HIGH-DIMENSIONAL PROBABILITY
An Introduction with Applications in Data Science
Reading Notes

Roman Vershynin, Second Edition

Xuanzhe Hou

October 2026

Contents

Appetizer: Using Probability to Cover a Set **1**

1 A Quick Refresher on Analysis and Probability **3**
1.1 Convex Sets and Functions . . . 3
1.2 Norms and Inner Products . . . 4
1.3 Random Variables and Random Vectors . . . 5
1.4 Union Bound . . . 6
1.5 Conditioning . . . 6
1.6 Probabilistic Inequalities . . . 7
1.7 Limit Theorems . . . 8

2 Concentration of Sums of Independent Random Variables **10**
2.1 Concentration inequalities and Gaussian tails . . . 10
2.2 Hoeffding Inequality . . . 11
2.3 Chernoff Inequality . . . 12
2.4 Application: Median-of-Means Estimator . . . 13
2.5 Application: Degrees of Random Graphs . . . 14
2.6 Subgaussian Distributions . . . 15
2.7 Subgaussian Hoeffding and Khintchine Inequalities . . . 16
2.8 Subexponential Distributions . . . 17
2.9 Bernstein Inequality . . . 19

3 Random Vectors in High Dimensions **20**
3.1 Concentration of the norm . . . 20
3.2 Covariance matrices and principal component analysis . . . 20
3.3 Examples of high-dimensional distributions . . . 21
3.4 Subgaussian distributions in higher dimensions . . . 22
3.5 Application: Grothendieck inequality and semidefinite programming . . . 23
3.6 Application: Maximum cut for graphs . . . 24
3.7 Kernel trick and tightening Grothendieck inequality . . . 24

4 Random Matrices **27**
4.1 A quick refresher on linear algebra . . . 27
4.2 Nets, covering, and packing . . . 28
4.3 Application: error correcting codes . . . 29
4.4 Upper bounds on subgaussian random matrices . . . 29
4.5 Application: community detection in networks . . . 30
4.6 Two-sided bounds on subgaussian matrices . . . 30
4.7 Application: covariance estimation and clustering . . . 30

5 Concentration Without Independence **32**
5.1 Concentration of Lipschitz functions on the sphere . . . 32
5.2 Concentration on other metric measure spaces . . . 33
5.3 Application: Johnson–Lindenstrauss lemma . . . 33
5.4 Matrix Bernstein inequality . . . 34
5.5 Application: community detection in sparse networks . . . 35

5.6 Application: covariance estimation for general distributions . . . 35

6 Quadratic Forms, Symmetrization, and Contraction 36
6.1 Decoupling . . . 36
6.2 Hanson–Wright inequality . . . 36
6.3 Symmetrization . . . 37
6.4 Random matrices with non-i.i.d. entries . . . 38
6.5 Application: matrix completion . . . 38
6.6 Contraction principle . . . 39

7 Random Processes 40
7.1 Basic concepts and examples . . . 40
7.2 Slepian, Sudakov–Fernique, and Gordon inequalities . . . 42
7.3 Application: sharp bounds for Gaussian matrices . . . 45
7.4 Sudakov inequality . . . 46
7.5 Gaussian width . . . 48
7.6 Application: random projections of sets . . . 51

8 Chaining 53
8.1 Dudley inequality . . . 53
8.2 Application: empirical processes . . . 54
8.3 VC dimension . . . 55
8.4 Application: statistical learning theory . . . 56
8.5 Generic chaining . . . 57
8.6 Chevet inequality . . . 58

9 Deviations of Random Matrices on Sets 59
9.1 Matrix deviation inequality . . . 59
9.2 Random matrices, covariance estimation, and Johnson–Lindenstrauss . . . 60
9.3 Random sections: the M^* bound and escape theorem . . . 60
9.4 Application: high-dimensional linear models . . . 61
9.5 Application: exact sparse recovery . . . 62
9.6 Deviations of random matrices for general norms . . . 64
9.7 Two-sided Chevet and Dvoretzky–Milman theorem . . . 64

Appetizer: Using Probability to Cover a Set

Convex combinations and empirical approximation

Definition 0.1 (Convex hull). A convex combination of $z_1, \dots, z_m \in \mathbb{R}^n$ is a vector

$$\sum_{i=1}^{m} \lambda_i z_i, \qquad \lambda_i \geq 0, \qquad \sum_{i=1}^{m} \lambda_i = 1.$$

The convex hull $\operatorname{conv}(T)$ is the set of all finite convex combinations of points of T.

Theorem 0.2 (Caratheodory; **Book 0.0.1**). Every point in the convex hull of $T \subset \mathbb{R}^n$ is a convex combination of at most $n+1$ points of T.

The number of points in Caratheodory's theorem depends on the ambient dimension. For an approximate representation, it can be bounded in terms of the desired accuracy alone.

Theorem 0.3 (Approximate Caratheodory; **Book 0.0.2**). Let $T \subset B_2^n$ and $x \in \operatorname{conv}(T)$. For every positive integer k, there are $x_1, \dots, x_k \in T$, with repetitions allowed, such that

$$\left\| x - \frac{1}{k} \sum_{j=1}^{k} x_j \right\|_2 \leq \frac{1}{\sqrt{k}}. \tag{0.1}$$

The weights are equal and the bound is independent of n.

Maurey's empirical method. Write $x = \sum_i \lambda_i z_i$ and define a random vector Z by $\mathbb{P}\{Z = z_i\} = \lambda_i$. Then $\mathbb{E}Z = x$. For independent copies $Z_1, \dots, Z_k$,

$$\begin{aligned} \mathbb{E} \left\| x - \frac{1}{k} \sum_{j=1}^{k} Z_j \right\|_2^2 &= \frac{1}{k^2} \mathbb{E} \left\| \sum_{j=1}^{k} (Z_j - x) \right\|_2^2 \\ &= \frac{1}{k} \mathbb{E} \left\| Z - \mathbb{E}Z \right\|_2^2 \leq \frac{1}{k} \mathbb{E} \left\| Z \right\|_2^2 \leq \frac{1}{k}. \end{aligned}$$

The cross terms vanish by independence and centering. Since the average squared error is at most $1/k$, at least one realization satisfies Equation (0.1).

Remark 0.4 (Existence from expectation). The expectation bound guarantees a realization whose squared error is at most the expectation. This use of a random construction to establish deterministic existence is the probabilistic method.

Covering and volume

Corollary 0.5 (Covering a polytope; **Book 0.0.3**). Let $P \subset B_2^n$ be a polytope with N vertices. For each positive integer k, P can be covered by at most N^k Euclidean balls of radius $1/\sqrt{k}$.

Proof. Use as centers all averages $k^{-1}\sum_{j=1}^{k} x_j$ of vertices, with repetition. There are at most N^k such averages. Since $P = \operatorname{conv}\{\text{vertices}\}$, approximate Caratheodory places every point of P within $1/\sqrt{k}$ of one center. ⊛

Theorem 0.6 (Volume of a polytope; **Book 0.0.4**). If $P \subset B_2^n$ has N vertices, then

$$\frac{\operatorname{Vol}(P)}{\operatorname{Vol}(B_2^n)} \le \left(3\sqrt{\frac{\log N}{n}}\right)^n. \tag{0.2}$$

Volume estimate. The preceding corollary gives

$$\operatorname{Vol}(P) \le N^k k^{-n/2} \operatorname{Vol}(B_2^n).$$

Minimizing $k \log N - (n/2)\log k$ over positive real k gives $k_0 = n/(2\log N)$. Taking $k = \lceil k_0 \rceil$ gives Equation (0.2) when $N \le e^{n/9}$. For $N > e^{n/9}$, the stated bound follows from $P \subset B_2^n$, since its right side exceeds one.

Remark 0.7 (High-dimensional interpretation; **Book 0.0.5**). If N grows subexponentially in n, the radius $3\sqrt{\log N/n}$ tends to zero. Equation (0.2) then bounds the volume of P by that of a Euclidean ball with this radius.

1 A Quick Refresher on Analysis and Probability

1.1 Convex Sets and Functions

Definition 1.1 (Convex set and convex function). A set $K \subset \mathbb{R}^n$ is convex when

$$\lambda x + (1-\lambda)y \in K \qquad (x, y \in K,\ 0 \le \lambda \le 1).$$

If K is convex, a function $f : K \to \mathbb{R}$ is convex when

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y). \tag{1.1}$$

A function is concave exactly when its negative is convex.

Proposition 1.2 (Elementary maximum principle). If $K = \operatorname{conv}\{x_1, \dots, x_m\}$ and f is convex on K, then

$$\sup_{x \in K} f(x) = \max_{1 \le i \le m} f(x_i).$$

Proof. For $x = \sum_i \lambda_i x_i$, repeated use of convexity gives $f(x) \le \sum_i \lambda_i f(x_i) \le \max_i f(x_i)$. ⊛

Example 1.3 (Norms are convex). For any norm and $0 \le \lambda \le 1$,

$$\|\lambda x + (1-\lambda)y\| \le \lambda \|x\| + (1-\lambda)\|y\|.$$

Consequently, every norm ball is convex.

Remark 1.4 (Extrema of convex functions). A convex function on a polytope attains a maximum at some vertex, but a minimum can lie in the interior. A concave function attains a minimum at some vertex.

1.2 Norms and Inner Products

The Euclidean norm and inner product are

$$\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{1/2}, \qquad \langle x, y\rangle = x^T y = \sum_{i=1}^n x_i y_i.$$

For $1 \le p < \infty$,

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, \qquad \|x\|_\infty = \max_i |x_i|.$$

Definition 1.5 (ℓ_p unit balls). The unit ball of $(\mathbb{R}^n, \|\cdot\|_p)$ is $B_p^n = \{x : \|x\|_p \le 1\}$. Two useful identifications are

$$B_\infty^n = [-1,1]^n, \qquad B_1^n = \operatorname{conv}\{\pm e_1, \dots, \pm e_n\}.$$

Theorem 1.6 (Minkowski, Holder, and duality). For $p \in [1, \infty]$,

$$\|x + y\|_p \le \|x\|_p + \|y\|_p.$$

If p, p' are conjugate, $1/p + 1/p' = 1$, then

$$|\langle x, y\rangle| \le \|x\|_p \|y\|_{p'}, \qquad \|x\|_p = \max_{\|y\|_{p'} \le 1} \langle x, y\rangle. \tag{1.2}$$

The case $p = p' = 2$ is Cauchy–Schwarz.

Proposition 1.7 (Comparison of finite-dimensional norms). If $1 \le p \le q \le \infty$, then

$$\|x\|_q \le \|x\|_p \le n^{1/p - 1/q} \|x\|_q.$$

Thus $B_p^n \subset B_q^n$ when $p \le q$.

Remark 1.8 (Dependence on p). For a fixed finite vector, $\|x\|_p$ decreases with p. For a fixed random variable on a probability space, $\|X\|_{L^p}$ increases with p. The normalizations differ: ℓ_p uses a sum, whereas expectation is an average.

1.3 Random Variables and Random Vectors

Definition 1.9 (Moments and L^p norms). For a random variable X,

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2, \qquad M_X(t) = \mathbb{E}e^{tX},$$

and for $1 \le p < \infty$,

$$\|X\|_{L^p} = (\mathbb{E}\,|X|^p)^{1/p}.$$

For $p = \infty$, $\|X\|_{L^\infty} = \operatorname{ess\,sup} |X|$. The space L^2 has inner product $\langle X, Y\rangle_{L^2} = \mathbb{E}(XY)$. Standard deviation and covariance are

$$\sigma(X) = \|X - \mathbb{E}X\|_{L^2}, \qquad \mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

Proposition 1.10 (Linearity and variance additivity). Expectation is linear without independence. If $X_1, \dots, X_n$ are uncorrelated, then

$$\mathrm{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \,\mathrm{Var}(X_i). \tag{1.3}$$

Independence is sufficient but not necessary for this identity.

Definition 1.11 (Random vector and covariance matrix). A random vector $X = (X_1, \dots, X_n)$ has mean $\mathbb{E}X = (\mathbb{E}X_1, \dots, \mathbb{E}X_n)$ and covariance matrix

$$\mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T]. \tag{1.4}$$

Its (i, j) entry is $\mathrm{Cov}(X_i, X_j)$.

Example 1.12 (Indicators). For an event E, the indicator $\mathbf{1}_E$ satisfies $\mathbb{E}\mathbf{1}_E = \mathbb{P}(E)$. Thus probabilities can be computed as expectations of indicators.

Remark 1.13 (Total variance). The scalar quantity $\mathbb{E}\,\|X - \mathbb{E}X\|_2^2$ records total dispersion and equals the trace of the covariance matrix. The matrix in Equation (1.4) also records directions and correlations.

1.4 Union Bound

Lemma 1.14 (Union bound; Book 1.4.1). For arbitrary events $E_1, \dots, E_n$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \le \sum_{i=1}^{n} \mathbb{P}(E_i). \tag{1.5}$$

Proof. Pointwise, $\mathbf{1}_{\cup_i E_i} \le \sum_i \mathbf{1}_{E_i}$. Take expectations. ⊛

Example 1.15 (No isolated vertices; Book 1.4.2). Let $G \sim G(n,p)$. For a fixed vertex, isolation has probability $(1-p)^{n-1}$. Hence

$$\mathbb{P}\{\text{some vertex is isolated}\} \le n(1-p)^{n-1} \le ne^{-p(n-1)}.$$

If $p \ge 4\log(n)/n$, this probability is at most $1/n$ for $n \ge 2$.

Remark 1.16 (A finite family of events). If each of N events has probability at most e^{-u}, their union has probability at most $e^{\log N - u}$. In particular, the bound becomes smaller than one when $u > \log N$, regardless of dependence among the events.

1.5 Conditioning

Definition 1.17 (Conditional probability and expectation). For events with $\mathbb{P}(F) > 0$,

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}.$$

For random variables, $\mathbb{E}[X \mid Y]$ is the conditional average of X given the information carried by Y.

Theorem 1.18 (Total expectation and total probability). For integrable X,

$$\mathbb{E}X = \mathbb{E}\big[\mathbb{E}(X \mid Y)\big]. \tag{1.6}$$

For an event E,

$$\mathbb{P}(E) = \mathbb{E}[\mathbb{P}(E \mid Y)].$$

If (F_i) partitions the sample space, then

$$\mathbb{P}(E) = \sum_i \mathbb{P}(E \mid F_i)\mathbb{P}(F_i). \tag{1.7}$$

Example 1.19 (Perfect cancellation; **Book 1.5.1**). Let X_i be independent Rademacher variables and let not all a_i vanish. Then

$$\mathbb{P}\left\{\sum_{i=1}^{n} a_i X_i = 0\right\} \le \frac{1}{2}.$$

Choose $a_n \neq 0$ and condition on $X_1, \dots, X_{n-1}$. The event fixes at most one value of X_n, and either sign has probability $1/2$. Average the conditional bound by Equation (1.6).

Remark 1.20 (Independence and conditioning). If X and Y are independent, the conditional law of X given Y equals the marginal law of X. In the preceding example, this leaves X_n as a Rademacher variable after the first $n-1$ coordinates have been fixed.

1.6 Probabilistic Inequalities

Theorem 1.21 (Jensen, Minkowski, and Holder). If f is convex and the expectations exist, then

$$f(\mathbb{E}X) \le \mathbb{E}f(X). \tag{1.8}$$

In particular, $\|\mathbb{E}X\| \le \mathbb{E}\,\|X\|$. For $p \ge 1$,

$$\|X+Y\|_{L^p} \le \|X\|_{L^p} + \|Y\|_{L^p}.$$

For conjugate p, p',

$$\mathbb{E}\,|XY| \le \|X\|_{L^p}\,\|Y\|_{L^{p'}}.$$

Lemma 1.22 (Integrated tail formula; **Book 1.6.1**). For every nonnegative random variable X,

$$\mathbb{E}X = \int_0^\infty \mathbb{P}\{X > t\}\,dt, \tag{1.9}$$

with both sides allowed to be infinite.

Proof. Use $x = \int_0^\infty \mathbf{1}_{\{t<x\}}\,dt$, substitute X for x, and apply Tonelli's theorem. ⊛

Proposition 1.23 (Markov inequality; **Book 1.6.2**). If $X \ge 0$ and $t > 0$, then

$$\mathbb{P}\{X \ge t\} \le \frac{\mathbb{E}X}{t}. \tag{1.10}$$

Corollary 1.24 (Chebyshev inequality; Book 1.6.3). If X has mean μ and variance σ^2, then for every $t > 0$,

$$\mathbb{P}\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}. \tag{1.11}$$

Markov and Chebyshev inequalities. Markov compares X with $t\mathbf{1}_{\{X \geq t\}}$. Chebyshev applies Markov to $(X - \mu)^2$. Applying Markov to $e^{\lambda X}$ instead gives $\mathbb{P}\{X \geq t\} \leq e^{-\lambda t}\mathbb{E}e^{\lambda X}$ for $\lambda > 0$.

Remark 1.25 (Moment monotonicity). On a probability space, $\|X\|_{L^p} \leq \|X\|_{L^q}$ for $1 \leq p \leq q \leq \infty$. Higher moments give greater weight to large values of $|X|$.

