BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 2: Hoeffding's Inequality and Its Applications

Lecturer: Long Zhao

2.1 Resources

- Vershynin (2018, Chapter 2&3): most of material.
- Bardenet et al. (2015): concentration for sampling without replacement.

2.2 Simplest Version

Let X_i be i.i.d. Bernoulli random variable with probability p and denote the corresponding sample mean as $\bar{p}_n$. Namely, $X_i \sim Bern(p)$ and $\bar{p}_n = \sum_{i=1}^n X_i/n$.

Theorem 2.1 (Hoeffding's Inequality for i.i.d. Bernoulli)

$$P(|\bar{p}_n - p| \geq \epsilon) \leq 2\exp(-2\epsilon^2 n)$$

holds for n *and* $\epsilon > 0$.

Unlike the central limit theorem (CLT), which requires $n \to \infty$, Hoeffding's inequality is non-asymptotic. This is the first non-trivial non-asymptotic result that I encountered.

2.2.1 Importance of Hoeffding's Inequality

Because of CLT, we know $\frac{\sqrt{n}(\bar{p}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1)$. At least for large enough n, it seems that we could use the tail behavior of $N(0,1)$ to approximate tail behavior of $|\bar{p}_n - p|$.

Proposition 2.2 *Let* $g \sim N(0,1)$*, then for all* $t > 0$*, we have*

$$P(g > t) \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}\exp(-t^2/2).$$

In particular, for $t \geq 1$*,*

$$P(g > t) \leq \frac{1}{\sqrt{2\pi}}\exp(-t^2/2).$$

Proof:

$$\begin{aligned} P(g > t) &= \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx \leq \frac{1}{\sqrt{2\pi}} \int_t^\infty (x/t)\exp(-x^2/2)dx \\ &= \frac{1}{t}\frac{1}{\sqrt{2\pi}} \int_t^\infty \exp(-x^2/2)d(x^2/2) = \frac{1}{t}\frac{1}{\sqrt{2\pi}} \exp(-t^2/2). \end{aligned}$$

Obviously, if $t \geq 1$, we have $1/t \leq 1$. ■

Thus, we have

$$\begin{aligned} P(|\bar{p}_n - p| \geq \epsilon) &= P\left(\left|\frac{\sqrt{n}(\bar{p}_n - p)}{\sqrt{p(1-p)}}\right| \geq \frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}\right) \approx P\left(|g| \geq \frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}\right) \\ &\leq 2\frac{1}{\sqrt{2\pi}} \exp(-\frac{\epsilon^2 n}{2p(1-p)}), \end{aligned}$$

which is exponential decay with same order $\epsilon^2 n$. It seems promising, as long as the approximation error of CLT is small compared to the exponential decay term. We actually know this approximation error from the Berry-Esseen Theorem introduced in Lecture 1.

Theorem 2.3 (Berry-Esseen) *Assume Y_i are i.i.d. with finite third moments, $\rho < \infty$. Then for all n,*

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}},$$

where $F_n(x)$ and $\Phi(x)$ are the c.d.f. of $\sqrt{n}\bar{Y}_n$ ($\bar{Y}_n = \sum_{i=1}^n Y_i/n$) and $N(0,1)$, respectively.

Based on the Berry-Esseen theorem, we know the approximation error of CLT could be as large as order $1/\sqrt{n}$. That is to say, the approximation error dominates the exponential decay term of $N(0,1)$. Hoeffding's inequality tells us that $\bar{p}_n$ shares similar tail behavior of standard normal distribution.

2.3 Proof of Simplest Version

We will leverage the following Lemma to prove Hoeffding's inequality.

Lemma 2.4 (Hoeffding's Lemma) *If $E(X) = \mu$, and $a \leq X \leq b$, then*

$$E(\exp(\lambda(X - \mu))) \leq \exp(\lambda^2(b-a)^2/8).$$

To prove Hoefdding's Lemma, one should utilize the convexity of $\exp(x)$ and then use numeric inequality to obtain the bound. For details, please see here.

Now, we begin the proof of Theorem 2.1.

Proof: It is enough to prove $P(\bar{p}_n - p \geq \epsilon) \leq \exp(-2\epsilon^2 n)$. For any $\lambda > 0$, we have

$$\begin{aligned} P(\bar{p}_n - p \geq \epsilon) &= P(\sum_{i=1}^{n}(X_i - p) \geq n\epsilon) \\ &\leq \exp(-\lambda n\epsilon) E\left(\exp(\lambda(\sum_{i=1}^{n}(X_i - p))\right) \\ &= \exp(-\lambda n\epsilon)\left(E(\exp(\lambda(X_1 - p))\right)^n \quad \text{(i.i.d.)} \\ &\leq \exp(-\lambda n\epsilon)\exp(\lambda^2 n/8) \quad \text{(Hoeffding's Lemma)} \end{aligned}$$

Since we have the freedom of choosing λ, we could choose $\lambda^\star = 4\epsilon$ to minimize $-\lambda\epsilon + \lambda^2/8$ and obtain $-2\epsilon^2$.

■

Remark 2.5 *To have a meaningful bound, we should have $\epsilon^2 n \gg 1$. That is to say, $\epsilon \gg \sqrt{1/n}$ or $\epsilon n \gg \sqrt{n}$. If we want to learn what will happen on a smaller scale, like $\epsilon n = 1$, we need to refer to the anti-concentration law, see Anti-concentration inequalities for more details. Anti-concentration is essential for Chernozhukov et al. (2017) which I have no understanding.*

Since we only use X_i bounded in the proof, it is easy to have the following generalized form.

Theorem 2.6 (Hoeffding's Inequality for Independent Bounded Random Variables) *If X_is are independent and $X_i \in [a_i, b_i]$ for $\forall i = 1, \ldots n$, we have*

$$P\left(\sum_{i=1}^{n}(X_i - EX_i) \geq \epsilon n\right) \leq \exp\left(-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

hold for n and $\epsilon > 0$.

Think about why we have n^2 instead of n here?

Remark 2.7 *To have the event $\sum_{i=1}^{n}(X_i - EX_i) \geq \epsilon n$ useful, we require $\sum_{i=1}^{n} EX_i \gg \epsilon n$. However, this might not always be the case. For example, $X_i \sim Bern(p_i)$ and $\sum_{i=1}^{n} p_i \to \lambda \ll \epsilon n$. In this case, one needs a new inequality called Chernoff's inequality. Please see Vershynin (2018, Chapter 2.3) for more details.*

2.4 Analysis of The Proof

2.4.1 Light tail

In the proof, we have $E(\exp(\lambda(X_i - EX_i)) = \exp(C\lambda^2)$ for all $\lambda \in \mathbb{R}$. Next, I want to answer two questions.

1. Do we need λ^2 in $E(\exp(\lambda(X_i - EX_i))$?
2. Do we need $E(\exp(\lambda(X_i - EX_i)) = \exp(C\lambda^2)$ for all $\lambda \in \mathbb{R}$?

To answer the first question, we need to recall the proof. For a fixed $\epsilon > 0$, we use Markov inequality to have

$$P\left(\sum_{i=1}^{n}(X_i - EX_i) \geq \epsilon n\right) \leq \min_{\lambda \in I} \exp(-\lambda n \epsilon) E\left(\exp(\sum_{i=1}^{n} \lambda(X_i - EX_i))\right),$$

Notice that the first term, $\exp(-\lambda n \epsilon)$ provides exponential decay when $\lambda > 0$. If the first term dominates the second term for some $\lambda > 0$, we have concentration. However, this is not possible if the second term is $\exp(C\lambda n)$ since one could choose ϵ small enough such that the first term never dominates the second. Meanwhile, as long as the second term is $\exp(C\lambda^q n)$ where $q > 1$, this is possible by smartly choosing λ. That is to say, higher order is essential but not necessarily quadratic.

The above analysis also answers the second question that we do not need $\lambda \in \mathbb{R}$. As long as I contains some positive parts, we still have exponential decay. It is just slower than Hoeffding's inequality because we might not be able to achieve the minimum of the quadratic function. In fact, this observation naturally leads to the Bernstein's inequality which we will cover later.

2.4.2 Independence and Linearity

Leveraging independence and linearity, we have $E(\exp(\sum_i X_i)) = \prod E(\exp(X_i))$ which is critical for the proof. However, it is not hopeless to move a little from independence. For simplicity, I will focus on bounding $E(\exp(X_1 + X_2 + X_3))$.

$$\begin{aligned} E(\exp(\lambda(X_1 + X_2 + X_3))) &= E\big(E(\exp(\lambda(X_1 + X_2 + X_3))|X_1, X_2)\big) \quad \text{Law of total expectation} \\ &= E\big(\exp(\lambda(X_1 + X_2))E(\exp(\lambda X_3)|X_1, X_2)\big) \end{aligned}$$

Say $0 \leq X_i \leq 1$ $(i = 1, 2, 3)$, then $0 \leq X_3|X_1, X_2 \leq 1$. Notice that Hoeffding's Lemma also holds for conditional probability $P(|X_1, X_2)$, we have

$$E(\exp(\lambda X_3)|X_1, X_2) \leq \exp(\lambda E(X_3|X_1, X_2)) \exp(\lambda^2/8)$$

Thus, we have

$$E(\exp(\lambda(X_1 + X_2 + X_3))) \leq E\big(\exp(\lambda(X_1 + X_2 + E(X_3|X_1, X_2)))\big) \exp(\lambda^2/8)$$

The simplest case is $E(X_3|X_1, X_2) = 0$. In this way, we eliminate one term in the summation with the same bound $\exp(\lambda^2/8)$ as the independent case. If we also have $E(X_2|X_1) = 0$, by conditional on X_1, we could obtain the same Hoeffding's inequality as the independent case. If we denote $S_k = \sum_{i=1}^{k} X_i$, then

$$E(X_k|X_1, \ldots X_{k-1}) = 0 \Rightarrow E(S_k|X_1, \ldots X_{k-1}) = S_{k-1},$$

namely S_k is a martingale if $E|S_k| < \infty$. Thus, it is natural to have the following concentration inequality for martingale.

Theorem 2.8 (Azuma-Hoeffding Inequality) *Let $\{Y_0, Y_1, \cdots\}$ be a martingale and $|Y_k - Y_{k-1}| \leq c_k$ almost surely. Then we have*

$$P(|Y_n - Y_0| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^n c_i^2}\right)$$

Next, we will use Azuma-Hoeffding Inequality to obtain concentration for the uniform sample without replacement (not independent). Abu-Mostafa et al. (2012) uses this concentration in its proof of the VC dimension.

Theorem 2.9 *Let $\mathcal{A} = \{a_1, \ldots, a_{2N}\}$ be a set of values with $a_i \in [0,1]$, and let $\mu = \frac{1}{2N}\sum_{i=1}^{2N} a_i$ be their mean. Let $\mathcal{D} = X_1, \ldots X_N$ be a sample of size N, sampled from $\mathcal{A}$ uniformly* **without** *replacement. Then*

$$P\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \geq \epsilon\right) \leq 2\exp(-2\epsilon^2 N).$$

Proof: Because it requires careful analysis to obtain $2\epsilon^2 N$, we only prove a weaker version with a different constant (< 2). Since sampling without replacement is not independent, our best chance is to construct a martingale and utilize Azuma-Hoeffding Inequality. Notice that

$$\begin{aligned} E\left(\sum_{i=1}^{k+1}(X_i - \mu)|X_1, \ldots, X_k\right) &= \sum_{i=1}^k (X_i - \mu) + \frac{2N\mu - \sum_{i=1}^k X_i}{2N-k} - \mu \\ &= \sum_{i=1}^k (X_i - \mu) - \frac{1}{2N-k}\sum_{i=1}^k (X_i - \mu) \\ &= \frac{2N-k-1}{2N-k}\sum_{i=1}^k (X_i - \mu). \end{aligned}$$

Thus, if we denote $S_k = \frac{1}{2N-k}\sum_{i=1}^k (X_i - \mu)$, then we have

$$E\left(S_{k+1}|X_1, \ldots, X_k\right) = E\left(\frac{1}{2N-k-1}\sum_{i=1}^{k+1}(X_i - \mu)|X_1, \ldots, X_k\right) = S_k.$$

Moreover, we could bound

$$|S_k - S_{k-1}| = \left|\frac{X_k - \mu}{2N-k} + \frac{\sum_{i=1}^{k-1}(X_i - \mu)}{(2N-k)(2N-k+1)}\right| \leq \frac{2}{N} = c_k \quad \text{if } k \leq N.$$

This inequality holds because $|X_i - \mu| \leq 1$ and $2N - k \geq N$. Thus, we have $\sum_{i=1}^N c_i^2 = 4/N$. Using

Azuma-Hoeffding inequality to have

$$P(|S_N - S_0| \geq \epsilon) \leq 2\exp(-\epsilon^2/2(4/N)) = \exp(-\epsilon^2 N/8),$$

where $S_N = \frac{1}{N}\sum_{i=1}^N (X_i - \mu)$ and $S_0 = 0$. ■

Remark 2.10 *Because of Doob's martingale inequality, it is possible to have concentration of* $P(\max_{1\leq i\leq k} Y_i \geq \epsilon)$ *where* Y_i *is a martingale. Bardenet et al. (2015) gives an example.*

2.5 Unbounded Independent Random Variables

In the proof of Hoeffding's inequality, we use $X_i \in [a_i, b_i]$ to generate a bound of $E(\exp(\lambda(X_i - \mu_i))$ in the form of $\exp(C\lambda^2)$ via Hoeffding's lemma. In this sense, as long as we have $\exp(C\lambda^2)$ as the upper bound, we do not require boundedness of random variables. In fact, if $X \sim N(\mu, \sigma^2)$, we have

$$E\exp(\lambda(X - \mu)) = \exp(\lambda^2\sigma^2/2)$$

hold for all λ. This naturally leads to our pursue of a larger class of random variables that could be unbounded but still have

$$E\exp(\lambda(X - \mu)) \leq \exp(C^2\lambda^2) \quad \forall\lambda \in \mathbb{R}.$$

We call them sub-Gaussian distribution which will be defined rigorously in the next lecture. Since the sum of sub-Gaussian distributions will concentrate around its expected value, they play a crucial role in high-dimensional probability and statistics.

2.6 Finite Hypothesis

In the Lecture 02 of Learning from Data Youtube Videos, they assume there are m hypotheses. Denote A_i the event that the difference between in-sample and out-of-sample performance of the ith hypothesis is larger than ϵ. By Hoeffding's inequality, we know

$$P(A_i) \leq 2\exp(-2\epsilon^2 n), \forall i = 1, \ldots, m.$$

Using union bound, we have

$$P(\cup A_i) \leq \sum_{i=1}^m P(A_i) = m\exp(-2\epsilon^2 n).$$

In some sense, m is the cost one pays for not knowing exactly which event happens. Thus, it has the fancy name of entropy cost. As long as m is polynomial in n, it will be dominated by the exponential term when n is large enough. Here is an example that m is exponential in n. Given n data points (1 or 0), there are in total 2^n possibilities. If the hypothesis set (potentially ∞ models) happens to contain models that could fit all 2^n cases, we could pick one for each case. Then the bound we have will be $2^n \exp(-2\epsilon^2 n)$ which is useless.

However, in most cases, we have infinite hypotheses, and the union bound no longer works. Then the job becomes how to use finite hypotheses to obtain a cover of infinite ones and then utilize union bound on the finite events. In the future class, we will first introduce a straight forward case of covering (operator norm of a random matrix) and then a complicated case (VC dimension).

2.7 Application of Hoeffdings' Inequality

Example 2.11 (Boosting Randomized Algorithms, Vershynin (2018) Exercise 2.2.8) *Suppose we have an algorithm that makes a decision at random and returns the correct answer with probability $1/2+\delta$ with some $\delta > 0$, which is just better than a random guess. To improve performance, we run the algorithm N times and take the majority vote. Show that, for any $\epsilon \in (0,1)$, the answer is correct with probability at least $1-\epsilon$, as long as*

$$N \geq \frac{1}{2\delta^2}\ln(\frac{1}{\epsilon})$$

Use X_i to denote the indicator function that the ith algorithm get the right answer. Then we have $EX_i = P(X_i = 1) = 1/2+\delta$ and X_is are independent. Now, the majority vote is wrong is equivalent to $\bar{X}_n - 1/2 \leq 0$, where $\bar{X}_n = \sum_{i=1}^n X_i/n$. Now, we are in business.

Proof: Using Hoeffding's inequality, we have

$$P\left(\bar{X}_n - (1/2+\delta) \leq -\delta\right) \leq \exp(-2n\delta^2).$$

Thus, for any $\epsilon > 0$, if $n \geq \ln(1/\epsilon)/2\delta^2$, we have $P(\bar{X}_n - 1/2 \leq 0) \leq \epsilon$. ■

Example 2.12 (Robust Estimation of Mean, Vershynin (2018) Exercise 2.2.9) *We want to estimate the mean μ of a random variable X from a sample $X_1, \ldots, X_n$ drawn independently from the distribution of X. We also know $Var(X) = \sigma^2 < \infty$. We want an $\epsilon-$accurate estimate, i.e. one that falls in the interval $(\mu-\epsilon, \mu+\epsilon)$ with probability $1-\delta$. How many sample do we need?*

It is tempting to use Hoeffding's inequality. Unfortunately, X might not be a bounded random variable. Since we have the variance of X, it is tempting to use Chebyshev inequality to have

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

If we choose $n = \frac{1}{\delta}\frac{\sigma^2}{\epsilon^2}$, we could bound the above probability by δ. However, n could be extremely large when δ is small.

Another interpretation of not using Hoeffding's inequality is that X might have heavy tail. If so, then sample mean might not be a good idea since it is sensitive to outliers. Thus, it is tempting to use median instead of mean as our estimation. Introduce $Y_i = 1_{X_i \geq \mu+\epsilon}$. Because X_i are i.i.d., we know Y_i are i.i.d. Bernoulli (bounded!) with probability $p(\epsilon) = P(X_i \geq \mu + \epsilon)$. Then

$$M_n \geq \mu + \epsilon \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \geq \mu+\epsilon} - \frac{1}{2} \geq 0 \Leftrightarrow \bar{Y}_n - \frac{1}{2} \geq 0,$$

where M_n is the median of n samples. Now, we could adopt the techniques used in the previous exercise to have

$$P(M_n \geq \mu + \epsilon) \leq \exp\left(-2n(1/2 - p(\epsilon))^2\right).$$

We could have exponential decay if $p(\epsilon) < 1/2$! Unfortunately, we do not have guaranteed. Sad. The good news is that, if we could construct some estimators that have $p(\epsilon) < 1/2$, we could take a median of them and enjoy the benefit of exponential decay. Because of Chebyshev inequality, we know the sample mean could achieve this goal. Thus, we will replace X_i with the sample mean $\bar{X}_{n,i}$ and then take a median of $\bar{X}_{n,i}$.

Proof: Because of Chebyshev inequality, if we choose $n \geq 4\sigma^2/\epsilon^2$, we have $p(\epsilon) = P(\bar{X}_{n,i} \geq \mu + \epsilon) \leq 1/4$. Here I abuse the notations a little to highlight the idea of the proof. Using Hoeffding's inequality, we have

$$P(M_N \geq \mu + \epsilon) \leq \exp\left(-2N(1/2 - p(\epsilon))^2\right) \leq \exp(-N/2),$$

where M_N is the median of N pieces of $\bar{X}_{n,i}$. We could choose $N = O(\ln(1/\delta))$ to make the RHS smaller than $1/2\delta$. In this way, we use $nN = O(\ln(1/\delta)\sigma^2/\epsilon^2)$ samples to obtain an estimator that belongs to $(\mu-\epsilon, \mu+\epsilon)$ with probability $1-\delta$. It is much smaller than the number of sample used based on Chebyshev inequality. ■

Example 2.13 (Probability Bound in Bertsimas and Sim (2004), Budget Uncertainty Set) . *If η_{ij}, $j \in J_i$ are independent and symmetrically distributed random variable in $[-1, 1]$, then*

$$P(\sum_{j\in J_i} \gamma_{ij}\eta_{ij} \geq \Gamma_i) \leq \exp\left(-\frac{\Gamma_i^2}{2|J_i|}\right),$$

where $|J_i|$ is the cardinal number of set J_i and $0 < \gamma_{ij} \leq 1$.

In a linear programing problem, the ith constraint $\sum_j \tilde{a}_{ij}x_j \leq b_i$ contains random coefficients $\tilde{a}_{ij} \in J_i$. Specifically,

$$\tilde{a}_{ij} = a_{ij} + \eta_{ij}\hat{a}_{ij},$$

where $\eta_{ij} \in [-1, 1]$ and $E\eta_{ij} = 0$. Budget uncertainty set requires the number of changed coefficients is not bigger than Γ_i, where Γ_i controls the level of robustness. When η_{ij}s are independent, it is possible to choose $\Gamma_i \ll |J_i|$ such that $\sum_j \tilde{a}_{ij} x_j > b_i$ only happens for a low probability. This violation probability is upper bounded by $P(\sum_{j \in J_i} \gamma_{ij}\eta_{ij} \geq \Gamma_i)$ which leads to the example above.

Proof: Hoeffding's inequality is perfect in this case because η_{ij} are independent and γ_{ij} is bounded. Utilizing Theorem 2.6, we have

$$P\left(\sum_{j \in J_i} (\gamma_{ij}\eta_{ij} - 0) \geq \Gamma_i\right) \leq \exp\left(-2\Gamma_i^2 / \sum_{j \in J_i} |2\gamma_{ij}|^2\right) \leq \exp(-\Gamma_i^2/2|J_i|)$$

■

Example 2.14 (High Dimensional Uniform Distribution) *Let $X_i \sim U(-1,1)$ and $X_i s$ are i.i.d. Let's consider a high-dimensional vector $\vec{X} = (X_1, \dots, X_p)$ where p is the dimension. Show that $\|\vec{X}\|_q^q$ concentrates.*

From $X_i \sim U(-1,1)$, we know $|X_i| \sim U(0,1)$ which is bounded. Moreover, $|X_i|^q$ is also bounded for $\forall q \geq 0$ and $E|X_i|^q = 1/(q+1)$. Thus, we could use Hoeffding's inequality to show that $\|\vec{X}\|_q^q = \sum_{i=1}^p |X_i|^q$ concentrate around the $1/(q+1)$.

Proof: By Hoeffding's inequality, we have

$$P\left(\left|\sum_{i=1}^p |X_i|^q - \frac{p}{q+1}\right| \geq \epsilon p\right) \leq 2\exp(-2\epsilon^2 p).$$

That is to say,

$$(1/(q+1) - \epsilon)p \leq \|\vec{X}\|_q^q \leq (1/(q+1) + \epsilon)p$$

is true with probability at least $1 - 2\exp(-2\epsilon^2 p)$. ■

Since $\|\vec{X}\|_2^2 = r$ is a high-dimensional ball which is easier to picture, we will choose $q = 2$ to understand the above result. If we take $p = 10,000$ and $\epsilon = 0.02$, then

$$P(0.31p \leq \|\vec{X}\|_2^2 \leq 0.36p) \geq 99.9\%.$$

In other words, $\vec{X}$ has some predictable behavior because of high dimensionality. Meanwhile, if $p = 2$, there is no such pattern exists. In fact, the author of Wainwright (2019) mentioned that the blessing of high dimensionality is concentration of measure (High-Dimensional Statistics I, Youtube Video).

We could normalize $\vec{X}$ to obtain $\vec{X}/\|\vec{X}\|_2$, which is on the high dimensional sphere. Is it a uniformly distributed on the sphere? Why? (Hint: think about 2 dimension case).

Example 2.15 (High Dimensional $N(0, I_p)$) *Let $\vec{X} \sim N(0, I_p)$, then we know X_is are i.i.d. $N(0,1)$. We also have $\|\vec{X}\|_2$ concentrates around $\sqrt{p}$.*

It is tempting to utilize $E(\exp(\lambda X_i)) = \exp(\lambda^2)$ to obtain the concentration. Unfortunately, $\|\vec{X}\|_2$ is about X_i^2 which have much heavier tail than Normal distribution. Thus, we need to have new concentration inequality (next class) to prove this result.

Using the technique of changing variables, we could prove that $\vec{X}/\|\vec{X}\|_2$ follows a uniform sphere distribution. (Are different coordinates independent?) Coupled with the concentration of $\|\vec{X}\|_2$, we know that $\vec{X}_i$ shall behave like Figure 2 (Right). This is the first high dimensional result that blow my mind. Hope you also find it amazing.

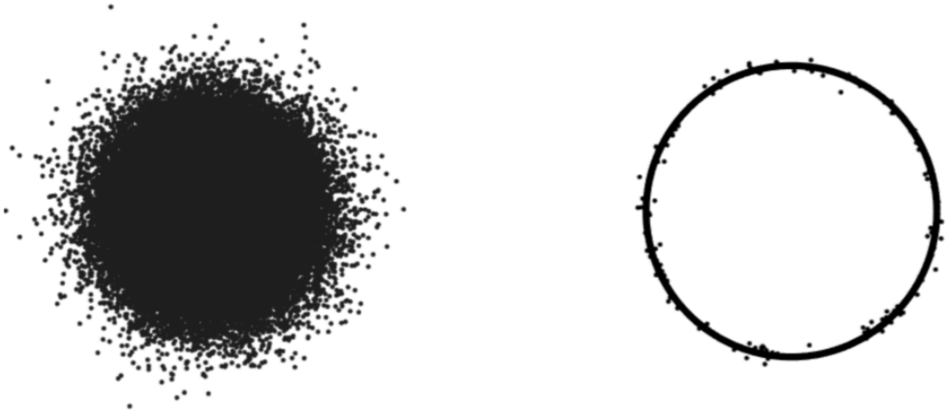


Figure 2.1: 2D $N(0, I_2)$ (Left) v.s. High Dimensional $N(0, I_p)$ (Right)

References

Abu-Mostafa YS, Magdon-Ismail M, Lin HT (2012) *Learning from data*, volume 4 (AMLBook New York, NY, USA:).

Bardenet R, Maillard OA, et al. (2015) Concentration inequalities for sampling without replacement. *Bernoulli* 21(3):1361–1385.

Bertsimas D, Sim M (2004) The price of robustness. *Operations research* 52(1):35–53.

Chernozhukov V, Chetverikov D, Kato K, et al. (2017) Central limit theorems and bootstrap in high dimensions. *The Annals of Probability* 45(4):2309–2352.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

& 4

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 3 & 4: Sub-Gaussian & Sub-Exponential Distributions

Lecturer: Long Zhao, longzhao@nus.edu.sg

3.1 Resources

- Vershynin (2018, Chapter 2&3): most of material.

3.2 Target

Let $X \sim N(0, I_p)$. We want to prove that $\|X\|_2$ concentrates around $\sqrt{p}$ when p is large. At first sight, it might seem unsurprising since $E\|X\|_2^2 = E\sum_{i=1}^p X_i^2 = \sum_{i=1}^p EX_i^2 = p$. Then $\|X\|_2$ is about the level of $\sqrt{p}$. However, based on Jensen's inequality, we only have $(E\|X\|_2)^2 \le E\|X\|_2^2 = p$. In fact, we know $\|X\|_2$ follows $\chi(p)$ where p is the degree of freedom. Based on the approximation of $\chi(p)$, we have

$$E\|X\|_2 = \sqrt{p-1} \times \left(1 - \frac{1}{4p} + O(\frac{1}{p^2})\right) \le \sqrt{p}.$$

Thus, $\|X\|_2$ is about $\sqrt{p}$ because $\|X\|_2^2$ concentrates around p which behaves like deterministic values. In other words, it behaves like law of large numbers:

$$\frac{1}{p}\sum_{i=1}^p X_i^2 \to 1 \Rightarrow \frac{1}{\sqrt{p}}\sqrt{\sum_{i=1}^p X_i^2} \to 1.$$

To establish the concentration of $\|X\|_2^2$, we need to introduce the sub-exponential distribution, which has heavier tails than the normal distribution. Here is the roadmap.

1. Definition of sub-Gaussian and sub-exponential.
2. Properties of sub-Gaussian (same bound as Hoeffding's inequality).
3. Properties of sub-exponential.
4. Connections between them.
5. Bernstein's inequality. (Concentration inequality for sub-Gaussian)

6. Concentration of $\|X\|_2$.

7. Almost orthogonal vectors.

3.3 (Only) The Tail Decay Matters

First of all, we need to define what is the tail of a distribution. Say X is a random variable, $P(|X| > t)$ for **large** t is the tail behavior. To see why it is the case, it is beneficial to analyze the moment generating function (MGF) of X

$$E\exp(\lambda X) = \int_{-\infty}^{\infty} \exp(\lambda x) dF(x) = \int_{-t}^{t} \exp(\lambda x) dF(x) + \int_{-\infty}^{-t} \exp(\lambda x) dF(x) + \int_{t}^{\infty} \exp(\lambda x) dF(x),$$

where $F(x)$ is the c.d.f. of X. For simplicity, we only analyze $\lambda > 0$. In this case, the first term is bounded by $\exp(\lambda t)$ and the second term is bounded by $\exp(-\lambda t)$. This means that they won't make the MGF into $+\infty$, but the third term could. For example, if X is a Pareto distribution whose $F(x)$ is

$$F(x) = \begin{cases} 1 - x^{-\alpha} & x \geq 1 \\ 0 & x < 1, \end{cases}$$

where $\alpha > 0$, we have

$$\int_{t}^{\infty} \exp(\lambda x) dF(x) = \int_{t}^{\infty} \alpha x^{-\alpha-1} \exp(\lambda x) dx = \infty, \quad \forall \lambda > 0.$$

More generally, if

$$\lim_{t} \exp(\lambda t) P(X > t) = \infty \quad \forall \lambda > 0,$$

we have $MGF(\lambda) = \infty$ for all $\lambda > 0$. Interestingly, this naturally leads to the definition of heavy-tail distribution.

Definition 3.1 *X is said to have a heavy (right) tail if its $MGF(\lambda) = \infty$ for all $\lambda > 0$.*

Based on our analysis in the last lecture, on the one hand, it is hopeless to have exponential concentration for the heavy-tail distribution. On the other hand, Hoeffding's inequality tells us that there is concentration for the distributions without tails (bounded variable). The purpose of this lecture is to build the bridge between these two extremes: introducing two families of distributions that have some tails but still have concentration.

Because of the importance of tail behavior, I will define sub-Gaussian and sub-exponential distributions in terms of $P(|X| > t)$:

- Sub-Gaussian: $P(|X| > t) \leq 2\exp(-c_1 t^2)$ for all $t \geq 0$.

- Sub-Exponential: $P(|X| > t) \leq 2\exp(-c_2 t)$ for all $t \geq 0$.

Try to prove the following properties

- sub-Gaussian distribution is also sub-exponential.
- If X is sub-Gaussian, then X^2 is sub-exponential.

3.4 Sub-Gaussian Distribution

3.4.1 Behavior of $N(0,1)$

Since the name contains Gaussian, we would like to investigate the tail behavior of $N(0,1)$ first.

Proposition 3.2 *Let $g \sim N(0,1)$, then for all $t > 0$, we have*

$$P(|g| > t) \leq 2\exp(-t^2/K^2)$$

Proof: From Proposition 2.2, we know that for $t \geq 1$

$$P(|g| > t) \leq \frac{2}{\sqrt{2\pi}}\exp(-t^2/2),$$

which is the form we want. As in our discussion about the tails, we care very little about $t \leq 1$. In fact, we could obtain the following trivial bound for $t \leq 1$

$$P(|g| > t) \leq 1 < 2\exp(-1/2) \leq 2\exp(-t^2/2).$$

More generally, for bounded t, we could choose K^2 so large that $2\exp(-t^2/K^2) > 1$ which leads to a trivial bound (probability is smaller than 1). Thus, we know that 2 in $\mathbf{2}\exp(-t^2/K^2)$ is not essential; as long as it is larger than 1, this trivial bound pass through. ■

Based on this trick, could you prove that sub-Gaussians is also sub-exponential distribution?

It is easy to show that for $N(0,\sigma^2)$, we also have tail bounded in the form of $2\exp(-t^2/K^2)$. It is less trivial for $N(\mu,\sigma^2)$, but we will not prove it rigorously here. Here is the intuition. If we only focus on $t \gg \mu$, then the existence of μ barely makes a difference. Therefore, we must have a similar bound.

Remark 3.3 *We pay very little attention to the specific form of constants. This is usually the case when dealing with concentration and high-dimensional probability/statistics.*

3.4.2 Sub-Gaussian Properties

The following proposition shows different properties of sub-Gaussian distribution. We will try to derive them from the tail behavior (property 1). Our ultimate target is property 5 because it plays a critical role in the proof of concentration.

Proposition 3.4 (Sub-Gausssian Properties) *Let X be a random variable. The following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*

$$P(|X| \geq t) \leq 2\exp(-t^2/K_1^2) \quad \forall t \geq 0$$

2. *The moments of X satisfy*

$$\|X\|_p = (E|X|^p)^{1/p} \leq K_2\sqrt{p} \quad \forall p \geq 1$$

3. *The MGF of X^2 satisfies*

$$E\exp(\lambda^2 X^2) \leq \exp(K_3^2\lambda^2) \quad \forall|\lambda| \leq \frac{1}{K_3}$$

4. *The MGF of X^2 is bounded at some point, namely*

$$E\exp(X^2/K_4^2) \leq 2.$$

Moreover, if $EX = 0$ then properties 1-4 are also equivalent to the following one.

5. *The MGF of X satisfies*

$$E\exp(\lambda X) \leq \exp(K_5^2\lambda^2) \quad \forall\lambda \in \mathbb{R}.$$

It is worth to go through all the calculations if $X \sim N(0,1)$.

$$\begin{aligned} E|X|^p &= \frac{2}{\sqrt{2\pi}}\int_0^\infty x^p \exp(-x^2/2)dx \\ &= \frac{2^{(p+1)/2}}{\sqrt{2\pi}}\int_0^\infty y^{(p-1)/2}\exp(-y)dy \quad (y = x^2/2) \\ &= \frac{2^{p/2}}{\sqrt{\pi}}\Gamma(\frac{p+1}{2}) \end{aligned}$$

Using numeric inequality, $\Gamma(x) \leq 3x^x$, we could get $\|X\|_p \leq K\sqrt{p}$.

$$
\begin{aligned}
E\exp(\lambda^2 X^2) &= \frac{2}{\sqrt{2\pi}}\int_0^\infty \exp(\lambda^2 x^2)\exp(-x^2/2)dx \\
&= \frac{2}{\sqrt{2\pi}}\int_0^\infty \exp(-(1/2-\lambda^2)x^2)dx
\end{aligned}
$$

Clearly, it blows up when $\lambda^2 \geq 1/2$. When $\lambda^2 < 1/2$, introduce $y = \sqrt{1-2\lambda^2}x$ ($y^2/2 = (1/2-\lambda^2)x^2$), then

$$
E\exp(\lambda^2 X^2) = \frac{2}{\sqrt{2\pi}}\int_0^\infty \exp(-y^2/2)dy \times (1-2\lambda^2)^{-1/2} = (1-2\lambda^2)^{-1/2}
$$

If we use a numeric inequality, $1/(1-x) \leq \exp(2x),\ \forall x \in [0,1/2]$, we have

$$
(1-2\lambda^2)^{-1/2} \leq \exp(2\lambda^2) \quad \forall 2\lambda^2 \leq 1/2.
$$

Thus, we have $E\exp(\lambda^2 X^2) \leq \exp(2\lambda^2) \leq \exp(2^2\lambda^2)$ for all $|\lambda| \leq 1/2$.

