

Lecture 16: Norms of random matrices

Lecturer: Yiqiao Zhong[Report an issue](#)*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.

Priority guide

High Core concepts and calculations to master; directly examined.

Mid Supporting concepts useful in other lectures; in scope for examinations.

Low Enrichment supporting understanding; will not be examined.

Further reading Optional self-study, references, extensions, and historical notes.

A star (*) marks optional material; priority labels give the scope of each block. Lecture-specific exemptions and the optional status of Part VI override the general exam guide.

High priority *Overview*

A matrix acts in infinitely many directions, but a finite net lets us control all those directions at once. We review singular values and norms, derive an operator-norm bound for independent subgaussian entries, and apply the same strategy to empirical covariance matrices. Keep track of which dimension belongs to the observations and which to the matrix; the optional Gaussian result gives sharper spectral edges.

Recap from earlier lectures

From Lectures 1, 14 and 15: directions and concentration

Reminder only: this block introduces no new exam requirements.

- **Uniform spherical directions.** Uniform spherical directions have subgaussian one-dimensional projections despite their dependent coordinates. Used in Section 3.
- **Norm concentration.** Independent unit-variance subgaussian coordinates give $\|X\|_2$ near $\sqrt{n}$. Used in Section 3.
- **Bernstein and union bounds.** Control one direction, then account for all directions in a finite set. Used in the proofs of Theorems 1 and 2; the auxiliary net lemma is optional.
- **SVD and operator norm.** The largest singular value is the largest stretch over unit inputs. Used in Section 1.

An asterisk in a topic heading marks optional material; the priority labels specify the reading scope.

Learning goals

- Relate singular values to operator and Frobenius norms.
- Use the operator-norm bound for a subgaussian random matrix.
- Interpret the sample-size and dimension dependence of covariance estimation.
- Choose a discretization radius; the net lemma and its proof remain optional.

High priority Introduction

For convenience, we refer to the Euclidean norm as $\|\cdot\|$ and $\|\cdot\|_2$. As before, we use c or C to denote a positive absolute constant. We write c_K or C_K to denote a scalar that depends only on K . In particular, if K is a constant, then c_K and C_K are both constants. We denote by $\mathbb{S}^{n-1}$ the unit sphere: $\{x \in \mathbb{R}^n : \|x\| = 1\}$.

1 Some linear algebra

1.1 A recap of singular value decomposition (SVD)

Mid priority A recap of singular value decomposition (SVD)

Let $A \in \mathbb{R}^{m \times n}$ be an $m \times n$ matrix. It has a decomposition $A = \sum_{i=1}^r \sigma_i u_i v_i^\top = U \Sigma V^\top$, where $r = \text{rank}(A)$, $U = [u_1, \dots, u_r] \in \mathbb{R}^{m \times r}$ has orthonormal columns, $V = [v_1, \dots, v_r] \in \mathbb{R}^{n \times r}$ has orthonormal columns, and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r \times r}$. We use $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ by convention.

Fact 1. $\sigma_i = \sqrt{\lambda_i(AA^\top)} = \sqrt{\lambda_i(A^\top A)}$, i.e., the square root of the i -th largest eigenvalue of $AA^\top$ or $A^\top A$; u_i and v_i are corresponding eigenvectors of $AA^\top$ and $A^\top A$, respectively. We have $A^\top u_i = V \Sigma U^\top u_i = \sigma_i v_i$. Similarly $Av_i = \sigma_i u_i$.

Fact 2. If A is symmetric ($A = A^\top$), then the singular values are the absolute eigenvalues, sorted in descending order.

Fact 3. $\sigma_1(A) = \max_{u: \|u\|=1} \|Au\|$ is called the **operator norm** of A .

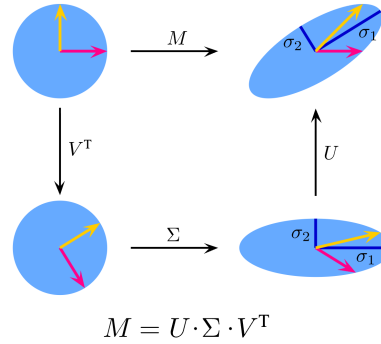


Figure 1: SVD in three steps: rotate the input, stretch along coordinate axes, then rotate the output. Here $V^\top$ rotates the input and Σ is the diagonal matrix of singular values; the colored arrows track the same two vectors. Figure: Georg-Johann, adapted by replacing the adjoint labels with $V^\top$, CC BY-SA 3.0.

Revisit the SVD construction in Lecture 1 to connect these three steps to the matrix factorization.