Remark 1.26 (Strict and weak tail probabilities). Replacing $>$ with $\geq$ leaves the integrated tail formula unchanged: the two probabilities differ only at atoms, a set of thresholds of Lebesgue measure zero.

1.7 Limit Theorems

Let $X_1, X_2, \ldots$ be IID with mean μ, and write $S_N = X_1 + \cdots + X_N$.

Theorem 1.27 (Strong law of large numbers; Book 1.7.1). If $\mathbb{E}\,|X_1| < \infty$, then

$$\frac{S_N}{N} \longrightarrow \mu \qquad \text{almost surely.}$$

Definition 1.28 (Normal distribution; Book 1.7.2). A standard normal random variable has density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

If $Z \sim N(0,1)$, $\mu \in \mathbb{R}$, and $\sigma > 0$, then $X = \mu + \sigma Z$ has distribution $N(\mu, \sigma^2)$ and density

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Theorem 1.29 (Lindeberg–Levy central limit theorem; Book 1.7.3). If $0 < \sigma^2 = \mathrm{Var}(X_1) < \infty$, then

$$\frac{S_N - N\mu}{\sigma\sqrt{N}} \Longrightarrow N(0,1).$$

Here $\Longrightarrow$ denotes convergence in distribution.

Example 1.30 (Bernoulli and binomial laws; Book 1.7.4). Fix $p \in (0,1)$. If $X_i \sim \mathrm{Ber}(p)$ independently, then $S_N \sim \mathrm{Binom}(N,p)$ with $\mathbb{E}S_N = Np$ and $\mathrm{Var}(S_N) = Np(1-p)$. Therefore

$$\frac{S_N - Np}{\sqrt{Np(1-p)}} \Longrightarrow N(0,1).$$

This special case is the de Moivre–Laplace theorem.

Definition 1.31 (Poisson distribution; Book 1.7.5). For $\lambda > 0$, $Z \sim \mathrm{Pois}(\lambda)$ when

$$\mathbb{P}\{Z = k\} = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots.$$

The distribution $\mathrm{Pois}(0)$ is the point mass at zero.

Theorem 1.32 (Poisson limit theorem; Book 1.7.6). Let $X_{N,i} \sim \mathrm{Ber}(p_{N,i})$ be independent for $1 \le i \le N$. If

$$\max_{i \le N} p_{N,i} \longrightarrow 0, \qquad \sum_{i=1}^{N} p_{N,i} \longrightarrow \lambda < \infty,$$

then $\sum_{i=1}^{N} X_{N,i} \Longrightarrow \mathrm{Pois}(\lambda)$.

Lemma 1.33 (Stirling approximation; Book 1.7.7). As $n \to \infty$,

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 + o(1)). \tag{1.12}$$

Lemma 1.34 (Elementary factorial bounds; Book 1.7.8). For every positive integer n,

$$\left(\frac{n}{e}\right)^n \le n! \le en\left(\frac{n}{e}\right)^n.$$

Remark 1.35 (Gamma function; Book 1.7.9). The factorial extends through

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt \quad (\mathrm{Re}\, z > 0), \qquad \Gamma(n+1) = n!.$$

The Stirling asymptotic extends to $\Gamma(z+1)$ as the real variable z tends to infinity.

The central limit theorem describes convergence at fixed thresholds. Thresholds that grow with N require separate tail estimates.

2 Concentration of Sums of Independent Random Variables

2.1 Concentration inequalities and Gaussian tails

Concentration inequalities bound the probability that a random variable deviates from a typical value. For deviations from the mean, they take the form

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \text{a rapidly decreasing function of } t.$$

The constants $C, c > 0$ in these estimates are absolute and may change between occurrences. For nonnegative quantities, $a \lesssim b$ means $a \leq Cb$, and $a \asymp b$ means that both $a \lesssim b$ and $b \lesssim a$ hold.

For S_N equal to the number of heads in N fair tosses, Chebyshev gives only

$$\mathbb{P}\{S_N \geq 3N/4\} \leq \frac{4}{N},$$

whereas this probability decays exponentially in N.

Proposition 2.1 (Gaussian tails; **Book 2.1.2**). If $g \sim N(0,1)$ and $t > 0$, then

$$\frac{t}{t^2+1}\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \leq \mathbb{P}\{g \geq t\} \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}. \tag{2.1}$$

In particular, for $t \geq 1$,

$$\mathbb{P}\{g \geq t\} \leq \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

Integration by parts. Write the tail as $\int_t^\infty \phi(x)\,dx$ and use $\phi'(x) = -x\phi(x)$. For the upper bound, replace 1 by x/t inside the integral. For the lower bound, integrate by parts once and bound the remainder using $x^{-2} \leq t^{-2}$ on $[t, \infty)$.

Theorem 2.2 (Berry–Esseen; **Book 2.1.4**). Let X_i be IID with mean μ, variance $\sigma^2 > 0$, and finite third absolute centered moment. If

$$Z_N = \frac{1}{\sigma\sqrt{N}}\sum_{i=1}^{N}(X_i - \mu), \qquad \rho = \frac{\mathbb{E}\,|X_1 - \mu|^3}{\sigma^3},$$

then, uniformly in $t \in \mathbb{R}$,

$$|\mathbb{P}\{Z_N \geq t\} - \mathbb{P}\{g \geq t\}| \leq \frac{\rho}{\sqrt{N}}.$$

Remark 2.3 (Accuracy of the normal approximation). An additive error of order $N^{-1/2}$ is larger than a tail probability of order e^{-cN}. To estimate such probabilities, a concentration inequality must bound the tail directly.

2.2 Hoeffding Inequality

Definition 2.4 (Rademacher variable). A Rademacher random variable ε takes values 1 and -1 with probability $1/2$ each. It is centered, has variance one, and satisfies

$$\mathbb{E}e^{\lambda\varepsilon} = \cosh\lambda \le e^{\lambda^2/2} \qquad (\lambda \in \mathbb{R}). \tag{2.2}$$

Theorem 2.5 (Rademacher Hoeffding; Book 2.2.1). Let $X_1, \dots, X_N$ be independent Rademacher variables and $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for $t \ge 0$,

$$\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \ge t\right\} \le \exp\left(-\frac{t^2}{2\|a\|_2^2}\right). \tag{2.3}$$

Exponential moments. For $\lambda \ge 0$, Markov and independence give

$$\begin{aligned}\mathbb{P}\left\{\sum_i a_i X_i \ge t\right\} &\le e^{-\lambda t}\mathbb{E}\exp\left(\lambda\sum_i a_i X_i\right)\\ &= e^{-\lambda t}\prod_i \mathbb{E}e^{\lambda a_i X_i}\\ &\le \exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_i a_i^2\right).\end{aligned}$$

The quadratic exponent is minimized at $\lambda = t/\|a\|_2^2$, which gives Equation (2.3).

Theorem 2.6 (Two-sided Hoeffding; Book 2.2.5). Under the same assumptions,

$$\mathbb{P}\left\{\left|\sum_i a_i X_i\right| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

Theorem 2.7 (Bounded Hoeffding; Book 2.2.6). Let $X_1, \dots, X_N$ be independent and satisfy $X_i \in [a_i, b_i]$ almost surely. Then, for every $t \ge 0$,

$$\mathbb{P}\left\{\sum_{i=1}^N (X_i - \mathbb{E}X_i) \ge t\right\} \le \exp\left(-\frac{2t^2}{\sum_{i=1}^N (b_i - a_i)^2}\right). \tag{2.4}$$

Bounded variables. Hoeffding's lemma bounds the moment generating function (MGF) of a centered variable supported on $[a_i, b_i]$ by

$$\mathbb{E}e^{\lambda(X_i - \mathbb{E}X_i)} \le \exp\left(\frac{\lambda^2(b_i - a_i)^2}{8}\right).$$

Substitution into the preceding argument and optimization over λ give the bound.

Remark 2.8 (Variance of a Rademacher sum). The denominator $\|a\|_2^2$ is the variance of a Rademacher sum. Thus a typical fluctuation has size $\|a\|_2$, and a deviation of size $u\|a\|_2$ has probability at most $2e^{-cu^2}$.

2.3 Chernoff Inequality

Let $X_i \sim \mathrm{Ber}(p_i)$ independently and set $S_N = \sum_i X_i$ and $\mu = \mathbb{E}S_N = \sum_i p_i$.

Theorem 2.9 (Chernoff upper tail; Book 2.3.1). For every $t \ge \mu$,

$$\mathbb{P}\{S_N \ge t\} \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t. \tag{2.5}$$

Bernoulli exponential moments. For $\lambda \ge 0$,

$$\mathbb{E}e^{\lambda X_i} = 1 + p_i(e^\lambda - 1) \le \exp(p_i(e^\lambda - 1)).$$

Hence

$$\mathbb{P}\{S_N \ge t\} \le \exp[-\lambda t + \mu(e^\lambda - 1)].$$

For $t \ge \mu$, optimize at $e^\lambda = t/\mu$ to obtain Equation (2.5).

Writing $t = (1+\delta)\mu$ gives

$$\mathbb{P}\{S_N \ge (1+\delta)\mu\} \le \exp\{-\mu[(1+\delta)\log(1+\delta) - \delta]\}. \tag{2.6}$$

Corollary 2.10 (Small relative deviations; Book 2.3.4). For $0 \le \delta \le 1$,

$$\mathbb{P}\{|S_N - \mu| \ge \delta\mu\} \le 2\exp\left(-\frac{\delta^2\mu}{3}\right). \tag{2.7}$$

Quadratic rate near the mean. Use the elementary lower bound $(1+\delta)\log(1+\delta)-\delta \geq \delta^2/3$ on $[0,1]$ in Equation (2.6). Repeat the MGF argument for $-S_N$ to control the lower tail and combine the two events.

Remark 2.11 (Small and large deviations). For $t-\mu$ of order at most μ, the exponent is quadratic, like a Gaussian tail. Far beyond the mean, the more accurate rate behaves like $t\log(t/\mu)$, which is stronger than a merely exponential tail.

Remark 2.12 (Poisson tails). The expression in Equation (2.5) also bounds Poisson tails. This is natural because a Poisson variable is a limit of sums of rare independent Bernoulli variables.

Remark 2.13 (Relative deviations). Equation (2.7) controls an absolute deviation $\delta\mu$, where δ is the deviation relative to the mean.

2.4 Application: Median-of-Means Estimator

The ordinary sample mean has variance σ^2/N. Under a finite variance assumption, Chebyshev's inequality gives a tail bound that decreases as the inverse square of the threshold. The median of several independent block means gives an error of order $N^{-1/2}$ with a Gaussian tail bound.

Definition 2.14 (Median-of-means construction). Split $N = BL$ observations into B disjoint blocks of length L. Let

$$\overline{X}_b = \frac{1}{L}\sum_{i=(b-1)L+1}^{bL} X_i, \qquad b = 1,\dots,B,$$

and define $\widehat{\mu}$ to be a median of $\overline{X}_1,\dots,\overline{X}_B$.

Theorem 2.15 (Median-of-means; **Book 2.4.1**). Let $X_1,\dots,X_N$ be independent copies of a random variable with mean μ and finite variance $\sigma^2 > 0$. For every $0 \leq t \leq \sqrt{N}$, one can choose a median-of-means estimator such that

$$\mathbb{P}\left\{|\widehat{\mu}-\mu| \geq \frac{t\sigma}{\sqrt{N}}\right\} \leq 2e^{-ct^2}. \tag{2.8}$$

Concentration of the median. Choose B of order t^2 and L of order N/t^2. Each block mean has variance σ^2/L, so Chebyshev bounds the probability that a block mean exceeds $\mu + t\sigma/\sqrt{N}$ by $1/4$. If the median exceeds this threshold, at least half of the independent block means do as well. Chernoff bounds this event by $e^{-cB} = e^{-ct^2}$. Treat the lower tail in the same way.

Remark 2.16 (Robustness). If more than half of the block means lie in an interval, their median lies in that interval regardless of the values of the remaining block means. This property allows the estimator to tolerate a few extreme observations.

Remark 2.17 (Choice of block sizes). The proof uses $B = t^2/4$ when this is an integer dividing N. In general, rounding the number of blocks and allowing unused observations or blocks of unequal sizes changes only the absolute constants, as in Exercise 2.16.

2.5 Application: Degrees of Random Graphs

In $G(n, p)$, every unordered pair of distinct vertices is connected independently with probability p. The degree d_i of vertex i is a binomial random variable with mean

$$d = (n-1)p.$$

Proposition 2.18 (Dense graphs are almost regular; **Book 2.5.1**). There is an absolute C such that, if $d \geq C \log n$, then with probability at least 0.99 every vertex satisfies

$$0.9d \leq d_i \leq 1.1d.$$

Chernoff inequality and the union bound. For a fixed vertex, Chernoff gives $\mathbb{P}\{|d_i - d| \geq 0.1d\} \leq 2e^{-cd}$. A union bound over all n vertices gives failure probability at most $2ne^{-cd}$. The condition $d \geq C \log n$ makes this at most 0.01.

Remark 2.19 (Sharp threshold scale). When the expected degree is below the threshold of order $\log n$, isolated vertices occur with substantial probability. Their degrees differ from d by d, so the stated bound cannot hold for every vertex.

Remark 2.20 (Dependence among degrees). Different degrees share edge indicators. Chernoff's inequality applies to the independent indicators within each degree, and the union bound then combines the resulting estimates over the vertices.

2.6 Subgaussian Distributions

Proposition 2.21 (Equivalent subgaussian properties; **Book 2.6.1**). For a random variable X, the following conditions are equivalent up to absolute changes in the parameter K:

1. $\mathbb{P}\{|X| \geq t\} \leq 2e^{-t^2/K^2}$ for all $t \geq 0$;
2. $\|X\|_{L^p} \leq K\sqrt{p}$ for all $p \geq 1$;
3. $\mathbb{E}e^{X^2/K^2} \leq 2$.

If $\mathbb{E}X = 0$, these are also equivalent to

$$\mathbb{E}e^{\lambda X} \leq e^{K^2\lambda^2} \qquad (\lambda \in \mathbb{R}). \tag{2.9}$$

Equivalence of the subgaussian properties. Integrate the tail of $|X|^p$ to obtain moment growth. Use the exponential series and the moment estimates to bound $\mathbb{E}e^{X^2/K^2}$. Markov applied to e^{X^2/K^2} returns a Gaussian tail. For centered X, expand $e^{\lambda X}$, eliminate the linear term, and control the remaining moments to obtain Equation (2.9).

Definition 2.22 (Subgaussian norm; **Book 2.6.4**). A random variable is subgaussian when the preceding conditions hold. Its subgaussian norm is

$$\|X\|_{\psi_2} = \inf\{K > 0 : \mathbb{E}e^{X^2/K^2} \leq 2\}. \tag{2.10}$$

Proposition 2.23 (Standard bounds in ψ_2 notation; **Book 2.6.6**). Every subgaussian X satisfies, for $t \geq 0$ and $p \geq 1$,

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp\left(-\frac{ct^2}{\|X\|_{\psi_2}^2}\right),$$
$$\|X\|_{L^p} \leq C\,\|X\|_{\psi_2}\sqrt{p}.$$

If $\mathbb{E}X = 0$, then for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}e^{\lambda X} \leq \exp(C\lambda^2\,\|X\|_{\psi_2}^2).$$

Each estimate characterizes $\|X\|_{\psi_2}$ up to absolute factors.

Example 2.24 (Basic subgaussian variables). Rademacher variables, centered bounded variables, and Gaussian variables are subgaussian. A constant is also subgaussian. In contrast, exponential and Poisson variables are not subgaussian because their tails decay more slowly than e^{-ct^2}.

Remark 2.25 (Norm properties). The functional in Equation (2.10) is a norm on the space of subgaussian random variables. In particular,

$$\|X+Y\|_{\psi_2} \le \|X\|_{\psi_2} + \|Y\|_{\psi_2}.$$

For independent centered variables, the subgaussian norm of the sum also satisfies a bound in terms of the sum of the squared norms.

Remark 2.26 (Centering and exponential moments). If Equation (2.9) holds for every positive and negative λ, differentiating at zero gives $\mathbb{E}X = 0$. For a general subgaussian variable, the moment generating function has the form $\mathbb{E}e^{\lambda X} = e^{\lambda \mathbb{E}X}\mathbb{E}e^{\lambda(X-\mathbb{E}X)}$.

2.7 Subgaussian Hoeffding and Khintchine Inequalities

2.7.1 Subgaussian norm of a sum

Proposition 2.27 (Subgaussian norm of an independent sum; **Book 2.7.1**). If $X_1, \dots, X_N$ are independent, centered, and subgaussian, then

$$\left\| \sum_{i=1}^N X_i \right\|_{\psi_2}^2 \le C \sum_{i=1}^N \|X_i\|_{\psi_2}^2. \tag{2.11}$$

Proof. Independence gives the product of the centered moment generating functions. Their bounds yield a quadratic exponent with coefficient $C\sum_i \|X_i\|_{\psi_2}^2$. The equivalent subgaussian properties then give the stated ψ_2 bound. ⊛

Theorem 2.28 (Subgaussian Hoeffding; **Book 2.7.3**). Under the same assumptions, for every $t \ge 0$,

$$\mathbb{P}\left\{ \left| \sum_{i=1}^N X_i \right| \ge t \right\} \le 2\exp\left(-\frac{ct^2}{\sum_i \|X_i\|_{\psi_2}^2} \right). \tag{2.12}$$

Applied to $a_i X_i$, Equation (2.12) has denominator $\sum_i a_i^2 \|X_i\|_{\psi_2}^2$.

