STAT210B - Theoretical Statistics

Berkeley University, Spring 2023
Prof. Song Mei

Last modified: September 15, 2023

Contents

1 Concentration Inequalities

Reference: Wainwright (2019), Ch. 2-3.

In this section we study concentration inequalities, which are probability statements about how random variables fluctuate around their means.

1.1 Bounds for Bounded Random Variables

The goal of this section is to study the *concentration phenomenon*, that is understand the behavior of $|X - \mu|$, where $X \sim \mathbb{P}_X$ and $\mu = \mathbb{E}_{X\sim\mathbb{P}_X}[X]$. As such we will study **tail bounds** of the form

$$\mathbb{P}_X(|X - \mu| \geq t) \leq g(t, n), \quad t > 0.$$

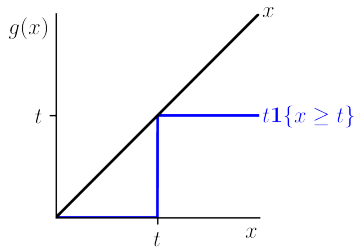
Proposition 1 (Markov's Inequality)**.** *Let X be a non-negative random variable. Then*

$$\forall\, t > 0, \qquad \mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Fix any $t > 0$. Then $X \geq t \cdot \mathbb{1}\{X \geq t\}$ almost surely. Hence, taking expectations of both sides

$$\mathbb{E}[X] \geq t \cdot \mathbb{E}[\mathbb{1}\{X \geq t\}] = t \cdot \mathbb{P}[X \geq t],$$

which was to be shown.



■

Note. *Note that even if $\mathbb{E}[X]$ is not finite the inequality above trivially holds.* ♦

Note. *In the following examples we will work with a sequence of iid random variables $Z_i, i = 1, \ldots, n, Z_i \sim \mathbb{P}_Z \in \mathcal{P}([0,1])$, where $\mathcal{P}([0,1])$ is the space of probability distributions with support over $[0,1]$. We use such space because it's convenient, as random variables with bounded support have all moments that are bounded.* ♦

Example 1 (Bound for the sample mean)**.** *Let's apply Markov's inequality to $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. We consider the non-negative random variable $|\overline{X} - \mu|$ and*

$$\mathbb{P}(|\overline{X} - \mu| \geq t) \leq \frac{\mathbb{E}[|\overline{X} - \mu|]}{t} \leq \frac{\sqrt{\mathbb{E}[(\overline{X} - \mu)^2]}}{t},$$

where the first inequality is the Markov's inequality and the second one is Jensen's inequality. To make the bound more explicit consider

$$\begin{aligned}\mathbb{E}[(\overline{X}-\mu)^2] &= \mathbb{E}[(\frac{1}{n}\sum_{i=1}^{n}(Z_i-\mu))^2] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[(Z_i-\mu)^2] + \underbrace{\frac{1}{n^2}\sum_{i\neq j}^{n}\mathbb{E}[(Z_i-\mu)(Z_j-\mu)]}_{=0,\ \textit{independence}} \\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}1 = \frac{1}{n}. \qquad (Z_i \text{ are bounded})\end{aligned}$$

Thus the bound becomes

$$\mathbb{P}(|\overline{X}-\mu|\geq t) \leq \frac{1}{\sqrt{n}\cdot t}. \tag{1.1}$$

We can now convert the bound in (1.1) into a statement of the form

$$|\overline{X}-\mu|\leq \kappa \qquad \textit{with probability at least } 1-\delta,$$

by simply requiring that $\frac{1}{\sqrt{n}t}=\delta$, solving for t, and plugging it back in (1.1). This yields

$$\kappa = \frac{1}{\sqrt{n}\delta} \quad \Longrightarrow \quad |\overline{X}-\mu|\leq \frac{1}{\sqrt{n}\delta} \qquad \textit{with probability at least } 1-\delta.$$

*Note that this statements holds for any n, thus it is a **finite sample** statement.* ♣ Next bound typically requires some extra assumptions, but is tighter.

Proposition 2 (Chebyshev's Inequality)**.** *Let X be a random variable with finite variance. Then*

$$\forall\, t>0, \quad \mathbb{P}(|X-\mathbb{E}[X]|\geq t) \leq \frac{\mathbb{V}(X)}{t^2}.$$

Proof. Apply Markov's Inequality

$$\mathbb{P}(|X-\mu|\geq t) = \mathbb{P}((X-\mu)^2\geq t^2) \leq \frac{\mathbb{E}[(X-\mu)^2]}{t^2}.$$

■

Example 2 (Bound for the sample mean II)**.** *If we apply Chebyshev's inequality to the previous example we get*

$$\mathbb{P}(|\overline{X}-\mu|\geq t) \leq \frac{1}{n\cdot t^2},$$

which implies

$$\kappa = \frac{1}{\sqrt{n\delta}} \quad \Longrightarrow \quad |\overline{X}-\mu|\leq \frac{1}{\sqrt{n\delta}} \qquad \textit{with probability at least } 1-\delta.$$

Note that Chebyshev's bound is tighter than Markov's as $\sqrt{\delta}\geq\delta$ since $\delta\in[0,1]$. ♣

Iterating up to the kth moment, we get the following bound.

Proposition 3 (kth moment inequality)**.** *If X is a random variable with $\mathbb{E}[|X|^k] < \infty$, then*

$$\forall\, t > 0, \quad \mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}[|X - \mu|^k]}{t^k}.$$

Example 3 (Bound for the sample mean III)**.** *Using a similar logic as before we can show that $\exists\, C_k \in (0, \infty)$ such that*

$$\mathbb{E}[|\overline{X} - \mu|^k] \leq \frac{C_k}{n^{k/2}},$$

and

$$|\overline{X} - \mu| \leq \frac{C_k}{\sqrt{n}\delta^{1/k}} \quad \textit{with probability at least } 1 - \delta.$$

♣

Another approach involves bounding the moment generating function

Proposition 4 (Chernoff's inequality)**.** *Let X be a random whose MGF exists at least in a neighborhood of 0. Then*

$$\forall\, t > 0, \quad \mathbb{P}(X \geq \mu + t) \leq \inf_{\lambda \geq 0} \{\mathbb{E}[e^{\lambda(X-\mu)}]/e^{\lambda t}\}.$$

Proof. By Markov's inequality, for $\lambda \geq 0$,

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}.$$

Since it holds for any $\lambda \geq 0$ it holds also for inf. The requirement $\lambda \geq 0$ is necessary to make the function $x \mapsto e^{\lambda x}$ non-decreasing.

Let $\tilde{\lambda} \leq 0$. Note also that

$$\mathbb{P}(X - \mu \leq -t) = \mathbb{P}(e^{\tilde{\lambda}(X-\mu)} \geq e^{-\tilde{\lambda}t}) \leq \inf_{\tilde{\lambda} \leq 0} \frac{\mathbb{E}[e^{\tilde{\lambda}(X-\mu)}]}{e^{-\tilde{\lambda}t}}$$

■

We will prove the next statement later on, when we show that all bounded random variables are sub-Gaussian.

Proposition 5 (MGF inequality)**.** *If $Z \sim \mathbb{P}_Z \in \mathcal{P}([0,1])$, then*

$$\forall\, \lambda \in \mathbb{R}, \quad \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq e^{\lambda^2/2}.$$

Example 4 (Bound for the sample mean IV)**.** *Going back to the previous example, we can apply Chernoff's inequality to get*

$$\mathbb{P}(\overline{X} - \mu \geq t) \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda(\overline{X}-\mu)}]}{e^{\lambda t}}.$$

Then,

$$\mathbb{E}[e^{\lambda(\overline{X}-\mu)}] = \mathbb{E}[e^{\frac{\lambda}{n}\sum_{i=1}^{n}(Z_i-\mu)}]$$

$$= \prod_{i=1}^{n} \mathbb{E}[e^{\frac{\lambda}{n}(Z_i-\mu)}] \qquad \text{(independence)}$$

$$\leq \prod_{i=1}^{n} e^{\frac{\lambda^2}{n^2 2}} = e^{\frac{\lambda^2}{2n}}. \qquad \text{(Proposition 4)}$$

Then

$$\mathbb{P}(\overline{X} - \mu \geq t) \leq \inf_{\lambda \geq 0} \{e^{\frac{\lambda^2}{2n} - \lambda t}\}$$

$$= e^{\frac{n^2 t^2}{2n} - nt^2} = e^{-\frac{nt^2}{2}}. \qquad (\lambda^\star = nt)$$

Similarly

$$\mathbb{P}(X - \mu \leq -t) \leq e^{-\frac{nt^2}{2}}.$$

Finally, by a union bound we have

$$\mathbb{P}(|\overline{X} - \mu| \geq t) = \mathbb{P}(\{\overline{X} - \mu \leq -t\} \cup \{\overline{X} - \mu \geq t\}) \leq 2e^{-\frac{nt^2}{2}}.$$

Then we get $t^\star = \sqrt{\frac{2\log(2/\delta)}{n}}$ *and*

$$\mathbb{P}(|\overline{X} - \mu| \geq t) \leq \sqrt{\frac{2\log(2/\delta)}{n}} \quad \textit{with probability at least } 1 - \delta.$$

♣

We can now compare the four inequalities we've seen so far.

Table 1: *High-Probability Bounds on* $|\overline{X} - \mu|$.

	Markov	Chebyshev	kth moment	Chernoff
require	$\mathbb{E}[X]$	$\mathbb{E}[X^2]$	$\mathbb{E}[X^k]$	MGF
bound	$\frac{1}{\sqrt{n}\cdot\delta}$	$\frac{1}{\sqrt{n}\cdot\sqrt{\delta}}$	$\frac{1}{\sqrt{n}\cdot\delta^{1/k}}$	$\frac{\sqrt{2\log(2/\delta)}}{\sqrt{n}}$

Note. *All the bounds share the same dependence in n but possess different dependencies in δ. The more you assume, the tighter the bound. Say $\delta = 0.001$, then $1/\delta = 1000$, whilst $\sqrt{2\log(2/\delta)} \approx 4$.*

The moment bound with an optimal choice of k is ***never worse*** *than the Chernoff's bound based on the moment generating function. Nonetheless, the Chernoff bound is most widely used in practice, possibly due to the ease of manipulating moment generating functions.* ♦

Example 5 (Dependence in δ)**.** *Consider the case of* $\{Z_i^{(s)}\}_{i\in[n],s\in[d]}$, $Z_i^{(s)} \overset{iid}{\sim} \mathbb{P}_Z^{(s)} \in \mathcal{P}([0,1])$. *Suppose our goal is to bound*

$$\sup_{s\in[d]} |\overline{X}^{(s)} - \mu_s|.$$

Then, we can apply first a union bound and then Markov's inequality, i.e.,

$$\begin{aligned}\mathbb{P}(\sup_{s\in[d]}|\overline{X}^{(s)}-\mu_s|\geq t) &\leq \sum_{s\in[d]}\mathbb{P}(|\overline{X}^{(s)}-\mu_s|\geq t) && \text{(union bound)}\\ &\leq d\cdot\frac{1}{t\sqrt{n}}. && \text{(Markov's inequality)}\end{aligned}$$

Alternatively, we can apply Chernoff's inequality

$$\begin{aligned}\mathbb{P}(\sup_{s\in[d]}|\overline{X}^{(s)}-\mu_s|\geq t) &\leq \sum_{s\in[d]}\mathbb{P}(|\overline{X}^{(s)}-\mu_s|\geq t) && \text{(union bound)}\\ &\leq d\cdot 2e^{-\frac{nt^2}{2}}. && \text{(Chernoff's inequality)}\end{aligned}$$

So

$$t^\star_{markov}=d\cdot\frac{1}{\delta\sqrt{n}},\qquad t^\star_{chernoff}=\frac{\sqrt{2\log(2d/\delta)}}{\sqrt{n}}.$$

In the second case the bound blows up slower as $d\to\infty$. The Chernoff's bound increases much more slowly (logarithmic in d) than the Markov bound (linear in d). ♣

1.2 Sub-Gaussian Random Variables

We now relax the assumption that $\mathbb{P}_Z\in\mathcal{P}([0,1])$.

Definition 1 (sub-Gaussian). *A random variable X with $\mu=\mathbb{E}[X]$ is sub-Gaussian with parameter σ, denoted with $\mathtt{sG}(\sigma)$, if*

$$\mathbb{E}[e^{\lambda(X-\mu)}]\leq e^{\lambda^2\sigma^2/2},\quad \forall\,\lambda\in\mathbb{R}\,.$$

Note (Gaussians and sub-Gaussians). *One may naturally ask why this condition is called "sub-Gaussian". To that end, we examine the following example regarding normal random variables. Informally, a sub-Gaussian random variable will have tails that decay at least as fast as the tails of a Gaussian random variable.*

(a) If $G\sim N(\mu,\sigma^2)$, then for $G-\mu=X\sim N(0,\sigma^2)$

$$\begin{aligned}\mathbb{E}[e^{\lambda(G-\mu)}]&=\mathbb{E}[e^{\lambda X}]\\ &=\int_{\mathbb{R}}e^{\lambda x}e^{-\frac{x^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{d}x\\ &=e^{\lambda^2\sigma^2/2}\int_{\mathbb{R}}e^{-\frac{(x-\lambda\sigma^2)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{d}x\\ &=e^{\lambda^2\sigma^2/2}.\end{aligned}$$

(b) If $G\sim N(0,1)$, then for ϕ the standard Gaussian density,

$$\lim_{t\to\infty}\frac{\mathbb{P}(G\geq t)}{\frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}}=1.$$

In addition,

$$\left(\frac{1}{t}-\frac{1}{t^3}\right)\phi(t) \leq \mathbb{P}(G \geq t) \leq \left(\frac{1}{t}-\frac{1}{t^3}+\frac{3}{t^5}\right)\phi(t),$$

which is known as the Mills' ratio.

♦

Proposition 6 (Hoeffding's Inequality I)**.** *If X is $\mathtt{sG}(\sigma)$, then for any $t \in \mathbb{R}$*

$$\mathbb{P}(|X-\mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Proof.

$$\begin{aligned}
\mathbb{P}(X-\mu \geq t) &\leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}} && \text{(Chernoff's Inequality)}\\
&\leq \inf_{\lambda \geq 0}\{e^{\lambda^2\sigma^2/2-\lambda t}\} && (X \sim \mathtt{sG}(\sigma))\\
&= e^{-\frac{t^2}{2\sigma^2}} && (\lambda^\star = \tfrac{t}{\sigma^2})
\end{aligned}$$

We found an upper bound for $X-\mu$, now we also want a lower bound. Note that, by definition, since $\lambda \in \mathbb{R}$, if X is sub-Gaussian so is $-X$. Hence by applying the same reasoning above to $-X$ (with mean is $-\mu$) we get

$$\mathbb{P}(X-\mu \leq -t) = \mathbb{P}(-X+\mu \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}.$$

Finally, by a union bound

$$\mathbb{P}(|X-\mu| \geq t) = \mathbb{P}(\{X-\mu \leq -t\} \cup \{X-\mu \geq t\}) \leq 2e^{-\frac{nt^2}{2\sigma^2}}.$$

■

Example 6 (Bound for the sample mean V)**.** *Going back to the previous example, we can now assume that $Z \sim \mathtt{sG}(\sigma)$ and work as before by first applying Chernoff's inequality*

$$\mathbb{P}(\overline{X}-\mu \geq t) \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda(\overline{X}-\mu)}]}{e^{\lambda t}}.$$

Then,

$$\begin{aligned}
\mathbb{E}[e^{\lambda(\overline{X}-\mu)}] &= \mathbb{E}[e^{\frac{\lambda}{n}\sum_{i=1}^n (Z_i-\mu)}]\\
&= \prod_{i=1}^n \mathbb{E}[e^{\frac{\lambda}{n}(Z_i-\mu)}] && \text{(independence)}\\
&\leq \prod_{i=1}^n e^{\frac{\sigma^2\lambda^2}{n^2 2}} = e^{\frac{\sigma^2\lambda^2}{2n}}. && \text{(sub-Gaussian)}
\end{aligned}$$

Then

$$\mathbb{P}(\overline{X}-\mu \geq t) \leq \inf_{\lambda \geq 0}\{e^{\frac{\sigma^2\lambda^2}{2n}-\lambda t}\}$$

$$= e^{\frac{n^2t^2}{2\sigma^2 n} - \frac{nt^2}{\sigma^2}} = e^{-\frac{nt^2}{2\sigma^2}}. \qquad (\lambda^\star = nt/\sigma^2)$$

Similarly, $-\overline{X} \sim \mathsf{sG}(\sigma)$*, thus*

$$\mathbb{P}(\overline{X} - \mu \leq -t) \leq e^{-\frac{nt^2}{2\sigma^2}}.$$

Finally, by a union bound we have

$$\mathbb{P}(|\overline{X} - \mu| \geq t) = \mathbb{P}(\{\overline{X} - \mu \leq -t\} \cup \{\overline{X} - \mu \geq t\}) \leq 2e^{-\frac{nt^2}{2\sigma^2}}.$$

Then we get $t^\star = \sqrt{\frac{2\log(2/\delta)}{n}}$ *and*

$$\mathbb{P}(|\overline{X} - \mu| \geq t) \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} \quad \textit{with probability at least } 1 - \delta.$$

♣

Sub-Gaussianity is **preserved by linear operations**.

Proposition 7 (Hoeffding's Inequality II)**.** *Suppose* $X_i \overset{ind}{\sim} \mathsf{sG}(\sigma_i), i = 1, \ldots, n$ *with mean* μ_i*. Then*

1. $\overline{X} \sim \mathsf{sG}(n^{-1}\sqrt{\sum_{i=1}^n \sigma_i^2})$,
2. *For any* $t \in \mathbb{R}$

$$\mathbb{P}(\overline{X} - \mu \geq t) \leq e^{-\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}}.$$

Proof. We just show that $\overline{X}$ is sub-Gaussian. The bound in (ii) follows from the traditional Hoeffding's inequality for sub-Gaussian random variables.

$$\mathbb{E}\left[e^{\lambda n^{-1}\sum_{i=1}^n (X_i - \mu_i)}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda n^{-1}(X_i - \mu)}\right] \qquad \text{(independence)}$$

$$\leq \prod_{i=1}^n e^{\lambda^2 \sigma_i^2/(2n^2)} = e^{\frac{\lambda^2}{2}\left(\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2\right)} \qquad (X_i \sim \mathsf{sG}(\sigma_i))$$

■

Proposition 8. *If* $X_i \sim \mathsf{sG}(\sigma)$ *and* $\kappa \in \mathbb{R}$*, then* $\kappa X_i \sim \mathsf{sG}(|\kappa|\sigma)$.

Example 7 (Bound for the sample mean VI)**.** *Let* $X_i \sim \mathsf{sG}(\sigma)$*, then, from the previous proposition we know that* $\overline{X} \sim \mathsf{sG}(\sqrt{\sigma^2/n})$*. Thus, by Hoeffding's inequality*

$$\mathbb{P}(\overline{X} - \mu \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}.$$

We can now answer to two questions:

1. *How to extract the order of* $\overline{X} - \mu$*?*

$$\delta = e^{-\frac{nt^2}{2\sigma^2}} \quad \Longrightarrow \quad t^\star = \sqrt{\frac{2\sigma^2}{n}\log(1/\delta)},$$

which implies that with probability at least $1-\delta$

$$\overline{X} - \mu \leq \sigma\sqrt{\frac{2\log(1/\delta)}{n}} = O(\sqrt{\log(1/\delta)}).$$

2. *How many samples such that* $\overline{X} - \mu \leq \varepsilon$ *with probability at least* $1-\delta$*?*

$$\delta = e^{-\frac{n\varepsilon^2}{2\sigma^2}} \quad \Longrightarrow \quad n^\star = \sigma^2\frac{\log(1/\delta)}{\varepsilon^2}.$$

♣

Example 8 (Some sub-Gaussian Random Variables)**.** *A list of popular sub-Gaussian random variables:*

1. *Gaussian random variables are* $\mathsf{sG}(\sigma)$ *where* $\sigma = \sqrt{\mathbb{V}(X)}$*.*

2. *Rademacher random variable is* $\mathsf{sG}(1)$*.*

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

where the last equality follows because the odd terms of the polynomial disappear. Our goal is to show that the MGF is upper bounded by $e^{\lambda^2/2}$*. This can be shown by Taylor's expansion*

$$e^{\lambda^2/2} = \sum_{k=0}^{\infty}\frac{\left(\lambda^2/2\right)^k}{k!} = \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{k!2^k}.$$

By using the fact that $(2k)! \geq k!2^k \equiv (2k)!!$ *we conclude that*

$$\mathbb{E}\left[e^{\lambda X}\right] = \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{k!2^k} = e^{\lambda^2/2}.$$

3. *Bounded RV with support on* $[a,b], a<b$ *are* $\mathsf{sG}(b-a)$*. Instead of direct calculation, we will use the* ***symmetrization trick****. This trick can be summarized in three steps:*

   (1) Let $Y' \overset{d}{=} X$ *be an independent copy of* X*.*

   (2) Apply Jenssen's inquality.

   (3) $W \sim$ *Rademacher,* $W\,(X-Y) \overset{d}{=} (X-Y)$

   To see this, let $\lambda \in \mathbb{R}$ *and* Y *be a copy of* X*. We are going to use two facts Then*

$$\begin{aligned}
\mathbb{E}[e^{\lambda(X-\mu)}] &\overset{(1)}{=} \mathbb{E}_X[e^{\lambda(X-\mathbb{E}_Y[Y])}] \\
&\leq \mathbb{E}_{X,Y}[e^{\lambda(X-Y)}] && \text{(Jensen's inequality)} \\
&= \mathbb{E}_{X,Y,W}[e^{\lambda W(X-Y)}] && \text{(symmetrization)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{X,Y}[\mathbb{E}_W[e^{\lambda(X-Y)W} \mid X, Y]] && \text{(LIE)}\\
&\leq \mathbb{E}_{X,Y}[e^{\frac{1}{2}(X-Y)^2\lambda^2}] && (W \sim \text{SG}(1))\\
&\leq e^{\frac{1}{2}(b-a)^2\lambda^2}, && \text{(bounded)}
\end{aligned}
$$

where the last line implies that $X \sim \mathsf{sG}(b-a)$

♣

Note. *Point (3) above relies on the following fact. Let* $X, Y \overset{iid}{\sim} F$. *Then* $Z = Y - X$ *is symmetric around 0. To show this we need to prove that* $f_z(z) = f_z(-z), \forall z \in \mathcal{Z}$. *So*

$$
\begin{aligned}
f_Z(z) &= \int_{-\infty}^{+\infty} f_X(x) f_Y(x+z)\,\mathrm{d}x && \text{(convolution)}\\
&= \int_{-\infty}^{+\infty} f_X(y-z) f_Y(y)\,\mathrm{d}y && (y = x+z)\\
&= \int_{-\infty}^{\infty} f_X(y) f_Y(y-z)\,\mathrm{d}y = f_Z(-z) && \text{(iid, thus exchangeable)}
\end{aligned}
$$

♦

1.3 Sub-Exponential Random Variables

Sometimes the requirement of sub-Gaussianity is very stringent. We are requiring that the probability in the tails decays as fast as e^{-t^2}. For example if $G \sim N(0,1)$, then G^2 is not sub-Gaussian anymore! To see this

$$
\mathbb{E}[e^{\lambda(G^2-1)}] = \int_{-\infty}^{\infty} e^{\lambda(z^2-1)} e^{-z^2/2}\,\mathrm{d}z = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}},
$$

for $\lambda < 1/2$. Whenever, $\lambda > 1/2$, the MGF of G^2 is infinite! This is what inspired the definition of sub-Exponential random variables. They are just sub-Gaussian but for a restricted range of λ.

Definition 2 (sub-Exponential). *A random variable* X *is sub-Exponential* $\mathsf{sE}(\nu, \alpha)$ *if*

$$
\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq e^{\frac{\lambda^2\nu^2}{2}}, \quad |\lambda| \leq \frac{1}{\alpha}.
$$

Note. *If* $X \sim \mathsf{sE}(\nu, 0)$ *then* $X \sim \mathsf{sG}(\nu)$. ♦

Example 9. *In the previous example* $G^2 \sim \mathsf{sE}(2, 4)$ *because*

$$
\mathbb{E}[e^{\lambda(G^2-1)}] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2}, \quad |\lambda| \leq \frac{1}{4}.
$$

♣

Proposition 9 (sub-Exponential Tail Bound). *If $X \sim \mathsf{sE}(\nu, \alpha)$, then*

$$\mathbb{P}[X - \mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{2v^2}} & \text{if } 0 \leq t \leq \frac{v^2}{\alpha} \\ e^{-\frac{t}{2\alpha}} & \text{for } t > \frac{v^2}{\alpha} \end{cases}$$

There is a condition that allows us to immediately verify that a variable is sub-Exponential.

