High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 0: Short Introduction

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/02/19)*

Keywords: concentration inequalities, expectation of suprema, uniform law of large numbers, random matrices, minimax lower bounds.

This lecture provides a short introduction on what we are interested in this course and why we are interested in them. We begin with the arguably simplest example.

Example 0.1 *Given n i.i.d random variables $X_1, \cdots, X_n$ with mean $\mu = \mathbb{E}[X_k]$, maybe the most common approach to infer μ is to use the sample mean $\frac{1}{n}\sum_{k=1}^n X_k$ as an estimator. Then a natural question arises: how well is the estimator? This question can be answer in different ways. For example,*

- *By the law of large numbers, it is known that $\frac{1}{n}\sum_{k=1}^n X_k$ converges to μ almost surely.*
- *Suppose the variance of the random variable is σ^2. The central limit theorem implies that*

$$\frac{\sum_{k=1}^n X_k}{\sigma\sqrt{n}} \to \text{ standard normal distribution},$$

  from which a confidence interval can be constructed (in the asymptotic sense).

- *Assuming the variance of the variable is σ^2, the mean square error (MSE) is*

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^n X_k - \mu\right)^2\right] \le \frac{\sigma^2}{n}.$$

This lecture considers another way to evaluate the performance of the estimator, which in more quantitative for the finite n case. More precisely, we consider the probability of $\frac{1}{n}\sum_{k=1}^n X_k$ deviating from μ by a small quantity,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n X_k - \mu\right| \ge t\right] \le \delta. \tag{0.1}$$

If δ is small, then it means with high probability $\frac{1}{n}\sum_{k=1}^n$ is close to μ.

Inequalities of the type (0.1) are known as **concentration inequalities**. Evidently, it is special case of the following more general form

$$\mathbb{P}\left[\left|f(X_1, \cdots, X_n) - \mathbb{E}\left[f(X_1, \cdots, X_n)\right]\right|\right] \le \delta.$$

This form of inequality covers many other important applications, including the generalization error analysis in statistical learning.

Example 0.2 *Given a pair of random variable (X, Y), a central task in statical learning is to find the relationship between X and Y. This is typically formed as the problem of finding a function (hypothesis) h in a function class $\mathcal{H}$ such that the population risk*

$$R(h) = \mathbb{E}\left[\mathcal{L}(h(X), Y)\right]$$

is minimized. Here $\mathcal{L}(\cdot, \cdot)$ represents certain loss function. However, since we do not know the distribution of by only have access to a set of i.i.d samples $X_1, \cdots, X_n$, a computationally tractable alternative is to minimize the empirical risk,

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k).$$

Letting h^* be the minimizer of $R(h)$ and $\hat{h}_n^*$ be the minimizer of $\widehat{R}_n(h)$, in order for $\hat{h}_n^*$ to generalize well for the entire distribution, we wish $R(\hat{h}_n^*)$ should be close to $R(h^*)$. This can be achieved if $\widehat{R}_n(h)$ is close to $R(h)$ for all $h \in \mathcal{H}$ since then they will have their minimizers close to each other. More precisely, we have

$$\begin{aligned} R(\hat{h}_n^*) - R(h^*) &= \left(R(\hat{h}_n^*) - \widehat{R}_n(\hat{h}_n^*)\right) + \left(\widehat{R}_n(\hat{h}_n^*) - \widehat{R}_n(h^*)\right) + \left(\widehat{R}_n(h^*) - R(h^*)\right) \\ &\leq \left|R(\hat{h}_n^*) - \widehat{R}_n(\hat{h}_n^*)\right| + \left|\widehat{R}_n(h^*) - R(h^*)\right| \\ &\leq 2\sup_{h\in\mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right|. \end{aligned}$$

Thus, in order to bound the generalization error $R(\hat{h}_n^*) - R(h^*)$, it suffices to bound

$$\sup_{h\in\mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right| = \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k) - \mathbb{E}\left[\mathcal{L}(h(X, Y))\right]\right| \tag{0.2}$$

Furthermore, in order to provide a high probability (upper) bound for (0.2)*, we can proceed in two steps: first show that*

$$f(Z_1, \cdots, Z_n) := \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k) - \mathbb{E}\left[\mathcal{L}(h(X, Y))\right]\right|, \quad \text{where } Z_k = (X_k, Y_k) \tag{0.3}$$

concentrates around its mean $\mathbb{E}\left[f(Z_1, \cdots, Z_n)\right]$and then provide a bound for $\mathbb{E}\left[f(Z_1, \cdots, Z_n)\right]$. Clearly, the form of f in this example is much more complicated than that in Example 0.1.

As noted in Example 0.2, we also need to bound the expectation of the supremum of a set of random variables for generalization error analysis. A general form is the following **expectation of suprema**:

$$\mathbb{E}\left[\sup_{t\in T} X_t\right],$$

where T is an index set. For the particular case as in (0.3), it is usually referred to as **uniform law of large numbers**. In addition to generalization error analysis, computing the expectation of

suprema also appears in other important applications, such as the estimation of the spectral norm of random matrices.

The concentration inequalities for random variables can be extended to **random matrices**. With a light abuse of notation, capital letters are also used to denote matrices. We will focus on the following type of inequality:

$$\mathbb{P}\left[\left\|\sum_{k=1}^{n}(X_k - \mathbb{E}\left[X_k\right])\right\|_2 \geq t\right] \leq \delta,$$

where $\|\cdot\|_2$ denotes the spectral norm of a matrix. It has applications in for example covariance matrix estimation, sparse linear regression.

For an estimation problem, there can be many different estimators. Thus, a natural question is which one is better or whether an estimator achieves the optimal performance. The answer to this question relies on the criterion that is used. For example, a minimum-variance unbiased estimator (MVUE) is an unbiased estimator that has lower variance than any other unbiased estimators. In this lecture, we consider the minmax framework, and study the **minimax lower bounds** over a family of estimation problems.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 1: Chernoff Method and Concentration Inequalities

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/02/26)*

Agenda:

- Preliminaries
- Sub-Gaussian distributions and Hoeffding inequality
- Sub-exponential distributions and Bernstein inequality
- Martingale methods and bounded differences inequality

1.1 Preliminaries

Theorem 1.1 *Let X be a non-negative random variable. Then,*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > t]\, dt.$$

Proof: We have

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty 1_{\{t<X\}} dt\right] = \int_0^\infty \mathbb{E}\left[1_{\{t<X\}}\right] dt = \int_0^\infty \mathbb{P}[X > t]\, dt,$$

as claimed. ■

Exercise 1.2 *Let X be a random variable and $p \in (0, \infty)$. Show that*

$$\mathbb{E}[|X|^p] = \int_0^\infty pt^{p-1}\mathbb{P}[|X| > t]\, dt.$$

Theorem 1.3 (Jensen's inequality) *If f is convex, then*

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

If f is concave, then

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

Proof: It suffices to prove the first inequality. Let $l(x)$ be the tangent line of $f(x)$ at $\mathbb{E}[X]$. Then,

$$\mathbb{E}[f(X)] \geq \mathbb{E}[l(X)] = l(\mathbb{E}[X]) = f(\mathbb{E}[X]),$$

where the first equality follows from the fact that $l(X)$ is a linear function and the second equality follows from that $l(x)$ is tangent to $f(x)$ at $\mathbb{E}[X]$. ■

Example 1.4 *Let $X_1, \cdots, X_n$ be standard Gaussian random variables, i.e., $X_k \sim \mathcal{N}(0, \sigma^2)$. Then,*

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \leq \sigma\sqrt{2\log n}.$$

Proof: First note that for any $\lambda > 0$

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda X_k)\right] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp(\lambda x)\exp(-x^2/2\sigma^2)dx \\ &= \exp(\sigma^2\lambda^2/2)\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}-\sigma\lambda\right)^2\right)dx \\ &= \exp(\sigma^2\lambda^2/2). \end{aligned} \tag{1.1}$$

It follows that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\max_k X_k\right)\right] = \mathbb{E}\left[\max_k \exp\left(\lambda X_k\right)\right] &\leq \sum_k \mathbb{E}\left[\exp\left(\lambda X_k\right)\right] \\ &\leq n\exp(\sigma^2\lambda^2/2) = \exp(\log(n)+\sigma^2\lambda^2/2).\end{aligned}$$

Thus, the application of Jensen's inequality yields that

$$\exp\left(\mathbb{E}\left[\lambda\max_k X_k\right]\right) \leq \mathbb{E}\left[\exp\left(\lambda\max_k X_k\right)\right] \leq \exp(\log(n)+\sigma^2\lambda^2/2),$$

which leads to

$$\mathbb{E}\left[\max_k X_k\right] \leq \frac{\log(n)}{\lambda} + \frac{\sigma^2\lambda}{2}.$$

Taking $\lambda = \sqrt{2\log(n)}/\sigma$ concludes the proof. ■

One goal of this course is to study the concentration of a random quantity (function of random variables/vectors/matrices) around its mean.

There are two probability results which can provide some instructions: law of large numbers and central limit theorem. As an example, consider the problem of estimating the unknown mean μ of a random variable from its i.i.d samples $X_1, \cdots, X_n$. A very natural way to estimate μ is to use the sample mean

$$f(X_1, \cdots, X_n) = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

It is not hard to see that $\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n X_k\right] = \mu$. Moreover, the law of large numbers tells us that

$$\frac{1}{n}\sum_{k=1}^n X_k \textit{ converges to } \mu \textit{ in probability when } n \textit{ goes to infinity.}$$

If the variance of the random variable exists, denoted σ^2, the central limit theorem implies that

$$\frac{\sum_{k=1}^{n}(X_k-\mu)}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1) \text{ when } n \text{ goes to infinity.}$$

However, both the law of large number and the central limit theorem are asymptotic results and they cannot tell us how well $f(X_1,\cdots,X_n)=\frac{1}{n}\sum_{k=1}^{n}X_k$ concentrates around μ when n is finite. That is, they cannot be used to bound the probability that $\frac{1}{n}\sum_{k=1}^{n}X_k$ deviates from its mean, namely

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|\geq t\right]. \tag{1.2}$$

In this course, we will study this kind of concentration problem directly. Indeed, we will begin with the bound for (1.2), and then move to concentration inequalities which involve more complicated f than the sum of random variables in order to tackle more difficult problems than mean estimation.

1.1.1 Markov and Chebshev Inequalities

One way to bound the tail probability of a random variable is to control its moments. The most elementary tail bound is Markov inequality which only uses the expectation (first moment) of the variable.

Theorem 1.5 (Markov inequality) *If X is a non-negative variable, then any $t>0$ one has*

$$\mathbb{P}[X>t]\leq\frac{\mathbb{E}[X]}{t}.$$

Proof: A simple calculation yields that

$$\mathbb{E}[X]\geq\mathbb{E}\left[X1_{\{X>t\}}\right]\geq t\mathbb{P}[X>t],$$

as claimed. ■

Using high order moments typically leads to stronger probability bounds, e.g., Chebshev inequality.

Theorem 1.6 (Chebshev inequality) *For a random variable with finite variance, there holds,*

$$\mathbb{P}[|X-\mathbb{E}[X]|>t]\leq\frac{\mathbb{E}\left[|X-\mathbb{E}[X]|^2\right]}{t^2}.$$

Proof: Apply Markov inequality to the random variable $|X-\mathbb{E}[X]|^2$ ■

Example 1.7 *Let X be a Bernoulli variable,*

$$X=\begin{cases}1 & \text{with probability } p\\ 0 & \text{with probability } 1-p.\end{cases}$$

Let $X_k,\ i = 1, \cdots, n$ be i.i.d copies of X, and define $S_n = \sum_{k=1}^{n} X_k$. For a positive number $p < \alpha < 1$, the application of Markov inequality gives

$$\mathbb{P}\left[S_n > \alpha n\right] \leq \frac{\mathbb{E}\left[S_n\right]}{\alpha n} = \frac{p}{\alpha},$$

while the application of Chebshev inequality gives

$$\begin{aligned}
\mathbb{P}\left[S_n > \alpha n\right] &= \mathbb{P}\left[S_n - pn > (\alpha - p)n\right] \\
&\leq \mathbb{P}\left[|S_n - pn| > (\alpha - p)n\right] \\
&\leq \frac{\mathbb{E}\left[|S_n - pn|^2\right]}{(\alpha - p)^2 n^2} \\
&= \frac{p(1-p)}{(\alpha - p)^2 n}.
\end{aligned}$$

This example shows that we can have a better bound (order of $1/n$ rather than a constant order) by Chebshev inequality. As can be seen later, by one of the main results in this lecture – Hoeffding inequality, we can establish a tail bound that decays exponentially fast.

There is a natural way to extend the Markov inequality to random variables with higher-order moments. For instance, if $\mathbb{E}\left[|X - \mathbb{E}\left[X\right]|^k\right]$ exists for some $k > 1$, then an application of the Markov inequality to the random variable $|X - \mathbb{E}\left[X\right]|^k$ yields that

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| > t\right] \leq \frac{\mathbb{E}\left[|X - \mathbb{E}\left[X\right]|^k\right]}{t^k}.$$

Of course, we can use other functions rather than a single moment of the random variable. The tight bounds that will be established next are indeed based on the *moment generating function* (MGF, a mixture of all moments),

$$\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right].$$

In the same spirit of the Markov or Chebshev inequality, we have

$$\mathbb{P}\left[X - \mathbb{E}\left[X\right] > t\right] = \mathbb{P}\left[e^{\lambda(X - \mathbb{E}[X])} > e^{\lambda t}\right] \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right], \quad \lambda > 0.$$

Note that in the above inequality, there is a free parameter $\lambda > 0$ to choose. The *Chernoff method* chooses λ in an interval $[0, b]$ (b can be infinite or finite up to the bound of moment generating function) such that the righthand side is minimized, leading to

$$\mathbb{P}\left[X - \mathbb{E}\left[X\right] > t\right] \leq \inf_{\lambda \in [0,b]} e^{-\lambda t}\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right]. \tag{1.3}$$

It is easy to see that the key in the application of the Chernoff method is to estimate $\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right]$. Indeed, one advantage of using moment generating function over the all possible polynomials is that the former one is a smooth function with the parameter λ and can be easily manipulated. Next we will study two different distributions based on the different behaviors of their moment generating functions, as well as the corresponding concentration inequalities.

1.2 Sub-Gaussian Distributions and Hoeffding Inequality

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a normal/Gaussian distribution of mean μ and variance σ^2. We have

$$\mathbb{P}\left[|X - \mu| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{1.4}$$

Exercise 1.8 *Prove* (1.4).

The above inequality shows the tail bound of normal distribution decays exponentially fast. Thus, it is interesting to see whether there are other distributions which exhibit similar behavior. The answer is affirmative, and this family of distributions are known as sub-Gaussian distributions. They are fully characterized by the behavior of their moment generating functions.

Definition 1.9 (Sub-Gaussian distribution) *A random variable X with mean μ is sub-Gaussian if there exists a positive number $\nu > 0$ such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\lambda^2\nu^2/2} \quad \textit{for all} \quad \lambda \in \mathbb{R}. \tag{1.5}$$

Remark 1.10 *Though here ν is NOT equivalent to the variance of a random variable, we can sometimes think of it as the variance to get some intuition.*

Example 1.11 (Gaussian distribution) *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian random variable. It follows from* (1.1) *that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\lambda^2\sigma^2/2} \quad \textit{for all} \quad \lambda \in \mathbb{R}.$$

Thus X is sub-Gaussian with parameter $\nu = \sigma$.

Example 1.12 (Rademacher variables) *A Rademacher random variable ε takes the values $\{-1, +1\}$ in the same probability. By taking expectations and using the power series expansion, we have*

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda X}\right] &= \frac{1}{2}\left(e^{-\lambda} + e^{\lambda}\right) \\
&= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{2^k k!} \\
&= e^{\lambda^2/2},
\end{aligned}$$

which shows that ε is a sub-Gaussian variable with parameter $\nu = \sigma = 1$.

Example 1.13 (Bounded random variables) *Let X be zero-mean, and supported on a closed interval $[a, b]$. We claim that*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2(b-a)^2/8}.$$

In other words, X is sub-Gaussian with parameter $(b-a)/2$. To show this, define $\psi(\lambda)$ (knowns as log-moment generating function) as

$$\psi(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right].$$

Then it suffices to show

$$\psi(\lambda) \leq \frac{\lambda^2 (b-a)^2}{8}.$$

First, it is not hard to see that

$$\psi'(\lambda) = \frac{\mathbb{E}\left[X e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]}$$

and

$$\begin{aligned}\psi''(\lambda) &= \frac{\mathbb{E}\left[e^{\lambda X}\right]\mathbb{E}\left[X^2 e^{\lambda X}\right] - \left(\mathbb{E}\left[X e^{\lambda X}\right]\right)^2}{\left(\mathbb{E}\left[e^{\lambda X}\right]\right)^2} \\ &= \mathbb{E}\left[X^2 \frac{e^{\lambda X}}{\mathbb{E}\left[e^{\lambda X}\right]}\right] - \left(\mathbb{E}\left[X \frac{e^{\lambda X}}{\mathbb{E}\left[e^{\lambda X}\right]}\right]\right)^2.\end{aligned}$$

It follows immediately that

$$\psi'(0) = 0.$$

Moreover, the expression for $\psi''(\lambda)$ implies that $\psi''(\lambda)$ is indeed the variance of X after a change of measure. Thus, by the variational definition o variance, we have

$$\psi''(\lambda) \leq \mathbb{E}\left[\left(X - \frac{b+a}{2}\right)^2 \frac{e^{\lambda X}}{\mathbb{E}\left[e^{\lambda X}\right]}\right] \leq \frac{(b-a)^2}{4}, \quad \forall \lambda \in \mathbb{R}.$$

Also noting that $\psi(0) = 0$, we finally have

$$\psi(\lambda) = \psi(0) + \psi'(0)\lambda + \frac{1}{2}\psi''(\xi)\lambda^2 \leq \frac{\lambda^2 (b-a)^2}{8},$$

which completes the proof.

1.2.1 Hoeffding Inequality

By the Chernoff method (see (1.3)) we can show that sub-Gaussian random variables have the same concentration properties as Gaussian random variables.

Theorem 1.14 (Hoeffding inequality) *Let X (with $\mathbb{E}[X] = \mu$) be a sub-Gaussian random variable with parameter ν. Then,*

$$\mathbb{P}\left[|X - \mu| > t\right] \leq 2e^{-\frac{t^2}{2\nu^2}}.$$

Proof: Inserting the sub-Gaussian property into(1.3) and optimizing the right hand side of the above inequality with respect to $\lambda > 0$ yields that

$$\mathbb{P}\left[X - \mu > t\right] \leq e^{-\frac{t^2}{2\nu^2}}.$$

Moreover, by considering $-X$, we can get

$$\mathbb{P}\left[X - \mu < -t\right] \leq e^{-\frac{t^2}{2\nu^2}},$$

which concludes the proof. ■

Chernoff bounds can be easily extended to sums of independent random variables because of the tensorization property of the moment generating functions in this situation, i.e., moment generating functions of sums of independent random variables become products of moment generating functions.

Proposition 1.15 *Let $X_1, \cdots, X_n$ be independent ν_k^2 sub-Gaussian random variables. Then $\sum_{k=1}^n X_k$ is a sub-Gaussian random variable with parameter $\nu = \sum_{k=1}^n \nu_k^2$.*

Proof: The moment generating function $\sum_{k=1}^n X_k$ can be upper bounded as

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\left(\sum_{k=1}^n X_k - \mathbb{E}\left[\sum_{k=1}^n X_k\right]\right)\right)\right] &= \mathbb{E}\left[\prod_{k=1}^n \exp\left(X_k - \mathbb{E}\left[X_k\right]\right)\right] = \prod_{k=1}^n \mathbb{E}\left[\exp\left(X_k - \mathbb{E}\left[X_k\right]\right)\right] \\ &\leq \prod_{k=1}^n \exp\left(\frac{\lambda^2\nu_k^2}{2}\right) = \exp\left(\frac{\lambda^2 \sum_{k=1}^n \nu_k^2}{2}\right),\end{aligned}$$

which completes the proof. ■

The follow general Hoeffding inequality follows immediately from Theorem 1.14 and Proposition 1.15.

Theorem 1.16 (General Hoeffding inequality) *Suppose $X_k,\ k = 1, \cdots, n$ are independent random variables, and X_k has mean μ_k and sub-Gaussian parameter ν_k. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\sum_{k=1}^n (X_k - \mu_k)\right| > t\right] \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}\right).$$

Example 1.17 *Suppose $X_k,\ k = 1, \cdots, n$ are independent random variables satisfying $\mathbb{E}\left[X_k\right] = \mu_k$ and $a \leq X_k \leq b$. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\sum_{k=1}^n (X_k - \mu_k)\right| > t\right] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

Example 1.18 *Let us revisit Example 1.7 using the Hoeffding inequality, yielding*

$$\mathbb{P}\left[S_n > \alpha n\right] = \mathbb{P}\left[\sum_{k=1}^n (X_k - p) \geq (\alpha - p)n\right] \leq \exp\left(-\frac{(\alpha - p)^2 n}{2}\right),$$

which decreases faster than what Chebshev inequality gives.

1.2.2 Equivalent Characterizations of sub-Gaussian Distribution[1]

We have shown that the sub-Gaussian property implies the exponential decay of the tail probability. In fact, the converse direction also holds true. Moreover, there are several equivalent characterizations of the sub-Gaussian distribution.

Theorem 1.19 *Let X be a mean zero random variable. Then the following four statements are equivalent.*

1. *X is sub-Gaussian satisfying,*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(c_1\lambda^2\nu^2\right) \quad \textit{for all} \quad \lambda \in \mathbb{R}.$$

2. *The tails of X satisfy*

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-\frac{t^2}{c_2\nu^2}\right) \quad \textit{for all } t \geq 0.$$

3. *The moments of X satisfy*

$$\|X\|_{L_p} := \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq c_3\nu\sqrt{p} \quad \textit{for all} \quad p \geq 1.$$

4. *The moment generating function of X^2 is bounded at some point[2],*

$$\mathbb{E}\left[\exp\left(\frac{X^2}{c_4\nu^2}\right)\right] \leq e.$$

Here, $c_i,\ i = 1, \cdots, 4$ are positive, absolute constants (see the notational remark in the syllabus).

Proof: We will proceed the proof in the following way: $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 1$.

$1 \Rightarrow 2$**:** We have established this above using the Chernoff method.

$2 \Rightarrow 3$**:** W.l.og, assume $c_2 = 1$. Then,

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= p\int_0^\infty t^{p-1}\mathbb{P}\left[|X| \geq t\right] dt \\
&\leq 2p\int_0^\infty t^{p-1}\exp\left(-\frac{t^2}{\nu^2}\right) dt \\
&= p\nu^p\int_0^\infty s^{\frac{p}{2}-1}e^{-s}ds \qquad (\text{letting } s = \frac{t^2}{\nu^2}) \\
&= p\nu^p\Gamma(p/2) \qquad (\Gamma(z) \text{ is a Gamma function}) \\
&\leq p\nu^p\left(p/2\right)^{p/2} \qquad (\Gamma(z) \leq z^z, \textbf{ check this! }).
\end{aligned}$$

Taking the p-th root on both sides and noting that $p^{1/p} \leq e$ (**check this!**) concludes the proof.

[1] This part can be skipped if you find it difficult.

[2] The constant e on the righthand side does not have any special meaning and can be replaced by any absolute constant (similar to different scales of a norm.

$3 \Rightarrow 4$**:** As above, we can assume $c_3 = 1$. Then

$$\mathbb{E}\left[\exp\left(\frac{X^2}{c_4\nu^2}\right)\right] = \sum_{p=0}^{\infty} \frac{\mathbb{E}\left[X^{2p}\right]}{p! c_4^p \nu^{2p}} \leq \sum_{p=0}^{\infty} \frac{\nu^{2p}(2p)^p}{p! c_4^p \nu^{2p}}$$
$$\leq \sum_{p=0}^{\infty} \left(\frac{2e}{c_4}\right)^p = \frac{1}{1-2e/c_4} \leq e \qquad (\text{use } p! \geq (p/e)^p, \textbf{ check this!})$$

provided $c_4 \geq 2e/(1-1/e)$.

$4 \Rightarrow 1$**:** Again, we can assume $c_4 = 1$. First noting that

$$\lambda x \leq \frac{\lambda^2\nu^2}{2} + \frac{x^2}{2\nu^2},$$

we have

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(\lambda^2\nu^2/2\right) \mathbb{E}\left[\exp\left(X^2/(2\nu^2)\right)\right] \leq \exp\left(\lambda^2\nu^2/2\right) \sqrt{\mathbb{E}\left[\exp\left(X^2/(\nu^2)\right)\right]}$$
$$\leq e^{1/2} \exp\left(\lambda^2\nu^2/2\right) \leq \exp\left(\lambda^2\nu^2\right)$$

provided $|\lambda| \geq 1/\nu$, where the second inequality follows from the Jensen inequality to the function $\sqrt{x}$. Thus, it remains to discuss the case $|\lambda| < 1/\nu$. In this situation, using the inequality $e^x \leq x + e^{x^2}$ (**check this!**) we have

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \underbrace{\mathbb{E}\left[\lambda X\right]}_{=0} + \mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] = \mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] = \mathbb{E}\left[\left(\exp\left(X^2/\nu^2\right)\right)^{(\lambda^2\nu^2)}\right]$$
$$\leq \left(\mathbb{E}\left[\exp\left(X^2/\nu^2\right)\right]\right)^{\lambda^2\nu^2}$$
$$\leq \exp\left(\lambda^2\nu^2\right),$$

where in the second inequality we utilize the Jensen inequality by noting that $\lambda^2\nu^2 < 1$. ■

Exercise 1.20 (Khintchine inequality) *Let $X_k,\ k = 1, \cdots, n$ be i.i.d, zero mean, unit variance sub-Gaussian random variables with parameter ν^2. Letting $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$, show that for any $p \in [2, \infty)$ we have*

$$\|a\|_2 \leq \|\sum_{k=1}^{n} a_k X_k\|_{L_p} \lesssim \nu\sqrt{p}\|a\|_2.$$

(See the notational remark in the syllabus for the meaning of $\lesssim$.)

At the end of this section we present the following lemma, where a very useful *decoupling technique* via the introduction of an independent random variable for auxiliary randomness is used in the proof. See Chapter 6.1 of of *High-dimensional probability: An introduction with applications in data science* by Roman Vershynin for the general decoupling technique.

Lemma 1.21 *Let X be mean zero sub-Gaussian random variable with parameter ν^2. Then*

$$\mathbb{E}\left[\exp\left(\lambda X^2\right)\right] \leq \frac{1}{[1-2\lambda\nu^2]_+^{1/2}},$$

where the equality holds for $X \sim \mathcal{N}(0, \nu^2)$.

Proof: When $X \sim \mathcal{N}(0, \nu^2)$, we can establish the equality by direction integral based on the pdf of the Gaussian distribution.

For a general sub-Gaussian variable X, let Z be an independent $\mathcal{N}(0,1)$ random variable. Noting that

$$\mathbb{E}\left[\exp\left(\lambda x Z\right)\right] = \exp\left(\frac{\lambda^2 x^2}{2}\right),$$

we have

$$\mathbb{E}\left[\exp\left(\lambda X^2\right)\right] = \mathbb{E}\left[\exp\left(\sqrt{2\lambda} X Z\right)\right] \leq \mathbb{E}\left[\exp\left(\lambda \nu^2 Z^2\right)\right] \leq \frac{1}{[1-2\lambda\nu^2]_+^{1/2}},$$

where the first inequality follows from the sub-Gaussian property of X and the second inequality follows from the the fact Z is $\mathcal{N}(0,1)$. ■

1.3 Sub-exponential Distributions and Bernstein Inequality

As we have seen from above, sub-Gaussian distribution is an extension of the Gaussian distribution. In contrast, sub-exponential distribution is an extension of the squared Gaussian distribution. For simplicity, let $X \sim \mathcal{N}(0,1)$ be standard normal distribution and let $Z = X^2$ be χ^2. Then,

$$\mathbb{E}\left[e^{\lambda(Z-1)}\right] = \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}, & \text{if } \lambda < \frac{1}{2} \\ \text{not exist}, & \text{otherwise.} \end{cases}$$

Thus, the moment generating function does not exist over the entire real line. Moreover, since $1 - x > e^{-x^2 - x}$ (**check this!**) for all $x < 1/2$, one has

$$\mathbb{E}\left[e^{\lambda(Z-1)}\right] \leq e^{4\lambda^2/2} \text{ for all } |\lambda| < \frac{1}{4}.$$

Compared with (1.5), we see that similar bound only holds in a local neighborhood of zero. This kind of condition defines the family of sub-exponential distributions.

Definition 1.22 (Sub-exponential distribution) *A random variable X with mean μ is sub-exponential if there are non-negative parameters (ν, b) such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\nu^2\lambda^2/2} \quad \textit{for all} \quad |\lambda| < 1/b.$$

Example 1.23 (χ^2-distribution) *We have shown that if $X \sim \mathcal{N}(0,1)$, then X^2 is sub-exponential with parameters $(\nu, b) = (2, 4)$.*

Example 1.24 (Exponential distribution) *Recall that X has exponential distribution with rate $a > 0$ if the pdf of X is given by*

$$f(x) = \begin{cases} ae^{-ax} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

A direct calculation shows that $\mathbb{E}[X] = \frac{1}{a}$. *For simplicity let* $a = 1$. *Then we have*

$$\mathbb{E}\left[\exp\left(\lambda(X-1)\right)\right] = \int_0^\infty e^{\lambda(x-1)}e^{-x}dx = \begin{cases} \frac{e^{-\lambda}}{1-\lambda} & \lambda < 1 \\ \infty & \lambda \geq 1. \end{cases}$$

The application of $1 - x > e^{-x^2-x}$ *for* $x < 1/2$ *yields that*

$$\mathbb{E}\left[\exp\left(\lambda(X-1)\right)\right] \leq e^{\lambda^2} \quad \textit{for all} \quad |\lambda| < \frac{1}{2}.$$

Bernstein condition based on the moments of X provides an indirect way to verify the sub-exponential property. More precisely, let X be random variable with mean μ and variance σ^2. We say Bernstein's condition with parameter b holds if

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \frac{1}{2}k!\sigma^2 b^{k-2} \quad \text{for } k = 3, 4, \cdots$$

Lemma 1.25 *If* X *satisfies the Bernstein condition, then* X *is sub-exponential with parameters* $(\sqrt{2}\sigma, 2b)$.

Proof: We have

$$\begin{aligned} \mathbb{E}\left[e^{\lambda(X-\mu)}\right] &= \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[\lambda^k(X-\mu)^k\right]}{k!} \\ &\leq 1 + \frac{\sigma^2\lambda^2}{2} + \frac{\sigma^2\lambda^2}{2}\sum_{k=1}^{\infty}(|\lambda|b)^k \\ &= 1 + \frac{\sigma^2\lambda^2}{2} + \frac{\sigma^2\lambda^2|\lambda|b}{2(1-|\lambda|b)} \quad \left(\forall\ |\lambda| < \frac{1}{b}\right) \\ &= 1 + \frac{\sigma^2\lambda^2/2}{1-|\lambda|b} \\ &\leq e^{\frac{\sigma^2\lambda^2/2}{1-|\lambda|b}} \\ &\leq e^{\frac{\sigma^2(\sqrt{2}\lambda)^2}{2}} \quad \forall\ |\lambda| \leq \frac{1}{2b}, \end{aligned} \tag{1.6}$$

which implies X is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$. ■

Exercise 1.26 *Let* X *be a random variable with* $\mathbb{E}[X] = \mu$. *Suppose* $|X - \mu| \leq b$. *Show that* X *satisfies the Bernstein condition.*

1.3.1 Bernstein Inequality

For sub-exponential distributions we can establish the Bernstein tail, which mixes the Gaussian tail and the exponential tail.

Theorem 1.27 (Bernstein inequality) *Suppose* X *is a sub-exponential variable with parameters* (ν, b). *Then*

$$\mathbb{P}\left[|X-\mu| > t\right] \leq 2\exp\left(-\frac{1}{2}\min\left(\frac{t^2}{\nu^2}, \frac{t}{b}\right)\right) = \begin{cases} 2e^{-\frac{t^2}{2\nu^2}}, & \textit{if } \ 0 \leq t \leq \frac{\nu^2}{b} \\ 2e^{-\frac{t}{2b}} & \textit{if } \ t > \frac{\nu^2}{b}. \end{cases}$$

Proof: We assume without loss of generality $\mu = 0$. The application of the Chernoff approach yields that

$$\mathbb{P}\left[X - \mu > t\right] \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda X}\right] \leq e^{-\lambda t + \nu^2\lambda^2/2}, \quad \forall 0 < \lambda \leq 1/b.$$

Optimizing the right hand side with respect to λ over $(0, 1/b]$ gives the one-sided tail bound. Consider $-X$ for the other tail bound. ∎

Example 1.28 *Let X be a random variable such that $|X - \mu| \leq b$. We know that it is also sub-exponential with parameters $(\sqrt{2}\sigma, b)$ where σ is the variance of X. Then the Bernstein inequality implies that*

$$\mathbb{P}\left[|X - \mu| > t\right] \leq \begin{cases} 2e^{-\frac{t^2}{4\sigma^2}}, & \text{if } \ 0 \leq t \leq \frac{\sigma^2}{b} \\ 2e^{-\frac{t}{2b}} & \text{if } \ t > \frac{\sigma^2}{b}, \end{cases}$$

while the application of the Hoeffding type bound gives

$$\mathbb{P}\left[|X - \mu| > t\right] \leq 2e^{-\frac{t^2}{2b^2}}.$$

It is evident that when t is sufficiently large, the Hoeffding type bound is better than the Bernstein type bound. However, it is worth noting that if t is small, the Bernstein type bound might be better than the Hoeffding type bound since it is possible that $\sigma^2 \ll b^2$.

For sub-exponential variable satisfying the Bernstein condition, we can actually establish the following slightly improved bound

$$\mathbb{P}\left[|X - \mu| > t\right] \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right). \tag{1.7}$$

Exercise 1.29 *Prove*(1.7). *(**Hint:** Apply the Chernoff method to the inequality* (1.6) *directly.)*

The Bernstein inequality shows that for small t the tail bound is of the Gaussian type while for large t the tail bound is of the exponential type. Though the region of the Gaussian tail is a bit restrictive for a single random variable (for example when t is very close to 0 we even have $e^{-t} \lesssim e^{-t^2}$), for the sum of independent random variables this region will increase as the number of random variables increases.

Proposition 1.30 *Suppose that $X_k,\ k = 1, \cdots, n$ are n independent variables, and that X_k is sub-exponential with parameters (ν_k, b_k). Then $\sum_{k=1}^n (X_k - \mu_k)$ is sub-exponential with parameters (ν_*, b_*), where*

$$\nu_*^2 = \sum_{k=1}^n \nu_k^2 \quad \text{and} \quad b_* = \max_{1 \leq k \leq n} b_k.$$

Moreover, if $X_k,\ k = 1, \cdots, n$ are i.d.d sub-exponential with parameters (ν, b), then $\sum_{k=1}^n (X_k - \mu)$ is sub-exponential with parameters $(\sqrt{n}\nu, b)$.

Proof: The moment generating function of $\sum_{k=1}^n (X_k - \mu_k)$ can be bounded as follows

$$\mathbb{E}\left[\exp\left(\lambda \sum_{k=1}^n (X_k - \mu_k)\right)\right] = \prod_{k=1}^n \mathbb{E}\left[\exp\left(\lambda(X_k - \mu_k)\right)\right] \leq \prod_{k=1}^n \exp\left(\lambda^2 \nu_k^2/2\right),$$

where the inequality is valid for all $\lambda < (\max_k b_k)^{-1}$. ■

The following general Bernstein inequality follows immediately from the last proposition.

Theorem 1.31 (General Bernstein inequality) *Suppose that $X_k,\ i = 1, \cdots, n$ are n independent variables, and that X_k is sub-exponential with parameters (ν_k, b_k). Then,*

$$\mathbb{P}\left[\left|\sum_{k=1}^n (X_k - \mu_k)\right| \geq t\right] \leq 2\exp\left(-\frac{1}{2}\min\left(\frac{t^2}{\nu_*^2}, \frac{t}{b_*}\right)\right) = \begin{cases} 2e^{-\frac{t^2}{2\nu_*^2}}, & \text{if } 0 \leq t \leq \frac{\nu_*^2}{b_*} \\ 2e^{-\frac{t}{2b_*}} & \text{if } t > \frac{\nu_*^2}{b_*}, \end{cases}$$

where

$$\nu_*^2 = \sum_{k=1}^n \nu_k^2 \quad \text{and} \quad b_* = \max_{1 \leq k \leq n} b_k.$$

Example 1.32 *Let $Z_k,\ k = 1, \cdots, n$ be i.i.d Chi-square variables. Noting that Z_k is sub-exponential with parameters $(2, 4)$, there holds*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n (Z_k - 1)\right| \geq t\right] \leq 2\exp\left(-\frac{n}{8}\min\left(t^2, t\right)\right).$$

Additionally, as a remark we roughly illustrates that why the exponential tail in the Bernstein inequality does not contradicts the central limit theorem. Let $X_k,\ k = 1, \cdots, n$ be i.i.d sub-exponential random variables with parameters (ν, b). Consider the rescaled random variable $\frac{1}{\sqrt{n}}\sum_{k=1}^n (X_k - \mu)$. The application of the Bernstein inequality yields that

$$\mathbb{P}\left[\left|\frac{1}{\sqrt{n}}\sum_{k=1}^n (X_k - \mu)\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\nu^2}} & 0 \leq t \leq \frac{\sqrt{n}\nu^2}{b} \\ 2e^{-\frac{\sqrt{n}t}{2b}} & t > \frac{\sqrt{n}\nu^2}{b}. \end{cases}$$

It is clear that the Gaussian tail bound region $0 \leq t \leq \frac{\sqrt{n}\nu^2}{b}$ increases linearly with respect to $\sqrt{n}$.

1.3.2 Equivalent Characterizations of sub-Exponential Distribution[3]

Under a generalized definition of sub-exponential distributions (in for example *High-dimensional probability: An introduction with applications in data science* by Roman Vershynin), we may establish the following equivalence.

Theorem 1.33 *Let X be a mean zero random variable. Then the following four statements are equivalent.*

[3]This part can be skipped if you find it difficult.

1. *X is sub-exponential satisfying,*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(c_1\lambda^2\nu^2\right) \quad \textit{for all} \quad |\lambda| \leq \frac{c_1'}{\nu}. \tag{1.8}$$

Note that if X satisfies Definition 1.22, then it will satisfy (1.8) *with* $\max(\nu, b)$*. However, the resulting Bernstein inequality will be weaker since both ν and b will be replaced by* $\max(\nu, b)$*.*

2. *The tails of X satisfy*

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-\frac{t}{c_2\nu}\right) \quad \textit{for all } t \geq 0.$$

3. *The moments of X satisfy*

$$\|X\|_{L_p} = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq c_3\nu p \quad \textit{for all} \quad p \geq 1.$$

4. *The moment generating function of $|X|$ is bounded at some point*[4]*,*

$$\mathbb{E}\left[\exp\left(\frac{|X|}{c_4\nu}\right)\right] \leq e.$$

Here, $c_i,\ i = 1, \cdots, 4$ and c_1' are positive, absolute constants.

Proof: We will proceed the proof in the following way: $2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 2$ and $1 \Leftrightarrow 3$.

$2 \Rightarrow 3$**:** W.l.o.g, we assume $c_2 = 1$. Then,

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= p\int_0^\infty t^{p-1}\mathbb{P}\left[|X| \geq t\right] dt \\
&\leq 2p\int_0^\infty t^{p-1}\exp\left(-t/v\right) dt \\
&= 2p\nu^p\Gamma(p) \\
&\leq 2p\nu^p p^p.
\end{aligned}$$

Taking a p-th root on both sides yields the result.

$3 \Rightarrow 4$**:** As above we assume $c_3 = 1$. Then,

$$\mathbb{E}\left[\exp\left(\frac{X}{c_4\nu}\right)\right] = \sum_{p=0}^\infty \frac{\mathbb{E}\left[|X|^p\right]}{p!c_4^p\nu^p} \leq \sum_{p=0}^\infty \frac{(\nu p)^p}{p!c_4^p\nu^p} \leq \sum_{p=0}^\infty \left(\frac{e}{c_4}\right)^p = \frac{1}{1 - e/c_4} \leq e$$

provided $c_4 \geq e/(1 - 1/e)$.

[4]The constant e on the righthand side does not have any special meaning and can be replaced by any absolute constant (similar to different scales of a norm.

$4 \Rightarrow 2$**:** Assume $c_4 = 1$. Applying the Markov inequality to $e^{X/\nu}$, it is easy to see that

$$\mathbb{P}[X \geq t] \leq e^{1-t/\nu}.$$

With the same result for the negative tail, we have

$$\mathbb{P}[|X| \geq t] \leq \min(2e^{1-t/\nu}, 1) \leq 2\exp\left(-\frac{2t}{5\nu}\right),$$

where in the second inequality we choose a constant c such that both $2e^{1-t/\nu} \leq 2e^{-ct/\nu}$ when t is greater than some threshold and $2e^{-ct/\nu} \geq 1$ when t is greater than the same threshold.