1.2 Matrix norms

High priority Matrix norms

There are many different matrix norms we encounter. Here are a few common norms. Let $A \in \mathbb{R}^{m \times n}$ denote a generic matrix.

- **Frobenius norm** $\|A\|_F = (\sum_{ij} A_{ij}^2)^{1/2}$. Essentially, we treat A as a vector of length mn and then use the Euclidean norm. Frobenius norm can be expressed using singular values $\|A\|_F^2 = \sum_{j=1}^r [\sigma_j(A)]^2$.

- Operator norm $\|A\|_{\text{op}} = \sigma_1(A)$, and by convention, simply $\|A\|$ unless specified otherwise. From the functional analysis perspective, we view A as a linear operator that maps a vector to another vector in the Euclidean space, and the induced norm is the operator norm. It can be shown that $\|A\|_{\text{op}} = \max_{u: \|u\|_2=1} \|Au\|_2$.
- Matrix p -norm $\|A\|_p = \max_{u: \|u\|_p=1} \|Au\|_p$. It generalizes the operator norm $p = 2$ to more general p ; in fact, it is the induced operator norm for ℓ^p spaces.
- Max norm $\|A\|_{\text{max}} = \max_{ij} |A_{ij}|$. Essentially, we treat A as a vector of length mn and consider the usual ℓ^∞ norm.

See also Wiki for a comprehensive comparison.

Operator norm. The operator norm is the largest singular value and controls the largest stretch over all unit input vectors. The Frobenius norm is the square root of the sum of squared singular values.

High priority Practice: Compare three matrix norms

Exercise prompt. For $A = \text{diag}(3, 4)$, compute $\|A\|_{\text{op}}$, $\|A\|_F$, and $\|A\|_{\text{max}}$. Explain what each norm measures and which values coincide.

See Exercise 1 in Section 7 for the full statement and complete solution.

1.3 Covering and packing numbers

High priority Covering and packing numbers

Definition 1 (ϵ -net). Let (T, d) be a metric space, then consider a subset $K \subset T$ and some $\epsilon > 0$, then a subset $\mathcal{N} \subset K$ is called an ϵ -net of K if every point x in K is within distance ϵ of a point in $\mathcal{N}$, i.e. $\forall x \in K, \exists x_0 \in \mathcal{N}$ such that $d(x_0, x) \leq \epsilon$. In other words, K is covered by balls of radius ϵ centered at every point in $\mathcal{N}$.

Definition 2 (covering number). The covering number $N(K, d, \epsilon)$ is the number of points in the smallest (i.e. containing the fewest points) ϵ -net of K , i.e., the least number of points that are needed to cover K with ϵ -balls among all possible ϵ -nets.

Proposition 1. Covering number of Euclidean unit ball $B_2^n = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$ in $\mathbb{R}^n$. $(\frac{1}{\epsilon})^n \leq N(B_2^n, d_2, \epsilon) \leq (1 + \frac{2}{\epsilon})^n$ for $0 < \epsilon \leq 1$, here $d_2(x, y) = \|x - y\|_2$ is Euclidean distance.

The upper bound follows from the packing proof in Exercise 6; the lower bound follows because covering a unit ball requires total covering-ball volume at least its volume. The same packing argument, with centers constrained to the sphere, also gives a sphere net of size at most $(1 + 2/\epsilon)^n$. See [2, Section 4.2].

Low priority Discretization lemma

Lemma 1. If $0 < \epsilon < 1$ and $\mathcal{N}$ is an ϵ -net of the unit sphere $\mathbb{S}^{n-1}$, then

$$\sup_{x \in \mathcal{N}} \|Ax\|_2 \leq \|A\| \leq \frac{1}{1 - \epsilon} \sup_{x \in \mathcal{N}} \|Ax\|_2.$$

If $0 < \epsilon < 1/2$ and in addition there is an ϵ -net $\mathcal{M}$ of $\mathbb{S}^{m-1}$, then

$$\sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \leq \|A\| \leq \frac{1}{1 - 2\epsilon} \cdot \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle.$$

Proof. $\forall x \in \mathbb{S}^{n-1}$, we find $x_0 \in \mathcal{N}$ s.t. $\|x - x_0\| \leq \epsilon$. Pick x to be the top singular vector v_1 , then we have

$$\begin{aligned}\|Ax - Ax_0\| &\leq \|A\| \cdot \|x - x_0\| \leq \epsilon \|A\|, \\ \|Ax_0\| &\geq \|Ax\| - \|Ax - Ax_0\| \geq \|A\| - \epsilon \|A\|, \\ \|A\| &\leq \frac{1}{1 - \epsilon} \|Ax_0\|.\end{aligned}$$

