A Guided Tour of Modern Statistics

Zixun Huang

May 13, 2026

Contents

Part I
Chapter 1: Classical Theory

1 Probability Theory and Stochastic Processes

1.1 Overview

In this chapter, we introduce the basic notation and fundamental tools of probability theory and measure theory that will be used throughout the subsequent chapters. The section is organized as follows:

1. probability space and random variables;
2. modes of convergence;
3. the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT);
4. martingales, Markov chains and Brownian motion.

This section mainly follows Introduction to Probability Theory (Tsinghua University), Stochastic Processes (Peking University), Stochastic Analysis (Peking University), Math C218A (UC Berkeley), Math C218B (UC Berkeley) and the book Probability: Theory and Examples by Rick Durrett [3], Sinho Chewi's notes of MATH C218A, and Introduction to Stochastic Processes [4].

1.2 Probability Space and Random Variables

1.2.1 Probability space

A probability space $(\Omega, \mathscr{F}, P)$ contains three elements:

- The space Ω: this is a non-empty set. It can be viewed as the set of all possible outcomes.
- The σ-field $\mathscr{F}$: this can be viewed as a collection of all the events.
- The probability measure P: this is a function from $\mathscr{F}$ to $[0,1]$. It gives a probability to each event.

Definition 1.1 ($\sigma-$Field)**.** *Suppose $\mathscr{F}$ is a non-empty collection of subsets of Ω.*

- *It is a field if it is closed under **complementation** and closed under **union**:*

$$A \in \mathscr{F} \implies A^c \in \mathscr{F}, \quad A_1, A_2 \in \mathscr{F} \implies A_1 \cup A_2 \in \mathscr{F}.$$

- *It is a **monotone class** if*

$$A_j \in \mathscr{F},\ A_j \subset A_{j+1},\ 1 \le j < \infty \implies \bigcup_j A_j \in \mathscr{F},$$

and

$$A_j \in \mathscr{F},\ A_j \supset A_{j+1},\ 1 \le j < \infty \implies \bigcap_j A_j \in \mathscr{F}.$$

- *It is a σ-field if it is closed under **complementation** and closed under **countable union**:*

$$A \in \mathscr{F} \implies A^c \in \mathscr{F}, \quad A_j \in \mathscr{F},\ 1 \le j < \infty \implies \bigcup_j A_j \in \mathscr{F}.$$

Definition 1.2 (Generated σ-Fields)**.** *Given any collection C of sets, the σ-field generated by C is the intersection of all σ-fields containing C.*

Definition 1.3 (Probability Measure)**.** *Suppose $\mathscr{F}$ is a σ-field on Ω. A probability measure P is a function from $\mathscr{F}$ to $[0,1]$ satisfying the following axioms:*

- $P[E] \geq 0$ *for all* $E \in \mathscr{F}$*;*
- $P[\Omega] = 1$*;*
- *If $\{E_j\}_j$ is a countable collection of pairwise disjoint sets in $\mathscr{F}$, then:*

$$P\left[\bigcup_j E_j\right] = \sum_j P[E_j].$$

These axioms imply the following consequences:

- $P[E^c] = 1 - P[E]$*;*
- $P[E \cup F] + P[E \cap F] = P[E] + P[F]$*;*
- *Continuity: if $E_n \uparrow E$ or $E_n \downarrow E$, then $P[E_n] \to P[E]$;*
- $P\left[\bigcup_j E_j\right] \leq \sum_j P[E_j]$.

Remark 1.4. *We care about the σ-field because that is where our random variables live. It defines the rules under which stochastic analysis can work. Sometimes we need to enlarge our space to a bigger σ-field and make sure the measure still fits — this is guaranteed by the following theorem.*

Theorem 1.5 (Carathéodory's Extension Theorem)**.** *Suppose $\mathscr{F}_0$ is a field and $\mathscr{F}$ is the σ-field generated by $\mathscr{F}_0$. Suppose μ is a probability measure on $\mathscr{F}_0$. Then there exists a unique probability measure on $\mathscr{F}$ that coincides with μ on $\mathscr{F}_0$.*

Remark 1.6. *The construction:*

$$\tau(A) = \inf\left\{\sum_{n=1}^{\infty} \mu(B_n) : B_n \in \mathscr{F},\ A \subset \bigcup_{n=1}^{\infty} B_n\right\}.$$

1.2.2 Random Variables

Definition 1.7 (Random Variables)**.** *A real-valued random variable is a function $X : \Omega \to \mathbb{R}$ such that*

$$X^{-1}(B) \in \mathscr{F}, \quad \forall B \in \mathscr{B}.$$

In other words, a random variable is just a ***measurable*** *function from $(\Omega, \mathscr{F})$ to $(\mathbb{R}, \mathscr{B})$.*

Definition 1.8 (Distribution)**.** *Each random variable X induces a probability measure μ on $(\mathbb{R}, \mathscr{B})$ by the following correspondence:*

$$\mu[B] = P[X^{-1}(B)] = P[X \in B], \quad \forall B \in \mathscr{B}.$$

The measure μ is called the law (or the distribution) of X, denoted by $\mathscr{L}(X)$; its associated distribution function is called the distribution function of X, denoted by F_X.

The concept of "expectation" is the same as integration in the probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Corollary 1.9. *If X takes only positive integer values, we have*

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{P}[X \geq n].$$

Definition 1.10 (p-th Moment)**.** *For any $p \in (0, \infty)$, define*

$$L^p(\Omega, \mathscr{F}, \mathbb{P}) = \{X \text{ random variable on } (\Omega, \mathscr{F}, \mathbb{P}) : \mathbb{E}[|X|^p] < \infty\}.$$

For $X \in L^p$, we call $\mathbb{E}[|X|^p]$ the p-th moment of X.

Definition 1.11 (Independent)**.** *The random variables $\{X_j, 1 \le j \le n\}$ are independent if, for any Borel sets $\{B_j, 1 \le j \le n\}$, we have*

$$\mathbb{P}\left[\bigcap_{j=1}^{n} \{X_j \in B_j\}\right] = \prod_{j=1}^{n} \mathbb{P}[X_j \in B_j].$$

The random variables $\{X_j, j \ge 1\}$ are independent if $\{X_j, 1 \le j \le n\}$ are independent for all n.

Theorem 1.12. *Suppose $\{A_j, 1 \le j \le n\}$ are independent and each A_j is a π-system. Denote by $\sigma(A_j)$ the σ-field generated by A_j for each j. Then $\{\sigma(A_j), 1 \le j \le n\}$ are independent.*

Example 1.13 (Kolmogorov's 0-1 Law)**.** *Let $\{X_n\}$ be a sequence of independent random variables. Let*

$$G_n = \sigma(X_k, k \ge n) \quad \text{and} \quad G_\infty = \bigcap_{n \ge 1} G_n.$$

Then G_∞ is trivial, i.e., for any $A \in G_\infty$, we have

$$\mathbb{P}[A] = 0 \text{ or } 1.$$

Define $S_n = \sum_{j=1}^{n} X_j$. It is clear that the following events are in G_∞:

$$\lim_{n\to\infty} S_n \text{ exists}, \quad \lim_{n\to\infty} \frac{S_n}{n} \text{ exists}, \quad \limsup_{n\to\infty} \frac{S_n}{n} > 0.$$

Whereas, the following event is not in G_∞:

$$\limsup_{n\to\infty} S_n > 0.$$

Proof. On the one hand, we have $A \in G_{n+1}$ for any n. Thus, A is independent of $\sigma(X_1, \ldots, X_n)$ for any n. Therefore, A is independent of $\sigma(X_n, n \ge 1)$. On the other hand, A is measurable with respect to $\sigma(X_n, n \ge 1)$. Therefore, A is independent of itself. This implies $\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2$, thus $\mathbb{P}[A] \in \{0, 1\}$. □

Now let's talk about Gaussians. The key thing to know is that Gaussians stay Gaussian under linear combinations and projections.

Proposition 1.14. *The following are equivalent:*

1. *The random vector $\xi = (\xi_1, \ldots, \xi_d)^\top$ follows a d-dimensional Gaussian distribution.*
2. *For any $a_1, \ldots, a_d$, the linear combination $\sum_{k=1}^{d} a_k \xi_k$ follows a one-dimensional Gaussian distribution.*

1.3 Convergence

1.3.1 Convergence: Almost Sure, in Probability, and in L^p

Definition 1.15 (Almost sure convergence (a.s.))**.** *The sequence of random variables $\{X_n\}$ converges a.s. to the random variable X if there exists a null set $\mathcal{N}$ such that*

$$\lim_{n\to\infty} X_n(\omega) = X(\omega), \quad \forall \omega \in \Omega \setminus \mathcal{N}.$$

Definition 1.16 (Convergence in probability)**.** *The sequence $\{X_n\}$ converges in probability to the random variable X if, for every $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0.$$

Definition 1.17 (Convergence in L^p))**.** *Assume $p \geq 1$. The sequence $\{X_n\}$ converges in L^p to the random variable X if $X_n \in L^p$, $X \in L^p$ and*

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

It is straightforward to prove that almost sure convergence ($a.s.$) implies convergence in probability, and convergence in L^p implies convergence in probability. However, almost sure convergence ($a.s.$) and L^p may not imply each other. Now we talk about an example: using probability methods to prove Weierstrass Theorem.

Example 1.18 (Weierstrass Theorem)**.** *Let f be a continuous function on $[0,1]$. Define the polynomial:*

$$f_n(x) = \sum_{j=0}^{n} \binom{n}{j} x^j (1-x)^{n-j} f\left(\frac{j}{n}\right).$$

This is called the Bernstein polynomial of degree n associated to f. Then we have

$$\sup_{x\in[0,1]} |f_n(x) - f(x)| \to 0, \quad n \to \infty.$$

Proof. Suppose $\{X_j, j \geq 1\}$ are i.i.d. Bernoulli variables with parameter $p \in [0,1]$: $\mathbb{P}[X_j = 1] = p, \mathbb{P}[X_j = 0] = 1 - p$, and set $S_n = \sum_{j=1}^n X_j$. Then we find

$$\mathbb{E}[X_j] = p, \quad \mathrm{var}(X_j) = p(1-p), \quad \mathbb{E}[S_n] = np, \quad \mathrm{var}(S_n) = np(1-p).$$

Moreover,

$$\mathbb{P}[S_n = j] = \binom{n}{j} p^j (1-p)^{n-j}, \quad \text{thus} \quad f_n(p) = \mathbb{E}[f(S_n/n)].$$

Note that f is bounded: $|f| \leq M$, and it is uniformly continuous: for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(y)| \leq \epsilon$ as long as $|x - y| \leq \delta$. Thus,

$$|f_n(p) - f(p)| \leq \mathbb{E}[|f(S_n/n) - f(p)|] \leq \epsilon + 2M\mathbb{P}[|S_n/n - p| > \delta].$$

Note that

$$\mathbb{P}[|S_n/n - p| > \delta] \leq \frac{\mathrm{var}(S_n)}{n^2\delta^2} = \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}.$$

Thus,

$$|f_n(p) - f(p)| \leq \epsilon + \frac{M}{2n\delta^2}.$$

□

We give a useful lemma to prove almost surely convergence.

Theorem 1.19 (Borel-Cantelli Lemma)**.**

- *For arbitrary sequence $\{E_n\}$, we have*

$$\sum_n \mathbb{P}[E_n] < \infty \quad \Rightarrow \quad \mathbb{P}[E_n \ i.o.] = 0.$$

- *If the events $\{E_n\}$ are independent, we have*

$$\sum_n \mathbb{P}[E_n] = \infty \quad \Rightarrow \quad \mathbb{P}[E_n \ i.o.] = 1.$$

1.3.2 Weak Convergence

Definition 1.20 (Weak Convergence)**.** *A sequence of measures $\{\mu_n\}$ converges weakly to a measure μ if*

$$\mu_n((a,b]) \to \mu((a,b]), \quad \text{for all continuity points } a, b \text{ of } \mu.$$

We denote by $\mu_n \Rightarrow \mu$.

Proposition 1.21. *Suppose $\{\mu_n\}$ and μ are probability measures. Then $\mu_n \Rightarrow \mu$ if and only if*

$$\lim_{n\to\infty} \int f(x)\, \mu_n[dx] \to \int f(x)\, \mu[dx], \quad \forall f \in C_b := \text{(bounded continuous functions)}.$$

1.3.3 Convergence in Distribution

Definition 1.22 (Convergence in Distribution)**.** *A sequence of random variables $\{X_n\}$ converges in distribution to a random variable X if distribution function $\mathscr{L}(X_n) \Rightarrow \mathscr{L}(X)$. We denote by $X_n \xrightarrow{d} X$, or $X_n \to X$ in distribution.*

Convergence in probability implies convergence in distribution. Furthermore, convergence in distribution is equivalent to the convergence of **characteristic functions**.

Definition 1.23 (Characteristic Function)**.** *For any random variable X with distribution μ, its characteristic function is defined to be:*

$$f : \mathbb{R} \to \mathbb{C}, \quad f(t) := \mathbb{E}\left[e^{itX}\right] = \int e^{itx}\, \mu[dx].$$

The distribution is uniquely identified by the characteristic function.

Theorem 1.24. *Suppose f is the characteristic function for the probability measure μ. For $x < y$, we have*

$$\mu[(x,y)] + \frac{1}{2}\mu[\{x\}] + \frac{1}{2}\mu[\{y\}] = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-itx} - e^{-ity}}{it} f(t)\, dt.$$

Theorem 1.25. *Let $\{\mu_n\}$ be a sequence of probability measures with characteristic functions $\{f_n\}$. Suppose that*

1. *f_n converges everywhere in $\mathbb{R}$ and defines the limiting function f;*
2. *f is continuous at $t = 0$.*

Then we have

1. *$\mu_n \xrightarrow{d} \mu$ where μ is a probability measure;*

2. the characteristic function of μ is f.

At the end of this subsection, we introduce some useful convergence results. Let (X_n) be a sequence of random variables and let X be a random variable such that $X_n \to X$ (a.s.):

$$\mathbb{P}(X_n \to X) = 1.$$

We state the convergence properties as follows:

- **(MON)** If $0 \leq X_n \uparrow X$, then $\mathbb{E}(X_n) \leq \mathbb{E}(X) < \infty$;
- **(FATOU)** If $X_n \geq 0$, then $\mathbb{E}(X_n) \leq \liminf \mathbb{E}(X_n)$;
- **(DOM)** If $|X_n(\omega)| \leq Y(\omega)$, $\forall(n,\omega)$, and $\mathbb{E}(Y) < \infty$, then
$$\mathbb{E}(|X_n - X|) \to 0,$$
- **(SCHEFFÉ)** If $\mathbb{E}(|X_n|) \to \mathbb{E}(|X|)$, then
$$\mathbb{E}(|X_n - X|) \to 0;$$
- **(BDD)** If there exists a constant K, such that $|X_n(\omega)| \leq K$, $\forall(n,\omega)$, then
$$\mathbb{E}(|X_n - X|) \to 0.$$

1.4 The Law of Large Numbers and the Central Limit Theorem

1.4.1 Strong Law of Large Numbers

We will give three versions (4th moment SLLN, 2nd moment SLLN and SLLN).

Theorem 1.26 (4th Moment SLLN)**.** *Let $(X_i, 1 \leq i < \infty)$ be IID, $EX = 0$, and $EX^4 < \infty$. Write $S_n = \sum_{i=1}^n X_i$. Then*

1. *$ES_n^4 \leq 3n^2 EX^4$*
2. *$S_n/n \to 0$ as $n \to \infty$.*

If $EX = \mu$, applying the theorem to $X - \mu$ shows that $S_n/n \to \mu$ a.s.

Proof.

$$\begin{aligned} ES_n^4 &= \sum_i \sum_j \sum_k \sum_l E[X_i X_j X_k X_l] \\ &= nEX^4 + \binom{4}{2}\binom{n}{2} E[X_1^2 X_2^2] \\ &= nEX^4 + 3n(n-1) \underbrace{(EX^2)^2}_{\leq EX^4} \end{aligned}$$

since $(EY)^2 \leq E(Y^2)$. Fix $\varepsilon > 0$.

$$\begin{aligned} P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) &\leq E\left|\frac{S_n}{n}\right|^4 \cdot \frac{1}{\varepsilon^4} \\ &\leq \varepsilon^{-4} n^{-4} \cdot 3n^2 EX^4 \\ &\leq 3\varepsilon^{-4} EX^4 n^{-2} \end{aligned}$$

This implies that

$$\sum_n P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) \leq \sum_n 3\varepsilon^{-4} EX^4 n^{-2} < \infty$$

By Borel-Cantelli Lemma 1.19, $S_n/n \to 0$ a.s. We used the fact that $s^4 = |s|^4$ and $s^2 = |s|^2$, but this does not work for the third moment: $s^3 \neq |s|^3$. □

Theorem 1.27 (2nd Moment SLLN)**.** *Given* $(X_i, 1 \leq i < \infty)$*, with* $EX_i \equiv 0$*, let* $\sup_i EX_i^2 = B < \infty$ *and the* X_i *be* ***orthogonal****,* $E(X_i X_j) = 0$*,* $j \neq i$*. (We are not assuming independence!) Write*

$$S_n = \sum_{i=1}^n X_i$$

Then $S_n/n \to 0$ *a.s.*

Proof. Since $\text{var}(S_n) \leq nB$, Chebyshev's inequality implies

$$P\left(\frac{|S_n|}{n} \geq \varepsilon\right) \leq \frac{nB}{n^2\varepsilon^2} = \frac{B}{n\varepsilon^2}$$

Take $n(j) = j^2$.

$$P\left(\left|\frac{S_{n(j)}}{n(j)}\right| \geq \varepsilon\right) \leq \frac{B}{\varepsilon^2}\frac{1}{j^2}$$

Use Borel-Cantelli 1.19.

$$\frac{S_{n(j)}}{n(j)} \to 0 \quad \text{a.s.} \quad \text{as } j \to \infty$$

It is enough to prove $D_j/j^2 \to 0$ a.s., for

$$D_j = \max_{j^2 \leq n < (j+1)^2} |S_n - S_{j^2}|$$

Then

$$D_j^2 = \max_{j^2 \leq n < (j+1)^2} (S_n - S_{j^2})^2$$

$$ED_j^2 \underset{\text{crude}}{\leq} \sum_{n=j^2}^{(j+1)^2-1} E(S_n - S_{j^2})^2$$

Since

$$E(S_n - S_{j^2})^2 = \text{var}\left(\sum_{j^2+1}^n X_i\right) \leq B(n - j^2)$$

Letting $n = j^2 + i$, we have

$$ED_j^2 \leq B\sum_{i=1}^{2j+1} i = \frac{1}{2}(2j+1)(2j+2)B$$

We have

$$P\left(\frac{D_j}{j^2} \geq \varepsilon\right) \leq \frac{ED_j^2}{\varepsilon^2 j^4} \in O(j^{-2})$$

Borel-Cantelli Lemma 1.19 implies that $D_j/j^2 \to 0$ as $j \to \infty$. □

Theorem 1.28 (SLLN)**.** *Let* (X_i) *be IID with* $E|X| < \infty$*. Then* $S_n/n \to EX$ *a.s. as* $n \to \infty$*.*

First prove some lemmas. We'll use 1.31 and truncate + center techniques.

Theorem 1.29. *Let (X_i) be independent, with $EX_i = 0$ and $\sigma_i^2 = \operatorname{var}(X_i) < \infty$. If $\sum_{i=1}^{\infty} \sigma_i^2 < \infty$, then $\sum_{i=1}^{\infty} X_i$ converges a.s.*

Proof. Define $M_k = \sup_{n>k} \left|\sum_{i=k+1}^{n} X_i\right|$. It is enough to show that $M_k \to 0$ a.s. as $k \to \infty$. Define also $W_k = \sup_{n_2>n_1>k} \left|\sum_{i=n_1+1}^{n_2} X_i\right|$ and note that $M_k \le W_k \le 2M_k$ and W_k decreases as k increases.

$$P\left(\sup_{k<n\le N}\left|\sum_{i=k+1}^{N} X_i\right| \ge \varepsilon\right) \underset{\substack{martingale\\ maximal\ ineq.}}{\le} \varepsilon^{-2} \operatorname{var}\left(\sum_{i=k+1}^{N} X_i\right) = \varepsilon^{-2} \sum_{i=k+1}^{N} \sigma_i^2$$

Taking $N \to \infty$, $P(M_k > \varepsilon) \le \varepsilon^{-2} \sum_{i=k+1}^{\infty} \sigma_i^2$.

$$P(W_k > \varepsilon) \le P\left(M_k > \frac{\varepsilon}{2}\right) \le 4\varepsilon^{-2} \sum_{i=k+1}^{\infty} \sigma_i^2 \to 0 \text{ as } k \to \infty$$

Taking $k \to \infty$, then $W_k \downarrow W_\infty$ for some W_∞ a.s. Then $P(W_\infty > \varepsilon) = 0$, which implies that $W_\infty = 0$ a.s., which implies that $W_k \downarrow 0$ a.s. and $M_k \to 0$ a.s. □

Lemma 1.30 (Deterministic Lemma (Kronecker))**.** *Let (x_n) be a sequence of reals, $S_n = \sum_{i=1}^{n} x_i$, $0 < a_n \uparrow \infty$ as $n \uparrow \infty$. If $\sum_i x_i/a_i$ converges, then $S_n/a_n \to 0$.*

Corollary 1.31. *Let (X_i) be independent, $EX_i = 0$, $EX_i^2 < \infty$, and $S_n = \sum_{i=1}^{n} X_i$. If $0 < a_n \uparrow \infty$ as $n \uparrow \infty$ and if $\sum_n EX_n^2/a_n^2 < \infty$, then $S_n/a_n \to 0$ a.s.*

Proof. Theorem 1.29 implies that $\sum_n X_n/a_n$ converges a.s. Then Lemma 1.30 implies that $S_n/a_n \to 0$ a.s. □

Proof of Theorem 1.28. The idea is to truncate, center, and then apply 9.4.

If $Z \ge 0$, then

$$EZ^k = \int_0^{\infty} kz^{k-1} P(Z \ge z)\, dz.$$

Define $Y_k = X_k 1_{(|X_k| \le k)}$. Then

$$\sum_k P(Y_k \ne X_k) = \sum_{k=1}^{\infty} P(|X| > k) \le \int_0^{\infty} P(|X| > x)\, dx = E|X| < \infty.$$

Then Borel-Cantelli Lemma 1.19 implies that $P(Y_k = X_k, \text{ ultimately}) = 1$. It is enough to prove that $(1/n)\sum_{k=1}^{n} Y_k \to EX$ a.s.

Center: define $X_k' = Y_k - EY_k$. *Claim:* $\sum_k \operatorname{var}(X_k')/k^2 < \infty$.

$$EY_k^2 = \int_0^{\infty} 2yP(|Y_k| > y)\, dy = \int_0^{\infty} \underbrace{2yP(k \ge |X_k| \ge y)1_{(y\le k)}}_{\text{Check this!}}\, dy$$

$$\le \int_0^{\infty} 2yP(|X_k| \ge y)1_{(y\le k)}\, dy$$

$$\sum_k \frac{\operatorname{var}(X_k)}{k^2} \le \sum_k \frac{EY_k^2}{k^2} \le \sum_k \frac{1}{k^2} \int_0^{\infty} 2yP(|X| \ge y)1_{(y\le k)}\, dy$$

$$= \int_0^{\infty} \underbrace{\left(\sum_k \frac{1}{k^2} 1_{(y\le k)} 2y\right)}_{G(y)} P(|X| \ge y)\, dy$$

Claim: $G(y) \le 4$, for all $0 < y < \infty$. Since $G(y) \le \sum_k 1/k^2 \le 2$ for $y \le 1$, this is true for $y \le 1$. Take $y > 1$.

$$\frac{1}{k^2} \le \int_{k-1}^{k} \frac{1}{x^2}\, dx$$

so

$$\sum_k \frac{1}{k^2} 1_{(y \le k)} = \sum_{k \ge \lceil y \rceil} \frac{1}{k^2} \le \int_{\lceil y \rceil - 1}^{\infty} \frac{1}{x^2}\, dx = \frac{1}{\lceil y \rceil - 1}$$

Since $y > 1$,

$$G(y) \le \frac{2y}{\lceil y \rceil - 1} \le 4.$$

Then

$$\sum_k \frac{\mathrm{var}(X_k')}{k^2} \le 4 \int_0^{\infty} P(|X| \ge y)\, dy = 4E|X| < \infty$$

Apply 1.31 to (X_n'): $(1/n)\sum_{i=1}^n X_i' \to 0$ a.s., so $(1/n)\sum_{i=1}^n (Y_i - EY_i) \to 0$ a.s. Note that

$$EY_i = EX 1_{(|X| \le i)} \to EX$$

as $i \to \infty$. By dominated convergence, $(1/n)\sum_{i=1}^n (EY_i - EX) \to 0$ a.s. Add the two equations to get $(1/n)\sum_{i=1}^n (Y_i - EX) \to 0$ a.s., which implies that $(1/n)\sum_{i=1}^n Y_i \to EX$ a.s. □

1.4.2 Central Limit Theorem

Theorem 1.32 (IID Central Limit Theorem)**.** *Let $(X_i, i \ge 1)$ be IID, $EX = \mu$, $\mathrm{var}(X) = \sigma^2 < \infty$. Then,*

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathrm{Normal}(0, \sigma^2).$$

Proof. WLOG take $\mu = 0$. It is enough to show

$$\underbrace{\phi_{S_n/\sqrt{n}}(t)}_{\text{Left}} \to \exp\left(-\frac{\sigma^2 t^2}{2}\right).$$

Also,

$$\phi_{S_n/\sqrt{n}}(t) = \left(\phi_X\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(1 + \frac{n(\phi_X(t/\sqrt{n}) - 1)}{n}\right)^n.$$

It is enough to show $n(\phi_X(t/\sqrt{n}) - 1) \to \sigma^2 t^2/2$. The bound for $n = 2$ and $EX = 0$ is

$$\left|\phi_X(s) - \left(1 - \frac{s^2\sigma^2}{2}\right)\right| = o(s^2).$$

Then, with $s = t/\sqrt{n}$,

$$\text{Left} = n\left(\frac{t^2}{n}\frac{\sigma^2}{2} + o\left(\frac{t^2}{n}\right)\right) = \frac{t^2\sigma^2}{2} + n \cdot o\left(\frac{t^2}{n}\right) \to \frac{t^2\sigma^2}{2}.$$ □

1.4.3 Law of the Iterated Logarithm

Theorem 1.33 (Law of the Iterated Logarithm)**.** *Let $\{X_n\}$ be independent, identically distributed random variables with zero means and unit variances. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\limsup_{n \to \infty} \frac{|S_n|}{\sqrt{2n \log\log n}} = 1 \quad a.s.$$

1.5 Stochastic Processes

Definition 1.34 (Usual Hypothesis)**.** *$(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, P)$ is a probability space equipped with a filtration, satisfying the usual conditions:*

1. *$\mathscr{F}_{t_1} \subset \mathscr{F}_{t_2}$ for $t_1 \leq t_2$, and $\mathscr{F}_t \subset \mathscr{F}$;*
2. *$(\Omega, \mathscr{F}, P)$ is a complete probability space;*
3. *$\mathscr{F}_0$ contains all P-null sets in $\mathscr{F}$;*
4. *(Right-continuity) For all $t > 0$, $\mathscr{F}_t = \bigcap_{s>t} \mathscr{F}_s$.*

Definition 1.35 (Stopping Time)**.** *A random variable $T : \Omega \longrightarrow [0, \infty]$ is a stopping time of the filtration $(\mathscr{F}_t)$ if $\{T \leq t\} \in \mathscr{F}_t$, for every $t \geq 0$. The σ-field of the past before T is then defined by*

$$\mathscr{F}_T = \{A \in \mathscr{F}_\infty : \forall t \geq 0,\ A \cap \{T \leq t\} \in \mathscr{F}_t\}.$$

Remark 1.36. *Easy to verify that $\mathscr{F}_T$ is indeed a σ-field. We have $\mathbb{E}[X_\tau] = \mathbb{E}X_0$ for stopping time τ and martingale $X_\cdot$ under some assumptions.*

Definition 1.37 (Adaptedness)**.** *A stochastic process $(X_t)_{t\in T}$ is said to be adapted to the filtration $(\mathscr{F}_t)_{t\in T}$ if, for every $t \in T$, X_t is a random variable on $(\Omega, \mathscr{F}_t)$, i.e., for all $a \in \mathbb{R}$,*

$$\{\omega : X_t(\omega) \leq a\} \in \mathscr{F}_t.$$

1.5.1 Markov Chains

Definition 1.38 (Markov Process)**.** *Consider a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in T}, P)$, where $(X_t)_{t\in T}$ is an adapted stochastic process with respect to $(\mathscr{F}_t)_{t\in T}$. We say $(X_t)_{t\in T}$ is a Markov process with respect to $(\mathscr{F}_t)_{t\in T}$ if for every bounded Borel function f and every $t > s$,*

$$E[f(X_t) \mid \mathscr{F}_s] = E[f(X_t) \mid \sigma(X_s)]\Big(= E[f(X_t) \mid X_s]\Big).$$

Remark 1.39. *Define $p(s, x; t, y)$ as the transition probability function. Easy to get Chapman-Kolmogorov Equation:*

$$\int_{\mathbb{R}^n} f(y)P(s, x; t, \mathrm{d}y) = \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} f(y)P(r, z; t, \mathrm{d}y) \right) P(s, x; r, \mathrm{d}z).$$

Theorem 1.40 (The MC Convergence Theorem)**.** *Suppose the chain is **irreducible** and **positive-recurrent**, so the stationary π exists. If the chain is also aperiodic, then $\mathbb{P}_{\mu_0}(X_n = j) \xrightarrow{n\to\infty} \pi(j)\ \forall j\ \forall \mu_0$.*

Using SLLN, we can easily get the following results.

Theorem 1.41 (Ergodic Theorem for Markov Chains)**.** *Consider an irreducible, positive-recurrent MC. Let π be the stationary distribution. Take $f : S \to \mathbb{R}$ such that $\sum_x \pi(x)|f(x)| < \infty$. Then,*

$$\frac{1}{t}\sum_{n=1}^{t} f(X_n) \xrightarrow{a.s.} \bar{f} := \sum_x \pi(x)f(x) \qquad \text{as } t \to \infty.$$

Remark 1.42. *Besides, some interesting techniques in the analysis of Markov processes are deeply connected to combinatorics. Using*

$$\sum_{n\geq 0}\frac{\binom{2n}{n}}{2^{2n}}\asymp\sum_{n\geq 0}\frac{1}{\sqrt{n}}\to\infty,\quad \sum_{n\geq 0}\frac{\binom{2n}{n}^2}{4^{2n}}\asymp\sum_{n\geq 0}\frac{1}{n}\to\infty,$$

we can get 1D, 2D random walks are recurrent. Another example is more complicated.

Example 1.43 (Catalan Number)**.** *Consider a simple symmetric random walk (S_n) starting at $S_0=0$. Let $\tau_i=\inf\{n\geq 1: S_n=i\}$. We derive the number of paths of length $2n$ from 0 to 0 that stay ≥ 0.*

For $n+i$ even, the set $\{S_n=i\}\setminus\{\tau_i=n\}=\{\tau_i<n,\, S_n=i\}$ consists of paths that visited i before time n. Splitting by S_{n-1} and applying reflection at time τ_i, the contributions from $S_{n-1}=i+1$ and $S_{n-1}=i-1$ are equal, giving

$$P(\tau_i<n,\, S_n=i)=P(S_{n-1}=i+1)=\frac{n-i}{n}\,P(S_n=i).$$

Therefore

$$P(\tau_i=n)=\frac{i}{n}\,P(S_n=i).$$

Set $i=1$ and $n=2k+1$. The event $\{\tau_1=2k+1\}$ forces $S_1,\dots,S_{2k}\leq 0$ with $S_{2k}=0$ (the last step is necessarily $+1$). The number of such paths of the first $2k$ steps is

$$\frac{1}{2k+1}\binom{2k+1}{k+1}.$$

Reflecting $S\mapsto -S$ counts instead the paths from 0 to 0 of length $2k$ staying ≥ 0. Simplifying:

$$\frac{1}{2k+1}\binom{2k+1}{k+1}=\frac{(2k)!}{(k+1)!\,k!}=\frac{1}{k+1}\binom{2k}{k}=C_k.$$

Another interesting topic is **Continuous-Time Markov Chains**. The continuous-time viewpoint is important in many problems, as it brings in the tools of ordinary differential equations.

Example 1.44 (Poisson Processes)**.** *A counting process $(X_t)_{t\geq 0}$ with $X_0=0$ is a* Poisson process of rate λ *if it has independent, stationary increments and*

$$\mathbb{P}\{X_{t+\Delta t}-X_t=1\}=\lambda\Delta t+o(\Delta t),\qquad \mathbb{P}\{X_{t+\Delta t}-X_t\geq 2\}=o(\Delta t).$$

Approach 1 (binomial limit) *Write $X_t=\sum_{j=1}^{n}(X_{jt/n}-X_{(j-1)t/n})$. For large n the summands are independent* Bernoulli$(\lambda t/n)$*, so*

$$\mathbb{P}\{X_t=k\}=\lim_{n\to\infty}\binom{n}{k}\Big(\frac{\lambda t}{n}\Big)^k\Big(1-\frac{\lambda t}{n}\Big)^{n-k}=e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

Approach 2 (ODE) *Set $P_k(t)=\mathbb{P}\{X_t=k\}$. The defining conditions give*

$$P_k'(t)=\lambda P_{k-1}(t)-\lambda P_k(t).$$

Substituting $f_k(t)=e^{\lambda t}P_k(t)$ reduces this to $f_k'=\lambda f_{k-1}$, $f_k(0)=0$. Since $f_0\equiv 1$, induction yields $f_k(t)=(\lambda t)^k/k!$.

Approach 3 (interarrival times) *Let T_n be the time between the $(n-1)$th and nth arrival. The T_i are i.i.d. and memoryless: $\mathbb{P}\{T_i \geq s+t \mid T_i \geq s\} = \mathbb{P}\{T_i \geq t\}$, so $T_i \sim \text{Exp}(\lambda)$. (Match rates: $\mathbb{E}[Y_n] = n/\lambda$ should give $\approx \lambda t$ arrivals by time t; or directly, $\mathbb{P}\{T_1 > t\} = \mathbb{P}\{X_t = 0\} = e^{-\lambda t}$.) Then $Y_n = T_1 + \cdots + T_n \sim \text{Gamma}(n, \lambda)$, and*

$$\mathbb{P}\{X_t = k\} = \mathbb{P}\{Y_k \leq t < Y_{k+1}\} = \int_0^t \frac{\lambda^k s^{k-1}}{(k-1)!} e^{-\lambda s} \cdot \lambda e^{-\lambda(t-s)}\, ds = e^{-\lambda t} \frac{(\lambda t)^k}{k!}.$$

1.5.2 Martingale

Definition 1.45 (Martingale)**.** *Let $(X_t)_{t\in T}$ be an adapted stochastic process on $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in T}, P)$. If for all $s < t$,*

$$E[X_t \mid \mathscr{F}_s] = X_s,$$

then $(X_t)_{t\in T}$ is called a martingale with respect to $(\mathscr{F}_t)_{t\in T}$, or $(X_t, \mathscr{F}_t)_{t\in T}$ is called a martingale. When the filtration $(\mathscr{F}_t)_{t\in T}$ is clear from context, we simply call $(X_t)_{t\in T}$ a martingale.

If $E[X_t \mid \mathscr{F}_s] \geq X_s$, then $(X_t, \mathscr{F}_t)_{t\in T}$ is called a submartingale; if $E[X_t \mid \mathscr{F}_s] \leq X_s$, then $(X_t, \mathscr{F}_t)_{t\in T}$ is called a supermartingale.

Definition 1.46 (Doob-Meyer Decomposition, rough version)**.** *Let $(X_t)_{t\geq 0}$ be an $(\mathscr{F}_t)_{t\geq 0}$-submartingale with RCLL paths. If $\{X_t\}_{t\geq 0}$ satisfies certain regularity conditions, then $(X_t)_{t\geq 0}$ can be uniquely decomposed as*

$$X_t = M_t + C_t, \quad t \geq 0$$

where $(M_t)_{t\geq 0}$ is an $(\mathscr{F}_t)_{t\geq 0}$-martingale, and $(C_t)_{t\geq 0}$ is a predictable, right-continuous, increasing process with $E\, C_t < \infty$, $\forall t \geq 0$, and $C_0 = 0$. $(C_t)_{t\geq 0}$ is called the compensator of $(X_t)_{t\geq 0}$.

Remark 1.47. *Easy to verify that if M_t is a martingale, then M_t^2 is a submartingale. Use $\langle M \rangle_t$ to denote the C_t and define the stochastic integral on general continuous martingale.*

A useful result (which we also use for the proof of SLLN) is as follows.

Theorem 1.48 (Doob's Martingale Maximal Inequality)**.** *Let $(\xi_t, \mathscr{F}_t)_{t\geq 0}$ be a martingale with RCLL paths (right-continuous with left limits). If for some $p \geq 1$ and all $t \geq 0$, $E|\xi_t|^p < \infty$, then for any $T < \infty$ and $a > 0$,*

$$P\left(\sup_{t\in[0,T]} |\xi_t| \geq a\right) \leq \frac{E|\xi_T|^p}{a^p}.$$

For any $a, b > 0$,

$$P\left(\sup_{t\in[0,T]} \xi_t \geq a\right) \leq \frac{E\, e^{b\xi_T}}{e^{ba}}.$$

Theorem 1.49 (Doob's Submartingale Convergence Theorem)**.** *Let $(\xi_n, \mathscr{F}_n)_{n\geq 0}$ be a submartingale with $\sup_n E|\xi_n| < \infty$. Then $\lim_{n\to\infty} \xi_n = \xi_\infty$ almost everywhere, and $E|\xi_\infty| < \infty$.*

1.5.3 Itô Calculus

Definition 1.50 (Brownian Motion)**.** *Let $(B_t)_{t\geq 0}$ be a stochastic process. If it satisfies the following three conditions:*

1. *$B_{t+s} - B_t$ follows a normal distribution $N(0, \sigma^2 s)$;*
2. *For any $0 < t_1 < t_2 < \cdots < t_n$, $B_0, B_{t_1} - B_0, \cdots, B_{t_n} - B_{t_{n-1}}$ are mutually independent;*
3. *For all ω, $B_t(\omega)$ is continuous in t.*

Then $(B_t)_{t\geq 0}$ is called a (one-dimensional) Brownian motion. In these notes, we always assume that the Brownian motion satisfies $\sigma^2 = 1$; when $B_0 = 0$, it is called a standard Brownian motion.

Remark 1.51 (History of Brownian Motion)**.** ***Einstein's derivation (1905).*** *Consider n particles in a fluid. Assume:*

1. *particles move independently;*
2. *increments over non-overlapping time intervals are independent.*

Let $\varphi(\Delta)$ be the symmetric density of displacement over a time interval τ, with $\varphi(\Delta) = \varphi(-\Delta)$.

Let $f(x,t)$ be the particle density. Then

$$f(x, t+\tau) = \int_{-\infty}^{+\infty} f(x-\Delta, t)\,\varphi(\Delta)\,\mathrm{d}\Delta.$$

Expand both sides: the left in τ, the right in Δ. By symmetry of φ, odd-order terms vanish. Keeping only leading terms and setting $D = \frac{1}{\tau}\int \frac{\Delta^2}{2}\,\varphi(\Delta)\,\mathrm{d}\Delta$, we obtain

$$\frac{\partial f}{\partial t} = D\frac{\partial^2 f}{\partial x^2}.$$

With initial condition $f(x,0) = n\,\delta(x)$, the solution is

$$f(x,t) = \frac{n}{\sqrt{4\pi D t}}\, e^{-\frac{x^2}{4Dt}}.$$

The ratio $f(x,t)/n$ is precisely the density of a Brownian motion. In particular, the root-mean-square displacement scales as $\sqrt{2Dt} \sim \sqrt{t}$.

Proposition 1.52 (Kolmogorov backward/Forward equation)**.** *Let $\{p(t,x,y),\, x \in \mathbb{R}^n,\, y \in \mathbb{R}^n\}$ be the transition density of an n-dimensional standard Brownian motion. Then*

$$\frac{\partial p(t,x,y)}{\partial t} = \frac{1}{2}\Delta_x p(t,x,y) = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2 p(t,x,y)}{\partial x_i^2} \tag{1}$$

$$\frac{\partial p(t,x,y)}{\partial t} = \frac{1}{2}\Delta_y p(t,x,y) = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2 p(t,x,y)}{\partial y_i^2} \tag{2}$$

Equation 1 is called the Kolmogorov backward equation (the first equation), and 2 is called the Kolmogorov forward equation (the second equation), also known as the Fokker–Planck equation.

Proposition 1.53 (Martingale Properties)**.** *For a standard $(\mathscr{F}_t)_{t\geq 0}$-Brownian motion $(B_t)_{t\geq 0}$,*

1. *$(B_t)_{t\geq 0}$ is an $(\mathscr{F}_t)_{t\geq 0}$-martingale;*
2. *$(B_t^2 - t)_{t\geq 0}$ is an $(\mathscr{F}_t)_{t\geq 0}$-martingale;*
3. *$(z_t = e^{cB_t - \frac{c^2 t}{2}})_{t\geq 0}$ is an $(\mathscr{F}_t)_{t\geq 0}$-martingale.*

Remark 1.54. *Based on the martingale properties, we define stochastic integral using $\mathrm{d}B_t$, which is well-known Itô Integral.*

Definition 1.55 (Itô Integral)**.** *For a stochastic process $(\phi_t)_{t\geq 0}$ in $\mathscr{L}_T^2$, let $(\phi_t^{(n)})_{t\geq 0}$ be a sequence of predictable simple (step) processes approximating it with respect to $(\mathscr{F}_t)_{t\geq 0}$, satisfying*

$$\|\phi^{(n)} - \phi\|^2 = E\int_0^T \left|\phi_t^{(n)} - \phi_t\right|^2 \mathrm{d}t \to 0, \quad \text{as } n \to \infty.$$

(At this point $\int_0^T \phi_t^{(n)}\,\mathrm{d}B_t$ is a Cauchy sequence in $L^2(\Omega)$.) Denote the limit of $\int_0^T \phi_t^{(n)}\,\mathrm{d}B_t$ in $L^2(\Omega)$ by $\int_0^T \phi_t\,\mathrm{d}B_t$, i.e.,

$$\int_0^T \phi_t\,\mathrm{d}B_t := L^2(\Omega)\text{-}\lim_{n\to\infty}\int_0^T \phi_t^{(n)}\,\mathrm{d}B_t,$$

which is called the Itô stochastic integral of $(\phi_t)_{0\le t\le T}$ with respect to $(B_t)_{t\ge 0}$.

