Exercise 33.2 (Hand)**.** (a) Compute $\lim_{x\to 0} \frac{x - \sin x}{x^3}$ by L'Hospital (how many rounds?) and check it against the Taylor expansion of sin. (b) Use Theorem 33.7 to bound the error of $\cos x \approx 1 - \frac{x^2}{2}$ on $|x| \leq 0.5$, and give the smallest n for which the degree-n Taylor polynomial of e^x at 0 is within 10^{-6} on $[0, 1]$ (bound $R_n \leq e/(n+1)!$).

Exercise 33.3 (Proof)**.** Prove the product rule and the quotient rule in full from Definition 33.1 (including the $(1/g)'$ lemma and the continuity step it needs). Then prove Ross 29.4–29.5 from the MVT: $f' \equiv 0$ on an interval implies f constant, and $f' = g'$ implies $f = g + \text{const}$.

Exercise 33.4 (Proof, ⋆)**.** (a) Prove Rolle's theorem and the MVT from Theorem 33.5(a) and Theorem 32.8(b), stating exactly where compactness enters. (b) Prove the contraction principle, Theorem 33.10(a): show the iterates are Cauchy via the geometric tail $\sum_k c^k$, use completeness, and prove uniqueness. Conclude with the error bound $d(x_k, x_*) \leq \frac{c^k}{1-c}\, d(x_1, x_0)$ — the a priori estimate you would use to decide how many iterations a fixed-point loop needs.

Deeper reading. Ross §28 (derivative, rules, chain rule), §29 (MVT and consequences), §30 (L'Hospital), §31 (Taylor's theorem, power series); Rudin, *Principles of Mathematical Analysis* ch. 9 (total derivative, chain rule, $\mathcal{C}^1$, contraction principle, inverse/implicit function theorems).

34 The Riemann Integral and the Fundamental Theorem

NEEDED FOR: Optional rigor layer: what $\int_a^b$ actually means, which functions have one, and the proofs of the rules (FTC, parts, substitution) used on every page of Part V. Not needed for ML.

34.1 Theory

Part V integrated freely; this lesson defines the integral it was using and proves the three rules everything else reduces to. The construction is Darboux's squeeze (Ross §32): trap the would-be area between sums that overshoot and sums that undershoot.

Definition 34.1 (Darboux sums and the integral)**.** Let f be bounded on $[a,b]$ and $P=\{a=t_0<t_1<\cdots<t_n=b\}$ a partition. With $M_k=\sup_{[t_{k-1},t_k]} f$ and $m_k=\inf_{[t_{k-1},t_k]} f$, the *upper* and *lower Darboux sums* are

$$U(f,P)=\sum_{k=1}^{n} M_k\,(t_k-t_{k-1}),\qquad L(f,P)=\sum_{k=1}^{n} m_k\,(t_k-t_{k-1}).$$

Refining P lowers U and raises L, and every lower sum is $\le$ every upper sum (Ross 32.2–32.4). Set $U(f)=\inf_P U(f,P)$ and $L(f)=\sup_P L(f,P)$; then $L(f)\le U(f)$, and f is *integrable* when they agree, the common value being $\int_a^b f$.

Theorem 34.2 (Cauchy criterion for integrability)**.** *f is integrable on $[a,b]$ iff for each $\varepsilon>0$ some partition has $U(f,P)-L(f,P)<\varepsilon$ (Ross 32.5). Moreover the same holds with "some partition" replaced by "every partition of small enough mesh," and the familiar* Riemann sums *$\sum_k f(x_k)(t_k-t_{k-1})$ (any tags x_k) then converge to $\int_a^b f$ — the Darboux and Riemann definitions agree (Ross 32.7–32.9).*

This is Theorem 31.3(b)'s spirit again: prove the integral exists without naming its value, by squeezing the gap. It immediately yields the two big classes:

Theorem 34.3 (Who is integrable)**.** *(a) Every monotone f on $[a,b]$ is integrable: on a mesh-δ uniform partition the gap telescopes, $U-L=\delta\sum_k(M_k-m_k)=\delta\,|f(b)-f(a)|$ (Ross 33.1). (b) Every continuous f on $[a,b]$ is integrable: f is* uniformly *continuous on the compact $[a,b]$ (Theorem 32.8(c)), so a fine enough mesh forces $M_k-m_k<\varepsilon/(b-a)$ on every subinterval and $U-L<\varepsilon$ (Ross 33.2). (c) Piecewise continuous or bounded piecewise monotone functions are integrable (patch (a) and (b) across the pieces; Ross 33.8) — square waves, quantizer staircases, and every signal Part V drew are covered.*

Note where each hypothesis of (b) works: compactness buys uniform continuity buys a single δ for the whole interval. That chain is this book's standard proof pattern, worth owning once and for all.

Theorem 34.4 (The integral's algebra)**.** *For integrable f,g on $[a,b]$ (Ross 33.3–33.6): linearity $\int(cf+g)=c\int f+\int g$; monotonicity $f\le g\Rightarrow\int f\le\int g$; the triangle inequality $\left|\int_a^b f\right|\le\int_a^b|f|$; and splitting $\int_a^b f=\int_a^c f+\int_c^b f$. The triangle inequality is the ML bound of Lesson 37 in embryo, and monotonicity is the one-line source of every "bound the integral by the sup" estimate in Lessons 31 and 35.*

The Fundamental Theorem, in both directions. Ross (§34) states the two halves separately, and they *are* different theorems: one integrates a derivative, the other differentiates an integral.

Theorem 34.5 (FTC I: $\int$ of a derivative)**.** *If g is continuous on $[a,b]$, differentiable on (a,b), and g' is integrable on $[a,b]$, then*

$$\int_a^b g'=g(b)-g(a).$$

Proof. Take any partition with $U(g', P) - L(g', P) < \varepsilon$ (Theorem 34.2). The MVT (Theorem 33.5(c)) on each subinterval produces x_k with $g(t_k) - g(t_{k-1}) = g'(x_k)(t_k - t_{k-1})$, so the telescoping sum $g(b) - g(a) = \sum_k g'(x_k)(t_k - t_{k-1})$ is a Riemann sum of g', trapped between $L(g', P)$ and $U(g', P)$ — as is $\int_a^b g'$. The two numbers differ by less than ε (Ross 34.1). □

Every antiderivative table lookup in calculus, and every "evaluate at the endpoints" in Part V, is this theorem. The MVT is its entire engine.

Theorem 34.6 (FTC II: derivative of an integral)**.** *Let f be integrable on $[a, b]$ and $F(x) = \int_a^x f(t)\,dt$. Then F is (uniformly) continuous on $[a, b]$; and at every point x_0 where f is continuous, $F'(x_0) = f(x_0)$ (Ross 34.3).*

Proof idea. Continuity: $|F(y) - F(x)| \le B|y - x|$ with $B = \sup|f|$ (Theorem 34.4). Differentiability at a continuity point: the difference quotient is the *average* of f over $[x_0, x]$, and averages of a function over shrinking intervals converge to its value wherever it is continuous — squeeze the average between $\inf f$ and $\sup f$ on the shrinking interval. □

Integration *smooths*: an integrable f with jumps still has continuous F, one degree better. (Iterate: each integration buys one derivative — the smoothness–decay staircase of Part V's transforms, seen in the time domain.) The two rules of calculus now cost one line each (Ross 34.2, 34.4):

Theorem 34.7 (Integration by parts; change of variable)**.** *(a) If u, v are continuous on $[a, b]$, differentiable on (a, b), with u', v' integrable, then*

$$\int_a^b u\,v' + \int_a^b u'\,v = u(b)v(b) - u(a)v(a)$$

— apply FTC I to $(uv)' = uv' + u'v$ (product rule, Theorem 33.3). (b) If u is $\mathcal{C}^1$ on an interval and f is continuous on u's range, then $\int_a^b f(u(x))\,u'(x)\,dx = \int_{u(a)}^{u(b)} f(u)\,du$ — both sides, as functions of b, have the same derivative $f(u(b))u'(b)$ by FTC II and the chain rule (Theorem 33.4), and agree at $b = a$; now use Ross 29.5 ($f' = g'$ forces $f = g + c$).

Parts is the single most-used identity downstream: it is how Fourier coefficients of derivatives were computed (Lesson 26), how every transform-table "differentiation in time" entry is proved (Lesson 39 does this honestly), and — in its distributional form — how Lesson 40 *defines* the derivative of δ.

The ceiling, and what comes next. The Riemann integral handles a bounded function on a bounded interval whose discontinuities are tame, and no more. Improper integrals ($\int_{-\infty}^{\infty}$, unbounded integrands) are bolted on as limits; worse, the class is not closed under pointwise limits — the very next lesson opens with a bounded increasing sequence of integrable functions whose limit has *no* Riemann integral. Lesson 35 rebuilds integration to fix exactly this, and nothing computed here changes value. For $\mathbb{R}^n$: the Darboux construction runs verbatim on boxes, and in the Lebesgue theory Fubini–Tonelli reduces $\int_{\mathbb{R}^n}$ to n one-dimensional integrals — which is how every double integral in this book is actually evaluated.

34.2 Practice: by hand

Example 34.8 (A Darboux computation from scratch)**.** $\int_0^b x^2\,dx$ via uniform partitions $t_k = kb/n$: on $[t_{k-1}, t_k]$, x^2 is increasing, so $M_k = t_k^2$, $m_k = t_{k-1}^2$, and

$$U(f, P_n) = \sum_{k=1}^{n} \frac{k^2 b^2}{n^2} \cdot \frac{b}{n} = \frac{b^3}{n^3} \cdot \frac{n(n+1)(2n+1)}{6} \longrightarrow \frac{b^3}{3},$$

with $L(f, P_n)$ the same sum one index lower, also $\to b^3/3$. Squeezed: $\int_0^b x^2 dx = b^3/3$. ✓ (And FTC I gets it from $(x^3/3)' = x^2$ in one line — the point of the theorem.)

Example 34.9 (How fast do Riemann sums converge?)**.** For $\mathcal{C}^1$ f with $|f'| \le M$ on $[a, b]$, the MVT gives $M_k - m_k \le M\,(t_k - t_{k-1})$ on each subinterval, so on a uniform mesh $h = (b-a)/n$:

$$0 \le U - L \le M \sum_k h \cdot h = \frac{M(b-a)^2}{n}.$$

A left-endpoint sum therefore approximates $\int_a^b f$ within $O(1/n)$ — and this is the honest error model for "sum $\times$ Δt" numerical integration of a sampled signal; smoother f and smarter tags (midpoint, trapezoid) buy $O(1/n^2)$, which is Taylor's theorem (Theorem 33.7) applied per subinterval.

> **Signals connection.** Part V's every $\int$: linearity and splitting are Theorem 34.4; evaluate-at-the-endpoints is FTC I; "the running integral of a bounded signal is continuous" is FTC II; every transform-property proof in Lesson 39 is parts or substitution. In the labs: numerically integrating a sampled waveform inherits the $O(1/n)$–$O(1/n^2)$ error ladder above, and energy computations $\int |x|^2$ move to Lesson 35's L^2 the moment limits enter.

34.3 Exercises

Exercise 34.1 (Hand)**.** Compute $\int_0^1 x\,dx$ directly from Darboux sums on the uniform partition (find $U(f, P_n)$, $L(f, P_n)$ exactly, then their common limit). Then show the Dirichlet-type function $f = \mathbf{1}_{\mathbb{Q}}$ on $[0, 1]$ has $U(f) = 1$, $L(f) = 0$: not integrable — every subinterval contains both rationals and irrationals.

Exercise 34.2 (Hand)**.** Evaluate using the theorems, naming each: (a) $\int_0^\pi t \sin t\,dt$ (parts); (b) $\int_0^1 \frac{x}{1+x^2}\,dx$ (substitution $u = 1 + x^2$); (c) $F'(x)$ for $F(x) = \int_0^{x^2} e^{-t^2} dt$ (FTC II plus chain rule).

Exercise 34.3 (Proof)**.** Prove Theorem 34.3(a) and (b) in full from Theorem 34.2, exhibiting the partitions. State precisely where uniform continuity (not mere continuity) is used in (b), and which theorem supplies it.

Exercise 34.4 (Proof, $\star$)**.** Prove FTC I (Theorem 34.5) in full. Then deduce integration by parts, and use FTC II to prove the change-of-variable formula as in Theorem 34.7(b). Finally, exhibit the one-line counterexample logic: why does FTC I *not* follow from FTC II alone when g' is integrable but not continuous? (Point to the hypothesis FTC II needs at x_0.)

Deeper reading. Ross §32 (Darboux and Riemann definitions), §33 (integrability classes and properties), §34 (both fundamental theorems, parts, change of variable); MIRA ch. 1 for the same story told as the on-ramp to measure theory.

35 Integration Done Right: Lebesgue, Almost Everywhere, and the Swap Theorems

NEEDED FOR: Optional rigor layer for Fourier analysis and statistical inference; the language of graduate probability and DSP. Not needed for ML.

35.1 Theory

Lesson 31 bought $\lim \int = \int \lim$ with uniform convergence on bounded intervals. Part V needs far more: transforms integrate over all of $\mathbb{R}$, and the sequences that matter (partial inversions, truncated spectra) converge anything but uniformly. The Riemann integral cannot deliver:

Example 35.1 (Riemann's ceiling)**.** Enumerate the rationals in $[0,1]$ as $r_1, r_2, \ldots$ and let $f_n = \mathbf{1}_{\{r_1,\ldots,r_n\}}$. Each f_n is Riemann integrable (finitely many spikes, integral 0), f_n increases pointwise to $f = \mathbf{1}_{\mathbb{Q}\cap[0,1]}$ — and f is not Riemann integrable at all: every subinterval contains rationals and irrationals, so all upper sums are 1 and all lower sums are 0. The class of Riemann integrable functions is not closed under the limits we care about.

Lebesgue's fix has two ingredients: measure the *sets* first, and integrate by slicing the *range* instead of the domain.

Definition 35.2 (Measure zero; almost everywhere)**.** $E \subseteq \mathbb{R}$ has *measure zero* if for every $\varepsilon > 0$ it can be covered by countably many intervals of total length $< \varepsilon$. A property holds *almost everywhere* (a.e.) if it holds off a set of measure zero. Every countable set is null (cover r_k by an interval of length $\varepsilon 2^{-k}$; total $\leq \varepsilon$).

Proposition 35.3 (Outer measure, and what an interval is worth)**.** *For $A \subseteq \mathbb{R}$ let $|A| = \inf\{\sum_k \ell(I_k) : A \subseteq \bigcup_k I_k\}$, the cheapest countable cover by open intervals (MIRA 2A). Then: (a) $|A| \leq |B|$ for $A \subseteq B$, and $|t + A| = |A|$ (translating a set translates its covers); (b) countable* sub*additivity, $|\bigcup_k A_k| \leq \sum_k |A_k|$: merge covers that are $\varepsilon 2^{-k}$-cheap; (c) $|[a,b]| = b - a$.*

Proof of (c), the one that costs something. $\leq$: the single cover $(a - \varepsilon,\, b + \varepsilon)$. $\geq$: let $\{I_k\}$ cover $[a,b]$. By Heine–Borel (Theorem 32.7) *finitely* many of the I_k already cover $[a,b]$, and for a finite cover, induction on the number of intervals gives $\sum \ell(I_k) \geq b - a$: some interval covers a, walk to just past its right endpoint, repeat. Compactness is the whole content — it converts an infinite combinatorial claim into a finite one (MIRA 2.12–2.14). □

So far, so plausible. The next theorem is the shock that shapes the entire subject.

Theorem 35.4 (You cannot measure every set)**.** *There is no function μ defined on* all *subsets of $\mathbb{R}$ that (i) gives every interval its length, (ii) is countably additive over disjoint sets, and (iii) is translation invariant (MIRA 2.22).*

Proof idea (MIRA 2.18). Declare $a \sim b$ on $[-1,1]$ when $a - b \in \mathbb{Q}$, and let V collect one point from each equivalence class (this choice *is* the Axiom of Choice at work). The countably many rational translates $r + V$, $r \in \mathbb{Q} \cap [-2,2]$, are pairwise disjoint, jointly cover $[-1,1]$, and all fit inside $[-3,3]$; so (i)–(iii) force

$$2 \;\leq\; \sum_r \mu(r+V) \;=\; \sum_r \mu(V) \;\leq\; 6.$$

But a countable sum of one constant is 0 or ∞. Contradiction either way. □

The repair is not to weaken additivity — countable additivity is exactly what Example 35.1 demands — but to shrink the *domain*: measure only a class of sets closed under everything analysis actually does.

Definition 35.5 (σ-algebras, Borel sets, Lebesgue measure)**.** A *σ-algebra* on X is a collection of subsets containing $\varnothing$ and closed under complements and *countable* unions (hence countable intersections). The *Borel sets* form the smallest σ-algebra on $\mathbb{R}$ containing the open sets — every interval, every closed set, every countable combination of them, and anything constructible without the Axiom of Choice (MIRA 2.23–2.29). On the Borel sets (and their enlargement by null sets, the *Lebesgue measurable* sets) outer measure *is* countably additive (MIRA 2D); this restricted $|\cdot|$ is *Lebesgue measure*. The pathology of Theorem 35.4 is quarantined — no set an engineer writes down ever meets it.

Example 35.6 (The Cantor set: null does not mean small-looking)**.** Delete the open middle third of $[0,1]$, then of each remaining interval, forever. What survives is the *Cantor set*: closed, Borel, with measure $\lim_n (2/3)^n = 0$ — yet *uncountable* (its points are exactly the base-3 expansions using digits 0 and 2, as many as there are binary sequences; MIRA 2D). So "null" is a statement about total length, not about cardinality or intricacy — a.e.-statements silently ignore uncountably many points at once. (The digit-censoring set of the exercises below is this construction in base 10.)

Definition 35.7 (Lebesgue measure and the integral, in three strokes)**.** (1) Measure sets by Lebesgue measure (Definition 35.5) — countably additive, which is the property that powers everything below. (2) A function is *measurable* if preimages of intervals are measurable sets ($\{t : f(t) > a\}$ measurable for every a suffices). (3) For $f \geq 0$,

$$\int f = \sup\Big\{ \textstyle\sum_i c_i\,|E_i| : \sum_i c_i \mathbf{1}_{E_i} \leq f \Big\},$$

the supremum over *simple* (finite-valued) minorants: slice the range into levels c_i, measure the level sets E_i. For general f, write $f = f^+ - f^-$ and set $\int f = \int f^+ - \int f^-$ when $\int |f| < \infty$; such f are called *(Lebesgue) integrable*, written $f \in L^1$. When the Riemann integral exists, the two integrals agree (MIRA 3B) — nothing you computed in Part V changes value.

Proposition 35.8 (Measurability is indestructible)**.** *Sums, products, $|f|$, and $\max(f,g)$ of measurable functions are measurable; and if each f_n is measurable, so are $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, and the pointwise limit wherever it exists (MIRA 2.46–2.53).*

Proof of the limit claims. $\{\sup_n f_n > a\} = \bigcup_n \{f_n > a\}$: a *countable* union of measurable sets, which is exactly what a σ-algebra tolerates. inf is the complement trick, and $\limsup_n f_n = \inf_m \sup_{n \geq m} f_n$ stacks the two. Compare Example 35.1, where an increasing limit escaped the Riemann class: here the class is closed under every countable limit operation — the design requirement, met by construction. □

Proposition 35.9 (Exactly whom Riemann could handle)**.** *A bounded f on $[a,b]$ is Riemann integrable iff its set of discontinuity points is null; and then the two integrals agree (MIRA 3.34). The integrable classes of Lesson 34 (continuous, monotone, piecewise) are special cases; $\mathbf{1}_{\mathbb{Q}}$, discontinuous everywhere, is the failure mode, exactly diagnosed. The Riemann integral was measure theory in disguise all along: its verdict on f is a statement about the measure of f's bad set.*

Two immediate dividends. First, $\mathbf{1}_{\mathbb{Q}\cap[0,1]}$ integrates to 0 (its level set is null): Example 35.1's ceiling is gone. Second:

Proposition 35.10 (Null sets are invisible)**.** *If $f = g$ a.e. then $\int f = \int g$ (over any set), so f and g have the same energy, the same Fourier transform, the same everything expressible by integrals. Conversely if $f \geq 0$ and $\int f = 0$ then $f = 0$ a.e.*

Proof. $f - g$ vanishes off a null set E; every simple minorant of $|f - g|$ is supported in E and so has integral $\leq \|\text{minorant}\|_\infty \cdot |E| = 0$. For the converse: the set $\{f > 1/n\}$ has measure zero (else the simple function $\frac{1}{n}\mathbf{1}_{\{f>1/n\}}$ would witness $\int f > 0$), and $\{f > 0\} = \bigcup_n \{f > 1/n\}$ is a countable union of null sets, hence null. □

This is why "the" inverse Fourier transform is only defined up to a.e. equality, and why nobody cares what a reconstructed signal does at the single point of a jump — Theorem 26.3's midpoint convention at discontinuities is a choice of representative, not a fact about the signal.

Theorem 35.11 (Monotone Convergence Theorem, MCT)**.** *If $0 \leq f_1 \leq f_2 \leq \cdots$ and $f_n \to f$ pointwise (a.e. suffices), then $\int f_n \to \int f$, finite or infinite.*

Proof. $\int f_n$ increases and never exceeds $\int f$, so its limit exists with $\lim \int f_n \leq \int f$. For $\geq$: fix a simple minorant $\sum_j c_j \mathbf{1}_{A_j} \leq f$ and a safety margin $t \in (0, 1)$. The sets $E_n = \{t : f_n \geq t \sum_j c_j \mathbf{1}_{A_j}\}$ increase with union everything, and countable additivity in its "continuity from below" form (MIRA 2.59) gives $|A_j \cap E_n| \to |A_j|$; hence $\int f_n \geq t \sum_j c_j |A_j \cap E_n| \to t \sum_j c_j |A_j|$. Let $t \uparrow 1$, then take the supremum over simple minorants (MIRA 3.11). □

Theorem 35.12 (Egorov: pointwise is uniform, off a sliver)**.** *If $|X| < \infty$ and measurable $f_n \to f$ pointwise on X, then for every $\varepsilon > 0$ there is a measurable $E \subseteq X$ with $|X \setminus E| < \varepsilon$ on which the convergence is* uniform *(MIRA 2.85).*

Proof idea. For each accuracy $1/m$, the sets $A_{N,m} = \{t : |f_k(t) - f(t)| < \frac{1}{m} \text{ for all } k \geq N\}$ increase to X as $N \to \infty$, so some $A_{N_m,m}$ misses less than $\varepsilon 2^{-m}$ of X (continuity from below); take $E = \bigcap_m A_{N_m,m}$. Lesson 31 drew a hard line between pointwise and uniform convergence; Egorov says that on finite-measure sets the line is only ε thick. □

Theorem 35.13 (Bounded Convergence Theorem)**.** *If $|X| < \infty$, $f_n \to f$ pointwise, and one constant c bounds every $|f_n|$, then $\int f_n \to \int f$.*

Proof. Split X by Egorov: off the good set E the error is at most $2c\,|X \setminus E|$; on it, at most $|X| \sup_E |f_n - f| \to 0$ (MIRA 3.26). The marching bump and the growing spike of Example 31.12 escape only through an infinite domain or unbounded height — exactly the two hypotheses. □

Theorem 35.14 (Fatou's lemma: the one-sided safety net)**.** *For measurable $f_n \geq 0$, $\int \liminf_n f_n \leq \liminf_n \int f_n$.*

Proof. $g_n = \inf_{k \geq n} f_k$ increases to $\liminf f_n$ with $g_n \leq f_n$; MCT on (g_n) gives $\int \liminf f_n = \lim \int g_n \leq \liminf \int f_n$. □

Equality can genuinely fail — the spike $n\mathbf{1}_{[0,1/n]}$ has $\int \liminf = 0 < 1 = \liminf \int$ — and the direction is the useful one: in the limit, mass can vanish (escape, concentrate) but never appear from nowhere. Fatou is the standard tool for certifying integrability *before* any convergence theorem applies.

Theorem 35.15 (Dominated Convergence Theorem, DCT — the workhorse)**.** *If $f_n \to f$ a.e. and there is a single $g \in L^1$ with $|f_n| \le g$ a.e. for all n, then $f \in L^1$ and*

$$\lim_n \int f_n = \int f.$$

Proof idea. Reduce to Theorem 35.13: since $\int g < \infty$, for any ε there is a set E of *finite* measure with $\int_{X\setminus E} g < \varepsilon$ (an integrable function keeps its mass on finite measure), and truncating where g is huge costs another ε (MIRA 3.28–3.29). On what remains the Bounded Convergence Theorem applies, and each g-controlled sliver adds at most 2ε to $|\int f_n - \int f|$ (MIRA 3.31). (Many texts route this through *Fatou's lemma* instead — $\int \liminf f_n \le \liminf \int f_n$ for $f_n \ge 0$, itself one line from MCT applied to $\inf_{k\ge n} f_k$; either road ends here.) □

The hypothesis is exactly calibrated to Example 31.12: the marching bumps $\mathbf{1}_{[n,n+1]}$ have no integrable dominator (their sup-envelope is $\mathbf{1}_{[1,\infty)}$, integral ∞), and neither do the spikes $n\mathbf{1}_{[0,1/n]}$ (envelope $\sim 1/t$). One integrable roof over the whole family forbids both escape routes — no uniformity, no bounded interval required.

Corollary 35.16 (The transform-side dividends)**.** *Let $x \in L^1(\mathbb{R})$, $X(j\omega) = \int x(t)e^{-j\omega t}dt$.*

(a) X is continuous, and $|X(j\omega)| \le \|x\|_1$ everywhere.

(b) If also $t\,x(t) \in L^1$, then X is differentiable with $\frac{dX}{d\omega} = \int(-jt)x(t)e^{-j\omega t}dt$: differentiation under the integral sign — Part V's "multiply by t $\leftrightarrow$ differentiate in ω" rule (dual of Theorem 27.3), now a theorem.

(c) The same argument gives: the Laplace transform $X(s)$ of Proposition 26.1 is differentiable in s — in the complex sense of Lesson 38 — at every interior point of its ROC.

Proof. (a) For $\omega_n \to \omega$, the integrands $x(t)e^{-j\omega_n t}$ converge pointwise and are all dominated by $|x| \in L^1$: DCT. The bound is the triangle inequality for integrals. (b) Difference quotients:

$$\frac{X(j(\omega+h)) - X(j\omega)}{h} = \int x(t)e^{-j\omega t}\,\frac{e^{-jht}-1}{h}\,dt.$$

The quotient $\frac{e^{-jht}-1}{h}$ tends to $-jt$ pointwise and is bounded in modulus by $|t|$ (the chord of the unit circle is shorter than the arc), so the integrands are dominated by $|t\,x(t)| \in L^1$: DCT again. (c) Identical, with e^{-st}; the domination $|t|e^{-\sigma t}|x(t)|$ is integrable slightly inside the ROC, which is why the statement is for interior points. □

Theorem 35.17 (Tonelli–Fubini: the double-integral license)**.** *For measurable $f(t,\tau)$ on $\mathbb{R}^2$: (Tonelli) if $f \ge 0$, the iterated integrals in either order and the double integral are all equal (possibly $+\infty$), unconditionally; (Fubini) if $\iint |f| < \infty$ — checkable by Tonelli — then the same equalities hold for f itself. (MIRA ch. 5.)*

Corollary 35.18 (Convolution certified)**.** *If $x, h \in L^1(\mathbb{R})$, then $(x*h)(t) = \int x(\tau)h(t-\tau)d\tau$ is defined for a.e. t, lies in L^1 with $\|x*h\|_1 \le \|x\|_1\|h\|_1$, and the convolution property (Theorem 27.4) holds honestly: $\widehat{x*h} = X(j\omega)H(j\omega)$.*

Proof. Tonelli on $|x(\tau)||h(t-\tau)| \geq 0$:

$$\iint |x(\tau)||h(t-\tau)|\,d\tau\,dt = \int |x(\tau)|\Big(\int |h(t-\tau)|dt\Big)d\tau = \|x\|_1\|h\|_1 < \infty,$$

(inner integral is shift-invariant). Finiteness of the double integral means the inner integral $\int |x(\tau)h(t-\tau)|d\tau$ is finite for a.e. t (Proposition 35.10 applied to the t-marginal), which is the "defined a.e." claim, and the displayed bound is $\|x * h\|_1 \leq \|x\|_1\|h\|_1$. Now Fubini applies to $x(\tau)h(t-\tau)e^{-j\omega t}$ (its modulus was just integrated), so the swap in Part V's proof of Theorem 27.4 — integrate over t first, substitute $u = t-\tau$, factor — is legal, line by line. □

The same two-step (Tonelli to check, Fubini to swap) certifies Parseval's derivation, the interchange $\frac{1}{T}\int_T \sum_k$ in Fourier series manipulations (sums are integrals with respect to counting measure — one theorem covers both), and every "exchange the order of integration" in statistical inference's expectation calculations.

Proposition 35.19 (The approximation ladder)**.** *For $f \in L^1$ and $\varepsilon > 0$ there are, each within ε of f in L^1: a* simple *function (finitely many values — straight from the integral's definition), a* step *function (constant on finitely many intervals — approximate each level set by intervals), and a* continuous *function of bounded support (round the corners) (MIRA 3.44, 3.47). Ladder proofs ride this constantly: verify an identity on steps, where it is a computation, then extend to all of L^1 by continuity of both sides — the density argument that Lessons 36 and 39 invoke by name.*

Theorem 35.20 (Riemann–Lebesgue lemma)**.** *If $x \in L^1(\mathbb{R})$ then $X(j\omega) \to 0$ as $|\omega| \to \infty$.*

Proof sketch. For $x = \mathbf{1}_{[a,b]}$, compute: $|X(j\omega)| = \left|\frac{e^{-j\omega a} - e^{-j\omega b}}{j\omega}\right| \leq \frac{2}{|\omega|} \to 0$. Finite combinations of such steps inherit the property; and step functions are dense in L^1 (Proposition 35.19), so for general x pick a step s with $\|x - s\|_1 < \varepsilon$ and use $|X| \leq |\hat{s}| + \|x - s\|_1$ (Corollary 35.16(a)'s bound applied to $x - s$). □

Read backwards, this is a design constraint: a transform that does *not* vanish at infinity (the ideal lowpass brick wall $\mathbf{1}_{[-\omega_c,\omega_c]}$, viewed as the transform of h) cannot come from an L^1 impulse response — so the ideal filter of Lesson 28 is necessarily unstable in the BIBO sense of Theorem 25.3. The sinc's $1/t$ tail is not sloppiness; it is forced.

35.2 Practice: by hand

Example 35.21 (Fubini's hypothesis is not decoration)**.** $f(t,\tau) = \dfrac{t^2 - \tau^2}{(t^2+\tau^2)^2}$ on $(0,1)^2$. Using $\frac{\partial}{\partial\tau}\frac{\tau}{t^2+\tau^2} = f(t,\tau)$:

$$\int_0^1\Big(\int_0^1 f\,d\tau\Big)dt = \int_0^1 \frac{dt}{1+t^2} = \frac{\pi}{4}, \qquad \int_0^1\Big(\int_0^1 f\,dt\Big)d\tau = -\frac{\pi}{4}.$$

The two orders *disagree*, which by Fubini convicts the modulus: $\iint |f| = \infty$ (check: on the wedge $\tau < t$, $|f| \gtrsim 1/(4t^2)$ near the corner). Whenever Part V swapped a double integral, Tonelli's finiteness check was silently passing; this example is what failure looks like.

Example 35.22 (DCT in one line, twice)**.** (i) $\lim_n \int_0^\infty e^{-t}\cos^n(t)\,dt = 0$: integrand $\to 0$ a.e. (only at $t \in \pi\mathbb{Z}$ does $|\cos t| = 1$, a null set), dominated by $e^{-t} \in L^1$. (ii) $\lim_{a\downarrow 0}\int_0^\infty e^{-at}\frac{\sin t}{t^2+1}dt = \int_0^\infty \frac{\sin t}{t^2+1}dt$: dominated by $\frac{1}{t^2+1}$. This move — "evaluate a hard limit by passing it through the integral" — is used constantly in Fourier analysis to regularize divergent-looking transforms (the e^{-at} is Abel regularization, and DCT is why it gives the right answer when one exists).

Statistical-inference connection (the big one). A probability space *is* a measure space with total measure 1; $\mathbf{E}[X]$ *is* the Lebesgue integral of X; "almost surely" *is* "almost everywhere"; B&T's technical footnotes (which events have probabilities, why $\mathbf{E}$ is linear even for wild random variables, when $\mathbf{E}[\lim] = \lim \mathbf{E}$) are Definition 35.7, MCT, and DCT verbatim. The strong law of large numbers in Lesson 21 is an a.e. statement; statistical inference's convergence-of-random-variables menagerie is Lesson 31's pointwise/uniform story with measure replacing the sup. This lesson is the standard rigor layer under all graduate probability and statistical signal processing.

35.3 Exercises

Exercise 35.1 (Hand)**.** Which sets are null? (a) $\mathbb{Z}$; (b) $\mathbb{Q}$; (c) $[0,1]$; (d) the set of $t \in [0,1]$ whose decimal expansion contains no digit 7 (cover by intervals at each digit depth: total length $(9/10)^n \to 0$ — an *uncountable* null set).

Exercise 35.2 (Hand)**.** Justify or refute with DCT/MCT: (a) $\lim_n \int_0^1 \frac{n\sin(t/n)}{t(1+t^2)}dt = \int_0^1 \frac{dt}{1+t^2}$; (b) $\lim_n \int_0^n \left(1-\frac{t}{n}\right)^n e^{t/2}dt = \int_0^\infty e^{-t/2}dt$; (c) $\lim_n \int_{\mathbb{R}} \frac{n\,dt}{1+n^2t^2} = 0$ (careful — compute it instead).

Exercise 35.3 (Proof, $\star$)**.** Prove Corollary 35.16(a) and (b) in full detail, stating exactly where DCT is invoked and what the dominating function is. Then use (b) twice to compute the transform of $t^2e^{-|t|}$ from the known pair $e^{-|t|} \leftrightarrow \frac{2}{1+\omega^2}$ (Example "Two-sided decay," Lesson 27).

Exercise 35.4 (Proof)**.** Using Tonelli as in Corollary 35.18, prove the DT statement: if $x, h \in \ell^1$ then $x * h \in \ell^1$ with $\|x*h\|_1 \le \|x\|_1\|h\|_1$ — i.e. the cascade-stability Exercise 25.3, which you proved by hand there, is a special case of Fubini with counting measure.

Deeper reading. MIRA ch. 1 (Riemann's limits), 2A–2D (outer measure, σ-algebras, non-measurable sets, Lebesgue measure), 2E (Egorov), 3A–3B (integration, MCT, BCT, DCT, Riemann $\leftrightarrow$ Lebesgue), 5A–5B (product measures, Tonelli/Fubini). For the probability dictionary: MIRA ch. 12 or any graduate probability text's opening chapter.

36 Signal Space: Banach and Hilbert Spaces, L^2, and Orthonormal Bases

NEEDED FOR: Optional rigor layer for Fourier analysis and statistical inference; the native language of graduate DSP. Not needed for ML.

36.1 Theory

Part I did linear algebra in $\mathbb{R}^n$; Part V quietly worked in vector spaces of *signals*, where dimension is infinite and a new question appears that finite dimensions never asks: do limits stay in the space?

Definition 36.1 (Normed space, Banach space)**.** A *norm* on a vector space V is a function $\|\cdot\| : V \to [0,\infty)$ with $\|v\| = 0$ iff $v = 0$, $\|\alpha v\| = |\alpha|\|v\|$, and the triangle inequality. It makes V a metric space via $d(u,v) = \|u - v\|$; V is a *Banach space* if it is *complete*: every Cauchy sequence converges to a point of V. The standing examples:

$\ell^1,\ \ell^2,\ \ell^\infty$:	sequences with $\|x\|_1 = \sum \|x[n]\|$, $\|x\|_2 = \big(\sum \|x[n]\|^2\big)^{1/2}$, $\|x\|_\infty = \sup \|x[n]\|$;
$L^1(\mathbb{R}),\ L^2(\mathbb{R})$:	signals with $\|x\|_1 = \int \|x\|$, $\|x\|_2 = \big(\int \|x\|^2\big)^{1/2}$ (Lebesgue, Lesson 35);
$C[a,b]$:	continuous functions with $\|x\|_\infty = \sup \|x\|$.

In L^p, "$x = 0$" means $x = 0$ a.e. (Proposition 35.10) — signals equal a.e. are the same vector. All of these are Banach spaces; for L^1, L^2 this is the Riesz–Fischer theorem (MIRA 7A) and it is the payoff of Lesson 35:

Theorem 36.2 (Completeness is Lebesgue's gift)**.** *L^2 is complete. With the Riemann integral in place of Lebesgue's it would not be: the truncated-rational sequence of Example 35.1 is L^2-Cauchy with no Riemann-integrable limit.*

Proof (MIRA 7.20). Let (x_n) be L^2-Cauchy; a convergent subsequence suffices. Pass to one with $\sum_k \|x_k - x_{k-1}\|_2 < \infty$ (take $x_0 = 0$; pick indices where the Cauchy gap drops below 2^{-k}). The increasing functions $g_m = \sum_{k\le m} |x_k - x_{k-1}|$ obey $\|g_m\|_2 \le \sum_k \|x_k - x_{k-1}\|_2$ (triangle inequality), so MCT (Theorem 35.11, applied to $g_m^2 \uparrow g^2$) puts the limit g in L^2; in particular $g < \infty$ a.e., so the telescoping series $\sum_k (x_k - x_{k-1})$ converges absolutely a.e. to some x with $|x| \le g$. Then $|x_k - x|^2 \le (2g)^2 \in L^1$, and DCT (Theorem 35.15) gives $\|x_k - x\|_2 \to 0$. Every tool is Lesson 35's: completeness is precisely what MCT and DCT buy. □

Uniform convergence (Lesson 31) is convergence in $\|\cdot\|_\infty$; "energy convergence" is convergence in $\|\cdot\|_2$. They are genuinely different topologies on signal space, and Gibbs is the statement that partial Fourier sums converge in the second but not the first.

Definition 36.3 (Hilbert space)**.** A *Hilbert space* H is an inner product space (Lesson 2's axioms, complex form as in Lesson 8: $\langle x, y\rangle = \int x\,\bar{y}$ on L^2, $\sum x[n]\overline{y[n]}$ on ℓ^2) that is complete in the induced norm $\|x\| = \langle x, x\rangle^{1/2}$. So: ℓ^2 and L^2 are Hilbert spaces; they are the homes of DT and CT finite-energy signals. Cauchy–Schwarz (Theorem 2.5) holds verbatim — its Lesson 2 proof used only the axioms.

Which norms come from inner products? Exactly those satisfying the parallelogram law $\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$ (expand the inner products). The sup-norm fails it (Exercise 36.1), so $C[a,b]$ with $\|\cdot\|_\infty$ and ℓ^∞ are Banach but not Hilbert: no orthogonality, no projections, no Fourier geometry. The 2-norms are not a convention; they are the only norms with geometry.

Theorem 36.4 (Projection theorem — Lesson 6, infinite-dimensional)**.** *Let M be a* closed *subspace of a Hilbert space H and $x \in H$. Then there is a unique $\hat{x} \in M$ closest to x, and it is characterized by orthogonality: $x - \hat{x} \perp M$.*

Proof. Let $d = \inf_{m\in M} \|x - m\|$ and pick $m_n \in M$ with $\|x - m_n\| \to d$. The parallelogram law, applied to $u = x - m_n$, $v = x - m_m$, gives

$$\|m_n - m_m\|^2 = 2\|x - m_n\|^2 + 2\|x - m_m\|^2 - 4\left\|x - \tfrac{m_n + m_m}{2}\right\|^2 \le 2\|x - m_n\|^2 + 2\|x - m_m\|^2 - 4d^2 \to 0,$$

using $\frac{m_n+m_m}{2} \in M$. So (m_n) is Cauchy; by completeness *of* H it converges, and because M is *closed* the limit $\hat{x}$ lies in M, with $\|x-\hat{x}\| = d$. Orthogonality and uniqueness are then Lesson 6's computation (Theorem 6.3) unchanged: for $m \in M$, $\|x-\hat{x}+tm\|^2$ is minimized at $t=0$, forcing $\langle x-\hat{x},\, m\rangle = 0$. □

Every hypothesis earned its keep: completeness produced the limit, closedness kept it in M. This theorem is the deepest single fact in the lesson — least squares (Theorem 6.4), LLMS estimation (Theorem 19.3), and Fourier truncation are all instances.

Theorem 36.5 (Orthonormal sequences: Bessel, then Parseval)**.** *Let $(e_k)_{k\in\mathbb{Z}}$ be orthonormal in a Hilbert space H and $x \in H$, with "Fourier coefficients" $a_k = \langle x,\, e_k\rangle$.*

(a) (Bessel) $\sum_k |a_k|^2 \le \|x\|^2$; in particular the coefficient sequence lies in ℓ^2.

(b) The partial sums $S_N = \sum_{|k|\le N} a_k e_k$ are the orthogonal projections of x onto the span of $e_{-N},\dots,e_N$, and $\sum_k a_k e_k$ always converges in H.

(c) The following are equivalent: (i) $\operatorname{span}(e_k)$ *is dense in H (the system is* complete*); (ii) $x = \sum_k a_k e_k$ for every x; (iii) Parseval, $\|x\|^2 = \sum_k |a_k|^2$, holds for every x.*

Proof. (b) first: $x - S_N \perp e_l$ for $|l| \le N$ by direct computation, so S_N is the projection (Theorem 36.4's characterization), and Pythagoras gives $\|x\|^2 = \|S_N\|^2 + \|x-S_N\|^2 = \sum_{|k|\le N}|a_k|^2 + \|x-S_N\|^2$. Letting $N\to\infty$ yields (a). Convergence of $\sum a_k e_k$: its partial sums are Cauchy because $\|\sum_{m\le|k|\le n} a_k e_k\|^2 = \sum_{m\le|k|\le n}|a_k|^2$ is a tail of the convergent series in (a) — completeness of H finishes (this is where Theorem 36.2 works for a living). (c): (i)⇒(ii): the limit $x' = \sum a_k e_k$ satisfies $x - x' \perp e_k$ for all k, hence $x - x' \perp$ a dense subspace, hence $x - x' \perp$ its closure $= H$, hence $x = x'$ (take the inner product with $x - x'$ itself). (ii)⇒(iii): continuity of the norm in the Pythagoras display above. (iii)⇒(i): if the span were not dense, Theorem 36.4 would produce $x \neq 0$ orthogonal to every e_k, violating Parseval. □

Theorem 36.6 (The harmonics are complete in $L^2[0,T]$)**.** *The normalized harmonics $e_k(t) = \frac{1}{\sqrt{T}} e^{jk\omega_0 t}$, $\omega_0 = 2\pi/T$, form a complete orthonormal system in $L^2[0,T]$. Consequently, for* every *finite-energy T-periodic signal — no differentiability, no Dirichlet conditions, jumps and all —*

$$\Big\| x - \sum_{|k|\le N} a_k e^{jk\omega_0 t} \Big\|_2 \longrightarrow 0 \qquad \textit{and} \qquad \frac{1}{T}\int_T |x|^2 = \sum_k |a_k|^2.$$

Proof sketch. Orthonormality is Lemma 26.2. Completeness reduces, by Theorem 36.5(c), to density of trigonometric polynomials in $L^2[0,T]$, which is proved in two standard steps: continuous functions are dense in L^2 (MIRA 3B, via simple and step functions), and trigonometric polynomials approximate continuous periodic functions uniformly (Fejér's theorem: the *averaged* partial sums $\frac{1}{N}\sum_{n<N} S_n$ use a nonnegative kernel and converge uniformly — Strichartz §1.2 or MIRA 11A — plotting the averaged partial sums shows exactly this Cesàro rescue of Gibbs). □

This upgrades Theorem 26.3 to its final form, and settles in what sense Fourier series "work for everything": in the energy norm, always; uniformly, only with smoothness (Corollary 31.14); pointwise a.e., true for L^2 but genuinely hard (Carleson 1966 — know the name, skip the proof). For the line instead of the circle, the same machinery yields:

Theorem 36.7 (Plancherel)**.** *The Fourier transform, defined on $L^1 \cap L^2$ and extended by continuity, is a bijection of $L^2(\mathbb{R})$ with $\int |x(t)|^2 dt = \frac{1}{2\pi}\int |X(j\omega)|^2 d\omega$ and $\langle x,\, y\rangle = \frac{1}{2\pi}\langle X,\, Y\rangle$. (MIRA ch. 11 / Strichartz §3.3; the DT sibling: the DTFT is, up to $\frac{1}{2\pi}$, the inverse of a Fourier* series*, so Theorem 36.6* is *its Parseval — Exercise 28.3 certified.)*

So the transform is a unitary change of basis of signal space — Lesson 11's Theorem 11.4, escaped from dimension N to dimension ∞. That was Part II's promise; this is the theorem that keeps it.

Definition 36.8 (Bounded operators)**.** A linear map $A : H \to H$ is *bounded* if $\|A\| := \sup_{x\neq 0}\|Ax\|/\|x\| < \infty$ — the operator norm, the infinite-dimensional largest singular value (Lesson 12). Bounded = continuous for linear maps (same ε–δ bookkeeping as always).