Proposition 10 (Bernstein's Condition). *Given a random variable X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}[X^2] - \mu^2$, we say that Bernstein's condition with parameter b holds if*

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

Moreover, if X satisfies the Bernstein's condition, then $X \sim \mathsf{sE}(\sqrt{2}\sigma, 2b)$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(X-\mu)}\right] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}\left[(X-\mu)^k\right]}{k!} && \text{(Taylor's expansion)} \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{1}{2} \frac{k!\sigma^2 b^{k-2}}{k!} && \text{(Bernstein's condition)} \\
&= 1 + \frac{\sigma^2}{2b^2} \Big(\sum_{k=0}^{\infty} (|\lambda| b)^k - 1 - |\lambda| b\Big) \\
&= 1 + \frac{\sigma^2}{2} \frac{\lambda^2}{1 - |\lambda| b} && \text{(geometric series)} \\
&\leq e^{\frac{\sigma^2\lambda^2}{2(1-|\lambda|b)}} && (1 + x \leq e^x) \\
&\leq e^{\sigma^2\lambda^2} && (|\lambda| \leq 1/2b)
\end{aligned}$$

■

Proposition 11. *Let X be a random variable such that $0 \leq X \leq b$ a.s. and variance σ^2. Then $X \sim \mathsf{sE}(\sqrt{2}\sigma, 2b)$.*

Proof.

$$|\mathbb{E}[(X-\mu)^k]| = \mathbb{E}[(X-\mu)^{k-2}(X-\mu)^2] \leq b^{k-2}\sigma^2 \leq \frac{1}{2} k! b^{k-2} \sigma^2,$$

where the last inequality holds for $k \geq 2$. Thus, X satisfies the Bernstein's condition, hence it is sub-Exponential. ■

Similarly to sub-Gaussian variables, the sub-Exponential property is maintained with linear transformations.

Proposition 12 (sum of sub-Exponential). *Let $X_i \overset{ind}{\sim} \mathsf{sE}(\nu_i, \alpha_i)$ with $\mu_i = \mathbb{E}[X_i]$, then we have that $\sum_{i=1}^n (X_i - \mu_i) \sim \mathsf{sE}(\sqrt{\sum_{i=1}^n \nu_i^2}, \max_{i \in [n]} \alpha_i)$.*

Proof.

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda\sum_{i=1}^n(X_i-\mu_i)}\right] &= \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i-\mu_i)}\right] && \text{(independence)}\\
&\leq \prod_{i=1}^n e^{-\frac{\nu_i^2\lambda^2}{2}}, \quad \forall\,|\lambda| \leq \min_{i\in[n]}\frac{1}{\alpha_i} && (X_i \sim \mathsf{sE}(\nu_i,\alpha_i))\\
&= e^{\lambda^2(\sum_{i=1}^n \nu_i^2)/2}, \quad \forall\,|\lambda| \leq \frac{1}{\max_{i\in[n]}\alpha_i}
\end{aligned}$$

■

Proposition 13 (Bernstein's Inequality)**.** *Let* $X_i \overset{iid}{\sim} \mathsf{sE}(\nu, b)$*, then*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\mu) \geq t\right) \leq \exp\left\{-\min\left(\frac{nt^2}{2\nu^2}, \frac{nt}{2b}\right)\right\}$$

Example 10 (Bound for the sample mean VII)**.** *Let* $X_i \sim \mathsf{sE}(\nu, b)$*, then, from the previous proposition we know that* $\overline{X} \sim \mathsf{sE}(\sqrt{\nu^2/n}, b)$*. We can now answer to two questions:*

1. *How to extract the order of* $\overline{X} - \mu$*?*

$$\delta = \exp\left\{-n\cdot\min\left\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\right\}\right\} \quad \Longrightarrow \quad t^\star = \max\left\{\nu\sqrt{\frac{2\log(1/\delta)}{n}}; \frac{2\log(1/\delta)}{n}b\right\},$$

which implies that with probability at least $1-\delta$

$$\overline{X} - \mu \leq \nu\sqrt{\frac{2\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{n}b = O(\sqrt{\log(1/\delta)}).$$

This means that the Bernstein's bound has ***the same order in*** n ***and*** δ *as the Hoeffding's bound. The latter is derived under the stronger assumption that the random variables are sub-Gaussian though!*

2. *How many samples such that* $\overline{X} - \mu \leq \varepsilon$ *with probability at least* $1-\delta$*?*

$$\delta = \exp\left\{-n\cdot\min\left\{\frac{\varepsilon^2}{2\nu^2}, \frac{\varepsilon}{2b}\right\}\right\} \quad \Longrightarrow \quad n^\star = \max\left\{\frac{2\nu^2}{\varepsilon^2}, \frac{2b}{\varepsilon}\right\}\log(1/\delta),$$

which is again of the same order as the sample size given by the Hoeffding's bound.

♣

Note (Hoeffding vs Bernstein)**.** *Suppose that* $X_i \overset{iid}{\sim} \mathbb{P}_X \in \mathcal{P}([0,b])$ *with* $\mathbb{V}(X_i) \leq \nu^2$*. Then* $X_i \sim \mathsf{sG}(b)$ *and* $X_i \sim \mathsf{sE}(\nu, b)$*.*

We can use either Hoeffding's or Bernstein's inequality and get

$$\mathbb{P}(\overline{X} - \mu \geq t) \leq e^{-\frac{nt^2}{2b^2}}, \qquad \mathbb{P}\left(\overline{X} - \mu \geq t\right) \leq \exp\left\{-\min\left(\frac{nt^2}{2\nu^2}, \frac{nt}{2b}\right)\right\}.$$

Note that $t \leq b$ (otherwise the statement is trivial!) and $\nu^2 \leq b^2$. This implies that $t/b \geq (t/b)^2$ and $(t/\nu)^2 \geq t/b)^2$ hence

$$\frac{t^2}{b^2} \leq \min\left(\frac{t^2}{\nu^2}, \frac{t}{b}\right) \implies \texttt{UB}_{\textit{Hoeff}} \geq \texttt{UB}_{\textit{Bern}}.$$

The Bernstein's bound is never worse; moreover, it is substantially better whenever $\sigma^2 << b^2$, as would be the case for a random variable that occasionally takes on large values, but has relatively small variance. Intuitively, it captures more of the Chebyshev's effect, i.e. that random variables with small variance should be tightly concentrated around their mean. ♦

The following is the tighter bound for bounded RVs

Proposition 14 (Bennett's Inequality)**.** *Let $X_i, i = 1, 2, \ldots, n$ be independent such that $X_i - \mathbb{E}[X_i] \leq b$, a.s. and with $\nu_i^2 = \mathbb{V}(X_i)$. Then*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left\{-\frac{\sum_{i=1}^{n}\nu_i^2}{b^2}h\left(\frac{bt}{\sum_{i=1}^{n}\nu_i^2}\right)\right\}$$

where $h(u) = (1+u)\log(1+u) - u$.

Proposition 15 (Alternative Characterizations)**.** *The following are alternative characterizations of a* $\mathsf{sG}(\sigma)$ *(left) and* $\mathsf{sE}(\nu,\alpha)$ *(right) random variable*

(i) $\mathbb{P}(|X|\geq t)\leq 2e^{-t^2/K_1^2}, \forall\, t\geq 0$

(ii) $\|X\|_{L^p}=(\mathbb{E}|X|^p)^{1/p}\leq K_2\cdot\sqrt{p}, p\in\mathbb{N}_+$

(iii) $\mathbb{E}[e^{\lambda^2X^2}]\leq e^{\lambda^2K_3^2}, \quad \forall|\lambda|\leq 1/K_3$

(iv) $\mathbb{E}[e^{X^2/K_4^2}]\leq 2$

Moreover, if $\mathbb{E}[X]=0$,

(v) $\mathbb{E}[e^{\lambda X}]\leq e^{\lambda^2K_5^2}$

(i) $\mathbb{P}(|X|\geq t)\leq 2e^{-t^2/K_1^2}, \forall\, t\geq 0$

(ii) $\|X\|_{L^p}=(\mathbb{E}|X|^p)^{1/p}\leq K_2\cdot p, p\in\mathbb{N}_+$

(iii) $\mathbb{E}[e^{\lambda|X|}]\leq e^{|\lambda|K_3}, \quad \forall\lambda\leq 1/K_3$

(iv) $\mathbb{E}[e^{|X|/K_4}]\leq 2$

Moreover, if $\mathbb{E}[X]=0$,

(v) $\mathbb{E}[e^{\lambda X}]\leq e^{\lambda^2K_5^2}, \quad |\lambda|\leq 1/K_5$

1.4 Maximal Inequality

The maximal inequality is the Jensen's inequality for maxima. It is useful to figure out the magnitude of the maximum of -not necessarily independent- random variables.

Proposition 16. *Let* $(X_i)_{i\in[n]}$ *be a sequence of random variables. For all convex and strictly increasing functions* $\psi:\mathbb{R}\to\mathbb{R}_+\cup\{0\}$*, we have*

$$\mathbb{E}\left[\max_{i\in[n]}X_i\right]\leq\psi^{-1}\left(\sum_{i=1}^n\mathbb{E}\left[\psi\left(X_i\right)\right]\right)\quad\text{and}\quad\mathbb{P}\left(\max_{i\in[n]}X_i\geq t\right)\leq\sum_{i=1}^n\mathbb{E}\left[\psi\left(X_i\right)\right]/\psi(t)$$

Proof.

$$\begin{aligned}
\mathbb{E}[\max_{i\in[n]}X_i]&=\mathbb{E}\left[\psi^{-1}\left(\max_{i\in[n]}\psi\left(X_i\right)\right)\right] && (\psi(\cdot)\text{ is strictly increasing})\\
&\leq\psi^{-1}\mathbb{E}\left[\left(\max_{i\in[n]}\psi\left(X_i\right)\right)\right] && (\text{Jensen})\\
&\leq\psi^{-1}\left(\mathbb{E}\left[\sum_{i=1}^n\psi\left(X_i\right)\right]\right) && (\psi(x)\geq 0,\forall\, x\in\mathbb{R})\\
&\leq\psi^{-1}\left(\sum_{i=1}^n\mathbb{E}\left[\psi\left(X_i\right)\right]\right) && (\text{independence})
\end{aligned}$$

The second result follows by Markov's inequality

$$\begin{aligned}
\mathbb{P}\left(\max_{i\in[n])}X_i\geq t\right)&=\mathbb{P}\left(\psi\left(\max_{i\in[n]}X_i\right)\geq\psi(t)\right) && (\psi(\cdot)\text{ is strictly increasing})\\
&\leq\frac{\mathbb{E}\left[\psi\left(\max_{i\in[n]}X_i\right)\right]}{\psi(t)} && (\text{Markov's inequality})\\
&\leq\frac{\sum_{i=1}^n\mathbb{E}\left[\psi\left(X_i\right)\right]}{\psi(\tau)} && (\psi(x)\geq 0,\forall\, x\in\mathbb{R})
\end{aligned}$$

■

Example 11 (Maximal inequality for sub-Gaussian)**.** *Let* $X_i \overset{iid}{\sim} \mathsf{sG}(\sigma)$ *and choose* $\psi(x) = e^{\lambda x}, \psi^{-1}(x) = \frac{1}{\lambda}\log(x)$*. We have*

$$\mathbb{E}[\max_{i\in[n]} X_i] \leq \frac{1}{\lambda}\log(\sum_{i=1}^{n}\mathbb{E}[e^{\lambda X_i}]) \overset{\mathsf{sG}}{\leq} \frac{1}{\lambda}(\log n + \frac{\lambda^2\sigma^2}{2}).$$

Since this bounds holds for all $\lambda > 0$ *(to guarantee that* $\psi(\cdot)$ *is strictly increasing), we can optimize over* λ *to get* $\lambda^\star = \sqrt{2\log n/\sigma^2}$ *implying that*

$$\mathbb{E}[\max_{i\in[n]} X_i] = O(\sqrt{\log n}).$$

If $X \sim N$ *this is a sharp bound. To derive a lower bound we need to work with the density functions.* ♣

1.5 Truncation Argument

Let $G_i \overset{\text{iid}}{\sim} N(0,1)$ and consider $X_i = G_i^4$. We know that $\mathbb{E}[X_i] = 3$ and we want to upper bound, say $|\overline{X} - 3|$. Each X_i are neither sub-Gaussian nor sub-Exponential. Technically, they are sub-Gamma, but we can do something different to bound them.

The key idea is the following

$$X_i \quad \longrightarrow \quad X_i\mathbb{1}(|X_i| \leq c) \quad \longrightarrow \quad \Big|\sum_{i=1}^{n} X_i\mathbb{1}(|X_i| \leq c) - \mathbb{E}[X_i\mathbb{1}(|X_i| \leq c)]\Big|$$

work with the truncated version of X_i and then go back. The advantage of the truncated version is that it is a bounded random variable.

Step 1 Find b_n such that

$$\mathbb{P}\left(\max_{i\in[n]} X_i \geq b_n\right) \leq \frac{\delta}{2}$$

In this case $b_n = O((\log n)^2)$.

Step 2 Find ε_n such that

$$\mathbb{E}[X_i\mathbb{1}(X_i \geq b_n)] \leq \varepsilon_n.$$

In this case we can choose $\varepsilon_n = O(n^{-c})$.[1]

[1] One needs to derive ε_n by direct calculation, e.g.

$$\mathbb{E}[X_i\mathbb{1}(X_i \geq b_n)] = \int_{b_n}^{\infty}\mathbb{P}(X_i \geq t)\,\mathrm{d} = t\int_{b_n}^{\infty}\mathbb{P}(G_i \geq t^{1/4})\,\mathrm{d}t \leq \int_{b_n}^{\infty} e^{-\sqrt{t}/2}\,\mathrm{d}t \approx e^{-c\sqrt{b_n}} \approx n^{-c},$$

where integration is by substitution and parts and last bit uses $b_n = O((\log n)^2)$.

Step 3 Apply Hoeffding/Bernstein and get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i\mathbb{1}\left(X_i \leq b_n\right) - \mathbb{E}\left[X_i\mathbb{1}\left(X_i \leq b_n\right)\right]\right) \leq t_n\right) \geq 1-\delta/2$$

Since $X_i\mathbb{1}(X_i \leq b_n)$ are bounded random variables, we know that they are going to be $\mathsf{sG}((\log n)^2)$. When $Z_i \overset{\text{iid}}{\sim} \mathsf{sG}(\sigma)$ we know that Hoeffding's inequality yields a bound of order $\sqrt{\sigma^2\log(1/\delta)/n}$, therefore $t_n = O((\log n)^2\sqrt{\log(1/\delta)/n})$.

Step 4 Combining the previous steps

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right]\right) \leq t_n + \varepsilon_n\right) \geq 1-\delta$$

The first thing to notice is that

$$\mathbb{E}[X_i] = \mathbb{E}[X_i\mathbb{1}(X_i \leq b_n)] + \mathbb{E}[X_i\mathbb{1}(X_i \geq b_n)] \implies \mathbb{E}[X_i] - \mathbb{E}[X_i\mathbb{1}(X_i \leq b_n)] \leq \varepsilon_n.$$

The second thing to notice is that we are working under the event $\{\max_{i\in[n]} X_i \geq b_n\}^c$, thus $X_i = X_i\mathbb{1}(X_i \leq b_n), \forall\, i \in [n]$.

1.6 Martingale Concentration

So far we considered only independent random variables, but what about more complicated structures like martingales $S_n = \sum_i X_i$ or $S_n = f(X_1, \ldots, X_n)$?.

Definition 3. $\{(Y_k, \mathcal{F}_k)\}_{k\geq 1}$ *is a martingale sequence if:*

1. $|Y_k|_{k\geq 1}$ *is* $\{\mathcal{F}_k\}_{k>1}$*-adapted which means that* $Y_k \in m(\mathcal{F}_k)$.
2. $\mathbb{E}\left[|Y_k|\right] < \infty \quad \forall\, k$
3. $\mathbb{E}\left[Y_{k+1} \mid \mathcal{F}_k\right] = Y_k \quad \forall\, k$

$\{(D_k, \mathcal{F}_k)\}_{k\geq 1}$ *is a martingale difference sequence (MDS) if and only if* $\left\{\left(\sum_{s=1}^{k} D_s, \mathcal{F}_k\right)\right\}$ *is a martingale sequence.*

Example 12. *We now give two examples of martingales:*

1. *Consider* $\{X_i\}_{i\geq 1}$ *such that* $\mathbb{E}\left[|X_i|\right] < \infty$ *and define* $Y_k = \sum_{s=1}^{k}\left(X_s - \mathbb{E}\left[X_s\right]\right)$, $\mathcal{F}_k = \sigma\left(X_{1:k}\right)$. *Adaptation to the filtration follows by construction of generated sigma-algebra and integrability follows by the assumption that* $\mathbb{E}[|X_i|] < \infty$. *As per the last requirement*

$$\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = \mathbb{E}[X_{k+1} - \mathbb{E}[X_{k+1}] \mid \mathcal{F}_k] + Y_k = Y_k.$$

2. *This example is about the* ***Doob's Martingale****. Let X_i be independent random variables, $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$, and $f : \mathcal{X}^n \to \mathbb{R}$ such that $\mathbb{E}[|f(X_{1:n})|] < \infty$. Our goal is to study $|f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]|$. Define $Y_k = \mathbb{E}[f(X_{1:n}) \mid \mathcal{F}_k]$ and $D_k = Y_{k+1} - Y_k$. We can see that $\{(Y_k, \mathcal{F}_k)\}_{k\geq 1}$ is a martingale. Adaptation and integrability are immediate. Regarding the third assumption*

$$\mathbb{E}\left[Y_{k+1} \mid \mathcal{F}_k\right] = \mathbb{E}\left[\mathbb{E}\left[f\left(X_{1:n}\right) \mid \mathcal{F}_{k+1}\right] \mid \mathcal{F}_k\right] = \mathbb{E}\left[f\left(X_{1:n}\right) \mid \mathcal{F}_k\right] = Y_k.$$

It is also immediate to show that $\{(D_k, \mathcal{F}_k)\}_{k\geq 1}$ is an MDS.

♣

The next concentration inequalities are very similar to the ones we have already seen. They have to be understood as bounds on either $Y_n - Y_0$ or on the telescoping decomposition $\sum_{k=1}^n D_k, D_k := Y_k - Y_{k-1}$. Note indeed that if $X_i, i \in [n]$ are independent random variables and we are interested in studying $f(X_1, X_2, \dots, X_n) \equiv f(X_{1:n})$, then martingales are useful to model the concentration of the new random variable $f(X_{1:n})$. Indeed, let $Y_k = \mathbb{E}[f(X_{1:n}) \mid \mathcal{F}_k], Y_0 = \mathbb{E}[f(X_{1:n})), Y_n = f(X_{1:n})$. Then

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^n \underbrace{(Y_k - Y_{k-1})}_{D_k},$$

where $\{(D_k, \mathcal{F}_k)\}_{k\geq 1}$ is the Doob's MDS.

Proposition 17 (Azuma-Bernstein)**.** *Let $\{(D_k, \mathcal{F}_k)\}_{k\geq 1}$ be a MDS such that*

$$\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 \nu_k^2/2} \quad a.s., \quad \forall\, |\lambda| \leq \frac{1}{\alpha_k}.$$

Then:

1. $\sum_{k=1}^n D_k \sim \mathsf{sE}(\sqrt{\sum_{k=1}^n \nu_k^2}, \max_{k\in[n]} \alpha_k)$
2. $$\mathbb{P}\left(\left|\sum_{k=1}^n D_k\right| \geq t\right) = 2\exp\left\{-\min\left\{\frac{t^2}{2\sum_k \nu_k^2}, \frac{t}{2\max_k \alpha_k}\right\}\right\}$$

Note. *Differently form before, now the statement on the MGF is not deterministic. Indeed, $\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right]$ is a random variable, thus the statement holds almost surely. We will make ν_k and α_k random variables too in the future. The only requirement is that they must be $m(\mathcal{F}_{k-1})$.* ♦

Proof.

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda \sum_{i=1}^n D_k}\right] &= \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n-1} D_k}\, \mathbb{E}[e^{\lambda D_n} \mid \mathcal{F}_{n-1}]\right] && \text{(tower property)} \\
&= \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n-1} D_k}\right] e^{\lambda^2 \nu_k^2/2}, \quad \forall\, |\lambda| \leq \frac{1}{\alpha_n} && \text{(assumption)}
\end{aligned}$$

$$\leq \prod_{k=1}^{n} e^{-\frac{\nu_k^2 \lambda^2}{2}}, \quad \forall\, |\lambda| \leq \min_{k\in[n]} \frac{1}{\alpha_k} \qquad \text{(repeat)}$$
$$= e^{\lambda^2 (\sum_{k=1}^{n} \nu_k^2)/2}, \quad \forall\, |\lambda| \leq \frac{1}{\max_{k\in[n]} \alpha_k}$$

■

Note. *In the proof we replaced the part in which we were leveraging independence with the part in which we leverage the tower property. Before $\sum_i D_i$ had independent summands, now they need not to be independent.* ♦

Proposition 18 (Azuma-Hoeffding). *Let $\{(D_k, \mathcal{F}_k)\}_{k\geq 1}$ be a MDS such that $\exists \{(a_k, b_k)\}_{k=1}^n$, where $a_k, b_k \in m(\mathcal{F}_{k-1})$ such that*

$$\forall\, k \in [n] \qquad a_k \leq D_k \leq b_k, \quad a.s. \qquad \text{and} \qquad |b_k - a_k| \leq L_k \in \mathbb{R}_+, \quad a.s.,$$

then

$$\sum D_k \sim \mathsf{sG}\left(\frac{1}{2}\sqrt{\sum_{k=1}^{n} L_k^2}\right) \qquad \text{and} \qquad \mathbb{P}\left(\left|\sum_{k=1}^{n} D_k\right| \geq t\right) \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}}.$$

Proof. Follows from Azuma-Bernstein with $\alpha_k \to \infty$. ■

Definition 4 (Bounded difference function). *A function f is $(L_1, \ldots, L_n) - BD$ if*

$$|f(X_{1:k-1}, X_k, X_{k+1:n}) - f(X_{1:k-1}, X_k', X_{k+1:n})| \leq L_k, \quad \forall k, X_{1:n}, X_{1:n}'.$$

If we change one coordinate at the time the function does not vary much.