In addition we pick $y = Ax/\|Ax\| \in \mathbb{S}^{m-1}$ such that $\langle Ax, y \rangle = \|Ax\| = \|A\|$ (the zero matrix is immediate). Then take $y_0 \in \mathcal{M}$ such that $\|y - y_0\| \leq \epsilon$. By $\langle Ax, y \rangle - \langle Ax_0, y_0 \rangle = \langle Ax, y - y_0 \rangle + \langle A(x - x_0), y_0 \rangle$, we follow the above steps and get the desired result. $\square$

Remark 1. This lemma says that if we don't care about the factor (usually $1/(1 - \epsilon)$ is a constant), then the operator norm $\|A\| = \sup_{\|u\|=1} \|Au\|$ is approximately the same as a discretized version.

High priority

Instructor's Note Check a finite set of directions

A net is like checking a surface on a fine grid. We first control the grid points by probability bounds, then pay a deterministic approximation factor to fill the gaps. Making the grid finer improves that factor but increases the number of events in the union bound.

Low priority Discretize directions

An epsilon-net approximates every unit direction within distance epsilon. Net cardinality is exponential in dimension, and its logarithm competes with concentration in uniform matrix bounds.

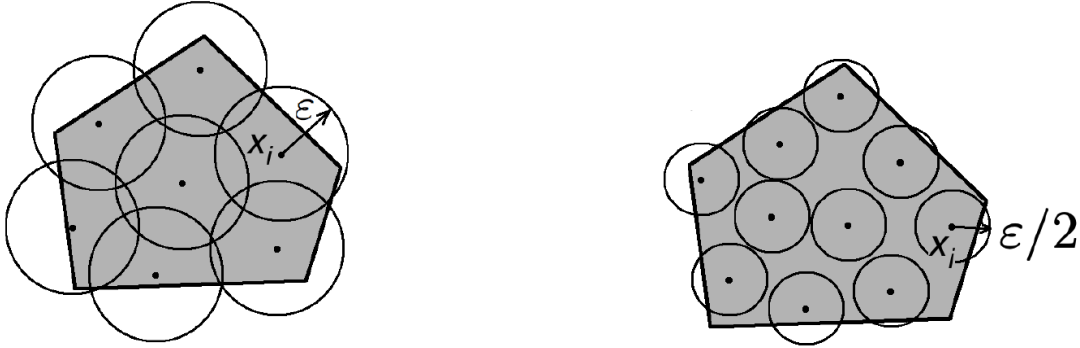


Figure 2: Left: radius- ϵ balls cover the set. Right: an ϵ -separated set has disjoint radius- $\epsilon/2$ balls. Compare the total volume of these disjoint balls with a containing region to bound the packing size. Figure: Vershynin, Figure 4.1, diagram excerpt.

Mid priority Practice: Choose a discretization radius

Exercise prompt. For the two-net bilinear bound, compare $\epsilon = 1/4$ and $\epsilon = 1/8$: give the approximation factors and the covering-number upper bounds in dimensions n and m . Use the approximation factor $(1 - 2\epsilon)^{-1}$ and the covering bound $(1 + 2/\epsilon)^d$ in dimension d .

See Exercise 2 in Section 7 for the full statement and complete solution.

2 Operator norm of a subgaussian random matrix

High priority *Operator norm of a subgaussian random matrix*

The goal is to control the largest stretch of a random matrix. For independent centered subgaussian entries with norm bounded by K , the relevant upper-bound scale is $K(\sqrt{m} + \sqrt{n})$; the theorem below states the failure probability and the additional deviation term explicitly.

We sometimes say $A \in \mathbb{R}^{m \times n}$ has independent subgaussian entries if all the entries (A_{ij}) are jointly independent, $\mathbf{E}A_{ij} = 0$, and subgaussian.

Theorem 1. *Let $A \in \mathbb{R}^{m \times n}$ have independent subgaussian entries. $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$. For any $t > 0$, $\|A\| \leq CK(\sqrt{m} + \sqrt{n} + t)$ with probability at least $1 - 2e^{-t^2}$ for a certain constant C .*

We need the following lemma, which is in fact an equivalent form of Hoeffding's inequality.