$1 \Rightarrow 3$ Using the numerical inequality $|x|^p \leq p^p(e^x + e^{-x})$ for all x and $p > 0$ (**check this!**) with $x = \frac{c_1' X}{\nu}$ and then taking the expectation yields

$$\begin{aligned}\mathbb{E}\left[\left|\frac{c_1' X}{\nu}\right|^p\right] &\leq \mathbb{E}\left[p^p\left(\exp\left(\frac{c_1' X}{\nu}\right) + \exp\left(\frac{-c_1' X}{\nu}\right)\right)\right] \\ &\leq 2p^p \exp\left(c_1 \frac{(c_1')^2}{\nu^2}\nu^2\right),\end{aligned}$$

which gives 3 after simplification.

$3 \Rightarrow 1$ Assume $c_3 = 1$ for simplicity. By Taylor's expansion we have

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &= 1 + \mathbb{E}[\lambda X] + \sum_{p=2}^{\infty} \frac{\lambda^p \mathbb{E}[X^p]}{p!} \\ &\leq 1 + \sum_{p=2}^{\infty} \frac{(\lambda p \nu)^p}{p!} \\ &\leq 1 + \sum_{p=2}^{\infty} (\lambda e \nu)^p \qquad (\text{use } p! \geq (p/e)^p) \\ &= 1 + \frac{(\lambda e\nu)^2}{1 - \lambda e \nu} \\ &\leq 1 + 2(\lambda e \nu)^2 \qquad (\text{assume } \lambda e\nu \leq 1/2) \\ &\leq \exp\left(2(\lambda e \nu)^2\right),\end{aligned}$$

which concludes the proof with $c_1 = 2e^2$ and $c_1' = 2e$. ■

1.4 Martingale Methods and Bounded Differences Inequality

Martingale is defined based on conditional expectation, which is about finding an equivalent version (by preserving expectation) of a random variable under given information pool (Wikipedia is a good source for more details). In general, a sequence $\{Z_k\}_{k=0}^{\infty}$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=0}^{\infty}$, the pair $\{(Z_k, \mathcal{F}_k)\}_{k=0}^{\infty}$ is called a martingale if for all $k \geq 0$,

$$\mathbb{E}[|Z_k|] < \infty, \quad \text{and} \quad \mathbb{E}[Z_{k+1}|\mathcal{F}_k] = Z_k.$$

However, to avoid the technical difficulties of filtration or σ-algebra, we proceed directly in terms of the marginal difference sequence.

Let $f(x_1, \cdots, x_n) : \mathcal{X}^n \to \mathbb{R}$ be a function and let $X_1, \cdots, X_n$ be independent random variables taking values in the sample space $\mathcal{X}$ (elements of $\mathcal{X}$ can be scalars, vectors, and so on). Define the following martingale difference sequence

$$\begin{aligned} D_1 &= \mathbb{E}\left[f(X_1, \cdots, X_n)\right], \quad D_n = f(X_1, \cdots, X_n), \\ D_k &= \mathbb{E}\left[f(X_1, \cdots, X_n)|X_1, \cdots, X_k\right] - \mathbb{E}\left[f(X_1, \cdots, X_n)|X_1, \cdots, X_{k-1}\right] \\ &= \mathbb{E}_{k+1}[f(X_1, \cdots, X_k, \underbrace{X_{k+1}, \cdots, X_n})] - \mathbb{E}_k[f(X_1, \cdots, X_{k-1}, \underbrace{X_k, X_{k+1}, \cdots, X_n})], \end{aligned} \tag{1.9}$$

where $\mathbb{E}_{k+1}[\cdot]$ means taking expectation with respect to $X_{k+1}, \cdots, X_n$ while keeping $X_1, \cdots, X_n$ unchanged, i.e.,

$$\mathbb{E}_{k+1}\left[f(X_1, \cdots, X_k, X_{k+1}, \cdots, X_n)\right] := \mathbb{E}\left[f(x_1, \cdots, x_k, X_{k+1}, \cdots, X_n)\right]|_{x_1 = X_1, \cdots, x_k = X_k}.$$

It is clear that

$$\mathbb{E}\left[D_k|X_1, \cdots, X_{k-1}\right] = 0,$$

and consequently $\mathbb{E}\left[D_k\right] = 0$. Moreover, we have

$$\sum_{k=1}^{n} D_k = f(X_1, \cdots, X_n) - \mathbb{E}\left[f(X_1, \cdots, X_n)\right],$$

which enables us to study the concentration of $f(X_1, \cdots, X_n)$ around its mean by studying the concentration of the sum $\sum_{k=1}^n D_k$.

Even though $D_k,\ k = 1, \cdots, n$ are not independent to each other, the martingale structure enables us to establish the sub-Gaussian tail once they are bounded.

Theorem 1.34 (Azuma-Hoeffding tail bound) *Let $\{D_k\}_{k=1}^n$ be the martingale difference sequence defined in* (1.9)*. Suppose that $A_k \le D_k \le B_k$ almost surely for all $k \ge 1$, where A_k and B_k are functions of $X_1, \cdots, X_{k-1}$. If $B_k - A_k \le L_k$, then for all $t \ge 0$, we have*

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \ge t\right] \le 2e^{-\frac{2t^2}{\sum_{k=1}^n L_k^2}}$$

Proof: Noting that $\mathbb{E}\left[D_k|X_1, \cdots, X_{k-1}\right] = 0$, repeating the argument in Example 1.13 for a conditional expectation yields that

$$\mathbb{E}\left[e^{\lambda D_k}|X_1, \cdots, X_{k-1}\right] \le \exp\left(\frac{\lambda^2 (B_k - A_k)^2}{8}\right) \le \exp\left(\frac{\lambda^2 L_k^2}{8}\right) \tag{1.10}$$

Consequently,

$$\begin{aligned} \mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}|X_1, \cdots, X_{n-1}\right]\right] \\ &= \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k} \mathbb{E}\left[e^{\lambda D_n}|X_1, \cdots, X_{n-1}\right]\right] \end{aligned}$$

$$\leq e^{\lambda^2 L_n^2/8}\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n-1}D_k}\right].$$

Thus, iterating this procedure yields $\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n}D_k}\right] \leq e^{\lambda^2\sum_{k=1}^{n}L_k^2/8}$, which means that $\sum_{k=1}^{n}D_k$ is sub-Gaussian with parameter $\nu^2 = \frac{\sum_{k=1}^{n}L_k^2}{4}$, and an application of the former Hoeffding inequality yields the desired tail bound. ■

Remark 1.35 *There are two key ingredients in the above proof: one is the sub-Gaussian type property but for the conditional expectation; the other one is the tensorization property of the moment generating function but for martingale difference sequence.*

Exercise 1.36 *Write out the details for the proof of* (1.10)*.*

Since Azuma-Hoeffding inequality actually shows the concentration of $f(X_1,\cdots,X_n)$ around its mean with the proviso that D_k are bounded, a natural question will be for which f the corresponding D_k are bounded. Next we are going to show that this is the case if f does not fluctuate with each argument too much, leading to the bounded difference inequality, i.e., the McDiarmid inequality. This result reveals a connection between stability and concentration: if a function $f(x_1,\cdots,x_n)$ is not too sensitive to any of its coordinates x_i, then it is anticipated that $f(X_1,\cdots,X_n)$ ($X_i,\ i=1,\cdots,n$ are independent or weakly independent) is close to its mean. This is also the first concentration result in this course that is beyond the sum of independent random variables, as well as a benchmark concentration inequality we will revisit a few times.

Theorem 1.37 (McDiarmid inequality/Bounded difference inequality) *Let $X_k, k=1,\cdots,n$ be independent random variables taking values in $\mathcal{X}$, where $\mathcal{X}$ is the sample space. Suppose that a function $f:\mathcal{X}^n\to\mathbb{R}$ satisfies the bounded difference property*

$$|f(x_1,\cdots,x_{k-1},x_k,x_{k+1},\cdots,x_n) - f(x_1,\cdots,x_{k-1},x_k',x_{k+1},\cdots,x_n)| \leq L_k$$

with parameters $(L_1,\cdots,L_n)$ for all $x_1,\cdots,x_n,x_k'\in\mathcal{X}$. Then

$$\mathbb{P}\left[|f(X)-\mathbb{E}\left[f(X)\right]|\geq t\right] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n}L_k^2}}.$$

Proof: Define D_k as in (1.9). By the last theorem we only need to show D_k is bounded. To this end, define

$$A_k = \inf_{x\in\mathcal{X}}\mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n}\right] - \mathbb{E}_k\left[f(X_1,\cdots,X_{k-1},\underbrace{X_k,X_{k+1},\cdots,X_n}\right]$$

and

$$B_k = \sup_{x\in\mathcal{X}}\mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n}\right] - \mathbb{E}_k\left[f(X_1,\cdots,X_{k-1},\underbrace{X_k,X_{k+1},\cdots,X_n}\right].$$

It is clear that $A_k\leq D_k\leq B_k$ almost surely. Moreover, we have

$$B_k - A_k = \sup_{x\in\mathcal{X}}\mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n}\right] - \inf_{x\in\mathcal{X}}\mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n}\right]$$

$$
\begin{aligned}
&\leq \sup_{x,y\in\mathcal{X}} \left| \mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n}) \right] - \mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},y,\underbrace{X_{k+1},\cdots,X_n}) \right] \right| \\
&= \sup_{x,y\in\mathcal{X}} \left| \mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n}) - f(X_1,\cdots,X_{k-1},y,\underbrace{X_{k+1},\cdots,X_n}) \right] \right| \\
&\leq L_k,
\end{aligned}
$$

as desired. ■

Exercise 1.38 *Show how to prove the result in Example 1.17 using the McDiarmid inequality.*

Example 1.39 (Rademacher complexity) *Let $\{\varepsilon_k\}_{k=1}^n$ be an i.i.d sequence of Rademacher variables, namely*

$$
\mathbb{P}\left[\varepsilon_k = 1\right] = \mathbb{P}\left[\varepsilon_k = -1\right] = \frac{1}{2},
$$

and let $\varepsilon = (\varepsilon_1,\cdots,\varepsilon_n)$. Given a subset $\mathcal{A}$ of $\mathbb{R}^n$, define the random variable

$$
Z = \sup_{a\in\mathcal{A}} \left[\sum_{k=1}^n a_k\varepsilon_k \right] = \sup_{a\in\mathcal{A}} \left[\langle a, \varepsilon\rangle \right].
$$

The Rademacher complexity, denoted $\mathcal{R}_n(\mathcal{A})$, is defined as the expectation of Z,

$$
\mathcal{R}_n(\mathcal{A}) = \mathbb{E}\left[Z\right].
$$

Here the random variable Z and its expectation measures the size of $\mathcal{A}$ based on the Rademacher sequence. Roughly speaking, it measures the "diameter" of the set in different directions randomly and then computes the average. (Again, when it is not clear how to do, do randomly). They also reflect how strong the set $\mathcal{A}$ looks like a random set defined by the Rademacher sequence. For example, if $\mathcal{A} = \{1,-1\}^n$, then it is equal to ε in certain sense.

We want to show that the McDiarmid inequality can be used to establish the concentration of Z. Define

$$
f(x_1,\cdots,x_n) = \sup_{a\in\mathcal{A}} \left[\sum_{k=1}^n a_k x_k \right], \quad x_k \in \{1,-1\}.
$$

it suffices to show that f satisfies the bounded difference property. To this end, we have

$$
\begin{aligned}
&f(x_1,\cdots,x_{k-1},x_k,x_{k+1},x_n) - f(x_1,\cdots,x_{k-1},x_k',x_{k+1},x_n) \\
&= \sup_{a\in\mathcal{A}} \left[\sum_{k=1}^n a_k x_k \right] - \sup_{a\in\mathcal{A}} \left[\sum_{j=1}^{k-1} a_j x_j + a_k x_k' + \sum_{j=k+1}^n a_j x_j \right] \\
&\leq \sup_{a\in\mathcal{A}} \left[\left(\sum_{k=1}^n a_k x_k \right) - \left(\sum_{j=1}^{k-1} a_j x_j + a_k x_k' + \sum_{j=k+1}^n a_j x_j \right) \right] \\
&= \sup_{a\in\mathcal{A}} a_k (x_k - x_k')
\end{aligned}
$$

$$\leq 2 \sup_{a \in \mathcal{A}} |a_k|,$$

where the last line follows from the fact $x_k,\ x_k' \in \{1, -1\}$. *Similarly, we have*

$$f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, x_n) - f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, x_n) \leq 2 \sup_{a \in \mathcal{A}} |a_k|.$$

Consequently,

$$|f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, x_n) - f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, x_n)| \leq 2 \sup_{a \in \mathcal{A}} |a_k|.$$

Thus, by the McDiarmid inequality we can see that Z *is sub-Gaussian with parameter* $\nu^2 = \sum_{k=1}^n \sup_{a \in \mathcal{A}} |a_k|^2$. *Later, we will show that this parameter can be sharpened to* $\sup_{a \in \mathcal{A}} \sum_{k=1}^n |a_k|^2$. ***To some extend, this has motivated the development of other machinaries for establishing the concentration inequality.***

Example 1.40 *Let* $X_k,\ k = 1, \cdots, n$ *be bounded random vectors in* $\mathbb{R}^d$ *satisfying* $\mathbb{E}[X_k] = 0$ *and* $\|X_k\|_2 \leq B$. *We want to study the concentration of*

$$\left\| \frac{1}{n} \sum_{k=1}^n X_k \right\|_2$$

around the mean $\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2\right]$. *Let* $f(x_1, \cdots, x_n) = \left\|\frac{1}{n}\sum_{k=1}^n x_k\right\|_2$, *where* $x_k \in \mathbb{R}^n$. *Then, by triangular inequality*

$$|f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n)| \leq \frac{1}{n}\|x_k - x_k'\|_2 \leq \frac{2B}{n}.$$

Thus, the application of the bounded difference inequality yields that

$$\mathbb{P}\left[\left|\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2 - \mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2\right]\right| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{2B^2}\right).$$

If we further assume $\mathbb{E}\left[\|X_k\|_2^2\right] \leq \sigma^2$. *Then*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2\right] \leq \left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2^2\right]\right)^{1/2} = \left(\frac{1}{n^2}\sum_{k=1}^n \mathbb{E}\left[\|X_k\|_2^2\right]\right)^{1/2} \leq \frac{\sigma}{\sqrt{n}}.$$

Consequently, we have

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2 \geq \frac{\sigma}{\sqrt{n}} + t\right] \leq 2\exp\left(-\frac{nt^2}{2B^2}\right).$$

Example 1.41 *As pointed out in a motivation example, the analysis of the generalization error in empirical risk minimization eventually boils down to the bound of quantity in the form of*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}[f(X)] \right|,$$

where $X_k, k = 1, \cdots, n$ are random vectors. Suppose $|f|_\infty < B$ for all $f \in \mathcal{F}$. Letting

$$g(x_1, \cdots, x_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(x_k) - \mathbb{E}\left[f(X)\right] \right|,$$

we have

$$\left| g(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - g(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n) \right| \leq \sup_{f \in \mathcal{F}} \left| \frac{1}{n} f(x_k) - \frac{1}{n} f(x_k') \right| \leq \frac{2B}{n}.$$

Then the application of the bounded difference inequality yields

$$\mathbb{P}\left[\left| \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(x_k) - \mathbb{E}\left[f(X)\right] \right| - \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(x_k) - \mathbb{E}\left[f(X)\right] \right| \right] \right| \geq t \right] \leq 2\exp\left(-\frac{nt^2}{2B^2} \right)$$

Remark 1.42 *Note that in Examples 1.39 and 1.41, we still need to compute the mean of the random quantity of interest, which will be another focus of the course.*

Remark 1.43 *The bounded difference inequality is very useful and the next two lectures are essentially about generalizing the bounded different inequality by considering different f and $(X_1, \cdots, X_n)$.*

Reading Materials

[1] Martin J. Wainwright, *High-dimensional statistics – A non-asymptotic viewpoint*, Chapters 2.1 and 2.2.

[2] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, Chapters 2.5, 2.6, 2.7 and 2.8.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 2: Herbst Argument and Entropy Method

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/03/13)*

Recap and Motivation: In Lecture 1 we have discussed the sub-Gaussian and sub-exponential distributions and the corresponding tail bounds for sums of independent random variables and functions satisfying the bounded different property. Our next goal is to extend the concentration results to other interesting functions. We will restrict our attention to the sub-Gaussian type tails while some of the techniques may also be applicable for establishing the Bernstein type bound.

Define the log-moment generating function of a random variable X as

$$\psi(\lambda) = \log \mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}[X])\right)\right]. \tag{2.1}$$

The sub-Gaussian property (see Definition 1.9 of Lecture 1) can be equivalently expressed as

$$\psi(\lambda) \lesssim \lambda^2\nu^2 \quad \text{for all} \quad \lambda \in \mathbb{R}. \tag{2.2}$$

By the Chernoff bound we know that the sub-Gaussian property immediately implies a Gaussian tail bound (they are indeed equivalent). Moreover, the sub-Gaussian property can be established for sums of independent random variables and functions obeying the bounded difference inequality. As already seen, the proofs rely essentially on the tensorization property (or a martingale difference sequence variant) of the log-moment generating function defining the sub-Gaussian property, i.e.,

$$\log \mathbb{E}\left[\exp\left(\lambda \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k])\right)\right] \le \sum_{k=1}^{n} \log \mathbb{E}\left[\exp\left(\lambda\left(X_k - \mathbb{E}[X_k]\right)\right)\right].$$

However, for more complicated functions $f(X_1, \cdots, X_n)$ arising from the applications than sums of independent random variables, the above tensorization property hardly holds. That is, the sub-Gaussian property in terms of the (log-)moment generating function overall does not tensorize well. To mitigate this issue, one idea is to introduce an alternative formulation of the sub-Gaussian property that behaves well under tensorization.

In this lecture we will study the sub-Gaussian property based on certain entropy function and establish a concentration inequalities for more general f. To motivate this, let us recap the calculus method that is used in the proof of the sub-Gaussian property for bounded random variables (see Example 1.13 of Lecture 1). First, a simple calculation yields that

$$\psi(0) = 0 \quad \text{and} \quad \psi'(0) = 0.$$

Thus in order to establish the sub-Gaussian property (2.2), it suffices to show that

$$\psi''(\lambda) \lesssim \nu^2 \quad \text{for all} \quad \lambda \in \mathbb{R}.$$

Noting that (2.2) is equivalent to

$$\psi(\lambda)/\lambda \lesssim \lambda\nu^2 \quad \text{for all} \quad \lambda \in \mathbb{R},$$

it also suffices to show that

$$\frac{d}{d\lambda}\left(\psi(\lambda)/\lambda\right) \lesssim \nu^2. \tag{2.3}$$

Though this is a trivial reformulation, it will lead to a more powerful method for proving concentration inequalities. Moreover, it turns out that (2.3) can be related to a type of entropy function that tensorizes well.

Agenda:

- Herbst argument and Tensorization
- Modified log-Sobolev inequality and Entropy method
- Gaussian concentration

2.1 Herbst Argument and Tensorization

Definition 2.1 *The entropy of a* **nonnegative** *random variable Z, denoted* $\mathrm{Ent}\left[Z\right]$*, is defined as*

$$\mathrm{Ent}\left[Z\right] = \mathbb{E}\left[Z\log\left(Z\right)\right] - \mathbb{E}\left[Z\right]\log\left(\mathbb{E}\left[Z\right]\right)$$

Remark 2.2 *Note that the entropy defined here should not be confused with the Shannon entropy which is roughly about on average how many bits are needed to store a random variable. In general, entropy can be defined for any convex function ϕ : $\mathrm{Ent}_\phi\left[Z\right] = \mathbb{E}\left[\phi(Z)\right] - \phi(\mathbb{E}\left[Z\right])$, which reflects to how extend Z behaves like its $\mathbb{E}\left[Z\right]$ based on a convex function.*

Exercise 2.3 *Show that* $\mathrm{Ent}\left[Z\right] \geq 0$.

Example 2.4 (Entropy of MGF of Gaussian) *Let $X \sim \mathcal{N}(0, \sigma^2)$. We have*

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda X}\right] &= \mathbb{E}\left[e^{\lambda X}\log\left(e^{\lambda X}\right)\right] - \mathbb{E}\left[e^{\lambda X}\right]\log\left(\mathbb{E}\left[e^{\lambda X}\right]\right) \\
&= \mathbb{E}\left[\lambda X e^{\lambda X}\right] - \mathbb{E}\left[e^{\lambda X}\right]\log\left(e^{\frac{\lambda^2\sigma^2}{2}}\right) \\
&= \frac{1}{2}\lambda^2\sigma^2\mathbb{E}\left[e^{\lambda X}\right] \quad \textit{for all} \quad \lambda \in \mathbb{R},
\end{aligned}$$

where we can use $d\mathbb{E}\left[e^{\lambda X}\right]/d\lambda = \mathbb{E}\left[Xe^{\lambda X}\right]$ to calculate the first term in the second line.

Lemma 2.5 (A useful identity) *Let $\psi(\lambda)$ be the log-moment generating function defined in* (2.1)*. We have*

$$\frac{\mathrm{Ent}\left[e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} = \lambda\psi'(\lambda) - \psi(\lambda).$$

Proof: The result follows from the definition and $\mathbb{E}\left[Xe^{\lambda X}\right] = \frac{d}{d\lambda}\mathbb{E}\left[e^{\lambda X}\right]$. Note that X is not necessarily mean zero though it is centered when defining $\psi(\lambda)$. ■

Example 2.6 (Entropy of MGF of bounded random variable) *Let X be mean zero and supported on $[a,b]$. By Lemma 2.5, we have*

$$\begin{aligned}\frac{\text{Ent}\left[e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} &= \lambda\psi'(\lambda) - \psi(\lambda) = [\lambda\psi'(\lambda) - \psi(\lambda)] - [0\cdot\psi'(0) - \psi(0)] \\ &= \int_0^\lambda \xi\psi''(\xi)d\xi \\ &\leq \frac{(b-a)^2}{4}\int_0^\lambda \xi d\xi \\ &= \frac{\lambda^2(b-a)^2}{8},\end{aligned}$$

where the inequality follows from the bound for $\psi''(\xi)$, see Example 1.13 of Lecture 1. Thus,

$$\text{Ent}\left[e^{\lambda X}\right] \leq \frac{\lambda^2(b-a)^2}{8}\mathbb{E}\left[e^{\lambda X}\right].$$

2.1.1 Herbst Argument

The last two examples reveal a connection between $\text{Ent}\left[e^{\lambda X}\right]$ and $\mathbb{E}\left[e^{\lambda X}\right]$ for certain sub-Gaussian random variables. It turns out this relation can be used to define the sub-Gaussian property, which follows from the Herbst argument.

Theorem 2.7 (Herbst) *Suppose that*

$$\text{Ent}\left[e^{\lambda X}\right] \leq \frac{\lambda^2\nu^2}{2}\mathbb{E}\left[e^{\lambda X}\right] \quad \textit{for all} \quad \lambda \geq 0. \tag{2.4}$$

Then X satisfies the bound

$$\psi(\lambda) = \log\mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}[X])\right)\right] \leq \frac{1}{2}\lambda^2\nu^2 \quad \textit{for all} \quad \lambda \geq 0. \tag{2.5}$$

Proof: The proof is indeed based on the argument sketched around (2.3). First note that

$$\lim_{\lambda\to 0}\frac{\psi(\lambda)}{\lambda} = 0. \qquad (\textbf{check this!})$$

Consequently,

$$\frac{\psi(\lambda)}{\lambda} = \int_0^\lambda \frac{d}{d\xi}\left(\frac{\psi(\xi)}{\xi}\right)d\xi.$$

Moreover, condition (2.4) can be used to provide an upper bound for $\frac{d}{d\xi}\left(\frac{\psi(\xi)}{\xi}\right)$ since there holds

$$\frac{d}{d\xi}\left(\frac{\psi(\xi)}{\xi}\right) = \frac{1}{\xi^2}\left(\xi\psi'(\xi) - \psi(\xi)\right) = \frac{1}{\xi^2}\frac{\text{Ent}\left[e^{\xi X}\right]}{\mathbb{E}\left[e^{\xi X}\right]} \leq \frac{1}{\xi^2}\frac{\xi^2\nu^2}{2} = \frac{\nu^2}{2} \quad \text{for all} \quad \xi \geq 0,$$

where the second equality follows from Lemma 2.5 and the inequality follows from (2.4). Inserting this bound into the integral yields that

$$\frac{\psi(\lambda)}{\lambda} \leq \frac{\lambda\nu^2}{2} \Rightarrow \psi(\lambda) \leq \frac{\lambda^2\nu^2}{2},$$

as claimed. ■

Exercise 2.8 *Show that* $\text{Ent}\left[e^{\lambda(X+c)}\right] = e^{\lambda c} \cdot \text{Ent}\left[e^{\lambda X}\right]$ *for any* $c \in \mathbb{R}$. *Note this implies if* X *satisfies* (2.4)*, so does* $X + c$. *That is why we do not need to center the random variable in* (2.4)*, but are still able to obtain a result for a centered random variable in* (2.5). *Indeed, the random variables we are interested are in the form of* $f(X_1, \cdots, X_n)$ *which are generally not mean zero.*

The following proposition follows immediately from Theorem 2.7, showing the sub-Gaussian property can be defined based on the relation between $\text{Ent}\left[e^{\lambda X}\right]$ and $\mathbb{E}\left[e^{\lambda X}\right]$.

Proposition 2.9 (sub-Gaussian property via Entropy) *Suppose*

$$\text{Ent}\left[e^{\lambda X}\right] \leq \frac{\lambda^2 \nu^2}{2} \mathbb{E}\left[e^{\lambda X}\right] \quad \textit{for all} \quad \lambda \in \mathbb{R}. \tag{2.6}$$

Then, X *is sub-Gassuain with parameter* ν.

Exercise 2.10 *Prove Proposition 2.9. (**Hint:** apply Theorem 2.7 to* $-X$ *and* $-\lambda$ *in the case when* (2.6) *holds for* $\lambda \leq 0$.*)*

2.1.2 Tensorization of Entropy

The entropy has a nice tensorization property for functions of independent variables which enables us to bound the entropy of random variables in the form of $g(X_1, \cdots, X_n)$ in a coordinate-wise manner. To present this property, let us first introduce a new notation. Let $X_1, \cdots, X_n$ be independent random variables. Given $g : \mathcal{X}^n \to [0, \infty)$, we define $\text{Ent}_k\left[g(X_1, \cdots, X_n)\right]$ as

$$\text{Ent}_k\left[g(X_1, \cdots, X_n)\right] = \text{Ent}\left[g(x_1, \cdots, x_{k-1}, X_k, x_{k+1}, \cdots, x_n)\right]|_{x_1 = X_1, \cdots, x_{k-1} = X_{k-1}, x_{k+1} = X_{k+1}, \cdots, x_n = X_n}.$$

In other words, $\text{Ent}_k\left[g(X_1, \cdots, X_n)\right]$ is the entropy of $g(X_1, \cdots, X_n)$ with respect to the variable to X_k only, while the others keep fixed. Note that $\text{Ent}_k\left[g(X_1, \cdots, X_n)\right]$ is still a *random variable*, a function of $X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n$.

Theorem 2.11 (Tensorization of entropy) *We have*

$$\text{Ent}\left[g(X_1, \cdots, X_n)\right] \leq \mathbb{E}\left[\sum_{k=1}^{n} \text{Ent}_k\left[g(X_1, \cdots, X_n)\right]\right],$$

where $X_1, \cdots, X_n$ *are independent.*

The tensorization property of entropy is an exact analogue of the tensorization property of the variance, see for example Chapter 2.1 of [2]. This property allows us to deduce a bound for functions of independent random variables from bounds for functions of each individual random variable, thus is very helpful for studying high dimensional problems. If we think $\text{Ent}_k[\cdot]$ as the way of quantifying the random in the k-th mode (when average out the other random variables), the theorem implies that the randomness of the joint distribution is less than or equal to the sum of the randomness of all the modes. **Compared to MGF, this property implies that for any form of g, we can control the entropy of $g(X_1, \cdots, X_n)$ by considering the entropy of each coordinate, which means entropy tensorizes better than MGF.** The proof of this theorem is based on the following variational form of entropy[1].

[1]Expressing a quantity in terms of a variational form is very widely used in maths. Typical examples include variational forms for norms, conjugate duality in optimization. We will see more of this technique later.

Lemma 2.12 (Variational formula of entropy) *Let $Z \geq 0$ be a nonnegative random variable. Then,*

$$\begin{aligned}\mathrm{Ent}\left[Z\right] &= \sup\left\{\mathbb{E}\left[ZX\right]: \ X \text{ is a random variable satisfying } \mathbb{E}\left[e^X\right] = 1\right\} \\ &= \sup\{\mathbb{E}\left[Z\left(\log Y - \log \mathbb{E}\left[Y\right]\right): \ Y \geq 0\right]\}.\end{aligned}$$

Proof: Note if letting $X = \log\left(Z/\mathbb{E}\left[Z\right]\right)$, then it is not hard to show that $\mathbb{E}\left[e^X\right] = 1$ and $\mathbb{E}\left[ZX\right] = \mathrm{Ent}\left[Z\right]$. Thus, it suffices to show that

$$\mathrm{Ent}\left[Z\right] - \mathbb{E}\left[ZX\right] \geq 0$$

for all X satisfying $\mathbb{E}\left[e^X\right] = 1$. Note that $\mathrm{Ent}\left[Z\right] - \mathbb{E}\left[ZX\right]$ can be expressed as

$$\mathrm{Ent}\left[Z\right] - \mathbb{E}\left[ZX\right] = \mathbb{E}\left[\left(e^{-X}Z\log\left(e^{-X}Z\right)\right)e^X\right] - \mathbb{E}\left[\left(e^{-X}Z\right)e^X\right]\log\mathbb{E}\left[\left(e^{-X}Z\right)e^X\right].$$

Since $\mathbb{E}\left[e^X\right] = 1$, if we define the new probability $d\mathbb{Q} = e^X d\mathbb{P}$ where $\mathbb{P}$ is the probability distribution defining Z and X, then $\mathrm{Ent}\left[Z\right] - \mathbb{E}\left[ZX\right]$ is indeed the entropy of $e^{-X}Z$ under the probability distribution, i.e.,

$$\mathrm{Ent}\left[Z\right] - \mathbb{E}\left[ZX\right] = \mathrm{Ent}_{\mathbb{Q}}\left[e^{-X}Z\right] \geq 0, \tag{2.7}$$

where the inequality follows from the nonnegative property of entropy.

The second equality follows simply from the fact that

$$\mathbb{E}\left[e^X\right] = 1 \Leftrightarrow \exists\ Y \geq 0 \text{ such that } X = \log Y - \log\mathbb{E}\left[Y\right].$$

The proof is complete now. ∎

Exercise 2.13 *If you are not happy with the $d\mathbb{Q} = e^X d\mathbb{P}$ argument in (2.7). Try to show it directly by following a similar argument for Jensen's inequality. That is, show that*

$$\mathbb{E}\left[f(Z)e^X\right] \geq f(\mathbb{E}\left[Ze^X\right])$$

hold for any convex function f and random variable X satisfying $\mathbb{E}\left[e^X\right] = 1$.

Proof: [of Theorem 2.11] Let $Z = g(X_1, \cdots, X_n)$ and define

$$U_k = \log\mathbb{E}\left[Z|X_1, \cdots, X_k\right] - \log\mathbb{E}\left[Z|X_1, \cdots, X_{k-1}\right].$$

Then we have

$$\mathrm{Ent}\left[Z\right] = \mathbb{E}\left[Z\left(\log\left(Z\right) - \log\mathbb{E}\left[Z\right]\right)\right] = \sum_{k=1}^{n}\mathbb{E}\left[ZU_k\right].$$

Thus, it suffices to show that $\mathbb{E}\left[ZU_k|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n\right] \leq \mathrm{Ent}_k\left[Z\right]$. To this end, let us fix

$$X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n$$

and consider

$$ZU_k = Z\left(\log\mathbb{E}\left[Z|X_1, \cdots, X_k\right] - \log\mathbb{E}\left[Z|X_1, \cdots, X_{k-1}\right]\right)$$

as a function of X_k. Noting that $\mathbb{E}_{X_k}\left[\mathbb{E}\left[Z|X_1, \cdots, X_k\right]\right] = \mathbb{E}\left[Z|X_1, \cdots, X_{k-1}\right]$ due to the independence of all the X_k, the application of Lemma 2.12 with respect to X_k immediately that

$$\mathbb{E}\left[ZU_k|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n\right] \leq \mathrm{Ent}_k\left[Z\right],$$

which completes the proof. ∎

Exercise 2.14 *Verify that the equality in the tensorization property holds for*

$$g(X_1, \cdots, X_n) = \exp\left(\lambda \sum_{k=1}^{n} X_k\right),$$

where $X_1, \cdots, X_n$ are independent.

Example 2.15 (Bounded difference inequality revisited) *In this example, we are going to show that the bounded difference inequality can be proved in an alternative way based on the Herbst argument (Theorem 2.7) and the tensorization property (Theorem 2.11). Recall that a function f satisfies the bounded difference property if*

$$|f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n)| \leq L_k$$

with parameters $(L_1, \cdots, L_n)$ over the range of the independent random variables $X = (X_1, \cdots, X_n)$. Thus, when fixing $X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n$, $f(X_1, \cdots, X_n)$ can be viewed as a bounded random variable which locates in an interval of length at most L_k. Then it follows from Example 2.6 that

$$\mathrm{Ent}_k\left[e^{\lambda f(X_1, \cdots, X_n)}\right] \leq \frac{L_k^2}{8} \mathbb{E}_k\left[e^{\lambda f(X_1, \cdots, X_n)}\right],$$

where $\mathbb{E}_k[\cdot]$ means taking expectation with respect to X_k only. Furthermore, letting $g(X_1, \cdots, X_n) = e^{\lambda f(X_1, \cdots, X_n)}$, the tensorization property implies that

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda f(X_1, \cdots, X_n)}\right] &\leq \sum_{k=1}^{n} \frac{L_k^2}{8} \mathbb{E}\left[\mathbb{E}_k\left[e^{\lambda f(X_1, \cdots, X_n)}\right]\right] \\
&= \left(\sum_{k=1}^{n} \frac{L_k^2}{8}\right) \mathbb{E}\left[e^{\lambda f(X_1, \cdots, X_n)}\right].
\end{aligned}$$

Thus, by the Herbst argument, we know that $f(X_1, \cdots, X_n)$ is sub-Gaussian with parameter $\nu^2 = \frac{\sum_{k=1}^{n} L_k^2}{4}$ and the tail bound in the bounded difference inequality follows immediately.

2.2 Modified Log-Sobolev Inequality and Entropy Method

As demonstrated in the last example, in order to apply the Herbst argument and the tensorization property to establish the sub-Gaussian tail, it remains to bound $\mathrm{Ent}_k\left[e^{\lambda f(X_1, \cdots, X_n)}\right]$. In a more general setting, this can be achieved by the modified log-Sobolev inequality (MLS), which is the last piece of the entropy method. In a nutshell, MLS controls the entropy of $e^{\lambda f(X)}$ by the fluctuation/gradient of the function f.

Lemma 2.16 (MLS) *Let $f : \mathcal{X} \to \mathbb{R}$ be a single variable function. Define*

$$D^- f(x) = f(x) - \inf_{z \in \mathcal{X}} f(z).$$

Then for any $\lambda \geq 0$ we have

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda f(X)}\right] &\leq \mathbb{E}\left[\phi\left(\lambda D^- f(X)\right) e^{\lambda f(X)}\right] \\
&\leq \frac{1}{2} \mathbb{E}\left[|\lambda D^- f(X)|^2 e^{\lambda f(X)}\right].
\end{aligned}$$

where $\phi(x) = e^{-x} + x - 1$.

Proof: First by basic calculus, it is not hard to show that for any nonnegative random variable Z there holds

$$\mathrm{Ent}\left[Z\right] = \inf_{t>0} \mathbb{E}\left[Z\log Z - Z\log t - Z + t\right].$$

Thus, letting $Z = e^{\lambda f(X)}$ yields that

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda f(X)}\right] &= \inf_{t>0} \mathbb{E}\left[\lambda f(X)e^{\lambda f(X)} - e^{\lambda f(X)}\log t - e^{\lambda f(X)} + t\right] \\
&\le \mathbb{E}\left[\lambda f(X)e^{\lambda f(X)} - e^{\lambda f(X)}\log\left(e^{\lambda \inf_z f(z)}\right) - e^{\lambda f(X)} + e^{\lambda \inf_z f(z)}\right] \\
&= \mathbb{E}\left[\left\{\lambda f(X) - \lambda \inf_z f(z) - 1 + e^{-\lambda f(X) + \lambda \inf_z f(z)}\right\} e^{\lambda f(X)}\right] \\
&= \mathbb{E}\left[\phi\left(\lambda D^- f(X)\right) e^{\lambda f(X)}\right].
\end{aligned}$$

The second inequality in the lemma simply follows from the fact $\phi(x) \le \frac{1}{2}x^2$ for $x \ge 0$. ■

When $f : \mathcal{X}^n \to \mathbb{R}$ is a multivariable function, applying the MLS conditionally to each $\mathrm{Ent}_k\left[e^{\lambda f(X_1,\cdots,X_n)}\right]$ leads to the following theorem which can be viewed as an generalization of the bounded difference inequality.

Theorem 2.17 (General bounded difference inequality) *Let $x = (x_1, \cdots, x_n)$ and define*

$$\begin{aligned}
D_k^- f(x) &= f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - \inf_{z\in\mathcal{X}} f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n), \\
D_k^+ f(x) &= \sup_{z\in\mathcal{X}} f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n).
\end{aligned}$$

Let $X_1, \cdots, X_n$ be i.i.d random variables. If $\sum_{k=1}^n |D_k^- f(x)|^2 \le \nu_1^2$, then we have

$$\mathbb{P}\left[f(X_1, \cdots, X_n) \ge \mathbb{E}\left[f(X_1, \cdots, X_n)\right] + t\right] \le \exp\left(-\frac{t^2}{2\nu_1^2}\right)$$

Similarly, if $\sum_{k=1}^n |D_k^+ f(x)|^2 \le \nu_2^2$, we have

$$\mathbb{P}\left[f(X_1, \cdots, X_n) \le \mathbb{E}\left[f(X_1, \cdots, X_n)\right] - t\right] \le \exp\left(-\frac{t^2}{2\nu_2^2}\right).$$

Proof: If we fix $X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n$ and consider $f(X_1, \cdots, X_n)$ as a function of X_k, the application of the MLS yields that

$$\mathrm{Ent}_k\left[e^{\lambda f(X_1,\cdots,X_n)}\right] \le \frac{1}{2}\mathbb{E}_k\left[|\lambda D_k^- f(X_1, \cdots, X_n)|^2 e^{\lambda f(X_1,\cdots,X_n)}\right], \quad \lambda \ge 0.$$

By the tensorization property of entropy, we have

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda f(X_1,\cdots,X_n)}\right] &\le \sum_{k=1}^n \mathbb{E}\left[\mathrm{Ent}_k\left[e^{\lambda f(X_1,\cdots,X_n)}\right]\right] \\
&\le \frac{1}{2}\mathbb{E}\left[\lambda^2\left(\sum_{k=1}^n |D_k^- f(X_1, \cdots, X_n)|^2\right) e^{\lambda f(X_1,\cdots,X_n)}\right]
\end{aligned}$$

$$\leq \frac{\lambda^2 \nu_1^2}{2} \mathbb{E}\left[e^{\lambda f(X_1,\cdots,X_n)}\right].$$

The upper tail bound follows immediately from the Herbst argument and the Chernoff method.

The lower tail bound can be established by considering $-f$. ∎

Remark 2.18 *Let*

$$D_k f(x) = \sup_z f(x_1,\cdots,x_{k-1},z,x_{k+1},\cdots,x_n) - \inf_z f(x_1,\cdots,x_{k-1},z,x_{k+1},\cdots,x_n).$$

Note the tail bound obtained by the bounded difference inequality (Theorem 1.41 of Lecture 1) is of the order $\exp\left(-t^2/\sum_{k=1}^n \|D_k f(x)\|_\infty^2\right)$*, while the tail bound established here is of the order* $\exp\left(-t^2/\|\sum_{k=1}^n |D_k f(x)|^2\|_\infty\right)$*. It is trivial that* $\left\|\sum_{k=1}^n |D_k f(x)|^2\right\|_\infty \leq \sum_{k=1}^n \|D_k f(x)\|_\infty^2$*. Moreover, there are cases* $\left\|\sum_{k=1}^n |D_k f(x)|^2\right\|_\infty$ *can be sufficiently smaller than* $\sum_{k=1}^n \|D_k f(x)\|_\infty^2$*. Thus, the bounds of Theorem 2.17 are an improvement over that in the bounded difference inequality. This is due to the fact entropy function tensorizes better than the moment generating function.*

Note that the upper and lower tail bounds here are essentially asymmetric: the upper bound is controlled by $\sum_{k=1}^n |D_k^- f(X)|^2$ *while the lower bound is controlled by* $\sum_{k=1}^n |D_k^+ f(X)|^2$*. There are problems where it is may be not clear how to bound one of them. However, if the function* f *satisfies a stronger condition, it is still possible to obtain a two-sided tail bound from the single bound of* $\sum_{k=1}^n |D_k^- f(X)|^2$*. The machinery needed to prove such bounds are discussed in the next lecture.*

Example 2.19 *Let* $\mathcal{A} \subset \mathbb{R}^{n\times n}$ *be a set of symmetric matrices whose entries are* $\{1,-1\}$*. For any* $A \in \mathcal{A}$*, we let* $\lambda_{\max}(A)$ *denote the largest eigenvalue of* A*. By variational formula of the largest eigenvalue of symmetric matrix, we know that*

$$\lambda_{\max}(A) = \sup_{\|v\|_2=1} \langle v, Av\rangle. \tag{2.8}$$

Let $\mathcal{A}' \subset \mathcal{A}$ *be the subset of matrices which can only differ from* A *in the* (i,j)*-th and* (j,i)*-th entries and let* A^- *be a matrix such that*

$$\lambda_{\max}(A^-) = \inf_{A'\in\mathcal{A}',\ A'_{ij}=A'_{ji}\in\{1,-1\}} \lambda_{\max}(A').$$

Letting v_A *be the unit vector where the equality in* (2.8) *is achieved, then we have*

$$\begin{aligned} D_{ij}^- \lambda_{\max}(A) &= \langle v_A, Av_A\rangle - \max_{\|v\|_2} \left\langle v, A^- v\right\rangle \\ &\leq \langle v_A, Av_A\rangle - \left\langle v_A, A^- v_A\right\rangle \\ &= \left\langle v_A, (A-A^-)v_A\right\rangle \\ &\leq 4|v_A(i)||v_A(j)|. \end{aligned} \tag{2.9}$$

Note the above bound only relies on A*. Thus, it follows that*

$$\begin{aligned} \sum_{1\leq i\leq j\leq n} |D_{ij}^- \lambda_{\max}(A)|^2 &\leq 16 \sum_{i,j=1}^n |v_A(i)|^2 |v_A(j)|^2 \\ &= 16 \left(\sum_{i=1}^n |v_A(i)|^2|\right) \left(\sum_{i=1}^n |v_A(j)|^2|\right) \end{aligned}$$

$$= 16. \tag{2.10}$$

Thus, if A is a random symmetric matrix with independent entries (upper triangular part) taking the value from $\{1, -1\}$ randomly, then by the general bounded difference inequality we can establish the following upper tail

$$\mathbb{P}\left[\lambda_{\max}(A) \geq \mathbb{E}\left[\lambda_{\max}(A)\right] + t\right] \leq e^{-t^2/32}.$$

Of course we still need to estimate the mean of $\lambda_{\max}(A)$. This will be discussed in the future by studying the mean of the supremum of an empirical process.

