

Lecture 1: Mathematical basics for statistics

Lecturer: Yiqiao Zhong

Report an issue

*Original lecture date: September 3, 2025. Course edition: Fall 2026.**Last update: October 19, 2026**These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.*

Welcome to STAT 709! *Mathematical statistics* is built from probability, linear algebra, analysis, and optimization.

Do not be discouraged if some of the results below are unfamiliar. You are not expected to know everything before taking this course. Use this lecture to identify gaps and warm up by working through the exercises.

This begins Part I. The part overview connects these lectures and includes a final self-check.

Learning goals

By the end of this lecture, you should be able to:

- use closure, compactness, and continuity to reason about whether an estimator or optimizer exists;
- manipulate $o(\cdot)$ and $O(\cdot)$ notation without hiding which quantities are held fixed;
- perform basic calculations with Gaussian vectors and Gaussian tails;
- use Stirling's approximation to estimate binomial coefficients and optionally recognize the appearance of entropy;
- understand basic properties of eigenvalues and singular values, and translate among quadratic forms, matrix norms, and PCA;
- apply standard norm inequalities and duality; and
- recognize a convex optimization problem and write its first-order optimality condition, with subgradients as an optional extension.

1 Sets, topology, and compactness**Mid priority** *Supporting language for existence arguments*

Throughout the course, the ambient space will usually be the Euclidean space $\mathbb{R}^d$, equipped with the Euclidean norm

$$\|x\|_2 = \left(\sum_{j=1}^d x_j^2 \right)^{1/2}.$$

For $x \in \mathbb{R}^d$ and $\epsilon > 0$, write

$$B_\epsilon(x) = \{y \in \mathbb{R}^d : \|y - x\|_2 < \epsilon\}$$

for the open ball of radius ϵ centered at x .

Definition 1 (Open and closed sets). A set $U \subset \mathbb{R}^d$ is **open** if, for every $x \in U$, there exists $\epsilon > 0$ such that $B_\epsilon(x) \subset U$. A set $K \subset \mathbb{R}^d$ is **closed** if its complement K^c is open.

Arbitrary unions and finite intersections of open sets are open. Dually, arbitrary intersections and finite unions of closed sets are closed. The restriction to *finite* intersections and unions matters. For example,

$$\bigcap_{n=1}^{\infty} (-1/n, 1/n) = \{0\},$$

which is not open.

Definition 2 (Closure, interior, and boundary). For $S \subset \mathbb{R}^d$:

$$\begin{aligned}\bar{S} &= \{x \in \mathbb{R}^d : B_\epsilon(x) \cap S \neq \emptyset \text{ for every } \epsilon > 0\}, \\ S^\circ &= \{x \in S : B_\epsilon(x) \subset S \text{ for some } \epsilon > 0\}, \\ \partial S &= \bar{S} \setminus S^\circ.\end{aligned}$$

These are called the **closure**, **interior**, and **boundary** of S , respectively.

Equivalently, $\bar{S}$ is the smallest closed set containing S . A point x belongs to $\bar{S}$ if and only if there is a sequence $(x_n)_{n \geq 1}$ in S such that $x_n \rightarrow x$. This sequential characterization is particularly convenient in Euclidean spaces.

Example 1. The rational numbers $\mathbb{Q}$ are countable and dense in $\mathbb{R}$: $\bar{\mathbb{Q}} = \mathbb{R}$. More generally, $\mathbb{Q}^d$ is countable and dense in $\mathbb{R}^d$. Thus, countability is a statement about the number of points, whereas density is a statement about how those points are distributed through the space.

Definition 3 (Countability). A set is **countable** if it is finite or can be placed in one-to-one correspondence with a subset of the natural numbers $\mathbb{N}$. A countable union of countable sets is countable, and every finite Cartesian product of countable sets is countable.

Definition 4 (Compactness). A set $K \subset \mathbb{R}^d$ is **compact** if every sequence in K has a subsequence that converges to a point in K .

Theorem 1 (Heine–Borel and the extreme-value theorem). For $K \subset \mathbb{R}^d$, the following statements are equivalent:

1. K is compact;
2. K is closed and bounded.

Moreover, if K is nonempty and compact and $f : K \rightarrow \mathbb{R}$ is continuous, then f attains both its minimum and its maximum on K .

Remark 1 (Why we care). An estimator is often defined as an optimizer,

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} L_n(\theta).$$

Before discussing uniqueness or computation, we should ask whether the minimum is attained. Compactness of Θ and continuity of L_n give a simple sufficient condition. If Θ is open or unbounded, additional arguments such as coercivity may be needed.

Example 2 (A noncompact parameter space). The continuous, strictly convex function $f(x) = e^x$ on $\mathbb{R}$ has infimum 0, but it has no minimizer. Strict convexity guarantees at most one minimizer; it does not guarantee that a minimizer exists.

2 Limits and asymptotic notation

High priority Definitions and calculation rules students must use fluently

Asymptotic notation compares deterministic sequences. Stochastic analogues such as $o_{\mathbb{P}}(\cdot)$ and $O_{\mathbb{P}}(\cdot)$ will be introduced later. Throughout this section, suppose that $b_n > 0$ for all sufficiently large n .

Definition 5 (Small- o , big- O , and asymptotic equivalence). For real sequences (a_n) and (b_n) :

$$\begin{aligned} a_n = o(b_n) &\iff \frac{a_n}{b_n} \longrightarrow 0, \\ a_n = O(b_n) &\iff \text{there exist } C < \infty \text{ and } N \text{ such that } |a_n| \leq Cb_n \text{ for } n \geq N, \\ a_n \sim b_n &\iff \frac{a_n}{b_n} \longrightarrow 1. \end{aligned}$$

We write $a_n = \Theta(b_n)$ if both $a_n = O(b_n)$ and $b_n = O(|a_n|)$.

Thus, $o(\cdot)$ is a relative statement of negligibility, $O(\cdot)$ is an eventual upper bound up to a constant, and $\sim$ is a relative-error approximation. In particular,

$$a_n = (1 + o(1))b_n \iff a_n \sim b_n,$$

provided $b_n \neq 0$ eventually.

Example 3. As $n \rightarrow \infty$,

$$\log n = o(n^\alpha) \text{ for every fixed } \alpha > 0, \quad \left(1 + \frac{x}{n}\right)^n \longrightarrow e^x$$

for every fixed $x \in \mathbb{R}$. Consequently,

$$\left(1 + \frac{x}{n}\right)^n = e^x(1 + o(1)).$$

The word “fixed” is important: the conclusion need not remain valid uniformly if $x = x_n$ is allowed to grow with n .