Proof:

1 ⇒ 2. Without loss of generality, take $K_1 = 1$. Then

$$
\begin{aligned}
E|X|^p &= E\int_0^\infty 1_{|X|^p > x}dx = \int_0^\infty P(|X|^p > x)dx \quad \text{(Fubini Theorem)} \\
&= \int_0^\infty P(|X|^p > t^p)dt^p \quad (x = t^p) \\
&= \int_0^\infty pt^{p-1}P(|X| > t)dt \leq \int_0^\infty 2pt^{p-1}\exp(-t^2)dt = \int_0^\infty p(t^2)^{p/2-1}\exp(-t^2)dt^2 \\
&= p\Gamma(p/2) \leq 3p(p/2)^{p/2} \quad (\Gamma(x) \leq 3x^x,\ \forall x \geq 1/2)
\end{aligned}
$$

Take power of $1/p$ to both sides to have

$$
\|X\|_p \leq (3p)^{1/p}/\sqrt{2}\sqrt{p} \leq 3\sqrt{p} \quad (3p/\sqrt{2} \leq 3^p, \forall p \geq 1).
$$

We also show that $K_2 = 3$.

1 ⇒ 3. Using the same technique above

$$\begin{aligned}
E\exp(\lambda^2X^2) &= \int_0^\infty P(\exp(\lambda^2X^2) > x)dx \\
&= \int_0^\infty P(\exp(\lambda^2X^2) > \exp(\lambda^2t^2))d\exp(\lambda^2t^2) \quad (x = \exp(\lambda^2t^2)) \\
&= \int_0^\infty P(|X| > t)\exp(\lambda^2t^2)\lambda^2dt^2 \\
&\le \int_0^\infty 2\exp(-y)\exp(\lambda^2y)\lambda^2dy \quad (y = t^2) \\
&= \int_0^\infty 2\lambda^2\exp(-(1-\lambda^2)y)dy = \frac{2\lambda^2}{1-\lambda^2} \quad (\text{when } \lambda^2 < 1) \\
&\le \exp(2\lambda^2) \quad (\text{when } \lambda^2 \le 1/2).
\end{aligned}$$

The last inequality holds because $1/(1-x) \le \exp(2x)$ holds for $x \le 1/2$. We also show that $K_3 = \sqrt{2}$.

1 ⇒ 5. I do not know how to do this. Please let me know if you figure it out.

3 ⇒ 5. Without loss of generality, assume $K_3 = 1$. If we use numeric inequality, $\exp(x) \le x + \exp(x^2)$, when $\lambda < 1$, we have

$$E\exp(\lambda X) \le E\lambda X + E\exp(\lambda^2X^2) \le \exp(\lambda^2).$$

When $\lambda \ge 1$, we know $\lambda X \le \lambda^2/2 + X^2/2$. Thus,

$$E\exp(\lambda X) \le \exp(\lambda^2/2)E\exp(X^2/2) \le \exp(\lambda^2/2)\exp(1/2) \le \exp(\lambda^2)$$

■

Noticing that $E\exp(\lambda X) = \exp(K_5^2\lambda^2)$ plays a key role in Hoeffding's inequality. We are expecting an almost identical concentration form. Before taking that adventure, let me first introduce a norm of sub-Gaussian distribution, which comes in handy soon.

3.4.3 Sub-Gaussian Norm

Definition 3.5 (Sub-Gaussian Norm) *The sub-Gaussian norm of X, denoted as $\|X\|_{\psi_2}$, is defined as*

$$\|X\|_{\psi_2} = \inf\{t > 0 : E\exp(X^2/t^2) \le 2\}$$

This definition of norm (why bounded by 2) is very weird to me. To understand why it is defined this way, we need to leverage the power of Orlicz spaces. A function $\psi : [0,\infty) \to [0,\infty)$ is called Orlicz function if ψ

is convex, increasing, and satisfies

$$\psi(0) = 0, \ \psi(x) \to \infty, \text{ as } x \to \infty.$$

Example 3.6 (Orlicz Functions)

- *$\psi(x) = x^p$, $p \geq 1$.*
- *$\psi(x) = \exp(x^2) - 1$.*
- *$\psi(x) = \exp(x) - 1$.*
- *$\psi(x) = \exp(x^q) - 1$, $q \geq 1$. If f and g are convex and g is non-decreasing. Then $g(f(x))$ is convex. Take $g(x) = \exp(x)$ and $f(x) = x^q$.*

For a given Orlicz function ψ, the Orlicz norm (show that it is indeed a norm) of a random variable X is defined as

$$\|X\|_\psi := \inf\{t > 0 : E\psi(|X|/t) \leq 1\}.$$

Example 3.7 (Orlicz Norms)

- *$\psi(x) = x^p$, $p \geq 1$. Because $\|X\|_\psi = (E|X|^p)^{1/p} = \|X\|_p$.*
- *$\psi(x) = \exp(x^2) - 1$. $\|X\|_{\psi_2} = \inf\{t > 0 : E\exp(X^2/t^2) \leq 2\}$.*
- *$\psi(x) = \exp(x) - 1$. $\|X\|_{\psi_1} = \inf\{t > 0 : E\exp(|X|/t) \leq 2\}$.*

Now, we know where does the upper bound of 2 come from. It is because $2 = 1 + 1$ (just kidding). The Orlicz space L_ψ consists of all random variables X with a finite Orlicz norm. With $\psi = x^p$, $p \geq 1$, we recover the L^p space.

Remark 3.8 *We could locate L_{ψ_2} in the hierarchy of the classical L^p spaces:*

$$L^\infty \subset L_{\psi_2} \subset L^p.$$

This means we successfully extended bounded random variables L^∞.

3.4.4 Sub-Gaussian Properties in terms of $\|X\|_{\psi_2}$

Proposition 3.9 *The properties in Proposition 3.4 could be written as*

1. *$P(|X| \geq t) \leq 2\exp(-t^2/\|X\|^2_{\psi_2})$ for all $t \geq 0$.*
2. *$\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p}$*
3. *$E\exp(X^2/\|X\|^2_{\psi_2}) \leq 2$ (Definition)*

4. If $EX = 0$ then $E\exp(\lambda X) \leq \exp(C\lambda^2 \|X\|_{\psi_2}^2)$ for all $\lambda \in \mathbb{R}$.

Here C are absolute constant that has nothing to do with X.

Proof: Since property 1 does not have constant, we prove it as following:

$$\begin{aligned} P(|X| \geq t) &= P\left(\exp\left(\frac{X^2}{\|X\|_{\psi_2}^2}\right) \geq \exp\left(\frac{t^2}{\|X\|_{\psi_2}^2}\right)\right) \\ &\leq \exp\left(-\frac{t^2}{\|X\|_{\psi_2}^2}\right) E\left(\exp\frac{X^2}{\|X\|_{\psi_2}^2}\right) \leq 2\exp\left(-\frac{t^2}{\|X\|_{\psi_2}^2}\right). \end{aligned}$$

For the others, we could construct $X_{new} = X/\|X\|_{\psi_2}$ which has $K_1 = 1$ and $K_3 = \sqrt{2}$. Based on the proof of Proposition 3.4, we have this proposition proved. ■

Since the last property matters the most for the concentration, our goal becomes showing $\|X\|_{\psi_2} < \infty$. The following two properties of the sub-Gaussian norm are useful.

Lemma 3.10 (Centering) *If X is a sub-Gaussian random variable, then $X - EX$ is also sub-Gaussian and*

$$\|X - EX\|_{\psi_2} \leq C\|X\|_{\psi_2},$$

where C is an absolute constant.

Proof: Since $\|X\|_{\psi_2}$ is a norm, we have

$$\|X - EX\|_{\psi_2} \leq \|X\|_{\psi_2} + \|EX\|_{\psi_2}.$$

By definition, for a constant a, $\|a\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}|a|$. Thus, we have

$$\begin{aligned} \|EX\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}|EX| &\leq \frac{1}{\sqrt{\ln 2}}E|X| \quad \text{(Jensen's inequality)} \\ &\leq C\|X\|_{\psi_2} \quad \text{(Property 2 of Proposition 3.9)}. \end{aligned}$$

■

Is this rigorous proof for the argument that $N(\mu, \sigma^2)$ is sub-Gaussian no matter what μ is?

Proposition 3.11 *[Sum of independent Sub-Gaussians] Let $X_1, \ldots X_N$ be independent, mean zero, sub-Gaussian random variables. Then $\sum_{i=1}^N X_i$ is also sub-Gaussian random variable, and*

$$\left\|\sum_{i=1}^N X_i\right\|_{\psi_2}^2 \leq C\sum_{i=1}^N \|X_i\|_{\psi_2}^2,$$

where C is an absolute constant.

Proof: Since it is the sum of independent variables, we will approach this problem with MGF. For any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned} E\exp(\lambda\sum_{i=1}^{N} X_i) &= \prod_{i=1}^{N} E\exp(\lambda X_i) \quad \text{independence} \\ &\le \prod_{i=1}^{N} \exp(C\lambda^2\|X_i\|_{\psi_2}^2) \quad \text{(property 4 of Proposition 3.9)} \\ &= \exp(C\lambda^2 K^2) \quad \text{where } K^2 = \textstyle\sum_{i=1}^{N}\|X_i\|_{\psi_2}^2. \end{aligned} \tag{3.1}$$

Based on the 4th property in Proposition 3.9, we know $\sum_{i=1}^{N} X_i$ is sub-Gaussian (If $E\exp(\lambda X) \le \exp(C_2\lambda^2)$ for all $\lambda \in \mathbb{R}$, then X is sub-Gaussian). More specifically, from the 4th property, we have

$$E\exp(\lambda\sum_{i=1}^{N} X_i) \le \exp(C\lambda^2\|\sum_{i=1}^{N} X_i\|_{\psi_2}^2)$$

Compare it with Equation 3.1, we have

$$\left\|\sum_{i=1}^{N} X_i\right\|_{\psi_2}^2 \le C_1 K^2.$$

Here we have C_1 because $\left\|\sum_{i=1}^{N} X_i\right\|_{\psi_2}^2$ might not be the smallest number such that the 4th property holds. ■

Since $\|\cdot\|_{\psi_2}$ is a norm, we naturally have

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

Taking squares to both sides to have

$$\left\|\sum_{i=1}^{N} X_i\right\|_{\psi_2}^2 \le \sum_{i=1}^{N}\sum_{j=1}^{N}\|X_i\|_{\psi_2}\|X_j\|_{\psi_2}.$$

The right-hand side has in total N^2 terms. Proposition 3.11 shows that if we have X_i independent, then we could use N terms to bound instead. This is similar to the behavior of variance,

$$E\left(\sum_{i=1}^{N} X_i\right)^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} E(X_iX_j) = \sum_{i=1}^{N} EX_i^2,$$

if X_is are independent.

3.4.5 General Hoeffding's Inequality

Theorem 3.12 *[General Hoeffding's Inequality 1] Let $X_1, \ldots, X_N$ be independent, mean zero, sub-Gaussian random variables. Then for every $t \geq 0$, we have*

$$P\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{N} \|X_i\|_{\psi_2}^2}\right).$$

Proof: By the first property of Proposition 3.9, we only need to show

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

This is exactly what Proposition 3.11 says.

■

Theorem 3.13 (General Hoeffding's Inequality 2) *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-Gaussian random variables and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then for every $t \geq 0$, we have*

$$P\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right),$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Proof: From Theorem 3.12, we know

$$P\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{N} \|a_i X_i\|_{\psi_2}^2}\right).$$

Because $\|\cdot\|_{\psi_2}$ is a norm, we have $\|a_i X_i\|_{\psi_2}^2 = a_i^2\|X_i\|_{\psi_2}^2$, thus,

$$\sum_{i=1}^{N} \|a_i X_i\|_{\psi_2}^2 = \sum_{i=1}^{N} a_i^2\|X_i\|_{\psi_2}^2 \leq K^2\sum_{i=1}^{N} a_i^2 = K^2\|a\|_2^2.$$

■

From the proof above, we could use the properties of $\|\cdot\|_{\psi_2}$ without going back to the Markov inequality.

3.5 Sub-Exponential Distribution

Although we have expanded concentration for sub-Gaussian distributions, we still leave out important ones that has heavier tail than sub-Gaussian but lighter tail than heavy-tail distribution. Specifically, we still can not get concentration result for $\|\vec{X}\|_2^2$, where $\vec{X} \sim N(0, I_p)$ because

$$P(X_i^2 \geq t) = P(X_i \geq \sqrt{t}) \leq 2\exp(-t/2) \quad \text{(see Section 3.4.1)},$$

which behaves like exponential distribution. Meanwhile, it is not hopeless to have concentration: say $X \sim Exp(1)$ $(EX = 1)$, we have

$$E\exp(\lambda(X-1)) = \int_0^\infty \exp(\lambda(x-1))\exp(-x)dx = \frac{\exp(-\lambda)}{1-\lambda} \quad \text{(when } \lambda < 1)$$

Next, we use Taylor expansion to show that it is reasonable[1] to believe that it could be bounded by $\exp(C\lambda^2)$ for $|\lambda| < c$.

$$\frac{\exp(-\lambda)}{1-\lambda} = (1 - \lambda + \frac{1}{2}\lambda^2 + \dots)(1 + \lambda + \lambda^2 + \dots) = (1 + \frac{1}{2}\lambda^2 + \dots) \leq \exp(C\lambda^2) \quad \text{(when } \lambda \text{ is small and } C \text{ is large)}.$$

Could we extend this argument in the following way? For any random variable X that $EX = 0$,

$$E\exp(\lambda X) \approx E(1 + \lambda X + \frac{1}{2}\lambda^2 X^2) = 1 + \frac{1}{2}\lambda^2\sigma^2 \leq \exp(1/2\lambda^2\sigma^2),$$

when λ is small enough. Thus, there always exists a small region around 0 that $E\exp(\lambda X) \leq \exp(C\lambda^2)$. This is totally wrong, but where is the issue?

Coming back to the exponential distribution, $E\exp(\lambda(X-1)) \leq \exp(C\lambda^2)$ is exactly what we need to establish concentration (Recall the analysis in Lecture 2: higher order of λ within an interval is enough.) Next, we will show rigorously concentration also exists for them.

Proposition 3.14 (Sub-Exponential Properties) *Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*

$$P(|X| \geq t) \leq 2\exp(-t/K_1), \quad \forall t \geq 0.$$

2. *The moments of X satisfy*

$$\|X\|_p = (E|X|^p)^{1/p} \leq K_2 p, \quad \forall p \geq 1.$$

[1] If you know how to prove this elegantly without using Tailor expansion, please let me know. However, do not put too much effort in it. It is just a numeric inequality.

3. *The MGF of $|X|$ satisfies*

$$E\exp(\lambda|X|) \le \exp(K_3\lambda), \quad 0 \le \lambda \le \frac{1}{K_3}$$

4. *The MGF of $|X|$ is bounded at some point, namely*

$$E\exp(|X|/K_4) \le 2.$$

Moreover, if $EX = 0$, then properties 1-4 are also equivalent to the following one.

5. *The MGF of X satisfies*

$$E\exp(\lambda X) \le \exp(K_5^2\lambda^2), \quad \forall|\lambda| \le \frac{1}{K_5}$$

Proof: Since the 5th one is the most important for the proof of concentration, we will only establish the route from 1 to 5.

$1 \Rightarrow 2$ & $1 \Rightarrow 3$. One could obtain them using the same technique as the sub-Gaussian case.

$2 \Rightarrow 5$. Without loss of generality, we let $K_2 = 1$. Using Taylor expansion, we have

$$\begin{aligned}
E\exp(\lambda X) &= E\left(1 + \lambda X + \sum_{i=2}^{\infty}\frac{(\lambda X)^i}{i!}\right) = 1 + \sum_{i=2}^{\infty}\frac{\lambda^i EX^i}{i!} \quad (EX = 0)\\
&\le 1 + \sum_{i=2}^{\infty}\frac{(\lambda i)^i}{(i/e)^i} \quad (\text{Stirling's Approximation: } i! \ge (i/e)^i)\\
&= 1 + \sum_{i=2}^{\infty}(e\lambda)^i = 1 + \frac{(e\lambda)^2}{1-e\lambda} \quad (e\lambda < 1)\\
&\le 1 + 2(e\lambda)^2 \quad (\text{if } e\lambda \le 1/2)\\
&\le \exp(2e^2\lambda^2) \quad (1 + x \le \exp(x)).
\end{aligned}$$

■

Now, we could define the norm of sub-exponential random variables.

Definition 3.15 (Sub-Exponential Norm) *The sub-exponential norm of X, denoted as $\|X\|_{\psi_1}$, is defined as*

$$\|X\|_{\psi_1} = \inf\{t > 0 : E\exp(|X|/t) \le 2\}.$$

Based on our previous discussion about the Orlicz norm, you shall know where does 2 come from.

3.5.1 Sub-Exponential Properties in terms of $\|X\|_{\psi_1}$

Proposition 3.16 *The properties in Proposition 3.14 could be written as*

1. $P(|X| \geq t) \leq 2\exp(-t/\|X\|_{\psi^1})$ *for all* $t \geq 0$.
2. $\|X\|_{L^p} \leq C\|X\|_{\psi_1} p$
3. $E\exp(|X|/\|X\|_{\psi_1}) \leq 2$ *(Definition)*
4. *If* $EX = 0$ *then* $E\exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_1}^2)$ *for all* $|\lambda| \leq c/\|X\|_{\psi_1}$.

Here c *and* C *are absolute constants that have nothing to do with* X.

Proof: Think about what are K_1 and K_2 for $X_{new} = X/\|X\|_{\psi_1}$. Following the proof of Proposition 3.14, we could obtain this proposition. ■

3.6 Connections between Sub-Gaussian and Sub-Exponential

Lemma 3.17

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2.$$

One could prove this using the definition of $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$.

Lemma 3.18 (Product of sub-Gaussian is sub-exponential) *Let* X *and* Y *be sub-Gaussian random variables. Then* XY *is sub-exponential. Moreover,*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$$

Proof: Without loss of generality, assume $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. Otherwise, we could introduce $X_{new} = X/\|X\|_{\psi_2}$ and $Y_{new} = Y/\|Y\|_{\psi_2}$. By definition of $\|\cdot\|_{\psi_2}$, we have

$$E\exp(X^2) \leq 2 \quad \text{and} \quad E\exp(Y^2) \leq 2.$$

Meanwhile,

$$\begin{aligned} E\exp(|XY|) &\leq E[\exp(X^2/2) \times \exp(Y^2/2)] \quad (|XY| \leq X^2/2 + Y^2/2) \\ &\leq \frac{1}{2}(E\exp(X^2) + E\exp(X^2)) \quad (\exp(X^2/2) \times \exp(Y^2/2) \leq \frac{1}{2}\left(\exp X^2 + \exp Y^2\right)) \\ &\leq 2. \end{aligned}$$

Thus, we have $\|XY\|_{\psi_1} \leq 1 = \|X\|_{\psi_2}\|Y\|_{\psi_2}$.

■

3.7 Bernstein's Inequality

This inequality provides the concentration inequality for sums of independent sub-exponential random variables.

Theorem 3.19 (Bernstein's Inequality 1) *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

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

Let us first analyze why we get the minimum of two terms. For sub-exponential distribution, we could bound the $P(\sum_{i=1}^N X_i \geq t)$ by

$$\exp(-\lambda t + C\lambda^2), \quad 0 < \lambda \leq b.$$

For $-\lambda t + C\lambda^2$, if $\lambda^\star = t/2C \leq b$, then it achieves $-t^2/4C$. This is basically the bound of Hoeffding's inequality. If not, namely $b < t/2C$, the smallest value will be achieved at $\lambda = b$ which leads to $-tb + Cb^2 < -tb/2 < 0$. Instead of quadratic in t, it is linear. Now we only need to nail down C and b, then we could prove Bernstein's inequality.

Proof: From Proposition 3.16, we know that

$$E\exp(\lambda X_i) \leq \exp(C_1\lambda^2\|X_i\|_{\psi_1}^2) \quad \forall|\lambda| \leq c_1/\|X_i\|_{\psi_1}.$$

This means that $b = c_1/\max_i \|X_i\|_{\psi_1}$. Meanwhile, the coefficient of λ^2 is $C = C_1\sum_i \|X_i\|_{\psi_1}^2$. Based on our analysis,

$$P(\sum_{i=1}^N X_i \geq t) \leq \exp\left(-\min(\frac{t^2}{4C_1}, \frac{tb}{2})\right).$$

We could use the following naive bound to make the form nicer[2],

$$b = \frac{c_1}{\max_i \|X_i\|_{\psi_1}} \geq \frac{2c}{\max_i \|X_i\|_{\psi_1}} \quad \text{and} \quad C = C_1\sum_i \|X_i\|_{\psi_1}^2 \leq \frac{1}{4c}\sum_i \|X_i\|_{\psi_1}^2,$$

$$\text{where } c = \min\left(\frac{c_1}{2}, \frac{1}{4C_1}\right),$$

[2] You could tell how careless we are about constants. Thus, the probability bounds (in nice forms) are usually very loose. However, you could tighten them significantly if you do not use these naive bounds.

which leads to a simpler bound

$$\exp\left(-c\min\left(\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right).$$

Do the same thing for $P(\sum_{i=1}^N X_i \leq -t)$ to obtain Bernstein's Inequality. ■

3.7.1 Analysis of Bernstein's Inequality

When t is large, the quadratic term will dominate the linear term, leading to a sub-exponential decay. This is on the same level of a single term X_i, which is also sub-exponential. When t is small enough, the linear term dominates, then we have sub-Gaussian decay. This is what CLT suggested since normal distribution has sub-Gaussian decay.

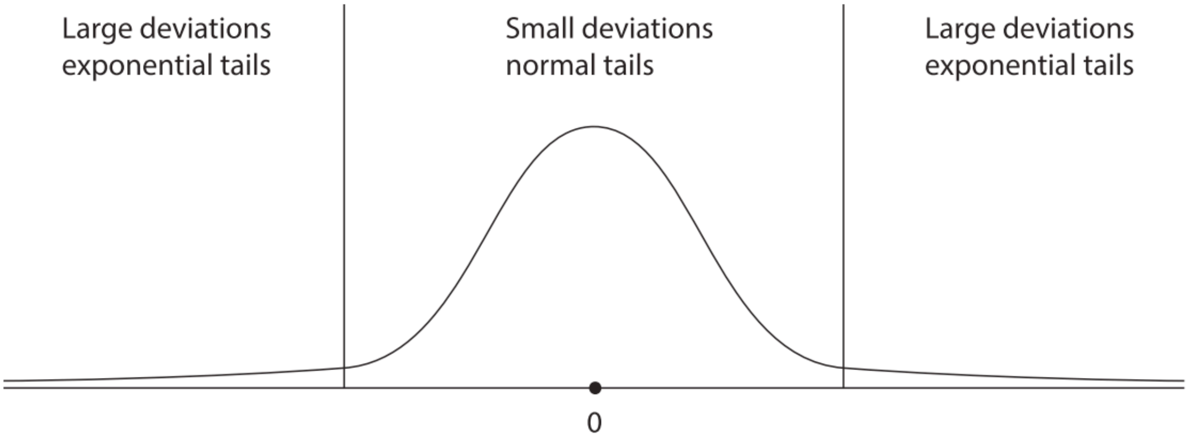


Figure 3.1: Bernstein's Inequality: sub-Gaussian for small deviation and sub-exponential for large deviation.

3.7.2 Other Forms of Bernstein's Inequality

The same proof technique of Theorem 3.19 could also handle $\sum_{i=1}^N a_i X_i$, which leads to the following version of Bernstein's inequality.

Theorem 3.20 (Bernstein's Inequality 2) *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then for every $t \geq 0$, we have*

$$P\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right),$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Proof: Notice that

$$\sum_{i=1}^{N} \|a_i X_i\|_{\psi_1}^2 \leq \left(\max_i \|X_i\|_{\psi_1}\right)^2 \sum_{i=1}^{N} |a_i|^2 = K^2 \|a\|_2^2$$
$$\max_i \|a_i X_i\|_{\psi_1} \leq \max_i |a_i| \max_i \|X_i\|_{\psi_1} = K\|a\|_\infty.$$

Replace X_i with $a_i X_i$ in Theorem 3.19 to obtain this theorem. ■

In the special case $a_i = 1/N$, we obtain Bernstein's inequality for averages:

Corollary 3.21 (Bernstein's Inequality for Averages) *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables. Then for every $t \geq 0$, we have*

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

where $K = \max_i \|X_i\|_{\psi_1}$.

This is a quantitative form of the law of large numbers.

3.8 Concentration of $\|X\|_2$

Let $X = (X_1, \ldots X_p)$ be a random vector in $\mathbb{R}^p$. Assume X_i are independent sub-Gaussian random variables with $EX_i^2 = 1$. Then we have

$$E\|X\|_2^2 = E\sum_{i=1}^{p} X_i^2 = \sum_{i=1}^{p} EX_i^2 = p.$$

Since X_i is sub-Gaussian, X_i^2 is sub-exponential. Thus, based on Bernstein's Inequality, we should have $\|X\|_2^2$ concentrates around p. Then we should also expect $\|X\|_2$ concentrates around $\sqrt{p}$.

To establish the relationship between $|\|X\|_2 - \sqrt{p}|$ and $|\|X\|_2^2 - p|$, we first prove a numeric inequality.

Lemma 3.22 *For all numbers $z \geq 0$ and $\delta \geq 0$, if $|z-1| \geq \delta$, then $|z^2 - 1| \geq \max(\delta, \delta^2)$.*

Proof: If $z \geq 1 + \delta$, square both sides to have $z^2 \geq 1 + 2\delta + \delta^2$. Thus,

$$z^2 - 1 \geq 2\delta + \delta^2 \geq \max(\delta, \delta^2).$$

If $z \leq 1 - \delta$, because $z \geq 0$, we must have $\delta \leq 1$. Square both sides to have $z^2 \leq 1 - 2\delta + \delta^2$ which implies

$$|z^2 - 1| = -(z^2 - 1) \geq 2\delta - \delta^2 \geq \delta = \max(\delta, \delta^2).$$

Here we use $\delta \leq 1$ repeatedly. ■

Theorem 3.23 (Concentration of the Norm) *Let $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be a random vector with independent, sub-Gaussian coordinates X_i that satisfy $EX_i^2 = 1$. Then*

$$\left\| \|X\|_2 - \sqrt{p} \right\|_{\psi_2} \leq CK^2,$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Proof: Since X_i is sub-Gaussian, we have X_i^2 is sub-exponential. Because of the property of centering, we also know $X_i^2 - 1$ is sub-exponential. More precisely,

$$\|X_i^2 - 1\|_{\psi_1} \leq C\|X_i^2\|_{\psi_1} = C\|X_i\|_{\psi_2}^2 \leq CK^2.$$

Apply Corollary 3.21 (notice that K has different meaning there) to $1/p\sum_i (X_i^2 - 1)$ to have

$$P\left(\left| \frac{1}{p} \sum_{i=1}^{p} (X_i^2 - 1) \right| \geq t \right) \leq 2\exp\left(-cp \min\left(\frac{t^2}{C^2K^4}, \frac{t}{CK^2} \right) \right).$$

If we choose C large enough (such that $CK^2 \geq 1$), the RHS of the above inequality could be bounded by

$$2\exp\left(-\frac{\tilde{c}p}{K^4} \min\left(t^2, t\right) \right).$$

Now, we will use Lemma 3.22 to transform the concentration in $\|X\|_2^2$ to $\|X\|_2$.

$$\begin{aligned} P\left(\left| \frac{1}{\sqrt{p}} \|X\|_2 - 1 \right| \geq \delta \right) &\leq P\left(\left| \frac{1}{p} \|X\|_2^2 - 1 \right| \geq \max(\delta, \delta^2) \right) \\ &= P\left(\left| \frac{1}{p} \sum_{i=1}^{p} (X_i^2 - 1) \right| \geq \max(\delta, \delta^2) \right) \\ &\leq 2\exp\left(-\frac{\tilde{c}p}{K^4} \delta^2 \right). \end{aligned}$$

The last inequality we use the fact that

$$\min\left(\max(\delta, \delta^2)^2, \max(\delta, \delta^2) \right) = \begin{cases} \min(\delta^2, \delta) = \delta^2 & \delta < 1 \\ \min(\delta^4, \delta^2) = \delta^2 & \delta \geq 1. \end{cases}$$

Thus, we have

$$P\left(\left| \|X\|_2 - \sqrt{p} \right| \geq t \right) \leq 2\exp\left(-\frac{\tilde{c}}{K^4} t^2 \right), \quad t = \sqrt{p}\delta.$$

Based on Proposition 3.9, this is equivalent to what we want to prove.

■

Example 3.24 *We could apply the above theorem to $X \sim N(0, I_p)$ to have $\|X\|_2$ concentrates around $\sqrt{p}$. Moreover, $X/\|X\|_2$ follows $Unif(S^{p-1})$ and is independent of $\|X\|_2$. Thus,*

$$X = \frac{X}{\|X\|_2} \times \|X\|_2 \approx Unif(\sqrt{p}S^{p-1}).$$

3.9 Almost Orthogonal Vectors

Let $\vec{X}_1$ and $\vec{X}_2$ be two independent random variable following $N(0, I_p)$. We want to argue that $\vec{X}_1/\|\vec{X}_1\|_2$ and $\vec{X}_2/\|\vec{X}_2\|_2$ are almost orthogonal. Namely,

$$\cos(\theta_{12}) \triangleq \left\langle \frac{\vec{X}_1}{\|\vec{X}_1\|_2}, \frac{\vec{X}_2}{\|\vec{X}_2\|_2} \right\rangle = \frac{<\vec{X}_1, \vec{X}_2>}{\|\vec{X}_1\|_2\|\vec{X}_2\|_2} \approx 0. \tag{3.2}$$

Based on Theorem 3.23, we know $\|\vec{X}_i\|_2 \approx \sqrt{p}$, thus the denominator is about p. Meanwhile,

$$\begin{aligned} E<\vec{X}_1, \vec{X}_2>^2 &= E\left(E\left(<\vec{X}_1, \vec{X}_2>^2 |\vec{X}_2\right)\right) = E\left(\vec{X}_2^T E(\vec{X}_1\vec{X}_1^T)\vec{X}_2\right) \\ &= E\vec{X}_2^T I_p \vec{X}_2 = E\|\vec{X}_2\|_2^2 = p. \end{aligned}$$

Thus, we are expecting $|<\vec{X}_1, \vec{X}_2>|$ is of the level $\sqrt{p}$. Then $\cos(\theta_{12})$ (Equation 3.2) is about $1/\sqrt{p}$ which is very small when p is large. To make a more rigorous argument, we need to show $<\vec{X}_1, \vec{X}_2>$ concentrates. Based on the Lemma 3.18, we have

$$\|X_{1i}X_{2i}\|_{\psi_1} \le \|X_{1i}\|_{\psi_2}\|X_{2i}\|_{\psi_2} = \frac{1}{\ln 2}.$$

where X_{1i} and X_{2i} are the ith coordinate of $\vec{X}_1$ and $\vec{X}_2$, respectively. Thus, we could use Bernstein inequality for

$$<\vec{X}_1, \vec{X}_2>= \sum_{i=1}^{p} X_{1i}X_{2i}. \quad \text{(Check all the conditions are satisfied!)}$$

Therefore, we have

$$P(|<\vec{X}_1, \vec{X}_2>| \ge t) \le 2\exp\left(-c\min(\frac{t^2}{p}, t)\right)$$

Because $|<\vec{X}_1, \vec{X}_2>| \le \epsilon(1-\delta)^2 p$ (event A), $\|\vec{X}_1\|_2 \ge (1-\delta)\sqrt{p}$ (event B), and $\|\vec{X}_2\|_2 \ge (1-\delta)\sqrt{p}$ (event

C) imply that $|\cos(\theta_{12})| \leq \epsilon$, we have

$$\begin{aligned} &P\left(|\cos(\theta_{12})| \leq \epsilon\right) \geq P(A \cap B \cap C) \\ \Rightarrow &P\left(|\cos(\theta_{12})| > \epsilon\right) \leq P(A^c \cup B^c \cup C^c) \leq P(A^c) + P(B^c) + P(C^c) \leq 2\exp(-c\epsilon^2(1-\delta)^2 p) + 2\exp(-Cp\delta^2) \\ &\qquad\qquad\qquad\quad \leq 4\exp(-C(\epsilon,\delta)p) \end{aligned}$$

The last inequality comes from 3 different concentrations. This is totally different from the 2-D case: one could show there $E\theta_{12} = \pi/4$ instead of $\pi/2$ in a high dimension setting.

Remark 3.25 (Gaussian Chaos Variables) *Let $Q \in \mathbb{R}^{p\times p}$ be a symmetric matrix, and let $w, \tilde{w}$ be independent zero-mean Gaussian random vectors with covariance matrix I_p. The random variable*

$$Z \triangleq w^T Q \tilde{w}$$

is known as a (decoupled) Gaussian chaos. If $Q = I_p$, we get back to $< w, \tilde{w} >$ which we have proved to be sub-exponential. Using the concentration of Lipschitz function of $N(0, I_p)$, one could prove that Z is also sub-exponential. For more details, please see Wainwright (2019, Example 2.31).

3.9.1 Exponential Many Almost Orthogonal Vectors

Now, let us get K independent random variables following $N(0, I_p)$, denoted as $\vec{X}_1, \ldots, \vec{X}_K$. Then we have $K(K-1)/2$ pairs each one has a very low probability of having a large $\cos(\theta)$. Utilizing union bound, we have

$$P\left(\max_{1\leq i<j\leq K} |\cos(\theta_{ij})| > \epsilon\right) \leq 2K(K-1)\exp(-C(\epsilon,\delta)p).$$

If we choose $K = 1/2\exp(C(\epsilon,\delta)p/2)$, we know the RHS is smaller than 1. This means that

$$P\left(\max_{1\leq i<j\leq K} |\cos(\theta_{ij})| \leq \epsilon\right) > 0.$$

Thus, we prove that there exist exponential many vectors that are almost orthogonal to each other. I personally find this way of proving existence fascinating! It is just one sentence in Terrace Tao's blog, but it takes me more than two days to figure out rigorously. In other words, almost orthogonality is much lower value in high dimension than orthogonality because only p vectors are orthogonal to each other.