Theorem 36.9 (Stable LTI filters are bounded operators on ℓ^2)**.** *If $h \in \ell^1$, then $x \mapsto h * x$ maps ℓ^2 into ℓ^2 with*

$$\|h * x\|_2 \;\le\; \Big(\max_{\omega} |H(e^{j\omega})|\Big)\, \|x\|_2 \;\le\; \|h\|_1\, \|x\|_2.$$

The operator norm is exactly $\max_\omega |H(e^{j\omega})|$.

Proof. By Plancherel (DT form) and the convolution property (Theorem 28.3):

$$\|h * x\|_2^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 |X(e^{j\omega})|^2 d\omega \le \big(\max|H|\big)^2 \frac{1}{2\pi}\int_{-\pi}^{\pi} |X|^2 = \big(\max|H|\big)^2 \|x\|_2^2,$$

and $\max|H| \le \sum |h[k]|$ by the triangle inequality. Sharpness: concentrate $|X|^2$ near the maximizing frequency. □

In frequency domain the filter is the *multiplication operator* $X \mapsto H{\cdot}X$: a "diagonal matrix" indexed by ω, with the transform as the diagonalizing unitary. That is Lesson 11's circulant story (Theorem 11.6), now exact in infinite dimensions — and it is the template for graduate DSP, where filters, decimators, channel models, and estimators are all operators on ℓ^2, compared by operator norm, and "design" means shaping $H(e^{j\omega})$ under constraints.

Statistical-inference connection. Random variables with finite variance form a Hilbert space, $L^2(\Omega, \mathbf{P})$, with $\langle X,\, Y\rangle = \mathbf{E}[X\bar{Y}]$ — Lesson 18 flagged that covariance behaves like an inner product (Theorem 18.3); it *is* one. The orthogonality principle of LLMS estimation (Theorem 19.3) is Theorem 36.4 verbatim, with M the span of the observations; conditional expectation $\mathbf{E}[X \mid Y]$ is the projection onto the (closed) subspace of all functions of Y. Wiener filters and Kalman filters — statistical inference's second half — are computed projections in this Hilbert space, nothing more. One theorem, learned once, running the whole estimation theory.

36.2 Practice: by hand

Example 36.10 (Best approximation, computed like Lesson 6)**.** Project $x(t) = t$ on $[-\pi, \pi]$ onto $\operatorname{span}\{1, \cos t, \sin t\}$ in L^2. The three are orthogonal with $\|1\|^2 = 2\pi$, $\|\cos\|^2 = \|\sin\|^2 = \pi$. Coefficients: $\langle t,\, 1\rangle = 0$, $\langle t,\, \cos\rangle = 0$ (odd integrand), $\langle t,\, \sin\rangle = \int_{-\pi}^{\pi} t \sin t\, dt = 2\pi$. So $\hat{x}(t) = \frac{2\pi}{\pi}\sin t = 2\sin t$ — the first term of the sawtooth's Fourier series, of course: partial Fourier sums *are* these projections (Theorem 36.5(b)). Residual energy: Pythagoras, $\|x - \hat{x}\|^2 = \|x\|^2 - \|2\sin t\|^2 = \frac{2\pi^3}{3} - 4\pi \approx 8.11$.

Example 36.11 (Why closedness matters)**.** In ℓ^2, let M be the subspace of sequences with finitely many nonzero terms — dense, not closed. The vector $x[n] = 2^{-n}$ $(n \geq 0)$ has no closest point in M: truncations get arbitrarily close but the infimum 0 is not attained. Theorem 36.4 does not fail; its hypothesis does. (In graduate DSP this is not pedantry: "the projection onto the space of causal finite-energy signals" exists because that space is closed; "onto the FIR filters of unbounded order" does not.)

36.3 Exercises

Exercise 36.1 (Hand)**.** Show the parallelogram law fails for $\|\cdot\|_\infty$ on $C[0,1]$ (try $u(t) = 1$, $v(t) = t$) and for $\|\cdot\|_1$ on $\mathbb{R}^2$ (try $u = (1,0)$, $v = (0,1)$). Conclude neither norm admits an inner product — so "project onto the closest point" is not even well-posed there (find two closest points to $(1,1)$ in $\operatorname{span}\{(1,0)\}$ under $\|\cdot\|_\infty$... there are many).

Exercise 36.2 (Hand)**.** In $L^2[0,1]$, apply Gram–Schmidt (Lesson 6's algorithm, verbatim, with $\langle f, g\rangle = \int_0^1 f\bar{g}$) to $1, t, t^2$. You will produce the first three (shifted) Legendre polynomials. Then compute the best line approximating e^t in energy, and compare its slope with the calculus tangent at any point — projection is a global fit, not a local one.

Exercise 36.3 (Proof, $\star$)**.** Prove that in the Pythagoras display of Theorem 36.5's proof, the error $\|x - S_N\|$ is *decreasing* in N, and deduce the "best approximation improves monotonically" claim that Lesson 26 made after Theorem 26.3. Then prove: if Parseval holds for x and y separately, the polarization identity forces $\langle x, y\rangle = \sum_k a_k \bar{b}_k$ (Parseval for inner products).

Exercise 36.4 (Proof)**.** Prove that the DT ideal lowpass operator (multiplication by $\mathbf{1}_{|\omega| \leq \omega_c}$ in frequency) is a bounded operator on ℓ^2 of norm 1, even though Lesson 35 showed its impulse response is not in ℓ^1 and Theorem 25.3 showed it is not BIBO stable. Moral: ℓ^2-boundedness and BIBO (ℓ^∞) stability are different norms on the same operator — graduate DSP tracks both.

Deeper reading. MIRA 6A–6B (Banach), 7A (completeness of L^p), 8A–8B (Hilbert spaces, orthonormal bases); Bowers–Kalton ch. 1 (normed spaces) and 7.1–7.2 (Hilbert space theory, operators). For Fejér and Carleson context: Strichartz §1.2.

37 Complex Analysis I: Analytic Functions, Power Series, and Contour Integrals

NEEDED FOR: Optional rigor layer: the ground floor under poles, residues, and inversion integrals (next lesson) — what "analytic" actually buys and why closed-curve integrals vanish. Not needed for ML.

37.1 Theory

Lesson 8 built the complex *numbers*; this lesson studies *functions* of $z = x + jy$, and the next one cashes the theory in for poles, residues, and ROCs. The plane $\mathbb{C}$ is the metric space $\mathbb{R}^2$ (Lesson 32): open discs $D(\alpha; r) = \{z : |z - \alpha| < r\}$, open sets, and *regions* (open and connected) are already defined, and Heine–Borel, compactness, and uniform convergence transfer unchanged (Bak–Newman ch. 1).

Definition 37.1 (Complex differentiability; analyticity)**.** f is *differentiable* at z if $f'(z) = \lim_{h\to 0} \frac{f(z+h)-f(z)}{h}$ exists — with $h \in \mathbb{C}$, so the limit must agree along *every* approach direction. f is *analytic* at z if it is differentiable on some open set containing z; analytic on all of $\mathbb{C}$ = *entire*. The sum, product, quotient (where defined), chain, and inverse rules all hold with the proofs of Lesson 33 verbatim — the algebra never noticed that h was real (Bak–Newman 2.4, 3.5).

Proposition 37.2 (Cauchy–Riemann)**.** *Write $f = u + jv$. If $f'(z)$ exists then, comparing the real-axis and imaginary-axis approaches, $f_y = jf_x$, i.e.*

$$u_x = v_y, \qquad u_y = -v_x \qquad \textit{(Bak–Newman 3.1).}$$

Conversely, if f_x, f_y exist near z, are continuous at z, and satisfy CR there, then f is differentiable at z — the proof is two applications of the real MVT (Theorem 33.5) to u and v (Bak–Newman 3.2).

CR is a genuine constraint, not bookkeeping: $f(z) = \bar{z}$ has $u = x, v = -y$, so $u_x = 1 \neq -1 = v_y$ — differentiable *nowhere*, though perfectly smooth as an $\mathbb{R}^2$ map. A complex derivative demands that the local linearization be multiplication by the complex number $f'(z)$: a rotation-plus-scaling, with no reflection and no shear. That single geometric demand is where all the coming miracles are hiding.

Theorem 37.3 (Power series: radius, differentiability, coefficients)**.** *For $\sum_{k\geq 0} C_k z^k$, let $L = \limsup_k |C_k|^{1/k}$ and $R = 1/L$ (with $R = \infty$ for $L = 0$, $R = 0$ for $L = \infty$). Then the series converges absolutely for $|z| < R$, diverges for $|z| > R$ (Bak–Newman 2.8), and on $|z| < R$ its sum f:*

(a) is differentiable, with $f'(z) = \sum_{k\geq 1} kC_k z^{k-1}$, again with radius R (2.9) — hence infinitely differentiable (2.10);

(b) determines its coefficients: $C_k = f^{(k)}(0)/k!$ (2.11), so two power series agreeing near 0 agree coefficient by coefficient.

The convergence is uniform on every closed sub-disc $|z| \leq r < R$ (M-test, Theorem 31.13, with $M_k = |C_k| r^k$) — which is what licenses term-by-term operations below.

$e^z = \sum z^k/k!$, $\sin z$, $\cos z$ are entire ($L = 0$); the geometric series $\sum z^k = \frac{1}{1-z}$ has $R = 1$ — and note *why*: there is a singularity on the boundary. That "radius of convergence = distance to the nearest trouble" slogan becomes a theorem in the next lesson, and it is the entire reason z-transform ROCs are annuli bounded by poles.

Definition 37.4 (Contour integrals; the ML formula)**.** For a smooth curve $C : z(t)$, $a \leq t \leq b$, and continuous f,

$$\int_C f\,dz = \int_a^b f(z(t))\,\dot{z}(t)\,dt.$$

The value is invariant under orientation-preserving reparametrization, flips sign on $-C$, and obeys the *ML formula* $\left|\int_C f\,dz\right| \leq \max_C |f| \cdot \text{length}(C)$ (Bak–Newman 4.5–4.10) — the triangle inequality (Theorem 34.4) plus arc length. If $f_n \to f$ uniformly on C, then $\int_C f_n \to \int_C f$ (4.11). And if $f = F'$ for some analytic F, the complex chain rule gives the FTC verbatim (4.12):

$$\int_C f\,dz = F(z(b)) - F(z(a)) \quad \Longrightarrow \quad \oint_C f\,dz = 0 \ \text{ for closed } C.$$

So $\oint_C z^k\,dz = 0$ for every $k \geq 0$ and every closed curve: z^k has the antiderivative $z^{k+1}/(k+1)$. The question with everything downstream at stake: does *every* analytic f have an antiderivative? Term-by-term, a power series does — and that is the honest core of:

Theorem 37.5 (Closed curve theorem)**.** *(a) (Rectangle Theorem, Bak–Newman 4.14) If f is entire and Γ bounds a rectangle, $\oint_\Gamma f\,dz = 0$.* Proof mechanism: *bisect the rectangle into four; the integral is the sum of the four sub-integrals, so one sub-rectangle carries at least a quarter of $|I|$; iterate and zoom onto a point z_0 (Lesson 32's nested-compact-sets theorem). Near z_0, $f(z) = f(z_0) + f'(z_0)(z - z_0) + o(|z - z_0|)$; the linear part integrates to zero exactly (it has an antiderivative), and the $o(\cdot)$ part is killed by ML on ever-smaller rectangles: perimeter$^2 \cdot o(1) \to 0$ beats the 4^n bookkeeping. (b) (Integral and Closed Curve Theorems, 4.15–4.16) Hence an entire f has an entire antiderivative $F(z) = \int_0^z f$ (integrate along an axis-parallel path; the Rectangle Theorem makes the value path-independent), and therefore $\oint_C f\,dz = 0$ for* every *closed curve C. The same argument runs verbatim in any open disc (Bak–Newman 6.1–6.3), which is the local version used from now on.*

One point deserves emphasis: differentiability at each point — a purely *local* hypothesis — has produced a *global* conclusion about every closed curve. That promotion of local to global is the signature of complex analysis, and the bisection in (a) is the same compactness engine as Heine–Borel's.

Theorem 37.6 (Cauchy integral formula; Taylor representation)**.** *Let f be analytic in $D(\alpha; r)$, let C_ρ be a counterclockwise circle in the disc enclosing a. Then*

$$f(a) = \frac{1}{2\pi j} \oint_{C_\rho} \frac{f(z)}{z - a}\,dz \qquad \textit{(Bak–Newman 5.3, 6.4)},$$

and f equals a power series $\sum_k C_k(z - \alpha)^k$ convergent on the largest *disc centered at α inside the domain, with $C_k = \frac{1}{2\pi j} \oint \frac{f(\omega)}{(\omega-\alpha)^{k+1}} d\omega = f^{(k)}(\alpha)/k!$ (5.5, 6.5–6.6).*

Proof mechanism. Apply the closed curve theorem to the difference quotient $g(z) = \frac{f(z)-f(a)}{z-a}$ (continuous at a, and the Rectangle Theorem survives that one delicate point — Bak–Newman 5.1): $\oint g = 0$ turns $\oint \frac{f(z)}{z-a} dz$ into $f(a) \oint \frac{dz}{z-a}$, and the direct computation on a circle gives $\oint \frac{dz}{z-a} = 2\pi j$ (5.4) — *the* integral of the subject, the only one ever evaluated by hand. For Taylor: expand $\frac{1}{\omega - z}$ as a geometric series in z/ω, uniformly convergent on the circle, and swap $\oint \sum = \sum \oint$ (Definition 37.4's uniform-limit rule). □

Corollary 37.7 (What analyticity buys)**.** *Let f be analytic on a region D.*

(a) (Infinitely differentiable, Bak–Newman 6.8) f', f'', ... all exist on D: one complex derivative implies all of them — fail-safe versions of everything Lesson 33 had to hypothesize.

(b) (Uniqueness theorem, 6.9) If $f(z_n) = 0$ on a sequence of distinct points with a limit inside *D, then $f \equiv 0$ on D. Hence two analytic functions agreeing on an arc, a segment, or any set with a limit point agree everywhere — identities proved on the real line (like $e^{z+w} = e^z e^w$) hold on all of $\mathbb{C}$ for free.*

(c) (Isolated zeros, of finite order) If $f \not\equiv 0$ and $f(z_0) = 0$, the Taylor series at z_0 has a first nonzero coefficient, say at order m: $f(z) = (z - z_0)^m h(z)$ with h analytic, $h(z_0) \neq 0$. The integer m is the order *of the zero — finite, and the zero is isolated. Poles will be exactly this picture for $1/f$, and "pole–zero plots" owe their existence to (b) and (c): zeros and poles are isolated dots of definite order, never smears.*

37.2 Practice: by hand

Example 37.8 (The one integral, and its failure mode)**.** On the unit circle $C : e^{j\theta}$, $0 \le \theta \le 2\pi$:

$$\oint_C z\,dz = \int_0^{2\pi} e^{j\theta}\, je^{j\theta} d\theta = j\int_0^{2\pi} e^{2j\theta} d\theta = 0, \qquad \oint_C \frac{dz}{z} = \int_0^{2\pi} \frac{je^{j\theta}}{e^{j\theta}} d\theta = 2\pi j,$$

$$\oint_C \bar{z}\,dz = \int_0^{2\pi} e^{-j\theta} je^{j\theta} d\theta = 2\pi j.$$

First: analytic with antiderivative $\Rightarrow$ zero. Second: analytic *except one point inside* $\Rightarrow 2\pi j$, the fingerprint the next lesson turns into residues. Third: $\bar{z}$ is not analytic anywhere, and its closed-curve integral has no obligation to vanish. Three integrals, the whole subject in miniature.

Example 37.9 (Radius-of-convergence drill)**.** (a) $\sum k^2 z^k$: $|C_k|^{1/k} = k^{2/k} \to 1$, $R = 1$. (b) $\sum z^k/k!$: $(1/k!)^{1/k} \to 0$, $R = \infty$. (c) $\sum 2^k z^{2k}$ (lacunary): $\limsup |C_k|^{1/k} = \sqrt{2}$, $R = 1/\sqrt{2}$ — lim sup, not lim, is doing the work. (d) The geometric expansion of $\frac{1}{z-1}$ around $\alpha = 2$: $\frac{1}{1+(z-2)} = \sum_k (-1)^k (z-2)^k$, $R = 1$ — exactly the distance from 2 to the singularity at 1 (Bak–Newman ch. 6 example). ✓

> **Signals connection.** Transforms $X(s)$, $X(z)$ are analytic functions of a complex variable on their ROCs (proved next lesson), so everything here applies verbatim: their Taylor expansions are the series Part V manipulated; the uniqueness theorem is why a transform identity checked on the $j\omega$-axis holds on the whole ROC (the axis has limit points!); isolated zeros of definite order are why pole–zero cancellation, minimum-phase factorization, and root loci are discrete bookkeeping. The formula $\oint \frac{dz}{z-a} = 2\pi j$ computed above is, one lesson from now, the inverse z-transform.

37.3 Exercises

Exercise 37.1 (Hand)**.** Using Cauchy–Riemann: (a) show $f(z) = z^2 = x^2 - y^2 + j\,2xy$ satisfies CR everywhere and read off $f'(z) = 2z$ from f_x; (b) show $f(z) = |z|^2 = x^2 + y^2$ satisfies CR at $z = 0$ only — differentiable at one point, analytic nowhere; (c) decide where $f(z) = \mathrm{Re}\, z$ is differentiable.

Exercise 37.2 (Hand)**.** Compute from the definition: (a) $\oint_C (z-a)^k\,dz$ on the circle $C : a + \rho e^{j\theta}$ for $k = -2, -1, 0, 1$ (which one is nonzero, and what is its value?); (b) the radius of convergence of $\sum_k \frac{z^k}{k}$, of $\sum_k k!\, z^k$, and of the Taylor series of $\frac{1}{1+z^2}$ around $\alpha = 0$ and around $\alpha = 1$ (distance to which singularities?).

Exercise 37.3 (Proof)**.** Prove Proposition 37.2 (necessity): compute the difference quotient along h real and $h = j\eta$, equate, and derive both CR equations. Then verify that under CR the Jacobian of $(x, y) \mapsto (u, v)$ is $\left(\begin{smallmatrix} u_x & -v_x \\ v_x & u_x \end{smallmatrix}\right)$ — a rotation-scaling with determinant $|f'(z)|^2 \ge 0$: analytic maps preserve orientation and angles wherever $f' \neq 0$.

Exercise 37.4 (Proof, $\star$)**.** Prove the uniqueness theorem (Corollary 37.7(b)) from the Taylor representation: first show that if all derivatives of f vanish at z_0 then $f \equiv 0$ on a disc; then run the connectedness argument (the set where all derivatives vanish is open, its complement is open, D is connected — Lesson 32). Conclude the isolated-zeros factorization $f(z) = (z - z_0)^m h(z)$ of Corollary 37.7(c).

Deeper reading. Bak–Newman chs. 1–3 (the plane, power series, analyticity and CR), ch. 4 (line integrals, Rectangle and Closed Curve Theorems), chs. 5–6 (Cauchy Integral Formula, Taylor representation, uniqueness, maximum-modulus — read the max-modulus section too; it is two pages and worth it).

38 Complex Analysis II: Zeros, Poles, Residues, and Why ROCs Work

NEEDED FOR: Optional rigor layer for Fourier analysis (inversion, Gaussians) and the s/z-plane; standard equipment for graduate DSP. Not needed for ML.

38.1 Theory

Part V's transform tables, ROC rules, and partial-fraction inversions all rest on the behavior of functions of a *complex* variable. Lesson 37 built the machine — analyticity, power series, Cauchy's theorem and formula; this lesson points it at singularities, which is where transforms actually live.

Definition 38.1 (Holomorphic functions, recalled)**.** f is *holomorphic* (equivalently *analytic*) on an open set $\Omega \subseteq \mathbb{C}$ if $f'(z) = \lim_{h\to 0} \frac{f(z+h)-f(z)}{h}$ exists for every $z \in \Omega$, with h approaching 0 from *every direction in* $\mathbb{C}$ — Definition 37.1, tested in practice by Cauchy–Riemann (Proposition 37.2). Recall the verdicts: polynomials, e^z, $\sin z$, $\cos z$ are entire (holomorphic on $\mathbb{C}$); rational functions are holomorphic off their poles; $\bar{z}$, $|z|^2$, $\operatorname{Re} z$ are holomorphic nowhere — "anti-rotating" behavior fails the directional limit.

Definition 38.2 (Contour integrals; the ML bound)**.** For a curve $\gamma : [a, b] \to \mathbb{C}$ and continuous f, $\int_\gamma f\,dz = \int_a^b f(\gamma(t))\gamma'(t)\,dt$. The one estimate used everywhere (*ML bound*): $\left|\int_\gamma f\,dz\right| \le \max_\gamma |f| \cdot \text{length}(\gamma)$ — the triangle inequality for integrals, nothing more.

Theorem 38.3 (Cauchy's theorem and integral formula)**.** *Let f be holomorphic on and inside a simple closed curve γ (traversed counterclockwise). Then*

$$\oint_\gamma f\,dz = 0, \qquad \text{and} \qquad f(z_0) = \frac{1}{2\pi j}\oint_\gamma \frac{f(z)}{z - z_0}\,dz \quad \text{for every } z_0 \text{ inside } \gamma.$$

Proof. Lesson 37 proved both: the vanishing integral is Theorem 37.5 (Goursat's bisection plus the local antiderivative), and the formula is Theorem 37.6, riding on the one hand-computed integral $\oint \frac{dz}{z-z_0} = 2\pi j$. □

The formula says a holomorphic function is determined inside a curve by its boundary values — and differentiating it under the integral sign (legal by Lesson 35!) gives $f^{(n)}(z_0) = \frac{n!}{2\pi j}\oint \frac{f(z)}{(z-z_0)^{n+1}}dz$ for every n: *one* complex derivative buys infinitely many, plus a convergent Taylor series in the largest disk avoiding singularities (Bak–Newman ch. 6). Two famous one-liners follow:

Theorem 38.4 (Liouville; Fundamental Theorem of Algebra)**.** *(a) A bounded entire function is constant: the derivative formula on a radius-R circle gives $|f'(z_0)| \le \max|f|/R \to 0$. (b) Every nonconstant polynomial p has a root in $\mathbb{C}$: else $1/p$ would be entire and bounded (it $\to 0$ at ∞), hence constant. Dividing out roots: p factors completely into linear factors.*

Part (b) is why *every* rational transform in Lesson 30 splits into partial fractions over $\mathbb{C}$ — the algebra that Part V used from a table is a consequence of complex analysis, not an accident of examples.

Isolated singularities, classified. Suppose f is analytic in a deleted neighborhood $0 < |z - z_0| < d$ but not at z_0 (Bak–Newman ch. 9). Exactly three things can happen:

1. *Removable*: some g analytic at z_0 agrees with f off z_0 — the singularity is a bookkeeping accident, as for $\frac{\sin z}{z}$ at 0. *Riemann's principle* detects it: if $\lim_{z\to z_0}(z - z_0)f(z) = 0$ — in particular whenever f is *bounded* near z_0 — the singularity is removable (Bak–Newman 9.3–9.4).

2. *Pole of order k*: $(z - z_0)^k f(z)$ has a finite nonzero limit; equivalently $f(z) = h(z)/(z - z_0)^k$ with h analytic, $h(z_0) \neq 0$ (9.5). This is the mirror image of Corollary 37.7(c): poles of f are zeros of $1/f$, of the same order, so pole–zero plots inherit the isolated-dots-of-definite-order geometry from the uniqueness theorem. Near a pole, $|f(z)| \to \infty$.

3. *Essential*: neither — and then the function goes berserk (Casorati–Weierstrass, 9.6): in *every* deleted neighborhood of z_0, f comes arbitrarily close to *every* complex value, as $e^{1/z}$ does at 0. No rational transform of Part V does this; every singularity of a rational $X(s)$ or $X(z)$ is a pole, which is why LTI bookkeeping never needs more than pole–zero arithmetic.

Definition 38.5 (Laurent series, poles, residues)**.** On an annulus $r < |z - z_0| < R$ where f is holomorphic, f has a unique expansion $f(z) = \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n$ (Bak–Newman ch. 9). If only finitely many negative powers appear, z_0 is a *pole* of order m (simple if $m = 1$); the coefficient c_{-1} is the *residue* $\operatorname{Res}_{z_0} f$. For a simple pole, $\operatorname{Res}_{z_0} f = \lim_{z\to z_0}(z - z_0)f(z)$; if $f = p/q$ with $q(z_0) = 0 \neq q'(z_0)$, this is $p(z_0)/q'(z_0)$; for an order-m pole, $\operatorname{Res}_{z_0} f = \frac{1}{(m-1)!} \lim_{z\to z_0} \frac{d^{m-1}}{dz^{m-1}} \big[(z - z_0)^m f(z)\big]$.

Theorem 38.6 (Residue theorem)**.** *If f is holomorphic on and inside the simple closed curve γ except at finitely many points $z_1, \ldots, z_k$ inside, then*

$$\oint_\gamma f\, dz = 2\pi j \sum_{i=1}^{k} \operatorname{Res}_{z_i} f.$$

Proof sketch. Shrink γ to tiny circles around each z_i (Cauchy's theorem on the Swiss-cheese region between); on each circle integrate the Laurent series term by term (uniform convergence on the circle, Theorem 31.11): every power $(z - z_i)^n$ integrates to 0 except $n = -1$, which gives $2\pi j\, c_{-1}$ — the $\oint \frac{dz}{z-z_0} = 2\pi j$ computation from Theorem 38.3, which is thus the only integral anyone ever actually evaluates in this subject. □

Example 38.7 (A transform pair, certified by contour)**.** Claim (Lesson 27's two-sided decay pair, read backwards):

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2a}{a^2 + \omega^2}\, e^{j\omega t}\, d\omega = e^{-a|t|}, \qquad a > 0.$$

Take $t > 0$. Close the contour with a large upper semicircle $|\omega| = R$: on it, $|e^{j\omega t}| = e^{-t \operatorname{Im}\omega} \leq 1$ and the rational factor is $O(1/R^2)$, so by ML the semicircle contributes $O(1/R) \to 0$ (for slower

$1/R$ decay one invokes *Jordan's lemma*, Bak–Newman ch. 11 — the $e^{j\omega t}$ decay rescues the estimate). The integrand's poles are at $\omega = \pm ja$; only $\omega = +ja$ is inside. Residue (p/q' rule): $\frac{2a\,e^{j(ja)t}}{2(ja)} = \frac{e^{-at}}{j}$. Hence the integral is $\frac{1}{2\pi} \cdot 2\pi j \cdot \frac{e^{-at}}{j} = e^{-at}$. For $t < 0$ close *downward* (that is where $e^{j\omega t}$ decays), pick up $\omega = -ja$ with a sign flip from orientation: e^{at}. Together: $e^{-a|t|}$. ✓ The "close the contour on the side where the exponential dies" rule is the entire mechanics of transform inversion.

Now the s- and z-planes make sense. Three facts, all direct consequences:

1. *Transforms are holomorphic on their ROCs.* Corollary 35.16(c) lets us differentiate $X(s) = \int x(t)e^{-st}dt$ in s, under the integral sign: X is holomorphic on the open ROC strip. Same for $X(z)$ on its annulus.

2. *The ROC is pole-free and pole-bounded.* Holomorphy fails exactly at poles, so the ROC cannot contain one; and for rational transforms the Taylor/Laurent expansion converges precisely up to the nearest singularity, so the ROC of Proposition 30.3 extends exactly to the nearest pole line/circle — Part V's "ROC bounded by poles" rule, now a theorem about radii of convergence.

3. *The z-transform* is *a Laurent series* — $X(z) = \sum_n x[n]z^{-n}$, in the variable z^{-1} — and Laurent coefficients on a given annulus are *unique*, recoverable by the contour integral $x[n] = \frac{1}{2\pi j}\oint_\gamma X(z)z^{n-1}dz$ (γ any circle in the ROC). That is why "$X(z)$ *plus its ROC*" determines $x[n]$ (Example 30.2's moral), why the same formula with different annuli gives different signals, and why the inversion tables of Lesson 30 have unique answers. The Laplace inversion (Bromwich) integral $x(t) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} X(s)e^{st}ds$ is the same statement on a vertical line in the ROC, evaluated exactly as in Example 38.7: close left for $t > 0$ (picking up the poles left of the ROC → causal modes $e^{p_i t}$), close right for $t < 0$.

Pole locations are therefore not bookkeeping; they are the singularity structure of a holomorphic function, and "causal and stable iff all poles in the open left half-plane / unit disk with the ROC outside them" is complex analysis wearing an engineering hat.

38.2 Practice: by hand

Example 38.8 (z-inversion by Laurent expansion, both ROCs)**.** $X(z) = \dfrac{1}{1-az^{-1}}$. On $|z| > |a|$: expand in powers of az^{-1} (geometric, converges there): $X(z) = \sum_{n\geq 0} a^n z^{-n}$, so $x[n] = a^n u[n]$. On $|z| < |a|$: rewrite $X = \frac{-a^{-1}z}{1-a^{-1}z}$ and expand in powers of $a^{-1}z$: $X(z) = -\sum_{m\geq 1} a^{-m}z^m$, so $x[n] = -a^n u[-n-1]$. Two annuli, two Laurent series, two signals — Example 30.2 *re-derived* instead of table-matched. Check the second against the inversion contour for $n = -1$: $x[-1] = \frac{1}{2\pi j}\oint z^{-2}X(z)\,dz$ with $z^{-2}X(z) = \frac{z^{-1}}{z-a}$ (use $X(z) = \frac{z}{z-a}$); on a circle inside $|z| < |a|$ the only enclosed pole is $z = 0$, with residue $\frac{1}{0-a} = -a^{-1}$. ✓ ($x[-1] = -a^{-1}$.)

Example 38.9 (Residue drill)**.** (a) $\mathrm{Res}_{z=j}\frac{1}{z^2+1} = \frac{1}{2j}$ (p/q'). (b) $\mathrm{Res}_{s=-2}\frac{s+1}{(s+2)^2(s+3)}$: order-2 pole, $\frac{d}{ds}\frac{s+1}{s+3}\big|_{s=-2} = \frac{2}{(s+3)^2}\big|_{-2} = 2$. (c) $\oint_{|z|=2}\frac{e^z}{z(z-1)}dz = 2\pi j\big(\mathrm{Res}_0 + \mathrm{Res}_1\big) = 2\pi j\,(-1+e)$. Note (b) is exactly a repeated-root partial-fraction coefficient — the residue *is* the partial-fraction numerator, which is why the two techniques never disagree.

Fourier / grad DSP connection. Fourier analysis evaluates inversion integrals and the Gaussian's self-transform by exactly these contour moves (shift the path, invoke Cauchy). In graduate DSP: *minimum-phase* systems are those whose H and $1/H$ are both holomorphic on $|z| > 1$ — log-magnitude and phase then determine each other (Hilbert transform relations, from Cauchy's formula); *spectral factorization* (splitting a power spectrum into causal × anticausal, the engine of Wiener filtering) is an exercise in allocating poles and zeros between $|z| < 1$ and $|z| > 1$; and the Paley–Wiener theorems say "causal ⇔ holomorphic in a half-plane" in full generality. None of that is startable without this lesson.

38.3 Exercises

Exercise 38.1 (Hand)**.** Verify the Cauchy–Riemann equations for $f(z) = e^z = e^x(\cos y + j \sin y)$ and their failure for $f(z) = \bar{z}$. Then explain in one sentence why $|X(j\omega)|$ is never holomorphic in ω (moduli are real-valued), and why that is consistent with $X(s)$ being holomorphic in s.

Exercise 38.2 (Hand)**.** Compute by residues: (a) $\int_{-\infty}^{\infty} \frac{d\omega}{(\omega^2+1)(\omega^2+4)} = \frac{\pi}{6}$; (b) $\mathrm{Res}_{z=0} \frac{\sin z}{z^4}$ (Laurent: $= -\frac{1}{6}$); (c) the inverse Laplace transform of $\frac{1}{(s+1)(s+2)}$ with ROC $\mathrm{Re}\, s > -1$, by Bromwich-plus-residues, and check against partial fractions.

Exercise 38.3 (Proof, ⋆)**.** Prove Liouville's theorem from the derivative form of the Cauchy formula (ML bound on a radius-R circle, $R \to \infty$), then deduce the Fundamental Theorem of Algebra, then deduce that every strictly proper rational function is a finite sum of terms $\frac{c_{ik}}{(s-p_i)^k}$. (You have now proved that the partial-fraction method of Lesson 30 always works.)

Exercise 38.4 (Proof)**.** Let $x[n]$ be causal with $\sum |x[n]| r^{-n} < \infty$ for some $r < 1$ (exponentially stable). Show $X(z)$ is holomorphic on $|z| > r$, and that $\log X$ is holomorphic there too provided X has no zeros in $|z| > r$ — the minimum-phase condition. (Differentiate under the sum; Lesson 35 licenses it.)

Deeper reading. Bak–Newman ch. 9 (isolated singularities, Laurent series — the Casorati–Weierstrass proof is four elegant lines), chs. 10–11 (the residue theorem, real-integral applications, Jordan's lemma), ch. 12 (further contour techniques); chs. 1–6 belong to Lesson 37.

39 The Fourier, Laplace, and z-Transforms as Operators: The Property Toolbox, Proved

NEEDED FOR: Optional rigor layer: every entry of Part V's transform tables re-derived with its hypotheses visible, on the spaces $(\ell^1, \ell^2, L^1, L^2)$ where it is true. Standard equipment for graduate DSP. Not needed for ML.

39.1 Theory

Part V's Theorem 27.3 was a toolbox; this lesson is its certificate. The plan: put each transform on its proper space, view it as an *operator* (Definition 36.8), and prove the table entries by naming, each time, the analysis theorem that licenses the manipulation. Nothing new is computed — what is new is knowing *why* each line is true and *when* it fails.

Where each transform lives. Two regimes per transform, and the distinction does real work (Kreyszig §§1.2, 3.1; Lessons 35 and 36):

1. *Absolutely summable/integrable* ($x \in \ell^1$ or L^1): the defining sum or integral converges absolutely at every frequency, and $X(e^{j\omega}) = \sum_n x[n]e^{-j\omega n}$ converges *uniformly* in ω (M-test, Theorem 31.13, $M_n = |x[n]|$), so X is *continuous* (Theorem 31.9). This is the regime of pointwise formulas.

2. *Finite energy* ($x \in \ell^2$ or L^2): the defining integral may not converge pointwise at all; the transform is defined as the L^2-limit of truncations, and Plancherel (Theorem 36.7) says the extension $\mathcal{F} : L^2 \to L^2$ is (up to the constant $\frac{1}{2\pi}$) *unitary* — Lesson 11's "the DFT is a rotation of $\mathbb{C}^N$," with $N = \infty$. Energy identities and inversion-in-norm live here.

The two regimes overlap on $\ell^1 \cap \ell^2$ (dense in ℓ^2), which is how every L^2 statement below is proved: verify it on the absolutely convergent core, then extend by continuity of the operators involved — the standard density argument, used once and reused forever.

Three structural operators. Fix the shift $(S_k x)[n] = x[n-k]$, the modulation $(M_{\omega_0} x)[n] = e^{j\omega_0 n}x[n]$, and (continuous time) $(S_{t_0}x)(t) = x(t-t_0)$. Each is linear and preserves every ℓ^p/L^p norm — shifts reindex, modulations have unimodular factors — so each is unitary on ℓ^2/L^2: the cleanest possible bounded operators (Definition 36.8, $\|S_k\| = \|M_{\omega_0}\| = 1$).

Theorem 39.1 (The toolbox, proved — DTFT and CTFT)**.** *Let $x \in \ell^1(\mathbb{Z})$ (respectively $L^1(\mathbb{R})$); each identity then extends to ℓ^2/L^2 by density. With $X = \mathcal{F}x$:*

(a) Linearity. *Termwise, from linearity of sums and integrals (Theorem 34.4); absolute convergence is inherited by finite linear combinations (triangle inequality).*

(b) Time shift $\mathcal{F}(S_k x) = e^{-j\omega k}X$. *Discrete: reindex $m = n - k$ in $\sum_n x[n-k]e^{-j\omega n}$ — legal because the series converges* absolutely*, and Theorem 31.6 says absolutely convergent series may be reordered. Continuous: substitute $\tau = t - t_0$ (Theorem 34.7(b), extended to $\mathbb{R}$ by Lesson 35). Read as operators: $\mathcal{F}S_k = M_{-k}\mathcal{F}$ — the transform* intertwines *shift with modulation. Its mirror image $\mathcal{F}M_{\omega_0} = S_{\omega_0}\mathcal{F}$ is the modulation property, proved by the same one line.*

(c) Scaling $x(at) \leftrightarrow \frac{1}{|a|}X(j\frac{\omega}{a})$*: the substitution $u = at$, with $|a|$ from the orientation flip when $a < 0$ — the same change-of-variable theorem, nothing more.*

(d) Differentiation in time $x'(t) \leftrightarrow j\omega X(j\omega)$*, valid when x is absolutely continuous with $x' \in L^1$ and $x(t) \to 0$ as $|t| \to \infty$: integrate by parts (Theorem 34.7(a)) on $[-T, T]$,*

$$\int_{-T}^{T} x'(t)e^{-j\omega t}dt = \left[x(t)e^{-j\omega t}\right]_{-T}^{T} + j\omega \int_{-T}^{T} x(t)e^{-j\omega t}dt,$$

and let the boundary term die as $T \to \infty$. The decay hypothesis is load-bearing*: drop it and the boundary term survives — the honest source of every "$+x(0^-)$" term in the one-sided Laplace differentiation rule of Lesson 30.*

(e) Differentiation in frequency $-jt\,x(t) \leftrightarrow X'(j\omega)$*: differentiate under the integral sign, licensed by Corollary 35.16 (domination by $|t\,x(t)| \in L^1$) — the Lebesgue lesson earning its keep.*

(f) Convolution $\mathcal{F}(x * h) = X \cdot H$*: with* $x, h \in \ell^1$*,*

$$\sum_n \Big(\sum_k x[k]h[n-k] \Big) e^{-j\omega n} = \sum_k x[k]e^{-j\omega k} \sum_m h[m]e^{-j\omega m},$$

where the swap of the double sum is exactly Fubini–Tonelli for sums (Lesson 35) — available because $\sum_n \sum_k |x[k]||h[n-k]| = \|x\|_1 \|h\|_1 < \infty$*. This is Proposition 25.2's reindexing done with a license. The multiplication property is the mirror statement.*

(g) Parseval/Plancherel $\sum_n |x[n]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{j\omega})|^2 d\omega$*: this is Theorem 36.7, i.e. Bessel plus completeness of the harmonics (Theorem 36.6) — the operator statement being that* $\frac{1}{\sqrt{2\pi}}\mathcal{F}$ *is unitary, with adjoint equal to its inverse. In particular transforms preserve inner products:* $\langle x, y \rangle = \frac{1}{2\pi} \langle X, Y \rangle$ *— matched filtering may be done in either domain, at identical SNR.*

Proposition 39.2 (Laplace and z: same proofs, plus ROC bookkeeping)**.** *For* $X(z) = \sum_n x[n] z^{-n}$ *on ROC* $r_1 < |z| < r_2$ *and* $X(s) = \int x(t) e^{-st} dt$ *on its strip (Lesson 30):*

(a) Every identity above holds verbatim with $e^{j\omega} \to z$*,* $j\omega \to s$*, because each proof used only absolute convergence — which is the* definition *of the ROC. On the ROC the DTFT/CTFT computations run unchanged.*

(b) Shift: $x[n-k] \leftrightarrow z^{-k} X(z)$*, ROC unchanged except possibly at* $0/\infty$ *(the factor* z^{-k} *is analytic off* 0*).* Exponential weighting: $a^n x[n] \leftrightarrow X(z/a)$*, ROC scaled by* $|a|$*;* $e^{s_0 t} x(t) \leftrightarrow X(s - s_0)$*, strip translated by* $\operatorname{Re} s_0$ *— substitute and watch where absolute convergence now holds. This is why pole locations move exactly with modulation/damping in Lesson 30's tables.*

(c) Differentiation in z: $n\,x[n] \leftrightarrow -z\,X'(z)$ *— term-by-term differentiation of a Laurent series inside its annulus, which Theorem 37.3(a) (via Lesson 38) makes unconditionally legal on the open ROC. No extra hypotheses: analyticity of transforms on their ROC is doing the work.*

(d) Convolution: $x * h \leftrightarrow XH$ *on (at least) the intersection of the ROCs — Fubini as in (f) above, with the intersection exactly where both series converge absolutely. The "at least" is real: pole–zero cancellation can enlarge it (Example 30.2's fine print).*

Theorem 39.3 (Convolution operators: the norm is the peak of the response)**.** *Let* $h \in \ell^1$ *and* $Tx = x * h$ *on* ℓ^2*. Then* T *is bounded (Theorem 36.9) and in fact*

$$\|T\| = \max_{\omega} |H(e^{j\omega})|.$$

Proof idea. Upper bound: by Theorem 39.1(f)–(g), $\|Tx\|_2^2 = \frac{1}{2\pi} \int |H|^2 |X|^2 \le (\max |H|)^2 \|x\|_2^2$ — Plancherel converts convolution to multiplication, where the estimate is trivial. Sharpness: concentrate X in an ever-narrower band around the maximizing ω^* (the max is attained: H is continuous on the compact circle, Theorem 32.8(b)); then $\|Tx\|_2 / \|x\|_2 \to |H(e^{j\omega^*})|$. □

This single identity is the rigorous content of "the worst-case gain of a filter is the peak of its magnitude response" — the spec every anti-aliasing and loop-shaping argument in the labs quotes. Note the pattern of proof, because it is the lesson's moral in one line: *diagonalize by Fourier, estimate in the diagonal domain, return by unitarity.* The transform tables are not a list of coincidences; they are the statement that $\mathcal{F}$ simultaneously diagonalizes every shift-invariant operator, with the properties (b), (f) saying exactly that.

39.2 Practice: by hand

Example 39.4 (One entry, end to end, with the licenses shown)**.** Claim: for $x[n] = a^n u[n]$, $|a| < 1$, the delayed signal $y = S_3 x$ has $Y(e^{j\omega}) = e^{-3j\omega}\frac{1}{1-ae^{-j\omega}}$ and the same energy as x. (i) $x \in \ell^1$ (geometric), so the DTFT converges absolutely: $X = \sum_{n\geq 0} a^n e^{-j\omega n} = \frac{1}{1-ae^{-j\omega}}$ (Proposition 31.5(a) with $r = ae^{-j\omega}$, $|r| = |a| < 1$). (ii) Shift property, reindexing licensed by absolute convergence: $Y = e^{-3j\omega}X$. (iii) Energy: $|e^{-3j\omega}| = 1$, so $\int |Y|^2 = \int |X|^2$, and by Parseval both equal $2\pi \sum_n |x[n]|^2 = \frac{2\pi}{1-|a|^2}$. Every step names its theorem; nothing is folklore. ✓

Example 39.5 (Watching a hypothesis fail)**.** $x[n] = \frac{\sin(\omega_c n)}{\pi n}$ (the ideal low-pass response, Lesson 28) is in ℓ^2 but *not* ℓ^1. Concretely: $X(e^{j\omega})$ is the discontinuous box — so X cannot be a uniform limit of the continuous partial sums (Theorem 31.9), confirming $x \notin \ell^1$; the partial sums converge only in L^2, and they Gibbs-overshoot at the edges forever (Corollary 31.10). Every "truncate the ideal filter" design decision in the labs is a negotiation with exactly this failure.

Grad DSP / ML connection. This lesson is the working grammar of graduate DSP: multirate identities, adaptive-filter convergence, and spectral factorization all argue "diagonalize, estimate, return." The operator view feeds forward: Lesson 41 treats $\mathcal{F}$ and convolutions as the standing examples of adjoints, and Part VII's landscape analyses (Lesson 45) diagonalize regularized deconvolution by exactly Theorem 39.3's move. In ML, circular convolution layers are diagonalized by the DFT (Lesson 11), and the $\max|H|$ operator norm is the Lipschitz constant that spectral-normalization schemes control.

39.3 Exercises

Exercise 39.1 (Hand)**.** For $x(t) = e^{-at}u(t)$, $a > 0$: derive $X(j\omega) = \frac{1}{a+j\omega}$, then obtain the transforms of (a) $x(t-2)$, (b) $e^{j\omega_0 t}x(t)$, (c) $x(3t)$, (d) $x'(t)$ (careful: x jumps at 0 — which hypothesis of Theorem 39.1(d) fails, and what extra term appears?), naming the property used each time; check (d) against the direct computation of $\mathcal{F}[-ae^{-at}u(t)]$ plus the jump's contribution.

Exercise 39.2 (Hand)**.** Verify Parseval numerically-by-hand for $x[n] = \{1, 1\}$ (i.e. $x[0] = x[1] = 1$): compute $\sum |x[n]|^2$ directly, then $\frac{1}{2\pi}\int_{-\pi}^{\pi} |1 + e^{-j\omega}|^2 d\omega$ by expanding $|1 + e^{-j\omega}|^2 = 2 + 2\cos\omega$. Then find $\|T\|$ for the two-tap averager $h = \{\frac{1}{2}, \frac{1}{2}\}$ via Theorem 39.3, and exhibit the frequency that (in the limit) attains it.