Useful calculation rules. If the expressions are well defined and the comparison sequences are eventually positive, then:

$$\begin{aligned} a_n = o(b_n), c_n = O(d_n) &\implies a_n c_n = o(b_n d_n), \\ a_n = O(b_n), c_n = O(d_n) &\implies a_n + c_n = O(b_n + d_n), \\ a_n = o(1) &\implies \frac{1}{1 + a_n} = 1 - a_n + O(a_n^2). \end{aligned}$$

The last identity is a first-order Taylor expansion.

Constants and uniformity. If a_n also depends on a parameter θ , the statement $a_n(\theta) = O(b_n)$ usually means that the constant may depend on θ . The stronger statement

$$\sup_{\theta \in \Theta} |a_n(\theta)| = O(b_n)$$

uses one constant uniformly over Θ . Statistical arguments involving a supremum over a parameter space generally require the uniform form.

Remark 2 (A common logical error). The statement $a_n = O(b_n)$ does not imply that a_n/b_n converges. For example, $a_n = (-1)^n$ is $O(1)$, but it has no limit.

3 Gaussian calculations

3.1 One-dimensional and multivariate Gaussian variables

High priority *Gaussian vectors and affine transformations*

The Gaussian distribution appears throughout mathematical statistics because it is stable under linear transformations and because normalized sums often approach it.

Definition 6 (Standard normal distribution). *A random variable G has the standard normal distribution, written $G \sim N(0, 1)$, if it has density*

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$

Its upper-tail function is

$$\bar{\Phi}(t) = \mathbb{P}(G > t) = \int_t^\infty \phi(x) dx, \quad t \in \mathbb{R}.$$

The identity $\int_{-\infty}^\infty e^{-x^2/2} dx = \sqrt{2\pi}$ ensures that ϕ integrates to one.

Definition 7 (Gaussian vector). *A random vector $X \in \mathbb{R}^d$ is Gaussian if every linear combination $v^\top X$ is a one-dimensional Gaussian random variable. We write $X \sim N(\mu, \Sigma)$ when $\mathbb{E} X = \mu$ and $\text{Cov}(X) = \Sigma$. The covariance matrix Σ is symmetric and positive semidefinite; it may be singular.*

If Σ is positive definite, X has the multivariate Gaussian density

$$f_X(x) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}, \quad x \in \mathbb{R}^d.$$

When Σ is singular, the same Gaussian-vector definition applies, but the distribution is supported on a lower-dimensional affine subspace and does not have a density with respect to d -dimensional Lebesgue measure.

The defining projection property makes the most useful calculation immediate: for every fixed $v \in \mathbb{R}^d$,

$$v^\top X \sim N(v^\top \mu, v^\top \Sigma v).$$

More generally, stacking several linear projections gives the following affine-transformation rule.

Proposition 1 (Affine transformations). *If $X \sim N(\mu, \Sigma)$, $A \in \mathbb{R}^{m \times d}$, and $b \in \mathbb{R}^m$, then*

$$AX + b \sim N(A\mu + b, A\Sigma A^\top).$$

In particular, if $G \sim N(0, I_d)$, then $AG \sim N(0, AA^\top)$.

Remark 3. *Uncorrelated random variables need not be independent in general. For a **jointly Gaussian** vector, however, **zero covariance does imply independence**. Thus, if $X \sim N(\mu, \Sigma)$ and Σ is diagonal, the coordinates of X are independent.*

Predict a projected variance. Consider the fixed two-dimensional Gaussian model

$$\mu = \begin{pmatrix} 0.5 \\ -0.3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1.6 & 0.9 \\ 0.9 & 1.1 \end{pmatrix}, \quad u_\theta = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}.$$

Before comparing directions, predict $\text{Var}(u_{45^\circ}^\top X)$. Does the positive covariance make it larger than either coordinate variance? Report a decimal value in squared coordinate units (scientific notation is also acceptable); give three decimal places.

The projection rule gives

$$\mathbb{E}[u_\theta^\top X] = 0.5 \cos \theta - 0.3 \sin \theta, \quad \text{Var}(u_\theta^\top X) = 1.6 \cos^2 \theta + 2(0.9) \sin \theta \cos \theta + 1.1 \sin^2 \theta.$$

At 45° , both coordinates of u_θ equal $1/\sqrt{2}$, so the mean is $0.2/\sqrt{2}$ and the variance is

$$\frac{1.6 + 2(0.9) + 1.1}{2} = 2.25.$$

The two covariance cross terms contribute 0.9: omitting them gives 1.35. The standard deviation is 1.5, whereas the requested variance is 2.25. The explanation is available whether or not you try the check.

Static comparison. These exact parameter calculations give the comparison without a simulation:

θ	0°	45°	90°
$\mathbb{E}[u_\theta^\top X]$	0.5	$0.2/\sqrt{2}$	-0.3
$\text{Var}(u_\theta^\top X)$	1.6	2.25	1.1

Rotating the direction changes the projections of the same observations, not the observations themselves. A finite projected sample illustrates the Gaussian law; its histogram need not match the reference density exactly. Positive semidefiniteness: every covariance satisfies $u^\top \Sigma u = \text{Var}(u^\top X) \geq 0$. If this variance is zero, the projection equals its mean almost surely: it is a point mass, not a density of positive width. The displayed covariance is positive definite, since $1.6 > 0$ and $\det \Sigma = 0.95 > 0$. For a related variance calculation, see Exercise 6.

3.2 Gaussian tails and Mills' ratio

Mid priority *Two-sided tail refinement and its proof*

Proposition 2 (Mills' ratio). *For every $t > 0$,*

$$\frac{t}{t^2 + 1} \phi(t) \leq \bar{\Phi}(t) \leq \frac{\phi(t)}{t}.$$

Consequently, the logarithm of the Gaussian upper tail has leading term $-t^2/2$ as $t \rightarrow \infty$.

Proof. Since $\phi'(x) = -x\phi(x)$,

$$\bar{\Phi}(t) = \int_t^\infty \phi(x) dx \leq \frac{1}{t} \int_t^\infty x\phi(x) dx = \frac{\phi(t)}{t}.$$

Integration by parts also gives

$$\bar{\Phi}(t) = \frac{\phi(t)}{t} - \int_t^\infty \frac{\phi(x)}{x^2} dx \geq \frac{\phi(t)}{t} - \frac{\bar{\Phi}(t)}{t^2}.$$

Rearranging yields the lower bound. □

Used again in Gaussian-type tail bounds and the optional thresholding lecture.