Remark 1.56. *Different choices of representative points will influence the result. We use left endpoints in the definition of the Itô integral. If we use midpoints instead, we obtain the Stratonovich integral.*

Proposition 1.57 (Properties of Stochastic Integrals)**.** .

1. *If τ is an $(\mathscr{F}_t)_{t\ge 0}$-stopping time with $\tau \le T$, then*

$$\int_0^\tau \phi_t\,\mathrm{d}B_t = \int_0^T \phi_t \mathbf{1}_{(0,\tau]}(t)\,\mathrm{d}B_t.$$

2. *$\left\{\xi_t \overset{\text{def}}{=} \int_0^t \phi_s\,\mathrm{d}B_s\right\}$ is an $(\mathscr{F}_t)_{t\in[0,T]}$-martingale.*

3. *$\left\{\eta_t \overset{\text{def}}{=} \left(\int_0^t \phi_s\,\mathrm{d}B_s\right)^2 - \int_0^t \phi_s^2\,\mathrm{d}s\right\}$ is an $(\mathscr{F}_t)_{t\in[0,T]}$-martingale, and*

$$E\left[\left(\int_s^t \phi_u\,\mathrm{d}B_u\right)^2 \middle| \mathscr{F}_s\right] = E\left[\int_s^t \phi_u^2\,\mathrm{d}u \middle| \mathscr{F}_s\right].$$

4. *If $(\phi_t)_{t\in[0,T]}$ is a bounded adapted process with respect to $(\mathscr{F}_t)_{t\in[0,T]}$, i.e., $\forall t$, $|\phi_t| < M < \infty$, then*

$$\left\{\zeta_t \overset{\text{def}}{=} e^{\int_0^t \phi_s\,\mathrm{d}B_s - \frac{1}{2}\int_0^t \phi_s^2\,\mathrm{d}s}\right\}$$

is an $(\mathscr{F}_t)_{t\in[0,T]}$-martingale.

Theorem 1.58 (Itô's Formula)**.** *Let $F(t,x)$ be once continuously differentiable in t and twice continuously differentiable in x. Then*

$$\begin{aligned} F(t,B_t) - F(0,B_0) &= \int_0^t \frac{\partial F(s,B_s)}{\partial s}\,\mathrm{d}s + \int_0^t \frac{\partial F(s,B_s)}{\partial x}\,\mathrm{d}B_s + \frac{1}{2}\int_0^t \frac{\partial^2 F(s,B_s)}{\partial x^2}\,\mathrm{d}s \\ &= \int_0^t F_t'(s,B_s)\,\mathrm{d}s + \int_0^t F_x'(s,B_s)\,\mathrm{d}B_s + \frac{1}{2}\int_0^t F_{xx}''(s,B_s)\,\mathrm{d}s. \end{aligned}$$

In differential form, this is written as

$$\mathrm{d}F(t,B_t) = F_t'(t,B_t)\,\mathrm{d}t + \frac{1}{2}F_{xx}''(t,B_t)\,\mathrm{d}t + F_x'(t,B_t)\,\mathrm{d}B_t.$$

Remark 1.59. *A useful trick to remember:*

$$(\mathrm{d}B_t)^2 = \mathrm{d}t.$$

1.5.4 Stochastic Differential Equations

At last, we talk about Stochastic Differential Equations (SDE). Consider a general form:

$$\mathrm{d}\xi_t = \mathbf{b}(t, \xi_t)\,\mathrm{d}t + \Sigma(t, \xi_t)\,\mathrm{d}\mathbf{B}_t.$$

Theorem 1.60 (Feynman-Kac Formula)**.** *Let*

$$\mathrm{d}\xi_t = \mathbf{b}(\xi_t)\,\mathrm{d}t + \Sigma(\xi_t)\,\mathrm{d}\mathbf{B}_t \tag{3}$$

and denote $\mathscr{L} = \frac{1}{2}\sum_{i,j=1}^n a_{ij}\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^n b_i(x)\frac{\partial}{\partial x_i}$*, where* $(a_{ij})_{i,j=1,\cdots,n} = \Sigma\Sigma^\top$*. Let* $f \in C_0^2(\mathbb{R}^n)$*,* $q(x) \in C(\mathbb{R}^n)$*, and* $q(x)$ *be bounded below.*

1. *Let* $v(t,x) = E\left(f(\xi_t)\exp\left(-\int_0^t q(\xi_s)\,\mathrm{d}s\right)\middle|\xi_0 = x\right)$*. Then*

$$\begin{cases} \dfrac{\partial v}{\partial t} = \mathscr{L}v - qv & t > 0,\ x \in \mathbb{R}^n \\ v(0,x) = f(x) & x \in \mathbb{R}^n \end{cases}.$$

2. *If* $w(t,x) \in C^{1,2}(\mathbb{R}\times\mathbb{R}^n)$ *is bounded on* $K \times \mathbb{R}^n$*,* $K \subset \mathbb{R}$*, and* w *is a solution to* (1)*, then* $w(t,x) = v(t,x)$*.*

Example 1.61 (Diffusion Model)**.** *Consider the forward OU process*

$$\mathrm{d}x_t = -\tfrac{1}{2}x_t\,\mathrm{d}t + \mathrm{d}B_t, \quad x_0 \sim p_0.$$

A direct computation gives $x_t|x_0 \sim N(\alpha_t x_0,\, \sigma_t^2 I)$ *where* $\alpha_t = e^{-t/2}$ *and* $\sigma_t^2 = 1 - e^{-t}$*. As* $t \to \infty$*, the distribution* p_t *converges exponentially to* $N(0, I)$ *— structure is destroyed and replaced by pure noise.*

By the Fokker–Planck equation, the density $p(x,t)$ *satisfies*

$$\frac{\partial p}{\partial t} = \nabla\cdot\left(\tfrac{1}{2}x\,p\right) + \tfrac{1}{2}\Delta p = \tfrac{1}{2}\nabla\cdot(x\,p + \nabla p)\,.$$

Reverse SDE. *Anderson (1982) showed that for a general forward SDE* $\mathrm{d}x_t = f(x,t)\,\mathrm{d}t + g(t)\,\mathrm{d}B_t$*, a reverse process is given by*

$$\mathrm{d}\tilde{x}_t = \left[f(\tilde{x}_t, t) - g^2(t)\,\nabla_x \log p(\tilde{x}_t, t)\right]\mathrm{d}t + g(t)\,\mathrm{d}\bar{B}_t, \quad t : T \to 0,$$

where $\bar{B}_t$ *is a backward Brownian motion.*

Proof. The forward SDE $\mathrm{d}x_t = f\,\mathrm{d}t + g\,\mathrm{d}B_t$ has Fokker–Planck equation

$$\frac{\partial p}{\partial t} = -\nabla\cdot(f\,p) + \tfrac{1}{2}g^2\Delta p.$$

Set $s = T - t$ and $q(x,s) = p(x, T-s)$. Since $\partial q/\partial s = -\partial p/\partial t$, the density q must satisfy

$$\frac{\partial q}{\partial s} = \nabla\cdot(f\,p) - \tfrac{1}{2}g^2\Delta p. \tag{$\star$}$$

Now we verify that the proposed reverse SDE, with drift $h = -f + g^2\nabla_x \log p$ and diffusion g, reproduces ($\star$). Its Fokker–Planck equation in s reads $\partial q/\partial s = -\nabla\cdot(h\,q) + \frac{1}{2}g^2\Delta q$. Substituting $q = p$ and using $p\,\nabla\log p = \nabla p$:

$$-\nabla\cdot(h\,p) + \tfrac{1}{2}g^2\Delta p = \nabla\cdot(f\,p) - \underbrace{\nabla\cdot(g^2\nabla p)}_{=\,g^2\Delta p} + \tfrac{1}{2}g^2\Delta p = \nabla\cdot(f\,p) - \tfrac{1}{2}g^2\Delta p,$$

which matches ($\star$). □

For our OU process, $f(x,t) = -x/2$ and $g = 1$, so the reverse SDE reads

$$\mathrm{d}\tilde{x}_t = \left[-\tfrac{1}{2}\tilde{x}_t - \nabla_x \log p(\tilde{x}_t, t)\right] \mathrm{d}t + \mathrm{d}\bar{B}_t.$$

If we knew the score function $\nabla_x \log p(x,t)$, we could simulate this backward from $\tilde{x}_T \sim N(0, I)$ and recover samples from p_0. The entire problem reduces to ***learning the score****.*

Denoising score matching *We want to minimize $L_t(\theta) = \mathbb{E}_{x\sim p_t}\left[\|s_\theta(x,t) - \nabla_x \log p(x,t)\|^2\right]$, but $\nabla_x \log p(x,t)$ is unknown. The key trick is integration by parts. Expand the square and note that the cross term satisfies*

$$\int \langle s_\theta,\, \nabla_x p\rangle \,\mathrm{d}x = -\int (\nabla \cdot s_\theta)\, p \,\mathrm{d}x.$$

Now use $p(x,t) = \int p_t(x|x_0)\, p_0(x_0)\,\mathrm{d}x_0$ and apply integration by parts again on the inner integral. After recombining, we get

$$L_t(\theta) = \mathbb{E}_{x_0\sim p_0}\, \mathbb{E}_{x_t|x_0}\left[\|s_\theta(x_t,t) - \nabla_{x_t} \log p_t(x_t|x_0)\|^2\right] + C,$$

where C is independent of θ. The crucial point: the intractable marginal score $\nabla \log p$ has been replaced by the conditional score $\nabla \log p_t(\cdot|x_0)$, which is explicitly computable. Since $x_t|x_0 \sim N(\alpha_t x_0,\, \sigma_t^2 I)$, we have $\nabla_{x_t} \log p_t(x_t|x_0) = -(x_t - \alpha_t x_0)/\sigma_t^2 = -\xi_t/\sigma_t$ where $x_t = \alpha_t x_0 + \sigma_t \xi_t$ and $\xi_t \sim N(0,I)$. The final training objective is therefore

$$\min_\theta\ \mathbb{E}_t\, \mathbb{E}_{x_0}\, \mathbb{E}_{\xi_t}\left[\left\| s_\theta(\alpha_t x_0 + \sigma_t \xi_t,\, t) - \frac{\xi_t}{\sigma_t}\right\|^2\right],$$

which only requires samples from p_0 and standard Gaussians.

2 Parameter Estimation

In this section, we study one of the central topics in statistics: parameter estimation. This is the beginning of our tour! The section is organized as follows:

1. Fundamentals
2. Methods of Estimation
3. Large Sample Theory

This section mainly follows STAT210A (UC Berkeley), STAT300A (Stanford University), STAT300B (Stanford University) and Mathematics Statistics (Peking University, taught by Fang Yao). I also refered to the book Theoretical Statistics [2].

2.1 Fundamentals

Not all data is relevant to a particular decision problem.

Definition 2.1 (Statistic)**.** *A statistic $T : \mathscr{X} \to \mathscr{T}$ is a function of the data.*

Definition 2.2 (Sufficient Statistic)**.** *A statistic is sufficient for a model $\mathscr{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$ if for all t, the conditional distribution $X \mid T(x) = t$ does not depend on θ.*

Theorem 2.3 (Neyman-Fisher Factorization Criterion (NFFC))**.** *Suppose each $\mathbb{P}_\theta \in \mathscr{P}$ has density $p(x;\theta)$ w.r.t. a common σ-finite measure μ, i.e., $\frac{d\mathbb{P}_\theta}{d\mu} = p(x;\theta)$. Then $T(X)$ is sufficient if and only if $p(x;\theta) = g_\theta(T(x))h(x)$ for some g_θ, h.*

Example 2.4. *The model $\{\mathbb{P}_\theta : \theta \in \Omega\}$ forms an **s-dimensional exponential family** if each $\mathbb{P}_\theta$ has density of the form:*

$$p(x;\theta) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) h(x)$$

- *$\eta_i(\theta) \in \mathbb{R}$ are called the **natural parameters**.*
- *$T_i(x) \in \mathbb{R}$ are its **sufficient statistics**, which follows from **NFFC**.*
- *$B(\theta)$ is the log-partition function because it is the logarithm of a normalization factor:*

$$B(\theta) = \log\left(\int \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x)\right) h(x)\, d\mu(x)\right) \in \mathbb{R}$$

- *$h(x) \in \mathbb{R}$: base measure.*

Exponential families are of particular interest to us, because many common distributions are exponential families (e.g., Normal, Binomial, and Poisson)

Definition 2.5 (Minimal Sufficiency)**.** *A sufficient statistic T is **minimal** if for every sufficient statistic T' and for every $x, y \in \mathscr{X}$, $T(x) = T(y)$ whenever $T'(x) = T'(y)$. In other words, T is a function of T' (there exists f such that $T(x) = f(T'(x))$ for any $x \in \mathscr{X}$).*

Theorem 2.6. *Let $\{p(x;\theta), \theta \in \Omega\}$ be a family of densities with respect to some measure μ. Suppose that there exists a statistic T such that for every $x, y \in \mathscr{X}$:*

$$p(x;\theta) = C_{x,y} p(y;\theta) \quad \Longleftrightarrow \quad T(x) = T(y)$$

for every θ and some $C_{x,y} \in \mathbb{R}$. Then T is a minimal sufficient statistic.

Definition 2.7 (Ancillary). *A statistic A is **ancillary** for $X \sim \mathbb{P}_\theta \in \mathscr{P}$ if the distribution of $A(X)$ does not depend on θ.*

Definition 2.8 (First-Order Ancillary). *A statistic A is **first-order ancillary** for $X \sim \mathbb{P}_\theta \in \mathscr{P}$ if $\mathbb{E}_\theta[A(X)]$ does not depend on θ.*

From this we define the concept of complete statistics.

Definition 2.9 (Complete Statistic). *A statistic T is **complete** for $X \sim \mathbb{P}_\theta \in \mathscr{P}$ if no non-constant function of T is first-order ancillary. In other words, if $\mathbb{E}_\theta[f(T(X))] = 0$ for all θ, then $f(T(X)) = 0$ with probability 1 for all θ.*

Remark 2.10. *Every last bit of information has been mined.*

Theorem 2.11 (Complete Statistics for Exponential Family). *$(T_1, \ldots, T_s)$ is complete for any s-dimensional full rank exponential family.*

Theorem 2.12 (Basu's Theorem). *If T is complete and sufficient for $\mathscr{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, and A is ancillary then $T(X) \perp\!\!\!\perp A(X)$.*

Example 2.13. $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ *(μ, σ^2 both unknown).*
Claim: $\bar{X} \perp\!\!\!\perp \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

Proof. Fix any $\sigma > 0$, and consider the submodel $\mathscr{P}_\sigma = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$. In each submodel, $\bar{X}$ is complete and sufficient, and $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is ancillary. By Basu's Theorem, $\bar{X} \perp\!\!\!\perp \sum_{i=1}^n (X_i - \bar{X})^2$ under $\mathcal{N}(\mu, \sigma^2)$ for any μ. Since σ is arbitrary, we have $\bar{X} \perp\!\!\!\perp \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ for the full model. □

2.2 Methods of Estimation

Now we introduce loss function to describe the quality of estimators quantitatively. Consider θ as the decision rule, and $R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(X))$ as the risk function.

Theorem 2.14 (Rao-Blackwell Theorem, K 3.28). *Suppose that T is sufficient for $\mathbb{P} = \{\mathbb{P}_\theta, \theta \in \Omega\}$, that $\delta(X)$ is an estimator for $g(\theta)$ for which $\mathbb{E}(\delta(X))$ exists, and that $R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(X)) < \infty$. If $L(\theta, \cdot)$ is convex, then*

$$R(\theta, \eta) \le R(\theta, \delta) \quad \text{for} \quad \eta(T(X)) = \mathbb{E}(\delta(X) \mid T(X)).$$

If $L(\theta, \cdot)$ is strictly convex, then $R(\theta, \eta) < R(\theta, \delta)$ for any θ unless $\eta(T'(x)) = \delta$ w.p. 1.

2.2.1 Unbiased Estimation

An estimator is unbiased if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$ (to estimate). Although uniformly best estimator does not exist, quite often, we can find a unbiased estimator with uniformly minimum risk, that is, an unbiased δ satisfying $R(\theta, \delta) \le R(\theta, \delta')$, $\forall \theta$ and any other unbiased estimators δ'. Such an estimator is called a **uniformly minimum risk unbiased estimator (UMRUE)**.

Consider $L(\theta, d) = (\theta - d)^2$, then an UMRUE becomes a **uniformly minimum variance unbiased estimator (UMVUE)**.

$$\begin{aligned}\mathbb{E}_\theta\left[(\theta - \delta(X))^2\right] &= \left(\mathbb{E}_\theta[\delta(X)] - \theta\right)^2 + \mathbb{E}_\theta\left\{\left(\delta(X) - \mathbb{E}_\theta[\delta(X)]\right)^2\right\} \\ &= \text{Bias}^2 + \text{Variance}.\end{aligned}$$

Theorem 2.15 (Lehmann-Scheffe Theorem). *If T is a complete and sufficient statistic, and $\mathbb{E}_\theta[h(T(X))] = g(\theta)$ (i.e., $h(T(x))$ is unbiased for $g(\theta)$), then $h(T(X))$ is*

1. *the only function of $T(X)$ that is unbiased for $g(\theta)$*
2. *an UMRUE under any convex loss function,*
3. *the unique UMRUE (up to a $\mathscr{P}$–null set) under any strictly convex loss function,*
4. *the unique UMVUE (up to a $\mathscr{P}$–null set).*

Example 2.16. *Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. First, note that if σ^2 is known, $\bar{X}$ is a complete sufficient statistic for μ and hence also the UMVUE. Consider the case when $\theta = (\mu, \sigma^2)$ is unknown.*

(a) The UMVUE for μ is $\bar{X}$.

(b) The UMVUE for σ^2 is $\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$.

(c) What is the UMVUE for σ? First, note that $X_i - \bar{X} \sim \mathcal{N}(0, \frac{n-1}{n}\sigma^2)$, and hence $\mathbb{E}[|X_i - \bar{X}|] = \sigma\sqrt{\frac{2}{\pi}}\sqrt{\frac{n-1}{n}}$. This implies

$$\frac{\sqrt{\pi n}}{\sqrt{2(n-1)}}|X_i - \bar{X}|$$

is unbiased for σ. At this point we could Rao-Blackwellize, but the math is messy. Instead, we will try to stumble upon the solution. Let

$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2.$$

We know that

$$S^2 \sim \sigma^2 \chi^2_{n-1}.$$

Thus,

$$\mathbb{E}(S) = \sigma \mathbb{E}(\chi_{n-1}).$$

Which in turn implies that

$$\frac{\mathbb{E}(S)}{\mathbb{E}(\chi_{n-1})} = \sigma,$$

meaning $\frac{S}{\mathbb{E}(\chi_{n-1})}$ is unbiased for σ and hence UMVU.

(d) What is the UMVUE for μ^2? Taking the expectation of the UMVUE for μ and squaring it yields

$$\mathbb{E}(\bar{X}^2) = \mu^2 + \sigma^2/n.$$

So,

$$\delta_n(X) = \bar{X}^2 - \frac{S^2}{n(n-1)}$$

is the UMVUE. Note that $\delta_n(X)$ may be negative even though it estimates a non-negative quantity. Indeed, δ_n is inadmissible and dominated by the biased estimator $\max(0, \delta_n(X))$.

If we know some prior knowledge of the estimators (eg: Equivariant, Invariant), we can design better estimators.

2.2.2 Bayes Estimators

Our optimality goal, given a measure Λ, is to find an estimator δ_Λ which minimizes the **average risk**,

$$r(\Lambda, \delta) = \int R(\theta, \delta)\, d\Lambda(\theta).$$

If Λ is a probability distribution on Ω, we call Λ the **prior** distribution. The estimator δ_Λ, if it exists, is called the **Bayes estimator** with respect to Λ, and the minimized average risk $r(\Lambda, \delta_\Lambda)$ is called the **Bayes risk**.

Theorem 2.17 (Bayes Estimators)**.** *Suppose $\Theta \sim \Lambda$, and $X|\Theta = \theta \sim P_\theta$. If*

1. *there exists δ_0 an estimator of $g(\theta)$ with finite risk for all θ, and*
2. *there exists a value $\delta_\Lambda(x)$ that minimizes*

$$\mathbb{E}\left[L(\Theta, \delta_\Lambda(X)) \mid X = x\right] \quad \text{for almost every } x,$$

then δ_Λ is a Bayes estimator with respect to Λ.

Example 2.18. *Suppose that $X \sim Bin(n, \theta)$ given $\Theta = \theta$ and that Θ has prior distribution $Beta(a, b)$. The prior density is given by*

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \mathbb{I}(0 < \theta < 1)$$

*The **likelihood** (model density) is given by*

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

The marginal density is given by

$$f(x) = \int f(x \mid \theta)\pi(\theta)\, d\theta.$$

The posterior density may be calculated using Bayes rule which states that

$$\textit{posterior} = \frac{\textit{joint}}{\textit{marginal}} = \frac{\textit{prior} \cdot \textit{likelihood}}{\textit{marginal}}.$$

In our notation, the posterior density is given by the formula

$$\begin{aligned} \pi(\theta \mid x) &= \frac{\pi(\theta) f(x \mid \theta)}{f(x)} \\ &= \frac{\pi(\theta) f(x \mid \theta)}{\int \pi(\theta') f(x \mid \theta')\, d\theta'} \end{aligned}$$

$$\begin{aligned} \pi(\theta \mid x) &\propto \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \sim Beta(x+a, n-x+b). \end{aligned}$$

Hence, the Bayes estimator of θ under the squared error loss is given by

$$\mathbb{E}[\Theta \mid X = x] = \frac{x+a}{n+a+b}.$$

This posterior mean may be expressed as

$$\frac{X+a}{n+a+b} = \frac{n}{n+a+b}\left(\frac{X}{n}\right) + \frac{a+b}{n+a+b}\left(\frac{a}{a+b}\right).$$

Hence, the Bayes estimate is a ***convex combination*** *of the sample proportion X/n (which is the UMVUE) and the prior mean $a/(a+b)$. Thus, the Bayes estimate modifies the sample estimate in light of prior information by* ***"shrinking" the sample estimate towards the prior mean****. (This is a commonly recurring property of Bayes estimators.) In addition, as the sample size n tends to infinity, the weight of the prior mean tends to zero, the empirical evidence increasingly outweighs the prior information, and the posterior mean becomes less distinguishable from the sample proportion.*

Theorem 2.19 (Unbiased Estimators)**.** *If δ is unbiased for $g(\theta)$ with $r(\Lambda, \delta) < \infty$ and $\mathbb{E}[g(\Theta)^2] < \infty$ then δ is not Bayes under squared error loss unless its average risk is zero, i.e.,*

$$\mathbb{E}[(\delta(X) - g(\Theta))^2] = 0,$$

where the expectation is taken over X and Θ.

Example 2.20. *Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, with $\sigma^2 > 0$ known. Is $\bar{X}$ Bayes under squared error for some choice of prior distribution? We know that, $\mathbb{E}(\bar{X} \mid \theta) = \theta$, i.e., $\bar{X}$ is an unbiased estimator of θ. Further, we have that the average risk under squared error,*

$$\mathbb{E}[(\bar{X} - \Theta)^2] = \frac{\sigma^2}{n} \neq 0,$$

which means that $\bar{X}$ is not the Bayes estimator under any prior distribution!

Definition 2.21 (Admissible)**.** *We say a policy δ is inadmissible if there exists a policy δ^* s.t. for all $\theta \in \mathscr{S}$, we have $R(\theta, \delta^*) \leq R(\theta, \delta)$ and that $\exists \theta \in \mathscr{S}$, the inequality holds strictly.*

Theorem 2.22 (Admissable)**.** *A unique Bayes estimator (a.s. for all P_θ) is admissible.*

Remark 2.23 (Why Consider Bayes Estimators)**.** *So why consider Bayes estimators? (1) All admissible estimators are limits of Bayes estimators (Wald, 1949). (2) Allow us to incorporate relevant prior information and experience into our estimators. (3) A general method for generating reasonable estimators under various optimality criteria.*

Remark 2.24 (How to Choose Priors)**.**

1. *Subjective. If prior knowledge about or experience with a model parameter is available, we can incorporate this information into the prior choice.*
2. *Objective. When no prior knowledge is available, we can choose a maximally noninformative or reference prior.*

2.2.3 Minimax Estimators

In minimax estimation, we collapse our risk function by looking at the worse-case risk. Given $X \sim \mathbb{P}_\theta$, where $\theta \in \Omega$, and a loss function $L(\theta, d)$, we want to find an estimator δ that minimizes the maximum risk:

$$\sup_{\theta \in \Omega} R(\theta, \delta).$$

Any such δ is called a **minimax** estimator.

Definition 2.25. *We say that a prior Λ is a* ***least favorable prior*** *if $r_\Lambda \geq r_{\Lambda'}$ for any other prior distribution Λ'.*

$$r_\Lambda = \inf_\delta r(\Lambda, \delta) = \inf_\delta \int_{\theta \in \Omega} R(\theta, \delta)\, d\Lambda(\theta).$$

Theorem 2.26 (Relationship with Bayes Estimators)**.** *Suppose δ_Λ is Bayes for Λ with*

$$r_{\Lambda\in\Omega} = \sup_\theta R(\theta, \delta_\Lambda)$$

That is, the Bayes risk of δ_Λ is the maximum risk of δ_Λ. Then,

1. *δ_Λ is minimax*
2. *Λ is a least favorable prior*
3. *If δ_Λ is the unique Bayes estimator for Λ (a.s. for all P_θ), then it is the unique minimax estimator.*

Corollary 2.27. *If a Bayes estimator δ_Λ has constant risk (that is, $R(\theta, \delta_\Lambda) = R(\theta', \delta_\Lambda)$ for all θ and θ'), then δ_Λ is minimax. Note that this is a sufficient but not necessary condition.*

Example 2.28. *Suppose $X \sim Binom(n, \theta)$ for some $\theta \in (0, 1)$ and that we use the squared error loss function. Is the sample proportion $\frac{X}{n}$ minimax? The risk of this estimator is*

$$R\left(\theta, \frac{X}{n}\right) = \frac{\theta(1-\theta)}{n}.$$

The graph of $R\left(\theta, \frac{X}{n}\right)$ versus θ looks like the following:

The risk has a unique maximum at $\theta = \frac{1}{2}$, so the worst-case risk is

$$\sup_{\theta\in\Omega} R\left(\theta, \frac{X}{n}\right) = R\left(\frac{1}{2}, \frac{X}{n}\right) = \frac{1}{4n}.$$

We can use the Corollary 2.27 to find a minimax estimator and then compare the risk of the minimax estimator with that of $\frac{X}{n}$. To find a minimax estimator, we will search for a prior such that the Bayes estimator has constant risk.

Recall the following useful fact. Under the prior distribution $Beta(a, b)$, the Bayes estimator under the squared error loss is

$$\delta_{a,b}(X) = \frac{X + a}{n + a + b}.$$

For any a and b,

$$\begin{aligned}
R(\theta, \delta_{a,b}) &= \mathbb{E}_\theta\left[\left(\frac{X+a}{n+a+b} - \theta\right)^2\right] \\
&= \frac{1}{(n+a+b)^2}\mathbb{E}_\theta\left[(X + a - (n+a+b)\theta)^2\right] \\
&= \frac{1}{(n+a+b)^2}\mathbb{E}_\theta\left[(X - n\theta - a(\theta-1) - \theta b)^2\right] \\
&= \frac{1}{(n+a+b)^2}\left(n\theta(1-\theta) + (a(\theta-1) + \theta b)^2\right).
\end{aligned}$$

This is a quadratic function of θ. To eliminate the θ dependence in $R(\theta, \delta_{a,b})$, we need the coefficients of the linear and quadratic terms to equal zero. The coefficient of θ^2 is

$$-n + (a+b)^2,$$

so we need $a + b = \sqrt{n}$ (since $a, b > 0$). The coefficient of θ is

$$n - 2a(a+b) = n - 2a\sqrt{n},$$

so we need $a = b = \frac{\sqrt{n}}{2}$. *With these choices of* a *and* b, *the risk of* $R(\theta, \delta_{a,b})$ *is constant, which implies that* $Beta\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$ *is a least favorable prior with constant risk. Then our Bayes estimator*

$$\delta_{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}(X) = \frac{X + \frac{\sqrt{n}}{2}}{n + \sqrt{n}},$$

is minimax with constant risk of

$$\frac{1}{4(\sqrt{n}+1)^2}.$$

Since the worst-case risk of $\frac{X}{n}$ *is* $\frac{1}{4n} > \frac{1}{4(\sqrt{n}+1)^2}$, *we can conclude that* $\frac{X}{n}$ *is not minimax.*

2.2.4 James Stein Estimator

Example 2.29 (James Stein Estimator)**.** *Let* $X_1, X_2, \ldots, X_p$ *be independent with* $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$ *for* $1 \leq i \leq p$. *For the sake of simplicity, say* $\sigma^2 = 1$. *Now our goal is to estimate* $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ *under the loss function:*

$$L(\theta, d) = \sum_{i=1}^{p} (d_i - \theta_i)^2$$

A natural estimator for θ *is* $X = (X_1, X_2, \ldots, X_p)$. *It can be shown that* X *is the UMRUE, the maximum likelihood estimator, a generalized Bayes estimator, and a minimax estimator for* θ. *So, it would be natural to think that* X *is admissible. However, counterintuitively, it turns out that this is not the case when* $p \geq 3$.

When $p \geq 3$, X *is dominated by the* ***James-Stein estimator*** *(and that too, strictly dominated):*

$$\delta(X) = (\delta_1(X), \delta_2(X), \ldots, \delta_p(X)) \text{ where}$$

$$\delta_i(X) = \left(1 - \frac{p-2}{\|X\|_2^2}\right) X_i.$$

It turns out that the James-Stein estimator is not itself admissible because it is dominated by the positive part James-Stein estimator

$$\delta_i(X) = \max\left(1 - \frac{p-2}{\|X\|_2^2}, 0\right) X_i.$$

Remark 2.30. *Intuitively, the problem with the estimate* X *is that* $\|X\|_2^2$ *is typically much larger than* $\|\theta\|_2^2$*:*

$$\mathbb{E}[\|X\|_2^2] = E\left[\sum_{i=1}^{p} X_j^2\right] = p + \sum_{i=1}^{p} \theta_i^2 = p + \|\theta\|_2^2$$

where p *is actually* $\sigma^2 p = p$ *in this case. So, we may view the J-S estimator as a method for correcting the bias in the size of* X. *It achieves this by shrinking each coordinate of* X *toward 0.*

1. *Suppose* $\theta_i \overset{iid}{\sim} \mathcal{N}(0, A)$ *then the Bayes estimator for* θ_i *is*

$$\delta_{A,i}(X) = \frac{X_i}{1 + \frac{1}{A}} = \left(1 - \frac{1}{A+1}\right) X_i$$

2. *In this step we must choose* A. *Marginalizing over* θ, *we see that* X *has the distribution,*

$$X_i \overset{iid}{\sim} \mathcal{N}(0, A+1)$$

We will use X and the knowledge of this marginal distribution to find an estimate of $\frac{1}{A+1}$. One could, in principle, use any estimate of A, and it is common to use a maximum likelihood estimate, but here we will used an unbiased estimate.

It can then be shown that

$$\mathbb{E}\left[\frac{1}{\|X\|_2^2}\right] = \frac{1}{(p-2)(A+1)}$$

($\frac{1}{A+1}\|X\|_2^2$ follows a χ_n^2 distribution). So

$$1 - \frac{p-2}{\|X\|_2^2}$$

must be UMVU for $1 - \frac{1}{A+1}$.

If we plug this estimator into our Bayes estimator we obtain the J-S estimator:

$$\delta(X_i) = \left(1 - \frac{p-2}{\|X\|_2^2}\right) X_i.$$

2.2.5 How to Derive an Estimator

Example 2.31 (Moment Estimation)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables from P_θ, $\theta \in \Theta \subset \mathbb{R}^k$, and assume $\mathbb{E}|X_1|^k < \infty$.*

- *Let $\mu_j = \mathbb{E}X_1^j$ be the jth moment of P and then*

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$$

becomes the jth sample moment, which is an unbiased estimator of μ_j, $j = 1, \ldots, k$.

- *Typically,*

$$\mu_j = h_j(\theta), \quad j = 1, \ldots, k \tag{4}$$

for some functions h_j on $\mathbb{R}^k$. By substituting μ_j's on the left-hand side of (4) by the sample moments $\hat{\mu}_j$, we obtain a moment estimator $\hat{\theta}$, i.e., $\hat{\theta}$ satisfies

$$\hat{\mu}_j = h_j(\hat{\theta}), \quad j = 1, \ldots, k,$$

which is a sample analogue of (4). This method of deriving estimators is called the method of moments.

- *Let $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_k)$ and $h = (h_1, \ldots, h_k)$. Then $\hat{\mu} = h(\hat{\theta})$. If the inverse function h^{-1} exists, then the unique moment estimator of θ is $\hat{\theta} = h^{-1}(\hat{\mu})$.*

- *When h^{-1} does not exist (i.e., h is not one-to-one), any solution of $\hat{\mu} = h(\hat{\theta})$ is a moment estimator of θ. If possible, we always choose a solution $\hat{\theta}$ in the parameter space Θ. In some cases, however, a moment estimator may not exist.*

Example 2.32 (Maximum Likelihood Estimation)**.** *Let $X \in \mathscr{X}$ be a sample with a PDF $f_{\boldsymbol{\theta}}$ w.r.t. a σ-finite measure, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$.*

- *For each $x \in \mathscr{X}$, $f_{\boldsymbol{\theta}}(x)$ considered as a function of $\boldsymbol{\theta}$ is called the likelihood function and denoted by $\ell(\theta)$.*

- *Let $\bar{\Theta}$ be the closure of Θ. A $\hat{\theta} \in \bar{\Theta}$ satisfying $\ell(\hat{\theta}) = \max_{\theta \in \bar{\Theta}} \ell(\theta)$ is called a maximum likelihood estimate (MLE) of θ. If $\hat{\theta}$ is a measurable function of X, then $\hat{\theta}$ is called a maximum likelihood estimator (MLE) of θ.*

2.3 Large Sample Theory

2.3.1 Delta Method

Theorem 2.33 (Delta Method)**.** *Let $r_n \to \infty$ and $\phi : \mathbb{R}^d \to \mathbb{R}^k$ be differentiable at θ and assume that $r_n(T_n - \theta) \xrightarrow{d} T$ for some random vector T. Then*

1. *$r_n(\phi(T_n) - \phi(\theta))$ converges in distribution to $\phi'(\theta)T$*
2. *$r_n(\phi(T_n) - \phi(\theta)) - r_n\phi'(\theta)(T_n - \theta)$ converges in probability to 0*

Here $\phi'(\theta) \in \mathbb{R}^{k\times d}$ is the Jacobian Matrix $[\phi'(\theta)]_{ij} = \frac{\partial \phi_i(\theta)}{\partial \theta_j}$.

Proof. By the definition of the derivative, we have that

$$\phi(t) = \phi(\theta) + \phi'(\theta)(t - \theta) + o(\|t - \theta\|),$$

i.e.

$$\phi(t) = \phi(\theta) + \phi'(\theta)(t - \theta) + R(\|t - \theta\|) \tag{5}$$

where $\lim_{h\to 0} \frac{R(h)}{h} = 0$. Since $r_n(T_n - \theta)$ converges in distribution, we know that $r_n(T_n - \theta) = O_p(1)$, which implies that $r_n\|T_n - \theta\| = O_p(1)$. We also have that $\|T_n - \theta\| = o_p(1)$, which implies $R(\|T_n - \theta\|) = o_p(\|T_n - \theta\|)$. Thus

$$r_n R(\|T_n - \theta\|) = r_n o_p(\|T_n - \theta\|) = o_p(r_n\|T_n - \theta\|) = o_p(O_p(1)) = o_p(1).$$

Using this along with (5), we have the second part of the theorem. Noting that $r_n\phi'(\theta)(T_n - \theta) \xrightarrow{d} \phi'(\theta)T$, and applying Slutsky's theorem, we get the first part as well. □

Theorem 2.34 (Second Order Delta Method)**.** *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable at θ, and $r_n(T_n - \theta) \xrightarrow{d} T$. Then if $\nabla\phi(\theta) = 0$, we have*

$$r_n^2(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \frac{1}{2}T^\top \nabla^2\phi(\theta)T.$$

Proof. By definition,

$$\phi(t) = \phi(\theta) + \nabla\phi(\theta)^\top(t - \theta) + \frac{1}{2}(t - \theta)^\top\nabla^2\phi(\theta)(t - \theta) + R(\|t - \theta\|),$$

where $R(h) = o(\|h\|^2)$. Since $\nabla\phi(\theta) = 0$, we actually have

$$\phi(t) = \phi(\theta) + \frac{1}{2}(t - \theta)^\top\nabla^2\phi(\theta)(t - \theta) + R(\|t - \theta\|). \tag{6}$$

Note $r_n^2 R(\|T_n - \theta\|) = r_n^2 o_p(\|T_n - \theta\|^2) = o_p(\|r_n(T_n - \theta)\|^2)$. Since $r_n(T_n - \theta)$ converges in distribution, so does $\|r_n(T_n - \theta)\|^2$, and so $\|r_n(T_n - \theta)\|^2 = O_p(1)$. Thus

$$r_n^2 R(\|T_n - \theta\|) = o_p(O_p(1)) = o_p(1). \tag{7}$$

Now by the continuous mapping theorem, we have that

$$\frac{1}{2}(r_n(T_n - \theta))^\top\nabla^2\phi(\theta)(r_n(T_n - \theta)) \xrightarrow{d} \frac{1}{2}T^\top\nabla^2\phi(\theta)T. \tag{8}$$

So combining (6), (7), (8) and using Slutsky's lemma, we get the desired convergence in distribution. □

Example 2.35. *Suppose $\theta \in (0,1)$, $X_i \sim$ Bernoulli(θ). To estimate θ, we may use the sample mean $\hat{\theta}_n = n^{-1}\sum_{i=1}^n X_i$. Clearly, $\mathbb{E}\hat{\theta}_n = \theta$, $Var(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n}$, $\left[\sqrt{n}(\hat{\theta}_n - \theta)\right]^2 \xrightarrow{d} \theta(1-\theta)\cdot\chi^2_{(1)}$. Instead of using mean squared error to measure the performance of $\hat{\theta}_n$, let us use the Kullback-Leibler (KL) divergence (or the log loss). This is*

$$D_{KL}(P \,\|\, Q) = \int dP \log\left(\frac{dP}{dQ}\right).$$

Let $P_t =$ Bernoulli(t), $t \in [0,1]$. So

$$D_{KL}(P_t \,\|\, P_\theta) = t\log\frac{t}{\theta} + (1-t)\log\frac{1-t}{1-\theta}.$$

Let $\phi(t) = D_{KL}(P_t \,\|\, P_\theta)$. Then

$$\phi'(t) = \log\frac{t}{1-t} - \log\frac{\theta}{1-\theta}.$$

Note $\phi'(\theta) = 0$. So we need the second derivative:

$$\phi''(t) = \frac{1}{t} + \frac{1}{1-t} = \frac{1}{t(1-t)},$$

and so $\phi''(\theta) = \frac{1}{\theta(1-\theta)}$. So by the second order Delta Method,

$$nD_{KL}(P_{\hat{\theta}_n} \,\|\, P_\theta) \xrightarrow{d} \frac{1}{2}\chi^2_{(1)}.$$

2.3.2 Fisher Information

Definition 2.36 (Operator norm)**.** $\|A\|_{\text{op}} := \sup_{\|u\|_2 \le 1} \|Au\|_2$.
Note: $A \in \mathbb{R}^{k\times d}$, $u \in \mathbb{R}^d$ and $\|Ax\|_2 \le \|A\|_{\text{op}}\|x\|_2$.

Before we do anything, we have to make several assumptions.

1. We have a "nice, smooth" model, i.e. the Hessian is Lipschitz-continuous. To be rigorous, the following must hold:
$$\left\|\nabla^2\ell_{\theta_1}(x) - \nabla^2\ell_{\theta_2}(x)\right\|_{\text{op}} \le M(x)\,\|\theta_1 - \theta_2\|_2 \qquad \mathbb{E}_\theta[M^2(x)] < \infty$$
2. The MLE, $\hat{\theta}_n \in \arg\max_{\theta\in\Theta} P_n\ell_\theta(x)$, is consistent, i.e. $\hat{\theta}_n \xrightarrow{p} \theta_0$ under P_{θ_0}.
3. Θ is a convex set.