It is worth noting that it seems we cannot use the general bounded difference inequality to establish a dimension free lower bound. To see this, first note that we can establish a similar bound to (2.9) for $D_{ij}^{+}\lambda_{\max}(A)$, but with the principal eigenvector of A being replaced by that of principal eigenvector of a A^{+} which only differs from A in the the (i,j)-th and (j,i)-th entries and in the meantime achieves the maximum principal eigenvalue, i.e.,

$$D_{ij}^{+}\lambda_{\max}(A) \leq 4|v_{A^{+}}(i)||v_{A^{+}}(j)|.$$

Noting that $v_{A^{+}}$ varies for different (i,j), we cannot proceed in the similar fashion as in (2.10). As can seen from the next lecture, we can establish the lower tail by a different technique.

From the general bounded difference inequality we can obtain the following proposition which is relatively easier to manage. Recall that a function $f : \mathcal{X}^n \to \mathbb{R}$ is *separately convex* if for each $i = 1, \cdots, n$, it is a convex function of its i-th coordinate when the rest of the coordinates are fixed.

Proposition 2.20 *Let $X_k, k = 1, \cdots, n$ be independent random variables taking values in an interval $[a,b]$ and let $f : [a,b]^n \to \mathbb{R}$ be a separately convex function which also satisfies the Lipschitz condition $|f(x) - f(y)| \leq L\|x - y\|_2$ for all $x, y \in [a,b]^n$. Then, for all $t \geq 0$,*

$$\mathbb{P}\left[f(X_1, \cdots, X_n) \geq \mathbb{E}\left[f(X_1, \cdots, X_n)\right] + t\right] \leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right)$$

Proof: For ease of presentation, we assume the partial derivatives of f exist (Otherwise we can adopt a standard approximation argument). Letting x_k' be the random variable at which

$$\inf_z f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n)$$

is achieved. Then it follows from the separately convex property of f that

$$\begin{aligned} D_k^{-} f(x) &= f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - \inf_z f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n) \\ &= f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n) \\ &\leq \partial_k f(x)(x_k - x_k'). \qquad \text{(use the separately convexity here)} \end{aligned}$$

It follows that

$$\sum_{k=1}^{n} |D_k^{-} f(x)|^2 \leq \sum_{k=1}^{n} \partial_k^2 f(x)(x_k - x_k')^2$$

$$\leq \sum_{k=1}^{n} \partial_k^2 f(x)(b-a)^2$$
$$= \|\nabla f(x)\|_2^2 (b-a)^2 \leq L^2 (b-a)^2,$$

where $\|\nabla f(x)\|_2 \leq L$ follows from the Lipschitz condition of f (**check this!**). The upper tail bound then follows immediately from Theorem 2.17. ■

Remark 2.21 *The lower tail cannot be established by considering $-f$ since it is a concave function.*

Example 2.22 (Sharper upper bounds on Rademacher complexity) *Let us revisit Example 1.43 of Lecture 1, which is about establishing an upper tail bound for*

$$Z = \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^{n} a_k \varepsilon_k \right],$$

where $\varepsilon_k, k = 1, \cdots, n$ are i.i.d Rademacher variables. Let

$$f(x_1, \cdots, x_n) = \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^{n} a_k x_k \right], \quad x_k \in \{1, -1\}.$$

*Since f is a supremum of a collection of linear function, it is a convex function and hence separately convex. Moreover, it is not hard to show that (**check this!**)*

$$|f(x_1, \cdots, x_n) - f(x_1', \cdots, x_n')| \leq \sup_{a \in \mathcal{A}} \|a\|_2 \|x - x'\|_2,$$

where $x = (x_1, \cdots, x_n)$ and $x' = (x_1', \cdots, x_n')$ That is, f is Lipschitz with parameter $\sup_{a \in \mathcal{A}} \|a\|_2$. Thus, it follows from Proposition 2.20 that

$$\mathbb{P}\left[f(\varepsilon_1, \cdots, \varepsilon_n) \geq \mathbb{E}\left[f(\varepsilon_1, \cdots, \varepsilon_n) \right] + t \right] \leq \exp\left(-\frac{t^2}{8 \sup_{a \in \mathcal{A}} \|a\|_2^2} \right).$$

Note that the quantity $\sup_{a \in \mathcal{A}} \|a\|_2^2$ (the squared Euclidean width of the set) used in the upper bound here may be substantially than $\sum_{k=1}^{n} \sup_{a \in \mathcal{A}} a_k^2$ established in Example 1.43 of Lecture 1.

2.3 Gaussian concentration

Lastly, we present a classical concentration inequality of standard Gaussian random variables. The proof of the inequality requires a type of Gaussian log-Sobolev inequality which is listed below without proof. Interested readers are referred to Chapter 3.4 of [2] or Chapter 5.3 of [3] for a proof.

Lemma 2.23 (Gaussian log-Sobolev inequality) *Let $X_1, \cdots, X_n$ be a collection of n independent standard Gaussian random variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Then*

$$\text{Ent}\left[f^2(X_1, \cdots, X_n) \right] \leq 2\mathbb{E}\left[\|\nabla f(X_1, \cdots, X_n)\|_2^2 \right].$$

To see why the above inequality is referred to as a type of log-Sobolev inequality, assume for simplicity f is single variable function. Then by the chain rule we have

$$\text{Ent}\left[e^{\lambda f(X)}\right] \leq \frac{1}{2}\mathbb{E}\left[|\lambda f'(X)|^2 e^{\lambda f(X)}\right], \quad \text{for all } \lambda \in \mathbb{R}$$

which is analogous to the one give in Lemma 2.16, but with the discrete gradient replaced by the calculus gradient. Similarly, when f is a multivariable function, we have (**check this!**)

$$\text{Ent}\left[e^{\lambda f(X_1,\cdots,X_n)}\right] \leq \frac{1}{2}\mathbb{E}\left[\|\lambda \nabla f(X_1,\cdots,X_n)\|_2^2\, e^{\lambda f(X_1,\cdots,X_n)}\right], \quad \text{for all } \lambda \in \mathbb{R} \tag{2.11}$$

Theorem 2.24 *Let $X_1,\cdots,X_n$ be a collection of n independent standard Gaussian random variables. Let $f:\mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function with parameter $L>0$. That is, for any $x,y \in \mathbb{R}^n$,*

$$|f(x)-f(y)| \leq L\|x-y\|_2.$$

Then $f(X_1,\cdots,X_n) - \mathbb{E}[f(X_1,\cdots,X_n)]$ is sub-Gaussian with parameter $\nu^2 = L^2$, and hence

$$\mathbb{P}\left[|f(X_1,\cdots,X_n) - \mathbb{E}[f(X_1,\cdots,X_n)]| \geq t\right] \leq 2\text{exp}\left(-\frac{t^2}{2L^2}\right).$$

Proof: We may assume f is differentiable (otherwise we can use an approximation argument). Then $\|\nabla f(X_1,\cdots,X_n)\|_2$ is bounded by L (**check this!**). It follows from (2.11) that

$$\text{Ent}\left[e^{\lambda f(X_1,\cdots,X_n)}\right] \leq \frac{\lambda^2 L^2}{2}\mathbb{E}\left[e^{\lambda f(X_1,\cdots,X_n)}\right].$$

Then claim follows immediately by the Herbst argument. ■

Example 2.25 (Gaussian complexity) *Let $X_1,\cdots,X_n$ be an i.i.d sequence of $\mathcal{N}(0,1)$ variables. Given a set $\mathcal{A} \in \mathbb{R}^n$, define the random variable*

$$Z = \sup_{a\in\mathcal{A}}\left[\sum_{k=1}^n a_k X_k\right] = \sup_{a\in\mathcal{A}}\langle a, X\rangle,$$

where $X=(X_1,\cdots,X_n)$. The Gaussian complexity, denoted $\mathcal{G}_n(\mathcal{A})$, is defined as the expectation of Z,

$$\mathcal{G}_n(\mathcal{A}) = \mathbb{E}[Z],$$

which is another way to measure the complexity of a set (cf. the Rademacher complexity).

*Define $f(x_1,\cdots,x_n) = \sup_{a\in\mathcal{A}}[\sum_{k=1}^n a_k x_k]$. It is not hard to see that f is a Lipschitz function with parameter $\sup_{a\in\mathcal{A}}\|a\|_2$ (**check this!**). Thus, by the Gaussian concentration inequality we know that $Z = \sup_{a\in\mathcal{A}}\langle a, X\rangle$ is sub-Gaussian with parameter $\nu^2 = \sup_{a\in\mathcal{A}}\|a\|_2^2$.*

Example 2.26 (Singular values of Gaussian random matrices) *Let $A \in \mathbb{R}^{n\times n}$ be a random Gaussian matrix whose entries obey the i.i.d standard Gaussian distribution. Let $\sigma_k(A)$ be the k-th largest singular value of A. By Weyl's theorem (this can be found in any standard linear algebra textbook), we have*

$$|\sigma_k(A) - \sigma_k(A')| \leq \|A-A'\|_F.$$

That is, $\sigma_k(A)$ is Lipschitz with parameter 1. Therefore, we can conclude that $\sigma_k(A)$ is sub-Gaussian with parameter $\nu^2 = 1$.

Reading Materials

[1] Martin Wainwright, *High-dimensional statistics – A non-asymptotic viewpoint*, Chapter 3.1.

[2] Ramon van Handel, *Probability in High Dimension*, Chapters 3.3, 3.4.

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Chapters 5.3, 5.4, 6.1, 6.3, 6.4, 6.6.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 3: Lipschitz Concentration and Transportation Method

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/03/27)*

Recap and Motivation: In this lecture, we take a reverse route by fixing the family of functions and then asking for what kind of random variables concentration phenomena will display. In particular, we will consider Lipschitz functions.

Definition 3.1 *Letting $(\mathcal{X}, d)$ be a (measurable) metric space, we say a function $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if $|f(x) - f(y)| \le Ld(x, y)$ for all $x, y \in \mathcal{X}$.*

Remark 3.2 *Note that in this lecture $\mathcal{X}$ can used to denote a single metric space or a product metric space, which should be clear from its context. In addition, the Lipschitz condition is indeed equivalent to $f(x) - f(y) \le Ld(x, y)$ for all $x, y \in \mathcal{X}$ since $d(x, y) = d(y, x)$.*

In the last lecture we have already seen a result about concentration of Lipchitz functions - Gaussian concentration. That is, letting $X_1, \cdots, X_n$ be i.i.d standard Gaussian random variables taking values in $\mathcal{X} = \mathbb{R}$, if $f : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz function with respect to the $\|\cdot\|_2$ norm, then $f(X_1, \cdots, X_n)$ presents the Gaussian type concentration. It is natural to study the general principal for the concentration phenomena of Lipschitz functions and answer the following question:

> *When $f(X_1, \cdots, X_n)$ is sub-Gaussian for f being a Lipschitz function under certain metric defined on $\bigotimes_{k=1}^n \mathcal{X}_k$?*

The answer to this question essentially relies on the distribution[1] of $(X_1, \cdots, X_n)$. As in the entropy method, there are two key ingredients in the analysis: 1) a new characterization of the sub-Gaussian property based on the Wasserstein distance (known as transportation lemma), 2) a tensorization property (known as Marton theorem) which can transfer the problem from the general n case to the $n = 1$ case.

In fact we can also express the bounded difference inequality as the concentration of Lipschitz functions under a properly chosen metric. Let $(X_1, \cdots, X_n)$ be a vector of independent random variables taking values in $\bigotimes_{k=1}^n \mathcal{X}_k := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. For any function $f : \bigotimes_{k=1}^n \mathcal{X}_k \to \mathbb{R}$ satisfying the bounded difference property

$$|f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n)| \le L_k,$$

$f(X_1, \cdots, X_n)$ has the Gaussian type concentration. To rephrase this result into a concentration result of Lipschitz functions, first define the following *weighted Hamming metric* on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$,

$$d_L(x, y) = \sum_{k=1}^n L_k 1_{\{x_k \neq y_k\}}, \quad \text{where } 1_{\{x_k \neq y_k\}} = \begin{cases} 1 & \text{if } x_k \neq y_k \\ 0 & \text{if } x_k = y_k. \end{cases}$$

[1] Or relies on the property of the probability measure on $\bigotimes_{k=1}^n \mathcal{X}_k$ induced by $(X_1, \cdots, X_n)$ since the distribution of the random variable $f(X_1(\omega), \cdots, X_n(\omega))$, where ω is in a probability space $(\Omega, \mathcal{F}, \mu)$, is completely determined by the probability measure on $\bigotimes_{k=1}^n \mathcal{X}_k$ induced by $(X_1, \cdots, X_n)$. Thus, we may equivalently ask: under what probability measure on $\bigotimes_{k=1}^n \mathcal{X}_k$, $f(x_1, \cdots, x_n)$ is sub-Gaussian?

Exercise 3.3 *Verify that $d_L(\cdot,\cdot)$ is a metric.*

Assuming f satisfies the bounded difference property. It follows that

$$\begin{aligned} f(x) - f(y) &= \sum_{k=1}^{n} (f(x_1, \cdots, x_{k-1}, x_k, y_{k+1}, \cdots, y_n) - f(x_1, \cdots, x_{k-1}, y_k, y_{k+1}, \cdots, y_n)) \\ &\leq \sum_{k=1}^{n} L_k 1_{\{x_k \neq y_k\}} \\ &= d_L(x, y). \end{aligned}$$

That is, f is 1-Lipschitz under with respect to the metric d_L. Therefore the bounded difference inequality can be rephrased as: Assume f is 1-Lipschitz under with respect to the metric d_L, then for any independent random variables $X_1, \cdots, X_n$, $f(X_1, \cdots, X_n)$ is sub-Gaussian.

Agenda:

- KL divergence, Wasserstein distance
- Transportation lemma and tensorization
- Talagrand concentration inequality
- Short summary

3.1 KL Divergence, Wasserstein Distance

In this section we introduce two notions, KL divergence and Wasserstein distance, to measure the divergence or distance between the two probability distributions. These two distances are not only useful here but actually widely used in machine learning. Of course there are other divergence measures which will be introduced in the due course.

3.1.1 KL Divergence

Definition 3.4 (Kullback-Leibler (KL) Divergence) *Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$, the KL divergence (or relative entropy) of $\mathbb{Q}$ with respect to $\mathbb{P}$ is defined as*

$$D(\mathbb{Q}\|\mathbb{P}) = \begin{cases} \mathrm{Ent}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] & \text{if } \mathbb{Q} \ll \mathbb{P} \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathrm{Ent}_{\mathbb{P}}[\cdot]$ means computing the entropy (see Definition 2.1 of Lecture 2) of $d\mathbb{Q}/d\mathbb{P}$ under the probability distribution $\mathbb{P}$.

Remark 3.5 *The KL divergence quantifies the difference of $\mathbb{P}$ and $\mathbb{Q}$ using the randomness of $\mathbb{Q}$ relative to $\mathbb{P}$. Thus, KL divergence is also known as relative entropy. Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$, $\mathbb{Q} \ll \mathbb{P}$ means $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, namely, there exists a (real-valued) nonnegative random variable $Y(x)$ with $\mathbb{E}_{\mathbb{P}}[Y] = \int_{\mathcal{X}} Y d\mathbb{P} = 1$ such that*

$$\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}[Y 1_A] \quad \text{for any measurable } A.$$

If $\mathbb{P}$ and $\mathbb{Q}$ have densities with respect to some underlying measure μ (e.g., take $\mu = \mathbb{P}+\mathbb{Q}$), denoted $p(x)$ and $q(x)$, then

$$\mathbb{Q}(A) = \int_A q(x)\mu(dx) = \int_A \frac{q(x)}{p(x)} p(x)\mu(dx) = \mathbb{E}_{\mathbb{P}}\left[\frac{q(x)}{p(x)}\right],$$

and thus Y can be chosen to be $Y(x) = q(x)/p(x)$.

Remark 3.6 (Equivalent definition of KL-divergence) *By the definition of entropy[2], we have*

$$\begin{aligned} D(\mathbb{Q}\|\mathbb{P}) &= \mathbb{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}} \log \frac{d\mathbb{Q}}{d\mathbb{P}}\right] - \mathbb{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] \log \mathbb{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{Q}}{d\mathbb{P}}\right]. \end{aligned} \tag{3.1}$$

If both $\mathbb{P}$ and $\mathbb{Q}$ have densities with respect to some underlying measure μ, then

$$D(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \mu(dx). \tag{3.2}$$

In particular, when x is a discrete space and $\mathbb{P}$ and $\mathbb{Q}$ are discrete probability distributions, we have

$$D(\mathbb{Q}\|\mathbb{P}) = \sum_{x\in\mathcal{X}} q(x) \log \frac{q(x)}{q(x)}. \tag{3.3}$$

Note that $D(\mathbb{Q}\|\mathbb{P})$ is not a metric ($D(\mathbb{Q}\|\mathbb{P}) \neq D(\mathbb{P}\|\mathbb{Q})$ in general, **give an example!**). However, we do have the following lemma.

Lemma 3.7 *$D(\mathbb{Q}\|\mathbb{P}) \geq 0$ and the equality holds if and only if $\mathbb{P} = \mathbb{Q}$ (almost everywhere).*

Proof: Use the definition in (3.1) and Jensen's inequality (noting that $\log x$ is strictly convex). ■

Example 3.8 *Let $\mathbb{P} = \mathcal{N}(\mu_1, \sigma^2)$ and $\mathbb{Q} = \mathcal{N}(\mu_2, \sigma^2)$. Then*

$$\begin{aligned} D(\mathbb{Q}\|\mathbb{P}) &= \mathbb{E}_{x\sim\mathbb{Q}}\left[\frac{-(x-\mu_2)^2}{2\sigma^2} + \frac{-(x-\mu_1)^2}{2\sigma^2}\right] \\ &= \mathbb{E}_{x\sim\mathbb{Q}}\left[\frac{\mu_1^2 - \mu_2^2 - 2(\mu_1-\mu_2)x}{2\sigma^2}\right] \\ &= \frac{\mu_1^2 - \mu_2^2 - 2(\mu_1-\mu_2)\mu_2}{2\sigma^2} \\ &= \frac{(\mu_1-\mu_2)^2}{2\sigma^2}. \end{aligned}$$

Exercise 3.9 *Compute the KL divergence between two multivariate Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$.*

The following lemmas establishes a connection between moment generating function and the KL divergence, showing the duality property between them.

[2]We will always assume $\mathbb{Q} \ll \mathbb{P}$ without specifying this next.

Lemma 3.10 (Duality between KL and MGF) *Let $X \sim \mathbb{P}$. Then,*

$$\log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right] = \sup_{\mathbb{Q} \ll \mathbb{P}} \left\{\mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - D(\mathbb{Q}\|\mathbb{P})\right\}.$$

Proof: Let $Z = f(X) - \log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right]$. Since $\mathbb{E}_{\mathbb{P}}\left[e^{Z}\right] = 1$, by the variational result in Lemma 2.12 of Lecture 2, we have

$$D(\mathbb{Q}\|\mathbb{P}) = \mathrm{Ent}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] \geq \mathbb{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}} Z\right] = \mathbb{E}_{\mathbb{Q}}\left[Z\right] = \mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - \log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right],$$

or equivalently that

$$\log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right] \geq \mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - D(\mathbb{Q}\|\mathbb{P}).$$

Since this holds for any $\mathbb{Q} \ll \mathbb{P}$, it follows that $\log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right] \geq \sup_{\mathbb{Q} \ll \mathbb{P}} \left\{\mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - D(\mathbb{Q}\|\mathbb{P})\right\}$.

In addition, if we define $\mathbb{Q}$ by

$$d\mathbb{Q} = \frac{e^{f(X)}}{\mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right]} d\mathbb{P},$$

a direct calculation can show that $\log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right] = \mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - D(\mathbb{Q}\|\mathbb{P})$ (**check this!**). This completes the proof. ■

Lemma 3.11 (Chain rule of KL) *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures that define the joint distribution of random variables (X_1, X_2). Then,*

$$D(\mathbb{Q}\|\mathbb{P}) = D(\mathbb{Q}_1\|\mathbb{P}_1) + \mathbb{E}_{\mathbb{Q}_1}\left[D\left(\mathbb{Q}_2(\cdot|X_1)\|\mathbb{P}_2(\cdot|X_1)\right)\right],$$

where $\mathbb{P}_1$ and $\mathbb{Q}_2$ are marginal distributions of X_1 under the joint distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively, and $\mathbb{P}_2(\cdot|X_1)$ and $\mathbb{Q}_2(\cdot|X_1)$ are the conditional distribution of X_2 given X_1 under the joint distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively.

Proof: It follows from the Bayes formula that

$$d\mathbb{P} = d\mathbb{P}_1 \cdot d\mathbb{P}_2(\cdot|X_1) \quad \text{and} \quad d\mathbb{Q} = d\mathbb{Q}_1 \cdot d\mathbb{Q}_2(\cdot|X_1).$$

Consequently,

$$\begin{aligned} D(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{Q}}{d\mathbb{P}}\right] &= \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{Q}_1}{d\mathbb{P}_1}\right] + \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{Q}_2(\cdot|X_1)}{d\mathbb{P}_2(\cdot|X_1)}\right] \\ &= \mathbb{E}_{\mathbb{Q}_1}\left[\log \frac{d\mathbb{Q}_1}{d\mathbb{P}_1}\right] + \mathbb{E}_{\mathbb{Q}_1}\left[\mathbb{E}_{\mathbb{Q}_2(\cdot|X_1)}\left[\log \frac{d\mathbb{Q}_2(\cdot|X_1)}{d\mathbb{P}_2(\cdot|X_1)}\right]\right] \\ &= D(\mathbb{Q}_1\|\mathbb{P}_1) + \mathbb{E}_{\mathbb{Q}_1}\left[D\left(\mathbb{Q}_2(\cdot|X_1)\|\mathbb{P}_2(\cdot|X_1)\right)\right] \end{aligned}$$

as claimed. ■

Remark 3.12 *To have a better understanding of the above lemma, let us particularly consider the case when (X_1, X_2) are real-valued and $\mathbb{P}$ and $\mathbb{Q}$ have continuous densities $p(x_1, x_2)$ and $q(x_1, x_2)$ with respect to the Lebesgue measure, respectively. Let $p_1(x_1)$ and $q_1(x_1)$ be the marginal distribution of X_1 and X_2 under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Let $p_2(x_2|x_1)$ and $q_2(x_2|x_1)$ be the conditional probability densities of X_2 given X_1 under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then*

$$p_2(x_2|x_1) = \frac{p(x_1, x_2)}{\int_{\mathbb{R}} p(x_1, x_2)dx_2} \quad \textit{and} \quad q_2(x_2|x_1) = \frac{q(x_1, x_2)}{\int_{\mathbb{R}} q(x_1, x_2)dx_2}.$$

It follows that

$$\begin{aligned}
D(\mathbb{Q}\|\mathbb{P}) &= \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{q(x_1, x_2)}{p(x_1, x_2)} dx_1 dx_2 \\
&= \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{q_2(x_2|x_1) \left(\int_{\mathbb{R}} q(x_1, x_2)dx_2\right)}{p_2(x_2|x_1) \left(\int_{\mathbb{R}} p(x_1, x_2)dx_2\right)} dx_1 dx_2 \\
&= \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{\left(\int_{\mathbb{R}} q(x_1, x_2)dx_2\right)}{\left(\int_{\mathbb{R}} p(x_1, x_2)dx_2\right)} dx_1 dx_2 + \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{q_2(x_2|x_1)}{p_2(x_2|x_1)} dx_1 dx_2 \\
&= \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{q_1(x_1)}{p_1(x_1)} dx_1 dx_2 + \int_{\mathbb{R}\times\mathbb{R}} q_1(x_1) q_2(x_2|x_1) \log \frac{q_2(x_2|x_1)}{p_2(x_2|x_1)} dx_1 dx_2 \\
&= \int_{\mathbb{R}} q_1(x_1) \log \frac{q_1(x_1)}{p_1(x_1)} dx_1 + \int_{\mathbb{R}} q_1(x_1) \left(\int_{\mathbb{R}} q_2(x_2|x_1) \log \frac{q_2(x_2|x_1)}{p_2(x_2|x_1)} dx_2 \right) dx_1 \\
&= D(\mathbb{Q}_1\|\mathbb{P}_1) + \mathbb{E}_{\mathbb{Q}_1} \left[D\left(\mathbb{Q}_2(\cdot|X_1)\|\mathbb{P}_2(\cdot|X_1)\right) \right].
\end{aligned}$$

Exercise 3.13 *Let $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2$ and $\mathbb{Q} = \mathbb{Q}_1 \times \mathbb{Q}_2$ two product probability measures that define the joint distribution of random variables (X_1, X_2). Show that*

$$D(\mathbb{Q}\|\mathbb{P}) = D(\mathbb{Q}_1\|\mathbb{P}_1) + D(\mathbb{Q}_2\|\mathbb{P}_2).$$

Exercise 3.14 *Generalize the result in Lemma 3.11 and Exercise 3.13 to the case of joint distributions of n random variables.*

3.1.2 Wasserstein Distance

Wasserstein distance is a distance defined over probability measures within the framework of optimal transport. Roughly speaking, it is the least cost of transporting/redistributing a source probability measure to a target probability measure. Optimal transport was first introduced in the Monge formulation. Then Kantorovich relaxed it by allowing the mass splitting in the source based on the notation of coupling.

Definition 3.15 (Coupling) *Let $\mathbb{P}$ and $\mathbb{Q}$ be two given probability measures on $\mathcal{X}$. We say a probability measures π on $\mathcal{X} \times \mathcal{X}$ is a coupling of $\mathbb{P}$ and $\mathbb{Q}$ if the marginal distributions of π in the first and second coordinate coincides with $\mathbb{P}$ and $\mathbb{Q}$, respectively. In addition, we denote by $\mathcal{C}(\mathbb{P}, \mathbb{Q})$ the set of all the couplings of $\mathbb{P}$ and $\mathbb{Q}$.*

Exercise 3.16 *Does the coupling always exist? Are there always more than two couplings for a fixed pair of probability measures?*

Definition 3.17 (Wasserstein Distance) *Let $(\mathcal{X}, d)$ be a metric space. Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$, their Wasserstein distance is defined as*

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} d(x, y) d\pi(x, y) = \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi \left[d(X, Y) \right] \tag{3.4}$$

If we view the joint distribution π as a transport plan, meaning a scheme for reshuffling the probability mass $\mathbb{P}$ to another probability mass $\mathbb{Q}$, and view $d(\cdot, \cdot)$ as the unit transport cost, then $\mathbb{E}_\pi \left[d(X, Y) \right]$ can be interpreted as the transport cost of associated with the plan π. Seeking the the transportation plan that minimizes the transport cost is the optimal transport problem. The solution to the optimal transport problem measures how far we have to move the mass of $\mathbb{P}$ and turn it into $\mathbb{Q}$ and thus is a natural way to define the distance between two probability measures.

Remark 3.18

1. *If d is a distance on $\mathcal{X}$, $W_1(\mathbb{P}, \mathbb{Q})$ is indeed a distance. Namely, it satisfies the three conditions required for a distance, especially the triangular inequality. A proof of this can be found in [4] and references therein.*

2. *It is evident that we can express the Wasserstein distance as*

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{(X, Y)} \{ \mathbb{E} \left[d(X, Y) \right], \ X \sim \mathbb{P}, \ Y \sim \mathbb{Q} \}.$$

3. *What we have in* (3.4) *is actually the* 1*-Wasserstein distance, hence there is subscript* 1 *in the notation. In general, we may also define p-Wasserstein distance as follows:*

$$W_p = \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \left(\mathbb{E}_\pi \left[[d(X, Y)]^p \right] \right)^{1/p}. \tag{3.5}$$

4. *Computing the Wasserstein distance is generally difficult and relies on some numerical solvers. A detailed discussion of the computations is beyond the scope of this lecture.*

If viewing (3.4) as an optimization problem (equality constrained) on the infinity dimensional probability measure space, we can compute its dual problem. In addition, since every probability measure corrresponds a linear functional over the function space on $\mathcal{X}$, we can also define the distance between probability measures based on the perspective of linear functional (similar to operator norm). This provides another duality for Wasserstein distance which plays an important role in this lecture.

Theorem 3.19 (Duality) *Under mild conditions, we have*

$$W_1(\mathbb{P}, \mathbb{Q}) \triangleq \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi \left[d(X, Y) \right] = \sup_{f \in \mathrm{Lip}(\mathcal{X}, d)} \left| \mathbb{E}_\mathbb{P} \left[f(X) \right] - \mathbb{E}_\mathbb{Q} \left[f(Y) \right] \right|. \tag{3.6}$$

The proof of this theorem can be found in [4]. Thus, we only consider a simple example.

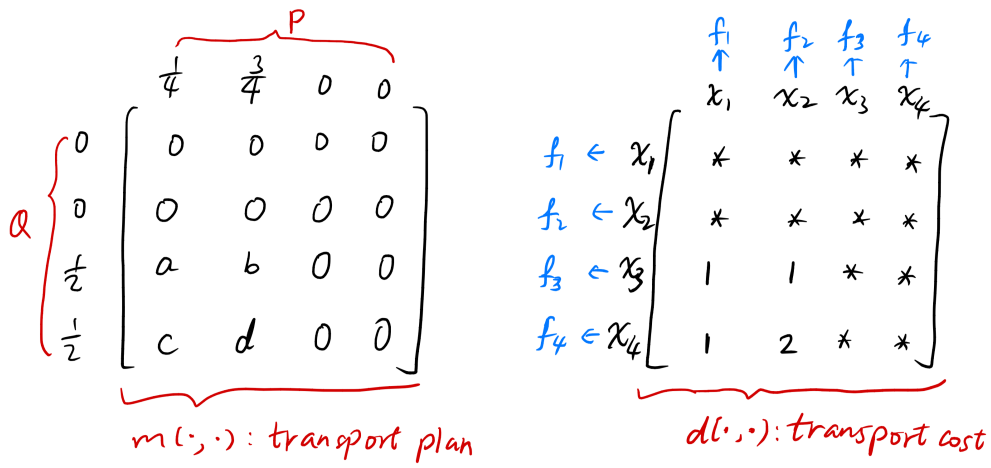


Figure 3.1: Problem setup of Example 3.20.

Example 3.20 (Discrete example) *Here we consider $\mathbb{P}$ and $\mathbb{Q}$ defined on a discrete space $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$, see Figure 3.1 for the problem setup. In order for π to be a coupling of $\mathbb{P}$ and $\mathbb{Q}$ we must have*

$$a + c = 1/4,\ b + d = 3/4,\ a + b = 1/2,\ c + d = 1/2,\ a, b, c, d \geq 0.$$

For example, if we take $\pi = \mathbb{P} \otimes \mathbb{Q}$, then

$$a = 1/8,\ b = 3/8,\ c = 1/8,\ d = 3/8,$$

with the total transport cost

$$c_1 = \mathbb{E}_\pi\left[d(X, Y)\right] = 1/8 + 3/8 + 1/8 + 6/8 = 11/8.$$

There are also exists other coupling of $\mathbb{P}$ and $\mathbb{Q}$ (or transport plan) in addition to $\pi_2 = \mathbb{P} \otimes \mathbb{Q}$. For example, we may let

$$a = 0,\ b = 1/2,\ c = 1/4,\ d = 1/4,$$

with the total transport cost

$$c_2 = \mathbb{E}_\pi\left[d(X, Y)\right] = 0 + 1/2 + 1/4 + 1/2 = 5/4 < c_1.$$

*In fact, this is the minimum total transport cost we can achieve over all the possible couplings of $\mathbb{P}$ and $\mathbb{Q}$ (**why?**), i.e.,*

$$\inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi\left[d(X, Y)\right] = 5/4.$$

Moreover, let $f = (f_1, f_2, f_3, f_3)$ be a function defined on $\mathcal{X}$. Then f is 1-Lipschitz under $d(\cdot, \cdot)$ if and only if

$$|f_1 - f_3| \leq d(x_1, x_3) = 1,\ |f_2 - f_3| \leq d(x_2, x_3) = 1,\ |f_1 - f_4| \leq d(x_1, x_4) = 1,\ |f_2 - f_4| \leq d(x_2, x_4) = 2.$$

Given a 1-Lipschitz f, letting π (with (a,b,c,d)) be any coupling of $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\begin{aligned}
|\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]| &= \left|\frac{1}{4}f_1 + \frac{3}{4}f_2 - \frac{1}{2}f_3 - \frac{1}{2}f_4\right| \\
&= |(a+c)f_1 + (b+d)f_2 - (a+b)f_3 - (c+d)f_4| \\
&= |a(f_1 - f_3) + b(f_2 - f_3) + c(f_1 - f_4) + d(f_2 - f_4)| \\
&\leq \mathbb{E}_{\pi}[d(X,Y)].
\end{aligned}$$

It follows that,

$$\sup_{f \in \mathrm{Lip}(\mathcal{X},d)} |\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]| \leq \inf_{\pi \in \mathcal{C}(\mathbb{P},\mathbb{Q})} \mathbb{E}_{\pi}[d(X,Y)].$$

The equality can be achieved for example by taking $f = (f_1, f_2, f_3, f_4) = (0, 1, 0, -1)$. Thus, we have verified the duality theorem using this example.

Next we study a special case of the Wasserstein distance with the trivial metric $d(x,y) = 1_{\{x \neq y\}}$. In this case it can be shown that the Wasserstein distance is none other than the total variation distance which itself is interesting in many applications.

Definition 3.21 (Total variation distance) *The total variation distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

Example 3.22 (Wasserstein distance for $d(x,y) = 1_{\{x \neq y\}}$) *In this case it is not hard to see that f is 1-Lipschitz if and only if*

$$|f(x) - f(y)| \leq 1.$$

Since the Wasserstein distance is invariant to constant offsets of the function, we have

$$W_1(\mathbb{P},\mathbb{Q}) = \sup_{0 \leq f \leq 1} |\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Y)]|.$$

Let $d\mathbb{P}(x) = p(x)\mu(dx)$ and $d\mathbb{Q}(x) = q(x)\mu(dx)$ for some underlying measure μ. Then we can rewrite the Wasserstein distance as

$$\begin{aligned}
W_1(\mathbb{P},\mathbb{Q}) &= \sup_{0 \leq f \leq 1} \left|\int_{\mathcal{X}} f(x)(p(x) - q(x))\mu(dx)\right| \\
&= \int_{\mathcal{X}} [p(x) - q(x)]_+ \mu(dx). \qquad (3.7)
\end{aligned}$$

On the other hand, since

$$|\mathbb{P}(A) - \mathbb{Q}(A)| = \left|\int_A (p(x) - q(x))\mu(dx)\right|,$$

it is not hard to see that the supremum is attained at $A_1 = \{x : \ p(x) \geq q(x)\}$ *or* $A_2 = \{x : \ q(x) \geq p(x)\}$ ***(show that*** $|\mathbb{P}(A) - \mathbb{Q}(A)|$ ***have the same value on these two sets!)****. It follows that*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \int_{\mathcal{X}} [p(x) - q(x)]_+ \, \mu(dx).$$

Therefore we have

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi [X \neq Y] = \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}. \tag{3.8}$$

In this case we can also construct the optimal coupling/transport plan explicitly. We refer interested readers to Chapter 4.2 of [2] for details.

3.2 Transportation Lemma and Tensorization

3.2.1 Transportation Lemma

There is a necessary and sufficient condition in terms of the two probability divergence measures to characterize the sub-Gaussian property of Lipschitz functions. We begin with a sub-Gaussian characterization in terms of the KL divergence.

Lemma 3.23 (Sub-Gaussian in terms of KL) *Letting* $X \sim \mathbb{P}$*, then* $f(X)$ *is* ν^2*-sub-Gaussian if and only if*

$$\mathbb{E}_{\mathbb{Q}} [f(Y)] - \mathbb{E}_{\mathbb{P}} [f(X)] \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})} \quad \textit{for all } \mathbb{Q} \ll \mathbb{P}.$$

Proof: By the definition $f(X)$ is sub-Gaussian if and only if

$$\log \mathbb{E}_{\mathbb{P}} \left[e^{\lambda (f(X) - \mathbb{E}_{\mathbb{P}}[f(X))]} \right] \leq \frac{\lambda^2 \nu^2}{2} \quad \text{for all } \lambda \in \mathbb{R}.$$

Then, by the duality between DL divergence and MGF, this is equivalent to

$$\lambda \left(\mathbb{E}_{\mathbb{Q}} [f(Y) - \mathbb{E}_{\mathbb{P}} [f(X)]] \right) - D(\mathbb{Q}\|\mathbb{P}) - \frac{\lambda^2 \nu^2}{2} \leq 0 \quad \text{for all } \lambda \in \mathbb{R} \text{ and } \mathbb{Q} \ll \mathbb{P}.$$

Taking the supremum of the left hand side yields the claim. ■

Lemma 3.24 (Transportation lemma) *Let* $\mathbb{P}$ *be a probability measure defined on a metric space* $(\mathcal{X}, d)$*. Then the following are equivalent:*

1. *Letting* $X \sim \mathbb{P}$*,* $f(X)$ *is* ν^2*-sub-Gaussian for every* $f \in \mathrm{Lip}\,(\mathcal{X}, d)$*.*
2. $W_1(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})}$ *for all probability measures* $\mathbb{Q} \ll \mathbb{P}$*.*

Proof: By the last lemma we can see that that the property 1 can be stated as

$$|\mathbb{E}_{\mathbb{Q}} [f(Y)] - \mathbb{E}_{\mathbb{P}} [f(X)]\,| \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})} \quad \text{for all } \mathbb{Q} \ll \mathbb{P}. \quad \text{for all } f \in \mathrm{Lip}\,(\mathcal{X}, d) \text{ and } \mathbb{Q} \ll \mathbb{P}.$$

Taking the supremum of the left hand side with respect to $f \in \mathrm{Lip}\,(\mathcal{X}, d)$ yields that the above expression is equivalent to

$$W_1(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})} \quad \text{for all } \mathbb{Q} \ll \mathbb{P},$$

which concludes the proof. ■

Exercise 3.25 *Lemma 3.24 works for f being 1-Lipschitz functions. What about general L-Lipschitz functions?*

Our first consequence of Lemma 3.24 is a useful inequality known as *Pinsker inequality*

Proposition 3.26 (Pinsker inequality) *Let $\mathbb{P}$ and $\mathbb{Q}$ are probability measures satisfying $\mathbb{Q} \ll \mathbb{P}$. Then*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{Q}\|\mathbb{P})}. \tag{3.9}$$

Proof: First we have shown that f is 1-Lipschitz with respect to $d(x, y) = 1_{\{x \neq y\}}$ if and only if $|f(x) - f(y)| \leq 1$. Thus, for any $X \sim \mathbb{P}$, we known that $f(X)$ is in an interval of length bounded by 1. Consequently, $f(X)$ is $\frac{1}{4}$-sub-Gaussian. Therefore, applying Lemma 3.24 yields the result since we have already shown that $W_1(\mathbb{P}, \mathbb{Q}) = \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}$ for the trivial metric. ■

Of course, if we could give an independent proof of Pinsker inequality (there are indeed direct proofs), we can use Lemma 3.24 to provide an alternative proof of the sub-Gaussian property of bounded variables (by taking $f = 1_{[a,b]}(x)$).

Exercise 3.27 *Let $X \sim \mathbb{P}$ be ν^2-sub-Gaussian. Show that $W_1(\mathbb{Q}, \mathbb{P}) \lesssim \sqrt{\nu^2 D(\mathbb{Q}\|\mathbb{P})}$.*

Due to Theorem 3.19, in order to show $f(X)$ is sub-Gaussian for $f \in \mathrm{Lip}\,(\mathcal{X}, d)$, it suffices to show that

$$\inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi \left[d(X, Y) \right] \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})}. \tag{3.10}$$

Inequalities of this type are usually called *transportation (cost) inequalities* which play an key role in establishing some useful concentration inequalities that cannot be established by the previous methods.

