

Lecture 16: Norms of random matrices

Lecturer: Yiqiao Zhong

Last update: 13:37:37, 2024-11-20

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

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)

Let $A \in \mathbb{R}^{m \times n}$ be a $m \times n$ -matrix. It has a decomposition $A = \sum_{i=1}^r s_i u_i v_i^\top = U \Sigma V^\top$, where $r = \text{rank}(A)$, $U = [u_1, \dots, u_r] \in \mathbb{R}^{m \times r}$ orthonormal column vectors, $V = [v_1, \dots, v_r] \in \mathbb{R}^{n \times r}$ orthonormal column vectors, 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 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 $\sigma_i(A) = |\lambda_i(A)|$.

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

1.2 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 $\|A\|_F = \sum_{j=1}^r [\sigma_j(A)]^2$.
- Operator norm $\|A\|_{\text{op}} = \sigma_1(A)$. In 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 $\|A\|_{\max} = \max_{ij} |A_{ij}|$. Essentially we treat A as a vector of length mn and consider the usual ℓ^∞ norm.

See also Wiki for a comprehensive comparison.

1.3 Covering and packing number

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 ϵ -net of K if every point x in K is within distance ϵ of some point N , i.e. $\forall x \in K$, $\exists x_0 \in \mathcal{N}$ such that $d(x_0, x) < \epsilon$. 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 $\mathcal{N}(K, d, \epsilon)$ of ϵ -net over K is the number of points in the smallest (i.e. contains least 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 $\mathbb{R}^n$. $(\frac{1}{\epsilon})^n \leq \mathcal{N}(B_2^n, d_{L^2}, \epsilon) \leq (1 + \frac{2}{\epsilon})^n$ for $\forall \epsilon \in [0, 1]$, here d_{L^2} is the L^2 -distance (Euclidean distance).

This result is proved by a volume argument. See Chap. 4.2 of [2] for more details.

Lemma 1. $\forall \epsilon \in (0, 1)$, $\mathcal{N}$ is an ϵ -net of 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 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

$$\|Ax - Ax_0\| \leq \|A\| \cdot \|x - x_0\| \leq \epsilon \|A\|,$$

$$\|Ax_0\| \geq \|Ax\| - \|Ax - Ax_0\| \geq \|A\| - \epsilon \|A\|,$$

$$\|A\| \leq \frac{1}{1 - \epsilon} \|Ax\|.$$

In addition we pick $y \in \mathbb{S}^{n-1}$ such that $\langle Ax, y \rangle = \|Ax\|^2 = \|A\|^2$. 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 $\|A\| = \sup_{\|u\|=2} \|Au\|$ is approximately the same as a discretized version.

2 Operator Norm of Subgaussian Random Matrix

The goal of the section is the following: Show

$$\|A\| \leq C(\sqrt{m} + \sqrt{n}),$$

if A_{ij} is independent subgaussian variable with high probability, where $C > 0$ is an absolute constant as usual.

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 $t > 0$, Then $\|A\| \leq CK(\sqrt{m} + \sqrt{n} + t)$ with probability at least $1 - 2e^{-t^2}$ for 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 $\mathbb{E}X_i = 0$, 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. Pick $\epsilon = \frac{1}{4}$, let $\mathcal{N}, \mathcal{M}$ be ϵ -net of $\mathbb{S}^{n-1}, \mathbb{S}^{m-1}$ respectively such that $|\mathcal{N}| < 9^n, |\mathcal{M}| < 9^m$. (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 production 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 following lemma. Now apply Lemma 2 to $\langle Ax, y \rangle$ and we have

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

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 $u > 0$. Now we unfix x and y using 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). *This theorem shows $\|A\|$ is roughly of order $\sqrt{nm} + \sqrt{n}$ with high probability. Assume $n \geq m$ without loss of generality (we can consider $A^\top$ otherwise), this means that $\sigma_1(A)$ is of order $\sqrt{n}\sigma$.*

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

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

Since $\|A\|_F^2 = \sum_{j=1}^m [\sigma_j(A)]^2$, we have $\frac{1}{m} \sum_{k=1}^m \sigma_k^2(A) \xrightarrow{P} n\sigma^2$. This means an average singular value is of order $\sqrt{n}\sigma$ with high probability. So the largest singular value is the same magnitude as average singular value.

If fact, as long as $n > (1 + \eta)m$ for some constant $\eta > 0$, all singular values of A are on the same magnitude $\sqrt{n}\sigma$.

3 Statistical application: Covariance estimation

Suppose $X_1, \dots, X_n$ are i.i.d. random vectors with $\mathbb{E}X_0 = 0$ and $\text{Cov}(X_0) = \Sigma$ exists. By Law of Large Numbers, with dimension m 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 m .