Exercise 39.3 (Proof)**.** Prove the convolution theorem for ℓ^1 sequences in full (Theorem 39.1(f)): show first that $x, h \in \ell^1$ implies $x * h \in \ell^1$ with $\|x * h\|_1 \leq \|x\|_1 \|h\|_1$ (Tonelli for nonnegative double sums), then justify the swap for the transform itself. Conclude the z-domain version on $\mathrm{ROC}_x \cap \mathrm{ROC}_h$.

Exercise 39.4 (Proof, $\star$)**.** Prove Theorem 39.3's sharpness half: for continuous H on $[-\pi, \pi]$ attaining $\max|H|$ at ω^*, construct x_N with $X_N = \sqrt{\pi N}\,\mathbf{1}_{[\omega^* - \frac{1}{N}, \omega^* + \frac{1}{N}]}$ (unit energy, by Parseval), and show $\|Tx_N\|_2^2 \to |H(\omega^*)|^2$ using continuity of $|H|^2$ at ω^*. Explain in one sentence why no single unit-energy x attains the norm when $|H|$ peaks at a single frequency — and which function *would*, if only it had finite energy.

Deeper reading. Kreyszig §§2.6–2.7 (bounded operators), §§3.1–3.6 (Hilbert space, orthonormal series — the frame for Parseval); Bak–Newman chs. 2, 9 for the analyticity facts used in Proposition 39.2; and back through Lessons 27–30 with this lesson open beside them — every table entry there now has an owner here.

40 Distributions: Making δ Honest

NEEDED FOR: Optional rigor layer for Fourier analysis (which develops exactly this) and statistical inference (white noise); essential for graduate DSP. Not needed for ML.

40.1 Theory

Part V used $\delta(t)$ under a treaty ("defined operationally by sifting"), and paid for it in caveats: δ is not a function (no function is 0 a.e. yet integrates to 1 — Proposition 35.10 forbids it), the square wave was differentiated only by analogy, and the impulse-train identity of Lemma 29.1 was "certified elsewhere." Distribution theory is the elsewhere. The single idea: *stop asking for the value of a signal at a point; ask only for its averages against smooth test signals.* Every measurement in engineering is such an average — no instrument reads $x(t_0)$; it reads $\int x\varphi$ for some aperture φ.

Definition 40.1 (Test functions)**.** $\mathcal{D}$ is the space of infinitely differentiable functions $\varphi : \mathbb{R} \to \mathbb{C}$ that vanish outside a bounded interval. It is not obvious such functions exist; the standard bump

$$\varphi(t) = \begin{cases} e^{-1/(1-t^2)}, & |t| < 1 \\ 0, & |t| \geq 1 \end{cases}$$

is C^∞ (all derivatives of $e^{-1/s}$ vanish as $s \downarrow 0$ faster than any power — Ross §31's classic example), and shifts, scalings, and sums of it generate a rich supply. A sequence $\varphi_n \to \varphi$ *in* $\mathcal{D}$ if all supports sit in one common bounded interval and $\varphi_n^{(k)} \to \varphi^{(k)}$ uniformly for every k — the strongest convergence imaginable, which is the point: the stronger the convergence on test functions, the *weaker* (more inclusive) the dual notion defined next.

Definition 40.2 (Distributions)**.** A *distribution* (generalized function) is a linear functional $f : \mathcal{D} \to \mathbb{C}$, written $\langle f, \varphi\rangle$, that is continuous: $\varphi_n \to \varphi$ in $\mathcal{D}$ implies $\langle f, \varphi_n\rangle \to \langle f, \varphi\rangle$. The set of distributions is $\mathcal{D}'$.

- Every locally integrable signal x *is* a distribution via $\langle x, \varphi\rangle = \int x(t)\varphi(t)\,dt$ — and two signals give the same functional iff they agree a.e., so nothing is lost and nothing spurious is added. (Note: this pairing is linear in φ, without the complex conjugate of Lesson 36's inner product — same bracket, different convention, standard in both subjects.)
- $\langle \delta, \varphi\rangle = \varphi(0)$, and $\langle \delta_{t_0}, \varphi\rangle = \varphi(t_0)$: *sifting is the definition*, not a derived property. Continuity is immediate. O&W's operational account and this one now agree by construction.

Proposition 40.3 (Continuity, checkably: the order of a distribution)**.** *A linear functional on $\mathcal{D}$ is continuous iff it is* bounded*: for every R there are M and m such that*

$$|\langle f, \varphi\rangle| \;\leq\; M \sum_{k \leq m} \|\varphi^{(k)}\|_\infty \qquad \textit{for all } \varphi \textit{ supported in } [-R, R]$$

(Strichartz §6.5). The least workable m is the order *of f: how many derivatives of the test signal the measurement is allowed to consult. Honest signals and δ have order 0 ($|\int x\varphi| \le \|x\|_{L^1[-R,R]}\|\varphi\|_\infty$; $|\varphi(0)| \le \|\varphi\|_\infty$); the dipole δ' has order exactly 1 — no multiple of $\|\varphi\|_\infty$ alone can bound $|\varphi'(0)|$, since a test function can wiggle steeply while staying small.* Positive *distributions ($\langle f,\,\varphi\rangle \ge 0$ whenever $\varphi \ge 0$) automatically have order 0: pin φ between $\pm c\psi$ for a plateau $\psi \equiv 1$ near $[-R,R]$ and use positivity on $c\psi \pm \varphi$ — they are exactly integration against positive measures, and δ' is not even a* difference *of two of them (Strichartz §6.4). Order is what makes the theory usable: to certify a formula as a distribution you verify one finite estimate, not a limit over every sequence.*

Example 40.4 (The principal value, made honest)**.** $1/t$ is not locally integrable, so it is not a distribution by pairing alone — but its *odd symmetry* rescues one:

$$\Big\langle \operatorname{pv}\tfrac{1}{t},\ \varphi\Big\rangle = \lim_{\epsilon\downarrow 0}\int_{|t|>\epsilon}\frac{\varphi(t)}{t}\,dt = \int_{|t|\le 1}\frac{\varphi(t)-\varphi(0)}{t}\,dt + \int_{|t|>1}\frac{\varphi(t)}{t}\,dt,$$

because $\int_{\epsilon<|t|\le 1}\frac{dt}{t} = 0$ (the two sides cancel), and the first integrand is bounded by $\|\varphi'\|_\infty$ (MVT, Theorem 33.5). The estimate exhibits order 1. This is the object living inside $\mathcal{F}u = \pi\delta(\omega) + \frac{1}{j\omega}$: the "$1/j\omega$" of Part V's integration property was never a function, and now it does not need to be.

Definition 40.5 (Operations, by moving them to the test function)**.** Every operation on distributions is defined by transposing what it would do to an honest signal under the integral sign (substitute, integrate by parts — then *define* the result to be that formula):

shift:	$\langle f(t-t_0),\,\varphi\rangle = \langle f,\,\varphi(t+t_0)\rangle$	
scale:	$\langle f(at),\,\varphi\rangle = \frac{1}{\lvert a\rvert}\langle f,\,\varphi(t/a)\rangle$	
multiply by smooth g :	$\langle gf,\,\varphi\rangle = \langle f,\,g\varphi\rangle$	($g\varphi \in \mathcal{D}$ ✓)
differentiate:	$\langle f',\,\varphi\rangle = -\langle f,\,\varphi'\rangle$	(integrate by parts).

Each right-hand side is a continuous functional of φ, so the definitions are legal; and when f is a differentiable signal, integration by parts says the new definition agrees with the classical one. Consequences, instantly:

Proposition 40.6 (The δ calculus, now theorems)**.** *(a) $\delta(at) = \frac{1}{|a|}\delta(t)$: both sides send $\varphi \mapsto \varphi(0)/|a|$.*

(b) $g(t)\,\delta(t-t_0) = g(t_0)\,\delta(t-t_0)$ for smooth g: both sides send $\varphi \mapsto g(t_0)\varphi(t_0)$.

(c) $u' = \delta$: $\langle u',\,\varphi\rangle = -\langle u,\,\varphi'\rangle = -\int_0^\infty \varphi'(t)\,dt = \varphi(0) = \langle\delta,\,\varphi\rangle$.

(d) δ' (the dipole*) sends $\varphi \mapsto -\varphi'(0)$; more generally $\langle\delta^{(k)},\,\varphi\rangle = (-1)^k\varphi^{(k)}(0)$.*

(e) Every *distribution has derivatives of all orders. Differentiation, the most fragile operation in classical analysis (Theorem 31.15's hypotheses; Weierstrass's nowhere-differentiable function), is* unconditionally legal *here — the cost being that the derivative may no longer be a function.*

Example 40.7 (Differentiating the square wave, honestly)**.** The T-periodic square wave x of Example 26.5 ($+1$ then -1) is classically differentiable a.e. with derivative 0 — useless. As a distribution, $x' = \sum_m \big(2\,\delta(t - t_m^{\uparrow}) - 2\,\delta(t - t_m^{\downarrow})\big)$: each jump of height J contributes $J\,\delta$ at the jump (the computation is (c) above, localized). The general rule: *distributional derivative = classical derivative a.e. + one impulse per jump, weighted by the jump.* This is the law behind Part V's "differentiate the triangle to get the square, differentiate the square to get impulses," and it converts the smoothness–decay ladder of Exercise 26.3 into an exact bookkeeping: each differentiation multiplies a_k by $jk\omega_0$, and a signal whose m-th derivative first shows impulses has $|a_k| \sim 1/k^m$.

Definition 40.8 (Convergence of distributions)**.** $f_n \to f$ *in* $\mathcal{D}'$ if $\langle f_n,\, \varphi\rangle \to \langle f,\, \varphi\rangle$ for every test function φ — convergence of all measurements. This is the weakest convergence in this booklet, and therefore the most forgiving.

Proposition 40.9 (Nascent deltas)**.** *Let* $k \in L^1(\mathbb{R})$ *with* $\int k = 1$*, and set* $k_\epsilon(t) = \frac{1}{\epsilon}k(t/\epsilon)$*. Then* $k_\epsilon \to \delta$ *in* $\mathcal{D}'$ *as* $\epsilon \downarrow 0$*.*

Proof. $\langle k_\epsilon,\, \varphi\rangle = \int k(s)\,\varphi(\epsilon s)\,ds \to \varphi(0)\int k = \varphi(0)$, by DCT (Theorem 35.15): the integrands are dominated by $\|\varphi\|_\infty |k| \in L^1$. — Every "limit of narrow pulses" picture of δ (O&W's rectangle of width Δ, the Gaussian of shrinking variance, the Cauchy kernels $\frac{\epsilon/\pi}{t^2+\epsilon^2}$) is this one proposition. Remarkably, decay of k is not needed if oscillation substitutes for it: $\frac{\sin(t/\epsilon)}{\pi t} \to \delta$ in $\mathcal{D}'$ too (its tails cancel by Riemann–Lebesgue, Theorem 35.20) — and *that* family is precisely the truncated Fourier inversion integral, which is why inversion recovers x: partial inversion = $x *$ (nascent δ) $\to x$. □

Example 40.10 (What distributions refuse to do)**.** There is no continuous way to multiply two arbitrary distributions: $\delta \cdot \delta$ would need $\varphi(0)\cdot$"$\delta(0)$," which does not exist — consistent with Part V's experience that LTI theory (linear!) handles impulses effortlessly while a nonlinear system fed an impulse ($y = x^2$ of Example 24.7) is meaningless. Similarly $\delta(t)u(t)$ is undefined (u is not smooth at exactly the wrong point). The rule of thumb: distributions may be added, shifted, scaled, differentiated, convolved with nice things, and multiplied by *smooth* functions — and that list is exactly the toolbox LTI theory uses.

Proposition 40.11 (Convolution, honestly — and why it smooths)**.** *For* $f \in \mathcal{D}'$ *(or* $\mathcal{S}'$*) and a smooth* ψ*, define*

$$(f * \psi)(t) = \langle f,\, \psi(t - \cdot)\rangle$$

— the measurement of f *through the aperture* ψ *slid to position* t*. This is an ordinary, infinitely differentiable* function *of* t*, with derivatives falling on the aperture,* $\frac{d}{dt}(f*\psi) = f*\psi'$*: convolving with a smooth kernel smooths* any *distribution (Strichartz §4.3). Three instant consequences:* $\delta * \psi = \psi$ *(*δ *is the convolution identity — Part V's defining slogan, now a computation);* $\delta' * \psi = \psi'$ *(differentiation* is *convolution with the dipole); and* $\mathcal{F}(f*\psi) = \hat{\psi}\,\mathcal{F}f$ *for tempered* f *(transpose the* $\mathcal{S}$ *identity). For LTI theory this closes the loop: "pass* x *through the filter" is now defined for any tempered input and smooth impulse response, and Proposition 40.9's nascent deltas explain the standard proof pattern — mollify (*$x * k_\epsilon$ *is smooth), verify classically, pass to the distributional limit.*

Tempered distributions and the Fourier transform. To transform a distribution we need test functions the transform maps to *themselves* — compact support dies (a compactly supported signal has an entire, never compactly supported, transform: Lesson 38). The right class trades support for decay:

Definition 40.12 (Schwartz class; tempered distributions)**.** $\mathcal{S}$ is the space of C^∞ functions all of whose derivatives decay faster than any power ($t^m e^{-t^2}$ and friends). The Fourier transform maps $\mathcal{S}$ onto $\mathcal{S}$, bijectively, with the inversion formula valid classically (Strichartz §3.2–3.3 — the proof is Corollary 35.16 run repeatedly: smoothness of φ buys decay of $\hat{\varphi}$ and vice versa; the transform trades the two, and $\mathcal{S}$ is where the trade is even). *Tempered distributions* $\mathcal{S}'$ are the continuous linear functionals on $\mathcal{S}$: everything of at most polynomial growth — all of Part V's signals, periodic signals, impulse trains, but not e^{t^2}. For $f \in \mathcal{S}'$, define

$$\langle \mathcal{F}f,\, \varphi \rangle = \langle f,\, \mathcal{F}\varphi \rangle.$$

Since $\mathcal{F}$ is a bijection of $\mathcal{S}$, $\mathcal{F}f$ is again a tempered distribution: *every* tempered distribution has a Fourier transform, and inversion holds in $\mathcal{S}'$. The convention matches Part V's: $\mathcal{F}\varphi(\omega) = \int \varphi(t) e^{-j\omega t} dt$.

Example 40.13 (The Gaussian, transformed — every tool at once)**.** The flagship inhabitant of $\mathcal{S}$ is the Gaussian, and its transform is this book's tools working in concert (Strichartz §3.5). For $\varphi(t) = e^{-at^2}$, $a > 0$: complete the square in the exponent,

$$-at^2 - j\omega t = -a\Big(t + \frac{j\omega}{2a}\Big)^2 - \frac{\omega^2}{4a},$$

so $\hat{\varphi}(\omega) = e^{-\omega^2/4a} \int e^{-a(t+j\omega/2a)^2} dt$. Sliding the integration path back to the real axis is a *contour shift*: e^{-az^2} is entire, so the integral around the rectangle with horizontal sides at heights 0 and $\omega/2a$ vanishes (closed curve theorem, Theorem 37.5), and the short vertical ends die as the rectangle widens (ML bound, using $|e^{-az^2}| = e^{-a(t^2-y^2)}$). What remains is the famous real integral, evaluated by squaring and switching to polar coordinates — Tonelli (Theorem 35.17) licensing the very step Exercise 17.3 used: $\int e^{-at^2} dt = \sqrt{\pi/a}$. Altogether

$$e^{-at^2} \;\longleftrightarrow\; \sqrt{\tfrac{\pi}{a}}\, e^{-\omega^2/4a},$$

a Gaussian again — and at $a = \frac{1}{2}$ a fixed point of $\mathcal{F}$ up to scale. Width $\sim \sqrt{a}$ in time against $\sim 1/\sqrt{a}$ in frequency is the uncertainty trade at its cleanest, and this pair seeds the nascent-δ family, the heat kernel, and every Gaussian-window spectrogram in the labs.

Theorem 40.14 (The impossible pairs, computed)**.** *In* $\mathcal{S}'$*:*

$$\mathcal{F}\delta = 1, \qquad \mathcal{F}1 = 2\pi\,\delta(\omega), \qquad \mathcal{F}e^{j\omega_0 t} = 2\pi\,\delta(\omega - \omega_0), \qquad \mathcal{F}\cos(\omega_0 t) = \pi\big(\delta(\omega-\omega_0) + \delta(\omega+\omega_0)\big).$$

Proof. First: $\langle \mathcal{F}\delta, \varphi\rangle = \langle \delta, \hat{\varphi}\rangle = \hat{\varphi}(0) = \int \varphi(t)\, dt = \langle 1, \varphi\rangle$. Second: $\langle \mathcal{F}1, \varphi\rangle = \langle 1, \hat{\varphi}\rangle = \int \hat{\varphi}(\omega)\, d\omega = 2\pi\varphi(0)$ by the inversion formula on $\mathcal{S}$ evaluated at $t = 0$. The third is the second plus the shift rule (Definition 40.5, transposed through $\mathcal{F}$ — the modulation property of Theorem 27.3, which holds in $\mathcal{S}'$ because it holds on $\mathcal{S}$); the fourth is Euler plus linearity. These are the δ-pairs of Example 27.2, no longer "certified by sifting" but proved. Note the engine: *a formula true on* $\mathcal{S}$ *transposes to all of* $\mathcal{S}'$ *for free* — every property in Theorem 27.3 (shift, modulation, differentiation $\mathcal{F}(f') = j\omega\mathcal{F}f$, convolution with an L^1 kernel) is valid for tempered distributions by this one argument. □

Theorem 40.15 (Impulse train ↔ impulse train (Poisson summation))**.** *In $\mathcal{S}'$, with $\omega_s = 2\pi/T$:*

$$\mathcal{F}\Big[\sum_{n\in\mathbb{Z}} \delta(t-nT)\Big] = \frac{2\pi}{T}\sum_{k\in\mathbb{Z}} \delta(\omega - k\omega_s), \qquad \textit{equivalently} \qquad \sum_n \varphi(nT) = \frac{1}{T}\sum_k \hat{\varphi}(k\omega_s) \ \ \forall \varphi \in \mathcal{S}.$$

Proof sketch. The train $p(t) = \sum_n \delta(t - nT)$ is T-periodic, and its Fourier *series* coefficients are all $a_k = \frac{1}{T}\int_{-T/2}^{T/2} p\, e^{-jk\omega_s t} dt = \frac{1}{T}$ (sifting on one period). Summing the resulting series $p = \frac{1}{T}\sum_k e^{jk\omega_s t}$ in $\mathcal{S}'$ (partial sums converge as distributions — test and regroup; Strichartz §7 does the epsilonics) and transforming term by term with Theorem 40.14 gives the claim. Pairing the first identity with φ yields the second — a nontrivial numerical identity between two ordinary sums, free of any δ's, which is the form used in analytic number theory and in filter-bank aliasing analysis alike. □

This retroactively proves Lemma 29.1, hence spectrum replication (Theorem 29.2), hence the sampling theorem (Theorem 29.3): the entire sampling story of Lesson 29 now stands on $\mathcal{S}'$ with no operational IOUs. Likewise white noise — statistical inference's "process with autocorrelation $\sigma^2\delta(\tau)$" — is honestly a distribution-valued object: its sample paths are not functions, but its averages against test apertures are perfectly well-defined random variables, which is exactly how the physical instrument sees it.

Where a distribution lives, and what one can be. (Strichartz §§6.1–6.3.) The *support* of f is the complement of the largest open set where all measurements vanish ($\langle f, \varphi\rangle = 0$ for every φ supported there). $\operatorname{supp}\delta = \operatorname{supp}\delta' = \{0\}$ — though δ' "sees" beyond its support infinitesimally: it consults φ' at 0, not just $\varphi(0)$. Compactly supported distributions are automatically tempered ($\mathcal{E}' \subset \mathcal{S}' \subset \mathcal{D}'$), and supports add under convolution, $\operatorname{supp}(f*g) \subseteq \operatorname{supp} f + \operatorname{supp} g$ — the rigorous form of "a length-L FIR filter smears support by at most L." Two structure facts then pin down exactly how wild a distribution can be:

Theorem 40.16 (Point support means impulses; the structure theorem)**.** *(a) If* $\operatorname{supp} f = \{t_0\}$*, then $f = \sum_{k=0}^{m} b_k\, \delta^{(k)}(t - t_0)$, a* finite *sum of impulses and their derivatives — infinite sums of $\delta^{(k)}$ at one point fail the order estimate of Proposition 40.3 (Strichartz §6.3). (b) Every tempered distribution is a* finite *sum $f = \sum_k \frac{d^k}{dt^k} f_k$ of distributional derivatives of continuous functions of polynomial growth (Strichartz §6.2). The impulse train, for instance, is the second derivative of one continuous piecewise-linear function of quadratic growth.*

Part (a) is why "whatever happens at a single instant" in system theory is always a finite train of impulses and dipoles, never anything stranger — and why sparse spike deconvolution may posit its unknown as $\sum_i a_i\delta(t-t_i)$ without loss of generality. Part (b) is the demystification: distributions are not new objects, only finitely many derivatives of ordinary continuous signals — distribution theory completes differential calculus exactly as Lesson 35 completed integration, and adds nothing it was not compelled to add.

40.2 Practice: by hand

Example 40.17 (Drill)**.** (a) $t\,\delta(t) = 0$ (Proposition 40.6(b) with $g = t$); hence the general solution of $t\,f = 0$ in $\mathcal{D}'$ is $f = c\,\delta$ — division by t is possible "up to impulses," the fact behind $\mathcal{F}u = \pi\delta(\omega) + \frac{1}{j\omega}$ (Exercise 40.3). (b) $\langle \delta', t\varphi\rangle = -\frac{d}{dt}(t\varphi)|_0 = -\varphi(0)$, i.e. $t\,\delta' = -\delta$: dipoles

differentiate what multiplies them. (c) $\frac{d}{dt}|t| = \operatorname{sign} t$, $\frac{d^2}{dt^2}|t| = 2\delta$: the triangle-wave/square-wave/impulse-train ladder of Part V in one line per rung. (d) $(x * \delta_{t_0}) = x(t - t_0)$: convolution with a shifted impulse shifts — Part V's Example usage, proved by transposition.

Fourier / statistical-inference / grad DSP connection. Osgood's Fourier lecture notes devote two chapters to exactly this material — $\mathcal{S}$, $\mathcal{S}'$, the pairs of Theorem 40.14, Poisson summation — so this lesson is direct course prep, not backstory. In statistical inference, generalized processes (white noise, ideal derivatives of Brownian motion) live in $\mathcal{S}'$. In graduate DSP, distribution theory is load-bearing wherever ideal elements meet analysis: sampling and multirate identities (Theorem 40.15 is the aliasing formula), PDE-based signal models, sparse spike deconvolution (the unknown is a sum of δ's), and the theory of generalized spectral densities.

40.3 Exercises

Exercise 40.1 (Hand)**.** Compute, by pairing with a test function: (a) the second distributional derivative of $\max(t, 0)$ (answer: δ); (b) $\frac{d}{dt}\left[e^{-t}u(t)\right]$ (answer: $\delta - e^{-t}u(t)$ — the product rule survives because e^{-t} is smooth); (c) $\delta'(-t)$ in terms of δ' (odd).

Exercise 40.2 (Hand)**.** Using Theorem 40.14 and linearity only, transform: (a) a $+1/-1$ square wave via its Fourier series (an impulse train in frequency with weights a_k); (b) $\operatorname{sign}(t) = 2u(t) - 1$ given $\mathcal{F}u = \pi\delta + \frac{1}{j\omega}$; (c) t (differentiate $\mathcal{F}1$: answer $2\pi j\,\delta'(\omega)$).

Exercise 40.3 (Proof, $\star$)**.** Derive $\mathcal{F}u = \pi\delta(\omega) + \frac{1}{j\omega}$ (principal value): write $u = \frac{1}{2} + \frac{1}{2}\operatorname{sign}(t)$, transform the constant with Theorem 40.14, and get $\mathcal{F}[\operatorname{sign}]$ as the $\mathcal{S}'$-limit of $\mathcal{F}[e^{-a|t|}\operatorname{sign}(t)] = \frac{-2j\omega}{a^2+\omega^2}$ as $a \downarrow 0$ (pair with φ and use DCT). Conclude the integration property of Theorem 27.3 — divide by $j\omega$, add $\pi X(0)\delta$ — which Part V stated without proof.

Exercise 40.4 (Proof)**.** Show that if $f_n \to f$ in $\mathcal{D}'$ then $f_n' \to f'$ in $\mathcal{D}'$ — differentiation is *continuous* on distributions (transpose and use the definition), in stark contrast to Example 31.16's second row. Apply it: the Fourier partial sums of the square wave converge in $\mathcal{D}'$, therefore so do their derivatives, to the impulse train of Example 40.7 — termwise differentiation of Fourier series is *always* legal distributionally.

Deeper reading. Strichartz §§1.1–1.4 (what distributions are), 2.1–2.4 (the calculus of distributions), 3.2–3.5 ($\mathcal{S}$, tempered distributions, the Fourier transform), 4.3 (convolution), 6.1–6.5 (support, structure theorems, point support, continuity and order); §7 for Poisson summation and sampling. O&W 1.4 and 2.5 are the operational account this lesson upgrades.

41 Dual Spaces, Riesz Representation, and Functional Derivatives

NEEDED FOR: Optional rigor layer; the calculus of graduate DSP, estimation, and machine learning in function space. Not needed for ML (but explains its gradients).

41.1 Theory

Lesson 40 was secretly a lesson about *dual spaces*; this lesson makes the pattern explicit and then builds the derivative that goes with it.

Definition 41.1 (Dual space)**.** For a normed space V, the *dual* V^* is the space of bounded (equivalently continuous) linear functionals $\ell : V \to \mathbb{C}$, with norm $\|\ell\| = \sup_{\|v\|\le 1} |\ell(v)|$. Examples already met: $\mathcal{D}'$ and $\mathcal{S}'$ are duals ($\mathcal{D}^*$, $\mathcal{S}^*$, with continuity in the appropriate convergences); "evaluate at t_0" is in $C[a,b]^*$ (norm 1) but is *not* a bounded functional on L^2 — point values are meaningless in L^2 (a.e.-classes!), which is precisely why $\delta \notin L^2$ and why sampling theory (Lesson 29) needed bandlimitedness to make evaluation continuous.

Theorem 41.2 (Riesz representation)**.** *Every bounded linear functional on a Hilbert space H is an inner product: for each $\ell \in H^*$ there is a unique $y \in H$ with*

$$\ell(x) = \langle x,\, y\rangle \quad \textit{for all } x \in H, \qquad \textit{and} \qquad \|\ell\| = \|y\|.$$

Proof. If $\ell = 0$ take $y = 0$. Otherwise $M = \ker \ell$ is a closed proper subspace (closed because ℓ is continuous). By the projection theorem (Theorem 36.4) pick a unit vector $e \perp M$ (project any $x_0 \notin M$ onto M and normalize the residual). Every $x \in H$ splits as $x = \big(x - \frac{\ell(x)}{\ell(e)}e\big) + \frac{\ell(x)}{\ell(e)}e$, first piece in M (apply ℓ: zero). Take inner products with e: $\langle x,\, e\rangle = \frac{\ell(x)}{\ell(e)}$, i.e. $\ell(x) = \langle x,\, \overline{\ell(e)}\, e\rangle$; set $y = \overline{\ell(e)}\, e$. Uniqueness: if $\langle x,\, y_1 - y_2\rangle = 0$ for all x, test $x = y_1 - y_2$. Norm equality: Cauchy–Schwarz one way, $x = y/\|y\|$ the other. □

So on a Hilbert space the dual is the space itself — the infinite-dimensional justification of a move Part I made without comment: representing the linear functional "score" as a weight *vector* ($\ell(x) = w \cdot x$), and the linear part of a loss change as a gradient vector. Riesz is what lets both survive in function space:

Definition 41.3 (Functional derivative)**.** Let $J : H \to \mathbb{R}$ be a functional on (an open subset of) a Hilbert space of signals. Its *Gâteaux derivative* at x in direction v is

$$\delta J(x; v) = \frac{d}{d\epsilon} J(x + \epsilon v)\Big|_{\epsilon=0},$$

— ordinary one-variable calculus along the line $x + \epsilon v$. If $v \mapsto \delta J(x; v)$ is a bounded linear functional, Riesz supplies a unique $\nabla J(x) \in H$ with

$$\delta J(x; v) = \langle v,\, \nabla J(x)\rangle \quad \text{for all directions } v,$$

the *functional gradient*. When only distributional pairing is available, physics writes the same object as the density $\frac{\delta J}{\delta x(t)}$: $\delta J(x; v) = \int \frac{\delta J}{\delta x(t)}\, v(t)\, dt$ — a functional derivative *is* a distribution in the perturbation, and ∇J is its Riesz representative when it has one. Steepest descent, stationarity ($\nabla J = 0$), and the chain rule all read exactly as in Lesson 7 (Proposition 7.3 is the finite-dimensional case).

Example 41.4 (The three computations that cover practice)**.** All on real L^2, all by expanding $J(x + \epsilon v)$ and reading the $O(\epsilon)$ term.

(a) *Energy.* $J(x) = \frac{1}{2}\|x\|_2^2$: $J(x+\epsilon v) = J(x) + \epsilon\langle v,\, x\rangle + O(\epsilon^2)$, so $\nabla J = x$. (The $\frac{1}{2}\|\cdot\|^2$-gradient-is-identity rule.)

(b) *Data misfit through a filter.* $J(x) = \frac{1}{2}\|h * x - y\|_2^2$:

$$\delta J(x; v) = \langle h * v,\ h * x - y\rangle = \langle v,\ \tilde{h} * (h * x - y)\rangle, \qquad \tilde{h}(t) = h(-t),$$

so $\nabla J = \tilde{h} * (h * x - y)$: *filter the residual with the time-reversed filter.* The middle step moved h across the inner product — the *adjoint* of convolution-by-h is convolution-by-$\tilde{h}$ (substitute $u = t - \tau$; in frequency: the adjoint of multiplication by H is multiplication by $\bar{H}$, Lesson 10's conjugate-transpose in its ω-indexed form). Time-reversed filters in backpropagation-through-convolution and in matched filtering both come from this adjoint.

(c) *Smoothness penalty.* $J(x) = \frac{1}{2}\int (x')^2$: $\delta J(x; v) = \int x'v' = -\int x''v$ (integration by parts, boundary terms zero for compactly supported perturbations — Lesson 40's move), so $\frac{\delta J}{\delta x(t)} = -x''(t)$. Derivative penalties have *second*-derivative gradients: the Laplacian smoothing of graphics and ML.

Theorem 41.5 (Fundamental lemma of the calculus of variations; Euler–Lagrange)**.** *(a) If g is locally integrable and $\int g\,\varphi = 0$ for all $\varphi \in \mathcal{D}(a, b)$, then $g = 0$ a.e. on (a, b) — i.e. the map from signals to distributions is injective, so a vanishing functional derivative pins down the function. (b) Consequently, a smooth minimizer of $J(x) = \int_a^b L\big(t, x(t), x'(t)\big)dt$ with fixed endpoints satisfies*

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} = 0.$$

Proof. (a) is MIRA 3B (pair against smoothed indicators of $\{g > \varepsilon\}$; the bump functions of Definition 40.1 exist precisely to run this argument). (b): at a minimizer, $0 = \delta J(x; \varphi) = \int \big(L_x\varphi + L_{x'}\varphi'\big) = \int \big(L_x - \frac{d}{dt}L_{x'}\big)\varphi$ for every test φ (integrate by parts; endpoint terms vanish since φ does); apply (a). □

Example 41.6 (Matched filter, by Cauchy–Schwarz)**.** Among unit-energy filters h, which maximizes the output $|(s * h)(0)|$ when the input is a known pulse s? $(s * h)(0) = \int s(\tau)h(-\tau)d\tau = \langle s,\, \tilde{h}\rangle$, and Cauchy–Schwarz (Theorem 2.5) gives $|\langle s,\, \tilde{h}\rangle| \le \|s\|\|h\|$ with equality iff $\tilde{h} \propto s$: the optimal filter is $h(t) = s(-t)/\|s\|$, the *time-reversed template* — the matched filter of every radar and every communication receiver. In statistical inference's Hilbert space of random variables, the identical argument (with the noise-whitened inner product) yields the SNR-optimal detector; this is Example 19.6's finite-dimensional geometry, grown up.

Example 41.7 (Tikhonov-regularized deconvolution)**.** Given noisy filtered data $y \approx h * x$, minimize

$$J(x) = \tfrac{1}{2}\|h * x - y\|_2^2 + \tfrac{\lambda}{2}\|x\|_2^2.$$

By Example 41.4(a)+(b): $\nabla J = \tilde{h} * (h * x - y) + \lambda x$. Setting $\nabla J = 0$ and Fourier-transforming (Plancherel makes this legitimate; $\mathcal{F}\tilde{h} = \bar{H}$): $\bar{H}(H\hat{X} - Y) + \lambda\hat{X} = 0$, so

$$\hat{X}(j\omega) = \frac{\overline{H(j\omega)}}{|H(j\omega)|^2 + \lambda}\, Y(j\omega).$$

At $\lambda = 0$ this is naive inverse filtering Y/H — explosive wherever $H \approx 0$; the λ floor is exactly what tames the ill-conditioning that Lesson 42 will quantify as a condition number. This formula (and its statistical twin, the Wiener filter, where λ becomes the noise-to-signal spectral ratio) is the single most-reused answer in graduate DSP.

Connections, backward and forward. Backward: machine learning's gradient descent (Lesson 7) is this lesson with $H = \mathbb{R}^n$; the LLMS orthogonality principle (Theorem 19.3) is $\nabla J = 0$ for $J = \mathbf{E}[(X - \hat{X})^2]$; least squares (Theorem 6.4) is Example 41.4(b) with matrices. Forward: Wiener and Kalman filtering (statistical inference) are projections computed via these gradients; adaptive filters (LMS, RLS) run stochastic descent on $J = \mathbf{E}|d - h * x|^2$ over the filter h — the gradient is again an adjoint-filtered residual; variational methods, splines, and physics-informed models all start from Theorem 41.5. When a deep-learning paper writes $\frac{\delta \mathcal{L}}{\delta f}$ for a loss over a function, it means Definition 41.3.

41.2 Practice: by hand

Example 41.8 (Drill)**.** (a) $J(x) = \langle x, g \rangle$ (linear): $\nabla J = g$, constant — as in Lesson 7's table. (b) $J(x) = \int c(t)x(t)^2 dt$, c smooth: $\nabla J = 2cx$ (pointwise weight). (c) $J(x) = \frac{1}{2}\|x'\|^2 + \frac{1}{2}\|x\|^2$: stationarity gives $-x'' + x = 0$ — Euler–Lagrange producing a differential equation, the pattern of all variational modeling. (d) Shortest path: $J(x) = \int_a^b \sqrt{1 + x'^2}\, dt$: $L_x = 0$, so $\frac{d}{dt}\frac{x'}{\sqrt{1+x'^2}} = 0$, so x' constant: straight lines — the sanity check every calculus-of-variations user runs once.

41.3 Exercises

Exercise 41.1 (Hand)**.** Compute functional gradients on L^2: (a) $J(x) = \langle x, g \rangle^2$; (b) $J(x) = \frac{1}{4}\int x^4$; (c) $J(x) = \frac{1}{2}\|h * x\|^2$ (use the adjoint). Answers: $2\langle x, g \rangle\, g$; x^3; $\tilde{h} * h * x$.

Exercise 41.2 (Hand)**.** Derive the Euler–Lagrange equation for the smoothing spline functional $J(x) = \frac{1}{2}\sum_i (x(t_i) - y_i)^2 + \frac{\lambda}{2}\int (x'')^2$ formally: $\frac{\delta J}{\delta x(t)} = \sum_i (x(t_i) - y_i)\,\delta(t - t_i) + \lambda x''''(t)$ — note the data term's derivative *is a train of impulses* (Lesson 40), which is why the minimizer is a piecewise cubic with knots exactly at the data.

Exercise 41.3 (Proof, $\star$)**.** Prove the adjoint identity used in Example 41.4(b): $\langle h * v, w \rangle = \langle v, \tilde{h} * w \rangle$ for real $h, v, w \in L^2$ with $h \in L^1$ (Fubini, Lesson 35, licenses the swap). Conclude that the operator norm of convolution-by-$\tilde{h}$ equals that of convolution-by-h (Theorem 36.9: $|\bar{H}| = |H|$).

Exercise 41.4 (Proof)**.** In Example 41.7, prove J is strictly convex for $\lambda > 0$ (expand $J(x + v) - J(x) - \langle v, \nabla J(x) \rangle = \frac{1}{2}\|h * v\|^2 + \frac{\lambda}{2}\|v\|^2 > 0$ for $v \neq 0$), so the stationary point is the unique global minimizer — the function-space copy of Lesson 6's least-squares uniqueness.

Deeper reading. Bowers–Kalton ch. 2 (dual spaces), 7.1–7.2 (Riesz, operators and adjoints); Strichartz §1 (distributions as duals); any calculus of variations opening chapter (e.g. Gelfand–Fomin ch. 1) for Theorem 41.5 in depth.

42 Numerical Linear Algebra: Conditioning, Stability, Factorizations, and Iteration

NEEDED FOR: Optional; the compute layer under every `solve`, `lstsq`, `eigh`, `svd`, and `fft` call in this booklet and under graduate DSP; its iteration half is the numerical side of Lesson 47's training dynamics. Not formally needed for ML — but it explains numerical mystery failures.

Parts I–V trusted `solve`, `lstsq`, `eigh`, `svd`, and `fft`. This lesson opens them up. It follows Trefethen–Bau (T&B) more closely than the rest of Part VI follows its sources, because the questions it answers are daily ones in graduate DSP, where the matrices are correlation matrices and blurring operators, often nearly singular by nature: *how many digits did I just lose, whose fault was it — the problem's or the algorithm's — and what is the routine actually doing?* The lesson has three movements. Theory I–II set up the two concepts the whole subject rests on, *conditioning* (a property of the problem) and *backward stability* (a property of the algorithm), and prove the law that combines them. Theory III–VII open the factorizations: Householder QR, LU with pivoting and Cholesky (`solve`), the least-squares algorithms (`lstsq`), the eigenvalue and SVD algorithms (`eigh`, `svd`), and the FFT, each with its cost, its stability theorem, and where feasible the proof. Theory VIII is iteration at scale: Krylov spaces, conjugate gradients, gradient descent, and preconditioning — the numerical face of Lesson 47.

42.1 Theory I: floating point and conditioning

Definition 42.1 (Floating point, in one box)**.** IEEE double precision stores a sign, an 11-bit exponent, and a 53-bit significand, so that the representable numbers near t are spaced $2^{-52}|t|$ apart and rounding gives relative error at most $\varepsilon_{\text{mach}} = 2^{-53} \approx 1.1 \times 10^{-16}$:

$$\text{fl}(t) = t(1+\epsilon), \qquad |\epsilon| \le \varepsilon_{\text{mach}}.$$

The *fundamental axiom of floating point arithmetic* (T&B 13.7) says each of $+, -, \times, /$ is exact up to the same relative fudge: $x \circledast y = (x * y)(1+\epsilon)$. Everything in this lesson follows from these two lines. Sixteen digits sounds like plenty; the subject exists because *subtraction of nearly equal numbers* promotes relative error catastrophically: $\text{fl}(1 + 10^{-13}) - 1$ keeps only 3 of its 16 digits (*catastrophic cancellation*). Single precision (`float32`, the default on GPUs and in most embedded DSP) has $\varepsilon \approx 6 \times 10^{-8}$; half precision, $\approx 5 \times 10^{-4}$. Every bound below scales with ε, which is why the same algorithm can be fine in double and useless in half.

Definition 42.2 (Condition number of a problem)**.** A *problem* is a map $f : X \to Y$ between normed spaces (data to solution). Its *relative condition number* at x is the worst-case amplification of relative error, in the limit of small perturbations:

$$\kappa(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \frac{\|f(x+\delta x) - f(x)\|/\|f(x)\|}{\|\delta x\|/\|x\|}, \qquad \text{so that} \qquad \kappa(x) = \frac{\|J(x)\|\,\|x\|}{\|f(x)\|}$$

when f is differentiable with Jacobian J. Conditioning is *the problem's fault*: no algorithm, however clever, can be expected to deliver more than $\log_{10}(1/(\kappa\varepsilon))$ correct digits, since merely rounding the input perturbs the exact answer by $\kappa\varepsilon$.

Example 42.3 (Four condition numbers, T&B Examples 12.1–12.5)**.** (a) $x \mapsto x/2$: $J = \frac{1}{2}$, $\kappa = \frac{(1/2)|x|}{|x|/2} = 1$. (b) $x \mapsto \sqrt{x}$: $\kappa = \frac{(1/2\sqrt{x})\,x}{\sqrt{x}} = \frac{1}{2}$. (c) $(x_1, x_2) \mapsto x_1 - x_2$ (with the ∞-norm on the data): $J = (1, -1)$, $\|J\|_\infty = 2$, so $\kappa = \dfrac{2\max(|x_1|, |x_2|)}{|x_1 - x_2|}$: subtraction of nearly equal numbers is an *ill-conditioned problem*, which is the honest statement of "catastrophic cancellation." (d) Roots of a polynomial from its coefficients: $(x-1)^2 = x^2 - 2x + 1$ perturbed to $x^2 - 2x + (1-\epsilon)$ has roots $1 \pm \sqrt{\epsilon}$ — a relative change ϵ in one coefficient moves the root by $\sqrt{\epsilon}$, so $\kappa = \infty$ at a double root and is astronomically large near one (Wilkinson's polynomial). This is why no serious eigenvalue routine forms the characteristic polynomial (Theory VI).

Theorem 42.4 (Conditioning of $Av = b$, T&B 12.1–12.2)**.** *Let A be square and nonsingular, with $\kappa(A) = \|A\|\,\|A^{-1}\|$ (in the 2-norm, $\kappa(A) = \sigma_1/\sigma_n$ by Lesson 12's SVD).*

(a) The problem $v \mapsto Av$ has $\kappa = \|A\|\|v\|/\|Av\| \le \kappa(A)$.

(b) The problem $b \mapsto v = A^{-1}b$ has $\kappa = \|A^{-1}\|\|b\|/\|v\| \le \kappa(A)$.

(c) The problem $A \mapsto v = A^{-1}b$ (b fixed) has $\kappa = \kappa(A)$ exactly.

Proof. (a) and (b) are Definition 42.2 with the linear map itself as Jacobian, plus $\|Av\| \le \|A\|\|v\|$ (for the inequality in (a)) and $\|b\| = \|Av\| \le \|A\|\|v\|$ (for (b)). For (c), perturb A to $A+\delta A$ and v to $v+\delta v$ in $Av = b$: $(A+\delta A)(v+\delta v) = b$ gives, to first order, $A\,\delta v = -\delta A\,v$, i.e. $\delta v = -A^{-1}\delta A\,v$, so $\frac{\|\delta v\|}{\|v\|} \le \|A^{-1}\|\|\delta A\| = \kappa(A)\frac{\|\delta A\|}{\|A\|}$. Equality is attained: with the SVD $A = U\Sigma V^{\mathsf{T}}$, take $\delta A = \epsilon\,u_n(v/\|v\|)^{\mathsf{T}}$ (so $\|\delta A\| = \epsilon$), for which $A^{-1}\delta A\,v = \epsilon\|v\|\,A^{-1}u_n = \epsilon\|v\|\,v_n/\sigma_n$ has norm exactly $\|A^{-1}\|\|\delta A\|\|v\|$. □

Three corollaries you will use constantly: $\kappa(A) \ge 1$ always; $\kappa(Q) = 1$ for unitary Q (all singular values equal 1) — *orthogonal transformations lose no digits*, the reason Theory III–VI are built from them; and $\kappa(A^{\mathsf{T}}A) = \kappa(A)^2$ (singular values square; Exercise 42.3). In the 2-norm the ill-conditioned directions are the small-σ ones: for a circulant (filter) matrix these are the near-zeros of $|H|$ in Example 41.7, and Theorem 36.9's $\max|H|$ is $\|A\|$.

42.2 Theory II: stability, backward error, and the accuracy law

Definition 42.5 (Accuracy, stability, backward stability, T&B 14)**.** Let $\tilde{f}$ be an algorithm (a floating point implementation) for the problem f. It is *accurate* if $\|\tilde{f}(x) - f(x)\|/\|f(x)\| = O(\varepsilon)$ for all x; *stable* if it gives *nearly* the right answer to *nearly* the right question, $\|\tilde{f}(x) - f(\tilde{x})\|/\|f(\tilde{x})\| = O(\varepsilon)$ for some $\tilde{x}$ with $\|\tilde{x}-x\|/\|x\| = O(\varepsilon)$; and *backward stable* if it gives *exactly* the right answer to nearly the right question:

$$\tilde{f}(x) = f(\tilde{x}) \quad \text{for some } \tilde{x} \text{ with } \frac{\|\tilde{x} - x\|}{\|x\|} = O(\varepsilon).$$

Here $O(\varepsilon)$ means $\le C\varepsilon$ for a constant C independent of the data x (but allowed to depend on the dimensions m, n — typically polynomially, and in practice, thanks to statistical cancellation of rounding errors, "$O(\varepsilon)$" is rarely worse than 100ε). Because all norms on a finite-dimensional space are equivalent, all three properties are *norm-independent* (T&B Theorem 14.1).