2.7.2 Khintchine inequality

Theorem 2.29 (Subgaussian Khintchine; **Book 2.7.5**). Let X_i be independent centered subgaussian variables with $\mathbb{E}X_i^2 = 1$, and put $K = \max_i \|X_i\|_{\psi_2}$. For $p \ge 2$ and $a \in \mathbb{R}^N$,

$$\|a\|_2 \le \left\| \sum_{i=1}^N a_i X_i \right\|_{L^p} \le CK\sqrt{p}\,\|a\|_2. \tag{2.13}$$

Moment bounds. At $p = 2$, independence, zero means, and unit variances give equality with $\|a\|_2$. Monotonicity of L^p norms gives the lower bound for $p \geq 2$. For the upper bound, first use the subgaussian sum estimate and then the moment characterization $\|Y\|_{L^p} \leq C \|Y\|_{\psi_2} \sqrt{p}$.

2.7.3 Maxima and centering

Proposition 2.30 (Maximum of subgaussians; Book 2.7.6). For $N \geq 2$, subgaussian variables $X_1, \ldots, X_N$, not necessarily independent, satisfy

$$\left\| \max_{i \leq N} |X_i| \right\|_{\psi_2} \leq C \sqrt{\log N} \max_i \|X_i\|_{\psi_2} . \tag{2.14}$$

Consequently, $\mathbb{E} \max_i |X_i| \leq C \sqrt{\log N} \max_i \|X_i\|_{\psi_2}$.

Maximum of a finite family. The union bound gives

$$\mathbb{P}\{\max_i |X_i| \geq t\} \leq 2N \exp(-ct^2/K^2).$$

The transition to a small probability occurs at $t \asymp K\sqrt{\log N}$. Integrating or converting this tail gives Equation (2.14).

Lemma 2.31 (Centering; Book 2.7.8). Every subgaussian variable satisfies

$$\|X - \mathbb{E}X\|_{\psi_2} \leq C \|X\|_{\psi_2} .$$

2.8 Subexponential Distributions

Subexponential variables have exponentially decaying or lighter tails. They arise naturally by squaring a subgaussian variable or multiplying two subgaussian variables.

Proposition 2.32 (Equivalent subexponential properties; Book 2.8.1). The following conditions are equivalent up to absolute changes in K:

1. $\mathbb{P}\{|X| \geq t\} \leq 2e^{-t/K}$ for all $t \geq 0$;
2. $\|X\|_{L^p} \leq Kp$ for all $p \geq 1$;
3. $\mathbb{E}e^{|X|/K} \leq 2$.

If $\mathbb{E}X = 0$, they are also equivalent to the local MGF estimate

$$\mathbb{E}e^{\lambda X} \leq e^{K^2\lambda^2} \qquad \text{for } |\lambda| \leq 1/K. \tag{2.15}$$

Definition 2.33 (Subexponential norm; Book 2.8.4).

$$\|X\|_{\psi_1} = \inf\{K > 0 : \mathbb{E}e^{|X|/K} \leq 2\}. \tag{2.16}$$

Lemma 2.34 (Square equivalence; **Book 2.8.5**). X is subgaussian if and only if X^2 is subexponential, and

$$\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2 . \tag{2.17}$$

Lemma 2.35 (Product of subgaussians; **Book 2.8.6**). If X and Y are subgaussian, with no independence assumption, then

$$\|XY\|_{\psi_1} \le \|X\|_{\psi_2} \|Y\|_{\psi_2} . \tag{2.18}$$

Proof. After normalization, use $|XY| \le (X^2 + Y^2)/2$ and Cauchy–Schwarz:

$$\mathbb{E}e^{|XY|} \le \left(\mathbb{E}e^{X^2}\mathbb{E}e^{Y^2}\right)^{1/2} \le 2.$$

Rescale to obtain Equation (2.18). ⊛

Example 2.36 (Canonical subexponential variables). The square of a Gaussian, the product of two subgaussians, centered exponential variables, and bounded variables are subexponential. A Gaussian itself is both subgaussian and subexponential; the latter classification is weaker.

Remark 2.37 (Exponential moments near zero). A centered subgaussian has a quadratic MGF bound for all λ. A centered subexponential has a quadratic MGF bound guaranteed near zero. Optimizing this local bound produces a Gaussian regime until the optimal λ hits the boundary, followed by an exponential regime. This is exactly the shape of Bernstein's inequality.

Remark 2.38 (Beyond ψ_1 and ψ_2). Orlicz ψ_α classes describe tail bounds of the form e^{-ct^α}. The subgaussian and subexponential cases correspond to $\alpha = 2$ and $\alpha = 1$, respectively.

2.9 Bernstein Inequality

Theorem 2.39 (Subexponential Bernstein; Book 2.9.1). Let $X_1, \dots, X_N$ be independent, centered, and subexponential. Then, for $t \geq 0$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\sum_i \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right]. \tag{2.19}$$

Optimization over the exponential parameter. The local MGF bound gives

$$\mathbb{E}\exp\left(\lambda\sum_i X_i\right) \leq \exp\left(C\lambda^2\sum_i \|X_i\|_{\psi_1}^2\right)$$

provided $|\lambda| \leq c/\max_i \|X_i\|_{\psi_1}$. Without the constraint, the quadratic exponent is optimized by $\lambda \asymp t/\sum_i \|X_i\|_{\psi_1}^2$. If this value is allowed, one gets the quadratic term in Equation (2.19); if it is too large, optimize at the boundary and obtain the linear term. Repeat for the negative sum.

Corollary 2.40 (Weighted Bernstein; Book 2.9.2). Under the same assumptions, put $K = \max_i \|X_i\|_{\psi_1}$. For $a \in \mathbb{R}^N$ and $t \geq 0$,

$$\mathbb{P}\left\{\left|\sum_i a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]. \tag{2.20}$$

For the normalized sum $N^{-1}\sum_i X_i$, the two regimes become

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_i X_i\right| \geq u\right\} \leq 2\exp[-cN\min(u^2/K^2, u/K)].$$

Small deviations are Gaussian in u; large deviations are exponential.

Theorem 2.41 (Bernstein for bounded variables; Book 2.9.5). Let X_i be independent and centered, with $|X_i| \leq K$ almost surely, and put $\sigma^2 = \sum_i \mathbb{E}X_i^2$. Then, for every $t \geq 0$,

$$\mathbb{P}\left\{\left|\sum_i X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right). \tag{2.21}$$

Remark 2.42 (Hoeffding versus Bernstein). Hoeffding depends only on coordinate ranges. Bernstein uses the actual sum of variances and is sharper when the variables are usually much smaller than their bound K. For very large deviations, the maximal bound K necessarily reappears.

3 Random Vectors in High Dimensions

3.1 Concentration of the norm

The squared Euclidean norm is the sum of the squared coordinates. If each coordinate has unit second moment, then $\mathbb{E}\,\|X\|_2^2 = n$. Under independence and subgaussian assumptions, the norm concentrates around $\sqrt{n}$.

Theorem 3.1 (Concentration of the norm; Book Theorem 3.1.1). Let $X = (X_1, \dots, X_n)$ have independent subgaussian coordinates satisfying $\mathbb{E}X_i^2 = 1$, and put $K = \max_i \|X_i\|_{\psi_2}$. Then

$$\big\|\|X\|_2 - \sqrt{n}\big\|_{\psi_2} \le CK^2.$$

Equivalently, for every $t \ge 0$,

$$\mathbb{P}\big\{\big|\|X\|_2 - \sqrt{n}\big| \ge t\big\} \le 2\exp\left(-\frac{ct^2}{K^4}\right).$$

Proof sketch. The centered squared norm is $\|X\|_2^2 - n = \sum_i (X_i^2 - 1)$. Each summand is subexponential, since the square of a subgaussian variable is subexponential. Bernstein's inequality therefore bounds its tail in the quadratic and linear regimes. The identity $|z - n| = |\sqrt{z} - \sqrt{n}|(\sqrt{z} + \sqrt{n})$ then gives a Gaussian tail for the norm.

When K is bounded, the vector lies with high probability in a shell of constant width around the sphere of radius $\sqrt{n}$.

3.2 Covariance matrices and principal component analysis

The second moment matrix and the covariance matrix are

$$M = \mathbb{E}XX^T, \qquad \Sigma = \mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T] = M - (\mathbb{E}X)(\mathbb{E}X)^T.$$

They coincide when X is centered.

Proposition 3.2 (Second moments; Book Proposition 3.2.1). For every $v \in \mathbb{R}^n$,

$$\mathbb{E}\,\langle X, v\rangle^2 = v^T M v, \qquad \mathbb{E}\,\|X\|_2^2 = \operatorname{tr} M.$$

If Y is an independent copy of X, then

$$\mathbb{E}\,\langle X, Y\rangle^2 = \|M\|_F^2\,.$$

Let $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$ be the eigenvalues of Σ with orthonormal eigenvectors $v_1, \dots, v_n$. The Rayleigh quotient gives

$$\lambda_k = \max_{\substack{\|v\|_2 = 1 \\ v \perp v_1, \dots, v_{k-1}}} v^T \Sigma v.$$

Thus the first principal component maximizes projected variance, and each later component maximizes the remaining variance after removing the previous directions.

Definition 3.3 (Isotropic random vector; Book Definition 3.2.5). A random vector $X \in \mathbb{R}^n$ is *isotropic* if

$$\mathbb{E}XX^T = I_n,$$

or, equivalently, $\mathbb{E}\langle X, v\rangle^2 = \|v\|_2^2$ for every $v \in \mathbb{R}^n$.

For a centered vector with positive definite covariance Σ, the transformed vector $\Sigma^{-1/2}X$ is isotropic. This transformation is called whitening.

3.3 Examples of high-dimensional distributions

3.3.1 Gaussian vectors

The standard normal $g \sim N(0, I_n)$ has density

$$(2\pi)^{-n/2}\exp(-\|x\|_2^2/2)$$

and is rotation invariant. Consequently, $\langle g, v\rangle \sim N(0, \|v\|_2^2)$ for every fixed v. More generally, $X \sim N(\mu, \Sigma)$ means

$$X \overset{d}{=} \mu + \Sigma^{1/2}g.$$

If Σ is nonsingular, its density is

$$\frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right].$$

For a jointly Gaussian vector, uncorrelated coordinates are independent.

3.3.2 The sphere and convex bodies

If θ is uniform on S^{n-1}, then $\theta \overset{d}{=} g/\|g\|_2$, and $\sqrt{n}\,\theta$ is isotropic. Fixed one-dimensional marginals become approximately Gaussian as the dimension increases. The projective central limit theorem makes this approximation uniform over directions.

Uniform measures on convex bodies provide further examples with dependent coordinates. Concentration for these measures depends on the geometry of the body.

Theorem 3.4 (Projective central limit theorem; Book Theorem 3.3.9). If X_n is uniform on S^{n-1}, then for every choice of unit direction $v_n \in S^{n-1}$,

$$\sqrt{n}\,\langle X_n, v_n\rangle \implies N(0,1) \qquad (n \to \infty).$$

Rotation invariance makes the approximation uniform over the chosen direction.

3.3.3 Parseval frames

Vectors $u_1, \dots, u_N \in \mathbb{R}^n$ form a Parseval frame precisely when any of the following equivalent identities holds:

$$\sum_{i=1}^N u_i u_i^T = I_n, \qquad \sum_{i=1}^N \langle u_i, x\rangle^2 = \|x\|_2^2, \qquad x = \sum_{i=1}^N \langle u_i, x\rangle\, u_i.$$

Equivalently, a random vector uniform on $\{\sqrt{N}u_i\}_{i=1}^N$ is isotropic. Thus Parseval frames give discrete isotropic distributions.

3.4 Subgaussian distributions in higher dimensions

Definition 3.5 (Subgaussian random vector; Book Definition 3.4.1). A random vector $X \in \mathbb{R}^n$ is subgaussian if every one-dimensional marginal is subgaussian. Its norm is

$$\|X\|_{\psi_2} = \sup_{v \in S^{n-1}} \|\langle X, v\rangle\|_{\psi_2}.$$

Independent centered subgaussian coordinates give a subgaussian vector whose ψ_2 norm is bounded by an absolute constant times the largest coordinate norm. The uniform distribution on the sphere also has subgaussian marginals, although its coordinates are dependent.

Theorem 3.6 (Uniform measure on the sphere; Book Theorem 3.4.5). If X is uniform on S^{n-1}, then for all $v \in S^{n-1}$ and $t \geq 0$,

$$\mathbb{P}\{\langle X, v\rangle \geq t\} \leq 2e^{-nt^2/2}, \qquad \|X\|_{\psi_2} \leq \frac{C}{\sqrt{n}}.$$

Remark 3.7 (Isotropy and subgaussian marginals). Isotropy controls second moments in every direction. Subgaussianity controls all moments and tails in every direction. An isotropic vector can still put a small amount of mass far from the origin.

Example 3.8 (An isotropic coordinate distribution). Let J be uniform on $\{1, \dots, n\}$, let ε be an independent Rademacher sign, and set $X = \varepsilon\sqrt{n}\, e_J$. Then $\mathbb{E}XX^T = I_n$ and $\|X\|_2 = \sqrt{n}$ deterministically, while $|\langle X, e_1\rangle| = \sqrt{n}$ with probability $1/n$. Its subgaussian norm therefore grows with n, despite the fixed norm and isotropy.

3.5 Application: Grothendieck inequality and semidefinite programming

Grothendieck's inequality compares a discrete sign optimization problem with a vector relaxation. In the real case, if

$$\left|\sum_{i,j} a_{ij} x_i y_j\right| \le 1 \quad \text{for all } x_i, y_j \in \{\pm 1\},$$

then

$$\left|\sum_{i,j} a_{ij} \langle u_i, v_j \rangle\right| \le K_G$$

for all unit vectors u_i, v_j in any real Hilbert space, where one may take $K_G \le 1.783$.

Definition 3.9 (Semidefinite program; Book Definition 3.5.4). A semidefinite program optimizes a linear function of a symmetric matrix subject to linear constraints and the cone constraint $Z \succeq 0$. A typical vector relaxation becomes

$$\max\{\langle A, Z\rangle : Z \succeq 0,\ Z_{ii} = 1\},$$

because every feasible Z is a Gram matrix $Z_{ij} = \langle u_i, u_j \rangle$ of unit vectors.

Theorem 3.10 (Positive semidefinite Grothendieck relaxation; Book Theorem 3.5.7). For an $n \times n$ symmetric positive semidefinite matrix A, let

$$\operatorname{int}(A) = \max_{x \in \{\pm 1\}^n} x^T A x, \qquad \operatorname{sdp}(A) = \max_{\substack{Z \succeq 0 \\ Z_{ii} = 1}} \langle A, Z \rangle .$$

Then

$$\operatorname{int}(A) \le \operatorname{sdp}(A) \le 2K \operatorname{int}(A), \qquad K \le 1.783.$$

The semidefinite relaxation replaces signs $x_i \in \{\pm 1\}$ by unit vectors. Gaussian hyperplane rounding converts a solution of this relaxation into signs.

3.6 Application: Maximum cut for graphs

Let G be a finite simple undirected graph with adjacency matrix A. For a sign vector x encoding the two sides of a partition,

$$\operatorname{cut}(G, x) = \frac{1}{4} \sum_{i,j} A_{ij}(1 - x_i x_j).$$

A uniformly random cut separates each edge with probability $1/2$, giving a $1/2$ approximation in expectation.

The semidefinite relaxation replaces the signs by unit vectors:

$$\operatorname{sdp}(G) = \max_{\substack{u_1,\dots,u_n \in \mathbb{R}^n \\ \|u_i\|_2 = 1}} \frac{1}{4} \sum_{i,j} A_{ij}(1 - \langle u_i, u_j \rangle).$$

In terms of the Gram matrix Z, this maximizes $\frac{1}{4}\sum_{i,j} A_{ij}(1 - Z_{ij})$ subject to $Z \succeq 0$ and $Z_{ii} = 1$. For an optimal family u_i, sample $g \sim N(0, I_n)$ and set $x_i = \operatorname{sign}\langle g, u_i \rangle$. For unit vectors u, v, the key identity is

$$\mathbb{E}\big[\operatorname{sign}\langle g, u\rangle \operatorname{sign}\langle g, v\rangle\big] = \frac{2}{\pi} \arcsin\langle u, v\rangle .$$

Theorem 3.11 (Gaussian hyperplane rounding; Book Theorem 3.6.4). The semidefinite relaxation followed by Gaussian hyperplane rounding produces a cut whose expected value is at least 0.878 times the SDP optimum, and hence at least 0.878 times the maximum cut.

Proof sketch. The contribution of an edge to the semidefinite objective depends on $\langle u_i, u_j \rangle$. Hyperplane rounding separates the two vectors with probability equal to their angle θ divided by π. The inequality

$$\frac{\theta}{\pi} \geq 0.878 \, \frac{1 - \cos\theta}{2} \qquad (0 \leq \theta \leq \pi)$$

bounds this probability below by 0.878 times the corresponding contribution to the relaxation. Summing over edges gives the result.