Mid priority *Tail calculations and the $\sqrt{\log d}$ noise scale*

The moment-generating function gives the convenient, slightly looser bound

$$\mathbb{P}(|G| \geq t) \leq 2e^{-t^2/2}, \quad t \geq 0.$$

Example 4 (The noise level in many coordinates). Let $G_1, \dots, G_d$ be standard normal random variables; *independence is not needed for this bound*. By a union bound,

$$\mathbb{P}\left(\max_{1 \leq j \leq d} |G_j| \geq t\right) \leq 2de^{-t^2/2}.$$

Therefore, with probability at least $1 - \delta$,

$$\max_{1 \leq j \leq d} |G_j| \leq \sqrt{2 \log(2d/\delta)}.$$

The scale $\sqrt{\log d}$ is a recurring high-dimensional noise threshold.

Instructor's Note *A direct Gaussian tail bound*

Moment-generating functions are a powerful, general-purpose method that we will study later. Here we can obtain the tail bound by combining Proposition 2 with a direct calculation using the Gaussian density.

Recall Mills' ratio: its upper bound is $\bar{\Phi}(t) \leq \phi(t)/t$ for $t > 0$.

For $t > 1$, the upper bound in Proposition 2 is tighter than the simpler bound $2e^{-t^2/2}$: by symmetry,

$$\mathbb{P}(|G| \geq t) = 2\bar{\Phi}(t) \leq \frac{2\phi(t)}{t} = \frac{\sqrt{2/\pi}}{t} e^{-t^2/2} < e^{-t^2/2} \leq 2e^{-t^2/2}.$$

For $0 < t < 1$, a change of variables gives a simple direct argument. In fact, this argument works for every $t \geq 0$ and improves the constant 2 to 1. Substitute $x = t + u$:

$$\begin{aligned} \bar{\Phi}(t) &= \int_0^\infty \phi(t+u) du \\ &= e^{-t^2/2} \int_0^\infty e^{-tu} \phi(u) du \\ &\leq e^{-t^2/2} \int_0^\infty \phi(u) du = \frac{1}{2} e^{-t^2/2}. \end{aligned}$$

The only inequality is $e^{-tu} \leq 1$, since $t, u \geq 0$; the last equality uses symmetry of the standard normal density. Multiplying by 2 gives

$$\boxed{\mathbb{P}(|G| \geq t) \leq e^{-t^2/2} \leq 2e^{-t^2/2}, \quad t \geq 0.}$$

This also covers $t = 0$ and $t = 1$, and establishes the bound used in the noise-scale calculation without moment-generating functions.

4 Combinatorics, Stirling's formula, and entropy

Mid priority *Approximating factorials, binomial masses, and combinatorial growth*

For integers $0 \leq k \leq n$,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad n! = 1 \cdot 2 \cdots n.$$

The binomial coefficient counts subsets of size k selected from n objects. In probability it also describes the mass function of a binomial random variable.

Theorem 2 (Stirling's formula). As $n \rightarrow \infty$,

$$n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n (1 + o(1)).$$

Equivalently,

$$\log(n!) = n \log n - n + \frac{1}{2} \log(2\pi n) + o(1).$$

Example 5 (A local approximation to the binomial distribution). Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$, n is even, and $k = k_n$ is an integer satisfying $|k| = O(\sqrt{n})$. Applying Stirling's formula to the three factorials gives

$$\mathbb{P}\left(\sum_{i=1}^n X_i = \frac{n}{2} + k\right) = 2^{-n} \binom{n}{n/2 + k} = (1 + o(1)) \sqrt{\frac{2}{\pi n}} \exp\left(-\frac{2k^2}{n}\right).$$

This is a local normal approximation to $\text{Binomial}(n, 1/2)$ near its mean.

Low priority Binary entropy

For $\gamma \in (0, 1)$, define the binary entropy function, using natural logarithms, by

$$h(\gamma) = -\gamma \log \gamma - (1 - \gamma) \log(1 - \gamma).$$

If $m = m_n$ and $m/n \rightarrow \gamma \in (0, 1)$, then

$$\log \binom{n}{m} = nh(m/n) + O(\log n) = (h(\gamma) + o(1))n.$$

More precisely, if m/n remains bounded away from 0 and 1, then

$$\binom{n}{m} = (1 + o(1)) \frac{\exp\{nh(m/n)\}}{\sqrt{2\pi n(m/n)(1 - m/n)}}.$$

Remark 4 (Why entropy appears). The quantity $h(\gamma)$ measures the exponential growth rate of the number of subsets with approximately γn selected elements. With base-2 logarithms, the same count is approximately $2^{nh_2(\gamma)}$, where $h_2(\gamma) = h(\gamma)/\log 2$. This combinatorial interpretation connects entropy to coding, likelihoods, and large-deviation probabilities.

5 Linear algebra for statistics

5.1 Trace, covariance, and quadratic forms

High priority Trace identities, covariance matrices, and quadratic-form calculations

For a square matrix M , define

$$\text{Tr}(M) = \sum_i M_{ii}.$$

If the matrix products are dimensionally compatible, the trace is invariant under cyclic permutations:

$$\text{Tr}(ABC) = \text{Tr}(BCA) = \text{Tr}(CAB).$$

It is not generally invariant under an arbitrary reordering of the factors.

Proposition 3 (Expected quadratic form). *Let $X \in \mathbb{R}^d$ have mean μ and covariance matrix Σ , and let $M \in \mathbb{R}^{d \times d}$. If the relevant second moments are finite, then*

$$\mathbb{E}[X^\top M X] = \text{Tr}(M\Sigma) + \mu^\top M \mu.$$

For a centered random vector, this reduces to $\mathbb{E}[X^\top M X] = \text{Tr}(M\Sigma)$.

Proof. Use $X^\top M X = \text{Tr}(M X X^\top)$ and $\mathbb{E}[X X^\top] = \Sigma + \mu\mu^\top$. $\square$

A real symmetric matrix A is **positive semidefinite**, written $A \succeq 0$, if $x^\top A x \geq 0$ for every $x \in \mathbb{R}^d$. It is positive definite, written $A \succ 0$, if the inequality is strict for every nonzero x . Every covariance matrix is positive semidefinite because

$$v^\top \Sigma v = \text{Var}(v^\top X) \geq 0.$$

5.2 Eigenvalues and perturbation

High priority *Spectral theorem and Rayleigh-quotient calculations*

Let $A \in \mathbb{R}^{d \times d}$ be symmetric. The spectral theorem gives an orthogonal matrix Q and real eigenvalues $\lambda_1(A) \geq \dots \geq \lambda_d(A)$ such that

$$A = Q \text{diag}\{\lambda_1(A), \dots, \lambda_d(A)\} Q^\top.$$

The Rayleigh quotient is

$$R_A(x) = \frac{x^\top A x}{x^\top x}, \quad x \neq 0.$$

It satisfies

$$\lambda_{\max}(A) = \max_{\|x\|_2=1} x^\top A x, \quad \lambda_{\min}(A) = \min_{\|x\|_2=1} x^\top A x.$$