Proposition 19 (Bounded difference inequality). *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is $L_{1:n} - BD$ and X_i are independent, then*

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

Proof. Just Azuma-Hoeffing with $\sum_{k=1}^{n} D_k = f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$, which is a MDS. We also need to find $\{(A_k, B_k)\}_{k=1}^n$ such that $|B_k - A_k| \leq L_k$ to apply Azuma-Hoeffding. Recall that $D_k = \mathbb{E}[f(X_{1:n}) \mid X_{1:k}] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}]$. Therefore we can bound it by simply setting

$$B_k = \sup_x \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}, X_k = x] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}]$$
$$A_k = \inf_x \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}, X_k = x] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}].$$

which gives us $A_k \leq D_k \leq B_k$ a.s. by construction and $A_k, B_k \in m(\mathcal{F}_{k-1})$. ■

Example 13 (U-statistics). *Let $X_i \overset{iid}{\sim} \mathbb{P}_X \in \mathcal{P}([a, b])$. We want to estimate $\theta := \mathbb{E}_{X,X'}[|X - X'|]$. An estimator is*

$$U(X_{1:n}) = \binom{n}{2}^{-1} \sum_{i<j} |X_i - X_j|.$$

It is natural to ask what is the estimation error $|U(X_{1:n}) - \theta|$*? To do so, we show that* $U(X_{1:n})$ *is a bounded-difference function.*

$$|U(X_{1:n}) - U(X_{1:k-1}, X_k, X_{k+1:n})| = \left| \binom{n}{2}^{-1} \sum_{i \neq j} \left(|X_i - X_j| - \left|X_i - X'_j\right| \right) \right| \leq \frac{4bn}{n^2 + o(n)} \approx \frac{4b}{n}.$$

Therefore U *is* $(\frac{4b}{n}, \ldots, \frac{4b}{n}) - BD$*, hence we can apply the BD inequality and get*

$$\mathbb{P}(|U(X_{1:n}) - \theta| \geq t) \leq 2 \exp\left\{ -\frac{t^2}{n\frac{8b^2}{n^2}} \right\} = 2 \exp\left\{ -\frac{nt^2}{8b^2} \right\}$$

So the estimation error is larger than $t = \sqrt{8 \log(2/\delta)/n}$ *with probability at least* $1 - \delta$*.* ♣

Example 14 (Supremum of empirical process)**.** *Consider a setting in which we observe samples from* $(Z_i)_{i \in [n]} \overset{iid}{\sim} \mathbb{P}_Z$ *and we are interested in estimating some parameter* $\theta = F(\mathbb{P}_Z) \in \Theta$*. To evaluate different estimators we use a loss function* $\ell : \mathcal{Z} \times \Theta \to [0, 1]$*. We consider the empirical risk* $\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(z_i; \theta)$ *and the population risk* $R(\theta) = \mathbb{E}[R_n(\theta)] = \mathbb{E}[\ell(Z; \theta)]$*. We also consider the excess risk, defined as*

$$\varepsilon(z_{1:n}) \equiv \sup_{\theta \in \Theta} \left\{ R(\theta) - \hat{R}_n(\theta) \right\}.$$

We claim that $\varepsilon(z_{1:n})$ *is* $\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right) - BD$*. To see this define* $\theta^\star := \arg\max_\theta \varepsilon(Z_{1:n})$*. Then*

$$\begin{aligned}
\varepsilon(Z_{1:n}) - \varepsilon(Z_{1:n-1}, Z'_n) &= R(\theta^\star) - \widehat{R}(\theta^\star; Z_{1:n}) - \sup_{\theta \in \Theta} \left\{ R(\theta) - \hat{R}_n(\theta; Z_{1:n-1}, Z'_n) \right\} \\
&\leq R(\theta^\star) - \widehat{R}(\theta^\star; Z_{1:n}) - R(\theta^\star) + \hat{R}_n(\theta^\star; Z_{1:n-1}, Z'_n) \\
&= \hat{R}_n(\theta^\star; Z_{1:n-1}, Z'_n) - \hat{R}_n(\theta^\star; Z_{1:n}) \\
&= \frac{1}{n}(\ell(\theta^\star; Z_n) - \ell(\theta^\star; Z'_n)) \leq \frac{1}{n},
\end{aligned}$$

where the last inequality follows from the bounded image of the loss function. Note that the same reasoning can be used to bound $\varepsilon(Z_{1:n-1}, Z'_n) - \varepsilon(Z_{1:n})$*. Finally,* Z_i *are iid, thus exchangeable hence the bound holds for all coordinates.*

Then $|\varepsilon(Z_{1:n}) - \mathbb{E}[\varepsilon(Z_{i:n})]| \leq \sqrt{\frac{2 \log(2/\delta)}{n}}$ *with probability at least* $1 - \delta$*.* ♣

1.6.1 Generalization of Martingale Concentration Inequalities

As we said above, Azuma-Bernstein inequality can be generalized to predictable scale parameters ν_k. By predictable we mean that $\nu_k \in m(\mathcal{F}_{k-1})$. The problem then is that statements like

$$\mathbb{P}\left(\left| \sum_{k=1}^n D_k \right| \geq t \right) = 2 \exp\left\{ -\min\left\{ \frac{t^2}{2\sum_k \nu_k^2}, \frac{t}{2 \max_k \alpha_k} \right\} \right\}$$

are not meaningful anymore because the LHS is deterministic and the RHS is a random variable.

Proposition 20 (Freedman's Inequality). *Let $\{(D_k, \mathcal{F}_k)\}_{k\geq 1}$ be a MDS and let $\{\nu_k\}_{k=1}^n$ be random variable such that $\nu_k \in m(\mathcal{F}_{k-1})$. If*

$$\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 \nu_k^2/2} \quad a.s., \quad \forall\, |\lambda| \leq \frac{1}{\alpha_k}.$$

Then $\forall\, |\lambda| \leq \frac{1}{\alpha_k}$

$$\left|\sum_{k=1}^n D_k\right| \leq \lambda \sum_{k=1}^n \nu_k^2 + \frac{1}{\lambda}\log(2/\delta)$$

with probability at least $1-\delta$. Further, if $\sum_{k=1}^n \nu_k^2 \leq V$ a.s., then

$$\left|\sum_{k=1}^n D_k\right| \leq \min\left\{\sqrt{2V \cdot \log(2/\delta)}, 2b\log(2/\delta)\right\}$$

with probability at least $1-\delta$.

Proof. The proof is very similar to the one we used for the Azuma-Bernstein inequality. The difference is that we don't look directly at the MGF of $\sum_k D_k$. In particular, in the other case we iteratively used the fact that

$$\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 \nu_k^2/2} \quad \text{a.s.}$$

Now we are going to use

$$\mathbb{E}\left[e^{\lambda D_k - \lambda^2 \nu_k^2/2} \mid \mathcal{F}_{k-1}\right] \leq 1.$$

Let $|\lambda| \in [0, 1/\alpha_k]$. Then

$$\mathbb{E}[e^{\sum_{k=1}^n (\lambda D_k - \lambda^2\nu_k^2/2)}] = \mathbb{E}[e^{\sum_{k=1}^{n-1}(\lambda D_k - \lambda^2\nu_k^2/2)}\, \mathbb{E}[e^{\lambda D_n - \lambda^2\nu_n^2/2} \mid \mathcal{F}_{n-1}]] \leq 1$$

Then, by Markov's inequality

$$\mathbb{P}\left(e^{\sum_{k=1}^n (\lambda D_k - \lambda^2\nu_k^2/2)} \geq 2\delta^{-1}\right) \leq \frac{\mathbb{E}\left[e^{\sum_{k=1}^n (\lambda D_k - \lambda^2\nu_k^2/2)}\right]}{2/\delta^{-1}} \leq \delta/2.$$

Finally

$$\sum_{k=1}^n (\lambda D_k - \lambda^2\nu_k^2/2) \geq \log(2/\delta) \quad \text{wp } \delta \quad \Longrightarrow \quad \sum_{k=1}^n D_k \leq \lambda \sum_{k=1}^n \nu_k^2/2 + \frac{1}{\lambda}\log(2/\delta) \quad \text{wp } 1-\delta/2.$$

The same logic can be applied to $-D_k$, noting that it is still a MDS and everything else follows (just think of absorbing the - into the λ) to give

$$\left|\sum_{k=1}^n D_k\right| \leq \lambda \sum_{k=1}^n \nu_k^2 + \frac{1}{\lambda}\log(2/\delta).$$

■

Proposition 21 (Doob's Maximal Inequality). *If $\{X_s\}_{s\geq 1}$ is a non-negative super-Martingale, i.e. if $\mathbb{E}[X_t \mid \mathcal{F}_s] \leq X_s, s < t$, then $\forall\, u > 0$*

$$\mathbb{P}\left(\max_{0\leq t\leq T} X_t \geq u\right) \leq \frac{\mathbb{E}[X_0]}{u}.$$

If $\{X_s\}_{s\geq 1}$ is a non-negative sub-Martingale, i.e. if $\mathbb{E}[X_t \mid \mathcal{F}_s] \geq X_s, s < t$, then $\forall\, u > 0$

$$\mathbb{P}\left(\max_{0\leq t\leq T} X_t \geq u\right) \leq \frac{\mathbb{E}[X_T]}{u}.$$

Note. *Intuition is that in a super-Martingale the expectation is roughly decreasing, whilst in a sub-Martingale is roughly increasing.* ♦

Proof. Define the stopping time $\tau := \inf\{t \geq 0 : X_t \geq u\}$ which describes the first time in which the martingale goes above u. Then

$$\begin{aligned}
\mathbb{P}\left(\max_{0\leq t\leq T} X_t \geq u\right) &= \mathbb{P}(X_\tau \geq u, \tau \leq T) \\
&\leq \frac{\mathbb{E}[X_\tau \mathbb{1}(\tau \leq T)]}{u} && \text{(Markov)} \\
&= \sum_{t=0}^{T} \frac{\mathbb{E}[X_\tau \mathbb{1}(\tau = t)]}{u} \\
&= \sum_{t=0}^{T} \frac{\mathbb{E}[\mathbb{E}[X_t \mid \mathcal{F}_0]\mathbb{1}(\tau = t)]}{u} && (\star) \\
&\leq \sum_{t=0}^{T} \frac{\mathbb{E}[X_0 \mathbb{1}(\tau = t)]}{u} && \text{(super-Martingale)} \\
&\leq \frac{\mathbb{E}[X_0]}{u} && (X \text{ is non-negative})
\end{aligned}$$

where we haven't checked $(\star)$. ■

Note. *Note that we've used a particular version of Markov's inequality. Let X be a non-negative random variable and let Y be another real-valued random variable. Then, let $B \in \mathcal{B}_{\mathbb{R}}$*

$$X\mathbb{1}(Y \in B) \geq t\mathbb{1}(X \geq t, Y \in B) \quad a.s.$$

and the same proof we used for Markov's inequality goes through. Thus

$$\mathbb{E}[X\mathbb{1}(Y \in B)] \geq \mathbb{E}[t\mathbb{1}(X \geq t, Y \in B)] \implies \mathbb{P}(\mathbb{1}(X \geq t, Y \in B)) \leq \frac{\mathbb{E}[X\mathbb{1}(Y \in B)]}{t}.$$

♦

We can use Doob's maximal inequality to strengthen Freedman's inequality.

Proposition 22 (Maximal Freedman's Inequality)**.** *Let $\{(D_k, \mathcal{F}_k)\}_{k\geq 1}$ be a MDS and let $\{\nu_k\}_{k=1}^n$ be random variable such that $\nu_k \in m(\mathcal{F}_{k-1})$. If*

$$\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 \nu_k^2/2} \quad a.s., \quad \forall\, |\lambda| \leq \frac{1}{\alpha_k}.$$

Then $\forall\, |\lambda| \leq \frac{1}{\alpha_k}$:

$$\forall\, K \in [n] \quad \left|\sum_{k=1}^{K} D_k\right| \leq \lambda \sum_{k=1}^{n} \nu_k^2 + \frac{1}{\lambda}\log(2/\delta)$$

with probability at least $1 - \delta$.

Proof. Consider the random variable $X_k = e^{\sum_{t=0}^{k}(\lambda D_t - \lambda^2 \nu_t^2/2)}$. We want to show that it is a super-martingale with respect to its natural filtration. Indeed $\forall\, |\lambda| \leq 1/\alpha_k$

$$\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = \mathbb{E}[e^{\sum_{t=0}^{k-1}(\lambda D_t - \lambda^2 \nu_t^2/2)} e^{\lambda D_k - \lambda^2 \nu_t^2} \mid \mathcal{F}_{k-1}] = X_{k-1}\, \mathbb{E}[e^{\lambda D_k - \lambda^2 \nu_t^2} \mid \mathcal{F}_{k-1}] \leq X_{k-1}.$$

Therefore, we can immediately apply Doob's maximal inequality and get

$$\mathbb{P}\left(\max_{0\le t\le T} X_t \ge \frac{2}{\delta}\right) \le \frac{\mathbb{E}[X_0]}{2/\delta} \le \delta/2.$$

We can rearrange this statement as follows

$$\mathbb{P}\left(\max_{0\le t\le T} X_t \ge \frac{2}{\delta}\right) = \mathbb{P}\left(\exists\, t \in [T] : X_t \ge \frac{2}{\delta}\right) = 1 - \mathbb{P}\left(\forall\, t \in [T] : X_t \le \frac{2}{\delta}\right),$$

which implies that

$$\forall\, k \in [T] \quad \sum_{t=0}^{k} (\lambda D_t - \lambda^2 \nu_t^2/2) \le \log(2/\delta) \quad \text{wp at least } 1-\delta.$$

We can do a similar reasoning with $-D_t$ and conclude the proof. ■

Note. *The previous version was for n fixed, whereas this one is for any n! More formally, let $\texttt{UB}(\delta, \lambda) := \lambda \sum_{k=1}^{n} \nu_k^2 + \frac{1}{\lambda}\log(2/\delta)$, then*

$$\underbrace{\mathbb{P}\left(\left|\sum_{k=1}^{K} D_k\right| \le \texttt{UB}(\delta,\lambda)\right) \ge 1-\delta}_{\textit{Freedman}} \quad \textit{vs} \quad \underbrace{\mathbb{P}\left(\forall\, K \in [n], \left|\sum_{k=1}^{K} D_k\right| \le \texttt{UB}(\delta,\lambda)\right) \ge 1-\delta}_{\textit{Maximal Freedman}}.$$

*Note that a statement on all $K \in [n]$ can still be obtained through Freedman's inequality and the union bound but such bound is going to be much looser. Indeed, by using the maximal Freedman's inequality we can refine the probabilistic statement on the **same** bound and get rid of the factor $1/K$. To see this, suppose we have applied Freedman's inequality. Then*

$$\begin{aligned}
\mathbb{P}\left(\forall\, K \in [n], \left|\sum_{k=1}^{K} D_k\right| \le \texttt{UB}(\delta,\lambda)\right) &= 1 - \mathbb{P}\left(\exists\, K \in [n], \left|\sum_{k=1}^{K} D_k\right| \ge \texttt{UB}(\delta,\lambda)\right) \\
&= 1 - \mathbb{P}\left(\bigcup_{K=0}^{n} \left|\sum_{k=1}^{K} D_k\right| \ge \texttt{UB}(\delta,\lambda)\right) \\
&\ge 1 - \sum_{K=0}^{n} \mathbb{P}\left(\left|\sum_{k=1}^{K} D_k\right| \ge \texttt{UB}(\delta,\lambda)\right) \ge 1 - \frac{\delta}{K}
\end{aligned}$$

♦

1.7 Gaussian Concentration

Definition 5 (Lipschitz functions). *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be $L-$Lipschitz with in the ℓ_p-norm if $\exists\, L > 0$ such that*

$$\forall\, x, y \in \mathbb{R}^n \qquad |f(x) - f(y)| \le L\|x - y\|_p.$$

Proposition 23 (Gaussian Concentration Inequality). *Let $X_i \overset{iid}{\sim} N(0,1)$ and $f : \mathbb{R}^n \to \mathbb{R}$ that is L-Lipschitz in ℓ_2-norm. Then*

1. $f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$ *is* $\mathsf{sG}(L)$
2. *By Hoeffding's inequality we have*

$$\mathbb{P}\left(|f\left(X_{1:n}\right) - \mathbb{E}\left[f\left(X_{1:n}\right)\right]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2L^2}\right)$$

To prove this statement we will also use the following lemma

Lemma 1. *For any convex function $\phi : \mathbb{R} \to \mathbb{R}$ and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:*

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(X)])] \leq \mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle \nabla f(X), Y\rangle\right)\right], \qquad X, Y \overset{iid}{\sim} N\left(0, I_n\right)$$

We first prove the lemma.

Proof.

$$\begin{aligned}
\mathbb{E}_X[\phi(f(X) - \mathbb{E}_X[f(X)])] &= \mathbb{E}_X[\phi(f(X) - \mathbb{E}_Y[f(Y)])] && (Y \text{ is a copy of } X)\\
&\leq \mathbb{E}_{X,Y}[\phi(f(X) - f(Y))] && (\text{Jensen})
\end{aligned}$$

Define the following random variable

$$Z(\theta) = X\cos\theta + Y\sin\theta.$$

The variable $Z(\theta)$ can be thought of as a **path** between X and Y. Indeed, when $\theta = 0$ we get $Z(\theta) = X$, whilst if $\theta = \pi/2$ we get $Z(\theta) = Y$. Therefore, as θ varies in the interval $[0, \pi, 2]$ we are moving from X to Y. The rv Z has some nice properties

$$\forall\, \theta \in [0, \pi, 2] \qquad Z(\theta) \overset{d}{=} X \overset{d}{=} Y, \quad Z'(\theta) \overset{d}{=} X \overset{d}{=} Y, \quad Z(\theta) \perp\!\!\!\perp Z'(\theta).$$

First, for a fixed θ, $Z(\theta)$ is a linear combination of standard normals, Then $\mathbb{E}[Z(\theta)] = 0$ and $\mathbb{V}(Z(\theta)) = \mathbb{V}(X)\cos^2\theta + \mathbb{V}(Y)\sin^2\theta = (\sin^2\theta + \cos^2\theta)I_n = I_n$ showing that $Z(\theta) \sim N(0, I_n)$. Consider now $Z'(\theta) = -X\sin\theta + Y\cos\theta$. Using a similar reasoning we can show that $Z'(\theta) \sim N(0, I_n)$. Independence comes from the fact that

$$\mathbb{E}[Z(\theta)Z'(\theta)] = -\mathbb{E}[X^2]\cos\theta\sin\theta + \mathbb{E}[XY](\cos^2\theta - \sin^2\theta) + \mathbb{E}[Y^2]\sin\theta\cos\theta = 0.$$

By the Fundamental Theorem of Calculus

$$g(x) = g(x_0) + \int_{x_0}^{x} g'(t)\,\mathrm{d}t.$$

Consider $g(\theta) = f(Z(\theta))$ and let $x = \pi/2$ and $x_0 = 0$, then

$$g(\pi/2) = g(0) + \int_0^{\pi/2} g'(t)\,\mathrm{d}t, \quad \text{where} \quad g'(\theta) = \langle \nabla(f(Z(\theta)), Z'(\theta)\rangle.$$

By definition of $g(\cdot)$ we also know that $g(\pi/2) = f(Y)$ and $g(0) = X$, thus

$$f(Y) - f(X) = \int_0^{\pi/2} \langle \nabla(f(Z(\theta)), Z'(\theta)\rangle\,\mathrm{d}\theta,$$

which implies

$$\mathbb{E}[\phi(f(Y) - f(X))] = \mathbb{E}\left[\phi\left(\int_0^{\pi/2} \langle \nabla(f(Z(\theta)), Z'(\theta)\rangle\,\mathrm{d}\theta\right)\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\phi\left(\frac{\pi}{2}\underbrace{\int_0^{\pi/2}\frac{2}{\pi}\langle\nabla(f(Z(\theta)),Z'(\theta)\rangle\,\mathrm{d}\theta}_{\int_0^{\pi/2} f(\theta)h(\theta)\,\mathrm{d}\theta=\mathbb{E}_\theta[h(\theta)]}\right)\right] && \text{(think of } \theta\sim U[0,\pi,2])\\
&\le \mathbb{E}\left[\int_0^{\pi/2}\frac{2}{\pi}\phi\left(\frac{\pi}{2}\langle\nabla(f(Z(\theta)),Z'(\theta)\rangle\right)\mathrm{d}\theta\right] && \text{(Jensen)}\\
&= \frac{2}{\pi}\int_0^{\pi/2}\mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle\nabla(f(Z(\theta)),Z'(\theta)\rangle\right)\right]\mathrm{d}\theta && \text{(Fubini)}\\
&= \mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle\nabla(f(X),Y\rangle\right)\right],
\end{aligned}
$$

where the last equality exploits the fact that the term $\mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle\nabla(f(Z(\theta)),Z'(\theta)\rangle\right)\right]$ is constant for all θ as $Z(\theta)$ and $Z'(\theta)$ are standard normals. ■

We now prove the Gaussian concentration inequality, based on the **interpolation method**. We will show a weaker version of the proposition, where we will get a factor larger than one in front of L.

Proof. Let's use the lemma with $\phi(\cdot)=\exp(\lambda\cdot)$.

$$
\begin{aligned}
\mathbb{E}_{X,Y}[\exp(\lambda(f(X)-\mathbb{E}[f(X)]))] &\le \mathbb{E}_{X,Y}\left[\exp\left(\lambda\frac{\pi}{2}\langle\nabla f(x),Y\rangle\right)\right] && \text{(lemma)}\\
&= \mathbb{E}_X\left[\mathbb{E}_Y\left[\exp\left(\lambda\frac{\pi}{2}\langle\nabla f(x),Y\rangle\mid X\right)\right]\right] && \text{(tower rule)}\\
&= \mathbb{E}_X\left[e^{\lambda^2\frac{\pi^2}{4}\|\nabla f(x)\|_2^2/2}\right] && (\mathbb{E}[e^{\langle\mu,Y\rangle}]=e^{\frac{1}{2}\|\mu\|_2^2})\\
&\le e^{\lambda^2\frac{\pi^2}{4}L^2/2}\sim \mathsf{sG}\left(\frac{\pi}{2}L\right),
\end{aligned}
$$

where the last step uses the IVT, the fact that f is Lipschitz and differentiable to show that $\nabla f(X)$ is bounded by L. ■

Let's see a bunch of applications.

Example 15 (Order Statistics)**.** *Let $X_i \overset{iid}{\sim} N(0,1)$ and define $f_k(X_{1:n})=X_{(k)}$. To study the concentration of this random variable we want to show that it is L-Lipschitz. Thus*

$$
|f_k(X_{1:n})-f_k(Y_{1:n})| = |X_{(k)}-Y_{(k)}| \le \sqrt{\sum_{k=1}^n (X_{(k)}-Y_{(k)})^2} \le \sqrt{\sum_{k=1}^n (X_k-Y_k)^2} = \|X-Y\|_2,
$$

where the last equality is called the sorting inequality. Therefore

$$
\left|X_{(k)}-\mathbb{E}\left[X_{(k)}\right]\right| \le O(\sqrt{\log(1/\delta)}).
$$

♣

Example 16 (Singular values of Gaussian random matrices)**.** *Let $X\in\mathbb{R}^{n\times d}$ where each element is $X_{ij}\sim N(0,1)$. Define $f_k(X)=\sigma_k(X)$ and $f_1(X)=\|X\|_{\mathsf{op}}$, where*

$$
\|X\|_{\mathsf{op}} = \inf_{v\in\mathbb{R}^d}\{\|Xv\| : \|v\|=1\}.
$$

♣

Example 17 (Gaussian Complexity)**.** *Given a set $\mathcal{A} \subseteq \mathbb{R}^n$, how can we measure its "size"? A reasonable size function S should satisfy at least $\mathcal{A} \subseteq \mathcal{B} \implies S(\mathcal{A}) \leq S(\mathcal{B})$. Two examples of S are the Euclidean width $D(\mathcal{A}) = \max_{x \in \mathcal{A}} \|x\|_2$ and the dimension.*

Let $\mathbf{w} := (w_1, \dots, w_n)' \in \mathbb{R}^n, w_i \overset{iid}{\sim} N(0,1)$. The Gaussian complexity (or "statistical dimension") of a set $\mathcal{A} \subseteq \mathbb{R}^n$ is defined as

$$\mathcal{G}(\mathcal{A}) := \mathbb{E}_{W \sim N(0, I_n)}[\sup_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \mathbf{w} \rangle].$$

We can think of $\mathbf{w}$ as being random noise and we want to understand how "aligned" $\mathcal{A}$ is with this noise. This follows from the geometric intuition that the inner product of two vector is 0 if they are orthogonal and it's maximized when they are parallel.

Define $f(\mathbf{w}) = \sup_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \mathbf{w} \rangle$. We want to show that it concentrates around $\mathcal{G}(\mathcal{A})$. To do so, we show that $f(\mathbf{w})$ is Lipschitz. Let $\mathcal{A}$ be compact for simplicity and define $\mathbf{a}^\star = \arg\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \mathbf{w} \rangle$. Then

$$\begin{aligned}
f(\mathbf{w}) - f(\mathbf{w}') &= \langle \mathbf{a}^\star, \mathbf{w} \rangle - \sup_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \mathbf{w} \rangle \\
&\leq \langle \mathbf{a}^\star, \mathbf{w} \rangle - \langle \mathbf{a}^\star, \mathbf{w} \rangle \\
&= \langle \mathbf{a}^\star, \mathbf{w} - \mathbf{w}' \rangle \\
&\leq \|\mathbf{a}^\star\|_2 \|\mathbf{w} - \mathbf{w}'\|_2 && \text{(Cauchy-Schwarz)} \\
&\leq \sup_{\mathbf{a} \in \mathcal{A}} \|\mathbf{a}\|_2 \|\mathbf{w} - \mathbf{w}'\|_2 && (\mathbf{a}^\star \text{ depended on } \mathbf{w}) \\
&= \mathcal{D}(\mathcal{A}) \|\mathbf{w} - \mathbf{w}'\|_2,
\end{aligned}$$

therefore $f(\mathbf{w}) \sim \mathsf{sG}(\mathcal{D}(\mathcal{A}))$. ♣

1.8 Extensions

Let $X_1, X_2, \dots, X_n$ be independent and suppose we are interested in a rv $Z = f(X_{1:n})$ where $f : \mathbb{R}^n \to \mathbb{R}$. How do study the concentration of Z?