3.7 Kernel trick and tightening Grothendieck inequality

3.7.1 Tensor powers and analytic functions

For tensor powers,

$$\left\langle u^{\otimes k}, v^{\otimes k} \right\rangle = \langle u, v \rangle^k .$$

For a power series $f(z) = \sum_{k \geq 0} a_k z^k$ with $\sum_{k \geq 0} |a_k| < \infty$, form two maps into the Hilbert direct sum of the tensor spaces:

$$\Phi(u) = \bigoplus_{k \geq 0} \sqrt{|a_k|}\, u^{\otimes k}, \qquad \Psi(v) = \bigoplus_{k \geq 0} \operatorname{sign}(a_k) \sqrt{|a_k|}\, v^{\otimes k}.$$

Here $u^{\otimes 0} = 1$. The coefficient condition makes these vectors well defined for $u, v \in B_2^n$, and the tensor identity gives

$$\langle \Phi(u), \Psi(v) \rangle = f(\langle u, v \rangle).$$

For every unit vector u,

$$\|\Phi(u)\|^2 = \|\Psi(u)\|^2 = \sum_{k \geq 0} |a_k|.$$

If the power series converges for every real z, the same construction is defined for all $u, v \in \mathbb{R}^n$, as in Book Lemma 3.7.7. Thus a nonlinear function of an inner product can be represented as an inner product in a larger Hilbert space, with the vector lengths determined by the coefficients.

3.7.2 Proof of Grothendieck inequality

Grothendieck inequality. Set $c = \log(1+\sqrt{2})$ and use $f(z) = \sin(cz)$. The sum of the absolute Taylor coefficients is

$$c + \frac{c^3}{3!} + \frac{c^5}{5!} + \cdots = \sinh c = 1.$$

The preceding construction therefore maps unit vectors u_i, v_j to unit vectors u_i', v_j' satisfying $\left\langle u_i', v_j' \right\rangle = \sin(c \langle u_i, v_j \rangle)$. Their finite family spans a finite dimensional Hilbert space. Let g be standard Gaussian on that space and round by its projections. Since $c < \pi/2$, the Gaussian rounding identity gives

$$\mathbb{E}\big[\operatorname{sign}\langle g, u_i'\rangle \operatorname{sign}\left\langle g, v_j'\right\rangle\big] = \frac{2c}{\pi}\langle u_i, v_j\rangle .$$

The assumption on the sign objective then implies

$$\frac{2c}{\pi}\left|\sum_{i,j} a_{ij}\langle u_i, v_j\rangle\right| = \left|\mathbb{E}\sum_{i,j} a_{ij}\operatorname{sign}\langle g, u_i'\rangle \operatorname{sign}\left\langle g, v_j'\right\rangle\right| \le 1.$$

Thus $K_G \le \pi/(2\log(1+\sqrt{2})) < 1.783$.

3.7.3 Kernels and feature maps

Definition 3.12 (Positive semidefinite kernel; Book Section 3.7.3). A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semidefinite kernel if, for every finite family $u_1, \dots, u_N \in \mathcal{X}$, the matrix $(K(u_i, u_j))_{i,j=1}^N$ is symmetric and positive semidefinite.

Theorem 3.13 (Moore–Aronszajn; Book Section 3.7.3). A kernel K is positive semidefinite if and only if there are a real Hilbert space H and a map $\Phi : \mathcal{X} \to H$ such that

$$K(u, v) = \langle \Phi(u), \Phi(v) \rangle \qquad (u, v \in \mathcal{X}).$$

The map Φ is called a feature map. The associated Hilbert space can be realized as a reproducing kernel Hilbert space.

Example 3.14 (Gaussian and polynomial kernels). For $u, v \in \mathbb{R}^n$, the functions

$$K(u, v) = \exp\left(-\frac{\|u - v\|_2^2}{2\sigma^2}\right), \qquad K(u, v) = (\langle u, v \rangle + r)^k$$

are positive semidefinite kernels when $\sigma, r > 0$ and k is a positive integer. Evaluating $K(u, v)$ gives the inner product of the feature vectors directly. This allows methods expressed through inner products to use nonlinear features.

4 Random Matrices

4.1 A quick refresher on linear algebra

4.1.1 Singular values and variational principles

Theorem 4.1 (Singular value decomposition; Book Theorem 4.1.1). Every real $m \times n$ matrix admits a decomposition

$$A = \sum_{i=1}^{\min(m,n)} s_i u_i v_i^T, \qquad s_1 \geq s_2 \geq \cdots \geq 0,$$

where the u_i and v_i are orthonormal systems. The s_i are the singular values of A; we set $s_i(A) = 0$ for $i > \min(m, n)$.

For a symmetric matrix, Courant–Fischer characterizes each eigenvalue by a max–min and a min–max over subspaces. The corresponding formula for singular values uses $\|Ax\|_2$:

$$s_k(A) = \max_{\dim E = k} \min_{x \in E \cap S^{n-1}} \|Ax\|_2 .$$

Thus a bound on $\|Ax\|_2$ over a subspace gives a bound on the corresponding singular value.

Definition 4.2 (Operator and Frobenius norms; Book Definition 4.1.8).

$$\|A\| = \max_{\|x\|_2 = 1} \|Ax\|_2 = \max_{\substack{\|x\|_2 = 1 \\ \|y\|_2 = 1}} |y^T A x|, \qquad \|A\|_F = \left(\sum_i s_i(A)^2 \right)^{1/2}.$$

Both norms are invariant under left and right multiplication by orthogonal matrices, and $\|A\| = s_1(A)$.

Theorem 4.3 (Best low-rank approximation; Book Theorem 4.1.13). For $1 \leq k \leq \min(m, n)$, the truncated SVD $A_k = \sum_{i \leq k} s_i u_i v_i^T$ is a best approximation in operator norm among matrices of rank at most k, and

$$\min_{\operatorname{rank}(B) \leq k} \|A - B\| = s_{k+1}(A).$$

4.1.2 Perturbation and approximate isometries

Weyl's inequality gives $|s_k(A) - s_k(B)| \le \|A - B\|$. Davis–Kahan adds eigenvector stability: if a target eigenvalue is separated from the rest by a gap δ, then the angle between the corresponding eigendirections is at most a constant times $\|A - B\| / \delta$.

For an $m \times n$ matrix A with $m \ge n$ and a number $\varepsilon \ge 0$, the following three statements are equivalent:

$$\left\|A^T A - I\right\| \le \varepsilon, \qquad |\, \|Ax\|_2^2 - \|x\|_2^2 \,| \le \varepsilon \, \|x\|_2^2 \text{ for all } x,$$

and every squared singular value of A belongs to $[1 - \varepsilon, 1 + \varepsilon]$.

4.2 Nets, covering, and packing

Definition 4.4 (Nets and covering numbers; Book Definitions 4.2.1–4.2.2). For $\varepsilon > 0$, an ε-net of a metric set (K, d) is a subset $\mathcal{N} \subset K$ such that every point of K lies within distance ε of some point of $\mathcal{N}$. The smallest possible cardinality is the covering number $N(K, d, \varepsilon)$.

Definition 4.5 (Packing number; Book Definition 4.2.4). An ε-packing is a subset of K whose distinct points have distance strictly greater than ε. Its largest possible cardinality is $P(K, d, \varepsilon)$.

A maximal separated set is automatically a net, and therefore

$$P(K, d, 2\varepsilon) \le N(K, d, \varepsilon) \le P(K, d, \varepsilon).$$

Volume comparison yields the fundamental Euclidean estimate

$$\varepsilon^{-n} \le N(B_2^n, \varepsilon) \le \left(1 + \frac{2}{\varepsilon}\right)^n,$$

and the upper bound also holds for S^{n-1}.

Volumetric net argument. Choose a maximal ε-separated set. The $\varepsilon/2$ balls centered at its points are disjoint, while slightly enlarged balls all lie inside a dilation of the original body. Comparing volumes bounds the number of centers. Maximality then supplies the covering property.

4.3 Application: error correcting codes

The logarithm of a covering number is the number of bits needed to describe a point to a prescribed resolution. In the Hamming cube, packing becomes error correction. If codewords are separated by more than $2r$, a decoder that chooses the nearest codeword can correct any r changed bits.

If $C(K, d, \varepsilon)$ denotes the minimum number of bits needed to encode every point of K within error ε, then Book Proposition 4.3.1 gives

$$\log_2 N(K, d, \varepsilon) \le C(K, d, \varepsilon) \le \lceil \log_2 N(K, d, \varepsilon/2) \rceil .$$

If k information bits are encoded into n bits, a suitable packing exists when

$$n - k \ge 2r \log_2\left(\frac{en}{2r}\right).$$

The proof constructs a maximal packing in Hamming distance and bounds its cardinality using the corresponding covering number.

4.4 Upper bounds on subgaussian random matrices

For an ε-net $\mathcal{N}$ of the sphere with $0 \le \varepsilon < 1$,

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

Using nets on both domain and range, with $0 \le \varepsilon < 1/2$, gives the bilinear version

$$\|A\| \le \frac{1}{1-2\varepsilon} \sup_{x \in \mathcal{N}_n,\, y \in \mathcal{N}_m} |y^T A x|.$$

Theorem 4.6 (Norm of a subgaussian matrix; Book Theorem 4.4.3). Let A be an $m \times n$ matrix with independent, centered, subgaussian entries, and assume $\max_{i,j} \|A_{ij}\|_{\psi_2} \le K$. For every $t \ge 0$, with probability at least $1 - 2e^{-t^2}$,

$$\|A\| \le CK(\sqrt{m} + \sqrt{n} + t).$$

Proof sketch. Fix unit vectors x, y. The bilinear form $y^T A x$ is a weighted sum of independent centered subgaussian random variables and therefore has a subgaussian tail. Choose nets of the two spheres with a fixed radius, such as $1/4$. A union bound over all pairs of net points then bounds the maximum bilinear form, and the net lemma bounds the operator norm. The cardinalities of the two nets account for the terms $\sqrt{n}$ and $\sqrt{m}$.

The same argument, after splitting the upper and lower triangular parts, yields $\|A\| \lesssim K(\sqrt{n} + t)$ for symmetric matrices with independent, centered, subgaussian entries on and above the diagonal. For every $t \ge 0$, the bound holds with probability at least $1 - 4e^{-t^2}$.

4.5 Application: community detection in networks

For even n, the stochastic block model $G(n,p,q)$ splits the vertices into two communities of size $n/2$. Each pair of vertices is connected independently, with probability p within a community and q between communities. Self-loops are included with probability p. The vector of all ones and the community label vector, with entries $+1$ and -1, are eigenvectors of the expected adjacency matrix and span an invariant subspace of dimension two. The observed adjacency matrix differs from its expectation by a random perturbation.

The spectral algorithm computes the eigenvector for the second largest eigenvalue of the observed adjacency matrix and rounds its signs. The bound on the norm of the random matrix controls the perturbation, and Davis–Kahan bounds the angle between the empirical and population eigenvectors. In the dense regime, if $G \sim G(n,p,q)$ and $\mu = \min(q, p-q) > 0$, the book's theorem bounds the number of misclassified vertices by C/μ^2 with failure probability at most $4e^{-n}$.

The error bound depends on the ratio of the perturbation norm to the gap separating the informative population eigenvalue from the rest of the spectrum.

4.6 Two-sided bounds on subgaussian matrices

Theorem 4.7 (Bounds on the singular values; Book Theorem 4.6.1). Let the rows of an $m \times n$ matrix A be independent, centered, isotropic, subgaussian vectors, and let their vector ψ_2 norms be at most K. For every $t \geq 0$, with probability at least $1 - 2e^{-t^2}$,

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t).$$

For a fixed $x \in S^{n-1}$,

$$\|Ax\|_2^2 - m = \sum_{i=1}^{m} \big(\langle A_i, x \rangle^2 - 1 \big).$$

Each summand is centered and subexponential, so Bernstein's inequality bounds the deviation for a fixed x. A net and the characterization of approximate isometries extend this bound to every unit vector. When the lower bound is positive, A is injective; when the error is small relative to $\sqrt{m}$, its condition number is close to one.

4.7 Application: covariance estimation and clustering

Let $X_1, \dots, X_m$ be independent copies of a random vector X, and write $\Sigma = \mathbb{E} X X^T$. The empirical second moment matrix is

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

where A is the data matrix. Whitening reduces the problem to controlling $m^{-1} A^T A - I$.

Theorem 4.8 (Subgaussian covariance estimation; Book Theorem 4.7.1). Suppose that for some $K \geq 1$, $\|\langle X, x\rangle\|_{\psi_2} \leq K \|\langle X, x\rangle\|_{L^2}$ for every $x \in \mathbb{R}^n$. Then

$$\mathbb{E} \|\Sigma_m - \Sigma\| \leq CK^2 \left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right) \|\Sigma\| .$$

For fixed K and $0 < \varepsilon < 1$, a sample size of order n/ε^2 therefore gives expected operator norm error at most $\varepsilon \|\Sigma\|$.

For the symmetric Gaussian mixture

$$X = \theta\mu + g, \qquad \theta \in \{\pm 1\}, \quad g \sim N(0, I_n),$$

where θ is a Rademacher variable independent of g, the population covariance is $I + \mu\mu^T$. Its top eigenvector points in the cluster direction. Empirical covariance concentration and eigenvector perturbation therefore yield a spectral clustering method. Book Theorem 4.7.5 states that if $m \geq Cn$ and $\|\mu\|_2 \geq C$, the method misclassifies at most one percent of the samples with probability at least 0.99.

5 Concentration Without Independence

5.1 Concentration of Lipschitz functions on the sphere

Definition 5.1 (Lipschitz constant; Book Definition 5.1.1). A map $f : (X, d_X) \to (Y, d_Y)$ is L-Lipschitz if

$$d_Y(f(x), f(y)) \le L d_X(x, y)$$

for all x, y. The smallest admissible L is $\|f\|_{\mathrm{Lip}}$.

Theorem 5.2 (Sphere concentration; Book Theorem 5.1.3). Let X be uniform on the sphere $\sqrt{n}S^{n-1}$, and let $f : \sqrt{n}S^{n-1} \to \mathbb{R}$ be Lipschitz. Then

$$\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \le C \|f\|_{\mathrm{Lip}}.$$

Consequently, for every $t \ge 0$,

$$\mathbb{P}\{|f(X) - \mathbb{E}f(X)| \ge t\} \le 2\exp\left(-\frac{ct^2}{\|f\|_{\mathrm{Lip}}^2}\right).$$

Theorem 5.3 (Isoperimetric inequalities; Book Theorems 5.1.4–5.1.5). Among measurable subsets of Euclidean space with fixed volume, balls minimize the volume of metric neighborhoods. Among subsets of a sphere with fixed surface measure, spherical caps minimize the surface measure of geodesic neighborhoods.

The isoperimetric inequality bounds the measure of neighborhoods. If a set $A \subset \sqrt{n}S^{n-1}$ has normalized area at least $1/2$, then its t-neighborhood has area at least $1 - 2e^{-ct^2}$ for every $t \ge 0$.

Concentration about a median. Write $L = \|f\|_{\mathrm{Lip}}$, let M be a median of $f(X)$, and set $A = \{f \le M\}$. The set A has measure at least one half. If x lies within distance t/L of A, the Lipschitz condition gives $f(x) \le M + t$. The bound on the measure of this neighborhood therefore bounds the upper tail. Applying the same argument to $-f$ gives the lower tail. Integrating these bounds controls the distance between the mean and the median.

The coordinates of X are dependent because $\sum_i X_i^2 = n$. The argument uses the growth of neighborhoods on the sphere to control $f(X)$.

5.2 Concentration on other metric measure spaces

For the standard Gaussian measure, half-spaces minimize the measure of neighborhoods among sets of a given measure. The same argument with a median therefore gives $\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \le C \|f\|_{\mathrm{Lip}}$ when $X \sim N(0, I_n)$ and the Lipschitz constant is taken in the Euclidean metric.

On $\{0,1\}^n$ with normalized Hamming distance, a uniform point satisfies a bound of order $\|f\|_{\mathrm{Lip}} / \sqrt{n}$. The same bound holds for a uniform random permutation with normalized Hamming distance.

On a compact connected Riemannian manifold, a positive lower bound for the Ricci curvature gives concentration at a scale proportional to the reciprocal square root of that bound. This applies to the special orthogonal group and, by projection, to Grassmannians.

For uniform points in $[0,1]^n$ and $\sqrt{n}B_2^n$, Lipschitz functions in the Euclidean metric concentrate at a scale proportional to $\|f\|_{\mathrm{Lip}}$, independently of n. More generally, if a density has the form $e^{-U(x)}$ with $\nabla^2 U(x) \succeq \kappa I$ for some $\kappa > 0$, concentration occurs at scale $\|f\|_{\mathrm{Lip}} / \sqrt{\kappa}$.

Theorem 5.4 (Talagrand convex concentration; Book Theorem 5.2.12). Let $X = (X_1, \dots, X_n)$ have independent coordinates with $|X_i| \le 1$ almost surely. Every convex Lipschitz function $f : [-1,1]^n \to \mathbb{R}$ satisfies Gaussian concentration at the scale $\|f\|_{\mathrm{Lip}}$.