3.2.2 Tensorization and Bounded Difference Inequality Revisited

Given metric spaces $(\mathcal{X}_k, d_k),\ k = 1, \cdots, n$, there are different ways to define a metric on $\bigotimes_{k=1}^n \mathcal{X}_k$, for example,

$$d_L(x, y) = \sum_{k=1}^n L_k d_k(x_k, y_k), \quad L_k > 0 \tag{3.11}$$

or

$$d_2(x, y) = \sqrt{\sum_{k=1}^n d_k(x_k, y_k)^2}. \tag{3.12}$$

Rather that considering the tensorizaton in a specific setting, the following theorem provides a general tensorization principle. The proof of the theorem is by induction and is omitted. Details of the proof can be found in [2] and [3].

Theorem 3.28 (Marton) *Let $\bigotimes_{k=1}^n \mathbb{P}_k$ be a product measure on a product measure space $\bigotimes_{k=1}^n \mathcal{X}_k$. Let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function, and let $c_k : \mathcal{X}_k \times \mathcal{X}_k \to \mathbb{R}_+$ be positive weight function. Suppose that for $k = 1, \cdots, n$ and for every probability measure $\mathbb{Q}_k$ that which is absolutely continuous with respect to $\mathbb{P}_k$*

$$\inf_{\pi \in \mathcal{C}(\mathbb{P}_k, \mathbb{Q}_k)} \phi\left(\mathbb{E}_\pi\left[c_k(X_k, Y_k)\right]\right) \leq 2\nu^2 D(\mathbb{Q}_k \| \mathbb{P}_k).$$

Then for any probability measure $\mathbb{Q}$ that is absolutely continuous with respect to $\bigotimes_{k=1}^n \mathbb{P}_k$, we have

$$\inf_{\pi \in \mathcal{C}(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q})} \sum_{k=1}^n \phi(\mathbb{E}_\pi\left[c_k(X_k, Y_k)\right]) \leq 2\nu^2 D\left(\mathbb{Q} \| \bigotimes_{k=1}^n \mathbb{P}_k\right). \tag{3.13}$$

Note that the left hand of (3.13) is not necessarily a W_1 distance. Thus, for different metric, we need to choose suitable ϕ and c_k in order to obtain a result associated with W_1 distance. The following proposition considers the $d_L(\cdot, \cdot)$ metric in (3.11).

Proposition 3.29 *Let $\bigotimes_{k=1}^n \mathbb{P}_k$ be a product measure on a product measure space $\bigotimes_{k=1}^n (\mathcal{X}_k, d_k)$. If for each univariate probability measure,*

$$W_1(\mathbb{Q}_k, \mathbb{P}_k) \leq \sqrt{2\nu^2 D(\mathbb{Q}_k \| \mathbb{P}_k)} \quad \textit{for all } \mathbb{Q}_k \ll \mathbb{P}_k. \tag{3.14}$$

Then

$$W_1\left(\mathbb{Q}, \bigotimes_{k=1}^n \mathbb{P}_k\right) \leq \sqrt{2\nu^2 \left(\sum_{k=1}^n L_k^2\right) D\left(\mathbb{Q} \| \bigotimes_{k=1}^n \mathbb{P}_k\right)} \quad \textit{for all } \mathbb{Q} \ll \bigotimes_{k=1}^n \mathbb{P}_k,$$

where the W_1 is defined using the distance $d_L(x, y) = \sum_{k=1}^n L_k d_k(x_k, y_k)$. Hence, for any 1-Lipschitz function $f : \bigotimes_{k=1}^n \mathcal{X}_k \to \mathbb{R}$ under this metric, $f(X)$ is sub-Gaussian if $X \sim \bigotimes_{k=1}^n \mathbb{P}_k$.

Proof: First we have

$$\begin{aligned}
W_1\left(\mathbb{Q}, \bigotimes_{k=1}^n \mathbb{P}_k\right) &= \inf_{\pi \in \mathcal{C}(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q})} \mathbb{E}_\pi\left[\sum_{k=1}^n L_k d_k(X_k, Y_k)\right] \\
&= \inf_{\pi \in \mathcal{C}(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q})} \sum_{k=1}^n L_k \mathbb{E}_\pi\left[d_k(X_k, Y_k)\right] \\
&\leq \sqrt{\sum_{k=1}^n L_k^2} \sqrt{\inf_{\pi \in \mathcal{C}(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q})} \sum_{k=1}^n \left(\mathbb{E}_\pi\left[d_k(X_k, Y_k)\right]\right)^2}.
\end{aligned}$$

Noting that the assumption implies

$$W_1(\mathbb{Q}_k, \mathbb{P}_k)^2 = \inf_{\pi \in \mathcal{C}(\mathbb{P}_k, \mathbb{Q}_k)} \left(\mathbb{E}_\pi\left[d_k(X_k, Y_k)\right]\right)^2 \leq 2\nu^2 D(\mathbb{Q}_k \| \mathbb{P}_k),$$

the claim follows immediately from Theorem 3.28 by taking $\phi(x) = x^2$ and $w_k(\cdot, \cdot) = d_k(\cdot, \cdot)$. ■

Example 3.30 (Bounded difference inequality revisited) *For any $X_k \sim \mathbb{P}_k$, under the metric $d_k(x_k, y_k) = 1_{\{x_k \neq y_k\}}$, the Pinsker inequality implies that*

$$W_1(\mathbb{Q}_k, \mathbb{P}_k) \leq \sqrt{\frac{1}{2} D(\mathbb{Q}_k \| \mathbb{P}_k)}.$$

Thus, for f being 1-Lipschitz under the metric $d_L(x, y) = \sum_{k=1}^n L_k 1_{\{x_k \neq y_k\}}$ (equivalent to the bounded difference property as mentioned at the beginning of this lecture), the application of Proposition 3.29 yields that

$$f(X_1, \cdots, X_k) \text{ is } \frac{\sum_{k=1}^n L_k^2}{4}\text{-sub-Gaussian},$$

which recovers the bounded difference inequality. More precisely, we have

$$\inf_{\pi \in \mathcal{C}\left(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q}\right)} \sum_{k=1}^n \left(\mathbb{E}_\pi \left[d_k(X_k, Y_k)\right]\right)^2 = \inf_{\pi \in \mathcal{C}\left(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q}\right)} \sum_{k=1}^n \left(\pi \left[X_k \neq Y_k\right]\right)^2 \leq \frac{1}{2} D\left(\mathbb{Q} \| \bigotimes_{k=1}^n \mathbb{P}_k\right). \tag{3.15}$$

3.3 Talagrand Concentration Inequality

Up to this point, the transportation method did not yield any new results yet. In this section we will use a type of asymmetric transportation cost inequalities based on a one-sided variant of the trivial metric to establish the following remarkable Talagrand concentration inequality.

Theorem 3.31 (Talagrand) *Let $f : \bigotimes_{k=1}^n \mathcal{X}_k \to \mathbb{R}$ be a function satisfying*

$$f(y) - f(x) \leq \sum_{k=1}^n a_k(y) 1_{\{x_k \neq y_k\}} \quad \text{for all } x, y.$$

Assume $\sup_y \left(\sum_{k=1}^n a_k^2(y)\right) \leq \nu^2$. Then, for any independent random variables $X = (X_1, \cdots, X_n)$ taking values in $\bigotimes_{k=1}^n \mathcal{X}_k$, $f(X)$ is ν^2-sub-Gaussian.

Proof: Assume $X \sim \bigotimes_{k=1}^n \mathbb{P}_k$. By Lemma 3.23, we need to show that

$$\left|\mathbb{E}_Q \left[f(Y)\right] - \mathbb{E}_{\bigotimes_{k=1}^n \mathbb{P}_k} \left[f(X)\right]\right| \leq \sqrt{2\nu^2 D\left(\mathbb{Q} \| \bigotimes_{k=1}^n \mathbb{P}_k\right)}. \tag{3.16}$$

Letting $\pi \in \mathcal{C}\left(\mathbb{Q}, \mathbb{P}\right)$, a simple calculation yields that (again the goal is to reduce the problem about the concentration of f to the problem of comparing the divergences of two probability measures)

$$\begin{aligned}
\mathbb{E}_Q \left[f(Y)\right] - \mathbb{E}_{\bigotimes_{k=1}^n \mathbb{P}_k} \left[f(X)\right] &= \mathbb{E}_\pi \left[f(Y) - f(X)\right] \\
&\leq \mathbb{E}_\pi \left[\sum_{k=1}^n c_k(Y) 1_{\{X_k \neq Y_k\}}\right] \\
&= \sum_{k=1}^n \mathbb{E}_\pi \left[c_k(Y) 1_{\{X_k \neq Y_k\}}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n} \mathbb{E}_\pi \left[c_k(Y) \pi \left[X_k \neq Y_k | Y \right] \right] \\
&\leq \sum_{k=1}^{n} \left(\mathbb{E}_\pi \left[(c_k(Y))^2 \right] \right)^{1/2} \left(\mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | Y \right] \right)^2 \right] \right)^{1/2} \\
&\leq \left(\mathbb{E}_\pi \left[\sum_{k=1}^{n} (c_k(Y))^2 \right] \right)^{1/2} \left(\sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | Y \right] \right)^2 \right] \right)^{1/2} \\
&\leq \nu \left(\sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | Y \right] \right)^2 \right] \right)^{1/2}.
\end{aligned}
$$

Similarly, we have (**check this!**)

$$
\mathbb{E}_{\bigotimes_{k=1}^{n} \mathbb{P}_k} \left[f(X) \right] - \mathbb{E}_Q \left[f(Y) \right] \leq \nu \left(\sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | X \right] \right)^2 \right] \right)^{1/2}.
$$

Therefore, in order to show (3.16), it suffices to show (since π is arbitrary above)

$$
\max \left\{ \inf_{\pi \in \mathcal{C}(\mathbb{Q}, \mathbb{P})} \sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | Y \right] \right)^2 \right], \inf_{\pi \in \mathcal{C}(\mathbb{Q}, \mathbb{P})} \sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | X \right] \right)^2 \right] \right\} \leq 2D \left(\mathbb{Q} \| \bigotimes_{k=1}^{n} \mathbb{P}_k \right), \tag{3.17}
$$

which actually holds. Thus the proof is complete. ■

Remark 3.32 *Here we have used* (3.17) *without proof. Note that* (3.17) *can be viewed as an asymmetric version of* (3.15)*. The details of proof can be found in [2] and [3], which uses a conditional version of Theorem 3.28 for tensorization.*

Corollary 3.33 (Talagrand) *Let $X = (X_1, \cdots, X_n)$ be a vector of independent random variables taking values in $[a, b]$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-Lipschitz with respect to the Euclidean norm. Then, $f(X)$ is $L^2(b-a)^2$-sub-Gaussian.*

Proof: The first order condition for convexity implies

$$
f(y) - f(x) \leq \nabla f(y)^\top (y - x) \quad \text{for all } x, y.
$$

Since $|x_k - y_k| \leq (b - a)$, we have

$$
\begin{aligned}
f(y) - f(x) &\leq \sum_{k=1}^{n} \partial_k f(y)(y_k - x_k) \\
&\leq \sum_{k=1}^{n} (b-a) \left| \partial_k f(y) \right| 1_{\{x_k \neq y_k\}}.
\end{aligned}
$$

The result follows immediately from Theorem 3.31 by further noting that $\|\nabla f(y)\|_2 \leq L$. ■

Remark 3.34 *a) Only upper tail can be established in this scenario based on the entropy method in Lecture 2. b) The convexity property of f is indispensable to establish the sub-Gaussian property. There exists nonconvex 1-Lipschitz functions Corollary 3.33 even fails for symmetric Bernoulli variables, see for example Problem 4.9 of [2].*

Example 3.35 (Rademacher complexity revisited) *In Example 2.22 of Lecture 2, we have established upper tail bound of the Rademacher complexity of a set $\mathcal{A}$ in terms of the width of the set $\sup_{a\in\mathcal{A}} \|a\|_2$. In the example we actually show that the function $f(x_1, \cdots, x_n) = \sup_{a\in\mathcal{A}} [\sum_{k=1}^n a_k x_k]$ is convex and $\sup_{a\in\mathcal{A}} \|a\|_2$-Lipschitz continuous. Thus, it follows from Corollary 3.33 that the Rademacher complexity $\sup_{a\in\mathcal{A}} [\sum_{k=1}^n a_k \varepsilon_k]$ is $4\sup_{a\in\mathcal{A}} \|a\|_2^2$ and hence*

$$\mathbb{P}\left[\left|\sup_{a\in\mathcal{A}}\left[\sum_{k=1}^n a_k\varepsilon_k\right] - \mathbb{E}\left[\sup_{a\in\mathcal{A}}\left|\sum_{k=1}^n a_k\varepsilon_k\right|\right]\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{8\sup_{a\in\mathcal{A}}\|a\|_2^2}\right).$$

3.4 Short Summary

We have discussed three methods for establishing concentration inequalities of functions of independent random variables:

- Chernoff method (Lecture 1)
- Entropy method (Lecture 2)
- Transportation method (Lecture 3).

In both the entropy method and the transportation method, the variational formulations (duality in the transportation method) play an important role in showing the tensorization property. Here are a list of concentration inequalities we have presented:

- Hoeffding inequality (Lecture 1, for sum of independent sub-Gaussian random variables)
- Bernstein equality (Lecture 1, for sum of independent sub-exponential random variables)
- Bounded difference inequality (Lecture 1, for function obeying bounded difference property)
- General bounded difference inequality (Lecture 2, for function obeying asymmetric bounded difference property)
- Gaussian concentration inequality (Lecture 2, Lipschitz function concentration of Gaussian random variables)
- Talagrand inequality (Lecture 3, Lipschitz and convex function concentration of bounded random variables)

Gaussian concentration can also be derived via transportation method discussed in this lecture. In addition, there is another method – geometric method based on isoperimetric inequalities we did not cover. This method works for certain types of probability measures such as Gaussian measure and uniform measure on the sphere.

Reading Materials

[1] Martin Wainwright, *High-dimensional statistics – A non-asymptotic viewpoint*, Chapter 3.3.

[2] Ramon van Handel, *Probability in High Dimension*, Chapters 4.

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Chapters 8.

[4] Gabriel Peyré and Marco Cuturi, *Computational optimal transport*, Chapter 2, 8.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 4: Expectation of Suprema: Finite Approximation

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/04/10)*

Recap and Motivation: As already mentioned previously, estimating the suprema of the form

$$\sup_{t\in T} X_t \tag{4.1}$$

arises in a wide range of contexts. Two representative examples are:

- The generalization error analysis in empirical risk minimization finally reduces to

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}\left[f(X)\right] \right|,$$

  where $\mathcal{F}$ is a set of functions. This is typically referred to as *Uniform Law of Large Numbers*.

- The spectral norm of a random matrix $W \in \mathbb{R}^{m\times n}$ can be expressed as

$$\|W\|_2 = \sup_{\|u\|_2=1, \|v\|_2=1} u^\top W v.$$

While the concentration of (4.1) around its mean for typical applications can be established by the concentration inequalities discussed in the last three lectures, computing the expectation

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \tag{4.2}$$

is by no means easy. This will be the focus in the next few lectures. We will first study the general form (4.2), and then give a particular treatment of this problem for uniform law of large numbers. When there is no conflict from the context, it is always assumed that $\mathbb{E}\left[X_t\right] = 0$.

Agenda:

- Finite maxima
- Gaussian complexity and Rademacher complexity
- Covering and packing
- Finite approximation bound

4.1 Finite Maxima

The problem here is to bound

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right],$$

where X_k is σ^2-sub-Gaussian. Maybe the most naive approach is to bound the supremum by a sum,

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \leq \mathbb{E}\left[\sum_{k=1}^{n} |X_k|\right] \leq n \max_{k=1,\cdots,n} \mathbb{E}\left[|X_k|\right] \lesssim n\sigma,$$

where the last inequality follows from the σ^2-sub-Gaussian property of each X_k. Of course, bounding a maximum by a sum is an exceedingly crude idea. We may consider a transform of $\max_{k=1,\cdots,n} X_k$ such that the gap between supreme and sum is seemingly not so large after the transform. Next we attempt to provide a bound based on the higher order moments,

$$\begin{aligned}
\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] &\leq \left(\left(\mathbb{E}\left[\max_{k=1,\cdots,n} |X_k|\right]\right)^p\right)^{1/p} \\
&\leq \left(\mathbb{E}\left[\max_{k=1,\cdots,n} |X_k|^p\right]\right)^{1/p} \\
&\leq \left(n \max_{k=1,\cdots,n} \mathbb{E}\left[|X_k|^p\right]\right)^{1/p} \\
&\lesssim n^{1/p}\sigma\sqrt{p}.
\end{aligned}$$

Minimizing the righthand side with respect to p yields that

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \lesssim \sigma\sqrt{\log n}.$$

Of course we can apply the moment generating function to estimate the maximum as in the development of the tail bound by the Chernoff method. More precisely, we have the following lemma.

Lemma 4.1 *Let $\{X_k\}_{k=1}^n$ be σ^2-sub-Gaussian random variables. Then*

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \leq \sigma\sqrt{2\log n}.$$

Proof: We have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \max_k X_k\right)\right] = \mathbb{E}\left[\max_k \exp\left(\lambda X_k\right)\right] &\leq \sum_k \mathbb{E}\left[\exp\left(\lambda X_k\right)\right] \\
&\leq n\exp(\sigma^2\lambda^2/2) = \exp(\log(n) + \sigma^2\lambda^2/2).
\end{aligned}$$

Thus, the application of Jensen's inequality yields that

$$\exp\left(\mathbb{E}\left[\lambda \max_k X_k\right]\right) \leq \mathbb{E}\left[\exp\left(\lambda \max_k X_k\right)\right] \leq \exp(\log(n) + \sigma^2\lambda^2/2),$$

which leads to

$$\mathbb{E}\left[\max_k X_k\right] \leq \frac{\log(n)}{\lambda} + \frac{\sigma^2\lambda}{2}.$$

Taking $\lambda = \sqrt{2\log(n)}/\sigma$ concludes the proof. ∎

It is evident that we cannot obtain a general low bound, for example, letting $X_1 = \cdots = X_n$. Nevertheless, the upper bound in the last lemma is indeed tight for independent Gaussian random variables. More details on lower bound will be provided in the sequel for suprema of random Gaussian processes.

Lemma 4.2 *Let $\{X_k\}_{k=1}^n$ be i.i.d $\mathcal{N}(0,\sigma^2)$ random variables. Then there exists a small absolute constant $c > 0$ such that*

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \geq c \cdot \sigma\sqrt{\log n}.$$

Proof: First we have

$$\begin{aligned}
\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] &= \mathbb{E}\left[\max\left\{\max_{k=1,\cdots,n} X_k, 0\right\}\right] + \mathbb{E}\left[\min\left\{\max_{k=1,\cdots,n} X_k, 0\right\}\right] \\
&\geq \mathbb{E}\left[\max\left\{\max_{k=1,\cdots,n} X_k, 0\right\}\right] + \mathbb{E}\left[\min\left\{X_1, 0\right\}\right] \\
&= \int_0^\infty \mathbb{P}\left[\max_{k=1,\cdots,n} X_k > t\right] dt + \mathbb{E}\left[\min\left\{X_1, 0\right\}\right] \\
&\geq \delta \cdot \mathbb{P}\left[\max_{k=1,\cdots,n} X_k > \delta\right] - \mathbb{E}\left[|X_1|\right] \\
&= \delta\left(1 - \left(\mathbb{P}\left[X_1 \leq \delta\right]\right)^n\right) - \mathbb{E}\left[|X_1|\right] \\
&= \delta\left(1 - \left(1 - \mathbb{P}\left[X_1 > \delta\right]\right)^n\right) - \mathbb{E}\left[|X_1|\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{P}\left[X_1 > \delta\right] &= \frac{1}{\sqrt{2\pi}\sigma}\int_\delta^\infty e^{-\frac{x^2}{2\sigma^2}} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_\delta^\infty e^{-\frac{(y+\delta)^2}{2\sigma^2}} dy \\
&\geq \frac{e^{-\delta^2/\sigma^2}}{c_1}
\end{aligned}$$

for some numerical constant $c_1 > 0$. Choosing $\delta = \sigma\sqrt{\log(n/c_1)}$ yields that $\mathbb{P}\left[X_1 > \delta\right] \geq 1/n$. Consequently,

$$\begin{aligned}
\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] &\geq \sigma\sqrt{\log(n/c_1)}(1 - (1 - 1/n)^n) - \sigma \\
&\geq (1 - 1/e)\sqrt{\log(n/c_1)}\sigma - \sigma,
\end{aligned}$$

which concludes the proof for sufficiently large n. ∎

4.2 Rademacher Complexity and Gaussian Complexity

This section studies $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ associated with Rademacher complexity and Gaussian complexity. Recall that give a set $T \subset \mathbb{R}^d$, the Rademacher complexity is defined as

$$\mathcal{R}(T) = \mathbb{E}\left[\sup_{t\in T} \langle \varepsilon, t\rangle\right], \quad \varepsilon = (\varepsilon_1, \cdots, \varepsilon_d).$$

while the Gaussian complexity of T is defined as

$$\mathcal{G}(T) = \mathbb{E}\left[\sup_{t\in T} \langle g, t\rangle\right], \quad g \sim \mathcal{N}(0, I_d),$$

Lemma 4.3 *We have*

$$\mathcal{R}(T) \lesssim \mathcal{G}(T) \lesssim \mathcal{R}(T)\sqrt{\log d}.$$

Proof: **Lower bound.** First we have

$$\begin{aligned}
\mathcal{R}(T) &= \mathbb{E}_\varepsilon\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k t_k\right] = \sqrt{\frac{\pi}{2}}\mathbb{E}_\varepsilon\left[\sup_{t\in T}\mathbb{E}_g\left[\sum_{k=1}^d \varepsilon_k |g_k| t_k\right]\right] \\
&\le \sqrt{\frac{\pi}{2}}\mathbb{E}_\varepsilon\left[\mathbb{E}_g\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k |g_k| t_k\right]\right] \\
&= \sqrt{\frac{\pi}{2}}\mathbb{E}\left[\sup_{t\in T}\sum_{k=1}^n g_k t_k\right] \\
&= \sqrt{\frac{\pi}{2}}\mathcal{G}(T),
\end{aligned}$$

where the third follows from the fact that $\varepsilon_k|g_k|$ has the same distribution with g_k.

Upper bound. To prove the upper bound, first note that the function

$$f(a_1, \cdots, a_d) := \mathbb{E}\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k t_k a_k\right]$$

is a convex function. Thus, the maximum of f over the region $\{(a_1, \cdots, a_d) : \ |a_k| \le 1, k = 1, \cdots, d\}$ must be achieved at the boundary. Then it follows that

$$\mathbb{E}\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k t_k a_k\right] \le \mathbb{E}\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k t_k\right] \quad \text{for all } |a_k| \le 1, k = 1, \cdots, d.$$

Consequently,

$$\begin{aligned}
\mathcal{G}(T) &= \mathbb{E}_g\left[\mathbb{E}_\varepsilon\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k |g_k| t_k\right]\right] \\
&= \mathbb{E}_g\left[\max_k |g_k| \cdot \mathbb{E}_\varepsilon\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k \frac{|g_k|}{\max_k |g_k|} t_k\right]\right]
\end{aligned}$$

$$\leq \mathbb{E}_g \left[\max_k |g_k| \cdot \mathcal{R}(T) \right]$$
$$\asymp \mathcal{R}(T)\sqrt{\log d},$$

as claimed. ■

Example 4.4 (Unit 2-norm ball $\mathbb{B}_2^d$) *Let $\mathbb{B}_2^d = \{t \in \mathbb{R}^d : \ \|t\|_2 \leq 1\}$. Then,*

$$\mathcal{R}(T) = \mathbb{E}\left[\sup_{\|t\|_2 \leq 1} \langle \varepsilon, t \rangle \right] = \mathbb{E}\left[\|\varepsilon\|_2 \right] = \sqrt{d}.$$

The same argument shows that

$$\mathcal{G}(T) = \mathbb{E}\left[\|g\|_2 \right] \leq \sqrt{\mathbb{E}\left[\|g\|_2^2 \right]} = \sqrt{d}.$$

Together with Lemma 4.3, we can conclude that $\mathcal{G}(\mathbb{B}_2^d) \asymp \sqrt{d}$. Therefore, the Rademacher complexity and the Gaussian complexity of $\mathbb{B}_2^d$ are essentially the same.

Example 4.5 (Unit 1-norm ball $\mathbb{B}_1^d$) *Let $\mathbb{B}_1^d = \{t \in \mathbb{R}^d : \ \|t\|_1 \leq 1\}$. Then,*

$$\mathcal{R}(T) = \mathbb{E}\left[\sup_{\|t\|_1 \leq 1} \langle \varepsilon, t \rangle \right] = \mathbb{E}\left[\|\varepsilon\|_\infty \right] = 1,$$

while

$$\mathcal{G}(T) = \mathbb{E}\left[\sup_{\|t\|_1 \leq 1} \langle g, t \rangle \right] = \mathbb{E}\left[\|g\|_\infty \right] \asymp \sqrt{\log d}.$$

Therefore, in this case, the Rademacher complexity and Gaussian complexity differ by the order $\sqrt{\log d}$. By Lemma 4.3, this difference turns out to be the worst possible.

4.3 Covering and Packing

For general random variables X_t and infinite number of elements in T, the first step can be made by approximating the supremum with a maximum of finite number of random variables. It should be not surprising that overall the bound for (4.2) should rely on the complexity or richness of the index set T. This section studies two closely related ways to measure the complexity of T, which indeed shows how to approximate T with a set of finite number of elements to achieve certain accuracy.

Definition 4.6 (ε-net and covering number) *Let (T, d) be a metric space. A set $N \subset T$ is called an ε-net of (T, d) if for every $t \in T$, there exists a $\pi(t) \in N$ such that $d(t, \pi(t)) \leq \varepsilon$. The covering number of (T, d), denoted $N(T, d, \varepsilon)$, is the smallest possible cardinality of an ε-net of (T, d). That is,*

$$N(T, d, \varepsilon) = \inf\{|N| : \ N \text{ is a } \varepsilon\text{-net of } (T, d)\}.$$

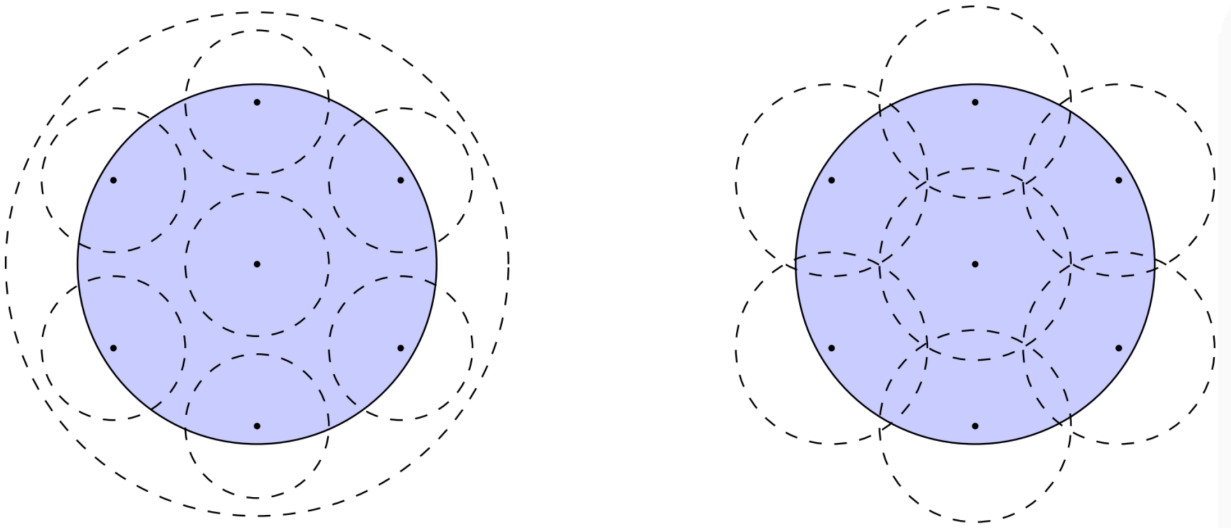


Figure 4.1: Covering (right) and packing (left) [2].

Note that N is a ε-net of T if and only if (see right of Figure 4.1)

$$T \subset \bigcup_{t \in N} B(t, \varepsilon), \quad \text{where } B(t, \varepsilon) = \{s \in T : \ d(t, s) \leq \varepsilon\}.$$

The covering number $N(T, d, \varepsilon)$ can be viewed as a measure of the complexity of T at the scale ε: The more complex T is, the more number of points we need to approximate it up to a certain precision. In addition, the logarithm of the covering number $\log N(T, d, \varepsilon)$ is often called the *metric entropy* of T as it is equivalent to the number of bits needed to encode every points of T up to a prescribed precision ε.

Example 4.7 *Consider the interval $T = [-1, 1]$ with the metric $d(t, t') = |t - t'|$. If we let*

$$N = \{t_k = -1 + 2(k-1)\varepsilon, \ k = 1, \cdots, k_{\max}\}$$

for the $k_{\max}$ such that $t_{k_{\max}} \leq 1$. it is not hard to see that N is an ε-net of T. Thus, we have

$$N(T, d, \varepsilon) \leq \frac{1}{\varepsilon} + 1.$$

Exercise 4.8 *Generalize the above result to the d-dimensional cube $T = [-1, 1]^d$ with $d(t, t') = \|t - t'\|_\infty$ and show that $N(T, d, \varepsilon) \leq \left(\frac{1}{\varepsilon} + 1\right)^d$.*

Definition 4.9 (ε-packing and packing number) *Let (T, d) be a metric space. A set $P \subset T$ is called an ε-packing of (T, d) if for every $t, t' \in P$ and $t \neq t'$, we have $d(t, t') > \varepsilon$. The packing number of (T, d), denoted $P(T, d, \varepsilon)$, is the largest possible cardinality of an ε-packing of (T, d). That is,*

$$P(T, d, \varepsilon) = \sup\{|P| : \ P \text{ is a } \varepsilon\text{-packing of } (T, d)\}.$$

The key idea, which was already hinted at above, is that the notion of packing is dual to the notion of covering (i.e., the typical primal-dual relationship between inf and sup), as given in the following lemma. This means that we can use covering and packing interchangeably. It is often the case that estimating one of them is easier than estimating the other in the applications.

Lemma 4.10 (Dual or equivalence between covering and packing) *For any $\varepsilon > 0$,*

$$P(T, d, 2\varepsilon) \leq N(T, d, \varepsilon) \leq P(T, d, \varepsilon).$$

Proof: **Upper bound.** Let P be a maximal ε-packing of (T, d). Then it is not hard to see that P is also a ε-net of (T, d); otherwise it will violates the maximality. The upper bound follows immediately.

Lower bound. Let $P = \{x_i\}$ be an 2ε-packing of (T, d) and let $N = \{y_j\}$ be a ε-net of (T, d). It can be argued that each closed $B(y_j, \varepsilon)$ ball can only contain one x_i due to the 2ε-separability of $\{x_i\}$. Since each x_i must be contained in one $B(y_j, \varepsilon)$, we must have $|P| \leq |N|$. The lower bound follows due to the arbitrary of P and N. ∎

The following lemma studies the covering of unit-norm balls in $\mathbb{R}^d$. The proof is based on a clever technique known as a volume argument.

Lemma 4.11 *Let $\|\cdot\|$ be a norm defined in $\mathbb{R}^d$ (e.g., 1-norm, 2-norm or infinity-norm). Let $\mathbb{B}^d$ be a unit $\|\cdot\|$ ball, i.e., $\mathbb{B}^d = \{t \in \mathbb{R}^d : \|t\| \leq 1\}$. Then*

$$\left(\frac{1}{\varepsilon}\right)^d \leq N(\mathbb{B}^d, \|\cdot\|, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^d.$$

Proof: **Lower bound.** Let $N = N(\mathbb{B}^d, \|\cdot\|, \varepsilon)$. We know that B can be covered with N balls of radius ε (see right of Figure 4.1 for an illustration under the 2-norm). Thus,

$$\mathrm{vol}(\mathbb{B}^d) \leq N \mathrm{vol}(\varepsilon \mathbb{B}^d).$$

Noting that $\mathrm{vol}(\varepsilon \mathbb{B}^d) = \varepsilon^d \mathrm{vol}(\mathbb{B}^d)$, the lower bound follows.

Upper bound. For the upper bound we consider $P = P(\mathbb{B}^d, \|\cdot\|, \varepsilon)$. Let $\{x_i\} \subset \mathbb{B}^d$ be the ε-packing of $\mathbb{B}^d$. Construct P balls $B(x_i, \varepsilon/2)$. Then we have

$$\bigcup_{x_i} B(x_i, \varepsilon/2) \subset \mathbb{B}^d + \frac{\varepsilon}{2}\mathbb{B}^d = \left(1 + \frac{\varepsilon}{2}\right)\mathbb{B}^d,$$

see the left of Figure 4.1. Thus, it follows that

$$P \cdot \mathrm{vol}\left(\frac{\varepsilon}{2}\mathbb{B}^d\right) \leq \mathrm{vol}\left(\left(1 + \frac{\varepsilon}{2}\right)\mathbb{B}^d\right).$$

It follows that $P \leq \left(1 + \frac{2}{\varepsilon}\right)^d$, and the upper bound follows by noting Lemma 4.10. ∎

The result in Lemma 4.11 for the unit 2-norm ball $\mathbb{B}_2^d$ will be very useful for studying the spectral norm of a random matrix. It follows that the metric entropy of $\mathbb{B}_2^d$ satisfies

$$\log N(\mathbb{B}_2^d, \|\cdot\|_2, \varepsilon) \sim d \log\left(\frac{1}{\varepsilon}\right),$$

which *scales linearly with respect to d*. For some spaces (of functions), the metric entropy scales exponentially with respect to d, hence suffering from the *curse of dimensionality*.

Remark 4.12 *Result and argument in Lemma 4.11 can be generalized to any set T in $\mathbb{R}^d$, see [3].*

So far, we have studied the covering number of various subsets of $\mathbb{R}^d$. Next, we turn to the metric entropy of Lipschitz functions. Consider the function class $\mathcal{F} = \{f \in \mathrm{Lip}\,([0,1], |\cdot|) : \ 0 \le f \le 1\}$. We have the following result.

Lemma 4.13 *There is a numerical constant $c > 0$ such that*

$$N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le e^{c/\varepsilon} \text{ for } \varepsilon < \frac{1}{2} \quad \text{and} \quad N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = 1 \text{ for } \varepsilon \ge \frac{1}{2}.$$

Proof: The claim $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = 1$ for $\varepsilon \ge \frac{1}{2}$ is trivial since $\left\|f - \frac{1}{2}\right\|_\infty \le \frac{1}{2}$ for each $f \in \mathcal{F}$. To prove the first one, we partition the horizontal axis into consecutive nonoverlapping intervals $I_1, \cdots, I_{\lceil 2/\varepsilon \rceil}$ of length $\varepsilon/2$, and partition the vertical axis into consecutive nonoverlapping intervals $J_1, \cdots, J_{\lceil 1/\varepsilon \rceil}$ of length ε, see Figure 4.2. For each $f \in \mathcal{F}$, we define $\pi(f)$ as follows:

$$\pi(f)(x) = \frac{\max J_\ell + \min J_\ell}{2} \quad \text{for } x \in I_k, \text{ where } J_\ell \text{ is the interval containing } f(\min I_k).$$

Let $N = \{\pi(f) : \ f \in \mathcal{F}\}$. We first show that N is an ε-net of $\mathcal{F}$. This follows from that $\forall x \in I_k$, we have

$$\begin{aligned} |f(x) - \pi(f)(x)| &\le |f(x) - f(\min I_k)| + \left| f(\min I_k) - \frac{\max J_\ell + \min J_\ell}{2} \right| \\ &\le |x - \min I_k| + \left| f(\min I_k) - \frac{\max J_\ell + \min J_\ell}{2} \right| \\ &\le \varepsilon. \end{aligned}$$

It remains to bound $|N|$. A trivial bound for $|N|$ is $|N| \le \lceil 1/\varepsilon \rceil^{\lceil 2/\varepsilon \rceil}$. If we explore the Lipschitz continuity of f more carefully, we can achieve the bound in the lemma, see [2] for details. ■

4.4 Finite Approximation Bound

At last, we consider the general case $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ and present a simple bound via one step of finite approximation. More precisely, the idea is approximating $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ by a finite maximum over a ε-net of T, together with the approximation error.

Theorem 4.14 *Assume X_t is σ^2-sub-Gaussian for every $t \in T$. Then,*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \inf_{\varepsilon > 0} \left\{ \mathbb{E}\left[\sup_{t\in T} \left(X_t - X_{\pi(t)}\right)\right] + \sqrt{2\sigma^2 \log N(T, d, \varepsilon)} \right\}.$$

Proof: Let $\varepsilon > 0$ and N be a ε-net of (T, d). For any t, let $\pi(t)$ be the point in N such that $d(t, \pi(t)) \le \varepsilon$. Then,

$$\sup_{t\in T} X_t = \sup_{t\in T} \left(X_t - X_{\pi(t)} + X_{\pi(t)}\right) \le \sup_{t\in T} \left(X_t - X_{\pi(t)}\right) + \sup_{t\in T} X_{\pi(t)}.$$

Taking the expectation on both sides and using the upper bound for the supremum of a finite number of sub-Gaussian random variables (Lemma 4.4 of Lecture 4) yields

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} \left(X_t - X_{\pi(t)}\right)\right] + \sqrt{2\sigma^2 \log N(T, d, \varepsilon)}.$$

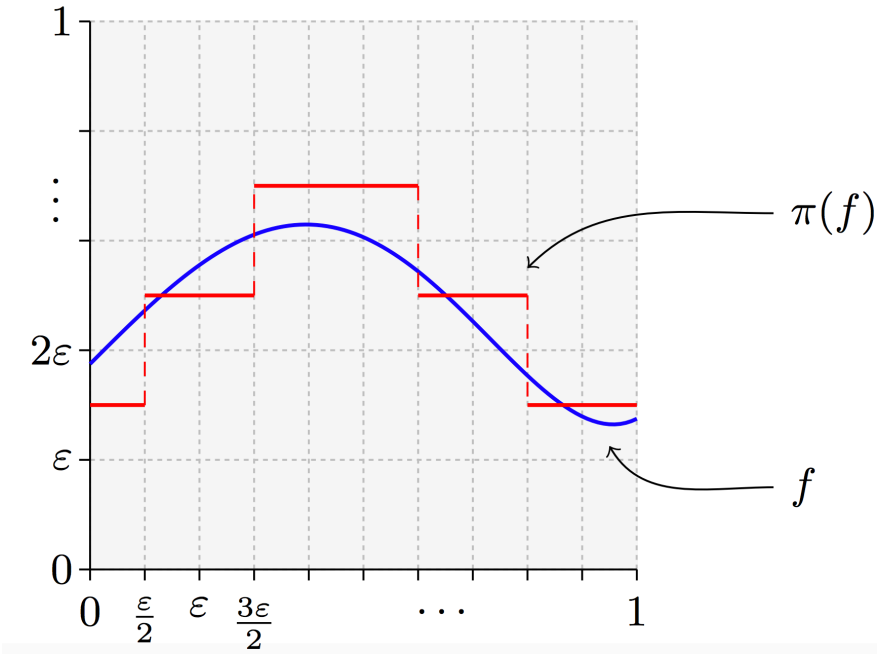


Figure 4.2: Illustration for Lemma 4.13 [2].

Noting that the first term indeed relies on ε since $d(t, \pi(t)) \leq \varepsilon$, taking the infimum over all $\varepsilon > 0$ concludes the proof. ■

There is a trade-off in the bound of Theorem 4.14. When ε decreases, the first term will potentially become smaller since X_t becomes closer to $X_{\pi(t)}$, but the second term increases as the covering number increases under a decreasing precision.

Example 4.15 (Maximum singular value of sub-Gaussian random matrix) *Let $W \in \mathbb{R}^{m \times n}$ be a random matrix with i.i.d σ^2-sub-Gaussian entries. We would like to estimate*

$$\mathbb{E}\left[\|W\|_2\right].$$

By the variational formula of the spectral norm,

$$\|W\|_2 = \sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} u^T W v,$$

*it is equivalent to bound $\mathbb{E}\left[\sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} u^T W v\right]$. Firstly note that $u^T W v$ is σ^2-sub-Gaussian for every $u \in \mathbb{B}_2^m$ and $v \in \mathbb{B}_2^n$ (**verify this!**). Let M be an ε-net of $\mathbb{B}_2^m$ and N be an ε-net of $\mathbb{B}_2^n$. By Theorem 4.14, we have*

$$\mathbb{E}\left[\|W\|_2\right] \leq \mathbb{E}\left[\sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} \left(u^T W v - \pi(u)^T W \pi(v)\right)\right] + \sqrt{2\sigma^2 \log|M||N|}, \quad \forall \varepsilon > 0.$$

- *By Lemma 4.11, we can choose $M \leq (1 + 2/\varepsilon)^m$ and $N \leq (1 + 2/\varepsilon)^n$.*

- *Because for any* $u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n$,

$$\begin{aligned} |u^T W v - \pi(u)^T W \pi(v)| &\le |(u - \pi(u))^T W v| + |\pi(u)^T W (v - \pi(v))| \\ &\le 2\varepsilon \|W\|_2, \end{aligned}$$

we have $\mathbb{E}\left[\sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} \left(u^T W v - \pi(u)^T W \pi(v)\right)\right] \le 2\varepsilon \mathbb{E}\left[\|W\|_2\right]$.

