

Notes for High-Dimensional Probability Second Edition by
Roman Vershynin

Gallant Tsao

July 30, 2025

Contents

0	Appetizer: Using Probability to Cover a Set	4
0.1	Covering Geometric Sets	5
1	A Quick Refresher on Analysis and Probability	7
1.1	Convex Sets and Functions	7
1.2	Norms and Inner Products	7
1.3	Random Variables and Random Vectors	7
1.4	Union Bound	8
1.5	Conditioning	9
1.6	Probabilistic Inequalities	9
1.7	Limit Theorems	11
2	Concentration of Sums of Independent Random Variables	13
2.1	Why Concentration Inequalities?	13
2.2	Hoeffding Inequality	14
2.3	Chernoff Inequality	16
2.4	Application: Median-of-means Estimator	18
2.5	Application: Degrees of Random Graphs	20
2.6	Subgaussian Distributions	21
2.6.1	The Subgaussian Norm	23
2.7	Subgaussian Hoeffding and Khintchine Inequalities	23
2.7.1	Subgaussian Hoeffding Inequality	24
2.7.2	Subgaussian Khintchine Inequality	24
2.7.3	Maximum of Subgaussians	25
2.7.4	Centering	26
2.8	Subexponential Distributions	26
2.8.1	Subexponential Properties	26
2.8.2	The Subexponential Norm	28
2.9	Bernstein Inequality	29
3	Random Vectors in High Dimensions	32
3.1	Concentration of the Norm	32
3.2	Covariance Matrices and PCA	33
3.2.1	Learning from the Covariance Matrix	33
3.2.2	Principle Component Analysis	34
3.2.3	Isotropic Distributions	35
3.3	Examples of High-dimensional Distributions	35
3.3.1	Standard Normal	35
3.3.2	General Normal	36
3.3.3	Uniform on the Sphere	37
3.3.4	Uniform on a Convex Set	38
3.3.5	Frames	38
3.4	Subgaussian Distributions in High Dimensions	40
3.4.1	Gaussian, Rademacher, and More	40
3.4.2	Uniform on the Sphere	40
3.4.3	Non-examples	41
3.5	Application: Grothendieck Inequality and Semidefinite Programming	42
3.5.1	Semidefinite Programming	44
3.6	Application: Maximum Cut for Graphs	46
3.6.1	A Simple 0.5-approximation Algorithm	46
3.6.2	Semidefinite Relaxation	47
3.7	Kernel Trick and Tightening of Grothendieck Inequality	48
3.7.1	Tensors	49
3.7.2	Proof of Theorem 3.5.1	51
3.7.3	Kernels and Feature Maps	51

4	Random Matrices	52
4.1	A Quick Refresher on Linear Algebra	52
4.1.1	Singular Value Decomposition	52
4.1.2	Min-max Theorem	53
4.1.3	Frobenius and Operator Norms	54
4.1.4	The Matrix Norms and the Spectrum	54
4.1.5	Low-rank Approximation	55
4.1.6	Perturbation Theory	55
4.1.7	Isometries	57
4.2	Nets, Covering, and Packing	57
4.2.1	Covering Numbers and Volume	59
4.3	Application: Error Correcting Codes	60
4.3.1	Metric Entropy and Complexity	61
4.3.2	Error Correcting Codes	61
4.4	Upper Bounds on Subgaussian Random Matrices	63
4.4.1	Computing the Norm on an ε net	63
4.4.2	The Norms of Subgaussian Random Matrices	63
4.4.3	Symmetric Matrices	65
4.5	Application: Community Detection in Networks	65
4.5.1	Stochastic Block Model	65
4.5.2	The Expected Adjacency Matrix Holds the Key	66
4.5.3	The Actual Adjacency Matrix is a Good Approximation	66
4.5.4	Perturbation Theory	67
4.5.5	Spectral Clustering	67
4.6	Two-sided Bounds on Subgaussian Matrices	68
4.7	Application: Covariance Estimation and Clustering	69
4.7.1	Application: Clustering of Point Sets	71
5	Concentration Without Independence	73
5.1	Concentration of Lipschitz Functions on the Sphere	73
5.1.1	Lipschitz Functions	73
5.1.2	Concentration via Isoperimetric Inequalities	73
5.1.3	Blow-up of Sets on the Sphere	74
5.1.4	Proof of Theorem 5.1.3	75
5.2	Concentration on Other Metric Measure Spaces	76
5.2.1	Gaussian Concentration	76
5.2.2	Hamming Cube	76
5.2.3	Symmetric Group	77
5.2.4	Riemannian Manifolds with Strictly Positive Curvature	77
5.2.5	Special Orthogonal Group	77
5.2.6	Grassmannian	78
5.2.7	Continuous Cube and Euclidean Ball	78
5.2.8	Densities of the Form $e^{-U(x)}$	78
5.2.9	Random Vectors with Independent Bounded Coordinates	79
5.3	Application: Johnson-Lindenstrauss Lemma	79
5.4	Matrix Bernstein Inequality	81
5.4.1	Matrix Calculus	81
5.4.2	Trace Inequalities	83
5.4.3	Proof of Matrix Bernstein Inequality	83
5.4.4	Matrix Hoeffding and Khintchine Inequalities	85
5.5	Application: Community Detection in Sparse Networks	86
5.6	Application: Covariance Estimation for General Distributions	88
5.7	Extra notes	90

6	Quadratic Forms, Symmetrization, and Contraction	85
6.1	Decoupling	85
6.2	Hanson-Wright Inequality	87
6.3	Symmetrization	87
6.4	Random Matrices with non-i.i.d. Entries	88
6.5	Application: Matrix Completion	89
6.6	Contraction Principle	89
7	Random Processes	92
7.1	Basic Concepts and Examples	92
7.1.1	Covariance and Increments	93
7.1.2	Gaussian Processes	93
7.2	Slepian, Sudakov-Fernique, and Gordon Inequalities	94
7.2.1	Gaussian Interpolation	95
7.2.2	Proof of Slepian Inequality	97
7.2.3	Sudakov-Fernique and Gordon Inequalities	98
7.3	Application: Sharp Bounds for Gaussian Matrices	99
7.4	Sudakov Inequality	99
7.4.1	Application for covering numbers in \mathbb{R}^n	100
7.5	Gaussian Width	101
7.5.1	Geometric Meaning of Width	102
7.5.2	Examples	103
7.5.3	Gaussian Complexity and Effective Dimension	104
7.6	Application: Random Projection of Sets	105
8	Chaining	106
8.1	Dudley Inequality	106
8.1.1	Variations and Examples	109
8.2	Application: Empirical Processes	110
8.3	VC Dimension	110
8.3.1	Definition and Examples	110
8.3.2	Pajor's Lemma	112
8.3.3	Sauer-Shelah Lemma	113
8.3.4	Growth Function	113
8.3.5	Covering Numbers via VC Dimension	115
8.3.6	VC Law of Large Numbers	116
8.4	Application: Statistical Learning Theory	119
8.5	Generic Chaining	119
8.5.1	A Makeover of Dudley's Inequality	119
8.5.2	The γ_2 Functional and Generic Chaining	119
8.5.3	Majorizing Measure and Comparison Theorems	121
8.6	Chevet Inequality	123
9	Deviations of Random Matrices on Sets	125
9.1	Matrix Deviation Inequality	125
9.2	Random Matrices, Covariance Estimation, and Johnson-Lindenstrauss	128
9.2.1	Singular Values of Random Matrices	128
9.2.2	Random Projections of Sets	129
9.2.3	Covariance Estimation for Low-dimensional Distributions	129
9.2.4	Johnson-Lindenstrauss Lemma for Infinite Sets	129
9.3	Random Sections: The M^* Bound and Escape Theorem	130
9.3.1	The M^* Bound	130
9.3.2	The Escape Theorem	131
9.4	Application: High-dimensional Linear Models	132
9.5	Application: Exact Sparse Recovery	132
9.6	Deviations of Random Matrices for General Norms	132
9.7	Two-sided Chevet Inequality and Dvoretzky-Milman Theorem	134
9.7.1	Two-sided Chevet's Inequality	134
9.7.2	Dvoretzky-Milman Theorem	135

5 Concentration Without Independence

This chapter mainly explores other approaches to concentration that do not rely on independence.