Accuracy is too much to ask of an ill-conditioned problem — rounding the input already spoils it — so stability is the right goal, and backward stability is the form in which most of numerical linear algebra achieves it.

Example 42.6 (What is and is not backward stable, T&B 15)**.** (a) *Subtraction is backward stable.* With $\mathrm{fl}(x_i) = x_i(1+\epsilon_i)$ and the axiom for $\ominus$,

$$\mathrm{fl}(x_1)\ominus\mathrm{fl}(x_2) = \big(x_1(1+\epsilon_1) - x_2(1+\epsilon_2)\big)(1+\epsilon_3) = x_1(1+\epsilon_4) - x_2(1+\epsilon_5), \qquad |\epsilon_4|, |\epsilon_5| \le 2\varepsilon + O(\varepsilon^2),$$

exactly the difference of the perturbed data $\tilde{x}_i = x_i(1+\epsilon_{3+i})$. Cancellation is an ill-conditioned *problem* computed by a backward stable *algorithm* — the two concepts separate cleanly. The same bookkeeping (compose factors $(1+\epsilon)$, move them freely between numerators and denominators at a cost of $O(\varepsilon^2)$) shows $+, \times, /$ and inner products backward stable. (b) *Outer product*

xy^{T} *is stable but not backward stable*: the computed matrix has small entrywise error but is almost never exactly rank one, so it is not $(x+\delta x)(y+\delta y)^{\mathsf{T}}$ for any small perturbation. Rule: when the solution space is bigger than the data space, backward stability is rare (an explicit computed U in an SVD, for instance, is only *stable*). (c) $x \mapsto x+1$ computed as $\mathrm{fl}(x) \oplus 1$ is stable but not backward stable: near $x=-1$ the absolute error $O(\varepsilon)$ is unbounded relative to x. (d) *Eigenvalues via the characteristic polynomial are unstable.* For $A = \mathrm{diag}(1+10^{-14}, 1)$ the coefficients of $\det(zI-A)$ carry $O(\varepsilon)$ errors and, by Example (d) above, the roots move by $O(\sqrt{\varepsilon}) \approx 10^{-8}$: eight digits lost on a perfectly conditioned problem. `np.roots(np.poly(A))` is a bug, not a method.

Theorem 42.7 (The accuracy law, T&B 15.1)**.** *If a backward stable algorithm is applied to a problem with condition number $\kappa(x)$, the computed answer satisfies*

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = O\big(\kappa(x)\,\varepsilon\big).$$

A backward stable algorithm loses about $\log_{10}\kappa$ digits, and nothing can do better in general.

Proof. Backward stability gives $\tilde{f}(x) = f(\tilde{x})$ with $\|\tilde{x}-x\|/\|x\| = O(\varepsilon)$. Definition 42.2 bounds the effect of that perturbation on the exact solution: $\|f(\tilde{x})-f(x)\|/\|f(x)\| \le (\kappa(x)+o(1))\|\tilde{x}-x\|/\|x\|$. Combine. □

This two-step reasoning — *condition the problem, then prove the algorithm backward stable* — is *backward error analysis*, and it is why the theorems below never estimate forward errors step by step: the condition number is a global property invisible to individual rounding errors, so forward analysis has to rediscover it the hard way. Two proofs of backward stability recur so often that we do them once, in full.

Theorem 42.8 (Back substitution is backward stable, T&B 17.1)**.** *Let $Rx = b$ (R upper triangular, nonsingular) be solved by back substitution, $x_j = \big(b_j - \sum_{k>j} x_k r_{jk}\big)/r_{jj}$ (subtractions carried out left to right), costing $\sim m^2$ flops. The computed $\tilde{x}$ satisfies*

$$(R+\delta R)\tilde{x} = b \quad \textit{exactly, with} \quad \frac{|\delta r_{ij}|}{|r_{ij}|} \le m\,\varepsilon + O(\varepsilon^2) \quad \textit{for every entry.}$$

(A componentwise *bound: zeros of R stay zero, and each entry is perturbed relative to itself.)*

Proof. $m=1$: $\tilde{x}_1 = b_1 \oslash r_{11} = \frac{b_1}{r_{11}}(1+\epsilon_1) = \frac{b_1}{r_{11}(1+\epsilon_1')}$ with $|\epsilon_1'| \le \varepsilon + O(\varepsilon^2)$ — the rounding error, moved into the denominator, is a relative perturbation of r_{11}. $m=2$: $\tilde{x}_2$ is as above with $r_{22}(1+\epsilon_1')$, and $\tilde{x}_1 = (b_1 \ominus (\tilde{x}_2 \otimes r_{12})) \oslash r_{11}$. The $\otimes$ is a factor $(1+\epsilon_2)$ on r_{12}; the $\ominus$ and $\oslash$ contribute factors $(1+\epsilon_3)(1+\epsilon_4)$ on the whole quotient, which move into the denominator as $r_{11}(1+2\epsilon_5)$, $|\epsilon_5| \le \varepsilon + O(\varepsilon^2)$:

$$\tilde{x}_1 = \frac{b_1 - \tilde{x}_2 r_{12}(1+\epsilon_2)}{r_{11}(1+2\epsilon_5)}.$$

So $\tilde{x}$ solves $(R+\delta R)x = b$ exactly with relative perturbations bounded entrywise by $\left(\begin{smallmatrix} 2 & 1 \\ 1 & \end{smallmatrix}\right)\varepsilon$. General m: the same three sources — one $\otimes$ per off-diagonal entry, one $\oslash$ per diagonal entry, and one $\ominus$ per subtraction, which (computing left to right) lands as one perturbation on the diagonal and one on each entry to its right — give the pattern $|\delta r_{ij}|/|r_{ij}| \le$ (row-dependent count $\le m)\varepsilon$; T&B write out the 5×5 case. The trick worth remembering: every rounding error is *reinterpreted* as a perturbation of the input, never propagated forward. □

Lemma 42.9 (Unitary factors do not amplify backward error)**.** *Suppose $B = Q_k \cdots Q_2 Q_1 A$ is computed right to left, and each floating point application is backward stable in the sense $\mathrm{fl}(Q_j C) = Q_j(C + E_j)$ with $\|E_j\| \le c\,\varepsilon \|C\|$ (true for a Householder reflector or a Givens rotation applied by the standard formulas, by the bookkeeping of Theorem 42.8). Then in the 2-norm the computed $\tilde{B} = Q_k \cdots Q_1 (A + E)$ with $\|E\| \le k\,c\,\varepsilon \|A\|\,(1 + O(\varepsilon))$. The conclusion fails if the Q_j are replaced by arbitrary matrices X_j: the error then inflates by $\prod \kappa(X_j)$.*

Proof. Induct on k: $\tilde{B} = Q_k(\tilde{B}_{k-1} + E_k)$ with $\tilde{B}_{k-1} = Q_{k-1} \cdots Q_1 (A + E')$, so $\tilde{B} = Q_k \cdots Q_1 \big(A + E' + Q_1^{\mathsf{T}} \cdots Q_{k-1}^{\mathsf{T}} E_k\big)$, and $\|Q_1^{\mathsf{T}} \cdots Q_{k-1}^{\mathsf{T}} E_k\| = \|E_k\| \le c\varepsilon \|\tilde{B}_{k-1}\| = c\varepsilon \|A\|(1 + O(\varepsilon))$ because unitary matrices preserve the 2-norm. Errors *add*; they never multiply. With non-unitary X_j the pull-back $X_1^{-1} \cdots X_{k-1}^{-1} E_k$ can be $\prod \|X_j^{-1}\|$ times larger. □

Lemma 42.9 is the master key of the lesson: Householder QR, Hessenberg and tridiagonal reduction, the QR algorithm, bidiagonalization, and the FFT are all "a product of unitary factors," and each inherits backward stability from it. Gaussian elimination is a product of *triangular* factors, and its stability story (Theory IV) is exactly the story of how large those factors can get.

42.3 Theory III: orthogonalization — Gram–Schmidt, Householder, and QR

Lesson 6 built Q by Gram–Schmidt; $A = \hat{Q}\hat{R}$ (*reduced*: $\hat{Q}$ is $m \times n$ with orthonormal columns, $\hat{R}$ is $n \times n$ upper triangular) exists for every A and is unique when A has full column rank and $r_{jj} > 0$ (T&B 7.1–7.2; the *full* QR appends $m - n$ orthonormal columns to $\hat{Q}$ and zero rows to $\hat{R}$). Numerically there are three ways to get there.

Proposition 42.10 (Classical vs. modified Gram–Schmidt, T&B 7–9)**.** *Classical GS (Lesson 6) forms $r_{ij} = q_i^{\mathsf{T}} a_j$ for all $i < j$ from the* original *column a_j, then $v_j = a_j - \sum_{i<j} r_{ij} q_i$; modified GS (MGS) subtracts one projection at a time, $v_j \leftarrow v_j - (q_i^{\mathsf{T}} v_j)\, q_i$ for $i = 1, \ldots, j-1$, updating v_j between projections. Identical in exact arithmetic, $\sim 2mn^2$ flops each; in floating point both deliver a product $\tilde{Q}\tilde{R} = A + O(\varepsilon\|A\|)$, but the computed $\tilde{Q}$ loses orthogonality: $\|\tilde{Q}^{\mathsf{T}}\tilde{Q} - I\|$ is typically $O(\kappa(A)^2\varepsilon)$ for classical and $O(\kappa(A)\varepsilon)$ for modified. Neither is backward stable as a QR factorization; MGS becomes a backward stable* least-squares *solver only when used as in Theorem 42.18(b).*

The mechanism: $\tilde{q}_i$ is orthogonalized against *computed* $\tilde{q}$'s, whose own errors it inherits; classical GS additionally subtracts projections computed from stale data, so cancellation errors compound. Householder's idea turns the process around: instead of orthogonalizing A's columns (*triangular orthogonalization*, $AR_1R_2\cdots = Q$), triangularize A with orthogonal matrices (*orthogonal triangularization*, $Q_n \cdots Q_1 A = R$), so that Lemma 42.9 applies.

Definition 42.11 (Householder reflector)**.** For $v \neq 0$, $F = I - 2\dfrac{vv^{\mathsf{T}}}{v^{\mathsf{T}}v}$ is the reflection across the hyperplane $v^{\perp}$: it is symmetric, orthogonal, and its own inverse ($F^2 = I$). Given x, the choice

$$v = \operatorname{sign}(x_1)\|x\|\, e_1 + x \qquad \text{gives} \qquad Fx = -\operatorname{sign}(x_1)\|x\|\, e_1 :$$

F maps x to a multiple of e_1, zeroing everything below the first entry. (Check: $v^{\mathsf{T}}x = \|x\|^2 + \operatorname{sign}(x_1)\|x\|\, x_1$ and $v^{\mathsf{T}}v = 2\|x\|^2 + 2\operatorname{sign}(x_1)\|x\|\, x_1 = 2v^{\mathsf{T}}x$, so $Fx = x - v \cdot 2v^{\mathsf{T}}x / v^{\mathsf{T}}v = x - v$.) Of the two reflections that send x onto the e_1-axis, this sign choice picks the one whose v is a *sum* of like-signed quantities, never a difference of nearly equal ones: it is the single line of Householder QR that exists for stability reasons.

Proposition 42.12 (Householder QR, T&B 10)**.** *For $k = 1, \dots, n$: take $x = A_{k:m,k}$, form v_k as above, and apply $F_k = I - 2v_k v_k^{\mathsf{T}}/v_k^{\mathsf{T}} v_k$ to the trailing block $A_{k:m,k:n}$. After n steps $Q_n \cdots Q_1 A = R$ with $Q_k = \operatorname{diag}(I_{k-1}, F_k)$, so $A = QR$, $Q = Q_1 \cdots Q_n$. Cost $\sim 2mn^2 - \frac{2}{3}n^3$ flops (about the same as Gram–Schmidt for $m \gg n$, two-thirds of it for $m = n$). Q is never formed: the n vectors v_k are stored, and $Q^{\mathsf{T}}b$ (apply $F_1, \dots, F_n$ in turn) or Qx (apply $F_n, \dots, F_1$) costs $O(mn)$. This is LAPACK's* `geqrf`*, hence* `np.linalg.qr`*; the compact form is* `scipy.linalg.qr(A, mode='raw')`*.*

Theorem 42.13 (Householder QR is backward stable, T&B 16.1–16.3)**.** *Let $\tilde{R}$ be the computed triangular factor and let $\hat{Q} = \hat{F}_1 \cdots \hat{F}_n$ be the* exactly *orthogonal product of the reflectors defined by the* computed *vectors $\tilde{v}_k$. Then*

$$\hat{Q}\tilde{R} = A + \delta A, \qquad \frac{\|\delta A\|}{\|A\|} = O(\varepsilon).$$

Consequently, solving a square nonsingular $Av = b$ by "$QR = A$; $y = Q^{\mathsf{T}}b$ via the reflectors; $Rv = y$ by back substitution" is backward stable, $(A + \Delta A)\tilde{v} = b$ with $\|\Delta A\|/\|A\| = O(\varepsilon)$, and the computed solution obeys $\|\tilde{v} - v\|/\|v\| = O(\kappa(A)\varepsilon)$.

Proof. The factorization statement is Lemma 42.9 with $Q_j = \hat{F}_j$: each reflector application is backward stable with respect to the block it acts on, and the $\hat{F}_j$ are exactly orthogonal. (The computed $\tilde{v}_j$ differ from the exact ones, but that only changes *which* orthogonal matrix we call $\hat{Q}$; the lemma never needed the exact Q.) For the solve, the three steps give exact equations $\hat{Q}\tilde{R} = A + \delta A$; $(\hat{Q} + \delta Q)\tilde{y} = b$ with $\|\delta Q\| = O(\varepsilon)$ (applying reflectors to b is backward stable, and the perturbation can be charged to $\hat{Q}$); and $(\tilde{R} + \delta R)\tilde{v} = \tilde{y}$ with $\|\delta R\|/\|\tilde{R}\| = O(\varepsilon)$ (Theorem 42.8). Compose:

$$b = (\hat{Q} + \delta Q)(\tilde{R} + \delta R)\tilde{v} = \big[A + \delta A + (\delta Q)\tilde{R} + \hat{Q}(\delta R) + (\delta Q)(\delta R)\big]\tilde{v} =: (A + \Delta A)\tilde{v}.$$

Each of the four terms is $O(\varepsilon\|A\|)$: $\|\tilde{R}\| \le \|\hat{Q}^{\mathsf{T}}\|\|A + \delta A\| = O(\|A\|)$, so $\|(\delta Q)\tilde{R}\| \le \|\delta Q\|\|\tilde{R}\| = O(\varepsilon\|A\|)$; $\|\hat{Q}\,\delta R\| \le \|\delta R\| = O(\varepsilon)\|\tilde{R}\| = O(\varepsilon\|A\|)$; and the product term is $O(\varepsilon^2\|A\|)$. The accuracy statement is Theorem 42.7 with Theorem 42.4(c). □

T&B's experiment makes the theorem vivid: factor a 50×50 $A = QR$ with random R; the computed $\tilde{Q}, \tilde{R}$ are each wrong in the *third* digit ($\|\tilde{Q} - Q\| \approx 9 \times 10^{-3}$: the individual factors are ill-conditioned functions of A), yet $\|\tilde{Q}\tilde{R} - A\|/\|A\| \approx 10^{-15}$. The factor errors are "diabolically correlated" (Wilkinson): backward stability guarantees the *product*, not the factors, and the proof above shows that the product is all a solver needs.

42.4 Theory IV: solving $Av = b$ — LU with pivoting and Cholesky (`solve`)

Gaussian elimination is *triangular triangularization*: $L_{m-1} \cdots L_1 A = U$ with each L_k unit lower triangular (subtract multiples $\ell_{jk} = u_{jk}/u_{kk}$ of row k from the rows below), so $A = LU$ with $L = L_1^{-1} \cdots L_{m-1}^{-1}$ unit lower triangular (its subdiagonal entries are just the multipliers) and U upper triangular; cost $\sim \frac{2}{3}m^3$ flops, half of Householder QR's $\frac{4}{3}m^3$ on a square matrix. Then $Av = b$ is two triangular solves, $Ly = b$ and $Uv = y$, m^2 each. The factors are not unitary, so Lemma 42.9 does not apply, and the whole stability question is *how big can L and U get?*

Example 42.14 (Why pivoting). $A = \begin{pmatrix} 10^{-20} & 1 \\ 1 & 1 \end{pmatrix}$. Without pivoting, $\ell_{21} = 10^{20}$ and $u_{22} = 1 - 10^{20} \cdot 1$, which rounds to -10^{20}: the "1" has vanished. The computed factors multiply to $\tilde{L}\tilde{U} = \begin{pmatrix} 10^{-20} & 1 \\ 1 & 0 \end{pmatrix}$ — the $(2,2)$ entry of A has been replaced by 0, a backward error of size 1 in a matrix of norm ~ 1. Solving $Av = (1,0)$ this way returns $\tilde{v} \approx (0,1)$ against the true $v \approx (-1,1)$, on a problem with $\kappa(A) \approx 2.6$. Swap the rows first (*partial pivoting*: at step k, bring the largest $|a_{jk}|$, $j \geq k$, to the pivot position): $\ell_{21} = 10^{-20}$, $u_{22} = 1 - 10^{-20} \to 1$, and everything is exact to working precision.

Theorem 42.15 (Stability of Gaussian elimination, T&B 22.1–22.2). *Let $PA = LU$ be computed with partial pivoting (P a permutation; then every $|\ell_{ij}| \leq 1$). The computed factors satisfy*

$$\tilde{L}\tilde{U} = PA + \delta A, \qquad \frac{\|\delta A\|}{\|A\|} = O(\rho\,\varepsilon), \qquad \rho = \frac{\max_{ij} |u_{ij}|}{\max_{ij} |a_{ij}|} \text{ (the growth factor)},$$

and the solve via the two triangular systems is backward stable with the same factor: $(A+\Delta A)\tilde{v} = b$, $\|\Delta A\|/\|A\| = O(\rho\varepsilon)$, hence $\|\tilde{v} - v\|/\|v\| = O(\rho\,\kappa(A)\,\varepsilon)$.

Proof sketch. Elimination is the same pattern of $\otimes$ and $\ominus$ operations as Theorem 42.8, and the same bookkeeping (Wilkinson's) gives the componentwise statement $|\delta A| \leq m\varepsilon|\tilde{L}||\tilde{U}| + O(\varepsilon^2)$. Without pivoting, $\||L||U|\|$ can be arbitrarily larger than $\|A\|$ (Example 42.14), and the algorithm is unstable by any standard. With partial pivoting $|\ell_{ij}| \leq 1$ gives $\|L\| = O(1)$ in any norm, and by definition of ρ, $\|U\| = O(\rho\|A\|)$; the constants absorb the dimension. The solve statement composes the three backward-stable pieces exactly as in Theorem 42.13. □

So `solve` (LAPACK `gesv`: partial-pivoted LU plus two triangular solves) is backward stable *if ρ is $O(1)$*. Two facts about ρ, one alarming and one reassuring. Worst case: for the $m \times m$ matrix with 1 on the diagonal, -1 below it, 1 in the last column and 0 elsewhere, no row swaps ever occur and the last column doubles at each step, so $\rho = 2^{m-1}$ (and this is the maximum possible, Exercise 42.6) — at $m = 60$ that is 10^{18}, and `np.linalg.solve` returns garbage on a matrix with $\kappa \approx m$. Formally, then, Gaussian elimination is backward stable only in the useless dimension-dependent sense (T&B 22.3). In practice: for random matrices ρ is observed to be $O(m^{1/2})$ (the maximum over three million random matrices of size up to 32 was 11.99), the density of ρ decays exponentially, and in more than fifty years of computing no naturally occurring problem with explosive growth is known. The mechanism (T&B 22): instability requires the column spaces of A to be skewed in an exponentially unlikely way; matrices from applications are not. Hence the honest summary of Definition 42.5's original text — `solve` "is backward stable in practice" — and the cheap insurance: after any `solve`, check the *backward error* $\|b - A\tilde{v}\|/(\|A\|\|\tilde{v}\|)$, which must be $O(\varepsilon)$ whatever κ is, and read κ separately (`np.linalg.cond`). A small residual never certifies a small error; it certifies that the algorithm did its job and hands the rest to κ.

Theorem 42.16 (Cholesky factorization, T&B 23.1–23.3). *Every symmetric positive definite A has a unique factorization $A = R^{\mathsf{T}}R$ with R upper triangular and positive diagonal. It is computed by the recursion*

$$A = \begin{pmatrix} a_{11} & w^{\mathsf{T}} \\ w & K \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ w/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & K - ww^{\mathsf{T}}/a_{11} \end{pmatrix} \begin{pmatrix} \alpha & w^{\mathsf{T}}/\alpha \\ 0 & I \end{pmatrix}, \qquad \alpha = \sqrt{a_{11}},$$

applied to the trailing block $K - ww^{\mathsf{T}}/a_{11}$ in turn; $\sim \frac{1}{3}m^3$ flops, half of LU. It needs no pivoting and is backward stable: the computed $\tilde{R}$ satisfies $\tilde{R}^{\mathsf{T}}\tilde{R} = A + \delta A$ with $\|\delta A\|/\|A\| = O(\varepsilon)$, and solving $Av = b$ by $R^{\mathsf{T}}y = b$, $Rv = y$ gives $(A + \Delta A)\tilde{v} = b$, $\|\Delta A\|/\|A\| = O(\varepsilon)$.

Proof. Existence. $a_{11} = e_1^{\mathsf{T}}Ae_1 > 0$, so α is real and the block identity is a direct multiplication. The trailing block $K' = K - ww^{\mathsf{T}}/a_{11}$ is again SPD: for $y \neq 0$ put $x = (-w^{\mathsf{T}}y/a_{11},\, y)$ and expand $x^{\mathsf{T}}Ax = a_{11}t^2 + 2t\, w^{\mathsf{T}}y + y^{\mathsf{T}}Ky$ with $t = -w^{\mathsf{T}}y/a_{11}$, which equals $y^{\mathsf{T}}K'y$; so $y^{\mathsf{T}}K'y = x^{\mathsf{T}}Ax > 0$. Induct: $K' = R_2^{\mathsf{T}}R_2$, and assembling gives $A = R^{\mathsf{T}}R$ with first row $(\alpha,\, w^{\mathsf{T}}/\alpha)$. *Uniqueness:* the form $R^{\mathsf{T}}R$ forces $r_{11} = \sqrt{a_{11}}$ and then the rest of row 1, and inductively every row. *Stability:* in the 2-norm, $\|R\|^2 = \|R^{\mathsf{T}}R\| = \|A\|$ (singular values of R are square roots of the eigenvalues of A), so in Theorem 42.15's bookkeeping $\||\tilde{R}^{\mathsf{T}}||\tilde{R}|\| \leq \|\tilde{R}\|^2 \approx \|A\|$: the growth factor is 1 by structure. (Intuition: the largest entry of an SPD matrix lies on the diagonal, and stays there for every trailing block.) The solve statement composes as in Theorem 42.13, with Theorem 42.8 for both triangular solves (Exercise 42.7). □

When the matrix is SPD — covariance and autocorrelation matrices, $A^{\mathsf{T}}A$, the Hessian in Newton's method (Lesson 47), every Kalman covariance update — `cho_factor`/`cho_solve` in `scipy.linalg` (or its `solve` with `assume_a='pos'`) is twice as fast as `solve` and has no growth-factor caveat; `np.linalg.cholesky` returns the lower factor $L = R^{\mathsf{T}}$ and raises `LinAlgError` if a trailing diagonal goes nonpositive — the standard numerical test for positive definiteness, and the reason "add λI" (diagonal loading, Lesson 45) is what people do when it fires.

42.5 Theory V: least squares — conditioning and the four algorithms (`lstsq`)

For A ($m \times n$, $m \geq n$, full column rank) and b, x minimizes $\|Ax - b\|$ iff $A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$ (Theorem 6.4); with $P = A(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$ the orthogonal projector onto range A and $A^{+} = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$ the *pseudoinverse*, $y = Pb$ is the fitted vector and $x = A^{+}b$. Three routes:

Algorithm	Flops (leading terms)	Idea
Normal equations	$mn^2 + \frac{1}{3}n^3$	Cholesky of $A^{\mathsf{T}}A$
Householder QR	$2mn^2 - \frac{2}{3}n^3$	$A = \hat{Q}\hat{R}$, solve $\hat{R}x = \hat{Q}^{\mathsf{T}}b$
SVD	$2mn^2 + 11n^3$	$A = \hat{U}\hat{\Sigma}V^{\mathsf{T}}$, $x = V\hat{\Sigma}^{-1}\hat{U}^{\mathsf{T}}b$

The normal equations are cheapest; the theorem that decides between them needs the conditioning of least squares, which is subtler than that of a square system because x depends on A *nonlinearly* through the geometry of range A.

Theorem 42.17 (Conditioning of least squares, T&B 18.1)**.** *Let $\kappa = \kappa(A) = \sigma_1/\sigma_n$, let θ be the angle between b and* range A *($\cos\theta = \|y\|/\|b\|$), and let $\eta = \|A\|\|x\|/\|y\| \in [1, \kappa]$ measure how far $\|y\|$ falls short of its maximum $\|A\|\|x\|$. The 2-norm relative condition numbers are*

	y	x
b	$\dfrac{1}{\cos\theta}$	$\dfrac{\kappa}{\eta\cos\theta}$
A	$\dfrac{\kappa}{\cos\theta}$	$\kappa + \dfrac{\kappa^2\tan\theta}{\eta}$

(the b-row exact, the A-row upper bounds). For $m = n$, $\theta = 0$ and the x-column reduces to Theorem 42.4(b),(c).

Proof. Row b. $y = Pb$ is linear with $\|P\| = 1$, so $\kappa_{b\to y} = \|P\|\|b\|/\|y\| = 1/\cos\theta$, attained by $\delta b \in \operatorname{range} A$. $x = A^+ b$ is linear with $\|A^+\| = 1/\sigma_n$, so $\kappa_{b\to x} = \dfrac{\|b\|}{\sigma_n\|x\|} = \dfrac{\sigma_1}{\sigma_n}\cdot\dfrac{\|y\|}{\|A\|\|x\|}\cdot\dfrac{\|b\|}{\|y\|} = \dfrac{\kappa}{\eta\cos\theta}$, attained by δb along the last left singular vector. *Entry* (A, x). Differentiate the normal equations $A^{\mathsf T}Ax = A^{\mathsf T}b$ along a perturbation δA, writing $r = b - Ax$ for the residual: $A^{\mathsf T}A\,\delta x + (\delta A^{\mathsf T}A + A^{\mathsf T}\delta A)x = \delta A^{\mathsf T}b$, i.e.

$$\delta x = (A^{\mathsf T}A)^{-1}\delta A^{\mathsf T} r \;-\; A^+\delta A\, x.$$

The second term is the square-system effect: $\|A^+\delta A\, x\|/\|x\| \le \|\delta A\|/\sigma_n = \kappa\,\|\delta A\|/\|A\|$. The first is the *tilt* of $\operatorname{range} A$ acting on the residual: $\|(A^{\mathsf T}A)^{-1}\| = 1/\sigma_n^2$, and $\|r\| = \|y\|\tan\theta = \|A\|\|x\|\tan\theta/\eta$, so $\|(A^{\mathsf T}A)^{-1}\delta A^{\mathsf T}r\|/\|x\| \le \dfrac{\sigma_1^2}{\sigma_n^2}\dfrac{\tan\theta}{\eta}\dfrac{\|\delta A\|}{\|A\|}$. Add. *Entry* (A, y). y depends on A only through $\operatorname{range} A$; a perturbation δA tilts $\operatorname{range} A$ by an angle at most $\delta\alpha \le \|\delta A\|/\sigma_n = \kappa\|\delta A\|/\|A\|$ (push the shortest semi-axis $\sigma_n u_n$ of the hyperellipse A(unit sphere) orthogonally to $\operatorname{range} A$), and as $\operatorname{range} A$ tilts, y moves on the sphere of radius $\|b\|/2$ centered at $b/2$ (because $y \perp b - y$ always), by at most $\|b\|\,\delta\alpha = \|y\|\,\delta\alpha/\cos\theta$. T&B 18 has the pictures. □

Read the (A, x) entry as a dial. When the fit is close ($\tan\theta \approx 0$) or $\eta \approx \kappa$, least squares is a κ-conditioned problem; when the residual is $O(1)$ and $\eta \ll \kappa$, it is a κ^2-conditioned problem, and no algorithm can beat $\kappa^2\varepsilon$. The algorithms sort themselves by whether they achieve the *problem's* $\kappa + \kappa^2\tan\theta/\eta$ or always pay κ^2.

Theorem 42.18 (Stability of least-squares algorithms, T&B 19.1–19.4)**.** *For the full-rank problem:*

(a) Householder QR (with or without column pivoting, with $\hat Q^{\mathsf T}b$ *formed explicitly or via the reflectors) is backward stable: the computed* $\tilde x$ *exactly minimizes* $\|(A + \delta A)\tilde x - b\|$ *for some* $\|\delta A\|/\|A\| = O(\varepsilon)$.

(b) Modified Gram–Schmidt is backward stable provided $\hat Q^{\mathsf T}b$ *is obtained implicitly by running MGS on the augmented matrix* $[A \;\; b]$ *and reading the* $(n+1)$*st column of* R*; the naive* $\tilde R^{-1}(\tilde Q^{\mathsf T}b)$ *is unstable.*

(c) The SVD route is backward stable.

(d) The normal equations are unstable*: Cholesky solves* $A^{\mathsf T}Ax = A^{\mathsf T}b$ *backward-stably (Theorem 42.16), but for the* κ^2*-conditioned matrix* $A^{\mathsf T}A$*, so the forward error is* $O(\kappa^2\varepsilon)$ *regardless of* θ *and* η.

By Theorems 42.7 and 42.17, (a)–(c) deliver $\|\tilde x - x\|/\|x\| = O\big((\kappa + \kappa^2\tan\theta/\eta)\varepsilon\big)$.

The proof of (a) is Theorem 42.13's composition argument applied to the $m \times n$ case (T&B 19); (b)–(d) are stated with pointers. The moral, in the form every numerical course teaches it, is Proposition 42.19.

Proposition 42.19 (Normal equations square the condition number)**.** $\kappa(A^{\mathsf T}A) = \kappa(A)^2$ *(Exercise 42.3). Hence solving least squares via* $A^{\mathsf T}A\hat x = A^{\mathsf T}b$ *computes a* κ^2*-conditioned problem: for* $\kappa(A) = 10^5$ *— routine for Vandermonde, wavelet, or correlation matrices — the normal*

equations leave ~ 6 *digits where QR leaves* ~ 11 *on a close fit. The rule:* form $A^{\mathsf{T}}A$ for theory, never for computation; call `lstsq`. *(Sample covariance matrices* $\frac{1}{n}X^{\mathsf{T}}X$ *are the same trap in statistics clothing; the SVD of the data matrix* X *is the stable route to PCA.) The normal equations are acceptable exactly when* κ *is modest or* $\tan\theta/\eta$ *is* $O(1)$ *— then the problem itself is* κ^2*-conditioned and nothing is lost.*

T&B's experiment (reproduced in the NumPy lab): fit $e^{\sin 4t}$ on 100 points by a degree-14 polynomial in the monomial basis, $\kappa = 2.3\times10^{10}$, $\theta = 3.7\times10^{-6}$, $\eta = 2.1\times10^{5}$, so the x-condition number is 3.2×10^{10} and the attainable accuracy $\sim 10^{-6}$. Householder QR: $x_{15} = 1.00000031\ldots$ (correct value 1); SVD: $0.99999998\ldots$; MGS on $[A\ \ b]$: $1.00000006\ldots$; naive MGS: 1.029 (10^{14} amplification — unstable); normal equations: 0.39 — *not one digit*, because $\kappa^2\varepsilon \approx 50$.

What `lstsq` does. `np.linalg.lstsq(A, b, rcond)` calls LAPACK `gelsd`: SVD by bidiagonalization plus divide-and-conquer (Theory VI), then $x = \sum_{\sigma_i > \texttt{rcond}\cdot\sigma_1} \frac{u_i^{\mathsf{T}}b}{\sigma_i} v_i$ — backward stable, and the only fully stable option when A is rank-deficient (then the minimum-norm solution is meant, and it depends on the *numerical* rank; Proposition 42.25). `scipy.linalg.lstsq` with `lapack_driver='gelsy'` is Householder QR with column pivoting: cheaper, stable for almost all problems. Solving the normal equations with `solve(Phi.T @ Phi, Phi.T @ y)` is the textbook move; its least-squares twin `lstsq` is the right call.

42.6 Theory VI: eigenvalues and the SVD (`eigh`, `svd`)

Proposition 42.20 (Eigenvalue solvers must iterate, T&B 24–25)**.** *The roots of any polynomial* $p(z) = z^m + c_{m-1}z^{m-1} + \cdots + c_0$ *are the eigenvalues of its companion matrix (the matrix with* 1*'s on the subdiagonal and* $-c_0, \ldots, -c_{m-1}$ *in the last column, Lesson 13), so an eigenvalue algorithm using finitely many* $+, -, \times, /, \sqrt[k]{}$ *would contradict Abel's theorem for* $m \geq 5$. Every *eigenvalue algorithm is therefore iterative — and by Example 42.6(d) the iteration must not go through the characteristic polynomial. The practical design, for symmetric* A*, has two phases: (1) a* finite *Householder similarity reduction* $A \mapsto Q^{\mathsf{T}}AQ = T$ *tridiagonal (*$\sim \frac{4}{3}m^3$ *flops; Hessenberg,* $\sim \frac{10}{3}m^3$*, for nonsymmetric* A*), backward stable by Lemma 42.9 (T&B 26.1); (2) an iteration on* T *costing* $O(m)$ *per step, converging in* $O(m)$ *steps:* $O(m^2)$ *for the eigenvalues alone,* $O(m^3)$ *with eigenvectors accumulated.*

Phase 2 is built from three classical ideas.

Theorem 42.21 (Rayleigh quotient, power, inverse, and Rayleigh-quotient iteration, T&B 27)**.** *Let* $A = A^{\mathsf{T}}$ *have eigenpairs* (λ_j, q_j)*,* q_j *orthonormal, and let* $r(x) = x^{\mathsf{T}}Ax/x^{\mathsf{T}}x$ *(the Rayleigh quotient: the least-squares eigenvalue estimate for* x*, i.e.* $\arg\min_\alpha \|Ax - \alpha x\|$*).*

(a) $\nabla r(x) = \frac{2}{x^{\mathsf{T}}x}(Ax - r(x)x)$*: the eigenvectors are exactly the stationary points of* r*, and* $r(x) - \lambda_J = O(\|x - q_J\|^2)$ *— an* ϵ*-accurate eigenvector gives an* ϵ^2*-accurate eigenvalue.*

(b) Power iteration $v \leftarrow Av/\|Av\|$ *converges to* $\pm q_1$ *with* $\|v^{(k)} \mp q_1\| = O(|\lambda_2/\lambda_1|^k)$ *if* $|\lambda_1| > |\lambda_2|$ *and* $q_1^{\mathsf{T}}v^{(0)} \neq 0$*; the Rayleigh quotient then converges as* $|\lambda_2/\lambda_1|^{2k}$*.*

(c) Inverse iteration *with shift* μ*,* $v \leftarrow (A - \mu I)^{-1}v$ *normalized, converges to the eigenvector* q_J *of the eigenvalue nearest* μ *at rate* $|\mu - \lambda_J|/|\mu - \lambda_K|$ *(*λ_K *the second nearest). The near-singularity of* $A - \mu I$ *is harmless: the solve's error lies mostly along* q_J*, which is the direction wanted.*

(d) Rayleigh quotient iteration *(RQI) — inverse iteration with $\mu = r(v^{(k)})$ refreshed every step — converges* cubically*: if $\|v^{(k)} - q_J\| = \epsilon$ then $\|v^{(k+1)} - q_J\| = O(\epsilon^3)$ and $|\lambda^{(k+1)} - \lambda_J| = O(|\lambda^{(k)} - \lambda_J|^3)$.*

Proof. (a) Differentiate the quotient; at $x = q_J$ the gradient vanishes, and a smooth function with zero gradient is quadratic near its stationary point. More concretely, if $x = \sum a_j q_j$ then $r(x) = \sum a_j^2 \lambda_j / \sum a_j^2$ is a weighted mean of the eigenvalues with weights a_j^2: if $|a_j/a_J| \le \epsilon$ for $j \ne J$, the mean is within $O(\epsilon^2)$ of λ_J. (b) $A^k v^{(0)} = \sum a_j \lambda_j^k q_j = \lambda_1^k \big(a_1 q_1 + \sum_{j \ge 2} a_j (\lambda_j/\lambda_1)^k q_j\big)$; normalize (Lesson 9's powers-of-A picture). The eigenvalue rate is (a). (c) is (b) applied to $(A - \mu I)^{-1}$, whose eigenvalues are $(\lambda_j - \mu)^{-1}$ with the same eigenvectors. (d) If $\|v^{(k)} - q_J\| = \epsilon$, (a) gives $|\lambda^{(k)} - \lambda_J| = O(\epsilon^2)$; one inverse-iteration step with this shift contracts the eigenvector error by the factor $|\lambda^{(k)} - \lambda_J| / |\lambda^{(k)} - \lambda_K| = O(\epsilon^2)$, giving $O(\epsilon^3)$; and (a) again gives $O(\epsilon^6)$ for the next eigenvalue. The constants are uniform near (λ_J, q_J). □

T&B's Example 27.1: $A = \left(\begin{smallmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 4 \end{smallmatrix}\right)$, $v^{(0)} = (1,1,1)/\sqrt{3}$: $\lambda^{(0)} = 5$, $\lambda^{(1)} = 5.2131\ldots$, $\lambda^{(2)} = 5.214319743184\ldots$ against the true $5.214319743377\ldots$ — ten digits in three steps, and three more steps would give about 270. Practice does the first step by hand.

Theorem 42.22 (The QR algorithm, T&B 28–29)**.** *The* pure QR algorithm *— $A^{(0)} = A$; factor $A^{(k-1)} = Q^{(k)} R^{(k)}$, set $A^{(k)} = R^{(k)} Q^{(k)}$ — produces orthogonally similar matrices, and the accumulated $\hat{Q}^{(k)}$ is exactly the orthogonal factor of A^k:*

$$A^{(k)} = (\hat{Q}^{(k)})^{\mathsf{T}} A \, \hat{Q}^{(k)}, \qquad \hat{Q}^{(k)} = Q^{(1)} \cdots Q^{(k)}, \qquad A^k = \hat{Q}^{(k)} \hat{R}^{(k)}.$$

The QR algorithm is *simultaneous power iteration on all of $\mathbb{R}^m$ at once, and its last column is inverse iteration. For symmetric A with $|\lambda_1| > \cdots > |\lambda_m|$, $A^{(k)}$ converges to* $\mathrm{diag}(\lambda_j)$ *with off-diagonal entries decaying like $|\lambda_{j+1}/\lambda_j|^k$ — linearly. The* practical *QR algorithm adds a shift ($A^{(k-1)} - \mu^{(k)} I = Q^{(k)} R^{(k)}$, $A^{(k)} = R^{(k)} Q^{(k)} + \mu^{(k)} I$) and* deflation *(split off a converged eigenvalue and continue on the smaller block). With the Rayleigh quotient shift $\mu^{(k)} = A_{mm}^{(k-1)}$ the last column undergoes RQI, so convergence is cubic in the generic case; the* Wilkinson shift *(the eigenvalue of the trailing 2×2 block nearer A_{mm}) breaks the symmetry that stalls RQI on matrices such as $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and guarantees convergence. Applied to the tridiagonal T, each step costs $O(m)$. The whole is backward stable: the computed $\tilde{\Lambda}$ and the exactly orthogonal $\hat{Q}$ built from the computed rotations satisfy $\hat{Q}\tilde{\Lambda}\hat{Q}^{\mathsf{T}} = A + \delta A$, $\|\delta A\|/\|A\| = O(\varepsilon)$ (T&B 29.1, via Lemma 42.9).*

Proof of the power-iteration identity. Induct: if $A^{k-1} = \hat{Q}^{(k-1)} \hat{R}^{(k-1)}$, then, using $A^{(k-1)} = (\hat{Q}^{(k-1)})^{\mathsf{T}} A \hat{Q}^{(k-1)}$,

$$A^k = A \hat{Q}^{(k-1)} \hat{R}^{(k-1)} = \hat{Q}^{(k-1)} A^{(k-1)} \hat{R}^{(k-1)} = \hat{Q}^{(k-1)} Q^{(k)} R^{(k)} \hat{R}^{(k-1)};$$

the last expression is a QR factorization of A^k, unique up to signs. Convergence and the cubic rate then follow from Theorem 42.21(b),(d) column by column (T&B 28.4, 29). □

This is `np.linalg.eigh` (LAPACK `syevd`: tridiagonalization plus a divide-and-conquer variant of phase 2; `eigvalsh` skips the eigenvectors and is the $O(m^2)$ path) and `np.linalg.eig` (`geev`: Hessenberg plus shifted QR for general matrices, returning possibly complex output). Pass a symmetric matrix to `eig` and you pay $\sim 2.5\times$ the flops and may receive eigenvalues with $O(\varepsilon)$ imaginary parts and non-orthogonal eigenvectors; always `eigh`. What backward stability buys is settled by a perturbation theorem.

Theorem 42.23 (Weyl's inequality; accuracy of `eigh`)**.** *For symmetric A and E, with eigenvalues in decreasing order, $|\lambda_j(A+E) - \lambda_j(A)| \le \|E\|_2$ for every j. Hence the eigenvalues computed by* ***`eigh`*** *satisfy*

$$|\tilde{\lambda}_j - \lambda_j| = O(\varepsilon\|A\|) \quad \text{for every } j:$$

absolute *accuracy at the level of the largest eigenvalue, i.e. relative accuracy $O(\varepsilon\,\lambda_{\max}/|\lambda_j|)$ — the small eigenvalues of a covariance matrix (the noise subspace of MUSIC, the discarded components of PCA) are accurate to $\varepsilon\lambda_{\max}$ and nothing better.*

Proof. Courant–Fischer: $\lambda_j(A) = \max_{\dim S=j}\,\min_{x\in S,\ \|x\|=1} x^{\mathsf{T}}Ax$. Indeed for any j-dimensional S, dimension counting (Lesson 5) gives $S \cap \operatorname{span}(q_j,\dots,q_m) \neq \{0\}$, and on a unit vector there $x^{\mathsf{T}}Ax \le \lambda_j$; while $S = \operatorname{span}(q_1,\dots,q_j)$ has min $= \lambda_j$. *Weyl:* for unit x, $x^{\mathsf{T}}(A+E)x \le x^{\mathsf{T}}Ax + \|E\|$; take min over S then max over S to get $\lambda_j(A+E) \le \lambda_j(A) + \|E\|$, and symmetrically with $A = (A+E) - E$. The accuracy statement is Weyl with $E = \delta A$ from Theorem 42.22. □

Eigen*vectors* are a different story: the angle between the computed $\tilde{q}_j$ and q_j is bounded by $\|E\|/\operatorname{gap}_j$ with $\operatorname{gap}_j = \min_{i\neq j}|\lambda_i - \lambda_j|$ (Davis–Kahan; Exercise 42.8 does the 2×2 case). Individual eigenvectors of clustered eigenvalues are ill-conditioned; only the *invariant subspace* they span is stable. For PCA this means: the top-k subspace is trustworthy when $\lambda_k \gg \lambda_{k+1}$, the individual axes inside it are not when $\lambda_1 \approx \lambda_2$. For *nonsymmetric* matrices even eigenvalues can be arbitrarily ill-conditioned (Bauer–Fike: $|\tilde{\lambda} - \lambda| \le \kappa(V)\|E\|$ for $A = V\Lambda V^{-1}$; the $m\times m$ Jordan block perturbed by ϵ in its corner has eigenvalues $\epsilon^{1/m}$ times roots of unity), which is why Part V's pole computations are done from factored forms, not from state matrices.