Combining all them together yields

$$\mathbb{E}\left[\|W\|_2\right] \le \frac{1}{1-2\varepsilon}\sqrt{2\sigma^2(m+n)\log(3/\varepsilon)}, \quad \forall \varepsilon > 0.$$

Taking ε *to be a small constant (e.g.,* $\varepsilon = 1/4$*) yields that*

$$\mathbb{E}\left[\|W\|_2\right] \lesssim \sigma(\sqrt{m} + \sqrt{n}).$$

As can be seen in the next lecture, this crude bound already captures the correct (tight) order of magnitude of the matrix norm.

A careful reader may find out that what have used in the above analysis is essentially the result

$$\|W\|_2 = \sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} u^T W v \le \frac{1}{1-2\varepsilon} \sup_{u\in M, v\in N} u^T W v.$$

Moreover, because the remaining term is of the same order with the target to bound but with a smaller factor, i.e.,

$$\mathbb{E}\left[\sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} \left(u^T W v - \pi(u)^T W \pi(v)\right)\right] \le 2\varepsilon \mathbb{E}\left[\|W\|_2\right],$$

optimal bound (in order) can be achieved for this case. But sometimes, the bound for the remaining term obtained via invoking the Lipschitz property is inefficient, so the finite approximate scheme only yields sub-optimal bound.

Example 4.16 (Uniform law of large numbers over Lipschitz functions (sub-optimal bound)) *Consider the metric space* $([0,1], |\cdot|)$ *and a probability measure* $\mathbb{P}$ *defined on it. Let* $\mathcal{F} = \{f \in \mathrm{Lip}\left([0,1], |\cdot|\right) : \ 0 \le f \le 1\}$. *Let* $X_1, \cdots, X_n \sim \mathbb{P}$ *be i.i.d samples. A key step when studying the uniform law of large numbers is to bound*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right] \right| \right].$$

For each of notation, let $Z_f = \frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]$. *Without loss of generality (**why we can make the simplification?**), we consider*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right].$$

*First note that Z_f is $1/4n$-sub-Gaussian since $f \in [0,1]$ (**check this!**). Let N be the ε-net of $(\mathcal{F}, \|\cdot\|_\infty)$, by Lemma 4.13, we have $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq e^{c/\varepsilon}$, for $\varepsilon < 1/2$. Thus, the application of Theorem 4.14 yields that*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right] \leq \inf_{0<\varepsilon<1/2}\left\{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(Z_f - Z_{\pi(f)}\right)\right] + \sqrt{\frac{c}{2n\varepsilon}}\right\}$$

Moreover, we have

$$\begin{aligned} Z_f - Z_{\pi(f)} &= \left(\frac{1}{n}\sum_{k=1}^n \left(f(X_k) - \pi(f)(X_k)\right)\right) + \mathbb{E}\left[\pi(f)(X) - f(X)\right] \\ &\leq \frac{2}{n}\sum_{k=1}^n \|f - \pi(f)\|_\infty \\ &\leq 2\varepsilon. \end{aligned}$$

It follows that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right] \leq \inf_{0<\varepsilon<1/2}\left\{2\varepsilon + \sqrt{\frac{c}{2n\varepsilon}}\right\} \asymp n^{-1/3}.$$

*However, this rate is **not** tight. Note that for a single function,*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]\right|\right] \leq \left(\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]\right)^2\right]\right)^{1/2} \lesssim n^{-1/2}.$$

*Thus, it would be more desirable if we can still get the same $n^{-1/2}$ rate for $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]\right|\right]$. In the next section, we will show that this is **indeed true**.*

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapters 5.1, 5.2 and 5.3.

[2] Ramon van Handel, *Probability in High Dimension*, Chapter 5.1, 5.2.

[3] Roman Vershynin, *High-Dimensional Probability: An introduction with applications in data science*, Chapter 4.2, 4.4.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 5: Expectation of Suprema: Chaining

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/04/10)*

Recap and motivation: Let $\{X_t : t \in T\}$ a set of σ_t^2-sub-Gaussian variables defined on a metric space (T, d) and assume $|T| < \infty$. We know that

$$\mathbb{E}\left[\max_{t\in T} X_t\right] \lesssim \max_t \sigma_t \sqrt{\log |T|}. \tag{5.1}$$

If X_t are independent and all the σ_t^2 are in the same order, for example $\sigma_t^2 \sim \sigma^2$, there is not much we can do since the bound $\sigma\sqrt{\log |T|}$ is tight. However, if σ_t^2 varies and X_t are related (all depending on t), it may be helpful to first cluster the random variables and bound $\mathbb{E}\left[\sup_{t_k\in T} X_{t_k}\right]$ based on the clusters. Suppose $X_t - X_s$ is $d(t,s)^2$-sub-Gaussian which reflects the closeness between X_t and X_s. We can cluster $\{X_t\}_{t\in T}$ based on the cluster of (T, d). Let T' be a set of representative points of $|T'|$ clusters ($|T'| < |T|$) and each point $t \in T$ is close to some point $\pi(t) \in T'$ within some accuracy $\sup_t d(t, \pi(t))$. Then, we immediately have

$$\begin{aligned}\mathbb{E}\left[\max_{t\in T} X_t\right] &\le \mathbb{E}\left[\max_{t\in T}\left(X_t - X_{\pi(t)}\right)\right] + \mathbb{E}\left[\max_{t'\in T'} X_{t'}\right] \\ &\lesssim \max_t d(t,\pi(t))\sqrt{\log |T|} + \max_{t'\in T'} \sigma_{t'}\sqrt{\log |T'|}.\end{aligned} \tag{5.2}$$

The number of terms in the maximum of the first term of (5.2) is the same with that in (5.1), but the sub-Gaussian parameter can be substantially small due to the clustering effect. The sub-Gaussian parameter in the maximum of the second term of (5.2) may be similar with that in (5.1), but the number of terms can be smaller. Overall, it is expected that the bound in (5.2) can improve the one in (5.1). *Of course, we can continue to do the same for the second term in* (5.2). *This leads to the chaining method, where the clustering is done through the ε-net.*

Agenda:

- The chaining method
- Examples
- Generic chaining

5.1 The Chaining Method

Definition 5.1 (Sub-Gaussian process) *A random process $\{X_t\}_{t\in T}$ defined on a metric space (T, d) is called sub-Gaussian if $\mathbb{E}[X_t] = 0$ and $X_t - X_s$ is $d(t,s)^2$-sub-Gaussian for all $t, s \in T$.*

Since the variations within the random process is determined by the metric space (T, d), it is expected to exploit the structure of (T, d) to bound $\mathbb{E}[\sup_t X_t]$. The chaining method will be our

focus of this lecture. It provides one way to exploit the structure of (T, d). (A more refined way (see [2] for example, equivalent to generic chaining which is presented in the last section but not required, may improve the chaining bound in some situations.)

Example 5.2 *Consider $X_t = \langle g, t\rangle$, where $g \in \mathcal{N}(0, I_d)$ and $t \in T \subset \mathbb{R}^d$. Since $X_t - X_s = \langle g, t - s\rangle$ is a $\|t - s\|_2^2$-Gaussian, $\{X_t\}_{t\in T}$ is certainly a sub-Gaussian process.*

Theorem 5.3 (Discrete Dudley inequality) *Let $\{X_t\}_{t\in T}$ be a separable sub-Gaussian process on the metric space (T, d). Then*[1]*,*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim \sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T, d, 2^{-k})}.$$

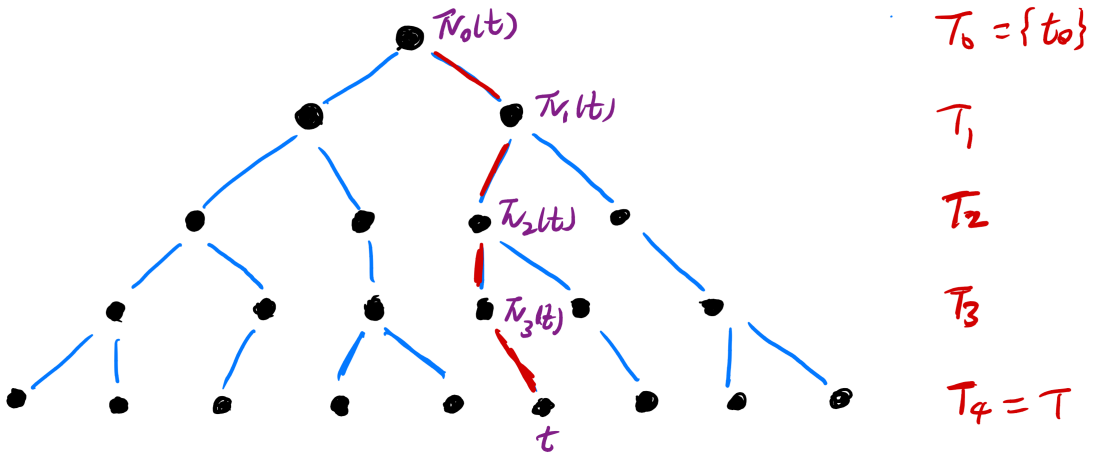


Figure 5.1: Illustration for chaining.

Proof: Without loss of generality we may assume $|T| < \infty$ since the separability of $\{X_t\}_{t\in T}$ implies that $\mathbb{E}\left[\sup_{t\in T} X_t\right] = \lim_{n\to\infty} \mathbb{E}\left[\sup_{t\in T_n} X_t\right]$ where T_n in an increasing finite subset of T.

Let $k_0 \in \mathbb{Z}$ such that $2^{-k_0} > \operatorname{diam}(T)$. Then any singleton $T_0 = \{t_0\}$ is an 2^{-k_0}-net of T. For $k > k_0$, let T_k be the 2^{-k}-net of T with covering number $N(T, d, 2^{-k})$. Moreover, since $|T| < \infty$, there exists a sufficiently large K such that $T_K = T$, see Figure 5.1. Thus, we have

$$X_t = X_{t_0} + \sum_{k=k_0+1}^{K} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right),$$

where $\pi_k(t)$ maps t to the nearest point in T_k. It follows that

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \sum_{k=k_0+1}^{K} \mathbb{E}\left[\sup_{t\in T}\left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right)\right]. \tag{5.3}$$

[1]The negative k in the sum denotes the approximation in the coarse (or large) scale with a small metric entropy. If $\operatorname{diam}(T) < \infty$, there exists a sufficiently small k_0 such that for all $k \le k_0$, $N(T, d, 2^k) = 1$ and thus $\log N(T, d, 2^k) = 0$.

First note that there are at most

$$|T_k||T_{k-1}| \leq |T_k|^2 = N(T,d,2^{-k})^2$$

terms in $\sup_{t\in T}\left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right)$. Moreover, since

$$d(\pi_k(t),\pi_{k-1}(t)) \leq d(\pi_k(t),t) + d(t,\pi_{k-1}(t)) \leq 3\times 2^{-k},$$

and $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ is $d(\pi_k(t),\pi_{k-1}(t))^2$-sub-Gaussian by the assumption, we have

$$\mathbb{E}\left[\sup_{t\in T}\left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right)\right] \lesssim 2^{-k}\sqrt{\log N(T,d,2^{-k})}.$$

Inserting this into (5.3) completes the proof. ∎

Remark 5.4 *A careful reader may find out that what we have actually established in Theorem 5.3 is that*

$$\mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right] \lesssim \sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})}.$$

This observation will be useful in one of the examples in the sequel.

Discrete Dudley inequality bounds $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ by a sum of (geometric structured) covering scales times the corresponding square root of metric entropies. The result can be written in an integral form since the sum can be viewed as a Riemann sum approximation to a certain integral.

Theorem 5.5 (Dudley integral) *Let $\{X_t\}_{t\in T}$ be a separable sub-Gaussian process on the metric space (T,d). Then*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}d\varepsilon.$$

Proof: The claim follows from

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})} &= 2\sum_{k\in\mathbb{Z}}\int_{2^{-(k+1)}}^{2^{-k}} \sqrt{\log N(T,d,2^{-k})}d\varepsilon \\
&\leq 2\sum_{k\in\mathbb{Z}}\int_{2^{-(k+1)}}^{2^{-k}} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon \\
&= 2\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}d\varepsilon,
\end{aligned}$$

which completes the proof. ∎

Remark 5.6 *In the proof we have shown that*

$$\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})} \leq 2\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}d\varepsilon.$$

Actually, we can also establish that

$$\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})} = \sum_{k\in\mathbb{Z}}\int_{2^{-k}}^{2^{-(k-1)}} \sqrt{\log N(T,d,2^{-k})}d\varepsilon \geq \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}d\varepsilon.$$

Thus nothing is lost in expressing the chaining bound as an integral rather than a sum, up to a constant factor.

Remark 5.7 . *It is worthing that we always have $N(T,d,\varepsilon) = 1$ when $\varepsilon \geq \mathrm{diam}(T)$. Thus, it is sufficient to take integral up to $\varepsilon = \mathrm{diam}(T)$.*

Remark 5.8 *It is not always the case that the bound by the Dudley integral is better than the one step discretization bound. Note that $N(T,d,\varepsilon)$ may approaches ∞ as ε approaches 0. Then the Dudley integral in an indefinite integral at the point 0. If $\sqrt{\log N(T,d,\varepsilon)}$ diverges very fast, the Dudley integral can be infinite. In this case the one step-discretization would still give a nontrivial bound even when the covering number is not integrable. Thus, sometimes, it is useful to combine the chaining method and the one step-discretization method to obtain a bound which mixes the Dudley integral (from a point strictly larger than zero) and the uniform one step-discretization bound, see for example Problem 5.11 in [2].*

5.2 Examples

5.2.1 Gaussian Complexity of $\mathbb{B}_2^d$

Recall that the Gaussian complexity of $\mathbb{B}_2^d$ is given by

$$\mathcal{G}(\mathbb{B}_2^d) = \mathbb{E}\left[\sup_{t\in\mathbb{B}_2^d}\langle g,t\rangle\right], \quad g\sim\mathcal{N}(0,I_d).$$

Letting $X_t = \langle g,t\rangle$, we known that $X_t - X_s$ is $\|t-s\|_2^2$-sub-Gaussian. Moreover, the covering number of $(\mathbb{B}_2^d, \|\cdot\|_2)$ at a scale $0<\varepsilon<1$ can be bounded by $(3/\varepsilon)^d$ (see Lecture 4). Thus, by the Dudley integral, we have

$$\begin{aligned}\mathcal{G}(\mathbb{B}_2^d) &\lesssim \int_0^1 \sqrt{\log N(\mathbb{B}_2^d, \|\cdot\|_2, \varepsilon)}d\varepsilon \\ &= \sqrt{d}\int_0^1 \sqrt{\log\frac{3}{\varepsilon}}d\varepsilon \lesssim \sqrt{d},\end{aligned}$$

which captures the correct order of $\mathcal{G}(\mathbb{B}_2^d)$, see Lecture 4.

5.2.2 Revisit Example 4.16

Consider the metric space $([0,1],|\cdot|)$ and a probability measure $\mathbb{P}$ defined on it. Let $\mathcal{F} = \{f\in \mathrm{Lip}\,([0,1],|\cdot|):\ 0\leq f\leq 1\}$ and let $X_1,\cdots,X_n\sim\mathbb{P}$ be i.i.d samples. We have shown in Example 4.16 of Lecture 4 that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]\right|\right] \lesssim n^{-1/3}$$

by the one-step discretization bound. In this example we improve this bound to the optimal $n^{-1/2}$ rate by the Dudley integral.

Define $Z_f = \frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}[f(X)]$. Then

$$Z_f - Z_g = \frac{1}{n}\sum_{k=1}^n (f(X_k) - g(X_k) - (\mathbb{E}[f(X_k)] - \mathbb{E}[g(X_k)])$$

is $\frac{1}{n}\|f-g\|_\infty^2$-sub-Gaussian. Thus, if we define

$$d(f,g) = n^{-1/2}\|f-g\|_\infty,$$

then $Z_f - Z_g$ is $d(f,g)^2$-sub-Gaussian. In addition, it is easily seen that (**check this!**)

$$N(\mathcal{F}, n^{-1/2}\|\cdot\|_\infty, \varepsilon) = N(\mathcal{F}, \|\cdot\|_\infty, n^{1/2}\varepsilon).$$

Thus, the application of the Dudley integral (also noting Remark 5.4 and $0 \in \mathcal{F}$) yields

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}[f(X)]\right|\right] &\lesssim \int_0^\infty \sqrt{\log N(\mathcal{F}, \|\cdot\|_\infty, n^{1/2}\varepsilon)}d\varepsilon \\
&= \frac{1}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)}d\varepsilon \\
&= \frac{1}{\sqrt{n}}\int_0^{1/2} \sqrt{\frac{c}{\varepsilon}}d\varepsilon \\
&\asymp \frac{1}{\sqrt{n}},
\end{aligned}$$

where the second line follows from the changes of variables, and the second to the last line follows from Lemma 4.13 of Lecture 4.

5.2.3 A Failure Example

Let $T = \left\{\frac{e_k}{\sqrt{1+\log k}} : k = 1, \cdots, n\right\}$, where e_k is the k-th canonical vector. Consider the Gaussian complexity of T,

$$\mathcal{G}(T) = \mathbb{E}\left[\sup_{t\in T}\langle g, t\rangle\right], \quad g \sim \mathcal{N}(0, I_d).$$

Note that $\mathcal{G}(T)$ can be explicitly written as

$$\mathcal{G}(T) = \mathbb{E}\left[\sup_{k=1,\cdots,n}\frac{g_k}{\sqrt{1+\log k}}\right].$$

Thus, it can be shown that there exists a universal constant $C > 0$ such that for all n,

$$\mathcal{G}(T) \le \mathbb{E}\left[\sup_{k=1,\cdots,n}\frac{|g_k|}{\sqrt{1+\log k}}\right]$$

$$
\begin{aligned}
&= \int_0^\infty \mathbb{P}\left[\sup_{k=1,\cdots,n} \frac{|g_k|}{\sqrt{1+\log k}} \geq t\right] dt \\
&= \int_0^a \mathbb{P}\left[\sup_{k=1,\cdots,n} \frac{|g_k|}{\sqrt{1+\log k}} \geq t\right] dt + \int_a^\infty \mathbb{P}\left[\sup_{k=1,\cdots,n} \frac{|g_k|}{\sqrt{1+\log k}} \geq t\right] dt \\
&\leq a + \sum_{k=1}^n \int_a^\infty \mathbb{P}\left[\frac{|g_k|}{\sqrt{1+\log k}} \geq t\right] dt \\
&\leq C \qquad (\textbf{complete this step by choosing } a \textbf{ properly!}) \ .
\end{aligned}
$$

However, we will show that the bound from the Dudley integral diverges as $n \to \infty$. Here, we consider the case $n = 2^{2^L}$. First note that the first m vectors in T is $1/\sqrt{\log m}$ separated. Thus, the packing number satisfies

$$P(T, \|\cdot\|_2, 1/\sqrt{\log m}) \geq m.$$

It follows that

$$
\begin{aligned}
&\int_0^\infty \sqrt{\log N(T, \|\cdot\|_2, \varepsilon)} d\varepsilon \\
&\geq \int_0^{\frac{1}{2\sqrt{\log(n)}}} \sqrt{\log N(T, \|\cdot\|_2, \varepsilon)} d\varepsilon \\
&+ \int_{\frac{1}{2\sqrt{\log(n)}}}^{\frac{1}{2\sqrt{\log\left(n^{1/2}\right)}}} \sqrt{\log N(T, \|\cdot\|_2, \varepsilon)} d\varepsilon \\
&+ \cdots\cdots \\
&+ \int_{\frac{1}{2\sqrt{\log\left(n^{1/2^{L-1}}\right)}}}^{\frac{1}{2\sqrt{\log\left(n^{1/2^L}\right)}}} \sqrt{\log N(T, \|\cdot\|_2, \varepsilon)} d\varepsilon \\
&\geq \int_0^{\frac{1}{2\sqrt{\log(n)}}} \sqrt{\log N\left(T, \|\cdot\|_2, \frac{1}{2\sqrt{\log n}}\right)} d\varepsilon \\
&+ \int_{\frac{1}{2\sqrt{\log(n)}}}^{\frac{1}{2\sqrt{\log\left(n^{1/2}\right)}}} \sqrt{\log N\left(T, \|\cdot\|_2, \frac{1}{2\sqrt{\log\left(n^{1/2}\right)}}\right)} d\varepsilon \\
&+ \cdots\cdots \\
&+ \int_{\frac{1}{2\sqrt{\log\left(n^{1/2^{L-1}}\right)}}}^{\frac{1}{2\sqrt{\log\left(n^{1/2^L}\right)}}} \sqrt{\log N\left(T, \|\cdot\|_2, \frac{1}{2\sqrt{\log\left(n^{1/2^L}\right)}}\right)} d\varepsilon \\
&\geq \int_0^{\frac{1}{2\sqrt{\log(n)}}} \sqrt{\log P\left(T, \|\cdot\|_2, \frac{1}{\sqrt{\log n}}\right)} d\varepsilon
\end{aligned}
$$

$$
\begin{aligned}
&+ \int_{\frac{1}{2\sqrt{\log(n)}}}^{\frac{1}{2\sqrt{\log\left(n^{1/2}\right)}}} \sqrt{\log P\left(T, \|\cdot\|_2, \frac{1}{\sqrt{\log\left(n^{1/2}\right)}}\right)} d\varepsilon \\
&+ \cdots\cdots \\
&+ \int_{\frac{1}{2\sqrt{\log\left(n^{1/2^{L-1}}\right)}}}^{\frac{1}{2\sqrt{\log\left(n^{1/2^{L}}\right)}}} \sqrt{\log P\left(T, \|\cdot\|_2, \frac{1}{\sqrt{\log\left(n^{1/2^{L}}\right)}}\right)} d\varepsilon \\
&\geq \frac{1}{2} + \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right) L \to \infty \text{ as } L \to \infty,
\end{aligned}
$$

where the third inequality follows from the relationship between covering number and packing number, and the last one uses the note that the first m vectors in T is $1/\sqrt{\log m}$ separated.

Thus, the Dudley inequality/integral is not able to capture the right bound for the Gaussian complexity of T in this example. Next we will present a method that works well for this example.

5.3 Generic Chaining

Before introducing generic chaining, we first reformulate the Dudley inequality into an equivalent form. To this end, we need to give the definition of *admissible sequence.* Let $\{T_k\}_{k=1}^{\infty}$ be a sequence of subsets of T. If

$$
|T_0| = 1, \quad |T_k| \leq 2^{2^k}, \ k = 1, 2, \cdots \tag{5.4}
$$

$\{T_k\}_{k=1}^{\infty}$ is called an *admissible sequence.*

Lemma 5.9 *We have*

$$
\int_0^{\infty} \sqrt{\log N(T, d, \varepsilon)} d\varepsilon \asymp \inf_{\{T_k\}_{k=0}^{\infty}} \sum_{k=0}^{\infty} 2^{k/2} \sup_{t \in T} d(t, T_k), \tag{5.5}
$$

where the infimum is taken over all the admissible sequence satisfying (5.4)*.*

Proof: First note that the righthand side (5.5) is equivalent to

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

Then letting $e_k(T) = \inf_{T_k} \sup_{t \in T} d(t, T_k)$, one can easily see that

$$
e_0(T) = \inf\{\varepsilon : \ N(T, d, \varepsilon) = 1\}, \quad e_k(T) = \inf\{\varepsilon : \ N(T, d, \varepsilon) \leq 2^{2^k}\} \text{ for } k \geq 1.
$$

Therefore, for $\varepsilon < e_k(T)$, $N(T, d, \varepsilon) \geq 2^{2^k} + 1$. It follows that

$$
\int_{e_{k+1}(T)}^{e_k(T)} \sqrt{\log N(T, d, \varepsilon)} d\varepsilon \gtrsim 2^{k/2}(e_k(T) - e_{k+1}(T))
$$

Consequently,

$$\begin{aligned}\int_0^{e_0(T)} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon &\gtrsim \sum_{k=0}^{\infty} 2^{k/2}(e_k(T)-e_{k+1}(T)) \\ &= \sum_{k=0}^{\infty} 2^{k/2}e_k(T) - \sum_{k=1}^{\infty} 2^{(k-1)/2}e_k(T) \\ &\geq \left(1-\frac{1}{\sqrt{2}}\right)\sum_{k=0}^{\infty} 2^{k/2}e_k(T),\end{aligned}$$

which completes the proof of one direction.

For the other direction, we have

$$\begin{aligned}\int_0^{\infty} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon &= \int_0^{e_0(T)} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon \\ &= \sum_{k=0}^{\infty}\int_{e_{k+1}(T)}^{e_k(T)} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon \\ &\leq \sum_{k=0}^{\infty} 2^{(k+1)/2}(e_k(T)-e_{k+1}(T)) \\ &\lesssim \sum_{k=0}^{\infty} 2^k e_k(T).\end{aligned}$$

Now the proof is complete. ∎

Remark 5.10 *The derivation above also reveals why it requires $|T_k| \leq 2^{2^k}$ in the admissible sequence. Basically, we would like to have a matching lower and upper bound for Dudely integral in the form of the righthand of* (5.5)*. See Remark 6.28 of [2] for details.*

The generic chaining will allow us to pull the supremum outside the sum and thus leads to a potentially smaller bound.

Theorem 5.11 (Generic chaining) *Let $\{X_t\}_{t\in T}$ be a separable sub-Gaussian process on the metric space (T,d). Then*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim \gamma(T,d) := \inf_{\{T_k\}_0^{\infty}} \sup_{t\in T} \sum_{k=0}^{\infty} 2^{k/2} d(t,T_k), \tag{5.6}$$

where the infimum is taken over all the admissible sequences.

Proof: As before, we can still assume $|T| < \infty$. Thus, it holds that

$$X_t - X_{t_0} = \sum_{k=1}^{K} \left(X_{\pi_k(t)} - X_{\pi_{k-1}}(t)\right),$$

where $\pi_k(t)$ maps t to the closest point in T_k. The overall goal is to show that

$$\mathbb{P}\left[\sup_{t\in T} |X_t - X_{t_0}| \gtrsim u\gamma(T,d)\right] \lesssim \exp\left(-u^2/2\right) \quad \text{for } u \geq c,$$

where $c > 0$ is an absolute constant. The claim will then follow immediately.

To this end, we will consider each term in the chaining sum and then take a uniform bound. Because $|X_{\pi(t)} - X_{\pi_{k-1}(t)}|$ is $d(\pi_k(t), \pi_{k-1}(t))^2$-sub-Gaussian, we have

$$\mathbb{P}\left[|X_{\pi(t)} - X_{\pi_{k-1}(t)}| \geq Cu2^{k/2} d(\pi_k(t), \pi_{k-1}(t))\right] \leq 2\exp\left(-u^2 2^k\right), \tag{5.7}$$

where $C > 0$ is an absolute and fixed constant. Let Ω_u be the event such that

$$|X_{\pi(t)} - X_{\pi_{k-1}(t)}| \leq Cu2^{k/2} d(\pi_k(t), \pi_{k-1}(t)) \quad \text{for all } t \in T \text{ and } k.$$

Since there are at most $|T_k||T_{k-1}|$ terms in $|X_{\pi(t)} - X_{\pi_{k-1}(t)}|$, we have

$$\mathbb{P}\left[\Omega_u^c\right] \leq 2\sum_{k\geq 1} 2^{2^{k+1}} \exp\left(-u^2 2^k\right).$$

Note that whenever Ω_u occurs, we have[2]

$$\sup_{t\in T} |X_t - X_{t_0}| \leq Cu \sup_{t\in T} \sum_{k=1}^{K} 2^{k/2} d(\pi_k(t), \pi_{k-1}(t)).$$

Consequently,

$$\mathbb{P}\left[\sup_{t\in T} |X_t - X_{t_0}| \geq Cu \sup_{t\in T} \sum_{k=1}^{K} 2^{k/2} d(\pi_k(t), \pi_{k-1}(t))\right] \leq 2\sum_{k\geq 1} 2^{2^{k+1}} \exp\left(-u^2 2^k\right)$$

Noting that $d(\pi_k(t), \pi_{k-1}(t)) \leq d(t, T_k) + d(t, T_{k-1})$, we have

$$\mathbb{P}\left[\sup_{t\in T} |X_t - X_{t_0}| \gtrsim u\gamma(T, d)\right] \leq 2\sum_{k\geq 1} 2^{2^{k+1}} \exp\left(-u^2 2^k\right).$$

Thus, it only remains to bound $\sum_{k\geq 1} 2^{2^{k+1}} \exp\left(-u^2 2^k\right)$. Noting that

$$u^2 2^k \geq u^2/2 + u^2 2^{k-1} \geq u^2/2 + 2^{k+1}$$

for $u \geq 2$, one can easily obtain that $\sum_{k\geq 1} 2^{2^{k+1}} \exp\left(-u^2 2^k\right) \lesssim \exp\left(-u^2/2\right)$. ∎

It is worth noting that the tail bound version of the Dudley integral and the generic chaining can also be established, see for example [2] or [3]. The generic chaining bound is not as convenient to use as the Dudley integral since constructing a good admissible sequence is not always easy. However, the difference between the generic chaining bound and the Dudley integral can look minor, but sometimes it is real. To see this, we revisit the failure example for the Dudley integral and still consider the case $n = 2^{2^L}$. For notational convenience, we let $t_k = e_k/\sqrt{1 + \log k}$. We construct an admissible sequence as follows:

$$T_0 = \{t_n\}, \quad T_k = \{t_2, \cdots, t_{2^{2^k}}, t_n\}, \ k = 1, \cdots, L-1.$$

[2]Basically, the argument here tensorize well without first triggering the sup in the first place.

Then give any $t \in T$, there exists a K such that $t_{2^{2^K}} < t \leq t_{2^{2^{K+1}}}$. It follows that

$$\sum_{k=0}^{\infty} 2^{k/2} d(t, T_k) = \sum_{k=0}^{K} 2^{k/2} d(t, T_k) \lesssim \sum_{k=0}^{K} 2^{(k-K)/2} = O(1).$$

Here t_n is included in T_k in order for $d(t, T_k) \asymp 2^{-K/2},\ k \leq K$ (independent of k). Because t is arbitrary, we can conclude that the generic chaining can capture the right magnitude in this special example.

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapter 5.3.

[2] Ramon van Handel, *Probability in High Dimension*, Chapter 5.3.

[3] Roman Vershynin, *High-Dimensional Probability: An introduction with applications in data science*, Chapter 8.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 6: Uniform Law of Large Numbers

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/04/23)*

Motivation: We are interested in bounding the random variable

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n}f(X_k)-\mathbb{E}\left[f(X)\right]\right|, \tag{6.1}$$

where $\mathcal{F}$ is a class of functions. That is, we want to estimate the deviation between $\frac{1}{n}\sum_{k=1}^{n}f(X_k)$ and $\mathbb{E}\left[f(X)\right]$ uniformly over the class $\mathcal{F}$ – hence the name of uniform laws of large numbers. Here, we ignore the measurability issue after taking the supremum. It is worth noting that we can view (6.1) as the (random) distance between the empirical probability measure $\mathbb{P}_n := \frac{1}{n}\sum_{k=1}^{n}\delta_{X_k}$ and the probability measure $\mathbb{P}$ where the distance is defined in the linear functional/operator sense (cf. Wasserstein distance where $\mathcal{F}$ is a particular class of functitons).

Recall that the quantity in (6.1) plays an important role in the generalization analysis of empirical risk minimization methods. Apart from this, it also has a close connection with the classical Glivenko-Cantelli theorem. Letting $X\sim\mathbb{P}$, the cumulative distribution function (CDF) $F(a)$ is given by $F(a)=\mathbb{P}\left[X\le a\right]$. Given a set of i.i.d samples $\{X_k\}_{k=1}^n$, we can estimate F by the empirical CDF,

$$\widehat{F}_n(a)=\frac{1}{n}\sum_{k=1}^{n}1_{(-\infty,a]}(X_k),$$

i.e., the empirical frequency over $(-\infty,a]$. Then it is natural to ask whether

$$\left|\widehat{F}_n(a)-F(a)\right| \text{ is small uniformly for all } a\in\mathbb{R}?$$

Letting $\mathcal{F}=\{1_{(-\infty,a]}(x):\ a\in\mathbb{R}\}$, since $\mathbb{E}\left[1_{(-\infty,a]}(X)\right]=F(a)$, we actually need to bound (6.1), where $\mathcal{F}$ is given by

$$\mathcal{F}=\{1_{(-\infty,a]},\ a\in\mathbb{R}\}. \tag{6.2}$$

Under some proper conditions (e.g., $\|f\|_\infty\le b$ for $f\in\mathcal{F}$), it is easy to show that the quantity in (6.1) concentrates around its mean, for example by bounded difference property. Thus, we only pay attention to its expectation

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right]. \tag{6.3}$$

In particularly, we focus on the case when $\mathcal{F}$ is a set of binary value functions in (6.1), with the Glivenko-Cantelli theorem as a special example. Define

$$d(f,g)=n^{-1/2}\|f-g\|_\infty.$$

By the discussion in Lecture 5.2.2, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n}f(X_k)-\mathbb{E}\left[f(X)\right]\right|\right] \lesssim \int_0^\infty \sqrt{\log N(\mathcal{F},\|\cdot\|_\infty, n^{1/2}\varepsilon)}d\varepsilon$$
$$= \frac{1}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\mathcal{F},\|\cdot\|_\infty,\varepsilon)}d\varepsilon. \tag{6.4}$$

However, for $\mathcal{F}$ given in (6.2), since

$$\|1_{(-\infty,a]}-1_{(-\infty,a']}\|_\infty = 1 \quad \text{whenever} \quad a\neq a',$$

we have $N(\mathcal{F},\|\cdot\|_\infty,\varepsilon)=\infty$ for $\varepsilon<1$. Thus, the bound in (6.4) is not quite meaningful. The symmetrization argument provides a way to overcome this pitfall, which allows us to use the Dudley integral based on covering under potentially a smaller distance through separating the sign (or "Gaussian part") out from its magnitude .

To motivate the symmetrization argument, consider the random variable $\sum_{k=1}^n X_k$ where X_k are independent mean zero random variables. When the magnitude of each X_k is of the order $O(1)$, a naive bound for $|\sum_{k=1}^n X_k|$ would be $O(n)$. However, by the central limit theorem, a more desirable bound would be $O(\sqrt{n})$. This is due to that the terms in the sum are independent and centered, so they are likely to have opposite signs, yielding the cancellation effect. Therefore, the random sign $\sum_{k=1}^n \operatorname{sign}(X_k)$ plays an essential role in the Gaussian tail while the magnitudes of X_k only determine the variance.

Agenda:

- Symmetrization
- VC Dimension and Sauer-Shelah Lemma
- Classical Glivenko-Cantelli Theorem

6.1 Symmetrization

As already mentioned, the symmetrization technique separates the sign (or "Gaussian part") of the process out from its magnitude and analyze each part sequentially. This allows us to provide bounds for (6.1) more efficiently.

Lemma 6.1 (Upper bound by symmetrization) *Let $\{X_k\}_{k=1}^n$ be i.i.d random variables. Then,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right] \leq 2\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k f(X_k)\right|\right],$$

where $\{\varepsilon_k\}_{k=1}^n$ is a collection of i.i.d Rademacher random variables.

Proof: Let $\{Y_k\}_{k=1}^n$ be i.i.d copies of $\{X_k\}_{k=1}^n$. We have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X)\right]\right)\right|\right] = \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}_Y\left[f(Y_k)\right]\right)\right|\right]$$

$$
\begin{aligned}
&= \mathbb{E}_X \left[\sup_{f\in\mathcal{F}} \left| \mathbb{E}_Y \left[\sum_{k=1}^n (f(X_k) - f(Y_k)) \right] \right| \right] \\
&\leq \mathbb{E}_X \left[\sup_{f\in\mathcal{F}} \mathbb{E}_Y \left[\left| \sum_{k=1}^n (f(X_k) - f(Y_k)) \right| \right] \right] \\
&\leq \mathbb{E}_{X,Y} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n (f(X_k) - f(Y_k)) \right| \right].
\end{aligned}
$$

where the third line follows from Jensen inequality. Noting that $f(X_k) - f(Y_k)$ is symmetric and thus has the same distribution with $\varepsilon_k(f(X_k) - f(Y_k))$, it follows that

$$
\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n (f(X_k) - \mathbb{E}\left[f(X)\right]) \right| \right] &\leq \mathbb{E}_{X,Y,\varepsilon} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n (\varepsilon_k(f(X_k) - f(Y_k))) \right| \right] \\
&\leq \mathbb{E}_{X,\varepsilon} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n \varepsilon_k f(X_k) \right| \right] + \mathbb{E}_{Y,\varepsilon} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n \varepsilon_k f(Y_k) \right| \right],
\end{aligned}
$$

which completes the proof since $\{Y_k\}_{k=1}^n$ are i.i.d copies of $\{X_k\}_{k=1}^n$. ■

Lemma 6.2 (Lower bound by symmetrization) *Let $\{X_k\}_{k=1}^n$ be i.i.d random variables. Then,*

$$
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n (f(X_k) - \mathbb{E}\left[f(X_k)\right]) \right| \right] \geq \frac{1}{2} \mathbb{E}_{X,\varepsilon} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n \varepsilon_k \left(f(X_k) - \mathbb{E}\left[f(X_k)\right]\right) \right| \right],
$$

where $\{\varepsilon_k\}_{k=1}^n$ is a collection of i.i.d Rademacher random variables.

Proof: We have

$$
\begin{aligned}
&\mathbb{E}_{X,\varepsilon} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n \varepsilon_k \left(f(X_k) - \mathbb{E}_X\left[f(X_k)\right]\right) \right| \right] \\
&= \mathbb{E}_{X,\varepsilon} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n \varepsilon_k \left(f(X_k) - \mathbb{E}_Y\left[f(Y_k)\right]\right) \right| \right] \\
&\leq \mathbb{E}_{X,Y,\varepsilon} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n \varepsilon_k \left(f(X_k) - f(Y_k)\right) \right| \right] \\
&= \mathbb{E}_{X,Y} \left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n (f(X_k) - f(Y_k)) \right| \right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n (f(X_k) - \mathbb{E}\left[f(X_k)\right]) \right| \right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \sum_{k=1}^n (f(Y_k) - \mathbb{E}\left[f(Y_k)\right]) \right| \right],
\end{aligned}
$$

which completes the proof since $\{Y_k\}_{k=1}^n$ are i.i.d copies of $\{X_k\}_{k=1}^n$. ■

Remark 6.3 *Note the right hand side in Lemma 6.2 cannot be replaced by $\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}} \left|\sum_{k=1}^n \varepsilon_k f(X_k)\right|\right]$ since a counter example can be easily constructed for the $n = 1$ case.*

To upper bound (6.3), by Lemma 6.1, it suffices to bound

$$\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k f(X_k)\right|\right]. \tag{6.5}$$

For this, we can first condition on $X=(x_1,\cdots,x_n)$ and bound

$$\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k f(x_k)\right|\right] \tag{6.6}$$

and then take expectation with respect to X. It follows that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right]\lesssim\sqrt{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}f^2(X_k)\right]}\sqrt{\log\Pi_{\mathcal{F}}(n)}, \tag{6.7}$$

where

$$\Pi_{\mathcal{F}}(n):=\max_{\{x_1,\cdots,x_n\}\subset\mathcal{X}}\left|\left\{(f(x_1),\cdots,f(x_n)):\ f\in\mathcal{F}\right\}.\right| \tag{6.8}$$

Note that when $|\mathcal{F}|=\infty$ in which case a direct bound uniform bound for (6.3) fails. In contrast, it is possible that $\Pi_{\mathcal{F}}(n)$ is finite (e.g., when $\mathcal{F}$ a class of binary value functions or for classification problems). In this case we can still work out an upper bound for (6.3) through (6.6) and obtain

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right]\lesssim\sqrt{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}f^2(X_k)\right]}\sqrt{\log\Pi_{\mathcal{F}}(n)}\leq\sqrt{n}b\sqrt{\log\Pi_{\mathcal{F}}(n)}. \tag{6.9}$$

Next we will focus on the case when $\mathcal{F}$ a class of binary value functions (and hence $\Pi_{\mathcal{F}}(n)$ is finite) It can be shown that the growth of $\Pi_{\mathcal{F}}(n)$ is determined by a notion called VC dimension. In other words, VC dimension provides a different way to quantify the complexity of the function class $\mathcal{F}$. Though we only discuss the VC dimension for the families of binary value functions, it can be extended to general classes of functions, see for example Chapter 7.3 of [2].