5.1 Concentration of Lipschitz Functions on the Sphere

For a random vector X in \mathbb{R}^n and a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. When does the random variable $f(X)$ concentrate, i.e.

$$f(X) \approx \mathbb{E}[f(X)] \text{ with high probability?}$$

If X is normal and f is linear, this is easy: $f(X)$ is normal (Corollary 3.3.2) and concentrates well (Proposition 2.1.2).

What about for general *nonlinear* functions f ? We can't expect good concentration for any f , but if f does not oscillate too wildly, we might get good concentration. Namely, we'll use Lipschitz functions to rule out these oscillations:

5.1.1 Lipschitz Functions

Definition 5.1.1. Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \rightarrow Y$ is called Lipschitz if there exists $L \in \mathbb{R}$ such that

$$d_Y(f(u), f(v)) \leq L \cdot d_X(u, v) \text{ for every } u, v \in X.$$

The infimum of all L in this definition is called the Lipschitz norm because of f and is denoted $\|f\|_{\text{Lip}}$.

If $\|f\|_{\text{Lip}} \leq 1$, f is called a contraction.

(Important) Technically the Lipschitz norm is only a seminorm, since it vanishes on nonzero constant functions. It's called a norm in the book for brevity.

The class of Lipschitz functions sits between differentiable and uniformly continuous:

$$f \text{ is differentiable} \implies f \text{ is Lipschitz} \implies f \text{ is uniformly continuous.}$$

Moreover, from Exercise 5.1,

$$\|F\|_{\text{Lip}} \leq \sup_{x \in \mathbb{R}^n} \|\nabla f(x)\|_2.$$

Example 5.1.2. Vectors, matrices, and norms define natural Lipschitz functions:

- (a) For a fixed vector $\theta \in \mathbb{R}^n$, the linear functional

$$f(x) = \langle x, \theta \rangle \text{ has Lipschitz norm } \|f\|_{\text{Lip}} = \|\theta\|_2.$$

- (b) More generally, any $m \times n$ matrix A , the linear operator

$$f(x) = Ax \text{ has Lipschitz norm } \|F\|_{\text{Lip}} = \|A\|.$$

- (c) For any norm $\|\cdot\|$ on \mathbb{R}^n , the function

$$f(x) = \|x\|$$

has Lipschitz norm equal to the smallest L such that

$$\|x\| \leq L\|x\|_2 \text{ for all } x \in \mathbb{R}^n.$$

Proof. Exercise 5.2. □

5.1.2 Concentration via Isoperimetric Inequalities

Any Lipschitz function on the Euclidean sphere $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ concentrates:

Theorem 5.1.3. Let $X \sim \text{Unif}(\sqrt{n}S^{n-1})$. Then for any Lipschitz function $f : \sqrt{n}S^{n-1} \rightarrow \mathbb{R}$ we have

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\text{Lip}}.$$

The theorem above works for the geodesic distance metric as well (Exercise 5.4).

Theorem 5.1.3 has been proved already for linear functions f . Theorem 3.4.5 tells us that X is a subgaussian random vector, and this by definition means that any linear function of X is a subgaussian random variable.

To fully prove Theorem 5.1.3, we need to argue that any Lipschitz function concentrates at least as well as a linear function. We'll use the area of their sublevel sets - regions of the sphere where $f(x) \leq a$ for a given level a . To do this, we'll use the *isoperimetric inequality*, namely the one for subsets on \mathbb{R}^n :

Theorem 5.1.4 (Isoperimetric inequality on \mathbb{R}^n). Among all subsets $A \subset \mathbb{R}^n$ with given volume, the Euclidean balls have minimal area. Moreover, for any $\varepsilon > 0$, the Euclidean balls minimize the volume of the ε -neighborhood of A , defined as

$$A_\varepsilon = \{x \in \mathbb{R}^n : \exists y \in A \text{ such that } \|x - y\|_2 \leq \varepsilon\} = A + \varepsilon B_2^n.$$

The figure below illustrates the isoperimetric inequality:

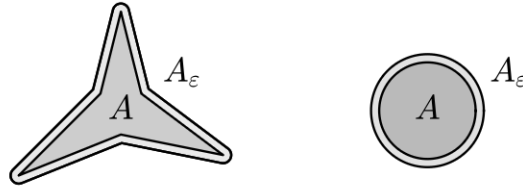


Figure 5.1 The isoperimetric inequality says that among all sets A with a given volume, Euclidean balls minimize the volume of their ε -neighborhood A_ε .

A similar isoperimetric inequality holds for subsets on S^{n-1} , and in this case the minimizers are the spherical caps - neighborhoods of a single point. To state this principle, let σ_{n-1} denote the normalized area on the sphere S^{n-1} (The $n - 1$ -dimensional Lebesgue measure).

Theorem 5.1.5 (Isoperimetric inequality on the sphere). Let $\varepsilon > 0$. Then among all subsets $A \subset S^{n-1}$ with given area $\sigma_{n-1}(A)$, the spherical caps minimize the area of the neighborhood $\sigma_{n-1}(A_\varepsilon)$, where

$$A_\varepsilon := \{x \in \mathbb{R}^n : \exists y \in S^{n-1} \text{ such that } \|x - y\|_2 \leq \varepsilon\}.$$

5.1.3 Blow-up of Sets on the Sphere

The isoperimetric inequality leads to a remarkable and counterintuitive result: if a set A covers at least half of the sphere in area, its ε -neighborhood A_ε will cover most of the sphere. To simplify things in view of Theorem 5.1.3, we'll operate on the sphere with radius \sqrt{n} .

Lemma 5.1.6 (Blow-up). Let $A \subset \sqrt{n}S^{n-1}$, and let σ denote the normalized area on that sphere. If $\sigma(A) \geq 1/2$, then for every $t \geq 0$,

$$\sigma(A_t) \geq 1 - 2 \exp(-ct^2).$$

Proof. Consider the hemisphere defined by the first coordinate:

$$H := \{x \in \sqrt{n}S^{n-1} : x_1 \leq 0\}.$$

By assumption, $\sigma(A) \geq 1/2 = \sigma(H)$, hence the isoperimetric inequality (Theorem 5.1.5) implies that

$$\sigma(A_t) \geq \sigma(H_t).$$

The neighborhood H_t of the hemisphere H is a spherical cap (a portion of a sphere cut off by a plane), and we could compute its area directly, but it is easier to use Theorem 3.4.5 instead, which states that a random vector $X \sim \text{Unif}(\sqrt{n}S^{n-1})$ is subgaussian, and $\|X\|_{\psi_2} \leq C$. Since σ is the uniform probability measure on the sphere, it follows that

$$\sigma(H_t) = P(X \in H_t).$$

Now, the definition of the neighborhood implies that

$$\{x \in \sqrt{n}S^{n-1} : x_1 \leq t/\sqrt{2}\} \subset H_t.$$

Thus

$$\sigma(H_t) \geq P(X_1 \leq t/\sqrt{2}) \geq 1 - 2\exp(-ct^2).$$

The last inequality holds because $\|X_1\|_{\psi_2} \leq \|X\|_{\psi_2} \leq C$. Then the lemma is proved because $\sigma(A_t) \geq \sigma(H_t)$. \square

Remark 5.1.7 (A more dramatic blow-up). The $1/2$ value for the area in Lemma 5.1.6 was arbitrary, and can be replaced with any constant, or even an exponentially small quantity (Exercise 5.3)!

Remark 5.1.8 (A zero-one law). The blow-up phenomenon we just saw can be quite counterintuitive at first. However, this is a typical phenomenon in high dimensions. It is similar to *zero-one laws* in probability theory, which basically say that events influenced by many random variables tend to have probabilities zero or one.