Btw, the current result is weaker than the previous one regarding $1/\sqrt{p}$. The reason is that I fail to prove the $| < \vec{X}_1, \vec{X}_2 > |$ concentrates around $\sqrt{p}$ as suggested in Vershynin (2018, Remark 3.2.5). If you know how to prove it, please let me know.

3.9.2 Motivation for My Research

In portfolio optimization, there are usually lots of risky assets (p is large). One attempt is dimension reduction from p dimensions to $k \ll p$. Somehow, it is easy to construct a basis, denoted $\hat{v}_i$s ($\|\hat{v}_i\|_2 = 1$), such that $\|\hat{v}_i - v_i\|_2 \approx 0$ for $i \leq k$ while $\|\hat{v}_i - v_i\|_2 \gg 0$ for $k < i \leq p$. Here v_i are the true value that $\hat{v}_i$ is trying to estimate. It is temping to throw away all $\hat{v}_i,\ k < i \leq p$ since they are terrible estimations. However, I realize the value of exact orthogonality, and argue that the space spanned by $\hat{v}_i,\ k < i \leq p$ is well estimated because the space is orthogonal to $\hat{v}_i, i \leq k$.

Next, I will prove this rigorously. Introduce $\hat{P}_{\perp k} = \sum_{i=k+1}^{p} \hat{v}_i \hat{v}_i^T$ and $P_{\perp k} = \sum_{i=k+1}^{p} v_i v_i^T$. They are the projection operators of the space spanned by $\hat{v}_i,\ k < i \leq p$ and $v_i,\ k < i \leq p$, respectively. Because they are basis with $\|\hat{v}_i\|_2 = 1 = \|v_i\|_2$, we have

$$\hat{P}_{\perp k} = I_p - \sum_{i=1}^{k} \hat{v}_i \hat{v}_i^T \quad P_{\perp k} = I_p - \sum_{i=1}^{k} v_i v_i^T$$
$$\Rightarrow \|\hat{P}_{\perp k} - P_{\perp k}\|_{op} = \left\| \sum_{i=1}^{k} v_i v_i^T - \sum_{i=1}^{k} \hat{v}_i \hat{v}_i^T \right\|_{op} \leq 2 \sum_{i=1}^{k} \|\hat{v}_i - v_i\|_2.$$

Thus, $\hat{P}_{\perp k}$ is also a good estimation of $P_{\perp k}$ and one shall not throw it away.

3.9.3 Similar Technique to Prove Existence [Optional]

For a high dimension bounded set, does there exist a linear transformation to a low dimension space such that two close balls sandwich the convex hull of its image? Milman used a random matrix as the linear transformation and proved the probability of that property happening is high. Then, there must exist such linear transformation.

Theorem 3.26 (Dvoretzky-Milman's Theorem. Theorem 11.3.3 (Vershynin 2018)) *Let A be an $m \times p$ Gaussian random matrix with i.i.d. $N(0,1)$ entries, $T \subset \mathbb{R}^p$ be a bounded set, and let $\epsilon \in (0,1)$. Suppose*

$$m \leq c\epsilon^2 d(T),$$

where $d(T)$ is the stable dimension of T. Then with probability at least 0.99, we have

$$(1-\epsilon)B \subset conv(AT) \subset (1+\epsilon)B,$$

where B is a Euclidean ball with radius $W(T)$.

Figure 3.9.3 is a visualization of $T = [-1,1]^7$ onto 2 dimension.

In my research, I consider T where the data cloud belongs. I want to argue that random directions $X_i \sim$

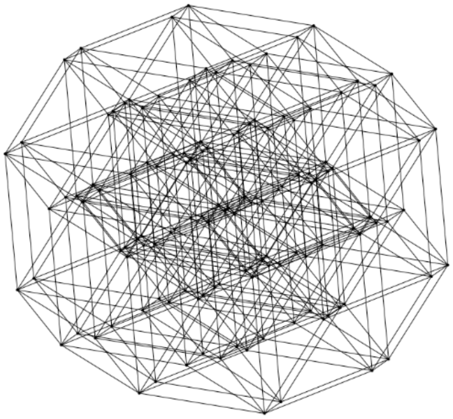


Figure 3.2: A Random Projection of a 7-dimension cube onto the plane.

$Unif(S^{p-1})$ ($\|X_i\|_2 = 1$ is essential for me) will only transform data to be close to a ball without any direction with extremely high variance. With the concentration of norm, we know that Theorem 3.26 with random directions also holds.

References

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 4: A Note on $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$

Lecturer: Long Zhao, longzhao@nus.edu.sg

4.1 Order of t

Based on the definition of $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$, we have

$$P(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{\|X\|_{\psi_2}^2}\right)$$
$$P(|X| \geq t) \leq 2\exp\left(-\frac{t}{\|X\|_{\psi_1}}\right).$$

Thus, I claim that proving concentration is equivalent to show corresponding norm is finite. However, there is a catch and I will illustrate it using the following example. Say, I want to have concentration of $\sum_{i=1}^n X_i$ where X_i is sub-Gaussian. Then we have

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2}^2 \leq \left(\sum_{i=1}^n \|X_i\|_{\psi_2}\right)^2 = \sum_{1\leq i,j\leq n} \|X_i\|_{\psi_2}\|X_j\|_{\psi_2} < \infty. \tag{4.1}$$

This is true without assuming independence of X_i. Meanwhile, this is clearly not concentration. What goes wrong? In concentration, we have $t = n\epsilon$ where $\epsilon > 0$. This means that t^2 is $O(n^2)$. The probability bound is meaningless if $\|\sum_{i=1}^n X_i\|_{\psi_2}^2$ is also $O(n^2)$. Unfortunately, based on Equation 4.1, without independence of X_i, $\|\sum_{i=1}^n X_i\|_{\psi_2}^2$ is order n^2. This shows the importance of following property of $\|\cdot\|_{\psi_2}$ when X_is are independent (and $EX_i = 0$),

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2}^2 \leq C\sum_{i=1}^n \|X_i\|_{\psi_2}^2.$$

Right now, the right-hand side is $O(n)$ which means that we have concentration when n is large.

In the future, we will use the following properties of $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$ to prove concentration.

1. $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$ are norms. Namely, we have triangle inequality, and $\|aX\|_\psi = |a|\|X\|_\psi$.
2. $\|X - EX\|_\psi \leq C\|X\|_\psi$. That is to say, centering will not change the distribution family.
3. Connection between $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$: $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$ & $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$
4. **If X_is are independent sub-Gaussian and $EX_i = 0$, $\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \leq C\sum_{i=1}^n \|X_i\|_{\psi_2}^2$.**

5. **If X_is are independent sub-exponential and $EX_i = 0$, $\|\sum_{i=1}^n X_i\|_{\psi_1} \approx C\max_i \|X_i\|_{\psi_1}$ for large deviations.** Bernstein's inequality gives a more precise relationship. However, I think the vague one is easier to remember and more intuitive. It states that the tail behavior is determined by the one with thickest tail.

   From the properties above, it is easy to derive the following one.

6. If X_is are independent sub-Gaussian and $EX_i = 0$,

$$\left\|\sum_{i=1}^n a_i X_i\right\|_{\psi_2}^2 \le C\sum_{i=1}^n a_i^2\|X_i\|_{\psi_2}^2 \le C\left(\sum_{i=1}^n a_i^2\right)\max_i \|X_i\|_{\psi_2}^2 = C\|a\|_2^2 \max_i \|X_i\|_{\psi_2}^2.$$

   This property implies that any linear combination of X_is is still sub-Gaussian. This will be handy in the future.

   Since we will use the concentration of the norm quite frequently in the future, I also list it here as a property.

7. $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ be a random vector with independent, sub-Gaussian coordinates X_i that satisfies $EX_i^2 = 1$. Then

$$\|\|X\|_2 - \sqrt{p}\|_{\psi_2} \le C(\max_i \|X_i\|_{\psi_2})^2.$$

From now on, we will use the above properties to simplify our argument, and we will take the Johnson-Lindenstrauss Lemma as a demo.

4.2 Application: Johnson-Lindenstrauss Lemma

Say we have n observations and each one is $p \gg 1$ dimension. Is it possible to find a d-dimension (d might be much smaller than p) transformation of data such that the pairwise (Euclidean) distance is maintained with small error. Mathematically speaking, let x_i and y_i be the ith original and transformed data point. Is it possible to have

$$(1-\epsilon)\|x_i - x_j\|_2 \le \|y_i - y_j\|_2 \le (1+\epsilon)\|x_i - x_j\|_2 \tag{4.2}$$

holds for every pair of i, j?

Let's first deal with the simplest case with fixed i and j ($i \neq j$), and we use $a = x_i - x_j$ and $Z = y_i - y_j$ to simplify the argument. If we only demand

$$(1-\epsilon)\|a\|_2 \le \|Z\|_2 \le (1+\epsilon)\|a\|_2$$

to hold with high probability, it is essentially the concentration of the norm, $\|Z\|_2$, around $\|x\|_2$. All we need is each coordinate of Z, denoted as Z_j, is sub-Gaussian and $EZ_j^2 = \|a\|_2^2/d$. To see the second requirement, notice that $E\|Z\|_2^2 = \sum_{j=1}^d EZ_j^2 = \|a\|_2^2$.

Obtain sub-Gaussian Z_j. Since a could be any vector in $\mathbb{R}^p$, the first requirement is not a trivial task. Luckily, based on property 6, we know that any linear combination of independent sub-Gaussian is still sub-Gaussian. Thus, for Z_j, we could create p independent sub-Gaussian random variables, $W_{j1}, \dots, W_{jp}$, such that

$$\left\|\sum_{i=1}^p a_i W_{ji}\right\|_{\psi_2}^2 \le C\|a\|_2^2 \max_i \|W_{ji}\|_{\psi_2}^2.$$

To simplify the notation, we introduce random vector $W_j = (W_{j1}, \dots, W_{jp})^T$ and then $Z_j = W_j^T a$.

Obtain $EZ_j^2 = \|a\|_2^2/d$. Because $Z_j = W_j a$, we could rewrite this requirement as following,

$$a^T a/d = \|a\|_2^2/d = EZ_j^2 = Ea^T W_j W_j^T a = a^T E(W_j W_j^T) a.$$

That is to say, all we need is

$$E(W_j W_j^T) = I_p/d.$$

Because of the independence among $W_{j1}, \dots, W_{jp}$, we could write

$$\begin{aligned} E(W_j W_j^T) &= \begin{pmatrix} EW_{j1}W_{j1} & EW_{j1}W_{j2} & \cdots & EW_{j1}W_{jp} \\ EW_{j2}W_{j1} & EW_{j2}W_{j2} & \cdots & EW_{j2}W_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ EW_{jp}W_{j1} & EW_{jp}W_{j2} & \cdots & EW_{jp}W_{jp} \end{pmatrix} \\ &= \begin{pmatrix} EW_{j1}^2 & EW_{j1}EW_{j2} & \cdots & EW_{j1}EW_{jp} \\ EW_{j2}EW_{j1} & EW_{j2}^2 & \cdots & EW_{j2}EW_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ EW_{jp}EW_{j1} & EW_{jp}EW_{j2} & \cdots & EW_{jp}^2. \end{pmatrix} \end{aligned}$$

Thus, as long as $EW_{ji} = 0$ and $EW_{ji}^2 = 1/d$, we have $E(W_j W_j^T) = I_p/d$.

Using i.i.d. W_{ji}. Since there is no other information about a, it is intuitive to use i.i.d. W_{ji} which is sub-Gaussian with $EW_{ji} = 0$ and $EW_{ji}^2 = 1/d$. Because we need to choose d, we have to isolate it. Therefore, we introduce $K = \sqrt{d}\|W_{ij}\|_{\psi_2}$ (Because we choose i.i.d. W_{ij}, we do not need to take $\max_{i,j}$), then

$$\left\|\sqrt{d} Z_j\right\|_{\psi_2}^2 = d\left\|Z_j\right\|_{\psi_2}^2 \le d\left(C\|W_{ij}\|_{\psi_2}^2\|a\|_2^2\right) = CK^2\|a\|_2^2, \quad \forall j = 1, \dots, d.$$

By property 7, we have

$$P\left(|\|Z\|_2 - \|a\|_2| \geq \epsilon\|a\|_2^2\right) \leq 2\exp\left(-\frac{c\epsilon^2\|a\|_2^4}{K^4\|a\|_2^4}d\right) = 2\exp\left(-\frac{c\epsilon^2}{K^4}d\right).$$

Thus, if we choose $d \geq C/\epsilon^2$, we could make this probability very small.

Remark 4.1 *It is unfortunate that we could not use* $\epsilon\|a\|_2$ *in place of* $\epsilon\|a\|_2^2$*. One could get around this issue by scaling down* x_i *such that* $\|a\|_2 \leq 1$ *for any pair of* i, j*. In this way,* $\|a\|_2 \geq \|a\|_2^2$ *and*

$$P\left(|\|Z\|_2 - \|a\|_2| \geq \epsilon\|a\|_2\right) \leq P\left(|\|Z\|_2 - \|a\|_2| \geq \epsilon\|a\|_2^2\right).$$

From fixed i, j to every pair. We have proved that the probability of violating Equation 4.2 for a given i, j decays exponentially. Since there are only $n(n-1)/2$ pairs, we could use a union bound to get the following probability bound of violation for any pair,

$$n(n-1)\exp\left(-\frac{c\epsilon^2}{K^4}d\right) \leq \exp(2\ln(n) - \frac{c\epsilon^2}{K^4}d).$$

Thus, if we choose $d = C\ln(n)/\epsilon^2$, we could have the above probability minimal.

Independent of p. Notice that d has nothing to do with p. Thus, it is possible that $d \ll p$ which means an extremely efficient dimension reduction that almost keeps the distance. This has been used in Chiong and Shum (2019) to reduce the size of a large choice set.

Choice of W_{ji}. Theoretically speaking, $N(0, 1/d)$ is a natural choice. However, it requires lots of computational power when calculating $W_j^T a$ when p is large. Li et al. (2006) and Chiong and Shum (2019) use the following distribution ($s \geq 1$)

$$\sqrt{d}W_{ji} = \sqrt{s}\begin{cases} 1 & \text{with probability } 1/2s \\ 0 & \text{with probability } 1 - 1/s \\ -1 & \text{with probability } 1/2s. \end{cases}$$

It is easy to have $EW_{ji} = 0$ and $EW_{ji}^2 = 1/d$. Since it is bounded, it is also sub-Gaussian.

Compare its $\|\cdot\|_{\psi_2}$ with $N(0,1)$'s. Will you choose s super large?

References

Chiong KX, Shum M (2019) Random projection estimation of discrete-choice models with large choice sets. *Management Science* 65(1):256–271.

Li P, Hastie TJ, Church KW (2006) Very sparse random projections. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 287–296.

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 5: High Probability Upper Bound of $\|A\|_{op}$

Lecturer: Long Zhao, longzhao@nus.edu.sg

Largest eigenvalue in ecology. May (1972). Counting equilibria in complex systems via random matrices

5.1 Resources

- Wainwright (2019, Chapter 4.1, Chapter 6). Motivation of infinite events and covariance matrix estimation.
- Vershynin (2018, Chapter 4).
- Terence Tao's Talk on Random Matrix. Unlike Vershynin (2018), he explains the covering idea exceptionally well.
- StatQuest: Principal Component Analysis (PCA). Simple introduction of PCA.
- 3Blue1Brown: Eigenvectors and eigenvalues. Simple introduction of eigenvectors and eigenvalues.

5.2 Objective

Here is the summary of what we have known.

1. Sum of independent sub-Gaussian or sub-exponential random variables concentrates. This is about the probability of **one** event.
2. We could use union bound to the probability of **finite** events. We have done this in our construction of almost orthogonal vectors and proof of Johnson-Lindenstrauss lemma.

This lecture's objective is to shed some light on how do we use concentration to bound the probability of infinite events. Before we proceed to solving the issue of infinite events, let us explore why it is important in the learning setting.

5.2.1 Why Infinite Events?

Assume that we are given n samples $\{X_i\}_{i=1}^n$ drawn i.i.d. according to a distribution $P_{\theta^\star}$, for some fixed but unknown $\theta^\star \in \Omega$. We could choose a cost function $L_\theta(X)$ that measures the 'fit' between a parameter θ and the sample X. The principle of empirical risk minimization is based on minimizing

$$\hat{R}_n(\theta, \theta^\star) \triangleq \frac{1}{n}\sum_{i=1}^{n} L_\theta(X_i).$$

This quantity is known as the empirical risk. The connection to $\theta^\star$ is through the samples X_1^n. Since X_1^n are random, the empirical risk is also random. One simple example of $L_\theta(X)$ is least-squares loss

$$L_\theta(y, x^T) = (y - x^T\theta)^2,$$

where $X = (y, x^T)$ and $y = x^T\theta^\star + \epsilon$. We naturally care about the expectation of the empirical risk, which is called the population risk,

$$R(\theta, \theta^\star) \triangleq E_{\theta^\star}\left(L_\theta(X)\right).$$

Say the parameter we get by minimizing empirical risk is $\hat{\theta} \in \Omega_0 \subset \Omega$, we are curious about how large is the excess risk, defined as

$$E(\hat{\theta}, \theta^\star) \triangleq R(\hat{\theta}, \theta^\star) - \inf_{\theta\in\Omega_0} R(\theta, \theta^\star).$$

For simplicity, assume that there exists a $\theta_0 \in \Omega_0$ such that $R(\theta_0, \theta^\star) = \inf_{\theta\in\Omega_0} R(\theta, \theta^\star)$. **By law of large numbers, θ_0 is the parameter chosen with infinite data points.** Then we could decompose the excess risk as

$$E(\hat{\theta}, \theta^\star) = \left(R(\hat{\theta}, \theta^\star) - \hat{R}_n(\hat{\theta}, \theta^\star)\right) + \left(\hat{R}_n(\hat{\theta}, \theta^\star) - \hat{R}_n(\theta_0, \theta^\star)\right) + \left(\hat{R}_n(\theta_0, \theta^\star) - R(\theta_0, \theta^\star)\right).$$

Since $\hat{\theta}$ minimizes empirical risk, we know the second term is non-positive. The third term is relatively easier than the first since θ_0 is fixed while $\hat{\theta}$ is random (why?). One way to handle the first term is to bound it by

$$\sup_{\theta\in\Omega_0}\left|R(\theta, \theta^\star) - \hat{R}_n(\theta, \theta^\star)\right|,$$

which is almost always (uncountable) infinite events.

5.2.2 How to Handle Infinite Events?

The idea of solving infinite events is to transform infinite into finite ones somehow. To do this, we need to use the idea of covering. You might find this idea of converting from infinite to finite familiar. In fact, the open cover definition of compactness states that any open cover has a finite version. What we want to use is a quantitative version of it.

Unlike previous lectures, I will not state the explicit assumptions first and then have rigorous proof. Instead, as the analysis advances, I will try to find the conditions we could leverage.

This is a more practical setting in research where you try to prove something with your own conditions. The unfortunate consequence is that we will have different assumptions from Vershynin (2018) and slightly different conclusions. However, I think the flow is more natural this way.

With a high probability upper bound of $\|A\|_{op}$, we could also understand the behavior of $\|\Sigma - \hat{\Sigma}\|_{op}$. Here Σ is the true covariance matrix and $\hat{\Sigma}$ is the sample covariance matrix. That is to say, with n observations, how good is the sample covariance matrix. This analysis plays a major role in minimum-variance portfolio optimization which tries to solve the following optimization

$$\begin{aligned} \min_{w} \quad & w^T \hat{\Sigma} w \\ \text{subject to} \quad & w^T \mathbf{1} = 1, \end{aligned}$$

where w is the portfolio weight and $w^T\mathbf{1} = 1$ means that one has to put all money in the market.

5.3 Preliminary on Matrices

The target is to introduce the operator norm of an $n \times p$ matrix, denoted as $\|\cdot\|_{op}$. Along the way, we will discuss about singular value decomposition, which has a close connection with $\|\cdot\|_{op}$.

5.3.1 Singular value decomposition

Let A be an $n \times p$ matrix. You could think A as a data matrix with n observations and each one has p dimensions. We could represent A in the following form

$$A = \sum_{i=1}^{\min(n,p)} s_i u_i v_i^T,$$

where u_i is the ith eigenvector of AA^T; v_i is the ith eigenvector of A^TA; s_i is the ith singular value which is equal to $\sqrt{\lambda_i}$ where λ_i is the ith eigenvalue of A^TA or AA^T. For simplicity, we rank $s_1 \geq s_2 \geq \ldots (\geq 0)$, and we assume $s_i = 0, \min(n,p) < i \leq \max(n,p)$.

Sometimes, the following matrix form is more convenient:

$$A = UDV^T,$$

where $U = (u_1, \ldots u_n)$, $V = (v_1, \ldots v_p)$, D is an $n \times p$ matrix with $D_{ii} = s_i$ and others being 0. Moreover, U and V are orthogonal matrices ($UU^T = U^TU = I_n$ and $VV^T = V^TV = I_p$) in $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively.

Now, it is easy to see

$$\begin{aligned} AA^T &= UDD^TU^T = UDiag_{n\times n}(s_i^2)U^T \\ A^TA &= VD^TDV^T = VDiag_{p\times p}(s_i^2)V^T. \end{aligned}$$

5.3.2 $\|A\|_{op}$

If we denote space $\mathbb{R}^p$ ($\mathbb{R}^n$) with Euclidean norm $\|\cdot\|_2$ as l_2^p (l_2^n), we could see A as an operator from l_2^p to l_2^n. Now, we could definite $\|A\|_{op}$

$$\|A\|_{op} = \max_{x\in\mathbb{R}^p/0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\in S_{p-1}} \|Ax\|_2, \tag{5.1}$$

where S_{p-1} is the unit sphere in $\mathbb{R}^p$ ($x \in S^{p-1} \Leftrightarrow \|x\|_2 = 1$). Intuitively speaking, $\|A\|_{op}$ measures the maximum length ratio between the vector after transformation and before. Noticing that $\|Ax\|_2 = \max_{y\in S^{n-1}} < Ax, y >$ (Cauchy-Schwarz inequality), we also have

$$\|A\|_{op} = \max_{x\in S^{p-1}, y\in S^{n-1}} < Ax, y > .$$

If we want to bound $\|A\|_{op}$, we need to deal with uncountable many events because S_{p-1} has uncountable many points. This is exactly why we want to handle it. [Here is a technical issue. We need $\{\|A\|_{op} \le x\}$ to be a measurable event. In other words, $\|A\|_{op}$ should be a random variable first, and then we could talk about probability. It is not a issue here because S^{p-1} is compact, there exists a point $x^\star$ that $\|A\|_{op} = \|Ax^\star\|_2$. Namely, $\|A\|_{op}$ is fundamentally "one event" about some norm $\|Ax^\star\|_2$. Unfortunately, we do not know where $x^\star$ is.]

Let we establish a natural connection between $\|A\|_{op}$ and s_1.

$$\|A\|_{op} = s_1.$$

To prove it conceptually, we shall decompose A as UDV^T. Because U and V are orthogonal matrices, they will not change the length ratio (they are just rotations). Thus, the largest change is s_1, which is the largest singular value.

5.4 Analysis of Bounding $\|A\|_{op}$

5.4.1 One Event: Fixed x

Let us first make sure that we know how to bound a single event from Equation 5.1. Namely, for a fixed $x \in S^{p-1}$, could we bound $P(\|Ax\|_2 > t)$? This naturally links to the concentration of the norm (property 7), which requires each coordinate to be independent, sub-Gaussian, with the same second moment. Let us write each coordinate explicitly.

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix} x = \begin{pmatrix} a_1^T x \\ \vdots \\ a_n^T x \end{pmatrix}$$

- To have **independence of coordinates**, we must have a_i independent. That is to say, we assume each row of data is independent, which is a natural assumption for the data generating process.
- To have **sub-Gaussian of coordinates for all** x, all we know now is to have each element of a_i independent sub-Gaussian. This is a very strong assumption. Luckily, we could relax this assumption to the sub-Gaussian random vectors.

  Definition 5.1 (Sub-Gaussian Random Vectors, Vershynin (2018) Definition 3.4.1) *A random vector $X \in \mathbb{R}^p$ is called sub-Gaussian if the one-dimensional marginals $< X, x >$ are sub-Gaussian random variables for all $x \in \mathbb{R}^p$. The sub-Gaussian norm of X is defined as*

  $$\|X\|_{\psi_2} = \sup_{x \in S^{p-1}} \| < X, x > \|_{\psi_2}.$$

  Although this definition is motivated by the fact that if $X \sim N(\mu, \Sigma)$, then $< X, x > \sim N(\mu^T x, x^T \Sigma x)$, it also serves us nicely.

- To have **same second moment**, we must have

  $$x^T \left(E a_i a_i^T \right) x = E(a_i^T x)^2 = E(a_j^T x)^2 = x^T \left(E a_j a_j^T \right) x$$

  holds for all $x \in S^{p-1}$ and i, j pairs. Thus, we must have

  $$E a_i a_i^T = E a_j a_j^T \quad \forall 1 \le i, j \le p.$$

  That is to say, we could have a_i and a_j following different distributions, but they need to share certain moment information. This is slightly weaker than the i.i.d. assumption for the data generating procedure.

  Because $x \in S^{p-1}$ implies $x^T x = 1$, it is tempting to transform different cases into the one with $E a_i a_i^T = I_p$. In fact, this is possible. Say $E b_i b_i^T = \Sigma$, then

  $$E \left(\Sigma^{-1/2} b_i \right) \left(\Sigma^{-1/2} b_i \right)^T = I_p.$$

  In this way, we could establish results for $\Sigma^{-1/2} b_i$ and then convert it back to b_i. Noticing the importance of random vectors with $E a_i a_i^T = I_p$, we have the following definition.

  Definition 5.2 (Isotropic random vectors, Vershynin (2018) Definition 3.2.1) *A random vec-*

tor $X \in \mathbb{R}^p$ is called isotropic if

$$EXX^T = I_p.$$

Now, let's summarize the conditions we have for A, each row a_i is **independent sub-Gaussian isotropy random vector**. Now, from the concentration of the norm we have

$$\|\|Ax\|_2 - \sqrt{n}\|_{\psi_2} \leq C \max_i \|a_i\|_{\psi_2} \triangleq CK^2, \text{ where } K = \max_i \|a_i\|_{\psi_2}.$$

Namely, we could bound $P(|\|Ax\|_2 - \sqrt{n}| > t)$ easily now.

5.4.2 Covering: from Infinite to Finite.

We want to answer the question that is it possible to find a finite set $\mathcal{N}$ such that

$$\max_{x \in S^{p-1}} \|Ax\|_2 \leq C \max_{x \in \mathcal{N}} \|Ax\|_2.$$

Because of compactness of S^{p-1}, we know $\|A\|_{op} = \|Ax^\star\|_2$ for some $x^\star \in S^{p-1}$. To have the bound above, all we need is to approach $x^\star$ as close as possible. Unfortunately, we have no idea where $x^\star$ is. Then the brute-force solution is to approach any point of S^{p-1} with certain error ϵ. Namely,

$$\forall x \in S^{p-1}, \ \exists y(x) \in \mathcal{N}, \text{ such that } \|x - y(x)\|_2 \leq \epsilon.$$

We call $\mathcal{N}$ that satisfies this property the ϵ-net (of S^{p-1}) and call the smallest cardinal of ϵ-net the covering number, denoted as $N(S^{p-1}, \epsilon)$ (Sometimes, we also need to specify the corresponding distant measure d. Here it is the Euclidean distance.) Now, we have

$$\|Ax\|_2 \leq \|Ay(x)\|_2 + \|A(x - y(x))\|_2 \leq \|Ay(x)\|_2 + \epsilon\|A\|_{op}.$$

(It is worth noticing that we might not be able to bound the difference term of $x - y(x)$ universally in other situations. Then we need to adopt a more advanced technique of covering: chaining to get around this issue. We will talk about this technique in our proof of the feasibility of learning.)

Taking the maximization with respect to $x \in S^{p-1}$ (think about why we could use $\max_{x \in \mathcal{N}} \|Ax\|_2$ instead of $\max_{x \in \mathcal{N}} \|Ay(x)\|_2$) to have

$$\|A\|_{op} \leq \max_{x \in \mathcal{N}} \|Ax\|_2 + \epsilon\|A\|_{op} \Rightarrow \|A\|_{op} \leq \frac{1}{1-\epsilon} \max_{x \in \mathcal{N}} \|Ax\|_2.$$

We almost achieved our goal of transforming infinite events into finite events. The only thing left is how large is $N(S^{p-1}, \epsilon)$ which we try to bound next.

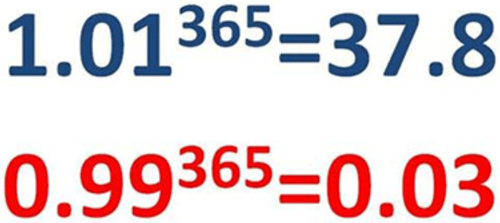
Let $\mathcal{P}$ be the maximal ϵ-separated subset of S^{p-1} that the distance between each pairs is larger than ϵ. Then

$\mathcal{P}$ is an ϵ-net of S^{p-1}, otherwise, it is no longer the maximal ϵ-separated subset (think!). If we draw a ball with $\epsilon/2$ at each point of $\mathcal{P}$, we know that those balls are disjoint. Meanwhile, the distance between any point from those balls and the origin is at most $1+\epsilon/2$. That is to say, those balls are contained in a bigger ball with radius $1+\epsilon/2$. Thus, the total volumes of those balls is smaller than the volume of $B(1+\epsilon/2)$,

$$N(S^{p-1},\epsilon) \le |\mathcal{P}| \le \frac{(1+\epsilon/2)^p}{(\epsilon/2)^p} = (\frac{2}{\epsilon}+1)^p,$$

which is exponential in p. Try to think about two things.

1. Why it does not matter even the number of events are exponential in p?
2. We could achieve a tighter bound using the fact that all balls are excluded from the ball with radius $1-\epsilon/2$. Why don't we use the tighter bound instead? Hint: think about $(1.01)^p - (0.99)^p$ for a very large p (the figure below feels familiar?). Following this, could you prove that almost all volume of a unit ball in high dimension concentrates around on the sphere? Isn't it amazing!



$$1.01^{365}=37.8$$

$$0.99^{365}=0.03$$

5.4.3 Take $\epsilon = 1/2$

By taking $\epsilon = 1/2$, we have

$$\|A\|_{op} \le 2\max_{\mathcal{N}} \|Ax\|_2, \quad N(S^{p-1},\frac{1}{2}) \le 5^p.$$

Then we have the following bound

$$\begin{aligned} P\left(\|A\|_{op} > 2\sqrt{n}+t\right) &\le P\left(\max_{x\in\mathcal{N}} \|Ax\|_2 \ge \sqrt{n}+t/2\right) \\ &\le N(S^{p-1},\frac{1}{2}) \times \exp\left(-c\frac{t^2}{K^4}\right) \\ &\le 5^p \exp\left(-c\frac{t^2}{K^4}\right). \end{aligned}$$

By choosing $t \ge \tilde{C}\max_i \|a_i\|_{\psi_2}\sqrt{p}$ with large enough $\tilde{C}$, we could make the probability above very small.

Thus,

$$\|A\|_{op} \leq 2\sqrt{n} + \tilde{C}K^2\sqrt{p}$$

with high probability when $\tilde{C}$ is large enough.

Remark 5.3 *One could make the probability bound explicit by let $t = K^2(C\sqrt{p} + u)$, then*

$$\begin{aligned} c\frac{t^2}{K^4} &\geq c\frac{K^4(Cp + u^2)}{K^4} \qquad ((a+b)^2 \geq a^2 + b^2) \\ &= cCp + cu^2. \end{aligned}$$

Choose C such that $\exp(cCp) \geq 5^p$*, then we have*

$$5^p \exp\left(-c\frac{t^2}{K^4}\right) \leq \exp(-cu^2).$$

5.5 An Elegant Take

The above bound is messy because the coefficient of $\sqrt{n}$ depends our choice of $\epsilon = 1/2$. Next, we will explore a more elegant way of bounding which focus on $\|A^TA/n - I_p\|_{op}$ instead of $\|A\|_{op}$ directly. The motivation is that

$$\|A^TA/n - I_p\|_{op} \leq \frac{L}{n} \Rightarrow \sqrt{n} - \sqrt{L} \leq s_i(A) \leq \sqrt{n} + \sqrt{L}.$$

That is to say, we will obtain a high-probability interval for all singular values.

5.5.1 One Event: Fixed x

Let $B = A^TA/n - I_p$, then B is a symmetric matrix and we have the following property[1] regarding $\|B\|_{op}$

$$\|B\|_{op} = \max_{x \in S^{p-1}} | < Bx, x > |$$

To see the possibility of bounding one event with fixed x, we shall expend $< Bx, x >$

$$< Bx, x >= \frac{1}{n}\|Ax\|_2^2 - 1 = \frac{1}{n}\sum_{i=1}^{n}\left((a_i^Tx)^2 - 1\right).$$

[1] I only know how to prove it using eigenvalue decomposition. If you know an intuitive way, please share with me.

Notice that $E(a_i^T x)^2 = 1$, we now face a similar situation when we try to prove the concentration of the norm. Since $a_i^T x$ is sub-Gaussian, $(a_i^T x)^2$ is sub-exponential. Thus, we shall bound $\|(a_i^T x)^2 - 1\|_{\psi_1}$:

$$\begin{aligned}\left\|(a_i^T x)^2 - 1\right\|_{\psi_1} &\leq C\|(a_i^T x)^2\|_{\psi_1} \quad \text{(centering)} \\ &= C\|a_i^T x\|_{\psi_2}^2 \quad (\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2) \\ &\leq C\|a_i\|_{\psi_2}^2 \quad \text{(definition of } \|a_i\|_{\psi_2}) \\ &\leq CK^2 \quad (K = \max_i \|a_i\|_{\psi_2}).\end{aligned}$$

Thus, we could apply Bernstein's inequality to bound one event.