6.2 VC Dimension and Sauer-Shelah Lemma

Definition 6.4 (Shattering and VC dimension) *Let $\mathcal{F}$ be a class of binary value functions. We say a set $(x_1,\cdots,x_n)\subset\mathcal{X}$ is shattered by $\mathcal{F}$ if*

$$\left|\left\{(f(x_1),\cdots,f(x_n)):\ f\in\mathcal{F}\right\}\right|=2^n.$$

*The VC dimension of $\mathcal{F}$, denoted $v(\mathcal{F})$ or simply v for short, is defined as the largest integer n for which **there exists** a collection of points $(x_1,\cdots,x_n)$ that is shattered by $\mathcal{F}$.*

Remark 6.5 *By the definition, when $n>v$, then for any collection of points $(x_1,\cdots,x_n)$,*

$$\left|\left\{(f(x_1),\cdots,f(x_n)):\ f\in\mathcal{F}\right\}\right|$$

must be exactly smaller than 2^n. In terms of the growth function in (6.8)*, the VC dimension is the largest integer n such that $\Pi_{\mathcal{F}}(n)=2^n$.*

Exercise 6.6 *If there exists n points that can be shattered, why for any $m < n$ there exists m points hat can also be shattered? If there does not exist n points that can be shattered, why for any $m > n$ there does not exist m points that can be shattered?*

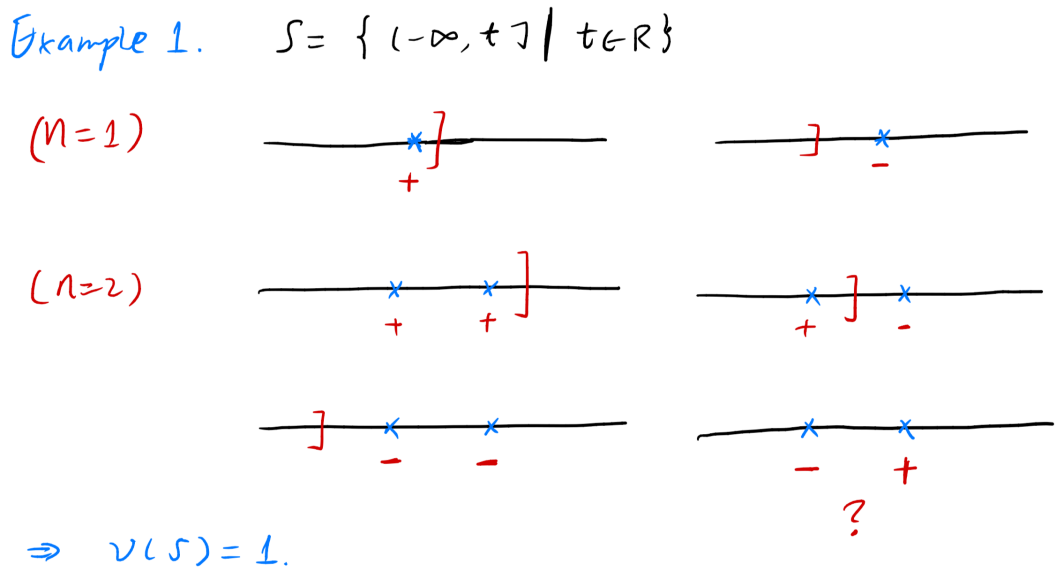


Figure 6.1: Example I

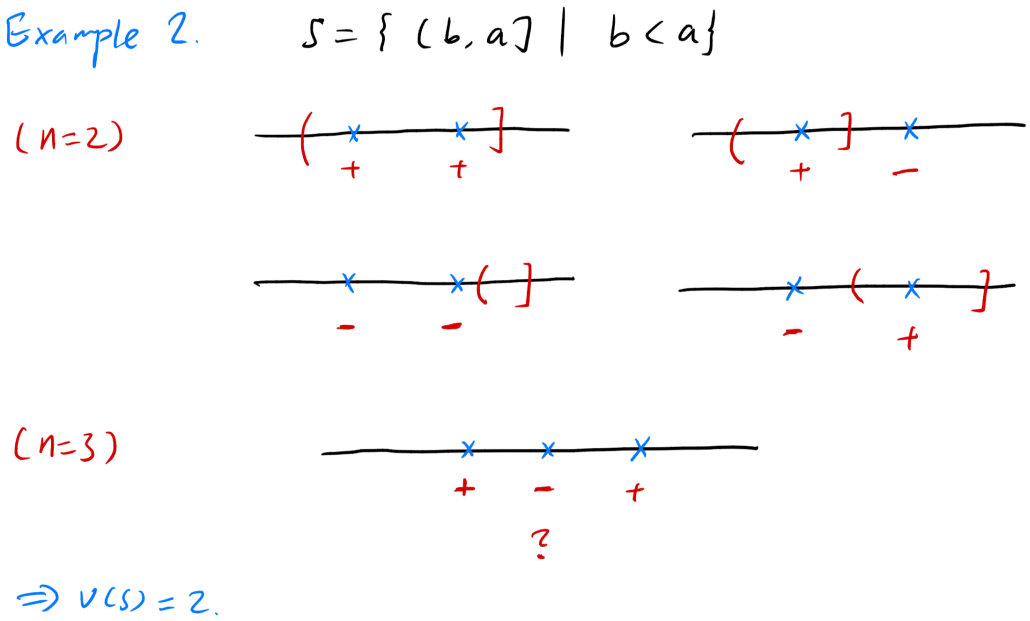


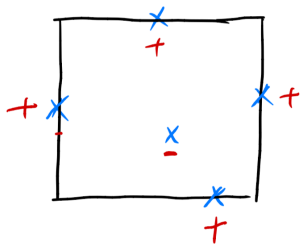
Figure 6.2: Example II

Example 3. $S = \{[a,b] \times [c,d] \mid a \le b,\ c \le d\}$

($n=4$) there exist four points on $\mathbb{R}^2$ that can be shattered.

[Try this]

($n=5$) For any given five points in $\mathbb{R}^2$, first find the smallest rectangle that contains all the points. Set the point inside the rectangle to be '−', and the others to be '+',



This configuration cannot be realized by S.

$\Rightarrow$ $v(S) = 4$.

Figure 6.3: Example III

Example 6.7 *Figures 6.1, 6.2 and 6.3 give three examples with finite VC dimension, where*

$$\mathcal{F} = \{1_S(x),\ S \in \mathcal{S}\}.$$

There also exists set $\mathcal{S}$ such that the VC dimension of $\mathcal{F}$ is infinite, see [3].

For the function class having a finite VC dimension, it turns out its growth function is of the polynomial order in n.

Lemma 6.8 (Sauer-Shelah) *For all $n \ge v$ and $(x_1, \cdots, x_n) \subset \mathcal{X}$, there holds*

$$\Pi_{\mathcal{F}}(n) := \max_{\{x_1,\cdots,x_n\}\subset\mathcal{X}} |\{(f(x_1), \cdots, f(x_k)) :\ f \in \mathcal{F}\}| \le \sum_{k=0}^{v} \binom{n}{k} \le \left(\frac{en}{v}\right)^v.$$

Proof: The second inequality follows directly from the combinatorial argument

$$\sum_{k=0}^{v} \binom{n}{k} \le \sum_{k=0}^{v} \binom{n}{k} \left(\frac{n}{v}\right)^{v-k}$$

$$
\begin{aligned}
&\leq \sum_{k=0}^{n} \binom{n}{k} \left(\frac{n}{v}\right)^{v-k} \\
&= \left(\frac{n}{v}\right)^{v} \sum_{k=0}^{n} \binom{n}{k} \left(\frac{v}{n}\right)^{k} \\
&= \left(\frac{n}{v}\right)^{v} (1+v/n)^{n} \\
&\leq \left(\frac{en}{v}\right)^{v}
\end{aligned}
$$

The first inequality follows from an inductive argument and the details will be omitted. Interested readers may find them in [1] and [3]. ■

Note that Lemma 6.8 is a truly deep result. For $n > v(\mathcal{F})$, though the definition of VC dimension implies that $|\{(f(x_1), \cdots, f(x_n)) : \ f \in \mathcal{F}\}| < 2^n$ for any $(x_1, \cdots, x_n)$, this does not exclude the possibility that there exists a $(x_1, \cdots, x_n)$ such that $|\{(f(x_1), \cdots, f(x_n)) : \ f \in \mathcal{F}\}| = 2^n - 1$. However, the Sauer-Shelah lemma says that this cannot be true.

Exercise 6.9 *For the three examples in Example 6.7, show that* $|\{(f(x_1), \cdots, f(x_n)) : \ f \in \mathcal{F}\}| \lesssim n^v$ *directly rather that using the Sauer-Shelah lemma.*

6.3 Classical Glivenko-Cantelli Theorem

In this section we return back to the problem of estimating $\mathbb{E}\left[\|\widehat{F}_n - F\|_\infty\right]$, where F and $\widehat{F}_n$ are CDF and empirical CDF, respectively. It corresponds to estimating (6.3) for $\mathcal{F} = \{1_{(-\infty,a]}(x) : \ a \in \mathbb{R}\}$. By Example 6.7, we first know that $v(\mathcal{F}) = 1$. It follows from the Sauer-Shelah lemma that $\Pi_n(\mathcal{F}) \lesssim n$. Together with (6.9), we have

$$
\mathbb{E}\left[\|\widehat{F}_n - F\|_\infty\right] \lesssim \sqrt{\frac{\log n}{n}}, \tag{6.10}
$$

Remark 6.10 *By certain central limit theorem (Kolmogorov theorem), one can directly show that the optimal rate for* $\|\widehat{F}_n - F\|_\infty$ *is* $1/\sqrt{n}$*. Next, we will remove the log-factor in* (6.10) *by more advanced technique.*

Let $\mathcal{F} = \{1_C, \ C \subset \mathcal{X}\}$ be the set of binary value functions defined on a probability space $(\mathcal{X}, \mathbb{P})$. For any $f, g \in \mathcal{F}$, we define

$$
\|f - g\|_{L^2(\mathbb{P})} = \left(\int_{\mathcal{X}} (f(x) - g(x))^2 d\mathbb{P}(x)\right)^{1/2}.
$$

Lemma 6.11 *There is a numerical constant* $c > 0$ *such that*

$$
N(\mathcal{F}, \|\cdot\|_{L^2(\mathbb{P})}, \varepsilon) \leq \left(\frac{c}{\varepsilon}\right)^{cv} \quad \text{for } \varepsilon < 1.
$$

where v *is the VC dimension of* $\mathcal{F}$*.*

The proof of Lemma 6.11 relies on the following lemma.

Lemma 6.12 *Let $f_1, \cdots, f_n$ be functions on $(\mathcal{X}, \mathbb{P})$. If*

$$\|f_i\|_\infty \le 1, \quad \|f_i - f_j\|_{L^2(\mathbb{P})} > \varepsilon \quad \textit{for all } i \neq j,$$

then there exists $m \asymp \varepsilon^{-4} \log n$ points $x_1, \cdots, x_m$ such that

$$\frac{1}{m} \sum_{k=1}^{m} |f_i(x_k) - f_j(x_k)|^2 > \varepsilon^2/4 \quad \textit{for all } i \neq j. \tag{6.11}$$

Proof: The proof of this lemma uses a very interesting probabilistic argument: we first choose m points randomly and then show (6.11) holds with high probability. Then there must exist such m deterministic points. More precisely, let $X_1, \cdots, X_m \sim \mathbb{P}$ be i.i.d samples. The application of Hoeffding inequality implies that

$$\mathbb{P}\left[\frac{1}{m}\sum_{k=1}^{m}\left(|f_i(X_k) - f_j(X_k)|^2 - \mathbb{E}\left[|f_i(X_k) - f_j(X_k)|^2\right]\right) \le -t\right] \le \exp\left(-\frac{mt^2}{2}\right).$$

Noting that

$$\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}|f_i(X_k) - f_j(X_k)|^2\right] = \mathbb{E}\left[|f_i(X_k) - f_j(X_k)|^2\right] = \|f_i - f_j\|^2_{L^2(\mathbb{P})} > \varepsilon^2,$$

we have

$$\mathbb{P}\left[\frac{1}{m}\sum_{k=1}^{m}|f_i(X_k) - f_j(X_k)|^2 \le \frac{\varepsilon^2}{4}\right] \le \exp\left(-\frac{m\varepsilon^4}{4}\right).$$

Now a union bound gives

$$\mathbb{P}\left[\frac{1}{m}\sum_{k=1}^{m}|f_i(X_k) - f_j(X_k)|^2 \ge \frac{\varepsilon^2}{4} \text{ for all } i \neq j\right] \ge 1 - n^2 \exp\left(-\frac{m\varepsilon^4}{4}\right) > 0$$

provided $m \asymp \varepsilon^{-4} \log n$. ∎

Proof: [of Lemma 6.11] Let $f_1, \cdots, f_n$ be an maximal ε-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mathbb{P})})$. By Lemma 6.12, there exist $m \asymp \varepsilon^{-4} \log n$ points $x_1, \cdots, x_m$ such that

$$\frac{1}{m}\sum_{k=1}^{m}|f_i(x_k) - f_j(x_k)|^2 > \varepsilon^2/4 \quad \text{for all } i \neq j.$$

Thus, letting $\mathcal{F}_n = \{f_1, \cdots, f_n\}$,

$$n = \left|\{f_i(x_1), \cdots, f_i(x_m) : \ f_i \in \mathcal{F}_n\}\right|.$$

Note that the VC dimension of $\mathcal{F}_n$ is less or equal than the VC dimension of $\mathcal{F}$. By the Sauer-Shelah lemma we have

$$n \le \left(\frac{em}{v}\right)^v \le \left(\frac{c\varepsilon^{-4}\log n}{v}\right)^v,$$

and the claim follows after some simple calculus. ∎

Theorem 6.13 (Glivenko-Cantelli) *We have* $\mathbb{E}\left[\|\widehat{F}_n - F\|_\infty\right] \lesssim \frac{1}{\sqrt{n}}$.

Proof: For fixed $(x_1, \cdots, x_n)$, let

$$Z_f = \frac{1}{\sqrt{n}} \sum_{k=1}^n \varepsilon_k f(x_k).$$

Noting that $Z_f - Z_g = \frac{1}{\sqrt{n}} \sum_{k=1}^n \varepsilon_k (f(x_k) - g(x_k))$ is $\frac{1}{n}\sum_{k=1}^n (f(x_k) - g(x_k))^2$-sub-Gaussian (see Lecture 1). Thus, if we define the metric

$$d(f,g) = \sqrt{\frac{1}{n}\sum_{k=1}^n (f(x_k) - g(x_k))^2},$$

then $Z_f - Z_g$ is $d(f,g)^2$-sub-Gaussian. Let $\widetilde{\mathcal{F}} = \{\mathcal{F}, 0\}$, namely we add a 0 function to $\mathcal{F}$. Note that we still have $v(\widetilde{\mathcal{F}}) = 1$ (**check this!**). Thus, Lemma 6.11 implies that

$$N(\widetilde{\mathcal{F}}, d, \varepsilon) \le \left(\frac{c}{\varepsilon}\right)^c, \quad \text{for } \varepsilon < 1,$$

where $c > 0$ is a universal constant. Moreover, it is easy to see that $d(f,g) \le 1$ for any $f, g \in \widetilde{F}$, and thus $\mathrm{diam}(\mathcal{F}) \le 1$. By the Dudley integral (also noting Remark 6.4) we have

$$\begin{aligned}
\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{k=1}^n \varepsilon_k f(x_k)\right|\right] &= \mathbb{E}_\varepsilon\left[\sup_{f\in\widetilde{F}}\left|\frac{1}{\sqrt{n}}\sum_{k=1}^n \varepsilon_k f(x_k) - 0\right|\right] \\
&\lesssim \int_0^1 \sqrt{\log N(\widetilde{\mathcal{F}}, d, \varepsilon)}\, d\varepsilon \\
&= O(1),
\end{aligned}$$

where $O(1)$ means a constant. Thus,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]\right|\right] \lesssim \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n \varepsilon_k f(X_k)\right|\right] \lesssim \frac{1}{\sqrt{n}}.$$

The proof is now complete. ■

Reading Materials

[1] Martin Wainwright, *High-dimensional statistics – A non-asymptotic viewpoint*, Chapter 4.

[2] Ramon van Handel, *Probability in High Dimension*, Chapters 7.1, 7.2.

[3] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, *Foundations of Machine Learning*, Chapter 3.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 7: Lower Bound on Gaussian Process

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/05/02)*

Recap and motivation: In the last few lectures, we have studied the upper bound for $\mathbb{E}\left[\sup_{t\in T} X_t\right]$. In this section we study the lower bound of $\mathbb{E}\left[\sup_{t\in T} X_t\right]$. It is clear that we cannot expect to obtain a nontrivial lower bound at the level of generality. For example, even in the case of finite maxima, we have seen that the additional assumption of independence is needed to obtain a meaningful lower bound. Otherwise, an extreme example would be $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ with $X_t = X$ for all t. Therefore, in this lecture we will *restrict our attention to the Gaussian process*. As before, we will always assume X_t is centered (i.e., $\mathbb{E}\left[X_t\right] = 0$ for all t, unless stated otherwise).

Definition 7.1 (Gaussian process) *The random process $\{X_t\}_{t\in T}$ is called a centered Gaussian process if the random variables $\{X_{t_1}, \cdots, X_{t_n}\}$ are centered and jointly Gaussian for all $n \geq 1$ and $t_1, \cdots, t_n \in T$.*

Recall that for the centered Gaussian random variable, its sub-Gaussian parameter is equal to its variance. Thus, if we define

$$d(t,s) = \sqrt{\mathbb{E}\left[(X_t - X_s)^2\right]} = \|X_t - X_s\|_{L_2}. \tag{7.1}$$

Then, a Gaussian process is a sub-Gaussian process on (T, d). Note d is usually referred to the *canonical metric* defined on T and it is indeed a pseudo-metric but it satisfies the triangle inequality.

Lower bounds will be presented in Section 3 of this lecture. The first two sections contain some technical results that are needed for the establishment of the lower bounds and are also of independent interest.

Agenda:

- Gaussian interpolation
- Gaussian comparison inequality
- Sudakov minoration inequality

7.1 Gaussian Interpolation

The proof of the Gaussian comparison inequality in the next section relies on a technique known as Gaussian interpolation. First we have the multidimensional version of the Gaussian integration by parts.

Lemma 7.2 (Gaussian integration by parts) *Let $X \sim \mathcal{N}(0, \Sigma)$, where Σ is an $n \times n$ variance matrix. Then,*

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

Proof: In the special 1-d case when $X \sim \mathcal{N}(0,1)$, the claim of the lemma reduces to

$$\mathbb{E}\left[Xf(X)\right] = \mathbb{E}\left[f'(X)\right],$$

which follows immediately after we apply the integration by part to

$$\mathbb{E}\left[f'(X)\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(x) e^{-\frac{x^2}{2}} dx.$$

In general, first note that letting $Z \sim \mathcal{N}(0, I_n)$, then X has the same distribution as $\Sigma^{1/2} Z$. Thus,

$$\mathbb{E}\left[X_i f(X)\right] = \sum_{k=1}^{n} \Sigma_{ik}^{1/2} \mathbb{E}\left[Z_k f(\Sigma^{1/2} Z)\right] = \sum_{k=1}^{n} \Sigma_{ik}^{1/2} \mathbb{E}\left[Z_k g(Z)\right],$$

where $g(z) = f(\Sigma^{1/2} z)$ and hence

$$\frac{\partial g}{\partial z_k}(z) = \sum_{j=1}^{n} \Sigma_{jk}^{1/2} \frac{\partial f}{\partial x_j}(\Sigma^{1/2} z).$$

Since the result for the special 1-d case implies (noting Z_k are independent)

$$\mathbb{E}\left[Z_k g(Z)\right] = \mathbb{E}\left[\frac{\partial g}{\partial z_k}(Z)\right] = \sum_{j=1}^{n} \Sigma_{jk}^{1/2} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(\Sigma^{1/2} Z)\right],$$

we have

$$\begin{aligned}
\mathbb{E}\left[X_i f(X)\right] &= \sum_{k=1}^{n} \Sigma_{ik}^{1/2} \mathbb{E}\left[Z_k g(Z)\right] \\
&= \sum_{k=1}^{n} \Sigma_{ik}^{1/2} \sum_{j=1}^{n} \Sigma_{jk}^{1/2} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(\Sigma^{1/2} Z)\right] \\
&= \sum_{j=1}^{n} \left(\sum_{k=1}^{n} \Sigma_{ik}^{1/2} \Sigma_{jk}^{1/2} \right) \mathbb{E}\left[\frac{\partial f}{\partial x_j}(\Sigma^{1/2} Z)\right] \\
&= \sum_{j=1}^{n} \Sigma_{ij} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(X)\right],
\end{aligned}$$

as desired. ■

Using the Gaussian integration by parts property, we are ready to present and prove the Gaussian interpolation result.

Lemma 7.3 *Let $X \sim \mathcal{N}(0, \Sigma^X)$ and $Y \sim \mathcal{N}(0, \Sigma^Y)$ be two independent n-dimensional Gaussian vectors. Define*

$$Z(t) = \sqrt{t} X + \sqrt{1-t} Y, \quad t \in [0,1].$$

Then for every smooth function f we have

$$\frac{d}{dt} \mathbb{E}\left[f(Z(t))\right] = \frac{1}{2} \sum_{i,j=1}^{n} \left(\Sigma_{ij}^X - \Sigma_{ij}^Y\right) \mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(t))\right].$$

Proof: By the chain rule we have

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[f(Z(t))\right] &= \sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\frac{dZ_i}{dt}\right] \\ &= \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\frac{X_i}{\sqrt{t}}\right] - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\frac{Y_i}{\sqrt{1-t}}\right].\end{aligned}$$

Considering the first term, as X and Y are independent, we can apply Lemma 7.2 to the $2n$-dimensional Gaussian random (X, Y) (**what is the covariance matrix?**) and obtain

$$\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\frac{X_i}{\sqrt{t}}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\Sigma^X_{ij}\mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(t))\right].$$

Since the second term can be bounded similarly, the proof is complete. ■

7.2 Gaussian Comparison Inequality

Theorem 7.4 (Sudakov-Fernique inequality) *Let $\{X_t\}_{t\in T}$ and $\{Y_t\}_{t\in T}$ be two mean zero separable Gaussian processes. Suppose*

$$\mathbb{E}\left[|X_t - X_s|^2\right] \geq \mathbb{E}\left[|Y_t - Y_s|^2\right] \quad \textit{for all } t, s \in T.$$

Then,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq \mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

This theorem is very intuitive: if $\{X_t\}_{t\in T}$ has larger pairwise variance than $\{Y_k\}_{t\in T}$, then $\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq \mathbb{E}\left[\sup_{t\in T} Y_t\right]$. It is enough to establish the theorem for two Gaussian vectors $X \sim \mathcal{N}(0, \Sigma^X)$ and $Y \sim \mathcal{N}(0, \Sigma^Y)$. Moreover, we can assume X and Y are independent; otherwise we can consider an independent copy of one of them.

Proof: For any $\beta > 0$ define

$$f_\beta(x) = \frac{1}{\beta}\log\sum_{k=1}^{n} e^{\beta x_k}.$$

It is not hard to see that (**check this!**)

$$\max_{k=1,\cdots,n} x_k \leq f_\beta(x) \leq \max_{k=1,\cdots,n} x_k + \frac{\log n}{\beta}.$$

Thus, $f_\beta(x) \to \max_{k=1,\cdots,n} x_k$ as $\beta \to \infty$. Moreover,

$$\frac{\partial f}{\partial x_k} = \frac{e^{\beta x_k}}{\sum_{k=1}^{n} e^{\beta x_k}} =: p_k(x), \quad \frac{\partial^2 f}{\partial x_k \partial x_j} = \beta\left(\delta_{kj}p_k(x) - p_k(x)p_j(x)\right),$$

where δ_{kj} equals 1 if $k = j$ and equals 0 otherwise. It follows from Lemma 7.3 that

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[f_\beta(Z(t))\right] &= \frac{1}{2}\sum_{k,j=1}^{n}\left(\Sigma_{kj}^X - \Sigma_{kj}^Y\right)\mathbb{E}\left[\frac{\partial^2 f_\beta}{\partial x_k \partial x_j}(Z(t))\right] \\ &= \frac{\beta}{2}\sum_{k=1}^{n}\left(\Sigma_{kk}^X - \Sigma_{kk}^Y\right)\mathbb{E}\left[p_k(Z(t))(1-p_k(Z(t)))\right] - \frac{\beta}{2}\sum_{k\neq j}\left(\Sigma_{kj}^X - \Sigma_{kj}^Y\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right].\end{aligned}$$

Noting that $1 - p_k(x) = \sum_{j\neq k} p_j(x)$, we have

$$\begin{aligned}\sum_{k=1}^{n}\left(\Sigma_{kk}^X - \Sigma_{kk}^Y\right)\mathbb{E}\left[p_k(Z(t))(1-p_k(Z(t)))\right] &= \sum_{k\neq j}\left(\Sigma_{kk}^X - \Sigma_{kk}^Y\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right] \\ &= \sum_{k\neq j}\left(\Sigma_{jj}^X - \Sigma_{jj}^Y\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right].\end{aligned}$$

It follows that

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[f_\beta(Z(t))\right] &= \sum_{k\neq j}\frac{\beta}{4}\left(\Sigma_{kk}^X - 2\Sigma_{kj}^X + \Sigma_{jj}^X\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right] - \sum_{k\neq j}\frac{\beta}{4}\left(\Sigma_{kk}^Y - 2\Sigma_{kj}^Y + \Sigma_{jj}^Y\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right] \\ &= \frac{\beta}{4}\sum_{k\neq j}\left(\mathbb{E}\left[|X_k - X_j|^2\right] - \mathbb{E}\left[|Y_k - Y_j|^2\right]\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right] \\ &\geq 0,\end{aligned}$$

where in the last line we have used the assumption. Thus $f_\beta(Z(t))$ is increasing in t, yielding

$$\mathbb{E}\left[f_\beta(X)\right] \geq \mathbb{E}\left[f_\beta(Y)\right].$$

Letting $\beta \to \infty$ concludes the proof. ■

There are also other types of Gaussian comparison inequalities such as the Slepian inequality or the Gordon inequality, which can be proved similarly, see for example [3]. The Gaussian comparison inequalities have many interesting applications. Here we give an example before presenting the application on the lower bound of the Gaussian process in the next section.

Example 7.5 (Spectral norm of Gaussian matrices) *Let $W \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d $\mathcal{N}(0,1)$ entries. By the the one-step discretization bound in Lecture 5 (see Example 5.15), we have*

$$\mathbb{E}\left[\|W\|_2\right] \leq C(\sqrt{m} + \sqrt{n}).$$

Next we can show that the bound can be sharpened to

$$\mathbb{E}\left[\|W\|_2\right] \leq \sqrt{m} + \sqrt{n}$$

by the Sudakov-Fernique inequality[1]*. We still begin with the variational form for $\|W\|_2$,*

$$\|W\|_2 = \sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} u^T W v =: \sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} Y_{uv}.$$

[1]However, note that the one-step discretization bound works for all the general sub-Gaussian matrices, not only the standard Gaussian matrices.

We have

$$\begin{aligned}\mathbb{E}\left[|Y_{uv}-Y_{ts}|^2\right] &= \mathbb{E}\left[\left(\sum_{ij} W_{kj}(u_i v_j - t_i s_j)\right)^2\right] \\ &= \sum_{ij}(u_i v_j - t_i s_j)^2 \\ &= \|uv^T - ts^T\|_F^2 \\ &\leq \|u-t\|_2^2 + \|v-s\|_2^2.\end{aligned}$$

If we construct another Gaussian process as follows,

$$X_{uv} = \langle g, u\rangle + \langle h, v\rangle, \quad g \sim \mathcal{N}(0, I_m),\ h \sim \mathcal{N}(0, I_n).$$

it is easy to see that $\mathbb{E}\left[|X_{uv}-X_{ts}|^2\right] = \|u-t\|_2^2 + \|v-s\|_2^2$. *Thus, applying the Sudakov-Fernique inequality yields*

$$\begin{aligned}\mathbb{E}\left[\sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} u^T W v\right] &\leq \mathbb{E}\left[\sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} \langle g, u\rangle + \langle h, v\rangle\right] \\ &= \mathbb{E}\left[\sup_{u\in\mathbb{B}_2^m} \langle g, u\rangle\right] + \mathbb{E}\left[\sup_{v\in\mathbb{B}_2^n} \langle h, v\rangle\right] \\ &\leq \sqrt{m} + \sqrt{n}.\end{aligned}$$

7.3 Sudakov Minoration Inequality

Theorem 7.6 (Sudakov minoration inequality) *Let* $\{X_t\}_{t\in T}$ *be a centered Gaussian process. Then*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \gtrsim \sup_{\varepsilon>0} \varepsilon\sqrt{\log N(T, d, \varepsilon)},$$

where d *is the canonical metric defined in* (7.1).

Proof: For any $\varepsilon > 0$, let P be ε-packing of T under the canonical metric with the packing number $P(T, d, \varepsilon)$. Let $X = \{X_t\}_{t\in P}$ and let $Y = \{Y_t\}_{t\in P}$ be a vector of length $P(T, d, \varepsilon)$ with i.i.d $\mathcal{N}(0, \frac{\varepsilon^2}{2})$ variables. Then,

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

Thus the Sudakov-Fernique inequality yields

$$\mathbb{E}\left[\sup_{t\in P} X_t\right] \geq \mathbb{E}\left[\sup_{t\in P} Y_t\right] \asymp \varepsilon\sqrt{\log P(T, d, \varepsilon)} \geq \varepsilon\sqrt{\log N(T, d, \varepsilon)},$$

where the last inequality follows from the relationship $N(T, d, \varepsilon) \leq P(T, d, \varepsilon)$. ■

Example 7.7 (Gaussian complexity of unit 2-ball $\mathbb{B}_2^d$) *In Lecture 5, we have seen that*

$$\mathcal{G}(\mathbb{B}_2^d) = \mathbb{E}\left[\sup_{t\in\mathbb{B}_2^d} \langle g, t\rangle\right] \asymp \sqrt{d},$$

where the lower bound is obtained via the comparison with the corresponding Rademacher complexity. Since $\langle g, t\rangle$ is a Gaussian process with the canonical metric given by

$$d(t,s) = \sqrt{\mathbb{E}\left[|\langle g, t-s\rangle|^2\right]} = \|t-s\|_2,$$

we can also use the Sudakov minoration inequality to get the lower bound,

$$\mathcal{G}(\mathbb{B}_2^d) \gtrsim \varepsilon\sqrt{\log N(\mathbb{B}_2^d, \|\cdot\|_2, \varepsilon)} \asymp \varepsilon\sqrt{d\log\frac{1}{\varepsilon}} \asymp \sqrt{d}$$

after choosing a proper ε.

Example 7.8 (Lower bound on spectral norm of Gaussian matrices) *In Example 7.5, we have seen that*

$$\mathbb{E}\left[\|W\|_2\right] \le \sqrt{m} + \sqrt{n}$$

for an $m \times n$ Gaussian random matrix. The Sudakov minoration inequality can be used to show that this bound is sharp in terms of the scaling. Recall that

$$\|W\|_2 = \sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} u^T W v =: \sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} Y_{uv}.$$

*The application of the Sudakov minoration inequality yields that (**complete the details!**)*

$$\begin{aligned}
\mathbb{E}\left[\|W\|_2\right] &\gtrsim \varepsilon\sqrt{\log N(\mathbb{B}_2^m \otimes \mathbb{B}_2^n, \|\cdot\|_F, \varepsilon)} \\
&\gtrsim \varepsilon\sqrt{\log\left(N(\mathbb{B}_2^m, \|\cdot\|_2, \varepsilon)\cdot N(\mathbb{B}_2^n, \|\cdot\|_2, \varepsilon)\right)} \\
&\asymp \varepsilon\sqrt{(m+n)\log\frac{1}{\varepsilon}} \\
&\gtrsim \sqrt{m} + \sqrt{n}
\end{aligned}$$

after choosing ε properly.

Sudakov minoration inequality can also be used in a reverse way to provide an upper bound for the metric entropy of a set.

Example 7.9 (Metric entropy of unit 1-ball $\mathbb{B}_1^d$ under the Euclidean distance) *We have already seen that*

$$\mathcal{G}(\mathbb{B}_1^d) = \mathbb{E}\left[\sup_{t\in\mathbb{B}_1^d} \langle g, t\rangle\right] \asymp \sqrt{\log d}.$$

Together with the Sudakov minoration inequality, we have

$$\log N(\mathbb{B}_1^d, \|\cdot\|_2, \varepsilon) \lesssim \frac{1}{\varepsilon^2} \log d.$$

Up to constant, this result matches the bound for the covering number of a convex hull of a finite set due to Maurey. Noting that

$$\log N(\mathbb{B}_2^d, \|\cdot\|_2, \varepsilon) \asymp d \log \frac{1}{\varepsilon},$$

we can thus see in a different way how the unit 1*-ball is much smaller than the unit* 2*-ball.*

Combining the Sudakov minoration inequality with the Dudley inequality/integral, we have for the Gaussian process $\{X_t\}_{t\in T}$

$$\sup_k 2^{-k}\sqrt{\log N(T, d, 2^{-k})} \lesssim \mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim \sum_k 2^{-k}\sqrt{\log N(T, d, 2^{-k})}.$$

In some situations, the upper and lower bounds are not as far apart as may appear at first sight because the term $2^{-k}\sqrt{\log N(T, d, 2^{-k})}$ behaves like a geometric sequence so that their sum is of the same order as the largest one (for example, consider Example 7.7). However, there are also cases where there is indeed a gap between these two bounds. It turns out the generic chaining bound is tight for Gaussian processes, i.e.,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \asymp \gamma(T, d),$$

see Section 3 of Lecture 6 for the definition of $\gamma(T, d)$. This is the notable Talagrand majorizing measure theorem. We will omit the details, see for example [2] and [3]. For *stationary* Gaussian process, Dudley integral is also tight.

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapter 5.4.

[2] Ramon van Handel, *Probability in High Dimension*, Chapter 6.1.

[3] Roman Vershynin, *High-Dimensional Probability: An introduction with applications in data science*, Chapter 7.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 8: Random Matrices and Applications

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/05/06)*

Motivation: The study of random matrices is directly motivated by the estimation of covariance matrices. Let $X \in \mathbb{R}^n$ be a *mean zero* random vector. Then the covariance matrix corresponding to X is given by

$$\Sigma = \mathbb{E}\left[XX^T\right].$$

However, since we typically do not know the distribution of X but only have access to m i.i.d samples $\{X_k\}_{k=1}^m$ of X, a natural estimator[1] of Σ is

$$\Sigma_m = \frac{1}{m}\sum_{k=1}^m X_k X_k^T.$$

Then we would like to know how close the random matrix Σ_m to its mean Σ, *particularly in terms of the matrix spectral norm.*

The approaches for studying a random matrix usually rely on the distribution of its elements. For example, if the random matrix has Gaussian entries, we can first establish the concentration results based on for example the Gaussian concentration inequality and then use the Gaussian comparison theorem to estimate the expectation. When the random matrix has certain sub-Gaussian properties, we can still utilize this property to obtain a desirable bound. When there is no explicit distributions associated with the elements of the random matrix, we will attempt to establish some useful bounds by imitating the chernoff method for random variables. This lecture will start directly from the settings of the sub-Gaussian properties. For the analysis of the specific Gaussian matrices, see Chapter 6.1 of [1]. Before proceeding, it is worth noting that we will study matrix concentration in terms the spectral norm rather than the Frobenius norm. This is largely due to that we are usually interested in the deviation of principle directions of the covariance matrix, and the bound based on spectral norm is sufficiently tighter than that based on the Frobenius norm (which is the sum of the errors in all directions). In addition, it is trivial that the matrix concentration bound in terms of Frobenius norm can be reduced to concentration result of random variables.

Agenda:

- Covariance matrix under sub-Gaussian assumption
- Application: Clustering based on PCA
- Matrix Bernstein inequality
- Application: Covariance matrix for general distributions
- Application: Sparse Recovery

[1] When the covariance matrix is known to have certain structure, a better estimator can be constructed based on that structure, see Chapter 6.5 of [1].

8.1 Covariance Matrix under sub-Gaussian Assumption

In this section we will consider the concentration of the covariance matrix Σ_m when X is a sub-Gaussian random vector, defined as follows.

Definition 8.1 (Sub-Gaussian random vector) *A mean zero random vector $X \in \mathbb{R}^n$ is sub-Gaussian with parameter σ^2 if for each $v \in \mathbb{S}^{n-1}$ (i.e., $\|v\|_2 = 1$), $\langle X, v\rangle$ is a sub-Gaussian random variable with parameter σ^2.*

Example 8.2 *Assume $X \in \mathbb{R}^n$ has i.i.d σ^2-sub-Gaussian entries. Then,*

$$\mathbb{E}\left[e^{\lambda\langle X,v\rangle}\right] = \mathbb{E}\left[\prod_{k=1}^{n} e^{\lambda v_k X_k}\right] \leq \prod_{k=1}^{n} e^{\frac{\lambda^2 v_k^2 \sigma^2}{2}} = e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{for } v \in \mathbb{S}^{n-1},$$

meaning $\langle X, v\rangle$ is σ^2-sub-Gaussian. Thus, X is a σ^2-sub-Gaussian random vector.

Example 8.3 *Let $X \sim \mathcal{N}(0, \Sigma)$. Then for any $v \in \mathbb{S}^{n-1}$, $v^T X \sim \mathcal{N}(0, v^T\Sigma v)$. Since $v^T\Sigma v \leq \|\Sigma\|_2$, we can conclude that X is a sub-Gaussian random vector with parameter at most $\|\Sigma\|_2$.*

The following lemma provides a characterization of the spectral norm of a symmetric matrix in terms of the ε-net. We have indeed seen this result for general matrices in Lecture 4.

Lemma 8.4 *Let $Z \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Assume $\varepsilon \in [0, 1/2)$ and let N be a ε-net of $\mathbb{S}^{n-1}$ under the $\|\cdot\|_2$ metric. Then*

$$\|Z\|_2 \leq \frac{1}{1-2\varepsilon}\sup_{v\in N}|\langle Zv, v\rangle|.$$

Proof: For any $x \in \mathbb{S}^{n-1}$, by the definition of ε-net, there exists a vector $\pi(x) \in N$ such that $\|x - \pi(x)\|_2 \leq \varepsilon$. It follows that

$$\langle Zx, x\rangle - \langle Z\pi(x), \pi(x)\rangle = \langle Z(x-\pi(x)), x\rangle + \langle Z\pi(x), x-\pi(x)\rangle,$$

and hence

$$|\langle Zx, x\rangle - \langle Z\pi(x), \pi(x)\rangle| \leq 2\varepsilon\|Z\|_2.$$

Consequently,

$$\|Z\|_2 = \sup_{x\in\mathbb{S}^{n-1}}|\langle Zx, x\rangle| \leq \sup_{x\in\mathbb{S}^{n-1}}\left(|\langle Z\pi(x), \pi(x)\rangle| + 2\varepsilon\|Z\|_2\right).$$

Then the proof is complete after rearrangement. ∎

We are now ready to state and prove the main result in this section, which concerns about the concentration of empirical covariance matrices.

Theorem 8.5 *Let $X \in \mathbb{R}^n$ be a mean zero σ^2-sub-Gaussian random vector and $\Sigma = \mathbb{E}\left[XX^T\right]$ be its covariance matrix. Let $\{X_k\}_{k=1}^m$ be i.i.d samples and define $\Sigma_m = \frac{1}{m}\sum_{k=1}^m X_k X_k^T$. Then,*

$$\mathbb{P}\left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \geq c_1\left\{\sqrt{\frac{n}{m}} + \frac{n}{m}\right\} + t\right] \leq c_2\exp\left(-c_3\min\{t, t^2\}m\right) \quad \text{for all } t \geq 0.$$

Here, $c_1, c_2, c_3 > 0$ are absolute numerical constants.

Proof: Let $Z = \Sigma_m - \Sigma$. Taking N to be a 1/4-net of $\mathbb{S}^{n-1}$, we have $|N| \le 9^n$ and

$$\|Z\|_2 \le 2 \sup_{v \in N} |\langle Zv, v \rangle| .$$

The overall strategy of the proof is to first consider a fixed $v \in N$ and then take a union bound.