Low priority *Courant–Fischer min–max characterization—not examined*

Theorem 3 (Courant–Fischer min–max theorem). *For $1 \leq k \leq d$,*

$$\begin{aligned} \lambda_k(A) &= \max_{\substack{V \subset \mathbb{R}^d \\ \dim(V)=k}} \min_{\substack{x \in V \\ \|x\|_2=1}} x^\top A x \\ &= \min_{\substack{V \subset \mathbb{R}^d \\ \dim(V)=d-k+1}} \max_{\substack{x \in V \\ \|x\|_2=1}} x^\top A x. \end{aligned}$$

Mid priority *Weyl’s eigenvalue perturbation bound*

Theorem 4 (Weyl’s perturbation bound). *If $A, E \in \mathbb{R}^{d \times d}$ are symmetric, then for every $1 \leq k \leq d$,*

$$|\lambda_k(A + E) - \lambda_k(A)| \leq \|E\|_{\text{op}},$$

where

$$\|E\|_{\text{op}} = \max_{\|x\|_2=1} \|Ex\|_2 = \max_{1 \leq j \leq d} |\lambda_j(E)|.$$

Thus, a small operator-norm perturbation cannot move any eigenvalue very far. This is the basic bridge from concentration of a sample covariance matrix to accuracy of its empirical eigenvalues.

The corresponding complex-valued results hold for **Hermitian** matrices, which satisfy $A = A^*$, where $A^* = \overline{A}^\top$ is the conjugate transpose.

Used again in random-matrix norms and PCA perturbation.

5.3 Singular value decomposition and matrix norms

High priority *Singular value decomposition*

Let $A \in \mathbb{R}^{n \times m}$ be nonzero with rank $\rho \geq 1$. Its thin singular value decomposition (SVD) is

$$A = USV^\top,$$

where $U \in \mathbb{R}^{n \times \rho}$ and $V \in \mathbb{R}^{m \times \rho}$ have orthonormal columns and

$$S = \text{diag}\{s_1(A), \dots, s_\rho(A)\}, \quad s_1(A) \geq \dots \geq s_\rho(A) > 0.$$

The nonzero eigenvalues of $A^\top A$ and $AA^\top$ are $s_1(A)^2, \dots, s_\rho(A)^2$.

Mid priority *Variational characterization and operator/Frobenius norms*

The largest singular value has the variational representations

$$s_1(A) = \max_{\|y\|_2=1} \|Ay\|_2 = \max_{\substack{\|x\|_2=1 \\ \|y\|_2=1}} x^\top Ay.$$

Two matrix norms used throughout the course are

$$\|A\|_{\text{op}} = s_1(A), \quad \|A\|_{\text{F}} = \left(\sum_{i,j} A_{ij}^2 \right)^{1/2} = \left(\sum_{j=1}^{\rho} s_j(A)^2 \right)^{1/2}.$$

They satisfy

$$\|A\|_{\text{op}} \leq \|A\|_{\text{F}} \leq \sqrt{\rho} \|A\|_{\text{op}}.$$

Low priority *Optimal low-rank approximation—not examined*

Theorem 5 (Eckart–Young–Mirsky). *For $1 \leq r < \rho$, define*

$$A_r = \sum_{j=1}^r s_j(A) u_j v_j^\top.$$

Then A_r is a best rank- r approximation to A in both operator and Frobenius norm:

$$\begin{aligned} \min_{\text{rank}(B) \leq r} \|A - B\|_{\text{op}} &= s_{r+1}(A), \\ \min_{\text{rank}(B) \leq r} \|A - B\|_{\text{F}} &= \left(\sum_{j=r+1}^{\rho} s_j(A)^2 \right)^{1/2}. \end{aligned}$$

5.4 PCA as variance maximization and low-rank approximation

Mid priority *Equivalent formulations of PCA*

Let $X \in \mathbb{R}^{n \times d}$ be a nonzero **centered data matrix**: row i is the feature vector for observation i , after subtracting the sample mean. No assumption that $n > d$ or that X has full rank is needed. Write

$$X = USV^\top, \quad V = [v_1, \dots, v_\rho].$$

Then:

- v_j is the j th **principal direction** or loading vector;
- $Xv_j = s_j u_j$ is the vector of j th principal-component scores;
- the empirical covariance matrix $\widehat{\Sigma} = X^\top X/n$ has eigenvectors v_j and eigenvalues s_j^2/n ; and
- $X_r = \sum_{j=1}^r s_j u_j v_j^\top$ is the best rank- r reconstruction of the centered data.

The first principal direction also solves

$$v_1 \in \arg \max_{\|v\|_2=1} \frac{1}{n} \|Xv\|_2^2 = \arg \max_{\|v\|_2=1} v^\top \widehat{\Sigma} v.$$

Thus, the eigenvalue view, the variance-maximization view, and the low-rank approximation view describe the same procedure.

6 Vector norms, inequalities, and duality

Mid priority *Norm calculations, inequalities, and duality*

A function $\|\cdot\| : \mathbb{R}^d \rightarrow [0, \infty)$ is a norm if, for all $x, y \in \mathbb{R}^d$ and scalars a :

1. $\|x\| = 0$ if and only if $x = 0$;
2. $\|ax\| = |a| \|x\|$;
3. $\|x + y\| \leq \|x\| + \|y\|$.

For $1 \leq p < \infty$,

$$\|x\|_p = \left(\sum_{j=1}^d |x_j|^p \right)^{1/p}, \quad \|x\|_\infty = \max_{1 \leq j \leq d} |x_j|.$$

The Euclidean inner product is $\langle x, y \rangle = x^\top y$. For nonzero x, y , the angle between them is defined by

$$\cos \theta = \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}.$$

Theorem 6 (Cauchy–Schwarz and Hölder). *For $x, y \in \mathbb{R}^d$,*

$$|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2.$$

More generally, if $1 \leq p, q \leq \infty$ and $1/p + 1/q = 1$, then

$$|\langle x, y \rangle| \leq \|x\|_p \|y\|_q.$$

Cauchy–Schwarz implies the triangle inequality for the Euclidean norm:

$$\begin{aligned} \|x + y\|_2^2 &= \|x\|_2^2 + \|y\|_2^2 + 2\langle x, y \rangle \\ &\leq \|x\|_2^2 + \|y\|_2^2 + 2\|x\|_2 \|y\|_2 = (\|x\|_2 + \|y\|_2)^2. \end{aligned}$$

Proposition 4 (Norm comparison). *If $1 \leq p \leq q \leq \infty$, then*

$$\|x\|_q \leq \|x\|_p \leq d^{1/p-1/q} \|x\|_q.$$

Important special cases are

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1, \quad \|x\|_1 \leq \sqrt{d} \|x\|_2, \quad \|x\|_2 \leq \sqrt{d} \|x\|_\infty.$$

The powers of d are not cosmetic. For the all-ones vector, $\|\mathbf{1}\|_1 = d$, $\|\mathbf{1}\|_2 = \sqrt{d}$, and $\|\mathbf{1}\|_\infty = 1$, so the dimension-dependent constants are sharp.