Theorem 42.24 (Computing the SVD, T&B 31)**.** *(a) Singular values obey Weyl's inequality too: $|\sigma_k(A+\delta A) - \sigma_k(A)| \le \|\delta A\|$. (b) Computing the SVD as the eigendecomposition of $A^{\mathsf{T}}A$ is* unstable *for the small singular values: a stable eigensolver returns $\tilde{\lambda}_k = \sigma_k^2 + O(\varepsilon\|A\|^2)$, so $\tilde{\sigma}_k = \sigma_k + O(\varepsilon\|A\|^2/\sigma_k)$ — the error is amplified by $\|A\|/\sigma_k$, up to $\kappa(A)$ for σ_n: the normal-equations disease again. (c) The stable route is Golub–Kahan* bidiagonalization*: Householder reflectors alternately on the left (zero a column below the diagonal) and the right (zero a row beyond the superdiagonal), $A = U_1 B V_1^{\mathsf{T}}$ with B upper bidiagonal, $\sim 4mn^2 - \frac{4}{3}n^3$ flops (or $\sim 2mn^2 + 2n^3$ for $m \gg n$ by first computing $A = QR$ and bidiagonalizing R, Lawson–Hanson–Chan); then a shifted-QR or divide-and-conquer iteration on B, $O(n^2)$ for the values. Backward stable by Lemma 42.9, hence by (a) $|\tilde{\sigma}_k - \sigma_k| = O(\varepsilon\|A\|)$ for every k.*

Proof. (a) The symmetric matrix $H = \left(\begin{smallmatrix} 0 & A^{\mathsf{T}} \\ A & 0 \end{smallmatrix}\right)$ has eigenvalues $\pm\sigma_k$ (and zeros), with eigenvectors built from the singular vectors: $H\left(\begin{smallmatrix} v_k \\ \pm u_k \end{smallmatrix}\right) = \pm\sigma_k \left(\begin{smallmatrix} v_k \\ \pm u_k \end{smallmatrix}\right)$. Apply Theorem 42.23 to H and δH, noting $\|\delta H\| = \|\delta A\|$ (Exercise 42.8). (b) is the arithmetic stated. (c) The standard algorithms are, in disguise, the eigenvalue algorithms applied to H without ever forming it; the stability is Lemma 42.9 plus (a). □

Proposition 42.25 (The SVD is the numerical microscope)**.** *For measured data, exact rank is meaningless (rounding fills every dimension); the usable notion is* numerical rank*: the number of singular values above the threshold $\sigma_1 \cdot \max(m,n)\cdot\varepsilon$ (*`np.linalg.matrix_rank`*'s default, justified by Theorem 42.24: below it, a computed σ_k is indistinguishable from 0) or above the measurement-noise level. Truncating at r, $A_r = \sum_{i\le r}\sigma_i u_i v_i^{\mathsf{T}}$ is the best rank-r approximation (Eckart–Young, Exercise 42.3), and solving $Av = b$ by $v = \sum_{i\le r}\frac{u_i^{\mathsf{T}}b}{\sigma_i}v_i$ (*truncated SVD*;* `pinv`*,* `lstsq`*'s* `rcond`*) or by damping $\frac{\sigma_i}{\sigma_i^2+\lambda}$ (*Tikhonov *— Example 41.7 with singular vectors in place of frequencies, Theorem 45.1) is how ill-posed deconvolution, system identification,*

and subspace methods (MUSIC/ESPRIT: signal subspace = top singular vectors of a data matrix, Course 3 Lab 6.4) are actually computed. `np.linalg.svd` *calls LAPACK* `gesdd`*; pass* `full_matrices=False` *unless you need the $m-n$ extra columns of U.*

42.7 Theory VII: the FFT as a stable factorization (`fft`)

Course 3 Lab 6.3 (the DFT in practice) derives the radix-2 FFT as a recursion. As linear algebra it is a *sparse factorization* of the DFT matrix: with $\Omega = \mathrm{diag}(W_N^k)_{k<N/2}$ and Π_N the even/odd permutation,

$$F_N = \begin{pmatrix} I & \Omega \\ I & -\Omega \end{pmatrix} \begin{pmatrix} F_{N/2} & \\ & F_{N/2} \end{pmatrix} \Pi_N, \qquad \text{so} \qquad \frac{F_N}{\sqrt{N}} = \Big(\prod_{s=1}^{\log_2 N} \frac{B^{(s)}}{\sqrt{2}} \Big) P,$$

a product of $\log_2 N$ butterfly stages $B^{(s)}$ and a bit-reversal permutation P, each stage having two nonzeros per row. Since $\left(\begin{smallmatrix} I & \Omega \\ I & -\Omega \end{smallmatrix}\right)^* \left(\begin{smallmatrix} I & \Omega \\ I & -\Omega \end{smallmatrix}\right) = 2I$, every $B^{(s)}/\sqrt{2}$ is unitary, and Theorem 11.4's $\kappa(F_N) = 1$ now has a reason at every stage.

Proposition 42.26 (Accuracy of the FFT)**.** *Computed in floating point with twiddle factors accurate to $O(\varepsilon)$, the FFT returns $\widetilde{Fx} = F(x+\delta x)$ with $\|\delta x\|/\|x\| = O(\varepsilon \log_2 N)$, hence $\|\widetilde{Fx} - Fx\|/\|Fx\| = O(\varepsilon \log_2 N)$ (Higham,* Accuracy and Stability of Numerical Algorithms*, Thm 24.2). The naive N^2 matrix–vector product has the bound $O(\varepsilon N)$: the FFT is not only $N/\log N$ times faster but more accurate, because each output touches only $\log_2 N$ operations instead of N.*

Proof. Lemma 42.9 with $k = \log_2 N$ unitary factors $B^{(s)}/\sqrt{2}$ (the scaling is bookkeeping): a butterfly is one complex multiplication by a twiddle factor known to relative accuracy $O(\varepsilon)$ and one complex add/subtract, each backward stable with a small constant c, so the stage error is $c\varepsilon$ times the current norm and the stages add. Permutations are exact. □

Proposition 42.27 (Structure is speed: circulant and Toeplitz)**.** *An LTI filter on a length-N window is a* Toeplitz *matrix (constant diagonals, Lesson 25); with periodic boundary it is* circulant*, and Lesson 11 proved every circulant is diagonalized by the DFT: $C = F^{-1}\Lambda F$ with $\Lambda = \mathrm{diag}(\hat{c})$ (Theorem 11.6). Consequences:*

- *circulant multiply / solve = FFT, pointwise op, inverse FFT: $O(N \log N)$ and stable — three backward-stable stages whose composition gives a solve with forward error $O(\kappa(C)\,\varepsilon \log N)$ (Exercise 42.4) — vs. $O(N^2)$/$O(N^3)$ dense; and $\kappa(C) = \max|\hat{c}|/\min|\hat{c}|$, read straight off the frequency response: an invertible-looking filter with a deep spectral notch is an ill-conditioned matrix, no SVD needed;*

- *Toeplitz systems — the normal equations of LMS/Wiener filtering, autocorrelation matrices of stationary processes (Course 3 Labs 6.4 and 6.8) — admit the $O(N^2)$ Levinson–Durbin recursion (the workhorse of linear prediction and speech coding; weakly stable for SPD Toeplitz, Cybenko) and $O(N \log^2 N)$ superfast variants; embedding a Toeplitz matrix in a circulant of twice the size makes Toeplitz* multiplication *an FFT, the trick behind every fast convolution in this booklet (Lesson 25) — and, in Theory VIII, the matrix–vector product inside an iterative Toeplitz solver.*

42.8 Theory VIII: iteration at scale — Krylov spaces, conjugate gradients, gradient descent, preconditioning

Factorizations cost $O(m^3)$ and touch every entry. When m is 10^5 and the matrix is sparse or structured, so that a matrix–vector product costs $O(m)$ or $O(m \log m)$, the right tool is an iteration that only ever *multiplies by* A. Everything such an iteration can know after n products lies in the *Krylov subspace*

$$\mathcal{K}_n = \operatorname{span}(b, Ab, A^2b, \dots, A^{n-1}b) = \{p(A)b : \deg p < n\},$$

and the best Krylov methods are the ones that extract the optimal element of $\mathcal{K}_n$ in a useful norm, at $O(m)$ extra work per step.

Proposition 42.28 (Arnoldi and Lanczos, T&B 33–36)**.** *Running modified Gram–Schmidt on* $b, Ab, A^2b, \dots$ as they are generated *(orthogonalize* Aq_n *against* $q_1, \dots, q_n$*) gives an orthonormal basis* Q_n *of* $\mathcal{K}_n$ *with* $AQ_n = Q_{n+1}\tilde{H}_n$*,* $\tilde{H}_n$ *upper Hessenberg (*Arnoldi*); its top block* $H_n = Q_n^{\mathsf{T}} A Q_n$ *is the orthogonal projection of* A *onto* $\mathcal{K}_n$*, whose eigenvalues (*Ritz values*) converge, fastest at the ends of the spectrum, to eigenvalues of* A*. For symmetric* A*,* H_n *is tridiagonal, so only* q_{n-1}, q_n *are needed: the three-term* Lanczos *recurrence,* $O(m)$ *per step — what* `scipy.sparse.linalg.eigsh` *(ARPACK) runs for the top-k eigenpairs of a matrix too large for* `eigh`*, and the exact-arithmetic engine of large-scale PCA. In floating point, Lanczos loses orthogonality among the* q_j *as soon as a Ritz value converges, producing spurious repeated "ghost" eigenvalues; production codes reorthogonalize or restart. Power iteration is the special case that keeps only* q_n*.*

For SPD systems the Krylov method is conjugate gradients, and for it we can prove everything.

Theorem 42.29 (Conjugate gradients, T&B 38.1–38.2)**.** *Let* A *be SPD,* $\|e\|_A = \sqrt{e^{\mathsf{T}} A e}$ *the* A*-norm, and run*

$$\begin{aligned} &x_0 = 0,\ r_0 = b,\ p_0 = r_0; \quad \text{for } n = 1, 2, \cdots: \\ &\alpha_n = \frac{r_{n-1}^{\mathsf{T}} r_{n-1}}{p_{n-1}^{\mathsf{T}} A p_{n-1}}, \qquad x_n = x_{n-1} + \alpha_n p_{n-1}, \qquad r_n = r_{n-1} - \alpha_n A p_{n-1}, \\ &\beta_n = \frac{r_n^{\mathsf{T}} r_n}{r_{n-1}^{\mathsf{T}} r_{n-1}}, \qquad p_n = r_n + \beta_n p_{n-1}. \end{aligned}$$

One product Ap_{n-1} *per step. As long as* $r_{n-1} \neq 0$*: (a) the iterates, search directions, and residuals all span the Krylov space,*

$$\mathcal{K}_n = \operatorname{span}(x_1, \dots, x_n) = \operatorname{span}(p_0, \dots, p_{n-1}) = \operatorname{span}(r_0, \dots, r_{n-1});$$

(b) the residuals are orthogonal, $r_n^{\mathsf{T}} r_j = 0$ *for* $j < n$*, and the search directions are* A*-conjugate,* $p_n^{\mathsf{T}} A p_j = 0$ *for* $j < n$*; (c)* x_n *is the unique minimizer of* $\|x_\star - x\|_A$ *over* $x \in \mathcal{K}_n$*, where* $x_\star = A^{-1}b$*; the error decreases monotonically and vanishes for some* $n \leq m$ *(in exact arithmetic). Equivalently,* x_n *minimizes* $\varphi(x) = \frac{1}{2} x^{\mathsf{T}} A x - b^{\mathsf{T}} x$ *over* $\mathcal{K}_n$*: CG is a descent method on Lesson 47's quadratic, with* $\nabla\varphi(x_n) = -r_n$*, whose search directions are chosen so that minimizing along a line minimizes over the whole Krylov space.*

Proof. (a) By induction from the three update lines: $x_n \in \operatorname{span}(p_j)$, $p_n \in r_n + \operatorname{span}(p_{n-1})$, $r_n \in r_{n-1} + A\operatorname{span}(p_{n-1})$, so all three spans equal $\operatorname{span}(b, \dots, A^{n-1}b)$. (b) Induct on n. $r_n^{\mathsf{T}} r_j = r_{n-1}^{\mathsf{T}} r_j - \alpha_n p_{n-1}^{\mathsf{T}} A r_j$: for $j < n-1$ both terms vanish by hypothesis (note $r_j \in \operatorname{span}(p_0, \dots, p_j)$); for $j = n-1$ the difference is zero exactly when $\alpha_n = r_{n-1}^{\mathsf{T}} r_{n-1} / p_{n-1}^{\mathsf{T}} A r_{n-1}$, and $p_{n-1}^{\mathsf{T}} A r_{n-1} = p_{n-1}^{\mathsf{T}} A p_{n-1} - \beta_{n-1} p_{n-1}^{\mathsf{T}} A p_{n-2} = p_{n-1}^{\mathsf{T}} A p_{n-1}$ by conjugacy — which is the algorithm's α_n. Likewise $p_n^{\mathsf{T}} A p_j = r_n^{\mathsf{T}} A p_j + \beta_n p_{n-1}^{\mathsf{T}} A p_j$: for $j < n-1$, $Ap_j \in \operatorname{span}(r_0, \dots, r_{j+1})$ is orthogonal to r_n and the second term vanishes by hypothesis; for $j = n-1$ use $Ap_{n-1} = (r_{n-1} - r_n)/\alpha_n$ to see the sum is zero exactly for the algorithm's β_n. (c) For $x = x_n - \Delta x \in \mathcal{K}_n$, with $e = x_\star - x = e_n + \Delta x$: $\|e\|_A^2 = \|e_n\|_A^2 + \|\Delta x\|_A^2 + 2 e_n^{\mathsf{T}} A \Delta x$, and $e_n^{\mathsf{T}} A \Delta x = r_n^{\mathsf{T}} \Delta x = 0$ by (b) since $\Delta x \in \mathcal{K}_n$. So $\|e\|_A^2 \geq \|e_n\|_A^2$ with equality iff $\Delta x = 0$. Monotonicity is $\mathcal{K}_n \subset \mathcal{K}_{n+1}$; termination is $\dim \mathcal{K}_n \leq m$. Finally $\|x_\star - x\|_A^2 = 2\varphi(x) + x_\star^{\mathsf{T}} b$, so minimizing either is the same, and $\nabla\varphi = Ax - b = -r$. □

Theorem 42.30 (CG convergence, T&B 38.3–38.5)**.** *Let $\kappa = \lambda_{\max}/\lambda_{\min}$ be the 2-norm condition number of the SPD matrix A, and $\mathcal{P}_n$ the polynomials of degree $\leq n$ with $p(0) = 1$. Then*

$$\frac{\|e_n\|_A}{\|e_0\|_A} = \min_{p \in \mathcal{P}_n} \frac{\|p(A)e_0\|_A}{\|e_0\|_A} \leq \min_{p \in \mathcal{P}_n} \max_{\lambda \in \Lambda(A)} |p(\lambda)| \leq 2\Big(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\Big)^n.$$

In particular CG converges to a fixed tolerance in $O(\sqrt{\kappa})$ iterations; if A has only n distinct eigenvalues it converges in n steps; and clustered spectra converge faster than κ alone predicts.

Proof. By Theorem 42.29(a), $x_n = q(A)b$ with $\deg q < n$, so $e_n = x_\star - q(A)Ax_\star = p(A)e_0$ with $p(\lambda) = 1 - \lambda q(\lambda) \in \mathcal{P}_n$; by (c), x_n minimizes $\|e\|_A$ over $\mathcal{K}_n$, which is the equality. Expanding $e_0 = \sum a_j q_j$ in eigenvectors, $\|p(A)e_0\|_A^2 = \sum a_j^2 \lambda_j p(\lambda_j)^2 \leq \max_j p(\lambda_j)^2 \|e_0\|_A^2$, the first inequality; and a polynomial in $\mathcal{P}_n$ vanishing on n distinct eigenvalues gives termination in n steps. For the last bound take the scaled Chebyshev polynomial $p(\lambda) = T_n\big(\gamma - \frac{2\lambda}{\lambda_{\max} - \lambda_{\min}}\big)/T_n(\gamma)$, $\gamma = \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} = \frac{\kappa+1}{\kappa-1}$: on $[\lambda_{\min}, \lambda_{\max}]$ the argument of the numerator lies in $[-1, 1]$ where $|T_n| \leq 1$, while $T_n(\gamma) = \frac{1}{2}(z^n + z^{-n})$ with $\gamma = \frac{1}{2}(z + z^{-1})$, i.e. $z = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$ (solve the quadratic), so $1/T_n(\gamma) \leq 2z^{-n}$. □

Gradient descent, and what the square root buys. *Steepest descent* on φ takes $p_n = r_n$ (no memory) with the exact line-search step $\alpha = r^{\mathsf{T}} r / r^{\mathsf{T}} A r$ (Exercise 42.10); Lesson 47's Theorem 47.1 is the fixed-step version. Its A-norm error contracts by at most $\frac{\kappa-1}{\kappa+1}$ per step (Kantorovich's inequality), so it needs $O(\kappa)$ iterations against CG's $O(\sqrt{\kappa})$: at $\kappa = 10^4$ and a 10^{-6} tolerance, about 7×10^4 steps against about 700. The difference is *memory*: CG's $p_n = r_n + \beta_n p_{n-1}$ is a momentum term, and CG is the optimal momentum method for a quadratic — the heavy-ball and Nesterov accelerations of ML optimizers achieve the same $\sqrt{\kappa}$ rate on quadratics, which is the reason they exist. Stochastic gradient descent (Lesson 47) and LMS (Course 3 Lab 6.8) are the noisy, memoryless versions, and their eigenvalue-spread slowdown is the same κ.

CG in floating point. The exact-arithmetic guarantees — orthogonal residuals, termination at step m — fail on a computer: like Lanczos, CG loses orthogonality. What survives (Greenbaum's analysis) is the rate: the A-norm error still decreases at essentially Theorem 42.30's rate, only delayed, and the recursively updated r_n tracks the true residual $b - Ax_n$ down to a floor of order $\varepsilon\kappa\|A\|\|x\|$. So CG is robust, but its *stopping test* must be on a quantity you trust: check $\|b - Ax_n\|$ directly every so often.

Preconditioning (T&B 40). Solve $M^{-1}Av = M^{-1}b$ instead, with M cheap to invert and $M^{-1}A$ "close to I" — for CG, symmetrize: $M = CC^{\mathsf{T}}$ and solve $(C^{-1}AC^{-\mathsf{T}})(C^{\mathsf{T}}v) = C^{-1}b$, which is SPD and needs only one solve with M per step. The goal is not literally a small $\kappa(M^{-1}A)$ but a *clustered* spectrum, which Theorem 42.30's polynomial bound rewards directly. Standard choices: $M = \mathrm{diag}(A)$ (Jacobi; T&B's example takes a 1000×1000 sparse matrix from five digits in 40 CG steps to fifteen in 30); incomplete Cholesky (run Theorem 42.16 but keep only A's sparsity pattern); one step of a classical iteration; a coarser or lower-order discretization of the same operator; and — the DSP gem — the *circulant preconditioner* for Toeplitz systems (Strang, T. Chan): M the circulant closest to T, applied by FFT in $O(N \log N)$, for which the eigenvalues of $M^{-1}T$ cluster at 1 and CG converges in a number of steps independent of N. The lab measures it. In ML the same word covers feature normalization, batch normalization, Adam's per-coordinate scaling, and whitening: all are attacks on κ, none is free of the trade-off between the cost of applying M^{-1} and the iterations it saves.

42.9 Theory IX: Lesson 42 in fixed point

Everything above assumed the floating point axiom $\mathrm{fl}(x) = x(1+\delta)$, $|\delta| \le \varepsilon$. Course 3's targets — the Q15 filters and `arm_rfft_q15` of Course 3 Labs 6.1–6.3, the int8 engines of ML accelerators — do not obey it, and the lesson's theorems change in two specific ways that decide what a 16-bit chip can and cannot compute.

Two rules replace the axiom. In a b-bit format with n fraction bits (Q15: $b = 16$, $n = 15$, range $[-1, 1)$, step $q = 2^{-15} \approx 3 \times 10^{-5}$):

(F1) *Rounding is absolute:* $\mathrm{fl}(x) = x + \delta$, $|\delta| \le q/2$. Relative to x the error is $q/2|x|$: a quantity of size 2^{-10} carries five good bits. Products are exact in double width (Q15×Q15 = Q30) and sums are exact, so a dot product accumulated at full width and rounded once suffers a *single* δ — Course 3 Lab 6.2's standing rule, and the reason the bounds below carry no factor m from accumulation.

(F2) *Range is bounded:* $|x| < 2^{\,b-1-n}$, or the result wraps or saturates — a nonlinear catastrophe no perturbation theory covers. Every intermediate quantity must be bounded *a priori*, by a proof, and the data scaled so the bound fits. Each bit of headroom spent on (F2) is a bit of precision lost to (F1); floating point escapes the trade-off with an exponent per number, block floating point (Course 3 Lab 6.3) with one per block.

Proposition 42.31 (The accuracy law in fixed point)**.** *Let a fixed-point algorithm for $Av = b$ run with every intermediate in range and produce $\tilde{v}$ with $(A + \delta A)\tilde{v} = b + \delta b$, each entry of δA and δb bounded by cq for a modest constant c — what the proofs of Theorems 42.8, 42.13, 42.15, and 42.16 deliver when each relative $(1+\delta)$ is replaced by an absolute $+\delta$. Then, in the ∞-norm and to first order,*

$$\frac{\|\tilde{v} - v\|}{\|v\|} \;\le\; \kappa(A)\, cq\Big(\frac{m}{\|A\|} + \frac{1}{\|b\|}\Big) \;\approx\; \kappa(A) \cdot c\,(m+1)\,q \qquad \textit{when } \|A\| \approx \|b\| \approx 1.$$

The accuracy law survives with ε replaced by q divided by the scale of the data*: data that fill the word see $\varepsilon_{\text{eff}} = q$; data scaled down by 2^{-h} for headroom see $2^h q$.*

Proof. Theorem 42.4 bounds the relative error by $\kappa(A)\big(\|\delta A\|/\|A\| + \|\delta b\|/\|b\|\big)$, and (F1) gives $\|\delta A\|_\infty \le mcq$, $\|\delta b\|_\infty \le cq$. □

The number to remember: $\log_2 \kappa$ of the 15 bits go to conditioning and the rest are the answer. $\kappa = 10^2$ leaves 8 bits (2.5 digits); $\kappa = 10^3$ leaves 5; the Hilbert matrices of the lab and the κ^2 of the normal equations are not computable in Q15 at all. This, more than speed, is why embedded signal processing prefers well-conditioned formulations — orthogonal transforms, biquad and lattice structures, normalized LMS — and why a 32-bit accumulator over a 16-bit datapath is the standard compromise.

Elimination: growth is overflow. Theorem 42.15's growth factor ρ was an accuracy nuisance in floating point; under (F2) it is a range requirement. The intermediate U needs $\lceil \log_2 \rho \rceil$ bits of headroom above the data, so A must be scaled down by that much and, by Proposition 42.31, the answer loses those bits. Partial pivoting helps twice — $|\ell_{ij}| \le 1$ keeps every multiplier in Q15 for free, and ρ is small in practice — but the worst case 2^{m-1} that was merely embarrassing in double is fatal at 16 bits.

Proposition 42.32 (Cholesky never overflows)**.** *Let A be SPD with unit diagonal (from any SPD matrix by $D^{-1/2}AD^{-1/2}$, $D = \operatorname{diag} A$ — exact if the scales are powers of two, otherwise a data perturbation covered by Proposition 42.31). Then $|a_{ij}| \le 1$, every entry of the Cholesky factor satisfies $|r_{ij}| \le 1$, and every partial sum $a_{kj} - \sum_{i<l} r_{ik} r_{ij}$ in Theorem 42.16's recurrence has magnitude ≤ 2: one headroom bit in the accumulator and the whole factorization runs in range. The one place (F1) bites is the division by r_{kk}, whose relative error is at most $q/2r_{kk} \le \frac{q}{2}\sqrt{\kappa(A)}$.*

Proof. $|a_{ij}| \le \frac{1}{2}(a_{ii} + a_{jj}) = 1$ is Exercise 42.7(a). $\sum_{i \le j} r_{ij}^2 = a_{jj} = 1$ gives $|r_{ij}| \le 1$; $\sum_{i<l} r_{ik} r_{ij}$ is an inner product of two vectors of norm ≤ 1, hence ≤ 1 by Cauchy–Schwarz, and $|a_{kj}| \le 1$ adds one. For the pivot, R^{-1} has diagonal entries $1/r_{kk}$, so $1/r_{kk} \le \|R^{-1}\| = \sigma_{\min}(R)^{-1} = \lambda_{\min}(A)^{-1/2}$, and $\lambda_{\max}(A) \ge a_{11} = 1$ gives $\lambda_{\min}(A)^{-1} \le \kappa(A)$. □

This is Theorem 42.16's "no growth" clause reread as a range guarantee, and it is why Cholesky, not LU, is the fixed-point factorization — and its Toeplitz cousin more so. The Levinson–Durbin recursion for the Toeplitz normal equations of Course 3 Labs 6.4 and 6.8 (autocorrelation scaled so $r_0 = 1$) produces reflection coefficients k_i with $|k_i| < 1$ and prediction-error powers $E_i = E_{i-1}(1 - k_i^2) \le 1$: every quantity Q15-native by construction, which is why speech codecs run LPC in 16-bit fixed point. Its failure mode is Cholesky's pivot again: $E_i \to 0$ (a nearly singular T, poles near the circle) and the relative error q/E_i explodes — Course 3 Lab 6.2's coefficient-quantization villain in its numerical-analysis costume.

Proposition 42.33 (Iteration in fixed point: floor and stall)**.** *Let $x_{k+1} = x_k + \alpha r_k$, $r_k = b - Ax_k$ (fixed-step steepest descent, Richardson, LMS in the mean) be a contraction in exact arithmetic, $\|e_{k+1}\| \le \theta \|e_k\|$ with $\theta < 1$, and let each step be rounded to Q15 with absolute error $\|\delta_k\| \le \eta$ ($\eta \approx q/2$ for the update plus $\alpha q/2$ for the residual). Then the error is bounded forever,*

$$\limsup_{k \to \infty} \|e_k\| \;\le\; \frac{\eta}{1 - \theta},$$

and for an SPD system with $\alpha = 2/(\lambda_{\max} + \lambda_{\min})$, where $\theta = (\kappa - 1)/(\kappa + 1)$ in the 2-norm, the floor is $\eta(\kappa + 1)/2 \approx \kappa q/4$ — the same κq as a direct solve. Moreover, once $|\alpha r_k|_i < q/2$ for every i the update rounds to zero and the iteration stalls *where it stands, which can happen before the floor is reached.*

Proof. $e_{k+1} = Ge_k + \delta_k$ with $\|G\| \le \theta$ gives $\|e_{k+1}\| \le \theta\|e_k\| + \eta$; unroll: $\|e_k\| \le \theta^k \|e_0\| + \eta(1 + \theta + \theta^2 + \cdots)$. The stall is (F1) applied to $x_k + \alpha r_k$. □

Two consequences. The memoryless methods — LMS above all — are the fixed-point workhorses precisely because Proposition 42.33 covers them; their known pathology, adaptation that freezes when $\mu e[n]x[n]$ drops below half a step, is the stall, and the cures (dither, sign-error updates, a wider accumulator for the weights) are all ways of keeping the update above $q/2$. CG's momentum recursion is not a step-by-step contraction, its finite-precision theory needs relative precision, and it is essentially unused below 32-bit float. The lab's postscript (Exercise 42.14) shows both the floor and the price of headroom: the solution x of a Toeplitz system with $\|T\|_\infty \approx 19$ can fill only a twentieth of the word if $b = Tx$ is to fit, and the relative floor is correspondingly worse than κq.

The FFT: half a bit per stage. Theory VII's factorization $F = A_\nu \cdots A_1 P$ has stages of 2-norm $\sqrt{2}$, not 1: a butterfly $a \pm \omega b$ can double a magnitude, so (F2) forces a scaling. Halve after every stage (the `arm_rfft_q15` convention, output X/N) and Lemma 42.9's argument returns in absolute form: rounding noise injected at stage j (variance $\sigma_B^2 \approx q^2/12$ per output, Course 3 Lab 6.2's model) has its variance halved by each of the $\nu - j$ later stages, so the output noise variance is at most $\sigma_B^2(1 + \frac{1}{2} + \frac{1}{4} + \cdots) \le 2\sigma_B^2$ regardless of N, while the signal, scaled by $1/N$, has variance σ_x^2/N. The output noise-to-signal ratio grows like $N = 2^\nu$: half a bit per stage, five bits at $N = 1024$ (Exercise 42.14 measures 3.2 dB per doubling). Scaling the input by $1/N$ once instead lets the noise grow like N^2, a full bit per stage; block floating point rescales only when a stage actually grows and recovers most of the difference. Floating point's $O(\varepsilon \log_2 N)$ of Proposition 42.26 costs $\log_2 \log_2 N < 4$ bits at any N you will meet — the precise sense in which the FFT is "harder" in fixed point.

Int8 inference is block floating point with a learned block. A quantized neural layer stores $W \approx s\hat{W}$ with $\hat{W}$ an int8 matrix and s a floating point scale — one per tensor, or one per output channel, which is block floating point with the row as the block; activations are quantized the same way, products accumulate in int32 (exact, by (F1)'s double-width rule), and the result is rescaled once. The error analysis is Theorem 42.4(a), the condition of *multiplication*, not of solution: the layer computes $(W + \delta W)(x + \delta x)$ with $|\delta W_{ij}| \le s_i/2$, so $\|\delta W\|_2 \le \|\delta W\|_F \le \frac{\max_i s_i}{2}\sqrt{mn}$ (worst case; $\approx s\sqrt{mn/12}$ under the uniform model), and the relative output error is at most $\frac{\|W\|\|x\|}{\|Wx\|}\left(\frac{\|\delta W\|}{\|W\|} + \frac{\|\delta x\|}{\|x\|}\right)$ — bounded by $\kappa(W)$ times the relative step but usually far less, since $\|Wx\|$ is rarely near its minimum. Per-channel scales exist because $\|\delta W\|/\|W\|$ is smallest when each row fills its own word — Proposition 42.31's "scale the data" applied row by row; and *quantization-aware training* inserts the rounding into the forward pass (with a straight-through gradient) so the optimizer finds weights whose Wx tolerate it. The int8 story is Lesson 42 in miniature: absolute error, a scale to fill the word, a condition number that says how much error survives the map, and structure (per-channel, per-block) to keep that number small.

Algorithm	double ($\varepsilon = 1.1 \times 10^{-16}$, relative)	Q15 ($q = 3 \times 10^{-5}$, absolute)
`solve` (LU, pivoting)	error $\sim \kappa\rho\varepsilon$	headroom $h = \lceil \log_2 \rho \rceil$ bits, error $\sim \kappa\, 2^h q$
Cholesky, Levinson	error $\sim \kappa\varepsilon$, no pivoting	in range at unit diagonal; error $\sim \kappa q$, pivot $q\sqrt{\kappa}$
`lstsq` via QR	$\kappa\varepsilon$ (normal eq.: $\kappa^2\varepsilon$)	normal eq. unusable; Levinson/NLMS instead
steepest descent, LMS	converges to $\kappa\varepsilon$	floor and stall at $\sim \kappa q$
CG	$\sqrt{\kappa}$ steps, floor $\kappa\varepsilon$	not used
`fft`	relative $O(\varepsilon \log_2 N)$	NSR $\propto N$: $\frac{1}{2}$ bit per stage

42.10 Practice: by hand

Example 42.34 (Conditioning without a computer). $A = \begin{pmatrix} 1 & 1 \\ 1 & 1.0001 \end{pmatrix}$: nearly dependent columns. $\sigma_1 \approx 2$, $\sigma_2 \approx 5 \times 10^{-5}$ (product = det = 10^{-4}, sum of squares = $\|A\|_F^2 \approx 4$), so $\kappa \approx 4 \times 10^4$. Perturb $b = (2, 2.0001)$ — solution $(1, 1)$ — to $b' = (2, 2.0002)$: solution jumps to $(0, 2)$. A 5×10^{-5} relative data change produced an $O(1)$ solution change: κ in action. This is what a near-flat spectral notch, a nearly coherent pair of sensors, or two nearly identical features (Lesson 6's duplicated-column `LinAlgError`) all look like to the arithmetic.

Example 42.35 (Householder QR by hand). $A = \begin{pmatrix} 2 & 3 \\ 1 & 0 \\ 2 & 3 \end{pmatrix}$. Column 1: $x = (2, 1, 2)$, $\|x\| = 3$, $v_1 = \text{sign}(2) \cdot 3e_1 + x = (5, 1, 2)$, $v_1^\mathsf{T} v_1 = 30$, so $F_1 = I - \frac{v_1 v_1^\mathsf{T}}{15}$. On x: $v_1^\mathsf{T} x = 15$, $F_1 x = x - v_1 = (-3, 0, 0)$. On the second column $a_2 = (3, 0, 3)$: $v_1^\mathsf{T} a_2 = 21$, $F_1 a_2 = a_2 - \frac{21}{15} v_1 = (-4, -1.4, 0.2)$. Step 2 on the trailing $(-1.4, 0.2)$: norm $\sqrt{1.96 + 0.04} = \sqrt{2}$, $v_2 = \text{sign}(-1.4)\sqrt{2}\, e_1 + (-1.4, 0.2)$, and F_2 sends it to $(+\sqrt{2}, 0)$. Hence

$$R = \begin{pmatrix} -3 & -4 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix}, \qquad q_1 = \frac{a_1}{r_{11}} = -\tfrac{1}{3}(2, 1, 2), \qquad q_2 = \frac{a_2 - r_{12} q_1}{r_{22}} = \tfrac{1}{3\sqrt{2}}(1, -4, 1),$$

and the check $R^\mathsf{T} R = \left(\begin{smallmatrix} 9 & 12 \\ 12 & 18 \end{smallmatrix}\right) = A^\mathsf{T} A$ passes. Gram–Schmidt (Lesson 6) would give the same q's up to sign with $r_{11} = +3$; the negative signs are the price of choosing v by addition. Note that Q was never written down — v_1, v_2 are Q.

Example 42.36 (Cholesky by hand). $A = \left(\begin{smallmatrix} 4 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 6 \end{smallmatrix}\right)$: $\alpha = 2$, first row of R is $(2, 1, 1)$; trailing block $K - ww^\mathsf{T}/a_{11} = \left(\begin{smallmatrix} 5 & 3 \\ 3 & 6 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 4 & 2 \\ 2 & 5 \end{smallmatrix}\right)$; repeat: $(2, 1)$, then $5 - 1 = 4$, $r_{33} = 2$. So $R = \left(\begin{smallmatrix} 2 & 1 & 1 \\ & 2 & 1 \\ & & 2 \end{smallmatrix}\right)$ and $R^\mathsf{T} R = A$. Had a trailing corner come out ≤ 0, A would not have been positive definite — the test is the algorithm.

Example 42.37 (The four condition numbers of a two-column regression, T&B 18.1). $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1.0001 \\ 1 & 1.0001 \end{smallmatrix}\right)$, $b = (2,\ 0.0001,\ 4.0001)$. Normal equations: $A^\mathsf{T} A = \left(\begin{smallmatrix} 3 & 3.0002 \\ 3.0002 & 3.00040002 \end{smallmatrix}\right)$, $A^\mathsf{T} b = (6.0002,\ 6.00060002)$, $\det A^\mathsf{T} A = 2 \times 10^{-8}$, and Cramer's rule gives exactly $x = (1, 1)$, $y = Ax = (2, 2.0001, 2.0001)$, $r = b - y = (0, -2, 2)$. Numbers: $\|b\| \approx 4.472$, $\|y\| \approx 3.464$, $\cos\theta \approx 0.775$, $\tan\theta = \|r\|/\|y\| \approx 0.816$; the eigenvalues of $A^\mathsf{T} A$ are ≈ 6.0004 and $\approx 3.3 \times 10^{-9}$, so $\sigma_1 \approx 2.45$, $\sigma_2 \approx 5.8 \times 10^{-5}$, $\kappa \approx 4.2 \times 10^4$; and $\eta = \|A\|\|x\|/\|y\| \approx 1.0$ (x points along the top right singular vector). Theorem 42.17: y from b: 1.3; x from b: 5.5×10^4; y from A: 5.5×10^4; x from A: $4.2 \times 10^4 + 1.8 \times 10^9 \cdot 0.82 \approx 1.5 \times 10^9$. The coefficients of this tiny regression are determined to about 7 digits by any stable algorithm — and here the normal equations do about as well, since $\tan\theta/\eta \approx 1$ puts the problem itself at κ^2. Exercise 42.12 changes b to a near-exact fit and watches the last entry collapse to κ.

Example 42.38 (Quadratic and cubic convergence, felt). $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, eigenpairs $(3, (1, 1)/\sqrt{2})$ and $(1, (1, -1)/\sqrt{2})$. Start at $x_0 = (2, 1)$: the angle θ_0 to q_1 has $\sin^2\theta_0 = 0.1$, and $r(x_0) = 14/5 = 2.8$ — exactly $\lambda_1 \cos^2\theta_0 + \lambda_2 \sin^2\theta_0 = 3 - 2(0.1)$, the weighted mean of Theorem 42.21(a): an angle error of 0.32 rad gave an eigenvalue error of 0.2. One RQI step: $(A - 2.8I)w = x_0$,

$A - 2.8I = \left(\begin{smallmatrix} -0.8 & 1 \\ 1 & -0.8 \end{smallmatrix}\right)$ with determinant -0.36, gives $w \propto (2.6, 2.8) \propto (13, 14)$, so $\tan\theta_1 = \frac{14-13}{14+13} = \frac{1}{27}$, $\sin^2\theta_1 = \frac{1}{730}$, and $r(x_1) = 3 - \frac{2}{730} = 2.99726$: the error fell from 0.2 to 0.0027, about $0.2^3/3$. Next step: $\tan\theta_2 \approx (1/27)^3 \cdot 1.1 \approx 4 \times 10^{-5}$, eigenvalue error $\approx 3 \times 10^{-9}$. Two steps, nine digits.

Example 42.39 (CG in two steps, steepest descent in many)**.** $A = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right)$ ($\lambda = 4, 2$, $\kappa = 2$), $b = (1, 0)$, $x_\star = (\frac{3}{8}, -\frac{1}{8})$. CG: $r_0 = p_0 = (1, 0)$, $\alpha_1 = \frac{1}{3}$, $x_1 = (\frac{1}{3}, 0)$, $r_1 = (0, -\frac{1}{3})$ ($\perp r_0$ ✓), $\beta_1 = \frac{1}{9}$, $p_1 = (\frac{1}{9}, -\frac{1}{3})$ ($p_1^\mathsf{T} A p_0 = \frac{3}{9} - \frac{1}{3} = 0$ ✓), $Ap_1 = (0, -\frac{8}{9})$, $\alpha_2 = \frac{1/9}{8/27} = \frac{3}{8}$, $x_2 = (\frac{1}{3} + \frac{1}{24}, -\frac{1}{8}) = (\frac{3}{8}, -\frac{1}{8}) = x_\star$: done in $m = 2$ steps, as Theorem 42.29(c) promises. Steepest descent from the same x_1 moves along r_1 with $\alpha = \frac{r_1^\mathsf{T} r_1}{r_1^\mathsf{T} A r_1} = \frac{1}{3}$ to $x_2 = (\frac{1}{3}, -\frac{1}{9})$, $r_2 = (\frac{1}{9}, 0)$: the residual shrinks by exactly $\frac{1}{3} = \frac{\kappa-1}{\kappa+1}$ per step and zigzags forever between the two axes — the worst case of Theorem 47.1, realized. Memory of one previous direction was the whole difference.

Grad DSP / ML connection. Estimation and adaptive filtering run on exactly four numerical kernels: least squares (QR/SVD, never the normal equations unless $\tan\theta/\eta$ is $O(1)$ — Theorem 42.17), SPD solves (Cholesky: Kalman updates, Wiener filters, Newton steps), eigen/SVD of covariance and data matrices (PCA, subspace tracking, MUSIC — `eigh`/`svd`, accurate to $\varepsilon\lambda_{\max}$ absolutely, Theorem 42.23), and Toeplitz/circulant products and solves (FFT, Levinson, preconditioned CG). Conditioning is not a corner case there: correlation matrices of oversampled or narrowband signals are *structurally* near-singular, and regularization (λ, diagonal loading, SVD truncation) is the standing repair — Lesson 41 gave its variational meaning, Lesson 45 its modeling meaning, this lesson its numerical one. On the ML side, training is Theory VIII with noise: the κ that throttles gradient descent, the $\sqrt{\kappa}$ that momentum buys back, and the preconditioners called normalization are all one subject. And on the embedded side (Course 3's platforms), `float32` multiplies every bound here by 10^8 relative to double: a $\kappa = 10^4$ least-squares problem that lost 4 of 16 digits now loses 4 of 7.