1. **Martingale concentration** by using the *Bounded difference inequality* we can show that
$$f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \sim \mathsf{sG}\left(\sum_{k=1}^n L_k^2\right)$$
Requirements:
   i) f is $(L_1, \dots, L_n) - BD$, i.e.
   $$|f(X_{1:k-1}, X_k, X_{k+1:n}) - f(X_{1:k-1}, X_k', X_{k+1:n})| \leq L_k, \quad \forall k, X_{1:n}, X_{1:n}'.$$
   ii) $X_1, X_2, \dots, X_n$ are independent

   Examples: U-statistics, supremum of empirical process

2. **Gaussian concentration** by using the *Gaussian concentration inequality* we can show that
$$f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \sim \mathsf{sG}\left(L\right)$$
Requirements:
   i) f is L-Lipschitz, i.e. $\forall\, x, y \in \mathbb{R}^n, |f(x) - f(y)| \leq L\|x - y\|_p$
   ii) $X_i \overset{\text{iid}}{\sim} N(0,1)$

   Examples: Gaussian complexity, singular values of Gaussian random matrix

There are two general principles here:

1. We require the **function** $f(\cdot)$ to be **stable with respect to perturbations of** X
2. Require the **measure** of X to be **well-behaved**

1.8.1 Convexity

See Wainwright Ch. 3 for a more technical introduction.

Proposition 24 (Convex I)**.** *Let $X_1, X_2, \ldots, X_n$ be independent and $X_i \in [a,b]$ a.s. Suppose that f is L-Lipschitz and separately convex, i.e. $\frac{\partial^2}{\partial X_i^2} f(X_{1:n}) \geq 0$. Then*

$$\mathbb{P}(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \geq t) \leq e^{-\frac{t^2}{4L^2(b-a)^2}}.$$

This is just a bound for the upper tail.

Note that if you don't have convexity, you can still show that f is $L(b-a) - BD$ and apply the martingale concentration bound but then you'd get a factor of n in the denominator so the bound gets larger.

Proposition 25 (Convex II)**.** *Let $X_1, X_2, \ldots, X_n$ be independent and $X_i \in [a,b]$. Suppose that f is L-Lipschitz and convex, i.e. $\nabla^2 f(X_{1:n})$ exists and positive definite. Then*

$$f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \sim \mathsf{sG}\left(L(b-a)\right).$$

By assuming convexity we also get the lower bound.

Example 18 (Rademacher Complexity)**.** *Consider a set $\mathcal{A} \subseteq \mathbb{R}^n$. Gaussian Complexity is defined as*

$$\mathcal{G}(\mathcal{A}) := \mathbb{E}_{W \sim N(0, I_n)}[\sup_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \mathbf{w} \rangle],$$

whereas the Rademacher complexity is

$$\mathcal{R}(\mathcal{A}) := \mathbb{E}_{\varepsilon_i \overset{iid}{\sim} U(\{-1,1\})}[\sup_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \boldsymbol{\varepsilon} \rangle].$$

We've seen that $\sup_{\mathbf{a}\in\mathcal{A}}\langle \mathbf{a}, \mathbf{w}\rangle$ *is* $\mathcal{D}(\mathcal{A})-$*Lipschitz and used the Gaussian concentration inequality to show that it concentrates around its expectation. Now we can no longer apply the same result, because* $\boldsymbol{\varepsilon}$ *are not Gaussian and the function is not Lipschitz.*

Let's consider some properties of these two measures.

Definition 6 (Dual norm)**.** *Let* $\|\cdot\|$ *be a norm in* $\mathbb{R}^n$*. The associated dual norm is defined as*

$$\|\mathbf{x}\|_* = \sup\left\{\langle \mathbf{x}, \mathbf{z}\rangle : \|\mathbf{z}\| \leq 1\right\}.$$

If we consider the ℓ_p-*norm, then*

$$\|\mathbf{x}\|_* \equiv \sup_{\mathbf{z}\in\mathcal{B}_p(1)} \langle \mathbf{x}, \mathbf{z}\rangle.$$

By Holder's inequality, the dual norm of the ℓ_p-*norm is the* ℓ_q-*norm, where* $\frac{1}{p}+\frac{1}{q}=1$ *that is* $q=\frac{p}{p-1}$*.*

Let $1<p<\infty$*, then define the* ℓ_p*-ball of radius* r *as*

$$\mathcal{B}_p(r) \equiv \{\mathbf{x}\in\mathbb{R}^n : \|\mathbf{x}\|_p \leq r\}$$

Then, the Gaussian complexity of $\mathcal{B}_p(r)$ *is*

$$\mathcal{G}(\mathcal{B}_p(r)) = \mathbb{E}\left[\sup_{\|\mathbf{x}\|_p\leq r}\langle \mathbf{x}, \mathbf{w}\rangle\right] = r\cdot\mathbb{E}\left[\sup_{\mathcal{B}_p(1)}\langle \mathbf{x}, \mathbf{w}\rangle\right] = r\cdot\mathbb{E}\left[\|\mathbf{w}\|_q\right] = O\left(n^{\frac{1}{q}}\right),$$

where the last equality follows from the fact that

$$\mathbb{E}[\|\mathbf{w}\|_q] \leq \left(\sum_{i=1}^n \underbrace{\mathbb{E}[|w_i|^q]}_{O(1)}\right)^{1/q} \approx O(n^{1/q}).$$

The Rademacher complexity of $\mathcal{B}_p(r)$ *is*

$$\mathcal{R}\left(\mathcal{B}_p(r)\right) = \mathbb{E}\left[\sup_{a\in\mathcal{B}_p(r)}\langle a, \boldsymbol{\varepsilon}\rangle\right] = \mathbb{E}\left[\|\boldsymbol{\varepsilon}\|_q\right] = \mathbb{E}\left[\left(\sum_{i=1}^n |\varepsilon_i|^q\right)^{\frac{1}{q}}\right] = n^{1/q}.$$

Therefore, both complexity are of the same order. Now suppose $p=1$*. Recall that the dual norm of the* ℓ_1-*norm is the* $\ell_\infty-$*norm. Thus*

$$\mathcal{R}\left(\mathcal{B}_1(1)\right) = \mathbb{E}\left[\sup_{\mathbf{a}\in\mathcal{B}_1(1)}\langle \mathbf{a}, \boldsymbol{\varepsilon}\rangle\right] = \mathbb{E}\left[\|\boldsymbol{\varepsilon}\|_\infty\right] = 1,$$

$$\mathcal{G}(\mathcal{B}_1(r)) = \mathbb{E}\left[\sup_{\mathbf{x}\in\mathcal{B}_1(1)}\langle \mathbf{x}, \mathbf{w}\rangle\right] = \mathbb{E}\left[\|\mathbf{w}\|_\infty\right] \approx \sqrt{2\log n}.$$

Define $f(\boldsymbol{\varepsilon}) = \sup_{\mathbf{a}\in\mathcal{A}}\langle \mathbf{a}, \boldsymbol{\varepsilon}\rangle$. We have already shown that this function is $\mathcal{D}(\mathcal{A})-$Lipschitz. Now, note that $\langle\cdot,\cdot\rangle$ is linear hence affine. Since the supremum of affine functions is convex, we have that $f(\boldsymbol{\varepsilon})$ is convex, $\mathcal{D}(\mathcal{A})-$Lipschitz, and bounded in $[-1,1]$. Therefore

$$f(\boldsymbol{\varepsilon}) \sim \mathsf{sG}(2\mathcal{D}(\mathcal{A})),$$

which is of the same order as the Gaussian complexity!

Later on in this class we will use the Rademacher complexity or the Gaussian complexity to upper bound other quantities of interest and show that they concentrate too! ♣

1.8.2 Log-Concavity

Definition 7 (Log-concave). *Let $f : \mathbb{R}^n \to \mathbb{R}$. We say that f is $\gamma-$strongly log-concave if $f(x) = \exp(-\psi(x))$ and $\psi(x)$ is $\gamma-$strongly convex, i.e., $\nabla^2\psi(x) \succeq \gamma I_n$ if exists.*

The name comes from the fact that $\log f(x)$ is strongly concave in the traditional sense.

Example 19. *Suppose we have a random vector X whose distribution belongs to the exponential family, i.e.,*

$$P_\theta(x) = \frac{1}{z(\theta)}\exp(\langle\theta, T(x)\rangle).$$

Suppose we have a prior $\pi(\theta) \sim N(0, I_n)$. The posterior distribution is

$$p(\theta \mid x) \propto P_\theta(x)\pi(\theta) \propto= \exp(-\psi(\theta))$$

where $\psi(\theta) = -\langle\theta, T(x)\rangle + \frac{1}{2}\|\theta\|_2^2 + \log z(\theta)$. Strong convexity of $\psi(\cdot)$ comes from the fact that the first element as null Hessian, the second has identity, and, as a result from properties of the exponential family, $\nabla^2 \log z(\theta) = \mathbb{V}(T(X))$ which is always positive definite. In particular, $\psi(\cdot)$ is $1-$strongly convex. ♣

Proposition 26 (log-concave). *Let $X \sim \mu$ be a random vector, where μ is $\gamma-$strongly log-concave. Suppose that f is L-Lipschitz. Then*

$$f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \sim \mathsf{sG}(L/\sqrt{\gamma}).$$

This is just a bound for the upper tail.

Note. *Note that this is the first concentration inequality in which we didn't require independence or X being a Martingale.*

When γ is equal to 1 it is equivalent to the Gaussian Concentration inequality, but we haven't assumed normality of the Xs. ♦

1.9 Efron-Stein Inequality

We first prove the following auxiliary lemma. It is basically a decomposition of the variance of Z in n terms. In each term the expectation $\mathbb{E}[X]$ is substituted by the conditional expectation of Z on all but one of the X_is.

Lemma 2. *Let $X_1, X_2, \dots, X_n$ be independent random variable and $Z = f(X_1, \dots, X_n)$ be a square integrable function. Let $X_{-\ell} = (X_1, \dots, X_{\ell-1}, X_{\ell+1}, \dots, X_n)'$. Then*

$$\mathbb{V}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[(Z - \mathbb{E}[Z \mid X_{-i}])^2].$$

Proof. First, define the martingale difference sequence $D_\ell := \mathbb{E}[Z \mid X_{1:\ell}] - \mathbb{E}[Z \mid X_{1:\ell-1}], \ell = 2, 3, \dots, n$ with $\mathbb{E}[Z \mid X_{1:n}] = Z$ since Z is $\sigma(X_{1:n})-$measurable and $D_1 := \mathbb{E}[Z \mid X_1] - \mathbb{E}[Z]$. When we sum over this mds all the terms simplify except for Z and $\mathbb{E}[Z]$. Therefore, we can use the telescopic sum trick to rewrite

$$Z - \mathbb{E}[Z] = \sum_{\ell=1}^{n} (\mathbb{E}[Z \mid X_{1:\ell}] - \mathbb{E}[Z \mid X_{1:\ell-1}]).$$

Thus

$$\mathbb{V}(Z) = \mathbb{E}[(Z - \mathbb{E}[Z])^2] = \mathbb{E}[(\sum_{\ell=1}^{n} D_\ell)^2].$$

By the law of iterated expectations, if $\{D_\ell\}_\ell$ is an mds with respect to the filtration $\{\mathcal{F}_\ell\}_\ell$, then it is also uncorrelated, indeed fix $j, k \in [n], j > k$. Then

$$\mathbb{E}[D_j D_k] = \mathbb{E}[\mathbb{E}[D_j D_k \mid \mathcal{F}_k]] = \mathbb{E}[\mathbb{E}[D_j \mid \mathcal{F}_k] D_k] = 0.$$

Therefore,

$$\mathbb{V}(Z) = \mathbb{E}[(\sum_{\ell=1}^{n} D_\ell)^2] = \sum_{\ell=1}^{n} \mathbb{E}[D_\ell^2].$$

Finally, note the following key fact

$$\begin{aligned} \mathbb{E}_{X_{\ell:n}}[Z \mid X_{1:\ell-1}] &= \mathbb{E}_{X_{\ell+1:n}}[\mathbb{E}_{X_\ell}[Z \mid X_{1:\ell-1}, X_{\ell+1:n}] \mid X_{1:\ell-1}] \\ &= \mathbb{E}_{X_{\ell+1:n}}[\mathbb{E}_{X_\ell}[Z \mid X_{1:\ell-1}, X_{\ell+1:n}] \mid X_{1:\ell}], \end{aligned}$$

where the first equality comes from the tower rule of expectations and the second one comes from the fact that the inner expectation is a function of $X_{-\ell}$ and $X_\ell \perp\!\!\!\perp X_j, \forall\, j \neq \ell$ so we can add it to the outer conditioning set. Thus

$$\begin{aligned} D_\ell^2 &= (\mathbb{E}_{X_{\ell+1:n}}[Z \mid X_{1:\ell}] - \mathbb{E}_{X_{\ell:n}}[Z \mid X_{1:\ell-1}])^2 \\ &= (\mathbb{E}_{X_{\ell+1:n}}[Z \mid X_{1:\ell}] - \mathbb{E}_{X_{\ell+1:n}}[\mathbb{E}_{X_\ell}[Z \mid X_{1:\ell-1}, X_{\ell+1:n}] \mid X_{1:\ell}])^2 \\ &= (\mathbb{E}_{X_{\ell+1:n}}[Z - \mathbb{E}_{X_\ell}[Z \mid X_{1:\ell-1}, X_{\ell+1:n}] \mid X_{1:\ell}])^2 \\ &\leq \mathbb{E}_{X_{\ell+1:n}}[(Z - \mathbb{E}_{X_\ell}[Z \mid X_{1:\ell-1}, X_{\ell+1:n}])^2 \mid X_{1:\ell}], \qquad \text{(Jensen's)} \end{aligned}$$

then taking the expectation with respect to $X_{1:\ell}$ on both sides we get

$$\forall\, \ell \in [n],\, \mathbb{E}[D_\ell^2] \leq \mathbb{E}[(Z - \mathbb{E}[Z \mid X_{-\ell}])^2] \implies \mathbb{V}(Z) \leq \sum_{\ell=1}^{n} \mathbb{E}[(Z - \mathbb{E}[Z \mid X_{-\ell}])^2],$$

which was to be shown. ■

Proposition 27 (Efron-Stein Inequality)**.** *Let $X_1, X_2, \dots, X_n$ be independent random variable and $Z = f(X_1, \dots, X_n)$ be a square integrable function. Moreover, if $X_1', \dots, X_n'$ are independent copies of $X_1, \dots, X_n$ and if we define, for every $i = 1, \dots, n$*

$$Z_i' = f\left(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n\right),$$

we have

$$\mathbb{V}(Z) \leq \frac{1}{2}\sum_{\ell=1}^{n}\mathbb{E}[(Z - Z'_\ell)^2] = \sum_{\ell=1}^{n}\mathbb{E}[(Z - Z'_\ell)_+^2] = \sum_{\ell=1}^{n}\mathbb{E}[(Z - Z'_\ell)_-^2],$$

where $x_+ = \max(x, 0)$ *and* $x_- = \max(-x, 0)$ *denote the positive and negative parts of a real number* x*, respectively.*

Proof. The first part of the Efron-Stein inequality follows from the fact that if X, Y are iid and, wlog, mean-zero, then

$$\frac{1}{2}\,\mathbb{E}[(X - Y)^2] = \frac{1}{2}\,\mathbb{E}[X^2 + Y^2 - 2XY] = \frac{1}{2}\,\mathbb{E}[X^2 + \mathbb{E}[Y^2|X] - 2X\,\mathbb{E}[Y \mid X]] = \mathbb{V}(X).$$

Thus, since Z and each Z'_ℓ are iid and using the previous lemma

$$\mathbb{V}(Z) \leq \sum_{\ell=1}^{n}\mathbb{E}[(Z - \mathbb{E}[Z \mid X_{-\ell}])^2] = \frac{1}{2}\sum_{\ell=1}^{n}\mathbb{E}[(Z - Z'_\ell)^2].$$

Let's now prove the other two equalities. Let Y be a rv symmetric around 0, then

$$\mathbb{E}[Y^2] = \int_{\infty}^{0} y^2\,\mathrm{d}F(y) + \int_{0}^{\infty} y^2\,\mathrm{d}F(y) = 2\int_{0}^{\infty} y^2\,\mathrm{d}F(y) = 2\int_{0}^{\infty}\max\{0, y\}^2\,\mathrm{d}F(y),$$

and

$$\mathbb{E}[Y^2] = \int_{\infty}^{0} y^2\,\mathrm{d}F(y) + \int_{0}^{\infty} y^2\,\mathrm{d}F(y) = 2\int_{-\infty}^{0} y^2\,\mathrm{d}F(y) = 2\int_{-\infty}^{0}\max\{0, -y\}^2\,\mathrm{d}F(y).$$

Since Z, Z' are iid, $X - X'$ is symmetric around 0 which concludes the proof. ■

Corollary 1 (Jackknife Bound)**.** *Let* $Z_{-\ell} := h_\ell(X_1, \ldots, X_{\ell-1}, X_{\ell+1}, \ldots, X_n)$ *for arbitrary measurable functions* $h_\ell : \mathcal{X}^{n-1} \to \mathbb{R}$*. Then*

$$\mathbb{V}(Z) \leq \sum_{\ell=1}^{n}\mathbb{E}[(Z - Z_{-\ell})^2]$$

Proof. For any rv X with finite second moment $\mathbb{E}[(X - \mathbb{E}[X])^2 \leq \mathbb{E}[(X - a)^2], \forall\, a \in \mathbb{R}$. Using this fact conditionally

$$\mathbb{E}[(Z - \mathbb{E}[Z \mid X_{-\ell}])^2 \mid X_{-\ell}] \leq \mathbb{E}[(Z - Z_{-\ell})^2 \mid X_{-\ell}],$$

where the second term is well-defined as long as h_ℓ are measurable with respect to $\sigma(X_{-\ell})$. Taking expectations wrt X on both sides and using the tower rule concludes the proof. ■

This corollary is useful to prove that the Jackknife estimator (Efron and Stein 1981) yields a conservative estimate of the variance of an estimator of the form $\widehat{\theta} = f(X)$.

Corollary 2. *If* f *has the bounded difference property with constants* $c_1, \ldots, c_n$*, then Efron-Stein implies that*

$$\mathbb{V}(Z) \leq \frac{1}{2}\sum_{i=1}^{n} c_i^2.$$

Intermezzo: Stochastic Processes and Uniform Convergence

Pointwise and Uniform Convergence

Definition 8 (Pointwise Convergent)**.** *Let (X, d) be a metric space and $f_n : X \to \mathbb{R}(n \in \mathbb{N})$ a sequence of functions. Then f_n converges pointwise to f if*

$$\forall\, x \in X, \forall\, \varepsilon > 0, \exists\, N_{\varepsilon,x} \in \mathbb{N} : d(f_n(x), f(x)) < \varepsilon, \forall\, n \geq N_{\varepsilon,x}$$

Example 20. *The sequence of functions $f_n(x) = x^n/n$ converges pointwise to zero on the interval $X = [-1, 1]$, because for each $x \in [-1, 1]$ one has $|x^n/n| \leq 1/n$, and thus*

$$\lim_{n\to\infty} \frac{x^n}{n} = 0$$

♣

Note. *Pointwise convergent sequence of functions might have some problems*

1. *The limit of a pointwise convergent sequence of continuous functions does not have to be continuous.*
2. *The derivatives of a pointwise convergent sequence of functions do not have to converge.*
3. *The integrals of a pointwise convergent sequence of functions do not have to converge.*

An example of 1. is $f_n(x) = x^n \mathbb{1}_{[0,1]}(x)$. If we fix $x \in [0, 1)$, then $\lim_{n\to\infty} f_n(x) = 0$, if instead $x = 1$, then $\lim_{n\to\infty} f_n(x) = 1$. ♦

Definition 9 (Uniform Convergent)**.** *Let (X, d) be a metric space and $f_n : X \to \mathbb{R}(n \in \mathbb{N})$ a sequence of functions. Then f_n converges uniformly to f if*

$$\forall\, \varepsilon > 0, \exists\, N_\varepsilon \in \mathbb{N} : d(f_n(x), f(x)) < \varepsilon, \forall\, n \geq N_\varepsilon, \forall\, x \in X,$$

Note. *Uniform convergence requires the same ε to be valid for all points in the domain. It can be thought of as requiring the sequence of functions to lay in a sleeve.*

If the functions are bounded, the convergence above is equivalent to require that

$$\sup_{x\in X} |f_n(x) - f(x)| \to 0.$$

Uniform convergence preserves many properties of the converging sequence:

1. *The limit of a uniformly convergent sequence of continuous functions is continuous.*
2. *The limit of a uniformly convergent sequence of integrable functions is integrable*
3. *The limit of a uniformly convergent sequence of differentiable functions is differentiable*

♦

Stochastic Processes and Empirical Processes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathbb{I}$ be an index set, and let S be the state space (usually a locally compact, complete metric space).

Definition 10 (Stochastic process)**.** *A stochastic process is a defined as a measurable function $i \mapsto X_i$, $X : \mathbb{I} \times \Omega \to S$ or, alternatively, $\forall\, i \in \mathbb{I}$, $X(i, \cdot) : \Omega \to S$. It can be characterized in the following ways:*

1. *$i \in \mathbb{I}$ fixed, then $X(i, \cdot) : \Omega \to S$ is an $S-$valued **random variable** (marginal).*
2. *$\omega \in \Omega$ fixed. Then $X(\cdot, \omega) : \mathbb{I} \to S$ is a **random trajectory**.*
3. *$X : \Omega \to S^{\mathbb{I}} = \{$space of S-valued functions with domain $\mathbb{I}\}$*

If $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with cumulative distribution function (c.d.f.) F then the empirical distribution function (e.d.f.) $\mathbb{F}_n : \mathbb{R} \to [0, 1]$ is defined as

$$\mathbb{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(X_i), \qquad \text{for } x \in \mathbb{R}.$$

In other words, for each $x \in \mathbb{R}$, the quantity $n\mathbb{F}_n(x)$ simply counts the number of X_i 's that are less than or equal to x. The e.d.f. is a natural unbiased (i.e., $\mathbb{E}[\mathbb{F}_n(x)] = F(x)$ for all $x \in \mathbb{R}$) estimator of F. The corresponding empirical process is

$$\mathbb{G}_n(x) = \sqrt{n}\,(\mathbb{F}_n(x) - F(x)), \qquad \text{for } x \in \mathbb{R}$$

Note that both $\mathbb{F}_n$ and $\mathbb{G}_n$ are stochastic processes (i.e., random functions) indexed by the real line ($\mathbb{I} = \mathbb{R}$ using the above notation). By the strong law of large numbers (SLLN), for every $x \in \mathbb{R}$, we can say that

$$\mathbb{F}_n(x) \overset{a.s.}{\to} F(x) \qquad \text{as } n \to \infty.$$

Also, by the central limit theorem (CLT), for each $x \in \mathbb{R}$, we have

$$\mathbb{G}_n(x) \overset{d}{\to} N(0, F(x)(1 - F(x))) \qquad \text{as } n \to \infty.$$

Two of the basic results in empirical process theory concerning $\mathbb{F}_n$ and $\mathbb{G}_n$ are the *Glivenko-Cantelli* and *Donsker* theorems. These results generalize the above two results to processes that hold for all x simultaneously.

2 Uniform Laws of Large Numbers

Reference: Wainwright (2019), Ch. 4.