For a function of independent coordinates, bounded differences give another concentration inequality. If changing coordinate i changes f by at most c_i, then for $t \ge 0$,

$$\mathbb{P}\{f(X) - \mathbb{E}f(X) \ge t\} \le \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right).$$

5.3 Application: Johnson–Lindenstrauss lemma

Theorem 5.5 (Johnson–Lindenstrauss; Book Theorem 5.3.1). Given N points in $\mathbb{R}^n$ and $0 < \varepsilon < 1$, let $1 \le m \le n$ be an integer satisfying

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

Let P be the orthogonal projection onto a uniformly random m-dimensional subspace and set $Q = \sqrt{n/m}\, P$. Then, with probability at least $1 - 2e^{-c\varepsilon^2 m}$,

$$(1-\varepsilon)\|x-y\|_2 \le \|Qx - Qy\|_2 \le (1+\varepsilon)\|x-y\|_2$$

simultaneously for all listed pairs x, y.

For one fixed vector z, rotational invariance turns a random projection of z into the first m coordinates of a random point on the sphere. Sphere concentration shows

$$\|Pz\|_2 \approx \sqrt{\frac{m}{n}}\, \|z\|_2 .$$

There are fewer than N^2 pairwise differences, so a union bound makes the estimate for one vector simultaneous. The target dimension depends only logarithmically on the number of points and not on the original ambient dimension.

5.4 Matrix Bernstein inequality

For a symmetric matrix, scalar functions such as e^X are defined spectrally. The Loewner order $X \preceq Y$ means $Y - X$ is positive semidefinite. Useful facts include

$$\|X\| \le a \iff -aI \preceq X \preceq aI,$$

and monotonicity of the largest eigenvalue and trace under Loewner order.

Theorem 5.6 (Matrix Bernstein; Book Theorem 5.4.1). Let $X_1, \dots, X_N$ be independent centered symmetric $n \times n$ random matrices with $\|X_i\| \le K$ almost surely. Define

$$\sigma^2 = \left\| \sum_{i=1}^N \mathbb{E} X_i^2 \right\|.$$

Then, for all $t \ge 0$,

$$\mathbb{P}\left\{ \left\| \sum_i X_i \right\| \ge t \right\} \le 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3} \right).$$

Trace exponential method. The trace of a matrix exponential bounds the exponential of its largest eigenvalue. Lieb's concavity theorem bounds successive conditional expectations in terms of the moment generating functions of the individual summands. Each of these is bounded by an exponential involving $\mathbb{E} X_i^2$. Choosing the Laplace parameter gives the upper tail, and applying the same argument to $-\sum_i X_i$ gives the lower tail. The factor n comes from bounding the trace by n times the largest eigenvalue.

The supporting analytic statements are the Golden–Thompson inequality $\operatorname{tr} e^{A+B} \le \operatorname{tr}(e^A e^B)$ and Lieb's concavity of $X \mapsto \operatorname{tr} \exp(H + \log X)$ for fixed symmetric H and positive definite X.

For fixed symmetric matrices A_i and independent Rademacher signs, matrix Hoeffding gives, for every $t \ge 0$,

$$\mathbb{P}\left\{ \left\| \sum_i \varepsilon_i A_i \right\| \ge t \right\} \le 2n \exp\left(-\frac{t^2}{2 \left\| \sum_i A_i^2 \right\|} \right),$$

while matrix Khintchine gives, for every $1 \le p < \infty$,

$$\left(\mathbb{E} \left\| \sum_i \varepsilon_i A_i \right\|^p \right)^{1/p} \le C \sqrt{p + \log n} \left\| \left(\sum_i A_i^2 \right)^{1/2} \right\|.$$

Hermitian dilation extends the statements to rectangular matrices.

5.5 Application: community detection in sparse networks

For a network whose expected degree is logarithmic in n, matrix Bernstein bounds the perturbation using the variances of the entries of the adjacency matrix.

For the sparse stochastic block model $p = a/n$, $q = b/n$ with $b < a < 3b$, the book proves that if

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

spectral clustering achieves at least 99 percent accuracy with probability at least 0.99. Matrix Bernstein bounds the difference between the adjacency matrix and its expectation. Davis–Kahan then bounds the change in the informative eigenvector, from which the number of misclassified vertices is estimated.

5.6 Application: covariance estimation for general distributions

Matrix Bernstein also applies to bounded samples from general distributions. Let $X \in \mathbb{R}^n$ with $n \geq 2$. If, for some $K \geq 1$,

$$\|X\|_2 \leq K(\mathbb{E}\,\|X\|_2^2)^{1/2}$$

almost surely, the empirical second moment matrix formed from m independent copies of X satisfies

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

The required sample size contains an additional factor $\log n$ compared with the subgaussian bound. Effective rank and stable rank can refine the dimension dependence by accounting for the spectrum of the covariance matrix.

6 Quadratic Forms, Symmetrization, and Contraction

6.1 Decoupling

Write a quadratic form as

$$X^T A X = \sum_{i,j} a_{ij} X_i X_j.$$

Even when the coordinates X_i are independent, the products in this sum are not. In the decoupled form $X^T A X'$, the vector X' is an independent copy of X. Conditional on either vector, the expression is a linear form in the other.

Theorem 6.1 (Decoupling; Book Theorem 6.1.1). Let A have zero diagonal, let X have independent centered coordinates, and let X' be an independent copy. For every convex function F for which the expectations exist,

$$\mathbb{E} F(X^T A X) \le \mathbb{E} F(4 X^T A X').$$

Proof sketch. Choose I by including each coordinate independently with probability $1/2$, using randomness independent of X. Keep terms whose first index lies in I and whose second index lies in I^c. Averaging over the selector gives one quarter of the original off-diagonal form, so Jensen's inequality reduces the comparison to this selected sum. Independence allows the I^c block to be replaced by an independent copy without changing its law. Centering makes the omitted terms conditionally mean zero, and another application of Jensen's inequality compares the selected sum with the full decoupled form.

Remark 6.2 (Diagonal terms). Terms $a_{ii} X_i^2$ involve only one coordinate. When the X_i are subgaussian, the centered diagonal terms form a sum of independent subexponential variables.

6.2 Hanson–Wright inequality

Proposition 6.3 (Norm of a subgaussian vector; Book Proposition 6.2.1). If $X \in \mathbb{R}^n$ is centered and subgaussian with $\|X\|_{\psi_2} \le K$, then, for every $t \ge 0$,

$$\mathbb{P}\{\|X\|_2 \ge C K(\sqrt{n} + t)\} \le e^{-t^2}.$$

The Frobenius and operator norms enter the bound for a quadratic form in different ways. The square of the Frobenius norm is the sum of squared singular values, whereas the operator norm is the largest singular value.

Theorem 6.4 (Hanson–Wright; Book Theorem 6.2.2). Let $X = (X_1, \dots, X_n)$ have independent centered subgaussian coordinates, and set $K = \max_i \|X_i\|_{\psi_2}$. For every $n \times n$ matrix A and $t \ge 0$,

$$\mathbb{P}\{|X^T A X - \mathbb{E} X^T A X| \ge t\} \le 2 \exp\left[-c \min\left(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|}\right)\right].$$

The transition between the two terms in the exponent occurs at the scale $t \asymp K^2 \|A\|_F^2 / \|A\|$. Below this scale the exponent is quadratic in t, and above it the exponent is linear in t.

Proof sketch. Split A into diagonal and off-diagonal parts. Bernstein controls the diagonal because $X_i^2 - \mathbb{E}X_i^2$ is subexponential. Decouple the off-diagonal part, then compare its MGF with a Gaussian bilinear form $g^T A g'$, where g and g' are independent standard Gaussian vectors in $\mathbb{R}^n$. The SVD reduces that form to a sum of independent products of Gaussian variables and gives

$$\mathbb{E}e^{\lambda g^T A g'} \le e^{\lambda^2 \|A\|_F^2} \quad \text{when } |\lambda| \le (2\|A\|)^{-1}.$$

Optimizing the Chernoff bound over the permitted values of λ produces the two terms in the minimum.

Example 6.5 (A rank-one quadratic form). For a fixed v, the statistic $X^T(vv^T)X = \langle X, v\rangle^2$ has $\left\|vv^T\right\| = \|v\|_2^2$ and $\left\|vv^T\right\|_F = \|v\|_2^2$. Hanson–Wright therefore gives a subexponential tail for the square of a subgaussian marginal.

6.3 Symmetrization

Introduce independent copies X_i' and independent Rademacher signs ε_i, independent also of the vectors. The differences $X_i - X_i'$ are symmetric, so signs can be inserted without changing their joint law.

Lemma 6.6 (Symmetrization; Book Lemma 6.3.2). For independent centered random vectors X_i in any normed space,

$$\frac{1}{2}\mathbb{E}\left\|\sum_i \varepsilon_i X_i\right\| \le \mathbb{E}\left\|\sum_i X_i\right\| \le 2\mathbb{E}\left\|\sum_i \varepsilon_i X_i\right\|.$$

For independent noncentered random vectors, the corresponding inequality is

$$\mathbb{E}\left\|\sum_i (X_i - \mathbb{E}X_i)\right\| \le 2\mathbb{E}\left\|\sum_i \varepsilon_i X_i\right\|.$$

Conditional on the vectors, the sum on the right depends only on the Rademacher signs. Its expectation can therefore be evaluated first over the signs and then over the vectors.

Proof sketch. For the upper bound, write $\mathbb{E}X_i$ as a conditional expectation of an independent copy and apply Jensen:

$$\mathbb{E}\left\|\sum_i (X_i - \mathbb{E}X_i)\right\| \le \mathbb{E}\left\|\sum_i (X_i - X_i')\right\|.$$

Insert Rademacher signs by symmetry, then use the triangle inequality to split the primed and unprimed sums. The lower comparison follows by conditioning and Jensen in the opposite direction.

6.4 Random matrices with non-i.i.d. entries

For matrices whose entries have different distributions, the maximum Euclidean norm of the rows gives a bound on the expected operator norm.

Theorem 6.7 (Matrices with nonidentically distributed entries; Book Theorem 6.4.1). Let A be an $n \times n$ symmetric matrix with $n \geq 2$ and independent centered entries on and above the diagonal. If A_i denotes row i, then

$$\mathbb{E} \max_i \|A_i\|_2 \leq \mathbb{E} \|A\| \leq C\sqrt{\log n}\, \mathbb{E} \max_i \|A_i\|_2 .$$

The displayed expectations are assumed to exist.

The lower bound follows from $\|A\| \geq \|Ae_i\|_2$. For the upper bound, decompose the matrix entry by entry, symmetrize, and apply matrix Khintchine. Thus the expected maximum row norm determines the expected operator norm up to a factor $\sqrt{\log n}$.

6.5 Application: matrix completion

Let X be an unknown rank-r $n \times n$ matrix with $n \geq 2$. Each entry is observed independently with probability $0 < p = m/n^2 \leq 1$, producing $Y_{ij} = \delta_{ij} X_{ij}$. Setting unobserved entries to zero and multiplying by p^{-1} gives the unbiased estimate $p^{-1}Y$. The estimator $\widehat{X}$ is its best rank-r approximation.

Theorem 6.8 (Spectral matrix completion; Book Theorem 6.5.1). If $m \geq n \log n$, then

$$\frac{1}{n} \mathbb{E} \left\| \widehat{X} - X \right\|_F \leq C \sqrt{\frac{rn \log n}{m}}\, \|X\|_\infty ,$$

where $\|X\|_\infty = \max_{i,j} |X_{ij}|$.

The sampling error $p^{-1}Y - X$ has independent entries with possibly different distributions. Hermitian dilation allows the preceding theorem to be applied to this error matrix and bounds its operator norm by the maximum row and column norms. Since both X and $\widehat{X}$ have rank at most r, the Frobenius norm of their difference is at most a constant times $\sqrt{r}$ times its operator norm. The quantity $n^{-1} \left\| \widehat{X} - X \right\|_F$ is the root mean square error per entry.

6.6 Contraction principle

Theorem 6.9 (Rademacher contraction; Book Theorem 6.6.1). For fixed vectors $x_1, \dots, x_N$ in a normed space, scalars a_i, and independent Rademacher signs,

$$\mathbb{E}\left\|\sum_i a_i \varepsilon_i x_i\right\| \le \|a\|_\infty \, \mathbb{E}\left\|\sum_i \varepsilon_i x_i\right\|.$$

After scaling so $\|a\|_\infty \le 1$, view the expectation as a convex function of $a \in [-1,1]^N$. A convex function on a cube reaches its maximum at a vertex. At a vertex, every a_i is a sign and can be absorbed into ε_i.

For independent centered random vectors $X_1, \dots, X_N$ in a normed space with $N \ge 2$, let g_i be independent standard Gaussians, independent also of the X_i. Gaussian symmetrization gives the comparison

$$\frac{c}{\sqrt{\log N}} \mathbb{E}\left\|\sum_i g_i X_i\right\| \le \mathbb{E}\left\|\sum_i X_i\right\| \le 3\mathbb{E}\left\|\sum_i g_i X_i\right\|.$$

The factor $\sqrt{\log N}$ comes from the expected maximum of the absolute Gaussian coefficients and is optimal in general.

7 Random Processes

7.1 Basic concepts and examples

Definition 7.1 (Random process; Book Definition 7.1.1). A random process indexed by a set T is a collection

$$(X_t)_{t\in T}$$

of random variables defined on the same probability space. The index need not represent time: in the high-dimensional setting, T is often a subset of $\mathbb{R}^n$ or a class of vectors, matrices, or functions.

Example 7.2 (Basic examples; Book Examples 7.1.2–7.1.5).

- If $T = \{1, \dots, n\}$, the process is simply the random vector $(X_1, \dots, X_n)$.
- A random walk has
$$X_k = \sum_{i=1}^{k} Z_i,$$
where the increments Z_i are independent and centered.
- Standard Brownian motion $(X_t)_{t\geq 0}$ starts at $X_0 = 0$, has almost surely continuous paths and independent increments, and satisfies $X_t - X_s \sim N(0, t-s)$ for $t \geq s$.
- When $T \subset \mathbb{R}^n$, the process is often called a random field; a spatially indexed measurement is a typical example.

7.1.1 Covariance and increments

Assume that $\mathbb{E}X_t = 0$. A process with nonzero means can be centered by replacing X_t with $X_t - \mathbb{E}X_t$.

Definition 7.3 (Covariance and canonical pseudometric; Book Section 7.1.1). The covariance function and the increments of $(X_t)_{t\in T}$ are

$$\Sigma(t,s) := \operatorname{Cov}(X_t, X_s) = \mathbb{E}X_tX_s, \qquad d(t,s) := \|X_t - X_s\|_{L_2} = \big(\mathbb{E}(X_t - X_s)^2\big)^{1/2}.$$

The function d is called the canonical metric, although it is more precisely a pseudometric: distinct indices can satisfy $d(t,s) = 0$.

Expanding the square gives

$$d(t,s)^2 = \Sigma(t,t) - 2\Sigma(t,s) + \Sigma(s,s). \tag{7.1}$$

If the zero random variable occurs as one of the process coordinates, then the covariance follows by polarization:

$$\Sigma(t,s) = \frac{1}{2}\big(d(t,0)^2 + d(s,0)^2 - d(t,s)^2\big).$$

The zero coordinate fixes the individual variances as $\Sigma(t,t) = d(t,0)^2$; increments alone determine only differences between coordinates.

Example 7.4 (Canonical metrics for Brownian motion and random walks). For Brownian motion,

$$d(t,s) = \sqrt{t-s}, \qquad t \geq s.$$

For the random walk above with $\mathbb{E}Z_i^2 = 1$,

$$d(k,\ell) = \sqrt{k-\ell}, \qquad k \geq \ell.$$

Thus the canonical geometry can be very different from the ambient Euclidean geometry of the index set.

7.1.2 Gaussian processes

Definition 7.5 (Gaussian process; Book Definition 7.1.9). A process $(X_t)_{t\in T}$ is Gaussian if every finite subcollection $(X_t)_{t\in T_0}$, $T_0 \subset T$, has a joint normal distribution. Equivalently, every finite linear combination $\sum_{t\in T_0} a_t X_t$ is normal.

A centered Gaussian process is determined in distribution by its covariance function. If the zero variable belongs to the process, equation (7.1) shows that its canonical pseudometric determines the distribution as well.

Theorem 7.6 (Concentration of Gaussian processes; Book Theorem 7.1.11). Let $(X_t)_{t\in T}$ be a Gaussian process with finite T. Then

$$\left\| \sup_{t\in T} X_t - \mathbb{E}\sup_{t\in T} X_t \right\|_{\psi_2} \leq C \sup_{t\in T} \sqrt{\operatorname{Var}(X_t)}.$$

Proof sketch. Represent the finite Gaussian vector $(X_t)_{t\in T}$ as $\mu + Ag$, where $\mu_t = \mathbb{E}X_t$ and g is a standard Gaussian vector. The map

$$g \longmapsto \max_{t\in T}(\mu_t + \langle a_t, g\rangle),$$

where a_t is the row of A corresponding to t, is Lipschitz with constant $\max_t \|a_t\|_2 = \max_t \sqrt{\operatorname{Var}(X_t)}$. Gaussian concentration gives the claim.