Theorem 2.37 (Fisher Information)**.** *Let $x_i \overset{\text{iid}}{\sim} P_{\theta_0}$, $\hat{\theta}_n$ be the MLE (i.e. $\nabla P_n\ell_{\hat{\theta}_n} = 0$) and assume the conditions stated above. Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}\big(0,\, (P_{\theta_0}\nabla^2\ell_{\theta_0})^{-1}P_{\theta_0}\nabla\ell_{\theta_0}\nabla\ell_{\theta_0}^\top\,(P_{\theta_0}\nabla^2\ell_{\theta_0})^{-1}\big).$$

Remark 2.38. *Let us rewrite the asymptotic variance. Given that $\nabla^2\ell_\theta = \nabla\left(\frac{\nabla p_\theta}{p_\theta}\right) = \frac{\nabla^2 p_\theta}{p_\theta} - \frac{\nabla p_\theta \nabla p_\theta^\top}{p_\theta^2}$:*

$$\mathbb{E}_\theta\left[\frac{\nabla^2 p_\theta}{p_\theta}\right] = \int \frac{\nabla^2 p_\theta}{p_\theta}\,p_\theta\,d\mu = \int \nabla^2 p_\theta\,d\mu = \nabla^2\int p_\theta\,d\mu = 0.$$

As a result:

$$\mathbb{E}_\theta[\nabla^2 \ell_\theta] = -\mathbb{E}_\theta\left[\left(\frac{\nabla p_\theta}{p_\theta}\right)\left(\frac{\nabla p_\theta}{p_\theta}\right)^T\right] = -\operatorname*{Cov}_\theta(\nabla \ell_\theta(x)).$$

We define the ***Fisher Information*** *as* $I_\theta := \mathbb{E}_\theta[\nabla \ell_\theta(x) \nabla \ell_\theta(x)^\top] = \operatorname{Cov}_\theta \nabla \ell_\theta$*, where the final equality holds because* $\mathbb{E}_\theta[\nabla \ell_\theta(x)] = 0$ *(*θ *maximizes* $\mathbb{E}_\theta[\ell_\theta(x)]$*). To show this, assume that we can swap* $\nabla, \mathbb{E}$*. Then,* $\nabla \ell_\theta(x) = \nabla \log p_\theta(x) = \dfrac{\nabla p_\theta(x)}{p_\theta(x)}$*. Using that result, we see that:*

$$\mathbb{E}_\theta[\nabla \ell_\theta] = \mathbb{E}\left[\frac{\nabla p_\theta}{p_\theta}\right] = \int \frac{\nabla p_\theta}{p_\theta}\, p_\theta \, d\mu = \int \nabla p_\theta \, d\mu = \nabla \int p_\theta \, d\mu = \nabla(1) = 0.$$

We now have a more compact representation of the asymptotic distribution described in the Theorem above.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0,\, I_{\theta_0}^{-1} I_{\theta_0} I_{\theta_0}^{-1}) = \mathsf{N}(0,\, I_{\theta_0}^{-1}).$$

Consider $I_\theta = -\nabla^2 \mathbb{E}[\ell_\theta(x)]$*. If the magnitude of the second derivative is "large," that implies that the log-likelihood is steep around the global maximum (making it "easy" to find). Alternatively, if the magnitude of* $-\nabla^2 \mathbb{E}[\ell_\theta(x)]$ *is "small," we do not have sufficient curvature to find the optimal* θ*.*

Theorem 2.39 (Cramer-Rao)**.** *Let* $g(\theta) = \mathbb{E}_\theta[\delta] \in \mathbb{R}$ *and* $I_\theta = \mathbb{E}_\theta[\nabla \ell_\theta (\nabla \ell_\theta)^\top] \succ 0$*, then*

$$\operatorname{Var}_\theta(\delta) \geq \nabla g(\theta)^\top I_\theta^{-1} \nabla g(\theta).$$

Proof. Set $\Psi(x) = \nabla_\theta \ell_\theta(x)$, we have that $\mathbb{E}_\theta[\Psi] = 0$, and that

$$\begin{aligned}
\mathbb{E}[(\delta - g(\theta))\Psi] &= \mathbb{E}[\delta \Psi] \\
&= \mathbb{E}[\delta \nabla \ell_\theta] \\
&= \mathbb{E}\left[\delta \frac{\nabla p_\theta}{p_\theta}\right] \\
&= \int \delta \nabla p_\theta \, d\mu(x).
\end{aligned}$$

Under good regularity conditions, we have that

$$\mathbb{E}[(\delta - g(\theta))\Psi] = \nabla \int \delta(x) p_\theta(x) \, d\mu(x) = \nabla g(\theta).$$

We take

$$\gamma = \nabla g(\theta), \quad C = I_\theta$$

to get the desired result. □

Corollary 2.40 (Cramer-Rao)**.** *If* $\hat{\theta} : \mathscr{X} \to \Theta$ *is unbiased, then*

$$\mathbb{E}\left[\|\hat{\theta} - \theta\|_2^2\right] \geq \operatorname{tr}(I_\theta^{-1})$$

and

$$\mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^\top] \succeq I_\theta^{-1}.$$

Proof. Take

$$\begin{aligned} g(\theta) &= v^\top \theta \\ \delta &= v^\top \hat{\theta}(X). \end{aligned}$$

Applying the Cramer-Rao theorem,

$$\mathbb{E}\Big[(v^\top(\hat{\theta}-\theta))^2\Big] \geq v^\top I_\theta^{-1} v$$

and

$$\mathbb{E}\Big[(v^\top(\hat{\theta}-\theta))^2\Big] = \mathbb{E}\Big[\mathrm{tr}\Big((\hat{\theta}-\theta)(\hat{\theta}-\theta)^\top v v^\top\Big)\Big] = v^\top \mathrm{Cov}(\hat{\theta})\, v.$$

□

Remark 2.41. *Proof does not give much intuition. The real theorem is Le Cam and Hajek's local asymptotic minimax theorem (discuss later).*

3 Hypothesis Testing

In this section, we study another central topic in statistics: Hypothesis Testing. The section is organized as follows:

1. Fundamentals
2. Multiple Testing and Error Rate Control
3. Causal Inference
4. Conformal Prediction
5. Watermark Detection

This section mainly follows STAT300A (Stanford University), STAT300C (Stanford University), STATS361 (Stanford University) and Mathematics Statistics (Peking University, taught by Fang Yao).

3.1 Fundamentals

Hypothesis testing is just a particular type of decision problem. As usual, we assume that the data is sampled according to $X \sim \mathbb{P}_\theta$ and that $\mathbb{P}_\theta$ belongs to the model $\mathscr{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$. In addition to the standard setup, we divide the models in $\mathscr{P}$ into two disjoint subclasses known as "hypotheses":

$$\begin{aligned} H_0 &: \theta \in \Omega_0 \subset \Omega \qquad \text{(null hypothesis)} \\ H_1 &: \theta \in \Omega_1 = \Omega \setminus \Omega_0 \quad \text{(alternative hypothesis)} \end{aligned}$$

Our goal is to infer which hypothesis is correct. This can be cast as classification, so our decision space is

$$\mathscr{D} = \{\text{accept } H_0,\ \text{reject } H_0\}.$$

	$\theta \in \Omega_0$	$\theta \in \Omega_1$
Reject H_0	1 (Type I Error)	0 (Good)
Accept H_0	0 (Good)	1 (Type II Error)

Table 1: Canonical loss function $L(\theta, d)$.

We have two types of error which induce a loss. A Type I error or false positive occurs when we reject H_0 when it is in fact true. Similarly a Type II error or false negative occurs when we accept H_0 when it is false. Define a test function $\phi(X)$ as

$$\phi(X) = \mathbb{P}\big(\delta_\phi(X, U) = \text{Reject } H_0 \,\big|\, X\big).$$

where U is as usual a uniform random variable independent of X.

Definition 3.1. *The **power function** of a test ϕ is $\beta(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\text{Reject } H_0)$.*

Remark 3.2. *If $\theta_0 \in \Omega_0$, then $\beta(\theta_0) = R(\theta_0, \delta_\phi) =$ Type I Error rate. For $\theta_1 \in \Omega_1$, then $\beta(\theta_1) = 1 - R(\theta_1, \delta_\phi) = 1 -$ Type II Error rate.*

Now we talk about Neyman-Pearson paradigm, which simply bounds it and focuses on minimizing the Type II error. Especially, we require a level α test

$$\sup_{\theta_0 \in \Omega_0} \mathbb{E}_{\theta_0} \phi(X) = \sup_{\theta_0 \in \Omega_0} \beta(\theta_0) \le \alpha.$$

that maximizes the power $\beta(\theta_1) = \mathbb{E}_{\theta_1}[\phi(X)]$ for each $\theta_1 \in \Omega_1$. Such a test is called **uniformly most powerful (UMP)**.

Consider the "simple" test

$$H_0 : X \sim p_0$$
$$H_1 : X \sim p_1$$

where p_0, p_1 denote the densities of $\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}$ with respect to some common measure μ, and we call $\mathbb{E}_{p_1}[\phi(X)]$ the **power of the test** ϕ. Our goal in the simple case can be compactly described as:

$$\begin{aligned} \max_{\phi} \quad & \mathbb{E}_{p_1}[\phi(X)] \\ \text{s.t.} \quad & \mathbb{E}_{p_0}[\phi(X)] \leq \alpha. \end{aligned}$$

3.1.1 Neyman-Pearson Lemma

Lemma 3.3 (Neyman-Pearson)**.**

(i) ***Existence.*** *For testing $H_0 : p_0$ vs. $H_1 : p_1$, there is a test $\phi(X)$ and a constant k such that:*

(a) $\mathbb{E}_{p_0}\phi(X) = \alpha$ (size = level).

(b) $\phi(x) = \begin{cases} 1 & \text{if } \dfrac{p_1(x)}{p_0(x)} > k \text{ (always reject if likelihood ratio is } > k\text{).} \\ 0 & \text{if } \dfrac{p_1(x)}{p_0(x)} < k \text{ (always accept if likelihood ratio is } < k\text{).} \end{cases}$

Such a test is called a likelihood ratio test (LRT)*.*

(ii) ***Sufficient.*** *If a test satisfies (a),(b) for some constant k, it is* most powerful *for testing $H_0 : p_0$ vs. $H_1 : p_1$ at level α. (Hence, the LRT from part (i) is most powerful.)*

(iii) ***Necessary.*** *If a test ϕ is MP at level α then it satisfies (b) for some k, and it also satisfies (a) unless there exists a test of size $< \alpha$ with power 1. (In the latter case, we did not need to expend all of budgeted Type I error.)*

Remark 3.4. *The proof is intuitive, and the construction is*

$$\phi(x) = \begin{cases} 1 & \text{if } r(x) > c_0, \\ \gamma & \text{if } r(x) = c_0, \\ 0 & \text{if } r(x) < c_0. \end{cases}$$

Example 3.5 (One parameter exponential family)**.** *Consider the case where $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x) \propto h(x)\exp(\theta T(x))$, and we are interested in testing*

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1.$$

We want an MP test at level α. The likelihood ratio is

$$\frac{\prod_i p_{\theta_1}(x_i)}{\prod_i p_{\theta_0}(x_i)} \propto \exp\left((\theta_1 - \theta_0)\sum_i T(x_i)\right).$$

Since the exponential is just a monotone transformation, an MP test will reject for large $(\theta_1 - \theta_0)\sum_i T(x_i)$. Assuming $\theta_1 > \theta_0$, we will reject for large $\sum_i T(x_i)$. That is, an MP test has the form

$$\phi(x) = \begin{cases} 1 & \text{if } \sum_i T(x_i) > k \\ \gamma & \text{if } \sum_i T(x_i) = k \\ 0 & \text{if } \sum_i T(x_i) < k, \end{cases}$$

where k, γ are chosen to satisfy the size constraint

$$\alpha = \mathbb{E}_{\theta_0}\phi(X) = \mathbb{P}_{\theta_0}\left[\sum_i T(X_i) > k\right] + \gamma \mathbb{P}_{\theta_0}\left[\sum_i T(X_i) = k\right].$$

Note that $\sum_i T(x_i)$ has no explicit θ dependence and that k, γ do not depend on θ_1 (assuming $\theta_1 > \theta_0$). This means ϕ is in fact UMP for testing

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta > \theta_0.$$

Here, H_1 is an example of a one-sided alternative*, which arises when the parameter values of interest lie on only one side of the real-valued parameter θ_0.*

3.1.2 Monotone Likelihood Ratio

In the above example, we were able to extend our MP test for a simple hypothesis to a UMP test for a one-sided hypothesis. This phenomenon is not unique to exponential families. We can get the same behavior whenever the models have a so-called monotone likelihood ratio.

Definition 3.6 (Families with monotone likelihood ratio (MLR))**.** *We say that the family of densities $\{p_\theta : \theta \in \mathbb{R}\}$ has **monotone likelihood ratio** in $T(x)$ if*

(i) $\theta \neq \theta'$ implies $p_\theta \neq p_{\theta'}$ (identifiability),

(ii) $\theta < \theta'$ implies $p_{\theta'}(x)/p_\theta(x)$ is a nondecreasing function of $T(x)$ (monotonicity).

Theorem 3.7 (Monotone Likelihood Ratio)**.** *Suppose $X \sim p_\theta(x)$ has MLR in $T(x)$ and we test $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$. Then*

(i) There exists a UMP test at level α of the form

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > k \\ \gamma & \text{if } T(x) = k \\ 0 & \text{if } T(x) < k, \end{cases}$$

where k, γ are determined by the condition $\mathbb{E}_{\theta_0}\phi(X) = \alpha$.

(ii) The power function $\beta(\theta) = \mathbb{E}_\theta\phi(X)$ is strictly increasing when $0 < \beta(\theta) < 1$.

3.1.3 Composite Null

Now we introduce a new strategy to deal with cases with a composite null. Consider the case with a simple alternative:

$$H_0 : X \sim f_\theta, \quad \theta \in \Omega_0$$
$$H_1 : X \sim g,$$

where g is known. We now impose a prior distribution Λ on Ω_0. So we consider the new hypothesis

$$H_\Lambda : X \sim h_\Lambda(x) = \int_{\Omega_0} f_\theta(x)\, d\Lambda(\theta),$$

where $h_\Lambda(x)$ is the marginal distribution of X induced by Λ. In order to reduce the problem to a simple versus simple case, let us test H_Λ against H_1. Let β_Λ be the power of the MP level-α test ϕ_Λ for testing H_Λ vs. g.

Definition 3.8 (Least favorable Distribution)**.** *Λ is a least favorable distribution if $\beta_\Lambda \leq \beta_{\Lambda'}$ for any prior Λ'.*

Hence, Λ will be the least favorable distribution if the MP test under Λ has smaller power than the MP test under any other prior distribution. The following theorem can help us to deal with the case of composite null by using the notion of least favorable distribution, which tells that if we choose Λ in the right way, we can get the MP.

Theorem 3.9 (MP and Least Favorable Distribution)**.** *Suppose ϕ_Λ is a MP level-α test for testing H_Λ against g. If ϕ_Λ is level-α for the original hypothesis H_0 (i.e., $\mathbb{E}_{\theta_0}\phi_\Lambda(x) \leq \alpha,\ \forall\, \theta_0 \in \Omega_0$), then*

1. *The test ϕ_Λ is MP for original $H_0 : \theta \in \Omega_0$ vs. g.*
2. *The distribution Λ is least favorable.*

3.1.4 Method of Undetermined Multipliers

Definition 3.10 (Unbiasedness)**.** *Let $\alpha \in [0, 1]$. A test ϕ is **unbiased** level-α if*

$$\forall\, \theta_1 \in \Omega_1 \quad \mathbb{E}_{\theta_1}\phi(X) \geq \alpha \qquad \text{and} \qquad \forall\, \theta_0 \in \Omega_0 \quad \mathbb{E}_{\theta_0}\phi(X) \leq \alpha.$$

Unbiasedness enforces the appealing property that the probability of rejection is greater under any alternative distribution than it is under any null distribution. A uniformly most powerful test is always unbiased if it exists.

Definition 3.11 (α-similarity)**.** *A test ϕ satisfying $\mathbb{E}_\theta\phi(X) = \alpha$ for all $\theta \in \omega$ is called α-**similar** on ω.*

The following lemma tells us we can find a UMPU test by looking only at α-similar tests.

Lemma 3.12. *If $\theta \mapsto \beta_\phi(\theta)$ is continuous (in θ) on Ω for all ϕ, and ϕ_0 is a UMP test amongst α-similar level-α tests, then ϕ_0 is UMPU at level α.*

Proof. Firstly, because ϕ_0 is UMP α-similar tests, it is at least as powerful as $\phi_\alpha(X) \equiv \alpha$, and the power of ϕ_0 on Ω_1 is therefore $\geq \alpha$. Hence, ϕ_0 is unbiased. □

Let us test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ when X is distributed according some member of the one-dimensional exponential family

$$p_\theta(x) = h(x) \exp\left(\theta T(x) - A(\theta)\right)$$

We have the $\alpha-$level condition

$$\beta_\phi(\theta_0) = \alpha,$$

and the derivative condition

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\beta_\phi(\theta_0) = 0.$$

As now we have multiple constraints, we introduce the method of undetermined multipliers.

Lemma 3.13 (MoUM)**.** *Suppose $F_1, \ldots, F_{m+1}$ are real-valued functions defined on a common domain U. We will maximize $F_{m+1}(u)$ subject to constraints of the form*

$$F_i(u) = c_i \quad \text{for } i = 1, \cdots, m$$

where $c_1, \ldots, c_m$ are known constants. To do this, it suffices to find u_0 that satisfies the constraints and maximizes

$$F_{m+1}(u) - \sum_{i=1}^{m} k_i F_i(u)$$

*for any choice of the **undetermined multipliers** $k_1, \ldots, k_m$.*

In this setting, $H_0 : \theta = \theta_0$. We will fix a simple alternative $\theta = \theta' \neq \theta_0$ and hope that our best test has no θ' dependence. We would like to maximize power $\int \phi(x) p_{\theta'}(x)\, d\mu(x)$ subject to

$$\int \phi(x) p_{\theta_0}(x)\, d\mu(x) = \alpha$$
$$\int \phi(x) \frac{d}{d\theta} p_{\theta_0}(x)\, d\mu(x) = 0.$$

For a 1-parameter exponential family, we have

$$p_\theta(x) = h(x) e^{\theta T(x) - A(\theta)} \quad \text{and} \tag{9}$$
$$\frac{d}{d\theta} p_\theta(x) = h(x) e^{\theta T(x) - A(\theta)} \left(T(x) - A'(\theta)\right) = p_\theta(x) \left(T(x) - \mathbb{E}_\theta[T(X)]\right). \tag{10}$$

Applying the reasoning from the previous section, we find that a most powerful test has rejection region defined by

$$p_{\theta'}(x) > k_1 p_{\theta_0}(x) + k_2 \frac{d}{d\theta} p_{\theta_0}(x)$$

for some values of k_1 and k_2, which is equivalent to

$$\frac{e^{(\theta' - \theta_0) T(x)}}{k_1' + k_2' T(x)} > \text{const}$$

with some rearranging.

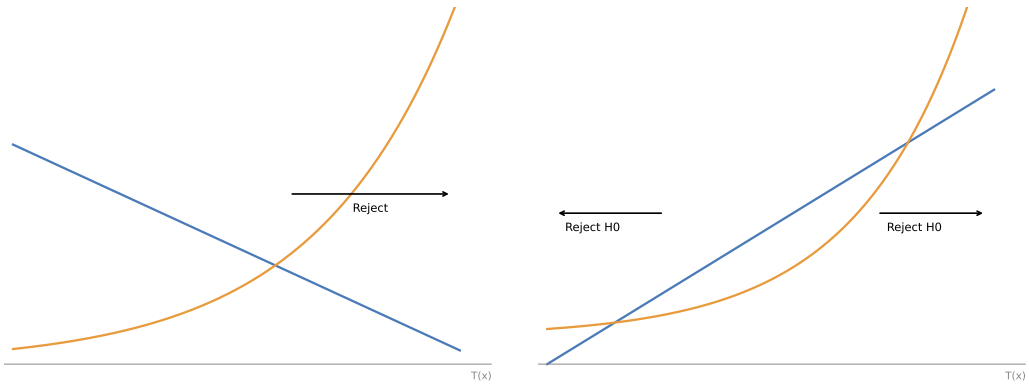

Figure 1: Rejection Regions

The first possibility will not give rise to an unbiased test, because the result would be a one-sided test with monotone power functions. Therefore any optimal ϕ is of the form

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > C_1 \text{ or } T(x) < C_2 \\ \gamma_i & \text{if } T(x) = C_i \\ 0 & \text{otherwise} \end{cases}.$$

A simplification is possible if $T(x)$ is symmetrically distributed under θ_0. Then the optimal test rejects whenever $|T(x)| > \text{const}$. Such tests are called **equitailed tests**.

3.1.5 UMP Invariant Tests

Example 3.14. *Suppose that we observe $X = (X_1, \ldots, X_d)$, with independent coordinates $X_i \sim \mathcal{N}(\theta_i, 1)$ and that we are interested in the hypothesis testing problem*

$$H_0 : \theta_1 = \ldots = \theta_d = 0 \quad vs. \quad H_1 : \exists i \in \{1, \ldots, d\} : \theta_i \neq 0.$$

If we transform our data so that $X' = OX$ for $O \in \mathbb{O}(d)$, the set of $d \times d$ orthogonal matrices, i.e., the set of square matrices such that $O^\top O = OO^\top = I$, then we have $X_i' \overset{ind}{\sim} \mathcal{N}(\theta_i', 1)$ for $\theta' = O\theta$, and our hypothesis testing problem can be seen to be equivalent to testing

$$H_0 : \theta_1' = \ldots = \theta_d' = 0 \quad vs. \quad H_1 : \exists i \in \{1, \ldots, d\} : \theta_i' \neq 0.$$

*Hence, when searching for a test, the **principle of invariance** would suggest constraining $\phi(X) = \phi(OX)\ \forall O \in \mathbb{O}(d)$. In this case, it can be checked that ϕ is invariant in this way iff it is a function of the magnitude of the vector of samples, i.e. of $T = \sum_i X_i^2$. This tells us that if we only care about invariant tests, our optimality goal is to search for UMP tests among functions of T.*

In this example, T is non-central chi-squared distributed with d degrees of freedom, i.e., $T \sim \chi_d^2(\psi^2)$ with a non-centrality parameter $\psi^2 = \sum_{i=1}^d \theta_i^2$. Thus, we can simplify the null and alternative hypotheses of our testing problem to

$$H_0 : \psi = 0 \quad vs. \quad H_1 : \psi > 0. \quad \textit{(one-sided test)}$$

Note that we have reduced both the relevant data and the relevant parameter space for our testing problem; these are common advantages of imposing invariance constraints. To derive our UMP invariant test, we will check that the $\chi_d^2(\psi^2)$ has monotone likelihood ratios in T. Note that the density of non-central chi-squared distribution has the form:

$$f(t; \psi) = e^{-\frac{\psi^2}{2}} \sum_{k=0}^{\infty} \frac{\left(\frac{\psi^2}{2}\right)^k}{k!} \cdot \frac{t^{\frac{d}{2}-1+k} e^{-\frac{t}{2}}}{2^{k+\frac{d}{2}} \Gamma\left(k + \frac{d}{2}\right)}.$$

The likelihood ratio can therefore be computed as

$$\frac{p_{\psi^2}(t)}{p_{\psi^2=0}(t)} = \frac{f(t; \psi)}{\frac{t^{\frac{d}{2}-1} e^{-t/2}}{2^{\frac{d}{2}} \Gamma\left(\frac{d}{2}\right)}} = e^{-\frac{\psi^2}{2}} \sum_k c_k \left(\frac{\psi^2}{2}\right)^k t^k.$$

where c_k are non-negative constants. We can see that each term in the sum above increases in t, and thus the ratio as a whole is increasing in t. (We only need to compare each parameter value with 0, because our null hypothesis is simple.) Thus this family has MLR, and the UMPI test rejects when T is large.

3.1.6 Confidence Regions

Definition 3.15 (Confidence region)**.** *A set $S(X)$ satisfying $\mathbb{P}_\theta[\theta \in S(X)] \geq 1 - \alpha$, for all $\theta \in \Omega$ is known as a $1 - \alpha$ **confidence region**, with a $1 - \alpha$ **confidence level**.*

Given a model $\mathscr{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, we begin by defining an appropriate collection of tests. For each $\theta_0 \in \Omega$, let $\Omega_0(\theta_0)$ be a set containing θ_0, where $\Omega_0(\theta_0) \subset \Omega$. Next, define ϕ_{θ_0} to be a level α test for $H_0 : \theta \in \Omega_0(\theta_0)$ vs. $H_1 : \theta \notin \Omega_0(\theta_0)$, and define $A(\theta_0)$ as the acceptance region of ϕ_{θ_0}. Because ϕ_{θ_0} is a level α test, and $\theta_0 \in \Omega_0(\theta_0)$, we have $P_\theta(X \in A(\theta)) \geq 1 - \alpha$ for all $\theta \in \Omega$.

Now consider the region $S(X) = \{\theta \in \Omega : X \in A(\theta)\}$. Since $P_\theta(\theta \in S(X)) = P_\theta(X \in A(\theta)) \geq 1 - \alpha$, $S(X)$ is a $1 - \alpha$ confidence region! Different choices of null sets $\Omega_0(\theta_0)$ will lead to different forms of confidence regions. For example,

1. **One-sided tests** $\Omega_0(\theta_0) = \{\theta : \theta \le \theta_0\}$ often yield **confidence bounds**

$$S(X) = \{\theta : u(X) \le \theta\}$$

where $u(X)$ denotes a data-dependent lower bound.

2. **Two-sided tests** $\Omega_0(\theta_0) = \{\theta_0\}$ often yield **confidence intervals**

$$S(X) = \{\theta : u(X) \le \theta \le v(X)\},$$

where $v(X)$ denotes a data-dependent upper bound.

Intuitively, we desire a $1-\alpha$ confidence region that is as narrow as possible, which we can achieve by minimizing the number of extraneous it contains. We make this notion precise in the following optimality property for confidence regions.

Definition 3.16 (UMA)**.** *A $1-\alpha$ confidence region $S(X)$ is **uniformly most accurate** if, among all $1-\alpha$ regions, $\mathbb{P}_\theta(\theta' \in S(X)) = \mathbb{P}_\theta(X \in A(\theta'))$ is minimized for all (θ, θ') satisfying $\theta \notin \Omega_0(\theta')$.*

3.2 Multiple Testing and Error Rate Control

3.2.1 Bonferroni's Test and Fisher's Test

We focus on global testing first with a number of null hypotheses $H_{0,i}$. We may care about the global null hypothesis

$$H_0 = \bigcap_{i=1}^{n} H_{0,i}.$$

We'll assume that our p-values have the super-uniform property that (here p_i is a random variable coming out of the test)

$$\mathbb{P}(p_i \le t) = t$$

under the null hypothesis.

Definition 3.17 (Bonferroni's Global Test)**.** *Let α be some desired significance level (for example 0.05), and suppose we have n different hypotheses. **Bonferroni's global test** rejects the global null if*

$$\min_i p_i \le \frac{\alpha}{n}.$$

Proposition 3.18 (Size of Bonferroni's Global Test)**.** *If the p-values are super-uniform then Bonferroni's procedure has **size** (that is, chance of Type I error) at most α.*

Proof. The probability of rejecting the null under the null hypothesis is, by a union bound and the super-uniform property

$$\begin{aligned}
\mathbb{P}(\text{reject}) &= \mathbb{P}\left(\bigcup_{i=1}^{n} \left\{p_i \le \frac{\alpha}{n}\right\}\right) \\
&\le \sum_{i=1}^{n} \mathbb{P}\left(p_i \le \frac{\alpha}{n}\right) \\
&\le n \cdot \frac{\alpha}{n} \\
&= \alpha.
\end{aligned}$$

□

Remark 3.19. *Notice that this does not require any independence assumptions, and in fact if we assume p-values are uniform and independent then*

$$\begin{aligned}\mathbb{P}(\textit{reject}) &= 1 - \mathbb{P}\left(\bigcap_{i=1}^{n}\left\{p_i \geq \frac{\alpha}{n}\right\}\right)\\ &= 1 - \left(1 - \frac{\alpha}{n}\right)^n\\ &\approx 1 - e^{-\alpha}\\ &\approx \alpha - \frac{\alpha^2}{2} + \cdots.\end{aligned}$$

Remark 3.20. *But if we're not under the global null, we will not "hug the uniform line $y = x$" anymore – sometimes (1) these sorted p-values will have a few that are extremely small and then the rest generally following the line (meaning maybe five hypotheses have very strong signals), and sometimes (2) all of the p-values will be overall deflated (too small). Bonferroni's method is really most useful in case (1), because it will successfully reject due to the very small p-values; it will not reject the null in case (2) and thus is not a very powerful test.*

Definition 3.21 (Fisher's Combination Test)**.** ***Fisher's combination test*** *rejects the global null using a more combined measure. Specifically, we consider the test statistic*

$$T = -\sum_{i=1}^{n} 2\log p_i,$$

and we reject the null if T is large.

Proposition 3.22. *Assume that $p_1, \cdots, p_n$ are **independent** and uniform (for example in meta-analysis, suppose we do not have overlapping patients among the studies). Then under the null hypothesis, we have $T \sim \chi^2_{2n}$ (that is, the chi-square distribution with $2n$ degrees of freedom).*

Now we need to discuss which test is better. Consider a particular "needle in a haystack" problem. Let $Y_i \sim N(\mu_i, 1)$ independently for $i = 1, \ldots, n$, and consider the global null

$$H_0 : \mu_1 = \cdots = \mu_n = 0 \quad \text{vs.} \quad H_1 : \exists\, i,\ \mu_i \neq 0.$$

Under H_0, Bonferroni's statistic $\max_i Y_i$ concentrates around $\sqrt{2\log n}$. So we have power only when some $\mu_i > \sqrt{2\log n}$ — the "needle in a haystack" threshold. Plotting power against $\mu_i/\sqrt{2\log n}$, there is a sharp phase transition at 1: power is $\approx \alpha$ below and ≈ 1 above.

Assume the size of "needle" $h = \sqrt{2r\log n}$. We will prove no α-level test with some nontrivial power for $r = 1 - \epsilon$. The point is to avoid having a composite alternative: instead, consider the Bayesian decision problem with null hypothesis

$$H_0 : \mu_i = 0 \text{ for all } i$$

and a "simple" alternative hypothesis

$$H_1 : \{\mu_i\} \sim \pi,$$

where π selects a coordinate uniformly at random and sets its mean to $\mu^{(n)}$, keeping all other means to zero.

The most powerful test in such a setting where H_0, H_1 are both simple hypotheses is the likelihood ratio test, and if we can show it has no power then everything else will have no power. Indeed, we have likelihoods

$$f_0(y) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_j^2\right), \quad f_1(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \mu)^2\right) \prod_{j:j\neq i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_j^2\right).$$

and we want to reject when $\frac{f_1}{f_0}$ is large. What's nice is that all the terms cancel out except the shifted mean:

$$L = \frac{1}{n}\sum_{i=1}^{n} \exp\left(Y_i\mu - \frac{1}{2}\mu^2\right).$$

Notice that this is different from Bonferroni — it's a softmax instead of looking at the maximum, since if $\mu = \infty$ this would be completely dominated by the maximum Y_i. This is nice because it's just a sum of iid terms of mean 1, and in fact if μ is small enough this likelihood concentrates:

Proposition 3.23. *Under the null hypothesis, if $\mu^{(n)} = (1-\epsilon)\sqrt{2\log n}$, then $L \to 1$ in probability as $n \to \infty$.*

Under such proposition, if we set threshold $T_n(\alpha)$ so that $\mathbb{P}_{\mathbb{H}_0}(L \geq T_n(\alpha)) = \alpha$, then

$$\begin{aligned}
\mathbb{P}(\text{type II error}) = \mathbb{P}_{H_1}(L \leq T_n(\alpha)) &= \int \mathbf{1}\{L \leq T_n(\alpha)\}\, dP_{H_1} \\
&= \int L \cdot \mathbf{1}\{L \leq T_n(\alpha)\}\, dP_{H_0} \\
&= \int \mathbf{1}\{L \leq T_n(\alpha)\}\, dP_{H_0} + \int (L-1)\mathbf{1}\{L \leq T_n(\alpha)\}\, dP_{H_0} \\
&\to (1-\alpha) + 0, \\
\implies \mathbb{P}(\text{type I error}) + \mathbb{P}(\text{type II error}) &\to 1.
\end{aligned}$$

Remark 3.24. *Consider Fisher's combination test with*

$$T = \sum_{i=1}^{n} Y_i^2 = \|\mathbf{Y}\|_2^2.$$

By CLT, we have

$$\frac{T - (n + \|\mu\|_2^2)}{\sqrt{2n + 4\|\mu\|_2^2}} \sim \mathcal{N}(0,1).$$

And the test is powerful when $\|\mu\| > \sqrt{2n}$ and us quite large compared to something like Bonferroni. In slightly different terminology, the main parameter that mattered here was proportional to the signal-to-noise ratio

$$SNR = \frac{\textit{signal power}}{\textit{expected noise power}} = \frac{\|\mu\|^2}{\sigma^2 n}$$

Our original model was that we had independent statistics X_i which are $N(0,1)$ under the null hypothesis and $N(\mu_i, 1)$ under the alternative, but now we want to extend to a setting where we have a small fraction of non-null hypotheses. Thus we will now use a simple model where

$$H_0 : X_i \overset{\text{iid}}{\sim} N(0,1), \quad H_1 : X_i \overset{\text{iid}}{\sim} (1-\varepsilon)N(0,1) + \varepsilon N(\mu, 1).$$

In the literature we historically parameterize

$$\varepsilon_n = n^{-\beta}, \quad \frac{1}{2} < \beta < 1,$$

$$\mu_n = \sqrt{2r\log n}, \quad 0 < r < 1.$$

It turns out that we have a threshold curve

$$\rho^*(\beta) = \begin{cases} \beta - \frac{1}{2} & \frac{1}{2} < \beta \leq \frac{3}{4}, \\ (1 - \sqrt{1-\beta})^2 & \frac{3}{4} \leq \beta \leq 1, \end{cases}$$

such that Neyman-Pearson has full power for $r > \rho^*(\beta)$ (that is, we can adjust the test so that the sum of type I and type II error probabilities approaches 0) and no power for $r < \rho^*(\beta)$ (that is, for any test, the limiting sum of type I and type II error is at least 1). A crude calculation, we get power if

$$\max_{\text{non-null}} X_i \approx \sqrt{2r \log n} + \sqrt{2 \log n^{1-\beta}} > \sqrt{2 \log n}.$$

3.2.2 Higher Criticism

In general, the whole point is that in global testing we do not know ϵ and μ and thus cannot use the NP test in the first place. We introduce Tukey's higher criticism.

$$HC_n^* = \max_{0<\alpha\le\alpha_0} \frac{F_n(\alpha) - \alpha}{\sqrt{\alpha(1-\alpha)/n}}.$$

Remark 3.25. *The process $\sqrt{n}(F_n(t) - t)$ will converge in distribution to a Brownian bridge, and the maximum of this rescaled Brownian bridge on $(1/n, \alpha_0)$ converges in probability to $\sqrt{2 \log\log n}$ (this is the law of the iterated logarithm). And a result of Donoho and Jin says that rejecting when $HC_n^* \ge \sqrt{(1+\varepsilon)2 \log\log n}$ gives us $\mathbb{P}_0(\text{type I}) + \mathbb{P}_1(\text{type II}) \to 0$ for any r above the detection threshold.*

Remark 3.26. *The main problem is that near $t = 0$, we are no longer approximately normal — $B(n,p)$ is nicely approximated by a Gaussian if np is not too small, but if np is small we get something that's Poisson, which has much heavier tails. The Burk-Jones statistic is an attempt to resolve this problem, but it is not fully effective. The idea is that for each t, we can test whether $n\hat{F}_n(t) < t$ using a likelihood ratio test*

$$\log LR_n(t) = \begin{cases} nD(\hat{F}_n(t), t) & 0 \le t \le F_n(t), \\ 0 & \text{otherwise}, \end{cases}$$

where $D(p_0, p_1) = p_0 \log \frac{p_0}{p_1} + q_0 \log \frac{q_0}{q_1}$, $q_i = 1 - p_i$ is the Kullback-Leibler divergence. We then define

$$BJ^+ = \max_{1\le i\le n/2} nD\left(p_{(i)}, \frac{i}{n}\right)$$

that is, at what significance level we detect a divergence between what we see and what we expect.

3.2.3 False Discovery Rate

We now turn to identifying which hypotheses are non-null and to controlling the false discovery rate.

	accepted	rejected	total
true	U	V	n_0
false	T	S	$n - n_0$
total	$n - R$	R	n

Definition 3.27 (Familywise Error Rate)**.** *The **familywise error rate (FWER)** is defined to be $\mathbb{P}(V \ge 1)$.*

Theorem 3.28 (Bonferroni's Method)**.** *Bonferroni's method controls FWER at level α; more specifically, if there are $n_0 \le n$ null hypotheses, we get*

$$\text{FWER} \le \mathbb{E}[V] = \frac{n_0}{n}\alpha.$$

Proof. Since V is nonnegative-integer-valued, we have $\mathbb{P}(V \geq 1) \leq \mathbb{E}[V]$. And letting $\mathscr{N}_0$ denote the set of null hypotheses, we have

$$\mathbb{E}[V] = \mathbb{E}\left[\sum_{i \in \mathscr{N}_0} \mathbf{1}\left\{p_i \leq \frac{\alpha}{n}\right\}\right] = \sum_{i \in \mathscr{N}_0} \mathbb{P}\left(p_i \leq \frac{\alpha}{n}\right) = n_0 \cdot \frac{\alpha}{n},$$

which completes the proof. □

Suppose we have n hypotheses $H_{(1)}, \ldots, H_{(n)}$ corresponding to the **ordered** p-values $p_{(1)} \leq \cdots \leq p_{(n)}$ (so we choose $H_{(1)}$ to be the hypotheses with the most surprising p-value, and so on). Now we will compare p-values with an adaptive threshold based on what we've seen so far.

What we do here is called a **Holm's procedure procedure**:

- First, compare with Bonferroni's threshold: if $p_{(1)} \leq \frac{\alpha}{n}$, then reject $H_{(1)}$ and move to the next step. Otherwise, reject nothing (accept all null hypotheses).
- Now in general for step i, if $p_{(i)} \leq \frac{\alpha}{n-i+1}$, then reject $H_{(i)}$ and go to the next step $(i+1)$. Otherwise, accept $\underline{H_{(i)}, H_{(i+1)}, \ldots, H_{(n)}}$ and stop.
- Finally, if $p_{(n)} \leq \alpha$, then we reject $H_{(n)}$; otherwise we accept it.

Notice that

$$\{V \geq 1\} = \{\text{procedure reached } i_0\} \cap \left\{p_{(i_0)} \leq \frac{\alpha}{n - i_0 + 1}\right\},$$

and $\mathbb{P}(V \geq 1) \leq \alpha$.

However, FWER is so stringent that we often return nothing if we require FWER control.

Definition 3.29 (False Discovery Proportion & False Discovery Rate)**.** *The* ***false discovery proportion (FDP)*** *is given by*

$$\text{FDP} = \frac{V}{\max(R, 1)} = \begin{cases} V/R & R \geq 1 \\ 0 & \textit{otherwise} \end{cases}$$

(this last case is just by convention so that we can control this quantity). In general this random variable is unobserved, and the ***false discovery rate (FDR)*** *is the expectation* $\text{FDR} = \mathbb{E}[\text{FDP}]$ *of this quantity.*

We'll now consider a procedure generally more powerful than Holm's procedure – instead of Hochberg's procedure, we consider a step-up procedure with critical values $\alpha_i = \frac{\alpha i}{n}$, which is far less conservative than $\frac{\alpha}{n-i+1}$.

Theorem 3.30 (Controlling FDR)**.** *Suppose our test statistics are independent (so p-values are independent). Then the Benjamini-Hochberg controls the FDR at level* α*. More precisely, we actually have the expression*

$$\text{FDR} = \frac{n_0}{n}\alpha \leq \alpha.$$

Proof. Let $V_i = \{H_i \text{ rejected}\}$ be the indicator function for hypothesis i being rejected. By definition, we have

$$\text{FDP} = \sum_{i \in \mathscr{H}_0} \frac{V_i}{R \vee 1}.$$

Now, it suffices to show that for any null i, we have $\mathbb{E}\left[\frac{V_i}{R\vee 1}\right] = \frac{\alpha}{n}$. (This is somehow "the only answer we can get" because the nulls are uniform and thus the random variables are exchangeable.) To prove this claim, notice that we can do casework over the value of R and write

$$\frac{V_i}{R\vee 1} = \sum_{k=1}^{n} \frac{V_i \mathbf{1}\{R=k\}}{k} = \sum_{k=1}^{n} \frac{\mathbf{1}\left\{p_i \le \frac{\alpha k}{n}\right\}\mathbf{1}\{R=k\}}{k},$$

since assuming $R=k$, we know the threshold for rejection is $\frac{\alpha k}{n}$. Notice that on the event $p_i \le \frac{\alpha k}{n}$, changing p_i to zero doesn't change the threshold, meaning whenever we reject H_i, the number (and identity) of rejections is the same. So we can write the above expression as

$$\sum_{k=1}^{n} \frac{\mathbf{1}\left\{p_i \le \frac{\alpha k}{n}\right\}\mathbf{1}\{R(p_i\to 0)=k\}}{k},$$

where this notation means that we set this null p-value to zero. Now we can take the expectation of this quantity conditioned on all other p-values – the only randomness is in p_i here, so

$$\begin{aligned}\mathbb{E}\left[\frac{V_i}{R\vee 1}\,\middle|\, p_1,\dots,p_{i-1},p_{i+1},\dots,p_n\right] &= \sum_{k=1}^{n} \frac{\frac{\alpha k}{n}\mathbf{1}\{R(p_i\to 0)=k\}}{k} \\ &= \sum_{k=1}^{n} \frac{\alpha}{n}\mathbf{1}\{R(p_i\to 0)=k\} \\ &= \frac{\alpha}{n}.\end{aligned}$$

Then finally taking the expectation over the last p-value yields the result. □

Consider dependency,

Theorem 3.31 (Guo-Rao '08)**.** *Let $S(n) = 1+\frac{1}{2}+\cdots+\frac{1}{n} \approx \log n + 0.577$ be the nth harmonic number. Then there are joint distributions where the FDR of the Benjamini-Hochberg procedure $BH(\alpha)$ is at least $\min(1, \alpha S(n))$.*

Theorem 3.32 (Benjamini-Yekutieli '01)**.** *Under dependence of p-values, the $BH(\alpha)$ procedure does control at level $\alpha S(n)$; in fact,*

$$\text{FDR} \le \alpha S(n)\cdot \frac{n_0}{n}.$$

The following proof is also by Professor Candés and a former student:

Proof. Let $\alpha_i = \frac{i\alpha}{n}$; much like in the proof before, it suffices to show that $\mathbb{E}\left[\frac{V_i}{R\vee 1}\right] = \frac{\alpha}{n}S(n)$. We again have

$$\frac{V_i}{R\vee 1} = \sum_{k=1}^{n} \frac{\mathbf{1}\{p_i \le \alpha_k\}\mathbf{1}\{R=k\}}{k},$$

and we look at where p_i can fall. Summing over the possible ranks it can take on, we have

$$\sum_{k=1}^{n}\sum_{\ell=1}^{k} \frac{\mathbf{1}\{\alpha_{\ell-1}\le p_i\le \alpha_\ell\}\mathbf{1}\{R=k\}}{k} = \sum_{\ell=1}^{k}\sum_{k\ge \ell} \frac{\mathbf{1}\{\alpha_{\ell-1}\le p_i\le \alpha_\ell\}\mathbf{1}\{R=k\}}{k}$$

just by swapping the order of summation. But now if we do the k-sum first, we're just looking at the probability of getting a particularly high number of rejections, so this simplifies to

$$\sum_{\ell=1}^{n} \frac{\mathbf{1}\{R\ge \ell\}}{R}\mathbf{1}\{p_i\in[\alpha_{\ell-1},\alpha_\ell]\}.$$

Everything so far has been an equality, so "nothing interesting" has happened yet. But now we can simplify the first fraction to be bounded by $\frac{1}{\ell}$,

$$\sum_{\ell=1}^{n} \frac{1}{\ell} \mathbf{1}\{p_i \in [\alpha_{\ell-1}, \alpha_\ell]\} = \sum_{\ell=1}^{n} \frac{1}{\ell} \frac{\alpha}{n} = S(n) \frac{\alpha}{n},$$

and then the rest of the proof proceeds as before. What's surprising is that the result of Guo and Rao shows that there are distributions of p-values for which this inequality is indeed tight! □

3.2.4 E-Values

Example 3.33. *Consider the following situation (which is real): research group A tests a medication and gets a promising but not conclusive result (whatever that means – perhaps it's risky to go to trial). Then research group B tests again on new data, but it's still not clear, so research group C tests next (again on new data). We then want to understand how to combine these test results together even when we aren't following a fixed plan.*

Another bad news is in many cases we can not get p-values easily. The e-value is a generic replacement of the p-value which will handle this problem of optional continuation.