5.1.4 Proof of Theorem 5.1.3

WLOG, we can assume that $\|f\|_{\text{Lip}} = 1$. Let M denote the median of $f(X)$, which by definition satisfies

$$P(f(X) \leq M) \geq \frac{1}{2} \text{ and } P(f(X) \geq M) \geq \frac{1}{2}.$$

Consider the sublevel set

$$A := \{x \in \sqrt{n}S^{n-1} : f(x) \leq M\}.$$

Since $P(X \in A) \geq \frac{1}{2}$, Lemma 5.1.6 implies that

$$P(X \in A_t) \geq 1 - 2\exp(-ct^2).$$

On the other hand, we claim that

$$P(X \in A_t) \leq P(f(X) \leq M + t).$$

Indeed, if $X \in A_t$ then $\|X - y\|_2 \leq t$ for some point $y \in A$. By definition, $f(y) \leq M$. Since f is Lipschitz with $\|f\|_{\text{Lip}} = 1$, it follows that

$$f(X) \leq f(y) + \|X - y\|_2 \leq M + t.$$

Combining the two bounds above, we conclude that

$$P(f(X) \leq M + t) \geq 1 - 2\exp(-ct^2).$$

Repeating the argument for $-f$, we obtain a similar bound for the probability that $f(X) \geq M - t$ (do). Combining the two, we get a similar bound for the probability that $|f(X) - M| \leq t$, and conclude that

$$\|f(X) - M\|_{\psi_2} \leq C.$$

Then we can replace the median by the mean, which follows by centering (Lemma 2.7.8). Therefore the proof is complete. \square

5.2 Concentration on Other Metric Measure Spaces

We can extend concentration from the sphere to other spaces as well. The proof of Theorem 5.1.3 relied on two ingredients:

- (a) an isoperimetric inequality,
- (b) a blow-up of its minimizers.

There are not unique to the sphere - many spaces satisfy them hence we can derive similar concentration results.

Remark 5.2.1. Concentration keeps the mean, median, and L^p norms close. Therefore, we can always replace the mean $\mathbb{E}[f(X)]$ with the median (Exercise 5.6), or, if the mean is nonnegative, with the L^p norm for any $p \geq 1$, though the constant may depend on p (Exercise 5.10).

5.2.1 Gaussian Concentration

The Gaussian measure of a Borel set $A \subset \mathbb{R}^n$ is defined as

$$\gamma_n(A) := P(X \in A) = \frac{1}{(2\pi)^{n/2}} \int_A e^{-\|x\|_2^2/2} dx$$

where $X \sim N(0, I_n)$ is the standard normal random vector in \mathbb{R}^n .

Theorem 5.2.2 (Gaussian isoperimetric inequality). Let $\varepsilon > 0$. Then among all sets $A \subset \mathbb{R}^n$ with given gaussian measure $\gamma_n(A)$, the half-spaces minimize the Gaussian measure of the neighborhood $\gamma_n(A_\varepsilon)$.

Theorem 5.2.3 (Gaussian concentration). Consider a random vector $X \sim N(0, I_n)$ and a Lipschitz function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ (with respect to the Euclidean metric). Then

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\text{Lip}}.$$

Example 5.2.4. Two special cases of Theorem 5.2.3 should already be familiar:

- (a) For linear functions f , it follows since $X \sim N(0, I_n)$ is subgaussian.
- (b) For the Euclidean norm $f(x) = \|x\|_2$, it follows from norm concentration (Theorem 3.1.1).

5.2.2 Hamming Cube

The method based on isoperimetry also works on the Hamming cube $(\{0, 1\}^n, d, \mathbb{P})$ (Definition 4.2.14), where $d(x, y)$ is the normalized Hamming distance:

$$d(x, y) = \frac{1}{n} |\{i : x_i \neq y_i\}|.$$

The measure \mathbb{P} is the uniform probability measure on the cube:

$$\mathbb{P}(A) = \frac{|A|}{2^n} \text{ for any } A \subset \{0, 1\}^n.$$

Theorem 5.2.5 (Concentration on the Hamming cube). Consider a random vector $X \sim \{0, 1\}^n$. Then for any function $f : \{0, 1\}^n \rightarrow \mathbb{R}$ we have

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

5.2.3 Symmetric Group

A similar result holds for the symmetric group S_n , a set of all $n!$ permutations of $\{1, \dots, n\}$. We can view the symmetric group as a metric measure space (S_n, d, \mathbb{P}) , where $d(\pi, \rho)$ is the normalized Hamming distance - the fraction of the symbols on which permutations π and ρ differ:

$$d(\pi, \rho) = \frac{1}{n} |\{i : \pi(i) \neq \rho(i)\}|.$$

The measure \mathbb{P} is the uniform probability measure on S_n :

$$\mathbb{P}(A) = \frac{|A|}{n!} \text{ for any } A \subset S_n.$$

Theorem 5.2.6 (Concentration on the symmetric group). Consider a random permutation $X \sim \text{Unif}(S_n)$ and a function $f : S_n \rightarrow \mathbb{R}$. Then

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

5.2.4 Riemannian Manifolds with Strictly Positive Curvature

(Feel free to skip this if not familiar with differential geometry)

A compact connected Riemannian manifold (M, g) comes with the geodesic distance $d(x, y)$, which is the shortest length of a curve connecting the points. Then we can define a metric measure space (M, d, \mathbb{P}) where \mathbb{P} is the uniform probability measure derived by normalizing the Riemannian volume.

Let $c(M)$ denote the infimum of the Ricci curvature tensor over all tangent vectors. Assuming $c(M) > 0$, then it can be proved that

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{c(M)}}$$

for any Lipschitz function $f : M \rightarrow \mathbb{R}$.

To give an example, $c(S^{n-1}) = n - 1$. Then the above gives another approach for the concentration inequality of the sphere.

5.2.5 Special Orthogonal Group

The special orthogonal group $\text{SO}(n)$ consists of all $n \times n$ orthogonal matrices with determinant 1. We can treat it as a metric measure space $(\text{SO}(n), \|\cdot\|_F, \mathbb{P})$, with distance given by the Frobenius norm $\|A - B\|_F$ and \mathbb{P} as the uniform probability measure.

Theorem 5.2.7 (Concentration on the special orthogonal group). Consider a random orthogonal matrix $X \sim \text{Unif}(\text{SO}(n))$ and a function $f : \text{SO}(n) \rightarrow \mathbb{R}$. Then

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

The result above can be deduced from the concentration on general Riemannian manifolds.

Remark 5.2.8 (Haar measure). To generate a random orthogonal matrix $X \sim \text{Unif}(\text{SO}(n))$, one way is to start with an $n \times n$ Gaussian random matrix G with $N(0, 1)$ independent entries, and compute its SVD $G = U\Omega V^T$. Then the matrix of left singular vectors is uniformly distributed in $\text{SO}(n)$.

The uniform probability distribution on $\text{SO}(n)$ is given by

$$\mu(A) := P(X \in A) \text{ for } A \subset \text{SO}(n).$$

This is the unique rotation-invariant probability measure on $\text{SO}(n)$, called the Haar measure.

5.2.6 Grassmannian

The Grassmannian manifold $G_{n,m}$ consists of all m -dimensional subspaces of \mathbb{R}^n . When $m = 1$, it can be identified with the sphere S^{n-1} . Therefore the concentration on the Grassmannian includes the concentration on the sphere.

We can treat $G_{n,m}$ as a metric space $(G_{n,m}, d, \mathbb{P})$, where the distance between subspaces is given by the operator norm

$$d(E, F) = \|P_E - P_F\|$$

where P_E and P_F are the orthogonal projections onto the subspaces. The probability measure is the Haar measure (Remark 5.2.8). A random subspace E can hence be computed by computing the image of the random $n \times m$ Gaussian random matrix with i.i.d. $N(0, 1)$ entries.