Definition 7.7 (Canonical Gaussian process; Book equation (7.2)). For $T \subset \mathbb{R}^n$ and $g \sim N(0, I_n)$, define

$$X_t := \langle g, t\rangle, \qquad t \in T.$$

Then

$$\|X_t - X_s\|_{L_2} = \|t-s\|_2,$$

so the canonical pseudometric is exactly Euclidean distance on T.

Lemma 7.8 (Finite Gaussian processes are canonical; Book Lemma 7.1.12). If $X \sim N(0, \Sigma)$ in $\mathbb{R}^n$, there are points $t_1, \dots, t_n \in \mathbb{R}^n$ such that

$$X \overset{d}{=} (\langle g, t_i \rangle)_{i=1}^n, \qquad g \sim N(0, I_n).$$

One may take t_i to be the rows of $\Sigma^{1/2}$.

Remark 7.9 (Expected supremum on a general index set). For a general index set, define

$$\mathbb{E} \sup_{t \in T} X_t := \sup_{T_0 \subset T,\, T_0 \text{ finite}} \mathbb{E} \max_{t \in T_0} X_t.$$

This agrees with the expected supremum for the usual separable processes.

7.2 Slepian, Sudakov–Fernique, and Gordon inequalities

The central question is to estimate $\mathbb{E} \sup_{t \in T} X_t$. For Brownian motion, the reflection principle gives the exact identity

$$\mathbb{E} \sup_{0 \le t \le t_0} X_t = \sqrt{\frac{2t_0}{\pi}},$$

but a general Gaussian process rarely admits a direct calculation. Comparison inequalities replace it by a simpler process with larger or smaller increments.

Theorem 7.10 (Slepian inequality; Book Theorem 7.2.2). Let $(X_t)_{t \in T}$ and $(Y_t)_{t \in T}$ be centered Gaussian processes. Suppose that, for all $s, t \in T$,

$$\mathbb{E} X_t^2 = \mathbb{E} Y_t^2, \qquad \mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2.$$

Then, for every $\tau \in \mathbb{R}$,

$$\mathbb{P}\left\{ \sup_{t \in T} X_t \ge \tau \right\} \le \mathbb{P}\left\{ \sup_{t \in T} Y_t \ge \tau \right\},$$

and consequently

$$\mathbb{E} \sup_{t \in T} X_t \le \mathbb{E} \sup_{t \in T} Y_t.$$

7.2.1 Gaussian interpolation

For independent centered Gaussian vectors $X \sim N(0, \Sigma^X)$ and $Y \sim N(0, \Sigma^Y)$, interpolate by

$$Z(u) := \sqrt{u}\, X + \sqrt{1-u}\, Y, \qquad 0 \le u \le 1. \tag{7.2}$$

Its covariance is a linear function of u:

$$\Sigma^{Z(u)} = u\Sigma^X + (1-u)\Sigma^Y.$$

Lemma 7.11 (Gaussian integration by parts; Book Lemmas 7.2.3–7.2.4). If $G \sim N(0,1)$ and the expectations exist, then

$$\mathbb{E}[Gf(G)] = \mathbb{E}f'(G).$$

More generally, if $X \sim N(0, \Sigma)$ in $\mathbb{R}^n$, then

$$\mathbb{E}[X_i f(X)] = \sum_{j=1}^n \Sigma_{ij}\, \mathbb{E}\frac{\partial f}{\partial x_j}(X), \qquad i = 1, \dots, n.$$

Lemma 7.12 (Gaussian interpolation formula; Book Lemma 7.2.5). For $Z(u)$ in (7.2) and twice differentiable $f : \mathbb{R}^n \to \mathbb{R}$,

$$\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}f(Z(u)) = \frac{1}{2}\sum_{i,j=1}^n (\Sigma^X_{ij} - \Sigma^Y_{ij})\mathbb{E}\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(u)),$$

provided all displayed expectations are finite.

Proof sketch. Differentiate $f(Z(u))$ by the chain rule. Terms containing X_i and Y_i come with factors $1/(2\sqrt{u})$ and $-1/(2\sqrt{1-u})$. Condition on the other vector and apply multivariate Gaussian integration by parts. The square roots cancel, giving the displayed sum of covariance differences and expected second derivatives.

7.2.2 Proof of Slepian inequality

Lemma 7.13 (Functional Slepian inequality; Book Lemma 7.2.6). Let X, Y be centered Gaussian vectors in $\mathbb{R}^n$ with equal coordinate variances and with

$$\mathbb{E}(X_i - X_j)^2 \le \mathbb{E}(Y_i - Y_j)^2 \qquad \text{for all } i, j.$$

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable and

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \ge 0 \qquad (i \ne j),$$

then $\mathbb{E}f(X) \ge \mathbb{E}f(Y)$ whenever both expectations exist.

Slepian inequality. Equal variances and increment domination imply $\Sigma^X_{ii} = \Sigma^Y_{ii}$ and $\Sigma^X_{ij} \geq \Sigma^Y_{ij}$. The interpolation derivative is therefore nonnegative for functions with nonnegative mixed second derivatives. To approximate the event $\{\max_i x_i < \tau\}$, take a smooth nonincreasing $h : \mathbb{R} \to [0,1]$ and

$$f(x) = \prod_{i=1}^{n} h(x_i).$$

For $i \neq j$,

$$\partial_{ij} f(x) = h'(x_i)h'(x_j) \prod_{k \notin \{i,j\}} h(x_k) \geq 0.$$

Thus $\mathbb{E}f(X) \geq \mathbb{E}f(Y)$. Passing to the indicator limit yields the stochastic comparison. Integrating the tails gives the comparison of expected suprema.

7.2.3 Sudakov–Fernique and Gordon inequalities

Theorem 7.14 (Sudakov–Fernique inequality; Book Theorem 7.2.8). Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be centered Gaussian processes. If

$$\mathbb{E}(X_t - X_s)^2 \leq \mathbb{E}(Y_t - Y_s)^2 \qquad \text{for all } s, t \in T,$$

then

$$\mathbb{E} \sup_{t\in T} X_t \leq \mathbb{E} \sup_{t\in T} Y_t.$$

Smooth approximation of the maximum. For finite $T = \{1, \dots, n\}$, approximate the maximum by

$$f_\beta(x) := \frac{1}{\beta} \log \sum_{i=1}^{n} e^{\beta x_i}, \qquad \max_i x_i \leq f_\beta(x) \leq \max_i x_i + \frac{\log n}{\beta}.$$

Writing

$$p_i(x) := \frac{e^{\beta x_i}}{\sum_k e^{\beta x_k}},$$

one has

$$\partial_i f_\beta = p_i, \qquad \partial_{ij} f_\beta = \beta(\delta_{ij} p_i - p_i p_j).$$

The interpolation identity simplifies to

$$\frac{\mathrm{d}}{\mathrm{d}u} \mathbb{E} f_\beta(Z(u)) = \frac{\beta}{4} \sum_{i \neq j} \left[\mathbb{E}(X_i - X_j)^2 - \mathbb{E}(Y_i - Y_j)^2 \right] \mathbb{E}[p_i(Z(u)) p_j(Z(u))] \leq 0.$$

Integrate from $u = 0$ to $u = 1$, and then send $\beta \to \infty$.

Theorem 7.15 (Gordon min–max comparison; Book Theorem 7.2.9). Let $(X_{ut})_{(u,t)\in U\times T}$ and $(Y_{ut})_{(u,t)\in U\times T}$ be centered Gaussian processes. Assume

$$\mathbb{E}(X_{ut}-X_{us})^2 \le \mathbb{E}(Y_{ut}-Y_{us})^2 \qquad \text{for every } u,t,s,$$

and

$$\mathbb{E}(X_{ut}-X_{vs})^2 \ge \mathbb{E}(Y_{ut}-Y_{vs})^2 \qquad \text{for } u\neq v \text{ and all } t,s.$$

Then

$$\mathbb{E}\inf_{u\in U}\sup_{t\in T} X_{ut} \le \mathbb{E}\inf_{u\in U}\sup_{t\in T} Y_{ut}.$$

If, in addition, $\mathbb{E}X_{ut}^2 = \mathbb{E}Y_{ut}^2$ for every (u,t), then for every $\tau\in\mathbb{R}$,

$$\mathbb{P}\left\{\inf_{u\in U}\sup_{t\in T} X_{ut}\ge\tau\right\} \le \mathbb{P}\left\{\inf_{u\in U}\sup_{t\in T} Y_{ut}\ge\tau\right\}.$$

Remark 7.16 (Variance and increment assumptions). Slepian's inequality compares the distributions of suprema under equal coordinate variances. Sudakov–Fernique compares their expectations using only the increment variances. Gordon's expectation comparison also uses only increment variances; its probability comparison requires equal coordinate variances. The increment comparison reverses across different outer indices because the outer operation is an infimum.

7.3 Application: sharp bounds for Gaussian matrices

Theorem 7.17 (Norm of a Gaussian matrix; Book Theorem 7.3.1). Let A be an $m\times n$ matrix with independent $N(0,1)$ entries. Then

$$\mathbb{E}\,\|A\| \le \sqrt{m}+\sqrt{n}.$$

Proof sketch. Express the operator norm as the supremum of the Gaussian process

$$X_{uv} := \langle Au, v\rangle, \qquad (u,v)\in \mathbb{S}^{n-1}\times\mathbb{S}^{m-1}.$$

For unit vectors $u,w\in\mathbb{R}^n$ and $v,z\in\mathbb{R}^m$,

$$\mathbb{E}(X_{uv}-X_{wz})^2 = \left\|uv^T - wz^T\right\|_F^2 \le \|u-w\|_2^2 + \|v-z\|_2^2.$$

Compare this with

$$Y_{uv} := \langle g,u\rangle + \langle h,v\rangle, \qquad g\sim N(0,I_n),\quad h\sim N(0,I_m),$$

whose increment variance is exactly the right-hand side. Sudakov–Fernique therefore gives

$$\mathbb{E}\,\|A\| \le \mathbb{E}\sup_{u,v} Y_{uv} = \mathbb{E}\,\|g\|_2 + \mathbb{E}\,\|h\|_2 \le \sqrt{n}+\sqrt{m}.$$

Corollary 7.18 (Gaussian matrix norm tail; Book Corollary 7.3.2). Under the same assumptions, for every $t \geq 0$,

$$\mathbb{P}\{\|A\| \geq \sqrt{m} + \sqrt{n} + t\} \leq 2e^{-ct^2}.$$

Indeed, vectorize A as a standard Gaussian vector in $\mathbb{R}^{mn}$. The map $A \mapsto \|A\|$ is 1-Lipschitz with respect to Frobenius distance because

$$|\|A\| - \|B\|| \leq \|A - B\| \leq \|A - B\|_F .$$

Gaussian concentration gives the stated tail bound.

Example 7.19 (Gaussian orthogonal ensemble and smallest singular values; Book Exercises 7.11 and 7.13). Let A be a symmetric $n \times n$ matrix whose entries on and above the diagonal are independent, with diagonal entries distributed as $N(0, 2)$ and off-diagonal entries as $N(0, 1)$. This normalization defines the Gaussian orthogonal ensemble used here. Gaussian comparison gives

$$\mathbb{E}\,\|A\| \leq 2\sqrt{n}$$

and, for every $t \geq 0$, the tail bound

$$\mathbb{P}\{\|A\| \geq 2\sqrt{n} + t\} \leq 2e^{-ct^2}.$$

For an $m \times n$ matrix B with independent $N(0, 1)$ entries and $m \geq n$, Gordon's expectation comparison gives

$$\mathbb{E}s_n(B) \geq \mathbb{E}\,\|g_m\|_2 - \mathbb{E}\,\|g_n\|_2 , \qquad g_k \sim N(0, I_k).$$

For $m \geq n \geq C$, the monotonicity of $\mathbb{E}\,\|g_k\|_2 - \sqrt{k}$ yields $\mathbb{E}s_n(B) \geq \sqrt{m} - \sqrt{n}$.

7.4 Sudakov inequality

Let $N(T, d, \varepsilon)$ denote the smallest cardinality of an ε-net in the canonical pseudometric d; equivalently, it is the smallest number of closed d-balls of radius ε that cover T. The quantity $\log N(T, d, \varepsilon)$ is the metric entropy at scale ε.

Theorem 7.20 (Sudakov inequality; Book Theorem 7.4.1). Let $(X_t)_{t\in T}$ be a centered Gaussian process with canonical pseudometric d. For every $\varepsilon \geq 0$,

$$\mathbb{E}\sup_{t\in T} X_t \geq c\varepsilon\sqrt{\log N(T, d, \varepsilon)}.$$

Proof sketch. Assume first that $N(T, d, \varepsilon) < \infty$, and choose a maximal ε-separated set $\mathcal{N} \subset T$. Maximality makes $\mathcal{N}$ an ε-net, so $|\mathcal{N}| \geq N(T, d, \varepsilon)$. On $\mathcal{N}$, compare X_t with

$$Y_t := \frac{\varepsilon}{\sqrt{2}} g_t,$$

where the g_t are independent standard normals. For distinct $s, t \in \mathcal{N}$,

$$\mathbb{E}(X_t - X_s)^2 = d(t, s)^2 \geq \varepsilon^2 = \mathbb{E}(Y_t - Y_s)^2.$$

Sudakov–Fernique and the lower bound for the maximum of independent Gaussian variables yield

$$\mathbb{E} \sup_{t \in T} X_t \geq \mathbb{E} \max_{t \in \mathcal{N}} Y_t \geq c\varepsilon \sqrt{\log |\mathcal{N}|}.$$

The case of infinite covering number follows by taking arbitrarily large finite separated subsets.

7.4.1 Applications to covering numbers in Euclidean space

Corollary 7.21 (Euclidean Sudakov inequality; Book Corollary 7.4.2). For every $T \subset \mathbb{R}^n$ and $\varepsilon > 0$,

$$\mathbb{E} \sup_{t \in T} \langle g, t \rangle \geq c\varepsilon \sqrt{\log N(T, \varepsilon)}, \qquad g \sim N(0, I_n),$$

where the covering number is computed in Euclidean distance.

Corollary 7.22 (Covering a polytope; Book Corollary 7.4.3). Let $P \subset B_2^n$ be a polytope with N vertices. Then, for every $\varepsilon > 0$,

$$N(P, \varepsilon) \leq N^{C/\varepsilon^2}.$$

Proof sketch. For a fixed g, a linear functional reaches its maximum over P at a vertex. If the vertices are $x_1, \ldots, x_N$, then

$$\mathbb{E} \sup_{x \in P} \langle g, x \rangle = \mathbb{E} \max_{i \leq N} \langle g, x_i \rangle \leq C \sqrt{\log N},$$

since $\|x_i\|_2 \leq 1$. Combine this upper bound with Euclidean Sudakov and solve for $N(P, \varepsilon)$.

Remark 7.23 (Infinite covering numbers). If $N(T, d, \varepsilon) = \infty$ for some $\varepsilon > 0$, arbitrarily large finite separated subsets give $\mathbb{E} \sup_{t \in T} X_t = \infty$.

7.5 Gaussian width

Definition 7.24 (Gaussian width; Book Definition 7.5.1). For a nonempty set $T \subset \mathbb{R}^n$, its Gaussian width is

$$w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle , \qquad g \sim N(0, I_n).$$

Proposition 7.25 (Properties of Gaussian width; Book Proposition 7.5.2). For nonempty sets $T, S \subset \mathbb{R}^n$, the following properties hold.

(a) $w(T) < \infty$ if and only if T is bounded.

(b) $w(UT + y) = w(T)$ for every orthogonal U and every $y \in \mathbb{R}^n$.

(c) $w(\operatorname{conv}(T)) = w(T)$.

(d) $w(T + S) = w(T) + w(S)$ and $w(aT) = |a| w(T)$.

(e)
$$w(T) = \frac{1}{2} w(T - T) = \frac{1}{2} \mathbb{E} \sup_{x,y \in T} \langle g, x - y \rangle .$$

(f)
$$\frac{1}{\sqrt{2\pi}} \operatorname{diam}(T) \le w(T) \le \frac{\sqrt{n}}{2} \operatorname{diam}(T).$$

(g) For every $m \times n$ matrix A,
$$w(AT) \le \|A\| \, w(T).$$

The diameter bounds. For any $x, y \in T$, symmetry of $T - T$ and the normal distribution give

$$w(T) \ge \frac{1}{2} \mathbb{E} \, |\langle g, x - y \rangle| = \frac{1}{\sqrt{2\pi}} \|x - y\|_2 .$$

Take the supremum over pairs. Conversely,

$$w(T) = \frac{1}{2} \mathbb{E} \sup_{x,y \in T} \langle g, x - y \rangle \le \frac{1}{2} \mathbb{E} \, \|g\|_2 \operatorname{diam}(T) \le \frac{\sqrt{n}}{2} \operatorname{diam}(T).$$

An interval and a Euclidean ball show that the two dimension dependences are both sharp up to absolute constants.