2.1 Uniform Convergence for CDFs: Glivenko-Cantelli

Suppose $X_i \overset{\text{iid}}{\sim} X \in [0,1]$, where X has CDF $F(t) = \mathbb{P}(X \leq t)$. Consider the empirical CDF

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq t).$$

For fixed t, $\{\widehat{F}_n(t)\}_{n\geq 1}$ is a sum of iid variates, thus bu LLN we have that

$$\lim_{n\to\infty} \widehat{F}_n(t) - F(t) \qquad \text{a.s.},$$

which shows that the ECDF is pointwise consistent for the true CDF. What if we want to prove some stronger form of convergence, say, in the $\ell_\infty-$norm

$$\sup_{t\in[0,1]} |\widehat{F}_n(t) - F(t)| = \|F_n - F\|_\infty.$$

Note that $\|F_n - F\|_\infty$ is a random object!!

Note. *Pointwise convergence does not necessarily imply uniform convergence. We can come up with examples such that*

$$\forall\, t \in [0,1] \quad \lim_{n\to\infty} \widehat{f}_n(t) = f(t) \quad a.s. \quad but \quad \|f_n - f\|_\infty \not\to 0 \quad a.s.$$

♦

Theorem 1 (Glivenko (1933)-Cantelli (1933) theorem)**.** *Let $X_i \overset{iid}{\sim} X$, then*

$$\|F_n - F\|_\infty = \sup_{t\in[0,1]} |\widehat{F}_n(t) - F(t)| \to 0 \qquad a.s.$$

In this course, we will prove a quantitative version of this theorem, that is

$$\mathbb{P}\left(\|F_n - F\|_\infty \geq \sqrt{\frac{8\log(n+1)}{n}} + \delta\right) \leq \exp\left(-\frac{n\delta^2}{2}\right).$$

This result also implies a.s. convergence by Borel-Cantelli lemma.

Definition 11. *Let V be a vector space. A functional $\gamma : V \to \mathbb{R}$ is continuous at F in the sup-norm if*

$$\forall\, \varepsilon > 0, \exists\, \delta > 0 : \|G - F\|_\infty < \delta \implies |\gamma(G) - \gamma(F)| < \varepsilon$$

Proposition 28 (Continuous Mapping Theorem)**.** *Let X_n be a sequence of random variables. Then*

$$X_n \xrightarrow{\mathrm{d}} X \Rightarrow g(X_n) \xrightarrow{\mathrm{d}} g(X), \quad X_n \xrightarrow{\mathrm{p}} X \Rightarrow g(X_n) \xrightarrow{\mathrm{p}} g(X), \quad X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X).$$

Proof. We show p, and then a.s. Note that showing the latter would suffice as $a.s. \implies p \implies d$, however it is constructive to show all of them.

Suppose $X_n \xrightarrow{p} X$ and consider the continuous function $g(\cdot)$. By definition of continuity of $g(\cdot)$ at X

$$\forall\, \varepsilon > 0, \exists\, \delta_\varepsilon > 0 : \|X_n - X\| < \delta_\varepsilon \implies |g(X_n) - g(X)| < \varepsilon.$$

Hence whenever $\|X_n - X\| < \delta_\varepsilon$ realizes so does $|g(X_n) - g(X)| < \varepsilon$ but not the viceversa, hence the latter has a probability of realizing no smaller than the former. Therefore

$$\mathbb{P}(|g(X_n) - g(X)| < \varepsilon) \geq \mathbb{P}(|X_n - X| < \delta_\varepsilon) \to 1,$$

where the last statement follows from the convergence in probability of $X_n \to X$

$$\forall\, M > 0, \quad \mathbb{P}(|X_n - X| \leq M) \to 1$$

and picking $M = \delta_\varepsilon$.

Suppose $X_n \xrightarrow{a.s.} X$ and consider the continuous function $g(\cdot)$. Using a similar intuition as for the first proof, recall that continuous functions preserve limits, i.e.

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \implies \lim_{n\to\infty} g(X_n(\omega)) = g(X(\omega))$$

Thus

$$\mathbb{P}(\{\omega : \lim_{n\to\infty} g(X_n(\omega)) = g(X(\omega))\}) \geq \mathbb{P}(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$$

■

An implication of GCT and the CMT is that if $\gamma(F)$ is a functional of F, where $\gamma(\cdot)$ is continuous with respect to the $\ell_\infty-$norm, then uniform convergence implies that

$$\gamma(\widehat{F}_n) \to \gamma(F).$$

Example 21 (Goodness-of-fit test)**.**

$$\gamma(F) = \int (F(t) - F_0(t))^2 \, \mathrm{d}F(t).$$

♣

2.2 Uniform Laws for more general function classes

Consider $X_i \overset{\texttt{iid}}{\sim} X$, where $X \sim \mathbb{P}$ and consider the set of integrable functions

$$\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}, \mathbb{E}_{X\sim\mathbb{P}}[|f(X)|] < \infty\}.$$

Consider an empirical process indexed by $\mathcal{F}$:

$$\left\{ f \in \mathcal{F} : \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E}[f(x)] \right) \right\}$$

sometimes also written as

$$\left\{ f \in \mathcal{F} : \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(\cdot) - \mathbb{E}[f(\cdot)] \right) \right\}$$

Formally, an empirical process G over a class of functions $\mathcal{F}$ is a function that maps elements $(\omega, f) \in \Omega \times \mathcal{F}$ to, typically, $\mathbb{R}$. It is just a stochastic process where the index set $\mathbb{I}$ is a set of functions rather than $\mathbb{Z}$ or $\mathbb{R}$. As such $G(\cdot, f)$ is a random variable that converges in distribution to $N(0, \mathbb{V} f)$ for each single $f \in \mathcal{F}$. The beauty of $G(\cdot, \cdot)$ is that is allows us to make statements on the whole class $\mathcal{F}$ rather than for a single element f at a time.

Define the empirical distribution of X_i as

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Define the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right|$$

which **measures the absolute deviation between the sample average and the population average, uniformly over the class $\mathcal{F}$**. Note that the supremum is over functions, so we cannot apply our classical results on asymptotic theory that we know. Note that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ is a random variable, where the randomness comes from X. Therefore, we can apply our traditional results to $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$!

Definition 12. *We say that a class of functions $\mathcal{F}$ is a Glivenko-Cantelli (GC) class for $\mathbb{P}$ if and only if*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$$

as $n \to \infty$, where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability under $\mathbb{P}$. The class of functions $\mathcal{F}$ satisfies a strong Glivenko-Cantelli law if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{a.s.} 0$.

Note. *If $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then if $\mathcal{F}_2$ is GC for $\mathbb{P}$, then $\mathcal{F}_1$ is GC for $\mathbb{P}$. This is verified because $\sup_{f \in \mathcal{F}_1} \leq \sup_{f \in \mathcal{F}_2}$.* ♦

Example 22 (GC classes)**.** *An example of a GC class is the one related to GC theorem. Indeed, we know that $\|\widehat{F}_n - F\|_\infty \to 0$ a.s. Remember that in this case the statement*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right|$$

is simply

$$\|\widehat{F}_n - F\|_\infty = \sup_{t\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \le t) - \mathbb{E}[\mathbb{1}(X \le t)] \right|.$$

We can thus think of $\|\widehat{F}_n - F\|_\infty$ *as the supremum of an empirical process indexed by* $\mathcal{F}$*, where the class of functions is*

$$\mathcal{F} = \{\mathbb{1}(X \le t), t \in [0,1]\} \equiv \{\mathbb{1}(\cdot \le t), t \in [0,1]\}.$$

An example of a class of functions that is not GC is given by

$$\mathcal{F} = \{\mathbb{1}_S \mid S \in \mathcal{S}\}, \quad \mathcal{S} = \{S \subseteq [0,1] : \#S \in \mathbb{N}\},$$

when we have a sample of $X_i \overset{iid}{\sim} \mathbb{P} \in \mathcal{P}([0,1])$*. To see this*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{S\in\mathcal{S}} \left| \underbrace{\frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \in S)}_{=1} - \underbrace{\mathbb{E}[\mathbb{1}(X_i \in S)]}_{=0} \right| = 1,$$

The first equality follows because, for fixed n*, the realization* $\{X_1, X_2, \ldots, X_n\} \in [0,1]^n$ *is a finite set so it belongs to* $\mathcal{S}$*. The second equality follows because we assumed* X_i *to have density, hence* $\mathbb{P}(S) = 0, \forall\, S \in \mathcal{S}$*.* ♣

Example 23 (Empirical risk minimization)**.** *Why we care about max of empirical process in stats and ML? They lie at the heart of methods based on empirical risk minimization.*

Data distribution: $(X_i)_{i\in[n]} \sim \mathbb{P} \in \{\mathbb{P}_\theta : \theta \in \Theta\}$.

Loss function: $\ell : X \times \Theta \to \mathbb{R}$.

Empirical risk: $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(X_i;\theta)$

Empirical risk minimizer: $\widehat{\theta} = \operatorname*{argmin}_\theta R_n(\theta)$

Population risk minimizer: $\theta^\star = \operatorname*{argmin}_\theta R(\theta)$

Excess risk: $\varepsilon = R(\widehat{\theta}) - R(\theta^\star)$

To bound the excess risk, we typically decompose it

$$\varepsilon = R(\widehat{\theta}) - R_n(\widehat{\theta}) + \underbrace{R_n(\widehat{\theta}) - R_n(\theta^\star)}_{\le 0} + R_n(\theta^\star) - R(\theta^\star)$$

The pink term is non-positive because $\widehat{\theta} \in \arg\min_\theta R_n(\theta)$*. The green term can be handled using standard concentration inequalities but we cannot handle the orange one despite it looks similar. Why? Well, we can't use Hoeffding's inequality on the second one because the summands in*

$$R_n(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell\left(X_i;\hat{\theta}\right)$$

are not independent because $\widehat{\theta}$ is data dependent, whilst the summands in

$$R_n(\theta^\star) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(X_i; \theta^\star\right)$$

are iid!

We just need to handle the orange term, that is

$$R(\hat{\theta}) - R_n(\hat{\theta}) = \mathbb{E}[\ell(X;\hat{\theta})] - \frac{1}{n}\sum_{i=1}^{n} \ell\left(X_i;\hat{\theta}\right) \leq \sup_{\theta\in\Theta}\left|\mathbb{E}[\ell(X;\theta)] - \frac{1}{n}\sum_{i=1}^{n} \ell\left(X_i;\theta\right)\right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$$

We can see that the orange term is upper bounded by the supremum of the empirical process associated to the class of functions $\mathcal{F} = \{\ell(\cdot;\theta) : \theta \in \Theta\}$ (note that for each θ we have a different loss function). Note that also the green term is dominated by this term! Therefore we conclude that the excess risk is at most $2\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$. ♣

2.3 Uniform Laws via Rademacher Complexity

The previous examples motivated why we care about the empirical process $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$. We saw that it upper bounds the excess error in a risk minimization problem, for example. As such we know try to work with this term and construct bounds for it.

Definition 13. *Consider a set $\mathcal{A} \subseteq \mathbb{R}^n$ and let $\varepsilon := (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)', \varepsilon_i \overset{iid}{\sim} U(\{-1, 1\})$. The Rademacher complexity of the set $\mathcal{A}$ is defined as*

$$\mathcal{R}(\mathcal{A}) := \mathbb{E}_\varepsilon[\sup_{\mathbf{a}\in\mathcal{A}} \langle \mathbf{a}, \varepsilon\rangle].$$

Note. *Sometimes Rademacher complexity is defined as $\mathcal{R}(\mathcal{A}) := \mathbb{E}_\varepsilon[\sup_{\mathbf{a}\in\mathcal{A}} |\langle \mathbf{a}, \varepsilon\rangle|]$, but this is just equal to*

$$\mathcal{R}(|\mathcal{A}|) = \mathbb{E}_\varepsilon[\sup_{\mathbf{a}\in|\mathcal{A}|} \langle \mathbf{a}, \varepsilon\rangle],$$

where $|A| := \mathcal{A} \cup (-\mathcal{A})$. In practice $\mathcal{R}(\mathcal{A})$ and $\mathcal{R}(|\mathcal{A}|)$ are of the same order so it does not matter with which one we work. ♦

The above definition regards sets, but we can have a similar definition for classes of functions! To achieve this let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$ be a class of functions and consider $(x_i)_{i\in[n]} \in \mathcal{X}$ be any fixed collections of points. Define the set of all vectors in $\mathbb{R}_n$ that can be obtained by applying a function $f \in \mathcal{F}$ to such collection as

$$\mathcal{F}(x_{1:n}) \equiv \{(f(x_1), \ldots, f(x_n) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n,$$

Note that the set $\mathcal{A} = \mathcal{F}(x_{1:n})/n$ in a set in $\mathbb{R}^n$, thus we can just apply the previous definition of Rademacher complexity to it.

Definition 14 (Empirical Rademacher Complexity of function class $\mathcal{F}$ on $x_{1:n}$)**.** *The Empirical Rademacher complexity of function class $\mathcal{F}$ is*

$$\mathcal{R}\left(\mathcal{F}(x_{1:n})/n\right) = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(x_i\right)\right|\right].$$

If we think of the collection of points $x_{1:n}$ as a random sample of X, then we obtain the following definition.

Definition 15 (Rademacher Complexity of function class $\mathcal{F}$ with measure $\mathbb{P}$)**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$ be a class of functions and consider $(X_i)_{i\in[n]} \overset{iid}{\sim} \mathbb{P} \in \mathcal{P}(\mathcal{X})$. Then, the Rademacher complexity of function class $\mathcal{F}$ with measure $\mathbb{P}$ is*

$$\mathcal{R}_n\left(\mathcal{F}\right) := \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right)\right|\right].$$

Note. *First, note that*

$$\begin{aligned}
\mathcal{R}_n\left(\mathcal{F}\right) &= \mathbb{E}_X\left[\mathcal{R}\left(\mathcal{F}(x_{1:n})/n\right)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right)\right| \mid X_{1:n}\right]\right] \\
&= \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right)\right|\right]. \qquad \text{(tower rule)}
\end{aligned}$$

Here is some intuition for why the Rademacher complexity is a measure of the complexity of a function class. First note that if $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then $\mathcal{R}(\mathcal{F}_1) \leq \mathcal{R}(\mathcal{F}_2)$. Note that the Rademacher complexity is the average of the maximum correlation between the vector $(f(X_1), \ldots, f(X_n))$ and the "noise vector" $(\varepsilon_1, \ldots, \varepsilon_n)$, where the maximum is taken over all functions $f \in \mathcal{F}$. The intuition is a natural one: a function class is extremely large–and, in fact, "too large" for statistical purposes–if we can always find a function that has a high correlation with a randomly drawn noise vector. Conversely, when the Rademacher complexity decays as a function of sample size, then it is impossible to find a function that correlates very highly in expectation with a randomly drawn noise vector.

♦

Example 24. *Let $\psi : \mathbb{R}^d \to \mathbb{R}^p$ be a fixed feature map and consider the linear function class $\mathcal{F} = \{f(x) = \langle\psi(x), \theta\rangle : ||\theta||_2 \leq B, \theta \in \mathbb{R}^p\}$. Note that if $p = 1$ it's a linear regression! Note that it is linear in the parameter θ, not necessarily in X! This should always be the first class we consider.*

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_{X_i,\varepsilon_i}\left[\sup_{||\theta||_2\leq B}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle\psi(X_i),\theta\rangle\right|\right] \\
&= \mathbb{E}_{X_i,\varepsilon_i}\left[\sup_{||\theta||_2\leq B}\left|\frac{1}{n}\left\langle\sum_{i=1}^{n}\varepsilon_i\psi(X_i),\theta\right\rangle\right|\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{X_i,\varepsilon_i}\left[\left|\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\psi(X_i)\right|\right|_2\right]\cdot B && (\ell_2-\ell_2 \text{ dual})\\
&\leq \mathbb{E}_{X_i,\varepsilon_i}\left[\left|\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\psi(X_i)\right|\right|_2^2\right]^{\frac{1}{2}}\cdot B && (\text{Jensen's})\\
&= \mathbb{E}_{X_i,\varepsilon_i}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}\varepsilon_i\varepsilon_j\langle\psi(X_i),\psi(X_j)\rangle\right]^{\frac{1}{2}}\cdot B\\
&= \mathbb{E}_{X_i,\varepsilon_i}\left[\frac{1}{n^2}\sum_{i=1}^{n}\varepsilon_i^2||\psi(X_i)||_2^2\right]^{\frac{1}{2}}\cdot B && (\ \mathbb{E}[\varepsilon_i\varepsilon_j]=0,\ i\neq j))\\
&= \frac{B}{\sqrt{n}}\,\mathbb{E}[||\psi(X)||_2^2]^{1/2}
\end{aligned}
$$

In this example, when B is large or when the expected norm of ψ is large, the Rademacher complexity is large. Whereas it gets lower as n increases. This aligns with our intuition for what makes $\mathcal{F}$ more complex. This is always true for classes of linear functions. ♣

The reason we introduce Rademacher complexity is because we can show that the supremum of an empirical processes - $||\mathbb{P}_n-\mathbb{P}||_{\mathcal{F}}$ in our notation - is almost equivalent to the Rademacher complexity. Moreover, we can also show ways of bounding the Rademacher complexity. This will give us a technique to bound $||\mathbb{P}_n-\mathbb{P}||_{\mathcal{F}}$, which is the object we care about. We start with showing the near equivalence of $||\mathbb{P}_n-\mathbb{P}||_{\mathcal{F}}$ and $\mathcal{R}_n(\mathcal{F})$, then we will show how $\mathcal{R}_n(\mathcal{F})$ provides bounds to $||\mathbb{P}_n-\mathbb{P}||_{\mathcal{F}}$.

2.3.1 Upper and Lower Bounding $\mathbb{E}\,\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}$ by $\mathcal{R}_n(\mathcal{F})$

In what follows we define the useful quantity

$$\|\mathbb{S}_n\|_{\mathcal{F}}=\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right)\right|$$

and the recentered function class

$$\overline{\mathcal{F}}=\{f-\mathbb{E}_X[f(X)],f\in\mathcal{F}\}.$$

Note. *Note that*

$$\mathbb{E}_{X,\varepsilon}\left[\|\mathbb{S}_n\|_{\mathcal{F}}\right]=\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right)\right|\right]=:\mathcal{R}_n(\mathcal{F})$$

and

$$\|\mathbb{P}_n-\mathbb{P}\|_{\overline{\mathcal{F}}}=\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}$$

because

$$\begin{aligned}\|\mathbb{P}_n - \mathbb{P}\|_{\overline{\mathcal{F}}} &:= \sup_{g\in\overline{\mathcal{F}}}\left|\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}[g(X_i)]\right| \\ &= \sup_{(f-\mathbb{E}\,f)\in\overline{\mathcal{F}}}\left|\frac{1}{n}\sum_{i=1}^n [f(X_i) - \mathbb{E}[f(X_i)]] - \mathbb{E}[f(X_i) - \mathbb{E}[f(X_i)]]\right| \\ &= \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]\right| =: \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}.\end{aligned}$$

♦

Lemma 3. *Let $\Phi(\cdot)$ be a convex non-decreasing function. Then*

$$\sup_{f\in\mathcal{F}}\Phi(\mathbb{E}[|f(X)|]) \le \mathbb{E}[\Phi(\sup_{f\in\mathcal{F}}|f(X)|)].$$

Proof. First of all, note that $|f(X)| \le |f(X)|$ a.s. $\forall\, f \in \mathcal{F}$, thus a fortiori $|f(X)| \le \sup_{f\in\mathcal{F}}|f(X)|$ a.s. Then, taking expectations on both sides we get $\mathbb{E}[|f(X)|] \le \mathbb{E}[\sup_{f\in\mathcal{F}}|f(X)|]$ which also holds for any $f \in \mathcal{F}$, therefore

$$\sup_{f\in\mathcal{F}}\mathbb{E}[|f(X)|] \le \mathbb{E}[\sup_{f\in\mathcal{F}}|f(X)|].$$

Since $\Phi(\cdot)$ is a non-decreasing function we get

$$\Phi(\sup_{f\in\mathcal{F}}\mathbb{E}[|f(X)|]) \le \Phi(\mathbb{E}[\sup_{f\in\mathcal{F}}|f(X)|]).$$

Moreover, since it is non-decreasing we can swap it with the sup on the LHS and since it is convex we can use Jensen's inequality on the RHS to swap it with the expectation and get

$$\sup_{f\in\mathcal{F}}\Phi(\mathbb{E}[|f(X)|]) \le \mathbb{E}[\Phi(\sup_{f\in\mathcal{F}}|f(X)|)].$$

■

Note. *The lemma above is a fancier version of the idea that the maximum of a sum can't exceed the sum of the maxima, simply because the second one "has more degrees of freedom/-control variables".* ♦

The next proposition gives an upper and a lower bound for $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$.

Proposition 29. *For any convex non-decreasing function $\Phi : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\|\mathbb{S}_n\|_{\overline{\mathcal{F}}}\right)\right] \overset{(a)}{\le} \mathbb{E}_X\left[\Phi\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right)\right] \overset{(b)}{\le} \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\|\mathbb{S}_n\|_{\mathcal{F}}\right)\right].$$

Note. *When applied with the convex non-decreasing function $\Phi(t) = t$, the above proposition gives yields the inequalities*

$$\frac{1}{2}\mathcal{R}_n(\overline{\mathcal{F}}) = \frac{1}{2}\mathbb{E}_{X,\varepsilon}\|\mathbb{S}_n\|_{\overline{\mathcal{F}}} \le \mathbb{E}_X\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right] \le 2\mathbb{E}_{X,\varepsilon}\|\mathbb{S}_n\|_{\mathcal{F}} = 2\mathcal{R}_n(\mathcal{F}).$$

Therefore the expected value of our empirical process is upper and lower bounded by the same object (up to a constant) which turns out to be the Rademacher complexity of the indexing class of functions.

The lower bound must involve $\overline{\mathcal{F}}$, whilst the upper bound can involve either $\mathcal{F}$ or $\overline{\mathcal{F}}$. For the upper bound we can also have $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \|\mathbb{P}_n - \mathbb{P}\|_{\bar{\mathcal{F}}} \leq 2\mathcal{R}_n(\overline{\mathcal{F}})$. ♦

Proof. Beginning with bound (b). Let Y_i be an independent copy of X_i. We have

$$\begin{aligned}
\mathbb{E}_X\left[\Phi\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right)\right] &:= \mathbb{E}_X\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_Y\left[f(Y_i)\right]\right|\right)\right] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}_{X,Y}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\right|\right)\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left\{f(X_i) - f(Y_i)\right\}\right|\right)\right] \\
&\overset{\text{(iii)}}{\leq} \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i)\right|\right)\right] \\
&\overset{\text{(iv)}}{\leq} \frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right)\right] + \frac{1}{2}\mathbb{E}_{Y,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i)\right|\right)\right] \\
&\overset{\text{(v)}}{=} \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right)\right] = \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\|\mathbb{S}_n\|_{\mathcal{F}}\right)\right] = \mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

where inequality (i) follows from Jensen's inequality to bring $\mathbb{E}_Y[\cdot]$ out of the absolute value and then applies the previous lemma to $\Phi(\sup \mathbb{E}_Y[|h(X)|]$; equality (ii) is from the symmetrization trick, which relies on the distribution of $f(X_i) - f(Y_i)$ being symmetric because $X_{i\in[n]}$ and $Y_{i\in[n]}$ are iid; step (iii) follows by the triangle inequality; step (iv) follows from Jensen's inequality and the convexity of Φ; step (v) follows since X and Y are i.i.d. samples.

Similarly, turning to the bound (a), we need to start with the demeaned class of functions $\overline{\mathcal{F}}$ because we need the expectation to pop out at the beginning of the proof. Thus

$$\begin{aligned}
\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\|\mathbb{S}_n\|_{\bar{\mathcal{F}}}\right)\right] &= \mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}|\ \frac{1}{n}\sum_{i=1}^n \varepsilon_i\left\{f(X_i) - \mathbb{E}_{Y_i}\left[f(Y_i)\right]\right\}\right)\right] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}|\ \frac{1}{n}\sum_{i=1}^n \varepsilon_i\left\{f(X_i) - f(Y_i)\right\}\right)\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_{X,Y}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \left\{f(X_i) - \mathbb{E}[X_i] + \mathbb{E}[Y_i] - f(Y_i)\right\}\right|\right)\right] \\
&\overset{\text{(iii)}}{\leq} \mathbb{E}_{X,Y}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]\right| + \frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}[f(Y_i)]\right|\right)\right] \\
&\overset{\text{(iv)}}{\leq} \frac{1}{2}\mathbb{E}_X\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \left\{f(X_i) - \mathbb{E}[f(X_i)]\right\}\right|\right)\right] \\
&\quad + \frac{1}{2}\mathbb{E}_Y\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \left\{f(Y_i) - \mathbb{E}[f(Y_i)]\right\}\right|\right)\right] \\
&= \mathbb{E}_X\left[\Phi\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right)\right]
\end{aligned}$$

where inequality (i) and (iv) follows from Jensen's inequality; equality (ii) from the symmetric distribution of $f(X_i) - f(Y_i)$ and adding and subtracting the same quantity; inequality (iii) follows from triangle inequality. ■

If we now recall that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ is a random variable, then we can simply use the results we obtained in the previous section to characterize its concentration.