Lemma 2. *Let $X_1, \dots, X_n$ be independent subgaussian random variables with $\mathbf{E}X_i = 0$, then $\sum_{i=1}^n X_i$ is also subgaussian with*

$$\left\| \sum_{i=1}^n X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^n \|X_i\|_{\psi_2}^2.$$

Proof of Theorem 1. Use the displayed net bound below as a tool; the general approximation lemma and its proof above are optional. Pick $\epsilon = \frac{1}{4}$, let $\mathcal{N}, \mathcal{M}$ be ϵ -nets of $\mathbb{S}^{n-1}, \mathbb{S}^{m-1}$ respectively such that $|\mathcal{N}| \leq 9^n$, $|\mathcal{M}| \leq 9^m$. (Note: $\mathbb{S}^{n-1}$ is the unit sphere: $\{x \in \mathbb{R}^n \mid \|x\| = 1\}$.) We can check from Lemma 1 that

$$\|A\| \leq 2 \max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the inner product notation (namely $\langle z, w \rangle = z^\top w$). Now we will have to control $\langle Ax, y \rangle$ for each $x \in \mathcal{N}, y \in \mathcal{M}$ with the preceding lemma. Now apply Lemma 2 to $\langle Ax, y \rangle$ and we have

$$\|\langle Ax, y \rangle\|_{\psi_2}^2 = \left\| \sum_{i,j} A_{i,j} x_j y_i \right\|_{\psi_2}^2 \leq CK^2 \sum_{i,j} x_j^2 y_i^2 = CK^2 \|x\|^2 \|y\|^2 = CK^2.$$

So by the tail bound (namely an equivalent definition of subgaussian variables), $\mathbf{P}(\langle Ax, y \rangle \geq u) \leq 2e^{-cu^2/K^2}$ for any $u > 0$. To control all net pairs at once, use the union bound inequality $\mathbf{P}(\cup_k B_k) \leq \sum_k \mathbf{P}(B_k)$, we find

$$\mathbf{P}(\max_{x,y} \langle Ax, y \rangle \geq u) \leq \sum_{x,y} \mathbf{P}(\langle Ax, y \rangle \geq u) \leq 9^{n+m} \cdot 2e^{-cu^2/K^2}.$$

If we pick $u = CK(\sqrt{n} + \sqrt{m} + t)$ with large enough C such that $cu^2/K^2 \geq 3(n+m) + t^2$. Thus

$$\mathbf{P}(\max_{x,y} \langle Ax, y \rangle \geq u) \leq 9^{n+m} 2e^{-3(n+m)-t^2} \leq 2e^{-t^2}.$$

This leads to $\mathbf{P}(\|A\| \geq 2u) \leq 2e^{-t^2}$. □

Remark 2 (Some heuristics about singular value magnitude). *The theorem supplies an upper bound at scale $K(\sqrt{m} + \sqrt{n})$. A matching lower bound needs nondegenerate variances; subgaussianity alone allows the zero matrix. For i.i.d. centered entries with variance $\sigma^2 > 0$, bounded standardized fourth moments and bounded K/σ , the Frobenius calculation below gives the complementary lower bound. Assume $n \geq m$ without loss of generality; together these estimates put the largest singular value at scale $\sigma\sqrt{n}$.*

Assuming $\sigma^2 = \mathbb{E}A_{ij}^2$, we expect $\|A\|_F^2$ to be close to its mean, that is,

$$\frac{\|A\|_F^2}{nm\sigma^2} = \frac{\sum_{i,j} A_{ij}^2}{nm\sigma^2} \xrightarrow{p} 1, \quad \text{as } n \rightarrow \infty$$

Since $\|A\|_F^2 = \sum_{j=1}^m [\sigma_j(A)]^2$, we have $m^{-1} \sum_{k=1}^m \sigma_k^2(A) = (1 + o_p(1))n\sigma^2$. Thus the root-mean-square singular value is $(1 + o_p(1))\sigma\sqrt{n}$. The largest singular value is at least this quantity, while the operator-norm theorem gives the matching upper scale. This calculation concerns the root mean square, not the arithmetic average.

For i.i.d. nondegenerate entries with controlled subgaussian norm relative to their standard deviation, lower singular-value theorems also put the smallest singular value on this scale when the aspect ratio stays away from one. This conclusion needs those extra hypotheses; the upper-bound theorem by itself does not prove it.

3 Statistical application: covariance estimation

High priority Statistical application: covariance estimation

Suppose $X_1, \dots, X_n$ are i.i.d. random vectors with $\mathbb{E}X_i = 0$ and covariance $\Sigma = \text{Cov}(X_i)$. By the law of large numbers, with dimension p being fixed, we have

$$\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \xrightarrow{p} \Sigma \quad \text{as } n \rightarrow \infty.$$

What is the property of the empirical covariance matrix $n^{-1} \sum_{i=1}^n X_i X_i^\top$ without assuming an asymptotic regime? The following theorem gives us an estimation error bound involving p .