5.5.2 Covering: from Infinite to Finite

Say $\|x - y(x)\|_2 \leq \epsilon$ and $y(x) \in S^{p-1}$, then

$$\begin{aligned}< Bx, x > &=< B(x - y(x)), x > + < By(x), x > \\ &=< B(x - y(x)), x > + < By(x), x - y(x) > + < By(x), y(x) > \\ &\leq \|B\|_{op}\epsilon + \|B\|_{op}\epsilon + < By(x), y(x) >\end{aligned}$$

For both sides, take maximization with respect to $x \in S^{p-1}$ to have

$$\begin{aligned}\|B\|_{op} &\leq 2\epsilon\|B\|_{op} + \max_{x\in\mathcal{N}} < By(x), y(x) > \\ \Rightarrow \|B\|_{op} &\leq \frac{1}{1-2\epsilon} \max_{x\in\mathcal{N}} < By(x), y(x) >\end{aligned}$$

By choosing $\epsilon < 1/2$, it is possible to convert infinite events into finite ones.

5.5.3 $\epsilon = 1/4$

By choosing $\epsilon = 1/4$, we have

$$\|B\|_{op} \leq 2 \max_{x\in\mathcal{N}} < By(x), y(x) >,$$

and $N(S^{p-1}, 1/4) \leq 9^p$. Then we have

$$\begin{aligned}P(\|B\|_{op} \geq \epsilon) &\leq P\left(\left(\max_{x\in\mathcal{N}} < By(x), y(x) >\right) \leq \frac{\epsilon}{2}\right) \\ &\leq 9^p P\left((< By(x), y(x) >) \geq \frac{\epsilon}{2}\right) \\ &\leq 9^p \exp\left(-c_1 \min(\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2})n\right).\end{aligned} \tag{5.2}$$

The following are some 'weird' techniques that focuses on relating $\|B\|_{op}$ and $s_i(A)$. Do not panic if you do not understand why it is set up this way.

Since $B = A^T A/n - I_p$, we have

$$\begin{aligned}\|B\|_{op} &= \max(s_1(A^T A/n) - 1, 1 - s_p(A^T A/n)) \quad (\text{Two cases to achieve } \max_{x \in S^{p-1}} \|Bx\|_2)\\ &= \max\left((s_1(A)/\sqrt{n})^2 - 1, 1 - (s_p(A)/\sqrt{n})^2\right) \quad (s_i(A^T A) = s_i(A)^2)\\ &\geq |(s_i(A)/\sqrt{n})^2 - 1| \quad (s_p(A) \leq s_i(A) \leq s_1(A))\end{aligned}$$

Denote $Z = s_i(A)/\sqrt{n}$. Since we have established some high-probability bound for $\|B\|_{op}$ which implies one for $|Z^2 - 1|$, we need to somehow transform it into bound of Z. This shall remind you about the following numeric inequality for $z \geq 0$,

$$|z - 1| \geq \delta \Rightarrow |z^2 - 1| \geq \max(\delta, \delta^2).$$

Namely, we will set $\epsilon = K^2 \max(\delta, \delta^2)$, then

$$\begin{aligned}P(|Z - 1| \geq K^2\delta) &\leq P\left(|Z^2 - 1|_{op} \geq K^2 \max(\delta, \delta^2)\right)\\ &\leq P(\|B\|_{op} \geq K^2 \max(\delta, \delta^2))\\ &\leq 9^p \exp\left(-c_1 \min(\max(\delta, \delta^2), \max(\delta, \delta^2)^2) n\right)\\ &= 9^p \exp\left(-c_1 \delta^2 n\right) \quad (\text{it is true for both } \delta \geq 1 \text{ and } \delta < 1).\end{aligned}$$

In order to have the bound small enough, we need to have

$$\delta = \tilde{C}\sqrt{\frac{p}{n}},$$

for large enough $\tilde{C}$. Thus, we have shown that

$$\sqrt{n} - \tilde{C}K^2\sqrt{p} \leq s_i(A) \leq \sqrt{n} + \tilde{C}K^2\sqrt{p}$$

with high probability.

Btw, one could also make the probability bound explicit by using the same trick as in Remark 5.3. Try it out!

5.5.4 IF $n \gg p$

We could have the following properties for $A/\sqrt{n}$,

$$1 - \tilde{C}K^2\sqrt{\frac{p}{n}} \leq s_i(\frac{1}{\sqrt{n}}A) \leq 1 + \tilde{C}K^2\sqrt{\frac{p}{n}}.$$

When $n \gg p$, we have all singular values are close to 1 with high probability. This means that for any $x \in \mathbb{R}^p$,

$$\left\| \frac{1}{\sqrt{n}} Ax \right\|_2 \approx \|x\|_2.$$

Do we accidentally find a transformation that keeps pairwise distance almost the same for any number of points? Is it stronger than the Johnson-Linderstrauss Lemma? Unfortunately, it is not stronger because it is inflating dimension from p to n ($n \gg p$)

5.5.5 From $a_i = \Sigma^{-1/2} b_i$ to b_i

Because we need isotropic random vectors to have concentration, we need to transform b_i into $a_i = \Sigma^{-1/2} b_i$ to have the high-probability bound. Now, let us investigate the original matrix $L = (b_1, \dots b_n)^T = \Sigma^{1/2} A$.

$$\|L\|_{op} = \|\Sigma^{1/2} A\|_{op} \leq \|\Sigma^{1/2}\|_{op} \|A\|_{op} = \|\Sigma\|_{op}^{1/2} \|A\|_{op}$$

This means that we have

$$\|L\|_{op} \leq \|\Sigma\|_{op}^{1/2} (\sqrt{n} + \tilde{C} K^2 \sqrt{p})$$

with high proability.

5.6 Application in Covariance Estimation

If we treat b_i as the ith observation of demeaned data, then

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} b_i b_i^T$$

is the sample covariance matrix. Because the true covariance matrix is Σ, then

$$\frac{1}{n} \sum_{i=1}^{n} b_i b_i^T - \Sigma = \hat{\Sigma} - \Sigma$$

is the estimation error of using sample covariance matrix to approximate the true covariance matrix. We care about how large the estimation error could be, namely $\|\hat{\Sigma} - \Sigma\|_{op}$. Since

$$\|\hat{\Sigma} - \Sigma\|_{op} \leq \|\Sigma\|_{op}^{1/2} \|\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I_p\|_{op} \|\Sigma\|_{op}^{1/2} = \|\Sigma\|_{op} \|B\|_{op}$$

Based on Equation 5.2, we have

$$\begin{aligned} P\left(\|\hat{\Sigma} - \Sigma\|_{op} \geq \epsilon \|\Sigma\|_{op}\right) &\leq P(\|B\|_{op} \geq \epsilon) \\ &\leq 9^p \exp\left(-c_1 \min(\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2}) n\right). \end{aligned}$$

We usually want to have a small ϵ and we could assume $\epsilon^2/K^4 \leq 1$ (namely, we are in the regime of small deviation), then we have

$$P\left(\|\hat{\Sigma} - \Sigma\|_{op} \geq \epsilon \|\Sigma\|_{op}\right) \leq 9^p \exp\left(-c_1 \frac{\epsilon^2}{K^4} n\right).$$

This means that if we choose $n = CK^4 p/\epsilon^2$ with large enough C, we could have the above probability very small. This means that for a large $p \times p$ matrix Σ, only $\Theta(p)$ observations are needed to get a good estimation.

5.6.1 Minimum-Variance Portfolio

The famous mean-variance portfolio optimization Markowitz (1952) proposes that the investor should make a tradeoff between benefit (expected return) and risk (variance). Mathematically speaking, one should solve the following optimization

$$\begin{aligned} \min_w \quad & w^T \Sigma w \\ \text{subject to} \quad & w^T \mu \geq \rho_{target} \\ & w^T \mathbf{1} = 1, \end{aligned}$$

where Σ is the covariance matrix, and μ is the expected return. Intuitively, among the portfolios that have a larger expected return than ρ_{target}, one should prefer the one with minimum variance (lowest risk). The biggest issue is that we do not know Σ and μ, which could only be estimated using data. μ is extremely hard to estimate. To get a sense of this, try to predict the price of GameStop on the next Monday. As stated in Jagannathan and Ma (2003),

"*The estimation error in the sample mean is so large that nothing much is lost in ignoring the mean altogether when no further information about the population mean is available.* "

Although I do not fully agree with this statement, but it is a good starting point to believe that most estimations of μ must contain a huge estimation error. If so, then the constraint of expected return is misleading and one choice is to ignore this constraint entirely to have

$$\begin{aligned} \min_w \quad & w^T \Sigma w \\ \text{subject to} \quad & w^T \mathbf{1} = 1. \end{aligned} \tag{5.3}$$

This is the minimum-variance optimization, which becomes quite popular nowadays. Another route to obtain minimum-variance optimization is to assume all stocks have the same expected return, namely $\mu \propto \mathbf{1}$ which renders the expected return constraint redundant.

Now, our analysis regarding $\hat{\Sigma} - \Sigma$ comes handy because we could bound $w^T \Sigma w$, the true variance of portfolio w, in the following way,

$$w^T \Sigma w = w^T (\Sigma - \hat{\Sigma}) w + w^T \hat{\Sigma} w \leq \|w\|_2^2 \|\Sigma - \hat{\Sigma}\|_{op} + w^T \hat{\Sigma} w. \tag{5.4}$$

We will proceed the analysis in three different scenarios.

1. $\|\Sigma - \hat{\Sigma}\|_{op}$ is negligible.
2. $\|\Sigma - \hat{\Sigma}\|_{op}$ is relatively small compared to $\|\Sigma\|_{op}$.

Case 1. In this case, we might want to minimize $w^T \hat{\Sigma} w$ as a proxy for minimizing $w^T \Sigma w$. Denote the corresponding portfolio $\hat{w}^\star$ and the true minimum-variance portfolio, $w^\star$. On average, we will be still disappointed because the expected in-sample variance will be smaller than the true variance as proved below

$$\begin{aligned}
(\hat{w}^\star)^T \hat{\Sigma} \hat{w}^\star &\leq (w^\star)^T \hat{\Sigma} w^\star && \text{(Definition of } \hat{w}^\star\text{)} \\
E\left((\hat{w}^\star)^T \hat{\Sigma} \hat{w}^\star\right) &\leq E\left((w^\star)^T \hat{\Sigma} w^\star\right) && \text{(Both sides take expectation)} \\
&= (w^\star)^T \Sigma w^\star && (E\hat{\Sigma} = \Sigma) \\
&\leq (\hat{w}^\star)^T \Sigma \hat{w}^\star && \text{(Definition of } w^\star\text{)}
\end{aligned}$$

But Equation 5.4 guarantees that such disappointment will not be very large. Unfortunately, this case is quite rare in portfolio optimization with large number of stocks (p is large).

Case 2. A common case is $n = 252$ (one-year daily return) and $p = 100$ (100 stocks). In this case, based on our analysis above, we will have $\|\Sigma - \hat{\Sigma}\|_{op}$ relatively small compared to $\|\Sigma\|_{op}$. If we only minimize $w^T \hat{\Sigma} w$, its optimal value might be comparable by the first term, $\|w\|_2^2 \|\Sigma - \hat{\Sigma}\|_{op}$, rendering the resulting portfolio untrustworthy. Thus, it is tempting to try

$$\begin{aligned}
\min_{w} \quad & w^T \hat{\Sigma} w + \lambda \|w\|_2^2 \\
\text{subject to} \quad & w^T \mathbf{1} = 1,
\end{aligned}$$

with λ to be chosen[2]. In this way, $w^T \hat{\Sigma} w + \lambda \|w\|_2^2$ serves as a proxy for the true upper bound. It also has some similarity with ridge regression. Thus, one might naturally want to replace $\|w\|_2^2$ with $\|w\|_1$ to obtain a LASSO-like portfolio.

[2] One could choose λ by cross-validation

If we stick with $n = 252$ but increase p to 500, then $\hat{\Sigma}$ is not full rank. This means that $(\hat{w}^\star)^T\hat{\Sigma}\hat{w}^\star = 0$. This makes $\hat{w}^\star$ not a meaningful choice at all. One could proceed with large λ to compensate. If we take $\lambda \to \infty$,

$$\begin{aligned} \min_{w} \quad & \|w\|_2^2 \\ \text{subject to} \quad & w^T\mathbf{1} = 1. \end{aligned}$$

By Cauchy inequality we know $1 = w^T\mathbf{1} \leq \|w\|_2\|\mathbf{1}\|_2$. Thus, the above optimization has optimal solution $w = \mathbf{1}/p$. This is the equally-weighted (EW) portfolio which obtains lots of attention since DeMiguel et al. (2009) show that none of 13 sophisticated portfolios could beat it consistently. EW should be your choice if you could not estimate Σ using data at all.

5.6.2 An Optimization Perspective

The estimation perspective of minimum-variance optimization ignores a key feature of the optimization: if one uses $\hat{\Sigma} = 1000\Sigma$, although the variance estimation will be terribly off, one could recover the optimal solution $w^\star$. That is to say, not all estimation errors are equally costly. However, $\|\cdot\|_{op}$ ignores this feature entirely.

Luckily, we could use the analytical solution to Equation 5.3 to see what directions are important.

$$w^\star = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}} \propto \Sigma^{-1}\mathbf{1} = \left(\sum_{i=1}^{p} \frac{1}{\lambda_i} v_i v_i^T\right)\mathbf{1} = \sum_{i=1}^{p} \frac{v_i^T\mathbf{1}}{\lambda_i} v_i,$$

where (v_i, λ_i) is the ith eigenpair of Σ. If we ignore the difference in $v_i^T\mathbf{1}$, the eigenvector with small[3] λ_i has a large effect on the solution. Then the question becomes, could we estimate these eigenvectors well? The following variant of Davis-Kahan theorem (Yu et al. 2015) shows that it depends on the $eigengap_i = \min(\lambda_{i-1} - \lambda_i, \lambda_i - \lambda_{i+1})$.

Theorem 5.4 *Let $(\hat{v}_i, \hat{\lambda}_i)$ be the ith eigenpair of $\hat{\Sigma}$. Then*

$$\|\hat{v}_i - v_i\|_2 \leq \frac{2^{3/2}\|\hat{\Sigma} - \Sigma\|_{op}}{eigengap_i}.$$

Some intuition from the Figure 5.1 might be helpful.

Thus, all we need for v_i with small λ_i is a large $eigengap_i$. Unfortunately, historical data (Figure 5.2) confirm that the opposite (tiny $eigengap_i$ for small eigenvalues) will happen. This means that eigenvectors with (only several) large eigenvalues **must be** well estimated while (lots) eigenvectors with small eigenvalues **could be** terrible estimation. For simplicity, I will call the former the top eigenvectors and the latter the bottom. In the future lectures, under stronger assumptions, I will show that bottom eigenvectors' estimations behave similarly to random directions (uniformly from S^{p-1}), which contains no information about the true

[3] More precisely, λ_i/λ_1 is small because λ_1 will cancel out. Namely, big or small is in relative term. This observation fixes the issue of 1000Σ shares the same solution with Σ

(a) Large height gap: easy to spot.

(b) Small height gap: hard to find.

Figure 5.1: Intuitions for Theorem 5.4

eigenvectors.

My research idea is quite simple. For the top ones, since they are reasonable estimations, then one will treat it as case 1. For the bottom ones, since they could be terrible, then one should be extremely conservative and adopt the EW approach. Then, one uses cross-validation to combine them to obtain the final portfolio. Amazingly, such a simple idea works exceptionally well empirically: it is comparable with the covariance estimated by a fancy high-dimensional statistical model.

Let us come back to $\|\Sigma - \hat{\Sigma}\|_{op}$. Intuitively speaking, for $\|\Sigma - \hat{\Sigma}\|_{op}$, a small error in top eigenvector might dominate a huge error in a bottom eigenvector. Thus, even though we have good high-probability bound for $\|\Sigma - \hat{\Sigma}\|_{op}$, focusing only on it misses an important feature of the problem. Similar argument could be made for robust optimization. If one is concerned about conservativeness in robust optimization, one choice is to disentangle benign and vicious error. In the minimum-variance case, because benign error is much larger in size than vicious error, this disentanglement significantly reduces the size of the uncertainty/ambiguity set without causing any trouble. However, because it is not easy to prove theoretically, it is hard to sell in academia.

Remark 5.5 *The bound provided in Theorem 5.4 might be loose since the left-hand side is naturally bounded by* 2*. However, this does not render the argument meaningless because one could get a much tighter bound using a more complex expression instead of* $\|\hat{\Sigma} - \Sigma\|_{op}$*. Since* $\|\hat{\Sigma} - \Sigma\|_{op}$ *is simpler to remember, it gets very popular. This is the benefit of digging into proofs of theorems.*

Remark 5.6 *With certain strong structure assumption, one might be able to get asymptotically accurate estimation of* Σ *when* $n/p \to c$ *but* $n \to \infty$*. In this case, replacing the estimation in place of* Σ *is a natural choice. Namely, the optimization perspective matters only when the error does not go to* 0*.*

5.7 Sharpe Bounds on Gaussian Matrices [Optional]

From the high probability bound in Section 5.5, we could show that

$$E\|A\|_{op} \leq \sqrt{n} + C\sqrt{p},$$

where C is a constant. If all entries of A is independent $N(0,1)$, then we could show that $C = 1$.

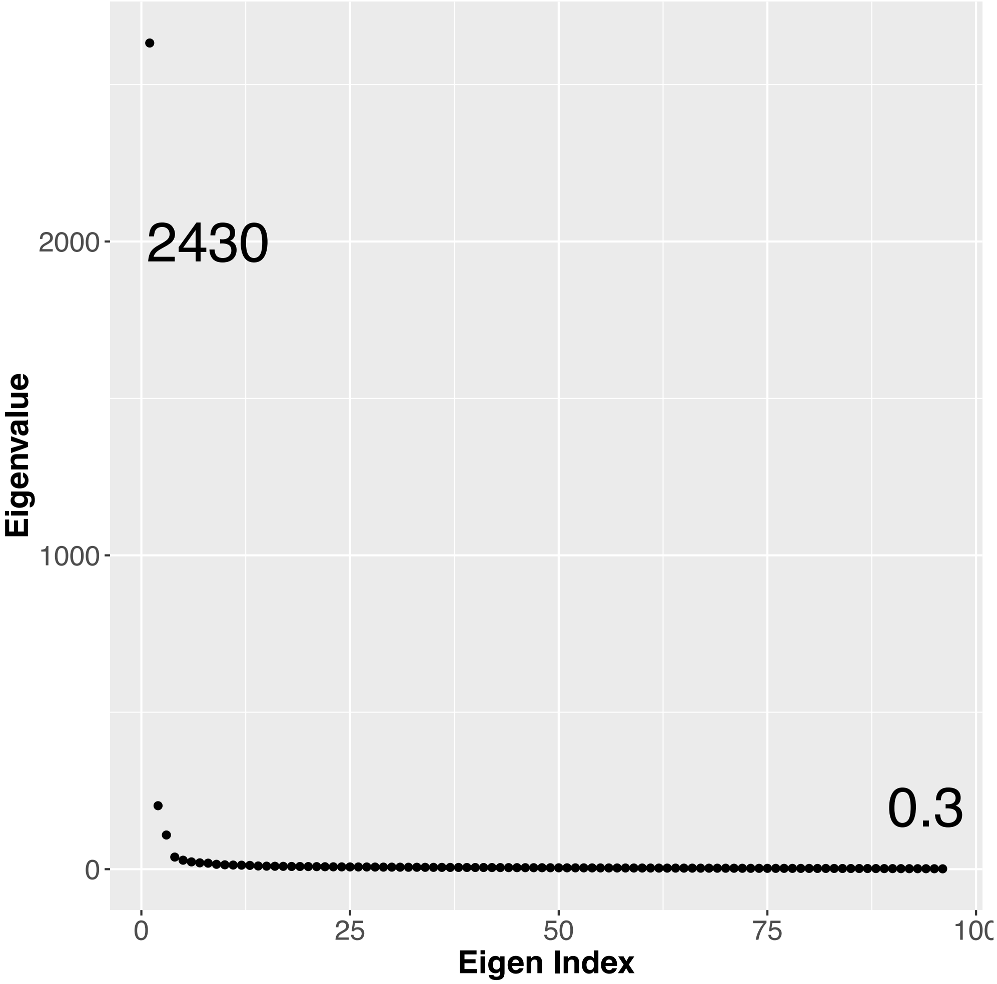


Figure 5.2: Distribution of λ_i

Theorem 5.7 (Norms of Gaussian Random Matrices) *Let A be an $n \times p$ matrix with independent $N(0,1)$ entries. Then*

$$E\|A\|_{op} \leq \sqrt{n} + \sqrt{p}.$$

The proof no longer uses the ϵ-net argument. Instead, it uses Slepian's inequality or more precisely Sudakov-Fernique's inequality (Vershynin 2018, Chapter 7) which is about comparing two Gaussian processes. Roughly speaking, it assures that

$$E\left(\sup_{t\in T} X_t\right) \leq E\left(\sup_{t\in T} Y_t\right),$$

if the two Gaussian processes $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ satisfy

$$\begin{aligned} &EX_t = 0 \quad \& \quad EY_t = 0 \quad \forall t \in T \\ &E(X_t - X_s)^2 \leq E(Y_t - Y_s)^2, \quad \forall t, s \in T. \end{aligned}$$

Unfortunately, we might not be able to cover these theorems in the class. However, let us still try to see why $\|A\|_{op}$ is connected with a Gaussian process.

Since A has Gaussian entries, we know $< Au, v >$ could be written as $X_{(u,v)}$ which is a Gaussian process. Based on the following relationship,

$$\|A\|_{op} = \sup_{u\in S^{p-1}, v\in S^{n-1}} < Au, v >,$$

we know that $\|A\|_{op} = \sup_{u\in S^{p-1}, v\in S^{n-1}} X_{(u,v)}$. This is how they connects.

In fact, we could also have a high-probability bound as following

Corollary 5.8 (Norms of Gaussian Random Matrices: Tails) *Let A be an $n \times p$ matrix with independent $N(0,1)$ entries. Then for every $t \geq 0$, we have*

$$P\left(\|A\|_{op} \geq \sqrt{n} + \sqrt{p} + t\right) \leq 2\exp(-ct^2).$$

The proof is based on the concentration of Lipschitz functions of $N(0, I_p)$ which I will cover in the future lecture borrowing an amazing proof by Maurey and Pisier.

5.8 Bounds for Structured Covariance Matrix [Optional]

To have faster rates for the estimation of covariance matrix, one might impose certain structure on it. Trivially, if we know the covariance matrix is identity matrix, there is no need to estimate it (fastest convergence). If we know the covariance matrix is diagonal, one could prove that the error is of order $\sqrt{(\log p)/n}$ (Wainwright 2019, Exercise 6.15) with new estimator $diag(\hat{\Sigma})$ where $\hat{\Sigma}$ is the sample covariance matrix. Diagonal matrix is a special case of sparse matrix with known non-sparse positions. More generally, if we do not know where the sparsity is, we could use the following estimator

$$T_\lambda(\hat{\Sigma})_{ij} = \begin{cases} \hat{\Sigma}_{ij} & \hat{\Sigma}_{ij} > \lambda \\ 0 & \hat{\Sigma}_{ij} \leq \lambda. \end{cases}$$

Basically, we apply the hard-thresholding operator T_λ to the sample covariance matrix to enforcing sparsity. It could be proved that the rate becomes

$$C\|A\|_{op}\sqrt{\frac{\log p}{n}},$$

where $\|A\|_{op}$ is a measure for sparsity of Σ. For more details, please see Wainwright (2019, Section 6.5).

References

DeMiguel V, Garlappi L, Uppal R (2009) Optimal versus naïve diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies* 22(5):1915–1953.

Jagannathan R, Ma T (2003) Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance* 58(4):1651–1684.

Markowitz H (1952) Portfolio selection. *The Journal of Finance* 7(1):77–91.

May RM (1972) Will a large complex system be stable? *Nature* 238(5364):413–414.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

Yu Y, Wang T, Samworth RJ (2015) A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika* 102(2):315–323.

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 6: A Note On Union Bound

Lecturer: Long Zhao, longzhao@nus.edu.sg

6.1 Resources

- van Handel (2014, Chapter 5). These lecture notes offer lots of deep insights about some techniques. Some chapters of Vershynin (2018) are also based on them.

6.2 Roadmap

Now, the focus of this course is providing information regarding $\sup_{t\in T} X_t$, where T is potentially an uncountable set. An example is

$$\|A\|_{op} = \sup_{t\in S^{p-1}} \|At\|_2,$$

where we could denote $X_t \triangleq At$. Let P be a maximal ϵ-separated set and $\pi(t) \in P$ satisfies $d(\pi(t), t) < \epsilon$. Then we have

$$\begin{aligned} \sup_{t\in T} X_t &= \sup_{t\in T} \left(X_t - X_{\pi(t)} + X_{\pi(t)}\right) \\ &\le \sup_{t\in T} \left(X_t - X_{\pi(t)}\right) + \sup_{t\in P} X_{\pi(t)} \quad (\sup(X+Y) \le \sup X + \sup Y). \end{aligned} \tag{6.1}$$

In the case of $\|A\|_{op}$, we use the above inequality to obtain the optimal upper bound up to a constant. We want to explore why such an easy idea could give an almost sharp bound. More specifically, we want to answer the following questions.

1. Since P is a finite set, we could use union bound to control the second term. Why the union bound works well for $\|A\|_{op}$?
2. In order to give a good bound, $\sup_{t\in T}\left(X_t - X_{\pi(t)}\right)$ should not dominate $\sup_{t\in T} X_t$. For $\|A\|_{op}$, we could show that they are of the same level, which leads to a decent bound.

6.3 When Union Bound Works Well

For the simplest case, let us only consider two events, A_1 and A_2. The union bound tells us that

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2).$$

Clearly, the bound is tight when $A_1 \cap A_2 = \emptyset$ (disjoint) and very loose when $A_1 = A_2$ (an extreme case of overlapping). Another interesting case is that A_1 and A_2 are independent (Is disjoint independent?).

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1)P(A_2) \tag{6.2}$$

Intuitively speaking, when $P(A_i)$ is small, $P(A_1)P(A_2)$ is a higher order term than $P(A_1) + P(A_2)$. Thus, the union bound should be descent. The following exercise provide a more precise lower bound when there are n independent events.

Exercise 6.1 *If $A_1, \ldots, A_n$ are independent events, then*

$$P(\cup_{i=1}^n A_i) \geq (1 - e^{-1}) \left(1 \wedge \sum_{i=1}^n P(A_i)\right)$$

Proof: Since A_is are independent events,

$$P(\cup_{i=1}^n A_i) = 1 - P(\cap_{i=1}^n A_i^c) = 1 - \prod_{i=1}^n P(A_i^c) = 1 - \prod_{i=1}^n (1 - P(A_i)).$$

We could bound $\prod_{i=1}^n (1 - P(A_i))$ as following

$$\begin{aligned}\prod_{i=1}^n (1 - P(A_i)) &\leq \prod_{i=1}^n \exp(-P(A_i)) \quad (1 - x \leq \exp(x)) \\ &= \exp\left(-\sum_{i=1}^n P(A_i)\right) \\ &= \exp\left(-\left[\sum_{i=1}^n P(A_i) \wedge 1\right]\right).\end{aligned}$$

Next, we would like to use the following numeric inequality to finish the proof.

$$h(x) \triangleq 1 - \exp(-x) - (1 - \exp(-1))x \geq 0 \quad \forall x \in [0, 1].$$

It is easy to check $h(0) = h(1) = 0$. Because $h(x)$ is concave, we have the above inequality holds for all $[0, 1]$.

Therefore,

$$1 - \exp\left(-\left[\sum_{i=1}^{n} P(A_i) \wedge 1\right]\right) \geq (1 - \exp(-1))\left(\sum_{i=1}^{n} P(A_i) \wedge 1\right).$$

■

To summarize,

- Lots of overlap: union bound is loose.
- Not so much overlap (disjoint or independent): union bound is descent.

Now we apply this insight to $\sup_{t\in T} X_t$. If X_t is continuous in some sense[1], for $t_1, t_2 \in T$ that $d(t_1, t_2)$ is small, we know two events that $X_{t_1} \leq x$ and $X_{t_2} \leq x$ have lots of overlaps. Thus, union bound is very loose.

If $d(t_1, t_2)$ is large, $X_{t_1} \leq x$ and $X_{t_2} \leq x$ are not closely related (kinda independence). This means that there might be little overlap between those two events. Then a union bound should work well. This is precisely the motivation for utilizing $\sup_{t\in P} X_t$ where any two points in P are at least ϵ separated.

Remark 6.2 *We use independence intuitively instead of rigorously. We never prove that $X_{t_1} \leq x$ and $X_{t_2} \leq x$ are independent when $d(t_1, t_2) \geq \delta$.*

6.4 $\sup_{t\in T}\left(X_t - X_{\pi(t)}\right)$

In the case of $\|A\|_{op}$,

$$\begin{aligned} X_t - X_{\pi(t)} &= \|At\|_2 - \|A\pi(t)\|_2 \leq \|A(t - \pi(t))\|_2 \quad \text{(Triangle Inequality of } \|\cdot\|_2) \\ &\leq \|A\|_{op}\|t - \pi(t)\|_2 \quad \text{(Definition of } \|A\|_{op}). \end{aligned} \tag{6.3}$$

Thus, we have

$$\sup_{t\in T}\left(X_t - X_{\pi(t)}\right) = \epsilon\|A\|_{op} = \epsilon \sup_{t\in T} X_t.$$

We got lucky for $\|A\|_{op}$ because no matter what $\epsilon < 1$ we choose, $\sup_{t\in T}\left(X_t - X_{\pi(t)}\right)$ will not dominate $\sup_{t\in T} X_t$. That is to say, the decomposition of Eq. 6.1 is great: first term could be absorbed; union bound works well for the second term. Thus, we could expecting a sharp bound for $\|A\|_{op}$.

However, lots of the times, we will not have this good luck. In Eq. 6.3, we leverage the $\|A\|_{op}$-Lipschitz property of $\|Ax\|_2$. The Lipschitz constant is the worst case for all possible pairs which is usually very large. van Handel (2014, Example 5.15) provides such an example where $E \sup_{t\in T} X_t \sim n^{-1/2}$ while $\sup_{t\in T}\left(X_t - X_{\pi(t)}\right) \sim n^{-1/3}$. This motivates us to avoid such worst case consideration.

[1] Two possible examples. 1. L-Lipschitz, $|X_t - X_s| \leq Ld(t, s)$. 2. Lipschitz with high probability. $\|X_t - X_s\|_{\psi_2} \leq Cd(t, s)$

References

van Handel R (2014) Probability in high dimension. Technical report, PRINCETON UNIV NJ, URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 7: Symmetrization 1 and Chaining

Lecturer: Long Zhao, longzhao@nus.edu.sg

7.1 Resources

- van Handel (2014, Chapter 5 and 7). Easiest to read among three, especially Chapter 5.
- Vershynin (2018, Chapter 8). Highlight key issues.
- Wainwright (2019, Chapter 4 and 5). Read it after you get the the other two.

7.2 Target

To honor the course name 'Introduction to Data Analytics,' I have to prove an essential theorem[1] in statistical learning theory:

Theorem 7.1 *[Excess Risk via VC Dimension] Assume that the target T is a Boolean function, and the hypothesis space $\mathcal{F}$ is a class of Boolean functions with finite VC dimension $vc(\mathcal{F})$. Then the excess risk is bounded as follows,*

$$E(\hat{f}, T) \leq C\sqrt{\frac{vc(\mathcal{F})}{n}}.$$

Here, we recall the notations from Lecture 5. We are given n samples $\{(X_i, T(X_i))\}_{i=1}^n$ where $X_1^n \triangleq \{X_i\}_{i=1}^n$ are drawn i.i.d. from some distribution P. Our target is to learn the boolean function T based on the samples. One way to do it is to minimize the empirical risk which is defined as

$$\hat{R}_n(f, T) = \frac{1}{n}\sum_{i=1}^{n} L(f(X_i), T(X_i)),$$

where L is some loss function. A popular choice is the squared loss defined as

$$L(x, y) = (x - y)^2.$$

Since we are dealing with Boolean functions f and T, we have $f(X) - T(X) \in \{-1, 0, 1\}$. For these values,

[1] We have seen it without proof in Learning from Data course.

$|x| = x^2$. Thus, we have

$$L(f(X), T(X)) = (f(X) - T(X))^2 = |f(X) - T(X)|.$$

I will stick to $L(\cdot, \cdot)$ without specifying its form unless it is required. We denote the boolean function that minimizes $\hat{R}_n(f, T)$ as $\hat{f}$. The expectation of the empirical risk is the population risk, mathematically speaking,

$$R(f, T) = E(L(f(X), T(X))).$$

The excess risk is defined as

$$E(\hat{f}, T) = R(\hat{f}, T) - \inf_{f \in \mathcal{F}} R(f, T).$$

For simplicity, we assume that there exists $f_0 \in \mathcal{F}$ such that $R(f_0, T) = \inf_{f \in \mathcal{F}} R(f, T)$. In this case, we have already proved that

$$E(\hat{f}, T) \leq 2E\left[\sup_{f \in \mathcal{F}} \left|\hat{R}_n(f, T) - R(f, T)\right|\right].$$

Then we have Theorem 7.1 proved if

$$2E\left[\sup_{f \in \mathcal{F}} \left|\hat{R}_n(f, T) - R(f, T)\right|\right] \leq C\sqrt{\frac{vc(\mathcal{F})}{n}}.$$

7.3 Analysis of $\sup_{f \in \mathcal{F}}$

Usually $|\mathcal{F}|$ is infinite, there is no way that we could use the union bound directly. Naturally, we want to use ϵ-net covering to transform infinite to finite. To determine which distance to use, we want to recall the case of $\|A\|_{op}$. There we use Euclidean norm $\|\cdot\|_2$ because of the following Lipschitz property:

$$\|Ax\|_2 - \|Ay\|_2 \leq \|Ax - Ay\|_2 \leq \|A\|_{op}\|x - y\|_2.$$

It is tempting to explore the possibility of Lipschitz property of two Boolean functions f and g. To simplify the notations, I introduce $L_f(X_i)$ as following,

$$L_f(X_i) \triangleq L(f(X_i), T(X_i)) - E(L(f(X_i), T(X_i))).$$

Then, $\hat{R}_n(f, T) - R(f, T)$ becomes

$$\hat{R}_n(f, T) - R(f, T) = \frac{1}{n}\sum_{i=1}^{n} L_f(X_i).$$

Now, we have

$$\begin{aligned}|\hat{R}_n(f,T)-R(f,T)|-|\hat{R}_n(g,T)-R(g,T)| &= \left|\frac{1}{n}\sum_{i=1}^{n}L_f(X_i)\right|-\left|\frac{1}{n}\sum_{i=1}^{n}L_g(X_i)\right| \\ &\leq \frac{1}{n}\left|\sum_{i=1}^{n}L_f(X_i)-L_g(X_i)\right| \quad (|x|-|y|\leq|x-y|) \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left|L_f(X_i)-L_g(X_i)\right|.\end{aligned}$$

If we use $L(x,y)=|x-y|$, we have

$$\begin{aligned}L(f(X_i),T(X_i))-L(g(X_i),T(X_i)) &= |f(X_i)-T(X_i)|-|g(X_i)-T(X_i)| \\ &\leq |f(X_i)-g(X_i)| \quad (|x|-|y|\leq|x-y|) \\ &\leq \|f-g\|_\infty.\end{aligned}$$

Taking expectation to both sides to have

$$EL(f(X_i),T(X_i))-EL(g(X_i),T(X_i))\leq\|f-g\|_\infty.$$

Use the definition of $L_f(X_i)$ to have

$$L_f(X_i)-L_g(X_i)\leq 2\|f-g\|_\infty,$$

which implies

$$|\hat{R}_n(f,T)-R(f,T)|-|\hat{R}_n(g,T)-R(g,T)|\leq 2\|f-g\|_\infty.$$

Thus, we have 2-Lipschitz regarding norm $\|\cdot\|_\infty$. Unfortunately, for any two different Boolean functions, we have

$$\|f-g\|_\infty=1.$$

This means that any ϵ-net ($\epsilon<1$) under $\|\cdot\|_\infty$ norm contains infinite functions. One way to get around this issue is to use L_2-norm

$$\|f-g\|_{L_2}=\left(E(f(X)-g(X))^2\right)^{\frac{1}{2}}\leq\|f-g\|_\infty.$$

It might be possible that ϵ-net under this smaller norm could be finite. However, we do not know why L_2 norm is the right choice (why not L_1?). Technically, it might not be easy to estimate $|\mathcal{N}(\epsilon)|$ since we have no idea about the distribution of X.