For any fixed $v \in N$, we have

$$\langle Zv, v \rangle = \frac{1}{m} \sum_{k=1}^{m} \left(\left(X_k^T v \right)^2 - \mathbb{E} \left[\left(X_k^T v \right)^2 \right] \right).$$

Since $X_k^T v$ is σ^2-sub-Gaussian, we have

$$\begin{aligned} \left\| \left(X_k^T v \right)^2 - \mathbb{E} \left[\left(X_k^T v \right)^2 \right] \right\|_{L_p} &\le \left\| \left(X_k^T v \right)^2 \right\|_{L_p} + \mathbb{E} \left[\left(X_k^T v \right)^2 \right] \\ &\lesssim \sigma^2 p, \end{aligned}$$

implying that $\left(X_k^T v \right)^2 - \mathbb{E} \left[\left(X_k^T v \right)^2 \right]$ is $c_4 \cdot \sigma^4$-sub-exponential (see Theorem 1.33 of Lecture 1). Then the application of the Bernstein inequality implies that

$$\mathbb{P} \left[|\langle Zv, v \rangle| \ge \frac{\delta}{2} \right] \lesssim \exp \left(-c_5 \min \left\{ \frac{\delta^2}{\sigma^4}, \frac{\delta}{\sigma^2} \right\} m \right).$$

Taking a union bound yields that

$$\begin{aligned} \mathbb{P} \left[\|Z\|_2 \ge \delta \right] &\le \mathbb{P} \left[\sup_{v \in N} |\langle Zv, v \rangle| \ge \frac{\delta}{2} \right] \\ &\lesssim 9^n \exp \left(-c_5 \min \left\{ \frac{\delta^2}{\sigma^4}, \frac{\delta}{\sigma^2} \right\} m \right) \\ &= \exp \left(n \log 9 - c_5 \min \left\{ \frac{\delta^2}{\sigma^4}, \frac{\delta}{\sigma^2} \right\} m \right) \end{aligned} \tag{8.1}$$

Let $\delta = \left(c_1 \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\} + t \right) \sigma^2$. Then,

$$\delta \ge \left(c_1 \frac{n}{m} + t \right) \sigma^2 \quad \text{and} \quad \delta^2 \ge \left(c_1^2 \frac{n}{m} + t^2 \right) \sigma^4.$$

Substituting them into (8.1) yields that

$$\mathbb{P} \left[\|Z\|_2 \ge \delta \right] \lesssim \exp \left(n \log 9 - c_5 \min \left\{ c_1 \frac{n}{m} + t, c_1^2 \frac{n}{m} + t^2 \right\} m \right).$$

The proof is complete if we take c_1 to be sufficiently large. ∎

Remark 8.6 *Since*

$$\begin{aligned} \mathbb{E} \left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \right] &= \int_0^\infty \mathbb{P} \left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \ge x \right] dx \\ &= \int_0^{c_1 \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\}} \mathbb{P} \left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \ge x \right] dx + \int_{c_1 \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\}}^\infty \mathbb{P} \left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \ge x \right] dx \end{aligned}$$

$$
\begin{aligned}
&\leq c_1 \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\} + \int_0^\infty \mathbb{P}\left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \geq c_1 \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\} + t \right] dt \\
&\leq c_1 \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\} + c_2 \int_0^\infty \exp\left(-c_3 \min\{t, t^2\} m \right) dt \\
&\lesssim \sqrt{\frac{n}{m}} + \frac{n}{m},
\end{aligned}
$$

it follows that

$$
\mathbb{E}\left[\|\Sigma_m - \Sigma\|_2 \right] \lesssim \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\} \sigma^2.
$$

Moreover, we have

$$
\mathbb{E}\left[\|\Sigma_m\|_2 \right] \lesssim \|\Sigma\|_2 + \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\} \sigma^2.
$$

Thus, an upper bound for $\mathbb{E}\left[\|\Sigma_m\|_2 \right]$ *can be derived from the concentration result under less stringent conditions. Note this bound cannot be obtained via methods discussed in the previous lectures since they only work for (sub)-Gaussian processes.*

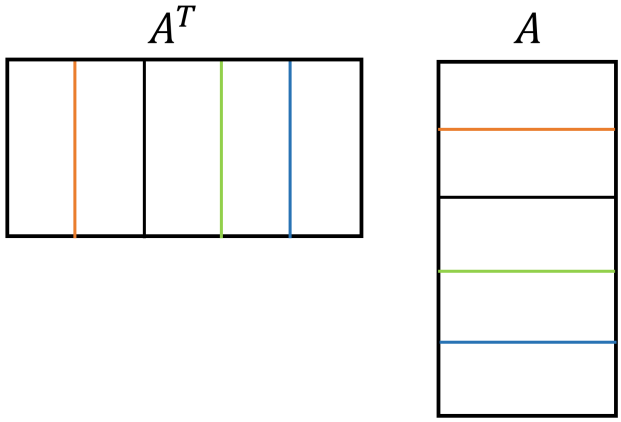


Figure 8.1: $\Sigma_m = \frac{1}{m} A^T A$.

Remark 8.7 *Assume* $\Sigma = I_n$ *and* X_k *is sub-Gaussian with parameter* $\sigma^2 = 1$*. Note that we can express* Σ_m *as* $\Sigma_m = \frac{1}{m} A^T A$*, where* $A^T = [X_1, \cdots, X_m]$ *(see Figure 8.3). Thus, Theorem 8.5 implies that, with high probability,*

$$
1 - c' \sqrt{\frac{n}{m}} \leq \frac{\sigma_{\min}(A)}{\sqrt{m}} \leq \frac{\sigma_{\max}(A)}{\sqrt{m}} \leq 1 + c' \sqrt{\frac{n}{m}}
$$

for some numerical constant $c' > 0$*, with the proviso that* $m \geq n$*. That is,* A *behaves more and more well-conditioned (like an orthogonal matrix) when* m/n *increases. This turns out to be very useful result itself.*

8.2 Application: Clustering Based on PCA

The PCA paradigm which first projects data onto a low dimensional subspace can be used for data clustering. For simplicity we consider the following Gaussian mixture model with two different means $\{-\mu, \mu\}$,

$$X = \varepsilon\mu + g, \tag{8.2}$$

where $\varepsilon \in \{1, -1\}$ is a Rademacher random variable, $\mu \in \mathbb{R}^n$ is deterministic and $g \in \mathcal{N}(0, I_n)$. In words, sampling from X will generate two clusters of data, obeying $\mathcal{N}(-\mu, I_n)$ and $\mathcal{N}(\mu, I_n)$ respectively, see Figure 8.4.

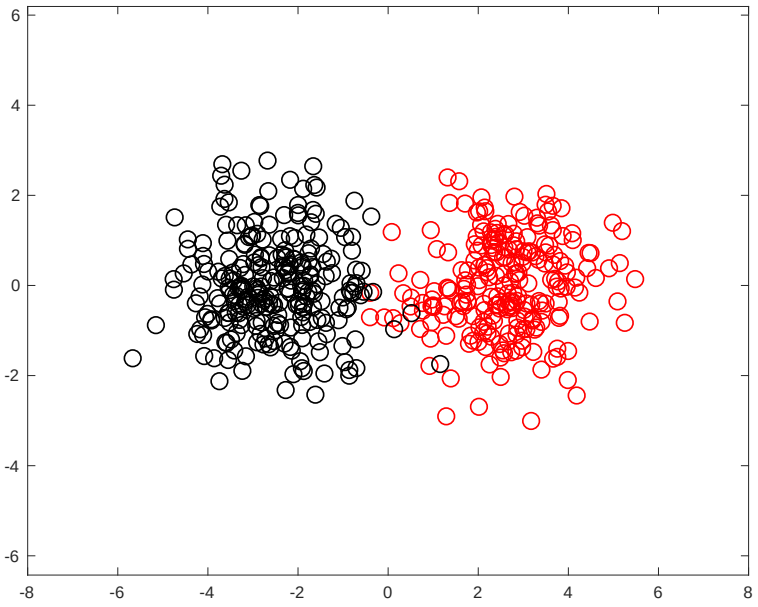


Figure 8.2: A simulation of points generated according to the Gaussian mixture model (8.2).

Suppose we are given a sample of m points $\{X_k\}_{k=1}^m$ drawn according to the Gaussian mixture model (see Figure 8.4 for a simulation) and want to identify which points belong to which cluster (i.e., determine they are generated from which mean). From the simulation, it is not hard to see that the data generated from X is stretch in the direction of μ, and the data points from different clusters have different inner product with μ. Assuming $\|\mu\|_2 \geq 1$, noting that

$$\langle \varepsilon\mu + g, \mu \rangle = \varepsilon\|\mu\|_2^2 + \langle g, \mu \rangle ,$$

where the size of $\langle g, \mu \rangle$ is about $\|\mu\|_2$, the sign of the inner product will coincide with ε, and hence can tell which mean the data point corresponds to. Indeed, if we define

$$Z_k = (\text{sign}(\langle \underbrace{\varepsilon_k\mu + g_k}_{X_k}, \mu/\|\mu\|_2 \rangle) \neq \varepsilon_k),$$

by the Hoeffding inequality, it can be shown that with high probability the number of misclassifications $\sum_{k=1}^m Z_k$ cannot exceed a fraction of m (**show this!**).

In the situation when we do not know μ but only have access to $\{X_k\}_{k=1}^m$, we can approximate μ by PCA since the principal direction of PCA captures the direction that the data points stretch the most. This gives the spectral algorithm for data clustering (here "spectral" refers to using the eigenvectors of a matrix for the task since the eigen-decomposition of a matrix is also known as spectral decomposition),

- Compute the covariance matrix $\Sigma_m = \frac{1}{m}\sum_{k=1}^m X_k X_k^T$.
- Compute the principal eigenvector q (of unit norm) of Σ_m, i.e., eigenvector corresponding to the largest eigenvalue of Σ_m.
- Partition the data points into two clusters based on the sign of $\langle X_k, q\rangle$ (data points with the same sign of $\langle X_k, q\rangle$ will be put into the same cluster).

Next we are going to show that q can be close to μ. To this end, first note that

$$\Sigma = \mathbb{E}\left[XX^T\right] = \mu\mu^T + I_n,$$

and the largest eigenvalue of Σ is $1 + \|\mu\|_2^2$, with the corresponding normalized eigenvector $\mu/\|\mu\|_2$. Since X is a sub-Gaussian random vector with the parameter proportional to $\|\mu\|_2^2$ (**check this!**), By Theorem 8.5, we have

$$\|\Sigma_m - \Sigma\|_2 \le \rho\|\mu\|_2^2, \tag{8.3}$$

for a sufficiently small $\rho > 0$ when $m \gtrsim n$ (the hidden constant relies on ρ). Noting the gap between the first and second largest eigenvalues of Σ is $\|\mu\|_2^2$, the Davis-Kahan theorem (**Consult a textbook on numerical linear algebra for this theorem!**) together with (8.3) implies that

$$\exists\ \theta \in \{1, -1\} \text{ such that } \|q - \theta(\mu/\|\mu\|_2)\|_2 \le \rho',$$

where $\rho' > 0$ is also a sufficiently small number (a multiple of ρ).

8.3 Matrix Bernstein Inequality

In the last section, we have studied the covariance matrix concentration based on the distributional information of the matrix elements (e.g, certain sub-Gaussian rows). When there is no distribution assumption to use, we can develop a matrix analogue of the chernoff method for random variables, treating the random matrix as a whole object. Both the matrix Hoeffding inequality and the matrix Bernstein inequality can be developed by this way. In this section we focus on the more widely used matrix Bernstein inequality.

8.3.1 Matrix Calculus

In this section we use $\mathbb{S}^{n\times n}$ to denote the set of $n \times n$ symmetric matrices and use $\mathbb{S}_+^{n\times n}$ to denote the set of $n \times n$ symmetric and positive definite matrices. In addition, we say $X \preceq Y$ or $Y \succeq X$ if $Y - X$ is positive semidefinite.

Definition 8.8 (Matrix Function) *Let $X \in \mathbb{S}^{n\times n}$ with the eigenvalue decomposition $X = Q\Lambda Q^T = \sum_{k=1}^n \lambda_k q_k q_k^T$. Given a function $f : \mathbb{R} \to \mathbb{R}$, we define $f(X)$ as*

$$f(X) = \sum_{k=1}^n f(\lambda_k) q_k q_k^T$$

In other words, we compute $f(X)$ by applying $f(\cdot)$ to the eigenvalues of X while the eigenvectors are kept unchanged.

Example 8.9 *Let $f(x) = a_0 + a_1 x + \cdots + a_j x^j$. Then,*

$$f(X) = a_0 I + a_1 X + \cdots + a_j X^j.$$

Example 8.10 *Let $f(x) = e^x$. Then,*

$$f(X) = e^X = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{X^k}{k!}.$$

Example 8.11 *Let $f(x) = \log x$. Then, for $X \in \mathbb{S}_+^{n\times n}$,*

$$e^{f(X)} = e^{\log X} = X.$$

Exercise 8.12 *Let X and Y be two matrices in $\mathbb{S}^{n\times n}$.*

1. *Show that if the matrices commute (i.e., $XY = YX$), then*

$$e^{X+Y} = e^X e^Y.$$

2. *Give an example of two matrices X and Y such that*

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

Note that the identity $e^{x+y} = e^x e^y$ plays an crucial rule in the proof of the concentration of the sum of random variables. Indeed, this identity allows us to tensorize, i.e., break the moment generating function of variable sum into the product of exponentials. Unfortunately, as we see in the above exercise, similar identity does not hold for matrices in general. Nevertheless, there are useful substitutes in terms of the matrix trace, which are stated below without proofs.

Lemma 8.13 (Golden-Thompson inequality) *For two matrices X and Y in $\mathbb{S}^{n\times n}$, we have*

$$\operatorname{trace}\left(e^{X+Y}\right) \leq \operatorname{trace}\left(e^X e^Y\right).$$

Lemma 8.14 (Lieb inequality) *Let $H \in \mathbb{S}^{n\times n}$. Define the function on the set $\mathbb{S}_+^{n\times n}$,*

$$f(X) = \operatorname{trace}\left(\exp\left(H + \log X\right)\right).$$

Then $f(X)$ is a concave function on $\mathbb{S}_+^{n\times n}$.

Remark 8.15 *The Jensen inequality still holds for random matrices since we can interpret $f(X)$ as a function of all the entries of X. Thus, letting X be a random matrix, we have*

$$\mathbb{E}\left[\operatorname{trace}\left(\exp\left(H+\log X\right)\right)\right] \leq \operatorname{trace}\left(\exp\left(H+\log \mathbb{E}\left[X\right]\right)\right)$$

Setting $X = e^Z$ yields that

$$\mathbb{E}\left[\operatorname{trace}\left(\exp\left(H+Z\right)\right)\right] \leq \operatorname{trace}\left(\exp\left(H+\log \mathbb{E}\left[e^Z\right]\right)\right). \tag{8.4}$$

This inequality will be used in the proof of the matrix Bernstein inequality.

Both the Golden-Thompson inequality and the Lieb inequality can be used to establish the matrix Bernstein inequality. We will use the Lieb inequality next as it tensorizes better and thus yields better parameter dependence.

8.3.2 Matrix Bernstein Inequality

Theorem 8.16 (Matrix Bernstein inequality) *Let $X_1, \cdots, X_m$ be independent, mean zero, $n \times n$ symmetric random matrices. Assume $\|X_k\|_2 \leq B$ almost surely for all k. Then, for any $t \geq 0$, we have*

$$\mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq t\right] \leq 2n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Bt/3}\right),$$

where $\sigma^2 = \left\|\sum_{k=1}^m \mathbb{E}\left[X_k^2\right]\right\|_2$ is the norm of the matrix variance of the sum.

Note that the matrix Bernstein is an exact analogue of the Bernstein inequality for random variables, see (1.8) in Lecture 1. Thus, the overall proof strategy is similar to the variable case. We start with establishing an inequality similar to (1.7) in Lecture 1.

Lemma 8.17 (Moment generating function of random matrix) *Let $X \in \mathbb{S}^{n\times n}$ be a mean zero random matrix which satisfies $\|X\|_2 \leq B$ almost surely. Then,*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \preceq \exp\left(g(\lambda)\mathbb{E}\left[X^2\right]\right) \quad \textit{where} \quad g(\lambda) = \frac{\lambda^2/2}{1 - B|\lambda|/3}$$

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

Proof: First it can be shown that (**check this!**)

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

Thus, for $|x| \leq B$, if $|\lambda| < 3/B$, then

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

It follows that

$$\exp\left(\lambda X\right) \preceq I + \lambda X + g(\lambda)X^2$$

when $\|X\|_2 \leq B$ and $|\lambda| < 3/B$. Taking expectation on both sides yields that

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \preceq I + g(\lambda)\mathbb{E}\left[X^2\right] \preceq \exp\left(g(\lambda)\mathbb{E}\left[X^2\right]\right),$$

as desired. ■

Proof: [Proof of Theorem 8.16] Noting that

$$\left\|\sum_{k=1}^{m} X_k\right\|_2 = \max\left\{\lambda_{\max}\left(\sum_{k=1}^{m} X_k\right), \lambda_{\max}\left(-\sum_{k=1}^{m} X_k\right)\right\},$$

it suffices to show that $\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^{m} X_k\right) \geq t\right] \leq n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Bt/3}\right)$, and the bound for $\mathbb{P}\left[\lambda_{\max}\left(-\sum_{k=1}^{m} X_k\right) \geq t\right]$ can be established in the same manner. To this end, for fixed $\lambda \geq 0$ and the application of the Markov inequality gives

$$\begin{aligned}
\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^{m} X_k\right) \geq t\right] &= \mathbb{P}\left[\exp\left(\lambda \cdot \lambda_{\max}\left(\sum_{k=1}^{m} X_k\right)\right) \geq \exp\left(\lambda t\right)\right] \\
&\leq \exp\left(-\lambda t\right)\mathbb{E}\left[\exp\left(\lambda \cdot \lambda_{\max}\left(\sum_{k=1}^{m} X_k\right)\right)\right] \\
&= \exp\left(-\lambda t\right)\mathbb{E}\left[\lambda_{\max}\left(\exp\left(\lambda \cdot \sum_{k=1}^{m} X_k\right)\right)\right] \\
&\leq \exp\left(-\lambda t\right)\mathbb{E}\left[\operatorname{trace}\left(\exp\left(\lambda \cdot \sum_{k=1}^{m} X_k\right)\right)\right]. \qquad (8.5)
\end{aligned}$$

To apply the Lieb inequality (8.4), letting $H = \lambda \sum_{k=1}^{m-1} X_k$ and $Z = \lambda X_m$, we have

$$\mathbb{E}\left[\operatorname{trace}\left(\exp\left(\lambda \cdot \sum_{k=1}^{m} X_k\right)\right)\right] \leq \mathbb{E}\left[\operatorname{trace}\left(\exp\left(\lambda \sum_{k=1}^{m-1} X_k + \log \mathbb{E}\left[e^{\lambda X_m}\right]\right)\right)\right]$$

Repeating this process yields that

$$\begin{aligned}
\mathbb{E}\left[\operatorname{trace}\left(\exp\left(\lambda \cdot \sum_{k=1}^{m} X_k\right)\right)\right] &\leq \operatorname{trace}\left(\exp\left(\sum_{k=1}^{m} \log \mathbb{E}\left[e^{\lambda X_k}\right]\right)\right) \\
&\leq \operatorname{trace}\left(\exp\left(\sum_{k=1}^{m} \log \exp\left(g(\lambda)\mathbb{E}\left[X_k^2\right]\right)\right)\right) \\
&= \operatorname{trace}\left(\exp\left(g(\lambda)\sum_{k=1}^{m} \mathbb{E}\left[X_k^2\right]\right)\right) \\
&\leq n\left\|\exp\left(g(\lambda)\sum_{k=1}^{m} \mathbb{E}\left[X_k^2\right]\right)\right\|_2 \\
&= n \cdot \exp\left(g(\lambda)\left\|\sum_{k=1}^{m} \mathbb{E}\left[X_k^2\right]\right\|_2\right)
\end{aligned}$$

$$= n \cdot \exp\left(g(\lambda)\sigma^2\right)$$

provided $|\lambda| \leq 3/B$, where in the second line we have used Lemma 8.17 for every $\mathbb{E}\left[e^{\lambda X_k}\right]$, the last line follows from the definition of σ^2. Plugging this bound into (8.5) gives

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^m X_k\right) \geq t\right] \leq n \cdot \exp\left(-\lambda t + g(\lambda)\sigma^2\right).$$

Note that this bound holds for all $0 < \lambda < 3/B$, and thus we can minimize the right side over this interval. Indeed, the minimum is attained at $\lambda = t/(\sigma^2 + Bt/3)$, yielding

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^m X_k\right) \geq t\right] \leq n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Bt/3}\right),$$

which is the desirable bound. ■

Theorem 8.16 gives a tail bound on $\|\sum_{k=1}^m X_k\|_2$ and this implies a bound on the expectation.

Theorem 8.18 (Matrix Bernstein in expectation) *Let $X_1, \cdots, X_m$ be independent, mean zero, $n \times n$ symmetric random matrices. Assume $\|X_k\|_2 \leq B$ almost surely for all k and let $\sigma^2 = \left\|\sum_{k=1}^m \mathbb{E}\left[X_k^2\right]\right\|_2$. Then,*

$$\mathbb{E}\left[\left\|\sum_{k=1}^m X_k\right\|_2\right] \lesssim \sigma\sqrt{\log n} + B\log n.$$

Proof: By Theorem 8.16, it is not hard to show that (**check this!**) there exists an absolute numerical constant $c > 0$ such that

$$\mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq c\left(\sigma\sqrt{\log n + u} + B(\log n + u)\right)\right] \leq 2e^{-u}.$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=1}^m X_k\right\|_2\right] &= \int_0^\infty \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq t\right] dt \\
&= \int_0^{c\left(\sigma\sqrt{\log n} + B\log n\right)} \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq t\right] dt + \int_{c\left(\sigma\sqrt{\log n} + B\log n\right)}^\infty \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq t\right] dt \\
&\leq c\left(\sigma\sqrt{\log n} + B\log n\right) \\
&\quad + \int_0^\infty \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq c\left(\sigma\sqrt{\log n + u} + B(\log n + u)\right)\right]\left(\frac{c\sigma}{2\sqrt{\log n + u}} + cB\right) du \\
&\leq c\left(\sigma\sqrt{\log n} + B\log n\right) \\
&\quad + \left(\frac{c\sigma}{2\sqrt{\log n}} + cB\right)\int_0^\infty \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq c\left(\sigma\sqrt{\log n + u} + B(\log n + u)\right)\right] du
\end{aligned}$$

$$\leq c\left(\sigma\sqrt{\log n}+B\log n\right)+\left(\frac{c\sigma}{\sqrt{\log n}}+cB\right)\int_0^\infty e^{-u}du$$
$$\lesssim \sigma\sqrt{\log n}+B\log n,$$

which complete the proof. ∎

The matrix Bernstein inequality can be extended to non-symmetric and non-square matrices.

Theorem 8.19 (Matrix Bernstein inequality for rectangular matrices) *Let $X_1,\cdots,X_m$ be independent, mean zero, $n_1\times n_2$ matrices. Assume $\|X_k\|_2\leq B$ almost surely for all k. Then, for any $t\geq 0$, we have*

$$\mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2\geq t\right]\leq 2(n_1+n_2)\exp\left(-\frac{t^2/2}{\sigma^2+Bt/3}\right),$$

where

$$\sigma^2=\max\left(\left\|\sum_{k=1}^m\mathbb{E}\left[X_kX_k^T\right]\right\|_2,\left\|\sum_{k=1}^m\mathbb{E}\left[X_k^TX_k\right]\right\|_2\right).$$

Proof: Apply Theorem 8.16 to the sum of $\begin{bmatrix}0 & X_k^T\\ X_k & 0\end{bmatrix}$. ∎

8.4 Application: Covariance Matrix for General Distributions

In the first section we have considered the covariance matrix problem when the random vector is sub-Gaussian. In this section we remove the sub-gaussian requirement and consider the case when the random vector has bounded ℓ_2-norm. In this situation, the Bernstein inequality will yield better result than simply using Theorem 8.5 with a crude estimation of the sub-Gaussian parameter based on the ℓ_2-norm of the random vector.

Theorem 8.20 *Let $X_1,\cdots,X_m\in\mathbb{R}^n$ be i.i.d zero mean random vectors with covariance $\Sigma=\mathbb{E}\left[X_kX_k^T\right]$. Assume $\|X_k\|_2\leq\sqrt{b}$ almost surely. Then for any $t>0$, the sample covariance matrix $\Sigma_m=\frac{1}{m}\sum_{k=1}^m X_kX_k^T$ satisfies*

$$\mathbb{P}\left[\|\Sigma_m-\Sigma\|_2\geq t\right]\leq 2n\cdot\exp\left(-\frac{mt^2/2}{b\|\Sigma\|_2+2bt/3}\right).$$

In addition, we have

$$\mathbb{E}\left[\|\Sigma_m-\Sigma\|_2\right]\lesssim\sqrt{\frac{b\|\Sigma\|_2\log n}{m}}+\frac{b\log n}{m}.$$

Proof: First note that if $\|X_k\|_2\leq\sqrt{b}$, there holds (**check this!**)

$$\|\Sigma\|_2=\left\|\mathbb{E}\left[X_kX_k^T\right]\right\|_2\leq b.$$

Letting $Z_k=\frac{1}{m}\left(X_kX_k^T-\Sigma\right)$, it follows that

$$\|Z_k\|_2\leq\frac{1}{m}\left\|X_kX_k^T\right\|_2+\frac{1}{m}\|\Sigma\|_2\leq\frac{2b}{m}.$$

Moreover, we have

$$\mathbb{E}\left[Z_k^2\right] = \frac{1}{m^2}\left(\mathbb{E}\left[(X_k X_k^T)^2\right] - \Sigma^2\right) \preceq \frac{1}{m^2}\mathbb{E}\left[\|X_k\|_2^2 X_k X_k^T\right] \preceq \frac{b}{m^2}\Sigma.$$

It follows that,

$$\sigma^2 = \left\|\sum_{k=1}^m \mathbb{E}\left[Z_k^2\right]\right\|_2 \leq \frac{b\|\Sigma\|_2}{m}.$$

Thus, applying Theorems 8.16 and 8.18 concludes the proof. ■

Example 8.21 *Let $X_k = \sqrt{n}e_{k_j}$, where e_{k_j} is the k_j-th canonical vector in $\mathbb{R}^n$ with k_j being sampled uniformly at random from $\{1, \cdots, n\}$. Then*

$$\mathbb{E}\left[X_k X_k^T\right] = \sum_{j=1}^n e_j e_j^T = I_n \quad \textit{and} \quad \|X_k\|_2 \leq \sqrt{n}.$$

Thus, by Theorem 8.20, we have

$$\mathbb{E}\left[\|\Sigma_m - I_n\|_2\right] \lesssim \sqrt{\frac{n\log n}{m}} + \frac{n\log n}{m}.$$

8.5 Application: Sparse Recovery

Consider the following underdetermined linear system (see Figure 8.3 for a pictorial illustration):

$$y = Ax^* + w, \tag{8.6}$$

where $A \in \mathbb{R}^{m\times n}$ is a fat matrix with $m < n$, y denotes the observation, x^* denotes the parameter to be estimated or signal to be reconstructed, and w denotes the measurement noise. *The goal is to infer or reconstruct x^* from the observation y.*

The linear model (8.6) arises in many statistical and signal processing applications. In statistics, (8.6) models the regime where the number of responses is fewer than the number of predictors (or covariates). In signal processing, it describes the problem where the number of measurements is less than size of the signal. Since the number of unknowns is larger than the number of equations, (8.6) does not admit a unique solution, in contrast to the classical least squares problem. Therefore, additional structures on the unknown vector x^* is needed to reduce the feasible space. In this section we will focus on the sparse solution, namely x^* only has a few nonzero entries.

Definition 8.22 (Sparse vector) *A vector $x \in \mathbb{R}^n$ is said to s-sparse if the number of nonzero entries in x is less than or equal to s. In other words, if we define*

$$\|x\|_0 = \#\left\{k \in \{1, \cdots, n\} : \ x_k \neq 0\right\}$$

which counts the number of nonzero entries in x, then x is s-sparse if $\|x\|_0 \leq s$.

In this lecture we will refer $\|\cdot\|_0$ as the ℓ_0-norm though it is technically not a norm. The notion of sparsity plays an important role in modern statistics, signal processing and machine learning, which characterizes a special type of low dimensional structure.

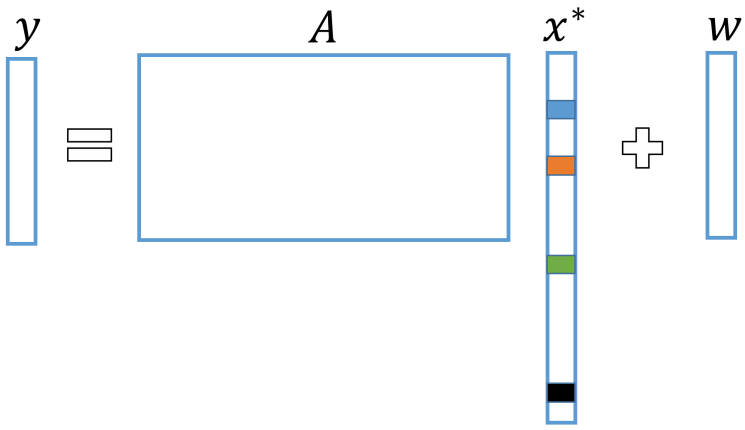


Figure 8.3: A pictorial illustration of (8.6).

.

- In statistics, especially in the context of variable selection, it means only a number of covariates play an important role (a typical example is genome expression).
- In signal processing or machine learning, it means the signal of interest has the sparse structure itself or under certain linear transform.

A basic question to answer is how and when one can reconstruct the sparse vector x^* when there are fewer observations. There have been many methods for sparse parameter estimation or sparse signal reconstruction, including both the convex and nonconvex methods. In this lecture, we study the most widely studied methods based on the ℓ_1-norm. *For simplicity, we only consider the noiseless case (i.e., $w = 0$). The noisy case can discussed in an overall similar way, see the references for details.*

8.5.1 Exact Recovery in the Noiseless Setting

We first study exact recovery problem for the noiseless case, i.e., when $w = 0$. Since we know x^* is a sparse signal it is natural to reconstruct it by seeking the sparsest vector which is consistent with the measurement, namely by solving the following ℓ_0-minimization problem:

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \quad \text{subject to} \quad Ax = y. \tag{8.7}$$

However, the ℓ_0 minimization problem is nonconvex and computationally intractable due to the combinatorial nature of ℓ_0-norm. In optimization, convex relaxation is a widely used technique to handle nonconvex problems. Here, the nearest convex relaxation of the ℓ_0-norm is the ℓ_1-norm which sums up the magnitudes of all the entries of a vector (i.e., $\|x\|_1 = \sum_{k=1}^n |x_k|$). Replacing the ℓ_0-norm with the ℓ_1-norm in the objective leads to the following ℓ_1-minimization,

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \quad \text{subject to} \quad Ax = y. \tag{8.8}$$

The ℓ_1-minimization problem is also known as *basis pursuit* in the literature. It is a convex problem which can be rewritten as a linear programming. It can be solved by the first order or the second

order methods. Indeed, the ℓ_1-minimization problem has spurred the significant development of the first order methods in optimization.

A central question in this section is when the ℓ_1-minimization is able to recover the target sparse solution x^*. To understand why the ℓ_1-minimization returns a sparse solution we first present the intuition and then give a rigorous analysis. Noting that (8.8) is trivially equivalent to

$$\min_{t\in\mathbb{R}} t \quad \text{subject to} \quad \|x\|_1 = t \text{ and } Ax = y.$$

That is, the solution to (8.8) can be found by gradually enlarge the ℓ_1-ball until the ball intersect with the solution set, see Figure 8.4. Since the ℓ_1-ball is pointy at its vertices (or the extreme sets in high dimension), the vertices will first touch the solution set. Noting the vertices have fewer nonzero entries, the ℓ_1-minimization tends to return a sparse solution.

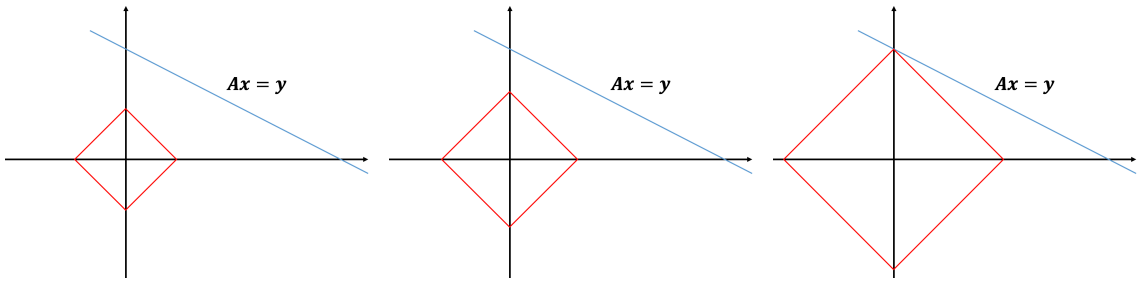


Figure 8.4: A pictorial illustration of ℓ_1-minimization.

.

There are several different conditions which have been developed for the guarantee analysis of the ℓ_1-minimization. In this lecture we will adopt the restricted isometry property proposed by Candes and Tao [2005].

Definition 8.23 (Restricted Isometry Property (RIP)) *Given an integer $s \in \{1, \cdots, n\}$, we say the matrix $A \in \mathbb{R}^{m\times n}$ $(m < n)$ satisfies the restricted isometry property with the constant δ_s if*

$$(1-\delta_s)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta_s)\|x\|_2^2 \tag{8.9}$$

holds for all s-sparse vectors x such that $\|x\|_0 \le s$.

The restricted isometry property basically means that every s columns of A, denoted A_S with $|S| = s$, form a nearly orthogonal matrix when δ_s is small since it can be easily seen that (8.9) is equivalent to

$$\|A_S A_S^T - I_s\|_2 \le \delta_s \tag{8.10}$$

for any subset S of cardinality at most s, where A_S denotes the sub-matrix formed by the columns of A in S.

We are now in position to present a rigorous analysis about when the ℓ_1 minimization is able to exactly reconstruct the target solution x^* based on the restricted isometry property of the matrix.

Theorem 8.24 (Exact recovery) *Let $y = Ax^*$, where x^* is a s-sparse vector (i.e., $\|x^*\|_0 \le s$). If the RIP constant of A of order $3s$ satisfies $\delta_{3s} < 1/3$, then the solution to (8.8) is x^*. That is, the ℓ_1 minimization is able to exactly recovery the sparse vector x^*.*

A careful reader may wonder when a matrix A satisfies the condition $\delta_{3s} < 1/3$. As we will see in the last section, certain random matrix satisfies this condition when $m \gtrsim s \log n$.

Proof: [Proof of Theorem 8.24] Let S denote the support of x^* and S^c denote the complement of S in $\{1, \cdots, n\}$. We first show that for any $x = x^* + h \in \mathbb{R}^n$, if $\|x\|_1 \leq \|x^*\|_1$, then there must hold

$$\|h_{S^c}\|_1 \leq \|h_S\|_1. \tag{8.11}$$

This follows from

$$\|x^*\|_1 \geq \|x\|_1 = \|x^* + h\|_1 = \|x^*_S + h_S\|_1 + \|h_{S^c}\|_1 \geq \underbrace{\|x^*_S\|_1}_{=\|x^*\|_1} - \|h_S\|_1 + \|h_{S^c}\|_1.$$

Thus it suffices to show the following nullspace property[2]: for any h in the nullspace of A (i.e., $Ah = 0$), if h satisfies (8.11), then we must have $h = 0$.

Next we are going to show that if $\delta_{3s} < 1/3$, the nullspace property holds. To this end, let $S_0 = S$ be the support of x^*, let S_1 be the first $2s$ largest entries (in magnitude) of h_{S^c}, let S_2 be the second $2s$ largest entries (in magnitude) of h_{S^c},and so on. Let $h_{S_j} \in \mathbb{R}^n$ be the vector such $h_{S_j}(i) = h(i)$ when $i \in S_j$ and $h(S_j)(i) = 0$. With a slight abuse of notion, we also use h_{S_j} to denote the vector segment supported on S_j. Noting that

$$0 = Ah = Ah_{S_0 \cup S_1} + \sum_{j \geq 2} Ah_{S_j},$$

we have

$$\begin{aligned} 0 &\geq \|Ah_{S_0 \cup S_1}\|_2 - \|\sum_{j \geq 2} Ah_{S_j}\|_2 \\ &\geq \|Ah_{S_0 \cup S_1}\|_2 - \sum_{j \geq 2} \|Ah_{S_j}\|_2 \\ &\geq \sqrt{1 - \delta_{3s}} \|h_{S_0 \cup S_1}\|_2 - \sqrt{1 + \delta_{3s}} \sum_{j \geq 2} \|h_{S_j}\|_2. \end{aligned} \tag{8.12}$$

Moreover, a simple calculation yields that

$$\begin{aligned} \sum_{j \geq 2} \|h_{S_j}\|_2 &\leq \sum_{j \geq 2} \sqrt{2s} \|h_{S_j}\|_\infty \\ &\leq \sum_{j \geq 2} \frac{\|h_{S_{j-1}}\|_1}{\sqrt{2s}} \\ &\leq \frac{1}{\sqrt{2s}} \|h_{S^c}\|_1 \end{aligned}$$

[2]The nullspace property for sparse recovery which basically means that the nullspace of A does not insects with the descent direction of the ℓ_1-norm at x^*. It is actually both sufficient and necessary for exact recovery of basis pursuit, see for example [1]. Theorem 8.24 gives a sufficient condition for this property to hold in terms of the RIP constant. There are also ensemble matrices which violate the RIP but for which the nullspace property holds. Actually, there are many works in the direction of studying the exact recovery directly based on the nullspace property.

$$\leq \frac{1}{\sqrt{2s}}\|h_S\|_1$$
$$\leq \frac{1}{\sqrt{2}}\|h_S\|_2$$
$$\leq \frac{1}{\sqrt{2}}\|h_{S_0\cup S_1}\|_2, \tag{8.13}$$

where the fourth line follows from (8.11). Inserting this inequality into (8.12) gives

$$\left(\sqrt{1-\delta_{3s}} - \frac{\sqrt{1+\delta_{3s}}}{\sqrt{2}}\right)\|h_{S_0\cup S_1}\|_2 \leq 0.$$

Since $\sqrt{1-\delta_{3s}} - \frac{\sqrt{1+\delta_{3s}}}{\sqrt{2}} > 0$ due to the assumption $\delta_{3s} < 1/3$, $\|h_{S_0\cup S_1}\|_2 = 0$ and thus $\|h\|_2 = 0$. ■

8.5.2 Random Matrices Satisfying RIP

Theorem 8.25 *Let A be an $m\times n$ matrix whose rows A_i are independent, isotropic (i.e., $\mathbb{E}\left[A_i^T A_i\right] = I_n$), sub-Gaussian vectors with parameter $\sigma^2 = 1$. Then, if*

$$m \gtrsim \delta^{-2} s \log n,$$

the matrix $A/\sqrt{m}$ satisfies the RIP with a small constant $0 < \delta < 1$ with probability at least $1 - c_2 \cdot \exp\left(-c_4\delta^2 m\right)$, where c_2 and c_4 are numerical constants.

Proof: Recall that, by (8.10), it is enough to show

$$\left\|\frac{1}{m}A_S^T A_S - I_s\right\|_2 \leq \delta$$

for all subsets S of cardinality s, where A_S denotes the sub-matrix constructed from the columns of A in S.

For a fixed subset S, first note that $A_i(S)$ is also σ^2-sub-Gaussian (**why?**) and it also satisfies $\mathbb{E}\left[A_i(S)^T A_i(S)\right] = I_s$. Thus, the application of Theorem 8.5 implies that

$$\mathbb{P}\left[\left\|\frac{1}{m}A_S^T A_S - I_s\right\|_2 \geq c_1\sqrt{\frac{s}{m}} + t\right] \leq c_2 \exp\left(-c_3 \min\{t, t^2\} m\right),$$

provided $m \geq s$. Let $t = \frac{\delta}{2}$. If $m \gtrsim c \cdot \delta^{-2} s \log n$ for a sufficiently large constant $c > 0$, then

$$\left\|\frac{1}{m}A_S^T A_S - I_s\right\|_2 \leq c_1\sqrt{\frac{s}{m}} + \frac{\delta}{2} \leq \delta$$

for all subsets S of cardinality s with probability at least

$$1 - \binom{n}{s} \cdot c_2 \exp\left(-c_3\delta^2 m\right) \geq 1 - c_2 \cdot \exp\left(s\log n - c_3\delta^2 m\right) \geq 1 - c_2 \cdot \exp\left(-c_4\delta^2 m\right),$$

which completes the proof. ■

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapters 6.2, 6.3, 6.4, 7.1, 7.2, 7.3.

[2] Roman Vershynin, *High-Dimensional Probability: An introduction with applications in data science*, Chapters 4.6, 4.7, 5.4, 5.6, 10.5, 10.6.

High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 9: Minimax Lower Bounds

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/05/15)*

Motivation: Consider a set of probability distributions define on $\mathcal{X}$ and indexed by Θ, denoted $\mathcal{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$. For example, θ can denote certain parameter of a distribution or the corresponding probability density function. Given a set of i.i.d data $(X_1, \cdots, X_n)$ sampled from $\mathbb{P}_\theta$ where θ is not known a prior, a fundamental statistical problem is to estimate θ from $\mathcal{D}$.

Let $\widehat{\theta} : (X_1, \cdots, X_n) \to \Theta$ be an estimation procedure. The concentration inequalities and other probability tools presented earlier can help establish an upper bound of the estimation error in terms of[1]

$$\Phi\left(\rho\left(\widehat{\theta}, \theta\right)\right),$$

where $\rho(\cdot, \cdot)$ is a (semi)metric defined on Θ and $\Phi : [0, \infty) \to [0, \infty)$ is an increasing function. As an example, for a univariate mean estimation problem, $\rho(\theta, \theta') = |\theta - \theta'|$ and $\Phi(t) = t^2$ yields the squared error. On the other hand, it is worth investigating whether the estimation error of $\widehat{\theta}$ is optimal. To this end, we study the lower bound of the estimation error based on the *minimax risk*, defined by

$$\mathfrak{M}_n(\Theta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\Phi\left(\rho\left(\widehat{\theta}, \theta\right)\right)\right], \tag{9.1}$$

where the subscript θ means that $X_1, \cdots, X_n$ are sampled from $\mathbb{P}_\theta$. That is, for a fixed estimation procedure we consider the worst case error by taking the supremum over all the distributions, and then study the smallest worst case error achievable by *any procedure*.

There are two methods for obtaining the minimax lower bound: Bayesian analysis and reduction to hypothesis testing. We will focus on the latter one since it is more versatile and can be applied to most situations. *To gain some intuition of the hypothesis testing method,consider the minimax risk of estimating a scalar parameter in terms of the risk function $|\theta - \theta'|$. Suppose there are two point θ_1 and θ_2 such that $|\theta_1 - \theta_2| \geq \delta$. If whatever method we use to test which point the observed data comes from the probability of testing error is a constant, then the estimation error for any procedure should be greater than a multiple of δ since with constant probability we are likely to mistaken one from the other.* Of course we can also consider the problem of testing multiple points. Thus, overall the problem is about how to choose the testing points such that they are as far away as possible while the probability of testing error for any testing method remains a constant.