Proposition 5 (Duality of ℓ_p norms). *If $1/p + 1/q = 1$, then*

$$\|x\|_p = \max_{\|y\|_q \leq 1} \langle x, y \rangle.$$

Thus, the dual norm of ℓ_p is ℓ_q ; in particular, ℓ_1 and ℓ_∞ are dual, while ℓ_2 is self-dual.

Remark 5 (Statistical interpretation). *Norms encode different notions of size. The ℓ_2 norm is tied to least-squares geometry, ℓ_1 regularization encourages sparse solutions, and ℓ_∞ controls the largest coordinatewise error. Duality explains why an ℓ_1 estimation error is naturally paired with an ℓ_∞ noise term.*

7 Convexity and optimality

High priority Convexity checks and first-order optimality calculations

Definition 8 (Convex sets and functions). *A set $C \subset \mathbb{R}^d$ is **convex** if*

$$tx + (1-t)y \in C \quad \text{for all } x, y \in C, \quad t \in [0, 1].$$

*A function $f : C \rightarrow \mathbb{R}$ is **convex** if*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

*for all $x, y \in C$ and $t \in [0, 1]$. It is **strictly convex** if the inequality is strict whenever $x \neq y$ and $t \in (0, 1)$.*

Geometrically, the graph of a convex function lies below every chord joining two points on the graph. Affine subspaces, Euclidean balls, and half-spaces are convex. A union of convex sets need not be convex, but an arbitrary intersection of convex sets is convex.

Predict the chord gap. Fix x, y and vary t from 0 to 1. When does the gap vanish? Here t weights y , replacing t by $1-t$ in the definition above.

At $z = (1-t)x + ty$, compare **A: average then evaluate**, $f(z)$, with **B: evaluate then average**, $(1-t)f(x) + tf(y)$. Convexity gives $B - A \geq 0$. The figure illustrates this inequality; it does not prove it for every convex function.

For $f(z) = \frac{1}{2}(z - \frac{1}{2})^2 + 0.35$, $x = -1$, $y = 1.5$ and $t = 0.5$, we have $z = 0.25$, $A = 0.38125$, $B = 1.1625$, and gap 0.78125. At $t = 0.8$ the gap is 0.5; at either endpoint $t = 0, 1$, or when $x = y$, the gap is zero.

Proposition 6 (First- and second-order characterizations). *Let $C \subset \mathbb{R}^d$ be open and convex.*

1. *If f is differentiable on C , then f is convex if and only if*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \text{for all } x, y \in C.$$

2. *If f is twice continuously differentiable on C , then f is convex if and only if $\nabla^2 f(x) \succeq 0$ for every $x \in C$. If $\nabla^2 f(x) \succ 0$ everywhere, then f is strictly convex.*

Proposition 7 (Smooth first-order optimality). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be differentiable and convex. A point x^* minimizes f if and only if*

$$\nabla f(x^*) = 0.$$

Proposition 8 (Existence and uniqueness). *If C is nonempty and compact and $f : C \rightarrow \mathbb{R}$ is continuous, a minimizer exists. If, in addition, C is convex and f is strictly convex, the minimizer is unique.*

Convexity also improves the geometry of optimization: every local minimizer of a convex function over a convex set is global. This structural fact does not by itself prescribe a runtime; computational guarantees also depend on how the problem is represented, the desired accuracy, and the available oracle or algorithm.

Example 6 (Least squares and regularization). For $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$,

$$L(\beta) = \|y - X\beta\|_2^2$$

is convex because $\nabla^2 L(\beta) = 2X^\top X \succeq 0$. It is strictly convex if and only if X has full column rank. For $\lambda > 0$,

$$L_{\lambda,q}(\beta) = \|y - X\beta\|_2^2 + \lambda \|\beta\|_q^q$$

is strictly convex when $q > 1$. When $q = 1$, the objective remains convex but is nonsmooth and need not be strictly convex.

7.1 Subgradients and nonsmooth optimality

Low priority Subgradients and nonsmooth optimality—not examined

Definition 9 (Subgradient). For a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, a vector $g \in \mathbb{R}^d$ is a **subgradient** of f at $x \in \mathbb{R}^d$ if

$$f(y) \geq f(x) + g^\top(y - x) \quad \text{for every } y \in \mathbb{R}^d.$$

The set of all subgradients is the **subdifferential** $\partial f(x)$.

If f is differentiable at x , then $\partial f(x) = \{\nabla f(x)\}$. For $f(x) = |x|$,

$$\partial|x| = \begin{cases} \{1\}, & x > 0, \\ [-1, 1], & x = 0, \\ \{-1\}, & x < 0. \end{cases}$$

Proposition 9 (First-order optimality). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex. A point x^* minimizes f if and only if

$$0 \in \partial f(x^*).$$

For the LASSO objective

$$L_{\lambda,1}(\beta) = \|y - X\beta\|_2^2 + \lambda \|\beta\|_1,$$

the optimality condition is

$$0 \in 2X^\top(X\hat{\beta} - y) + \lambda \partial \|\hat{\beta}\|_1.$$

Coordinatewise, the nonsmooth interval at zero creates a range of correlations for which $\hat{\beta}_j = 0$ is optimal. This is the basic mechanism behind sparsity.

Key ideas

- Compactness and continuity help establish existence; they do not imply uniqueness.
- Gaussian affine transformations and tail bounds are reusable probability tools.
- Quadratic forms, eigenvalues and singular values connect linear algebra to statistical variation.
- Norm inequalities and convexity organize estimation and optimization arguments.

Looking ahead

Lecture 2 uses sets and countability to define measurable events and probability measures. This gives the Gaussian and empirical examples a common probability language. Continue to Lecture 2.

8 Extensions and further reading

Low priority *Function spaces and orthogonal expansions—not examined*

Finite-dimensional vectors suggest an infinite-dimensional analogue. Let $I \subset \mathbb{R}$, and let h be a nonnegative density on I satisfying $\int_I h(x) dx = 1$. For $1 \leq q < \infty$, define

$$\|f\|_{L_q(h)} = \left(\int_I |f(x)|^q h(x) dx \right)^{1/q}.$$

For square-integrable functions, define

$$\langle f, g \rangle_{L_2(h)} = \int_I f(x)g(x)h(x) dx.$$

Strictly speaking, two functions that agree except on a set of measure zero represent the same element of $L_q(h)$. Lecture 2 will supply the measure-theoretic language behind this statement.