Theorem 5.2.9 (Concentration on the Grassmannian). Consider a random subspace $X \sim \text{Unif}(G_{n,m})$ and a function $f : G_{n,m} \rightarrow \mathbb{R}$. Then

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

Proof. The proof is a bit involved: Express the Grassmannian as the quotient via the special orthogonal group:

$$G_{n,m} = \text{SO}(n)/(\text{SO}(m) \times \text{SO}(n-m))$$

and use the fact that concentration carries over to quotients. \square

5.2.7 Continuous Cube and Euclidean Ball

Theorem 5.2.10 (Concentration on the continuous cube and ball). Let T be either the cube $[0, 1]^n$ or the ball $\sqrt{n}B_2^n$. Consider a random vector $X \sim \text{Unif}(T)$ and a Lipschitz function $f; T \rightarrow \mathbb{R}$, where the Lipschitz norm is with respect to the Euclidean distance. Then

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\text{Lip}}.$$

Proof. Exercises 5.12 & 5.13. \square

5.2.8 Densities of the Form $e^{-U(x)}$

The push forward method from the previous section can be applied to many other distributions in \mathbb{R}^n . For example, suppose a random vector X has a density of the form

$$f(x) = e^{-U(x)}$$

for some function $U : \mathbb{R}^n \rightarrow \mathbb{R}$. For example, $X \sim N(0, I_n)$, the normal density gives $U(x) = \|x\|_2^2 + c$ where c is constant (dependent on n but not on x), and Gaussian concentration holds for X .

In general, we would expect that if U has curvature at least like $\|x\|_2^2$, then there would be at least Gaussian concentration. As the theorem below shows, this depends on the Hessian of U :

Theorem 5.2.11. Consider a random vector X in \mathbb{R}^n whose density has the form $e^{-U(x)}$ for some function $U : \mathbb{R}^n \rightarrow \mathbb{R}$. Assume there exists $\kappa > 0$ such that

$$\nabla^2 U(x) \succcurlyeq \kappa I_n \text{ for all } x \in \mathbb{R}^n.$$

Then any Lipschitz function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfies

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{\kappa}}.$$

Proof. The proof uses semigroup methods, which are not covered in the text. \square

5.2.9 Random Vectors with Independent Bounded Coordinates

There is a remarkable partial generalization of Theorem 5.2.10 for random vectors X with independent coordinates that have arbitrary bounded distributions (not just uniform). By scaling, we can assume WLOG that $|X_i| \leq 1$.

Theorem 5.2.12 (Talagrand concentration inequality). Consider a random vector in \mathbb{R}^n , $X = (X_1, \dots, X_n)$ whose coordinates are independent and satisfy $|X_i| \leq 1$ almost surely. Then for any Lipschitz function $f : [-1, 1]^n \rightarrow \mathbb{R}$,

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\text{Lip}}.$$

5.3 Application: Johnson-Lindenstrauss Lemma

Suppose we have N data points in \mathbb{R}^n where the dimension n is very large. Can we reduce the dimension without losing the geometry of the data? The simplest way is to project onto a low-dimensional subspace

$$E \subset \mathbb{R}^n, \dim(E) := m \ll n,$$

see Figure 5.2 below. How shall we choose the subspace E , and how small should its dimension m be?

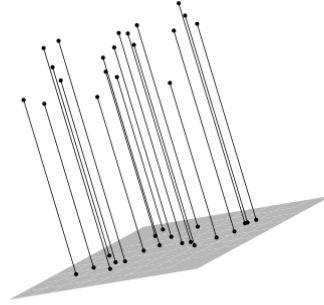


Figure 5.2 Johnson-Lindenstrauss Lemma reduces dimension of the data by random projection onto a low-dimensional subspace.

The Johnson-Lindenstrauss Lemma states that the geometry of data is well preserved if we choose E to be a *random subspace* of dimension

$$m \asymp \log N.$$

Here we say that E is a random m -dimensional subspace in \mathbb{R}^n uniformly distributed in $G_{n,m}$, i.e.

$$E \sim \text{Unif}(G_{n,m}),$$

if E is a random m -dimensional subspace of \mathbb{R}^n whose distribution is rotation invariant, i.e.

$$P(E \in \mathcal{E}) = P(U(E) \in \mathcal{E})$$

for any fixed subset $\mathcal{E} \in G_{n,m}$ and $n \times n$ orthogonal matrix U .

Theorem 5.3.1 (Johnson-Lindenstrauss Lemma). Let \mathcal{X} be a set of N points in \mathbb{R}^n and $\varepsilon > 0$. Assume that

$$m \geq C\varepsilon^{-2} \log N.$$

Let P be the orthogonal projection in \mathbb{R}^n onto a random m -dimensional subspace $E \sim \text{Unif}(G_{n,m})$. Then, with probability at least $1 - 2\exp(-c\varepsilon^2 m)$, the scaled projection $Q = \sqrt{n/m}P$ is an approximate isometry on \mathcal{X} :

$$(1 - \varepsilon)\|x - y\|_2 \leq \|Qx - Qy\|_2 \leq (1 + \varepsilon)\|x - y\|_2 \text{ for all } x, y \in \mathcal{X}.$$

The proof will be based on concentration of Lipschitz functions on the sphere in Section 5.1. We use it to examine how the random projection P acts on the fixed vector $x - y$, then take the union bound over all N^2 differences $x - y$.

Lemma 5.3.2 (Random Projection). Let P be a projection in \mathbb{R}^n onto a random m -dimensional subspace $E \sim \text{Unif}(G_{n,m})$. Fix any $z \in \mathbb{R}^n$ and $\varepsilon > 0$. Then,

(a) $(\mathbb{E} [\|Pz\|_2^2])^{1/2} = \sqrt{\frac{m}{n}} \|z\|_2$.

(b) With probability at least $1 - 2 \exp(-c\varepsilon^2 m)$, we have

$$(1 - \varepsilon) \sqrt{\frac{m}{n}} \|z\|_2 \leq \|Pz\|_2 \leq (1 + \varepsilon) \sqrt{\frac{m}{n}} \|z\|_2.$$

Proof. Without loss of generality, we may assume that $\|z\|_2 = 1$. Now switch the view. A random m -dimensional subspace E can be obtained by randomly rotating some fixed subspace, such as the coordinate subspace \mathbb{R}^m . But instead of fixing z and randomly rotate \mathbb{R}^m , we can fix the subspace $E = \mathbb{R}^m$ and randomly rotate z , which makes z uniformly distributed on the sphere:

$$z \sim \text{Unif}(S^{n-1}).$$

By rotation invariance, Pz has the same distribution!

(a) Since P is the projection onto the first m coordinates in \mathbb{R}^n ,

$$\mathbb{E} [\|Pz\|_2^2] = \mathbb{E} \left[\sum_{i=1}^m z_i^2 \right] = \sum_{i=1}^m \mathbb{E} [z_i^2] = m \mathbb{E} [z_1^2],$$

because the coordinates z_i of the random vector $z \sim \text{Unif}(S^{n-1})$ are identically distributed. To compute $\mathbb{E} [z_1^2]$, note that $\sum_{i=1}^n z_i^2 = 1$. Taking expectations on both sides, we obtain

$$\sum_{i=1}^n \mathbb{E} [z_i^2] = 1 \implies \mathbb{E} [z_1^2] = \frac{1}{n}.$$

Then, putting this into the equation above, we have

$$\mathbb{E} [\|Pz\|_2^2] = \frac{m}{n}.$$

(b) $x \mapsto \|Px\|_2$ is a Lipschitz function on S^{n-1} with Lipschitz norm bounded by 1. Then from Exercise 5.5, the concentration inequality gives

$$P \left(\left| \|Px\|_2 - \sqrt{\frac{m}{n}} \right| \geq t \right) \leq 2 \exp(-cnt^2).$$

(We replaced $\mathbb{E} [\|x\|_2]$ by $(\mathbb{E} [\|x\|_2^2])^{1/2}$ in the concentration inequality using Remark 5.2.1). Choosing $t := \varepsilon \sqrt{m/n}$, we complete the proof. \square

Proof of Johnson-Lindenstrauss Lemma. Consider the difference set

$$\mathcal{X} - \mathcal{X} := \{x - y : x, y \in \mathcal{X}\}.$$

We would like to show that, with required probability, the inequality

$$(1 - \varepsilon) \|z\|_2 \leq \|Qz\|_2 \leq (1 + \varepsilon) \|z\|_2$$

holds for all $z \in \mathcal{X} - \mathcal{X}$. Since $Q = \sqrt{n/m}P$, this inequality is equivalent to

$$(1 - \varepsilon) \sqrt{\frac{m}{n}} \|z\|_2 \leq \|Pz\|_2 \leq (1 + \varepsilon) \sqrt{\frac{m}{n}} \|z\|_2.$$