42.11 Practice: NumPy — the trust-but-verify lab

Two listings, one per movement. Run with `uv run --with "numpy>=2" --with scipy python`. Comments give the printed numbers (last digits vary by platform).

```
import numpy as np
from scipy.linalg import solve, cho_factor, cho_solve, lu, hilbert
rng = np.random.default_rng(1)
eps = np.finfo(float).eps                              # 2.2e-16

# ---- 1. backward vs forward error: solve on Hilbert matrices ------------
for n in (4, 8, 12):
    A = hilbert(n); x = np.ones(n); b = A @ x
    xt = np.linalg.solve(A, b)
    bwd = np.linalg.norm(b - A@xt) / (np.linalg.norm(A)*np.linalg.norm(xt))
    print(n, f"kappa={np.linalg.cond(A):.1e}",
          f"fwd={np.abs(xt-x).max():.1e}", f"bwd={bwd:.1e}")
# kappa 1.6e4 / 1.5e10 / 1.7e16 ; fwd 5e-14 / 2e-7 / 0.2 ; bwd <= 1e-16 always:
# the algorithm did its job every time (backward error ~ eps); kappa did the rest.
```

```
# ---- 2. the growth-factor matrix: solve fails at kappa ~ 60 -------------
def wilk(m):
    A = -np.tril(np.ones((m, m)), -1) + np.eye(m); A[:, -1] = 1; return A
for m in (20, 40, 60):
    A = wilk(m); A[:, -1] += 1e-10*rng.standard_normal(m)  # break exact integers
    x = np.ones(m); xt = np.linalg.solve(A, A@x)
    P, L, U = lu(A)
    print(m, f"kappa={np.linalg.cond(A):.1e}", f"rho={np.abs(U).max():.1e}",
          f"err={np.abs(xt-x).max():.1e}")
# rho = 5e5, 6e11, 6e17 = 2^(m-1); err 6e-11, 2e-10, 31 (!) with kappa only ~ m.
# Then: random 60x60 matrices give rho ~ 5-10 -- the instability is real but rare.

# ---- 3. least squares, four ways (T&B Lecture 19) -----------------------
m, n = 100, 15
t = np.arange(m)/(m-1); A = np.vander(t, n, increasing=True)
b = np.exp(np.sin(4*t)); b /= 2006.787453080206      # exact x[14] = 1
x_ref = np.linalg.lstsq(A, b, rcond=None)[0]
y = A @ x_ref; theta = np.arcsin(np.linalg.norm(b-y)/np.linalg.norm(b))
kap = np.linalg.cond(A); eta = np.linalg.norm(A,2)*np.linalg.norm(x_ref)/np.linalg.
    norm(y)
print(f"kappa={kap:.2e} theta={theta:.2e} eta={eta:.2e}",
      f"cond(x|A)={kap + kap**2*np.tan(theta)/eta:.1e}")
Q, R = np.linalg.qr(A)                                # Householder
x_qr = np.linalg.solve(R, Q.T@b)
def mgs(A):
    V = A.astype(float).copy(); n = V.shape[1]
    Q = np.zeros_like(V); R = np.zeros((n, n))
    for j in range(n):
        for i in range(j):                            # subtract one at a time
            R[i, j] = Q[:, i] @ V[:, j]; V[:, j] -= R[i, j]*Q[:, i]
        R[j, j] = np.linalg.norm(V[:, j]); Q[:, j] = V[:, j]/R[j, j]
    return Q, R
Qm, Rm = mgs(A); x_mgs = np.linalg.solve(Rm, Qm.T@b)   # naive: unstable
Qa, Ra = mgs(np.column_stack([A, b]))                 # augmented: stable
x_aug = np.linalg.solve(Ra[:n, :n], Ra[:n, n])
try:                                                  # normal equations
    x_ne = cho_solve(cho_factor(A.T@A), A.T@b)
except np.linalg.LinAlgError:                         # kappa^2 eps > 1: A^T A is
    print("Cholesky refuses: A^T A not numerically SPD") # not even numerically SPD
    x_ne = np.linalg.solve(A.T@A, A.T@b)              # LU, as MATLAB's \ falls back
U, s, Vt = np.linalg.svd(A, full_matrices=False)
x_svd = Vt.T @ ((U.T@b)/s)
for name, xx in [("householder", x_qr), ("mgs naive", x_mgs),
                 ("mgs augmented", x_aug), ("normal eqs", x_ne), ("svd", x_svd)]:
    print(f"{name:14s} x15 = {xx[14]:.14f}")
# householder 0.99999996, mgs naive 0.970 (!), mgs augmented 0.99999999,
# normal eqs -0.43 (no digits: kappa^2 eps ~ 100), svd 0.99999996.
print("orthogonality loss:", np.linalg.norm(Q.T@Q-np.eye(n)),
      np.linalg.norm(Qm.T@Qm-np.eye(n)))              # 2e-15 vs 3e-7 ~ kappa*eps

# ---- 4. eigh vs eig vs Rayleigh quotient iteration ----------------------
S = rng.standard_normal((6, 6)); S = S + S.T
w, V = np.linalg.eigh(S); we = np.linalg.eig(S)[0]
```

```
print("eig imag parts:", np.abs(we.imag).max(), " eigh residual:",
      np.linalg.norm(S@V - V*w))                      # 0.0 (lucky) or ~1e-16 ; 7e
    -15
A3 = np.array([[2., 1, 1], [1, 3, 1], [1, 1, 4]]); v = np.ones(3)/np.sqrt(3)
lam = v@A3@v
for k in range(3):                                    # RQI, T&B Example 27.1
    v = np.linalg.solve(A3 - lam*np.eye(3), v); v /= np.linalg.norm(v)
    lam = v@A3@v; print(k+1, f"{lam:.15f}")
# 5.213114754098361 / 5.214319743184 / 5.214319743377535 : cubic.

# ---- 5. singular values: svd vs eigh(A^T A) ------------------------------
U0, _ = np.linalg.qr(rng.standard_normal((50, 50)))
V0, _ = np.linalg.qr(rng.standard_normal((50, 50)))
sig = np.logspace(0, -12, 50); A = (U0*sig) @ V0.T
s_svd = np.linalg.svd(A, compute_uv=False)
s_eig = np.sqrt(np.abs(np.linalg.eigvalsh(A.T@A)))[::-1]
print("svd rel err on sigma_min:", abs(s_svd[-1]-sig[-1])/sig[-1],
      " via A^T A:", abs(s_eig[-1]-sig[-1])/sig[-1])   # 3e-6 (< eps*kappa) vs 1e4
    (!)
print("matrix_rank:", np.linalg.matrix_rank(A))       # 50 is what the threshold
    says

# ---- 6. FFT accuracy vs naive DFT ----------------------------------------
for N in (2**8, 2**12, 2**16):
    x = rng.standard_normal(N) + 1j*rng.standard_normal(N)
    X = np.fft.fft(x); err_fft = np.linalg.norm(np.fft.ifft(X) - x)/np.linalg.norm(x
    )
    line = f"N={N:6d} fft roundtrip {err_fft:.1e} ~ {eps*np.log2(N):.1e}"
    if N <= 2**12:
        F = np.exp(-2j*np.pi*np.outer(np.arange(N), np.arange(N))/N)
        line += f"   naive {np.linalg.norm(F@x - X)/np.linalg.norm(X):.1e}"
    print(line)
# fft roundtrip 2.7e-16 / 3.6e-16 / 4.3e-16: creeps up with log N, well under
# the bound; the naive DFT is 100-1000x worse (3e-14, 5e-13) and 1000x slower.
```

```
import numpy as np
from scipy.linalg import toeplitz
rng = np.random.default_rng(2)

def cg(matvec, b, tol=1e-10, maxit=2000, precond=None):
    """Conjugate gradients on an SPD operator; returns (x, residual history)."""
    x = np.zeros_like(b); r = b.copy(); z = precond(r) if precond else r
    p = z.copy(); rz = r@z; hist = []
    for k in range(maxit):
        Ap = matvec(p); alpha = rz/(p@Ap)
        x += alpha*p; r -= alpha*Ap
        hist.append(np.linalg.norm(b - matvec(x)))     # TRUE residual, not r
        if hist[-1] < tol*np.linalg.norm(b): break
        z = precond(r) if precond else r
        rz_new = r@z; p = z + (rz_new/rz)*p; rz = rz_new
    return x, hist
```

```
def steepest(matvec, b, tol=1e-10, maxit=20000):
    x = np.zeros_like(b); r = b.copy(); hist = []
    for k in range(maxit):
        Ar = matvec(r); x += (r@r)/(r@Ar)*r; r = b - matvec(x)
        hist.append(np.linalg.norm(r))
        if hist[-1] < tol*np.linalg.norm(b): break
    return x, hist

# ---- Toeplitz autocorrelation of an AR(1) process: r_k = rho^|k| --------
N, rho = 1024, 0.95
r = rho**np.arange(N); T = toeplitz(r); b = rng.standard_normal(N)
kappa = np.linalg.cond(T); print(f"kappa(T) = {kappa:.0f}")   # ~1500 = ((1+rho)/(1-
    rho))^2
# Toeplitz matvec by FFT (embed in a circulant of size 2N):
c = np.concatenate([r, [0], r[:0:-1]]); C = np.fft.fft(c)
tmv = lambda v: np.fft.ifft(C*np.fft.fft(np.concatenate([v, np.zeros(N)]))).real[:N]
print("matvec check:", np.abs(tmv(b) - T@b).max())            # ~1e-13

x_cg, h_cg = cg(tmv, b); x_sd, h_sd = steepest(tmv, b, maxit=60000)
sk = np.sqrt(kappa); bound_cg = np.log(2e10)/np.log((sk+1)/(sk-1))
bound_sd = np.log(1e10)/np.log((kappa+1)/(kappa-1))
print("CG iterations:", len(h_cg), " Chebyshev bound:", int(bound_cg))   # 332 vs
    462
print("SD iterations:", len(h_sd), " (kappa-1)/(kappa+1) bound:", int(bound_sd))
print("CG err:", np.abs(x_cg - np.linalg.solve(T, b)).max())  # 2e-9
# CG 332 steps (bound 462) ; steepest descent 15992 (bound 17460): sqrt(kappa).

# ---- Strang's circulant preconditioner: T's Toeplitz row, wrapped -------
cs = r.copy(); half = N//2; cs[half+1:] = r[1:half][::-1]; Cs = np.fft.fft(cs).real
print("preconditioner spectrum:", Cs.min(), Cs.max())         # SPD: all positive
pre = lambda v: np.fft.ifft(np.fft.fft(v)/Cs).real           # M^{-1} v in O(N log N
    )
x_pcg, h_pcg = cg(tmv, b, precond=pre)
print("PCG iterations:", len(h_pcg))                          # 3 (!), independent
    of N
# Eigenvalues of M^{-1} T cluster at 1 (Chan-Strang): try N=4096 -- same count.

# ---- Lanczos in ten lines: top eigenvalues without eigh ------------------
def lanczos(matvec, n, k):
    q = rng.standard_normal(n); q /= np.linalg.norm(q); Q = [q]; a, bb = [], []
    q_prev = np.zeros(n); beta = 0.0
    for j in range(k):
        w = matvec(q) - beta*q_prev; alpha = q@w; w -= alpha*q
        w -= np.array(Q).T @ (np.array(Q) @ w)             # full
    reorthogonalization
        beta = np.linalg.norm(w); a.append(alpha); bb.append(beta)
        q_prev, q = q, w/beta; Q.append(q)
    Tk = np.diag(a) + np.diag(bb[:-1], 1) + np.diag(bb[:-1], -1)
    return np.linalg.eigvalsh(Tk)
ritz = lanczos(tmv, N, 30); true = np.linalg.eigvalsh(T)
print("top-3 Ritz:", ritz[-3:], "\ntop-3 true:", true[-3:])   # agree to 5-7 digits
    at k=30
# Delete the reorthogonalization line and rerun: ghost copies of the top
```

```
# eigenvalue appear -- loss of orthogonality made visible.
```

Complexity notes. Listing 1: `solve` is $\frac{2}{3}m^3$; `qr` $2mn^2$; `lstsq`/`svd` $2mn^2 + O(n^3)$; `eigh` $\frac{4}{3}m^3$ plus $O(m^3)$ for vectors; the naive DFT N^2 against the FFT's $5N \log_2 N$ flops. Listing 2: every CG step is one Toeplitz matvec ($O(N \log N)$ by FFT) plus $O(N)$ vector work, so the whole preconditioned solve is $O(N \log N)$ — against $\frac{1}{3}N^3$ for Cholesky, N^2 for Levinson; Lanczos with full reorthogonalization is $O(Nk^2)$ plus k matvecs.

42.12 Exercises

Exercise 42.1 (Hand)**.** Predict the digits: you solve a least-squares problem with $\kappa(A) = 10^6$ in double precision on a close fit ($\tan\theta \approx 0$). How many correct digits does QR-based `lstsq` leave? The normal equations? Would single precision ($\varepsilon \approx 6 \times 10^{-8}$) with QR beat double precision with normal equations? What changes if the fit is poor, $\tan\theta/\eta \approx 1$? (Answers: $\sim$10; $\sim$4; no: $\kappa\varepsilon \approx 6 \times 10^{-2}$ leaves $\sim$1 digit, worse than the normal equations' $\sim$4 — stability cannot buy back precision that the format never had; for a poor fit the problem itself is κ^2-conditioned and all stable methods drop to ~ 4.)

Exercise 42.2 (Hand)**.** A filter has frequency response magnitudes between 0.01 and 2.5 on the DFT grid. What is the condition number of its length-N circulant matrix? How does Tikhonov regularization with $\lambda = 10^{-2}$ change the effective ratio max / min of $\frac{|H|}{|H|^2+\lambda}$-damped inversion gains? (Compute the new worst gain $\max \frac{|H|}{|H|^2+\lambda} \le \frac{1}{2\sqrt{\lambda}} = 5$.)

Exercise 42.3 (Proof, $\star$)**.** Prove $\kappa(A^\mathsf{T}A) = \kappa(A)^2$ from the SVD, and prove Eckart–Young for the operator norm: if $\operatorname{rank} B \le r$ then $\|A - B\| \ge \sigma_{r+1}$ (find a unit vector in $\operatorname{span}(v_1, \dots, v_{r+1}) \cap \operatorname{null} B$ — dimension counting, Lesson 5 — on which A acts with norm $\ge \sigma_{r+1}$).

Exercise 42.4 (Proof)**.** Show that the DFT matrix satisfies $\kappa(F) = 1$ (unitary up to scale, Theorem 11.4) — transforms lose no digits — and that for a circulant $C = F^{-1}\Lambda F$ the three-stage solve (FFT, divide by $\hat{c}$, inverse FFT), each stage backward stable, yields a computed $\tilde{v}$ with relative error $O(\kappa(C)\varepsilon \log N)$: structure does not evade conditioning, it only evades cost. (Compose the three backward errors as in Theorem 42.13; the middle stage is a diagonal solve, Theorem 42.8 with $m = 1$ per entry.)

Exercise 42.5 (Hand)**.** Householder-triangularize $A = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \\ 2 & 0 \end{smallmatrix}\right)$ by hand as in Example 42.35: write v_1, $F_1 a_1$, $F_1 a_2$, then v_2 and R, and check $R^\mathsf{T}R = A^\mathsf{T}A$. Then explain, using the sign rule of Definition 42.11, what would go wrong numerically for the column $x = (1, 10^{-9}, 10^{-9})$ if v were chosen as $x - \|x\|e_1$.

Exercise 42.6 (Hand)**.** Run Gaussian elimination with partial pivoting by hand on the 4×4 matrix of Theorem 42.15's worst case (ones on the diagonal and in the last column, -1 below the diagonal), verify that no rows are ever swapped and that the last column doubles at each step, and read off $\rho = 8$. Then prove $\rho \le 2^{m-1}$ for every matrix: with $|\ell_{ij}| \le 1$, each elimination step can at most double the largest entry of the trailing block.

Exercise 42.7 (Proof)**.** Complete Theorem 42.16: (a) show that the largest entry in magnitude of an SPD matrix lies on the diagonal (consider $x = e_i \pm e_j$), so no growth is possible; (b) derive the backward stability of the Cholesky *solve* from that of the factorization and Theorem 42.8, following the four-term composition in the proof of Theorem 42.13.

Exercise 42.8 (Proof, $\star$)**.** (a) Verify the eigenvector formula for $H = \left(\begin{smallmatrix} 0 & A^{\mathsf{T}} \\ A & 0 \end{smallmatrix}\right)$ in Theorem 42.24 and deduce Weyl's inequality for singular values. (b) Davis–Kahan in 2×2: for $A = \operatorname{diag}(\lambda_1, \lambda_2)$ and symmetric $E = \left(\begin{smallmatrix} 0 & \epsilon \\ \epsilon & 0 \end{smallmatrix}\right)$, compute the rotation angle of the eigenvectors of $A + E$ exactly and show $\tan 2\theta = 2\epsilon/(\lambda_1 - \lambda_2)$: the eigenvector error is $\|E\|/\text{gap}$ while the eigenvalue error is only $O(\epsilon^2/\text{gap})$. Why does the second fact not contradict Weyl?

Exercise 42.9 (Hand)**.** Do the first RQI step of T&B's Example 27.1 by hand: $A = \left(\begin{smallmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 4 \end{smallmatrix}\right)$, $v^{(0)} \propto (1,1,1)$. Check $\lambda^{(0)} = 5$, solve $(A - 5I)w = (1,1,1)$ by elimination (answer $w = (3,4,6)$), and compute $\lambda^{(1)} = r(w) = 318/61 = 5.21311\ldots$ against the true $5.21431974\ldots$: the error fell from 0.214 to 0.0012 in one step.

Exercise 42.10 (Proof)**.** (a) Show $\nabla\varphi(x) = -r$ for $\varphi(x) = \frac{1}{2}x^{\mathsf{T}}Ax - b^{\mathsf{T}}x$, derive the exact line-search step $\alpha = r^{\mathsf{T}}r / r^{\mathsf{T}}Ar$ of steepest descent, and write the full three-line iteration. (b) Prove that CG applied to a matrix with exactly n distinct eigenvalues terminates in n steps (Theorem 42.30: exhibit the polynomial). (c) T&B 38.2: A is 805×805 symmetric with eigenvalues $1.00, 1.01, \ldots, 9.00$ and also $10, 12, 16, 24$. How many CG steps guarantee a 10^{-6} reduction of $\|e\|_A$? (Kill the four outliers with four factors, then apply the Chebyshev bound with $\kappa = 9$ to the rest.)

Exercise 42.11 (NumPy)**.** (a) Reproduce T&B Figure 22.2 roughly: factor 10^4 random 32×32 matrices with `scipy.linalg.lu`, histogram the growth factors, and report the maximum. Then factor the worst-case matrix of Listing 1 at $m = 30, 60, 100$. (b) In Listing 2, remove the reorthogonalization line from `lanczos`, run $k = 60$, and count how many Ritz values sit within 10^{-6} of the top eigenvalue — the ghosts. (c) Rerun the preconditioned CG with $N = 4096$ and $\rho = 0.99$ ($\kappa \approx 4 \times 10^4$) and tabulate CG vs. PCG iteration counts.

Exercise 42.12 (Hand)**.** In Example 42.37 replace b by $(2,\ 2.0001,\ 2.0001)$ (an exact fit) and, without recomputing x, recompute θ, η, and the four condition numbers. Which entry collapses, and to what? What does this say about when the normal equations are safe? (The (A, x) entry drops from 1.5×10^9 to $\kappa \approx 4.2 \times 10^4$; the normal equations still pay κ^2, so they are now 10^5 times *worse* than QR.)

Exercise 42.13 (Proof)**.** (a) Proposition 42.32 keeps the factorization in range; show that the back substitution $Rv = y$ need not be, by bounding $\|v\|/\|y\|$ in terms of $\kappa(A)$, and state the headroom it requires. (b) In Proposition 42.33, show that a stall occurs only when $\|r_k\|_\infty < q/2\alpha$, and deduce that the stalled error can be as large as $\|A^{-1}\|_\infty q/2\alpha \approx \kappa q/4$ for $\alpha \approx 2/\lambda_{\max}$ — the floor and the stall are the same κq. (c) Verify the FFT noise claim: with independent rounding noises of variance σ_B^2 injected at each stage output, and each later stage computing $(u \pm t)/2$, show the output variance is $\sigma_B^2 \sum_{j=0}^{\nu-1} 2^{-j} < 2\sigma_B^2$.

Exercise 42.14 (NumPy)**.** Lesson 42 in Q15, with $\text{q15}(x) = \text{round}(2^{15}x)/2^{15}$. (a) For the AR(1) Toeplitz T of Listing 2 with $N = 64$, $\rho = 0.9$ ($\kappa = 319$) and a solution x scaled so that $b = Tx$ fits in $[-1, 1)$, solve the quantized system $\text{q15}(T)\, v = \text{q15}(b)$ in double and compare the relative error with $\kappa \cdot 2^{-16}$. (b) Run fixed-step steepest descent, $\alpha = 2/(\lambda_{\max} + \lambda_{\min})$, rounding the residual and the iterate to Q15 at every step; plot the error against the double-precision run and locate the floor. Explain, from the scale of x, why it is worse than (a). (c) Write a radix-2 decimation-in-time FFT that rounds to Q15 and halves after every stage; for $N = 64, \ldots, 4096$ and a uniform random input, report the output noise-to-signal ratio in dB against `np.fft.fft(x)/N` and read off the growth per doubling.

Deeper reading. Trefethen–Bau: Lectures 12–15 (conditioning, floating point, stability), 16–17 (Householder and back-substitution stability, with the only fully detailed proof in the book), 7–10 (Gram–Schmidt and Householder), 11 and 18–19 (least squares: algorithms, conditioning, stability — the polynomial-fit experiment), 20–23 (Gaussian elimination, pivoting, growth factors, Cholesky), 24–31 (eigenvalue problems, the two phases, Rayleigh quotient iteration, the QR algorithm, computing the SVD), 32–38 and 40 (Krylov iterations, Arnoldi/Lanczos, GMRES, conjugate gradients, preconditioning). Higham, *Accuracy and Stability of Numerical Algorithms*, ch. 24 for the FFT bound; Golub–Van Loan, *Matrix Computations*, for everything at production depth; Levinson–Durbin and circulant preconditioners in any statistical DSP text, e.g. Hayes ch. 5 and Ng, *Iterative Methods for Toeplitz Systems*. Fixed point: Oppenheim–Schafer, *Discrete-Time Signal Processing*, ch. 9 (finite-register effects in filters and FFTs); Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference (CVPR 2018) and Gholami et al., "A survey of quantization methods for efficient neural network inference (2021) for int8.

43 Transforms of Distributions: Moment Generating Functions, Characteristic Functions, and the Central Limit Theorem

NEEDED FOR: Optional rigor layer for Lesson 21 (the central limit theorem, proved) and for statistical inference (transforms, Chernoff bounds); not needed for ML.

43.1 Theory

Lesson 21 proved Chebyshev and the weak law from two numbers, μ and σ^2, and then *stated* the central limit theorem. The central limit theorem is a statement about the *shape* of the distribution of a sum, and shape is awkward to push through a sum directly: the density of $X_1 + X_2$ is the convolution of the two densities (B&T 4.1), and an n-fold convolution is hopeless by hand. The remedy is the one Part V uses for convolution — transform, so that convolution becomes multiplication.

Definition 43.1 (Moment generating function (transform))**.** The *moment generating function* (MGF; B&T say *transform*) of a random variable X is

$$M_X(s) = \mathbf{E}\big[e^{sX}\big] = \begin{cases} \displaystyle\sum_x e^{sx}\, p_X(x), & X \text{ discrete}, \\ \displaystyle\int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx, & X \text{ continuous}, \end{cases} \qquad s \in \mathbb{R},$$

defined at those s where the expectation is finite. Always $M_X(0) = \mathbf{E}[1] = 1$. We say X *has an MGF* if $M_X(s) < \infty$ for all s in some open interval $(-s_0, s_0)$, $s_0 > 0$; every statement below assumes this.

Not every random variable has one: e^{sx} grows exponentially, so a density with polynomial tails gives $M_X(s) = \infty$ for every $s \neq 0$ (Exercise 43.4). Statistical inference sidesteps this with the *characteristic function* $\mathbf{E}[e^{j\omega X}]$ — the Fourier transform of the density (Lesson 27), finite for every random variable since $|e^{j\omega x}| = 1$ — and everything below goes through with $j\omega$ in place of s. We stay with M_X because it keeps every computation real.

Proposition 43.2 (Transforms of the standard distributions)**.** *(a) Bernoulli(p):* $M_X(s) = 1-p+pe^s$.

(b) Poisson(λ): $M_X(s) = \exp\big(\lambda(e^s-1)\big)$.

(c) Exponential(λ): $M_X(s) = \dfrac{\lambda}{\lambda - s}$ *for* $s < \lambda$ *(and* $+\infty$ *for* $s \geq \lambda$*).*

(d) $\mathcal{N}(0,1)$*:* $M_X(s) = e^{s^2/2}$ *for every* s*.*

Proof. (a) $(1-p)e^0 + p\,e^s$. (b) and (c) are the definition plus one series or one integral:

$$\sum_{k\geq 0} e^{sk}\,\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k\geq 0}\frac{(\lambda e^s)^k}{k!} = e^{-\lambda}e^{\lambda e^s}, \qquad \int_0^\infty e^{sx}\lambda e^{-\lambda x}\,dx = \lambda\int_0^\infty e^{-(\lambda-s)x}\,dx = \frac{\lambda}{\lambda-s},$$

the integral requiring $\lambda - s > 0$; otherwise the integrand does not decay and the integral diverges. (d) Complete the square in the exponent, $sx - \frac{x^2}{2} = -\frac{1}{2}(x-s)^2 + \frac{s^2}{2}$:

$$M_X(s) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{sx-x^2/2}\,dx = e^{s^2/2}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x-s)^2/2}\,dx = e^{s^2/2},$$

the last integral being the total mass of a $\mathcal{N}(s,1)$ density. □

Proposition 43.3 (Three rules)**.** *Let* X *have an MGF on* $(-s_0, s_0)$*.*

(a) Affine. $M_{aX+b}(s) = e^{sb}M_X(as)$*. In particular* $\mathcal{N}(\mu,\sigma^2)$ *has* $M(s) = e^{\mu s + \sigma^2 s^2/2}$*.*

(b) Moments. M_X *is infinitely differentiable on* $(-s_0, s_0)$*,*

$$M_X(s) = \sum_{k=0}^{\infty}\frac{\mathbf{E}[X^k]}{k!}\,s^k \quad (|s| < s_0), \qquad \text{and} \qquad M_X^{(k)}(0) = \mathbf{E}[X^k] \quad (k = 0,1,2,\dots).$$

In particular $M_X'(0) = \mathbf{E}[X]$*,* $M_X''(0) = \mathbf{E}[X^2]$*, and* $\operatorname{var}(X) = M_X''(0) - M_X'(0)^2$*.*

(c) Independence. *If* X *and* Y *are independent, then* $M_{X+Y}(s) = M_X(s)\,M_Y(s)$*; hence for* $X_1,\dots,X_n$ *i.i.d. with common transform* M*,* $M_{X_1+\cdots+X_n}(s) = M(s)^n$*.*

Proof. (a) $\mathbf{E}[e^{s(aX+b)}] = \mathbf{E}[e^{sb}e^{(as)X}] = e^{sb}\,\mathbf{E}[e^{(as)X}]$. For the normal, write $X = \mu + \sigma Z$ with $Z \sim \mathcal{N}(0,1)$ and apply Proposition 43.2(d): $M_X(s) = e^{\mu s}e^{(\sigma s)^2/2}$.

(b) Expand $e^{sX} = \sum_{k\geq 0} s^k X^k/k!$ and take expectations term by term; that is the series in (b). The one step needing justification is the exchange of $\mathbf{E}$ with the infinite sum, and it is legal precisely because M_X is finite near 0: for $|s| < s_0$ the partial sums are dominated in absolute value by

$$e^{|s|\,|X|} \;\leq\; e^{sX} + e^{-sX},$$

which has finite expectation, so the dominated convergence theorem of Lesson 35 permits the exchange (B&T take it on faith; so may you on a first pass). The same domination shows the series converges absolutely for $|s| < s_0$, so it is a power series with radius of convergence at least s_0; inside that radius a power series is infinitely differentiable term by term, and differentiating k times at $s = 0$ leaves exactly $\mathbf{E}[X^k]$.

(c) For fixed s, e^{sX} and e^{sY} are functions of independent variables, hence independent (Exercise 18.4), so Theorem 16.4 gives $\mathbf{E}[e^{sX}e^{sY}] = \mathbf{E}[e^{sX}]\,\mathbf{E}[e^{sY}]$. Induct on n for the i.i.d. case. □

Rule (c) is the whole point: a sum of independent variables — a convolution of densities — becomes a product of transforms. Rule (b) explains the name and gives a second reading of the CLT's hypothesis: near $s = 0$ a transform is $1 + \mathbf{E}[X]s + \mathbf{E}[X^2]s^2/2 + \cdots$, so *the first two moments are all a transform remembers to second order.* The first payoff is an inequality.

Proposition 43.4 (Chernoff bound)**.** *For every $a \in \mathbb{R}$ and every $s \geq 0$ with $M_X(s) < \infty$,*

$$\mathbf{P}(X \geq a) \leq e^{-sa} M_X(s), \qquad \textit{hence} \qquad \mathbf{P}(X \geq a) \leq e^{-\phi(a)}, \quad \phi(a) = \sup_{s \geq 0} \big(sa - \ln M_X(s)\big).$$

For $X_1, \ldots, X_n$ i.i.d. with mean μ and transform M,

$$\mathbf{P}(M_n \geq a) \leq e^{-n\phi(a)}, \qquad \textit{with } \phi(a) > 0 \textit{ whenever } a > \mu :$$

the tail of the sample mean decays exponentially *in n, not like Chebyshev's $1/n$.*

Proof. For $s > 0$, $x \mapsto e^{sx}$ is increasing, so $\{X \geq a\} = \{e^{sX} \geq e^{sa}\}$, and Markov's inequality (Theorem 21.1(a)) on the nonnegative variable e^{sX} gives $\mathbf{P}(e^{sX} \geq e^{sa}) \leq \mathbf{E}[e^{sX}]/e^{sa}$; at $s = 0$ the bound reads $\mathbf{P}(X \geq a) \leq 1$. Taking the best s gives the second form. For the sample mean, $\mathbf{P}(M_n \geq a) = \mathbf{P}(X_1 + \cdots + X_n \geq na) \leq e^{-sna} M(s)^n = e^{-n(sa - \ln M(s))}$ by rule (c). Finally $g(s) = sa - \ln M(s)$ has $g(0) = 0$ and, by rule (b), $g'(0) = a - M'(0)/M(0) = a - \mu > 0$, so $g(s) > 0$ for small $s > 0$ and $\phi(a) \geq g(s) > 0$. □

Two facts about transforms are quoted without proof. Both are inversion theorems for the Laplace transform of a density and belong to Lesson 38's inversion integrals; B&T state the first and use the second.

Theorem 43.5 (Uniqueness (B&T's inversion property))**.** *If $M_X(s) = M_Y(s) < \infty$ for all s in some interval $(-s_0, s_0)$, then X and Y have the same CDF.*

Theorem 43.6 (Continuity theorem)**.** *Let $Y_1, Y_2, \ldots$ and Y have MGFs finite on a common interval $(-s_0, s_0)$, and suppose $M_{Y_n}(s) \to M_Y(s)$ for every $s \in (-s_0, s_0)$. Then $\mathbf{P}(Y_n \leq y) \to \mathbf{P}(Y \leq y)$ at every y where F_Y is continuous.*

Uniqueness is what makes "pattern matching" legitimate: if a sum's transform is recognizably Poisson, the sum *is* Poisson (Example 43.9). The continuity theorem is the limit version — convergence of transforms implies convergence of CDFs ("convergence in distribution") — and it is exactly the bridge the CLT needs, converting a calculus limit of transforms into a probabilistic statement. Its characteristic-function form is Lévy's continuity theorem; the MGF form is due to Curtiss (see *Deeper reading*). One calculus lemma, and we are ready.

Lemma 43.7. *If $a_n \to a$ in $\mathbb{R}$, then $\left(1 + \frac{a_n}{n}\right)^n \to e^a$.*

Proof. Since (a_n) is bounded, $1 + a_n/n > 0$ for all large n, and $\ln(1 + x)/x \to 1$ as $x \to 0$ (the derivative of ln at 1 is 1). Hence

$$n \ln\left(1 + \frac{a_n}{n}\right) = a_n \cdot \frac{\ln(1 + a_n/n)}{a_n/n} \;\longrightarrow\; a \cdot 1 = a,$$

where the ratio is read as 1 whenever $a_n = 0$. Exponentiate; exp is continuous. □

Theorem 43.8 (Central limit theorem — Theorem 21.3, now proved)**.** *Let*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

(the sum, standardized to mean 0 and variance 1). Then for every z,

$$\mathbf{P}(Z_n \le z) \xrightarrow[n\to\infty]{} \Phi(z),$$

the standard normal CDF — regardless of the distribution of the X_i.

Proof. We prove the theorem under the assumption that the X_i have an MGF, finite on some $(-s_0, s_0)$. (The general case, finite variance only, runs the identical argument on the characteristic function, where finiteness is automatic.)

Step 1: standardize. Put $Y_i = (X_i - \mu)/\sigma$. The Y_i are i.i.d. with $\mathbf{E}[Y_i] = 0$ and $\mathbf{E}[Y_i^2] = \operatorname{var}(Y_i) = 1$, and $Z_n = (Y_1 + \cdots + Y_n)/\sqrt{n}$. By rule (a), $M_Y(s) = e^{-\mu s/\sigma} M_X(s/\sigma)$, finite on $(-\sigma s_0, \sigma s_0)$; rename that interval $(-s_0, s_0)$.

Step 2: the transform of Z_n. By rule (c) and then rule (a),

$$M_{Z_n}(s) = M_{Y_1+\cdots+Y_n}\big(s/\sqrt{n}\big) = \big(M_Y(s/\sqrt{n})\big)^n,$$

finite for $|s| < s_0\sqrt{n}$, in particular on $(-s_0, s_0)$ for every n. This is the heart of the proof: sum $\to$ product $\to$ n-th power.

Step 3: Taylor at 0. By rule (b), M_Y is twice differentiable at 0 with $M_Y(0) = 1$, $M_Y'(0) = \mathbf{E}[Y] = 0$, and $M_Y''(0) = \mathbf{E}[Y^2] = 1$. Taylor's theorem to second order (remainder in Peano form) gives

$$M_Y(t) = 1 + \frac{t^2}{2} + r(t), \qquad \frac{r(t)}{t^2} \to 0 \ \text{ as } t \to 0.$$

Step 4: the limit. For $s = 0$ both $M_{Z_n}(0)$ and e^0 equal 1. For $s \neq 0$ put $t = s/\sqrt{n}$, so $t \to 0$ as $n \to \infty$, and write

$$M_{Z_n}(s) = \Big(1 + \frac{s^2}{2n} + r\big(s/\sqrt{n}\big)\Big)^n = \Big(1 + \frac{a_n}{n}\Big)^n, \qquad a_n = \frac{s^2}{2} + n\,r(t) = \frac{s^2}{2} + s^2 \cdot \frac{r(t)}{t^2},$$

using $n = s^2/t^2$. Since $r(t)/t^2 \to 0$, $a_n \to s^2/2$, and Lemma 43.7 gives

$$M_{Z_n}(s) \longrightarrow e^{s^2/2} = M_Z(s), \qquad Z \sim \mathcal{N}(0,1),$$

by Proposition 43.2(d) — for every s, in particular for every $s \in (-s_0, s_0)$.

Step 5: from transforms to CDFs. Apply the continuity theorem (Theorem 43.6) with $Y_n = Z_n$ and $Y = Z$, whose transforms are all finite on $(-s_0, s_0)$: $\mathbf{P}(Z_n \le z) \to \Phi(z)$ at every continuity point of Φ — which is every z, since Φ is continuous. □

What the proof says. After standardization every distribution with an MGF looks like $1 + t^2/2 + o(t^2)$ near $t = 0$; the sum sees each summand only through its transform at the tiny argument $s/\sqrt{n}$, where nothing beyond the second moment survives, and raising to the n-th power turns $1 + \frac{s^2/2}{n}$ into $e^{s^2/2}$ — the Gaussian transform. The distribution of the X_i enters only through $r(t)$, which is killed (all of it, for every distribution) by the $o(1/n)$-per-factor scaling. That is why the limit is universal, and also why convergence is slower for skewed summands: the third-moment term in $r(t)$ is what the finite-n error is made of (the Berry–Esseen theorem makes this quantitative — the CDF error is at most an absolute constant times $\mathbf{E}|Y|^3/\sqrt{n}$).

43.2 Practice: by hand

Example 43.9 (Sums by transform)**.** (a) $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\nu)$, independent: $M_{X+Y}(s) = e^{\lambda(e^s-1)}e^{\nu(e^s-1)} = e^{(\lambda+\nu)(e^s-1)}$, the Poisson$(\lambda+\nu)$ transform, so $X+Y \sim \text{Poisson}(\lambda+\nu)$ by uniqueness — no convolution sum. (b) $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ independent:

$$\prod_i e^{\mu_i s + \sigma_i^2 s^2/2} = e^{(\sum_i \mu_i)s + (\sum_i \sigma_i^2)s^2/2},$$

so the sum is $\mathcal{N}(\sum_i \mu_i, \sum_i \sigma_i^2)$ — the one-dimensional case of Theorem 20.5(a), with no vector machinery. (c) n independent Bernoulli(p): $(1-p+pe^s)^n$ is the Binomial(n,p) transform; one derivative at $s=0$, $n\,pe^s(1-p+pe^s)^{n-1}\big|_{s=0} = np$, recovers the mean.

Example 43.10 (Chernoff for a Gaussian tail)**.** For $Z \sim \mathcal{N}(0,1)$ and $a > 0$: $sa - \ln M_Z(s) = sa - s^2/2$, maximized at $s = a$, so $\mathbf{P}(Z \geq a) \leq e^{-a^2/2}$. Compare the one-sided Chebyshev bound $1/(1+a^2)$ of Exercise 21.3 (Lesson 21): at $a = 3$, $e^{-4.5} \approx 0.011$ against 0.10; the truth, $1 - \Phi(3) \approx 0.00135$, has Chernoff's shape (exponential in a^2), not Chebyshev's (polynomial).

Fourier / statistical-inference connection. $M_X(s) = \int e^{sx} f_X(x)\,dx$ is the two-sided Laplace transform of the density (Lesson 30, with e^{sx} in place of e^{-st}), and the characteristic function is its Fourier transform (Lesson 27); the uniqueness theorem is the inversion integral of Lesson 38 applied to a density, and the continuity theorem is its limit form. Statistical inference builds on exactly these tools: Chernoff becomes Hoeffding and the large-deviations rate $\phi(a)$; the CLT's characteristic-function proof extends to random vectors (Lesson 20's Gaussian vectors) and to non-identically distributed summands (Lindeberg).

43.3 Exercises

Exercise 43.1 (Hand)**.** (a) From $M_X(s) = \lambda/(\lambda - s)$ for $X \sim \text{Exponential}(\lambda)$, compute $M_X'(0)$ and $M_X''(0)$ and read off $\mathbf{E}[X]$, $\mathbf{E}[X^2]$, and $\text{var}(X)$. (b) Write down the transform of $S_n = X_1 + \cdots + X_n$ for i.i.d. Exponential(λ) summands and obtain $\mathbf{E}[S_n]$ and $\text{var}(S_n)$ from it. (c) Derive the transform of Uniform(a,b) (continuous) from the definition and check $M(s) \to 1$ as $s \to 0$.

Exercise 43.2 (Hand)**.** Let $S \sim \text{Binomial}(100, \frac{1}{2})$, so $M_S(s) = \left(\frac{1+e^s}{2}\right)^{100}$. Bound $\mathbf{P}(S \geq 60)$ three ways — Chebyshev (Theorem 21.1(b), and the one-sided form of Exercise 21.3), the Chernoff bound with the best s (find it by setting the derivative of $-60s + 100\ln\frac{1+e^s}{2}$ to zero), and the CLT approximation — and rank the three against each other.

Exercise 43.3 (Proof)**.** Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbf{P}(X_i = 1) = \mathbf{P}(X_i = -1) = \frac{1}{2}$ (a fair coin scored ± 1), and $Z_n = (X_1 + \cdots + X_n)/\sqrt{n}$. (a) Show $M_X(s) = \cosh s$ and $M_{Z_n}(s) = \left(\cosh(s/\sqrt{n})\right)^n$. (b) Prove directly, without Theorem 43.8, that $M_{Z_n}(s) \to e^{s^2/2}$ for every s. (Hint: from the series, $1 + \frac{t^2}{2} \leq \cosh t \leq e^{t^2/2}$ for all t — the upper bound uses $(2k)! \geq 2^k k!$ — then sandwich with Lemma 43.7.)

Exercise 43.4 (Proof, $\star$)**.** (a) Show that if M_X is finite on $(-s_0, s_0)$, then $\mathbf{E}[|X|^k] < \infty$ for every $k \geq 0$. (Hint: for $0 < s < s_0$, $|x|^k \leq \frac{k!}{s^k} e^{s|x|} \leq \frac{k!}{s^k}\left(e^{sx} + e^{-sx}\right)$.) (b) Let X have density $f(x) = c/(1+x^4)$ on $\mathbb{R}$ (c the normalizing constant). Show $\mathbf{E}[X^2] < \infty$ but $M_X(s) = \infty$ for every $s \neq 0$. Conclude that the proof of Theorem 43.8 given above does not cover this X, even though the theorem itself does.

Deeper reading. B&T 4.4 (transforms: Examples 4.22–4.33 and the inversion property), 5.1–5.4 (inequalities, WLLN, convergence in probability, CLT), 5.5 (strong law); B&T Problems 5.2 (the Chernoff bound) and 5.12 (the CLT proof as an outline) are this lesson's Theory in the book's own words. For the two quoted theorems: J. H. Curtiss, "A note on the theory of moment generating functions," *Ann. Math. Statist.* 13 (1942) 430–433 (uniqueness and continuity in MGF form); Durrett, *Probability: Theory and Examples*, Chapter 3 (characteristic functions, Lévy's continuity theorem, the CLT proved that way, and the Berry–Esseen rate).

Part VII: Convex Optimization and Information Theory for ML, Signals, and Sensors

Every part of this booklet has quietly leaned on the same two unnamed disciplines. Lesson 6 minimized $\|Ax - b\|$; Course 3's Wiener filter (Lab 6.8) minimized a mean-square error; its filter labs (6.1–6.2) asked for the smallest ripple a length-N filter can achieve; its detection labs (6.7–6.9) asked how distinguishable two noisy hypotheses are; its converter and audio labs (3.4, 9.3) traded bits against distortion in converters and codecs. Each time, the question underneath was one of two: *among all designs meeting the constraints, which is best — and how would we certify it?* (optimization), and *how much can the data reveal at all, whatever the algorithm?* (information). Part VII names the two disciplines and installs their working cores:

1. What makes a set or a function convex, why does convexity turn searching into certifying, and how are least squares, LASSO, logistic regression, and Chebyshev filter design all the same kind of problem? (Lesson 44.)

2. What do penalty terms actually do to a solution — and why is the ridge term the exact cure for Lesson 42's ill-conditioning? (Lesson 45.)

3. What is a dual problem, what do multipliers mean, and how do the KKT conditions certify an answer and price a constraint? (Lesson 46 — where water-filling is solved by hand, to be spent in Lesson 50.)

4. What do training loops compute, why do they slow down exactly when eigenvalues spread, and how does a corner in $\|x\|_1$ become an exact zero? (Lesson 47.)

5. What, in bits, does a measurement carry? (Lesson 48: entropy, divergence, mutual information, and the inequalities — Jensen, data processing, Fano — that police them.)

6. How fast can hypotheses be distinguished, how many bits can a noisy channel or a noisy sensor deliver, and what floor does Fisher information put under every estimator? (Lesson 50.)

7. How few bits keep a source within a distortion budget, and what does that say about quantizers, codecs, and learned representations? (Lesson 51.)

Two sources, both self-contained here at working depth: Boyd–Vandenberghe, *Convex Optimization* (BV) for Lessons 31–38, and Polyanskiy–Wu, *Information Theory: From Coding to Learning* (PW) for Lessons 48–51. The ground rule is Part VI's: a proof appears when it teaches a mechanism you will reuse (Jensen, Cauchy–Schwarz, the KKT balance of forces); a theorem is stated with a pointer when its proof is a course of its own (Shannon's coding theorems). Prerequisites: Parts I, III, and V, with Lessons 19–21 (Part IV) wanted for the information half; the Course 3 labs make every example land harder but are not assumed. Conventions: log is base 2 when the unit is bits and natural when calculus is being done — each use says which; problems are minimizations unless stated; $h_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy, a function worth memorizing.

44 Convex Models: Sets, Functions, and Problems

NEEDED FOR: ML (losses, regularizers, margins), DSP (filter design), inference and sensors (estimation under constraints).

44.1 Theory

Definition 44.1 (Convex set)**.** $C \subseteq \mathbb{R}^n$ is *convex* if for all $x, y \in C$ and $t \in [0,1]$, $tx+(1-t)y \in C$: the segment between any two points stays inside.

The examples are the sets this booklet already lives in. Affine sets $\{x : Ax = b\}$ (solution sets of Lesson 5) are convex: $A(tx + (1-t)y) = tb + (1-t)b = b$. Halfspaces $\{x : a^\mathsf{T} x \le b\}$ are convex by linearity of the dot product. Norm balls $\{x : \|x - x_0\| \le r\}$ are convex by the triangle inequality — for *every* norm, so ℓ_1 diamonds and ℓ_∞ boxes count, not just Euclidean balls. The positive semidefinite matrices form a convex cone: $z^\mathsf{T}(tP + (1-t)Q)z \ge 0$ termwise (the set Lesson 10 certified covariances into, Theorem 20.3).

Proposition 44.2 (Intersection preserves convexity)**.** *An arbitrary intersection of convex sets is convex. In particular a polyhedron $\{x : Ax \le b,\ Cx = d\}$ — an intersection of halfspaces and hyperplanes — is convex, so* any *list of linear constraints defines a convex feasible set.*

Proof. If x, y lie in every C_α, then $tx + (1-t)y \in C_\alpha$ for each α (convexity of that one set), hence in the intersection. □

Definition 44.3 (Convex function)**.** $f : C \to \mathbb{R}$ on a convex domain is *convex* if for all $x, y \in C$, $t \in [0,1]$,

$$f(tx + (1-t)y) \ \le \ tf(x) + (1-t)f(y),$$

i.e. chords lie above the graph; *strictly convex* if the inequality is strict for $x \ne y$, $t \in (0,1)$; *concave* if $-f$ is convex.

Theorem 44.4 (First-order characterization: tangents are global underestimators)**.** *A differentiable f on a convex domain is convex if and only if*

$$f(y) \ \ge \ f(x) + \nabla f(x)^\mathsf{T}(y - x) \qquad \textit{for all } x, y.$$

Proof. ($\Rightarrow$) Convexity gives $f(x + t(y-x)) \le (1-t)f(x) + tf(y)$, so for $t \in (0,1]$

$$\frac{f(x + t(y-x)) - f(x)}{t} \ \le \ f(y) - f(x).$$

The left side tends to the directional derivative $\nabla f(x)^\mathsf{T}(y-x)$ as $t \downarrow 0$ (Lesson 7's definition of the gradient), and the right side does not move. ($\Leftarrow$) Let $z = tx + (1-t)y$ and write the tangent inequality at z twice: $f(x) \ge f(z) + \nabla f(z)^\mathsf{T}(x - z)$ and $f(y) \ge f(z) + \nabla f(z)^\mathsf{T}(y - z)$. Multiply by t and $1-t$ and add: the gradient terms cancel because $t(x-z) + (1-t)(y-z) = 0$, leaving $tf(x) + (1-t)f(y) \ge f(z)$. □

This one inequality is the engine of the whole part: it converts a *local* object (the gradient at x) into a *global* statement (a bound valid at every y). Everything that follows — global optimality, duality bounds, convergence proofs — is Theorem 44.4 wearing different clothes.

Theorem 44.5 (Second-order characterization)**.** *A twice-differentiable f on a convex domain is convex iff its Hessian satisfies $\nabla^2 f(x) \succeq 0$ everywhere.*

Proof. Restrict to a line: $g(t) = f(x + tv)$ has $g''(t) = v^\mathsf{T} \nabla^2 f(x + tv)\, v$. If $\nabla^2 f \succeq 0$ then every $g'' \geq 0$, so every restriction is a convex scalar function (its derivative is nondecreasing, which gives the tangent inequality by the mean value theorem), and convexity on every line is convexity. Conversely if $v^\mathsf{T} \nabla^2 f(x) v < 0$ for some x, v, the restriction along v is locally concave, violating the chord inequality nearby. □

So $f(x) = \frac{1}{2}\|Ax - b\|^2$ is convex — Hessian $A^\mathsf{T} A \succeq 0$ (Lesson 6) — and a quadratic $\frac{1}{2} x^\mathsf{T} Q x + c^\mathsf{T} x$ is convex exactly when $Q \succeq 0$, which the spectral theorem (Theorem 10.4) reads as "no negative eigenvalue directions."

Proposition 44.6 (The calculus of convexity)**.** *(a) Nonnegative combinations: if f_1, f_2 are convex and $\alpha, \beta \geq 0$, then $\alpha f_1 + \beta f_2$ is convex. (b) Affine substitution: if f is convex, so is $x \mapsto f(Ax + b)$. (c) Pointwise maximum: if $f_1, \dots, f_m$ are convex, so is $f(x) = \max_i f_i(x)$ — and likewise a supremum over an arbitrary family.*

Proof. (a) and (b) are direct substitutions into the chord inequality. For (c),

$$f(tx + (1-t)y) = \max_i f_i(tx + (1-t)y) \leq \max_i \big[t f_i(x) + (1-t) f_i(y)\big] \leq t f(x) + (1-t) f(y),$$

since a max of sums is at most the sum of maxes. □

Part (c) is the workhorse of signal processing models: a *worst-case* quantity — maximum passband ripple over frequencies, worst residual over measurements, largest eigenvalue of a symmetric matrix (a sup of the convex functions $x^\mathsf{T} P z \dots$ over unit z) — is convex *because* it is a sup of convex functions, even though it is not differentiable. Convex analysis is the calculus that survives corners.