An orthonormal sequence $(\psi_j)_{j \geq 0}$ satisfies

$$\langle \psi_j, \psi_k \rangle_{L_2(h)} = \begin{cases} 1, & j = k, \\ 0, & j \neq k. \end{cases}$$

It is an **orthonormal basis** if its finite linear combinations are dense in $L_2(h)$. Completeness is essential: an arbitrary orthonormal sequence need not span the space.

For an orthonormal basis, every $f \in L_2(h)$ has the expansion

$$f = \sum_{j=0}^{\infty} \langle f, \psi_j \rangle_{L_2(h)} \psi_j$$

with convergence in $L_2(h)$, and Parseval's identity gives

$$\|f\|_{L_2(h)}^2 = \sum_{j=0}^{\infty} \langle f, \psi_j \rangle_{L_2(h)}^2.$$

Truncating the expansion is an orthogonal projection and gives the best approximation within the span of the retained basis functions. This is the function-space counterpart of least squares and truncated SVD.

Fourier basis. On $I = [-\pi, \pi]$ with $h(x) = 1/(2\pi)$, the functions

$$1, \quad \{\sqrt{2} \cos(kx)\}_{k \geq 1}, \quad \{\sqrt{2} \sin(kx)\}_{k \geq 1}$$

form an orthonormal basis of $L_2(h)$.

Hermite basis. For the standard Gaussian density $h = \phi$, the normalized probabilists' Hermite polynomials begin with

$$H_0(x) = 1, \quad H_1(x) = x, \quad H_2(x) = \frac{x^2 - 1}{\sqrt{2}}, \quad H_3(x) = \frac{x^3 - 3x}{\sqrt{6}}.$$

They form an orthonormal basis of $L_2(\phi)$. If $G \sim N(0, 1)$ and $f(G)$ is square integrable, then

$$f(G) = \sum_{j=0}^{\infty} \mathbb{E}[f(G)H_j(G)]H_j(G)$$

in mean square. This decomposes f into components of increasing polynomial degree and is useful in Gaussian approximation and the analysis of nonlinear statistics.

Low priority *Proof sketches for selected results—not examined*

The Gaussian integral. Let $I = \int_{-\infty}^{\infty} e^{-x^2/2} dx$. By Tonelli's theorem (introduced later) and polar coordinates,

$$\begin{aligned} I^2 &= \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2} dx dy \\ &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta = 2\pi. \end{aligned}$$

Hence $I = \sqrt{2\pi}$.

Weyl's bound from Rayleigh quotients. For every unit vector x ,

$$x^\top Ax - \|E\|_{\text{op}} \leq x^\top (A + E)x \leq x^\top Ax + \|E\|_{\text{op}}.$$

Applying the Courant–Fischer characterization to these inequalities yields Theorem 4.

Why truncated SVD is optimal. Let B have rank at most r . Its null space has codimension at most r . The span of the first $r + 1$ right singular vectors of A must therefore contain a unit vector x with $Bx = 0$. Hence

$$\|A - B\|_{\text{op}} \geq \|(A - B)x\|_2 = \|Ax\|_2 \geq s_{r+1}(A).$$

The truncated SVD attains this lower bound. A related orthogonality argument gives the Frobenius-norm formula.

Further reading *Historical notes, references, and self-study resources*

Stirling's approximation dates to the early eighteenth century, with the form developed by de Moivre and the constant $\sqrt{2\pi}$ identified by Stirling; see the MacTutor historical account. The normal approximation to the binomial distribution is correspondingly known as the de Moivre–Laplace theorem. Least squares was developed around the turn of the nineteenth century by Legendre and Gauss. The geometric method now called principal component analysis traces to Pearson and was later developed systematically by Hotelling.

For analysis and function spaces, see Rudin [3, 2]. Tao's text [4] reviews probability and linear algebra in preparation for random matrix theory. Vershynin [5] and Wainwright [6] connect Gaussian tails, matrix norms, and convexity to modern high-dimensional statistics. Boyd and Vandenberghe [1] is a standard reference for convex optimization.

Useful online supplements include:

- MIT OpenCourseWare 18.06SC: Linear Algebra;

- Boyd and Vandenberghe’s freely available convex-optimization text;
- Vershynin’s *High-Dimensional Probability*, second-edition draft;
- Tao’s *Topics in Random Matrix Theory* draft; and
- Seeing Theory, a visual and interactive introduction to probability and statistics.

9 Exercises and complete solutions

In the future HTML version, this section should be rendered as an expandable block. In the PDF, all contents are shown.

Mid priority Supporting practice: topology and optimizer existence

Exercise 1 (Topology and existence). Let $S = (0, 1] \cup \{2\} \subset \mathbb{R}$.

1. Find $\overline{S}$, S° , and ∂S .
2. Give a continuous function on the bounded open set $(0, 1)$ whose infimum is not attained.
3. Give a continuous function on the closed but unbounded set $[0, \infty)$ whose infimum is not attained.

Solution.

1. Every neighborhood of 0 intersects S , so 0 is added by closure. The point 2 is already in S and is isolated. Thus

$$\overline{S} = [0, 1] \cup \{2\}, \quad S^\circ = (0, 1), \quad \partial S = \{0, 1, 2\}.$$
2. Take $f(x) = x$ on $(0, 1)$. It is continuous and $\inf_{x \in (0, 1)} f(x) = 0$, but no point of the domain has value 0.
3. Take $f(x) = e^{-x}$ on $[0, \infty)$. It is continuous and has infimum 0, approached as $x \rightarrow \infty$, but $e^{-x} > 0$ for every finite x . These examples show why both closedness and boundedness appear in the Heine–Borel theorem.

□

High priority Core practice: asymptotic notation

Exercise 2 (Asymptotic notation). Suppose all comparison sequences below are eventually positive.

1. Show that $\log n = o(n^\alpha)$ for every fixed $\alpha > 0$.
2. If $a_n = O(b_n)$ and $c_n = o(d_n)$, prove that $a_n c_n = o(b_n d_n)$.
3. Give an example in which $a_n = O(b_n)$ but a_n/b_n does not converge.
4. For $a_n(t) = t/n$, explain the difference between $a_n(t) = O(n^{-1})$ for every fixed $t \in \mathbb{R}$ and a uniform $O(n^{-1})$ statement over all $t \in \mathbb{R}$.

Solution.

1. Treating n as a continuous variable, repeated use of l'Hôpital's rule is unnecessary:

$$\lim_{x \rightarrow \infty} \frac{\log x}{x^\alpha} = \lim_{x \rightarrow \infty} \frac{1/x}{\alpha x^{\alpha-1}} = \lim_{x \rightarrow \infty} \frac{1}{\alpha x^\alpha} = 0.$$

Restricting x to the integers gives the result.