If we step back and think the problem through the n data points, then there are no longer infinite functions:

the **predicted** label of each observation, namely $f(X_i)$, could be 0 or 1, which means that in total, there are at most 2^n possibilities. In other words, we could not distinguish two functions that generate the same labels on the n data points. This leads to infinite many boolean functions collapsing into finite ones. What is the difference between these two perspectives? The former faces the randomness of X_1^n while the latter is conditional on X_1^n.

How to generate conditional expectation in $E\sup_{f\in\mathcal{F}}()$? The only randomness is X_1^n, if we conditional on them, sup becomes deterministic, and we come back to unconditional expectation. We need to introduce some new randomness that makes conditioning on X_1^n possible. Now, we introduce the powerful tool of symmetrization.

7.4 Symmetrization

To motivate the idea of symmetrization (a new origin of randomness), let us think about CLT. Under mild conditions of i.i.d. random variables $\{Z_j\}_{i=1}^n$, we have

$$\sum_{i=1}^n (Z_j - E(Z_j)) \approx O(\sqrt{n}).$$

Why is the summation of n terms only about $\sqrt{n}$? Well, because the positive ones cancel out most of the negative ones (because of independence). Thus, the random sign of values (positive v.s. negative) is helping the CLT. It is tempting to believe that if we introduce independent random signs for each term, CLT still holds[2]. In other words, this kind of new randomness does not break the canceling phenomenon.

We also want to facilitate CLT in the attempt to control $\hat{R}_n(f,T) - R(f,T)$. To see this, we could rewrite it as

$$\hat{R}_n(f,T) - R(f,T) = \frac{1}{n}\sum_{i=1}^n L_f(X_i).$$

Let $\{\epsilon_i\}_{i=1}^n$ be i.i.d. symmetric Bernoulli distribution (or Rademacher distribution) that are also independent to $\{X_i\}_{i=1}^n$. In this way, ϵ_i could represent the random sign of the term $L_f(X_i)$. We want to somehow build the following connection,

$$E_X \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n L_f(X_i)\right| \ ?\leq?\ CE_X E_\epsilon \left(\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right|\right).$$

It is worth to notice that, the RHS is an inequality w.r.t two different randomnesses. In this way, we successfully create the conditional expectation regarding X_1^n which is about finite events! The following Proposition shows that there is indeed such kind of relationship.

[2] $\epsilon_i Z_j$s are i.i.d. and still satisfy mild conditions. Thus, CLT still works. Here, ϵ_is are random signs.

Proposition 7.2

$$E_X \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n L_f(X_i)\right| \leq 2E_X E_\epsilon \left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right|\right] \tag{7.1}$$

Proof: We will start with the following bound,

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n L_f(X_i)\right| &= \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (L_f(X_i) - EL_f(X_i))\right| \quad (EL_f(X_i)=0)\\
&= \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (L_f(X_i) - E_Y L_f(Y_i))\right| \quad (Y_i \sim X_i \text{ but independent})\\
&\leq \sup_{f\in\mathcal{F}} E_Y \left|\sum_{i=1}^n (L_f(X_i) - L_f(Y_i))\right| \quad (|\cdot| \text{ is convex})\\
&\leq E_Y \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (L_f(X_i) - L_f(Y_i))\right| \quad (E\sup \leq \sup E).
\end{aligned} \tag{7.2}$$

Notice that for any two independent copy of the same random variable, Z and $\tilde{Z}$, $Z-\tilde{Z}$ is a symmetric distribution meaning that it has the same distribution as $\tilde{Z}-Z$. Let ϵ be an independent symmetric Bernoulli distribution. Then

$$Z-\tilde{Z} \sim \epsilon(Z-\tilde{Z}).$$

This property implies that $L_f(X_i) - L_f(Y_i)$ shares the same distribution as $\epsilon_i(L_f(X_i) - L_f(Y_i))$. Taking expectation w.r.t. E_X to both sides of Eq. 7.2 to have

$$\begin{aligned}
E_X \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n L_f(X_i)\right| &\leq E_{X,Y} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (L_f(X_i) - L_f(Y_i))\right|\\
&= E_{X,Y,\epsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i(L_f(X_i) - L_f(Y_i))\right| \quad (\epsilon_i(L_f(X_i) - L_f(Y_i)) \sim L_f(X_i) - L_f(Y_i))\\
&\leq E_{X,Y,\epsilon} \sup_{f\in\mathcal{F}} \left[\left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right| + \left|\sum_{i=1}^n \epsilon_i L_f(Y_i)\right|\right] \quad (|x-y| \leq |x|+|y|)\\
&\leq E_{X,\epsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right| + E_{Y,\epsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(Y_i)\right| \quad (\sup(A+B) \leq \sup A + \sup B)\\
&= 2E_X E_\epsilon \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right| \quad (Y \sim X).
\end{aligned}$$

■

Remark 7.3 *Notice that in the proof, $L_f(X)$ could be any function of X. Namely, we do not require that it is a loss function. All we need is the independence among X_is. Think about where do we use such* *independence.*

Remark 7.4 *Given $x_1^n \triangleq (x_1, \dots, x_n)$ and a function class (not necessary Boolean) $\mathcal{F}$. The Empirical Rademacher complexity is defined by*

$$E_\epsilon \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right|.$$

Since $E\epsilon_i = 0$, one could interpret $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i)$ as the sample covariance between ϵ and $f(x)$. Since ϵ_is are i.i.d. Symmetric Bernoulli distribution, there is no true pattern in any realizations. Intuitively speaking, the higher covariance $\mathcal{F}$ could obtain with such pure random realizations of ϵ_is, the larger $\mathcal{F}$ is. Thus, Rademacher complexity is a way to measure how large $\mathcal{F}$ is. Similarly, one could replace ϵ_i with $g_i \sim N(0,1)$ to introduce the Gaussian complexity.

You might wonder that is it possible to lower bound $E_X \sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} L_f(X_i)|$ by its symmetrization counterpart. The following proposition shows that it is doable.

Proposition 7.5

$$\frac{1}{2} E_X E_\epsilon \left[\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i L_f(X_i) \right| \right] \le E_X \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} L_f(X_i) \right|$$

Proof: The core idea is similar to Proposition 7.2, we start with the following inequality.

$$\begin{aligned} \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i L_f(X_i) \right| &= \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i (L_f(X_i) - E_Y L_f(Y_i)) \right| \quad (Y \sim X \;\&\; EL_f(Y_i) = 0) \\ &\le \sup_{f\in\mathcal{F}} E_Y \left| \sum_{i=1}^{n} \epsilon_i (L_f(X_i) - L_f(Y_i)) \right| \quad (|\cdot| \text{ is convex}) \qquad (7.3) \\ &\le E_Y \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i (L_f(X_i) - L_f(Y_i)) \right| \quad (\sup E \le E \sup). \end{aligned}$$

Notice that $L_f(X_i) - L_f(Y_i)$ is symmetric distribution. Thus, $\epsilon_i(L_f(X_i) - L_f(Y_i))$ follows the same distribution as $L_f(X_i) - L_f(Y_i)$. Because of this, we could take expectation with respect to ϵ and X to both sides and then take away ϵ.

$$\begin{aligned}
E_{X,\epsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right| &\leq E_{X,Y,\epsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i (L_f(X_i) - L_f(Y_i))\right| \\
&= E_{X,Y} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (L_f(X_i) - L_f(Y_i))\right| \quad (\epsilon_i(L_f(X_i) - L_f(Y_i)) \sim L_f(X_i) - L_f(Y_i)) \\
&\leq E_{X,Y} \sup_{f\in\mathcal{F}} \left[\left|\sum_{i=1}^n L_f(X_i)\right| + \left|\sum_{i=1}^n L_f(Y_i)\right|\right] \quad (|x-y| \leq |x| + |y|) \\
&\leq E_X \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n L_f(X_i)\right| + E_Y \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n L_f(Y_i)\right| \quad (\sup(A+B) \leq \sup A + \sup B) \\
&= 2E_X \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n L_f(X_i)\right| \quad (Y \sim X)
\end{aligned}$$

■

It seems that our intuition that the cost of introducing random signs is tolerable is correct: based on previous two propositions, almost no cost is paid since the ratio is bounded by 1/2 and 2. Even more, given X_1^n,

$$\sum_{i=1}^n \frac{L_f(X_i)}{\sqrt{n}} \cdot \epsilon_i \quad \text{(You will see why } \sqrt{n} \text{ in a moment)}$$

is a linear combination of ϵ_is which are independent (mean-zero) bounded variables (sub-Gaussian). Based on the property of sub-Gaussian, we know

$$\begin{aligned}
\left\|\sum_{i=1}^n \frac{L_f(X_i)}{\sqrt{n}} \cdot \epsilon_i\right\|_{\psi_2}^2 &\leq C\left[\sum_{i=1}^n \left(\frac{L_f(X_i)}{\sqrt{n}}\right)^2\right] \max_{i=1,\ldots,n} \|\epsilon_i\|_{\psi_2}^2 \quad \text{(Property 6 (A Note on } \|\cdot\|_{\psi_2})) \\
&= \frac{C}{\log 2}\left[\sum_{i=1}^n \left(\frac{L_f(X_i)}{\sqrt{n}}\right)^2\right] \quad (\|\epsilon_i\|_{\psi_2}^2 = \frac{1}{\log 2}) \\
&\leq \frac{C}{\log 2} \quad (L_f(X_i)^2 \leq 1).
\end{aligned} \tag{7.4}$$

Thus, given X_1^n,

$$\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right|$$

is the maximum of **finite sub-Gaussian** mean-zero random variables (Are they independent? Why?). Next, we will address how to bound the expectation of such maximum.

7.5 Tail Behavior of $\max_{j=1,\ldots N} |Z_j|$, Z_j Sub-Gaussian

Here $Z_j = \sum_{i=1}^n \epsilon_i L_{f_j}(X_i)$ where f_j is a boolean function. The randomness of Z_j belongs to ϵ_1^n only because it is conditioned on X_1^n. N represents the number of possible predictions from function family $\mathcal{F}$ when conditional on X_1^n. Namely, $N = |\mathcal{F}|_{X_1^n}|$.

Let Z_js be sub-Gaussian random variables with $\|Z_j\|_{\psi_2} \leq K$. We are curious about the tail behavior of $\max_{j=1,\ldots,N} |Z_j|$.

$$\begin{aligned}
P(\max_{j=1,\ldots N} |Z_j| \geq t) &= P(\cup_{j=1}^N (|Z_j| \geq t)) \\
&\leq \sum_{j=1}^N P(|Z_j| \geq t) \quad \text{(Union bound)} \\
&\leq \sum_{j=1}^N 2\exp(-t^2/K^2) \quad \text{(Definition of } \|\cdot\|_{\psi_2}) \\
&= 2N\exp(-t^2/K^2).
\end{aligned}$$

If we choose $t = K(\sqrt{\log(N)} + u)$, then we have

$$\begin{aligned}
& t^2 \geq K^2(\log(N) + u^2) \\
\Rightarrow \quad & P(\max_{j=1,\ldots N} |Z_j| \geq t) \leq 2N\exp(-t^2/K^2) \leq 2N\exp(-\log(n) - u^2) = 2\exp(-u^2).
\end{aligned}$$

Thus, for n sub-Gaussian variables, the cost one pays for $\max_{j=1,\ldots n} Z_j$ is about $K\sqrt{\log(n)}$.

7.6 $E\max_{j=1,\ldots N} |Z_j|$, Z_j Sub-Gaussian

One could use the tail bound to bound the expectation directly. However, there exists another way that is less numerical and involves a new trick. Thus, I will provide this detour approach which starts with a different perspective of the union bound. $E\max_{j=1,\ldots,N} Z_j$.

7.6.1 Different Perspective of Union Bound

The union bound is

$$P(\cup_{j=1}^N A_j) \leq \sum_{j=1}^N P(A_j).$$

If we use the $P(A_j) = E1_{A_j}$, we have

$$E \sup_{j=1,\ldots N} 1_{A_j} \leq \sum_{j=1}^{N} E1_{A_j}.$$

We could obtain the union bound by using the following numeric inequality:

$$Z_j \geq 0 \;\Rightarrow\; \max_{j=1,\ldots N} Z_j \leq \sum_{j=1}^{N} Z_j.$$

If we use this inequality for any non-negative random variables Z_j, we have

$$E \max_{j=1,\ldots N} Z_j \leq \sum_{i=1}^{N} EZ_j. \tag{7.5}$$

If Z_j follows the same distribution (might not be independent), then we have

$$E \max_{j=1,\ldots N} Z_j \leq NEZ_j.$$

Most of the time, it is a loose bound (could you find a case that it is tight?). One way to think about it is that $\max_{j=1,\ldots N} Z_j$ should be closely related to the tail behavior, however, we are not exploiting it here at all. Our usual way of introducing tail behavior is to use the MGF. Because MGF is always positive, it is possible to bound $\max_{j=1,\ldots N} Z_j$ in the following way.

$$\begin{aligned}
E \max_j Z_j &= \frac{1}{\lambda} E \log \exp(\lambda \max_j Z_j) \quad (\log \exp x = x)\\
&\leq \frac{1}{\lambda} \log E(\exp(\lambda \max_j Z_j)) \quad (\log x \text{ is concave.})\\
&= \frac{1}{\lambda} \log E(\max_j \exp(\lambda Z_j)) \quad (\exp(x) \text{ is monotone increasing})\\
&\leq \frac{1}{\lambda} \log \left(\sum_{j=1}^{N} E \exp(\lambda Z_j) \right) \quad (\text{Eq. } 7.5)
\end{aligned}$$

7.6.2 $E \max_{j=1,\ldots N} Z_j$, Z_j Sub-Gaussian

If $EZ_j = 0$ and $\max_j \|Z_j\|_{\psi_2} = K$, we have

$$E \exp(\lambda Z_j) \leq \exp(CK^2\lambda^2),$$

which implies

$$E \max_j Z_j \leq \frac{1}{\lambda} \left(\log(N) + CK^2\lambda^2 \right).$$

Find $\lambda^\star \geq 0$ that minimizes the RHS to obtain

$$E \max_i Z_j \leq 2\sqrt{CK^2 \log N}. \tag{7.6}$$

Thus, the expectation of the maximum of n mean-zero sub-Gaussian random variables (do we require independence here?) are bounded by $2\sqrt{C \log N} \times K$ where K indicates the thickness of the tail. This form is very similar to the one about its tail behavior.

7.6.3 From Z_j to $|Z_j|$

The above argument fails for $|Z_j|$ since $E|Z_j| \neq 0$. Luckily, based on Vershynin (2018, Eq. 2.15), we know

$$\|Z_j\|_{L^p} \leq C_1 \|Z_j\|_{\psi_2} \sqrt{p} \quad \forall p \geq 1.$$

If we take $p = 1$, we have $E|Z_j| \leq C_1 \|Z_j\|_{\psi_2} \leq C_1 K$. By the definition of $\|\cdot\|_{\psi_2}$, $|Z_j|$ shares the same $\|\cdot\|_{\psi_2}$ as Z_j. Thus, $|Z_j| - E|Z_j|$ is mean-zero sub-Gaussian with $\|\cdot\|_{\psi_2}$ upper bounded by K. Based on the previous derivation, we have

$$E \sup_{j=1,\ldots,N} |Z_j| - C_1 K \leq E \sup_{j=1,\ldots,N} \Big[|Z_j| - E|Z_j|\Big] \leq 2K\sqrt{C \log N}.$$

Thus, we have

$$E \sup_{j=1,\ldots,N} |Z_j| \leq C_1 K + 2K\sqrt{C \log N} \leq C_2 K \sqrt{\log N}.$$

To summarize, the expectation of $E \max |Z_j|$ and $E \max Z_j$ are both of the size $\sqrt{\log N}$ which also agrees with the tail behavior.

7.6.4 When the Bound is Good?

The following inequality plays a key role in the above derivation,

$$\max_{i=1,\ldots,N} Z_j \leq \frac{1}{\lambda} \log \sum_{i=1}^{N} \exp(\lambda Z_j) \left(\leq \max_{i=1,\ldots N} Z_j + \frac{\log N}{\lambda} \right).$$

It is tempting to have $\log N/\lambda \to 0$, to have two bounds matched. However, our final choice of λ is of order $\log N$. This happens because $\log x$ is concave rendering $E[\log(\cdot)] \leq \log E[\cdot]$. In other words, the upper bound is not an upper bound for $\log E[\cdot]$.

When the maximum $\exp(\lambda Z)$ does not dominate others, the bound will be loose. If all Z_js follow the same distribution, this is likely to happen when there are strong positive correlation. Thus, we are expecting this bound works fairly well when Z_js are i.i.d. Indeed, one could show that

$$c\sqrt{\log N} \leq E[\max_{i=1,\ldots,N} Z_j] \leq C\sqrt{\log n},$$

when Z_js are i.i.d. $N(0,1)$.

7.7 A Simple Bound of $E_\epsilon \sup_{f\in\mathcal{F}} |\sum_{i=1}^n \epsilon_i L_f(X_i)|$

Recall that $|\sum_{i=1}^n \epsilon_i L_f(X_i)|$ is sub-Gaussian, and $\sup_{f\in\mathcal{F}}$ is in fact a maximum of at most 2^n random variables, we know that

$$\begin{aligned} E_\epsilon \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right| &\leq C\sqrt{n}\sqrt{\log 2^n} \quad \left(\left\|\sum_{i=1}^n \epsilon_i L_f(X_i)\right\|_{\psi_2} \lesssim \sqrt{n},\ \textit{Eq. } 7.4\right) \\ &= \tilde{C}n. \end{aligned}$$

This is a trivial result since $|\epsilon_i L_f(X_i)| \leq 1$. However, if we could replace the trivial cardinality bound of 2^n to a tighter bound, then we are in business. This is where the VC dimension comes into play.

Lemma 7.6 (Sauer-Shelah Lemma. Vershynin (2018), Theorem 8.3.16.) *Let $\mathcal{F}$ be a class of Boolean functions on an n-points set Ω. Then*

$$|\mathcal{F}| \leq \left(\frac{en}{d}\right)^d,$$

where $d = vc(\mathcal{F})$.

With this tighter bound, we could have the following

$$\begin{aligned} E_\epsilon \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i L_f(X_i)\right| &\leq C\sqrt{n}\sqrt{\log\left(\frac{en}{d}\right)^d} \quad \left(\left\|\sum_{i=1}^n \epsilon_i L_f(X_i)\right\|_{\psi_2} \lesssim \sqrt{n}\right) \\ &= \tilde{C}\sqrt{vc(\mathcal{F})n\log n}. \end{aligned}$$

Divide both sides by n and recall the definitions of $\hat{R}$, R and L_f, we have

$$E\left[\sup_{f\in\mathcal{F}} \left|\hat{R}_n(f,T) - R(f,T)\right|\right] \leq C\sqrt{\frac{vc(\mathcal{F})\log n}{n}}. \tag{7.7}$$

Unfortunately, we are still off the bound with a $\log n$ factor. Let us think about how to close the gap. Fundamentally speaking, it is our way of using union bound (or $Z_j \geq 0 \Rightarrow \max_i Z_j \leq \sum_i Z_j$) that leads to the problem. It must be the case that for some $f \neq g$, $\sum_{i=1}^n \epsilon_i L_f(X_i)$ and $\sum_{i=1}^n \epsilon_i L_g(X_i)$ are highly positively correlated rendering a loose bound. In fact, we could take a difference of them as

$$\sum_{i=1}^{n} \epsilon_i L_f(X_i) - \sum_{i=1}^{n} \epsilon_i L_g(X_i) = \sum_{i=1}^{n} \epsilon_i (L_f(X_i) - L_g(X_i)).$$

Clearly, if f and g agrees on most of X_is, then they will have a high correlation. It is worth to notice that this difference is also a sub-Gaussian variable. However, to get a deeper understanding, we need to handle the $EL(f(X_i), T(X_i))$ and $EL(g(X_i), T(X_i))$ (recall the definition of $L_f(X_i)$ and $L_g(X_i)$). It is not easy to do so, because conditioning on X_1^n does not help. Luckily, if we refine the form of symmetrization, we could get rid of them too.

7.8 Better Form of Symmetrization

The change of symmetrization form is driven mainly by technical reasons. If you find it confusing, you could jump to the conclusion and ignore it.

Proposition 7.7

$$E_X \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} W_i(f) - EW_i(f) \right| \le 2E_X E_\epsilon \left[\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i W_i(f) \right| \right], \tag{7.8}$$

where $W_i(f) = L(f(X_i), T(X_i))$.

Proof: We first replace $EW_i(f)$ with $E\tilde{W}_i(f)$ where $\tilde{W}_i(f)$ is an independent copy of $W_i(f)$. Then proceed use the identical techniques as in Proposition 7.2. ■

Now, we get rid of term $EL(f(X_i), T(X_i))$ and we could focus on

$$\sum_{i=1}^{n} \epsilon_i W_i(f) - \sum_{i=1}^{n} \epsilon W_i(g) = \sum_{i=1}^{n} \epsilon \left(W_i(f) - W_i(g) \right).$$

Moreover, if we leverage $L(f(X_i), T(X_i)) = |f(X_i) - T(X_i)|$, we have

$$\begin{aligned} W_i(f) - W_i(g) = L(f(X_i), T(X_i)) - L(g(X_i), T(X_i)) &= |f(X_i) - T(X_i)| - |g(X_i) - T(X_i)| \\ &\le |f(X_i) - g(X_i)| \quad (||x| - |y|| \le |x-y|) \\ &= |(f-g)(X_i)|. \end{aligned}$$

Thus, we could bound the ψ_2 norm of $\sum_{i=1}^n \epsilon_i (W_i(f) - W_i(g))$,

$$\left\| \sum_{i=1}^{n} \epsilon_i (W_i(f) - W_i(g)) \right\|_{\psi_2}^2 \le C \sum_{i=1}^{n} (W_i(f) - W_i(g))^2 \le C \sum_{i=1}^{n} ((f-g)(X_i))^2.$$

If we define $d(f,g) = \sqrt{\sum_{i=1}^n ((f-g)(X_i))^2}$, we have

$$\left\| \sum_{i=1}^n \epsilon_i (W_i(f) - W_i(g)) \right\|_{\psi_2} \leq Cd(f,g).$$

This is an amazing result because we connect the tail behavior of the distance to the distance between their indexes. In some sense, this indicates some continuity with respect to the indexes. We have exploited such continuity using covering in the last two lectures.

Let me make it more abstract to highlight the idea. Previously, we are dealing with

$$\max_{j=1,\ldots N} Z_j,$$

where Z_js are sub-Gaussian. Now, we also know $Z_j - Z_j$ is also sub-Gaussian (structure!), and the smaller $|i-j|$ is, the thinner the tail becomes. This means that Z_j and Z_{j+1} are highly correlated. It is tempting to use covering (at most choose one) to avoid utilizing a loose union bound.

7.9 Chaining

Given the analysis above, it is time to bound $\sup_{t \in T} Z_t$ with structure

$$\|Z_t - Z_s\|_{\psi_2} \leq Kd(t,s) \quad \& \quad EZ_s = 0. \tag{7.9}$$

I will first provide some simple examples to indicate that this structure is not rare. Then we will move to the analysis, which eventually will lead to the idea of chaining.

7.9.1 Simple Examples

Although $\sum_{i=1}^n \epsilon_i W_i(f)$ does not qualify because of $EW_i(f) \neq 0$, it motivates the following simple example.

$$Z_t = \langle a, t \rangle, \quad \text{where } a \in \mathbb{R}^n \text{ is sub-Gaussian with } Ea = 0.$$
$$\Rightarrow \|Z_t - Z_s\|_{\psi_2} = \|\langle a, t-s \rangle\|_{\psi_2} \leq \|a\|_{\psi_2} \|t-s\|_2.$$

Since bounded variable is sub-Gaussian, we know that Z_t also follows this structure when Z_t is Lipschitz w.r.t. t. Mathematically speaking,

$$|Z_t - Z_s| \leq Ld(t,s) \Rightarrow \|Z_t - Z_s\|_{\psi_2} \leq Cd(t,s).$$

This observation leads to another example of $Z_t = \|At\|_2 - E\|At\|_2$ which is $2\|A\|_{op}$-Lipschitz.

7.9.2 Analysis of Finite $|T|$

One could think about Eq. 7.9 as some continuity property. To tighten the union bound, we are eager to use ϵ-separated set, $P(\epsilon)$, to make sure that the union bound will not apply to the events with large overlaps. In other words, we are tempting to use following inequality

$$\sup_{t\in T} Z_t \le \sup_{t\in T}\left(Z_t - Z_{\pi(t)}\right) + \sup_{t\in P(\epsilon)} Z_{\pi(t)}. \tag{7.10}$$

There are three problems.

1. How to choose ϵ? This turns out to be irrelevant because chaining is about choosing lots of different ϵs.
2. How to handle the first term? This is critical because it could involves $|T| = \infty$.
3. How to handle the second term? The requirement that $EZ_t = 0$ is use to handle this. The idea is to choose ϵ so large that $P(\epsilon) = t_0$, namely one point. In this way, $\sup_{t\in P(\epsilon)}$ is a fake sup.

If we have Lipschitz property, we could pay the Lipschitz constant for the first term. However, the Lipschitz constant describes the worst possible case for any pair in T, which could be larger than $\sup_{t\in T} Z_t$. In other words, we might not be able to afford to use the Lipschitz constant to handle the first term. The situation seems worse when there is no Lipschitz property (Eq. 7.9 does not imply Lipschitz) because Lipschitz property might be the only tool for us to bound infinite T.

If so, let us think about how to handle the first term regarding finite T. It might be easy to expand the finite case to countable infinite. Based on Section 7.6.2, we know that

$$E\sup_{t\in T}\left(Z_t - Z_{\pi(t)}\right) \le 2CK\sqrt{\log|T|}\epsilon \quad (\|Z_t - Z_{\pi(t)}\|_{\psi_2} \le Kd(t,\pi(t)) \le K\epsilon).$$

This is a good direct bound for finite $|T|$. Unfortunately, it explodes when $|T| \to \infty$. One way to delay the problem is to introduce a $\epsilon/2$-separated set $P(\epsilon/2)$ and then utilize Eq. 7.10 again as following,

$$\sup_{t\in T}\left(Z_t - Z_{\pi(t)}\right) \le \sup_{t\in T}\left(Z_t - Z_{\pi'(t)}\right) + \sup_{t\in T}\left(Z_{\pi'(t)} - Z_{\pi(t)}\right)$$

The first term could be bounded as

$$E\sup_{t\in T}\left(Z_t - Z_{\pi'(t)}\right) \le 2CK\sqrt{\log|T|}\frac{\epsilon}{2}.$$

Since this bound still involves $|T|$, we still need to think how to handle it. Bounding the second term takes some work. First of all, we need to count the number of sub-Gaussian variables. There are at most $|P(\epsilon/2)| \times |P(\epsilon)| \le |P(\epsilon/2)|^2$ pairs of $(\pi'(t), \pi(t))$. Thus, we know there are at most $|P(\epsilon/2)|^2$ sub-Gaussian variables. Since

$$\begin{aligned} d(\pi'(t), \pi(t)) &\leq d(\pi'(t), t) + d(t, \pi(t)) \quad \text{(Triangle inequality of } d(\cdot,\cdot)\text{)} \\ &\leq \epsilon/2 + \epsilon \quad \text{(Definition of } \pi'(t) \text{ and } \pi(t)\text{)}, \end{aligned}$$

we know the sub-Gaussian norm is bounded by

$$\|Z_{\pi'(t)} - Z_{\pi(t)}\|_{\psi_2} \leq 3/2K\epsilon.$$

Now we could utilize Section 7.6.2 to bound the second term

$$E \sup_{t\in T} \left(Z_{\pi'(t)} - Z_{\pi(t)}\right) \leq 3CK\sqrt{2\log|P(\epsilon/2)|}\epsilon.$$

It is possible that

$$\sqrt{\log|P(\epsilon/2)|} \ll \sqrt{\log|T|}, \tag{7.11}$$

this means that by introducing a new separate set, we reduce the bound significantly. To get some intuition regarding Eq. 7.11, one could think under the assumption that $|T| \to \infty$. In this case, the RHS goes to infinity. Meanwhile, the LHS might stay finite (think about S^{p-1}).

Since we gain some hope by introducing another $\epsilon/2$-separated set, we shall continue doing this to eliminate the existence of $|T|$ in the upper bound. When $|T|$ is finite, $P(\epsilon/2^k)$ will eventually become $|T|$ once $k \geq K_0$ (or large enough). Thus, in the end, we could avoid bounding the first term because $\sup_{t\in T} Z_t - Z_{\pi^{K_0}(t)} = 0$.

Before we proceed to the countable infinite case, we shall take a look at the bound. It is in the following form,

$$C_2K\sum_{i=1}^{K_0}\sqrt{\log|P(\epsilon/2^i)|}\frac{\epsilon}{2^i} = 2C_2K\sum_{i=1}^{K_0}\sqrt{\log|P(\epsilon/2^i)|}\left(\frac{\epsilon}{2^i} - \frac{\epsilon}{2^{i+1}}\right)$$

Now, it looks like a Riemann summation. Notice that $|P(\epsilon)|$ is monotone decreasing, we could bound this summation by some integral in the following form

$$E\sup_{t\in T} Z_t \leq C_3K\int_0^\infty \sqrt{\log|N(T,d,\epsilon)|}d\epsilon.$$

This is the Dudley's integral inequality. It is worth to notice that I replace ϵ-separated set with ϵ-net. The reason is that ϵ-separated set is easy to motive but ϵ-net also works.

Remark 7.8 *The upper bound of the integral is not ∞. It is $diam(T) = \sup_{x,y} d(x,y)$. This happens because for any $\epsilon > diam(T)$, $N(T,d,\epsilon) = 1$.*

Remark 7.9 *Dudley's inequality is not correct for $E\sup_{t\in T}|Z_t|$ because $E|Z_{t_0}| \neq 0$. However, everything still goes through for $E\sup_{t\in T}|Z_t - Z_{t_0}|$ (why?). If we could find $Z_{t_0} = 0$, we could bound $E\sup_{t\in T}|Z_t|$. To* *highlight the main idea, I will not solve this issue but provide alert when it occurs.*

7.9.3 Finite to Infinite.

Let us first deal with the countable infinite case. Let T_k be the first k elements of T. We could apply the above inequality to every T_k to obtain a bound. Notice that $|N(T, d, \epsilon)|$ is monotone increasing[3] in T (larger set, higher points needed to cover). Thus for every T_k we have

$$E \sup_{t \in T_k} Z_t \leq C_3 K \int_0^\infty \sqrt{\log |N(T, d, \epsilon)|} d\epsilon.$$

Since LHS is increasing w.r.t. k, we could take $\lim_{k \to \infty}$ to both sides to obtain

$$E \sup_{t \in T} Z_t \leq C_3 K \int_0^\infty \sqrt{\log |N(T, d, \epsilon)|} d\epsilon.$$

That is to say, we have handled the countable infinite case. For the uncountable infinite case, we need the following assumption to make it work.

Definition 7.10 (Separable Process) *A random process $\{Z_t\}_{t \in T}$ is called separable if there is a countable set $T_0 \subset T$ such that*

$$Z_t \in \lim_{s \to t, s \in T_0} X_s \quad \forall t \in T.$$

If we are dealing with a separable process Z_t, we have

$$\sup_{t \in T} Z_t = \sup_{t \in T_0} Z_t.$$

Thus, we know how to handle sup of separable processes.

7.10 Application of Dudley's Inequality

We first apply Dudley's inequality to the excess risk which will remove the $\log n$ factor. Then we use it for L-Lipschitz functions to show that it is the covering number that matters.

7.10.1 Excess Risk

We would like to use Dudley's inequality to remove the $\log n$ factor in Eq. 7.12. We will leverage the following result about $|N(\epsilon)|$.

[3] This is not always true. Think about Vershynin (2018, Exercise 4.2.10). The exercise also shows that this technical issue is minor since there is an approximate version.

Theorem 7.11 (Covering Number via VC Dimension, Theorem 8.3.18 Vershynin (2018)) *Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ). Then for every $\epsilon \in (0,1)$, we have*

$$|N(\mathcal{F}, L^2(\mu), \epsilon)| \leq \left(\frac{2}{\epsilon}\right)^{Cd},$$

where $d = vc(\mathcal{F})$.