Theorem 2. Assume $X_0 \in \mathbb{R}^p$ has mean zero and covariance Σ , and is subgaussian relative to its covariance: $\|\langle X_0, u \rangle\|_{\psi_2} \leq K\sqrt{u^\top \Sigma u}$ for every $u \in \mathbb{R}^p$. Suppose that $X_1, \dots, X_n$ are i.i.d. copies of X_0 and $X = [X_1, \dots, X_n]^\top \in \mathbb{R}^{n \times p}$. Then for every $t \geq 0$, with probability at least $1 - 2e^{-c_K t^2}$,

$$\left\| \frac{1}{n} X^\top X - \Sigma \right\| \leq \max\{\delta, \delta^2\} \|\Sigma\| \quad \text{where } \delta = C_K \sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}. \quad (1)$$

Proof. First take $\Sigma = I_p$. A $1/4$ -net $\mathcal{N}$ of the unit sphere has at most 9^p points. Set $B = X^\top X/n - I_p$. For every fixed unit u , the centered squares $\langle X_i, u \rangle^2 - 1$ have subexponential norms bounded by C_K . Bernstein therefore gives

$$\mathbb{P}(|u^\top B u| > s) \leq 2 \exp[-c_K n \min(s^2, s)].$$

For symmetric B , $\|B\| \leq 2 \max_{u \in \mathcal{N}} |u^\top B u|$; Exercise 5 proves this version of the net argument. A union bound thus gives

$$\mathbb{P}(\|B\| > 2s) \leq 2 \exp[p \log 9 - c_K n \min(s^2, s)].$$

Put $r = (p + t^2)/n$ and $s = C_K(\sqrt{r} + r)$, increasing C_K to absorb $p \log 9$. This gives the claimed form $\max(\delta, \delta^2)$ after adjusting constants. For general positive definite Σ , set $Z_i = \Sigma^{-1/2} X_i$. The stated relative assumption gives the same K for Z_i . Since

$$\hat{\Sigma} - \Sigma = \Sigma^{1/2} (n^{-1} \sum_i Z_i Z_i^\top - I) \Sigma^{1/2},$$

its norm is at most $\|\Sigma\|$ times the whitened error. If Σ is singular, every direction in its kernel has zero projection almost surely; restrict to its range and apply the argument there, using rank at most p . $\square$

Examples of subgaussian random vectors. A random vector is subgaussian iff $\langle X, u \rangle$ is subgaussian for all unit vectors u . Some examples of subgaussian vectors:

1. $X \sim N(0, I_n)$;
2. $X \sim \text{Unif}(\{\pm 1\}^n)$;
3. $X \sim \text{Unif}(\sqrt{n} \mathbb{S}^{n-1})$.

For independent normal coordinates, every unit projection is standard normal. For independent Rademacher coordinates, Hoeffding's inequality gives a uniform subgaussian bound on unit projections. As explained in the previous lecture, the third distribution is not a product measure, but it has Gaussian-like fixed projections, as explained by the coupling if $X \sim \text{Unif}(\sqrt{n} \mathbb{S}^{n-1})$ then

$$X \stackrel{d}{=} \sqrt{n} \frac{G}{\|G\|}, \quad \text{where } G \sim N(0, I_n),$$

and $\|G\|$ concentrates around $\sqrt{n}$.

Relative covariance error. Under the stated relative subgaussian condition, covariance error scales with $\|\Sigma\|_{\text{op}}$ and $\sqrt{p/n} + p/n$. Uniform tail and covariance bounds are needed when converting this rate into a growing-dimension consistency claim.

High priority *Practice: A deterministic covariance obstruction*

Exercise prompt. If $p > n$ and $\Sigma = I_p$, prove $\|\hat{\Sigma} - I_p\|_{\text{op}} \geq 1$.

See Exercise 3 in Section 7 for the full statement and complete solution.

3.1 Discussion: Effects of dimensions

Mid priority *Discussion: Effects of dimensions*

For simplicity, assume the common mean $\mu := \mathbb{E} X_i$ is zero. Denote $X = [X_1, X_2, \dots, X_n]^T \in \mathbb{R}^{n \times p}$, and then empirical covariance is given by $\hat{\Sigma} = \frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$. Classical theory says that if p is fixed, then $\hat{\Sigma} \xrightarrow{p} \Sigma$ as $n \rightarrow \infty$. To understand how dimension p affects estimation consistency, below we treat p as an arbitrary number.