7.5.1 Geometric meaning of Gaussian width

For a unit vector θ, the directional width of T is the width of the smallest slab orthogonal to θ that contains T:

$$\sup_{x,y\in T} \langle \theta, x-y\rangle .$$

Thus width measures the average size of a set as viewed from a random direction.

Definition 7.26 (Spherical width; Book Definition 7.5.4). The spherical width of $T \subset \mathbb{R}^n$ is

$$w_s(T) := \mathbb{E} \sup_{x\in T} \langle \theta, x\rangle , \qquad \theta \sim \mathrm{Unif}(\mathbb{S}^{n-1}).$$

Lemma 7.27 (Gaussian versus spherical width; Book Lemma 7.5.5). For every bounded $T \subset \mathbb{R}^n$,

$$\left(\sqrt{n} - \frac{C}{\sqrt{n}}\right) w_s(T) \le w(T) \le \sqrt{n}\, w_s(T).$$

Proof sketch. Write $g = r\theta$, where $r = \|g\|_2$ and $\theta = g/\|g\|_2$. The direction is uniform on the sphere and independent of r, so

$$w(T) = \mathbb{E}\,\|g\|_2 \; w_s(T).$$

Use $\sqrt{n} - C/\sqrt{n} \le \mathbb{E}\,\|g\|_2 \le \sqrt{n}$.

7.5.2 Examples of Gaussian width

Example 7.28 (Ball, sphere, cube, and cross-polytope; Book Examples 7.5.6–7.5.8). Duality of norms gives the basic calculations

$$w(\mathbb{S}^{n-1}) = w(B_2^n) = \mathbb{E}\,\|g\|_2 = \sqrt{n} + O(n^{-1/2}),$$
$$w(B_\infty^n) = \mathbb{E}\,\|g\|_1 = \sqrt{\frac{2}{\pi}}\, n,$$
$$w(B_1^n) = \mathbb{E}\,\|g\|_\infty \asymp \sqrt{\log(2n)}.$$

The spherical widths of B_2^n and $\mathbb{S}^{n-1}$ equal 1.

Example 7.29 (Finite point sets; Book Example 7.5.9). Every finite $T \subset \mathbb{R}^n$ satisfies

$$w(T) \le C\sqrt{\log |T|}\, \mathrm{diam}(T).$$

Translate one point to the origin, rescale into a Euclidean ball, and apply the Gaussian maximal inequality to the finitely many linear functionals.

Remark 7.30 (Width of the cube and cross-polytope). For $n \geq 2$, the cube has 2^n vertices and width of order n, comparable to the width of its circumscribed Euclidean ball $\sqrt{n}B_2^n$. The cross-polytope has only $2n$ vertices and width of order $\sqrt{\log n}$. Its inscribed ball $n^{-1/2}B_2^n$ has width of order 1, whereas its circumscribed ball B_2^n has width of order $\sqrt{n}$. Thus the width of the cross-polytope is much closer to that of its inscribed ball, up to the logarithmic factor.

Example 7.31 (Matrices with bounded operator norm; Book Exercises 7.18–7.19). Under the Frobenius inner product, the nuclear norm is dual to the operator norm. Consequently, the Gaussian width of the operator norm unit ball in the space of $n \times n$ matrices is the expected nuclear norm of a standard Gaussian matrix, which is of order $n^{3/2}$.

7.5.3 Gaussian complexity and effective dimension

Definition 7.32 (Gaussian complexity; Book Section 7.5.3). For bounded $T \subset \mathbb{R}^n$, define

$$\gamma(T) := \mathbb{E} \sup_{x \in T} |\langle g, x \rangle|, \qquad h(T) := \left(\mathbb{E} \sup_{x \in T} |\langle g, x \rangle|^2 \right)^{1/2}.$$

The quantity $\gamma(T)$ is the Gaussian complexity. Directly,

$$w(T) \leq \gamma(T) \leq h(T).$$

Lemma 7.33 (Equivalent width parameters; Book Lemma 7.5.11). For every bounded $T \subset \mathbb{R}^n$,

$$\gamma(T - T) = 2w(T),$$

and, for every $y \in T$,

$$h(T) \asymp \gamma(T) \asymp w(T) + \|y\|_2.$$

In particular, if $0 \in T$, then

$$h(T) \asymp \gamma(T) \asymp w(T).$$

Proof sketch. The difference set is symmetric, so its width equals its absolute Gaussian supremum; the identity $w(T - T) = 2w(T)$ gives the first claim. For the second, the map $z \mapsto \sup_{x \in T} |\langle z, x \rangle|$ is Lipschitz with constant $\sup_{x \in T} \|x\|_2$. Gaussian concentration compares its L_2 and L_1 norms. The bound $\gamma(T) \geq \mathbb{E}|\langle g, y \rangle|$ for $y \in T$ controls the radius by $\gamma(T)$.

Gaussian width is translation invariant. The quantities $\gamma(T)$ and $h(T)$ also depend on the position of T relative to the origin.

Definition 7.34 (Effective dimension; Book Definition 7.5.12). For a bounded set T with $\operatorname{diam}(T) > 0$, the book's effective dimension is

$$d_{\text{eff}}(T) \equiv d(T) := \frac{h(T-T)^2}{\operatorname{diam}(T)^2} \asymp \frac{w(T)^2}{\operatorname{diam}(T)^2}.$$

Let $\dim(T)$ be the dimension of the smallest affine space containing T. Then

$$d_{\text{eff}}(T) \le \dim(T),$$

with equality for a Euclidean ball in a subspace.

Example 7.35 (Ellipsoids; Book Exercise 7.23). For a nonzero matrix A and the ellipsoid $A(B_2^n)$,

$$w(A(B_2^n)) \asymp \|A\|_F, \qquad d_{\text{eff}}(A(B_2^n)) = \frac{\|A\|_F^2}{\|A\|^2}.$$

Thus the effective dimension of the ellipsoid equals the stable rank of A.

7.6 Application: random projections of sets

For a finite set T, a sufficiently high-dimensional random projection behaves approximately like multiplication by $\sqrt{m/n}$. For a large or infinite set, the projected diameter also depends on spherical width. For example, the orthogonal projection of a Euclidean ball has the same diameter as the original ball.

Theorem 7.36 (Diameter of a random projection; Book Theorem 7.6.1). Let $T \subset \mathbb{R}^n$ be bounded, and let P be the orthogonal projection onto a uniformly random m-dimensional subspace of $\mathbb{R}^n$. Then

$$\mathbb{E}\operatorname{diam}(PT) \asymp w_s(T) + \sqrt{\frac{m}{n}}\operatorname{diam}(T).$$

All implicit constants are absolute.

Upper bound. Write Q for the $m \times n$ coordinate projection after a uniformly random rotation. Normalize $\operatorname{diam}(T) \le 1$. Then

$$\operatorname{diam}(QT) = \sup_{x \in T-T} \|Qx\|_2 = \sup_{x \in T-T} \sup_{z \in \mathbb{S}^{m-1}} \langle Qx, z\rangle.$$

Approximate $\mathbb{S}^{m-1}$ by a $1/2$-net $\mathcal{N}$ with $|\mathcal{N}| \le 5^m$, losing only an absolute factor. For fixed $z \in \mathcal{N}$, the vector $Q^T z$ is uniform on $\mathbb{S}^{n-1}$, hence

$$\mathbb{E}\sup_{x \in T-T} \left\langle Q^T z, x \right\rangle = w_s(T-T) = 2w_s(T).$$

Spherical concentration for this 1-Lipschitz supremum and a union bound over $\mathcal{N}$, followed by tail integration, give $\sqrt{m/n}\operatorname{diam}(T)$. Exercise 7.26 gives the lower bound.

Remark 7.37 (Projection dimension and effective dimension; Book Remark 7.6.2). Because a sum of nonnegative terms is equivalent to their maximum,

$$\mathbb{E}\,\mathrm{diam}(PT) \asymp \max\left\{w_s(T), \sqrt{\frac{m}{n}}\,\mathrm{diam}(T)\right\}.$$

The terms balance when

$$m \asymp \frac{n\,w_s(T)^2}{\mathrm{diam}(T)^2} \asymp \frac{w(T)^2}{\mathrm{diam}(T)^2} \asymp d_{\mathrm{eff}}(T).$$

Thus

$$\mathbb{E}\,\mathrm{diam}(PT) \asymp \begin{cases} \sqrt{m/n}\,\mathrm{diam}(T), & m \gtrsim d_{\mathrm{eff}}(T), \\ w_s(T), & m \lesssim d_{\mathrm{eff}}(T). \end{cases}$$

When m exceeds the effective dimension, the projected diameter has order $\sqrt{m/n}\,\mathrm{diam}(T)$. For smaller m, its order is $w_s(T)$.

Example 7.38 (Finite sets and the Euclidean ball; Book Exercise 7.22). For a finite set, $d_{\mathrm{eff}}(T) \lesssim \log|T|$, so dimensions above $\log|T|$ give projected diameter of order $\sqrt{m/n}\,\mathrm{diam}(T)$. For $T = B_2^n$, $w_s(T) = 1$ and $\mathrm{diam}(T) = 2$, consistent with the diameter of its orthogonal projection.

Example 7.39 (Gaussian projections; Book Exercise 7.25). For an $m \times n$ matrix G with independent $N(0,1)$ entries and a bounded set $T \subset \mathbb{R}^n$,

$$\mathbb{E}\,\mathrm{diam}(GT) \asymp w(T) + \sqrt{m}\,\mathrm{diam}(T).$$

Example 7.40 (Norms of projected matrices; Book Exercise 7.27). Let A be an $n \times k$ matrix, let P be the orthogonal projection onto a uniformly random m-dimensional subspace of $\mathbb{R}^n$, and let G be an $m \times n$ standard Gaussian matrix. Applying the projection estimates to the ellipsoid $A(B_2^k)$ gives

$$\mathbb{E}\,\|PA\| \asymp \frac{1}{\sqrt{n}}\,\|A\|_F + \sqrt{\frac{m}{n}}\,\|A\|, \qquad \mathbb{E}\,\|GA\| \asymp \|A\|_F + \sqrt{m}\,\|A\|.$$

8 Chaining

8.1 Dudley inequality

Definition 8.1 (Subgaussian increments; Book Definition 8.1.1). A process $(X_t)_{t\in T}$ has K-subgaussian increments with respect to a metric d if

$$\|X_t - X_s\|_{\psi_2} \le K d(t,s) \qquad \text{for every } s,t \in T.$$

Theorem 8.2 (Dudley entropy integral; Book Theorem 8.1.3). Let $(X_t)_{t\in T}$ be a centered process with K-subgaussian increments on (T,d). Interpreting the expected supremum through finite subsets as above,

$$\mathbb{E}\sup_{t\in T} X_t \le CK\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,\mathrm{d}\varepsilon.$$

The integrand vanishes for $\varepsilon \ge \operatorname{diam}(T)$.

The discrete version is

$$\mathbb{E}\sup_{t\in T} X_t \le CK\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})}.$$

Chaining with dyadic nets. Choose a 2^{-k}-net T_k at each scale and a map $\pi_k : T \to T_k$ that sends each point to a nearest point of the net. The coarsest net consists of a single point t_0, and

$$X_t - X_{t_0} = \sum_k \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right).$$

At scale k, each increment has subgaussian norm of order $K2^{-k}$ and there are at most $|T_k||T_{k-1}|$ pairs. The expected maximum is therefore bounded by a constant times $K2^{-k}\sqrt{\log|T_k|}$. Summing over scales and choosing nearly minimal nets gives the result.

For the canonical Gaussian process $X_t = \langle g,t\rangle$ on $T\subset\mathbb{R}^n$, the canonical metric is Euclidean, so

$$w(T) \le C\int_0^\infty \sqrt{\log N(T,\varepsilon)}\,\mathrm{d}\varepsilon.$$

Covering numbers measure the number of approximating points at each scale. The bound can be improved by retaining the approximation distances for each individual point throughout the sum.

8.2 Application: empirical processes

8.2.1 The Monte Carlo method

Let $X_1, \dots, X_n$ be i.i.d. with law μ. For an integrable function f, the Monte Carlo estimator

$$\widehat{I}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

is unbiased for $I(f) = \int f \, \mathrm{d}\mu$. The square root of its mean squared error is

$$\left(\mathbb{E} |\widehat{I}_n(f) - I(f)|^2 \right)^{1/2} = \sqrt{\frac{\mathrm{Var}(f(X))}{n}}.$$

The corresponding error over a class of functions is described by an empirical process.

Let $X_1, \dots, X_n$ be i.i.d. and let F be a class of functions. The centered empirical process is

$$Z_f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X), \qquad f \in F.$$

A uniform law of large numbers bounds $\sup_{f \in F} |Z_f|$.

For functions bounded in the sup norm, Hoeffding gives subgaussian increments under

$$d(f, g) = \frac{1}{\sqrt{n}} \|f - g\|_\infty .$$

Dudley's inequality bounds the expected supremum in terms of covering numbers of the function class.

Theorem 8.3 (Lipschitz law of large numbers; Book Theorem 8.2.3). For i.i.d. $X_i \in [0, 1]$ and the class of L-Lipschitz functions on $[0, 1]$,

$$\mathbb{E} \sup_{\|f\|_{\mathrm{Lip}} \le L} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \le \frac{CL}{\sqrt{n}}.$$

The empirical measure $\mu_n = n^{-1} \sum_i \delta_{X_i}$ rewrites the same quantity as

$$\sup_{f \in F} \left| \int f \, \mathrm{d}\mu_n - \int f \, \mathrm{d}\mu \right| .$$

Thus an empirical process measures the distance between empirical and population distributions through a chosen class of test functions.

8.3 VC dimension

Definition 8.4 (Shattering and VC dimension; Book Definition 8.3.1). A nonempty Boolean function class $F \subset \{0,1\}^\Omega$ *shatters* a finite set $S \subset \Omega$ if every labeling $S \to \{0,1\}$ is realized by the restriction of some $f \in F$. The VC dimension $\mathrm{vc}(F)$ is the largest cardinality of a shattered set.

Intervals on the line have VC dimension 2. Affine half-spaces in $\mathbb{R}^d$ have VC dimension of order d. VC dimension describes how many different labelings a function class can realize on finite sets.

Lemma 8.5 (Pajor; Book Lemma 8.3.7). If Ω and $F \subset \{0,1\}^\Omega$ are finite, then

$$|F| \le \#\{\Lambda \subseteq \Omega : \Lambda \text{ is shattered by } F\},$$

where the empty set is included. Counting these shattered subsets gives the Sauer–Shelah bound.

Lemma 8.6 (Sauer–Shelah; Book Lemma 8.3.9). If F is restricted to an n-point domain and $d = \mathrm{vc}(F) \le n$, then

$$|F| \le \sum_{k=0}^{d} \binom{n}{k} \le \left(\frac{en}{d}\right)^d.$$

For $d = 0$, the sum and the final expression are both interpreted as 1.

The growth function

$$\Pi_F(n) = \sup_{|S|=n} |F|_S|$$

equals 2^n for $n \le d$ and is at most polynomial in n for $n > d$.

Theorem 8.7 (VC entropy; Book Theorem 8.3.13). For a Boolean class of VC dimension d, any probability measure μ, and $0 < \varepsilon < 1$,

$$N(F, L^2(\mu), \varepsilon) \le \left(\frac{2}{\varepsilon}\right)^{Cd}.$$

The proof first takes a finite ε-separated subclass, samples a finite domain large enough to preserve separation, and then applies Sauer–Shelah on that random domain. This removes dependence on the size of the original domain.

Theorem 8.8 (VC law of large numbers; Book Theorem 8.3.15). Let F be a nonempty Boolean class of finite VC dimension d, and let $X, X_1, \dots, X_n$ be i.i.d. points in its domain, with $n \ge 1$. Then

$$\mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \le C \sqrt{\frac{\max\{1, d\}}{n}}.$$

Proof sketch. Symmetrize the empirical process. Conditional on the sample, the remaining randomness is Rademacher and the natural metric is empirical L^2. Apply Dudley conditionally to the increments and insert the uniform VC covering bound. The value at one fixed function contributes at most $C/\sqrt{n}$. Finally average over the sample.

For the class of intervals $(-\infty, t]$, the theorem becomes the Glivenko–Cantelli estimate

$$\mathbb{E}\,\|F_n - F\|_\infty \le \frac{C}{\sqrt{n}}$$

for the empirical distribution function.

8.4 Application: statistical learning theory

Let $X, X_1, \dots, X_n$ be i.i.d. points in a domain Ω, and let $T : \Omega \to \{0,1\}$ be a target function. The labeled sample is $(X_i, T(X_i))_{i=1}^n$. For a Boolean hypothesis $f : \Omega \to \{0,1\}$, the squared loss

$$\ell_f(x) = (f(x) - T(x))^2 = \mathbf{1}_{\{f(x)\neq T(x)\}}$$

is the indicator of misclassification. Define the population and empirical risks by

$$R(f) = \mathbb{E}\ell_f(X) = \mathbb{P}\{f(X) \neq T(X)\}, \qquad R_n(f) = \frac{1}{n}\sum_{i=1}^n \ell_f(X_i).$$

Let F be a nonempty hypothesis class whose population risk attains its minimum. Write f^* for a population minimizer and f_n^* for an empirical minimizer:

$$f_n^* \in \arg\min_{f\in F} R_n(f), \qquad f^* \in \arg\min_{f\in F} R(f).$$

The deterministic comparison

$$R(f_n^*) - R(f^*) \le 2\sup_{f\in F} |R_n(f) - R(f)|$$

reduces the expected excess risk to a uniform bound on the empirical process.