Definition 44.7 (Convex optimization problem)**.**

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots, m, \\ & Ax = b, \end{array}$$

with $f_0, \dots, f_m$ convex and the equality constraints affine. The feasible set is then convex (Proposition 44.2 plus the fact that sublevel sets $\{f_i \leq 0\}$ of convex functions are convex). Its optimal value is $p^\star$.

Theorem 44.8 (Local implies global; the optimality condition)**.** *For a convex problem: (a) any locally optimal x is globally optimal. (b) If f_0 is differentiable, a feasible $x^\star$ is optimal iff*

$$\nabla f_0(x^\star)^\mathsf{T} (y - x^\star) \;\geq\; 0 \qquad \textit{for every feasible } y.$$

(c) If moreover the problem is unconstrained, (b) reduces to $\nabla f_0(x^\star) = 0$.

Proof. (a) Suppose x is locally optimal but some feasible y has $f_0(y) < f_0(x)$. The segment $x_t = (1-t)x + ty$ is feasible (convex feasible set) and $f_0(x_t) \leq (1-t) f_0(x) + t f_0(y) < f_0(x)$ for every $t \in (0, 1]$ — arbitrarily close to x, contradicting local optimality. (b) If the condition holds, Theorem 44.4 gives $f_0(y) \geq f_0(x^\star) + \nabla f_0(x^\star)^\mathsf{T}(y - x^\star) \geq f_0(x^\star)$ for all feasible y: done. If it fails for some y, then moving from $x^\star$ toward y decreases f_0 to first order while staying feasible. (c) Take $y = x^\star - \nabla f_0(x^\star)$; the condition forces $-\|\nabla f_0(x^\star)\|^2 \geq 0$. □

Part (c) closes a loop: setting the gradient of $\frac{1}{2}\|Ax-b\|^2$ to zero gave the normal equations (Corollary 7.4), and convexity is the reason that stationary point was the *global* minimum — a fact Lesson 6 asserted geometrically and can now certify in one line.

Recognition is the skill. Almost no problem announces itself in the standard form; it is *rewritten* into it. Three rewrites cover a remarkable fraction of practice:

- *Epigraph trick* — minimize a max by naming the max: $\min_x \max_i |a_i^\mathsf{T} x - b_i|$ becomes $\min_{x,t} t$ subject to $-t \le a_i^\mathsf{T} x - b_i \le t$. The nondifferentiable objective became $2m$ linear inequalities and a linear objective: a linear program (LP). This is exactly how minimax (equiripple) FIR design is posed — Course 3 Lab 6.1's Parks–McClellan specification is an LP in the coefficients, since $H(e^{j\omega})$ is linear in h at each frequency.
- *Absolute values split* — $|u| \le t$ is the pair $u \le t$, $-u \le t$; so $\min \sum_i |a_i^\mathsf{T} x - b_i|$ and anything built from $\|\cdot\|_1$ or $\|\cdot\|_\infty$ is an LP.
- *Composition with affine maps* (Proposition 44.6b) — the logistic loss $\sum_i \log(1 + e^{-y_i a_i^\mathsf{T} w})$ is convex in w because $s \mapsto \log(1+e^{-s})$ is convex in the scalar s (second derivative $e^s/(1+e^s)^2 > 0$) and $s = y_i a_i^\mathsf{T} w$ is affine. The same two-line argument certifies least squares, ridge, LASSO, SVMs with hinge loss, and maximum-likelihood fitting of any log-concave model (BV §7.1) — which is why "the training problem is convex" was true for every classical ML model before deep networks.

ML / DSP / sensors connection. Convexity is the boundary between "run the solver and trust the answer" and "pray over the initialization." For classical ML it means the fitted model is a property of the data and the loss, not of the optimizer's starting point; for deep learning it is the lost paradise whose vocabulary (loss surfaces, saddle points, conditioning) still frames every training discussion. For DSP, the epigraph trick turns worst-case specs — ripple, overshoot, sidelobe level — into constraints a solver can certify, and Lesson 42's structured-matrix machinery is what makes those solvers fast. For sensors and estimation, Part IV's estimators reappear as convex programs: least squares (Theorem 6.4), LLMS (Theorem 19.4), and, in Lesson 50, detector and experiment design.

44.2 Practice: by hand

Example 44.9 (Chebyshev fit as an LP, end to end)**.** Fit a scalar line $y \approx c_1 u + c_0$ to $(u, y) \in \{(0,0),(1,1),(2,3)\}$, minimizing the *worst* error. Epigraph form: minimize t subject to $|c_0 - 0| \le t$, $|c_0 + c_1 - 1| \le t$, $|c_0 + 2c_1 - 3| \le t$. At the optimum the extreme residuals alternate in sign and tie (the equioscillation of Course 3 Lab 6.1, in miniature): try $c_1 = \frac{3}{2}$, $c_0 = -\frac{1}{4}$, giving residuals $-\frac{1}{4}, +\frac{1}{4}, -\frac{1}{4}$ — so $t = \frac{1}{4}$, and no line can do better because the middle point pulls opposite to the outer two. Compare least squares, which would trade a larger worst error for smaller total square error.

Example 44.10 (A convexity certificate in two lines)**.** Is $f(x_1, x_2) = e^{x_1} + e^{x_2} + (x_1 - x_2)^2$ convex? e^{x_i} are convex (second derivative $e^{x_i} > 0$), $(x_1 - x_2)^2$ is a convex quadratic composed with the affine map $x \mapsto x_1 - x_2$, and sums of convex functions are convex (Proposition 44.6a,b). No Hessian computation needed — the calculus of convexity is usually faster than the calculus of derivatives.

44.3 Exercises

Exercise 44.1 (Hand)**.** Which are convex? (a) $\{x \in \mathbb{R}^2 : x_1 x_2 \geq 1,\ x_1 > 0\}$; (b) $\{x : \|x\|_2 \geq 1\}$; (c) $\{x : \max(x_1, x_2) \leq 1\}$. (Answers: (a) yes — it is $\{x_1 > 0, x_2 \geq 1/x_1\}$ and $1/x_1$ is convex there, so this is a region above a convex curve's graph; check the chord inequality directly; (b) no — the segment from $(1, 0)$ to $(-1, 0)$ passes through the hole; (c) yes — intersection of two halfspaces.)

Exercise 44.2 (Hand)**.** Rewrite $\min_x \sum_{i=1}^{k} |a_i^\mathsf{T} x - b_i|$ as an LP with variables (x, u), $u \in \mathbb{R}^k$: state the objective and all $2k$ inequalities. Then rewrite $\min_x \|Ax - b\|_\infty$ the same way with a single extra scalar t.

Exercise 44.3 (Proof, $\star$)**.** Prove Jensen's inequality by induction from the definition: if f is convex, $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$, then $f\big(\sum_i \theta_i x_i\big) \leq \sum_i \theta_i f(x_i)$; conclude $f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)]$ for a discrete random variable X (write the expectation as a convex combination, Lesson 15). This inequality is the sole engine of Lesson 48's $D \geq 0$, and through it of essentially all of information theory — prove it once, carefully.

Exercise 44.4 (Proof)**.** Show that the sublevel set $\{x : f(x) \leq \alpha\}$ of a convex function is convex, and give a nonconvex function all of whose sublevel sets are convex ($f(x) = \sqrt{|x|}$), so the converse fails. (Such f are called quasiconvex; BV §3.4.)

Deeper reading. BV 2.1–2.3 and 2.5 (convex sets, separating hyperplanes), 3.1–3.2 (convex functions and the calculus that preserves them), 4.1–4.4 (problem classes: LP, QP, QCQP, SOCP).

45 Regularized Fitting and Inverse Problems

NEEDED FOR: ML (ridge, LASSO, robustness), DSP and imaging (deconvolution, restoration), sensors (calibration, fusion).

45.1 Theory

Lesson 42 diagnosed the disease: when A has small singular values, the least-squares solution multiplies measurement noise by $1/\sigma_i$ and the answer drowns (deconvolution shows it happening). Regularization is the cure, and it is a *modeling* act, not a numerical trick: it states, in the objective, which solutions were plausible before the data arrived,

$$\min_x \ \tfrac{1}{2}\|Ax - y\|^2 + \lambda R(x),$$

with the data term demanding "explain the measurements," R pricing implausible answers, and λ the exchange rate between the two currencies.

Theorem 45.1 (Ridge regression: existence, uniqueness, and the SVD filter factors)**.** *For $\lambda > 0$ and $R(x) = \frac{1}{2}\|x\|^2$ (Tikhonov, Example 41.7), the minimizer is unique for every A:*

$$x_\lambda = (A^\mathsf{T} A + \lambda I)^{-1} A^\mathsf{T} y.$$

In the SVD $A = U\Sigma V^{\mathsf{T}}$ (Theorem 12.1) with singular values σ_i,

$$x_\lambda = \sum_i \underbrace{\frac{\sigma_i}{\sigma_i^2 + \lambda}}_{\textit{filter factor}} (u_i^{\mathsf{T}} y)\, v_i, \qquad \textit{versus} \qquad x_{\mathrm{LS}} = \sum_i \frac{1}{\sigma_i}(u_i^{\mathsf{T}} y)\, v_i.$$

Each data component $u_i^{\mathsf{T}} y$ is passed with gain $\sigma_i/(\sigma_i^2 + \lambda)$: $\approx 1/\sigma_i$ where $\sigma_i^2 \gg \lambda$ (strong directions untouched), $\approx \sigma_i/\lambda$ where $\sigma_i^2 \ll \lambda$ (weak directions muted instead of amplified). As $\lambda \downarrow 0$, x_λ tends to the minimum-norm least-squares solution; as $\lambda \to \infty$, to 0.

Proof. The objective is convex with Hessian $A^{\mathsf{T}}A + \lambda I \succ 0$ (add $\lambda > 0$ to each nonnegative eigenvalue of $A^{\mathsf{T}}A$), so it is strictly convex: the stationary point (Theorem 44.8c) is the unique global minimum, and setting the gradient $A^{\mathsf{T}}(Ax-y)+\lambda x$ to zero gives the formula. Substituting the SVD: $A^{\mathsf{T}}A + \lambda I = V(\Sigma^{\mathsf{T}}\Sigma + \lambda I)V^{\mathsf{T}}$ and $A^{\mathsf{T}}y = V\Sigma^{\mathsf{T}}U^{\mathsf{T}}y$, so the v_i coordinate of x_λ is $(\sigma_i^2 + \lambda)^{-1}\sigma_i(u_i^{\mathsf{T}}y)$. The two limits read off the filter factor: $\sigma_i/(\sigma_i^2 + \lambda) \to 1/\sigma_i$ for $\sigma_i > 0$ and $\to 0$ for $\sigma_i = 0$, which is exactly the minimum-norm pseudoinverse solution. □

The condition number of the ridge system is $(\sigma_{\max}^2+\lambda)/(\sigma_{\min}^2+\lambda)$ — capped near $\sigma_{\max}^2/\lambda$ no matter how singular A is. That is Definition 42.2's κ being *designed*, not suffered; the frequency-domain deconvolution with a damped gain $\overline{G}/(|G|^2 + \lambda)$, is Theorem 45.1 in the Fourier basis, where convolution's singular vectors are complex exponentials (Lesson 11).

Theorem 45.2 (Weighted least squares is Gaussian maximum likelihood)**.** *If $y = Ax + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \Sigma)$, $\Sigma \succ 0$, then the maximum-likelihood estimate of x is*

$$\hat{x} = \arg\min_x (Ax - y)^{\mathsf{T}}\Sigma^{-1}(Ax - y),$$

i.e. least squares in the whitened variables $\Sigma^{-1/2}y$, $\Sigma^{-1/2}A$.

Proof. The density of y given x (Lesson 20) is $c\exp\big(-\frac{1}{2}(y - Ax)^{\mathsf{T}}\Sigma^{-1}(y - Ax)\big)$ with c independent of x; maximizing the log-likelihood is minimizing the quadratic form. Factoring $\Sigma^{-1} = \Sigma^{-1/2}\Sigma^{-1/2}$ (spectral theorem, Theorem 10.4) exhibits it as $\|\Sigma^{-1/2}(Ax - y)\|^2$. □

So "weight each sensor by its inverse variance" is not a heuristic: it is the likelihood speaking. A high-noise sensor gets a small weight; *correlated* sensor errors get rotated apart before weighting. This is the estimation-side twin of the Wiener filter (Course 3 Lab 6.8) and the workhorse of every calibration and fusion problem downstream.

Priors, formally. Add a prior $x \sim \mathcal{N}(0, (\lambda^{-1})I \cdot s^2)$ and maximize the posterior instead: by Bayes' rule (Theorem 14.9) the log posterior is the log likelihood plus $-\frac{\lambda}{2s^2}\|x\|^2$ plus constants — ridge *is* Gaussian MAP estimation, and λ is a noise-to-prior-variance ratio, which is why good λ's track the noise floor (Exercise 45.4 makes this exact). Swap the Gaussian prior for a Laplacian $\propto e^{-\alpha\|x\|_1}$ and MAP becomes the LASSO.

The ℓ_1 penalty and why corners make zeros. With $R(x) = \|x\|_1 = \sum_i |x_i|$ the problem stays convex (Proposition 44.6) but loses differentiability at $x_i = 0$ — on purpose. Take the cleanest case $A = I$ (denoising): the problem separates into scalars,

$$\min_{x_i} \ \tfrac{1}{2}(x_i - y_i)^2 + \lambda|x_i|,$$

For $x_i > 0$ the derivative is $x_i - y_i + \lambda$, vanishing at $x_i = y_i - \lambda$, valid only if $y_i > \lambda$; symmetrically for $x_i < 0$; and $x_i = 0$ is optimal whenever the one-sided slopes disagree in sign, i.e. $|y_i| \le \lambda$. The solution is the *soft threshold*

$$x_i = \mathrm{soft}_\lambda(y_i) = \mathrm{sign}(y_i)\max(|y_i| - \lambda, 0) :$$

an entire interval of data maps to *exactly* zero. A smooth penalty can only shrink coordinates toward zero (ridge multiplies by $1/(1+\lambda)$, never reaching it); the corner is what makes sparsity an *output* of the optimization rather than a rounding step. Lesson 47 upgrades this scalar fact into an algorithm for general A.

Robustness is a loss choice. Squared error prices a residual r at r^2: one gross outlier outbids a hundred honest measurements. The ℓ_1 loss $\sum_i |r_i|$ prices linearly, so the fit resists outliers (its scalar optimum is a median, not a mean — Exercise 45.3); Huber's loss

$$\phi_{\mathrm{hub}}(r) = \begin{cases} r^2, & |r| \le M \\ M(2|r| - M), & |r| > M \end{cases}$$

is quadratic where the Gaussian story is believable and linear where it is not, and is convex (BV §6.1.2). Choosing among them is a statement about the failure modes of your sensor, encoded where the solver can act on it.

ML / imaging / sensors connection. Every entry of the modern fitting menu is this lesson's template with a different (R, loss) pair: ridge and LASSO regression, elastic net, total-variation denoising in imaging (Course 3 Lab 9.4's restoration problems with $R = \|\nabla x\|_1$), robust regression in computer vision, and the data-plus-prior structure of every Kalman update (the Kalman filter of Course 3 Lab 6.6 minimizes exactly a weighted-least-squares-plus-prior objective at each step). The bias–variance story is Lesson 42's κ made quantitative: raising λ buys conditioning with bias, and the U-shaped error curve of regularized deconvolution is the trade being paid.

45.2 Practice: by hand

Example 45.3 (Filter factors on a 2×2 problem)**.** $A = \mathrm{diag}(1,\ 0.1)$, $y = (1, 1)$, so $\sigma_1 = 1, \sigma_2 = 0.1$. Least squares: $x = (1, 10)$ — the weak direction blew up a unit measurement into 10. Ridge with $\lambda = 0.01$: filter factors $\frac{1}{1.01} \approx 0.99$ and $\frac{0.1}{0.02} = 5$, so $x_\lambda \approx (0.99,\ 5)$: the strong coordinate is barely touched, the weak one is halved. With $\lambda = 0.1$: $(0.91,\ 0.91)$ — now both are shrunk; λ has crossed σ_1^2's scale and is biasing directions the data actually supported. The art is $\sigma_{\min}^2 \ll \lambda \ll \sigma_{\max}^2$ when the spectrum allows it.

Example 45.4 (Two sensors, one constant)**.** Sensors report $y_1 = 10$ (variance 1) and $y_2 = 16$ (variance 9). Weighted least squares (Theorem 45.2) minimizes $(x-10)^2 + \frac{1}{9}(x-16)^2$, giving

$$\hat{x} = \frac{10 + 16/9}{1 + 1/9} = \frac{106}{10} = 10.6.$$

The clean sensor dominates $9:1$ — and the posterior variance, $(1 + \frac{1}{9})^{-1} = 0.9$, is *smaller than either sensor's alone*: fusion helped even though the second sensor was nine times noisier. (Compare Lesson 19: this is LLMS estimation with the prior removed.)

45.3 Exercises

Exercise 45.1 (Hand)**.** For the 2×2 example above, find the λ minimizing the actual error $\|x_\lambda - x_{\text{true}}\|$ when $x_{\text{true}} = (1, 1)$ and y was generated with noise $(0, +0.9)$ added to Ax_{true} — i.e. the second measurement is mostly noise. (Set up x_λ by filter factors and search a few values: $\lambda \approx 0.02$–0.1 does well; the point is that the best λ is near the noise-to-signal scale, not near machine epsilon.)

Exercise 45.2 (Hand)**.** Soft-threshold the vector $y = (3, -0.4, 0.2, -2)$ with $\lambda = 0.5$ and with $\lambda = 1$; then apply ridge shrinkage $y/(1 + \lambda)$ with the same λ's. Which coordinates become exactly zero under each rule, and which merely small?

Exercise 45.3 (Hand)**.** Show that $\hat{c} = \arg\min_c \sum_i |y_i - c|$ is any median of $\{y_i\}$: compute the one-sided derivatives of the objective in c and find where the count of y_i above c balances the count below. This is why ℓ_1 data terms resist outliers: the optimum moves with the *ranks*, not the values.

Exercise 45.4 (Proof, $\star$)**.** Prove Theorem 45.1's filter-factor formula from scratch (substitute the SVD, use $V^\mathsf{T}V = I$), then prove the MAP claim: for $y = Ax + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, prior $x \sim \mathcal{N}(0, \tau^2 I)$, the posterior mode minimizes $\frac{1}{2\sigma^2}\|Ax - y\|^2 + \frac{1}{2\tau^2}\|x\|^2$, so $\lambda = \sigma^2/\tau^2$. Conclude: doubling the measurement noise should quadruple... no — work out exactly how λ should move, and why.

Deeper reading. BV 6.1–6.4 (norm approximation, least-norm, regularized and robust approximation — the Huber discussion is 6.1.2), 6.5 (function fitting), 7.1 (maximum likelihood, including logistic regression as a convex ML problem). For the numerical side, Lesson 42; for the statistical side, Lessons 19 and 20.

46 Duality, KKT Conditions, and Certificates

NEEDED FOR: ML (SVMs, kernels, sparsity certificates), communications (power allocation), sensors (resource trade-offs, constrained estimation).

46.1 Theory

A solver's answer to "minimize f_0" is a number $f_0(\hat{x})$ — an *upper* bound on $p^\star$, witnessed by $\hat{x}$. How would you ever know you are close, without knowing $p^\star$? You need lower bounds, and duality is a machine that manufactures them.

Definition 46.1 (Lagrangian and dual function)**.** For the problem of Definition 44.7 written with general constraints $f_i(x) \le 0$ $(i = 1, \dots, m)$ and $h_j(x) = 0$ $(j = 1, \dots, p)$, the *Lagrangian* is

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x),$$

and the *dual function* is $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$, with $\lambda \succeq 0$ required and ν free.

Theorem 46.2 (Weak duality, and concavity of the dual)**.** *(a) For every $\lambda \succeq 0$ and every ν, $g(\lambda, \nu) \le p^\star$. (b) g is concave in (λ, ν) — always, even when the primal problem is not convex.*

Proof. (a) Let x be any feasible point. Then $\lambda_i f_i(x) \le 0$ (nonnegative multiplier times nonpositive constraint) and $\nu_j h_j(x) = 0$, so $L(x,\lambda,\nu) \le f_0(x)$. Hence $g(\lambda,\nu) = \inf_z L(z,\lambda,\nu) \le L(x,\lambda,\nu) \le f_0(x)$; take the infimum over feasible x. (b) For each fixed x, $L(x,\lambda,\nu)$ is *affine* in (λ,ν). So g is a pointwise infimum of affine functions, which is concave by (the concave version of) Proposition 44.6c. □

So every choice of nonnegative multipliers is a *certificate*: if a solver returns $\hat{x}$ and someone hands you (λ,ν) with $g(\lambda,\nu) = f_0(\hat{x}) - 10^{-9}$, then $\hat{x}$ is provably within 10^{-9} of optimal — no faith in the solver required. The *dual problem* is to find the best certificate: maximize $g(\lambda,\nu)$ over $\lambda \succeq 0$; its optimal value $d^\star \le p^\star$, and the gap $p^\star - d^\star$ is the *duality gap*.

Theorem 46.3 (Strong duality; Slater's condition)**.** *If the primal problem is convex and some strictly feasible point exists ($f_i(\tilde{x}) < 0$ for all i, $A\tilde{x} = b$), then $d^\star = p^\star$: the best certificate is exact. (Proof: BV §5.3.2, a separating-hyperplane argument; the geometry — the set of achievable (constraint, objective) pairs is convex, and a nonvertical supporting hyperplane at the optimum* is *the multiplier vector — is worth absorbing even without the proof.)*

Theorem 46.4 (KKT conditions)**.** *Suppose $f_0,\dots,f_m$ are differentiable. If strong duality holds and $x^\star, (\lambda^\star,\nu^\star)$ are primal and dual optimal, then*

$$
\begin{aligned}
&\textit{(i) } f_i(x^\star) \le 0, \quad h_j(x^\star) = 0 && \textit{(primal feasibility)}\\
&\textit{(ii) } \lambda_i^\star \ge 0 && \textit{(dual feasibility)}\\
&\textit{(iii) } \lambda_i^\star f_i(x^\star) = 0 \ \textit{ for each } i && \textit{(complementary slackness)}\\
&\textit{(iv) } \nabla f_0(x^\star) + \textstyle\sum_i \lambda_i^\star \nabla f_i(x^\star) + \sum_j \nu_j^\star \nabla h_j(x^\star) = 0 && \textit{(stationarity).}
\end{aligned}
$$

Conversely, for a convex *problem, any (x,λ,ν) satisfying (i)–(iv) is primal–dual optimal with zero gap: KKT points are certified optima.*

Proof. Forward direction: with zero gap,

$$f_0(x^\star) = g(\lambda^\star,\nu^\star) = \inf_z L(z,\lambda^\star,\nu^\star) \le L(x^\star,\lambda^\star,\nu^\star) = f_0(x^\star) + \sum_i \lambda_i^\star f_i(x^\star) \le f_0(x^\star),$$

using feasibility at the last step. So both inequalities are equalities. Equality in the second says $\sum_i \lambda_i^\star f_i(x^\star) = 0$; each term is ≤ 0, so each is 0: that is (iii). Equality in the first says $x^\star$ minimizes $L(\cdot,\lambda^\star,\nu^\star)$ over all z, and L is differentiable, so its gradient vanishes there: that is (iv). Converse: (iv) says x minimizes the convex function $L(\cdot,\lambda,\nu)$ (convex since $\lambda \succeq 0$ weights convex f_i), so $g(\lambda,\nu) = L(x,\lambda,\nu) = f_0(x) + \sum \lambda_i f_i(x) = f_0(x)$ by (iii). A feasible point whose objective equals a dual value is optimal by weak duality. □

Read (ii)–(iv) as mechanics. At the optimum, the objective's downhill force $-\nabla f_0$ is balanced by constraint forces $\lambda_i \nabla f_i$ pushing back along the walls' normals; a wall can only push ($\lambda_i \ge 0$), and only if you are touching it ($\lambda_i = 0$ whenever $f_i(x^\star) < 0$). Every fact about SVMs' support vectors, active sets, and sparse solutions is this picture.

Theorem 46.5 (Multipliers are prices)**.** *Let $p^\star(u)$ be the optimal value when the constraints are relaxed to $f_i(x) \le u_i$. For a convex problem with strong duality, $p^\star(u)$ is convex in u, and*

$$p^\star(u) \ \ge \ p^\star(0) - \lambda^{\star\mathsf{T}} u \qquad \textit{for all } u;$$

if $p^\star$ is differentiable at 0, then $\lambda_i^\star = -\partial p^\star / \partial u_i$.

Proof. Fix u and let x be feasible for the relaxed problem. Then $f_0(x) \geq L(x, \lambda^\star, \nu^\star) - \sum_i \lambda_i^\star f_i(x) \geq g(\lambda^\star, \nu^\star) - \lambda^{\star\mathsf{T}} u = p^\star(0) - \lambda^{\star\mathsf{T}} u$, using $f_i(x) \leq u_i$, $\lambda^\star \succeq 0$, and strong duality at $u = 0$. Minimize the left side over feasible x. The global affine lower bound touching at $u = 0$ forces the gradient claim when the derivative exists (and convexity of $p^\star$ follows by the usual perturbation argument, BV §5.6.1). □

This is the theorem that makes duality an *engineering* tool: $\lambda_i^\star$ is the exact marginal value of relaxing constraint i. If the power budget's multiplier is 0.3 bits per joule, a slightly bigger battery is worth exactly that much capacity — computed for free, from the solve you already did.

Water-filling, by hand. The classic worked KKT example, and one we will spend in Lesson 50: allocate power P across n parallel channels with noise levels N_i to maximize total capacity,

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n \log\bigl(1 + p_i/N_i\bigr) \\ \text{subject to} & p \succeq 0, \quad \mathbf{1}^\mathsf{T} p = P. \end{array}$$

KKT (with multiplier ν on the sum and $\mu_i \geq 0$ on $-p_i \leq 0$): stationarity gives $\frac{1}{N_i + p_i} = \nu - \mu_i$; complementary slackness kills μ_i wherever $p_i > 0$. So active channels satisfy $N_i + p_i = 1/\nu$ — a common water level — and channels with $N_i \geq 1/\nu$ get nothing:

$$p_i^\star = \max\bigl(1/\nu - N_i,\ 0\bigr), \qquad \nu \text{ chosen so } \textstyle\sum_i p_i^\star = P.$$

Pour water of total volume P into a tank whose floor is the noise profile N_i: it settles at a common surface, and channels whose floor pokes above the surface stay dry. One picture, one page of KKT, and it is optimal — certified, not plausible.

ML / communications / sensors connection. The SVM's margin problem has a dual whose variables live one-per-training-point; complementary slackness says only points touching the margin get $\lambda_i > 0$ — the *support vectors* — and the data enter the dual only through inner products, which is the door kernels walk through. The LASSO's optimality conditions certify which coefficients are zero (the dual constraint is slack there). In communications, water-filling above is how OFDM loads subcarriers (Course 3 Labs 6.7–6.9); in sensor design, multipliers price the watt/bit/false-alarm trade-offs that define an operating point — the sensitivity theorem is why those prices are believable numbers rather than folklore.

46.2 Practice: by hand

Example 46.6 (Projection onto the nonnegative axis, with forces)**.** Minimize $\frac{1}{2}(x-a)^2$ subject to $-x \leq 0$. Lagrangian $L = \frac{1}{2}(x-a)^2 - \lambda x$; KKT: $x - a - \lambda = 0$, $\lambda \geq 0$, $x \geq 0$, $\lambda x = 0$. Case $x > 0$: then $\lambda = 0$, $x = a$ — consistent iff $a > 0$. Case $x = 0$: then $\lambda = -a$ — consistent iff $a \leq 0$. So $x^\star = \max(a, 0)$ and $\lambda^\star = \max(-a, 0)$: the multiplier is precisely how hard the wall must push to stop the objective, and Theorem 46.5 predicts $dp^\star/du = -\lambda^\star$ — relaxing $x \geq 0$ to $x \geq -u$ helps at rate $|a|$ exactly when $a < 0$. Check it: for $a < 0$ and small $u \geq 0$, $p^\star(u) = \frac{1}{2}(a+u)^2$; differentiate at $u = 0$ and see.

Example 46.7 (An equality-constrained solve)**.** Minimize $x_1^2 + x_2^2$ subject to $x_1 + x_2 = 1$. Stationarity: $(2x_1, 2x_2) + \nu(1, 1) = 0$, so $x_1 = x_2 = -\nu/2$; feasibility forces $x_1 = x_2 = \frac{1}{2}$,

$\nu = -1$. The dual function: $g(\nu) = \inf_x\big(x_1^2 + x_2^2 + \nu(x_1 + x_2 - 1)\big) = -\frac{\nu^2}{2} - \nu$, maximized at $\nu = -1$ with $d^\star = \frac{1}{2} = p^\star$: zero gap, as Theorem 46.3 promised (affine constraints need no strict feasibility). Geometrically this is Lesson 6's projection of the origin onto a line, and ν is the length of the perpendicular force — duality keeps recovering Part I's geometry with prices attached.

46.3 Exercises

Exercise 46.1 (Hand)**.** Solve $\min \frac{1}{2}(x-3)^2$ s.t. $x \le 1$ by KKT (both cases), report $(x^\star, \lambda^\star)$, and verify the sensitivity prediction by computing $p^\star(u)$ for the relaxed constraint $x \le 1 + u$ and differentiating at $u = 0$. (Answer: $x^\star = 1$, $\lambda^\star = 2$, $p^\star(u) = \frac{1}{2}(2-u)^2$, $\frac{d}{du}p^\star(0) = -2$.)

Exercise 46.2 (Hand)**.** Water-fill $P = 6$ over noise floors $N = (1, 2, 5)$ (natural log): find $1/\nu$, the allocation, and which channel is dry. Then raise P until the dry channel first turns on. (Answer: try all-wet: $1/\nu = (6+8)/3 = 14/3$, $p = (11, 8, -1)/3$ — infeasible; two-wet: $1/\nu = (6+3)/2 = 4.5$, $p = (3.5, 2.5, 0)$, and channel 3 stays dry until $1/\nu$ reaches 5, i.e. $P = (5-1) + (5-2) = 7$.)

Exercise 46.3 (Proof, $\star$)**.** Prove weak duality and the concavity of g (Theorem 46.2) without looking, then derive the dual of the LP $\min c^\mathsf{T}x$ s.t. $Ax \le b$: show $g(\lambda) = -b^\mathsf{T}\lambda$ if $A^\mathsf{T}\lambda + c = 0$ and $-\infty$ otherwise, so the dual is $\max -b^\mathsf{T}\lambda$ s.t. $A^\mathsf{T}\lambda + c = 0$, $\lambda \succeq 0$. (LP duality, which Part I's Farkas-flavored geometry has been circling all along.)

Exercise 46.4 (Proof)**.** Verify that the water-filling allocation satisfies all four KKT conditions, and use Theorem 46.5 to interpret ν: show that $dC^\star/dP = \nu$, the marginal capacity per watt — the number a link designer actually quotes.

Deeper reading. BV 5.1–5.2 (dual function and problem), 5.3–5.4 (geometric and saddle-point views), 5.5 (KKT — water-filling is §5.5.3), 5.6 (perturbation and sensitivity), 5.7 (examples worth reading with a pencil).

47 Algorithms: Gradient, Newton, and Proximal Steps

NEEDED FOR: ML (training dynamics, conditioning), embedded DSP and sensors (real-time and large-scale solving).

47.1 Theory

Lessons 31–36 say what the answer means; this lesson is how machines find it, and — more valuable day to day — *why they are slow when they are slow.* The template is

$$x_{k+1} = x_k + t_k \Delta x_k,$$

a direction Δx_k and a step size t_k (chosen by a line search: backtrack from $t = 1$ until the decrease matches a fixed fraction of what the tangent predicted, BV §9.2). Gradient descent takes $\Delta x = -\nabla f(x)$ — Lesson 7's steepest descent. Its behavior is completely computable on the problem that locally approximates every smooth problem: a quadratic.

Theorem 47.1 (Gradient descent on a quadratic: mode-by-mode contraction)**.** *Let* $f(x) = \frac{1}{2}x^{\mathsf{T}}Qx - b^{\mathsf{T}}x$ *with* $Q \succ 0$*, eigenvalues* $0 < \lambda_{\min} = \lambda_1 \le \cdots \le \lambda_n = \lambda_{\max}$*, and minimizer* $x^\star = Q^{-1}b$*. With fixed step* t*, the error* $e_k = x_k - x^\star$ *obeys, in the eigenbasis of* Q *(Theorem 10.4),*

$$e_{k+1}^{(i)} = (1 - t\lambda_i)\, e_k^{(i)} :$$

each eigenmode contracts independently by its own factor. Convergence for all initializations iff $t < 2/\lambda_{\max}$*; the best fixed step* $t^\star = 2/(\lambda_{\min} + \lambda_{\max})$ *yields worst-mode factor*

$$\max_i |1 - t^\star\lambda_i| = \frac{\kappa - 1}{\kappa + 1}, \qquad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Proof. $\nabla f(x) = Qx - b = Q(x - x^\star)$, so $e_{k+1} = e_k - tQe_k = (I - tQ)e_k$. Diagonalize $Q = U\Lambda U^{\mathsf{T}}$ and write e in the eigenbasis: the iteration matrix is diagonal with entries $1 - t\lambda_i$. All factors have modulus < 1 iff $0 < t\lambda_i < 2$ for every i, i.e. $t < 2/\lambda_{\max}$. The optimal t balances the two extreme modes, $1 - t\lambda_{\min} = -(1 - t\lambda_{\max})$, giving $t^\star$ and, on substitution, the factor $(\kappa - 1)/(\kappa + 1)$. □

Everything practical is in that factor. $\kappa = 10$: factor 0.82, fine. $\kappa = 10^4$ (a badly scaled feature matrix, Lesson 42): factor 0.9998 — ten thousand iterations per digit. And it is the *same* eigenvalue spread that throttles the LMS filter's adaptation (Course 3 Lab 6.8: LMS is exactly stochastic gradient descent on Lesson 19's MSE quadratic, and its modes converge at rates set by the input correlation matrix's eigenvalues). Preconditioning, feature normalization, momentum, whitening: all are attacks on κ. For general smooth strongly convex f ($mI \preceq \nabla^2 f \preceq MI$) the same linear convergence holds with $\kappa = M/m$ (BV §9.3) — the quadratic is not a toy, it is the local truth.

Newton's method. Gradient descent is blind to curvature; Newton's step

$$\Delta x_{\text{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

solves the local quadratic model exactly — on a true quadratic it lands on $x^\star$ in one step from anywhere ($x - Q^{-1}(Qx - b) = Q^{-1}b$).

Proposition 47.2 (Newton is affinely invariant)**.** *Let* T *be invertible and* $\tilde{f}(y) = f(Ty)$*. Newton iterates for* $\tilde{f}$ *started at* $y_0 = T^{-1}x_0$ *satisfy* $y_k = T^{-1}x_k$ *for all* k*: Newton's method* cannot see *linear changes of coordinates, so no preconditioning is needed and* κ *disappears from its convergence story.*

Proof. $\nabla\tilde{f}(y) = T^{\mathsf{T}}\nabla f(Ty)$ and $\nabla^2\tilde{f}(y) = T^{\mathsf{T}}\nabla^2 f(Ty)\,T$, so

$$\Delta y = -\big(T^{\mathsf{T}}\nabla^2 f\, T\big)^{-1} T^{\mathsf{T}}\nabla f = -T^{-1}\nabla^2 f^{-1}\nabla f = T^{-1}\Delta x,$$

and induction carries the correspondence forward. □

The price is solving $\nabla^2 f\, \Delta x = -\nabla f$ each step — a structured linear system, which is where Lesson 42 earns its keep (Cholesky on the Hessian, or exploiting bandedness/circulance; BV Appendix C). Near the optimum Newton converges *quadratically* — digits double per iteration — which is why interior-point solvers finish in tens of iterations regardless of scale. With

equality constraints $Ax = b$, stationarity of the Lagrangian plus linearized feasibility give the *KKT system*

$$\begin{bmatrix} \nabla^2 f(x) & A^\mathsf{T} \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ w \end{bmatrix} = - \begin{bmatrix} \nabla f(x) \\ Ax - b \end{bmatrix},$$

Lesson 46's balance of forces, refreshed at every iterate. Inequalities are handled by the *barrier method*: replace $f_i(x) \le 0$ by a cost $-\frac{1}{t}\sum_i \log(-f_i(x))$ that soars at the boundary, solve by Newton, increase t, repeat; the central path it follows has duality gap exactly m/t (BV §11.2), so the barrier's own multipliers certify how far you are from optimal at every moment. That is the algorithm inside CVXPY.

Corners, handled exactly: proximal steps. For composite objectives $f(x) + \lambda\|x\|_1$ the gradient half is smooth and the corner half has a closed-form minimizer — Lesson 45's soft threshold. The *proximal gradient* method (ISTA) alternates them:

$$x_{k+1} = \mathrm{soft}_{\lambda t}\big(x_k - t\nabla f(x_k)\big),$$

a gradient step on the smooth part, then the exact scalar shrinkage on each coordinate. Because the threshold sets small coordinates to *exactly* zero, the iterates are genuinely sparse from early on — the algorithm output inherits the model's corner, which no amount of plain gradient descent on a smoothed $\|x\|_1$ would achieve. Convergence holds for $t \le 1/\lambda_{\max}(A^\mathsf{T}A)$ when $f = \frac{1}{2}\|Ax - y\|^2$, by the same eigenvalue logic as Theorem 47.1.

ML / DSP connection. Training a model *is* this lesson: SGD is Theorem 47.1 with noisy gradients (and Lesson 21's averaging is why the noise washes out); learning-rate limits are $t < 2/\lambda_{\max}$ wearing a tuning-guide costume; ill-conditioned features slow training by exactly $(\kappa - 1)/(\kappa + 1)$; batch normalization is preconditioning. On the DSP side, LMS (Course 3 Lab 6.8) is SGD avant la lettre, ISTA is how sparse deconvolution and compressed-sensing reconstructions actually run, and the KKT system's block structure is why real-time MPC and array-processing solvers can hit kilohertz rates on embedded hardware (Course 3's platforms).

47.2 Practice: by hand

Example 47.3 (Watching the slow mode)**.** $f(x) = \frac{1}{2}(10x_1^2 + x_2^2)$: $\lambda = (10, 1)$, $\kappa = 10$. From $x_0 = (1, 1)$ with $t = 0.1$: x_1-mode factor $1 - 1 = 0$, x_2-mode factor $1 - 0.1 = 0.9$. One step kills the stiff coordinate completely — and then the algorithm crawls along the valley floor at 0.9 per step: $x_k = (0,\, 0.9^k)$. The picture (steep walls, shallow floor) is the universal cartoon of ill-conditioned descent; optimal $t^\star = 2/11$ would balance the modes at factor $9/11 \approx 0.82$.

Example 47.4 (One proximal step, by hand)**.** $f = \frac{1}{2}\|x - y\|^2$, $y = (3, -0.4, 0.2)$, $\lambda = 1$, $t = 1$: gradient step lands on y itself, and $\mathrm{soft}_1(y) = (2,\, 0,\, 0)$. One iteration, two exact zeros, and the survivor pulled toward zero by exactly λt — denoiser, sparsifier, and proximal map are one operation.

47.3 Exercises

Exercise 47.1 (Hand)**.** For $f(x) = \frac{1}{2}(100x_1^2 + x_2^2)$: find the largest stable fixed step, the optimal fixed step, and the number of iterations of optimal-step gradient descent needed to shrink

the slow mode by 10^{-3}. Then report Newton's iteration count for the same task. (Answers: $t < 2/100$; $t^\star = 2/101$; factor $99/101$ needs $\approx \ln(10^{-3})/\ln(99/101) \approx 345$ steps; Newton: one.)

Exercise 47.2 (Hand)**.** Run two ISTA iterations by hand on $\min_x \frac{1}{2}(x-4)^2+|x|$ with $t = \frac{1}{2}$ from $x_0 = 0$, then solve the problem exactly by the soft-threshold formula and compare. (Iterates: $x_1 = \text{soft}_{1/2}(2) = 1.5$, $x_2 = \text{soft}_{1/2}(1.5 + \frac{1}{2}(4 - 1.5)) = \text{soft}_{1/2}(2.75) = 2.25$; exact answer $\text{soft}_1(4) = 3$, and the iterates are climbing toward it at rate $1 - t = \frac{1}{2}$.)

Exercise 47.3 (Proof, $\star$)**.** Prove that $\text{prox}_{\alpha|\cdot|}(z) = \arg\min_x \frac{1}{2}(x-z)^2+\alpha|x|$ is the soft threshold, by the three-case analysis of Lesson 45 — then prove the vector statement: for $R(x) = \alpha\|x\|_1$ the prox acts coordinatewise. (Separability of both the quadratic and the ℓ_1 norm is the whole proof; note where it would *fail* for $\|x\|_2$.)

Exercise 47.4 (Proof)**.** Prove Proposition 47.2's claim that gradient descent is *not* affinely invariant: exhibit the transformed gradient step and show it equals T^{-1} times the original step only when $TT^\mathsf{T} = I$. (Moral: gradient descent depends on the inner product you silently chose; Newton does not. Whitening data is choosing a better inner product.)

Deeper reading. BV 9.1–9.5 (descent, exact rates, Newton), 10.1–10.2 (equality-constrained Newton and the KKT system), 11.1–11.3 (barrier and central path); Appendix C for the linear algebra inside each iteration. Proximal methods are post-BV: this lesson's ISTA is the entry point.

48 Entropy, KL Divergence, and Mutual Information

NEEDED FOR: ML (cross-entropy losses, information gain), inference and sensors (uncertainty accounting), communications (the currency of Lessons 50–51).

48.1 Theory

Optimization measured cost in whatever units the objective carried. Information theory's first move is more radical: it puts a *unit on uncertainty itself*, and then proves that the unit is conserved, spent, and lost in lawful ways.

Definition 48.1 (Entropy)**.** For a discrete random variable X with pmf p (Lesson 15),

$$H(X) = -\sum_x p(x)\log_2 p(x) = \mathbf{E}\big[\log_2 \tfrac{1}{p(X)}\big] \quad \text{bits},$$

with $0\log 0 = 0$. $\log \frac{1}{p(x)}$ is the *surprise* of the outcome x — rare outcomes surprise more — and H is the average surprise, equivalently the average number of yes/no questions an optimal strategy needs to learn X (Lesson 51 makes "questions" into "code bits" exactly).

Definition 48.2 (KL divergence)**.** For pmfs p, q on the same alphabet,

$$D(P\|Q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = \mathbf{E}_P\Big[\log\frac{p(X)}{q(X)}\Big],$$

($+\infty$ if p puts mass where q has none). It is the average log-likelihood advantage of the true model P over a rival Q, per observation — not a distance ($D(P\|Q) \neq D(Q\|P)$ in general), but the exact exchange rate at which data accumulates evidence, as Lesson 50 will prove.

Theorem 48.3 (Information inequality, and the two extremes of entropy)**.** *(a) $D(P\|Q) \geq 0$, with equality iff $P = Q$. (b) For an alphabet of size K: $0 \leq H(X) \leq \log_2 K$, with the maximum exactly at the uniform distribution and the minimum exactly at deterministic X.*

Proof. (a) $-D(P\|Q) = \sum_x p(x) \log \frac{q(x)}{p(x)} \leq \log \sum_x p(x) \frac{q(x)}{p(x)} \leq \log 1 = 0$, where the first step is Jensen's inequality (Exercise 44.3) applied to the concave log: $\mathbf{E}[\log Z] \leq \log \mathbf{E}[Z]$ with $Z = q(X)/p(X)$. Equality in Jensen for a strictly concave function forces Z constant, and a constant likelihood ratio with both sides summing to 1 forces $P = Q$. (b) Take Q uniform: $0 \leq D(P\|Q) = \log_2 K - H(X)$. And $H \geq 0$ because every term $-p \log p \geq 0$, vanishing only at $p \in \{0, 1\}$. □

One inequality, and it never stops paying: everything below — cross-entropy's lower bound, conditioning reducing entropy, data processing, channel converses, compression limits — is Theorem 48.3a applied to a well-chosen pair.

Cross-entropy, and what ML actually minimizes. Splitting the log in the *cross-entropy* $H(P, Q) = -\mathbf{E}_P[\log q(X)]$ gives

$$H(P, Q) = H(P) + D(P\|Q) \;\geq\; H(P).$$

Train a classifier by minimizing cross-entropy between the empirical labels and the model's predictive q: the $H(P)$ term is fixed by the data, so the training loss is *exactly* $D(P\|Q)$ plus a constant — the model is being fitted in KL, and maximum likelihood (Lesson 45) is the same statement read right to left. The loss on your training dashboard has been an information quantity all along.