Definition 3.34. *Recall that a null hypothesis $\mathscr{H}_0$ is basically a collection of probability measures. An* e-variable E for testing $\mathscr{H}_0$ *is a nonnegative random variable such that*

$$\sup_{P_0 \in \mathscr{H}_0} \mathbb{E}_{P_0}[E] \leq 1.$$

A realization of an e-variable is called an e-value. *Meanwhile, a* p-variable for testing $\mathscr{H}_0$ *is a nonnegative random variable that satisfies*

$$\sup_{P_0 \in \mathscr{H}_0} \mathbb{P}_{P_0}(P \leq \alpha) \leq \alpha$$

for all $\alpha \in (0,1)$, and a realized value of a p-variable is a p-value.

Proposition 3.35. *For any e-value E, the random variable E^{-1} is a conservative p-value (meaning that it is a p-value with wiggle room).*

Proof. Indeed, we have

$$\mathbb{P}\left(\frac{1}{E} \leq \alpha\right) = \mathbb{P}\left(E \geq \frac{1}{\alpha}\right) \leq \frac{\mathbb{E}[E]}{1/\alpha} \leq \alpha.$$

□

Example 3.36. *We'll be looking at* ***safety under optional continuation*** *in this lecture: suppose $(X_1, Z_1), (X_2, Z_2), \ldots$ is our data, where Z_i is some "side information" (for example, whether we have enough money to keep running experiments). Suppose that the data comes in batches of size $n_1, n_2, \ldots$, and $N_t = \sum_{i=1}^{t} n_i$ is the size of the data we have so far.*

We'll establish an e-value E_1 on the first batch. From there, we evaluate an e-value E_2 on the next batch, but only if the outcome is in a certain range (promising but not conclusive) and the external factors take on certain values (things that we cannot plan) – otherwise we stop early. Then depending on the outcomes and external factors up until the second batch, we decide whether or not to compute E_3, and so on. But the point is that after τ total data batches, **the final result we report is the product**

$$V_\tau = \prod_{i=1}^{\tau} E_i.$$

In particular, we're allowed to choose whether to continue on depending on whether each individual E_i is above some threshold of our choice, and the point is that we'll still be able to control the type I error:

Theorem 3.37. *Regardless of the stop-continue rule, as long as τ is a stopping time, V_τ is itself an e-value. More formally, let $\mathscr{F}_t$ be a filtration. Suppose that for all t, the conditional e-variable E_t is a nonnegative random variable which is $\mathscr{F}_t$-measurable, and such that for all $P_0 \in \mathscr{H}_0$ we have $\mathbb{E}_{P_0}[E_t|\mathscr{F}_{t-1}] = 1$. (This is easy to check in practice.) Then $V_t = \prod_{i\le t} E_i$ is a nonnegative supermartingale under the null. Thus by the optional stopping theorem, for any stopping time τ with respect to the filtration, $V_\tau = \prod_{t=1}^{\tau} E_t$ is an e-value. (In particular, V_t is an e-value for each fixed t.)*

Now we focus on multiple testing.

Definition 3.38. *In the **e-BH procedure**, we reject the hypotheses of the largest $\hat{k}$ e-values, where*

$$\hat{k} = \max\left\{i : \frac{ie_{(i)}}{n} \ge \frac{1}{\alpha}\right\}.$$

Remark 3.39. *We should be careful that $\frac{1}{p}$ is not an e-value in general (in fact its expectation need not be finite), even though $\frac{1}{e}$ is always a p-value.*

Theorem 3.40 (Wang–Ramdas '20)**.** *The e-BH procedure has FDR at most $\frac{n_0\alpha}{n}$, and no independence assumption is required.*

3.3 Causal Inference

3.3.1 Randomized Controlled Trials

A traditional view holds that statistics is concerned with inferring correlations or associations among variables. From this perspective, causal inference appears to have no place in statistics. In this section, however, we will discuss several statistical methods for estimating causal effects in both randomized experiments and observational studies.

We define the causal effect of a treatment via potential outcomes. For a binary treatment $w \in \{0, 1\}$, we define potential outcomes $Y_i(1)$ and $Y_i(0)$ corresponding to the outcome the i-th subject would have experienced had they respectively received the treatment or not. The **causal effect of the treatment on the i-th unit** is then

$$\Delta_i = Y_i(1) - Y_i(0).$$

The fundamental problem in causal inference is that only one treatment can be assigned to a given individual, and so only one of $Y_i(0)$ and $Y_i(1)$ can ever be observed. Thus, Δ_i can never be observed. Now, although Δ_i itself is fundamentally unknowable, we can (perhaps remarkably) use **randomized experiments** to learn certain properties of the Δ_i. In particular, large randomized experiments let us recover the average treatment effect (ATE)

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

When an RCT is unavailable, we can use observational data (X_i, Y_i, W_i) to construct an estimator for the ATE. Considering the **propensity score**

$$e(x) = \mathbb{P}\left[W_i = 1 \mid X_i = x\right],$$

and we construct the inverse-propensity weighting estimator

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_iY_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right).$$

Remark 3.41. *IPW relies on the unconfoundedness assumption*

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid e(X_i).$$

Many real-world policies assign treatment by a sharp rule on a running variable (e.g., scholarships above a GPA cutoff). Units just above and below the cutoff are nearly identical in all other respects, so the discontinuity creates a local "as-if randomized" experiment.

3.4 Conformal prediction

3.4.1 Fundamentals

Suppose $X_1, \dots, X_n$ are iid samples from some probability distribution P, and $Y_1, \dots, Y_m$ are iid samples from some other probability distribution Q. Our goal is to test the hypothesis that $P = Q$ even if we don't know either of the distributions. The strategy we learn in early statistics is that we should choose a test statistic T (for example we can let $T = |\overline{X} - \overline{Y}|$) and reject the null based on whether T is unusually large. The question we need to ask is "how large does T need to be," and typically this is where permutation tests are introduced: we compare T to how the statistic would look **if we were to permute the data**.

Formally, we introduce the following (randomized) **permutation distribution**: choose M uniform permutations $\sigma_1, \dots, \sigma_M$ from S_{n+m} (the set of permutations acting on a list of $n+m$ objects), and we will have these permutations act on the vector

$$Z = (X_1, \dots, X_n, Y_1, \dots, Y_m).$$

Specifically, the σ_is will shuffle the entries of Z, and then we evaluate the test statistic on the permuted entry. So we always take the average of the first n entries and the average of the last m entries after permuting, and we find their absolute difference; in other words, we compare

$$T(Z) = \left|\overline{Z_{1:n}} - \overline{Z_{n+1:n+m}}\right|$$

to the corresponding values of $T(Z_\sigma)$, and the p-value will essentially be the relative rank of $T(Z)$ compared to $T(Z_{\sigma_1})$ through $T(Z_{\sigma_M})$:

$$p = \frac{1 + \sum_{i=1}^{M} \mathbf{1}\{T(Z_{\sigma_i}) \geq T(Z)\}}{1 + M}.$$

Under the null, this is either uniformly distributed on $\left\{\frac{1}{M+1}, \frac{2}{M+1}, \dots, \frac{M+1}{M+1}\right\}$ or biased upward due to ties (because under the null we have exchangeability of the vector Z). Thus it is indeed a p-value, and notice that this doesn't rely on us needing to know the distribution of T at all. Given a set of permutations S, consider the quantity

$$p = \frac{1 + \sum_{\sigma \in S} \mathbf{1}\{T(Z_\sigma) \geq T(Z)\}}{1 + |S|}.$$

If the variables are exchangeable, then Z_σ has the same distribution as Z and thus $T(Z_\sigma)$ has the same distribution as $T(Z)$. We are interested in what sets of permutations S yield a valid p-value. Here are the answers:

S	p-value?
all permutations (S_n)	Yes
iid samples from S_n	Yes
an arbitrary fixed subset of S_n	No
iid samples from an arbitrary fixed subset	No
a subgroup of S_n	Yes
iid samples from a subgroup of S_n	Yes

Example 3.42 (Conformal Prediction)**.** *In predictive inference, we take some set of training samples $(X_1, Y_1), \dots, (X_n, Y_n)$ which are iid from some distribution P. We then get a test sample X_{n+1}, and our goal is to construct, from the training samples, a prediction interval for Y_{n+1} with prescribed coverage. That is, we want some C so that*

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha.$$

Note that this is not a confidence interval – this is an observation we have not seen yet, and it's interesting that we can solve this problem at all. In fact, the way we can do so is through permutation tests! We hypothesize a value of Y_{n+1} and test for exchangeability by computing

$$p^y = \frac{1}{(n+1)!} \sum_{\sigma \in S_{n+1}} \mathbf{1}\{T(Z^y_\sigma) \geq T(Z^y)\},$$

where $Z^y = \{(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)\}$. Then setting $y \in C(X_{n+1})$ if and only if $p^y \geq \alpha$, we claim this is a valid prediction interval regardless of our choice of T. Indeed, Y_{n+1} itself will be in the confidence interval if and only if $p^{Y_{n+1}} \geq \alpha$, but we've already shown that $p^{Y_{n+1}} \overset{\text{sto}}{\geq} U$ and thus we are done.

The motivation for conformal prediction is that we want some uncertainty in our prediction and some way of quantifying accuracy.

3.4.2 Approaches

Full conformal prediction is typically done as follows: we fit a model $\hat{\mu}(\cdot) = \mathscr{A}(Z^y)$ to $Z^y = ((X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y))$ which satisfies the symmetry assumption (for example in random forests, it doesn't matter what order we pass in the data, and this is true of most algorithms), and we define $T(Z^y) = |y - \hat{\mu}(X_{n+1})|$ to be the residual. Replacing T with what we have, we thus get a p-value

$$p^y = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{|Y_i - \hat{\mu}(X_i)| \geq |y - \hat{\mu}(X_{n+1})|\} = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{R^y_i \geq R^y_{n+1}\}.$$

and again we include y in the interval if and only if $p^y \geq \alpha$. This is all computationally intensive – every time we fit the model we might need to do a lot of computation, and in fact the prediction interval can be a disjoint set of intervals instead.

We introduce **split conformal prediction** then. This is a special case where we have n data points and we do sample splitting: we learn a model $\hat{\mu}$ with the first split (also called a "fold"), and on the second split we calculate out-of-sample residuals (that is, learn the distribution of the residuals $R_i = |Y_i - \hat{\mu}(X_i)|$). Then the test residual relates to this second split by keeping track of quantiles, and the point is that we separately do training and calibration and form our interval from points that have all not been used for training.

Formally, we compute a score function $S(x, y) = |y - \hat{\mu}(x)|$ by fitting a model on an independent training set. Once we have this, we use a distinct calibration set of size n to find typical size of residuals

$$S_i = S(X_i, Y_i), \quad S^y_{n+1} = S(X_{n+1}, y).$$

The point is that if $y = Y_{n+1}$ these points should all be indistinguishable (they're from the same distribution), and now we include y if

$$p^y = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}\{S_i \geq S^y_{n+1}\} \geq \alpha.$$

And this result now holds conditionally on the training set, since for all purposes $\hat{\mu}$ is fixed.

Remark 3.43. *Other methods to estimate $\hat{\mu}$ include Leave-One-Out and Cross-Validation.*

3.5 Application: Watermark Detection

TODO

4 Classical Statistical Model

Up to this point, we have acquired many statistical tools, and it is now time to step back and look at the whole picture. First, we collect a set of data points (x_i, y_i). We then select an appropriate statistical model for the data, typically a parameterized one, and find the optimal parameters. Finally, we validate the model and use it to make predictions at new data points. In this section, we will discuss several statistical models. The section is organized as follows:

1. Regression
2. Classification
3. Trees and Weak Learners
4. Cross-Validation
5. Time Series

This section mainly follows STAT254 (UC Berkeley), Statistics Learning (PKU, taught by Fang Yao) and STAT153 (UC Berkeley). I would also like to extend my heartfelt thanks to my STAT254 professor, Ryan Giordano, who has offered me countless insights and inspiration in the field of statistics.

4.1 Regression

4.1.1 Fundamentals

In general, we can define the "loss" of guessing $\hat{y}$ when the true value was is y as

$$\mathscr{L}(\hat{y}, y) = \mathscr{L}(f(\mathbf{x}), y)$$

We want to find $f(\cdot)$ so that $\mathscr{L}(f(\mathbf{x}), y)$ is as small as possible. But the loss for a particular f may depend on $\mathbf{x}_{\text{new}}$, since different $\mathbf{x}_{\text{new}}$ are associated with different distributions of y_{new}. It follows that it doesn't make sense to minimize the loss — instead we want to minimize the *risk*,

$$\mathscr{L}(f) := \mathbb{E}\left[\mathscr{L}(f(\mathbf{x}), y)\right] \qquad \text{(Risk)}.$$

So we would like to find

$$f^\star(\cdot) := \underset{f}{\operatorname{argmin}}\, \mathscr{L}(f) = \underset{f}{\operatorname{argmin}}\, \mathbb{E}\left[\mathscr{L}(f(\mathbf{x}), y)\right].$$

We'll consider square loss first:

$$l(\mathbf{x}, y) = \frac{1}{2}(y - f(\mathbf{x}))^2 \quad \hat{\mathscr{L}}(f) = \frac{1}{N}\sum_{i=1}^{N} l(\mathbf{x}_i, y_i).$$

We cannot actually compute f^*, a general question is how to calculate the gap between f^* and $\hat{f}$.

$$0 \le \mathscr{L}(\hat{f}) - \mathscr{L}(f^*) = \underbrace{\mathscr{L}(\hat{f}) - \hat{\mathscr{L}}(\hat{f})}_{\text{Difficult term!}} + \underbrace{\hat{\mathscr{L}}(\hat{f}) - \hat{\mathscr{L}}(f^*)}_{\text{Negative term}} + \underbrace{\hat{\mathscr{L}}(f^*) - \mathscr{L}(f^*)}_{\text{LLN term}}.$$

It is called generalization error and the term $\mathscr{L}(\hat{f}) - \hat{\mathscr{L}}(\hat{f})$ can be bounded $\sup_f[\mathscr{L}(f) - \hat{\mathscr{L}}(f)]$, which will be discussed in the later sections.

We begin with linear regression model:

$$y_n = \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_P x_{nP} + \varepsilon_n = \underbrace{\mathbf{x}_n^\top \boldsymbol{\beta}}_{f(\mathbf{x}_n)} + \epsilon_n, \quad \text{For } n = 1, \ldots, N.$$

where we use N to denote the number of data and P to denote the number of features. To get the optimal estimator $\hat{\boldsymbol{\beta}}$, we let

$$0 = \frac{\partial}{\partial \boldsymbol{\beta}} \mathscr{L}(\boldsymbol{\beta}) \implies \mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0 \implies \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

In the general case, the optimal function is $f^*(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$, and the linear model approximates this conditional expectation as $\mathbb{E}[y \mid \mathbf{x}] = \boldsymbol{\beta}^\top \mathbf{x}$. To expand the function space and improve expressiveness, we sometimes apply a feature mapping to $\mathbf{x}$. A simple example of a feature mapping is the polynomial mapping.

Splines For a partition $\zeta_0, \ldots, \zeta_K$ of the x-axis, define the indicator regressors

$$\boldsymbol{z}_n^0 = \begin{pmatrix} \mathrm{I}(\zeta_0 \le x < \zeta_1) \\ \vdots \\ \mathrm{I}(\zeta_{K-1} \le x < \zeta_K) \end{pmatrix}.$$

Regressing $y_n \sim \boldsymbol{z}_n^0$ produces a *piecewise constant* fit, returning the average of training responses within each interval. Indicator regressors give discontinuous fits with zero derivative. To obtain a piecewise linear fit, augment with

$$\boldsymbol{z}_n^1 = \begin{pmatrix} \mathrm{I}(\zeta_0 \le x < \zeta_1)(x - \zeta_0) \\ \vdots \\ \mathrm{I}(\zeta_{K-1} \le x < \zeta_K)(x - \zeta_{K-1}) \end{pmatrix}.$$

Regressing $y_n \sim \boldsymbol{z}_n^0 + \boldsymbol{z}_n^1$ yields a separate first-order Taylor approximation within each segment. Generalizing to degree p,

$$\boldsymbol{z}_n^p = \begin{pmatrix} \mathrm{I}(\zeta_0 \le x < \zeta_1)(x - \zeta_0)^p \\ \vdots \\ \mathrm{I}(\zeta_{K-1} \le x < \zeta_K)(x - \zeta_{K-1})^p \end{pmatrix},$$

gives a p-th order polynomial fit within each bucket.

Bias and Variance Trade-off Let's decompose this target $\mathbb{E}\left[(\hat{f}(\boldsymbol{x}_{\text{new}}) - y_{\text{new}})^2\right]$.

$$\begin{aligned}
\mathbb{E}\left[(\hat{f}(\boldsymbol{x}_{\text{new}}) - y_{\text{new}})^2\right] &= \mathbb{E}\left[(\hat{f}(\boldsymbol{x}_{\text{new}}) - \bar{f}(\boldsymbol{x}_{\text{new}}) + \bar{f}(\boldsymbol{x}_{\text{new}}) - f^\star(\boldsymbol{x}_{\text{new}}) + f^\star(\boldsymbol{x}_{\text{new}}) - y_{\text{new}})^2\right] \\
&= \underbrace{\mathbb{E}\left[(\hat{f}(\boldsymbol{x}_{\text{new}}) - \bar{f}(\boldsymbol{x}_{\text{new}}))^2\right]}_{\text{“variance”}} + \underbrace{\mathbb{E}\left[(\bar{f}(\boldsymbol{x}_{\text{new}}) - f^\star(\boldsymbol{x}_{\text{new}}))^2\right]}_{\text{“bias”}} + \underbrace{\mathbb{E}\left[(f^\star(\boldsymbol{x}_{\text{new}}) - y_{\text{new}})^2\right]}_{\text{“irreducible error”}}.
\end{aligned}$$

Define $M_{xx} = \mathbb{E}[\mathbf{x}^\top \mathbf{x}]$ and $M_{xy} = \mathbb{E}[\mathbf{x}^\top y]$. Then $\boldsymbol{\beta} = M_{xx}^{-1} M_{xy}$. By CLT,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = \left(\frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top\right)^{-1} \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n (y_n - \boldsymbol{x}_n^\top \boldsymbol{\beta}^*), \quad \sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \to \mathscr{N}(0, M_{xx}^{-1} U M_{xx}^{-1}),$$

where $U := \mathrm{Cov}(\mathbf{x}_n \epsilon_n)$. We have

$$N \cdot \text{Variance} \approx \mathrm{tr}(M_{xx}^{-1/2} U M_{xx}^{-/2}) = \sigma^2 \,\mathrm{tr}(I_P) = P\sigma^2 \implies \text{Variance} \approx \frac{P\sigma^2}{N}.$$

4.1.2 Ridge (L_2) and Lasso (L_1) Regression

When $P \gg N$, $\mathbf{X}^\top\mathbf{X}$ is non-invertible, and the model is prone to overfitting—a common phenomenon in this regime. A simple remedy is to introduce a penalty term. We begin with the L_2 penalty.

$$\mathscr{L}_{ridge}(\boldsymbol{\beta}, \lambda) = \|\mathbf{Y} - \mathbf{X}^\top\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \implies \hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Assume $\mathbf{Y} = \mathbf{X}^\top\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Let the singular value decomposition give $\mathbf{X}^\top\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^\top$ with $\boldsymbol{\Sigma}^2 = \mathrm{diag}(d_1, \ldots, d_P)$, so that $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1} = \mathbf{V}(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^\top$.

A direct computation gives

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{ridge}}] - \boldsymbol{\beta}^* = -\lambda\mathbf{V}(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^\top\boldsymbol{\beta}^*, \qquad \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) = \sigma^2\mathbf{V}\,\mathrm{diag}\left(\frac{d_j}{(d_j+\lambda)^2}\right)\mathbf{V}^\top,$$

which yields

$$\text{Bias: } \|\mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{ridge}}] - \boldsymbol{\beta}^*\|_2^2 = \sum_{j=1}^{P}\frac{\lambda^2}{(d_j+\lambda)^2}(\mathbf{v}_j^\top\boldsymbol{\beta}^*)^2, \qquad \text{Variance: } \mathrm{tr}\,\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) = \sigma^2\sum_{j=1}^{P}\frac{d_j}{(d_j+\lambda)^2}.$$

As λ increases, the variance decreases while the bias increases.

Another choice is L_1 penalty, which we called LASSO,

$$\mathscr{L}_{lasso}(\boldsymbol{\beta}, \lambda) = \|\mathbf{Y} - \mathbf{X}^\top\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Remark 4.1. *LASSO favors **sparse** solutions, whereas Ridge spreads weight more evenly across coefficients. This makes LASSO a natural tool for variable selection: by tuning λ, one can identify the salient predictors as those with nonzero coefficients.*

4.1.3 Kernel Methods

Consider the Ridge Regression, by some Mathematics tricks

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_P)^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{X}(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{Y}.$$

Note that we have replaced an $P \times P$ inverse with an $N \times N$ inverse, and may reduce computation (in some case). $\mathbf{X}\mathbf{X}^\top = (x_i^\top x_j)_{i,j}$, replace $x_i^\top x_j$ into $k(x_i, x_j)$, and we have kernel function.

Definition 4.2 (Reproducing Kernel Hilbert Space)**.** *Let $\mathscr{X}$ be an arbitrary set and $\mathscr{H}$ a Hilbert space of real-valued functions on $\mathscr{X}$. We say $\mathscr{H}$ is a* reproducing kernel Hilbert space *(RKHS) if there is a kernel $k : \mathscr{X} \times \mathscr{X} \mapsto \mathbb{R}$ such that*

- $\forall x \in \mathscr{X},\ k(\cdot, x) \in \mathscr{H}$.
- Reproducing property: $\forall x \in \mathscr{X},\ f \in \mathscr{H},\ \langle f, k(\cdot, x)\rangle_{\mathscr{H}} = f(x)$.

For a kernel $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, we define an integral operator $\mathscr{T}_k : L_\mu^2(\mathscr{X}) \to L_\mu^2(\mathscr{X})$ as follows

$$\mathscr{T}_k f(x) = \mathbb{E}_{x' \sim \mu}[k(x, x')f(x')].$$

Theorem 4.3 (Mercer's theorem)**.** *Let k be a continuous kernel on a **compact set** $\mathscr{X}$. There exist an orthonormal basis $\{e_j\}_{j=1}^\infty$ of $L_\mu^2(\mathscr{X})$ such that $\forall x, x' \in \mathscr{X}$,*

$$k(x, x') = \sum_{j=1}^{\infty}\lambda_j e_j(x)e_j(x').$$

The convergence is uniform on $\mathscr{X} \times \mathscr{X}$ and absolute for each $(x, x') \in \mathscr{X} \times \mathscr{X}$.

Remark 4.4. *From this perspective, an RKHS can be viewed as a subspace of L_2. The faster the eigenvalues λ_j decay, the smaller the RKHS—and correspondingly, the smoother the functions it contains.*

This theorem ensures the existence of an eigendecomposition of a kernel k, i.e., the corresponding integral operator $\mathscr{T}_k$. Note that $(\lambda_j)_{j\geq 1}$ and $(e_j)_{j\geq 1}$ are the eigenvalues and eigenfunctions of the integral operator $\mathscr{T}_k$ in the sense that

- $\mathscr{T}_k e_j = \lambda_j e_j$, i.e., $\mathbb{E}_{x'\sim\mu}[k(x,x')e_j(x')] = \lambda_j e_j(x)$.
- $\langle e_i, e_j\rangle_{L^2_\mu(\mathscr{X})} = \mathbb{E}_{x\sim\mu}[e_j(x)e_i(x)] = \delta_{i,j}$.

Feature map. Mercer's theorem gives a feature map for the kernel k. Let

$$\Phi : \mathscr{X} \to \ell^2, \qquad \Phi(x) = \left(\sqrt{\lambda_1}e_1(x), \sqrt{\lambda_2}e_2(x), \ldots, \sqrt{\lambda_j}e_j(x), \ldots\right)^\top.$$

Then,

$$k(x,x') = \sum_{j=1}^{\infty} \sqrt{\lambda_j}e_j(x)\sqrt{\lambda_j}e_j(x') = \langle \Phi(x), \Phi(x')\rangle_{\ell^2}.$$

Theorem 4.5 (Spectral representation of RKHS)**.** *Let k be a continuous kernel on a compact set $\mathscr{X}$, and $\{e_j\}$ be the orthonormal basis given in Mercer's theorem. Define*

$$\mathscr{H} = \left\{ f = \sum_j a_j e_j : \sum_j \frac{a_j^2}{\lambda_j} < \infty \right\},$$

with the inner product

$$\left\langle \sum_j a_j e_j, \sum_j b_j e_j \right\rangle_{\mathscr{H}} = \sum_j \frac{a_j b_j}{\lambda_j}.$$

Then, $\mathscr{H}$ is the RKHS $\mathscr{H}_k$.

4.2 Classification

4.2.1 Fundamentals

Let us now turn to classification problems. Like regression, we assume we get IID observations $z_n = (\boldsymbol{x}_n, y_n)$, but now with the difference that y takes one of a set of unordered distinct values, which I will call $\mathscr{Y}$. The simplest case is where y takes one of two values, which I will sometimes call "binary classification". As before, we wish to use $\boldsymbol{x}$ to guess what y is. As we will see shortly, there are more choices to what f should even be for classification than there are for regression.

Ultimately, we need a mapping $\boldsymbol{x} \mapsto \mathscr{Y}$, so we might take $\hat{y}(\boldsymbol{x}) \in \mathscr{Y}$. Necessarily, the loss function must take in two values in $\mathscr{Y}$—the guess and the truth—and return a real number. Formally, $\mathscr{L} : \mathscr{Y} \times \mathscr{Y} \mapsto \mathbb{R}$, which can be fully represented as a $|\mathscr{Y}| \times |\mathscr{Y}|$ matrix. The simplest case of this is called "zero–one" loss:

$$\mathscr{L}_{\mathrm{ZO}}(\hat{y}, y) = \mathrm{I}(\hat{y} \neq y) = \sum_{c\in\mathscr{Y}} \mathrm{I}(\hat{y} = c)\,\mathrm{I}(y \neq c).$$

Note that the maximizer of the zero–one loss has a closed form:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}(\boldsymbol{x},y)}\left[\mathscr{L}_{\mathrm{ZO}}(\hat{y},y)\right] &= \sum_{c\in\mathscr{Y}} \mathrm{I}(\hat{y}=c)\,\mathbb{E}_{\mathbb{P}(\boldsymbol{x},y)}\left[\mathrm{I}(y\neq c)\right] \\ &= \sum_{c\in\mathscr{Y}} \mathrm{I}(\hat{y}=c)\big(1-\mathbb{P}(y=c,\boldsymbol{x})\big) \\ &= 1-\sum_{c\in\mathscr{Y}} \mathrm{I}(\hat{y}=c)\big(1-\mathbb{P}(y=c\mid\boldsymbol{x})\big)\mathbb{P}(\boldsymbol{x}).\end{aligned}$$

Since $\mathbb{P}(\boldsymbol{x})$ does not depend on c, the best choice for $\hat{y}$ is

$$\hat{y}^*(\boldsymbol{x}) = \operatorname*{argmax}_{c\in\mathscr{Y}} \mathbb{P}(y=c\mid\boldsymbol{x}).$$

The natural generalization of zero–one loss and functions that return categories is a function that returns a probability. For binary regression with $y\in\{0,1\}$, we can take $\pi(\boldsymbol{x})$ to approximate $\mathbb{P}(y=1\mid\boldsymbol{x})$, and so take $\pi:\boldsymbol{x}\mapsto[0,1]$.

Cross Entropy Loss If $\pi(\boldsymbol{x})\in(0,1)$ then we can in fact use other losses. One common choice is the negative log likelihood:

$$\begin{aligned}\mathscr{L}_{\mathrm{CE}}(\pi,y) &= -\log\mathbb{P}(y\mid\pi) \\ &= -\log\pi^y(1-\pi)^{1-y} \\ &= -y\log\pi-(1-y)\log(1-\pi) \\ &= -y\log\frac{\pi}{1-\pi}-\log(1-\pi).\end{aligned}$$

This is also known as the "cross-entropy loss," since the maximum likelihood estimator is also the minimizer of the Kullback–Liebler divergence.

Logistic Regression Suppose we have $f\in\mathbb{R}$. Then

$$\pi(f)=\frac{\exp(f)}{1+\exp(f)} =: \mathrm{Expit}(f).$$

You can readily see that $\mathrm{Expit}(f)\in(0,1)$. The expit function is invertible with inverse sometimes called the "logit":

$$\mathrm{Logit}(\pi) := \log\frac{\pi}{1-\pi}.$$

Note that if you plug this formula into the cross-entropy loss,

$$\begin{aligned}\mathscr{L}_{\mathrm{CE}}(\mathrm{Expit}(f),y) &= -y\,\mathrm{Logit}(\mathrm{Expit}(f))+\log(1-\mathrm{Expit}(f)) \\ &= -yf+\log\left(1-\frac{\exp(f)}{1+\exp(f)}\right) \\ &= -yf-\log(1+\exp(f)),\end{aligned}$$

And if we further take $f(\boldsymbol{x})=\boldsymbol{\beta}^\top\boldsymbol{x}$, we get

$$\mathscr{L}_{\mathrm{CE}}(\mathrm{Expit}(\boldsymbol{\beta}^\top\boldsymbol{x}),y) = -y\boldsymbol{\beta}^\top\boldsymbol{x}-\log(1+\exp(\boldsymbol{\beta}^\top\boldsymbol{x})),$$

which is **logistic regression**.

Generative modeling We have noted that the optimal classifier takes the form of the maximum of $\mathbb{P}(y = c \mid \boldsymbol{x})\mathbb{P}(\boldsymbol{x})$. In our reasoning above we ignored $\mathbb{P}(\boldsymbol{x})$ since it does not depend on c, and modeled $\mathbb{P}(y = c \mid \boldsymbol{x})$, e.g. using the expit function and linear models. However, we could condition in the other direction to write

$$\hat{y} = \operatorname*{argmax}_c \mathbb{P}(y = c \mid \boldsymbol{x})\mathbb{P}(\boldsymbol{x}) = \operatorname*{argmax}_c \mathbb{P}(\boldsymbol{x} \mid y = c)\mathbb{P}(y = c).$$

When we have many more data than categories, $\mathbb{P}(y = c)$ is easy to estimate:

$$\mathbb{P}(y = c) \approx \frac{1}{N}\sum_{n=1}^{N} \mathrm{I}(y_n = c).$$

We can then attempt to model the conditional density $\mathbb{P}(\boldsymbol{x} \mid y = c)$ by forming density estimates the sets of points $\mathscr{X}_c := \{\boldsymbol{x}_n : y_n = c\}$. Density estimation is difficult in high dimensions—arguably more difficult than estimating good decision boundaries—but generative modeling can be particularly effective in low dimensions, or when something special is known about the structure of $\mathbb{P}(\boldsymbol{x} \mid y = c)$.

Note that the estimation $\mathbb{P}(\boldsymbol{x} \mid y = c)$ does not admit a ready formulation as a direct minimization of classification error. Estimation of densities is typically considered an **"unsupervised learning"** problem for this reason, even though here we would use the unsupervised learning estimate in service of a supervised learning problem.

4.2.2 Support Vector Machines

Suppose we have binary classification data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$, where $y_n \in \{-1, 1\}$, and we want to find a linear decision boundary:

$$\hat{y}_n = \operatorname{sign}(\boldsymbol{x}_n^T \hat{\boldsymbol{\beta}}).$$

Such a vector $\hat{\boldsymbol{\beta}}$ would be normal to a hyperplane that separates our predictions. Support Vector Machines add three layers on top of this:

- **Regularization**: find the separating hyperplane with maximum margin.
- **Soft margin**: allow for margin mistakes (e.g. not linearly separable).
- **Kernel**: expressive (possibly infinite) feature mapping.

For finite data that are linearly separable, there always are infinitely many separating hyperplanes. Which one does perceptron converge to? If it does, con implicit, so let's make it implicit and ask for a particular one. As a form of regularization, we'll ask for the one with maximum margin:

$$\hat{\boldsymbol{\beta}}_{\text{hard}} = \operatorname*{argmax}_{\|\boldsymbol{\beta}\|_2 = 1,\, \gamma} \gamma \quad \text{subject to} \quad y_n \boldsymbol{x}_n^T \boldsymbol{\beta} \geq \gamma \quad \forall n = 1, \ldots, N.$$

Proposition 4.6.

$$\hat{\boldsymbol{\beta}}_{hard} = \operatorname*{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2 \quad \textit{subject to} \quad y_n \boldsymbol{x}_n^T \boldsymbol{\beta} \geq 1 \quad \forall n = 1, \ldots, N.$$

Soft Margin SVM The hard margin SVM is restricted to linearly separable data. We now introduce a soft margin (linear) SVM, which trades separability with $\|\boldsymbol{\beta}\|_2$:

$$\hat{\boldsymbol{\beta}}_{\text{soft}}(\lambda) = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{n=1}^{N} \max(1 - y_n \boldsymbol{x}_n^T \boldsymbol{\beta},\, 0) + \lambda \|\boldsymbol{\beta}\|_2^2.$$

This way, we may allow classification (or margin) mistakes for the purpose of reducing $\|\boldsymbol{\beta}\|_2^2$. This brings us to the familiar form of minimizing loss plus penalty.

Connection to hard margin SVM To see how soft margin SVM relates to hard margin, add slack variable $\xi_n \geq \max(1 - y_n \boldsymbol{x}_n^T \boldsymbol{\beta}, 0)$ and observe that

$$\hat{\boldsymbol{\beta}}_{\text{soft}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_2^2 + \frac{1}{\lambda} \sum_{n=1}^{N} \xi_n \quad \text{subject to} \quad y_n \boldsymbol{x}_n^T \boldsymbol{\beta} \geq 1 - \xi_n,\ \xi_n \geq 0 \quad \forall n = 1, \ldots, N.$$

We also divided by λ. From this point we may intuitively believe that, for linearly separable data, $\hat{\boldsymbol{\beta}}_{\text{hard}}$ is obtained as the limit of $\hat{\boldsymbol{\beta}}_{\text{soft}}(\lambda)$ as $\lambda \to 0^+$, since the slack variables $\xi_1, \ldots, \xi_N$ are pushed to zero.

Interpretation as penalty for margin mistakes The slack variable ξ_n is nonzero if we make a margin mistake on $(\boldsymbol{x}_n, y_n)$. It is greater than one if we make a classification mistake. The sum $\sum_{n=1}^{N} \xi_n$ quantifies the total over all margin mistakes, which we can control by adjusting the penalty λ.

Kernels Using convex optimization theory, one may show that the soft margin SVM solution will satisfy

$$\hat{\boldsymbol{\beta}}_{\text{soft}} = \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n \quad ; \quad \text{where} \quad \alpha_n \neq 0 \text{ if and only if } y_n \boldsymbol{x}_n^T \hat{\boldsymbol{\beta}}_{\text{soft}} \leq 1.$$

If we used an expressive feature transformation $\varphi(\boldsymbol{x}_n) \in \mathbb{R}^Q$, the soft margin SVM $\hat{\boldsymbol{\beta}}_{\text{kernel}} \in \mathbb{R}^Q$ would satisfy

$$\hat{\boldsymbol{\beta}}_{\text{kernel}} = \sum_{n=1}^{N} \alpha_n y_n \varphi(\boldsymbol{x}_n).$$

The prediction on a new observation would be

$$y = \text{sign}\left(\varphi(\boldsymbol{x})^T \hat{\boldsymbol{\beta}}_{\text{kernel}}\right) = \sum_{n=1}^{N} \alpha_n y_n \varphi(\boldsymbol{x})^T \varphi(\boldsymbol{x}_n) =: \sum_{n=1}^{N} \alpha_n y_n k(\boldsymbol{x}, \boldsymbol{x}_n).$$

4.3 Trees and Weak Learners

4.3.1 Classification and Regression Trees

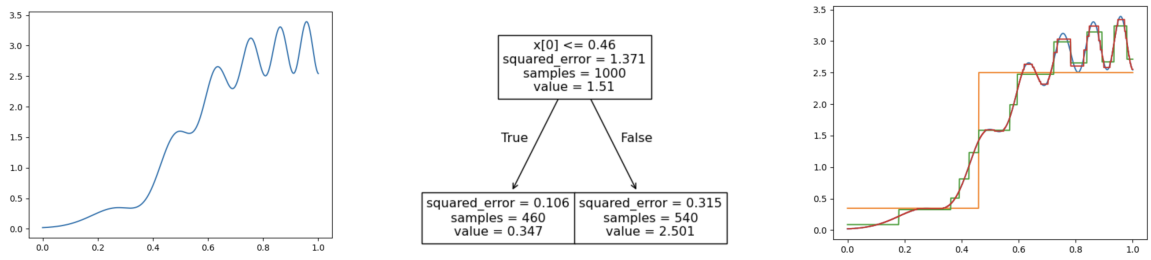

Figure 2: Classification and Regression Trees

Classification and Regression Trees (CART) is a tree-based method that recursively splits the data into two groups based on feature values, producing a model that can predict either a category or a numerical value.

In order to choose a more optimal tree, we "trim" the tree back by regularizing the size of the tree. Given a starting tree T and a cost complexity parameter α, we find the subtree that minimizes

$$\text{Cost complexity}\,(T') = \sum_{t \in T'} \sum_{n \in t} (y_n - \hat{y}_t)^2 + \alpha |T'|, \tag{11}$$

where $t \in T'$ means t is a leaf node of T', $n \in t$ means $\boldsymbol{x}_n$ is assigned to leaf node t, and $|T'|$ counts the number of leaf nodes.

4.3.2 Bootstrapping

Let $Y_1, \ldots, Y_N$ be i.i.d. observations from an unknown distribution F, and let $\hat{\mu} = s(\boldsymbol{Y})$ be an estimator of a parameter μ. The *bootstrap* approximates the sampling distribution of $\hat{\mu}$ by replacing F with the empirical distribution

$$\hat{F}(y) \;=\; \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(Y_i \le y),$$

and resampling from the data itself. For $b = 1, \ldots, R$, draw $\boldsymbol{Y}^{*b} = (Y_1^{*b}, \ldots, Y_N^{*b})$ **with replacement** from $\{Y_1, \ldots, Y_N\}$ and compute the bootstrap replicate $\hat{\mu}^{*b} = s(\boldsymbol{Y}^{*b})$. The empirical distribution of $\{\hat{\mu}^{*1}, \ldots, \hat{\mu}^{*R}\}$ then serves as a proxy for the sampling distribution of $\hat{\mu}$.

Remark 4.7. *Bootstrapping's power is computational rather than informational: it does not see more than the data, but it extracts what closed-form derivations would discard.*

4.3.3 Bagging

Suppose we're using squared error loss, and we have a learner that is approximately **unbiased but high variance**. That is, for a particular fixed x, we have $\hat{f}(x; \mathscr{D})$, where $\hat{f}(\cdot; \mathscr{D})$ depends on the training data $\mathscr{D}$, which satisfies

$$\mathbb{E}_{\mathscr{D}}\left[\hat{f}(x; \mathscr{D})\right] \approx f^*(x) \quad \text{and} \quad \text{Var}_{\mathscr{D}}\left(\hat{f}(x; \mathscr{D})\right) \text{ is large.}$$

Here, $\mathbb{E}\left[y \mid x\right] = f^*(x)$ is the best possible prediction we could make.

One might think of deep regression trees as being a possible example. Intuitively, one might expect such a predictor to be over-fitting the particular dataset at hand.

Recall that we showed earlier that the expected squared error decomposes into

$$\mathbb{E}_{\mathscr{D}}\left[(\hat{f}(x; \mathscr{D}) - f^*(x))^2\right] = \underbrace{\mathbb{E}_{\mathscr{D}}\left[\left(\hat{f}(x; \mathscr{D}) - \mathbb{E}_{\mathscr{D}}\left[\hat{f}(x; \mathscr{D})\right]\right)^2\right]}_{\text{Variance}} + \underbrace{\left(\mathbb{E}_{\mathscr{D}}\left[\hat{f}(x; \mathscr{D})\right] - f^*(x)\right)^2}_{\text{Bias}}.$$

The variance only increases the expected squared error, so if we could get a lower-variance estimator without changing the bias, we'd improve our loss. Ideally, we might imagine getting B different, new datasets $\mathscr{D}_1, \ldots, \mathscr{D}_B$, and then using

$$\hat{f}_{\text{Ideal}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}(x; \mathscr{D}_b) \approx \mathbb{E}_{\mathscr{D}}\left[\hat{f}(x; \mathscr{D})\right].$$

Of course, if we could actually do this, we should just combine the datasets into one giant dataset. But this idea motivates the feasible "bagging" estimator:

$$\hat{f}_{\text{Bagging}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}(x; \mathscr{D}_b^*),$$

where $\mathscr{D}_b^*$ is a *bootstrap* sample from the original dataset. The bootstrap distribution is a random distribution *conditional on the original dataset* $\mathscr{D}$, with probability distribution

$$\mathbb{P}\left((x_n^*, y_n^*) = (x_n, y_n)\right) = \frac{1}{N} \quad \text{for} \quad n = 1, \ldots, N.$$

The problem is, of course, that bootstrap draws are not draws from $\mathscr{D}$. Why does bagging work? In my opinion, the theory is not very clear, but it appears in practice that bagging tends to improve highly variable and expressive estimators, but can make less variable estimators worse.

4.3.4 Boosting

Algorithm 1 Forward Stagewise Additive Modeling

1: Initialize $\hat{f}_0(\cdot) = 0$.
2: **for** $m = 1, \ldots, M$ **do**
3: Compute

$$\hat{\theta}_m, \hat{\phi}_m \;:=\; \operatorname*{argmin}_{\theta,\phi} \frac{1}{N} \sum_{n=1}^{N} \mathscr{L}\Big(\hat{f}_{m-1}(\boldsymbol{x}_n) + \theta\phi(\boldsymbol{x}_n),\, y_n\Big)\,.$$

4: Set $\hat{f}_m(\cdot) = \hat{f}_{m-1}(\cdot) + \hat{\theta}_m\hat{\phi}_m(\cdot)$.
5: **end for**
6: **return** $\hat{f}_M(\cdot) = \sum_{m=1}^{M} \hat{\theta}_m\hat{\phi}_m(\cdot)$.