All you need to understand about the above theorem[4] is that it provides some bound for $|N(\epsilon)|$ which we could use in Dudley's inequality as following.

$$E_\epsilon \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i W_i(f)\right| \leq C\sqrt{n}\int_0^\infty \sqrt{\log|N(\epsilon)|}d\epsilon \quad (\left\|\sum_{i=1}^n \epsilon_i W_i(f-g)\right\|_{\psi_2} \lesssim \sqrt{n})$$

$$\leq C_1\sqrt{n}\int_0^\infty \sqrt{vc(\mathcal{F})\log(2/\epsilon)}d\epsilon$$
$$= C_2\sqrt{vc(\mathcal{F})n}. \quad (\int_0^\infty \sqrt{\log(2/\epsilon)}d\epsilon \lesssim 1)$$

Divide both sides by n and recall the definitions of $\hat{R}$, R and L_f, we have

$$E\left[\sup_{f\in\mathcal{F}}\left|\hat{R}_n(f,T) - R(f,T)\right|\right] \leq C\sqrt{\frac{vc(\mathcal{F})}{n}}. \tag{7.12}$$

Remark 7.12 *Is it possible to use Dudley's inequality without symmetrization? Why?*

7.10.2 Uniform Law of Large Numbers

The above example heavily relies on the fact that the function class is Boolean with finite VC dimension. In the following example, we focus on Lipschitz function on $[0,1]$ as the ϵ-net of Lipschitz function is well behaved.

Theorem 7.13 *Let $X, X_1, \ldots, X_n$ be i.i.d. random variables taking values in [0,1]. $\mathcal{F}$ is the class of Lipschitz functions*

$$\mathcal{F} \triangleq \{f : [0,1] \to \mathbb{R}, \|f\|_{Lip} \leq L\}.$$

[4] It is not rigorous here because Dudley's inequality is for $\sup_{t\in T} Z_t$ but I use it for $\sup_{t\in T}|Z_t|$. This technical issue could be solved (Vershynin 2018, Exercise 8.3.25).

Then

$$E \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - Ef(X) \right| \leq \frac{CL}{\sqrt{n}}.$$

First of all, I want to introduce the following notation to simplify the analysis,

$$X_f \triangleq \frac{1}{n}\sum_{i=1}^{n} f(X_i) - Ef(X).$$

Now the problem becomes $E\sup_{f\in\mathcal{F}}|X_f|$. Since we no longer have finite events when conditional on the X_1^n, symmetrization is no longer necessary. The following Lemma shows that some bound of $|N(\mathcal{F}, \|\cdot\|_\infty, \epsilon)|$ is obtainable.

Lemma 7.14 (Example 5.10 Wainwright (2019))

$$\log\left(|N(\mathcal{F}, \|\cdot\|_\infty, \epsilon)|\right) \lesssim \frac{L}{\epsilon}.$$

Equipped with this bound, we could leverage the Dudley's inequality if we can show the $\|\cdot\|_{\psi_2}$ of the increments is close related to $\|\cdot\|_\infty$.

Proof: 1. Increments. For any $f, g \in \mathcal{F}$, we have

$$X_f - X_g = \frac{1}{n}\sum_{i=1}^{n} Z_j \quad \text{where } Z_j \triangleq (f-g)(X_i) - E(f-g)(X).$$

Because Z_js are independent,

$$\begin{aligned}\|X_f - X_g\|_{\psi_2}^2 &= \frac{1}{n^2}\left\|\sum_{i=1}^{n} Z_j\right\|_{\psi_2}^2 \quad (\|\cdot\|_{\psi_2} \text{ is a norm})\\ &\leq \frac{C}{n^2}\sum_{i=1}^{n}\|Z_j\|_{\psi_2}^2 \quad (\text{Property of } \|\cdot\|_{\psi_2}).\end{aligned}$$

In fact, Z_j is bounded, as shown below

$$\begin{aligned}|Z_j| = |(f-g)(X_i) - E(f-g)(X)| &\leq |(f-g)(X_i)| + |E(f-g)(X)|\\ &\leq \|f-g\|_\infty + E\|f-g\|_\infty \quad (\text{Definition of } \|\cdot\|_\infty)\\ &= 2\|f-g\|_\infty.\end{aligned}$$

Thus, we know

$$\|Z_j\|_{\psi_2} \leq C\|f-g\|_\infty.$$

This implies

$$\begin{aligned}
&\|X_f - X_g\|_{\psi_2}^2 \le \frac{1}{n^2} \times nC^2 \|f-g\|_\infty^2 \\
\Rightarrow \quad &\|X_f - X_g\|_{\psi_2} \le \frac{C}{\sqrt{n}} \|f-g\|_\infty.
\end{aligned}$$

Moreover, we clearly have $EX_f = 0$. Now, we could use Dudley's inequality[5].

$$\begin{aligned}
E \sup_{f\in\mathcal{F}} |X_f| &\le \frac{C_4}{\sqrt{n}} \int_0^{diam(\mathcal{F})} \sqrt{\log\left(|N(\mathcal{F}, \|\cdot\|_\infty, \epsilon)|\right)} d\epsilon \\
&\le \frac{C_4}{\sqrt{n}} \int_0^{2L} \sqrt{\frac{L}{\epsilon}} d\epsilon \\
&\le \frac{C_5 L}{\sqrt{n}}.
\end{aligned}$$

In the proof, we use the following bound

$$\begin{aligned}
diam(\mathcal{F}) &= \sup_{f,g\in\mathcal{F}} \|f-g\|_\infty \le \sup_{f\in\mathcal{F}} \|f\|_\infty + \sup_{g\in\mathcal{F}} \|g\|_\infty \\
&= 2\sup_{f\in\mathcal{F}} \|f\|_\infty \le 2L.
\end{aligned}$$

The last inequality is non-trivial. It uses the fact that we could force $f(0) = 0$ without changing X_f:

$$X_f = \frac{1}{n}\sum_{i=1}^n f(X_i) - Ef(X) = \frac{1}{n}\sum_{i=1}^n (f(X_i) - f(0)) - E(f(X) - f(0)).$$

Therefore,

$$\begin{aligned}
\|f\|_\infty &= \sup_{x\in[0,1]} |f(x)| = \sup_{x\in[0,1]} |f(x) - f(0)| \quad (f(0) = 0) \\
&\le \sup_{x\in[0,1]} L|x| \quad (f(x) \text{ is L-Lipschitz}) \\
&\le L.
\end{aligned}$$

■

References

van Handel R (2014) Probability in high dimension. Technical report, PRINCETON UNIV NJ, URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

[5] Once again, it is not rigorous here: Dudley's inequality is for $\sup_{t\in T} Z_t$ but I use it for $\sup_{t\in T} |Z_t|$. We could get around this issue as well (Vershynin 2018, Remark 8.1.5).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 8: A Note on Symmetrization

Lecturer: Long Zhao, longzhao@nus.edu.sg

8.1 Resources

- Vershynin (2018, Chapter 6).
- Concentration of Measure [Terence Tao's Blog]

8.2 Objective

In our proof of excess risk via VC dimension, we saw that symmetrization drastically simplifies the problem. In fact, symmetrization is a standard tool. For example, Gao (2020) uses it to prove the finite-sample guarantees for Wasserstein distributionally robust optimization. This note's objective is to provide a systematic view of symmetrization and another fantastic example that symmetrization is critical.

8.3 Symmetric Distribution

A random variable is symmetric if X and $-X$ have the same distribution. Symmetric Bernoulli and $N(0, \sigma^2)$ are symmetric. Next, let me introduce some ways to construct symmetric distributions.

Proposition 8.1 (Constructing Symmetric Distributions) *Let X be a random variable and ξ be an independent symmetric Bernoulli random variable.*

1. *ξX and $\xi|X|$ are symmetric random variables, and they follow the same distribution.*
2. *If X is symmetric, ξX and $\xi|X|$ have the same distribution as X.*
3. *Let X' be an independent copy of X. Then $X - X'$ is symmetric.*

Proof: **1**. By definition, we only need to prove ξX and $-\xi X$ follow the same distribution to show ξX is symmetric. Let me proceed with standard way first.

$$\begin{aligned}
P(\xi X \leq x) &= P(\xi X \leq x, \xi = -1) + P(\xi X \leq x, \xi = 1) \\
&= \frac{1}{2}P(X \geq -x) + \frac{1}{2}P(X \leq x) \\
&= \frac{1}{2}P(-X \leq x) + \frac{1}{2}P(-X \geq -x) \\
&= P(-X \leq x, \xi = 1) + P(-X \geq -x, \xi = -1) \\
&= P(\xi(-X) \leq x, \xi = 1) + P(\xi(-X) \leq x, \xi = -1) \\
&= P(\xi(-X) \leq x).
\end{aligned}$$

In fact, we could prove it using independence more elegantly. Because ξ and X are independent, we know their joint distribution (ξ, X). Since ξ is symmetric, (ξ, X) has the same distribution as $(-\xi, X)$. Therefore, ξX and $-\xi X$ follow the same distribution. **It is worth to notice that we only use ξ is symmetric here.** Similarly, we have $\xi|X|$ is symmetric.

Next, we use standard way to prove ξX and $\xi|X|$ follow the same distribution. Without loss of generality, we only focus on $x \geq 0$,

$$\begin{aligned}
P(\xi|X| \leq x) &= P(\xi|X| \leq x, \xi = -1) + P(\xi|X| \leq x, \xi = 1) \\
&= \frac{1}{2}P(|X| \geq -x) + \frac{1}{2}P(|X| \leq x) \\
&= \frac{1}{2} + \frac{1}{2}P(-x \leq X \leq x) \\
&= \frac{1}{2}\left(P(X < -x) + P(X \geq -x)\right) + \frac{1}{2}P(-x \leq X \leq x) \\
&= \frac{1}{2}\left(P(X < -x) + P(-x \leq X \leq x)\right) + \frac{1}{2}P(X \geq -x) \\
&= \frac{1}{2}P(X \leq x) + \frac{1}{2}P(X \geq -x) \\
&= P(\xi X \leq x).
\end{aligned}$$

2. If X is symmetric, we have $P(X \leq x) = P(X \geq -x)$. Thus,

$$P(\xi X \leq x) = \frac{1}{2}P(X \leq x) + \frac{1}{2}P(X \geq -x) = P(X \leq x).$$

3. By independence, we know the joint distribution of (X, X') and (X', X) are the same. Thus, $X - X'$ has the same distribution as $X' - X$.

■

8.4 Properties of Symmetrization

Next Lemma is **essential** for symmetrization because it build some connections between X_i and $\epsilon_i X_i$.

Lemma 8.2 (Symmetrization) *Let $X_1, \dots X_N$ be independent, mean zero, random vectors in a normed space. Then*

$$\frac{1}{2} E \left\| \sum_{i=1}^N \epsilon_i X_i \right\| \leq E \left\| \sum_{i=1}^N X_i \right\| \leq 2E \left\| \sum_{i=1}^N \epsilon_i X_i \right\| .$$

Proof: **Upper bound**. Because every norm is convex (why?), we have

$$\begin{aligned} \left\| \sum_{i=1}^N X_i \right\| = \left\| \sum_{i=1}^N X_i - 0 \right\| &= \left\| \sum_{i=1}^N X_i - E_{X'} \sum_{i=1}^N X_i' \right\| \\ &\leq E_{X'} \left\| \sum_{i=1}^N X_i - \sum_{i=1}^N X_i' \right\| \quad \text{(Jensen's inequality)} \\ &= E_{X'} \left\| \sum_{i=1}^N (X_i - X_i') \right\| . \end{aligned}$$

Take expectation with respect to E_X to have

$$\begin{aligned} E \left\| \sum_{i=1}^N X_i \right\| &\leq E \left\| \sum_{i=1}^N (X_i - X_i') \right\| \\ &= E \left\| \sum_{i=1}^N \epsilon_i (X_i - X_i') \right\| \quad (X_i - X_i' \text{ and } \epsilon_i (X_i - X_i') \text{ share the same distribution}) \\ &\leq E \left\| \sum_{i=1}^N \epsilon_i X_i \right\| + E \left\| \sum_{i=1}^N \epsilon_i X_i' \right\| \quad (\|a+b\| \leq \|a\| + \|b\|) \\ &= 2E \left\| \sum_{i=1}^N \epsilon_i X_i \right\| \quad (\epsilon_i X_i \text{ and } \epsilon X_i' \text{ share the same distribution}). \end{aligned}$$

Lower Bound. All we need to prove is

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

Based on the prove above, 2 might come from

$$E \left\| \sum_{i=1}^N X_i - \sum_{i=1}^N X_i' \right\| \leq E \left\| \sum_{i=1}^N X_i \right\| + E \left\| \sum_{i=1}^N X_i' \right\| = 2E \left\| \sum_{i=1}^N X_i \right\| .$$

Meanwhile

$$\begin{aligned}E\left\|\sum_{i=1}^N X_i - \sum_{i=1}^N X_i'\right\| &= E\left\|\sum_{i=1}^N \epsilon_i(X_i - X_i')\right\| \quad (X_i - X_i' \text{ and } \epsilon_i(X_i - X_i') \text{ share the same distribution})\\ &= E_{X,\epsilon}E_{X'}\left\|\sum_{i=1}^N \epsilon_i X_i - \sum_{i=1}^N \epsilon_i X_i'\right\| \quad \text{(independence)}\\ &\geq E_{X,\epsilon}\left\|\sum_{i=1}^N \epsilon_i X_i - E_{X'}\sum_{i=1}^N \epsilon_i X_i'\right\| \quad \text{(Jensen's inequality)}\\ &= E_{X,\epsilon}\left\|\sum_{i=1}^N \epsilon_i X_i\right\|.\end{aligned}$$

■

Remark 8.3 *Where do we use independence in the proof?*

It is worth noticing that we only use the convexity of the norm in the proof. Thus, if X_i is a random matrix and $\|\cdot\|$ is the operator norm, the result still holds. Moreover, it is easy to get the following Corollaries using similar tricks under convexity.

Corollary 8.4 *Let $F : \mathbb{R}_+ \to \mathbb{R}$ be an increasing, convex function.*

$$EF\left(\frac{1}{2}\left\|\sum_{i=1}^N \epsilon_i X_i\right\|\right) \leq EF\left(\left\|\sum_{i=1}^N X_i\right\|\right) \leq EF\left(2\left\|\sum_{i=1}^N \epsilon_i X_i\right\|\right).$$

Proof: Using the composition property of convex function, we know $F(\|\cdot\|)$ is still convex. Following the proof above, we could obtain the proof. ■

Corollary 8.5 *Let $X_1, \ldots X_N$ be independent, mean zero* **random variables***. Show that $\sum_i X_i$ is sub-Gaussian if and only if $\sum_i \epsilon_i X_i$ is sub-Gaussian. Moreover,*

$$c\left\|\sum_{i=1}^N \epsilon_i X_i\right\|_{\psi_2} \leq \left\|\sum_{i=1}^N X_i\right\|_{\psi_2} \leq C\left\|\sum_{i=1}^N \epsilon_i X_i\right\|_{\psi_2}.$$

Proof: Because $f(x) = \exp(\lambda x),\ \lambda > 0$ is an increasing convex function, we have

$$\begin{aligned}\exp\lambda\left(\sum_{i=1}^N X_i\right) = \exp\lambda\left(\sum_{i=1}^N X_i - 0\right) &= \exp\lambda\left(\sum_{i=1}^N X_i - E_{X'}\sum_{i=1}^N X_i'\right)\\ &\leq E_{X'}\exp\lambda\left(\sum_{i=1}^N X_i - \sum_{i=1}^N X_i'\right) \quad \text{(Jensen's inequality)}\\ &= E_{X'}\exp(\lambda\left(\sum_{i=1}^N (X_i - X_i')\right).\end{aligned}$$

Take expectation with respect to E_X to have

$$\begin{aligned}
E\exp\lambda\left(\sum_{i=1}^N X_i\right) &\le E\exp\lambda\left(\sum_{i=1}^N (X_i - X_i')\right)\\
&= E\exp\lambda\left(\sum_{i=1}^N \epsilon_i(X_i - X_i')\right) \quad (X_i - X_i' \text{ and } \epsilon_i(X_i - X_i') \text{ share the same distribution})\\
&= E\left[E\left(\exp\lambda\left(\sum_{i=1}^N \epsilon_i X_i\right)\Big|\epsilon\right) \times E\left(\exp\lambda\left(\sum_{i=1}^N -\epsilon_i X_i'\right)\Big|\epsilon\right)\right] \quad (\epsilon_i X_i \text{ and } \epsilon_i X_i' \text{ are independent cond}\\
&= \left[E\exp\lambda\left(\sum_{i=1}^N \epsilon_i X_i\right)\right]^2 \quad (-\epsilon_i X_i' \text{ and } \epsilon_i X_i \text{ share same distribution})\\
&\le E2\lambda\left(\sum_{i=1}^N \epsilon_i X_i\right) \quad (x^2 \text{ is convex}).
\end{aligned}$$

Using similar tricks as above, one could prove a lower bound. Combined together to have

$$E\exp\left(\frac{1}{2}\lambda\sum_{i=1}^N \epsilon_i X_i\right) \le E\exp\left(\lambda\sum_{i=1}^N X_i\right) \le E\exp\left(2\lambda\sum_{i=1}^N \epsilon_i X_i\right).$$

By the equivalence of sub-Gaussian properties, the inequality above is all we need. ■

Remark 8.6 *The symmetrization result gives some bounds on the expectation. If we also have the concentration around the expectation, we could obtain high probability (lower and upper) bounds.*

8.5 Application: $E\|A\|_{op}$ without Sub-Gaussian Assumption

The target is to prove the following theorem.

Theorem 8.7 *[Operator Norm of Random Matrix with non-i.i.d. Entries] Let A be an $n \times n$ symmetric random matrix whose entries on and above the diagonal are independent, mean zero random variables. Then*

$$E\|A\|_{op} \le C\sqrt{\log n} \times E\max_i \|a_i\|_2,$$

where a_i denote the rows of A.

Let us first think about how good is the bound. Since

$$\|A\|_{op} = \max_{x\in S^{n-1}} \|Ax\|_2 = \max_{x\in S^{n-1}} \sqrt{\sum_i (a_i^T x)^2} \ge \max_{x\in S^{n-1}} |a_i^T x| = \|a_i\|_2 \quad \forall i,$$

we know that

$$E\|A\|_{op} \geq E \max_i \|a_i\|_2.$$

Thus, the bound is sharp up to the logarithmic factor.

8.5.1 Analysis

Lemma 8.2 tells us that we could approach this theorem using symmetrization. To accomplish this goal, we need to figure out two parts.

1. Is it possible to bound $E\|\sum_i \epsilon A_i\|_{op}$, where A_i are deterministic matrices?
2. How to decompose A into $\sum_i A_i$ to leverage the bound from part 1?

8.5.2 Bound for $E\|\sum_i \epsilon A_i\|_{op}$

This bound comes from the following two theorems that are for symmetric Bernoulli random variables. (You are not required to know how to prove them at all. They are not trivial, and the proofs involve the trace inequalities.)

Theorem 8.8 (Matrix Hoeffding's Inequality) *Let $\epsilon_1, \dots, \epsilon_N$ be independent symmetric Bernoulli random variables and let $A_1, \dots A_N$ be symmetric $n \times n$ matrices (deterministic). For any $t \geq 0$, we have*

$$P\left(\left\|\sum_{i=1}^N \epsilon_i A_i\right\|_{op} \geq t\right) \leq 2n \exp(-t^2/2\sigma^2),$$

where $\sigma^2 = \|\sum_{i=1}^N A_i^2\|_{op}$

Theorem 8.9 (Matrix Khintchine's inequality) *Let $\epsilon_1, \dots, \epsilon_N$ be independent symmetric Bernoulli random variables and let $A_1, \dots A_N$ be symmetric $n \times n$ $(n \geq 2)$ matrices (deterministic). Then we have*

$$E\left\|\sum_{i=1}^N \epsilon_i A_i\right\|_{op} \leq C\sqrt{\log n}\left\|\sum_{i=1}^N A_i^2\right\|_{op}^{1/2}$$

Denote $X = \left\|\sum_{i=1}^N \epsilon_i A_i\right\|_{op}$ and $\sigma^2 = \|\sum_{i=1}^N A_i^2\|_{op}$. Then the inequality becomes

$$EX \leq C(\sqrt{\log n})\sigma.$$

Coupled with the tail bound from Theorem 8.8, this becomes a standard problem: using tail bound to bound expectation. Since this I never show it rigorously before, I present a detailed proof here.

Proof: From Theorem 8.8, we know

$$P(X \geq t\sigma) \leq 2n\exp(-t^2/2).$$

To get rid of n in the probability bound, we utilize the following numeric inequality

$$x \leq a + (x-a)_+.$$

Plug in $x = X$, $a = \sqrt{2(\log n)\sigma}$ and then take expectation to have

$$\begin{aligned} EX &\leq \sqrt{2(\log n)}\sigma + E(X - \sqrt{2(\log n)}\sigma)_+ \\ &= \sqrt{2(\log n)}\sigma + \sqrt{2}\sigma E(X/\sqrt{2}\sigma - \sqrt{\log n})_+. \end{aligned}$$

Meanwhile, the second term could be calculated as following

$$\begin{aligned} \int_0^\infty P\left(\left(X/\sqrt{2}\sigma - \sqrt{\log n}\right) \geq t\right) dt &= \int_0^\infty P\left(X \geq \sqrt{2}\sigma(t + \sqrt{\log n})\right) dt \\ &\leq \int_0^\infty 2n\exp\left(-(t + \sqrt{\log n})^2\right) dt \\ &= \int_0^\infty 2n\exp\left(-(t^2 + 2(\sqrt{\log n})t + \log n)\right) dt \\ &= \int_0^\infty 2\exp\left(-(t^2 + 2(\sqrt{\log n})t)\right) dt \\ &\leq C. \quad \text{(Do the calculation!)} \end{aligned}$$

Then we have

$$EX \leq \sqrt{2(\log n)}\sigma + C\sqrt{2}\sigma \leq C'(\sqrt{\log n})\sigma$$

■

8.5.3 Decompose A into $\sum_i A_i$

Since we want to arrive at a_i somehow, it is tempting to decompose A into the summation of n separately rows. For example,

$$A_1 = \begin{pmatrix} a_1^T \\ 0 \end{pmatrix}$$

Unfortunately, this matrix is not symmetric, which means that we could not utilize Theorem 8.9. Additional to A_i being symmetric, we also hope that $\|\sum_i A_i^2\|_{op}$ is easy to calculate. In this case, we might want to have $\sum_i A_i^2$ diagonal, which implies its operator norm is the largest absolute value on the diagonal. To be more restrictive, we might want to have each term, namely A_i^2 diagonal.

Let us explore the possibility to have a symmetric matrix B such that B^2 is diagonal.

$$B_{ij} = \sum_k b_{ik}b_{kj} = 0 \quad \text{(When } i \neq j\text{)}.$$

The simplest case is that each term $b_{ik}b_{kj} = 0$ which implies that at least one of them is 0. Thus, if B only has element (i_0, j_0) and (j_0, i_0) nonzero and identical (to ensure B is symmetric matrix), we have B^2 diagonal. Let me give you an 3×3 matrix as an example.

$$B = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \Rightarrow \quad B^2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1. \end{pmatrix}$$

Thus, we could decompose A into $\sum_{i \leq j} A_{ij}$, where

$$A_{ij} = a_{ij}(e_i e_j^T + e_j e_i^T).$$

With some careful calculation, one could obtain

$$\sum_{i \leq j} A_{ij}^2 = \sum_{i=1}^{n} \|a_i\|_2^2 e_i e_i^T.$$

This implies that

$$\left\| \sum_{i \leq j} A_{ij}^2 \right\|_{op} = \max_i \|a_i\|_2^2.$$

Now, we have Theorem 8.7 proved.

8.6 Concentration for Lipschitz Function of $X \sim N(0, I_p)$

The proof of this concentration does not utilize symmetrization but tries to create an independent copy of X, denoted as Y. Then it utilizes a smooth transition from $F(X)$ to $F(Y)$ to obtain the result. Since the proof is so elegant, the transition idea is handy, and the result is significant, I present it here. Vershynin (2018, Chapter 5) also discuss the concentration of Lipschitz function of $U(S^{p-1})$ random vectors. They use isoperimetric inequalities, which are quite hard to understand for me, making me quite frustrated. I hope the proof below could ease your mind a little.

Theorem 8.10 *Let $X \sim N(0, I_p)$ and $F : \mathbb{R}^p \to \mathbb{R}$ be 1-Lipschitz function (i.e. $|F(x) - F(y)| \leq \|x - y\|_2$ for all $x, y \in \mathbb{R}^p$). Then for any $\lambda \geq 0$, one has*

$$P\left(|F(X) - EF(X)| \geq \lambda\right) \leq 2\exp(-C\lambda^2)$$

for some absolute constants C.

Before we move to the proofs, let us simplify the problem a little without losing generality. First of all, we could assume $EF(X) = 0$ because otherwise we could introduce $\tilde{F}(X) = F(X) - EF(X)$. Secondly, we could only require F to be differentiable. Otherwise, we could use a sequence of such functions to approximate F. For more details, you could refer to here. Then we know $\|\nabla F(x)\|_2 \leq 1$ because $F(x)$ is 1-Lipschitz.

Proof: Based on our previous experience of proving concentration, we only need to prove

$$E\exp(\lambda F(X)) \leq \exp(C\lambda^2), \quad \forall \lambda \in \mathbb{R}.$$

In fact, we only need to prove $\lambda \geq 0$ and $\lambda < 0$ follows an identical procedure.

Let Y be an independent copy of X. By Jensen inequality, we have

$$E\exp(\lambda(-F(Y))) \geq \exp(\lambda E(-F(Y))) = \exp(0) = 1.$$

Thus, we have

$$\begin{aligned} E\exp(\lambda(F(X)) &\leq E\exp(\lambda(F(X)))E\exp(\lambda(-F(Y))) \\ &= E\exp(\lambda(F(X) - F(Y))) \quad (X \text{ and } Y \text{ are independent}). \end{aligned}$$

If we could write $F(X) - F(Y)$ as $\int f(X_t)dt$, then we could use Jensen's inequality again as following,

$$E\exp(\lambda(F(X) - F(Y))) = E\exp(\lambda \int f(X_t)dt) \leq \int E\exp(\lambda f(X_t))dt,$$

which could give us an potential useful upper bound. **To use Jensen's inequality, we need to make $\int dt$ into a probability, namely $\int dt = 1$.**It is tempting to introduce $X_t = (1-t)X + tY$, then $X_0 = Y$ and $X_1 = X$,

$$F(X) - F(Y) = \int_0^1 \frac{dF(X_t)}{dt}dt = \int_0^1 \langle \nabla F(X_t), Y - X\rangle dt.$$

Unfortunately, it is not easy to calculate

$$E\exp(\lambda\langle \nabla F(X_t), Y - X\rangle),$$

because X_t is closely related to $Y - X$. Noticing that $Y - X = dX_t/dt$, we are trying to find a representation such that X_t and dX_t/dt are independent. Luckily, this is possible in the following way:

$$\begin{aligned} & X_\theta = \sin(\theta)X + \cos(\theta)Y \\ \Rightarrow \quad & X_0 = Y \;\&\; X_{\pi/2} = X \\ \Rightarrow \quad & \frac{dX_\theta}{d\theta} = \cos(\theta)X - \sin(\theta)Y. \quad (\text{For simplicity, I denote it as } X'_\theta) \end{aligned}$$

Because $\langle (\sin(\theta), \cos(\theta)), (\cos(\theta), -\sin(\theta)) \rangle = 0$, we have X_θ and X'_θ are independent. (Here we utilize the fact that X and Y are independent $N(0, I_p)$). Moreover, since $\sin^2(\theta) + \cos^2(\theta) = 1$, we have both X_θ and X'_θ follow $N(0, I_p)$. Now, we know how to bound

$$E \exp(\lambda \langle \nabla F(X_\theta), X'_\theta \rangle),$$

because conditional on X_θ, $\langle \nabla F(X_\theta), X'_\theta \rangle$ follows $N(0, \|\nabla F(X_\theta)\|_2^2)$. Since we know $\|\nabla F(x)\|_2 \leq 1$,

$$E \exp(\lambda \langle \nabla F(X_\theta), X'_\theta \rangle) = E \exp(\lambda^2 \|\nabla F(X_\theta)\|_2^2 / 2) \leq \exp(\lambda^2/2).$$

Combining things together, we have

$$\begin{aligned}
E \exp(\lambda(F(X) - F(Y))) &= E \exp(\lambda \int_0^{\pi/2} \langle \nabla F(X_\theta), X'_\theta \rangle dt) \\
&= E \exp(\lambda \frac{2}{\pi} \int_0^{\pi/2} \frac{\pi}{2} \langle \nabla F(X_\theta), X'_\theta \rangle dt) \quad \text{(Make sure that } \textstyle\int dt = 1 \text{ for Jensen's inequality)} \\
&\leq \frac{2}{\pi} \int_0^{\pi/2} E \exp(\lambda \frac{2}{\pi} \langle \nabla F(X_\theta), X'_\theta \rangle dt) \quad \text{(Jensen's Inequality)} \\
&\leq \frac{2}{\pi} \int_0^{\pi/2} \exp(\lambda^2 \pi^2 / 8) dt \quad \text{(Above inequality with } \lambda\pi/2\text{)} \\
&= \exp(\lambda^2 \pi^2 / 8)
\end{aligned}$$

■

Remark 8.11 *The above proof does not generate the best constant. To obtain it, one shall use log-Sobolev inequalities (entropy techniques). I will cover a tiny bit in the following section and please see van Handel (2014, Chapter 3) or Wainwright (2019, Chapter 3) for more details.*

8.7 Entropy Method (log-Soblev) [Optional]

For any convex function $\phi : \mathbb{R} \to \mathbb{R}$, we could define the following quantity of a random variance $X \sim P$

$$\mathbb{H}_\phi(X) \triangleq E\phi(X) - \phi(EX).$$

Based on Jensen's inequality, we have $\mathbb{H}_\phi(X) \geq 0$. If we choose $\phi(u) = u^2$, then we have

$$\mathbb{H}_\phi(X) = EX^2 - (EX)^2 = Var(X).$$

Another choice is $\phi(u) = -\log u$, then

$$\mathbb{H}_\phi(\exp(\lambda X)) = -\lambda EX + \log E\exp(\lambda X) = \log\Big(E\exp(\lambda(X - EX))\Big).$$

An equivalent property of sub-Gaussian is about the quantity within the log function:

$$E\exp(\lambda(X - EX)) \leq \exp(C\lambda^2).$$

Moreover, Hoeffding's lemma is also about this quantity.

The Entropy method uses $\phi(u) = u\log u$, this leads to

$$\mathbb{H}_\phi(\exp(\lambda X)) = E(\lambda X\exp(\lambda X)) - E\exp(\lambda X)\log E\exp(\lambda X). \tag{8.1}$$

My intuition is that one should not gain much using $u\log u$ instead of $-\log u$ since the first term of Eq. 8.1 is still very hard to handle. However, a simple trick on the second term makes a huge difference.[1] I will illustrate this point using the following Lemma

Lemma 8.12 (Entropy Bound for Univariate Functions) *Let $X, Y \sim P$ be a pair of i.i.d. variates. Then for any function $g : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{H}(\exp(\lambda g(X))) \leq \lambda^2 E\Big((g(X) - g(Y))^2\exp(\lambda g(X))1_{g(X)\geq g(Y)}\Big) \quad \forall\lambda > 0.$$

If one uses this Lemma coupled with the Lemma about Tensorization of Entropy (roughly speaking, the entropy of $\exp(\lambda f(X_1, \dots, X_n))$ is bounded by a summation of n univariate entropies.), he/she could obtain some concentration result of L-Lipschitz functions of $(X_1, \dots, X_n)$.

The following proof is straightforward because we know that creating an independent copy of X is necessary. In some sense, this shows the power of introducing such an independent copy.

[1] This is also a surprise to me. I used to think the second term does not matter much.

Proof:

$$
\begin{aligned}
\mathbb{H}(\exp(\lambda g(X))) &= E\Big(\lambda g(X)\exp(\lambda g(X))\Big) - E\exp(\lambda g(X))\log E\exp(\lambda g(X)) \quad \text{(Definiton of } \mathbb{H}(\cdot))\\
&= E_X\Big(\lambda g(X)\exp(\lambda g(X))\Big) - E_X\exp(\lambda g(X))\log E_Y\exp(\lambda g(Y)) \quad (Y\sim X)\\
&\le E_X\Big(\lambda g(X)\exp(\lambda g(X))\Big) - E_X\exp(\lambda g(X))E_Y\lambda g(Y) \quad (-\log(x) \text{ is convex })\\
&= E_X\Big(\lambda g(X)\exp(\lambda g(X))\Big) - E_{X,Y}\Big(\lambda g(Y)\exp(\lambda g(X))\Big) \quad (X,Y \text{ are independent})\\
&= E_{X,Y}\Big(\lambda(g(X)-g(Y))\exp(\lambda g(X))\Big)\\
&= \frac{1}{2}E_{X,Y}\Big(\lambda(g(X)-g(Y))[\exp(\lambda g(X))-\exp(\lambda g(Y))]\Big) \quad (\text{Switch } X \text{ and } Y)\\
&= E_{X,Y}\Big(\lambda(g(X)-g(Y))\big[\exp(\lambda g(X))-\exp(\lambda g(Y))\big]1_{g(X)\ge g(Y)}\Big) \quad \text{(Symmetry, think!)}
\end{aligned}
$$

Noticing that $\exp(x)$ is convex, we have

$$\frac{\exp(s)-\exp(t)}{s-t} \le \exp(s) \quad s \ge t.$$

(Convexity tells us that the slope of a left-hand-side secant is smaller than the tangent.) Multiply both sides by $(s-t)^2$ to obtain

$$(s-t)(\exp(s)-\exp(t)) \le \exp(s)(s-t)^2.$$

Apply this inequality to $g(X)$ and $g(Y)$ to obtain the result. ■

8.8 Related Form: Gaussian Multiplier [Optional]

The Gaussian multiplier is to multiply X_i by g_i where $g_i \sim N(0,1)$. Gaussian multiplier coupled with bootstrapping has become a popular hypothesis test tool, especially since Chernozhukov et al. (2017). I will use the mean-shifting hypothesis test to demonstrate the idea.

Here is a mean-shifting **vector** **(**$\in \mathbb{R}^p$**)** model with a shift of size δ_n happening at period m,

$$X_i = \mu + \delta_n 1_{i>m} + \xi_i, \quad i = 1,\dots n.$$

Here, ξ_i is i.i.d. mean-zero noise random vectors with unknown covariance structure Σ. If data follow this model, we are curious about whether there is a change (is δ_n not equal to 0). If there is one, where does it happen (what is m?). Statistically speaking, we want to first have a way to test

$$H_0 : \delta_n = 0 \quad v.s. \quad H_a : \delta_n \ne 0 \text{ and there exists an } m < n.$$

If we reject H_0, we want to have a good estimation of m. This hypothesis test tries to answer the stationarity question with a simple mean-shifting model. Since stationarity is assumed[2] in lots of applications, answering this question is of great value.