Definition 48.4 (Conditional entropy and mutual information)**.** $H(X \mid Y) = \sum_y p(y) H(X \mid Y = y)$, the average residual uncertainty after seeing Y; and

$$I(X; Y) = D\big(P_{XY} \,\|\, P_X P_Y\big) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

The three faces are equal by expanding logs (Exercise 48.2 walks one identity), and each earns its keep: the KL form makes $I \geq 0$ automatic (Theorem 48.3a with the pair $(P_{XY}, P_X P_Y)$) and shows $I(X; Y) = 0$ iff X, Y independent (Lesson 16's product criterion, Theorem 16.4); the second face reads "information is uncertainty destroyed"; the third says the destruction is symmetric. Two corollaries fall out free: $H(X \mid Y) \leq H(X)$ — *conditioning reduces entropy on average* — and the chain rule $H(X, Y) = H(X) + H(Y \mid X)$ (split $\log \frac{1}{p(x,y)}$).

Theorem 48.5 (Data-processing inequality)**.** *If $X \to Y \to Z$ is a Markov chain (Lesson 23: Z depends on X only through Y), then*

$$I(X; Z) \;\leq\; I(X; Y).$$

No processing of Y — deterministic or random, clever or dumb — can increase the information it carries about X.

Proof. Expand $I(X; Y, Z)$ by the chain rule for mutual information ($I(X; Y, Z) = I(X; Z) + I(X; Y \mid Z) = I(X; Y) + I(X; Z \mid Y)$, each line the entropy chain rule applied twice). Markovity says X and Z are conditionally independent given Y, so $I(X; Z \mid Y) = 0$; and $I(X; Y \mid Z) \geq 0$ as a (conditional) mutual information. Therefore $I(X; Z) \leq I(X; Y)$. □

Theorem 48.6 (Fano's inequality)**.** *Let X take K values, let $\hat{X} = g(Y)$ be* any *estimator of X from Y, and let $P_e = \mathbf{P}(\hat{X} \neq X)$. Then*

$$H(X \mid Y) \;\leq\; h_2(P_e) + P_e \log_2(K-1).$$

Contrapositive, the useful direction: if $H(X \mid Y)$ is large, no algorithm — present or future — can guess X from Y reliably: $P_e \;\geq\; \big(H(X \mid Y) - 1\big)/\log_2 K$.

Proof. Let $E = \mathbf{1}\{\hat{X} \neq X\}$. Chain rule twice on $H(E, X \mid Y)$:

$$H(X \mid Y) \leq H(E, X \mid Y) = H(E \mid Y) + H(X \mid E, Y) \leq h_2(P_e) + H(X \mid E, Y).$$

Given $E = 0$, $X = g(Y)$ is determined: zero entropy. Given $E = 1$, X is one of the $K-1$ values other than $g(Y)$: entropy at most $\log_2(K-1)$ (Theorem 48.3b). So $H(X \mid E, Y) \leq (1 - P_e) \cdot 0 + P_e \log_2(K-1)$. □

Fano is the theorem that turns information accounting into *impossibility proofs*: compute (or bound) the information a measurement chain delivers, and Fano converts any shortfall into a floor under the error of every conceivable decoder. Lesson 50's channel converse is exactly this move.

Continuous variables. For a density f (Lesson 17), the *differential entropy* $h(X) = -\int f \log f$ plays the same algebraic role — KL, mutual information, chain rules, and DPI all go through with integrals — but h itself loses two comforts: it can be negative (a uniform on $[0, \frac{1}{2}]$ has $h = -1$ bit), and it changes under rescaling ($h(aX) = h(X) + \log_2 |a|$). Mutual information, being a KL, keeps all its properties and its invariance — the robust citizen of the continuous world. The one continuous entropy fact this booklet needs constantly: among all densities with variance σ^2, the Gaussian uniquely maximizes h, with $h = \frac{1}{2}\log_2(2\pi e\sigma^2)$ — proved in Lesson 50 (Lemma 50.6), spent there on AWGN capacity and in Lesson 51 on the Gaussian rate-distortion converse.

ML / inference / sensors connection. Cross-entropy loss *is* $H(P) + D(P\|Q)$; the information gain that splits a decision tree is $I(\text{feature}; \text{label})$; variational inference optimizes a KL by construction. In inference, D is the error exponent of Lesson 50 and the natural metric on model space (Fisher information is its curvature — Lesson 50). For sensors, mutual information is the honest scorecard: a sensor is worth what $I(\Theta; Y)$ says, DPI warns that downstream DSP can only spend that budget, never grow it, and Fano converts the budget into a floor on any detector's error — three sentences that justify an entire measurement-design methodology.

48.2 Practice: by hand

Example 48.7 (The binary entropy function)**.** $X \sim \text{Ber}(p)$: $H = h_2(p)$, with $h_2(\frac{1}{2}) = 1$ bit, $h_2(0.11) \approx 0.5$, $h_2(0.01) \approx 0.081$. Steep near the ends: certainty is cheap to dent. And $D(\text{Ber}(\frac{1}{4})\|\text{Ber}(\frac{1}{2})) = \frac{1}{4}\log_2\frac{1/4}{1/2} + \frac{3}{4}\log_2\frac{3/4}{1/2} = -\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{2} \approx 0.189$ bits: each coin flip of a quarter-biased coin builds 0.189 bits of evidence against the fair-coin hypothesis — a number Lesson 50 will convert into error rates.

Example 48.8 (A binary symmetric channel, fully accounted)**.** $X \sim \mathrm{Ber}(\frac{1}{2})$, $Y = X \oplus N$ with $N \sim \mathrm{Ber}(\epsilon)$ independent. $H(Y) = 1$ (still uniform by symmetry); $H(Y \mid X) = h_2(\epsilon)$ (given X, the only randomness is the flip). So

$$I(X;Y) = 1 - h_2(\epsilon) :$$

1 bit at $\epsilon = 0$, 0.5 bits at $\epsilon \approx 0.11$, 0 at $\epsilon = \frac{1}{2}$ — a coin-flip channel carries nothing, and (check $h_2(1) = 0$) a channel that *always* lies carries a full bit, because consistent lying is a code.

48.3 Exercises

Exercise 48.1 (Hand)**.** Compute, in bits: $H(X)$ for $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$; then design a prefix code (0, 10, 110, 111) and check its average length equals H exactly. (Answer: $H = 1.75$; dyadic probabilities are the case where Lesson 51's bound is met with no slack.)

Exercise 48.2 (Hand)**.** From the joint pmf $p(x, y)$ with rows $x \in \{0, 1\}$: $p(0,0) = \frac{1}{2}$, $p(0,1) = \frac{1}{4}$, $p(1,0) = 0$, $p(1,1) = \frac{1}{4}$, compute $H(X)$, $H(Y)$, $H(X,Y)$, $H(X \mid Y)$, and $I(X;Y)$, and verify both identities $I = H(X) - H(X \mid Y) = H(X) + H(Y) - H(X,Y)$. (Answers: $H(X) = h_2(\frac{1}{4}) \approx 0.811$, $H(Y) = 1$, $H(X,Y) = 1.5$, $H(X \mid Y) = 0.5$, $I \approx 0.311$.)

Exercise 48.3 (Proof, $\star$)**.** Prove $D(P\|Q) \geq 0$ and both extremes of entropy (Theorem 48.3) from Jensen alone, then prove the chain rule $H(X,Y) = H(X) + H(Y \mid X)$ and deduce $H(X_1, \ldots, X_n) \leq \sum_i H(X_i)$ with equality iff independent.

Exercise 48.4 (Proof)**.** Prove the DPI (Theorem 48.5) for the deterministic special case $Z = g(Y)$ directly from the chain-rule expansion, and give an example where processing *loses* information strictly: X uniform on $\{0, 1, 2, 3\}$, $Y = X$, $Z = Y \bmod 2$ — compute $I(X;Y)$ and $I(X;Z)$. (2 bits vs 1 bit.)

Deeper reading. PW 1.1 (entropy), 2.1–2.5 (divergence, chain rules, data processing — stated there for general alphabets and worth seeing in that generality), 3.1–3.6 (mutual information and Fano), 5.1 (convexity of information measures, the bridge to this part's optimization half).

49 The Asymptotic Equipartition Property and Lossless Compression

NEEDED FOR: Information theory made operational: why entropy is the compression limit, and the typical-set picture used by every coding and learning bound downstream. Sensors and comms (bit budgets); ML (the compression view of learning).

49.1 Theory

Lesson 48 *defined* $H(X)$; this lesson proves it is not just a formula but a physical quantity: the number of bits an i.i.d. source costs, no more and no less. Two theorems do it — one about codes (Kraft–McMillan), one about probability (the AEP) — and the AEP's concentration picture is the one to keep: it returns in hypothesis testing (Lesson 50) and channel coding wearing different clothes.

Codes. A *variable-length code* is an injective $f : \mathcal{A} \to \{0,1\}^*$; extend it to strings symbol by symbol. Injectivity per symbol is not enough for streams: the extension must be injective too (*uniquely decodable*), and the practical subclass is the *prefix codes* — no codeword a prefix of another, decodable on the fly (PW 10.6–10.8).

Theorem 49.1 (Kraft–McMillan)**.** *If f is uniquely decodable with codeword lengths $l(a)$, then*

$$\sum_{a \in \mathcal{A}} 2^{-l(a)} \le 1,$$

and conversely any lengths satisfying this inequality are the lengths of some prefix *code (PW 10.9).*

Proof idea. Converse (construction): the dyadic-interval / binary-tree picture — Theorem 51.1 in Lesson 51 draws it. Direct (the McMillan half, and the reason "uniquely decodable" buys nothing over "prefix"): let $Z = \sum_a 2^{-l(a)}$ and expand

$$Z^n = \sum_{a_1,\dots,a_n} 2^{-[l(a_1)+\cdots+l(a_n)]} = \sum_m N_n(m)\, 2^{-m},$$

where $N_n(m)$ counts length-n strings whose encodings have total length m. Unique decodability forces $N_n(m) \le 2^m$ (distinct strings, distinct bit sequences), so $Z^n \le \sum_{m \le n\, l_{\max}} 1 = n\, l_{\max}$. A number with $Z^n = O(n)$ satisfies $Z \le 1$ — exponential beats linear (Lesson 31). □

Theorem 49.2 (Entropy is the operating point)**.** *Let $X \sim p$ on a finite alphabet.*

(a) (Prefix codes) $L^*(X) = \min \mathbf{E}[l(X)]$ over prefix codes satisfies $H(X) \le L^*(X) < H(X)+1$: the lower bound is Theorem 51.2's KL identity ($\bar{L} - H = D(p \,\|\, 2^{-l}/Z) - \log_2 Z$), the upper is Shannon's lengths $l(a) = \lceil \log_2 \frac{1}{p(a)} \rceil$, which pass Kraft. By Theorem 49.1 the same bounds govern all *uniquely decodable codes.*

(b) (Huffman) The greedy merge-the-two-least-likely construction produces a prefix code of minimal *average length (PW 10.11) — so the optimum in (a) is not just bounded but constructible by hand.*

(c) (Drop all constraints) Even the best unconstrained *injective labeling f^* — sort outcomes by probability, assign $\varnothing, 0, 1, 00, 01, \dots$ — obeys $H(X) - \log_2\big[e\,(H(X)+1)\big] \le \mathbf{E}[l(f^*(X))] \le H(X)$ (PW 10.2–10.3): entropy is the scale even without decodability. Single-symbol overheads are $O(\log H)$ at worst and vanish per symbol under blocking, by what follows.*

Probability: where the AEP comes from. Encode *blocks* $S^n = (S_1, \dots, S_n)$, i.i.d. $\sim p$. The random variable $\frac{1}{n} \log_2 \frac{1}{p(S_1)\cdots p(S_n)} = \frac{1}{n}\sum_i \log_2 \frac{1}{p(S_i)}$ is an empirical average of i.i.d. terms with mean $H(S)$ — so the weak law of large numbers (Theorem 21.2, Lesson 21) drives it to $H(S)$. That one observation, dressed up, is:

Theorem 49.3 (AEP; typical sets)**.** *For $\delta > 0$ define the* entropy-typical set

$$T_\delta^n = \Big\{ s^n : \Big| \tfrac{1}{n} \log_2 \tfrac{1}{p(s^n)} - H(S) \Big| \le \delta \Big\}.$$

Then (PW 11.6): (i) $\mathbf{P}[S^n \in T_\delta^n] \to 1$; (ii) $|T_\delta^n| \le 2^{n(H(S)+\delta)}$, since each member has $p(s^n) \ge 2^{-n(H(S)+\delta)}$ and probabilities sum to at most 1; and (iii) on T_δ^n the distribution is nearly flat: $p(s^n) = 2^{-nH(S)\pm n\delta}$.

Read it as geometry: of the $|\mathcal{A}|^n$ possible strings, essentially all probability sits on a set of only $\approx 2^{nH}$ strings, each about equally likely — *equipartition*. A fair-coin source is the degenerate case (every string typical); a biased coin ($p = 0.1$, $H \approx 0.47$) concentrates on the $\binom{n}{0.1n} \approx 2^{0.47n}$ strings with $\approx 10\%$ ones — a vanishing fraction of 2^n. Compression is now obvious:

Theorem 49.4 (Shannon's source coding theorem)**.** *For i.i.d. S^n encoded into nR bits with error probability ϵ_n (PW 11.1, 11.3):*

$$R > H(S) \;\Rightarrow\; \epsilon_n \to 0 \qquad \textit{and} \qquad R < H(S) \;\Rightarrow\; \epsilon_n \to 1.$$

Proof idea. Achievability: enumerate T_δ^n with $n(H+\delta) \le nR$ bits and flag the atypical remainder; the error is $\mathbf{P}[S^n \notin T_\delta^n] \to 0$ (Theorem 49.3(i)–(ii)). Converse: 2^{nR} codewords can carry at most 2^{nR} strings, but every subset capturing probability bounded away from 0 must contain $\gtrsim 2^{n(H-\delta)}$ typical strings, each of probability $\le 2^{-n(H-\delta)}$ — with $R < H$ the books do not balance, and the captured probability collapses. Entropy is a wall, not a benchmark. □

Note the anatomy, because it recurs: an *achievability* half (a construction meeting the limit) and a *converse* half (an accounting argument that no scheme beats it), meeting at a single number determined by an information quantity. Channel capacity (Lesson 50) and rate–distortion (Lesson 51) have exactly this shape, with the AEP's counting at the core of both.

49.2 Practice: by hand

Example 49.5 (Huffman, end to end)**.** $p = (0.4, 0.3, 0.2, 0.1)$ on $\{a, b, c, d\}$. Merge d, c (0.3), then the pair with b (0.6), then all (1.0); reading the tree: $a \to 0$? No — the two least-likely at the second step are $b\,(0.3)$ and $cd\,(0.3)$. One valid outcome: $a = 0$, $b = 10$, $c = 110$, $d = 111$, with $\bar{L} = 0.4(1) + 0.3(2) + 0.2(3) + 0.1(3) = 1.9$ bits, against $H = -\sum p \log_2 p \approx 1.846$ bits: inside the $[H, H+1)$ window of Theorem 49.2, overhead ≈ 0.054 bit, and by (a) the KL identity says the overhead *is* $D(p\|q)$ for the implied dyadic model $q = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$. ✓

Example 49.6 (Typical-set head count)**.** Bernoulli source, $p = 0.1$, $n = 1000$: $H = 0.469$ bits, so $|T_\delta^n| \approx 2^{469}$ typical strings out of 2^{1000} — a fraction 2^{-531} of the space carrying probability $\to 1$. Fixed-rate coding at $R = 0.5 > H$ succeeds with room to spare; at $R = 0.4 < H$ it must fail. And the flat-code alternative (10 bits per symbol block of 10, no model) spends 1000 bits where ≈ 469 suffice: the price of ignoring the distribution is $1 - H(0.1) \approx 0.53$ bit per symbol, forever.

Sensors / ML connection. Edge bit budgets (Course 3's Module 8 telemetry and logging) are Theorem 49.4 in practice: an event stream with entropy rate H cannot be logged losslessly below H bits per event, and Huffman / arithmetic coders get within a bit (per block, within $1/n$). In ML, the cross-entropy loss of Lesson 48 *is* the code length of the model-as-code reading (Theorem 49.2(a)'s KL identity), so "lower perplexity" literally means "compresses the data better"; and the typical-set picture — exponentially few sequences carry all the mass — is the standing intuition behind generalization bounds and sampling from generative models alike.

49.3 Exercises

Exercise 49.1 (Hand)**.** For $p = (0.5, 0.25, 0.15, 0.1)$: compute $H(X)$; build a Huffman code (show the merge tree); compute $\bar{L}$ and the overhead $\bar{L} - H$; verify Kraft holds with your lengths; and exhibit the implied dyadic model q and check $\bar{L} - H = D(p\|q)$ numerically to two decimals.

Exercise 49.2 (Hand)**.** Bernoulli(0.2) source, blocks of $n = 500$. Compute $H(S)$; the typical-set size exponent $n(H \pm \delta)$ for $\delta = 0.02$; the range of Hamming weights a typical string can have (use $\frac{1}{n}\log_2 \frac{1}{p(s^n)} = \frac{w}{n}\log_2 \frac{1}{p} + \frac{n-w}{n}\log_2 \frac{1}{1-p}$ and solve for w/n); and the fixed rate R you would provision for $\epsilon \to 0$.

Exercise 49.3 (Proof)**.** Prove the McMillan direction of Theorem 49.1 in full: justify $N_n(m) \leq 2^m$ from unique decodability, carry out the Z^n expansion, and complete the "$Z^n = O(n) \Rightarrow Z \leq 1$" step with an honest limit argument (Lesson 31).

Exercise 49.4 (Proof, $\star$)**.** Prove the AEP (Theorem 49.3) from the WLLN, then both halves of Theorem 49.4. For the converse, show precisely: if $B_n \subseteq \mathcal{A}^n$ with $|B_n| \leq 2^{nR}$, $R < H - 2\delta$, then $\mathbf{P}[S^n \in B_n] \leq \mathbf{P}[S^n \notin T_\delta^n] + |B_n|\, 2^{-n(H-\delta)} \to 0$ — the two-line accounting that closes every converse of this type.

Deeper reading. PW 10.1–10.3 (codes, Kraft–McMillan, optimal lengths, Huffman), 11.1–11.2 (fixed-length codes, AEP, source coding theorem), ch. 12 for what replaces the WLLN when the source has memory (entropy rate, Shannon–McMillan); then Lesson 51 for the lossy sequel.

50 Testing, Channels, and Sensor Information

NEEDED FOR: Communications (capacity, coding), inference (detection, estimation floors), sensors and radar (design as information maximization), ML (generalization's information bounds).

50.1 Theory

Course 3 Labs 6.7–6.9 built detectors — matched filter, threshold, CFAR — and measured them by SNR. This lesson supplies the deeper ledger: how fast evidence accumulates (divergence), the hard floor under every estimator (Fisher), and the ceiling on every channel and sensor (capacity). All three are the same mathematics.

Theorem 50.1 (Neyman–Pearson lemma)**.** *Observations $Y \sim P$ or $Y \sim Q$; a test decides. Among all tests with false-alarm probability $\mathbf{P}_Q(\text{decide } P) \leq \alpha$, the likelihood-ratio test*

$$\text{decide } P \iff L(y) = \frac{p(y)}{q(y)} \geq \tau$$

(with τ set so the false-alarm constraint is met, randomizing on the boundary if needed) minimizes the miss probability $\beta = \mathbf{P}_P(\text{decide } Q)$.

Proof. Let $\phi^\star(y) \in \{0, 1\}$ be the LRT's probability of deciding P, and ϕ any competitor. Pointwise,

$$\big(\phi^\star(y) - \phi(y)\big)\big(p(y) - \tau q(y)\big) \;\geq\; 0,$$

because where $p \geq \tau q$ the LRT has $\phi^\star = 1 \geq \phi$, and where $p < \tau q$ it has $\phi^\star = 0 \leq \phi$. Sum (or integrate) over y:

$$\big[(1-\beta^\star)-(1-\beta)\big] - \tau\big[\alpha^\star - \alpha_\phi\big] \geq 0 \implies \beta - \beta^\star \;\geq\; \tau\,(\alpha^\star - \alpha_\phi) \;\geq\; 0$$

whenever the competitor spends no more false alarm than the LRT ($\alpha_\phi \leq \alpha^\star$). Four lines, and detector design is over: only the threshold remains to be chosen. □

This is why Course 3 Labs 6.7–6.9's receivers all compute likelihood ratios — the matched filter is the LRT for a known signal in Gaussian noise (Example 41.6 derived it as a functional gradient; here it is optimal among *all* detectors, not just linear ones).

How fast does evidence accumulate? With n i.i.d. observations the log-likelihood ratio adds: $\log L_n = \sum_{k=1}^{n} \log \frac{p(Y_k)}{q(Y_k)}$. Under P, the WLLN (Theorem 21.2) drives $\frac{1}{n}\log L_n \to \mathbf{E}_P \log \frac{p}{q} = D(P\|Q)$: divergence *is* the per-sample drift of the evidence. The precise statement is:

Theorem 50.2 (Stein's lemma — KL is the testing exponent)**.** *For i.i.d. observations and any fixed false-alarm level $\alpha \in (0,1)$, the optimal miss probability satisfies*

$$-\frac{1}{n}\log \beta_n^\star \;\longrightarrow\; D(P\|Q)\,:$$

misses die like $e^{-nD(P\|Q)}$ (nats), no faster, no slower. (Proof: PW §14.5 — achievability is the LLN drift above pushing $\log L_n$ past any fixed threshold; the converse is a change-of-measure argument. Symmetrically, testing with both error exponents positive trades them along a curve whose corner cases are $D(P\|Q)$ and $D(Q\|P)$ — the Chernoff regime, PW ch. 16.)

So the asymmetric-looking KL has an exact operational meaning: $D(P\|Q)$ prices *mistaking P-data for Q*. For Gaussians of common variance, $D(\mathcal{N}(\mu_1,\sigma^2)\|\mathcal{N}(\mu_0,\sigma^2)) = \frac{(\mu_1-\mu_0)^2}{2\sigma^2}$ nats (Exercise 50.2) — one half the matched-filter SNR of Course 3 Labs 6.7–6.9: the two ledgers agree where they overlap, and KL keeps working where SNR stops meaning anything (non-Gaussian noise, unequal variances, discrete data).

The estimation floor. Testing bounds decisions; Fisher information bounds *continuous* estimates.

Definition 50.3 (Score and Fisher information)**.** For a family of densities $p_\theta(y)$ smooth in the parameter $\theta \in \mathbb{R}$, the *score* is $s_\theta(y) = \frac{\partial}{\partial\theta}\ln p_\theta(y)$ and the *Fisher information* is $J(\theta) = \mathbf{E}_\theta[s_\theta(Y)^2]$ — the mean squared sensitivity of the log-likelihood to the parameter. Under the usual regularity (differentiation through the integral, certified by Theorem 35.15), $\mathbf{E}_\theta[s_\theta(Y)] = \int \frac{\partial p_\theta}{\partial\theta}\,dy = \frac{\partial}{\partial\theta}\int p_\theta\,dy = 0$.

Theorem 50.4 (Cramér–Rao lower bound)**.** *Any unbiased estimator $\hat\theta(Y)$ ($\mathbf{E}_\theta\hat\theta = \theta$ for all θ) satisfies*

$$\operatorname{var}(\hat\theta) \;\geq\; \frac{1}{J(\theta)},$$

and for n i.i.d. observations, $\geq 1/(nJ(\theta))$.

Proof. Differentiate the unbiasedness identity $\int \hat{\theta}(y)\, p_\theta(y)\, dy = \theta$ in θ: $\mathbf{E}_\theta[\hat{\theta}\, s_\theta] = 1$. Since $\mathbf{E}s_\theta = 0$, this says $\operatorname{cov}(\hat{\theta}, s_\theta) = 1$. Cauchy–Schwarz in the covariance inner product (Theorem 18.3):

$$1 = \operatorname{cov}(\hat{\theta}, s_\theta)^2 \le \operatorname{var}(\hat{\theta})\, \operatorname{var}(s_\theta) = \operatorname{var}(\hat{\theta})\, J(\theta).$$

Independence adds Fisher informations (the score of a product is a sum of scores, and cross-terms vanish by zero mean), giving the n-sample form. □

For $Y \sim \mathcal{N}(\theta, \sigma^2)$: $s_\theta = (y-\theta)/\sigma^2$, $J = 1/\sigma^2$, and the sample mean achieves σ^2/n exactly — the CRLB is tight, and Lesson 19's "the mean is the best estimator" becomes a theorem about *all* unbiased estimators, not just linear ones. Fisher information is also the local curvature of KL: $D(P_\theta \| P_{\theta+\delta}) = \frac{J(\theta)}{2}\delta^2 + o(\delta^2)$ (PW §2.6) — testing and estimation are priced by the same geometry, KL globally, Fisher infinitesimally. In sensor language: J is per-measurement resolving power, the CRLB is the spec-sheet floor no clever post-processing beats (that is the DPI, Theorem 48.5, in estimation clothing), and Course 3 Lab 6.6's Kalman filter is this accounting run recursively for linear-Gaussian dynamics.

Channels and capacity. A *channel* is a conditional distribution $P_{Y|X}$ — exactly what a sensor is: state in, noisy reading out. Using it n times to carry one of M messages at rate $R = \frac{1}{n}\log_2 M$ bits/use:

Theorem 50.5 (Shannon's channel coding theorem)**.** *For a memoryless channel, every rate below*

$$C = \max_{P_X} I(X;Y)$$

is achievable with error probability $\to 0$ as $n \to \infty$, and every rate above C forces error bounded away from zero. (PW chs. 18–19. The converse is Fano's inequality, Theorem 48.6, plus the DPI — information that never entered cannot be decoded out; the achievability is Shannon's random-coding argument, probabilistically brilliant and practically silent about which *code — sixty years of coding theory live in that gap.)*

Two capacities carry most of engineering. *BSC(ϵ)*: by Lesson 48's accounting $I(X;Y) = H(Y) - h_2(\epsilon) \le 1 - h_2(\epsilon)$, achieved by uniform input, so $C = 1 - h_2(\epsilon)$ bits/use. *AWGN*, $Y = X + Z$, $Z \sim \mathcal{N}(0, N)$, power constraint $\mathbf{E}X^2 \le P$: here the maximization needs the fact promised in Lesson 48 —

Lemma 50.6 (Gaussians maximize entropy)**.** *Among densities with variance $\le \sigma^2$, $h(Y) \le \frac{1}{2}\log_2(2\pi e\sigma^2)$, with equality only for $\mathcal{N}(m, \sigma^2)$.*

Proof. Let f have variance $\le \sigma^2$ (mean m, wlog $m = 0$) and g be the $\mathcal{N}(0, \sigma^2)$ density. Then

$$0 \le D(f\|g) = \int f \log_2 \frac{f}{g} = -h(Y) - \int f \log_2 g.$$

But $\log_2 g(y) = -\frac{1}{2}\log_2(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\log_2 e$ is a *quadratic* in y, so $-\int f \log_2 g$ depends on f only through its second moment, and is at most its value under g itself: $-\int f \log_2 g \le -\int g \log_2 g = \frac{1}{2}\log_2(2\pi e\sigma^2)$. Rearranged: $h(Y) \le \frac{1}{2}\log_2(2\pi e\sigma^2)$, with equality iff $D(f\|g) = 0$, i.e. $f = g$ (Theorem 48.3). □

Then $I(X;Y) = h(Y) - h(Y \mid X) = h(Y) - h(Z) \le \frac{1}{2}\log_2\big(2\pi e(P+N)\big) - \frac{1}{2}\log_2(2\pi eN)$, giving

$$C = \frac{1}{2}\log_2\Big(1 + \frac{P}{N}\Big) \text{ bits per real use,}$$

achieved by Gaussian input. Information grows *logarithmically* in power: the first watt is worth more than the next ten. For n parallel Gaussian channels with noises N_i and a total power budget — exactly Lesson 46's water-filling problem, whose KKT solution is now revealed as the capacity-achieving allocation. That is the two halves of Part VII meeting: information theory says what the objective *means*; convex duality solves it and prices the budget.

Sensors as channels; measurements as experiments. A sensor suite maps state θ to readings Y through $P_{Y|\theta}$; "which measurements should we take?" becomes "maximize the information delivered." In the linear-Gaussian case $y_i = a_i^{\mathsf{T}}\theta + z_i$, both criteria of this lesson agree on the answer's shape: the Fisher information matrix is $J = \sum_i a_i a_i^{\mathsf{T}}/\sigma^2$, the CRLB generalizes to $\mathrm{cov}(\hat{\theta}) \succeq J^{-1}$, and with a Gaussian prior the mutual information is $I(\theta;Y) = \frac{1}{2}\log_2\det(I + \Sigma_\theta J)$ (Lesson 20's Gaussian machinery). Choosing how often to fire each candidate measurement to maximize $\log\det J$ is *D-optimal experiment design* — concave in the firing frequencies, hence a convex program (BV §7.5), solvable and certifiable by Lessons 36–38. Radar dwell scheduling, array geometry, calibration-sequence design: all instances.

Communications / radar / ML connection. Course 3 Labs 6.7–6.9's link budgets are this lesson quantified: BER curves are testing exponents, capacity is the ceiling that coding (LDPC, turbo) has now essentially reached, and $\frac{1}{2}\log_2(1 + \mathrm{SNR})$ is why every additional dB buys less throughput. Radar detection is Neyman–Pearson verbatim (CFAR sets τ adaptively to hold α). In ML, Fano-style arguments give minimax lower bounds — *no* learner beats them — and PW Part VI builds statistical learning theory on exactly the D, I, J triad of this lesson.

50.2 Practice: by hand

Example 50.7 (BSC capacities, felt)**.** $C(\epsilon) = 1 - h_2(\epsilon)$: $C(0) = 1$, $C(0.01) \approx 0.919$, $C(0.11) \approx 0.5$, $C(0.5) = 0$. An 11% bit-flip rate costs *half* the channel — noise is expensive out of all proportion to its rate, because the decoder must pay to locate the flips, not just count them.

Example 50.8 (Three cheap sensors vs one good one)**.** Binary state through BSC(0.1) sensors. One sensor: error 0.1; majority of three: $3(0.1)^2(0.9) + (0.1)^3 = 0.028$. Information accounting: one sensor delivers $1 - h_2(0.1) \approx 0.531$ bits; three deliver at most (and, being independent, exactly) $3\times$ the per-sensor mutual information $= 1.59$ bits about a 1-bit state — redundancy that the majority vote spends imperfectly (it discards the unanimity/2–1 distinction). The LRT uses all of it; with symmetric sensors majority *is* the LRT, and the gap to zero error is closed only exponentially, at Stein's rate nD.

Example 50.9 (AWGN arithmetic)**.** SNR $= 15$ (≈ 11.8 dB): $C = \frac{1}{2}\log_2 16 = 2$ bits/real use. To get 3 bits: SNR $= 63$, four times the power for one more bit. At low SNR, $C \approx \frac{\mathrm{SNR}}{2\ln 2}$ — linear in power, so spreading power over bandwidth is free there: the design logic of spread-spectrum and of every energy-starved sensor link.

50.3 Exercises

Exercise 50.1 (Hand)**.** Compute C for BSC(ϵ), $\epsilon \in \{0, 0.1, 0.5, 0.9\}$, and explain in one sentence why $\epsilon = 0.9$ matches $\epsilon = 0.1$. Then: how many uses of BSC(0.11) are needed, at minimum, to convey 500 bits reliably? (Answer: $C \approx \frac{1}{2}$, so ≥ 1000 uses — and Theorem 50.5 says some code gets arbitrarily close.)

Exercise 50.2 (Hand)**.** Show $D\big(\mathcal{N}(\mu_1, \sigma^2) \,\|\, \mathcal{N}(\mu_0, \sigma^2)\big) = \frac{(\mu_1 - \mu_0)^2}{2\sigma^2}$ nats by direct integration (the quadratic terms cancel; only the cross term survives). Conclude: a radar return displaced by one noise standard deviation builds evidence at $\frac{1}{2}$ nat per pulse, and Stein's lemma prices $\beta_n \approx e^{-n/2}$.

Exercise 50.3 (Proof, $\star$)**.** Prove the Cramér–Rao bound (Theorem 50.4) including the score's zero mean and the additivity of J over independent samples. Then compute J for $Y \sim \mathrm{Ber}(\theta)$ (answer: $J = \frac{1}{\theta(1-\theta)}$) and check that the sample frequency achieves it — the CRLB is tight for both of this booklet's favorite models.

Exercise 50.4 (Proof)**.** Prove Lemma 50.6 without looking. Then show where it was spent twice in this lesson (once on $h(Y)$, once — in Lesson 51 — on the error's entropy), and read PW §5.5 to state the *Gaussian saddle point*: for fixed signal and noise variances, Gaussian input is best and Gaussian noise is worst, so $\frac{1}{2}\log_2(1 + P/N)$ is simultaneously a guarantee against the worst noise and a ceiling for the best code — the reason link budgets may assume Gaussian everything without apology.

Deeper reading. PW 14.1–14.5 (Neyman–Pearson and Stein), 15.1–15.2 and ch. 16 (large deviations, Chernoff exponents), 19.1–19.3 (capacity and the coding theorem), 20.1–20.4 (Gaussian channels, water-filling), 2.6 and 29.1–29.2 (Fisher information and Cramér–Rao); BV 7.3 (detector design as an LP — worth seeing once) and 7.5 (experiment design). On the bench: Course 3 Labs 6.6–6.9 are the practice this theory grades.

51 Rate-Distortion, Compression, and Representation Learning

NEEDED FOR: DSP and audio/image coding (quantizers, codecs), ML (bottlenecks, learned representations), sensors (bit budgets at the edge).

51.1 Theory

Lesson 48 measured information; this lesson spends it. Two regimes: lossless — reproduce exactly, pay H — and lossy — reproduce within distortion D, pay the rate-distortion function. Lesson 49 proved the lossless side's limit theorems; the Kraft/entropy recap below keeps this lesson self-contained.

Theorem 51.1 (Kraft inequality)**.** *A binary prefix code (no codeword a prefix of another — decodable on the fly, as in Exercise 48.1) with lengths $l_1, \ldots, l_K$ exists iff $\sum_i 2^{-l_i} \leq 1$.*

Proof. ($\Rightarrow$) Map each codeword c of length l to the dyadic interval $[0.c,\, 0.c + 2^{-l}) \subset [0, 1)$ of its binary expansion. The prefix property says no codeword's interval contains another's, so the intervals are disjoint, and their total length $\sum_i 2^{-l_i}$ fits in $[0, 1)$. ($\Leftarrow$) Sort $l_1 \leq \cdots \leq l_K$

and greedily assign intervals left to right; the inequality guarantees room at every step, and dyadic alignment preserves the prefix property. (PW §10.3 draws the same proof on the binary tree.) □

Theorem 51.2 (Entropy is the price of lossless coding)**.** *For a source $X \sim p$ and any prefix code with lengths $l(x)$, the expected length obeys $\bar{L} = \mathbf{E}[l(X)] \geq H(X)$; and there is a prefix code (Shannon's $l(x) = \lceil \log_2 \frac{1}{p(x)} \rceil$) with $\bar{L} < H(X) + 1$. Coding blocks of n symbols squeezes the overhead to $\frac{1}{n}$: entropy is exactly the asymptotic bits-per-symbol cost.*

Proof. Lower bound: let $q(x) = 2^{-l(x)}/Z$ with $Z = \sum_x 2^{-l(x)} \leq 1$ (Kraft). Then

$$\bar{L} - H(X) = \sum_x p(x)\Big[l(x) + \log_2 p(x)\Big] = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} - \log_2 Z = D(P\|Q) - \log_2 Z \;\geq\; 0,$$

both terms nonnegative (Theorem 48.3; $-\log_2 Z \geq 0$ since $Z \leq 1$). Upper bound: the ceiling lengths satisfy Kraft ($\sum 2^{-\lceil \log 1/p \rceil} \leq \sum p = 1$), so the code exists, and each length exceeds $\log_2 \frac{1}{p(x)}$ by less than 1. □

Note what the proof *is*: a code is a probability model q in disguise ($q = 2^{-l}$), and its overhead over entropy is exactly $D(p\|q)$ — you pay KL for coding with the wrong model. That identity is the engine of arithmetic coding and of the compression-view of ML (a language model is a code; its log-loss *is* its compressed size; PW ch. 13). The block-coding claim is the AEP: by the WLLN (Theorem 21.2) applied to $\frac{1}{n} \log_2 \frac{1}{p(X_1)\cdots p(X_n)} \to H(X)$, long i.i.d. strings concentrate on $\approx 2^{nH}$ *typical* sequences of near-equal probability — index those and you are done (PW §11.2).

Quantization: analog sources meet finite bits. A b-bit uniform quantizer with step Δ rounds x to its bin center; modeling the error as uniform on $(-\frac{\Delta}{2}, \frac{\Delta}{2}]$ — honest when the signal moves through many bins per sample, dubious otherwise (Course 3 Lab 9.3 added dither precisely to launder this assumption) — gives

$$\mathbf{E}[e^2] = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} e^2 \, de = \frac{\Delta^2}{12}.$$

Each extra bit halves Δ, quartering the error power: 6.02 dB per bit, the number behind every ADC spec (Course 3 Lab 3.4) and audio word-length argument (Course 3 Lab 9.3). But scalar rounding is not the end of the story — the true frontier is:

Definition 51.3 (Rate-distortion function)**.** For source X and distortion measure $d(x, \hat{x})$,

$$R(D) = \min_{P_{\hat{X}|X}:\, \mathbf{E}[d(X,\hat{X})] \leq D} I(X; \hat{X}),$$

and Shannon's rate-distortion theorem (PW ch. 25) says $R(D)$ is exactly the minimal bits/symbol at which block coders can hit expected distortion D. It is the constrained-information twin of capacity: nonincreasing and convex in D, and $R(0) = H(X)$ for discrete sources — lossless coding is its endpoint.

Theorem 51.4 (Gaussian rate-distortion)**.** *For $X \sim \mathcal{N}(0, \sigma^2)$ with squared-error distortion,*

$$R(D) = \tfrac{1}{2} \log_2 \frac{\sigma^2}{D} \quad (0 < D \leq \sigma^2), \qquad \textit{equivalently} \qquad D(R) = \sigma^2 \, 2^{-2R}.$$

Proof of the converse (the achievability is a test-channel construction, PW §26.1). For any reproduction rule with $\mathbf{E}(X-\hat{X})^2 \le D$:

$$I(X;\hat{X}) = h(X) - h(X \mid \hat{X}) \;\ge\; h(X) - h(X-\hat{X}),$$

since conditioning reduces (differential) entropy and shifting by the known $\hat{X}$ does not change it. Now spend Lemma 50.6 on the error: $h(X-\hat{X}) \le \frac{1}{2}\log_2(2\pi e D)$. With $h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2)$,

$$I(X;\hat{X}) \;\ge\; \tfrac{1}{2}\log_2 \frac{\sigma^2}{D}. \qquad \square$$

Halving the distortion always costs exactly half a bit per sample; 6 dB per bit again — but now as an *unbeatable* frontier rather than a quantizer model. The gap between a real scalar quantizer and $D(R)$ (a factor $\frac{\pi e}{6} \approx 1.53$ dB at high rate, with entropy coding) is the fee for coding samples one at a time; transform and vector coders exist to shrink it. For *correlated* Gaussian vectors, diagonalize the covariance (Theorem 20.3) and the problem splits into independent scalar sources with variances λ_i; the optimal bit allocation solves a convex problem whose KKT conditions are *reverse water-filling*: all components are coded to a *common* distortion level D^*, and components with $\lambda_i \le D^*$ get zero bits (PW §26.1) — Lesson 46's picture upside down, poured into eigenvalues instead of channels.

Transform coding, and why codecs look the way they do. Reverse water-filling is a recipe: (1) rotate into decorrelated coordinates — the KLT, which is PCA (Lesson 20) and is approximated at zero covariance-knowledge cost by the DCT for smooth signals (Course 3 Lab 9.4's JPEG, Course 3 Lab 9.3's MDCT); (2) quantize each coordinate to the common distortion level, spending $\frac{1}{2}\log_2(\lambda_i/D^*)$ bits where it is due and *zero* on weak components — which is why most of a JPEG block's coefficients are simply dropped; (3) entropy-code the result (Theorem 51.2 collects the last free bits). Every mainstream codec is these three steps plus perceptual weighting of the distortion measure.

Representation learning as rate-distortion. An autoencoder's bottleneck enforces a rate constraint; its reconstruction loss is a distortion; training sketches a point on an $R(D)$ curve for the data distribution — with the network's expressiveness standing in for Shannon's optimal coder. The *information bottleneck* makes the analogy an objective: compress X into Z keeping what matters for a task Y,

$$\min_{P_{Z|X}} \; I(X;Z) - \beta\, I(Z;Y),$$

rate against *task-relevant* information rather than reconstruction. The DPI (Theorem 48.5) is the standing warning: $I(Z;Y) \le I(X;Y)$ always — representations budget information, they never mint it. And the granddaddy abstraction, *metric entropy* (PW ch. 27: the number of ϵ-balls needed to cover a function class, i.e. the bits needed to name a hypothesis to accuracy ϵ), is how these ideas price learning itself: classes that compress well generalize, a theme PW Part VI develops into theorems.

DSP / ML / sensors connection. Course 3 Lab 9.3's quantization-noise arithmetic, Course 3 Lab 3.4's oversampling-for-resolution trades, and Course 3 Lab 9.4's transform codecs are all points on or gaps from this lesson's curves — $\Sigma\Delta$ conversion, in this language,

buys rate with bandwidth and spends it via noise shaping. At the sensing edge, $R(D)$ answers the design question "how many bits must the radio carry so the fusion center's estimate stays within spec?" before any hardware is chosen. In ML, the compression view of generalization — minimum-description-length, bottlenecks, the code = model identity in Theorem 51.2's proof — is the field's most durable explanation of why simple models learned from finite data can be trusted.

51.2 Practice: by hand

Example 51.5 (An ADC's bit budget)**.** A 12-bit ADC at 48 kHz emits 576 kbit/s raw. If the analog front end leaves only ≈ 8.5 effective bits (Course 3: noise floors are specs too), the honest information rate is ≈ 408 kbit/s, and entropy coding of the actual signal distribution recovers still more — transmitting raw ADC words ships noise at premium prices.

Example 51.6 (Gaussian $D(R)$ numbers)**.** $\sigma^2 = 1$: $R = 1$ bit/sample gives $D = 0.25$; $R = 2$ gives 0.0625. Real 2-bit scalar quantizers: the best possible (Lloyd–Max, with levels placed where the Gaussian lives) achieves ≈ 0.118, and a lazy uniform-over-$\pm 4\sigma$ quantizer measures ≈ 0.34 — the frontier, the best scalar coder, and an actual design, in one column of numbers. To halve any distortion achieved *on* the frontier: half a bit, always.

Example 51.7 (Bernoulli rate-distortion endpoints)**.** For Ber(p), $p \le \frac{1}{2}$, Hamming distortion: $R(D) = h_2(p) - h_2(D)$ for $0 \le D < p$, and $R(D) = 0$ for $D \ge p$ — because guessing the constant 0 already achieves distortion p with zero bits. Sanity: $R(0) = h_2(p) = H(X)$, the lossless price, as promised.

51.3 Exercises

Exercise 51.1 (Hand)**.** For $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ build the Shannon code $l = \lceil \log_2 1/p \rceil$, verify Kraft holds with equality, and compare $\bar{L}$ to H. Then perturb to $p = (0.45, 0.3, 0.15, 0.1)$: compute H, the Shannon lengths, and the overhead — and identify which term of Theorem 51.2's proof the overhead is. ($H \approx 1.79$; lengths $(2, 2, 3, 4)$; $\bar{L} = 2.35$; the overhead is $D(p\|2^{-l})$ plus the slack in Kraft.)

Exercise 51.2 (Hand)**.** A sensor node observes $\mathcal{N}(0, \sigma^2)$ samples and may transmit 3 bits/sample. What distortion does the frontier allow? What does a 3-bit uniform quantizer's $\Delta^2/12$ model predict with range $\pm 4\sigma$? Explain the ordering. (Frontier: $\sigma^2 2^{-6} \approx 0.016\sigma^2$; uniform: $\Delta = \sigma$, $\Delta^2/12 \approx 0.083\sigma^2$ — the model is above the frontier, as it must be.)

Exercise 51.3 (Proof, $\star$)**.** Prove Theorem 51.4's converse from scratch, naming every inequality used (conditioning reduces entropy; translation invariance; Lemma 50.6). Then deduce the "half a bit per factor of two" rule and compute the bits per sample needed to reproduce a Gaussian source at 40 dB SNR. (10^4 ratio: $\frac{1}{2}\log_2 10^4 \approx 6.6$ bits.)

Exercise 51.4 (Proof)**.** Prove both directions of Kraft (Theorem 51.1) — draw the binary tree for $(1, 2, 3, 3)$ — and then show that *any* uniquely decodable code obeys the same inequality is harder (McMillan; PW §10.3): just verify the claim on the non-prefix but decodable code $\{0, 01, 011\}$.

Deeper reading. PW 10.1 and 10.3 (codes, Kraft, Huffman), 11.1–11.2 (block coding and the AEP), ch. 24 (quantization, including optimal and high-resolution theory), 25.1 and 26.1 (the rate-distortion theorem; Bernoulli and Gaussian evaluations, reverse water-filling), 27.1–27.2 and 27.6 (metric entropy and its rate-distortion connection); PW 4.8 for the learning links. On the bench, Course 3 Labs 3.4, 9.3, and 9.4 are where these curves meet silicon.