For any fixed z , Lemma 5.3.2 states that the above holds with probability at least $1 - 2 \exp(-c\varepsilon^2 m)$. It remains to take a union bound over $z \in \mathcal{X} - \mathcal{X}$. It follows that the bound above holds simultaneously for all $z \in \mathcal{X} - \mathcal{X}$, with probability at least

$$1 - |\mathcal{X} - \mathcal{X}| \cdot 2 \exp(-c\varepsilon^2 m) \geq 1 - N^2 \cdot 2 \exp(-c\varepsilon^2 m).$$

If $m \geq C\varepsilon^{-2} \log N$ then this probability is at least $1 - 2 \exp(-c\varepsilon^2 m/2)$, as claimed. Hence the proof is done. \square

Remark 5.3.3 (Non-adaptive, dimension-free). A remarkable feature of the JL lemma is that the dimension reduction map A is *non-adaptive*, meaning it does not depend on the data. Note also that the ambient dimension n of the data plays no role. With more tool, we will develop more advanced versions of the JH lemma (Exercise 9.37-9.39).

Remark 5.3.4 (Optimality). The JL lemma makes such a striking dimension reduction from N to $n = O(\log N)$. Can we go even smaller, say $n = o(\log N)$? Exercise 5.15 shows that we can't - the log dimension is the best we can do, even with nonlinear maps.

5.4 Matrix Bernstein Inequality

We extend generalized concentration inequalities from sums of independent random variables to sums of independent random matrices. We'll make a matrix version of Bernstein inequality (Theorem 2.9.5) by replacing random variables by random matrices, and absolute value by the operator norm. No need for independence of entries, rows, or columns within each random matrix!

Theorem 5.4.1 (Matrix Bernstein inequality). Let X_1, \dots, X_N be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all i . Then for every $t \geq 0$,

$$P\left(\left\|\sum_{i=1}^N X_i\right\| \geq t\right) \leq 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right)$$

where $\sigma^2 = \|\sum_{i=1}^N \mathbb{E}[X_i^2]\|$ is the operator norm of the matrix variance of the sum.

We can rewrite the RHS of the inequality as the mixture of subgaussian and subexponential tail, like in the scalar Bernstein inequality:

$$P\left(\left\|\sum_{i=1}^N X_i\right\| \geq t\right) \leq 2n \exp\left[-c \cdot \min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right].$$

The proof is similar to that of the scalar version: Repeat the MGF argument, swapping scalars with matrices. However, there is a big problem: Matrix multiplication is not commutative! Therefore we need some matrix calculus knowledge first.

5.4.1 Matrix Calculus

For an $n \times n$ symmetric matrix X , operations such as inversion or squaring only affect eigenvalues. For example, if the spectral decomposition of X is $X = \sum_{i=1}^n \lambda_i u_i u_i^T$, then

$$X^{-1} = \sum_{i=1}^n \frac{1}{\lambda_i} u_i u_i^T, \quad X^2 = \sum_{i=1}^n \lambda_i^2 u_i u_i^T, \quad 2I_n - 5X^3 = \sum_{i=1}^n (2 - 5\lambda_i^3) u_i u_i^T.$$

This suggests that for symmetric matrices, applying arbitrary functions on the matrices is equivalent to applying them to the eigenvalues:

Definition 5.4.2 (Functions of matrices). For a function $f : \mathbb{R} \rightarrow \mathbb{R}$ and an $n \times n$ symmetric matrix X with spectral decomposition as above, define

$$f(X) := \sum_{i=1}^n f(\lambda_i) u_i u_i^T.$$

This definition agrees with matrix addition and multiplication, and with Taylor series (Exercise 5.16). Of course, matrices can be compared with each other via a partial ordering:

Definition 5.4.3 (Loewner order). We write $X \succcurlyeq 0$ if X is a symmetric positive semidefinite matrix. We write $X \succeq Y$ and $Y \preceq X$ if $X - Y \succeq 0$.

This is a partial ordering because there are matrices for which neither $X \succeq Y$ nor $Y \succeq X$ holds.

Proposition 5.4.4 (Simple properties of Loewner order). We have

- (a) (Eigenvalue monotonicity) $X \preceq Y$ implies $\lambda_i(X) \leq \lambda_i(Y)$ for all i .
- (b) (Trace monotonicity) For a (weakly) increasing function $f : \mathbb{R} \rightarrow \mathbb{R}$,

$$X \preceq Y \implies \text{tr}(f(X)) \leq \text{tr}(f(Y)).$$

- (c) (Operator norm) For any $a \geq 0$,

$$\|X\| \leq a \iff -aI_n \preceq X \preceq aI_n.$$

- (d) (Upgrading scalar to matrix inequalities) For functions $f, g : \mathbb{R} \rightarrow \mathbb{R}$,

$$f(x) \leq g(x) \forall x \text{ with } |x| \leq a \implies f(X) \preceq g(X) \forall X \text{ with } \|X\| \leq a.$$

Proof. (a) If $X \preceq Y$, then $Y - X \succeq 0$ hence all eigenvalues of $Y - X$ are greater than equal to 0, and the result follows.

(b) The eigenvalues of $f(X)$ are $f(\lambda_i(X))$. The same can be said for $f(Y)$. By part (a) and the assumption, $f(\lambda_i(X)) \leq f(\lambda_i(Y))$. Summing these gives the result since the trace is the sum of the eigenvalues.

(c) From Remark 4.1.12, $\|X\| \leq a$ implies $u^T X u \leq a$ for all unit vectors u . Therefore $u^T (aI_n - X) u \geq 0$ for all u , meaning $aI_n - X \succeq 0$, thus $X \preceq aI_n$. For the other inequality, again from Remark 4.1.12, $u^T X u \geq -a$ for all unit vectors u . Following the exact procedure above gives $X \succeq -aI_n$.

(d) By considering $g - f$, we can assume that $f = 0$. If $\|X\| \leq a$, then all eigenvalues of X satisfy $|\lambda_i| \leq a$, which implies $g(\lambda_i) \geq 0$ by assumption. So, by definition, $g(X)$ has nonnegative eigenvalues $g(\lambda_i)$ and so $g(X) \succeq 0$. \square

Remark 5.4.5 (Operator norm as matrix absolute value). (c) of Proposition 5.4.4 looks quite familiar... It is a matrix version of the basic fact about absolute values: for $x \in \mathbb{R}$,

$$|x| \leq a \iff -a \leq x \leq a.$$

This makes the operator norm $\|\cdot\|$ a natural matrix version of the absolute value $|\cdot|$, and that's why it appears in the matrix Bernstein inequality (Theorem 5.4.1).

Remark 5.4.6 (Matrix monotonicity). Can we strengthen trace monotonicity (Proposition 5.4.4 (b)) to matrix monotonicity, i.e.

$$X \preceq Y \implies f(X) \preceq f(Y) \text{ for any weakly increasing } f : \mathbb{R} \rightarrow \mathbb{R}?$$

If X and Y commute, yes - but in general, no (Exercise 5.17).

However, some functions, like $1/x$ and $\log x$ on $[0, \infty)$, are matrix monotone, meaning that the above holds even for non-commuting matrices:

$$0 \preceq X \preceq Y \implies X^{-1} \succeq Y^{-1} \succeq 0 \text{ and } \log X \preceq \log Y$$

whenever X is invertible (Exercise 5.18).

5.4.2 Trace Inequalities

Here is another identity that works for real numbers but not for matrices in general: $e^{x+y} = e^x e^y$ for scalars, but in Exercise 5.19, there are $n \times n$ symmetric matrices X, Y such that

$$e^{X+Y} \neq e^X e^Y.$$

This is unfortunate, because when using the exponential moment method, we relied on this property to split the MGF via independence.