Theorem 2 (Concentration of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$)**.** *Suppose $\forall f \in \mathcal{F}, \|f\|_\infty \leq b$, i.e., the function class is uniformly bounded. Then with probability $1-\delta$,*

$$\frac{1}{2}\mathcal{R}_n(\overline{\mathcal{F}}) - b\sqrt{\frac{2\log(2/\delta)}{n}} \leq \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + b\sqrt{\frac{2\log(2/\delta)}{n}}$$

Note. *The theorem shows that if $\mathcal{R}_n(\mathcal{F}) = o(1)$, as $n \to \infty$, then $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$. To have almost sure convergence we need summability of the bounds $\mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq t) \leq \delta(n,t)$ in order to apply Borel-Cantelli.* ♦

Proof. First, we show that the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ has the bounded difference property with coefficients $\left(\frac{2b}{n}, \cdots, \frac{2b}{n}\right)$. Recall that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \|\mathbb{P}_n - \mathbb{P}\|_{\overline{\mathcal{F}}}$, so we work with the latter to simplify things. Then, pick $\bar{f}, \bar{h} \in \overline{\mathcal{F}}$ and perturb the first component of $x_{1:n}$

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \sup_{h\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \bar{h}(y_i)\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(y_i)\right| \\ &\leq \frac{1}{n}\left|\bar{f}(x_1) - \bar{f}(y_1)\right| \leq \frac{2b}{n}.\end{aligned}$$

The function $\|\mathbb{P}_n - \mathbb{P}\|_{\overline{\mathcal{F}}}$ is invariant to permutations of x, thus we conclude that such function has the bounded difference property with coefficients $(\frac{2b}{n}, \cdots, \frac{2b}{n})$.

Therefore, by the bounded difference inequality (Proposition 19), we have that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \sim \mathsf{sG}(2b/\sqrt{n})$ and thus with probability $1-\delta$ we have that

$$\left|\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right]\right| \leq b\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Applying Proposition 29, with $\Phi(\cdot)$ equal to the identity function we conclude the proof. ■

Proposition 30. *For any distribution $\mathbb{P}$ and any function class $\mathcal{F}$,*

$$\mathbb{E}[||\mathbb{P}_n - \mathbb{P}||_{\mathcal{F}}] \leq 2\mathcal{R}_n(\overline{\mathcal{F}}).$$

Proof. Just apply Proposition 29 with $\overline{\mathcal{F}}$ by noting that $\overline{\overline{\mathcal{F}}} = \overline{\mathcal{F}}$. ■

3 Bounding the Rademacher Complexity

In this section we will see various techniques to bound the Rademacher complexity. In turn, this is useful to bound the supremum of empirical processes.

3.1 Bounds of $\mathcal{R}_n(\mathcal{F})$ via Maximal Inequality

Lemma 4 (Maximal inequality). *Suppose Θ is finite, if $\forall\, \theta \in \Theta$, $X_\theta \sim \mathsf{sG}(\sigma)$ and mean-zero, we have*

$$\mathbb{E}\left[\max_{\theta\in\Theta} X_\theta\right] \le \sqrt{2\sigma^2 \log|\Theta|}$$

where $|\Theta|$ is the cardinality of set Θ.

However when $|\Theta| = \infty$, this inequality no longer holds. In this case, we want to apply some structure to reduce to the maximal inequality case. For example, we will learn the metric entropy method in the next chapter, where we will approximate the infinite set Θ by a finite set Θ_ε where $|\Theta_\varepsilon| < \infty$ and $\sup_{\theta\in\Theta_\varepsilon} X_\theta \to \sup_{\theta\in\Theta} X_\theta$ as $\varepsilon \to 0$.

In order to apply maximal inequality, we need to transform our (potentially infinite) function class into a class with finite cardinality.

First, note the following useful fact. We defined

$$\mathcal{F}(x_{1:n}) \equiv \{(f(x_1), \dots, f(x_n) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n,$$

which allows us to rewrite

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right| = \sup_{v\in\mathcal{F}(x_{1:n})} |\langle \varepsilon, v\rangle|/n.$$

Therefore

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f\left(X_i\right)\right|\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f\left(X_i\right)\right| \mid X_{1:n}\right]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\sup_{v\in\mathcal{F}(X_{1:n})}\left|\frac{1}{n}\langle\varepsilon, v\rangle\right| \mid X_{1:n}\right]\right] \\
&\le \sup_{x_{1:n}\in\operatorname{supp}\mathbb{P}} \mathbb{E}_\varepsilon\left[\sup_{v\in\mathcal{F}(x_{1:n})} |\langle\varepsilon, v\rangle/n|\right].
\end{aligned}$$

In the last step we gain tractability because we reduced the problem from involving a potentially infinite class $\mathcal{F}$ to a possibly finite one, i.e. $\mathcal{F}(x_{1:n})$, but we lose information about the form of $\mathbb{P}$. However, we will show that in many cases the loss of information is not so large.

Example 25. *Sometimes $|\mathcal{F}| = \infty$ but $|\mathcal{F}(x_{1:n})| < \infty$. For example, consider the function class $\mathcal{F} = \{\mathbb{1}\{\cdot \le t\} : t \in \mathbb{R}\}$. It's clear that $|\mathcal{F}| = \infty$. Without loss of generality, we assume $x_1 < x_2 < \cdots < x_n$. Then we directly get*

$$\begin{aligned}\mathcal{F}(x_{1:n}) &= \{(\mathbb{1}\{x_1 \le t\}, \mathbb{1}(x_2 \le t), \dots \mathbb{1}\{x_n \le t\}) : t \in \mathbb{R}\} \\ &= \{(0,0,\dots 0), (1,0,\dots 0), (1,1,0,\dots 0), \dots (1,1,\dots,1)\}.\end{aligned}$$

This shows $|\mathcal{F}(x_{1:n})| = n+1, \forall\, x_{1:n} \in \mathbb{R}^n$. ♣

Lemma 5. *Consider a class of functions $\mathcal{F}$ such that $|\mathcal{F}(x_{1:n})| < \infty$, then*

$$\mathcal{R}_n(\mathcal{F}) \le \sup_{x_{1:n}} D_{\mathcal{F}}(x_{1:n}) \cdot \sqrt{\frac{2\log\left(\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})|\right)}{n}}$$

Proof. We showed that

$$\mathcal{R}_n(\mathcal{F}) \le \sup_{x_{1:n} \in \operatorname{supp}\mathbb{P}} \mathbb{E}_\varepsilon \left[\sup_{v \in \mathcal{F}(x_{1:n})} |\langle \varepsilon, v\rangle / n| \right].$$

Let's focus on the inner expectation. It's known that $\varepsilon_i \sim \mathsf{sG}(1)$. By independence of ε_is, for any $v \in \mathbb{R}^n$, we directly have $\mathbb{E}[\langle \varepsilon, v\rangle] = 0$ and $\frac{1}{n}\sum_{i=1}^n v_i \varepsilon_i \sim \mathsf{sG}(\|v\|_2/n)$. To apply the maximal inequality we want the sub-Gaussian parameter to be independent from v, thus we set

$$\sigma_n = \sup_{v \in \mathcal{F}(x_{1:n})} \frac{1}{n}\|v\|_2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sqrt{\sum_{i=1}^n f(x_i)^2}.$$

In this case we are taking the maximum over M $\mathsf{sG}(\sigma_n)$ random variables with mean 0. Hence, applying maximal inequality, we have

$$\mathbb{E}_\varepsilon \left[\sup_{v \in \mathcal{F}(x_{1:n})} \left| \frac{1}{n}\langle \varepsilon, v\rangle \right| \right] \le \sigma_n \sqrt{2\log(|\mathcal{F}(x_{1:n})|)} = \underbrace{\sqrt{n}\sigma_n}_{D_{\mathcal{F}}(x_{1:n})} \sqrt{\frac{2\log(|\mathcal{F}(x_{1:n})|)}{n}},$$

where $D_{\mathcal{F}}(x_{1:n}) = \sup_{f \in \mathcal{F}} \left(\frac{1}{n}\left|\sum_{i=1}^n f(x_i)^2\right|\right)^{1/2}$ is termed the ℓ_2- radius of the set $\mathcal{F}(x_{1:n})/\sqrt{n}$. Taking the sup on both sides of the inequality gives the desired result. ■

Example 26. *Consider the function class $\mathcal{F} = \{\mathbb{1}\{x \le t\} : t \in \mathbb{R}\}$. We already established that in this case $\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})| \le n+1$. Besides, $\sup_{x_{1:n}} D_{\mathcal{F}}(x_{1:n}) = \sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n 1^2}{n}} = 1$. Therefore,*

$$\mathcal{R}_n(\mathcal{F}) \le \sqrt{\frac{2\log(n+1)}{n}}.$$

By Proposition 29 and Hoeffding's inequality for bounded random variables we also get that with probability $1-\delta$

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_t \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \le t) - F(t) \right| \lesssim 2\sqrt{\frac{2\log(n+1)}{n}} + \sqrt{\frac{\log(2/\delta)}{n}}.$$

This example gives a proof of the Glivenko-Cantelli Theorem. ♣

3.2 Bounds of $\mathcal{R}_n(\mathcal{F})$ via Polynomial Discrimination

When we look at the upper bound for $\mathcal{R}_n(\mathcal{F})$

$$\Delta := \sup_{x_{1:n}} D_{\mathcal{F}}(x_{1:n}) \sqrt{\frac{1}{n} 2 \log\left(\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})|\right)}, \quad D_{\mathcal{F}}(x_{1:n}) = \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \left| \sum_{i=1}^{n} f(x_i)^2 \right|\right)^{1/2},$$

we notice that $D_{\mathcal{F}}(x_{1:n})$ is typically bounded, e.g. it is $O(1)$ when f are bounded. We then want to have $\sqrt{\frac{2\log(|\mathcal{F}(x_{1:n})|)}{n}} \to 0$ as $n \to \infty$ to ensure that $\mathcal{R}_n(\mathcal{F}) = o(1)$. We will accomplish this by controlling the structure of $\mathcal{F}$ so that we can manage the cardinality of $\mathcal{F}(x_{1:n})$.

In what follows we restrict our attention to $\mathcal{F} \subseteq \{f : x \to \{\pm 1\}\}$, then $|\mathcal{F}(x_{1:n})| \leq 2^n$. This is because $\mathcal{F} \subseteq \{\pm 1\}^n$ (note that we distinguish elements in $\mathcal{F}$ by their images!). This family of functions is useful in ML when doing classification.

In this case, we have two frequent behaviors of $|\mathcal{F}(x_{1:n})|$.

a) If $|\mathcal{F}(x_{1:n})| \lesssim (n+1)^d \lesssim O\left(n^d\right)$

$$\mathcal{R}_n(\mathcal{F}) \lesssim \Delta = O\left(\sqrt{\frac{d \log n}{n}}\right). \qquad \text{(parametric rate)}$$

b) If $|\mathcal{F}(x_{1:n})| \lesssim O(c^n)$

$$\mathcal{R}_n(\mathcal{F}) \lesssim \Delta = O\left(\sqrt{\frac{n \log c}{n}}\right) = O(\sqrt{\log c}). \qquad \text{(exponential rate)}$$

This latter bound is not great. This is because $\mathcal{R}_n(\mathcal{F})$ has $\|f\|_\infty$ as a trivial bound. Indeed if we are dealing with functions such that $\|f\|_\infty = M < \infty$, then $\mathcal{R}(\mathcal{F}) \leq M < \sqrt{\log c}$ for some c and M.

This gives us an intuition that we need to restrict $|\mathcal{F}(x_{1:n})|$ to have at most polynomial growth in n and avoid it being exponential in n

Definition 16 (Polynomial discrimination)**.** *A class $\mathcal{F}$ of binary-valued functions with domain $\mathcal{X}$ has polynomial discrimination of order $\nu \geq 1$, i.e. $\mathcal{F} \in PD(\nu)$, if*

$$\forall\, n \in \mathbb{N}, \quad \forall\, x_{1:n} \in \mathcal{X}^n, \quad |\mathcal{F}(x_{1:n})| \leq (n+1)^\nu.$$

Lemma 6. *Suppose that $\mathcal{F} \in PD(\nu)$. Then for all positive integers n and any collection of points $x_{1:n} = (x_1, \ldots, x_n)$,*

$$\mathcal{R}_n(\mathcal{F}) \leq \sup_{x_{1:n}} D_{\mathcal{F}}(x_{1:n}) \sqrt{\frac{2\nu \log(n+1)}{n}},$$

where $D(x_{1:n}) := \sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$ is the ℓ_2-radius of the set $\mathcal{F}(x_{1:n})/\sqrt{n}$.

Proof. Direct application of Lemma 5. ■

Note. *Any bounded function class with polynomial discrimination is Glivenko–Cantelli. This follows from the fact that* $\|f\|_\infty \le b$ *implies* $\sup_{x_{1:n}} D_{\mathcal{F}}(x_{1:n}) \le b$. *Then by Theorem 2 we conclude that* $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \le o(1)$ *with probability* $1 - o(1)$. ♦

Example 27. *Some PD and non-PD classes:*

1) If $\mathcal{F} = \{\langle \psi(x), \theta\rangle + b : \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$, *then* $|\mathcal{F}(x_{1:n})| = +\infty$.

2) If $\mathcal{F} = \{\mathbb{1}\{\langle \psi(x), \theta\rangle \ge b\} : \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$, *then* $\mathcal{F}$ *is* $PD(d+1)$. ♣

3.3 Bounds of $\mathcal{R}_n(\mathcal{F})$ via the Vapnik–Chervonenkis dimension

In certain cases, we can verify by direct calculation that a given function class has polynomial discrimination. More broadly, it is of interest to develop techniques for certifying this property in a less laborious manner. The theory of Vapnik–Chervonenkis (VC) dimension provides one such class of techniques.

In particular, we will show that if a function class $\mathcal{F}$ has "VC dimension ν", then it is $PD(\nu)$, and by Lemma 6 we know $\mathcal{R}_n(\mathcal{F}) \lesssim \sqrt{\frac{\nu \log(n+1)}{n}}$.

Definition 17 (Shattering). *Given a function class* $\mathcal{F} \subseteq \{f : \mathcal{X} \to \{0,1\}\}$, *we say* $x_{1:n}$ *is shattered by* $\mathcal{F}$ *if* $|\mathcal{F}(x_{1:n})| = 2^n$.

Definition 18 (VC Dimension). *The VC dimension of* $\mathcal{F}$, *denoted* $\nu(\mathcal{F})$ *is the largest positive integer* n *such that there exist some collection of points* $x_{1:n}$ *that is shattered by* $\mathcal{F}$.

Example 28. *We now see three examples:*

1. $\mathcal{F} = \{\mathbb{1}(\cdot \le t) : t \in \mathbb{R}\}$. *We claim that* $\nu(\mathcal{F}) = 1$. *To see this*

   - *Let* $n = 1$. *Then* $\mathcal{F}(x_1) = \{0,1\}$, *so* $|\mathcal{F}(x_1)| = 2 = 2^1$.
   - *Let* $n = 2$. *Then* $\mathcal{F}(x_{1:2}) = \{(0,0),(1,0),(1,1)\}$, *so* $|\mathcal{F}(x_{1:2})| \le 3 < 2^2$. *This is because it is not possible to get* $(0,1)$ *and* $(1,0)$ *for the same collection* $x_{1:2}$ *given the shape of the functions in* $\mathcal{F}$.

   Therefore $\nu(\mathcal{F}) = 1$, *which is the largest integer for which we can find a collection of points such that* $|\mathcal{F}(x_{1:n})| = 2^n$.

2. $\mathcal{F} = \{\mathbb{1}(s \le \cdot \le t) : s < t \in \mathbb{R}\}$. *We claim that* $\nu(\mathcal{F}) = 2$. *To see this*

   - *Let* $n = 1$. *Then* $\mathcal{F}(x_1) = \{0,1\}$, *so* $|\mathcal{F}(x_1)| = 2 = 2^1$.
   - *Let* $n = 2$. *Then* $\mathcal{F}(x_{1:2}) = \{(0,0),(1,0),(0,1),(1,1)\}$, *so* $|\mathcal{F}(x_{1:2})| \le 4 = 2^2$.
   - *Let* $n = 3$. *Then at most* $\mathcal{F}(x_{1:3}) = 7 < 2^3$.
     This is because it is not possible to get $(1,1,0),(1,0,1)$, *and* $(0,1,1)$ *for the same collection* $x_{1:3}$ *given the shape of the functions in* $\mathcal{F}$.

3. *Example 4.21 in Wainwright (2019). Suppose $\phi_1, \cdots, \phi_p : \mathcal{X} \to \mathbb{R}$ are some feature functions and let*

$$\mathcal{F} = \left\{ \mathbb{1}\left(\sum_{i=1}^{p} a_i \phi_i(x) \leq b \right) : a_i, b \in \mathbb{R} \right\}.$$

Then $v(\mathcal{F}) \leq p + 1$.

♣

Note. *In general, if a function class $\mathcal{F}$ can be linearly parametrized with k parameters, then $\nu(\mathcal{F}) = k$. This is because the parameters directly control the VC dimension. If the function class is not linear, we typically don't use the VC dimension to upper bound the Rademacher complexity.* ♦

In the above definition, we naturally have, $\forall n > v(\mathcal{F})$,

$$\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})| \leq 2^n - 1.$$

However, we actually have a much better upper bound when $n > v(\mathcal{F})$ given as follows:

Proposition 31 (Vapnik and Chervonenkis (1971), Sauer (1972), Shelah (1972))**.** *Let $\mathcal{F}$ be a function class with VC dimension ν. If $n > \nu$, then*

$$\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})| \leq \sum_{i=1}^{\nu} \binom{n}{i} \leq \min\left\{ (n+1)^{\nu}, \left(\frac{en}{\nu}\right)^{\nu} \right\}.$$

Note (*PD* and *VC*)**.** *Note that if $\mathcal{F} \in VC(\nu)$, then by Proposition 31 we immediately have that $\mathcal{F} \in PD(\nu)$. Thus we can apply Lemma 6 and get*

$$\mathcal{R}_n(\mathcal{F}) \lesssim \sqrt{\frac{2\nu \log(n+1)}{n}}.$$

♦

Example 29. *Let $X_{i\in[n]} \overset{iid}{\sim} \mathbb{P}$. Consider the class of functions*

$$\mathcal{F} = \left\{ \mathbb{1}\left(\sum_{i=1}^{p} a_i \phi_i(x) \leq b \right) : a_i, b \in \mathbb{R} \right\}.$$

Then with probability $1 - \delta$

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \leq C\sqrt{\frac{(p+1)\log(n+1)}{n}} + k\sqrt{\frac{\log(2/\delta)}{n}},$$

where the bound on the expectation of the empirical process comes from Proposition 6 and the concentration part comes from Hoeffding's inequality for bounded random variables. This bound can be sharpened by removing the $\log(n+1)$. ♣

3.4 Bounds of $\mathcal{R}_n(\mathcal{F})$ via Metric Entropy

This is yet another method to upper bound $\mathcal{R}_n(\mathcal{F})$ Given $X_\theta \sim \mathsf{sG}(\sigma), \forall\, \theta \in \mathcal{T}$, we hope to give an upper bound of $\mathbb{E}[\sup_{\theta\in\mathcal{T}} X_\theta]$ even when $|\mathcal{T}| = \infty$. The idea is to construct a set $\mathcal{T}_\varepsilon$ that approximates $\mathcal{T}$ and it has finite cardinality, i.e. $|\mathcal{T}_\varepsilon| < \infty$. Formally,

$$\forall\, \theta \in \mathcal{T}, \exists\, \theta' \in \mathcal{T}_\varepsilon \quad \text{s.t.} \quad \rho(\theta, \theta') \le \varepsilon.$$

For any $\mathcal{T}_\varepsilon \subseteq \mathcal{T}$, we have

$$\mathbb{E}[\sup_{\theta\in\mathcal{T}} X_\theta] \le \underbrace{\mathbb{E}[\sup_{\theta'\in\mathcal{T}_\varepsilon} X_{\theta'}]}_{\sigma\sqrt{\log|\mathcal{T}_\varepsilon|}} + \underbrace{\mathbb{E}[\sup_{\substack{\theta\in\mathcal{T},\\ \theta'\in\mathcal{T}_\varepsilon,\\ \rho(\theta,\theta')\le\varepsilon}} (X_\theta - X_{\theta'})]}_{L(\varepsilon)},$$

so there is a trade-off in ε. If ε is smaller, $\mathcal{T}_\varepsilon$ gets finer and it cardinality increases, but the other terms gets smaller as $\rho(\theta, \theta')$ shrinks with ε. Then the question is, given $\mathcal{T}$ and a metric ρ on $\mathcal{T}$, how to find $\mathcal{T}_\varepsilon$ and bound $|\mathcal{T}_\varepsilon|$? This leads to the next subsection.

Definition 19 (Distance). *We say $(\mathcal{T}, \rho)$ is a metric space, if $\rho : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ satisfies all the following conditions:*

- *(Non-negativity) $\rho(\theta, \theta') \ge 0\ \forall\, \theta, \theta' \in \mathcal{T}$, with $\rho(\theta, \theta') = 0 \iff \theta = \theta'$.*
- *(Symmetry) $\rho(\theta, \theta') = \rho(\theta, \theta), \forall\, \theta, \theta' \in \mathcal{T}$.*
- *(Triangle inequality) $\rho(\theta, \theta') \le \rho(\theta, \theta'') + \rho(\theta', \theta''), \forall\, \theta, \theta', \theta'' \in \mathcal{T}$.*

Example 30. *A metric space can have elements that are finite-dimensional, e.g. $\mathcal{T} = \mathbb{R}^d$. If $\mathbb{R}^d$ is equipped with either the Euclidean norm $\rho(\theta, \theta') = \|\theta - \theta'\|_2$ or the Hamming norm $\rho(\theta, \theta') = d^{-1}\sum_{i=1}^d \mathbb{1}(\theta_i \ne \theta_i')$, then it is a metric space.*

However, a metric space can also be a space of functions, for example the space of square-integrable functions with respect to the measure $\mu \in \mathcal{P}([0, 1])$, typically defined as

$$\mathcal{T} = L^2(\mu, [0, 1]) = \{f : [0, 1] \to \mathbb{R}, \int f^2(x)\mu(\,\mathrm{d}x) < \infty\}.$$

It is a metric space when it is equipped with the $L^2(\mu)$ norm

$$\rho(f, g) = \|f - g\|_{L^2(\mu)} = \left(\int (f(x) - g(x))^2 \mu(\,\mathrm{d}x)\right)^{\frac{1}{2}}$$

or the $L^\infty-$norm

$$\rho(f, g) = \|f - g\|_\infty = \operatorname{ess\,sup}_{x\in[0,1]} |f(x) - g(x)|$$

or the $L^2(\mathbb{P}_n)$ norm

$$\rho(f, g) = \|f - g\|_{L^2(\mathbb{P}_n)} = \left(\frac{1}{n}\sum_{i=1}^n (f(x) - g(x))^2\right)^{1/2}.$$

♣

Definition 20 (Covering set). *The set $\mathcal{T}_\varepsilon = \{\theta^1, \cdots, \theta^N\} \subseteq \mathcal{T}$ is called a ε-cover of the set $\mathcal{T}$ with metric ρ, if $\forall\, \theta \in \mathcal{T}$, $\exists\, \theta' \in \mathcal{T}_\varepsilon$, s.t. $\rho(\theta, \theta') \leq \varepsilon$. The ε-covering number of $(\mathcal{T}, \rho)$ is defined as:*

$$N(\varepsilon; \mathcal{T}, \rho) = \inf\{n \in \mathbb{N} : |\mathcal{T}_\varepsilon| = n, \mathcal{T}_\varepsilon \text{ is } \varepsilon\text{-covering of } \mathcal{T}\}.$$

Finally $\log N(\varepsilon; \mathcal{T}, \rho)$ is the metric entropy of $\mathcal{T}$ with metric ρ of size ε.