For boosting, we first find the $\hat{f}_1(\cdot)$ that best predicts y_n, then the one that makes the biggest improvement when added to $\hat{f}_1(\cdot)$, and so on. Note that at each step all we need to be able to do is make a *small improvement*, since over M steps these small improvements accumulate. For example, we might take $\hat{\phi}_m(\cdot)$ to simply be a "stump": a regression tree with a single split. In this sense, boosting can improve on **high-biased estimators** by adding up a large number of them in an **increasingly expressive** way.

4.4 Cross-Validation

We spend some time developing theory to control the generalization error, $\hat{\mathscr{L}}(\hat{f}) - \mathscr{L}(\hat{f})$. In practice, however, this theory is not useful because we cannot actually compute the complexity. More importantly, the worst-case bounds we computed appear in practice to be *too loose* and may suggest bad advice (see, e.g., Belkin et al. (2019)). In practice, we most often *estimate* the risk $\mathscr{L}(\hat{f})$ using held-out data.

Recall that the estimator $\hat{\mathscr{L}}(\hat{f}) = \frac{1}{N}\sum_{n=1}^{N} \mathscr{L}(\hat{f}(\boldsymbol{x}_n), y_n)$ is not considered a reliable estimator of $\mathscr{L}(\hat{f})$ because $\hat{f}$ is not independent of $(\boldsymbol{x}_n, y_n)$. But suppose we have an actual new dataset, $\mathscr{D}_{\text{new}} = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_M, y_M))$. Then we could just estimate:

$$\hat{\mathscr{L}}_{\text{new}}(\hat{f}) \;\approx\; \frac{1}{M}\sum_{m=1}^{M} \mathscr{L}(\hat{f}(\boldsymbol{x}_m), y_m).$$

This *is* a consistent estimate of $\mathscr{L}(\hat{f})$ since we did not use $\mathscr{D}_{\text{new}}$ to train $\hat{f}$, and so $\hat{f}$ is independent of $\boldsymbol{x}_m, y_m$.

One might ask, if you had $\mathscr{D}_{\text{new}}$, why didn't you just use it to train $\hat{f}$? Conversely, you might imagine taking the data you do have, and splitting it (randomly) into a "training set" $\mathscr{D}_{\text{train}}$ and a "test set" $\mathscr{D}_{\text{test}}$, using only $\mathscr{D}_{\text{train}}$ to compute $\hat{f}$, and $\mathscr{D}_{\text{test}}$ to estimate $\hat{\mathscr{L}}_{\text{test}}(\hat{f})$ using the above formula. There are costs and benefits to doing this:

- You have less data to train $\hat{f}$.
- You have a good estimate of $\mathscr{L}(\hat{f})$.

Putting more data in $\mathscr{D}_{\text{train}}$ ameliorates the first cost, at the risk of making your estimate of $\mathscr{L}(\hat{f})$ worse.

A clever way to use nearly all the data to compute $\hat{f}$ and still get a reasonable test set error is to *repeat* the above procedure many times, each time with a different test set.

As an extreme, let $\hat{f}_{-n}$ denote an estimate of $\hat{f}$ with the n-th datapoint removed. Since you have removed only one datapoint, we might hope that $\hat{f}_{-n} \approx \hat{f}$, so we have paid little estimation cost. Of course, the estimate

$$\hat{\mathscr{L}}_{\text{test}}(\hat{f}_{-n}) \;=\; \mathscr{L}(\hat{f}_{-n}(\boldsymbol{x}_n), y_n)$$

uses a single datapoint, and so is very variable, if nearly unbiased. However, we can do this for *each* datapoint and average the results to reduce the variance of the test error estimate. This is called the "leave-one-out CV estimator," or LOO-CV:

$$\hat{\mathscr{L}}_{\text{LOO}} \;:=\; \frac{1}{N}\sum_{n=1}^{N} \mathscr{L}(\hat{f}_{-n}(\boldsymbol{x}_n), y_n).$$

Of course, you can do the same thing leaving two points out, or three, or so on, giving "leave-k-out CV" estimators. In the extreme where you leave a fraction N/K points out, you get "K-fold" CV estimators.

How many should you leave out? There seems to be a striking lack of useful theory. The rules of thumb are:

- By leaving out more datapoints you "bias" your estimate if your estimate of $\hat{f}$ depends strongly on how many datapoints you have.
- By leaving out more datapoints you reduce the "variance" of your estimate of $\mathscr{L}(\hat{f})$ by making the test set larger.

So there appears to be a sort of meta-bias-variance tradeoff in the estimation of $\mathscr{L}(\hat{f})$. Common practice is five- or ten-fold cross-validation.

4.5 Time Series

TODO

Part II

Chapter 2: High-Dimensional Statistics

5 Concentration Inequalities

We now turn to the high-dimensional setting, beginning with concentration inequalities. The section is organized as follows:

1. Basic Concentration Inequalities
2. Random Vectors
3. Random Matrices
4. Concentration of Lipschitz Functions
5. Gaussian Concentration

This section mainly follows STAT210B (UC Berkeley, taught by Song Mei), High-dimensional probability (PKU, taught by Zhihua Zhang) and Introduction to Machine Learning (PKU, taught by Lei Wu), STAT300B (Stanford), CS839 (U Wisconsin–Madison). I also referred to the book High-dimensional probability: An introduction with applications in data science [5] and the book High-Dimensional Statistics: A Non-Asymptotic Viewpoint [6].

5.1 Basic Concentration Inequalities

Suppose we have a random variable $X \sim \mathbb{P}_X$, sampled from the distribution $\mathbb{P}_X$. Let $\mu = \mathbb{E}_{X\sim\mathbb{P}_X}[X]$ be its expectation. In general, $|x-\mu|$ could be very large. However, in many scenarios (especially when X takes a special form), $|x-\mu|$ is very small with high probability.

Lemma 5.1 (Markov's Inequality)**.** *Let X be a nonnegative random variable. Then for all $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Lemma 5.2 (Chernoff's Inequality)**.** *For all $t > 0$, we have*

$$\mathbb{P}(X \geq \mu + t) \leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}} = e^{-h(t)},$$

where

$$h(t) = \sup_{\lambda} \lambda t - \log \mathbb{E}[e^{\lambda(X-\mu)}].$$

Lemma 5.3 (Union Bound)**.** *Suppose we have a collection of events $\{E_s\}_{s\in[d]}$. If $\mathbb{P}(E_s^c) \leq \frac{\delta}{d}$ for all s, then*

$$\mathbb{P}\left(\bigcup_{s\in[d]} E_s\right) \geq 1 - \delta.$$

5.1.1 Sub-Gaussians

Now we discuss an important type of random variables.

Definition 5.4 (Sub-Gaussian)**.** *A random variable with $\mu = \mathbb{E}[X]$ is σ-**sub-Gaussian** if there is a positive number $\sigma > 0$ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2\sigma^2/2} \qquad \forall \lambda \in \mathbb{R}.$$

Proposition 5.5 (Hoeffding's Inequality)**.** *Suppose X_i, $i = 1, \ldots, n$ are independent, where X_i has mean μ_i and is σ_i-sub-Gaussian. Then*

1. $\sum_{i=1}^n X_i$ *has mean* $\sum_{i=1}^n \mu_i$ *and is sub-Gaussian with parameter* $\sqrt{\sum_{i=1}^n \sigma_i^2}$.

2.
$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Example 5.6 (Bounded random variable)**.** *Let $X \in P([a,b])$. We claim that X is $(b-a)$-sub-Gaussian.*

Proof. Instead of a direct calculation, we use a series of tricks.

Trick 1: Let $X' \overset{d}{=} X$ with X, X' independent. Then

$$\mathbb{E}_X[e^{\lambda(X-\mu)}] = \mathbb{E}_X[e^{\lambda X - \mathbb{E}_X[X']}]$$

Trick 2: Use Jensen's inequality to get $e^{-\lambda \mathbb{E}[X']} \leq \mathbb{E}[e^{-\lambda X'}]$. This gives

$$\leq \mathbb{E}_{X,X'}\, \mathbb{E}[e^{\lambda(X-X')}]$$

Trick 3: Introduce $\varepsilon \sim \mathrm{Unif}(\{\pm 1\})$ with ε independent of (X, X'). Then $\varepsilon(X - X') \overset{d}{=} X - X'$.

$$= \mathbb{E}_{\varepsilon,X,X'}\, \mathbb{E}[e^{\lambda\varepsilon(X-X')}]$$

Using the tower property of conditional expectation,

$$= \mathbb{E}_{X,X'}\Big[\mathbb{E}_\varepsilon[e^{\lambda\varepsilon(X-X')} \mid X, X']\Big]$$

By the 1-sub-Gaussianity of ε,

$$\leq \mathbb{E}_{X,X'}[e^{\lambda^2(X-X')^2/2}]$$

Since $(X - X')^2 \leq (b-a)^2$ by the boundedness of X, X',

$$\leq e^{\lambda^2(b-a)^2/2}.$$ □

Remark 5.7. *Get $\frac{b-a}{2}$-sub-Gaussian using the following techniques. Suppose $m \leq X \leq M$. We claim*

$$\psi_X(\lambda) := \log \mathbb{E}e^{\lambda X} \leq \tfrac{1}{8}\lambda^2(M-m)^2.$$

Proof. WLOG assume $\mathbb{E}[X] = 0$. Then $\psi_X(0) = 0$ and $\psi_X'(0) = \mathbb{E}[X] = 0$. The second derivative is

$$\psi_X''(\lambda) = \mathrm{Var}_Q(X) \leq \tfrac{1}{4}(M-m)^2,$$

where Q is the tilted measure $dQ/dP \propto e^{\lambda X}$ (supported on $[m, M]$), and the bound is Popoviciu's inequality. By Taylor expansion at $\lambda = 0$,

$$\psi_X(\lambda) = \tfrac{1}{2}\psi_X''(\xi)\lambda^2 \leq \tfrac{1}{8}\lambda^2(M-m)^2.$$ □

Theorem 5.8 (Equivalent Conditions for Sub-Gaussians)**.** *Let X be a random variable. Then the following are equivalent:*

(i) The tails of X satisfy
$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{\kappa_1^2}\right) \qquad \forall t \geq 0.$$

(ii) The moments of X satisfy
$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \leq \kappa_2\sqrt{p}, \qquad \forall p \geq 1.$$

(iii) The moment generating function of X^2 satisfies

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(\kappa_3^2 \lambda^2) \qquad \forall \lambda \text{ such that } |\lambda| \leq \frac{1}{\kappa_3}.$$

(iv) The moment generating function of X^2 is bounded at some point:

$$\mathbb{E}[\exp(X^2/\kappa_4^2)] \leq 2.$$

Moreover, if $\mathbb{E}[X] = 0$, then properties (i)–(iv) are also equivalent to

(v) The moment generating function of X satisfies

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\kappa_5^2 \lambda^2/2) \qquad \forall \lambda \in \mathbb{R}.$$

Here, $\kappa_1, \ldots, \kappa_5$ are universal constants.

5.1.2 Sub-Exponentials

Let $G \sim \mathcal{N}(0,1)$. Then G^2 is not sub-Gaussian. This is because $\mathbb{E}[G^2] = 1$, and

$$\begin{aligned}\mathbb{E}[e^{\lambda(G^2-1)}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(z^2-1)} e^{-z^2/2}\, dz \\ &= \begin{cases} \dfrac{e^{-\lambda}}{\sqrt{1-2\lambda}} & \lambda < 1/2 \\ \infty & \lambda \geq 1/2. \end{cases}\end{aligned}$$

We can still derive a good but weaker tail bound for this kind of random variable.

Definition 5.9. *A random variable X is (ν, α)-**sub-exponential** if*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2 \nu^2/2} \qquad \forall |\lambda| \leq \frac{1}{\alpha}.$$

We can see from this definition that sub-Gaussian variables are sub-exponential with any $\alpha > 0$.

Example 5.10. *If $G \sim \mathcal{N}(0,1)$, then G^2 is $(2,4)$-sub-exponential.*

Proof. We want to show that

$$\mathbb{E}[e^{\lambda(G^2-1)}] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} \qquad \forall |\lambda| \leq \frac{1}{4}.$$

We can do this by comparing Taylor series. □

Proposition 5.11 (Bernstein Condition)**.** *Suppose X has mean μ and variance σ^2. Suppose that $\mathbb{E}[(X-\mu)^k] \leq \frac{1}{2} k!\, \sigma^2 b^{k-2}$ for all $k \geq 2$. Then X is $(\sqrt{2}\sigma, 2b)$-sub-exponential.*

Proof. We just need to show that the moment generating function is bounded. Do a Taylor expansion:

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}[(X-\mu)^k]}{k!} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2}.\end{aligned}$$

This is a geometric series, so we can simplify it:

$$\leq 1 + \frac{\lambda^2\sigma^2/2}{1 - b|\lambda|}$$
$$\leq e^{(\lambda^2\sigma^2/2)/(1-b|\lambda|)}.$$

When $|\lambda| \leq \frac{1}{2b}$,

$$\leq e^{\lambda^2(\sqrt{2}\sigma)^2/2}. \qquad \square$$

Lemma 5.12 (Bernstein's Inequality)**.** *Let $\{X_i\}_{i\in[n]}$ be independent with $\mathbb{E}[X_i] = \mu_i$ and X_i (ν_i, α_i)-sub-exponential. Then $\sum_{i=1}^n (X_i - \mu_i)$ is sub-exponential with parameters $\nu_* = \sqrt{\sum_{i=1}^n \nu_i^2}$ and $\alpha_* = \max_i \alpha_i$. Moreover,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \begin{cases} e^{-nt^2/(2\nu_*^2)} & t \leq \nu_*^2/\alpha_* \\ e^{-nt/(2\alpha_*)} & t > \nu_*^2/\alpha_*. \end{cases}$$

Theorem 5.13 (Strong Bernstein's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent, mean-zero, with $|X_i| \leq K$ for all i. Set $\sigma^2 = \sum_{i=1}^n \mathbb{E}[X_i^2]$. Then*

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

Remark 5.14. *How do we extract the order of $\frac{1}{n}\sum_{i=1}^n X_i - \mu$? Set $\delta = \exp\left(-n\min\left\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\right\}\right)$, and solve for t to get*

$$t = \max\left\{\nu\sqrt{\frac{2\log(1/\delta)}{n}},\ b\cdot\frac{2\log(1/\delta)}{n}\right\}.$$

This tells us that

$$\frac{1}{n}\sum_{i=1}^n X_i - \mu \leq \max\left\{\nu\sqrt{\frac{2\log(1/\delta)}{n}},\ b\cdot\frac{2\log(1/\delta)}{n}\right\} \qquad \text{with probability at least } 1-\delta.$$

Example 5.15. *Let $Y = \sum_{i=1}^n Z_i^2$ with $Z_i \sim \mathcal{N}(0,1)$. Then $Y \sim \chi^2(n)$. Last time, we showed that Z_i^2 is $\mathrm{sE}(2,4)$, so $Y \sim \mathrm{sE}(2\sqrt{n}, 4)$. By Bernstein's inequality,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i^2 - 1\right| \geq t\right) \leq 2e^{-nt^2/8} \qquad \forall t \leq 1.$$

Here is a problem: Suppose we have $\{u_1, u_2, \ldots, u_N\} \subseteq \mathbb{R}^d$ with a high dimension d. Can we find an $F : \mathbb{R}^d \to \mathbb{R}^m$ with some small m such that the distances are preserved? That is, we want

$$1 - \delta \leq \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + \delta, \qquad \forall i, j \in [N].$$

How small can we make m? The Johnson–Lindenstrauss lemma says that we can achieve this by random projection.

Lemma 5.16 (Johnson–Lindenstrauss)**.** *Let $X \in \mathbb{R}^{m\times d}$ have entries $X_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and let $F : \mathbb{R}^d \to \mathbb{R}^m$ be defined as $F(u) = \frac{1}{\sqrt{m}} X \cdot u$. Then for any fixed $\{u_1, \ldots, u_N\} \subseteq \mathbb{R}^d$, as long as $m \gtrsim \frac{1}{\varepsilon^2}\log\left(\frac{N}{\delta}\right)$, with probability $1-\delta$ we have*

$$1 - \varepsilon \leq \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + \varepsilon, \qquad \forall i, j \in [N].$$

Proof. Denote $Y_{i,j} = \frac{\|F(u_i)-F(u_j)\|_2^2}{\|u_i-u_j\|_2^2}$. We claim that $Y_{i,j} \sim \chi^2(m)/m$. Then Bernstein's inequality will give

$$\mathbb{P}(|Y_{i,j}-1| \geq t) \leq 2e^{-mt^2/8} \qquad \forall t \leq 1.$$

Using a union bound on all $N(N-1) \leq N^2$ pairs $i \neq j$, we get

$$\mathbb{P}(\exists\, i,j \in [N] \text{ s.t. } |Y_{i,j}-1| \geq t) \leq 2N^2 e^{-mt^2/8} \qquad \forall t \leq 1.$$

Setting the right-hand side equal to δ, we can solve for m to get

$$m \geq \frac{8}{t^2}\log\left(\frac{2N^2}{\delta}\right) = \frac{C}{t^2}\log\left(\frac{N}{\delta}\right).$$

Now let's verify the claim that $Y_{i,j} = \frac{\|F(u_i)-F(u_j)\|_2^2}{\|u_i-u_j\|_2^2} \sim \chi^2(m)/m$. Note that

$$\frac{1}{\sqrt{m}}X(u_i-u_j) \sim N\left(0, \frac{\|u_i-u_j\|_2^2}{m} I_m\right),$$

which implies that

$$\frac{\|X(u_i-u_j)\|_2^2}{m} \sim \frac{\|u_i-u_j\|_2^2 \cdot \chi^2(m)}{m}.$$

This proves the claim. □

Theorem 5.17 (Equivalent Conditions for sub-Exponentials)**.** *The following statements are equivalent:*

(a)

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t/\kappa_1), \qquad \forall t \geq 0.$$

(b)

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \leq \kappa_2 p, \qquad \forall p \geq 1.$$

(c)

$$\mathbb{E}[\exp(\lambda|X|)] \leq \exp(\kappa_3\lambda) \qquad \forall \lambda \text{ s.t. } 0 \leq \lambda \leq \frac{1}{\kappa_3}.$$

(d)

$$\mathbb{E}[\exp(|X|/\kappa_4)] \leq 2.$$

Moreover, if $\mathbb{E}[X] = 0$, then (a)–(d) are equivalent to

(e)

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\kappa_5^2/2) \qquad \forall |\lambda| \leq \frac{1}{\kappa_5}.$$

Here, $\kappa_1, \ldots, \kappa_5$ are universal constants.

Example 5.18. *Let $X_1 \sim \mathrm{sG}(\sigma_1)$ and $X_2 \sim \mathrm{sG}(\sigma_2)$ be not necessarily independent with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$. We claim that $X_1X_2 \sim \mathrm{sE}(K\sigma_1\sigma_2, K\sigma_1\sigma_2)$ for some universal K.*

Lemma 5.19 (Bennett's Inequality)**.** *Let $(X_i)_{i\in[n]}$ be independent, where $X_i - \mathbb{E}[X_i] \leq b$ a.s., and $\nu_i^2 := \mathrm{Var}(X_i)$ for all $i \in [n]$. Then*

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{\sum_{i=1}^n \nu_i^2}{b^2}\, h\left(\frac{bt}{\sum_{i=1}^n \nu_i^2}\right)\right),$$

where $h(u) = (1+u)\log(1+u) - u$.

5.1.3 Maximal Inequality

Lemma 5.20. *Let $(X_i)_{i\in[n]}$ be a sequence of random variables. For any convex, strictly increasing $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$, we have*

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \leq \psi^{-1}\left(\sum_{i=1}^{n} \mathbb{E}[\psi(X_i)]\right),$$
$$\mathbb{P}\left(\max_{i\in[n]} X_i \geq t\right) \leq \sum_{i=1}^{n} \frac{\mathbb{E}[\psi(X_i)]}{\psi(t)}.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\max_{i\in[n]} X_i\right] &= \mathbb{E}\left[\psi^{-1}\left(\max_{i\in[n]} \psi(X_i)\right)\right] && (\psi \text{ strictly increasing}) \\
&\leq \psi^{-1}\left(\mathbb{E}\left[\max_{i\in[n]} \psi(X_i)\right]\right) && (\text{Jensen, } \psi^{-1} \text{ concave}) \\
&\leq \psi^{-1}\left(\sum_{i=1}^{n} \mathbb{E}[\psi(X_i)]\right) && (\max \leq \text{sum}).
\end{aligned}$$

□

Example 5.21. *For $X_i \sim \mathrm{sG}(\sigma)$, take $\psi(u) = e^{\lambda u}$. Optimizing over λ, we get*

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \leq \sigma\sqrt{2\log n}.$$

This gives an important intuition: n sub-Gaussian random variables have maximum of order $\sqrt{\log n}$.

Remark 5.22 (Lower Bound on the Gaussian Maximum)**.** *The bound $\mathbb{E}[\max_{i\in[n]} X_i] \leq \sigma\sqrt{2\log n}$ is tight: for $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, one also has $\mathbb{E}[\max_{i\in[n]} X_i] \gtrsim \sqrt{\log n}$. The argument is short. Let $A = \{\exists\, i : X_i \geq c\sqrt{\log n}\}$; then*

$$\mathbb{E}\Big[\max_i X_i\Big] \geq c\sqrt{\log n} \cdot \mathbb{P}(A) - \mathbb{E}[X_1^-] \cdot \mathbb{P}(A^c),$$

where $\mathbb{E}[X_1^-] = 1/\sqrt{2\pi}$ is an absolute constant, so it suffices to show $\mathbb{P}(A) \geq \frac{1}{2}$. By independence and $1 - x \leq e^{-x}$,

$$\mathbb{P}(A) \geq 1 - \exp\Big(-n \cdot \mathbb{P}(X_1 \geq c\sqrt{\log n})\Big),$$

and the Gaussian lower tail bound $\mathbb{P}(X_1 \geq r) \gtrsim \frac{1}{r} e^{-r^2/2}$ gives $\mathbb{P}(X_1 \geq c\sqrt{\log n}) \gtrsim 1/(n^{c^2/2}\sqrt{\log n}) \geq 1/(2n)$ for any $c < \sqrt{2}$. So the $\sqrt{\log n}$ rate is unavoidable, not an artifact of the proof.

5.2 Random Vectors

5.2.1 Random Vectors

Theorem 5.23 (Concentration of the Norm)**.** *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, sub-Gaussian coordinates X_i that satisfy $\mathbb{E}X_i^2 = 1$. Then*

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

where $K = \max_i \|X_i\|_{\psi_2}$ and C is an absolute constant.

Proof. For simplicity, we assume that $K \geq 1$. We shall apply Bernstein's deviation inequality for the normalized sum of independent, mean-zero random variables

$$\frac{1}{n}\|X\|_2^2 - 1 = \frac{1}{n}\sum_{i=1}^n (X_i^2 - 1).$$

Since the random variable X_i is sub-Gaussian, $X_i^2 - 1$ is sub-exponential, and more precisely

$$\begin{aligned} \|X_i^2 - 1\|_{\psi_1} &\leq C\|X_i^2\|_{\psi_1} \\ &= C\|X_i\|_{\psi_2}^2 \\ &\leq CK^2. \end{aligned}$$

Applying Bernstein's inequality, we obtain for any $u \geq 0$ that

$$\mathbb{P}\left\{\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq u\right\} \leq 2\exp\left(-\frac{cn}{K^4}\min(u^2, u)\right). \tag{3.1}$$

(Here we used that $K^4 \geq K^2$ since we assumed that $K \geq 1$.)

This is a good concentration inequality for $\|X\|_2^2$, from which we are going to deduce a concentration inequality for $\|X\|_2$. To make the link, we can use the following elementary observation that is valid for all numbers $z \geq 0$:

$$|z - 1| \geq \delta \implies |z^2 - 1| \geq \max(\delta, \delta^2). \tag{3.2}$$

We obtain for any $\delta \geq 0$ that

$$\begin{aligned} \mathbb{P}\left\{\left|\tfrac{1}{\sqrt{n}}\|X\|_2 - 1\right| \geq \delta\right\} &\leq \mathbb{P}\left\{\left|\tfrac{1}{n}\|X\|_2^2 - 1\right| \geq \max(\delta, \delta^2)\right\} \\ &\leq 2\exp\left(-\frac{cn}{K^4}\cdot\delta^2\right). \end{aligned}$$

Changing variables to $t = \delta\sqrt{n}$, we obtain the desired sub-Gaussian tail

$$\mathbb{P}\left\{\left|\|X\|_2 - \sqrt{n}\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{K^4}\right) \quad \text{for all } t \geq 0. \tag{3.3}$$

This is equivalent to the conclusion of the theorem. □

Remark 5.24 (Deviation)**.** *Theorem 5.23 states that with high probability, X takes values very close to the sphere of radius $\sqrt{n}$. In particular, with high probability (say, 0.99), X even stays within* constant distance *from that sphere.*

Definition 5.25 (Isotropic Random Vectors)**.** *A random vector X in $\mathbb{R}^n$ is called* isotropic *if*

$$\Sigma(X) = \mathbb{E}\, XX^\top = I_n,$$

where I_n denotes the identity matrix in $\mathbb{R}^n$.

Lemma 5.26 (Isotropic Random Vectors)**.** *Let X be an isotropic random vector in $\mathbb{R}^n$. Then*

$$\mathbb{E}\,\|X\|_2^2 = n.$$

Moreover, if X and Y are two independent isotropic random vectors in $\mathbb{R}^n$, then

$$\mathbb{E}\,\langle X, Y\rangle^2 = n.$$

Remark 5.27 (Almost Orthogonality of Independent Vectors)**.** *Let us normalize the random vectors X and Y in Lemma 5.26, setting*

$$\overline{X} := \frac{X}{\|X\|_2} \quad \text{and} \quad \overline{Y} := \frac{Y}{\|Y\|_2}.$$

Lemma 5.26 is basically telling us that $\|X\|_2 \asymp \sqrt{n}$, $\|Y\|_2 \asymp \sqrt{n}$, and $\langle X, Y\rangle \asymp \sqrt{n}$ with high probability, which implies that

$$|\langle \overline{X}, \overline{Y}\rangle| \asymp \frac{1}{\sqrt{n}}.$$

Thus, in high-dimensional spaces independent and isotropic random vectors tend to be almost orthogonal.

Definition 5.28 (Sub-Gaussian Random Vectors)**.** *A random vector X in $\mathbb{R}^n$ is called* sub-Gaussian *if the one-dimensional marginals $\langle X, x\rangle$ are sub-Gaussian random variables for all $x \in \mathbb{R}^n$. The* sub-Gaussian norm *of X is defined as*

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

A good example of a sub-Gaussian random vector is a random vector with independent, sub-Gaussian coordinates:

Lemma 5.29 (Sub-Gaussian Distributions with Independent Coordinates)**.** *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean-zero, sub-Gaussian coordinates X_i. Then X is a sub-Gaussian random vector, and*

$$\|X\|_{\psi_2} \le C \max_{i \le n} \|X_i\|_{\psi_2}.$$

Theorem 5.30 (Uniform Distribution on the Sphere is Sub-Gaussian)**.** *Let X be a random vector uniformly distributed on the Euclidean sphere in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$:*

$$X \sim \mathrm{Unif}\left(\sqrt{n}\, S^{n-1}\right).$$

Then X is sub-Gaussian, and

$$\|X\|_{\psi_2} \le C.$$

5.2.2 Grothendieck's Inequality

Theorem 5.31 (Grothendieck's Inequality)**.** *Consider an $m \times n$ matrix (a_{ij}) of real numbers. Assume that, for any numbers $x_i, y_j \in \{-1, 1\}$, we have*

$$\left| \sum_{i,j} a_{ij} x_i y_j \right| \le 1.$$

Then, for any Hilbert space H and any vectors $u_i, v_j \in H$ satisfying $\|u_i\| = \|v_j\| = 1$, we have

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

where $K \le 1.783$ is an absolute constant.

Proof. Let $K = \frac{\pi}{\log(1+\sqrt{2})}$. We first prove that there exists a Hilbert space H and maps $\Psi, \Phi : S(\mathbb{R}^n) \to S(H)$ such that for all $u, v \in \mathbb{R}^n$:

$$\langle u, v\rangle = \frac{2K}{\pi} \arcsin\langle \Psi(u), \Phi(v)\rangle.$$

Proof. We first describe tensor inner products: for $T = u^{\otimes 3} \in \mathbb{R}^{n\times n\times n}$ defined by $T_{ijl} = u_i u_j u_l$, we have $\langle T, S\rangle = \sum_{ijl} T_{ijl} S_{ijl}$. By Taylor expansion,

$$\begin{aligned}
\arcsin(\langle u, v\rangle) &= \langle u, v\rangle + \frac{1}{3!}\langle u, v\rangle^3 + \frac{3}{5!}\langle u, v\rangle^5 + \cdots \\
&= \langle u, v\rangle + \frac{1}{3!}\langle u^{\otimes 3}, v^{\otimes 3}\rangle + \frac{3}{5!}\langle u^{\otimes 5}, v^{\otimes 5}\rangle + \cdots \\
&= \left\langle \begin{bmatrix} u \\ \sqrt{\frac{1}{3!}}\, u^{\otimes 3} \\ \sqrt{\frac{3}{5!}}\, u^{\otimes 5} \\ \vdots \end{bmatrix}, \begin{bmatrix} v \\ \sqrt{\frac{1}{3!}}\, v^{\otimes 3} \\ \sqrt{\frac{3}{5!}}\, v^{\otimes 5} \\ \vdots \end{bmatrix} \right\rangle.
\end{aligned}$$

We define

$$\Psi(u) = \begin{bmatrix} u \\ \sqrt{\frac{1}{3!}}\, u^{\otimes 3} \\ \sqrt{\frac{3}{5!}}\, u^{\otimes 5} \\ \vdots \end{bmatrix}, \qquad \Phi(v) = \begin{bmatrix} v \\ \sqrt{\frac{1}{3!}}\, v^{\otimes 3} \\ \sqrt{\frac{3}{5!}}\, v^{\otimes 5} \\ \vdots \end{bmatrix},$$

which proves the claim. Next we prove the original inequality:

$$\begin{aligned}
\sum_{i,j} B_{ij}\langle u_i, v_j\rangle &= K \sum_{i,j} B_{ij} \cdot \frac{2}{\pi} \arcsin\langle \Psi(u_i), \Phi(v_j)\rangle \\
&= K \sum_{i,j} B_{ij}\, \mathbb{E}_{w\sim \text{Gaussian}}[\text{sign}\langle w, \Phi(u_i)\rangle \cdot \text{sign}\langle w, \Psi(v_j)\rangle] \\
&\le K \max_{x_i, y_j \in \{\pm 1\}} \left| \sum_{i,j} B_{ij} x_i y_j \right|.
\end{aligned}$$

This completes the proof. □

Theorem 5.32. *Consider an $n \times n$ symmetric, positive-semidefinite matrix A. Let* $\text{INT}(A)$ *denote the maximum in the integer optimization problem*

$$\text{INT}(A) = \max\left\{ \sum_{i,j=1}^{n} A_{ij} x_i x_j \ : \ x_i = \pm 1 \text{ for } i = 1, \ldots, n \right\},$$

and $\text{SDP}(A)$ *denote the maximum in the semidefinite program*

$$\text{SDP}(A) = \max\left\{ \sum_{i,j=1}^{n} A_{ij}\langle X_i, X_j\rangle \ : \ \|X_i\|_2 = 1 \text{ for } i = 1, \ldots, n \right\}.$$

Then

$$\text{INT}(A) \le \text{SDP}(A) \le 2K \cdot \text{INT}(A),$$

where $K \le 1.783$ is the constant in Grothendieck's inequality.

5.3 Random Matrices

5.3.1 Covering Number

Definition 5.33 (ε-net)**.** *Let (T, d) be a metric space. Consider a subset $K \subset T$ and let $\varepsilon > 0$. A subset $N \subseteq K$ is called an* ε-net of K *if every point in K is within distance ε of some point of N, i.e.*

$$\forall x \in K \ \ \exists\, x_0 \in N : \ d(x, x_0) \leq \varepsilon.$$

Equivalently, N is an ε-net of K if and only if K can be covered by balls with centers in N and radii ε.

Definition 5.34 (Covering Numbers)**.** *The smallest possible cardinality of an ε-net of K is called the* covering number *of K and is denoted $N(K, d, \varepsilon)$. Equivalently, $N(K, d, \varepsilon)$ is the smallest number of closed balls with centers in K and radii ε whose union covers K.*

Definition 5.35 (Packing Numbers)**.** *A subset N of a metric space (T, d) is* ε-separated *if $d(x, y) > \varepsilon$ for all distinct points $x, y \in N$. The largest possible cardinality of an ε-separated subset of a given set $K \subset T$ is called the* packing number *of K and is denoted $P(K, d, \varepsilon)$.*

Lemma 5.36 (Equivalence of Covering and Packing Numbers)**.** *For any set $K \subset T$ and any $\varepsilon > 0$, we have*

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

Remark 5.37 (Entropy Numbers)**.** *The corresponding entropy numbers are the logarithms of the covering and packing numbers, respectively. (Discuss in next section).*

5.3.2 Sub-Gaussian Matrices

Theorem 5.38 (Norm of matrices with sub-Gaussian entries)**.** *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-Gaussian random variables. Then, for any $t > 0$ we have*

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

with probability at least $1 - 2\exp(-t^2)$. Here $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$.

Proof. This proof is an example of an ε-net argument. We need to control $\langle Ax, y\rangle$ for all vectors x and y on the unit sphere. To this end, we will discretize the sphere using a net (approximation step), establish a tight control of $\langle Ax, y\rangle$ for fixed vectors x and y from the net (concentration step), and finish by taking a union bound over all x and y in the net.
Step 1: Approximation. Choose $\varepsilon = 1/4$. We can find an ε-net N of the sphere S^{n-1} and an ε-net M of the sphere S^{m-1} with cardinalities

$$|N| \leq 9^n \quad \text{and} \quad |M| \leq 9^m.$$

The operator norm of A can be bounded using these nets as follows:

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

Step 2: Concentration. Fix $x \in N$ and $y \in M$. Then the quadratic form

$$\langle Ax, y\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} x_i y_j$$

is a sum of independent, sub-Gaussian random variables. The sum is sub-Gaussian, and

$$\|\langle Ax, y\rangle\|_{\psi_2}^2 \le C\sum_{i=1}^n\sum_{j=1}^m \|A_{ij}x_iy_j\|_{\psi_2}^2 \le CK^2\sum_{i=1}^n\sum_{j=1}^m x_i^2y_j^2$$
$$= CK^2\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{j=1}^m y_j^2\right) = CK^2.$$

We can restate this as the tail bound

$$\mathbb{P}\{\langle Ax, y\rangle \ge u\} \le 2\exp(-cu^2/K^2), \qquad u \ge 0.$$

Step 3: Union bound. Next, we unfix x and y using a union bound. Suppose the event $\max_{x\in N, y\in M}\langle Ax, y\rangle \ge u$ occurs. Then there exist $x \in N$ and $y \in M$ such that $\langle Ax, y\rangle \ge u$. Thus the union bound yields

$$\mathbb{P}\left\{\max_{x\in N,\, y\in M}\langle Ax, y\rangle \ge u\right\} \le \sum_{x\in N,\, y\in M}\mathbb{P}\{\langle Ax, y\rangle \ge u\}.$$

Using the tail bound and the size estimates on N and M, we bound the probability above by

$$9^{n+m}\cdot 2\exp(-cu^2/K^2).$$

Choose

$$u = CK(\sqrt{n} + \sqrt{m} + t).$$

Then $u^2 \ge C^2K^2(n + m + t^2)$, and if the constant C is chosen sufficiently large, the exponent is large enough, say $cu^2/K^2 \ge 3(n + m) + t^2$. Thus

$$\mathbb{P}\left\{\max_{x\in N,\, y\in M}\langle Ax, y\rangle \ge u\right\} \le 9^{n+m}\cdot 2\exp(-3(n + m) - t^2) \le 2\exp(-t^2).$$

Finally, combining this with the bound from Step 1, we conclude that

$$\mathbb{P}\{\|A\| \ge 2u\} \le 2\exp(-t^2).$$

Recalling our choice of u, we complete the proof. □

Now we are going to prove sharper and two-sided bounds on the entire spectrum of A and relax the independence of entries to just independence of rows.

Lemma 5.39 (Approximate Isometries)**.** *Let A be an $m \times n$ matrix and $\delta > 0$. Suppose that*

$$\|A^\top A - I_n\| \le \max(\delta, \delta^2).$$

Then

$$(1 - \delta)\|x\|_2 \le \|Ax\|_2 \le (1 + \delta)\|x\|_2 \qquad \text{for all } x \in \mathbb{R}^n.$$

Consequently, all singular values of A are between $1 - \delta$ and $1 + \delta$:

$$1 - \delta \le s_n(A) \le s_1(A) \le 1 + \delta.$$

Theorem 5.40 (Two-sided bound on sub-Gaussian matrices)**.** *Let A be an $m \times n$ matrix whose rows A_i are independent, mean zero, sub-Gaussian isotropic random vectors in $\mathbb{R}^n$. Then for any $t \ge 0$ we have*

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

with probability at least $1 - 2\exp(-t^2)$. Here $K = \max_i \|A_i\|_{\psi_2}$.

We will prove a slightly stronger conclusion, namely that

$$\left\| \frac{1}{m} A^\top A - I_n \right\| \le K^2 \max(\delta, \delta^2) \quad \text{where} \quad \delta = C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}} \right).$$

Using the approximate-isometry lemma 5.39, one can quickly check that this implies the singular value bounds.

Proof. We will prove the stronger conclusion using an ε-net argument. This will be similar to the proof for sub-Gaussian matrices, but we now use Bernstein's concentration inequality instead of Hoeffding's.

Step 1: Approximation. We can find a $\frac{1}{4}$-net N of the unit sphere S^{n-1} with cardinality $|N| \le 9^n$. The operator norm can be evaluated on N:

$$\left\| \frac{1}{m} A^\top A - I_n \right\| \le 2 \max_{x \in N} \left| \left\langle \left(\frac{1}{m} A^\top A - I_n \right) x, x \right\rangle \right| = 2 \max_{x \in N} \left| \frac{1}{m} \|Ax\|_2^2 - 1 \right|.$$

To complete the proof it suffices to show that, with the required probability,

$$\max_{x \in N} \left| \frac{1}{m} \|Ax\|_2^2 - 1 \right| \le \frac{\varepsilon}{2} \quad \text{where} \quad \varepsilon := K^2 \max(\delta, \delta^2).$$

Step 2: Concentration. Fix $x \in S^{n-1}$ and express $\|Ax\|_2^2$ as a sum of independent random variables:

$$\|Ax\|_2^2 = \sum_{i=1}^m \langle A_i, x \rangle^2 =: \sum_{i=1}^m X_i^2,$$

where A_i denote the rows of A. By assumption, A_i are independent, isotropic, and sub-Gaussian random vectors with $\|A_i\|_{\psi_2} \le K$. Thus $X_i = \langle A_i, x \rangle$ are independent sub-Gaussian random variables with $\mathbb{E} X_i^2 = 1$ and $\|X_i\|_{\psi_2} \le K$. Therefore $X_i^2 - 1$ are independent, mean-zero, sub-exponential random variables with

$$\|X_i^2 - 1\|_{\psi_1} \le CK^2.$$

Thus we can use Bernstein's inequality and obtain

$$\begin{aligned}
\mathbb{P}\left\{ \left| \frac{1}{m} \|Ax\|_2^2 - 1 \right| \ge \frac{\varepsilon}{2} \right\} &= \mathbb{P}\left\{ \left| \frac{1}{m} \sum_{i=1}^m X_i^2 - 1 \right| \ge \frac{\varepsilon}{2} \right\} \\
&\le 2 \exp\left[-c_1 \min\left(\frac{\varepsilon^2}{K^4}, \frac{\varepsilon}{K^2} \right) m \right] \\
&= 2 \exp\left[-c_1 \delta^2 m \right] && \text{(since } \varepsilon/K^2 = \max(\delta, \delta^2)\text{)} \\
&\le 2 \exp\left[-c_1 C^2 (n + t^2) \right].
\end{aligned}$$

The last bound follows from the definition of δ and the inequality $(a+b)^2 \ge a^2 + b^2$ for $a, b \ge 0$.