In this lecture we discuss two standard techniques for establishing the lower bounds of the minimax risks based on testing, including the Le Cam and Fano methods. Roughly speaking, Le Cam method is based on binary testing and Fano methods are based on multiway hypothesis testing. There is another method which is not covered in this lecture, known as Assouad method, for lower bounding the minimax risk. Assouad method is based on the multiple binary hypothesis testing when the risk function is separable, see for example Chapter 8 and 9 of [2].

Agenda:

[1] We use $\widehat{\theta}$ to denote $\widehat{\theta}(X_1, \cdots, X_n)$ for simplicity.

- Reduction to hypothesis testing
- Some divergence measures
- Le Cam method
- Fano methods

9.1 Reduction to Hypothesis Testing

Let $\{\theta_1, \cdots, \theta_m\}$ be a 2δ-packing of the space Θ under the (semi)metric ρ, i.e., $\rho(\theta_i, \theta_j) \geq 2\delta$ for all $i \geq j$. Define $\mathbb{P}_j^n = \mathbb{P}_{\theta_j} \times \cdots \times \mathbb{P}_{\theta_j}$. First, by the Markov inequality we have

$$\mathbb{E}_{\theta_j}\left[\Phi\left(\rho(\widehat{\theta}, \theta_j)\right)\right] \geq \Phi(\delta) \cdot \mathbb{P}_j^n\left[\Phi\left(\rho(\widehat{\theta}, \theta_j)\right) \geq \Phi(\delta)\right] \geq \Phi(\delta) \cdot \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta_j) \geq \delta\right],$$

where we note that $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$, and $\mathbb{P}_j^n$ indicates that $(X_1, \cdots, X_n)$ are sampled from $\mathbb{P}_{\theta_j}$. In addition, the second inequality is due to the fact that Φ is increasing. It follows that

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta\left[\Phi\left(\rho\left(\widehat{\theta}, \theta\right)\right)\right] \geq \max_{\theta_j} \mathbb{E}_{\theta_j}\left[\Phi\left(\rho(\widehat{\theta}, \theta_j)\right)\right] \geq \Phi(\delta)\left(\frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta_j) \geq \delta\right]\right).$$

Next we will show $\frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta) \geq \delta\right]$ can be lower bounded by hypothesis testing error.

In the hypothesis testing, a test function is a map from a set of i.i.d data sampled from one of $\{\mathbb{P}_{\theta_j},\ j = 1, \cdots, m\}$ to $\{1, \cdots, m\}$, which is used to infer from which probability distribution the data comes from. Given an estimation procedure $\widehat{\theta}$, we can define a test function naturally as follows:

$$\widehat{\Psi}(X_1, \cdots, X_n) = \arg\min_{\ell \in [m]} \rho(\widehat{\theta}(X_1, \cdots, X_n), \theta_\ell),$$

where the tier is broken arbitrarily. Since $\{\theta_1, \cdots, \theta_m\}$ is a 2δ-packing of Θ, it is clear that (see Figure 9.1)

$$\rho(\widehat{\theta}, \theta_j) < \delta \quad \Rightarrow \quad \widehat{\Psi} = j.$$

Thus, when $(X_1, \cdots, X_n)$ are sampled from $\mathbb{P}_{\theta_j}$, we have[2]

$$\mathbb{P}_j^n\left[\widehat{\Psi} \neq j\right] \leq \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta_j) \geq \delta\right].$$

Consequently,

$$\frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta_j) \geq \delta\right] \geq \frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n\left[\widehat{\Psi} \neq j\right].$$

[2]We also use $\widehat{\Psi}$ to denote $\widehat{\Psi}(X_1, \cdots, X_n)$ for simplicity.

Moreover, we have

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta\left[\Phi\left(\rho\left(\widehat{\theta},\theta_j\right)\right)\right]\geq\Phi(\delta)\left(\frac{1}{m}\sum_{j=1}^m\mathbb{P}_j^n\left[\widehat{\Psi}\neq j\right]\right).$$

Taking the infimum over all estimation procedures $\widehat{\theta}$ on the lefthand side and the infimum over all test functions yields the following proposition.

Proposition 9.1 *Under the setup of the above test problem, the minimax risk* (9.1) *is lower bounded as*

$$\mathfrak{M}_n(\Theta)\geq\Phi(\delta)\inf_{\Psi}\left(\frac{1}{m}\sum_{j=1}^m\mathbb{P}_j^n\left[\Psi(X_1,\cdots,X_n)\neq j\right]\right),\tag{9.2}$$

where the infimum ranges over all test functions. Note that δ is parameter that is free to choose and it denotes the minimum distance between θ_i and θ_j for all $i\neq j$.

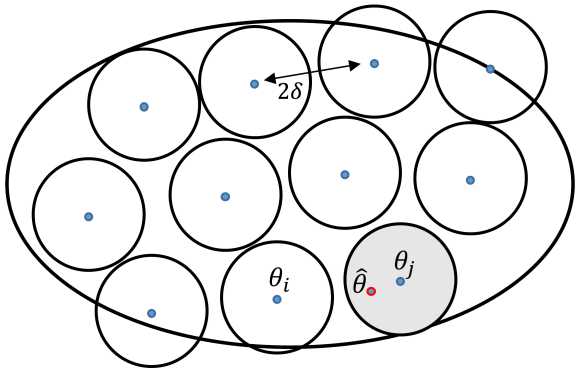


Figure 9.1: An illustration of 2δ-packing.

Consider a joint distribution (J,Z^J), where J is uniform distributed in $\{1,\cdots,m\}$ and given $J=j$, $Z^j=(X_1,\cdots,X_n)$ obeys the distribution of $\mathbb{P}_j^n$. It is clear that the joint distribution obeys

$$\mathbb{Q}\left[Z^J\in\cdot,J=j\right]=\frac{1}{m}\mathbb{P}_j^n\left[Z^j\in\cdot\right],$$

and the marginal distributions are given by

$$\mathbb{Q}_J\left[J=j\right]=\frac{1}{m}\quad\text{and}\quad\mathbb{Q}_Z\left[Z^J\in\cdot\right]=\frac{1}{m}\sum_{j=1}^m\mathbb{P}_j^n\left[Z^j\in\cdot\right].$$

Moreover, for any test function Ψ, we have

$$\mathbb{Q}\left[\Psi(Z^J)\neq J\right]=\sum_{j=1}^m\mathbb{Q}\left[\Psi(Z^J)\neq J,J=j\right]$$

$$= \sum_{j=1}^{m} \mathbb{Q}\left[\Psi(Z^J) \neq j, J = j\right]$$
$$= \frac{1}{m} \sum_{j=1}^{m} \mathbb{P}_j^n(\Psi(Z^j) \neq j).$$

Therefore, we can rewrite (9.2) as

$$\mathfrak{M}_n(\Theta) \geq \Phi(\delta) \inf_{\Psi} \mathbb{Q}\left[\Psi(Z^J) \neq J\right], \tag{9.3}$$

which will be used in the sequel for conciseness.

Remark 9.2 *In words, reduction to hypothesis testing lower bounds the best achievable estimation error by a multiple of the failing probability of test. It is not hard to imagine that the smallest mis-test probability fundamentally relies on how close $\mathbb{P}_j^n$ are, which enables us to provides a bound independent of the test function. Moreover, the lower bound in* (9.2) *or* (9.3) *is a function of the separation δ, which trades off between $\Phi(\delta)$ (increases as δ increases) and the probability of test error $\inf_\Psi \mathbb{Q}\left[\Psi(Z^J) \neq J\right]$ (relying on δ implicitly, decreases as δ increases).* In order to obtain a desirably large lower bound, one usually chooses the largest[3] δ such that $\inf_\Psi \mathbb{Q}\left[\Psi(Z^J) \neq J\right]$ is greater than a constant[4] (for example 1/2) and then uses the corresponding $\Phi(\delta)$ to provide lower bound. *As we have explained in the motivation part, the intuition is that if the parameters are far away (i.e, by choosing the largest possible δ) but it is still difficult to distinguish the related distributions from the observations (i.e., probability of testing error is constant), then the estimation error must be lower bounded by related function of the parameter distance since we can mistaken one for the other. Next, we will present two concrete methods: the Le Cam and Fano methods.*

9.2 Some Divergence Measures

We first take a detour and present some inequalities for divergence measures and their consequences for product distributions. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions defined on $\mathcal{X}$. Assume they have densities $p(x)$ and $q(x)$ respectively with respect to some underlying base measure μ. The three related divergences are

- KL divergence: $D(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \mu(dx)$,
- TV distance: $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \mu(dx)$,
- Hellinger distance: $H^2(\mathbb{P}\|\mathbb{Q}) = \int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mu(dx)$.

Recall that KL divergence and TV distance have also been mentioned in Lecture 3. The three divergence measures are related as follows.

Lemma 9.3 *For two distributions $\mathbb{P}$ and $\mathbb{Q}$, we have*

1. $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{Q}\|\mathbb{P})}$,

[3] As can be seen δ may rely on other parameters, such as the number of samples.

[4] That is, choose the largest possible δ that the testing problem is still sufficiently challenging.

2. $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq H(\mathbb{P}\|\mathbb{Q})\sqrt{1 - \frac{H^2(\mathbb{P}\|\mathbb{Q})}{4}}$.

Proof: The proof for the first inequality can be found in Lecture 3. The second inequality can be proved by the Cauchy-Schwarz inequality (**check this!**). ■

Recall that $\mathbb{P}^n$ (respectively, $\mathbb{Q}^n$) is the product distribution on the product space $\mathcal{X}^n$ (i.e., the distribution of n i.i.d random variables). It is desirable to express the distance between $\mathbb{P}^n$ and $\mathbb{Q}^n$ in terms of $\mathbb{P}$ and $\mathbb{Q}$. For TV distance, it is difficult to express $\|\mathbb{P}^n - \mathbb{Q}^n\|_{\mathrm{TV}}$ in terms of $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}$. For KL divergence and the Hellinger distance we have the following lemma.

Lemma 9.4 *For two distributions $\mathbb{P}$ and $\mathbb{Q}$ and the corresponding , we have*

1. $D(\mathbb{Q}^n\|\mathbb{P}^n) = nD(\mathbb{Q}\|\mathbb{P})$,
2. $H^2(\mathbb{P}^n\|\mathbb{Q}^n) \leq nH^2(\mathbb{P}\|\mathbb{Q})$.

Proof: The first inequality can be proved directly using the fact that the density functions for $\mathbb{P}^n$ and $\mathbb{Q}^n$ are $p(x_1)\cdots p(x_n)$ and $q(x_1)\cdots q(x_n)$ respectively. Additionally, it can be shown that (**check this!**)

$$\frac{1}{2}H^2(\mathbb{P}^n\|\mathbb{Q}^n) = 1 - \left(1 - \frac{1}{2}H^2(\mathbb{P}\|\mathbb{Q})\right)^n.$$

Then the second inequality follows immediately since $(1-x)^n \geq 1 - nx$ for $x \in [0,1]$. ■

9.3 Le Cam Method

Le Cam method provides lower bounds on the minimax using the simple binary hypothesis testing. This section explores this connection based on the total variation distance.

Lemma 9.5 *In the case of binary hypothesis testing, we have*

$$\inf_{\Psi} \mathbb{Q}\left[\Psi(Z^J) \neq J\right] = \frac{1}{2}\left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}\right),$$

where $\mathbb{P}_1^n$ and $\mathbb{P}_2^n$ are product distributions corresponding to θ_1 and θ_2, respectively.

Proof: For any test function Ψ defined on $\mathcal{X}^n$, let

$$A = \{(x_1, \cdots, x_n) \in \mathcal{X}^n : \ \Psi(x_1, \cdots, x_n) = 1\}.$$

and A^c be the complementary on which $\Psi = 2$. Then we have

$$\begin{aligned}
\sup_{\Psi} \mathbb{Q}\left[\Psi(Z^J) = J\right] &= \sup_{A} \frac{1}{2}\left(\mathbb{P}_1^n[A] + \mathbb{P}_2^n[A^c]\right) \\
&= \frac{1}{2} + \frac{1}{2}\sup_{A}\left(\mathbb{P}_1^n[A] - \mathbb{P}_2^n[A]\right) \\
&= \frac{1}{2} + \frac{1}{2}\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}.
\end{aligned}$$

Noting that $\sup_\Psi \mathbb{Q}\left[\Psi(Z^J) = J\right] = 1 - \inf_\Psi \mathbb{Q}\left[\Psi(Z^J) \neq J\right]$, the claim follows immediately. ■

Combining the above lemma and Proposition 9.1 together yields the following minimax risk bound.

Proposition 9.6 *We have*

$$\mathfrak{M}_n(\Theta) \geq \frac{\Phi(\delta)}{2}\left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}\right)$$

for ***any pair*** *of distributions θ_1 and θ_2 satisfying $\rho(\theta_1, \theta_2) \geq 2\delta$.*

Note that as δ decreases $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}$ decreases, and the binary hypothesis testing problem becomes more challenging. In practice, we roughly attempt to choose the largest possible δ such that $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}$ is a small constant so that we can still mistaken the θ_1 and θ_2 (yielding the lower bound of the estimation error depending on δ) .

Example 9.7 *Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$ be a family of normal distributions $\mathcal{N}(\theta, \sigma^2)$ with fixed variance σ^2. We study the minimax risk of estimating θ from i.i.d samples $\{X_k\}_{k=1}^n$ drawn from $\mathbb{P}_\theta$. We consider two parameters $\theta_1 = 0$ and $\theta_2 = \theta$ satisfying $\theta = 2\delta$. In order to apply the Le Cam method, we need to bound $\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\mathrm{TV}}$. Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ defined over $\mathcal{X}$, respectively with their probability densities $p(x)$ and $q(x)$ under some base measure μ, it can be easily shown that (****check this!****)*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left(\int_{\mathcal{X}} \frac{p^2(x)}{q(x)}\mu(dx) - 1\right).$$

Using this result for $\mathbb{P}_0^n$ and $\mathbb{P}_\theta^n$ on $\mathcal{X} = \mathbb{R}^n$ yields that

$$\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left(\exp\left(n\theta^2/\sigma^2\right) - 1\right) = \frac{1}{4}\left(\exp\left(4n\delta^2/\sigma^2\right) - 1\right).$$

Taking $\delta = \frac{1}{2}\frac{\sigma}{\sqrt{n}}$ yields that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[|\widehat{\theta} - \theta|^2\right] \geq \frac{\delta^2}{2}\left(1 - \sqrt{e-1}/2\right) \geq \frac{\delta^2}{6} = \frac{1}{24}\frac{\sigma^2}{n}.$$

*The scale σ^2/n is sharp, and the sample mean $\widehat{\theta} = \frac{1}{n}\sum_{k=1}^n X_k$ satisfies this bound (****check this!****).*

9.4 Fano Methods

The Fano methods provide lower bounds based on the multiway hypothesis testing and the Fano inequality in information theory.

9.4.1 Information Theory Basics

Information theory is essentially about studying the information or randomness stored in probability distributions. Here we provide some basic materials in information theory that is needed for lower bounding the minimax risk. More details about information can be found in the book *Elements of Information Theory*. The fundamental notion in information theory is Shannon entropy.

Definition 9.8 (Shannon entropy) *Let $X \sim \mathbb{Q}$ where $\mathbb{Q}$ is a probability distribution on $\mathcal{X}$ with density $q(x)$ with respect to some base measure μ. The Shannon entropy of[5] X is*

$$H(X) = -\int_{\mathcal{X}} q(x) \log q(x) \mu(dx). \tag{9.4}$$

Lets gain some intuition of Shannon entropy by looking at discrete random variables. When X is a discrete random variable, we can take $\mathcal{X}$ as a finite set and take μ as a counting measure on $\mathcal{X}$. In this case, the definition (9.4) reduces to the discrete entropy[6]

$$H(X) = -\sum_{x \in \mathcal{X}} q(x) \log q(x). \tag{9.5}$$

It measures on average how many bits are needed to store the probability distribution of $\mathcal{X}$. More precisely, to represent the probability for $X = x$ (i.e., $q(x)$), we need $\log 1/(q(x))$ bits since it corresponds to $1/q(x)$ possibilities. Thus, on average we need $H(x)$ bits to store the distribution of X. In the simplest uniform distribution case $q(x) = 1/|\mathcal{X}|$, $\log |\mathcal{X}|$ bits are needed to denote all $|\mathcal{X}|$ possibilities.

Lemma 9.9 *For **discrete entropy**, we have $0 \leq H(X) \leq \log |\mathcal{X}|$.*

It is worth noting that for differential entropy (i.e., entropy of continuous random variables), $H(X) \geq 0$ is not always true since $q(x)$ can be greater than 1 (for example consider a uniform distribution over a small interval). The upper bound $\log |\mathcal{X}|$ is achieved by the uniform distribution on $\mathcal{X}$, i.e., $\mathbb{Q}(X = x) = \frac{1}{|\mathcal{X}|}$.

Proof: The lower bound $H(X) \geq 0$ follows from $q(x) \leq 1$ and the upper bound follows from Jensen inequality. ■

We can also define the conditional entropy, which is the amount of information left in a random variable after observing another.

Definition 9.10 (Conditional entropy) *Given a pair of random variables(X, Y) on $(\mathcal{X}, \mathcal{Y})$ with joint distribution $\mathbb{Q}_{X,Y}$, the conditional entropy of $X|Y$ is defined as*

$$H(X|Y) = \mathbb{E}_Y \left[-\int_{\mathcal{X}} q(x|Y) \log q(x|Y) \mu(dx) \right].$$

In addition, given two random variables, we can define the mutual information between them.

Definition 9.11 (Mutual information) *Given a pair of random variables(X, Y) on $(\mathcal{X}, \mathcal{Y})$ with joint distribution $\mathbb{Q}_{X,Y}$, let $\mathbb{Q}_X$ and $\mathbb{Q}_Y$ denote the respect marginal distributions. The mutual information of X and Y is defined as*

$$I(X, Y) = D(\mathbb{Q}_{X,Y} \| \mathbb{Q}_X \mathbb{Q}_Y).$$

[5]Shannon entropy is actually a function of probability distributions since there are many different random variables obeying the same distribution. Here we just follow the standard practice in information theory and treat it as a function of random variables.

[6]Note that, for continuous random variables, the Shannon entropy is often referred to as the differential entropy.

We first note that $I(X,Y) \geq 0$, and $I(X,Y) = 0$ if and only if X and Y are independent. Thus, it can be thought as a way to measure the amount of dependence between X and Y. When X and Y are independent, $I(X;Y) = 0$.

We have the following properties about entropy, conditional entropy and mutual information.

Lemma 9.12 *We have*

1. $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$,
2. $H(X,Y|Z) = H(X|Z) + H(Y|X,Z) = H(Y|Z) + H(X|Y,Z)$,
3. $I(X,Y) = H(X) + H(Y) - H(X,Y)$
4. $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$,
5. $H(X|Y) \leq H(X)$, $H(Y|X) \leq H(Y)$,
6. $H(Y|X) = 0$ *if* $Y = f(X)$, *i.e., when* Y *is a function of* X.

Proof: Whenever it is possible, we will assume the existence of the (conditional) density functions in the proofs for conciseness.

The first two identities are known as the chain rule for entropy. We only prove the first equality in 1 and 2 since the other two can be proved similarly. Noting that $q(y|x) = \frac{q(x,y)}{q(x)}$, we have

$$\begin{aligned} H(Y|X) &= -\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} q(y|x) \log q(y|x) \mu(dy) \right) q(x) \mu(dx) \\ &= -\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \frac{q(x,y)}{q(x)} \log \frac{q(x,y)}{q(x)} \mu(dy) \right) q(x) \mu(dx) \\ &= H(X,Y) - H(X). \end{aligned}$$

Similarly, noting that $q(y|x,z) = \frac{q(x,y,z)}{q(x,z)} = \frac{q(x,y|z)q(z)}{q(x,z)} = \frac{q(x,y|z)}{q(x|z)}$ and $q(x,z) = q(x|z)q(z)$, we have

$$\begin{aligned} H(Y|X,Z) &= -\int_{\mathcal{X}} \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \frac{q(x,y|z)}{q(x|z)} \log \frac{q(x,y|z)}{q(x|z)} \mu(dy) \right) q(x|z) q(z) \mu(dx) \mu(dz) \\ &= -\int_{\mathcal{X}} \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} q(x,y|z) \log q(x,y|z) \mu(dy) \right) q(z) \mu(dx) \mu(dz) \\ &\quad + \int_{\mathcal{X}} \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} q(x,y|z) \log q(x|z) \mu(dy) \right) q(z) \mu(dx) \mu(dz) \\ &= -\int_{\mathcal{X}} \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} q(x,y|z) \log q(x,y|z) \mu(dy) \right) q(z) \mu(dx) \mu(dz) \\ &\quad + \int_{\mathcal{Z}} \left(\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} q(x,y|z) \mu(dy) \right) \log q(x|z) \mu(dx) \right) q(z) \mu(dz) \\ &= H(X,Y|Z) - H(X|Z). \end{aligned}$$

Expanding the expression for $I(X,Y)$,

$$I(X,Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} q(x,y) \log \frac{q(x,y)}{q(x)q(y)} \mu(dx) \mu(dy),$$

yields 3 straightforwardly.

Combining 1 and 2 together yields 4, and 5 follows from 4 directly. Note that 5 means the conditional entropy is always less than or equal to the entropy. That is, considering the entropy under certain condition only decreases the uncertainty of a random variable. Moreover, if X and Y are independent, then $H(X|Y) = H(X)$, so in this situation observing Y will not reduce the uncertainty in X.

When $Y = f(X)$, $Y|(X = x)$ is deterministic or it is a discrete random variable only taking one value at $f(x)$. Evidently, we have $H(Y|X) = 0$. This means there is no uncertainty in Y once X is observed and hence $H(Y|X) = 0$. ■

Now we are ready to present and prove the Fano inequality in information theory. Let X be random variable on a finite set $\mathcal{X}$. Assume we observe a different random variable Y, and want to estimate $\mathbb{Q}[\Psi(Y) \neq X]$, where $\mathbb{Q}$ is the joint distribution of X and Y, and $\Psi(\cdot)$ is a test function.

Lemma 9.13 (Fano inequality) *We have*

$$\mathbb{Q}[\Psi(Y) \neq X] \geq \frac{H(X|Y) - \log 2}{\log |\mathcal{X}|}. \tag{9.6}$$

Proof: Let E be the random variable such that $E = 1$ if $\Psi(Y) \neq X$ and $E = 0$ otherwise. The proof follows by expanding $H(X, E|Y)$ in two different ways given in 2 of Lemma 9.12.

Letting $h = -p \log p - (1 - p) \log(1 - p)$, we have

$$\begin{aligned} H(X, E|Y) &= H(E|Y) + H(X|E, Y) \\ &= \underbrace{H(E|Y)}_{\leq H(E)} + \underbrace{\mathbb{Q}[E = 1] H(X|E = 1, Y)}_{\leq \mathbb{Q}[E=1] \log(|\mathcal{X}|-1)} + \underbrace{\mathbb{Q}[E = 0] H(X|E = 0, Y)}_{=0} \\ &\leq h(\mathbb{Q}[\Psi(Y) \neq X]) + \mathbb{Q}[\Psi(Y) \neq X] \log(|\mathcal{X}| - 1), \end{aligned}$$

where we have used the fact that conditioned on $E = 1, Y = y$, X can only take $|\mathcal{X}| - 1$ possible values and conditioned on $E = 0, Y = y$, $X = \Psi(y)$ is deterministic. On the other hand,

$$H(X, E|Y) = H(X|Y) + H(E|X, Y) = H(X|Y),$$

where $H(E|X, Y) = 0$ due to 6 of Lemma 9.12. Combining the above two inequalities together and further noting $h(p) \leq \log 2$ for all $p \in [0, 1]$ concludes the proof. ■

9.4.2 Fano Lower Bound on Minimax Risk

Recall that the minimax risk can be lower bounded by $\Phi(\delta) \inf_\Psi \mathbb{Q}\left[\Psi(Z^J) \neq J\right]$, where the random variable Z^J is generated by first sampling J uniformly from $[m] = \{1, \cdots, m\}$ and then generating Z^J according to $\mathbb{P}_j^n$ (here $\mathbb{P}_j^n, j = 1, \cdots, m$ are the product distributions which corresponds to the 2δ-separated set $\{\theta_j\}_{j=1}^m$), see Section 9.1 for details. Intuitively, $\mathbb{Q}\left[\Psi(Z^J) \neq J\right]$ should relate to the dependence between Z^J and J. For example, if Z^J is independent of J, it would be impossible to tell J from Z^J. Since $I(Z^J, J)$ provides one way to characterize the dependence between Z^J and J in terms of the KL divergence. It is reasonable to bound $\mathbb{Q}\left[\Psi(Z^J) \neq J\right]$ by $I(Z^J, J)$ and then provide a minimax lower bound based on it. Indeed, we have the following theorem.

Theorem 9.14 *Under the setting for the construction of J and Z^J in Section 9.1, we have*

$$\mathfrak{M}_n(\Theta) \geq \Phi(\delta)\left(1 - \frac{I(Z^J, J) + \log 2}{\log m}\right). \tag{9.7}$$

Proof: It suffices to show that

$$\mathbb{Q}\left[\Psi(Z^J) \neq J\right] \geq 1 - \frac{I(Z^J, J) + \log 2}{\log m}. \tag{9.8}$$

To this end, letting $X = J$ and $Y = Z^J$ in (9.6) and further noting $H(J|Z^J) = H(J) - I(Z^J, J) = \log m - I(Z^J, J)$ shows (9.8). ■

In order to apply Theorem 9.14, we need to further upper bound $I(Z^J, J)$. The local Fano method and global Fano method establish the lower minimax risk bound by upper bounding $I(Z^J, J)$ in different ways.

9.4.3 Local Fano Method

The mutual information can be written in terms of the component distributions $\{\mathbb{P}_j^n\}_{j=1}^m$ and the mixture distribution $\mathbb{Q}_Z = \frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n$ as follows

$$I(Z^J, J) = \frac{1}{m}\sum_{j=1}^m D(\mathbb{P}_j^n \| \mathbb{Q}_Z). \tag{9.9}$$

Letting $p_j^n(x_1, \cdots, x_n)$ be the density of $\mathbb{P}_j^n$ under some base measure $\mu(dx_1 \cdots dx_n)$ and noting that $\frac{1}{m}$ is the density of $\mathbb{Q}_J$ under the counting measure $\mu(dj)$, the density of the joint distribution $\mathbb{Q}$ under the base measure $\mu(dx_1 \cdots dx_n)\ \mu(dj)$ is given by $\frac{1}{m}p_j^n(x_1, \cdots, x_n)$, and the density of $\mathbb{Q}_Z$ is given by $\frac{1}{m}\sum_{j=1}^m p_j^n(x_1, \cdots, x_n)$. Thus a simple calculation yields,

$$\begin{aligned}
I(Z^J, J) &= \int_{\mathcal{X}^n \times [m]} \frac{1}{m} p_j^n(x_1, \cdots, x_n) \log \frac{\frac{1}{m}p_j^n(x_1, \cdots, x_n)}{\left(\frac{1}{m}\sum_{i=1}^m p_i^n(x_1, \cdots, x_n)\right)\frac{1}{m}} \mu(dx_1 \cdots dx_n)\mu(dj) \\
&= \frac{1}{m}\sum_{j=1}^m \int_{\mathcal{X}^n} p_j^n(x_1, \cdots, x_n) \log \frac{p_j^n(x_1, \cdots, x_n)}{\left(\frac{1}{m}\sum_{i=1}^m p_i^n(x_1, \cdots, x_n)\right)} \mu(dx_1 \cdots dx_n) \\
&= \frac{1}{m}\sum_{j=1}^m D(\mathbb{P}_j^n \| \mathbb{Q}_Z),
\end{aligned}$$

which proves (9.9). In addition, we have

$$\begin{aligned}
D(\mathbb{P}_j^n \| \mathbb{Q}_Z) &= D(\mathbb{P}_j^n \| \frac{1}{m}\sum_{i=1}^m \mathbb{P}_i^n) \\
&= \int_{\mathcal{X}^n} p_j^n(x_1, \cdots, x_n) \log \frac{p_j^n(x_1, \cdots, x_n)}{\frac{1}{m}\sum_{i=1}^m p_i^n(x_1, \cdots, x_n)} \mu(dx_1 \cdots dx_n) \\
&= -\int_{\mathcal{X}^n} p_j^n(x_1, \cdots, x_n) \log \frac{\frac{1}{m}\sum_{i=1}^m p_i^n(x_1, \cdots, x_n)}{p_j^n(x_1, \cdots, x_n)} \mu(dx_1 \cdots dx_n)
\end{aligned}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\int_{\mathcal{X}^n} p_j^n(x_1,\cdots,x_n)\log\frac{p_j^n(x_1,\cdots,x_n)}{p_i^n(x_1,\cdots,x_n)}\mu(dx_1\cdots dx_n)$$
$$= \frac{1}{m}\sum_{i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n),$$

where the fourth line follows from the Jensen inequality. Inserting this inequality into (9.9) yields

$$I(Z^J, J) \leq \frac{1}{m^2}\sum_{j,i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n). \tag{9.10}$$

Therefore, we have the following proposition.

Proposition 9.15 *Under the setting for the construction of J and Z^J in Section 9.1, we have*

$$\mathfrak{M}_n(\Theta) \geq \Phi(\delta)\left(1 - \frac{\frac{1}{m^2}\sum_{j,i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n) + \log 2}{\log m}\right). \tag{9.11}$$

To apply the bound in (9.11), we need to construct a family of distributions $\{\mathbb{P}_j\}_{j=1}^m$ corresponding to $\{\theta_j\}_{j=1}^m$ such that

- $\rho(\theta_j, \theta_\ell) \geq 2\delta$, and m can be as large as possible,
- $D(\mathbb{P}_j^n\|\mathbb{P}_i^n)$ is sufficiently small.

Due to the second constraint, we cannot construct a packing of the entire space Θ; otherwise, $\max_{i,j} D(\mathbb{P}_j^n\|\mathbb{P}_i^n)$ would be large. Instead, the *local Fano method* construct a packing of local subset by first construct a packing set of a fixed radius and then shrinking the packing sets by δ, which leaves the packing number unchanged but gives us the room to choose a δ that is sufficiently small such that $D(\mathbb{P}_j^n\|\mathbb{P}_i^n)$ can be sufficiently small such that $1 - \frac{\frac{1}{m^2}\sum_{j,i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n)+\log 2}{\log m}$ is larger than a small constant. Let illustrate this with two examples.

Example 9.16 *We consider the mean estimation of multivariate normal distributions (in contrast to Example 9.7) $\mathcal{N}(\theta, \sigma^2 I_d)$, where $\theta \in \mathbb{R}^d$. It is not hard to show that the mean squared error of the sample mean estimator is of the order $\frac{d\sigma^2}{n}$ (**check this!**). In this example we will show that the minimax risk of the means squared error is $\gtrsim \frac{d\sigma^2}{n}$*

To this end, let $\{x_1,\cdots,x_m\}$ be a 1/2 packing of the unit ℓ_2-ball with $\log m \geq d\log 2$. Define $\theta_j = 4\delta x_j$. Then it is trivial that $\|\theta_i - \theta_j\|_2 \geq 2\delta$ and $\|\theta_i - \theta_j\|_2 \leq 8\delta$. In addition, we have

$$D(\mathbb{P}_j^n\|\mathbb{P}_i^n) = nD(\mathbb{P}_j\|\mathbb{P}_i) = nD(\mathcal{N}(\theta_j,\sigma^2 I_d)\|\mathcal{N}(\theta_i,\sigma^2 I_d)) = \frac{n}{2\sigma^2}\|\theta_j - \theta_i\|_2^2 \leq \frac{32n\delta^2}{\sigma^2}.$$

It follows that

$$\frac{\frac{1}{m^2}\sum_{j,i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n) + \log 2}{\log m} \leq \frac{\frac{32n\delta^2}{\sigma^2} + \log 2}{d\log 2} \lesssim \frac{1}{2},$$

if we choose $\delta^2 \asymp \frac{d}{n}\sigma^2$. Thus, we conclude that

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathbb{R}^d}\mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] \gtrsim \frac{d\sigma^2}{n}.$$

Example 9.17 *Consider the model $Y = A\theta^* + w$, where $A \in \mathbb{R}^{n\times d}$ is fixed and $rank(A) = d$, and $w \sim \mathcal{N}(0, \sigma^2 I_n)$. We want to lower bound the minimax risk when estimating θ^* from Y under the (semi)metric*

$$\rho(\theta, \theta') = \frac{\|A(\theta - \theta')\|_2}{\sqrt{n}}.$$

Define the set $S = \{x \in \text{range}(A) : \|x\|_2 = 1\}$. We can construct a $1/2$-packing of S with the packing number m satisfying $\log m \geq d \log 2$. Let $\{x_1, \cdots, x_m\}$ denote the packing set, the goal is to construct a set $\{\theta_1, \cdots, \theta_m\}$ such that $\rho(\theta_i, \theta_j) \geq 2\delta$. To this end, define θ_j to be the vector such that $A\theta_j = 4\delta\sqrt{n}x_j$. Then, it is easy to verify that

$$\rho(\theta_i, \theta_j) = \frac{\|A(\theta_i - \theta_j)\|_2}{\sqrt{n}} = 4\delta\|x_i - x_j\|_2,$$

and consequently, $2\delta \leq \rho(\theta_i, \theta_j) \leq 8\delta$.

Note that the observations $Y = (Y_1, \cdots, Y_n) \sim \mathcal{N}(A\theta, \sigma^2 I_n)$. By the divergence property of multivariable Gaussian distribution, we have

$$D(\mathbb{P}_j^n \| \mathbb{P}_i^n) = \frac{1}{2\sigma^2}\|A(\theta_j - \theta_i)\|_2^2 \leq \frac{32n\delta^2}{\sigma^2}.$$

It follows that

$$\frac{\frac{1}{m^2}\sum_{j,i=1}^m D(\mathbb{P}_j^n \| \mathbb{P}_i^n) + \log 2}{\log m} \leq \frac{\frac{32n\delta^2}{\sigma^2} + \log 2}{d \log 2} \lesssim \frac{1}{2},$$

if we choose $\delta^2 \asymp \frac{d}{n}\sigma^2$. Thus, we conclude that

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^d} \mathbb{E}\left[\frac{\|A(\widehat{\theta} - \theta)\|_2^2}{n}\right] \gtrsim \frac{d\sigma^2}{n}.$$

*This bound is sharp in order which can be achieved for example by the least-squares estimator (**check this!**).*

9.4.4 Global Fano Method

Recall from (9.9) that

$$I(Z^J, J) = \frac{1}{m}\sum_{j=1}^m D(\mathbb{P}_j^n \| \mathbb{Q}_Z), \quad \mathbb{Q}_Z = \frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n.$$

Thus, if we can construct a packing of $\mathcal{P}^n$ in terms of the KL divergence, it is likely to bound $I(Z^J, J)$ using the packing of the all the distributions. This leads to the global Fano method, also known as Yang-Barron method.

Lemma 9.18 *Let N_{KL} be the ε-covering number of $\mathcal{P}^n$ under the square root KL divergence. Then we have*

$$I(Z^J, J) \leq \inf_{\varepsilon>0}\left\{\varepsilon^2 + \log N_{\text{KL}}\right\}. \tag{9.12}$$

Proof: We first claim that

$$\frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n \| \mathbb{Q}_Z) \le \frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n \| \mathbb{Q}), \quad \text{for any } \mathbb{Q}.$$

That is, the average distribution minimizes the KL divergence. Indeed, we have

$$\begin{aligned}
\frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n \| \mathbb{Q}_Z) &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}_Z}\right] = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log\left(\frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\frac{d\mathbb{Q}}{d\mathbb{Q}_Z}\right)\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\right] + \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{Q}}{d\mathbb{Q}_Z}\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\right] - \mathbb{E}_{\mathbb{Q}_Z}\left[\log \frac{d\mathbb{Q}_Z}{d\mathbb{Q}}\right] \\
&\le \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\right]. \\
&= \frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n \| \mathbb{Q}).
\end{aligned}$$

Consequently,

$$I(Z^J, J) \le \frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n | \mathbb{Q}) \le \max_{j=1,\cdots,m} D(\mathbb{P}_j^n | \mathbb{Q})$$

for any $\mathbb{Q}$. Thus, it suffices to obtain a bound by a particular $\mathbb{Q}$.

To this end, let $\{\mathbb{Q}_1, \cdots, \mathbb{Q}_N\}$ be a ε-net of $\mathcal{P}^n$ under the square-root KL distance and define $\mathbb{Q} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{Q}_k$. By construction, there exists a $\mathbb{Q}_{k_j}$ such that $D(\mathbb{P}_j^n \| \mathbb{Q}_{k_j}) \le \varepsilon^2$. Then,

$$\begin{aligned}
D(\mathbb{P}_j^n \| \mathbb{Q}) &= \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\right] \\
&= \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{\frac{1}{N}\sum_{k=1}^{N} d\mathbb{Q}_k}\right] \\
&\le \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{\frac{1}{N} d\mathbb{Q}_{k_j}}\right] \\
&\le \varepsilon^2 + \log N.
\end{aligned}$$

Since this bound holds for any $\mathbb{P}_j^n$ and any $\varepsilon > 0$, the claim follows. ∎

Combing Lemma 9.18 with Theorem 9.14 yields the following proposition.

Proposition 9.19 *Under the setting for the construction of J and Z^J in Section 9.1, we have*

$$\mathfrak{M}_n(\Theta) \ge \Phi(\delta)\left(1 - \frac{\left(\varepsilon^2 + \log N_{\mathrm{KL}}\right) + \log 2}{\log m}\right). \tag{9.13}$$

Recall that m in (9.13) is the number of θ_j such that $\rho(\theta_i, \theta_j) \geq 2\delta$, so it relies on δ and when δ is prescribed we may choose $\{\theta_j\}_{j=1}^m$ to be global packing of Θ so that m is maximized. Note that there are two parameters ε and δ to be determined in (9.13). A typical way to choose them is

- choose ε such that $\varepsilon^2 \geq \log N_{\mathrm{KL}}$,
- choose largest possible δ such that $\log m \geq 4\varepsilon^2 + 2\log 2$,

so that $1 - \frac{\left(\varepsilon^2 + \log N_{\mathrm{KL}}\right) + \log 2}{\log m} \geq \frac{1}{2}$.

Example 9.20 *Consider the family of density functions*

$$\mathcal{F} = \{f : [0,1] \to [c_0, c_1] : \ \|f''\|_\infty \leq c_2 \text{ and } \int_0^1 f(x)dx = 1\},$$

where $0 < c_0 < 1 < c_1, c_2 > 1$ are constants. We study the minimax risk of estimating a density function from i.i.d data $X_1, \cdots, X_n \sim \mathbb{P}_f$ under the Hellinger distance

$$\rho(f, g) = H(f\|g) := H(\mathbb{P}_f \| \mathbb{P}_g) = \sqrt{\int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx}.$$

Note that

$$\begin{aligned} D(\mathbb{P}_f \| \mathbb{P}_g) &= \int_0^1 f(x) \log \frac{f(x)}{g(x)} dx \\ &\leq \int_0^1 f(x) \left(\frac{f(x)}{g(x)} - 1\right) dx \\ &= \int_0^1 \frac{(f(x) - g(x))^2}{g(x)} dx \\ &\leq \frac{1}{c_0} \int_0^1 (f(x) - g(x))^2 dx, \end{aligned}$$

and

$$\begin{aligned} \rho(f, g)^2 &= \int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx \\ &\leq \frac{1}{4c_0^2} \int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \left(\sqrt{f(x)} + \sqrt{g(x)}\right)^2 dx \\ &= \frac{1}{4c_0^2} \int_0^1 (f(x) - g(x))^2 \, dx. \end{aligned}$$

Therefore, both the squared KL divergence and $\rho(\cdot, \cdot)$ can be bounded by the L_2 distance. Consequently, in order to apply the global Fano method, we only need to understand the metric entropy in the L_2-norm. Since $f \in \mathcal{F}$ is second order smooth with $\|f''\|_\infty \leq c_2$, it can be shown that (See Example 5.11 of [1]),

$$\log N(\mathcal{F}, \|\cdot\|_2, \alpha) \asymp \left(\frac{1}{\alpha}\right)^{1/2}.$$

Since $D(\mathbb{P}_f^n \| \mathbb{P}_g^n) = nD(\mathbb{P}_f \| \mathbb{P}_g)$, $\sqrt{D(\mathbb{P}_f^n \| \mathbb{P}_g^n)} \le \varepsilon$ *if* $\sqrt{D(\mathbb{P}_f \| \mathbb{P}_g)} \le \varepsilon/\sqrt{n}$. *It follows that,*

$$\log N_{\mathrm{KL}} \asymp \left(\frac{\sqrt{n}}{\varepsilon}\right)^{1/2}$$

Thus, in order for $\varepsilon^2 \ge \log N_{\mathrm{KL}}$, *we may choose* $\varepsilon^2 \asymp (n)^{\frac{1}{5}}$. *Moreover, since* $\log m \asymp \left(\frac{1}{\delta}\right)^{1/2}$, *for the above choice of* ε, *we may choose* $\delta \asymp n^{-\frac{2}{5}}$ *such that* $\log m \ge 4\varepsilon^2 + 2\log 2$. *Finally, it can be concluded that*

$$\inf_{\widehat{f}} \sup_{f} H^2(\widehat{f} \| f) \gtrsim n^{-\frac{4}{5}}.$$

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapters 15.

[2] John Duchi, *Lecture notes for Statistics 311/Electrical Engineering 377: Information Theory and Statistics*, Chapter 7, 10.

[3] Francis Bach, *Learning Theory from First Principles*, Chapter 11.1.