Theorem 2. Assume that $X_0 \in \mathbb{R}^p$ is a subgaussian random vector, namely $\sup_{u \in \mathbb{S}^{p-1}} \|\langle X_0, u \rangle\|_{\psi_2} \leq K$. Suppose that $X_1, \dots, X_n$ are i.i.d. copies of X_0 and $X = [X_1, \dots, X_n]^T \in \mathbb{R}^{n \times p}$. Then for every $t \geq 0$, with probability at least $1 - 2e^{-c_K t^2}$,

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

Proof. Without loss of generality $\Sigma = I_p$ (consider $\tilde{X}_i = \Sigma^{-1/2} X_i$ otherwise), we take $\frac{1}{4}$ -net $\mathcal{N}$ of $\mathbb{S}^{p-1}$ with $|\mathcal{N}| \leq 9^p$.

For fixed unit vector u , by the assumption $\|\langle X_i, u \rangle\|_{\psi_2} \leq K$. Note that

$$\frac{1}{n} \|Xu\|^2 - 1 = \frac{1}{n} \sum_{i=1}^n (\langle X_i, u \rangle^2 - 1)$$

so $\frac{1}{n} \|Xu\|^2 - 1$ is the average of n subexponential variables. By Bernstein's inequality, we have that for any $u \in \mathcal{N}$ (exercise),

$$\mathbf{P} \left(\left| \frac{1}{n} \|Xu\|^2 - 1 \right| > \frac{1}{2} \max\{\delta, \delta^2\} \right) \leq 2 \exp(-c_K n \delta^2)$$

By the union bound, we have

$$\mathbf{P} \left(\max_{u \in \mathcal{N}} \left| \frac{1}{n} \|Xu\|^2 - 1 \right| \geq \frac{\epsilon}{2} \right) \leq 9^p \cdot \mathbf{P} \left(\left| \frac{1}{n} \|Xu\|^2 - 1 \right| \geq \frac{\epsilon}{2} \right).$$

By Lemma ??,

$$\left\| \frac{1}{n} X X^T - \Sigma \right\| \leq 2 \max_{u \in \mathcal{N}} \left(\frac{1}{n} \langle X, u \rangle^2 - 1 \right)$$

so

$$\mathbf{P} \left(\left\| \frac{1}{n} X X^T - \Sigma \right\| > \epsilon \right) \leq \mathbf{P} \left(2 \max_{u \in \mathcal{N}} \left(\frac{1}{n} \langle X, u \rangle^2 - 1 \right) > \frac{\epsilon}{2} \right) \leq 9^p \cdot \mathbf{P} \left(\left| \frac{1}{n} \|Xu\|^2 - 1 \right| \geq \frac{\epsilon}{2} \right)$$

Choosing $\epsilon = \max\{\delta, \delta^2\}$ with appropriate C_K leads to the upper bound in (1). $\square$

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

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

The first two distributions satisfy the subgaussian definition because they are product measures and Hoeffding's inequality. As explained in the previous lecture, the third distribution is not a product measure, but it is *approximately* a product measure, because 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 $\sqrt{n}$.

3.1 Discussion: Effects of dimensions

For simplicity, we assume $\mu = 0$. 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.

Recall $\|A\|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$, we start with some simple calculations:

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

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

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

The part inside the expectation is the expected average squared singular value. By Theorem 2, $\left\| \hat{\Sigma} - \Sigma \right\| = \sigma_1(\hat{\Sigma} - \Sigma) \leq \sqrt{\frac{Cp}{n}}$ w.h.p. Thus, our bound on $\left\| \hat{\Sigma} - \Sigma \right\|$ cannot be improved in general.

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 $p_n/n \rightarrow 0$, then $\left\| \hat{\Sigma}_n - \Sigma_n \right\| \xrightarrow{p} 0$. (consistency under operator norm)
- if $p_n/n \rightarrow c > 0$, assuming X has jointly independent entries and $\text{Tr}(\Sigma) = p_n$, then for certain $c' > 0$, $\liminf_{n \rightarrow \infty} \mathbb{E} \left[\left\| \hat{\Sigma}_n - \Sigma_n \right\|^2 \right] \geq c'$. As a consequence, we cannot have $\left\| \hat{\Sigma}_n - \Sigma_n \right\| \xrightarrow{p} 0$ (**exercise**), (inconsistent under operator norm)

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. Thus the following *always* holds

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

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 $\mathbb{P}(\left\| \hat{\Sigma}_n - \Sigma_n \right\| \geq C_0) \rightarrow 1$.

Remark 4. Usually, $p_n/n \rightarrow 0$ is often considered “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 “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.

4 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. Let $\sigma_{\min}(A)$ denote the smallest singular value of a matrix A .

Theorem 3 (Gordon's theorem). *Let $A = (A_{ij}) \in \mathbb{R}^{n \times p}$ 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 $\sigma_{\min}$ (exercise). Combining the two results, we get the following corollary.

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

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

Remark 5. $\sqrt{n} \pm \sqrt{p}$ is actually the right edge. We cannot improve much on this lower/upper bounds.

5 Further reading

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

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.