Step 3: Union bound. Now we can unfix $x \in N$ using a union bound. Recalling that N has cardinality bounded by 9^n, we obtain

$$\mathbb{P}\left\{ \max_{x \in N} \left| \frac{1}{m} \|Ax\|_2^2 - 1 \right| \ge \frac{\varepsilon}{2} \right\} \le 9^n \cdot 2 \exp\left[-c_1 C^2 (n + t^2) \right] \le 2 \exp(-t^2)$$

if we chose the absolute constant C large enough. As we noted in Step 1, this completes the proof. □

Theorem 5.41 (Matrix Bernstein's Inequality)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^{d\times d}$ be n independent symmetric random matrices satisfying, for all $1 \le i \le n$, $\mathbb{E}(X_i) = 0$, $\|X_i\|_{op} \le b$, and*

$$\max_i \|\mathbb{E}(\sum_i X_i X_i^T)\|_{op} \|\mathbb{E}(\sum_i X_i^T X_i)\|_{op} \le \sigma^2$$

. Then we have:

$$P\left(\left\|\sum_i X_i\right\|_{op} \ge t\right) \le 2d \exp\left(\frac{-t^2/2}{\sigma^2 + bt/3}\right)$$

Proof. We directly apply Lieb's theorem: $f(A) = \mathrm{tr}(\exp(H + \log A))$ is a convex function on the space of matrices of the same size as A. By Jensen's inequality:

$$\mathbb{E}\left[\mathrm{tr}\left(\exp\left(\sum_{i=1}^n \theta X_i\right)\right)\right] \le \mathrm{tr}\left(\exp\left(\sum_{i=1}^n \log \mathbb{E} e^{\theta X_i}\right)\right)$$

Using a Markov-like relaxation technique, we obtain the following concentration inequality for the largest eigenvalue:

$$P(\lambda_{\max}(Y) \ge t) \le \inf_{\theta \ge 0} \mathrm{tr}\left(\exp\left(\sum_{i=1}^n \log \mathbb{E} e^{\theta X_i}\right)\right) e^{-\theta t} \tag{12}$$

We next estimate the log moment generating function $\mathbb{E}e^{\theta X_i}$. From the condition $\|X_i\|_{op} \le b$, we have:

$$e^{\theta X} = 1 + \theta X + X^T f(X) X \preceq 1 + \theta X + f(b) X^2$$

where:

$$f(b) = \frac{e^{\theta b} - 1 - b}{b^2} = \sum_{k=2}^{\infty} \frac{(\theta b)^{k-2}}{3^{k-2}} \cdot \frac{\theta^2}{2} \le \frac{\frac{\theta^2}{2}}{1 - \frac{\theta b}{3}}$$

Therefore:

$$\log \mathbb{E} e^{\theta X} \le \frac{\frac{\theta^2}{2}}{1 - \frac{\theta b}{3}} \mathbb{E} X^2 \tag{13}$$

Finally, combining (12), (13) and the conditions:

$$P(\lambda_{\max}(Y) \ge t) \le \inf_{\theta \ge 0} \mathrm{tr}\left(\exp\left(\frac{\frac{\theta^2}{2}}{1 - \frac{\theta b}{3}} \sum_{i=1}^n \mathbb{E} X_i^2\right)\right) e^{-\theta t}$$

$$\le \inf_{0 \le \theta < \frac{3}{b}} d \exp\left(\frac{\frac{\theta^2}{2}}{1 - \frac{\theta b}{3}} \sigma^2\right) e^{-\theta t}$$

$$\le d \exp\left(\frac{-t^2/2}{\sigma^2 + bt/3}\right) \quad \left(\text{choosing } \theta = \frac{t}{\sigma^2 + \frac{bt}{3}}\right)$$

Therefore:

$$P\left(\left\|\sum_i X_i\right\|_{op} \ge t\right) \le 2d \exp\left(\frac{-t^2/2}{\sigma^2 + bt/3}\right)$$

□

5.3.3 Application: Community Detection in Networks

We consider a graph G containing n nodes, and we partition the node set A into k classes of equal size. Intuitively, if we can find such a clustering, the probability of an edge between two nodes in the same class should be relatively high, while the probability of an edge between two nodes in different classes should be relatively low. Based on this intuition, we give the following definition of the adjacency matrix A of the graph:

$$A_{i,j} \sim \begin{cases} \text{Bernoulli}(p), & \text{if } i,j \text{ are in the same class} \\ \text{Bernoulli}(q), & \text{if } i,j \text{ are in different classes} \end{cases}$$

where $p > q$. We assume that the observed graph is generated according to this probabilistic rule, and we hope to determine, conditional on this graph, the adjacency matrix A^* of the clustered graph:

$$A^*_{i,j} = \begin{cases} 1, & \text{if } i,j \text{ are in the same class} \\ 0, & \text{if } i,j \text{ are in different classes} \end{cases}$$

Note that $A - \frac{p+q}{2}$ has positive expectation on within-class edges and negative expectation on between-class edges. Therefore, we can consider the following optimization formulation:

$$\arg \max_{Y \in \mathbb{R}^{n \times n}} \left\langle A - \frac{p+q}{2}, Y \right\rangle = \sum_{i,j} \left(A_{i,j} - \frac{p+q}{2} \right) Y_{i,j}$$

$$\begin{aligned} \text{s.t.} \quad & Y_{i,j} \in \{0,1\} \\ & Y_{ii} = 1 \\ & \sum_{i,j} Y_{ij} = \frac{n}{k} \\ & \text{rank}(Y) = k \end{aligned}$$

Clearly this problem is non-convex and discontinuous, making it difficult to optimize. We consider relaxing the problem into the following convex semidefinite programming problem:

$$\hat{A} = \arg \max_{Y \in \mathbb{R}^{n \times n}} \left\langle A - \frac{p+q}{2}, Y \right\rangle = \sum_{i,j} \left(A_{i,j} - \frac{p+q}{2} \right) Y_{i,j}$$

$$\begin{aligned} \text{s.t.} \quad & Y_{i,j} \in [0,1] \\ & Y_{ii} = 1 \\ & Y \succeq 0 \end{aligned}$$

Theorem 5.42 (Community Detection)**.** *If $p \geq \frac{1}{n}$, then with probability $\geq 1 - 2\left(\frac{2}{e}\right)^n$:*

$$\frac{1}{n^2} \|\hat{A} - A\|_1 \lesssim \sqrt{\frac{p}{(p-q)^2 n}}$$

Remark 5.43. *We observe that the larger the gap between p and q, the smaller the error. Moreover, only when $\frac{(p-q)^2}{p} \lesssim \frac{1}{n}$, i.e., $p \lesssim \frac{1}{n}$, the right-hand side is $\gtrsim 1$ and the conclusion becomes trivial. This means that as long as $p \gtrsim \frac{1}{n}$ — that is, our graph can be sparse — the conclusion still holds. This answers why the title of the paper is "Community Detection in Sparse Networks". In fact, as long as the network has enough nodes, we can recover the community structure from a sparse graph!*

Proof. Since $\hat{A}$ is the optimal solution of the semidefinite program, plugging it into (8) we have:

$$\left\langle A - \frac{p+q}{2}, \hat{A} \right\rangle \geq \left\langle A - \frac{p+q}{2}, A^* \right\rangle \tag{14}$$

$$\implies 0 \geq \left\langle A - \frac{p+q}{2}, A^* - \hat{A} \right\rangle \geq \langle A - \mathbb{E}A, A^* - \hat{A} \rangle + \left\langle \mathbb{E}A - \frac{p+q}{2}, A^* - \hat{A} \right\rangle \tag{15}$$

We estimate the two terms separately:

$$\left\langle \mathbb{E}A - \frac{p+q}{2}, A^* - \hat{A} \right\rangle \leq \frac{p-q}{2} \sum_{i,j} |A^*_{i,j} - \hat{A}_{i,j}| = \frac{p-q}{2} \|A^* - \hat{A}\|_1 \tag{16}$$

Combining (14), (15) and applying Grothendieck's Inequality:

$$\frac{p-q}{2} \|A^* - \hat{A}\|_1 \leq \langle A - \mathbb{E}A, \hat{A} - A^* \rangle \tag{17}$$

$$\leq |\langle A - \mathbb{E}A, \hat{A} \rangle| + |\langle A - \mathbb{E}A, A^* \rangle| \tag{18}$$

$$\leq 2 \max_{Y_{ii}=1,\, Y \succeq 0} |\langle A - \mathbb{E}A, Y \rangle| \tag{19}$$

$$\leq 2K \max_{x,y \in \{\pm 1\}^n} \left| \sum_{i,j} (A_{ij} - \mathbb{E}A_{ij}) x_i y_j \right| \tag{20}$$

Looking at (19), this is a sum of n^2 independent random variables. We treat it as n^2 matrices of size 1×1 and apply Lemma 1. Let $Z_{ij} = (A_{ij} - \mathbb{E}A_{ij}) x_i y_j$, $Z = \sum_{i,j} Z_{ij}$, $\|Z_{ij}\|_{op} = |Z_{ij}| \leq 1$, $\mathrm{var}(Z_{ij}) \leq p(1-p) \leq p$:

$$P(|Z| \geq t) \leq 2 \exp\left(-\frac{t^2}{pn^2 + t/3} \right) \leq 2 \exp\left(-\frac{ct^2}{pn^2 + t} \right)$$

Taking a union bound over all i, j, we obtain:

$$P\left(\max_{(x,y) \in \{\pm 1\}^n \times \{\pm 1\}^n} \left| \sum_{i,j} (A_{ij} - \mathbb{E}A_{ij}) x_i y_j \right| \geq t \right) \leq 2 \exp\left(-\frac{ct^2}{pn^2 + t} \right) 4^n$$

We choose t sufficiently large: $t = c(\sqrt{pn^3} + n)$, so that the right-hand side is $\geq (2e)^{-n}$ for a sufficiently large constant c. Therefore, with probability $\geq 1 - (2e)^{-n}$:

$$\max_{(x,y) \in \{\pm 1\}^n \times \{\pm 1\}^n} \left| \sum_{i,j} (A_{ij} - \mathbb{E}A_{ij}) x_i y_j \right| \lesssim \sqrt{pn^3} + n \lesssim \sqrt{pn^3}$$

Combining (16), (20) and applying the lemma:

$$\frac{p-q}{2} \|A^* - \hat{A}\|_1 \lesssim \sqrt{pn^3}$$

That is:

$$\frac{1}{n^2} \|A^* - \hat{A}\|_1 \lesssim \sqrt{\frac{p}{(p-q)^2 n}}$$

□

Remark 5.44. *This upper bound can be optimized to* $e^{-\Omega((p-q)^2 n/p)}$.

By Theorem 5.42, we know that the solution obtained from the convex-relaxed semidefinite program is close to the true solution with high probability. One may also wonder how to determine the true clustering result from $\hat{Y}$. We consider the case $k = 2$: without loss of generality, set $Y^* = \begin{pmatrix} \mathbf{1} & 0 \\ 0 & \mathbf{1} \end{pmatrix}$, let u^* be the top singular vector of $Y^* - 1/2$, $u^* = \left(\sqrt{\frac{1}{n}}, \dots, \sqrt{\frac{1}{n}}, -\sqrt{\frac{1}{n}}, \dots, -\sqrt{\frac{1}{n}}\right)$. Let $\hat{u}$ be the top singular vector of $\hat{Y} - 1/2$; the indices of positive and negative components of $\hat{u}$ form the two classes, giving the clustering result. This is guaranteed by the following theorem:

Theorem 5.45. *If* $\frac{1}{n}\|A^* - \hat{A}\|_1 \le \frac{1}{20}$, *then:*

$$\frac{1}{n} \min_{\alpha \in \{\pm 1\}} \#\{i : \operatorname{sgn}(\hat{u}_i) \neq \alpha \cdot \operatorname{sgn}(u_i^*)\} \lesssim \frac{1}{n^2}\|A^* - \hat{A}\|$$

5.4 Concentration of Lipschitz Functions

Theorem 5.46 (Gaussian Concentration)**.** *Let* $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ *and* $f : \mathbb{R}^n \to \mathbb{R}$ *such that* f *is* L-***Lipschitz*** *in* $\|\cdot\|_2$, *i.e.*

$$|f(x) - f(y)| \le L\|x - y\|_2 \qquad \forall x, y \in \mathbb{R}^n.$$

Then

1. $f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$ *is* $\mathrm{sG}(L)$.
2. $$\mathbb{P}(|f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]| \ge t) \le 2\exp\left(-\frac{t^2}{2L^2}\right).$$

Theorem 5.47 (Concentration of Convex Lipschitz Functions)**.** *Suppose that*

1. f *is* L*-Lipschitz and convex:*
$$\nabla^2 f(x) \succ 0 \quad \text{if } \nabla^2 f \text{ exists}$$
2. $(X_i)_{i \in [n]}$ *independent with* $X_i \in [a, b]$ *a.s.*

Then $f(X_{1:n}) - \mathbb{E}[f]$ *is* $\mathrm{sG}(L(b-a))$, *so*

$$\mathbb{P}(|f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]| \ge t) \le 2\exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

5.5 Martingale Concentration Inequalities

5.5.1 Martingale Concentration

Now we discuss martingale case. Recall that $\{D_k\}_{k \ge 1}$ is a martingale difference sequence if $\{\sum_{k=1}^n D_k\}_{n \ge 1}$ is a martingale with respect to $\{\mathscr{F}_k\}_{k \ge 1}$.

Theorem 5.48 (Martingale Concentration Inequality)**.** *Let* $\{(D_k, \mathscr{F}_k)\}$ *be a martingale difference sequence. If*

$$\mathbb{E}[e^{\lambda D_k} \mid \mathscr{F}_{k-1}] \le e^{\lambda^2 \nu_k^2/2} \quad \text{a.s. } \forall \lambda \le \frac{1}{\alpha_k},$$

then

1. $\sum_{k=1}^n D_k$ *is* $\mathrm{sE}\left(\sqrt{\sum_{k=1}^n \nu_k^2}, \max_{k \le n} \alpha_k\right)$.

2.

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} D_k\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t^2}{2\sum_{k=1}^{n}\nu_k^2}, \frac{t}{2\alpha_*}\right\}\right)$$

Corollary 5.49 (Azuma-Hoeffding Inequality)**.** *Let $\{(D_k, \mathscr{F}_k)\}$ be a martingale difference sequence. Suppose there exists $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in (a_k, b_k)$ a.s., where b_k, a_k are $\mathscr{F}_{k-1}$-measurable and $|b_k - a_k| \leq L_k$. Then*

1. *$\sum_{k=1}^n D_k$ is $\mathrm{sG}\left(\sqrt{\sum_{k=1}^n L_k^2}/2\right)$.*
2. $$\mathbb{P}\left(\left|\sum_{k=1}^{n} D_k\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right).$$

Definition 5.50. *$f(x_1, \ldots, x_n)$ is a **bounded difference function** if for all $k \in [n]$, $x_{1:n}, x_k'$,*

$$|f(x_{1:k-1}, x_k, x_{k+1:n}) - f(x_{1:k-1}, x_k', x_{k+1:n})| \leq L_k.$$

This is a condition on how much the function changes if we change 1 coordinate. Here is a corollary of the Azuma-Hoeffding inequality.

Corollary 5.51 (McDiarmid's Inequality)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is $L_{1:n}$ bounded and $X_{1:n}$ has independent components. Then for all $t \geq 0$,*

$$\mathbb{P}(|f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}\right).$$

5.5.2 Gaussian Complexity

Gaussian complexity is a very important notion in compressed sensing. Suppose we have a set $A \subseteq \mathbb{R}^n$. How do we measure its "size"? A reasonable size function S should at least satisfy $S(A) \leq S(B)$ if $A \subseteq B$. Here are some reasonable size functions:

1. Euclidean width: $D(A) = \max_{a \in A} \|a\|_2$.
2. Dimension: A line has dimension 1, and a plane has dimension 2.

Definition 5.52. *Given a set A, let $W = (W_1, \ldots, W_n)^\top \in \mathbb{R}^n$ with $W_i \overset{iid}{\sim} \mathscr{N}(0, 1)$. The **Gaussian complexity** or "**statistical dimension**" of A is*

$$\mathscr{G}(A) := \mathbb{E}_{W \sim N(0, I_n)}\left[\sup_{a \in A} \langle a, W\rangle\right].$$

Note that if we don't take the supremum in the expectation, the quantity would be 0. This quantity is always nonnegative.

Example 5.53. *Let $B_p(r) = \{x \in \mathbb{R}^n : \|x\|_p \leq r\}$. Then*

$$\mathscr{G}(B_p(r)) = \mathbb{E}\left[\sup_{\|x\|_p \leq r} \langle x, W\rangle\right].$$

If q is the conjugate exponent of p, so $\frac{1}{p} + \frac{1}{q} = 1$, this is the variational representation of the $\|\cdot\|_q$ norm:

$$r\,\mathbb{E}[\|W\|_q] \approx r n^{1/q}.$$

Note that if $p_1 \le p_2$, then $q_1 \ge q_2$, so $\mathscr{G}(B_{p_1}(r)) \le \mathscr{G}(B_{p_2}(r))$.
We want to show that $f(W) := \sup_{a\in A}\langle a, W\rangle$ concentrates. Fix $w, w' \in \mathbb{R}^n$. Then

$$f(w) - f(w') = \sup_{a\in A}\langle a, w\rangle - \sup_{a\in A}\langle a, w'\rangle.$$

Denote $a^* = \arg\max_a \langle a, w\rangle$:

$$\begin{aligned} f(w) - f(w') &= \langle a^*, w\rangle - \sup_{a\in A}\langle a, w'\rangle \\ &= \inf_{a\in A}\langle a^*, w\rangle - \langle a, w'\rangle \\ &\le \langle a^*, w - w'\rangle \\ &\le \|a_*\|\|w - w'\|_2 \\ &\le D(A)\|w - w'\|_2. \end{aligned}$$

The other side can be proven similarly, so f is $D(A)$-Lipschitz. Concentration says that $f(W)$ is $\mathrm{sG}(D(A))$.

Example 5.54. *If we let $A = B_2(r)$, then*

$$\mathbb{E}[f(W)] = \mathscr{G}(B_2(r)) = r\sqrt{n},$$

since $D(A) = r$.

6 Function Spaces

Data in the real world often takes the form $y = f(x) + \epsilon$, and a central question is to understand the function space in which f lives. In this section, we examine several perspectives on function spaces. The section is organized as follows:

1. Rademacher Complexity
2. Metric Entropy Method
3. Glivenko-Cantelli Theorem and Donsker Theorem
4. Information Theory

This section mainly follows STAT210B (UC Berkeley, taught by Song Mei), High-dimensional probability (PKU, taught by Zhihua Zhang), Introduction to Machine Learning (PKU, taught by Lei Wu), STAT300B (Stanford), STAT364 (Yale). I also referred to the book High-Dimensional Statistics: A Non-Asymptotic Viewpoint [6].

6.1 Rademacher Complexity

6.1.1 Empirical Process Theory

Suppose $(X_i)_{i\in[n]} \overset{\text{iid}}{\sim} X \sim \mathbb{P}$, and suppose we have a **function class** $\mathscr{F} \subseteq \{f : \mathscr{X} \to \mathbb{R} : \mathbb{E}[|f(X)|] < \infty\}$.

Definition 6.1 (Empirical Process)**.** *The **empirical process** indexed by $\mathscr{F}$ is*

$$\left\{ \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right) : f \in \mathscr{F} \right\}.$$

Define

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right|.$$

Here, $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ is the **empirical measure**. This is the object we will study for the next portion of the course. If there is only one function f, we can deal with this using the law of large numbers and concentration inequalities. We will learn how to deal with this object using empirical process theory.

Why do we care about the maximum of the empirical process in statistics and machine learning? Recall the following setup:

Data distribution	$(X_i)_{i\in[n]} \overset{\text{iid}}{\sim} \mathbb{P}$
Loss function	$L : \mathscr{X} \times \Theta \to \mathbb{R}$
Empirical risk	$\widehat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i;\theta)$
Population risk	$R(\theta) = \mathbb{E}_{X\sim\mathbb{P}}[\ell(X;\theta)]$
Empirical risk minimizer	$\hat{\theta} = \arg\min_\theta \widehat{R}(\theta)$
Population risk minimizer	$\theta_* = \arg\min_\theta R(\theta)$
Excess risk	$E = R(\hat{\theta}) - R(\theta_*)$

We train $\hat{\theta}$ on the empirical risk, so we want the empirical risk to be close to the population risk. So to make sure training on our training data is accurate, we want to make the excess risk small. The excess risk has the following decomposition:

$$E = \underbrace{(R(\hat{\theta}) - \widehat{R}_n(\hat{\theta}))}_{\text{Gap}} + \underbrace{(\widehat{R}_n(\hat{\theta}) - \widehat{R}_n(\theta_*))}_{\leq 0} + \underbrace{(\widehat{R}_n(\theta_*) - R(\theta_*))}_{\text{bound using Hoeffding}}$$

The Gap is

$$\text{Gap} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(X;\hat{\theta}) - \ell(X_i;\hat{\theta})].$$

We cannot use the strong law of large numbers to examine this because the $\ell(X_i;\hat{\theta})$ are not independent random variables. We can fix this by replacing $\hat{\theta}$ by the sup over θ:

$$\leq \sup_{\theta} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(X;\theta) - \ell(X_i;\theta)] \right|.$$

Here, $f(X) = \ell(X;\theta)$, so we want to look at the function class $\mathscr{F} = \{\ell(\cdot;\theta) : \theta \in \Theta\}$.

Definition 6.2 (Glivenko-Cantelli Class)**.** *We say that $\mathscr{F}$ is a **Glivenko-Cantelli class** for $\mathbb{P}$ if*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \xrightarrow{P} 0.$$

Example 6.3. *The Glivenko-Cantelli theorem says that $\mathscr{F}_1 = \{\mathbb{1}_{\{x\leq t\}}\}_{t\in\mathbb{R}}$ is a Glivenko-Cantelli class for any $\mathbb{P} \in \mathscr{P}(\mathbb{R})$.*

6.1.2 Rademacher Complexity Bounds

Recall that the Rademacher complexity of a set $A \subseteq \mathbb{R}^n$ is

$$\mathscr{R}(A) := \mathbb{E}_{\varepsilon \overset{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})} \left\{ \sup_{a\in A} \langle a, \varepsilon \rangle \right\}$$

Definition 6.4 (Rademacher Complexity)**.** *Given a function class $\mathscr{F}$ and a fixed data set $(x_i)_{i\in[n]} \subseteq \mathscr{X}$, let*

$$\mathscr{F}(x_{1:n}) := \{(f(x_1), \ldots, f(x_n)) : f \in \mathscr{F}\} \subseteq \mathbb{R}^n.$$

*The **Rademacher complexity** of the function class $\mathscr{F}$ and the data set $(x_i)_{i\in[n]}$ is*

$$\mathscr{R}(\mathscr{F}(x_{1:n})/n) := \mathbb{E}_{\varepsilon \overset{iid}{\sim} Unif(\{\pm 1\})} \left\{ \sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right\}.$$

If we write $\mathscr{A} = \pm\mathscr{F}(x_{1:n})/n$, then we can relate Rademacher complexity of sets and function classes by

$$\widetilde{\mathscr{R}}(A) = \mathscr{R}(\mathscr{F}(x_{1:n})/n),$$

where $\widetilde{\mathscr{R}}$ denotes the Rademacher complexity of a set.

Definition 6.5 (Rademacher Complexity)**.** *Given a function class $\mathscr{F}$ and a distribution $\mathbb{P} \in \mathscr{P}(\mathscr{X})$, let $(X_i)_{i\in[n]} \overset{iid}{\sim} \mathbb{P}$. The **Rademacher complexity** of the function class $\mathscr{F}$ is*

$$\mathscr{R}(\mathscr{F}) := \mathbb{E}_{X_i \overset{iid}{\sim} \mathbb{P}}[\mathscr{R}(\mathscr{F}(X_{1:n})/n)].$$

First, observe that if $\mathscr{F}_1 \subseteq \mathscr{F}_2$, then $\mathscr{R}_n(\mathscr{F}_1) \leq \mathscr{R}_n(\mathscr{F}_2)$, so this is a measure of the size of a function class.

Example 6.6. *Let $\psi : \mathbb{R}^d \to \mathbb{R}^p$ be a fixed feature map, and consider the function class*

$$\mathscr{F} = \{f(x) = \langle \psi(x), \theta \rangle : \|\theta\|_2 \leq B\}.$$

Then the Rademacher complexity of this function class is

$$\begin{aligned}
\mathscr{R}_n(\mathscr{F}) &= \mathbb{E}_{X_i, \varepsilon_i} \left[\sup_{\|\theta\|_2 \leq B} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle \psi(X_i), \theta \rangle \right| \right] \\
&= \mathbb{E}_{X_i, \varepsilon_i} \left[\sup_{\|\theta\|_2 \leq B} \left| \varepsilon_i \left\langle \frac{1}{n} \sum_{i=1}^n \psi(X_i), \theta \right\rangle \right| \right] \\
&= \mathbb{E}_{X_i, \varepsilon_i} \left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \psi(X_i) \right\|_2 \right] \cdot B
\end{aligned}$$

Using Cauchy-Schwarz,

$$\begin{aligned}
&\leq \mathbb{E}_{X_i, \varepsilon_i} \left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \psi(X_i) \right\|_2^2 \right]^{1/2} \cdot B \\
&= \mathbb{E}_{X_i, \varepsilon_i} \left[\frac{1}{n^2} \sum_{i=1}^n \varepsilon_i^2 \|\psi(X_i)\|_2^2 \right]^{1/2} \cdot B \\
&= \frac{B}{\sqrt{n}} \mathbb{E}[\|\psi(X)\|_2^2]^{1/2}.
\end{aligned}$$

Remark 6.7. *Why introduce Rademacher complexity?*

1. *We will show that*
$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \approx \mathscr{R}_n(\mathscr{F}).$$
2. *The Rademacher complexity is easier to upper bound. We will have tools to upper bound it, such as*
   - *VC dimension,*
   - *Metric entropy methods.*

Proposition 6.8 (Rademacher Complexity Bounds)**.** *For any function class $\mathscr{F}$ and distribution $\mathbb{P}$,*

$$\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}] \leq 2\mathscr{R}_n(\mathscr{F}).$$

Proof. Let $Y_i \overset{\text{iid}}{\sim} X_i$ be independent of X_i. Then

$$\begin{aligned}
\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}] &= \mathbb{E} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)] \right| \right] \\
&= \mathbb{E}_{X_{1:n}} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_{Y_{1:n}}[f(Y_i)] \right| \right] \\
&\leq \mathbb{E}_{X_{1:n}, Y_{1:n}} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(Y_i)) \right| \right]
\end{aligned}$$

We can introduce a Rademacher random variable without changing the distribution.

$$
\begin{aligned}
&= \mathbb{E}_{X_{1:n},Y_{1:n},\varepsilon_{1:n}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i)-f(Y_i))\right|\right]\\
&\le \mathbb{E}_{X_{1:n},Y_{1:n},\varepsilon_{1:n}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|-\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right|\right]\\
&\le 2\mathscr{R}_n(\mathscr{F}).
\end{aligned}
$$

□

Define

$$\|\mathbb{S}_n\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|.$$

Here is an upgraded version.

Proposition 6.9. *For every convex, nondecreasing function $\Phi:\mathbb{R}\to\mathbb{R}$,*

$$\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\tfrac{1}{2}\|\mathbb{S}_n\|_{\overline{\mathscr{F}}}\right)\right] \le \mathbb{E}_X[\Phi(\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}})] \le \mathbb{E}_{X,\varepsilon}[\Phi(2\|\mathbb{S}_n\|_{\mathscr{F}})],$$

where $\overline{\mathscr{F}} = \{f-\mathbb{E}[f] : f\in\mathscr{F}\}$.

Remark 6.10. *Making $\Phi(t)=t$ retrieves the bound on $\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}$ in terms of Rademacher complexity. We can also take the upper bound to also be $\overline{\mathscr{F}}$ because $\mathbb{E}[\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}] = \mathbb{E}[\|\mathbb{P}_n-\mathbb{P}\|_{\overline{\mathscr{F}}}]$.*

Suppose that for all $f\in\mathscr{F}$, $\|f\|_\infty \le b$. Then $\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}$ is $(2b/n,\dots,2b/n)$-bounded difference. The bounded difference inequality then gives that $\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}$ is $\mathrm{sG}(b/\sqrt{n})$. In other words,

$$\left|\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}} - \mathbb{E}[\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}]\right| \le b\sqrt{\frac{\log(2/\delta)}{n}} \quad \text{with probability } 1-\delta.$$

How do we upper bound the Rademacher complexity? Let's take a higher level picture and try to bound $\mathbb{E}[\sup_{\theta\in\Theta} X_\theta]$. In many cases, X_θ is sub-Gaussian for each fixed θ.

The simplest case is when Θ is finite. In this case, we have a **maximal inequality**: If for all $\theta\in\Theta$, $X_\theta\in\mathrm{sG}(\sigma)$, then

$$\mathbb{E}\left[\max_{\theta\in\Theta} X_\theta\right] \le \sigma\sqrt{2\log|\Theta|}.$$

However, typically, this set Θ is infinite, so the maximal inequality cannot handle this case.

6.1.3 VC Dimension

Definition 6.11 (VC Dimension)**.** *Suppose $\mathscr{F}\subseteq\{F:\mathscr{X}\to\{0,1\}\}$ is binary valued. We say that $x_{1:n}$ is **shattered** by $\mathscr{F}$ if $|\mathscr{F}(x_{1:n})| = 2^n$. The **VC dimension**, $\nu(\mathscr{F})$, is the largest n such that there exists $x_{1:n}$ shattered by $\mathscr{F}$.*

Note that $|\mathscr{F}(X_{1:n})| \le 2^n$ always. So we want $\mathscr{F}$ to be able to distinguish between points in a maximal sense.

Proposition 6.12 (Vapnik-Chervonenkis, Sauer-Shelah)**.** *For $\mathscr{F}$ with VC dimension ν,*

$$\sup_{x_{1:n}}|\mathscr{F}(x_{1:n})| \le \sum_{i=1}^{\nu}\binom{n}{i} \le \min\left\{(n+1)^\nu, \left(\frac{ne}{\nu}\right)^\nu\right\}.$$

By this proposition, we immediately have

$$\mathscr{R}_n(\mathscr{F}) \le D\sqrt{\frac{\nu \log(n+1)}{n}}.$$

Example 6.13. *Let $\phi_1, \dots, \phi_p : \mathscr{X} \to \mathbb{R}$ be linear (which you can think of as feature maps), and consider $\mathscr{F} = \{\mathbb{1}_{\{\sum_{i=1}^p a_i \phi_i(x) \le b\}} : a_i, b \in \mathbb{R}\}$. Then $\nu(\mathscr{F}) \le p+1$. If $\mathscr{F} = \{\mathbb{1}_{\{\sum_{i=1}^p a_i \phi_i(x) \le b\}} : a_i, b \in \mathbb{R}\}$ and $(X_i)_{i\in[n]} \overset{iid}{\sim} \mathbb{P}$, then*

$$\sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \lesssim \sqrt{\frac{(p+1)\log n}{n}}.$$

This $\log n$ factor can be eliminated later by the chaining method.

6.2 Metric Entropy Method

6.2.1 Entropies of Function Classes

We want to understand the ε-covering number for $T \subseteq \mathbb{R}^d$. The intuition is that

$$\log N(\varepsilon; T, \rho) \asymp \log \frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B_\rho(\varepsilon))},$$

Lemma 6.14 (Volume Bounds)**.**

$$\frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B_\rho(\varepsilon))} \le N(\varepsilon; T, \rho) \le M(\varepsilon; T, \rho) \le \frac{\mathrm{Vol}(T + B_\rho(\varepsilon/2))}{\mathrm{Vol}(B_\rho(\varepsilon/2))},$$

where $T + B_\rho(\varepsilon/2) = \{a + b : a \in T, b \in B_\rho(\varepsilon/2)\}$.

Now we discuss a more complicated example.

Definition 6.15 (Hölder Space)**.** *Let $\mathscr{X} \subset \mathbb{R}^d$ be bounded. For $\alpha > 0$ let $\underline{\alpha}$ be the greatest integer strictly smaller than α. Define*

$$\|f\|_\alpha = \max_{0 \le k_\cdot \le \underline{\alpha}} \sup_x |D^k f(x)| + \max_{k_\cdot = \underline{\alpha}} \sup_{x \ne y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}}.$$

The Hölder space $C^\alpha(\mathscr{X})$ consists of continuous $f : \overline{\mathscr{X}} \to \mathbb{R}$ with $\|f\|_\alpha < \infty$; its unit ball is $C^\alpha(\mathscr{X})_1 = \{f : \|f\|_\alpha \le 1\}$.

Theorem 6.16 (Hölder Space Entropy Bounds)**.** *Let $\mathscr{X} \subset \mathbb{R}^d$ be bounded, convex with nonempty interior. There exists a constant K depending only on α and d such that*

$$\log N\big(\varepsilon, C^\alpha(\mathscr{X})_1, \|\cdot\|_\infty\big) \le K\, \lambda(\mathscr{X}^1) \big(1/\varepsilon\big)^{d/\alpha}.$$

Proof. Fix $\delta = \epsilon^{\frac{1}{\alpha}} \le 1$ and a $\delta-$net $x_1, \cdots, x_m$. For each $k = (k_1, \cdots, k_d)$ with $k_\cdot = \sum k_i \le \lfloor \alpha \rfloor$, define the vector

$$A_k f = \left(\left\lfloor \frac{D^k f(x_1)}{\delta^{\alpha - k_\cdot}} \right\rfloor, \dots, \left\lfloor \frac{D^k f(x_m)}{\delta^{\alpha - k_\cdot}} \right\rfloor \right).$$

If $Af = Ag$, then we have $\|f - g\|_\infty \lesssim \epsilon$. Now we count the number of vectors. Each column of the matrix can have at most $(2\delta^{-\alpha} + 2)^{(\beta+1)^d}$ different values.

Assume WLOG that $x_1, \dots, x_m$ have been chosen and ordered such that for each $j > 1$ there is an index $i < j$ with $\|x_i - x_j\| < 2\delta$. Then use the crude bound obtained previously for the first

column only. For each later column, indexed by x_j, there exists a previous x_i with $\|x_i - x_j\| < 2\delta$. By Taylor's theorem,

$$D^k f(x_j) = \sum_{k.+l. \le \beta} D^{k+l} f(x_i) \frac{(x_i - x_j)^l}{l!} + R,$$

where $|R| \lesssim \|x_i - x_j\|^{\alpha - k.}$. Thus with $B_k f = \delta^{\alpha - k.} A_k f$,

$$\begin{aligned}\left| D^k f(x_j) - \sum_{k.+l. \le \beta} B_{k+l} f(x_i) \frac{(x_i - x_j)^l}{l!} \right| &\lesssim \sum_{k.+l. \le \beta} \left| B_{k+l} f(x_i) - D^{k+l} f(x_i) \right| \frac{\|x_i - x_j\|^l}{l!} + \delta^{\alpha - k.} \\ &\lesssim \sum_{k.+l. \le \beta} \delta^{\alpha - k. - l.} \cdot \frac{\delta^{l.}}{l!} + \delta^{\alpha - k.} \lesssim \delta^{\alpha - k.}.\end{aligned}$$

Thus given the values in the ith column of Af, the values $D^k f(x_j)$ range over an interval of length proportional to $\delta^{\alpha - k.}$. It follows that the values in the jth column of Af range over integers in an interval of length proportional to $\delta^{k. - \alpha} \delta^{\alpha - k.} = 1$. Consequently, there exists a constant C depending only on α and d such that

$$\# Af \le (2\delta^{-\alpha} + 2)^{(\beta+1)^d} C^{m-1}.$$

The theorem follows upon replacing δ by $\varepsilon^{1/\alpha}$ and m by its upper bound $\lambda(\mathscr{X}^1)\varepsilon^{-d/\alpha}$, respectively, taking logarithms, and bounding $\log(1/\varepsilon)$ by a constant times $(1/\varepsilon)^{d/\alpha}$. □

6.2.2 One-step Discretization Bound

Now we are going to discuss the metric entropy method for obtaining bounds on empirical processes. We have a metric space (T, ρ), and we want to control

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \quad \text{or} \quad \mathbb{E}\left[\sup_{\theta \in T} |X_\theta|\right],$$

where X_θ is usually mean 0 and sub-Gaussian. We introduced the metric entropy is $\log N(\varepsilon; T, \rho)$, where $N(\varepsilon; T, \rho) = \inf\{N : |T_\varepsilon| = N, T_\varepsilon \text{ is an } \varepsilon\text{-cover}\}$ is the ε-covering number.

Here is the one-step discretization bound that the maximal inequality gives us:

Lemma 6.17. *If $X_\theta \sim \mathrm{sG}(\sigma)$ for all $\theta \in T$, then*

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta \in T} |X_\theta|\right] &\lesssim \inf_{\varepsilon} \inf_{\varepsilon\text{-cover } T_\varepsilon} \mathbb{E}\left[\sup_{\theta \in T_\varepsilon} |X_\theta|\right] + \mathbb{E}\left[\sup_{\rho(\theta, \widetilde{\theta}) \le \varepsilon} |X_\theta - X_{\widetilde{\theta}}|\right] \\ &\lesssim \inf_{\varepsilon} \sigma \sqrt{\log(N(\varepsilon; T, \rho))} + \mathbb{E}\left[\sup_{\rho(\theta, \widetilde{\theta}) \le \varepsilon} |X_\theta - X_{\widetilde{\theta}}|\right]\end{aligned}$$

Example 6.18 (Gaussian Complexity)**.** *Consider $W_i \overset{iid}{\sim} N(0,1)$, so $\langle W, \theta \rangle \sim \mathrm{sG}(\|\theta\|_2)$. Then we know that*

$$\mathscr{G}(B_2(1)) = \mathbb{E}\left[\sup_{\theta \in B_2(1)} \langle W, \theta \rangle\right] = \mathbb{E}[\|W\|_2] \asymp \sqrt{d}.$$

Here is another way to get this computation:

$$\begin{aligned}
\mathscr{G}(B_2(1)) &\le C\left[\sup_{\theta\in B_2(1)} \underbrace{\|\theta\|_2}_{=1} \underbrace{\sqrt{\log N(\varepsilon; B_2(1), \|\cdot\|_2)}}_{\le\sqrt{d\log(1+2/\varepsilon)}} + \mathbb{E}_W\left[\sup_{\|\theta-\theta'\|_2\le\varepsilon} |W_\theta - W_{\theta'}|\right]\right] \\
&\le C\left[\sqrt{d\log(1+2/\varepsilon)} + \mathbb{E}_W\left[\sup_{\|\theta-\widetilde{\theta}\|_2\le\varepsilon} \langle W, \theta-\theta'\rangle\right]\right] \\
&= C\left[\sqrt{d\log(1+2/\varepsilon)} + \mathbb{E}_W\left[\sup_{\|r\|_2\le\varepsilon} \langle W, r\rangle\right]\right] \\
&= C\left[\sqrt{d\log(1+2/\varepsilon)} + \varepsilon\,\mathbb{E}_W \underbrace{\left[\sup_{\|\widetilde{r}\|_2\le 1} \langle W, \widetilde{r}\rangle\right]}_{\mathscr{G}(B_2(1))}\right].
\end{aligned}$$

This tells us that

$$\mathscr{G}(B_2(1)) \le C\sqrt{d\log(1+2/\varepsilon)} + C\varepsilon\mathscr{G}(B_2(1)).$$

If we take $\varepsilon \le \frac{1}{2C}$*, then we get*

$$\mathscr{G}(B_2(1)) \le 2C\sqrt{d\log(1+4C)} \asymp \sqrt{d},$$

which is the same order as before.

6.2.3 Chaining Method

We have been using the bound

$$\mathbb{E}\left[\sup_{\theta\in T}|X_\theta|\right] \lesssim \inf_{\varepsilon} \underbrace{\inf_{\varepsilon\text{-cover } T_\varepsilon} \mathbb{E}\left[\sup_{\theta\in T_\varepsilon}|X_\theta|\right]}_{\text{bdd by covering number}} + \underbrace{\mathbb{E}\left[\sup_{\rho(\theta,\widetilde{\theta})\le\varepsilon}|X_\theta - X_{\widetilde{\theta}}|\right]}_{\text{how to give tight control?}}$$

Controlling the right term can require ad-hoc arguments. The chaining method gives a way to bound this effectively.

Definition 6.19 (Sub-Gaussian Process)**.** *$\{X_\theta\}_{\theta\in T}$ is a **sub-Gaussian process** with respect to ρ on T if*

$$\mathbb{E}[e^{\lambda(X_\theta - X_{\theta'})}] \le e^{\lambda^2\rho(\theta,\theta')^2/2},$$

or, equivalently, $X_\theta - X_{\theta'}$ is $\mathrm{sG}(\rho(\theta,\theta'))$.

Example 6.20. *Let $T \subseteq \mathbb{R}^d$ with $\rho = \|\cdot\|_2$. Look at $X_\theta = \langle W, \theta\rangle$, where $W \sim N(0, I_d)$. To bound the Gaussian complexity, we want to bound $\mathbb{E}[\sup_{\theta\in T} X_\theta]$. Then $X_\theta - X'_\theta = \langle W, \theta-\theta'\rangle \sim N(0, \|\theta-\theta'\|_2^2) \sim \mathrm{sG}(\|\theta-\theta'\|_2)$.*

Proposition 6.21 (Chaining Methods)**.** *Let $\{X_\theta, \theta\in T\}$ be a mean 0 sub-Gaussian process with metric ρ. Then if $D = \sup_{\theta,\widetilde{\theta}\in T}$,*

$$\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}}(X_\theta - X_{\widetilde{\theta}})\right] \le \inf_{\varepsilon\le D} 2\left[\sup_{\rho(r,r')\le\varepsilon}(X_r - X_{r'})\right] + 32\underbrace{\int_\varepsilon^D \sqrt{\log N(u;T,\rho)}\,du}_{=:J(\varepsilon;D;T,\rho)}.$$

Here, $J(\varepsilon; D; T, \rho)$ is known as **Dudley's entropy integral**.

Remark 6.22. *This gives an upper bound for $\mathbb{E}[\sup_{\theta\in T} X_\theta]$ because by the 0 mean condition and Jensen's inequality,*

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right] &= \mathbb{E}\left[\sup_{\theta,\theta'\in T}(X_\theta - \mathbb{E}_{\theta'}[X_{\theta'}])\right] \\ &\le \mathbb{E}\left[\sup_{\theta,\widetilde{\theta}}(X_\theta - X_{\widetilde{\theta}})\right].\end{aligned}$$

Proof. Take a sequence of ε-coverings corresponding to $\varepsilon_m = D/2^m$ for $m = 0, 1, 2, 3, \dots, L$. Let U_m be the minimal ε_m-covering of T, so $|U_m| \le N(\varepsilon_m; T_\rho)$. Then define the projection operation $\pi_m(\theta) = \arg\min_{\beta\in U_m} \rho(\theta, \beta)$.

This allows us to bound

$$\begin{aligned}|X_\theta - X_{\widetilde{\theta}}| \le |X_\theta - X_{\pi_2(\theta)}| + |X_{\pi_2(\theta)} - X_{\pi_1(\theta)}| + |X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}| \\ + |X_{\pi_1(\widetilde{\theta})} - X_{\pi_2(\widetilde{\theta})}| + |X_{\pi_2(\widetilde{\theta})} - X_{\widetilde{\theta}}|.\end{aligned}$$

Then we can take the expectation of $\sup_{\theta,\widetilde{\theta}}$ on both sides. What is the purpose of having all these interpolation points? The first and the last terms have infinitely many choices, so these are the discretization terms, while the middle terms have only finitely many choices, so we can apply the maximal inequality.