Nevertheless, there are useful substitutes for the missing identity. In particular, this subsection covers two of them, both belonging to the rich family of *trace inequalities*.

Theorem 5.4.7 (Golden-Thompson inequality). For any $n \times n$ symmetric matrices A and B ,

$$\text{tr}(e^{A+B}) \leq \text{tr}(e^A e^B).$$

Note that this does not hold for three or more matrices (we can find counterexamples)!

Theorem 5.4.8 (Lieb inequality). Let H be an $n \times n$ symmetric matrix. Define the function on matrices

$$f(X) := \text{tr}(\exp(H + \log X)).$$

Then f is concave on the space of PSD $n \times n$ symmetric matrices.

If X is a random matrix, then Lieb and Jensen inequalities imply that

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X]).$$

Applying this with $X = e^Z$, we obtain the following:

Lemma 5.4.9 (Lieb inequality for random matrices). Let H be a fixed $n \times n$ symmetric matrix and Z be a random $n \times n$ symmetric matrix. Then

$$\mathbb{E}[\text{tr}(\exp(H + Z))] \leq \text{tr}(\exp(H + \log \mathbb{E}[e^Z])).$$

5.4.3 Proof of Matrix Bernstein Inequality

(Step 1: Reduction of MGF) To bound the norm of the sum

$$S := \sum_{i=1}^N X_i,$$

we need to control the largest and smallest eigenvalues of S . Consider the largest eigenvalue

$$\lambda_{\max}(S) := \max_i \lambda_i(S)$$

and note that

$$\|S\| = \max_i |\lambda_i(S)| = \max(\lambda_{\max}(S), \lambda_{\max}(-S)).$$

To bound $\lambda_{\max}(S)$, we'll use the exponential moment method again. Fix $\lambda > 0$. Via the typical procedure,

$$P(\lambda_{\max}(S) \geq t) = P(e^{\lambda \cdot \lambda_{\max}} \geq e^{\lambda t}) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda \cdot \lambda_{\max}}].$$

Then by Definition 5.4.2, the eigenvalues of $e^{\lambda S}$ are $e^{\lambda \cdot \lambda_i(S)}$ so

$$E := \mathbb{E}[e^{\lambda \cdot \lambda_{\max}(S)}] = \mathbb{E}[\lambda_{\max}(e^{\lambda S})].$$

Since the eigenvalues of $e^{\lambda S}$ are all positive, the maximal eigenvalue is bounded by the sum of all eigenvalues, which is the trace. Therefore

$$E \leq \mathbb{E}[\text{tr}(e^{\lambda S})].$$

(Step 2: Application of Lieb Inequality) To use the Lieb inequality (Lemma 5.4.9), we separate the last term from the sum S :

$$E \leq \mathbb{E} \left[\text{tr} \left(\exp \left(\sum_{i=1}^{N-1} \lambda X_i + \lambda X_N \right) \right) \right].$$

Condition on $(X_i)_{i=1}^{N-1}$ and apply Lemma 5.4.9 for the fixed matrix $H := \sum_{i=1}^{N-1} \lambda X_i$ and the random matrix $Z := \lambda X_N$. We get

$$E \leq \mathbb{E}[\text{tr}(\exp \left(\sum_{i=1}^{N-1} \lambda X_i + \log \mathbb{E}[e^{\lambda X_N}] \right))].$$

Then we continue the same procedure above: separate λX_{N-1} and apply Lemma 5.4.9, and do the same thing for N times to get

$$E \leq \text{tr} \left(\exp \left[\sum_{i=1}^N \log \mathbb{E}[e^{\lambda X_i}] \right] \right).$$

(Step 3: MGF of the individual terms) We'll bound the matrix-values MGF via the following lemma:

Lemma 5.4.10 (Matrix MGF). Let X be an $n \times n$ symmetric mean zero random matrix such that $\|X\| \leq K$ almost surely. Then

$$\mathbb{E}[\exp(\lambda X)] \preceq \exp(g(\lambda)\mathbb{E}[X^2]) \text{ where } g(\lambda) = \frac{\lambda^2/2}{1 - |\lambda|K/3}$$

provided that $|\lambda| < 3/K$.

Proof. First, we can bound the (scalar) exponential function by the first few terms via Taylor expansion:

$$e^z \leq 1 + z + \frac{1}{1 - |z|/3} \cdot \frac{z^2}{2}, \quad |z| < 3.$$

(To get this inequality, write $e^Z = 1 + z + z^2 \sum_{p=2}^{\infty} z^{p-2}/p!$ and use the bound $p! \geq 2 \cdot 3^{p-2}$). Next, apply this inequality to $z = \lambda x$. If $|x| \leq K$ and $|\lambda| < 3/K$ then we obtain

$$e^{\lambda x} \leq 1 + \lambda x + g(\lambda)x^2,$$

where $g(\lambda)$ is the same as how we defined in the statement.

Then we can upgrade this to a matrix inequality using Proposition 5.4.4 (d). If $\|X\| \leq K$ and $|\lambda| < 3/K$, then

$$e^{\lambda X} \preceq I + \lambda X + g(\lambda)X^2.$$

Taking expectations on both sides, since $\mathbb{E}[X] = 0$,

$$\mathbb{E}[e^{\lambda X}] \preceq I + g(\lambda)\mathbb{E}[X^2].$$

To complete the proof of the lemma, let's use the inequality $1 + z \leq e^z$. We can transform this into a matrix inequality via Proposition 5.4.4 (d) and get $I + Z \preceq e^Z$ holds for all matrices Z , and in particular for $Z = g(\lambda)\mathbb{E}[X^2]$. \square

(Step 4: Completion of the proof) Using Lemma 5.4.10, we obtain

$$E \leq \text{tr} \left(\exp \left(\sum_{i=1}^N \log \mathbb{E}[e^{\lambda X_i}] \right) \right) \leq \text{tr}(\exp(g(\lambda)Z)), \text{ where } Z := \sum_{i=1}^N \mathbb{E}[X_i^2].$$

where we used matrix monotonicity of $\ln x$ (Remark 5.4.6) to take logarithms on both sides, summed up the results, and used trace monotonicity (Proposition 5.4.4 (b)) to take traces of the exponential of both sides.

Since the trace of $\exp(g(\lambda)Z)$ is a sum of n positive eigenvalues, it is bounded by n times the maximum eigenvalue, hence

$$\begin{aligned} E &\leq n\lambda_{\max}(\exp(g(\lambda)Z)) \\ &= m \exp(g(\lambda)\lambda_{\max}(Z)) \\ &= n \exp(g(\lambda)\|Z\|) \quad (\text{Since } Z \succeq 0) \\ &= n \exp(g(\lambda)\sigma^2) \quad (\text{By definition of } \sigma). \end{aligned}$$

Plugging in this bound for $E = \mathbb{E}[e^{\lambda \cdot \lambda_{\max}(S)}]$ into the original equation gives

$$P(\lambda_{\max}(S) \geq t) \leq n \exp(-\lambda t + g(\lambda)\sigma^2).$$

The above is a bound that holds for any $\lambda > 0$ as long as $|\lambda| < 3/K$, so we can minimize it in λ . Better yet, instead of computing the exact minimum (which can be quite ugly), we can choose the following value: $\lambda = t/(\sigma^2 + Kt/3)$, and substituting this value back gives

$$P(\lambda_{\max}(S) \geq t) \leq n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Repeating the argument for $-S$, we will get the same bound as the above, and summing up the two bounds completes the proof. \square

Remark 5.4.11 (Matrix Bernstein Inequality: expectation). Matrix Bernstein inequality gives a high-probability bound. It can be turned into a simpler (but less informative) expectation bound in a standard way. Using Theorem 5.4.1 and the integrated tail formula (Lemma 1.6.1), we can deduce that (Exercise 5.20)

$$\mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \lesssim \left\|\sum_{i=1}^N \mathbb{E}[X_i^2]\right\|^{1/2} \sqrt{\log(2n)} + K \log(2n)$$

where the \lesssim symbol hides an absolute constant factor. In the scalar case ($n = 1$), an expectation bound is trivial: the variance of sum formula gives

$$\mathbb{E}\left[\left|\sum_{i=1}^N X_i\right|\right] \leq \left(\mathbb{E}\left[\left|\sum_{i=1}^N X_i\right|^2\right]\right)^{1/2} = \left(\sum_{i=1}^N \mathbb{E}[X_i^2]\right)^{1/2}.$$

Remark 5.4.12 (The logarithmic price). For the equation in Remark 5.4.11, the high-dimensional version differs the 1-dimensional one by just a logarithmic factor. This is a surprisingly small price for high dimensions! Moreover, this price is in essentially optimal - Exercise 5.28 gives an example of why we can't get rid of it.