Assume also uniformly bounded coordinate fourth moments, as follows from the uniform subgaussian assumptions above. Recall $\|A\|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$; independence across observations gives

$$\mathbb{E} \left[\left\| \hat{\Sigma} - \Sigma \right\|_F^2 \right] = \frac{1}{n} \sum_{j,k=1}^p \text{Var}((X_i)_j (X_i)_k) \leq \frac{Cp^2}{n}.$$

In particular, this upper bound cannot be improved for simple cases like $X_i \sim \mathcal{N}(0, I_p)$. Since $\left\| \hat{\Sigma} - \Sigma \right\|_F^2 = \sum_{j \leq p} [\sigma_j(\hat{\Sigma} - \Sigma)]^2$ where σ_j denotes the j -th singular value, the above upper bound is equivalent to

$$\mathbb{E} \left\{ \frac{1}{p} \sum_{j=1}^p \left[\sigma_j(\hat{\Sigma} - \Sigma) \right]^2 \right\} \leq \frac{Cp}{n}.$$

The part inside the expectation is the average squared singular value. By Theorem 2, $\|\hat{\Sigma} - \Sigma\| = O_P(\sqrt{p/n} + p/n)$ when K and $\|\Sigma\|$ are uniformly bounded. This comparison suggests the scale of the error; the following Gaussian example establishes inconsistency in probability. Gaussian random-matrix limits below give a concrete sharpness example.

Inconsistency of covariance estimation. As before, in order to study the asymptotic behavior, let us imagine there is a limiting process where we collect more and more data, and the dimension p_n depends on n . We emphasize the dependency on n and write Σ_n and $\hat{\Sigma}_n$.

We have the following conclusions:

- If $p_n/n \rightarrow 0$, with uniformly bounded relative subgaussian constant and $\sup_n \|\Sigma_n\| < \infty$, then $\|\hat{\Sigma}_n - \Sigma_n\| \rightarrow_p 0$. Choose $t_n \rightarrow \infty$ slowly enough that $t_n/\sqrt{n} \rightarrow 0$ in the covariance bound.
- For a concrete inconsistent case, take $X_i \sim N(0, I_{p_n})$ and $p_n/n \rightarrow c > 0$. The Gaussian spectral-edge limit is $\lambda_{\max}(\hat{\Sigma}_n) \rightarrow (1 + \sqrt{c})^2$ almost surely. Hence $\liminf \|\hat{\Sigma}_n - I\| \geq 2\sqrt{c} + c > 0$. This uses the random-matrix limit discussed in Lecture 17. A lower bound on $\mathbb{E} \|\hat{\Sigma} - \Sigma\|^2$ alone would not establish inconsistency in probability.

Moreover, consider a more ‘extreme’ scenario where the dimension is larger than the sample size, then we actually have a *deterministic* result

- If $p > n$, then $\text{rank}(\hat{\Sigma}) \leq n$ (why?), which means that $\hat{\Sigma}$ is singular: its range lies in the span of the n observation vectors. The best rank- n approximation bound (the Eckart–Young theorem) then gives the following *always* holds

$$\|\hat{\Sigma}_n - \Sigma_n\| \geq \min_{M: \text{rank}(M) \leq n} \|M - \Sigma_n\| = \sigma_{n+1}(\Sigma_n).$$

Remark 3. For the second bullet point, we actually have a stronger result for many distributions: there exists a constant $C_0 > 0$ such that $\mathbb{P}(\|\hat{\Sigma}_n - \Sigma\| \geq C_0) \rightarrow 1$.

Remark 4. Usually, $p_n/n \rightarrow 0$ is often considered the “low-dimensional” regime where consistency generally holds under regularity conditions. In particular, $p_n \equiv p$ is the classical “fixed-dimensional” or “large-sample” regime where the LLN and the CLT are valid.

In contrast, $p_n/n \rightarrow c > 0$ is the “high-dimensional” regime where new phenomena exist.

The obstacles posed by high dimensionality can be overcome if we know additional structure. For example, if we know Σ_n is a sparse matrix, meaning that it contains only a few nonzero entries, then we can recover consistency by denoising the empirical covariance matrix [1]. Developing *regularized* methods is very typical of high-dimensional statistical research, as we will see later in this course.

Mid priority Practice: Exact Gaussian Frobenius error

Exercise prompt. For independent $N(0, I_p)$ observations, compute $\mathbb{E} \|\hat{\Sigma} - I_p\|_F^2$.

See Exercise 4 in Section 7 for the full statement and complete solution.

4 Gordon’s theorem*

Low priority Gordon’s theorem

If the distribution is standard normal instead of the general subgaussian assumption, then we have a rather accurate result on the expected extreme singular values. For $A \in \mathbb{R}^{n \times p}$ with $n \geq p$, let $\sigma_{\min}(A) = \inf_{\|u\|_2=1} \|Au\|_2$. This includes zero if A is rank deficient.