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in T} |X_\theta - X_{\widetilde{\theta}}|\right] \le \mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in T} |X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}|\right] + 2\mathbb{E}\left[\sup_{\theta\in T} |X_{\pi_2(\theta)} - X_{\pi_1(\theta)}|\right] \\ + \cdots + 2\mathbb{E}\left[\sup_{\theta\in T} |X_{\pi_L(\theta)} - X_{\pi_{L-1}(\theta)}|\right] + 2\mathbb{E}\left[\sup_{\theta\in T} |X_\theta - X_{\pi_L(\theta)}|\right].\end{aligned}$$

These terms on the right correspond to $\varepsilon_0, \varepsilon_1, \dots, \varepsilon_{L-1}, \varepsilon_*$, respectively. This process will define a Riemann sum. For the remaining details, see the textbook. □

Example 6.23 (Gaussian Complexity)**.** *We want to bound the Gaussian complexity $\mathscr{G}(B_2(1)) = \mathbb{E}[\sup_{\theta\in B_2(1)}\langle W, \theta\rangle]$ using chaining. We get the bound*

$$\begin{aligned}\mathscr{G}(B_2(1)) &\le C\int_0^2 \sqrt{\underbrace{\log N(u; B_2(1), \|\cdot\|_2)}_{\le d\log(2/u+1)}}\, du \\ &\le C\int_0^2 \sqrt{d\log(2/u+1)}\, du \\ &= C\sqrt{d}\underbrace{\int_0^2 \sqrt{\log(2/u+1)}\, du}_{C'} \\ &\asymp \sqrt{d}.\end{aligned}$$

6.2.4 Examples of Rademacher Complexity Bounds

We begin with the proposition.

Proposition 6.24. *Let $\mathscr{R}_n(\mathscr{F}) := \mathbb{E}_{\varepsilon_i, X_i}\left[\sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]$. Then*

1.

$$\mathscr{R}_n(\mathscr{F}) \;\lesssim\; \frac{D_{\mathbb{P}}}{\sqrt{n}} \int_0^1 \sup_Q \sqrt{\log N(D_Q\, u;\, \mathscr{F},\, L^2(Q))}\, du,$$

2.

$$\mathscr{R}_n(\mathscr{F}) \;\lesssim\; \frac{D_\infty}{\sqrt{n}} \inf_\varepsilon \varepsilon \;+\; \frac{1}{\sqrt{n}} \int_\varepsilon^1 \sup_Q \sqrt{\log N(D_\infty\, u;\, \mathscr{F},\, L^\infty)}\, du,$$

where $D_{\mathbb{P}} = \sup_{f\in\mathscr{F}} \|f\|_{L^2(\mathbb{P})}$ *and* $D_\infty = \sup_{f\in\mathscr{F}} \|f\|_\infty$.

Example 6.25. *Let* $\mathscr{F} = \{f_\theta(x) = 1 - e^{-\theta x},\ x \in [0,1] : \theta \in [0,1]\}$ *be a parametric function class. Then taking the derivative gives us*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \;\le\; \sup_{\theta\in[\theta_1,\theta_2]} \underbrace{|xe^{-\theta x}|}_{\le x} |\theta_1 - \theta_2| \;\le\; |\theta_1 - \theta_2|.$$

The covering number for the unit interval with $|\cdot|$ *is bounded as*

$$N(\varepsilon;\, [0,1],\, |\cdot|) \;\le\; \frac{1}{2\varepsilon} + 1,$$

so we get a covering number bound for the parametric function class

$$N(\varepsilon;\, \mathscr{F},\, L^\infty) \;\le\; N(\varepsilon;\, [0,1],\, |\cdot|) \;\le\; \frac{1}{2\varepsilon} + 1.$$

Using the chaining bound with $D_\infty = \sup_{f\in\mathscr{F}} \|f\|_\infty = \sup_{f\in\mathscr{F}} \sup_{x\in[0,1]} |1 - e^{-\theta x}| \le 1$,

$$\begin{aligned}
\mathscr{R}_n(\mathscr{F}) &\le \frac{D_\infty}{\sqrt{n}} \int_0^1 \sqrt{\log N(uD_\infty;\, \mathscr{F},\, L^\infty)}\, du \;=\; \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log N(u;\, \mathscr{F},\, L^\infty)}\, du \\
&= \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log\big(\tfrac{1}{2u} + 1\big)}\, du \\
&\lesssim \frac{C}{\sqrt{n}}.
\end{aligned}$$

Example 6.26 (Lipschitz parameterization)**.** *Consider a function class* $\mathscr{F} = \{f_\theta : \mathscr{X} \to \mathbb{R} : \theta \in B_2^d(1)\}$ *with* $\|f_0(x)\|_\infty = c_0 = 0$. *If* $|f_{\theta_1}(x) - f_{\theta_2}(x)| \le L\|\theta_1 - \theta_2\|_2$, *then we can use the bound*

$$\log N(\varepsilon;\, \mathscr{F},\, L^\infty) \;\le\; \log N\big(\varepsilon;\, B_2^d(1),\, \|\cdot\|_2\big) \;\lesssim\; d\log\big(\tfrac{1}{\varepsilon}\big) \;\lesssim\; d\log\big(\tfrac{1}{\varepsilon} + 1\big)$$

to get

$$\mathscr{R}_n(\mathscr{F}) \;\lesssim\; L\,\frac{D_\infty}{\sqrt{n}} \int_0^1 \sqrt{\log N(\varepsilon;\, \mathscr{F},\, L^\infty)}\, d\varepsilon,$$

where $D_\infty = \sup_\theta \|f_\theta\|_\infty \le c_0 + L = L$. *Continuing,*

$$\begin{aligned}
\mathscr{R}_n(\mathscr{F}) &\lesssim \frac{L}{\sqrt{n}} \int_0^1 \sqrt{d\log(1/u)}\, du \\
&\lesssim L\sqrt{\frac{d}{n}}.
\end{aligned}$$

Example 6.27 (Rademacher complexity of Lipschitz functions on $[0,1]^d$)**.** *For* $L > 0$ *and* $d \in \mathbb{Z}_{\ge 1}$, *consider the function class*

$$\mathscr{F}_L^d = \big\{g : [0,1]^d \to \mathbb{R} \;:\; g(\mathbf{0}) = 0,\ |g(\mathbf{x}) - g(\mathbf{y})| \le L\|\mathbf{x} - \mathbf{y}\|_\infty\big\}.$$

With $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathbb{P}$ *and* $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{Unif}(\{\pm 1\})$*, we wish to bound the Rademacher complexity*

$$\mathscr{R}_n(\mathscr{F}_L^d) \;=\; \mathbb{E}_{\mathbf{x}_i,\varepsilon_i}\left[\sup_{f\in\mathscr{F}_L^d}\left|\tfrac{1}{n}\sum_{i=1}^n \varepsilon_i f(\mathbf{x}_i)\right|\right]$$

independently of $\mathbb{P}$*, using two methods.*

The two ingredients common to both arguments are the metric entropy bound

$$\log N\big(\varepsilon;\, \mathscr{F}_L^d,\, \|\cdot\|_\infty\big) \;\lesssim\; \big(\tfrac{L}{\varepsilon}\big)^d,$$

and the fact that, conditionally on $(\mathbf{x}_i)$*, the quantity* $\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(\mathbf{x}_i)$ *is sub-Gaussian with parameter* $L/\sqrt{n}$*.*

Method 1: one-step discretization. Let $\mathscr{N}$ *be an* ε*-cover of* $\mathscr{F}_L^d$*. Splitting the supremum across the cover and its residual,*

$$\mathscr{R}_n(\mathscr{F}_L^d) \;\le\; \varepsilon + \sqrt{\tfrac{2\sigma^2}{n}\,\log|\mathscr{N}|} \;\le\; \varepsilon + \sqrt{\tfrac{2L^2}{n}\,\big(\tfrac{L}{\varepsilon}\big)^d} \;\asymp\; \varepsilon + \frac{L^{1+d/2}}{\sqrt{n}}\,\varepsilon^{-d/2}.$$

Choosing $\varepsilon = L\,n^{-1/(d+2)}$ *to balance the two terms yields*

$$\mathscr{R}_n(\mathscr{F}_L^d) \;\lesssim\; L\,n^{-\frac{1}{d+2}}.$$

Method 2: Dudley's entropy integral. The chaining bound gives

$$\mathscr{R}_n(\mathscr{F}_L^d) \;\lesssim\; \varepsilon + \frac{1}{\sqrt{n}}\int_\varepsilon^{2L}\big(\tfrac{L}{u}\big)^{d/2}du \;=\; \varepsilon + \frac{L}{\sqrt{n}}\int_{\varepsilon/L}^{2} u^{-d/2}\,du,$$

whose evaluation depends on d*:*

d	$\displaystyle\int_{\varepsilon/L}^{2} u^{-d/2}\,du$	*optimal* ε	*rate*
1	$\le 2\sqrt{2}$	0	$\mathscr{R}_n \lesssim L/\sqrt{n}$
2	$\log 2 + \log L - \log\varepsilon$	$L/\sqrt{n}$	$\mathscr{R}_n \lesssim \frac{L}{\sqrt{n}}(1+\log n)$
≥ 3	$\le \frac{2}{d-2}(L/\varepsilon)^{d/2-1}$	$L\,n^{-1/d}$	$\mathscr{R}_n \lesssim L\,n^{-1/d}$

Example 6.28 (Boolean Function Classes)**.** *Consider a Boolean function class* $\mathscr{F} \subseteq \{f : \mathscr{X} \to \{0,1\}\}$*, VC theory tells us that* $\mathscr{F}$ *has PD(*ν*), where* $\nu = \mathrm{VC}(\mathscr{F})$*. Using the maximal inequality, we have the bound*

$$\mathscr{R}_n(\mathscr{F}) \lesssim \sqrt{\frac{\nu\log(n+1)}{n}}.$$

We have mentioned that the log factor in this bound makes the bound not tight.

Proposition 6.29 (Covering Number for Boolean Function Class)**.** *For a boolean function class with* $\nu = \mathrm{VC}(\mathscr{F})$*,*

$$\sup_{\mathbb{P}} \log\big(N(\varepsilon;\mathscr{F},L^2(\mathbb{P}))\big) \lesssim \nu\log\left(\frac{e}{\varepsilon}\right)$$

for $\varepsilon < 1$*.*

For a sharp but difficult proof of this bound, see theorem 2.6.4 from [Van der Vaart and Wellner, 1996]. A weaker but easier version of this bound can be found in the notes [Sen, Theorem 7.9].

If we use the chaining argument, we get the bound

$$\mathscr{F}_n(\mathscr{F}) \lesssim \frac{1}{\sqrt{n}}\int_0^1 \sqrt{\nu\log(e/\varepsilon)}\,d\varepsilon \propto \sqrt{\frac{\nu}{n}}.$$

6.3 Glivenko-Cantelli Theorem and Donsker Theorem

We now generalize the preceding discussion to arbitrary function classes.

6.3.1 Glivenko-Cantelli Theorem

Recall that a class $\mathscr{F}$ is Glivenko-Cantelli when

$$\|\mathbb{P}_n - P\|_{\mathscr{F}} \xrightarrow{a.s.} 0,$$

where

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i), \quad \|\mathbb{P}_n - P\|_{\mathscr{F}} = \sup_f \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X).$$

Theorem 6.30 (Glivenko–Cantelli)**.** *Let $\mathscr{F}$ be a P-measurable class of measurable functions with envelope F such that $PF < \infty$. Let $\mathscr{F}_M$ be the class of functions $f1\{F \le M\}$ when f ranges over $\mathscr{F}$. Then*

$$\|\mathbb{P}_n - P\|_{\mathscr{F}} \to 0$$

almost surely if and only if

$$n^{-1}\log N(\varepsilon, \mathscr{F}_M, L_1(\mathbb{P}_n)) \to 0$$

in outer probability, for every $\varepsilon > 0$ and $M > 0$. In that case both convergences are also in outer mean. In particular, the class $\mathscr{F}$ is Glivenko–Cantelli if

$$\log N(\varepsilon, \mathscr{F}, L_1(\mathbb{P}_n)) = o_P(n)$$

for every $\varepsilon > 0$.

Proof. First, we prove $\|P_n - P\|_{\mathscr{F}} \to 0$ in mean. Using the symmetrization lemma

$$\begin{aligned}\mathbb{E}\|P_n - P\|_{\mathscr{F}} &\le 2\mathbb{E}_X\mathbb{E}_\epsilon \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathscr{F}} \\ &\le 2\mathbb{E}_X\mathbb{E}_\epsilon \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathscr{F}_M} + 2PF\{F > M\}.\end{aligned}$$

Use $\epsilon-$net $\mathscr{G}$ to cover $\mathscr{F}_M$.

$$\mathbb{E}_\epsilon \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathscr{F}_M} \le \mathbb{E}_\epsilon \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathscr{G}} + \varepsilon.$$

The cardinality of $\mathscr{G}$ can be chosen equal to $N(\epsilon, \mathscr{F}_M, L_1(\mathbb{P}_n))$. Use $\psi_2(x) = e^{x^2} - 1$, and use the maximal inequality, we have

$$\begin{aligned}\mathbb{E}_\epsilon \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathscr{G}} + \varepsilon &\le \sqrt{1 + \log N(\varepsilon, \mathscr{F}_M, L_1(\mathbb{P}_n))}\ \sup_{f\in\mathscr{G}} \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\psi_2|X} + \varepsilon \\ &\text{(Hoeffding)} \le \sqrt{1 + \log N(\varepsilon, \mathscr{F}_M, L_1(\mathbb{P}_n))}\sqrt{\frac{6}{n}}M + \varepsilon \xrightarrow{P} \epsilon.\end{aligned}$$

$\mathbb{E}\|P_n - P\|_{\mathscr{F}}$ converges almost surely to 0 by the following lemma.

Lemma 6.31. *Let $\mathscr{F}$ be a class of measurable functions with envelope F such that $PF < \infty$. Define a filtration by letting $\mathscr{S}_n$ be the σ-field generated by all measurable functions $h : \mathscr{X}^\infty \to \mathbb{R}$ that are permutation-symmetric in their first n arguments. Then*

$$\mathbb{E}(\|\mathbb{P}_n - P\|_{\mathscr{F}} \mid \mathscr{S}_{n+1}) \geq \|\mathbb{P}_{n+1} - P\|_{\mathscr{F}}, \quad a.s.$$

Furthermore, there exist versions of the measurable cover functions $\|\mathbb{P}_n - P\|_{\mathscr{F}}$ that are adapted to the filtration. Any such versions form a reverse submartingale and converge almost surely to a constant.

Another direction follows from another side of symmetrization lemma and Sudakov's inequality.

$$\frac{1}{2}\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\big(f(X_i) - Pf\big)\right\|_{\mathscr{F}} \leq \mathbb{E}\|\mathbb{P}_n - P\|_{\mathscr{F}} \to 0.$$

$$\frac{1}{\sqrt{n}}\sup_{\varepsilon>0}\varepsilon\sqrt{\log N(\varepsilon, \mathscr{F}, L_2(\mathbb{P}_n))} \leq 3\mathbb{E}_\xi\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i f(X_i)\right\|_{\mathscr{F}}.$$

And we get the results. □

Remark 6.32. *The covering number $N(\epsilon, \mathscr{F}, L_1(\mathbb{P}_n))$ is similar to the VC dimension. If covering number$= 2^n$, $\mathscr{F}$ is not GC class. In some cases, $L_2(\mathbb{P}_n)$ is difficult to calculate and we can use L^∞ to replace $L_2(\mathbb{P}_n)$.*

6.3.2 Donsker Theorem

Recall that a class $\mathscr{F}$ is P-Donsker when $\mathbb{G}_n \xrightarrow{d} \mathbb{G}$. Here we define F as the envelope of $\mathscr{F}$.

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - Pf).$$

Theorem 6.33 (Donsker)**.** *Let $\mathscr{F}$ be a class of measurable functions that satisfies the uniform entropy bound*

$$\int_0^\infty \sup_Q \sqrt{\log N\big(\varepsilon\|F\|_{Q,2}, \mathscr{F}, L_2(Q)\big)}\,d\varepsilon < \infty.$$

Let the classes

$$\mathscr{F}_\delta = \{f - g : f, g \in \mathscr{F},\ \|f - g\|_{P,2} < \delta\}$$

and $\mathscr{F}_\infty^2$ be P-measurable for every $\delta > 0$. If $P^*F^2 < \infty$, then $\mathscr{F}$ is P-Donsker.

6.4 Information Theory

To prove lower bounds, we turn to information theory, which lets us reduce estimation to hypothesis testing and gives us two main tools for the job: Le Cam's method and Fano's inequality.

6.4.1 Fundamentals

Entropy: We begin with a central concept in information theory: the entropy. Let P be a distribution on a finite (or countable) set $\mathscr{X}$, and let p denote the probability mass function associated with P. That is, if X is a random variable distributed according to P, then $P(X = x) = p(x)$. The *entropy of X* (or of P) is defined as

$$H(X) := -\sum_x p(x)\log p(x). \tag{21}$$

Because $p(x) \le 1$ for all x, it is clear that this quantity is positive. We will show later that if $\mathscr{X}$ is finite, the maximum entropy distribution on $\mathscr{X}$ is the uniform distribution, setting $p(x) = 1/|\mathscr{X}|$ for all x, which has entropy $\log(|\mathscr{X}|)$.

Mutual information: The *mutual information* $I(X;Y)$ between X and Y is the KL-divergence between their joint distribution and their products (marginal) distributions. More mathematically,

$$I(X;Y) := \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{22}$$

We can rewrite this in several ways. First, using Bayes' rule, we have $p(x,y)/p(y) = p(x \mid y)$, so

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(y)p(x \mid y) \log \frac{p(x \mid y)}{p(x)} \\ &= -\sum_x \sum_y p(y)p(x \mid y) \log p(x) + \sum_y p(y) \sum_x p(x \mid y) \log p(x \mid y) \\ &= H(X) - H(X \mid Y). \end{aligned}$$

Similarly, we have $I(X;Y) = H(Y) - H(Y \mid X)$, so mutual information can be thought of as the amount of entropy removed (on average) in X by observing Y. We may also think of mutual information as measuring the similarity between the joint distribution of X and Y and their distribution when they are treated as independent

$$I(X;Y) = D_{\mathrm{KL}}\left(P_{XY} \,\|\, P_X \times P_Y\right) \ge 0.$$

Moreover, we have $I(X;Y) > 0$ unless X and Y are independent.

Entropies of continuous random variables For continuous random variables, we may define an analogue of the entropy known as *differential entropy*, which for a random variable X with density p is defined by

$$h(X) := -\int p(x) \log p(x)\, dx. \tag{23}$$

Proposition 6.34. *Let X be a random vector on $\mathbb{R}^n$ with a density, and assume that $\mathrm{Cov}(X) = \Sigma$. Then for $Z \sim \mathsf{N}(0, \Sigma)$, we have*

$$h(X) \le h(Z).$$

Proof. Without loss of generality, we assume that X has mean 0. Let P be the distribution of X with density p, and let Q be multivariate normal with mean 0 and covariance Σ; let Z be this random variable. Then

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx &= -h(X) + \int p(x) \left[\frac{n}{2} \log(2\pi) - \frac{1}{2} x^\top \Sigma^{-1} x\right] dx \\ &= -h(X) + h(Z), \end{aligned}$$

because Z has the same covariance as X. As $0 \le D_{\mathrm{KL}}(P\|Q)$, we have $h(Z) \ge h(X)$ as desired. □

6.4.2 Data Processing Inequalities

A standard problem in information theory (and statistical inference) is to understand the degradation of a signal after it is passed through some noisy channel (or observation process). The simplest of

such results, which we will use frequently, is that we can only lose information by adding noise. In particular, assume we have the Markov chain

$$X \to Y \to Z.$$

Then we obtain the classical *data processing inequality.*

Proposition 6.35. *With the above Markov chain, we have $I(X;Z) \le I(X;Y)$.*

Proof. We expand the mutual information $I(X;Y,Z)$ in two ways:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y \mid Z) \\ &= I(X;Y) + \underbrace{I(X;Z \mid Y)}_{=0}, \end{aligned}$$

where we note that the final equality follows because X is independent of Z given Y:

$$I(X;Z \mid Y) = H(X \mid Y) - H(X \mid Y,Z) = H(X \mid Y) - H(X \mid Y) = 0.$$

Since $I(X;Y \mid Z) \ge 0$, this gives the result. □

There are related data processing inequalities for the KL-divergence—which we generalize in the next section—as well. In this case, we may consider a simple Markov chain $X \to Z$. If we let P_1 and P_2 be distributions on X and Q_1 and Q_2 be the induced distributions on Z, that is, $Q_i(A) = \int \mathbb{P}(Z \in A \mid x)\, dP_i(x)$, then we have

$$D_{\mathrm{KL}}(Q_1 \| Q_2) \le D_{\mathrm{KL}}(P_1 \| P_2),$$

the basic KL-divergence data processing inequality. A consequence of this is that, for any function f and random variables X and Y on the same space, we have

$$D_{\mathrm{KL}}(f(X) \| f(Y)) \le D_{\mathrm{KL}}(X \| Y).$$

6.4.3 Minimax Lower Bound

In statistical decision theory, we have a class of distributions $\mathscr{P}$ and a parameter/function of distributions $\theta : \mathscr{P} \to \Theta$. If this is a one to one mapping, we write $\mathscr{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Then we have **statistical estimators**, which are mappings $\widehat{\theta} : \mathscr{X} \to \Theta$. Suppose there is a **semimetric** $\rho(\theta, \theta') : \Theta \times \Theta \to \mathbb{R}$, such as

$$\rho(\theta, \theta') = \|\theta - \theta'\|_2, \qquad \rho(f, f') = \|f - f'\|_{L^2}.$$

If $\Phi : [0,\infty) \to [0,\infty)$ is increasing, the **risk** is

$$R(\widehat{\theta}; \theta(P)) = \mathbb{E}_{X \sim P}[\Phi(\rho(\widehat{\theta}(X); \theta(P)))].$$

In this framework, the **loss function** is $\ell = \Phi \circ \rho$.

Definition 6.36. *The **minimax risk** with n samples is*

$$\mathscr{M}_n(\theta(\mathscr{P}), \Phi \circ \rho) = \inf_{\widehat{\theta} : \mathscr{X} \to \Theta} \sup_{P \in \mathscr{P}} R(\widehat{\theta}; \theta(P)).$$

The inf and the sup mean that we are taking the best estimator for the worst model.

(a) If $R(\widehat{\theta})$ achieves $\mathscr{M}_n$, it is good enough.

(b) If $R(\widehat{\theta}) \gg \mathscr{M}_n$, we should either find a better estimator or a sharper lower bound.

The idea is to find a testing problem easier than the estimation problem. A lower bound for the testing problem will imply a lower bound for estimation.

Step 1: Construct a 2δ-separated set of Θ in the ρ-metric. So we require $\rho(\theta^i, \theta^j) \geq 2\delta$ for all $i \neq j$. This is the same as a packing, except we allow $\geq$ instead of $>$. If our separated set is $\{\theta^1, \theta^2, \dots, \theta^M\}$, we get $\{\mathbb{P}_{\theta^1}, \mathbb{P}_{\theta^2}, \dots, \mathbb{P}_{\theta^M}\}$.

Step 2: Sample $(J, Z) \in [M] \times \mathscr{X}$. The joint distribution is

$$\begin{cases} J \sim \text{Unif}(\{1, 2, \dots, M\}) \\ Z \mid J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

Step 3: Let $\mathbb{Q}$ be the joint distribution of (J, Z). Then the marginal distribution of Z is

$$\overline{\mathbb{Q}} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}.$$

Our testing problem is that we want to find a $\psi : \mathscr{X} \to [M]$ such that $\mathbb{Q}(\psi(Z) \neq J)$ is small. If $M = 2$, this is standard binary hypothesis testing. The testing error is

$$\mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2}\big[\underbrace{\mathbb{P}_{\theta^1}(\psi(Z) \neq 1)}_{\text{Type I error}} + \underbrace{\mathbb{P}_{\theta^2}(\psi(Z) \neq 2)}_{\text{Type II error}}\big].$$

This is different from the traditional hypothesis testing setup in that instead of fixing the Type I error and minimizing the Type II error, we want to minimize the average of these errors.

Proposition 6.37 (From estimation to testing)**.** *Let Φ be increasing and $\{\theta^1, \dots, \theta^M\}$ be 2δ-separated for $\delta > 0$. Then*

$$\mathscr{M}_n(\theta(\mathscr{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}(\psi(Z) \neq J).$$

This works for all $\delta > 0$, so we can pick the δ which gives the best lower bound. In general, $\Phi(\delta)$ is increasing with δ, but the testing error $\inf_\psi \mathbb{Q}(\psi(Z) \neq J)$ is decreasing with δ. We can choose $\delta = \delta_n$ such that $\inf_\psi \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2}$; any constant would work here. Then the minimax lower bound will be

$$\mathscr{M}_n \geq \frac{1}{2}\Phi(\delta_n).$$

Proof. Fix P and $\widehat{\theta}$. By Markov's inequality,

$$\begin{aligned} \mathbb{E}[\Phi(\rho(\widehat{\theta}, \theta))] &\geq \Phi(\delta)\mathbb{P}(\Phi(\rho(\widehat{\theta}, \theta)) \geq \Phi(\delta)) \\ &= \Phi(\delta)\mathbb{P}(\rho(\widehat{\theta}, \theta) \geq \delta). \end{aligned}$$

We now want to relate this probability with the testing error. We have

$$\begin{aligned} \sup_{P \in \mathscr{P}} \mathbb{P}(\rho(\widehat{\theta}, \theta) \geq \delta) &\geq \sup_{\theta \in \{\theta^1, \dots, \theta^M\}} \mathbb{P}_\theta(\rho(\widehat{\theta}, \theta) \geq \delta) \\ &\geq \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}(\rho(\widehat{\theta}, \theta^j) \geq \delta) \\ &= \mathbb{Q}(\rho(\widehat{\theta}, \theta^J) \geq \delta). \end{aligned}$$

Define a test ψ via $\widehat{\theta}$: Let

$$\psi(z) = \arg\min_{L \in [M]} \rho(\widehat{\theta}(Z), \theta^L).$$

This gives the θ^j which is the closest to our estimate $\widehat{\theta}(Z)$. With this definition,

$$\{\psi(Z) \neq J\} \subseteq \{\rho(\widehat{\theta}(Z), \theta^J) \geq \delta\}.$$

This means we can lower bound the above $\mathbb{Q}$ probability:

$$\inf_{\widehat{\theta}} \mathbb{Q}(\rho(\widehat{\theta}(Z), \theta^J) \geq \delta) \geq \inf_{\psi} \mathbb{Q}(\psi(Z) \neq J).$$

□

Remark 6.38. *How do we choose $\{\theta^1, \ldots, \theta^M\}$? Moreover, how do we lower bound $\inf_\psi \mathbb{Q}(\psi(Z) \neq J)$? Here are two general methods.*

1. *$M = 2$: Le Cam's Method*
2. *$M \geq 3$: Fano's Method*

Some Divergence Measures Recall the definition of the total variation distance, the KL divergence, and the Hellinger distance

$$\begin{aligned}
\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} &= \frac{1}{2} \int_{\mathscr{X}} |p(x) - q(x)| \, dx, \\
D(\mathbb{P}\|\mathbb{Q}) &= \int_{\mathscr{X}} p(x) \log \frac{p(x)}{q(x)} \, dx, \\
\mathbb{H}^2(\mathbb{P}\|\mathbb{Q}) &= \int_{\mathscr{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx.
\end{aligned}$$

These have the following relationships:

$$\begin{aligned}
\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} &\leq \sqrt{\tfrac{1}{2} D(\mathbb{P}\|\mathbb{Q})}, \\
\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} &\leq \sqrt{\mathbb{H}^2(\mathbb{P}\|\mathbb{Q})} \underbrace{\sqrt{1 - \frac{\mathbb{H}^2(\mathbb{P}\|\mathbb{Q})}{4}}}_{\leq 1}, \\
\mathbb{H}^2(\mathbb{P}\|\mathbb{Q}) &\leq \frac{1}{2} D(\mathbb{P}\|\mathbb{Q}).
\end{aligned}$$

Le Cam's Two Points Method Take $M = 2$. Then $J \sim \mathrm{Unif}(\{0, 1\})$, and $Z \mid J = j \sim \mathbb{P}_j$, and $\overline{\mathbb{Q}} = \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$. We claim that

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2} \left(1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}}\right).$$

Proof. For any ψ, we can find an A such that

$$\psi(x) = \begin{cases} 1 & x \in A \\ 0 & x \in A^c. \end{cases}$$

Then

$$\begin{aligned}
\mathbb{Q}(\psi(Z) = J) &= \frac{1}{2}\mathbb{P}_1(A) + \frac{1}{2}\mathbb{P}_0(A^c) \\
&= \frac{1}{2}(\mathbb{P}_1(A) - \mathbb{P}_0(A)) + \frac{1}{2}.
\end{aligned}$$

If we take the supremum over all ψ, we get

$$\begin{aligned}\sup_{\psi} \mathbb{Q}(\psi(Z) = J) &= \sup_{A} \frac{1}{2}(\mathbb{P}_1(A) - \mathbb{P}_0(A)) + \frac{1}{2} \\ &= \frac{1}{2}\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}} + \frac{1}{2}.\end{aligned}$$

The probability of the bad event is then

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2} - \frac{1}{2}\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}.$$

□

This gives the following theorem.

Theorem 6.39 (Le Cam's Two Points Lower Bound)**.** *For all $\delta > 0$ and $\mathbb{P}_0, \mathbb{P}_1 \in \mathscr{P}$ with $\rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta$,*

$$\mathscr{M}_n(\theta(\mathscr{P}), \Phi \circ \rho) \geq \frac{\Phi(\delta)}{2}\left(1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}\right).$$

Example 6.40 (Gaussian Location Family, $d = 1$)**.** *Our model is $\mathscr{P} = \{\mathbb{P}_\theta = \mathsf{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$, where σ is known. We have the semimetric $\rho(\theta', \theta) = |\theta' - \theta|$ and $\Phi(t) = t^2$. Our sample is $X_{1:n} \sim \mathbb{P}_\theta^n$. The true minimax risk is $\mathscr{M}_n = \frac{\sigma^2}{n}$. Here is a lower bound by Le Cam's method:*

Consider $\mathbb{P}_{2\delta}$ and $\mathbb{P}_0$, so $\rho(2\delta, 0) \geq 2\delta$. Then

$$\mathscr{M}_n(\theta(\mathscr{P}); |\theta - \theta'|^2) \geq \frac{\delta^2}{2}\left(1 - \|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\mathrm{TV}}\right),$$

where the n only appears in the bound as the fact that the measures are product measures. We want to lower bound $1 - \|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\mathrm{TV}}$ by $1/2$. We have by Pinsker's inequality and the tensorization property of K-L divergence

$$\begin{aligned}\|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\mathrm{TV}}^2 &\leq \frac{1}{2} D(\mathbb{P}_{2\delta}^n \| \mathbb{P}_0^n) \\ &= \frac{1}{2} n D(\mathbb{P}_{2\delta} \| \mathbb{P}_0) \\ &= \frac{1}{2} n \cdot \frac{(2\delta)^2}{2\sigma^2} \\ &= \frac{n\delta^2}{\sigma^2}.\end{aligned}$$

Now choose $\frac{n\delta^2}{\sigma^2} = \frac{1}{2}$, so $\delta_n^2 = \frac{\sigma^2}{2n}$. Then $\|\mathbb{P}_{2\delta_n}^n - \mathbb{P}_0^n\|_{\mathrm{TV}} \leq \frac{1}{2}$, and we get the minimax lower bound

$$\mathscr{M}_n \geq \frac{\delta_n^2}{2} \cdot \frac{1}{2} = \frac{\sigma^2}{16n}.$$

Up to constants, this is sharp.

Here is the problem with Le Cam's method. If we take $\theta \in \mathbb{R}^d$ with $\mathbb{P}_\theta = \mathsf{N}(0, \sigma^2 I_d)$ for $d \geq 2$, then we will get the lower bound

$$\mathscr{M}_n \geq \frac{\sigma^2}{16n},$$

even though the actual minimax risk is $\mathscr{M}_n = \sigma^2 \frac{d}{n}$.

Let

$$\mathbb{Q} : \begin{cases} J \sim \text{Unif}(\{1, 2, \ldots, M\}) \\ Z \mid J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

Lemma 6.41.

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) \geq 1 - \frac{I(Z; J) + \log 2}{\log M}.$$

The proof is in Section 15.4 and requires some ideas such as the entropy. This does not require any restriction on the $\mathbb{P}_{\theta^j}$. This lower bound gives us

Proposition 6.42. *Let $\{\theta^1, \ldots, \theta^M\}$ be 2δ-separated in the semimetric ρ. Then*

$$\mathscr{M}_n(\theta(\mathscr{P}); \Phi \circ \rho) \geq \Phi(\delta) \left(1 - \frac{I(Z; J) + \log 2}{\log M}\right).$$

When using this lower bound, we will find δ_n such that

$$1 - \frac{I(Z; J) + \log 2}{\log M} \geq \frac{1}{2}.$$

Then we will get

$$\mathscr{M}_n \geq \frac{1}{2} \Phi(\delta_n).$$

So we need to upper bound $I(Z; J)$.

A simple upper bound is given by

$$\begin{aligned} I(J; Z) &= \frac{1}{M} \sum_{j=1}^{M} D\left(\mathbb{P}_{\theta^j} \,\middle\|\, \frac{1}{M} \sum_{\ell=1}^{M} \mathbb{P}_{\theta^\ell}\right) \\ &\leq \frac{1}{M^2} \sum_{j,\ell=1}^{M} D(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^\ell}) \\ &\leq \max_{j,\ell} D(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^\ell}). \end{aligned}$$

Where we have used Jensen's inequality to show that the K-L divergence is convex in the second argument.

Example 6.43 (Gaussian location family, $d \geq 2$)**.** *Our model is $\mathscr{P} = \{\mathbb{P}_\theta = \mathsf{N}(0, \sigma^2 I_d) : \theta \in \mathbb{R}^d\}$, where σ is known. Our semimetric is $\rho(\theta', \theta) = \|\theta' - \theta\|_2$ with $\Phi(t) = t^2$. The true minimax risk is*

$$\mathscr{M}_n = \inf_{\widehat{\theta}} \sup_{\theta} \mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] = \sigma^2 \frac{d}{n}.$$

The lower bound by Fano's method gives

$$\begin{aligned} \mathscr{M}_n &\geq \Phi(\delta) \left(1 - \frac{I(Z; J) + \log 2}{\log M}\right) \\ &\geq \Phi(\delta) \left(1 - \frac{\max_{j,k} D(\mathbb{P}_{\theta^j}^n \| \mathbb{P}_{\theta^k}^n) + \log 2}{\log M}\right). \end{aligned}$$

Our goal is to find the largest $\delta_n, M, \{\theta^1, \ldots, \theta^M\}$ such that

(a) $\|\theta^j - \theta^k\|_2 \geq 2\delta_n$

(b)

$$\frac{\max_{j,k} D(\mathbb{P}_{\theta^j}^n \| \mathbb{P}_{\theta^k}^n) + \log 2}{\log M} \leq \frac{1}{2}.$$

Here is our construction: Let $\varepsilon_n = \sigma\sqrt{\frac{d}{n}}$ *and* $\delta_n = \frac{1}{100}\varepsilon_n = \frac{1}{100}\sigma\sqrt{\frac{d}{n}}$. *Let* $\{\theta^1, \ldots, \theta^M\}$ *be a maximal* $2\delta_n$ *packing of* $B(0, \varepsilon_n) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq \varepsilon_n\}$.

By a volume argument, we can get upper and lower bounds of M*:*

$$\log M \asymp d \log\left(c\frac{\varepsilon_n}{\delta_n}\right) \asymp c \cdot d.$$

To upper bound the K-L divergence on top, we have

$$\begin{aligned}
\max_{j,k} D(\mathbb{P}_{\theta^j}^n \| \mathbb{P}_{\theta^k}^n) &= n \max_{j,k} D(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}) \\
&= n \max_{j,k} \frac{\|\theta^j - \theta^k\|_2^2}{2\sigma^2} \\
&\leq \frac{n\varepsilon_n^2}{2\sigma^2} \\
&= c \cdot d.
\end{aligned}$$

Our quantities only depend on the ratio between ε_n *and* δ_n*, so we can adjust the constant in front of* δ_n *to get the desired upper bound of* $\frac{1}{2}$.

We then get

$$\mathscr{M}_n \geq \Phi(\delta_n)\frac{1}{2} = \frac{1}{2} \cdot \left(\frac{1}{100}\right)^2 \sigma^2 \frac{d}{n} = c\sigma^2 \frac{d}{n}.$$

The bound on $I(J; Z)$ by the max of the K-L divergences is generally only good when we have a parametric problem. For nonparametric problems, we want to use a better bound.

Lemma 6.44 (Yang-Barron's Bound)**.** *Let* $N_{\mathrm{KL}}(\varepsilon; \mathscr{P})$ *be an* ε*-cover of* $\mathscr{P}$ *in* $\sqrt{D_{\mathrm{KL}}}$. *Then*

$$I(Z; J) \leq \inf_{\varepsilon > 0} \varepsilon^2 + \log N_{\mathrm{KL}}(\varepsilon; \mathscr{P})$$

To apply this bound, we have two steps:

1. Choose $\varepsilon_n > 0$ such that
$$\varepsilon_n^2 \geq \log N_{\mathrm{KL}}(\varepsilon_n; \mathscr{P}).$$
2. Choose the largest $\delta_n > 0$ such that
$$\log M(\delta_n; \rho, \Omega) \geq 4\varepsilon_n^2 + 2\log 2.$$

7 High-Dimensional Statistics

7.1 Concentration of Sample Covariance

7.2 Sparse Linear Regression

7.3 High-Dimensional Principal Component Analysis

7.4 Low-Rank Matrix Recovery

7.5 Non-Parametric Estimation

8 Ramdom Matrix Theory

In addition to non-asymptotic results, we will need asymptotic analysis, which is more delicate. The section is organized as follows:

1. Density of eigenvalues in classical ensembles of random matrices
2. Semi-Circle Law and Marchenko–Pastur Law
3. BBP Transition
4. CLT for Eigenvalues
5. Spectrum Separation
6. Replica Method

This section mainly follows STATC206B (UC Berkeley, taught by Vadim Gorin) and STAT260 (UC Berkeley, taught by Song Mei, 2021). I also refered to the book [1].

8.1 Density of Eigenvalues in Classical Ensembles of Random Matrices

We begin our tour from an important type of matrix GβE ($\beta = 1, 2, 4$). Let X be $N \times N$ matrix with i.i.d. entries:

$$\begin{cases} a)\mathscr{N}(0,1) \\ b)\mathscr{N}(0,1) + i\mathscr{N}(0,1) \\ c)\mathscr{N}(0,1) + i\mathscr{N}(0,1) + j\mathscr{N}(0,1) + k\mathscr{N}(0,1) \end{cases}$$

Set $M = \frac{1}{2}(X + X^*)$, and we have the main theorem.

Theorem 8.1 (Density of Eigenvalues)**.** *Let the eigenvalues of M be $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$, and they have density*

$$\frac{1}{Z} \prod_{i<j} |\lambda_j - \lambda_i|^\beta \cdot \prod_{i=1}^{N} \exp(-\frac{\lambda_i^2}{2}) \, \mathrm{d}\lambda_1 ... \, \mathrm{d}\lambda_N.$$

where $\beta = 1, 2, 4$ for a), b), c) and

$$Z = \frac{(2\pi)^{N/2}}{N!} \prod_{j=0}^{N} \frac{\Gamma(1 + (j+1)\beta/2)}{\Gamma(1 + \beta/2)}$$

is the partition function (normalization factor).

Proof. We prove $\beta = 1$ here and leave $\beta = 2$ to the homework.

Step 1 Claim density of $M = \frac{1}{2}(X + X^*) \sim \exp(-\frac{1}{2}\operatorname{tr}(M^2))$.

Indeed,

$$\operatorname{tr}(M^2) = \sum_{i,j} m_{ij}^2 = \sum_{i=1}^{N} \underbrace{x_{ii}^2}_{\mathscr{N}(0,1)} + \frac{1}{2} \sum_{i<j} \underbrace{(x_{ij} + x_{ji})^2}_{\mathscr{N}(0,2)}$$

implies that the density of $M \propto \exp(-\frac{1}{2}\operatorname{tr}(M^2))$

Step 2 We can calculate that $\exp(-\frac{1}{2}\operatorname{tr}(M^2)) = \prod_{i=1}^{N} \exp(-\frac{\lambda_i^2}{2})$.

We derive it immediately by diagonalizing the the matrix M.

Step 3 Each symmetric matrix is determined by its eigenvalues and eigenvectors, i.e. there exists an almost bijection π

$$\pi : \underbrace{\mathscr{W}_N}_{\lambda_1<\lambda_2<...<\lambda_N} \times \underbrace{O(N)}_{\text{Orthogonal Bases}} \to \underbrace{\mathscr{H}_N}_{\text{Symmetric Matrix}}$$

We say "almost" because indeed the map is not injective: we can multiply the eigenvalue by ± 1. In other words, if the eigenvalues are unique, $\pi^{-1}(M)$ has exactly 2^N elements.

Now we give the key proposition of the proof.

Proposition 8.2 (Jacobian)**.** *Consider the map $\pi : (\Lambda, O) \mapsto B$, where $\pi((\Lambda, O)) = O\Lambda O^*$. Then the Jocobian of the map is $\prod_{i<j} |\lambda_j - \lambda_i|$.*

Remark 8.3. *It is an important technique to transform an intractable density calculation into a tractable one and a Jocobian.*

Proof. It is sufficient to calculate by taking $O = \mathrm{Id}$, as when we transform O to $A \cdot O$ (where A is orthogonal), the **uniform measure** on $O(N)$ and the Lebesgue measure on $\mathscr{H}(N)$ remain unchanged.