5.4.4 Matrix Hoeffding and Khintchine Inequalities

Theorem 5.4.13 (Matrix Hoeffding inequality). Let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables and A_1, \dots, A_N be any (fixed) symmetric $n \times n$ matrices. Then for any $t > 0$,

$$P\left(\left\|\sum_{i=1}^N \varepsilon_i A_i\right\| \geq t\right) \leq 2n \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

where $\sigma^2 = \left\|\sum_{i=1}^N A_i^2\right\|$.

Proof. Exercise 5.21. \square

Theorem 5.4.14 (Matrix Khintchine inequality). Let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables and A_1, \dots, A_N be any (fixed) symmetric $n \times n$ matrices. Then for every $p \in [1, \infty)$, we have

$$\left(\mathbb{E} \left[\left\| \sum_{i=1}^N \varepsilon_i A_i \right\|^p \right] \right)^{1/p} \leq C \sqrt{p + \log n} \left\| \sum_{i=1}^N A_i^2 \right\|^{1/2}.$$

Proof. Exercise 5.22. Use the matrix Hoeffding inequality. \square

Remark 5.4.15 (Non-symmetric, rectangular matrices). Matrix concentration inequalities easily extend to rectangular matrices using the *Hermitian dilation* introduced in Exercise 4.14. Replace each matrix X_i with the symmetric block matrix

$$\begin{bmatrix} 0 & X_i \\ X_i^T & 0 \end{bmatrix}$$

and apply usual matrix concentration. We can get the matrix Bernstein (Exercise 5.23) and Khintchine (Exercise 5.24) inequalities for rectangular matrices this way.

5.5 Application: Community Detection in Sparse Networks

In section 4.5, the method of *spectral clustering* was introduced, which is a basic method for community detection in networks. We showed that it works for relatively dense networks, where the expected average degree is $\gtrsim \sqrt{n}$. Now, using the matrix Bernstein inequality, we will show that spectral clustering actually works for much sparser networks, with an expected average degree as low as $O(\log n)$.

Theorem 5.5.1 (Spectral clustering for sparse stochastic block model). Let $G \sim G(n, p, q)$ where $p = a/n, q = b/n$ and $b < a < 3b$. Assume that

$$(a - b)^2 \geq Ca \log n.$$

Then, with probability at least 0.99, the spectral clustering algorithm identifies the communities of G with 99% accuracy, i.e. misclassifying at most $0.01n$ vertices.

Proof. We'll follow that same proof as that from Section 4.5, but with a sharper error bound.

Step 1: Decomposition. Again, let's split A into the deterministic and random parts:

$$A = D + R \text{ where } D = \mathbb{E}[A].$$

Before, the analysis is mostly on the deterministic matrix D , where the second largest eigenvector has ± 1 coefficients representing community membership. Now we have to analyze the random part

$$R = A - \mathbb{E}[A].$$

Let's decompose it entry by entry, keeping symmetry in mind. We can write R as a sum of independent, mean-zero random matrices Z_{ij} that isolate entries (i, j) and (j, i) :

$$R = \sum_{i \leq j} Z_{ij}, \text{ where } Z_{ij} = \begin{cases} R_{ij}(e_i e_j^T + e_j e_i^T) & \text{if } i \neq j, \\ R_{ii} e_i e_i^T & \text{if } i = j. \end{cases}$$

Step 2: Bounding the error. Since $A_{ij} \in \{0, 1\}$,

$$|R_{ij}| \leq 1 \implies \|Z_{ij}\| = \|R_{ij}\| \leq 1 \implies \|R_{ij} Z_{ij}\| \leq 1.$$

Then by applying the matrix Bernstein inequality (Remark 5.4.11) combined with Markov's inequality, we obtain with probability at least 0.99:

$$\|R\| \lesssim \sigma \sqrt{\log n} + \log n \text{ where } \sigma^2 = \left\| \mathbb{E} \left[\sum_{i \leq j} Z_{ij}^2 \right] \right\|.$$

Let's compute σ^2 . A quick check shows that Z_{ij}^2 is a diagonal matrix:

$$Z_{ij}^2 = \begin{cases} R_{ij}^2(e_i e_i^T + e_j e_j^T) & \text{if } i < j, \\ R_{ii}^2 e_i e_i^T & \text{if } i = j. \end{cases}$$

Then, by symmetry,

$$\sum_{i \leq j} Z_{ij}^2 = \sum_{i \leq j} R_{ij}^2(e_i e_i^T + e_j e_j^T) + \sum_i R_{ii}^2 e_i e_i^T = \sum_{i=1}^n \left(\sum_{j=1}^n R_{ij}^2 \right) e_i e_i^T.$$

This is a diagonal matrix, and so is its expectation. Thus

$$\sigma^2 = \left\| \mathbb{E} \left[\sum_{i \leq j} Z_{ij}^2 \right] \right\| = \max_{i=1, \dots, n} \sum_{j=1}^n \mathbb{E} [R_{ij}^2]$$

since the operator norm of a diagonal matrix is the maximal absolute value of its entries (Exercise 4.3 (b)). Recall that $R_{ij} = A_{ij} - \mathbb{E} [A_{ij}]$. In the stochastic block model, A_{ij} is either $\text{Ber}(p)$ or $\text{Ber}(q)$. So $\mathbb{E} [R_{ij}^2] = \text{Var}(A_{ij}) \leq p$ since $p > q$. Thus

$$\sigma^2 \leq np = a,$$

and substituting this into the initial bound for $\|R\|$, we get

$$\|R\| \lesssim \sqrt{a \log n} + \log n \lesssim \sqrt{a \log n}$$

because the assumption implies that $a \gtrsim \log n$.

Step 3: Applying Davis-Kahan. Let's apply Theorem 4.1.15 (see Exercise 4.16) to D and A , focusing on the second largest eigenvalue. As we noted in Section 4.5.4, the separation between $\lambda_2(D)$ of D and the rest of the spectrum is

$$\delta = \min(\lambda_2(D), \lambda_1(D) - \lambda_2(D)) = \min\left(\frac{p-q}{2}, q\right) n = \frac{a-b}{2}$$

since $a \leq 3b$ by assumption. Using the bound on R from the end of Step 2, the Davis-Kahan inequality guarantees that for some $\theta \in \{-1, 1\}$, the distance between the *unit* eigenvectors of D and A (denoted with bars) satisfies

$$\|\bar{u}_2(D) - \theta \bar{u}_2(A)\|_2 \leq \frac{2\|R\|}{\delta} \leq \frac{C_1 \sqrt{a \log n}}{a-b} < \frac{1}{10}$$

if we choose the constant C in the assumption of the theorem to be large enough. Multiply both sides by \sqrt{n} to get

$$\|u_2(D) - \theta u_2(A)\|_2 \lesssim \frac{\sqrt{n}}{10}.$$

Since all coefficients of $u_2(D)$ are ± 1 and correctly identify community membership, it follows that at least 99% of the coefficients in $\theta u_2(A)_j$ have the same sign as $u_2(D)_j$, and thus correctly identify the community membership. \square

Remark 5.5.2 (Sparsity). The sparsest graphs for which Theorem 5.5.1 is nontrivial have expected average degree

$$\frac{n(p+q)}{2} = \frac{a+b}{2} \asymp \log n,$$

That's way sparser than the bound of $O(\sqrt{n})$ that we have achieved previously (Remark 4.5.3)!