Note. *The covering number is the cardinality of the smallest set of points such that the union of the balls of radius ε centered on such points cover $\mathcal{T}$. Note that by definition of $\varepsilon-$cover we get*

$$\mathcal{T} \subseteq \bigcup_{\theta \in \mathcal{T}_\varepsilon} \mathcal{B}(\theta, \varepsilon).$$

We can see that the metric entropy of a set increases as ε decreases or as the size increases. ♦

Example 31 (Parametric vs Non-parametric). *For parametric families*

$$\log N(\varepsilon) \asymp d \log(1 + 1/\varepsilon),$$

whilst for non-parametric families

$$\log N(\varepsilon) \asymp \frac{c}{\varepsilon^\alpha}, \quad \alpha > 0.$$

In the first case as ε gets smaller the entropy increases with logarithmic rate, whereas in the non-parametric case it blows up quickly. ♣

Example 32. *Consider the set $\mathcal{T} = [-1, 1]$ and the metric $\rho(\theta, \theta') = |\theta - \theta'|$, then $N(\varepsilon; \mathcal{T}, \rho) \leq \frac{1}{\varepsilon} + 1$. For $\mathcal{T} = [-1, 1]^d$ and $\rho(\theta, \theta') = \|\theta - \theta'\|_\infty$, we have $N(\varepsilon; T, \rho) \leq (\frac{1}{\varepsilon} + 1)^d$. The intuition for the latter is how to cover a square with a grid of smaller squares of side 2ε.* ♣

It is not always the case we can use a constructive approach. If the set lives in an Eucledian space there is a general approach to upper and lower bound the covering number.

Definition 21 (ε-packing). *A set $\widetilde{\mathcal{T}}_\varepsilon = \{\theta^1, \cdots, \theta^M\} \subseteq \mathcal{T}$ is an ε-packing of $\mathcal{T}$ if $\forall\ \theta, \theta' \in \widetilde{\mathcal{T}}_\varepsilon, \theta \neq \theta'$, $\rho(\theta, \theta') > \varepsilon$. The ε-packing number of $\mathcal{T}$ is defined as*

$$M(\varepsilon; \mathcal{T}, \rho) = \sup\{M : |\widetilde{\mathcal{T}}_\varepsilon| =, \widetilde{\mathcal{T}}_\varepsilon \text{ is } \varepsilon\text{-packing of } \mathcal{T}\}.$$

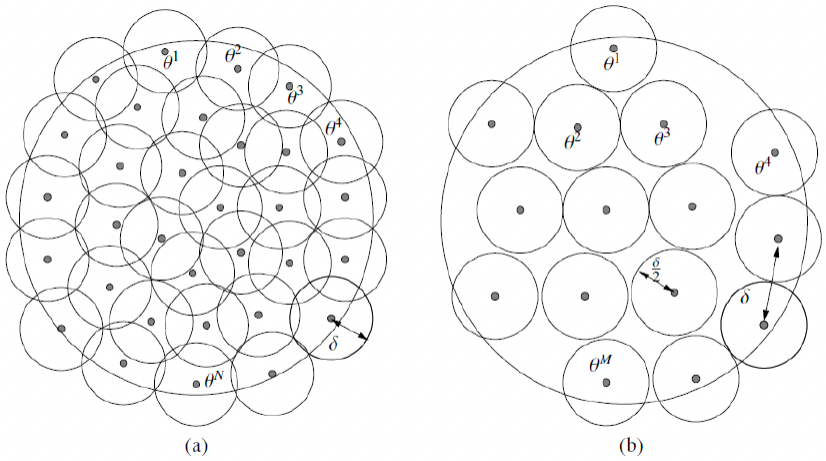
Note. *Note that $\forall\, \theta \neq \theta'$, $\mathcal{B}_\rho(\theta, \frac{\varepsilon}{2}) \cap \mathcal{B}_\rho(\theta', \frac{\varepsilon}{2}) = \emptyset$. So the points in $\widetilde{\mathcal{T}}_\varepsilon$ are such that* ♦

Lemma 7. *Let $\mathcal{T} \subseteq \mathbb{R}^n$. For any $\varepsilon > 0$, we have*

$$M(2\varepsilon; \mathcal{T}, \rho) \overset{\text{(i)}}{\leq} N(\varepsilon; \mathcal{T}, \rho) \overset{\text{(ii)}}{\leq} M(\varepsilon; \mathcal{T}, \rho).$$

Proof. (ii) a maximal ε-packing gives an ε-covering. To see this suppose that $\widetilde{\mathcal{T}}_\varepsilon$ is a maximal $\varepsilon-$packing of $\mathcal{T}$. Being maximal means that no other point can be added to $\widetilde{\mathcal{T}}_\varepsilon$ without violating the definition of $\varepsilon-$packing. Then, for each point θ in $\mathcal{T} \setminus \widetilde{\mathcal{T}}_\varepsilon$ there exists a point θ' in $\widetilde{\mathcal{T}}_\varepsilon$ such that $\rho(\theta, \theta') < \varepsilon$. Hence $\widetilde{\mathcal{T}}_\varepsilon$ is also a $\varepsilon-$cover.

Figure 1: *(a): δ-covering; (b): δ-packing.*



(i) Suppose $\mathcal{T}$ admits a 2ε-packing with size M, all ε-covering should have size at least M. This is because a 2ε packing does not cover $\mathcal{T}$ and all its points have non-intersecting neighborhoods of radius ε, therefore all the points in $\mathcal{T}$ are distant at least 2ε. As such to construct an $\varepsilon-$cover of $\mathcal{T}$ we need at least M points (equality holds on the real line for example). ■

Now we give the lower and upper bound of the covering and packing number.

Lemma 8. *Let $\mathcal{T} \subseteq \mathbb{R}^n$. For any $\varepsilon > 0$, we have*

$$\frac{\mathrm{Vol}(\mathcal{T})}{\mathrm{Vol}(\mathcal{B}_\rho(\varepsilon))} \overset{\text{(i)}}{\leq} N(\varepsilon; \mathcal{T}, \rho) \leq M(\varepsilon; \mathcal{T}, \rho) \overset{\text{(ii)}}{\leq} \frac{\mathrm{Vol}(\mathcal{T} + \mathcal{B}_\rho(\varepsilon/2))}{\mathrm{Vol}(\mathcal{B}_\rho(\varepsilon/2))},$$

where $\mathcal{T} + \mathcal{B}_\rho(\varepsilon/2) = \{a + b : a \in \mathcal{T}, b \in \mathcal{B}_\rho(\varepsilon/2)\}$.

Proof. The second inequality is directly obtained by Lemma 7. Now we prove the first and the last inequality.

(i) Since $\mathcal{T} = \bigcup_{\theta \in \mathcal{T}_\varepsilon} \mathcal{B}_\rho(\theta, \varepsilon)$, then

$$\mathrm{Vol}(\mathcal{T}) \leq \mathrm{Vol}\left(\bigcup_{\theta \in \mathcal{T}_\varepsilon} \mathcal{B}_\rho(\theta, \varepsilon)\right) \leq \sum_{\theta \in \mathcal{T}_\varepsilon} \mathcal{B}_\rho(\theta, \varepsilon) = |\mathcal{T}_\varepsilon| \cdot \mathrm{Vol}(\mathcal{B}_\rho(\varepsilon)),$$

where the last equality uses the fact that ρ is translation invariant.

(ii) Since $\bigcup_{\theta \in \widetilde{\mathcal{T}}_\varepsilon} \mathcal{B}_\rho(\theta, \varepsilon/2) \subseteq \mathcal{T} + \mathcal{B}_\rho(\varepsilon/2)$, we have

$$\sum_{\theta \in \widetilde{\mathcal{T}}_\varepsilon} \mathcal{B}_\rho(\theta, \varepsilon/2) = \mathrm{Vol}\left(\bigcup_{\theta \in \widetilde{\mathcal{T}}_\varepsilon} \mathcal{B}_\rho(\theta, \varepsilon/2)\right) \leq \mathrm{Vol}\left(\mathcal{T} + \mathcal{B}_\rho(\varepsilon/2)\right).$$

The fact that

$$|\sum_{\theta \in \widetilde{\mathcal{T}}_\varepsilon} \mathcal{B}_\rho(\theta, \varepsilon/2) = \widetilde{\mathcal{T}}_\varepsilon \cdot \mathrm{Vol}(\mathcal{B}_\rho(\varepsilon/2))$$

concludes the proof.

■

Example 33. *In this example, let $\rho = \|\cdot\|_p$, and $\mathcal{T} = \mathcal{B}_p(1) = \{x \in \mathbb{R}^d : \|x\|_p \le 1\}$. Then*

$$\mathrm{Vol}(\mathcal{T}) = c_{d,p}, \quad \mathrm{Vol}(\mathcal{B}_p(\varepsilon)) = c_{d,p}\varepsilon^d, \quad \mathrm{Vol}(\mathcal{T} + \mathcal{B}_p(\varepsilon/2)) = \mathrm{Vol}(\mathcal{B}_p(1+\varepsilon/2)) = c_{d,p}(1+\varepsilon/2)^d,$$

where $c_{d,p}$ is a constant that only depends on d and p. By Lemma 8, we have

$$N(\varepsilon; \mathcal{T}, \rho) \le \frac{\mathrm{Vol}(\mathcal{T} + \mathcal{B}_p(\varepsilon/2))}{\mathrm{Vol}(\mathcal{B}_p(\varepsilon/2))} = \frac{(1+\varepsilon/2)^d}{(\varepsilon/2)^d} = (2/\varepsilon + 1)^d,$$
$$N(\varepsilon; \mathcal{T}, \rho) \ge \frac{\mathrm{Vol}(\mathcal{B}_p(1))}{\mathrm{Vol}(\mathcal{B}_p(\varepsilon))} = \frac{1^d}{\varepsilon^d} = (1/\varepsilon)^d.$$

Therefore,

$$d \log(1/\varepsilon) \le \log N(\varepsilon; \mathcal{T}, \rho) \le d \log(2/\varepsilon + 1).$$

♣

Proposition 32 (One-step Discretization Bound)**.** *Let $(\mathcal{T}, \rho)$ be a metric space, $\mathcal{T}_\varepsilon$ be a $\varepsilon-$cover of $\mathcal{T}$, and let $\{X_\theta, \theta \in \mathcal{T}\}, X_\theta \overset{iid}{\sim} \mathsf{sG}(\sigma), \mathbb{E}[X_\theta] = 0$. Then for all $\varepsilon \le \mathrm{diam}(\mathcal{T})$*

$$\mathbb{E}[\sup_{\theta \in \mathcal{T}} |X_\theta|] \le 2\sqrt{\sigma^2 \log N(\varepsilon; \mathcal{T}, \rho)} + \mathbb{E}[\sup_{\substack{\theta \in \mathcal{T}, \\ \theta' \in \mathcal{T}_\varepsilon, \\ \rho(\theta, \theta') \le \varepsilon}} |X_\theta - X_{\theta'}|].$$

Proof. Note that

$$\sup_{\theta \in \mathcal{T}} X_\theta \le \sup_{\theta \in \mathcal{T}_\varepsilon} X_\theta + \sup_{\substack{\theta \in \mathcal{T}, \\ \theta' \in \mathcal{T}_\varepsilon, \\ \rho(\theta, \theta') \le \varepsilon}} (X_\theta - X_{\theta'})$$

and then take expectations on both sides. The first term can now be upper bounded by the maximal inequality and this concludes the proof. ■

We now give the same statement but in the version offered in Wainwright (2019).

Definition 22 (SG process)**.** *$\{X_\theta\}_{\theta \in \mathcal{T}}$ is a sub-Gaussian process with ρ on $\mathcal{T}$ if for any $\theta, \theta' \in \mathcal{T}$, $X_\theta - X_{\theta'}$ is $SG(\rho(\theta, \theta'))$, i.e.,*

$$\mathbb{E}[e^{\lambda(X_\theta - X_{\theta'})}] \le e^{\lambda^2 \rho(\theta, \theta')^2/2}, \ \forall \lambda \in \mathbb{R}.$$

Example 34. *Let $\mathcal{T} \subseteq \mathbb{R}^d$, $\rho = \|\cdot\|_2$, and define $X_\theta = \langle w, \theta \rangle$, where $w \sim N_d(\mathbf{0}, I_d)$. Then $X_\theta - X_{\theta'} = \langle \theta - \theta', w \rangle \sim N(0, \|\theta - \theta'\|_2^2) \sim \mathsf{sG}(\|\theta - \theta'\|_2)$.* ♣

Proposition 33 (One-step Discretization Bound II)**.** *Let $(\mathcal{T}, \rho)$ be a metric space, $\mathcal{T}_\varepsilon$ be a $\varepsilon-$cover of $\mathcal{T}$, and let $\{X_\theta, \theta \in \mathcal{T}\}$ be a mean-zero sub-Gaussian process. Then for all $\varepsilon \le \mathrm{diam}(\mathcal{T})$ such that $N(\varepsilon; \mathcal{T}, \rho) \ge 10$*

$$\mathbb{E}[\sup_{\theta, \widetilde{\theta} \in \mathcal{T}} (X_\theta - X_{\widetilde{\theta}})] \le 6\sqrt{\mathrm{diam}(\mathcal{T})^2 \log N(\varepsilon; \mathcal{T}, \rho)} + 2\,\mathbb{E}[\sup_{\substack{\theta, \widetilde{\theta} \in \mathcal{T}, \\ \rho(\theta, \theta') \le \varepsilon}} (X_\theta - X_{\theta'})].$$

Example 35 (Gaussian Complexity)**.** *In this example we show how to bound $\mathcal{G}(\mathcal{B}_2(1))$ by one-step discretization. Consider $w_i \overset{iid}{\sim} N(0,1)$, then $\langle w, \theta\rangle \sim SG(\|\theta\|_2)$. We can show that the Gaussian complexity of $\mathcal{B}_2(1)$ has order*

$$\mathcal{G}(\mathcal{B}_2(1)) = \mathbb{E}\left[\sup_{\theta\in\mathcal{B}_2(1)} \langle w, \theta\rangle\right] = \mathbb{E}[\|w\|_2] \approx \sqrt{d}.$$

Now we will show the above result by the one-step discretization method. First, by 32, we can obtain that

$$\begin{aligned}
\mathcal{G}(\mathcal{B}_2(1)) &\le \sup_{\theta\in\mathcal{B}_2(1)} \|\theta\|_2 \cdot \sqrt{\log(N(\varepsilon; \mathcal{B}_2(1), \|\cdot\|_2))} + \mathbb{E}\left[\sup_{\|\theta-\theta'\|_2\le\varepsilon} |\langle w, \theta\rangle - \langle w, \theta'\rangle|\right] \\
&\le \sqrt{\log(N(\varepsilon; \mathcal{B}_2(1), \|\cdot\|_2))} + \mathbb{E}\left[\sup_{\|\theta-\theta'\|_2\le\varepsilon} |\langle w, \theta\rangle - \langle w, \theta'\rangle|\right] \\
&\le \sqrt{d\log(2/\varepsilon+1)} + \mathbb{E}\left[\sup_{\|\theta-\theta'\|_2\le\varepsilon} |\langle w, \theta\rangle - \langle w, \theta'\rangle|\right].
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}\left[\sup_{\|\theta-\theta'\|_2\le\varepsilon} |\langle w, \theta\rangle - \langle w, \theta'\rangle|\right] &= \mathbb{E}\left[\sup_{\|\theta-\theta'\|_2\le\varepsilon} |\langle w, \theta-\theta'\rangle|\right] \\
&= \mathbb{E}\left[\sup_{\|r\|_2\le\varepsilon} |\langle w, r\rangle|\right] = \varepsilon\,\mathbb{E}\left[\sup_{\|r\|_2\le 1} |\langle w, r\rangle|\right] = \varepsilon\mathcal{G}(\mathcal{B}_2(1)).
\end{aligned}$$

Therefore,

$$\mathcal{G}(\mathcal{B}_2(1)) \le \sqrt{d\log(2/\varepsilon+1)} + \varepsilon\mathcal{G}(\mathcal{B}_2(1)) \Longrightarrow \mathcal{G}(\mathcal{B}_2(1)) \le \frac{1}{1-\varepsilon}\sqrt{d\log(2/\varepsilon+1)}.$$

Since ε was arbitrary, we can choose it and fix it to $1/2$ to get

$$\mathcal{G}(\mathcal{B}_2(1)) \le 2\sqrt{d\log(5)} \asymp \sqrt{d}.$$

♣

Example 36 (Operator Norm)**.** *If $W_{ij} \overset{iid}{\sim} N(0,1)$ and $W \in \mathbb{R}^{n\times d}$, then*

$$\mathbb{E}[\|W\|_{\mathsf{op}}] \le \sqrt{n} + \sqrt{d}$$

♣

Example 37 (Lipschitz Functions, Wainwright 5.6)**.** *Consider $\mathcal{F}_L = \{g : [0,1] \to \mathbb{R} : g(0) = 0, |g(y) - g(x)| \le L|y - x|\}$, then*

$$\log N(\delta; \mathcal{F}_L, L^\infty) \asymp \frac{L}{\delta}.$$

Let's now compute the Rademacher complexity of the class $\mathcal{F}_L$. First, note that each $\varepsilon_i X_i$ is mean-zero and sub-Gaussian. The reason is that ε_i are mean-zero, thus by iterated expectations we get that $\varepsilon_i X_i$ is mean-zero too. Moreover, $\varepsilon_i X_i$ are absolutely bounded by L because f are L-Lipschitz on the interval $[0,1]$

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}_L) &= \mathbb{E}_{X_i,\varepsilon_i}\left[\sup_{f\in\mathcal{F}_L}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}_L}|X_f|\right] && (X_f \sim \mathsf{sG}(L/\sqrt{n})) \\
&\lesssim \frac{L}{\sqrt{n}}\sqrt{\log N(\delta;\mathcal{F}_L,L^\infty)} + \mathbb{E}\left[\sup_{\substack{g,f\in\mathcal{F}_L \\ \|f-g\|_\infty\le\delta}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(f\left(x_i\right)-g\left(x_i\right)\right)\right|\right] \\
& && \text{(One-step discretization)} \\
&\lesssim \frac{L}{\sqrt{n}}\sqrt{\frac{L}{\delta}} + \delta
\end{aligned}$$

Again, δ was arbitrary, thus we can optimize over it. For example, say $L=1$, then $\delta^\star = n^{-1/3}$. By rescaling everything by L (note the the Rademacher complexity is scalable), we get

$$\mathcal{R}_n(\mathcal{F}_L) \lesssim \frac{2L}{n^{1/3}}.$$

♣

3.5 Bounds of $\mathcal{R}_n(\mathcal{F})$ via Chaining

In this section, we introduce the chaining method to obtain a tighter control of discretization error in Proposition 32 and have thus a tighter bound of $\mathbb{E}[\sup_{\theta\in\mathcal{T}}|X_\theta|]$. To formally establish our result, we need to recall the definition of sub-Gaussian process.

Definition 23 (SG process). *$\{X_\theta\}_{\theta\in\mathcal{T}}$ is a sub-Gaussian process with ρ on $\mathcal{T}$ if for any $\theta,\theta'\in\mathcal{T}$, $X_\theta - X_{\theta'}$ is $SG(\rho(\theta,\theta'))$, i.e.,*

$$\mathbb{E}[e^{\lambda(X_\theta - X_{\theta'})}] \le e^{\lambda^2\rho(\theta,\theta')^2/2},\ \forall\lambda\in\mathbb{R}.$$

Note. *The requirement of $\{X_\theta\}_{\theta\in\mathcal{T}}$ being a sub-Gaussian process is neither stronger nor weaker than asking that each X_θ is sub-Gaussian. However, the former implies that each $X_\theta - X_{\theta'} \sim \mathsf{sG}\left(\rho(\theta,\theta')\right)$.* ♦

Our goal is to consider the following bound

$$\mathbb{E}\left[\sup_{\theta\in T}|X_\theta|\right] \le \inf_{\varepsilon}\left\{\mathbb{E}\left[\sup_{\theta\in T_\varepsilon}|X_\theta|\right] + \underbrace{\mathbb{E}\left[\sup_{\rho(\theta,\tilde\theta)\le\varepsilon}|X_\theta - X_{\tilde\theta}|\right]}_{\text{discretization error}}\right\}$$

and have a better control on the discretization error. Regarding the first one, we already know how to control it via the maximal inequality.

Proposition 34 (Chaining Upper Bound)**.** *Let $(\mathcal{T}, \rho)$ be a metric space and let $\{X_\theta, \theta \in \mathcal{T}\}$ be a mean-zero sub-Gaussian process $\mathsf{sG}(\rho(\theta, \theta'))$ and let the diameter of $\mathcal{T}$ be defined as $D := \sup_{\theta,\widetilde{\theta}\in\mathcal{T}} \rho(\theta, \widetilde{\theta})$. Then, $\forall\, \varepsilon \in [0, D]$*

$$\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} (X_\theta - X_{\widetilde{\theta}})\right] \leq \inf_{\varepsilon\leq D}\left[2\,\mathbb{E}\left[\sup_{\rho(r,r')\leq\varepsilon} (X_r - X_{r'})\right] + 32 J(\varepsilon; D; \mathcal{T}, \rho)\right],$$

where $J(\varepsilon; D; \mathcal{T}, \rho) = \int_\varepsilon^D \sqrt{\log N(u; \mathcal{T}, \rho)}du$ is defined as Dudley's entropy integral.

Note. *Typically the term $32J(\varepsilon; D; \mathcal{T}, \rho)$ is much smaller than $D\sqrt{\log N(\varepsilon; \mathcal{T}, \rho)}$ that we get from the one-step discretization bound.*

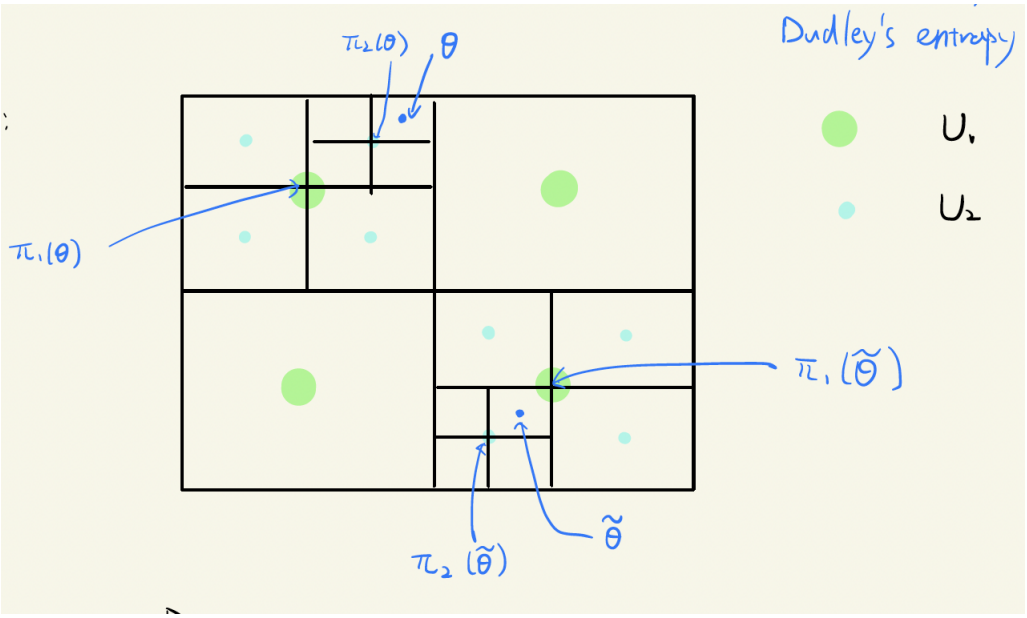
If you are interested in bounding $\mathbb{E}[\sup_\theta X_\theta]$ you can still use this bound because

$$\mathbb{E}[\sup_{\theta\in\mathcal{T}} X_\theta] = \sup_{\theta'\in\mathcal{T}} \mathbb{E}[\sup_{\theta\in\mathcal{T}}(X_\theta - \mathbb{E}[X_{\theta'}])] \leq \mathbb{E}[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} (X_\theta - X_{\widetilde{\theta}})],$$

where in the first equality we used the fact that X_θ are mean-zero, whilst in the equality we just swapped the integration and the sup *operator.*

A similar bound -with different constants- can be obtained for $\mathbb{E}[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} |X_\theta - X_{\widetilde{\theta}}|]$ because nothing would change in the proof. ♦

Proof. We start by constructing a sequence of finer and finer $\varepsilon-$covers of $\mathcal{T}$. Let $\varepsilon_m = \frac{D}{2^m}$ for $m = 1, 2, \ldots, L$. Let $\mathcal{U}_m$ be a minimal ε_m-covering of $\mathcal{T}$, then $|\mathcal{U}_m| = N(\varepsilon_m; \mathcal{T}, \rho)$. Also, define $\pi_m(\theta) = \arg\min_{\beta\in\mathcal{U}_m} \rho(\theta, \beta)$, i.e., $\pi_m(\theta)$ is the closest point in $\mathcal{U}_m$ to θ. The picture below shows graphically some different covers under the $\rho_\infty-$norm and $\pi_m(\theta)$.