Then we take $O = \exp(B) = \mathrm{Id} + B + ...$ and we can deduce that $B + B^* = 0$ as $OO^* = \mathrm{Id}$. Then the map π can be written as

$$\begin{aligned}
((\lambda_1, ..., \lambda_N), \exp(B)) &\mapsto \exp(B) \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix} \exp(-B) \\
&\mapsto (\mathrm{Id} + B) \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix} (Id - B) + o(B) \\
&\mapsto \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix} \\
&+ \begin{pmatrix} 0 & b_{12}(\lambda_2 - \lambda_1) & b_{13}(\lambda_3 - \lambda_1) & \cdots & b_{1n}(\lambda_n - \lambda_1) \\ b_{21}(\lambda_1 - \lambda_2) & 0 & b_{23}(\lambda_3 - \lambda_2) & \cdots & b_{2n}(\lambda_n - \lambda_2) \\ b_{31}(\lambda_1 - \lambda_3) & b_{32}(\lambda_2 - \lambda_3) & 0 & \cdots & b_{3n}(\lambda_n - \lambda_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{n1}(\lambda_1 - \lambda_n) & b_{n2}(\lambda_2 - \lambda_n) & b_{n3}(\lambda_3 - \lambda_n) & \cdots & 0 \end{pmatrix} + o(B)
\end{aligned}$$

As B has only $\frac{n(n-1)}{2}$ free parameter, so the Jacobian of the map is $\prod_{i<j} |\lambda_j - \lambda_i|$. □

Step 4 We now calculate Z:

$$Z = \int_{\mathbb{R}^N} \prod_{i<j} |\lambda_j - \lambda_i|^\beta \cdot \prod_{i=1}^N \exp(-\frac{\lambda_i^2}{2}) \, \mathrm{d}\lambda_1 ... \, \mathrm{d}\lambda_N = \frac{(2\pi)^{N/2}}{N!} \prod_{j=0}^N \frac{\Gamma(1 + (j+1)\beta/2)}{\Gamma(1 + \beta/2)}$$

□

8.2 Semi-Circle Law and Marchenko–Pastur Law

8.2.1 Trace Calculation

Theorem 8.4 (Tridiagonal Matrix Equivalent)**.** *Consider the real symmetric tridiagonal matrix*

$$T_\beta = \begin{pmatrix} \mathcal{N}(0,1) & \frac{1}{\sqrt{2}}\chi_{\beta(n-1)} & \cdots & 0 \\ \frac{1}{\sqrt{2}}\chi_{\beta(n-1)} & \mathcal{N}(0,1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{N}(0,1) \end{pmatrix}$$

where the second diagonal entries are $\chi_{\beta(n-1)}, \chi_{\beta(n-2)}, \cdots \chi_\beta \cdot$ $\chi_k = \sqrt{\chi_k^2} = \sqrt{\sum_{i=1}^k \mathcal{N}(0,1)^2}$*, whose density is*

$$\frac{1}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})} x^{k-1} e^{-\frac{x^2}{2}}.$$

Then the eigenvalues of T_β *have the same distribution of* $G\beta E$*.*

Proof. We present for $\beta = 1$. By linear algebra, we can choose an orthogonal matrix U_1, and transform the matrix M to $U_1 M U_1^\top$, whose the first row is

$$(x_{11}, \sqrt{\sum_{k=2}^n x_{ik}^2}, 0, ..., 0).$$

Note that the eigenvalues do not change after the transformation.

We can inductively do the same operation on the rest sub-matrix. □

Corollary 8.5. *For* $\beta = 1, 2, 4$*, the law of eigenvalues* $\sim \prod_{i<j}(x_i - x_j)^\beta \prod_{i=1}^N \exp(-\frac{x_i^2}{2})$.

Theorem 8.6. *In fact, the corollary is true for any* $\beta > 0$*.*

Now we begin to prove the semi-circle law. First we give the key result of trace calculation.

Theorem 8.7 (Semi-Circle Law)**.** *For each* $k > 0, \beta > 0$*, we have*

$$\frac{1}{N^{\frac{k}{2}+1}} \operatorname{tr}(T_\beta^k) = \frac{1}{N^{\frac{k}{2}+1}} \sum_{i=1}^N X_i^k \xrightarrow{N\to\infty, \textit{in probability}} \begin{cases} 0 & , \quad k \textit{ is odd;} \\ (\frac{\beta}{2})^k \operatorname{Cat}_{\frac{k}{2}} & , \quad k \textit{ is even.} \end{cases}$$

where Cat_k *is the k-the catalan number defined as*

$$\operatorname{Cat}_k = \frac{1}{k+1}\binom{2k}{k}.$$

Proof. By SLLN, we have

$$\lim_{N\to\infty} \frac{\chi_{\beta(N-1)}}{\sqrt{2N}} = \sqrt{\frac{\beta}{2}},$$
$$\lim_{N\to\infty} \frac{\chi_{\beta(N-\alpha N)}}{\sqrt{2N}} = \sqrt{\frac{\beta(1-\alpha)}{2}}.$$

So by the definition of the trace, we have

$$\begin{aligned}\operatorname{tr}\left(\frac{T_\beta}{\sqrt{N}}\right)^k &= \sum_{m=1}^{N} \sum_{\substack{\text{k-step paths}\\ \text{from } m \text{ to } m}} \left(\text{ some } \frac{\chi_{\beta(N-i)}}{\sqrt{2N}}\right)\left(\text{ some } \frac{\mathscr{N}(0,1)}{\sqrt{N}}\right) \\ &= \sum_{m=1}^{N} (\#\text{k-step paths from } m \text{ to } m)\sqrt{\frac{\beta}{2}\left(\frac{N-m}{N}\right)^k} + o(N).\end{aligned}$$

The second equation follows from that if there exists "some $\frac{\mathscr{N}(0,1)}{\sqrt{N}}$", the term vanishes. When $2 \nmid k$, #k-step paths from m to m is zero. When $2 \mid k$, #k-step paths from m to $m = \binom{k}{\frac{k}{2}}$. Approximate the summation by integration and we have

$$\frac{1}{N}\operatorname{tr}\left(\frac{T_\beta}{\sqrt{N}}\right)^k \to \binom{k}{\frac{k}{2}}\left(\frac{\beta}{2}\right)^{\frac{k}{2}} \int_0^1 x^{\frac{k}{2}}\,\mathrm{d}x = \operatorname{Cat}_{\frac{k}{2}}\left(\frac{\beta}{2}\right)^{\frac{k}{2}}$$

□

Now we introduce the semi-circle distribution.

Definition 8.8 (Wiegner Semi-Circle Distribution)**.** *The density of the distribution is*

$$\mu(x) = \frac{1}{2\pi}\sqrt{4-x^2}.$$

We use moment method to recover the proof and give the definition.

Corollary 8.9 (Moment Analysis)**.** *Let m_k are moments of T_β and we have*

$$m_u = \begin{cases} 0 & , \quad u \text{ is odd;} \\ \frac{1}{\frac{u}{2}+1}\binom{u}{\frac{u}{2}} & , \quad u \text{ is even.} \end{cases}$$

Theorem 8.10 (Moment of Semi-Circle Distribution)**.** *m_k are moments of semi-circle law and*

$$m_k = \int_{-2}^{2} \mu(x)x^k\,\mathrm{d}x.$$

Proof. (Method I) We just calculate

$$m_k = \int \frac{1}{\sqrt{2\pi}}\sqrt{4-x^2}x^k\,\mathrm{d}x.$$

Let $x = 2\cos\theta$, and by inductive calculation, we derive the results. □

However, we are not satisfied for the calculation is not so intuitive. We then introduce another proof.

Proof. (Method II) Introduce generating function:

$$G(z) = \sum_{k=0}^{\infty} m_k z^{-k-1},$$

$m_0 = 1$. By classical results, we have

$$\sum_{n=0}^{\infty} \operatorname{Cat}_n x^n = \frac{1-\sqrt{1-4x}}{2x} =: C(x).$$

Now we can derive the expression of $G(z)$ using $C(x)$.

$$G(z) = \sum_{k=0}^{\infty} m_k z^{-k-1} = \sum_{l=0}^{\infty} \mathrm{Cat}_l \, z^{-2l-1} = \frac{z - \sqrt{z^2-4}}{2}.$$

Next we introduce two propositions.

Proposition 8.11 (Stieltjes Transform)**.**

$$G(z) = \int \frac{\mu(x)}{z-x} \, \mathrm{d}x.$$

Proof. By Taylor expansion, we have

$$\begin{aligned}
\int \frac{\mu(x)}{z-x} \, \mathrm{d}x &= \frac{1}{z} \int \frac{\mu(x)}{1 - \frac{x}{z}} \, \mathrm{d}x \\
&= \frac{1}{z} \int \mu(x) \sum_{k=0}^{\infty} \left(\frac{x}{z}\right)^k \\
&= \sum_{k=0}^{\infty} m_k z^{-k-1} \\
&= G(z)
\end{aligned}$$

□

Proposition 8.12.

$$\mu(x_0) = -\frac{1}{\pi} \lim_{y_0 \to 0^+} \Im(G(x_0 + iy_0)).$$

Remark 8.13. *We can just add a perturbation onto the imagine axis and obtain the information of the point.*

Proof. By proposition 8.11, we have

$$-\frac{1}{\pi}\Im(G(x_0+iy_0)) = -\frac{1}{\pi} \int \Im \frac{1}{x_0 + iy_0 - x} \mu(x) \, \mathrm{d}x \tag{24}$$

$$= -\frac{1}{\pi} \int \Im \frac{x_0 - x - iy_0}{(x-x_0)^2 + y_0^2} \mu(x) \, \mathrm{d}x \tag{25}$$

$$= \int \frac{1}{\pi} \frac{y_0}{(x-x_0)^2 + y_0^2} \mu(x) \, \mathrm{d}x. \tag{26}$$

Notice that $\frac{1}{\pi} \frac{y_0}{(x-x_0)^2+y_0^2}$ is a "good" kernel, and thus as $y_0 \to 0$

$$\int \frac{1}{\pi} \frac{y_0}{(x-x_0)^2 + y_0^2} \mu(x) \, \mathrm{d}x \to \mu(x_0).$$

□

Return to the Theorem 8.10, by proposition 8.12

$$\mu(x_0) = -\frac{1}{\pi} \lim_{y \to 0^+} \Im \left(\frac{1}{2} \left[(x+iy) - \sqrt{(x+iy)^2 - 4} \right] \right) = \begin{cases} 0 & , \quad |x| > 2; \\ \frac{\sqrt{4-x^2}}{2\pi} & , \quad |x| \le 2. \end{cases}$$

□

Now we are prepared to formally state and prove the semi-circle law.

Theorem 8.14 (Semi-Circle Law)**.** *Let $\lambda_1 < ... < \lambda_N$ be eigenvalues of $G\beta E$. Set $x_i = \lambda_i \sqrt{\frac{2}{\beta N}}$ and let μ_N be their empirical measure: $\mu_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \delta_{x_i}$. Then*

$$\lim_{N\to\infty} \mu_N = \frac{1}{2\pi}\sqrt{4-x^2}\mathbb{1}_{|x|\le 2} = \mu_{circle}(x)$$

weakly in probability, which means $\forall$ bounded continuous function $f(x)$

$$\lim_{N\to\infty} \int_{\mathbb{R}} f(x)\mu_N(x) = \int_{\mathbb{R}} f(x)\mu(x)\,\mathrm{d}x.$$

Proof. We prove the result in four steps.

Step 1 The results hold for $f(x) = x^k$.

Step 2 The results hold for any polynomials.

Step 3 Take $L > 2$, the results hold for $f(x) = \mathbb{1}_{|x|>L} x^k$. We have

$$\begin{aligned}
\left|\int f(x)\mu_N(\mathrm{d}x)\right| &= \left|\int x^k \mu_N(\mathrm{d}x)\right| \\
&\le L^{-2m} \int_{|x|\ge L} x^{2k+2m}\mu_N(\mathrm{d}x) \\
&\to L^{-2m} \int x^{2k+2m}\mu(\mathrm{d}x) \\
&\le L^{-2m} 2^{2k+2m} = 2^{2k}\left(\frac{2}{L}\right)^m \to 0.
\end{aligned}$$

Step 4 By step 3, we can restrict the support set of f on a compact set, i.e. $[-4, 4]$. Apply the Weierstrass theorem and we get the proof. □

At last, we give three generalizations of the semi-circle law. Next theorem tells us the gaussian assumption of the semi-circle law is not neccessary.

Theorem 8.15 (Generalization)**.** *Let $z_{ij}, i < j$ be i.i.d. random variables with finite moments. $\mathbb{E}z_{ij} = 0$ and $\mathbb{E}z_{ij}^2 = \frac{1}{2}$. Let Y_i be i.i.d. with finite moments. Then the semi-circle law still holds for the matrix with entries Y_i and z_{ij}.*

8.2.2 Marchenko–Pastur Law

Now we consider the case where the matrix is not square. Let X be an $N \times S$ random matrix whose elements are Gaussian with parameter $\beta > 0$ (different form according to β). Assume that $N, S \to \infty$ in such a way that $N/S \to \gamma^2 \in (0, 1)$. Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ denote the eigenvalues of $\frac{1}{S}XX^*$, and define the empirical spectral measure

$$\rho^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i}.$$

Theorem 8.16 (Marchenko–Pastur Law). *As $N \to \infty$ (with $N/S \to \gamma^2 \in (0,1)$), the empirical spectral measure ρ^N converges weakly, in probability, to the Marchenko–Pastur distribution μ_{MP} with density*

$$\frac{\mathrm{d}\mu_{\mathrm{MP}}}{\mathrm{d}x}(x) = \frac{1}{2\pi\beta\gamma^2}\,\frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{x}\,\mathbf{1}_{[\lambda_-,\,\lambda_+]}(x),$$

where

$$\lambda_+ = \beta(1+\gamma)^2, \qquad \lambda_- = \beta(1-\gamma)^2.$$

Proof. We calculate the moments for ρ^N and μ_{MP} separately

Step 1 We claim that

$$\lim_{N,S\to\infty} \frac{1}{NS^k}\operatorname{tr}[(XX^*)^k] = \beta^k \sum_{r=0}^{k-1} \frac{\gamma^{2r}}{r+1}\binom{k}{r}\binom{k-1}{r}.$$

Notice that as $S \to \infty$

$$\frac{\chi_{aS}}{\sqrt{S}} \to \sqrt{a},$$

we have

$$\begin{aligned}
\mathrm{LHS} &= \lim_{S,N\to\infty} \frac{1}{NS^k} \sum_{m=1}^{N} \sum_{\substack{\text{k-step paths}\\ \text{from } m \text{ to } m}} (\text{diagonal entries})(\text{sub-diagonal entries}) \\
&= \lim_{S,N\to\infty} \frac{1}{NS^k} \sum_{m=1}^{N} \sum_{i=0}^{[k/2]} \left(\chi_{\beta(N-m)}\chi_{\beta(S-m+1)}\right)^{2i} \chi_{S+N-2m-2}^{2(k-2i)} \binom{k}{2i}\binom{2i}{i}
\end{aligned}$$

By replacing χ by its limit, we have

$$\frac{\chi_{\beta(N-m)}\chi_{\beta(S-m+1)}}{S} \to \beta\sqrt{(\gamma^2 - \frac{m}{S})(1-\frac{m}{S})}$$

$$\frac{\chi^2_{\beta(S+N-2m+2)}}{S} \to \beta(1+\gamma^2 - 2\frac{m}{S})$$

Then we approximate the sum by integral

$$\begin{aligned}
\mathrm{LHS} &= \frac{\beta^k}{\gamma^2} \sum_{i=0}^{[k/2]} \binom{k}{2i}\binom{2i}{i} \int_0^{\gamma^2} (\gamma^2 - t)^i (1-t)^i (1+\gamma^2-2t)^{k-2i}\,\mathrm{d}t \\
&= \frac{\beta^k}{\gamma^2} \sum_{i=0}^{[k/2]} \binom{k}{2i}\binom{2i}{i} \int_0^{\gamma^2} u^i (1-\gamma^2+u)^i (1-\gamma^2+2u)^{k-2i}\,\mathrm{d}u
\end{aligned}$$

We claim that

$$\frac{1}{\gamma^2} \sum_{i=0}^{[k/2]} \binom{k}{2i}\binom{2i}{i} \int_0^{\gamma^2} u^i (1-\gamma^2+u)^i (1-\gamma^2+2u)^{k-2i}\,\mathrm{d}u = \sum_{r=0}^{k-1} \frac{\gamma^{2r}}{r+1}\binom{k}{r}\binom{k-1}{r}.$$

and we obtain the proof.

The proof of the claim: Denote the left hand side as A. First, we observe that

$$\mathrm{A} = \text{The sum of coefficients of all the } x^i y^i \text{ in } \int_0^{\gamma^2} [xu + y(1-\gamma^2+u) + (1-\gamma^2+2u)]^k\,\mathrm{d}u \quad (27)$$

Denote B as the integral of (27)

$$B = \frac{1}{\gamma^2}\frac{[(x+1)\gamma^2+(y+1)]^{k+1}-[(y+1)(1-\gamma^2)]^{k+1}}{(x+y+2)(k+1)}$$
$$= \frac{1}{k+1}\sum_{l=0}^{k}[(y+1)(1-\gamma^2)]^l[(x+1)\gamma^2+y+1]^{k-l}.$$

To calculate the coefficients, we can let $x = \frac{1}{y}$ and calculate the constant coefficient.

$$\mathrm{B} = \frac{1}{k+1}\sum_{l=0}^{k}[(y+1)(1-\gamma^2)]^l\left(\left(\frac{1}{y}+1\right)\gamma^2+y+1\right)^{k-l}$$
$$= \frac{(y+1)^k}{k+1}\sum_{l=0}^{k}(1-\gamma^2)^l(\frac{1}{y}\gamma^2+1)^{k-l}$$
$$= \frac{(y+1)^k}{k+1}\frac{(\frac{1}{y}\gamma^2+1)^{k+1}-(1-\gamma^2)^{k+1}}{\frac{1}{y}\gamma^2+\gamma^2}$$
$$= \frac{y(y+1)^{k-1}}{(k+1)\gamma^2}\left(\left(\frac{1}{y}\gamma^2+1\right)^{k+1}-(1-\gamma^2)^{k+1}\right)$$

Then the constant coefficient is

$$\frac{1}{(k+1)\gamma^2}\sum_{r=0}^{k-1}\binom{k-1}{r}\binom{k+1}{r+1}\gamma^{2r+2} = \sum_{r=0}^{k-1}\frac{\gamma^{2r}}{r+1}\binom{k}{r}\binom{k-1}{r}.$$

Step 2 We claim that $\lambda_+ = \beta(1+\gamma)^2$ and $\lambda_- = \beta(1-\gamma)^2$.

Define the density of the limit distribution

$$\mu(x) = \frac{1}{2\pi\gamma^2}\frac{\sqrt{(\lambda_+-x)(x-\lambda_-)}}{x}\mathbb{1}_{[\lambda_-,\lambda_+]}$$

and $\mu_N = \rho^N$.

The calculation below is inspired by Zhidong Bai's book. First, we calculate the moment of $\mu(x)$,

for $m \geq 1$,

$$\begin{aligned}
\mathbb{E}_\mu x^m &= \int x^m \mu(x)\,\mathrm{d}x \\
&= \frac{1}{2\pi\gamma^2}\int x^{m-1}\sqrt{(\lambda_+ - x)(x-\lambda_-)}\,\mathrm{d}x \\
&= \frac{1}{2\pi\gamma^2}(2\gamma)^2\int_{-1}^{1}(1+\gamma^2+2\gamma u)^{m-1}\sqrt{1-u^2}\,\mathrm{d}u \\
&\text{(by setting } x = \beta(1+\gamma^2+2\gamma u)) \\
&= \beta^m \frac{1}{\pi}\sum_{l=0}^{[(m-1)/2]}\binom{m-1}{2l}(1+\gamma^2)^{m-1-2l}(4\gamma^2)^l\int_{-1}^{1}u^{2l}\sqrt{1-u^2}\,\mathrm{d}u \\
&= \beta^m \sum_{l=0}^{[(m-1)/2]}\binom{m-1}{2l}\frac{1}{l+1}\binom{2l}{l}(1+\gamma^2)^{m-1-2l}\gamma^{2l} \\
&= \beta^m \sum_{l=0}^{[(m-1)/2]}\sum_{s=0}^{m-1-l}\frac{(m-1)!}{l!(l+1)!s!(m-1-2l-s)!}\gamma^{2(l+s)} \\
&= \beta^m \frac{1}{m}\sum_{r=0}^{m-1}\gamma^{2r}\binom{m}{r}\sum_{l=0}^{r\wedge(m-1-r)}\binom{r}{l}\binom{m-r}{m-r-l-1} \\
&= \beta^m \frac{1}{m}\sum_{r=0}^{m-1}\binom{m}{r}\binom{m}{r+1}\gamma^{2r} \\
&= \beta^m \sum_{r=0}^{m-1}\frac{1}{r+1}\binom{m}{r}\binom{m-1}{r}\gamma^{2r}.
\end{aligned}$$

Next, we notice that

$$\mathbb{E}_{\mu_N} x^m = \frac{1}{N}\sum_{i=1}^{N}\lambda_i^N = \frac{1}{N}\operatorname{tr}\left[\left(\frac{XX^*}{S}\right)^m\right]$$

and

$$\lim_{N\to\infty}\frac{1}{N}\operatorname{tr}\left[\left(\frac{XX^*}{S}\right)^m\right] = \beta^m\sum_{r=0}^{m-1}\frac{1}{r+1}\binom{m}{r}\binom{m-1}{r}\gamma^{2r}.$$

So we have

$$\lim_{N\to\infty}\mathbb{E}_{\mu_N}x^m = \mathbb{E}_\mu x^m.$$

Similar to the proof of semi-circle law, we have ρ^N converges (weakly, in probability) to the Marchenko-Pastur law. □

8.3 BBP Transition

In the spiked model $\Sigma = I + \theta\, vv^*$, Baik, Ben Arous, and Péché (2005) discovered a phase transition for $\lambda_{\max}$ of Σ. In general form, it is stated as

Theorem 8.17 (Spiked Random Matrices)**.** *Let $\lambda_1 < \lambda_2 < \cdots < \lambda_N$ are N eigenvalues of*

$$C = \sqrt{N}avv^* + \sqrt{\frac{2}{\beta}}G\beta E,$$

where $\|v\| = 1$ Then there exists a_{crit} such that

1. *If* $a > a_{crit} = 1$, *then*

$$\lim_{N\to\infty} \frac{\lambda_N}{\sqrt{N}} = a + \frac{1}{a} > 2.$$

2. *If* $a < a_{crit}$, *then*

$$\lim_{N\to\infty} \frac{\lambda_N}{\sqrt{N}} = 2.$$

By the theorem, we can easily get

Corollary 8.18.

$$\hat{a} = \frac{1}{2}\left(\frac{\lambda_N}{\sqrt{N}} + \sqrt{\frac{\lambda_N^2}{N} - 4}\right) \overset{N\to\infty}{\longrightarrow} a.$$

Theorem 8.19 (Recover v)**.** *In the setting of Theorem 8.17. Let* $\hat{v}$ *denote the eigenvector corresponding to* λ_N *and let* ϕ *be the angle between* v *and* $\hat{v}$*,*

$$\lim_{N\to\infty} \sin\phi = \frac{1}{a} \wedge 1.$$

Proof. We prove the results in three steps.

Step 1 a, λ_N and ϕ are unchanged under orthogonal/unitary transformations of matrix C. We do two transformations:

- Rotate v ti be the first basis vector $(1, 0^{N-1})$.
- Rotate in the orthogonal complement of $(1, 0^{N-1})$, so that the $(N-1)\times(N-1)$ bottom random corner of GβE becomes diagonal, while the law of the first row and column of GβE preserved.

Step 2 Now we brought C into

$$\hat{C} = \begin{pmatrix} a\delta_N + \mathscr{N}(0, \frac{2}{\beta}) & \xi_2 & \xi_3 & \cdots & \xi_N \\ \bar{\xi}_2 & \mu_2 & & & \\ \bar{\xi}_3 & & \mu_3 & & \\ \vdots & & & \ddots & \\ \bar{\xi}_N & & & & \mu_N \end{pmatrix}$$

and $\mu_2, \cdots, \mu_N$ be the eigenvalues of $\sqrt{\frac{2}{\beta}} G\beta E$ and $\xi_2, \cdots, \xi_N \overset{i.i.d}{\sim} \sqrt{\frac{2}{\beta}}$ corresponding Normal distribution. We find an eigenvector $(x_1, \cdots, x_N)$ of $\tilde{C}$ with eigenvalues λ

$$\begin{cases} x_1\left(a\sqrt{N} + \mathscr{N}\left(0, \frac{2}{\beta}\right)\right) + \sum_{i=1}^N \xi_i x_i = \lambda x_1 \\ x_1\bar{\xi}_2 + \mu_2 x_2 = \lambda x_2 \Rightarrow x_2 = \frac{x_1\bar{\xi}_2}{\lambda - \mu_2} \\ \vdots \\ x_1\bar{\xi}_N + \mu_N x_N = \lambda x_N \Rightarrow x_N = \frac{x_1\bar{\xi}_N}{\lambda - \mu_N} \end{cases} \Longrightarrow \begin{cases} x_2 = \frac{\bar{\xi}_2}{\lambda - \mu_2} \\ \vdots \\ x_N = \frac{\bar{\xi}_N}{\lambda - \mu_N} \end{cases}$$

Plug the $2 \sim N$ equations into the first line and we get

$$x_1\left[a\sqrt{N} + \mathscr{N}\left(0, \frac{2}{\beta}\right) - \lambda + \sum_{i=1}^N \frac{\xi_i\bar{\xi}_i}{\lambda - \mu_i}\right] = 0. \qquad (*)$$

By interpolating of eigenvalues, C has 1 eigenvalue larger than μ_N. Only this eigenvalue has chance to become larger than $2\sqrt{N}$ as $\frac{\mu_N}{\sqrt{N}} \to 2$. Denote $y = \frac{\lambda_N}{\sqrt{N}}$ and investigate $(*)$ for $y > 2$.

$$a + \frac{\mathcal{N}\left(0, \frac{2}{\beta}\right)}{\sqrt{N}} - y + \frac{1}{N}\sum_{i=2}^{N} \frac{\xi_i \bar{\xi}_i}{y - \frac{\mu_i}{\sqrt{N}}} = 0. \tag{$**$}$$

By LLN, we have

$$\frac{1}{N}\sum_{i=2}^{N} \frac{\xi_i \bar{\xi}_i}{y - \frac{\mu_i}{\sqrt{N}}} \approx \frac{1}{N}\sum_{i=2}^{N} \frac{1}{y - \frac{\mu_i}{\sqrt{N}}} \to G(y) = \frac{1}{2}(y - \sqrt{y^2 - 4}).$$

Then we have

$$a - y + \frac{1}{2}(y - \sqrt{y^2 - 4}) = 0 \Longrightarrow y = a + \frac{1}{a},$$

and this proves Theorem 8.17.

Step 3 For Theorem 8.19, we notice that the eigenvalues are

$$(1, 0, \cdots, 0) \text{ and } \left(1, \frac{\bar{\xi}_2}{\lambda - \mu_2}, \cdots, \frac{\bar{\xi}_N}{\lambda - \mu_N}\right),$$

and we have

$$\cos^2 \phi = \frac{1}{1 + \sum_{i=2}^{N} \frac{\xi_i \bar{\xi}_i}{(\lambda - \mu_i)^2}} \to \frac{1}{1 + \frac{1}{2}\frac{y}{\sqrt{y^2-4}} - \frac{1}{2}} = \frac{2\sqrt{y^2 - 4}}{y + \sqrt{y^2 - 4}}.$$

As $y = a + \frac{1}{a}$, we have

$$\sin^2 \phi \to \frac{1}{a^2}.$$

□

8.4 CLT for Eigenvalues

Similar to the CLT for random variables, we have the following theorem.

Theorem 8.20. *Let f be analytic in a small neighboring of $[-2, 2]$. Let $\lambda_1 \leq \cdots \leq \lambda_N$ be e.v. of $\sqrt{\frac{2}{\beta}} G\beta E, \beta > 0$. Then*

$$\sum_{i=1}^{N} f\left(\frac{\lambda_i}{\sqrt{N}}\right) - N \int f(x) \frac{1}{2\pi} \sqrt{4 - x^2}$$

converges to dist to a Gaussian r.v. ξ_f jointly other several $f = f_1, \ldots, f_K$ with

$$\mathbb{E}[\xi_f] = \left(\frac{2}{\beta} - 1\right) m(f), \quad \operatorname{Cov}(\xi_f, \xi_g) = \frac{2}{\beta} C(f, g)$$

$$m(f) = \frac{1}{4}(f(2) + f(-2)) - \frac{1}{2\pi}\int_{-2}^{2} \frac{f(x)}{\sqrt{4 - x^2}}\, \mathrm{d}x$$

$$C(f, g) = \frac{1}{4\pi}\int_{-2}^{2}\int_{-2}^{2} \frac{(f(x) - f(y))(g(x) - g(y))}{(x - y)^2} \frac{4 - xy}{\sqrt{4 - x^2}\sqrt{4 - y^2}}\, \mathrm{d}x\, \mathrm{d}y.$$

or

$$C\left(\frac{1}{z - x}, \frac{1}{w - x}\right) = -\frac{1}{2(z - w)^2}\left(1 - \frac{zw - 4}{\sqrt{z^2 - 4}\sqrt{w^2 - 4}}\right) \quad \text{for } z, w \text{ out of } [-2, 2]$$

where $m(f)$ and $C(f, g)$ do not depend on β.

Proof. The main idea is fancy moments method.

Lemma 8.21 (Wick's formula)**.**

$$\mathbb{E}[\eta_1, \dots, \eta_m] = \sum_{\text{perfectmatchings}} \prod_{(i,j)\in \text{mathching}} \sigma_{ij}$$

Proof. By Laplacian's transform

$$\mathbb{E}[e^{t_1\eta_1+\cdots+t_m\eta_m}] = \exp\left(\frac{1}{2}\sum_{i,j=1}^{m} t_i t_j \sigma_{ij}\right).$$

Hence, we take derivative and get the expectation

$$\mathbb{E}[\eta_1 \cdots \eta_m] = \frac{\partial^m}{\partial t_1 \cdots \partial t_m}\left[\exp\left(\frac{1}{2}\sum_{i,j=1}^{m} t_i t_j \sigma_{ij}\right)\right]\Big|_{t_1=\cdots=t_m=0}.$$

Equivalently, this is the coefficient of $t_1 \cdots t_m$ in Taylor's expansion.

$$= \sum_{ij} 2^{-\frac{m}{2}} \prod_{ij} \sigma_{ij}.$$

Notice that each perfect match appears $2^{\frac{m}{2}}$ times, we prove the Lemma 8.21. □

Takeaway Moments can be reconstructed by applying a differential operator to Laplace transform.

Lemma 8.22. *Introduce an operator*

$$\mathscr{D}_a = \prod_{i,j}(z_i - z_j)^{-1}\left(\sum_{i=1}^{N} T_{a,i}\right)\prod_{i,j}(z_i - z_j)$$

$$T_{a,i} f(z_1, \dots, z_N) = f(z_1, \dots, z_{i-1}, z_i + a, z_{i+1}, \dots, z_N)$$

Then for $\lambda_1, \cdots, \lambda_N$ eigenvalues of GUE, we have

$$\underbrace{\mathbb{E}\prod_{k=1}^{M}\left[\sum_{i=1}^{N} e^{a_k\lambda_i}\right]}_{\text{Moments of linear statistic for } f(\lambda)=e^{a\lambda}} = \underbrace{\mathscr{D}_{a_m}\cdots\mathscr{D}_{a_1}}_{\text{They all commute}} \exp\left(\sum_{i=1}^{N}\frac{z_i^2}{2}\right)\Big|_{z_1=\cdots=z_N=0}.$$

Proof. From Lecture 6,

$$\mathbb{E}\exp(\mathrm{Tr}(\mathrm{GUE}\cdot \mathrm{Z})) = \mathbb{E}B_{\lambda_1,\dots,\lambda_N}(z_1, \dots, z_N) = \exp\left(\sum_{i=1}^{N}\frac{\lambda_i^2}{2}\right)$$

where $Z = \mathrm{diag}(z_1, \dots, z_N)$ and $\lambda_1, \dots, \lambda_N$ denote the e.v. of GUE.

Then act with $\mathscr{D}$

$$B_{\lambda_1,\dots,\lambda_N}(z_1, \dots, z_N) = \prod_{K=1}^{N}(K!)\cdot\frac{\det[\exp(\lambda_i z_j)]}{\prod_{i,j}(\lambda_i - \lambda_j)(z_i - z_j)}.$$

$$\mathscr{D} = \prod_{i,j}(z_i - z_j)^{-1}\left(\sum_{i=1}^{N} T_{a,i}\right)\prod_{i,j}(z_i - z_j).$$

So B is eigenfunction of $\mathscr{D}$ with eigenvalue $\sum_{i=1}^N \exp(a\lambda_i)$. Hence,

$$\mathbb{E}\left[\prod_{k=1}^m \left(\sum_{i=1}^N e^{a_k\lambda_i}\right)\right] B_{\lambda_1\cdots\lambda_N}(z_1,\cdots,z_N) = \mathscr{D}_{a_m}\cdots\mathscr{D}_{a_1}\exp\left(\sum_{i=1}^N \frac{\lambda_i^2}{2}\right).$$

Plug $z_1,\ldots,z_N=0$ and notice $B(0,\cdots,0)=1$. □

Lemma 8.23.

$$\mathscr{D}f(z_1)\cdots f(z_N) = f(z_1)\cdots f(z_N)\frac{a^{-1}}{2\pi i}\oint_{\{z_1,\ldots,z_N\}}\left[\prod_{j=1}^N \frac{v+a-z_j}{v-z_j}\right]\frac{f(v+a)}{f(v)}$$

The contour encloses $z_1,\ldots,z_N$ with no singularities of f.

Proof.

$$\begin{aligned}\mathscr{D}f(z_1)\cdots f(z_N) &= \sum_{i=1}^N\left[\prod_{j\neq i}\frac{z_i+a-z_j}{z_i-z_j}\right]\frac{f(z_i+a)}{f(z_i)}\cdot f(z_1)\cdots f(z_N)\\ &= \frac{a^{-1}}{2\pi i}\oint_{\{z_1,\ldots,z_N\}}\prod_{i=1}^N\frac{v+a-z_j}{v-z_j}\cdot\frac{f(v+a)}{f(v)}\,\mathrm{d}v f(z_1)\cdots f(z_N)\end{aligned}$$

□

Corollary 8.24.

$$\begin{aligned}\mathbb{E}\left[\prod_{K=1}^m\sum_{i=1}^N e^{a_K\lambda_i}\right] =& \frac{(a_1\cdots a_m)^{-1}}{(2\pi i)^m}\oint\sum_{k=1}^m\left[\frac{(v_k+a_k)^N}{v_k^N}\exp\left(\frac{a_k^2}{2}+a_kv_k\right)\right]\\ &\prod_{k<l}\frac{v_k-v_l+a_k-a_l}{v_k-v_l-a_l}\cdot\frac{v_k-v_l}{v_k-v_l+a_k}\,\mathrm{d}v_1\cdots\mathrm{d}v_k.\end{aligned}$$

Proof. We compute $\mathscr{D}_{a_m}\cdots\mathscr{D}_{a_1}\exp(\frac{z_i^2}{2})\cdots\exp(\frac{z_N^2}{2})$. By sequentially applying Lemma 8.23 and then setting $z_1=\cdots=z_N=0$ at the end □

Now we prove Theorem 8.20 for $\beta=2$, $f(\frac{\lambda}{\sqrt{N}})=\exp(a\cdot\frac{\lambda}{\sqrt{N}})$

Step 1: Expectation By Corollary 8.24 with $m=1$

$$\mathbb{E}\left[\sum\exp\left(a\cdot\frac{\lambda_i}{\sqrt{N}}\right)\right] = \frac{(\frac{a}{\sqrt{n}})^{-1}}{2\pi i}\oint_{\{0\}}\left(\frac{v+\frac{a}{\sqrt{N}}}{v}\right)^N\exp\left(\frac{a^2}{2N}+\frac{a}{\sqrt{N}}v\right)\mathrm{d}v$$

No steepest descent needed. Set $v=u\sqrt{N}$ to get

$$\begin{aligned}&\frac{N}{a}\frac{1}{2\pi i}\oint\exp\left(N\log\left(1+\frac{a}{Nu}\right)+au+\frac{2a^2}{N}\right)\\ =&\frac{N}{a}\frac{1}{2\pi i}\oint\exp\left(a\left(u+\frac{1}{u}\right)+\frac{a^2}{2N}\left(1-\frac{1}{u^2}\right)+O(N^{-2})\right)\mathrm{d}u\\ =&\frac{N}{a}\frac{1}{2\pi i}\oint\exp\left(a\left(u+\frac{1}{u}\right)\right)\mathrm{d}u+\frac{a}{4\pi i}\oint\exp\left(a\left(u+\frac{1}{u}\right)\right)\left(1-\frac{1}{u^2}\right)\mathrm{d}u+O(N^{-1}).\end{aligned}$$

We only need to prove

$$\frac{N}{a}\frac{1}{2\pi i}\oint \exp\left(a\left(u+\frac{1}{u}\right)\right)\mathrm{d}u = N\int_{-2}^{2}\exp(ax)\frac{1}{2\pi}\sqrt{4-x^2}\,\mathrm{d}x.$$

We transform the contour to the unit circle and change variables $x = u + \frac{1}{u}$ ($u = \frac{1}{2}(x \pm i\sqrt{4-x^2}), \mathrm{d}u = \frac{1}{2}(1 \pm i\frac{-x}{\sqrt{4-x^2}})\,\mathrm{d}x$)

$$\begin{aligned}\frac{a^{-1}}{2\pi i}\oint \exp\left(a\left(u+\frac{1}{u}\right)\right)\mathrm{d}u &= \frac{a^{-1}}{2\pi i}\left(\int_{-2}^{2}e^{ax}\frac{1}{2}(1+i\frac{-x}{\sqrt{4-x^2}})\,\mathrm{d}x + \int_{-2}^{2}e^{ax}\frac{1}{2}(1-i\frac{-x}{\sqrt{4-x^2}})\,\mathrm{d}x\right)\\ &= \frac{a^{-1}}{2\pi}\int_{-2}^{2}e^{ax}\frac{x}{\sqrt{4-x^2}}\,\mathrm{d}x \overset{\text{by parts}}{=} \frac{a^{-1}}{2\pi}\int_{-2}^{2}(ae^{ax})\sqrt{4-x^2}\,\mathrm{d}x\end{aligned}$$

Conclusion:

$$\mathbb{E}\left[\sum_{i=1}^{N}\exp\left(a\frac{\lambda_i}{\sqrt{N}}\right)\right] = N\int_{-2}^{2}e^{ax}\frac{1}{2\pi}\sqrt{4-x^2}\,\mathrm{d}x + O\left(N^{-1}\right)$$

Step 2: Variance

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{N}\exp\left(a_1\frac{\lambda_i}{\sqrt{N}}\right)\sum_{i=1}^{N}\exp\left(a_2\frac{\lambda_i}{\sqrt{N}}\right)\right] - \mathbb{E}\left[\sum_{i=1}^{N}\exp\left(a_1\frac{\lambda_i}{\sqrt{N}}\right)\right]\mathbb{E}\left[\sum_{i=1}^{N}\exp\left(a_2\frac{\lambda_i}{\sqrt{N}}\right)\right]\\ =&N\frac{(a_1a_2)^{-1}}{(2\pi i)^2}\iint\prod_{k=1}^{2}\left(\frac{v_k+\frac{a_k}{\sqrt{N}}}{v_k}\right)^N\exp\left(\frac{a_k^2}{2N}+\frac{a_kv_k}{N}\right)\cdot\left[\frac{v_1-v_2+\frac{a_1}{\sqrt{N}}-\frac{a_2}{\sqrt{N}}}{(v_1-v_2-\frac{a_2}{\sqrt{N}})(v_1-v_2+\frac{a_1}{\sqrt{N}})}-1\right]\mathrm{d}v_1\,\mathrm{d}v_2.\\ \approx&N^2\frac{(a_1a_2)^{-1}}{2\pi i^2}\iint\exp\left(a_1\left(v_1+\frac{1}{v_1}\right)+a_2\left(v_2+\frac{1}{v_2}\right)\right)\left[\frac{1+\frac{a_1-a_2}{N(u_1-u_2)}}{(1-\frac{a_2}{N(u_1-u_2)})(1+\frac{a_1}{N(u_1-u_2)})}\right]\mathrm{d}u_1\,\mathrm{d}u_2.\\ =&N^2\frac{(a_1a_2)^{-1}}{(2\pi i)^2}\iint\exp\left(a_1\left(v_1+\frac{1}{v_1}\right)+a_2\left(v_2+\frac{1}{v_2}\right)\right)\left[\frac{a_1a_2}{N^2(u_1-u_2)^2}\right]\mathrm{d}u_1\,\mathrm{d}u_2.\end{aligned}$$

Conclusion: Variance $\to \frac{1}{(2\pi i)^2}\iint\exp\left(a_1\left(v_1+\frac{1}{v_1}\right)+a_2\left(v_2+\frac{1}{v_2}\right)\right)\frac{\mathrm{d}u_1\,\mathrm{d}u_2}{(u_1-u_2)^2}$.

Match this with formula in Theorem 1. Hint: Either directly with the 1st formula for $C(f,g)$, or

$$\sum_{i=1}^{N}f(\lambda_i) = \frac{1}{2\pi i}\oint_{\text{around all }\lambda_i}f(z)\sum_{i=1}^{N}\frac{1}{z-\lambda_i}\,\mathrm{d}z$$

and use it to compute with $C(\frac{1}{z-x},\frac{1}{w-x})$ in Theorems.

Step 3: Gaussianity We need to show that

$$\mathbb{E}\left[\prod_{k=1}^{m}\left(\sum_{i=1}^{N}\exp\left(a_k\frac{\lambda_i}{\sqrt{N}}\right)-\mathbb{E}\sum_{i=1}^{N}\exp\left(a_k\frac{\lambda_i}{\sqrt{N}}\right)\right)\right] \to \text{expressions of the Wick's formula in Lemma 8.21.}$$

All of them have exactly the same $\prod_{k=1}^{m}$ part, but cross term part $\prod_{k<l}$ varies. As in Steps 1,2, we change variables $u_k = \sqrt{N}v_k$ and use

$$\text{crossterm} = 1 + \frac{a_ka_l}{N^2(u_k-u_l)^2} + O\left(N^{-3}\right)$$

The summation over 2^m integrals leads to cancellation at parts involving 1. The next term $\to$ perfect matching. $\square$

8.5 Spectrum Separation

TODO

8.6 Replica Method

TODO

Part III

Chapter 3: Other Topics in Statistics

9 Computational Statistics

We now turn to several topics in computational mathematics that are closely related to statistics. This section is organized as follows:

1. Numerical Linear Algebra
2. Numerical Analysis
3. Optimization
4. Sampling Methods and Simulation
5. Optimal Transport

This section mainly follows ...

9.1 Numerical Linear Algebra

9.2 Numerical Analysis

9.3 Optimization

9.4 Sampling Methods and Simulation

9.5 Optimal Transport

10 Deep Learning Theory

10.1 Approximation Theory

10.1.1 Universal Approximation

10.1.2 Kernel Methods and Random Feature Models

10.1.3 Benefits of Depth

10.2 Optimization Theory

10.2.1 Neural Tangent Kernel

10.2.2 Margin Maximization and Implicit Bias

10.2.3 Edge of Stability

10.3 Classic Models for Learning Theory

10.3.1 Linear Models

10.3.2 Statistical Query

References

[1] Zhidong Bai, Jack W Silverstein, et al. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[2] David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. CRC Press, 1979.

[3] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[4] Gregory F Lawler. *Introduction to stochastic processes*. Chapman and Hall/CRC, 2018.

[5] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[6] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.