Theorem 3 (Gordon’s theorem). Let $A = (A_{ij}) \in \mathbb{R}^{n \times p}$, $n \geq p$, be a random matrix with entries $A_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, then

$$\sqrt{n} - \sqrt{p} \leq \mathbb{E} \sigma_{\min}(A) \leq \mathbb{E} \sigma_{\max}(A) \leq \sqrt{n} + \sqrt{p}.$$

Recall that we use $\|A\|_{\text{op}}$ or $\|A\|$ to denote the operator norm of a matrix A .

Proposition 2. *Under the above assumptions, for any $t > 0$, with probability at least $1 - 2e^{-ct^2}$ where $c > 0$ is a constant,*

$$|\sigma_{\max}(A) - \mathbb{E} \sigma_{\max}(A)| \leq t.$$

Proof of Proposition 2. Viewing $\sigma_{\max}(A)$ as a function F of np Gaussian random variables, i.e.

$$F(A_{11}, A_{12}, \dots, A_{np}) := \sigma_{\max}(A).$$

Let $\tilde{A} = (\tilde{A}_{ij})$ be any matrix, then

$$\begin{aligned} \|A\| &= \max_{u \in \mathbb{S}^{p-1}} \|Au\| \leq \max_{u \in \mathbb{S}^{p-1}} [\|Au - \tilde{A}u\| + \|\tilde{A}u\|] \quad (\text{by triangle inequality}) \\ &\leq \max_{u \in \mathbb{S}^{p-1}} [\|A - \tilde{A}\| + \|\tilde{A}\|] \quad (\text{by definition of operator norm}) \\ &= \|A - \tilde{A}\| + \|\tilde{A}\|. \end{aligned}$$

Similarly, we can prove

$$\|\tilde{A}\| \leq \|A - \tilde{A}\| + \|A\|.$$

Thus, we derive the Lipschitz property of F :

$$|F(\tilde{A}) - F(A)| \leq \|\tilde{A} - A\| \leq \|\tilde{A} - A\|_F.$$

By the Gaussian concentration inequality in the previous lecture, we obtain

$$\|\sigma_{\max}(A) - \mathbb{E} \sigma_{\max}(A)\|_{\psi_2} \leq C.$$

By adjusting the constants, we get the claimed result. $\square$

Following a similar strategy, we can prove a similar bound where $\sigma_{\max}$ is replaced by $\sigma_{\min}$ (see Exercise 7). Combining the two results, we get the following corollary.

Corollary 1. *Under the previous assumptions, for any $t > 0$, with probability at least $1 - 4e^{-ct^2}$,*

$$\sqrt{n} - \sqrt{p} - t \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \sqrt{n} + \sqrt{p} + t.$$

Remark 5. *The quantities $\sqrt{n} - \sqrt{p}$ and $\sqrt{n} + \sqrt{p}$ describe the lower and upper Gaussian spectral edges when the dimensions grow proportionally. For a wide matrix, apply the result to its transpose to describe the nonzero singular values; its null space still makes the minimum stretch over all input directions zero.*

5 Further reading

Further reading *Further reading*

This lecture note is adapted mostly from Chapter 4 of [2]. Section 4 is taken from Section 7.3 of the same book.

Key ideas

- The operator norm is the largest stretch over unit directions.
- An epsilon-net reduces a uniform directional problem to finitely many tests; its proof route is optional.
- Sample-covariance accuracy depends on both dimension and sample size.
- Dimension-free intuition can fail when the dimension grows with the sample size.

6 Detailed reference

High priority *Operator norm*

The operator norm is the largest singular value and controls the largest stretch over all unit input vectors. The Frobenius norm is the square root of the sum of squared singular values. Return to the earlier explanation.

Low priority *Discretize directions*

An epsilon-net approximates every unit direction within distance epsilon. Net cardinality is exponential in dimension, and its logarithm competes with concentration in uniform matrix bounds. Return to the earlier explanation.

High priority *Relative covariance error*

Under the stated relative subgaussian condition, covariance error scales with $\|\Sigma\|_{\text{op}}$ and $\sqrt{p/n} + p/n$. Uniform tail and covariance bounds are needed when converting this rate into a growing-dimension consistency claim. Return to the earlier explanation.

Looking ahead

Lecture 17 uses covariance and operator-norm bounds to study PCA, regression and community detection. The ratio of dimension to sample size becomes part of the statistical problem. Continue to Lecture 17.

7 Exercises and complete solutions

High priority *Compare three matrix norms*

Exercise 1 (Compare three matrix norms).

For $A = \text{diag}(3, 4)$, compute $\|A\|_{\text{op}}$, $\|A\|_F$, and $\|A\|_{\text{max}}$. Explain what each norm measures and which values coincide.

Solution.