For simplicity, let's assume $p = 1$ to get some insights. If $\xi_i = 0$, it will be trivial because we see a jump in data. Since ξ_i is i.i.d. mean-zero, we might hope LLN or concentration kicks in when we are thinking about

$$\frac{1}{s}\sum_{i=1}^{s} X_i \quad \text{and} \quad \frac{1}{n-s}\sum_{i=s+1}^{n} X_i.$$

If there is a big difference between the two averages, we are confident that there is a shift and the shift is likely to happen at the largest difference point. The latter intuition needs some technical adjustment because the variance of the mean difference at different s could be drastically different. The following quantity handles this issue,

$$Z_n(s) = \sqrt{\frac{s(n-s)}{n}} \left(\frac{1}{s}\sum_{i=1}^{s} X_i - \frac{1}{n-s}\sum_{i=s+1}^{n} X_i \right).$$

When p fixed and $n \to \infty$, CLT will kick in for H_0 and one could do hypothesis test (which uses Σ^{-1}). $Z_n(s)$ is called CUSUM (cumulative sum) statistics.

When $p \gg n$, things become much more interesting. Even if we assume ξ_i is multivariate normal with an unknown covariance matrix, Σ, there is no way we could estimate Σ^{-1} (without strong structure assumptions like sparse or low rank) well enough to establish a good hypothesis test. One possibility is to bootstrap data from H_0 and build a confidence interval to get around the issue of estimating Σ. However, how could we bootstrap data from H_0? Here the Gaussian multiplier comes into play. $g_i X_i$ has constant mean $E(g_i X_i | X_i) = E(g_i) X_i = 0$. That is to say, given X_i, by multiplying different $g_i \sim N(0, 1)$, we approximately bootstrap data from H_0 and could build a confidence interval under H_0. The approximation comes from the fact that $g_i X_i$ given X_i does not share the covariance structure as X_i in H_0. Luckily, Yu and Chen (2020) show that this approximation error is mild.

I am incredibly passionate about Yu and Chen (2020), and the first author is a close friend of mine. If you are interested in research utilizing this method, please let me know.

References

Chernozhukov V, Chetverikov D, Kato K, et al. (2017) Central limit theorems and bootstrap in high dimensions. *The Annals of Probability* 45(4):2309–2352.

Gao R (2020) Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382* .

van Handel R (2014) Probability in high dimension. Technical report, PRINCETON UNIV NJ, URL https://web.math.princeton.edu/~rvan/APC550.pdf.

[2] For example, if one uses sample mean to estimate the expected value, one assumes stationarity in mean.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

Yu M, Chen X (2020) Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 9: Lower Bound of $E\sup_{t\in T} X_t$

Lecturer: Long Zhao, longzhao@nus.edu.sg

Because I treat statistical learning as an application of general techniques, you might find it unsatisfactory. You might want to take a look at Raklin (2020) which provides a systematic view of statistical learning. Here are some wonderful talks by him (Robustness, Stochastics, Uncertainty). Here is the MIT course Statistical Learning Theory and Application.

9.1 Resources

- van Handel (2014, Chapter 6.1). Easiest to read among three. The idea is super clear.
- Vershynin (2018, Chapter 7). Detailed approach with some examples.
- Wainwright (2019, Chapter 5.4). Too brief for a beginner. Try to read it at last.

9.2 Roadmap

In the last lecture, we focus on providing an upper bound for $E\sup_{t\in T} X_t$ using chaining when X_t has sub-Gaussian increments. In this lecture, we try to get a lower bound. It is important because together with the upper bound, we get a complete understanding of the quantity. More specifically,

1. Some might argue that the upper bound is loose. Therefore, the theoretical analysis is not insightful. A large lower bound renders this argument infeasible.
2. Sometimes (mainly in CS and statistics), it is important to show how good is the upper bound. If we have a lower bound comparable to the upper bound (up to a constant or log), we could establish that the upper bound is extremely good, indicating an excellent understanding of the underlying quantity.

Let $P(T, d, \epsilon)$ be a maximal ϵ-separated set of T with distance d. For simplicity, I will use $P(\epsilon)$ instead because T and d will be fixed. Naturally, we have the following lower bound

$$\sup_{t\in T} X_t \geq \sup_{t\in P(\epsilon)} X_t.$$

There are two great things about the RHS.

1. It is a maximal of **finite** random variables.

2. Those random variables **might** not be highly correlated, since

$$\|X_t - X_s\|_{\psi_2} \leq Kd(t,s), \quad \text{(sub-Gaussian increments)} \tag{9.1}$$

   and for any $t, s \in P(\epsilon)$, we have $d(t,s) \geq \epsilon$. It is worth noticing that this might not hold for ϵ-net.

There are also two bad news.

1. Even there is only finite events, we have no tools to establish a lower bound. The union bound could only provide an upper bound.
2. Eq. 9.1 is an inequality which means that there is no guarantee that X_t and X_s are not highly correlated given $d(t,s) \geq \epsilon$.

Thus, it should be extremely difficult to establish a lower bound for processes with sub-Gaussian increments. However, it might be possible for Gaussian process (I will formally define it later) since we have

$$X_t - X_s \sim N(0, E(X_t - X_s)^2),$$

which resolves[1] the second concern if we use $d(t,s) = \sqrt{E(X_t - X_s)^2}$. To handle the first issue, we need to recall that long long time ago, when we talk about how tight the union bound is, we actually develop a lower bound for **independent** events.

$$P(\cup_{i=1}^n A_i) \geq (1 - e^{-1})\left(1 \wedge \sum_{i=1}^n P(A_i)\right), \quad \text{(See A Note on Union Bound)}.$$

Thus, if we could lower bound $\sup_{t\in P(\epsilon)} X_t$ by the supreme of **finite independent** random variables, then it is possible to achieve a lower bound for $\sup_{t\in T} X_t$. It turns out that this is possible for Gaussian processes and the tools are called Gaussian comparison inequalities. Roughly speaking, these inequalities guarantees that

$$E \sup_{t\in P(\epsilon)} X_t \geq E \sup_{i=1\ldots|P(\epsilon)|} Z_i, \tag{9.2}$$

where Z_i are i.i.d. and follow $N(0, \epsilon^2/2)$. The following Lemma, establish an explicit lower bound of the RHS.

Lemma 9.1 (Maxima of i.i.d. $N(0,\sigma^2)$) *Let $X_1, \ldots, X_n$ be i.i.d. $N(0,\sigma^2)$ random variables, then*

$$E\left[\max_{i=1\ldots,n} X_i\right] \geq c\sigma\sqrt{\log n}.$$

The proof will leverage the pdf of Normal distribution, one could refer to van Handel (2014, Lemma 6.4)

[1] This is a subtle point. We use $\|\cdot\|_{\psi_2}$ to describe the tail behavior if we do not know the distribution. If we know it is normal distribution, there is no need to use $\|\cdot\|_{\psi_2}$.

for a simple proof with bad constant c or van Handel (2014, Problem 5.1) for a careful proof with a good constant c.

If we apply Lemma 9.1 to Eq. 9.2, we could have

$$E \sup_{i=1\ldots|P(\epsilon)|} Z_i \geq c\epsilon\sqrt{\log|P(\epsilon)|} \geq c\epsilon\sqrt{\log|N(\epsilon)|}.$$

Since the above inequality holds for any ϵ, we have

$$E\sup_{t\in T} X_t \geq c\max_{\epsilon>0} \epsilon\sqrt{\log|N(\epsilon)|}.$$

This lower bound is called Sudakov's inequality. Combining the result from chaining, we know that

$$c\max_{\epsilon>0} \epsilon\sqrt{\log|N(\epsilon)|} \leq E\sup_{t\in T} X_t \leq C\sum_k \sqrt{\log\left|N\left(\frac{\epsilon}{2^k}\right)\right|}\frac{\epsilon}{2^k}.$$

We could view the lower bound as the largest single term in the upper bound summation. If the summation behaves like a geometric series[2], then the upper and lower bound are of the same scale. The following visualization might be helpful.

9.3 Gaussian Process

Definition 9.2 (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 $n\geq 1$, $t_1,\cdots,t_n\in T$.*

Based on the above definition, we have (X_t, X_s) follows a joint normal distribution implying that $X_t - X_s$ is also a normal distribution. Thus, we have

$$X_t - X_s \sim N(0, E(X_t - X_s)^2),$$

which leads to a natural distance measure on T.

Definition 9.3 (Natural Distance) *A Gaussian process $\{X_t\}_{t\in T}$ is sub-Gaussian on (T,d) for the natural distance $d(t,s)=\sqrt{E(X_t-X_s)^2}$.*

This means that we could apply Dudley's inequality to the Gaussian process X_t regarding the natural distance.

[2] For example, $\sum_{k\geq 1} 2^{-k} = 1$ while largest term is 2^{-1}.

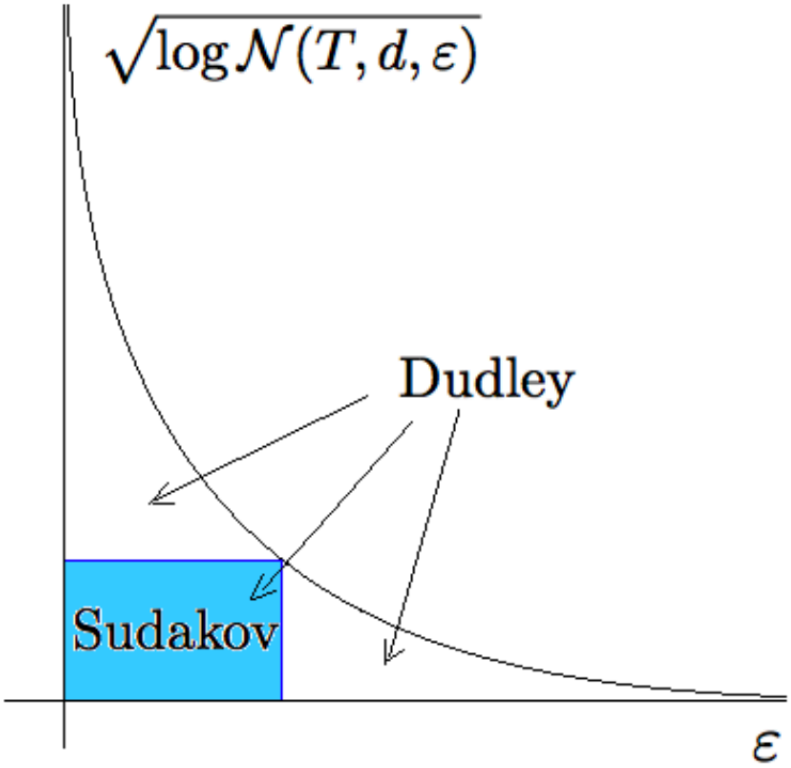


Figure 9.1: Sudakov's inequality v.s. Dudley's inequality.

9.4 Comparison Inequalities

The goal of comparison inequalities is to compare the sup of two Gaussian processes.

9.4.1 Simple Examples

One trivial comparison is that

$$X_t = \gamma Y_t \ (\gamma \geq 1) \Rightarrow E \sup_{t \in T} X_t \geq E \sup_{t \in T} Y_t.$$

The above inequality still holds for an independent copy of $\{Y_t\}_{t \in T}$ denoted as $\tilde{Y}_t$. The intuition is that if a Gaussian process is more volatile (higher variance), then the larger its sup will be.

The following example is less trivial. Let all X_is are independent and all Y_is are independent. Moreover, $X_i \sim N(0, \sigma^2_{X_i})$ and $Y_i \sim N(0, \sigma^2_{Y_i})$ where $\sigma^2_{X_i} \geq \sigma^2_{Y_i}$. That is to say, X_i has a higher variance than Y_i. Then it is reasonable to believe that

$$E \max_{i=1,\dots,n} X_i \geq E \max_{i=1,\dots,n} Y_i.$$

It is possible to prove by

$$P(\max_{i=1,\dots,n} X_i \leq \tau) \geq P(\max_{i=1,\dots,n} Y_i \leq \tau) \quad \forall \tau > 0.$$

The intuition is still that X_is are more likely to grab larger values than Y_is since they are more volatile.

9.4.2 General Case

The following theorem quantifies what means **more volatile** for Gaussian processes and confirms our intuition is indeed correct.

Theorem 9.4 (Slepian-Fernique) *Let* $X \sim N(0, \Sigma^X)$ *and* $Y \sim N(0, \Sigma^Y)$ *be* n*-dimensional Gaussian vectors. Suppose that we have*

$$E(X_i - X_j)^2 \geq E(Y_i - Y_j)^2 \quad \forall i, j = 1, \dots, n.$$

Then

$$E \max_{i=1,\dots,n} X_i \geq E \max_{i=1,\dots,n} Y_i.$$

The requirement is that any pair difference has a higher variance (more volatile). Before we prove the theorem, we would like to show two interesting applications of it.

9.5 Sudakov's Inequality

Although I have covered Sudakov's inequality in Section 9.2, it is beneficial to restate it formally as follows.

Theorem 9.5 (Sudakov) *For a Gausssian process* $\{X_t\}_{t\in T}$*, we have*

$$E\sup_{t\in T} X_t \geq c \sup_{\epsilon>0} \epsilon\sqrt{\log|N(\epsilon)|}$$

Proof: As analyzed in Section 9.2, we only need to show that

$$E \sup_{t\in P(\epsilon)} X_t \geq E \sup_{i=1,\dots,N} Y_i, \quad (N \triangleq |P(\epsilon)|) \tag{9.3}$$

where Y_is are i.i.d. normal distribution. To see the above inequality is a form of comparison inequality, we label the points in $P(\epsilon)$ as $t_1, \ldots, t_N$, then

$$\sup_{t\in P(\epsilon)} X_t = \sup_{i=1,\ldots,N} X_{t_i}.$$

To use Theorem 9.4, we need to calculate the pairwise variance

$$\begin{aligned} E(X_{t_i} - X_{t_j})^2 &\geq \epsilon^2 \quad \text{(Definition of } P(\epsilon)) \\ &= \epsilon^2/2 + \epsilon^2/2 = EY_i^2 + EY_j^2 \quad (Y_i\text{s are i.i.d.}) \\ &= E(Y_i - Y_j)^2. \end{aligned}$$

Thus, if $Y_i \sim N(0, \epsilon^2/2)$, we have Eq. 9.3 proved. Then we only need to use Lemma 9.1 to obtain the Sudakov's inequality. ■

9.6 Sharp Bounds on Gaussian Matrices [Optional]

In lecture 5, we proved a high-probability bound for $n \times p$ random matrices with independent, isotropic, sub-Gaussian rows. The bound we obtain there is

$$\sqrt{n} + C\sqrt{p}.$$

Next, we want to use Theorem 9.4 to show that $C = 1$ for Gaussian random matrices.

Theorem 9.6 (Operator Norms of Gaussian Random Matrices) *Let A be an $n \times p$ matrix with independent $N(0,1)$ entries. Then*

$$E\|A\|_{op} \leq \sqrt{n} + \sqrt{p}.$$

Remark 9.7 *Since $\|A\|_{op} \leq \|A\|_F = \sum_{i=1}^n \sum_{j=1}^p a_{ij}^2$ (largest singular value $\leq$ summation of all singular values), we know $\|\cdot\|_{op}$ (as a function of a_{ij}) is 1-Lipschitz. Since we already proved the concentration of Lipschitz functions of Gaussian random variables, we could establish tail bound for $\|A\|_{op}$ as following*

$$P(\|A\|_{op} \geq \sqrt{m} + \sqrt{n} + t) \leq 2\exp(-ct^2).$$

I will only highlight the core steps of the proof here. For a complete treatment, please see Vershynin (2018, Theorem 7.3.1.).

1. First of all, we need to connect $\|A\|_{op}$ to the supreme of a Gaussian process. This is promising from the following matrix equality.

$$\|A\|_{op} = \max_{u\in S^{p-1}, v\in S^{n-1}} \langle Au, v\rangle.$$

It is worth to notice that, since a_{ij} are independent $N(0,1)$, we have

$$\langle Au, v\rangle \sim N(0,1) \quad (\|u\|_2 = \|v\|_2 = 1).$$

Thus, if we define $X_{uv} \triangleq \langle Au, v\rangle$, we have

$$\|A\|_{op} = \sup_{uv\in T} X_{uv},$$

where $T = S^{p-1} \times S^{n-1}$, and X_{uv} is a Gaussian process (linear combinations of $N(0, I_{np})$ is still joint Gaussian).

2. It is tempting to use Theorem 9.4 to upper bound $\sup_{(u,v)\in T} X_{uv}$. However, the sup here is about uncountable many random variables while Theorem 9.4 is about finite ones. How to handle this issue? Recall that in chaining, we also face this problem. The idea is to use limit to move from finite to countable many and then use separable space to move to uncountable many. Luckily, these procedures still go through. Unluckily (for you), I will not present rigorous proof here. Please figure it yourself.

3. To use Theorem 9.4, we need to bound increments of X_{uv}. With careful analysis and bounding, we could prove

$$E(X_{uv} - X_{wz})^2 \le \|u-w\|_2^2 + \|v-z\|_2^2 \quad \text{(Non-trivial)}.$$

Proof: By definition, we have

$$X_{uv} = \langle Au, v\rangle = \sum_{ij} A_{ij}u_j v_i.$$

Thus,

$$\begin{aligned}
E(X_{uv} - X_{wz})^2 &= E(\sum_{ij} A_{ij}u_j v_i - w_j z_i)^2 \\
&= \sum_{ij}(u_j v_i - w_j z_i)^2 \quad (A_{ij} \text{ independent and } EA_{ij} = 0,\ EA_{ij}^2 = 1) \\
&= \sum_{ij} u_j^2 v_i^2 - 2\sum_{ij} u_j v_i w_j z_i + \sum_{ij} w_j^2 z_i^2 \\
&= \left(\sum_j u_j^2\right)\left(\sum_i v_i^2\right) - 2\left(\sum_j u_j w_j\right)\left(\sum_i v_i z_i\right) + \left(\sum_j w_j^2\right)\left(\sum_i z_i^2\right) \\
&= 2 - 2\left(\sum_j u_j w_j\right)\left(\sum_i v_i z_i\right) \quad (\|u\|_2 = \|v\|_2 = \|w\|_2 = \|z\|_2 = 1).
\end{aligned}$$

Denote

$$a \triangleq \sum_j u_j w_j \quad b \triangleq \sum_i v_i z_i \Rightarrow E(X_{uv} - X_{wz})^2 = 2 - 2ab$$

Then by Cauchy-Schwartz inequality we have

$$|a| \le \|u\|_2\|w\|_2 = 1 \quad |b| \le \|v\|_2\|z\|_2 = 1.$$

This implies that

$$\begin{aligned}
(1-a)(1-b) \ge 0 &\Rightarrow 1 - a - b + ab \ge 0 \\
&\Rightarrow 4 - 2a - 2b \ge 2 - 2ab \quad \text{(Multiply 2 to both sides and then rearrange)} \\
&\Rightarrow \|u\|_2^2 - 2u^T w + \|w\|_2 + \|v\|_2^2 - 2v^T z + \|z\|_2^2 \ge E(X_{uv} - X_{wz})^2 \\
&\quad (\|u\|_2 = \|v\|_2 = \|w\|_2 = \|z\|_2 = 1) \\
&\Rightarrow \|u - v\|_2^2 + \|v - z\|_2^2 \ge E(X_{uv} - X_{wz})^2.
\end{aligned}$$

■

4. Construct Gaussian process Y_{uv} such that

$$E(Y_{uv}-Y_{wz})^2 = \|u-w\|_2^2 + \|v-z\|_2^2.$$

This is possible by setting Y_{uv} as following

$$Y_{uv} \triangleq \langle g,u\rangle + \langle h,v\rangle,$$

where $g\sim N(0,I_p)$ and $h\sim N(0,I_n)$.

5. Use Theorem 9.4 to obtain the upper bound.

$$\begin{aligned}
E\|A\|_{op} &= E\sup_{(u,v)\in T} X_{uv} \le E\sup_{(u,v)\in T} Y_{uv}\\
&= E\sup_{u\in S^{p-1}}\langle g,u\rangle + E\sup_{v\in S^{n-1}}\langle h,v\rangle\\
&= E\|g\|_2 + E\|h\|_2\\
&\le \sqrt{E\|g\|_2^2} + \sqrt{E\|h\|_2^2} \quad \text{(Jensen's inequality)}\\
&= \sqrt{p}+\sqrt{n}.
\end{aligned}$$

9.7 Proof of Slepian-Fernique Theorem

The core idea is similar to our proof of concentration of Lipschitz function of Gaussian variables: interpolate from X to Y by introducing

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Y.$$

Then $Z(0) = X$ and $Z(1) = Y$. In this way, we translate a 'global' inequality (Theorem 9.4) into a local behavior

$$\frac{dE\max_{i=1,\dots,n} Z_i(t)}{dt} \ge 0,$$

which is potentially easier to prove leveraging the properties of normal distribution and calculus.

Intuitively speaking, if we want to show that position A is higher than position B, one way to do it is to prove something stronger: there exists a path from B to A such that the path is always going up. In this way, a global behavior (A is higher than B) becomes a local behavior (path is always up).

Remark 9.8 *You might wonder why we use $\sqrt{t}X$ instead of tX or $t^{1/3}X$. The reason is that*

$$\left.\frac{dZ(t)}{dt}\right|_{\textit{Coefficient of } X} \times \frac{dZ(t)}{dX} = \frac{1}{2\sqrt{t}}\times\sqrt{t} = \frac{1}{2}.$$

The RHS has nothing to do with t, which simplifies our proof. Recall our choice of $\cos(\theta)X+\sin(\theta)Y$ in the proof of concentration of Lipschitz function of Gaussian variables. Basically, we want to choose a specific interpolation that is easy to use.

We start with the simplest univariate normal distribution.

Lemma 9.9 *[Univariate Gaussian Integration by Parts]. Let $X \sim N(0,1)$. Then for any differentiable function $f:\mathbb{R}\to\mathbb{R}$ we have*

$$Ef'(X) = EXf(X).$$

Proof: Since we could extend the result from f with bounded support to arbitrary f using standard approximation argument[3], we only need to deal with f with bounded support.

Denote the density of $N(0,1)$ as

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

Then we could express $Ef'(X)$ as

$$\begin{aligned} Ef'(X) &= \int_{\mathbb{R}} f'(x)p(x)dx = -\int_{\mathbb{R}} f(x)p'(x)dx \quad \text{(Integral by parts + f bounded support)} \\ &= \int_{\mathbb{R}} xf(x)p(x)dx \quad (p'(x) = -xp(x)) \\ &= EXf(X). \end{aligned}$$

■

Lemma 9.10 (Multivariate Gaussian Integration by Parts) *Let $X \sim N(0,\Sigma)$. Then for any differentiable function $f:\mathbb{R}^n\to\mathbb{R}$ we have*

$$EXf(X) = \Sigma E\nabla f(X). \tag{9.4}$$

Proof: Noticing the equality above is a vector equality: it holds for each coordinate. That is to say, we want to prove

$$EX_i f(X) = (\Sigma E\nabla f(X))_i. \tag{9.5}$$

Noticing that other coordinates ($X_j,\ j\neq i$) are not independent from X_i, the LHS is still a multivariate integral. To leverage the result of univariate normal distribution, we have to introduce independence. This

[3] Terry Tao's blog 254A, Notes 2: The central limit theorem (1.Reductions) is an example of this argument.

leads to utilizing $Z \sim N(0, I_n)$ which implies $\Sigma^{1/2}Z \sim X \sim N(0,\Sigma)$. Moreover, define $g(Z) \triangleq f(\Sigma^{1/2}Z)$, then we only need to prove

$$EZ_i g(Z) = (\Sigma^{1/2} E\nabla f(\Sigma^{1/2}Z))_i \quad \text{(Multiplying } \Sigma^{-1/2} \text{ to Eq. 9.4)}.$$

For the LHS, we could conditional on $Z_j,\ j \neq i$ and utilize Lemma 9.9 to obtain

$$\begin{aligned} EZ_i g(Z) &= E\frac{\partial g(Z)}{\partial Z_i} = E\frac{\partial f(\Sigma^{1/2}Z)}{\partial Z_i} \quad \text{(Definition of } g) \\ &= (\Sigma^{1/2} E\nabla f(\Sigma^{1/2}Z))_i \quad \text{(Chain Rule)} \end{aligned}$$

■

Lemma 9.11 (Gaussian Interpolation) *Consider two independent Gaussian random vectors $X \sim N(0,\Sigma^X)$ and $Y \sim N(0,\Sigma^Y)$. Define the interpolation Gaussian vector*

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Y, \quad t \in [0,1].$$

Then for any twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\frac{dEf(Z(t))}{dt} = \frac{1}{2}\sum_{i,j=1}^n (\Sigma^X_{ij} - \Sigma^Y_{ij})E\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(t))\right].$$

Proof: Using the chain rule, we have

$$\begin{aligned} \frac{dEf(Z(t))}{dt} &= \sum_{i=1}^n E\frac{\partial f}{\partial x_i}(Z(t))\frac{dZ_i}{dt} \\ &= \frac{1}{2}\sum_{i=1}^n E\frac{\partial f}{\partial x_i}(Z(t))\left(\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}}\right). \end{aligned}$$

To utilize Lemma 9.10, we introduce

$$g_i(X) = \frac{\partial f}{\partial x_i}(\sqrt{t}X + \sqrt{1-t}Y),$$

which is a function of X conditional on Y. Since X and Y are independent, $X \sim N(0,\Sigma^X)$ conditional on

Y. Now, we have

$$\begin{aligned} E\frac{\partial f}{\partial x_i}(Z(t))X_i &= EX_i g_i(X) = (\Sigma^X E\nabla g_i(X))_i \quad \text{(Eq. 9.5)} \\ &= \sum_{j=1}^n \Sigma_{ij}^X E\frac{\partial g_i(X)}{\partial x_j} \quad \text{(Definition of matrix multiplication)} \\ &= \sum_{j=1}^n \Sigma_{ij}^X E\frac{\partial^2 f(Z(t))}{\partial x_i \partial x_j}\frac{\partial Z(t)}{\partial x_j} \quad \text{(Chain rule)} \\ &= \sum_{j=1}^n \Sigma_{ij}^X E\frac{\partial^2 f(Z(t))}{\partial x_i \partial x_j}\sqrt{t} \end{aligned}$$

This implies that

$$\sum_{i=1}^n E\frac{\partial f}{\partial x_i}(Z(t))\frac{X_i}{\sqrt{t}} = \sum_{i=1,j=1}^n \Sigma_{ij}^X E\frac{\partial^2 f(Z(t))}{\partial x_i \partial x_j}.$$

Do similar things to the term involving Y to have this Lemma proved.

■

All we need to do is to approximate $\max_{i=1,\ldots,n} X_i$ using twice differentiable functions. The following soft-max function is enticing

$$\max_{i=1,\ldots,n} X_i \le f_\lambda(X) \triangleq \frac{1}{\lambda}\log\sum_{i=1}^n \exp(\lambda X_i) \le \max_{i=1,\ldots n} X_i + \frac{\log n}{\lambda}.$$

Indeed, using Lemma 9.11, we could show that

$$\frac{dEf_\lambda(Z(t))}{dt} = \frac{\lambda}{4}\sum_{i\neq j}\left[E(X_i - X_j)^2 - E(Y_i - Y_j)^2\right] Ep_i(Z(t))p_j(Z(t)),$$

where

$$p_i(x) = \frac{\partial f_\lambda(x)}{\partial x_i} > 0.$$

Since the conditions of Theorem 9.4 are

$$E(X_i - X_j)^2 \ge E(Y_i - Y_j)^2 \quad \forall i, j,$$

we know that $\frac{dEf_\lambda(Z(t))}{dt} \ge 0$. This implies

$$Ef_\lambda(X) \ge Ef_\lambda(Y).$$

Taking $\lambda \to \infty$ and using dominated convergence theorem to change the order of limit and expectation to have

$$E \max_{i=1,\dots n} X_i \geq E \max_{i=1,\dots n} Y_i.$$

References

Raklin A (2020) Mathematical statistics: A non-asymptotic approach. URL https://www.mit.edu/~rakhlin/courses/mathstat/rakhlin_mathstat_sp20.pdf.

van Handel R (2014) Probability in high dimension. Technical report, PRINCETON UNIV NJ, URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 10: Compressed Sensing

Lecturer: Long Zhao

I skipped lots of proofs in this lecture note. They are not required for the test. I just want to show how all the things we have learned could be enhanced and then applied to a super complex problem.

10.1 Resources

- Vershynin (2018, Chapter 8-10). You are recommended to read it after this lecture. I will try my best to 'chain' the materials together: there are just too many things in these three chapters.
- Wainwright (2019, Chapter 7). It handles the noisy case very well. Here is his talk on this topic which might be more accessible.
- Compressed Sensing lecture from MIT 6.854 provides an alternative proof without heavy machinery like generic chaining.

10.2 Motivation

Compressed sensing could speed MRI[1] scan drastically (10 times faster). In this talk as well as this article, Prof. Donoho used compressed sensing to persuade the U.S. senate why mathematics is essential.

"The cost-benefit ratio of mathematical research has been off-scale. The federal government spends about \$250 million per year on mathematics research. Yet in the U.S., there are 40 million MRI scans per year, incurring tens of billions in Medicaid, Medicare, and other federal costs. The financial benefits of the roughly 10-to-1 productivity improvements now being seen in MRI could soon far exceed the annual NSF budget for mathematics research."

10.3 Problem

We want to recover $x^\star \in \mathbb{R}^p$ from some measurements that are linear combinations of $x^\star$, namely $a_i^T x^\star$. If there is no structure, then we need to have $n = p$ linearly independent a_is to obtain $x^\star$. Meanwhile, if we know $x^\star$ is sparse[2], namely $\|x^\star\|_0 = s \ll p$, is it possible to recover $x^\star$ using much fewer measurements?

[1] Here is an excellent video about how MRI works. Here is the reasoning why images should be sparse after some transformation.

[2] This talk nicely introduces the wavelet transformation and shows that images after wavelet transformation tend to be quite sparse.

For simplicity, we write all constraints $a_i^T x = y_i, \ i = 1, \ldots, n$ in a matrix form,

$$Ax = y.$$

If $n < p$, there are infinite solutions. Intuitively, we want to choose one with the lowest $\|x\|_0$. However, this is not computational feasible (NP- hard). One might consider a convex relaxation of $\|\cdot\|_0$ leading to the following convex optimization

$$\begin{aligned} \min_x \quad & \|x\|_1 \\ \text{subject to} \quad & Ax = y. \end{aligned} \tag{10.1}$$

Denote the solution to the above optimization problem as $\hat{x}$. the following Lemma shows that certain **geometric property** could lead to $\hat{x} = x^\star$.

Lemma 10.1 (Restricted Null-space Property) *Denote the null-space (or kernel) of A as $\ker(A)$ and define*

$$\mathbb{C}(S) = \{z \big| \ \|z_{S^c}\|_1 \leq \|z_S\|_1\},$$

where S is a s-element subset of $\{1, \ldots, p\}$ representing the support of $x^\star$. Moreover, z_S is a vector consisting of the coordinates belong to set S. If $\ker(A) \cap \mathbb{C}(S) = \{0\}$, we have $\hat{x} = x^\star$.

Remark 10.2 $\mathbb{C}(S)$ *is a cone. To see this, take $p = 2$ and $S = \{2\}$. Then*

$$\mathbb{C}(S) = \{z \big| \ |z_1| \leq |z_2|\}.$$

Moreover, if $z \in ker(A) \cap \mathbb{C}(S)$ and $z \neq 0$, we have $z/\|z\|_2 \in ker(A) \cap \mathbb{C}(S)$. Thus, $\ker(A) \cap \mathbb{C}(S) = \{0\}$ *is equivalent to* $\ker(A) \cap \mathbb{C}(S) \cap S^{p-1} = \emptyset$.

Remark 10.3 *In fact, we also have $\hat{x} = x^\star \Rightarrow$* $\ker(A) \cap \mathbb{C}(S) = \{0\}$. *This could be proved by construction. Please see Wainwright (2019, Page 202) for details.*

Proof: Let $h = \hat{x} - x^\star$. We only need to prove $h = 0$. Clearly $h \in \ker(A)$. Moreover,

$$\begin{aligned} \|h_{S^c}\|_1 &= \|\hat{x}_{S^c}\|_1 \quad (S \text{ is the support of } x^\star) \\ &= \|\hat{x}\|_1 - \|\hat{x}_S\|_1 \quad (\text{Definition of } \|\cdot\|_1) \\ &\leq \|x^\star\|_1 - \|\hat{x}_S\|_1 \quad (\hat{x} \text{ is optimal solution to Eq. 10.1 while } x^\star \text{ is feasible}) \\ &= \|x_S^\star\|_1 - \|\hat{x}_S\|_1 \quad (S \text{ is the support of } x^\star) \\ &\leq \|x_S^\star - \hat{x}_S\|_1 \quad (\text{Triangle inequality}) \\ &= \|h_S\|_1 \quad (\text{Definiton of } h). \end{aligned}$$

Thus, we have $h \in \mathbb{C}(S)$. Since $\ker(A) \cap \mathbb{C}(S) = \{0\}$, $h = 0$.

■

If we choose a fixed A matrix, then $\ker(A)$ will be fixed. Intuitively speaking, since $\ker(A)$ is large[3] when $n \ll p$, then one could adversarially choose S such that the restricted null-space property does not hold rendering $\hat{x} \neq x^\star$. Thus, our best bet might be a random A which might lead to $\hat{x} = x^\star$ with high probability even when $n \ll p$.

We want to choose just enough n such that the restricted null-space property holds with a high probability. In other words, we need to connect Ax with n somehow when $x \in S^{p-1}$.

10.4 Link Ax and n when $x \in S^{p-1}$

Notice that Ax is a length n random vector. Our goal now is to connect a random vector with its length. We actually have done something similar before. Do you recall anything?

Theorem 10.4 (Concentration of the Norm) *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, sub-Gaussian coordinates X_i that satisfy $EX_i^2 = 1$. Then*

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

where $K = \max_i \|X_i\|_{\psi_2}$.

To leverage the above theorem, we need to choose A such that each row, a_i^T is independent, sub-Gaussian, and isotropic. In this way,

$$E(a_i^T x)^2 = x^T E(a_i a_i^T) x = x^T x = \|x\|_2^2 = 1.$$

Thus, we have

$$\big\| \|Ax\|_2 - \sqrt{n} \big\|_{\psi_2} \leq CK^2,$$

where $K = \max_i \|a_i\|_{\psi_2}$. By the definition of $\| \cdot \|_{\psi_2}$, we have

$$P\left(\big| \|Ax\|_2 - \sqrt{n} \big| \leq CK^2 u\right) \geq 1 - 2\exp(-u^2) \tag{10.2}$$

is true for any **fixed or deterministic** $x \in S^{p-1}$. **The following argument is problematic,** but I still present here to motive why we need to introduce $\sup_{x \in T}$. Meanwhile, I hope you could be cautious and figure out why yourself.