5.6 Application: Covariance Estimation for General Distributions

In Section 4.7, we learned how to estimate the covariance matrix of a subgaussian distribution in \mathbb{R}^n from a sample of size $O(n)$. Now, we drop the subgaussian assumption, making this work for much broader distributions, even discrete ones. The trade-off is just a logarithmic oversampling factor!

For notation, we denote the sample covariance matrix as

$$\Sigma_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^T.$$

If X has zero mean, then Σ is the covariance matrix of X , and Σ_m is the sample covariance matrix.

Theorem 5.6.1 (General covariance estimation). Let X be a random vector in \mathbb{R}^n ($n \geq 2$). Assume that for some $K \geq 1$,

$$\|X\|_2 \leq K(\mathbb{E}[\|X\|_2^2])^{1/2} \text{ almost surely.}$$

Then for every positive integer m , we have

$$\mathbb{E}[\|\Sigma_m - \Sigma\|] \leq C \left(\sqrt{\frac{K^2 n \log n}{m}} + \frac{K^2 n \log n}{m} \right) \|\Sigma\|.$$

Proof. By Proposition 3.2.1 (b), we have $\mathbb{E}[\|X\|_2^2] = \text{tr}(\Sigma)$, hence the condition in the theorem becomes

$$\|X\|_2^2 \leq K^2 \text{tr}(\Sigma) \text{ almost surely.}$$

Apply the expected version of the matrix Bernstein inequality (Remark 5.4.11) for the sum of i.i.d. mean zero random matrices $X_i X_i^T - \Sigma$ and get

$$\mathbb{E}[\|\Sigma_m - \Sigma\|] = \frac{1}{m} \mathbb{E} \left[\left\| \sum_{i=1}^m (X_i X_i^T - \Sigma) \right\| \right] \lesssim \frac{1}{m} (\sigma \sqrt{\log n} + M \log n)$$

where

$$\sigma^2 = \left\| \sum_{i=1}^m \mathbb{E}[(X_i X_i^T - \Sigma)^2] \right\| = m \mathbb{E}[(X X^T - \Sigma)^2]$$

and M is any number chosen so that

$$\|X X^T - \Sigma\| \leq M \text{ almost surely.}$$

To complete the proof, it remains to bound σ^2 and M .

Let's start with σ^2 . By expanding the square,

$$\mathbb{E}[(X X^T - \Sigma)^2] = \mathbb{E}[(X X^T)^2] - \Sigma^2 \lesssim \mathbb{E}[(X X^T)^2]. \quad (*)$$

Furthermore, the assumption at the beginning of the proof gives

$$(X X^T)^2 = \|X\|^2 X X^T \lesssim K^2 \text{tr}(\Sigma) X X^T.$$

Taking expectations on both sides, we get

$$\mathbb{E}[(X X^T)^2] \lesssim K^2 \text{tr}(\Sigma) \Sigma.$$

Substituting this bound into (*), we get a bound for σ^2 :

$$\sigma^2 \leq K^2 m \text{tr}(\Sigma) \|\Sigma\|.$$

On the other hand, bounding M is easier:

$$\begin{aligned}\|XX^T - \Sigma\| &\leq \|X\|_2^2 + \|\Sigma\| \quad (\text{By triangle inequality}) \\ &\leq K^2 \text{tr}(\Sigma) + \|\Sigma\| \quad (\text{By assumption}) \\ &\leq 2K^2 \text{tr}(\Sigma) =: M. \quad (\|\Sigma\| \leq \text{tr}(\Sigma) \text{ and } K \geq 1).\end{aligned}$$

Substitute the bounds for σ and M into the overall bound, we get

$$\mathbb{E} [\|\Sigma_m - \Sigma\|] \leq \frac{1}{m} \left(\sqrt{K^2 m \text{tr}(\Sigma) \|\Sigma\|} \cdot \log n + 2K^2 \text{tr}(\Sigma) \cdot \log n \right).$$

Finally, plugging in the bound $\text{tr}(\Sigma) \leq n\|\Sigma\|$ completes the proof. \square

Remark 5.6.2 (Sample complexity). Theorem 5.6.1 shows that for any $\varepsilon \in (0, 1)$, we can estimate the covariance matrix with a small relative error:

$$\mathbb{E} [\|\Sigma_m - \Sigma\|] \leq \varepsilon \|\Sigma\|,$$

as long as the sample size is

$$m \asymp \varepsilon^{-2} n \log n.$$

Compared to the sample complexity $m \asymp \varepsilon^{-2} n$ for subgaussian distributions (Remark 4.7.2), dropping the subgaussian assumption costs just a small logarithmic oversampling factor! In general, this factor cannot be dropped (Exercise 5.28).

Remark 5.6.3 (Low-dimensional distributions). At the end of proof of Theorem 5.6.1, we used a rough bound $\text{tr}(\Sigma) \leq n\|\Sigma\|$. But instead, we can express the conclusion via the *effective rank* of Σ :

$$r = r(\Sigma) = \frac{\text{tr}(\Sigma)}{\|\Sigma\|}$$

and get a sharper bound

$$\mathbb{E} [\|\Sigma_m - \Sigma\|] \leq C \left(\sqrt{\frac{K^2 r \log n}{m}} + \frac{K^2 r \log n}{m} \right) \|\Sigma\|.$$

It shows that a sample of size

$$m \asymp \varepsilon^{-2} r \log n$$

is enough to estimate the covariance matrix. Since $r \leq n$, this sample size is at least as small as the value that we had estimated above. It is even much smaller for *approximately low-dimensional* distributions that concentrate near lower-dimensional subspaces.

Remark 5.6.4 (Effective and stable rank of a matrix). What does the effective rank from Remark 5.6.3 really tell us about a PSD matrix Σ ? TO get an idea, write it as the sum of eigenvalues divided by the largest one:

$$r(\Sigma) = \frac{\sum_{i=1}^n \lambda_i(\Sigma)}{\max_i \lambda_i(\Sigma)}.$$

This is always bounded by the actual rank (number of nonzero eigenvalues) and can be much smaller for “approximately” low-rank matrices - ones having only a few large eigenvalues. A related idea is the *stable rank*, defined for any matrix A

$$s(A) = \frac{\|A\|_F^2}{\|A\|^2} = \frac{\sum_{i=1}^n \sigma_i^2(A)}{\max_i \sigma_i^2(A)} = r(A^T A) = r(AA^T)$$

where σ_i denotes the singular values. Both are “soft” versions of rank that are stable under small changes. For some more intuition, see Exercise 5.26.

Remark 5.6.5 (High-probability guarantees). We covered expectation bounds, but our argument actually gives a more informative high-probability guarantee:

$$\|\Sigma_m - \Sigma\| \leq C \left(\sqrt{\frac{K^2 r (\log n + u)}{m}} + \frac{K^2 r (\log n + u)}{m} \right) \|\Sigma\|$$

with probability at least $1 - 2e^{-u}$. Here $r = \text{tr}(\Sigma)/\|\Sigma\| \leq n$ is the effective rank (Exercise 5.26).

Remark 5.6.6 (Boundedness assumption). The boundedness assumption in Theorem 5.6.1 might seem strong, but it cannot be dropped in general: if X is isotropic but zero with high probability, the sample is likely to consist entirely of zeros, making covariance estimation impossible (Exercise 5.27). However, this assumption can still be relaxed (Exercise 6.34). In practice, it is usually enforced by truncation - dropping a small percentage of samples with the largest norm.

5.7 Extra notes

There are lots of other concentration theorems not went over in the text. A very useful one is the McDiarmid inequality, which generalizes the Hoeffding inequality:

Theorem 5.7.1 (McDiarmid inequality). Let $X = (X_1, \dots, X_N)$ be a random vector with independent entries. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a measurable function. Assume that the value of $f(x)$ can change by at most $c_i > 0$ under an arbitrary change of a single coordinate of $x \in \mathbb{R}^n$. Then for any $t > 0$,

$$P(f(X) - \mathbb{E}[f(X)] \geq t) \leq \exp \left(-\frac{2t^2}{\sum_{i=1}^N c_i^2} \right).$$