The singular values are 4 and 3, so operator norm is 4 and Frobenius norm is $\sqrt{16 + 9} = 5$. The largest absolute entry is 4. The Frobenius norm accumulates all squared entries; the operator norm measures the largest directional stretch.

□

Mid priority *Choose a discretization radius*

Exercise 2 (Choose a discretization radius).

For the two-net bilinear bound, compare $\epsilon = 1/4$ and $\epsilon = 1/8$: give the approximation factors and the covering-number upper bounds in dimensions n and m . Use the approximation factor $(1 - 2\epsilon)^{-1}$ and the covering bound $(1 + 2/\epsilon)^d$ in dimension d .

Solution.

The factors $1/(1 - 2\epsilon)$ are 2 and 4/3. Volumetric bounds give product net sizes at most 9^{n+m} and 17^{n+m} . The finer net has a better deterministic factor but a larger union-bound cost. □

High priority *A deterministic covariance obstruction*

Exercise 3 (A deterministic covariance obstruction).

If $p > n$ and $\Sigma = I_p$, prove $\|\hat{\Sigma} - I_p\|_{\text{op}} \geq 1$.

Solution.

The range of $\hat{\Sigma} = n^{-1} \sum_i X_i X_i^\top$ lies in the span of n sample vectors, so its rank is at most n . Choose a unit vector v in its nontrivial kernel. Then $(\hat{\Sigma} - I)v = -v$, whose norm is 1, and the operator norm is at least 1. □

Mid priority *Exact Gaussian Frobenius error*

Exercise 4 (Exact Gaussian Frobenius error).

For independent $N(0, I_p)$ observations, compute $\mathbb{E} \|\hat{\Sigma} - I_p\|_F^2$.

Solution.

For a diagonal entry the summand X_j^2 has variance 2. For $j \neq k$, $X_j X_k$ has mean zero and variance 1 by independence. Averaging n independent samples divides these variances by n . Summing p diagonal and $p(p-1)$ off-diagonal entries gives $(p^2 + p)/n$. □

High priority *A net for a symmetric quadratic form*

Exercise 5 (A net for a symmetric quadratic form).

If B is symmetric and $\mathcal{N}$ is an ϵ -net of the unit sphere with $\epsilon < 1/2$, prove

$$\|B\| \leq (1 - 2\epsilon)^{-1} \max_{v \in \mathcal{N}} |v^\top B v|.$$

Solution.

Choose a unit vector x with $|x^\top B x| = \|B\|$ and a net point v with $\|x - v\| \leq \epsilon$. Then $x^\top B x - v^\top B v = (x - v)^\top B x + v^\top B(x - v)$ has absolute value at most $2\epsilon\|B\|$. Rearrange to obtain the claim. □

High priority *A volumetric covering argument*

Exercise 6 (A volumetric covering argument).

Prove that an ϵ -net of the Euclidean unit ball in $\mathbb{R}^n$ can be chosen with at most $(1 + 2/\epsilon)^n$ points.

Solution.

Take a maximal ϵ -separated set in the ball. The radius- $\epsilon/2$ balls about its points are disjoint and contained in the radius- $1 + \epsilon/2$ ball. Comparing volumes bounds its cardinality by $(1 + 2/\epsilon)^n$. Maximality ensures

every point in the unit ball lies within epsilon of a selected point. Existence of a finite maximal set follows from the same packing bound. $\square$

Low priority *The smallest singular value is Lipschitz*

Exercise 7 (The smallest singular value is Lipschitz).

For matrices A and B of the same size, define $\sigma_{\min}(A) = \inf_{\|x\|=1} \|Ax\|$. Show $|\sigma_{\min}(A) - \sigma_{\min}(B)| \leq \|A - B\|_F$.

Solution.

For every unit x , $\|Ax\| \geq \|Bx\| - \|A - B\|_{\text{op}}$. Taking infima gives $\sigma_{\min}(A) \geq \sigma_{\min}(B) - \|A - B\|_{\text{op}}$. Exchange A and B , then use operator norm at most Frobenius norm. Gaussian concentration therefore applies to this function. $\square$

Mid priority *A positive expected error need not prevent consistency*

Exercise 8 (A positive expected error need not prevent consistency).

Construct nonnegative $Z_n \rightarrow_p 0$ with $\mathbb{E} Z_n^2 = 1$ for every n .

Solution.

Take $Z_n = \sqrt{n}$ with probability $1/n$ and zero otherwise. For any fixed positive ϵ , $\mathbb{P}(Z_n > \epsilon) = 1/n$ for large n , yet $\mathbb{E} Z_n^2 = n/n = 1$. Rare large errors explain why an expectation lower bound alone does not rule out convergence in probability. $\square$

References

- [1] Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. 2008.
- [2] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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.