2. There are $C < \infty$ and N_1 such that $|a_n|/b_n \leq C$ for $n \geq N_1$. Also, $|c_n|/d_n \rightarrow 0$. Therefore,

$$\frac{|a_n c_n|}{b_n d_n} = \frac{|a_n|}{b_n} \frac{|c_n|}{d_n} \leq C \frac{|c_n|}{d_n} \rightarrow 0.$$

3. Let $a_n = (-1)^n$ and $b_n = 1$. Then $|a_n| \leq b_n$, but $a_n/b_n = (-1)^n$ has no limit.
4. For each fixed t , $|a_n(t)| \leq |t|n^{-1}$, so the hidden constant may be $|t|$. Uniformly over all $t \in \mathbb{R}$,

$$\sup_{t \in \mathbb{R}} |a_n(t)| = \infty,$$

so no finite constant works. Uniformity would hold on a bounded set such as $|t| \leq M$, with constant M .

□

High priority Core practice: Gaussian tails and a union bound

Exercise 3 (Gaussian maxima). Let $G_1, \dots, G_d$ be standard normal random variables, not necessarily independent.

1. Use the Gaussian moment-generating function to prove $\mathbb{P}(G_1 \geq t) \leq e^{-t^2/2}$ for $t \geq 0$.
2. Show that, for $\delta \in (0, 1)$,

$$\mathbb{P}\left(\max_{1 \leq j \leq d} |G_j| > \sqrt{2 \log(2d/\delta)}\right) \leq \delta.$$

Identify exactly where independence would have been used, if it were needed.

Solution. Markov's inequality (Lecture 4) gives $\mathbb{P}(Y \geq a) \leq \mathbb{E} Y/a$ for $Y \geq 0$ and $a > 0$. Completing the square in the normal density yields

$$\mathbb{E} e^{\lambda G_1} = e^{\lambda^2/2} \int_{\mathbb{R}} \phi(x - \lambda) dx = e^{\lambda^2/2}.$$

1. For any $\lambda > 0$, Markov's inequality and $\mathbb{E} e^{\lambda G_1} = e^{\lambda^2/2}$ give

$$\mathbb{P}(G_1 \geq t) = \mathbb{P}(e^{\lambda G_1} \geq e^{\lambda t}) \leq e^{-\lambda t + \lambda^2/2}.$$

For $t > 0$, the exponent is minimized at $\lambda = t > 0$, yielding $e^{-t^2/2}$. At $t = 0$, use the trivial one-sided bound $\mathbb{P}(G_1 \geq 0) \leq 1 = e^0$ separately. By symmetry, $\mathbb{P}(|G_1| \geq t) \leq 2e^{-t^2/2}$.

2. The union bound gives

$$\mathbb{P}\left(\max_j |G_j| > t\right) \leq \sum_{j=1}^d \mathbb{P}(|G_j| > t) \leq 2de^{-t^2/2}.$$

Substituting $t = \sqrt{2 \log(2d/\delta)}$ makes the right-hand side equal to δ . Independence was not used: the union bound is valid for arbitrarily dependent events.

□

Mid priority *Supporting practice: Stirling's approximation*

Exercise 4 (Stirling's formula and the central binomial coefficient). For even n , use Stirling's formula to show

$$\binom{n}{n/2} \sim 2^n \sqrt{\frac{2}{\pi n}}.$$

Solution. Applying Stirling's formula to $n!$ and $(n/2)!$ gives

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n, \quad \left\{\left(\frac{n}{2}\right)!\right\}^2 \sim \pi n \left(\frac{n}{2e}\right)^n.$$

Therefore,

$$\binom{n}{n/2} = \frac{n!}{\{(n/2)!\}^2} \sim \frac{\sqrt{2\pi n}}{\pi n} 2^n = 2^n \sqrt{\frac{2}{\pi n}}.$$

□

Low priority *Optional practice: entropy approximation—not examined*

Exercise 5 (Entropy approximation for binomial coefficients). Let $m = m_n \in \{1, \dots, n-1\}$ and $p_n = m/n$. Assume that p_n remains in a fixed compact subinterval of $(0, 1)$.

1. Use Stirling's formula to show

$$\log \binom{n}{m} = nh(p_n) - \frac{1}{2} \log\{2\pi np_n(1-p_n)\} + o(1).$$

2. Explain the combinatorial meaning of the leading term $nh(p_n)$.

Solution. Stirling's logarithmic form gives

$$\begin{aligned} \log \binom{n}{m} &= \log(n!) - \log(m!) - \log((n-m)!) \\ &= n \log n - m \log m - (n-m) \log(n-m) \\ &\quad - \frac{1}{2} \log\{2\pi np_n(1-p_n)\} + o(1). \end{aligned}$$

Because $m = np_n$ and $n-m = n(1-p_n)$, the first line of leading terms becomes

$$-np_n \log p_n - n(1-p_n) \log(1-p_n) = nh(p_n).$$

The uniform separation of p_n from 0 and 1 ensures that the Stirling remainders for all three factorials are $o(1)$.

The quantity $\binom{n}{m}$ counts binary strings with exactly m ones, or subsets of size m . Thus $h(p_n)$ is the exponential growth rate per coordinate of this collection. □

High priority *Core practice: quadratic forms and PCA*

Exercise 6 (Quadratic forms and population PCA). Let $X \in \mathbb{R}^d$ have mean μ and covariance Σ , and let $M \in \mathbb{R}^{d \times d}$.

1. Derive $\mathbb{E}[X^\top MX] = \text{Tr}(M\Sigma) + \mu^\top M\mu$.
2. Suppose $\mu = 0$. For a unit vector v , interpret $\mathbb{E}[(v^\top X)^2]$ and show that it is maximized by a leading eigenvector of Σ .

Solution. Using the cyclic property of trace,

$$X^\top MX = \text{Tr}(X^\top MX) = \text{Tr}(MX X^\top).$$

Taking expectations,

$$\mathbb{E}[X^\top MX] = \text{Tr}\{M \mathbb{E}(XX^\top)\} = \text{Tr}\{M(\Sigma + \mu\mu^\top)\} = \text{Tr}(M\Sigma) + \mu^\top M\mu.$$

If $\mu = 0$, then

$$\mathbb{E}[(v^\top X)^2] = \text{Var}(v^\top X) = v^\top \Sigma v.$$

The Rayleigh-quotient characterization gives

$$\max_{\|v\|_2=1} v^\top \Sigma v = \lambda_{\max}(\Sigma),$$

and the maximizers are unit vectors in the leading eigenspace. This is population PCA: choose the direction with largest projected variance. $\square$

Mid priority *Supporting practice: norm inequalities and duality*

Exercise 7 (Norm duality). Let $1 < p < \infty$ and let $q = p/(p-1)$.