Theorem 8.9 (VC generalization; Book Theorem 8.4.5). For Boolean classification with a hypothesis class F of finite VC dimension,

$$\mathbb{E}R(f_n^*) \le R(f^*) + C\sqrt{\frac{\mathrm{vc}(F)}{n}}.$$

A larger class can reduce $R(f^*)$, while a larger VC dimension increases the estimation bound. Since f_n^* depends on the sample, the supremum over F allows the same bound to apply to the selected function.

8.5 Generic chaining

An admissible sequence consists of subsets $T_k \subset T$ satisfying $|T_0| = 1$ and $|T_k| \le 2^{2^k}$ for $k \ge 1$.

Definition 8.10 (γ_2 functional; Book Definition 8.5.1).

$$\gamma_2(T,d) = \inf_{(T_k)} \sup_{t\in T} \sum_{k=0}^{\infty} 2^{k/2} d(t,T_k),$$

where the infimum is over admissible sequences and $d(t,T_k) = \inf_{s\in T_k} d(t,s)$.

Theorem 8.11 (Generic chaining; Book Theorem 8.5.2). A centered process with K-subgaussian increments satisfies

$$\mathbb{E}\sup_{t\in T} X_t \le CK\gamma_2(T,d).$$

The supremum in γ_2 is taken after summing the approximation distances for each point. In Dudley's bound, a supremum is taken separately at every scale. Changing this order can give a smaller bound.

Theorem 8.12 (Majorizing measures; Book Theorem 8.5.5). For a centered Gaussian process with canonical metric d,

$$c\gamma_2(T,d) \le \mathbb{E}\sup_{t\in T} X_t \le C\gamma_2(T,d).$$

Generic chaining proves the upper bound; the majorizing measure theorem gives the lower bound. In particular, if $(X_x)_{x\in T}$ is a centered process and

$$\|X_x - X_y\|_{\psi_2} \le K\|x-y\|_2 \quad (x,y\in T\subset\mathbb{R}^n),$$

then

$$\mathbb{E}\sup_{x\in T} X_x \le CKw(T).$$

8.6 Chevet inequality

For a bounded nonempty set $T \subset \mathbb{R}^n$, define its radius by

$$\operatorname{rad}(T) := \sup_{x \in T} \|x\|_2 .$$

Theorem 8.13 (Subgaussian Chevet; Book Theorem 8.6.1). Let A have independent centered subgaussian rows with vector ψ_2 norm at most K. For bounded $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$,

$$\mathbb{E} \sup_{x \in T,\, y \in S} \langle Ax, y \rangle \le CK\big[w(T)\operatorname{rad}(S) + w(S)\operatorname{rad}(T)\big].$$

The increment bound compares the bilinear process with the sum of two canonical Gaussian processes. Their expected suprema are proportional to $w(T)\operatorname{rad}(S)$ and $w(S)\operatorname{rad}(T)$, respectively.

9 Deviations of Random Matrices on Sets

9.1 Matrix deviation inequality

Let m, n be positive integers, and let A be an $m \times n$ matrix with independent isotropic subgaussian rows A_i. Set $K = \max_i \|A_i\|_{\psi_2}$ and define the deviation process

$$Z_x = \|Ax\|_2 - \sqrt{m}\,\|x\|_2 .$$

Theorem 9.1 (Matrix deviation inequality; Book Theorem 9.1.1). For every bounded $T \subset \mathbb{R}^n$,

$$\mathbb{E} \sup_{x \in T} |Z_x| \le CK^2 \gamma(T), \qquad \gamma(T) = \mathbb{E} \sup_{x \in T} |\langle g, x \rangle|.$$

Theorem 9.2 (Increment estimate; Book Theorem 9.1.2). For all $x, y \in \mathbb{R}^n$,

$$\|Z_x - Z_y\|_{\psi_2} \le CK^2 \|x - y\|_2 .$$

Proof sketch. For a unit vector, Bernstein's inequality applied to $\langle A_i, x \rangle^2 - 1$ gives concentration of $\|Ax\|_2$. For two vectors of equal norm, their sum and difference express the difference of squared norms as a sum of products of subgaussian random variables. Homogeneity and a geometric decomposition then extend the estimate to vectors with unequal norms. Generic chaining applies to the resulting increment bound in the Euclidean metric and gives the bound in terms of $\gamma(T)$.

For every $u \ge 0$, the corresponding tail bound holds with probability at least $1 - 2e^{-u^2}$:

$$\sup_{x \in T} |Z_x| \le CK^2 \big[w(T) + u \operatorname{rad}(T) \big].$$

A bound for the squared norms follows by multiplying $|\,\|Ax\|_2 - \sqrt{m}\,\|x\|_2\,|$ by the corresponding sum.

9.2 Random matrices, covariance estimation, and Johnson–Lindenstrauss

Taking $T = S^{n-1}$ gives $\gamma(T) \asymp \sqrt{n}$ and recovers the two-sided bound on singular values from Chapter 4. Choosing T to be an ellipsoid or a set of differences gives the following estimates.

For a scaled subgaussian projection $P = A/\sqrt{n}$,

$$\mathbb{E}\,\mathrm{diam}(PT) \le \sqrt{\frac{m}{n}}\,\mathrm{diam}(T) + CK^2 w_s(T),$$

where w_s is spherical width.

Theorem 9.3 (Covariance estimation with effective rank; Book Theorem 9.2.2). Let $X \in \mathbb{R}^n$ have a nonzero second moment matrix $\Sigma = \mathbb{E}XX^T$. Assume that, for some $K \ge 1$,

$$\|\langle X, v\rangle\|_{\psi_2} \le K\,\|\langle X, v\rangle\|_{L^2} \qquad \text{for every } v \in \mathbb{R}^n.$$

For independent copies $X_1, \dots, X_m$ of X, define

$$\Sigma_m = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T, \qquad r = \frac{\mathrm{tr}\,\Sigma}{\|\Sigma\|}.$$

Then

$$\mathbb{E}\,\|\Sigma_m - \Sigma\| \le CK^4\left(\sqrt{\frac{r}{m}} + \frac{r}{m}\right)\|\Sigma\|.$$

The rank parameter r can be far smaller than n when the spectrum decays. It arises from the Gaussian width of the ellipsoid $\Sigma^{1/2}S^{n-1}$.

For a bounded set $X \subset \mathbb{R}^n$, the scaled matrix $Q = A/\sqrt{m}$ satisfies an additive Johnson–Lindenstrauss estimate

$$\left|\,\|Qx - Qy\|_2 - \|x - y\|_2\,\right| \le \frac{CK^2 w(X)}{\sqrt{m}}$$

uniformly over $x, y \in X$ with probability at least 0.99. A finite normalized difference set has width $O(\sqrt{\log|X|})$, recovering the classical lemma; the same formula applies to infinite structured sets.

9.3 Random sections: the M^* bound and escape theorem

Theorem 9.4 (M^* bound; Book Theorem 9.3.1). Let $E = \ker A$. For a bounded set $T \subset \mathbb{R}^n$,

$$\mathbb{E}\,\mathrm{diam}(T \cap E) \le \frac{CK^2 w(T)}{\sqrt{m}}.$$

If $x, y \in T \cap E$, then $A(x - y) = 0$. Apply matrix deviation to $T - T$; on the kernel, the random image term disappears and the deterministic $\sqrt{m}\,\|x - y\|_2$ term controls the diameter. For $T = B_1^n$ with $n \ge 2$, the bound uses $w(B_1^n) \asymp \sqrt{\log n}$.

Theorem 9.5 (Escape theorem; Book Theorem 9.3.4). Let $T \subset S^{n-1}$. If

$$m \geq CK^4 w(T)^2,$$

then, with probability at least $1 - 2e^{-cm/K^4}$,

$$T \cap \ker A = \varnothing.$$

A random subspace of sufficiently large codimension therefore avoids T. The required codimension is proportional to the square of its Gaussian width.

9.4 Application: high-dimensional linear models

Consider

$$y = Ax + w,$$

with $m \ll n$. The solutions of the noiseless equation $Ax = y$ form an affine space whose direction is $\ker A$. A known constraint $x \in T$ restricts the possible solutions. In the noiseless case, choose any $\widehat{x} \in T$ satisfying $A\widehat{x} = y$.

Theorem 9.6 (Constrained recovery; Book Theorem 9.4.4). Under the standard isotropic subgaussian measurement model,

$$\mathbb{E} \left\| \widehat{x} - x \right\|_2 \leq \frac{CK^2 w(T)}{\sqrt{m}}.$$

Both x and $\widehat{x}$ lie in the same affine section $T \cap (x + \ker A)$, so a translated M^* argument bounds their distance. For fixed K, the bound is small relative to $\operatorname{diam}(T)$ when the number of measurements is sufficiently large compared with the effective dimension of T.

Replacing T by $\operatorname{conv} T$ makes the feasibility problem convex. Since $w(\operatorname{conv} T) = w(T)$, the same error bound applies to this convex relaxation.

9.4.1 Sparse vectors

Let $n \geq 2$ and $1 \leq s \leq n$. An s-sparse vector x with $\|x\|_2 \leq 1$ belongs to $\sqrt{s} B_1^n$. The convex feasibility program

$$Ax' = y, \qquad \|x'\|_1 \leq \sqrt{s}$$

therefore obeys

$$\mathbb{E} \left\| \widehat{x} - x \right\|_2 \leq CK^2 \sqrt{\frac{s \log n}{m}}.$$

The tighter relaxation

$$T_{n,s} = \sqrt{s} B_1^n \cap B_2^n$$

also imposes $\|x'\|_2 \leq 1$. Its Gaussian width is of order $\sqrt{s \log(en/s)}$, improving the error bound to $CK^2 \sqrt{s \log(en/s)/m}$.

9.4.2 Low-rank matrices

A matrix has rank r when exactly r singular values are nonzero. The nuclear norm $\|X\|_* = \sum_i s_i(X)$ is the analogue of the ℓ_1 norm for these singular values. Let X be a $d \times d$ matrix of rank at most r with $\|X\|_F \le 1$, where $1 \le r \le d$. Suppose the observations are

$$y_i = \langle G_i, X \rangle, \qquad i = 1, \dots, m,$$

where the G_i are independent $d \times d$ matrices with i.i.d. $N(0,1)$ entries and $\langle G_i, X \rangle = \operatorname{tr}(G_i^T X)$ is the Frobenius inner product. Since $\|X\|_* \le \sqrt{r}\,\|X\|_F$, the convex feasibility program

$$\langle G_i, X' \rangle = y_i \quad (i = 1, \dots, m), \qquad \|X'\|_* \le \sqrt{r}$$

contains X. Every feasible estimate $\widehat{X}$ satisfies

$$\mathbb{E}\left\|\widehat{X} - X\right\|_F \le C\sqrt{\frac{rd}{m}}.$$

The bound is small when the number of measurements is large compared with rd, which can be much smaller than d^2.

9.5 Application: exact sparse recovery

Basis pursuit solves

$$\min_{x'} \|x'\|_1 \quad \text{subject to} \quad Ax' = y.$$

Theorem 9.7 (Uniform exact recovery; Book Theorem 9.5.1). Let A have independent isotropic subgaussian rows, with $K = \max_i \|A_i\|_{\psi_2}$. For $n \ge 2$ and $1 \le s \le n$, if

$$m \ge CK^4 s \log n,$$

then, with probability at least $1 - 2e^{-cm/K^4}$, basis pursuit exactly recovers every s-sparse vector from $y = Ax$.

Let $h = \widehat{x} - x$ and $S = \operatorname{supp}(x)$. Optimality gives the cone inequality

$$\|h_{S^c}\|_1 \le \|h_S\|_1, \qquad \|h\|_1 \le 2\sqrt{s}\,\|h\|_2.$$

If $h \neq 0$, the normalized direction $h/\|h\|_2$ lies both in the kernel and in a spherical set of width $O(\sqrt{s \log n})$. The escape theorem rules out that intersection.

Definition 9.8 (Restricted isometry property; Book Definition 9.5.5). A matrix A has RIP with parameters (α, β, s), where $0 < \alpha \le \beta$ and s is a positive integer, if

$$\alpha \|v\|_2 \le \|Av\|_2 \le \beta \|v\|_2$$

for every s-sparse vector v.

Theorem 9.9 (RIP implies exact recovery; Book Theorem 9.5.6). Suppose A has RIP with parameters $(\alpha, \beta, (1+\lambda)s)$, where λs is an integer and

$$\lambda > \left(\frac{\beta}{\alpha}\right)^2.$$

Then basis pursuit uniquely recovers every s-sparse vector from $y = Ax$.

Proof sketch. Let $h = \widehat{x} - x$, and let $I_0 = \operatorname{supp}(x)$. Partition the remaining coordinates into blocks $I_1, I_2, \ldots$ of size λs, ordered by decreasing magnitude of h, with the last block possibly smaller. Set $I_{01} = I_0 \cup I_1$. Since $Ah = 0$, RIP and the cone inequality give

$$\alpha \|h_{I_{01}}\|_2 \le \beta \sum_{i \ge 2} \|h_{I_i}\|_2 \le \frac{\beta}{\sqrt{\lambda}} \|h_{I_0}\|_2 \le \frac{\beta}{\sqrt{\lambda}} \|h_{I_{01}}\|_2 .$$

The assumption on λ forces $h_{I_{01}} = 0$, and the ordering of the remaining coordinates then gives $h = 0$.

Theorem 9.10 (Random matrices satisfy RIP; Book Theorem 9.5.7). Let A have independent isotropic subgaussian rows, with $K = \max_i \|A_i\|_{\psi_2}$, and let $1 \le s \le n$. If

$$m \ge CK^4 s \log(en/s),$$

then, with probability at least $1 - 2e^{-cm/K^4}$, A has RIP with parameters

$$(\alpha, \beta, s) = (0.9\sqrt{m}, 1.1\sqrt{m}, s).$$

The proof fixes one support, applies the bound on singular values, and takes a union bound over the $\binom{n}{s}$ possible supports.

Remark 9.11 (Escape theorem and restricted isometry). The escape theorem is applied to the intersection of the unit sphere with the cone containing possible errors $\widehat{x} - x$. The restricted isometry argument bounds the action of A on each sparse coordinate subspace and decomposes the error into blocks of coordinates.

9.6 Deviations of random matrices for general norms

The Euclidean norm can be replaced by a positively homogeneous, subadditive function f. Support functions

$$f(z) = \sup_{y \in S} \langle z, y \rangle$$

are the main examples.

Theorem 9.12 (General Gaussian deviation; Book Theorem 9.6.3). Let A be an $m \times n$ matrix with i.i.d. $N(0,1)$ entries. Let $f : \mathbb{R}^m \to \mathbb{R}$ be positively homogeneous, subadditive, and bounded on the unit sphere, with $f(z) \le b \|z\|_2$ for every $z \in \mathbb{R}^m$ and some $b \ge 0$, and let $T \subset \mathbb{R}^n$. Then

$$\mathbb{E} \sup_{x \in T} |f(Ax) - \mathbb{E} f(Ax)| \le Cb\gamma(T).$$

The increment estimate is

$$\|[f(Ax) - \mathbb{E} f(Ax)] - [f(Ay) - \mathbb{E} f(Ay)]\|_{\psi_2} \le Cb \|x - y\|_2 .$$

For vectors x, y of equal length, the Gaussian images of the normalized sum and difference directions are independent; homogeneity extends the result to unequal lengths. Generic chaining completes the proof.

9.7 Two-sided Chevet and Dvoretzky–Milman theorem

Theorem 9.13 (Two-sided Chevet; Book Theorem 9.7.1). Let A be an $m \times n$ matrix with i.i.d. $N(0,1)$ entries, and let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be bounded. Then

$$\mathbb{E} \sup_{x \in T} \left| \sup_{y \in S} \langle Ax, y \rangle - w(S) \|x\|_2 \right| \le C\gamma(T) \operatorname{rad}(S).$$

The inner supremum is the support function of S. The theorem therefore follows from the deviation result for general norms and sharpens the one-sided Chevet inequality of Chapter 8.

Theorem 9.14 (Dvoretzky–Milman form; Book Theorem 9.7.2). Let A be an $m \times n$ matrix with i.i.d. $N(0,1)$ entries and let $T \subset \mathbb{R}^n$ be bounded. With probability at least 0.99,

$$r_- B_2^m \subset \operatorname{conv}(AT) \subset r_+ B_2^m, \qquad r_\pm = w(T) \pm C\sqrt{m} \operatorname{rad}(T),$$

with the inner inclusion meaningful when $r_- \ge 0$.

Duality for support functions converts uniform control of $\sup_{x \in T} \langle Ax, y \rangle$ over $y \in S^{m-1}$ into the two set inclusions. In particular, when $\operatorname{rad}(T) > 0$ and

$$m \le c \frac{w(T)^2}{\operatorname{rad}(T)^2},$$

the two radii are close, so the convex hull of the projected set is approximately a Euclidean ball.