[3] $\dim(\ker(A)) \geq p - n$

10.4.1 Wrong Argument

Denote event $\ker(A) \cap S^{p-1} \neq \emptyset$ as B, and then decompose the LHS into two cases,

$$\begin{aligned} P(\big|\|Ax\|_2 - \sqrt{n}\big| \leq CK^2u) = P(\big|\|Ax\|_2 - \sqrt{n}\big| \leq CK^2u \text{ and } B)+ \\ P(\big|\|Ax\|_2 - \sqrt{n}\big| \leq CK^2u \text{ and } B^c) \end{aligned}$$

If $\ker(A) \cap S^{p-1} \neq \emptyset$, we could choose $\tilde{x} \in \ker(A) \cap S^{p-1}$ which has $\|A\tilde{x}\|_2 = 0$ and $\|\tilde{x}\|_2 = 1$. Replacing x with $\tilde{x}$ in the above inequality and then the first term becomes

$$P(\big|\|A\tilde{x}\|_2 - \sqrt{n}\big| \leq CK^2u \text{ and } B) \leq P(\sqrt{n} \leq CK^2u \text{ and } B).$$

It is worth to notice that the $\sqrt{n} \leq CK^2u$ has no randomness. That is to say, if we choose $n > C^2K^4u^2$, then $P(\sqrt{n} \leq CK^2u \text{ and } B) = 0$ resulting in $P(||\|A\tilde{x}\|_2 - \sqrt{n}| \leq CK^2u \text{ and } B) = 0$. Then we know

$$\begin{aligned} P(B^c) &\geq P(\big|\|A\tilde{x}\|_2 - \sqrt{n}\big| \leq CK^2u \text{ and } B^c) \quad (\text{Any event } A \text{ and } B,\ P(A) \geq P(A \cap B)) \\ &= P(\big|\|A\tilde{x}\|_2 - \sqrt{n}\big| \leq CK^2u) \quad (P(\big|\|A\tilde{x}\|_2 - \sqrt{n}\big| \leq CK^2u \text{ and } B) = 0) \\ &\geq 1 - 2\exp(-u^2) \quad (\text{Eq. } 10.2) \end{aligned}$$

Using the definition of event B, above inequality means that when $n > C^2K^4u^2$, we have $\ker(A) \cap S^{p-1} = \emptyset$ with at least $1 - 2\exp(-u^2)$ probability.

We 'proved' something unbelievable: $\ker(A)$, a non-degenerate linear space, does not intersect with S^{p-1} with high probability! This is just impossible. Where is the problem?

10.4.2 The Issue and Its Solution

The concentration (Eq. 10.2) is only valid for deterministic $x \in S^{p-1}$. Is $\tilde{x}$ deterministic? Or is $\ker(A)$ deterministic for a random matrix A? Clearly not! Thus, we could not replace x by $\tilde{x}$ and proceed.

We have encountered similar issues before, could you recall? In the case of excess risk, the optimal parameter that minimizes the empirical risk is random (because data are random). Then we utilize a uniform bound $\sup_{x \in T}$ to control this randomness. We will do the same thing here. Let $T \triangleq \mathbb{C}(S) \cap S^{p-1}$, then $\forall \tilde{x} \in \ker(A) \cap T)$, we have

$$\sqrt{n} = ||\|A\tilde{x}\|_2 - \sqrt{n}| \leq \sup_{x \in T} ||\|Ax\|_2 - \sqrt{n}| = \sup_{x \in T} |X_x|,$$

where $X_x \triangleq \|Ax\|_2 - \sqrt{n}$. If we could get a high-probability bound of $\sup_{x \in T} |X_x|$, then we could proceed like Section 10.4.1 and get a large enough n such that $T \cap \ker(A) = \emptyset$ with high probability. Here is the road-map forward.

1. X_x has sub-Gaussian increments. We only know how to control $\sup_{x \in T} X_x$ when this property is true.

Thus, the first step is to show that X_x indeed has this property.

2. Generic chaining. Now, it is tempting to use Dudley's inequality. Unfortunately, It is not easy to calculate $N(\epsilon)$. Instead, we will sketch the idea of generic chaining, which improves Dudley's inequality. Although generic chaining involves another hard-to-compute quantity of T, the next step shows that it is possible to get around it.

3. A powerful comparison theorem. Amazingly, the generic chaining provides sharp upper and lower bounds for the Gaussian processes. This leads to a power comparison theorem that links X_x (hard to control) to another Gaussian process Y_x.

4. Finally, we need to resolve the unknown support of x, namely S, in $T = \mathbb{C}(S) \cap S^{p-1}$. We will get a larger set T_U containing all possible Ts. In this way, we could upper bound $\sup_{x\in T} |Y_x|$ by $\sup_{x\in T_U} |Y_x|$.

Both the tail and expectation bound utilize similar core ideas. However, the tail bound involves more careful treatment. **To highlight the main ideas, we will focus on expectation bound.**

10.5 X_x Has Sub-Gaussian Increments

Theorem 10.5 (Sub-Gaussian Increments) *Let A be an $n \times p$ matrix whose rows a_i^T are independent, isotropic and sub-Gaussian random vectors in $\mathbb{R}^p$. Then the random process X_x has sub-Gaussian increments, namely*

$$\|X_x - X_y\|_{\psi_2} \le CK^2 \|x - y\|_2 \quad \forall x, y \in \mathbb{R}^p.$$

Here $K = \max_i \|a_i\|_{\psi_2}$.

Remark 10.6 *This theorem is stronger than we need: it holds for any $x, y \in \mathbb{R}^p$ while we only need $x, y \in S^{p-1}$.*

This result is non-trivial. Even the sub-case, $x, y \in S^{p-1}$, takes three pages to prove. However, because the proof does not involve new tools, I will not present it here, and the proof is not required for the test. If you are interested, you could refer to Vershynin (2018, Chapter 9).

10.6 Generic Chaining

We could write the Dudley's inequality as

$$E \sup_{t\in T} X_t \lesssim \sum_{k=\kappa+1}^{\infty} \epsilon_{k-1} \sqrt{\log |T_k|}, \tag{10.3}$$

where $\epsilon_k = 2^{-k}$ and $|T_k| = N(t, d, \epsilon_k)$. It is fixing ϵ_k and operating with the ϵ-net. In generic chaining, we fix the cardinality of T_k and operate with the smallest possible ϵ_k. More specifically, we fix subsets $T_k \subset T$ such that

$$|T_0| = 1, \quad |T_k| \leq 2^{2^k}, \quad k = 1, 2, \ldots. \tag{10.4}$$

Such sequence of sets $(T_k)_{k=0}^{\infty}$ is called an *admissible sequence*. You might wonder why we choose 2^{2^k}, it is to make $\sqrt{\log |T_k|}$ into a geometric sequence which makes it possible to obtain a lower bound of $E \sup_{t \in T} X_t$ of the same order. Let

$$\epsilon_k = \sup_{t \in T} d(t, T_k),$$

where $d(t, T_k)$ is the distance from t to the set T_k. Then each T_k is an ϵ_k-net of T. Thus, we could write Eq. (10.3) as

$$E \sup_{t \in T} X_t \lesssim \sum_{k=\kappa+1}^{\infty} 2^{k/2} \sup_{t \in T} d(t, T_k). \tag{10.5}$$

The sharper bound is to pull the $\sup_{t \in T}$ out of the summation which leads to the following quantity.

Definition 10.7 (Talagrand's γ_2 Functional) *Let (T, d) be a metric space. Let $(T_k)_{k=0}^{\infty}$ be an admissible sequence (Eq. 10.4). The γ_2 functional of T is defined as*

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

Since γ_2 functional has $\sup_{t \in T}$ outside of summation, it is smaller than the Dudley's summation from the RHS of Eq. (10.3).

Theorem 10.8 (Generic Chaining Bound) *Let $(X_t)_{t \in T}$ be a mean zero random process on a metric space (T, d) with sub-Gaussian increments. Then*

$$E \sup_{t \in T} X_t \leq C K \gamma_2(T, d).$$

Remark 10.9 *Intuitively speaking, I do not know why generic chaining works better than chaining. Technically speaking, the generic chaining provides a more accurate bound for*

$$\left| X_{\pi^k(t)} - X_{\pi^{k-1}(t)} \right|.$$

Please feel free to share with me your intuition about this part. For the technical proof, please see Vershynin (2018, Chapter 8.5.2.).

Well, we put lots of effort into obtaining a tighter bound. However, calculating $\gamma_2(T, d)$ is not simpler than

$N(T, d, \epsilon)$ at all. How is generic chaining better suited for our purpose? The next section will demonstrate that we could use $\gamma_2(T, d)$ as an intermediate quantity to something that is much easier to calculate.

10.7 Talagrand's Comparison Inequality

Last lecture, we use Slepian-Fernique theorem to lower bound $\sup_{t\in T} Y_t$ for Gaussian process Y_t. The following theorem shows that the lower bound could be sharpened by $\gamma_2(T, d)$.

Theorem 10.10 (Talagrand's Majorizing Measure Theorem) *Let $(Y_t)_{t\in T}$ be a mean zero Gaussian process on a set T. Consider the canonical metric defined on T, i.e. $d_Y(t, s) = \sqrt{E(Y_t - Y_s)^2}$. Then*

$$c\gamma_2(T, d_Y) \leq E \sup_{t\in T} Y_t \leq C\gamma_2(T, d_Y).$$

The upper bound is given by Theorem 10.8. The lower bound is much harder to obtain (I do not know how to do it). As mentioned in Vershynin (2018), it is proved using 'a far-reaching, multi-scale strengthening of Sudakov's inequality'.

Notice that Theorem 10.8 holds for any process X_t that has sub-Gaussian increments w.r.t. some distance metric $d(\cdot, \cdot)$. If it happens to be $d_Y(\cdot, \cdot)$, we could link it to $E \sup_{t\in T} Y_t$ using Theorem 10.10 this leads to the following Corollary.

Corollary 10.11 (Talagrand's Comparison Inequality) *Let $(X_t)_{t\in T}$ be a mean zero random process on a set T and let $(Y_t)_{t\in T}$ be a mean zero Gaussian process. Assume that for all $t, s \in T$, we have*

$$\|X_t - X_s\|_{\psi_2} \leq K d_Y(t, s).$$

Then

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

We have made a great process towards an upper bound that is easy to calculate (Gaussian process now). We could make the upper bound even easier if we force

$$Y_t = \langle g, t\rangle,$$

where $g \sim N(0, I_p)$ and $t \in T \subset \mathbb{R}^p$. Then we have

$$\begin{aligned} d_Y(t, s) &= \sqrt{E(Y_t - Y_s)^2} = \sqrt{E\langle g, t - s\rangle^2} = \|t - s\|_2 \\ E \sup_{t\in T} Y_t &= E \sup_{t\in T} \langle g, t\rangle \triangleq w(T). \end{aligned}$$

$w(T)$ is the Gaussian-width of set T. This quantity is much easier to calculate. For example, we could take $T = B_1^p(r)$ defined as

$$B_1^p(1) \triangleq \{t \in \mathbb{R}^p \big| \, \|t\|_1 \leq r\},$$

namely, the L_1-ball with radius r. We have

$$\begin{aligned} E \sup_{t \in T} Y_t = E \sup_{t \in T} \langle g, t \rangle &\leq E\|g\|_\infty \|t\|_1 \quad \text{(Hölder's Inequality)} \\ &\leq rE \max_{i=1,\ldots,p} |g_i| \quad \text{(Definition of } T \text{ and } \|g\|_\infty) \\ &\leq Cr\sqrt{\log(p)} \quad \text{(Maximum of finite sub-Gaussian)} \end{aligned} \tag{10.6}$$

10.8 Handling Unknown S in $\mathbb{C}(S)$

Recall the definition of T in Section 10.4.2, we have

$$T = \mathbb{C}(S) \cap S^{p-1}$$

where S is the support of $x^\star$ and S^{p-1} is the unit sphere in $\mathbb{R}^p$. Unfortunately, we do not know S which means that we do not know T let alone taking a sup w.r.t. it. What we could do is to find T_U such that all possible T is a subset of it. The following lemma gives a candidate of such T_U.

Lemma 10.12

$$\|x\|_1 \leq 2\sqrt{s} \quad \forall x \in \mathbb{C}(S) \cap S^{p-1}$$

Proof: For any $x \in \mathbb{C}(S) \cap S^{p-1}$, we have

$$\begin{aligned} \|x\|_1 = \|x_S\|_1 + \|x_{S^c}\|_1 &\quad \text{(Definition of } \|\cdot\|_1) \\ \leq 2\|x_S\|_1 &\quad \text{(Definition of } \mathbb{C}(S)) \\ \leq 2\sqrt{s}\|x_S\|_2 &\quad (\|y\|_1 \leq \sqrt{s}\|y\|_2 \text{ for length } s \text{ vector}) \\ \leq 2\sqrt{s} &\quad (\|x_S\|_2 \leq \|x\|_2 = 1). \end{aligned}$$

■

Thus, we could use $T_U = B_1^p(2\sqrt{s})$ which leads to

$$w(T_U) = E \sup_{t \in T_U} Y_t \leq C\sqrt{s\log(p)} \quad \text{(Eq. 10.6)}.$$

If you believe me that there exists a tail version of Corollary 10.11 of similar form, then we have the following theorem.

Theorem 10.13 (Exact Sparse Recovery) *Suppose the rows a_i^T of A are independent, isotropic, and sub-Gaussian random vectors, and let $K \triangleq \max_i \|a_i\|_{\psi_2}$. If $n \geq CK^4 s \log p$, we have*

$$P(\hat{x} = x^\star) \geq 1 - 2\exp(-cm/K^4).$$

Since $s \ll p$, the number we needed is also much smaller than p. This shows the power of compressed sensing.

References

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

BDC6307: Introduction to Data Analytics **Spring 2021, NUS**

Lecture 11: Semicircle Law

Lecturer: Long Zhao

11.1 Resources

- van Handel (2014, Chapter 9.1). The technical part of this lecture closes follows this chapter.
- Benaych-Georges and Knowles (2016) addresses more advanced local law. Very mathematical reading.
- Random Matrices: Theory and Practice - Lecture 1 to 9 provides a thorough analysis of random matrix. Livan et al. (2018) is his book with coauthors on this topic.

11.2 Motivation

Eugene Wigner introduces random matrices to model the nuclei of heavy atoms. Because the number of atoms is enormous, and it is virtually impossible to calculate the levels of energy precisely and label them accordingly. Instead, Wigner takes a different perspective: he models these atoms' Hamiltonian using a random matrix whose entries are i.i.d. $N(0,1)$. Amazingly, this leads to a beautiful description of the corresponding eigenvalues (energy levels), namely, the semicircle law.

Since a random matrix could serve as the model for the data generating process, it has been used to understand behaviors in the high-dimensional setting. For example, Bloemendal et al. (2016), Bun et al. (2017), Johnstone and Paul (2018) utilize spike matrix, which relaxes the entry-wise independence assumption to investigate the principal components in the high-dimensional setting. Moreover, Wainwright (2019) uses Marchenko-Pastur law (a close cousin of semicircle law) as an inspiring example to highlight the importance of statistics in the high-dimensional setting. It has also been used in portfolio optimization to disentangle signals from pure noise Laloux et al. (1999, 2000), Plerou et al. (2002). I also use the delocalization phenomenon to justify a new dimension reduction tool for portfolio optimization.

It has also been used to model interactions within a complex system (May 1972); investigate properties of certain random graphs (Erdős et al. 2013).

11.3 Problem

Definition 11.1 *We call X a Wigner matrix if it satisfies the following property. X be an $n \times n$* **symmetric** *matrix whose entries X_{ij} are independent random variables with $EX_{ij} = 0$, $EX_{ij}^2 = 1$, and $E|X_{ij}|^3 \leq C$.*

We are interested in the eigenvalue distribution of X when n is large. First of all, just like CLT, we want to scale X correctly such that its eigenvalues will not explode as n goes to infinity. Notice that

$$E\sum_{i=1}^{n}\lambda_i^2 = ETr(X^TX) = E\sum_{ij}X_{ij}^2 = n^2.$$

Thus, we are expecting λ_i^2 is about n (dividing both sides by n) which means that $X/\sqrt{n}$ should be the correct scaling. Next, we define the spectral distribution of $X/\sqrt{n}$ as

$$\mu_n \triangleq E\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i(X/\sqrt{n})}\right).$$

It is worth noticing that the random variable inside the expectation is (almost surely) 1 at n points while 0 otherwise. Thus, taking expectation will smooth such a spiky random variable. Surprisingly, the limit distribution as $n \to \infty$ follows an unusual law.

Theorem 11.2 (Wigner's Semicircle Law) *μ_n converges in distribution to the Wigner's semicircle distribution*

$$\mu_{sc} \triangleq \frac{1}{2\pi}\sqrt{4-x^2}1_{|x|\le 2}dx.$$

Figure 11.1 shows the semicircle distribution. It is just amazing, isn't it? How is it possible? In this lecture, we will prove this law.

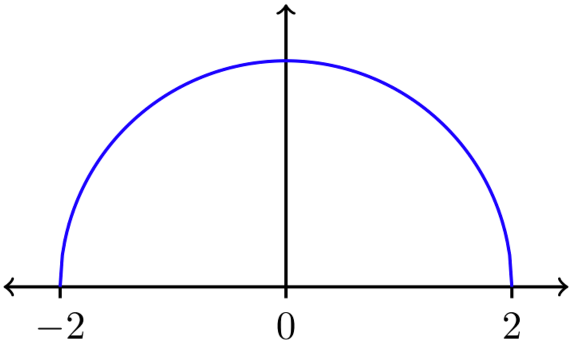


Figure 11.1: Semicircle Distribution

Theorem 11.2 is an asymptotic result and it could be strengthened into a non-asymptotic high probability description of both eigenvalues and eigenvectors[1].

[1] See Benaych-Georges and Knowles (2016) for more details. The proof is quite complicated, and I am still trying to figure out the magic.

Theorem 11.3 (Eigenvalue Rigidity) *With high probability,*

$$|\lambda_i - \gamma_i| \lesssim n^{-2/3}(i \wedge (n+1-i))^{-1/3},$$

where γ_i is the corresponding quantile from the semicircle distribution.

Theorem 11.4 (Complete Eigenvector Delocalization) *With high probability, the ith eigenvector u_i behaves like a random direction.*

To see why delocalization makes sense, let us think about random matrix Y with i.i.d. $N(0,1)$ entries. Based on the rotation invariant of multi-normal distribution, we know that Y is also distribution invariant under rotation (namely orthogonal matrix). Thus, the distribution of an eigenvector is also rotation invariant implying that it should come from the uniform distribution of a sphere. In other words, one eigenvector is a random direction.

In this lecture, we try to prove Theorem 11.2. There are four steps.

0. Briefly review complex numbers which facilitate the understanding of next step.
1. Translate convergence in distribution to some other convergence.
2. Prove the case for $Y_{ij} \sim N(0,1)$.
3. Use Lindeberg device to expand it to general case.

11.4 Review of Complex Numbers

When I was learning complex analysis, I got extremely frustrated because I rarely see complex numbers in real life except for Fourier transformation. To avoid causing you to feel the same way about the following lecture, I want to quote from Freeman Dyson, a Nobel Price Laureate of Physics, about complex numbers.

“ It turns out that the Schrödinger equation describes correctly everything we know about the behavior of atoms. It is the basis of all of chemistry and most of physics. And that **square root of minus one means that nature works with complex numbers and not with real numbers.**[2] ”

To show that fascinating things could happen in the complex world, let me prove the magic equality $\exp(\pi i) = -1$. Euler used this equality to prove that God exists (because it is so beautiful it must be created by God). The proof comes from $\exp(i\pi)$ via dynamics, in 3.14 minutes.

[2] This quote is from his famous birds and frogs article talking about two different types of researchers.

11.5 Green Function and Stieltjes Transform

In the proof of CLT, we translate convergence in distribution into point-wise convergence of the characteristic function. We also need something similar here. First of all, let us try the characteristic function. Will it work? Unfortunately, there is no linear relationship of the independent random variables that characteristic function could utilize. This leads to the Green function,

$$G(z,H) \triangleq (H-zI)^{-1},$$

where H is the matrix of interest and I is the identity matrix. Moreover, $z \in \mathbb{C}/\{\lambda_1(H),\ldots,\lambda_n(H)\}$. Notice that Green function is not a real-value function. Instead, it is a $n \times n$ matrix function! There are two great things regarding $G(z,H)$.

1. It is closely related to eigenvalues and eigenvectors. We could eigen decompose $H - zI$ as

   $$H - zI = \sum_{i=1}^{n} (\lambda_i - z) u_i u_i^T.$$

   This leads to

   $$(H-zI)^{-1} = \sum_{i=1}^{n} \frac{u_i u_i^T}{\lambda_i - z}.$$

   Let us think about the case $Re(z) = \lambda_i$ and $Im(z) \approx 0$. The ith term of the RHS will be much larger than other terms. This enable us to know the local behavior of λ_i and u_i. This is the foundation of the local laws from Benaych-Georges and Knowles (2016).

2. $\frac{\partial G(z,H)}{\partial H_{ij}}$ is easy to calculate. Since $I = G(z,H)(H-zI)$, we could take derivative w.r.t. H_{ij} to both sides to have

   $$0_{n\times n} = \frac{\partial G(z,H)}{\partial H_{ij}}(H-zI) + G(z,H)E_{ij},$$

   where $0_{n\times n}$ is a n-by-n matrix with all 0s and E_{ij} is a n-by-n with two 1 at position (i,j) and (j,i) while all other entries are 0s. Reorganize the equality to have

   $$\frac{\partial G(z,H)}{\partial H_{ij}} = -G(z,H)E_{ij}G(z,H).$$

   From here, it is easy to calculate $\frac{\partial^k G(z,H)}{\partial H_{ij}^k}$. You might wonder why do we need these derivatives[3]? There are roughly two parts. 1. If $X \sim N(0,1)$, we have $EXf(X) = Ef'(X)$ which involves f'. That is to say, we need this relationship when $X_{ij} \sim N(0,1)$; 2. In the Lindeberg device, we use Taylor expansion which uses derivatives to control the local difference.

[3] Here is a small catch. In the proof, we will handle $H = X/\sqrt{n}$ and the derivatives are w.r.t. X_{ij} instead of H_{ij}. However, chain rule tells us this is a trivial calculation once we know derivatives regarding H_{ij}.

It turns out that Green function $G(z,H)$ is overkill for Theorem 11.2: we could only use the expectation of $G(z,H)$'s trace. This leads to the following transform.

Definition 11.5 (Stieltjes Transform) *The Stieltjes transform S_μ of a probability measure μ on $\mathbb{R}$ is the function $S_\mu : \mathbb{C}/\mathbb{R} \to \mathbb{C}$ defined as*

$$S_\mu(z) \triangleq \int \frac{1}{u-z}\mu(du).$$

For the spectral distribution μ_n, its Stieltjes transformation is

$$\begin{aligned}\int \frac{1}{u-z}\mu_n(du) &= \int \frac{1}{u-z} E\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(X/\sqrt{n})}(du)\right) \quad \text{(Definition of } \mu_n\text{)}\\ &= \frac{1}{n} E \int \frac{1}{u-z} \sum_{i=1}^{n} \delta_{\lambda_i(X/\sqrt{n})}(du)\\ &= \frac{1}{n} E \sum_{i=1}^{n} \frac{1}{\lambda_i(X/\sqrt{n})-z}\\ &= \frac{1}{n} E\ Tr((X/\sqrt{n}-zI)^{-1})\\ &= \frac{1}{n} E\ Tr(G(z,H)) \quad (H = X/\sqrt{n}-zI).\end{aligned} \tag{11.1}$$

The following Lemma shows how to recover μ from $S_\mu(z)$.

Lemma 11.6 (Inversion Formula for Stieltjes Transformation) *For any bounded continuous function f*

$$\int f(x)\mu(dx) = \lim_{\epsilon\downarrow 0} \int \frac{1}{\pi} f(x) Im(S_\mu(x+i\epsilon))dx.$$

Proof: We could write the imaginary part of $(u-x-i\epsilon)^{-1}$ as

$$\frac{1}{\pi} Im\left(\frac{1}{u-x-i\epsilon}\right) = \frac{1}{\pi}\frac{\epsilon}{(u-x)^2+\epsilon^2} = \rho_\epsilon(x-u),$$

where $\rho_\epsilon(x)$ is the probability density function of the Cauchy distribution with mean 0 and scale parameter ϵ. Thus, we have

$$\begin{aligned}\int \frac{1}{\pi} f(x) Im(S_\mu(x+i\epsilon))dx &= \int f(x)\rho_\epsilon(x-u)dx\mu(du) \quad \text{(Definition of } S_\mu(z)\text{)}\\ &= Ef(X+Z_\epsilon),\end{aligned}$$

where $X \sim \mu$ and $Z_\epsilon \sim$ Cauchy$(0,\epsilon)$ are independent. Since $Z_\epsilon \to 0$ in probability and f is bounded, we know we could exchange the order of limit and expectation[4]. ■

Using the above Lemma, one could prove that point-wise convergence in Stieltjes transform is equivalent to convergence in distribution. Now we only need to prove that $S_{\mu_n}(z) \to S_{\mu_{sc}}(z)$ for all $z \in \mathbb{C}/\mathbb{R}$.

11.6 $Y_{ij} \sim N(0,1)$

We frequently use the following property of $N(0,1)$:

$$EZf(Z) = Ef'(Z). \tag{11.2}$$

This property comes handy for $G(z,H)$ since its partial derivatives w.r.t. H_{ij} are easy to calculate. To utilize this property, we need to create multiplication somehow. Luckily, there is a simple multiplication equality,

$$(H - zI)G(z,H) = (H - zI)(H - zI)^{-1} = I.$$

We could rearrange the equality as

$$HG(z,H) = zG(z,H) + I.$$

Taking trace and then expectation, we have

$$ETr(HG(z,H)) = zETr(G(z,H)) + n \quad (Tr(I) = n)$$

$$= znS_{\mu_n^N}(z) + n \quad (\text{Eq. } 11.1),$$

where we use μ_n^N to emphasize the $N(0,1)$ assumption. The above equality has two good things: 1. The LHS has multiplication of $HG(z,H)$ where we could apply property 11.2; 2. The RHS involves $S_{\mu_n^N}(z)$ which should converges to something denoted as $S(z)$. If LHS also converge to a function of S_μ, then we could solve the equation to obtain it. It turns out that the LHS converge to $-S(z)^2$ (see van Handel (2014) for more details) and we have

$$-S(z)^2 = 1 + S(z).$$

Solve it to obtain

$$S(z) = -\frac{z}{2} \pm \frac{1}{2}\sqrt{z^2 - 4}.$$

[4] One might use the last property of the list here.

Then we have

$$
\begin{aligned}
& Im(S(x+i\epsilon)) = -\frac{\epsilon}{2} \pm Im\sqrt{x^2-\epsilon^2+2i\epsilon-4} \\
\Rightarrow \quad & \lim_{\epsilon\downarrow 0} Im(S(x+i\epsilon)) = \pm\frac{1}{2}Im\sqrt{x^2-4} = \pm\frac{1}{2}\sqrt{4-x^2}1_{|x|\le 2}.
\end{aligned}
$$

Because $S(z)$ is the limit of $S_{\mu_n^N}(z)$ which is Stieltjes transformation of a probability, we must have $S(z)$ take the $+$ branch in the quadratic solution. This means that μ_n^N indeed converge to μ_{sc}.

11.7 Lindeberg Device

We have proved the semicircle law for matrix Y with i.i.d. $N(0,1)$ entries. We want to show that the difference between $S_{\mu_n^N}(z)$ and $S_{\mu_n}(z)$ goes to 0 as $n \to \infty$. Mathematically speaking, we care about

$$
S_{\mu_n^N} - S_{\mu_n}(z) = \frac{1}{n}ETr\Big(G(z, Y/\sqrt{n}) - G(z, X/\sqrt{n})\Big).
$$

It is tempting to use Lindeberg device to switch Y_{ij} to X_{ij} one by one. In this way, there are in total $n(n+1)/2$ terms (because of both are symmetric matrix). Thus, we need to make each term of order $1/n^{3/2}$. By utilizing and bounding $\frac{\partial^k G(z,H)}{\partial H_{ij}^k}$, we could show that it is achieved with two moments matching (mean and variance) and a finite third moments. Now, we could use Lindeberg device to prove that μ_n also converges to semicircle law. Again, for more details, please see van Handel (2014).

11.8 Marchenko-Pastur Law

Let the $n \times p$ matrix X be the data with i.i.d. entries[5] with mean 0 and variance $\sigma^2 < \infty$. Denote the sample covariance matrix as $\Sigma_n = \frac{1}{n}X^TX$. Then we have

$$
\begin{aligned}
E\Sigma_n &= \frac{1}{n}\sum_{i=1}^n Ex_ix_i^T \quad \text{(Definition of } \Sigma_n\text{, where } x_i \text{ is ith row)} \\
&= Ex_1x_1^T \quad (x_i \text{ share the same second moment}) \\
&= \sigma^2 I_p \quad \text{(Entry-wise independence and mean 0)}
\end{aligned}
$$

Marchenko-Pastur law states that when $p/n \to \alpha > 0$, the spectral distribution of Σ_n could be drastically different from σ^2.

Theorem 11.7 (Marchenko-Pastur Law) *Assume $p/n \to \alpha$, then the eigenvalue distribution of μ_m con-*

[5] The identical distribution assumption could be relaxed. Same first and second moment and bounded higher moments should be enough.

verges in distribution to the following distribution

$$\mu(dx) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{[(x-\gamma_-)(\gamma_+-x)]_+}}{\alpha x}dx + (1-\alpha^{-1})_+\delta_0(dx),$$

where $\gamma_\pm = \sigma^2(1\pm\sqrt{\alpha})^2$.

Figure 11.2 are borrowed from Wainwright (2019). When $\alpha = 0$, we are in the classical setting that p fixed and $n \to \infty$. By LLN, we have $\Sigma_n \to \sigma^2 I_p$ which implies the spectral distribution converges to point mass σ^2. Plugging $\alpha = 0$ to the Marchenko-Pastur law, one could also obtain$\gamma_\pm = \sigma^2$ and $\mu(dx) = \delta_{\sigma^2}(dx)$.

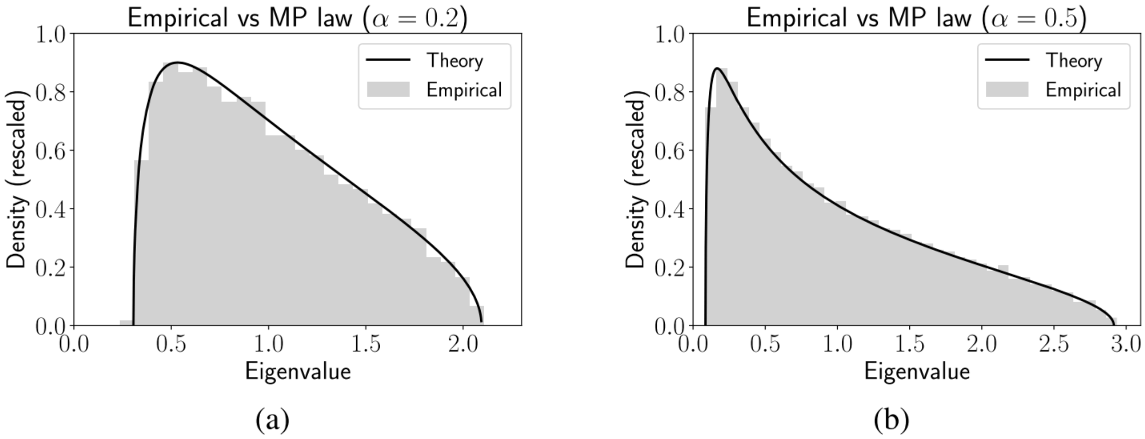


Figure 11.2: Marchenko-Pastur Law with $\alpha = 0.2$ and $\alpha = 0.5$.

The generating procedure of X contains almost no information since X_{ij} are entry-wise i.i.d. It is tempting to view the Marchenko-Pastur law as the spectral distribution of no information. Then any eigenvalues that significantly deviate from Marchenko-Pastur law could be seen as informative. This argument has been applied to the correlation matrix with some success (Laloux et al. 1999, 2000, Plerou et al. 2002).

In my opinion, X could serve as a model for non-stationarity: new data are generated from historical data plus X. This might lead to some interesting results.

11.9 Relaxing Entry-wise Independence

It is very hard to argue real world data has entry-wise independence. Thus, it is temping to relax it somehow. One possible way is to introduce several factors that link different entries. For example, the ith row x_i is generated as

$$x_i = z_i + \sum_{i=1}^{s_+} y_i f_i,$$

where z_i has entry-wise independence; f_i are deterministic vectors that link different entries and s_+ is the number of factors; y_i are random loadings of f_i that are also independent from z_i. This data generating process indicate the covariance matrix is a spike-matrix whose eigenvalues are all the same except s_+ ones. With this structure, one could still describe the behavior of eigenvalues and eigenvectors. For more details, please see Bloemendal et al. (2016).

References

Benaych-Georges F, Knowles A (2016) Lectures on the local semicircle law for wigner matrices. *arXiv preprint arXiv:1601.04055* .

Bloemendal A, Knowles A, Yau HT, Yin J (2016) On the principal components of sample covariance matrices. *Probability theory and related fields* 164(1-2):459–552.

Bun J, Bouchaud JP, Potters M (2017) Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports* 666:1–109.

Erdős L, Knowles A, Yau HT, Yin J, et al. (2013) Spectral statistics of erdős–rényi graphs i: Local semicircle law. *The Annals of Probability* 41(3B):2279–2375.

Johnstone IM, Paul D (2018) Pca in high dimensions: An orientation. *Proceedings of the IEEE* 106(8):1277–1292.

Laloux L, Cizeau P, Bouchaud JP, Potters M (1999) Noise dressing of financial correlation matrices. *Physical review letters* 83(7):1467.

Laloux L, Cizeau P, Potters M, Bouchaud JP (2000) Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance* 3(03):391–397.

Livan G, Novaes M, Vivo P (2018) *Introduction to random matrices: theory and practice*, volume 26 (Springer).

May RM (1972) Will a large complex system be stable? *Nature* 238(5364):413–414.

Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Guhr T, Stanley HE (2002) Random matrix approach to cross correlations in financial data. *Physical Review E* 65(6):066126.

van Handel R (2014) Probability in high dimension. Technical report, PRINCETON UNIV NJ, URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).