1. Prove

$$\|x\|_p = \max_{\|y\|_q \leq 1} \langle x, y \rangle$$

and exhibit a maximizing y when $x \neq 0$.

2. State maximizing choices for the endpoint pairs $(p, q) = (1, \infty)$ and $(p, q) = (\infty, 1)$.
3. Use Cauchy-Schwarz to derive $\|x\|_1 \leq \sqrt{d} \|x\|_2$ and $\|x\|_2 \leq \sqrt{d} \|x\|_\infty$.

Solution. Hölder's inequality gives $|\langle x, y \rangle| \leq \|x\|_p \|y\|_q \leq \|x\|_p$, so it remains to attain the upper bound. For $x \neq 0$, define

$$y_j = \frac{\text{sign}(x_j) |x_j|^{p-1}}{\|x\|_p^{p-1}}.$$

Because $(p-1)q = p$,

$$\|y\|_q^q = \frac{\sum_j |x_j|^p}{\|x\|_p^p} = 1, \quad \langle x, y \rangle = \frac{\sum_j |x_j|^p}{\|x\|_p^{p-1}} = \|x\|_p.$$

For $p = 1$, take $y_j = \text{sign}(x_j)$, which has $\|y\|_\infty \leq 1$. For $p = \infty$, choose an index j^* with $|x_{j^*}| = \|x\|_\infty$ and take $y = \text{sign}(x_{j^*}) e_{j^*}$.

Finally,

$$\|x\|_1 = \langle x, \mathbf{1} \rangle \leq \|x\|_2 \|\mathbf{1}\|_2 = \sqrt{d} \|x\|_2,$$

and

$$\|x\|_2^2 = \sum_{j=1}^d x_j^2 \leq d \|x\|_\infty^2.$$

$\square$

Mid priority *Supporting practice: spectral perturbation*

Exercise 8 (Spectral perturbation). Let A and $\hat{A} = A + E$ be symmetric $d \times d$ matrices with $\|E\|_{\text{op}} \leq \epsilon$. Let $\hat{v}_1$ be a unit leading eigenvector of $\hat{A}$.

1. Bound $|\lambda_k(\hat{A}) - \lambda_k(A)|$ for every k .
2. Show that

$$\hat{v}_1^\top A \hat{v}_1 \geq \lambda_1(A) - 2\epsilon.$$

Interpret the result when A is a population covariance and $\hat{A}$ is a sample covariance.

Solution. Weyl's bound immediately gives

$$|\lambda_k(\hat{A}) - \lambda_k(A)| \leq \epsilon \quad \text{for every } k.$$

For the second claim, compare the Rayleigh value at the same unit vector under A and $\hat{A}$, then compare their leading eigenvalues. Identify the reason for each transition; the complete argument follows.

1. Substitute the perturbation.

$$\hat{v}_1^\top A \hat{v}_1 = \hat{v}_1^\top \hat{A} \hat{v}_1 - \hat{v}_1^\top E \hat{v}_1.$$

Use $A = \hat{A} - E$; the vector is held fixed in both terms.

2. Use the eigenvector identity and norm bound.

$$\hat{v}_1^\top A \hat{v}_1 \geq \lambda_1(\hat{A}) - \epsilon.$$

Since $\hat{A} \hat{v}_1 = \lambda_1(\hat{A}) \hat{v}_1$ and $\|\hat{v}_1\|_2 = 1$, its Rayleigh value is $\lambda_1(\hat{A})$. Cauchy-Schwarz and the definition of operator norm give $|\hat{v}_1^\top E \hat{v}_1| \leq \|\hat{v}_1\|_2 \|E \hat{v}_1\|_2 \leq \|E\|_{\text{op}} \leq \epsilon$.

3. Apply Weyl to the leading eigenvalue.

$$\hat{v}_1^\top A \hat{v}_1 \geq \lambda_1(\hat{A}) - \epsilon \geq \lambda_1(A) - 2\epsilon.$$

Theorem 4 gives $\lambda_1(\hat{A}) \geq \lambda_1(A) - \epsilon$. This uses the mid-priority theorem statement, not its optional Courant–Fischer proof. Thus, operator-norm accuracy of the sample covariance implies eigenvalue accuracy and guarantees that the empirical leading direction captures population variance within 2ϵ of the optimum. Controlling the angle between the eigenvectors requires an additional eigengap condition and a result such as the Davis–Kahan theorem (an extension, not needed here). $\square$

Mid priority *Supporting practice: soft thresholding*

Exercise 9 (One-dimensional LASSO). For $y \in \mathbb{R}$ and $\lambda > 0$, find the unique minimizer of

$$f(\beta) = (y - \beta)^2 + \lambda|\beta|.$$

You may use the subgradient optimality condition, or work separately on $\beta < 0$, $\beta = 0$, and $\beta > 0$. Explain why the solution can be exactly zero.

Solution. One can solve the problem without subgradients. On $\beta > 0$, differentiation gives $2(\beta - y) + \lambda = 0$, so $\beta = y - \lambda/2$, which is positive exactly when $y > \lambda/2$. If $\beta < 0$, differentiation gives $2(\beta - y) - \lambda = 0$, so $\beta = y + \lambda/2$, which is negative exactly when $y < -\lambda/2$. At $\beta = 0$, the left and right derivatives are $-2y - \lambda$ and $-2y + \lambda$, respectively. They straddle zero exactly when $|y| \leq \lambda/2$. Combining the three cases,

$$\hat{\beta} = \text{sign}(y) \left(|y| - \frac{\lambda}{2} \right)_+.$$

Optional alternate route (subgradients; not examined). The condition

$$0 \in 2(\beta - y) + \lambda \partial|\beta|$$

produces the same three cases; at $\beta = 0$, it becomes $0 \in -2y + \lambda[-1, 1]$. This is equivalent to $|y| \leq \lambda/2$.

This is the soft-thresholding rule. The one-sided derivatives straddle zero for a whole interval of data values, $|y| \leq \lambda/2$, giving the exact optimizer $\hat{\beta} = 0$. The squared term is strictly convex, so the minimizer is unique. $\square$

References

- [1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [2] W. Rudin. *Real and Complex Analysis*. Higher Mathematics Series. McGraw-Hill Education, 1987.
- [3] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill, New York, 3 edition, 1976.
- [4] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.
- [5] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- [6] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

What changed

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 14, 2026. Added issue-reporting links to the webpage and PDF. The mathematics and priority assignments are unchanged in this update.