The reason why we do this is because we want to use the $\{\pi_m(\theta), \pi_m(\widetilde{\theta})\}_{m=1}^{L}$ to construct an interpolating path that connects any two θ and $\widetilde{\theta}$.

By the triangle inequality,

$$|X_\theta - X_{\widetilde{\theta}}| \le |X_\theta - X_{\pi_L(\theta)}| + \sum_{i=1}^{L-1} |X_{\pi_{i+1}(\theta)} - X_{\pi_i(\theta)}| + |X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}| + \sum_{i=1}^{L-1} |X_{\pi_{i+1}(\widetilde{\theta})} - X_{\pi_i(\widetilde{\theta})}| + |X_{\pi_L(\widetilde{\theta})} - X_{\widetilde{\theta}}|.$$

Then taking the sup over $\mathcal{T}$ and expectation yields

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} \left|X_\theta - X_{\widetilde{\theta}}\right|\right] \le \mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} \left|X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}\right|\right] \\ + 2\sum_{l=1}^{L-1} \mathbb{E}\left[\sup_{\theta\in\mathcal{T}} \left|X_{\pi_{l+1}(\theta)} - X_{\pi_l(\theta)}\right|\right] \\ + 2\,\mathbb{E}\left[\sup_{\theta\in\mathcal{T}} \left|X_\theta - X_{\pi_L(\theta)}\right|\right].\end{aligned}$$

Let's start with the first term $\mathbb{E}[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} |X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}|]$. This might seem a term involving a supremum over a set of infinite cardinality. However, since $(\pi_1(\theta), \pi_1(\widetilde{\theta}))$ can take at most $|\mathcal{U}_1|^2 \le N(D/2;\mathcal{T},\rho)^2$ distinct values and each term is $\mathsf{sG}(D)$ because the distances $|X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}|$ are bounded, we can apply the maximal inequality. Therefore

$$\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} \left|X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}\right|\right] \lesssim D\sqrt{2\log N(D/2;\mathcal{T},\rho)}.$$

For the second terms $\mathbb{E}[\sup_{\theta\in\mathcal{T}} |X_{\pi_{l+1}(\theta)} - X_{\pi_l(\theta)}|]$, since $(\pi_{l+1}(\theta), \pi_l(\theta))$ can take at most $|\mathcal{U}_l||\mathcal{U}_{l+1}| \le N(D/2^{l+1};\mathcal{T},\rho)^2$ distinct values and each term is $\mathsf{sG}(D/2^l)$, then by the maximal inequality,

$$\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} |X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}|\right] \lesssim \frac{D}{2^{l-1}}\sqrt{\log N(D/2^l;\mathcal{T},\rho)}.$$

Finally, we can't use the same trick on the last term because the first term is not projected. However,

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{T}} |X_\theta - X_{\pi_L(\theta)}|\right] \le \mathbb{E}\left[\sup_{\rho(\theta,\widetilde{\theta})\le D/2^{L-1}} |X_\theta - X_{\widetilde{\theta}}|\right].$$

Therefore, the final bound is

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in\mathcal{T}} |X_\theta - X_{\widetilde{\theta}}|\right] &\le 2\,\mathbb{E}\left[\sup_{\rho(\theta,\widetilde{\theta})\le D/2^{L-1}} |X_\theta - X_{\widetilde{\theta}}|\right] + c\cdot\sum_{l=1}^{L}\frac{D}{2^l}\sqrt{\log N(D/2^{l-1};\mathcal{T},\rho)} \\ &\le 2\,\mathbb{E}\left[\sup_{\rho(\theta,\widetilde{\theta})\le D/2^{L}} |X_\theta - X_{\widetilde{\theta}}|\right] + c\int_{D/2^L}^{D}\sqrt{\log N(u;\mathcal{T},\rho)}du.\end{aligned}$$

This holds for any L and we can take $\varepsilon = \frac{D}{2^L}$. ■

Example 38. *Let's revisit some examples to see the gains through the chaining argument.*

$$\begin{aligned}\mathcal{G}(\mathcal{B}_2(1)) &= \mathbb{E}[\sup_{\theta\in\mathcal{B}_2(1)} \langle w,\theta\rangle] \\ &\le \inf_{\varepsilon}\left\{\varepsilon\mathcal{G}(\mathcal{B}_2(1)) + \int_{\varepsilon}^{1}\sqrt{\log N(u;\mathcal{B}_2(1),\|\cdot\|_2)}du\right\}\end{aligned}$$

$$\leq \inf_{\varepsilon} \left\{ \int_0^1 \sqrt{d \log(2/u+1)} du \right\}$$
$$= \sqrt{d} \int_0^1 \sqrt{\log(2/u+1)} du \asymp \sqrt{d}.$$

where the last equality arbitrarily fixes $\varepsilon = 0$. Note that the integrand is still integrable as the function goes to $-\infty$ as $u \to 0$ sufficiently slow.

The operator norm $\mathbb{E}[\|W\|_{\mathtt{op}}] \asymp \sqrt{n+d}$.

For the class of functions $\mathcal{F}_L = \{g : [0,1] \to \mathbb{R} : g(0) = 0, |g(y) - g(x)| \leq L|y-x|\}$ we get that $\mathcal{R}(\mathcal{F}_L) \asymp \frac{cL}{\sqrt{n}}$.

In the previous case we got orders of $\sqrt{d}$, $\sqrt{n} + \sqrt{d}$, and $\frac{2L}{n^{1/3}}$, respectively. ♣

Note. *From the previous example it is clear that if the index set of the supremum process is somehow parametric (first two examples), the chaining argument does not buy us much. However, if the class is non-parametric then it yields much sharper bounds when compared to the one-step discretization bound.* ♦

3.6 Applications of the Chaining Method

In this section, we study the Rademacher complexity of the function class $\mathcal{F} \subseteq L^p(\mathbb{P})$ for $1 \leq p \leq \infty$, where $L^p(\mathbb{P})$ is the L^p space with respect to some probability measure $\mathbb{P}$. Let $\varepsilon_i \overset{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})$ denote Rademacher random variables. Recall that the Rademacher complexity of $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\varepsilon_i, X_i} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$

When the value of f is binary (such as with indicator functions), $\mathcal{F}(X_{1:n})$ is a finite set (of cardinality at most 2^n), and we can apply the VC dimension theory or use the maximal inequality. For general function classes, we consider the metric entropy method. To apply Proposition 34 successfully, in general, we have to:

- Define the metric ρ on the function space $\mathcal{F}$ of interest.
- Show that X_f is a sub-Gaussian process with respect to metric ρ.
- Find an upper bound for the covering number $N(u; \mathcal{F}, \rho)$.
- (Optional) Find an upper bound for the the discretization error.

3.6.1 Useful Metrics on the Function Space.

Definition 24 ($L^2(\mathbb{P})$ metric). *For any $f, g \in \mathcal{F}$, define*

$$\|f - g\|_{L^2(\mathbb{P})}^2 = \int_{\mathcal{X}} (f(x) - g(x))^2 \mathbb{P}(\,\mathrm{d}x).$$

Definition 25 (L^∞ metric)**.** *For any $f, g \in \mathcal{F}$, define*

$$\|f - g\|_{L^\infty} = \sup_{x \in \mathcal{X}} |f(x) - g(x)|.$$

Definition 26 ($L^2(\mathbb{P}_n)$ metric)**.** *Suppose $\mathbb{P}_n$ is the empirical measure with respect to $X_{1:n}$. For any $f, g \in \mathcal{F}$, define*

$$\|f - g\|^2_{L^2(\mathbb{P}_n)} = \int_{\mathcal{X}} (f(x) - g(x))^2 \mathbb{P}_n(\,\mathrm{d}x) = \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2.$$

Note. *There are some remarks on the base measures in the above definitions. In Definition 24, $\mathbb{P}$ is the base measure, and in most cases, $\mathbb{P}$ refers to some probability measures. In Definition 25, in general we don't need to write down the base measure $\mathbb{P}$, and we can assume that the base measure is the Lebesgue measure. Definition 26 can be viewed as a special case of Definition 24, with the empirical measure $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ is the Dirac measure. In Definition 26, we consider fixed $X_{1:n}$, and if $X_{1:n}$ are random, $\|\cdot\|_{L^2(\mathbb{P}_n)}$ is a random variable or a random function of $\{X_{1:n}\}$.* ♦

Lemma 9. *The $L^2(\mathbb{P}_n)$ metric is equivalent to $\|\cdot\|_2$ on $n^{-1/2}\mathcal{F}(X_{1:n}) \subseteq \mathbb{R}^n$.*

Proof. We defined

$$n^{-1/2}\mathcal{F}(X_{1:n}) = \Big\{ \frac{1}{\sqrt{n}} (f(X_1), ..., f(X_n))' \in \mathbb{R}^n : f \in \mathcal{F} \Big\},$$

for any $f, g \in \mathcal{F}$, we have $n^{-1/2} f(X_{1:n}), n^{-1/2} g(X_{1:n}) \in n^{-1/2}\mathcal{F}(X_{1:n})$ and

$$\|n^{-1/2} f(X_{1:n}) - n^{-1/2} g(X_{1:n})\|_2^2 = \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2 = \|f - g\|^2_{L^2(\mathbb{P}_n)}.$$

■

Note. *Lemma 9 provides another point of view of the $L^2(\mathbb{P}_n)$ metric. Indeed, despite it being a norm in the space $\mathcal{F}$ of functions, it is related to the ℓ_2-norm in the Euclidean space.* ♦

If we are dealing with a parametric function space, that is a class of functions were elements are indexed by a parameter, then we can define another metric induced by the metric of its parameter space.

Definition 27 (Metric on the parametric function space)**.** *Let $\mathcal{F} = \{f_\theta : \theta \in \mathcal{T} \subseteq \mathbb{R}^d\}$ be a function space parameterized by θ and supported in a metric space $(\mathcal{T}, \rho)$. The induced metric $\widetilde{\rho}$ on $\mathcal{F}$ is*

$$\rho(f_\theta, f_{\tilde{\theta}}) = \widetilde{\rho}(\theta, \tilde{\theta}).$$

Next, we discuss the relationships of the four metrics in Definitions 24, 25, 26, 27.

Lemma 10. *For any base measure $\mathbb{P}$ and for any $f, g \in \mathcal{F}$,*

$$\|f - g\|_{L^2(\mathbb{P})} \le \|f - g\|_{L^\infty}.$$

Proof. By definition,

$$\|f-g\|_{L^2(\mathbb{P})}^2 = \int_{\mathcal{X}} (f(x)-g(x))^2 \mathbb{P}(dx) \leq \sup_{x\in\mathcal{X}} (f(x)-g(x))^2 = \|f-g\|_{L^\infty}^2,$$

which implies the desired result. ■

Corollary 3. *For any $X_{1:n}$,*

$$\|f-g\|_{L^2(\mathbb{P}_n)} \leq \|f-g\|_{L^\infty}.$$

Lemma 11. *For any parametric function space $\mathcal{F} = \{f_\theta : \theta \in \mathcal{T} \subseteq \mathbb{R}^d\}$ with induced metric ρ, suppose that there exists a function Γ such that for any $x \in \mathcal{X}$,*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \Gamma(x) \cdot \rho(\theta_1, \theta_2).$$

Then, we have

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^2(\mathbb{P})} \leq \|\Gamma\|_{L^2(\mathbb{P})} \cdot \rho(\theta_1, \theta_2),$$

and

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^\infty} \leq \|\Gamma\|_{L^\infty} \cdot \rho(\theta_1, \theta_2).$$

Proof. By definition,

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^2(\mathbb{P})}^2 = \int_{\mathcal{X}} (f_{\theta_1}(x) - f_{\theta_2}(x))^2 \mathbb{P}(dx) \leq \int_{\mathcal{X}} \Gamma(x)^2 \cdot \rho(\theta_1,\theta_2)^2 \mathbb{P}(dx) = \|\Gamma\|_{L^2(\mathbb{P})}^2 \cdot \rho(\theta_1,\theta_2)^2,$$

which implies the first inequality. The proof of the second inequality is similar:

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^\infty} = \sup_{x\in\mathcal{X}} |f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \sup_{x\in\mathcal{X}} |\Gamma(x)| \cdot \rho(\theta_1,\theta_2) = \|\Gamma\|_{L^\infty} \cdot \rho(\theta_1,\theta_2).$$

■

The next example shows how we can bound a norm in the function space with a norm in the Euclidean space.

Example 39. *Consider the function class*

$$\mathcal{F} = \left\{ f_\theta(x) = 1 - e^{-\theta x}, x \in [0,1] : \theta \in [0,1] \right\}.$$

For fixed $x \in [0,1]$, if we view $f_\theta(x)$ as a function of θ, then the first derivative is $xe^{-\theta x}$, which is upper bounded by x. This implies that

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq x \cdot |\theta_1 - \theta_2|.$$

By Lemma 11,

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^2(\mathbb{P})} \leq \|x\|_{L^2(\mathbb{P})} \cdot |\theta_1 - \theta_2|,$$

and

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^\infty} \leq |\theta_1 - \theta_2|.$$

♣

The relationship of the metrics lead to the relationship of covering numbers, according to Lemma 12.

Lemma 12. *If ρ_1, ρ_2 are two metrics on $\mathcal{T}$ such that $\rho_1(\theta_1, \theta_2) \leq \rho_2(\theta_1, \theta_2)$ for any $\theta_1, \theta_2 \in \mathcal{T}$, then we have*

$$N(\varepsilon; \mathcal{T}, \rho_1) \leq N(\varepsilon; \mathcal{T}, \rho_2).$$

Proof. Suppose $\mathcal{T}_\varepsilon$ is an $\varepsilon-$cover under ρ_2. Then it means that $\forall\, \theta \in \mathcal{T}, \exists\, \theta' \in \mathcal{T}_\varepsilon : \rho_2(\theta, \theta') \leq \varepsilon.$. Then, by assumption, we get that $\rho_1(\theta, \theta') \leq \rho_2(\theta, \theta') \leq \varepsilon$, which shows that $\mathcal{T}_\varepsilon$ is an $\varepsilon-$cover under ρ_1 too. ■

Corollary 4.

$$N(\varepsilon; \mathcal{F}, L^2(\mathbb{P}_n)) \leq N(\varepsilon; \mathcal{F}, L^\infty), \quad \textit{and} \quad N(\varepsilon; \mathcal{F}, L^2(\mathbb{P})) \leq N(\varepsilon; \mathcal{F}, L^\infty).$$

If $|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \Gamma(x) \cdot \rho(\theta_1, \theta_2)$, then

$$N(\varepsilon; \mathcal{F}, L^\infty) \leq N(\varepsilon; \mathcal{T}, \|\Gamma\|_{L^\infty} \cdot \rho), \quad \textit{and} \quad N(\varepsilon; \mathcal{F}, L^2(\mathbb{P})) \leq N(\varepsilon; \mathcal{T}, \|\Gamma\|_{L^2(\mathbb{P})} \cdot \rho).$$

Proof. (i) Consequence of Lemma 12 and Lemma 10. (ii) Consequence of Lemma 11 and Lemma 10. ■

3.6.2 Rademacher Complexity is a sub-Gaussian Process

We now check whether the Rademacher complexity is a sub-Gaussian process and under which metric. We start with the empirical Rademacher complexity, treating the $X_{1:n}$ as fixed.

Lemma 13. *The empirical Rademacher complexity $\mathcal{R}(\mathcal{F}(X_{1:n}/n)$ is a sub-Gaussian process with metric $L^2(\mathbb{P}_n)$ or L^∞.*

Proof. Define $X_f := n^{-1/2} \sum_{i=1}^n \varepsilon_i f(x_i)$. Pick any $f, g \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda(X_f - X_g)\right\} \mid X_{1:n}\right] &= \mathbb{E}\left[\exp\left\{n^{-1/2}\lambda \sum_{i=1}^n \varepsilon_i (f(X_i) - g(X_i))\right\} \mid X_{1:n}\right] \quad \text{(definition of } X_f \text{ and } X_g\text{)} \\
&= \prod_{i=1}^n \mathbb{E}\left[\exp\left\{n^{-1/2}\lambda \varepsilon_i (f(X_i) - g(X_i))\right\} \mid X_i\right] \qquad \text{(by independence of } \varepsilon_i\text{)} \\
&\leq \prod_{i=1}^n \exp\left\{\frac{\lambda^2}{2n}(f(X_i) - g(X_i))^2\right\} \qquad \text{(by Hoeffding's inequality)} \\
&= \exp\left\{\frac{\lambda}{2} \cdot \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2\right\} = \exp\left\{\frac{\lambda}{2} \|f - g\|^2_{L^2(\mathbb{P}_n)}\right\}.
\end{aligned}$$

We conclude that $(X_f)_{f \in \mathcal{F}}$ is a sub-Gaussian process with metric $\|\cdot\|_{L^2(\mathbb{P}_n)}$. Also, since

$$\|f - g\|^2_{L^2(\mathbb{P}_n)} \leq \|f - g\|^2_{L^\infty},$$

$(X_f)_{f \in \mathcal{F}}$ is a sub-Gaussian process with metric $\|\cdot\|_{L^\infty}$. ■

Proposition 35. *Let $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{X,\varepsilon}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]$. Then*

$$\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right] \lesssim \mathcal{R}_n(\mathcal{F}) \lesssim \frac{D_{\mathbb{P}}}{\sqrt{n}} \sup_Q \int_0^1 \sqrt{\log N(\|\mathcal{F}\|_Q \cdot u; \mathcal{F}, L^2(Q))}\, \mathrm{d}u.$$

and

$$\sup_{\mathbb{P}} \mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right] \lesssim \mathcal{R}_n(\mathcal{F}) \lesssim D_\infty \cdot \inf_{\varepsilon}\left[\varepsilon + \frac{1}{\sqrt{n}}\int_\varepsilon^1 \sqrt{\log N(\|\mathcal{F}\|_\infty \cdot u; \mathcal{F}, L^\infty)}\,\mathrm{d}u\right],$$

where $D_{\mathbb{P}} := \sup_{f\in\mathcal{F}} \|f\|_{L^2(\mathbb{P})}$ *and* $D_\infty := \sup_{f\in\mathcal{F}} \|f\|_\infty$.

Note. *There is nothing specific about the metric used in the two results above other than the fact that they make the empirical Rademacher complexity a sub-Gaussian process.*

In some cases it is not hard to upper bound the covering number of $\mathcal{F}$*, thus we resort to the version*

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{\|\mathcal{F}\|_{\mathbb{P}_n}}{\sqrt{n}}\int_0^1 \sqrt{\log N(\|\mathcal{F}\|_{\mathbb{P}_n} \cdot u; \mathcal{F}, L^2(\mathbb{P}_n))}\,\mathrm{d}u,$$

which will be derived as an intermediate step in the proofs below. ♦

Proof. By the previous Lemma, we know that the empirical Rademacher complexity is a sub-Gaussian process with metric $L^2(\mathbb{P}_n)$. Therefore by Proposition 34 with $\varepsilon = 0$, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |X_f|\right] \lesssim \int_0^{\|\mathcal{F}\|_{\mathbb{P}_n}} \sqrt{\log N(u; \mathcal{F}, L^2(\mathbb{P}_n))}\,\mathrm{d}u,$$

where $\|\mathcal{F}\|_{\mathbb{P}_n}$ is defined as $\|\mathcal{F}\|_{\mathbb{P}_n} := \sup_{f,g\in\mathcal{F}} \|f-g\|_{L^2(\mathbb{P}_n)}$. Therefore,

$$\mathcal{R}(\mathcal{F}(X_{1:n})/n) = \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}} |X_f|\right] \lesssim \frac{1}{\sqrt{n}}\int_0^{\|\mathcal{F}\|_{\mathbb{P}_n}} \sqrt{\log N(u; \mathcal{F}, L^2(\mathbb{P}_n))}\,\mathrm{d}u.$$

Since the covering number $N(u; \mathcal{F}, L^2(\mathbb{P}_n))$ is not easy to compute, we want to find an upper bound of it:

$$\begin{aligned}
\mathcal{R}(\mathcal{F}(X_{1:n})/n) &\lesssim \frac{1}{\sqrt{n}}\int_0^{\|\mathcal{F}\|_{\mathbb{P}_n}} \sqrt{\log N(u; \mathcal{F}, L^2(\mathbb{P}_n))}\,\mathrm{d}u \\
&\lesssim \frac{\|\mathcal{F}\|_{\mathbb{P}_n}}{\sqrt{n}}\int_0^1 \sqrt{\log N(\|\mathcal{F}\|_{\mathbb{P}_n} \cdot v; \mathcal{F}, L^2(\mathbb{P}_n))}\,\mathrm{d}v && (u = \|\mathcal{F}\|_{\mathbb{P}_n} \cdot v) \\
&\lesssim \frac{\|\mathcal{F}\|_{\mathbb{P}_n}}{\sqrt{n}}\sup_Q \int_0^1 \sqrt{\log N(\|\mathcal{F}\|_Q \cdot u; \mathcal{F}, L^2(Q))}\,\mathrm{d}u \\
&\lesssim \frac{\|\mathcal{F}\|_{\mathbb{P}}}{\sqrt{n}}\sup_Q \int_0^1 \sqrt{\log N(\|\mathcal{F}\|_Q \cdot u; \mathcal{F}, L^2(Q))}\,\mathrm{d}u && (\|\mathcal{F}\|_{L^2(\mathbb{P}_n)} \le \|\mathcal{F}\|_{L^2(\mathbb{P})})
\end{aligned}$$

where the supremum is over all the probability measures Q. By taking expectations on both sides

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}[\mathcal{R}(\mathcal{F}(X_{1:n})/\sqrt{n})] \lesssim \frac{\|\mathcal{F}\|_{\mathbb{P}}}{\sqrt{n}}\sup_Q \int_0^1 \sqrt{\log N(\|\mathcal{F}\|_Q \cdot u; \mathcal{F}, L^2(Q))}\,\mathrm{d}u.$$

To get exactly the first result it is sufficient to note that $\|\mathcal{F}\|_{\mathbb{P}} \le D_{\mathbb{P}}$, because

$$\|\mathcal{F}\|_{\mathbb{P}} = \sup_{f,g\in\mathcal{F}}\int_{\mathcal{X}} (f(x)-g(x))^2\mathbb{P}(\,\mathrm{d}x) \le 2\sup_{f\in\mathcal{F}}\int_{\mathcal{X}} f(x)^2\mathbb{P}(\,\mathrm{d}x) = D_{\mathbb{P}},$$

which is basically saying that the diameter of the set is not larger than twice the radius.

The other result can be obtained in a similar fashion by not imposing $\varepsilon = 0$ when applying the Chaining bound with the L^∞ norm, i.e.,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |X_f|\right] \lesssim \inf_{\varepsilon}\left\{\mathbb{E}\left[\sup_{\|f-g\|_\infty\le\varepsilon}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i)-g(X_i))\right|\right] + \int_\varepsilon^{\|\mathcal{F}\|_\infty} \sqrt{\log N(u; \mathcal{F}, L^\infty)}\,\mathrm{d}u\right\}$$

$$
\begin{aligned}
&\lesssim \inf_{\varepsilon}\left\{\varepsilon + \int_{\varepsilon}^{\|\mathcal{F}\|_\infty} \sqrt{\log N(u;\mathcal{F},L^\infty)}\,\mathrm{d}u\right\} \\
&= \inf_{\varepsilon}\left\{\varepsilon + \|\mathcal{F}\|_\infty \int_{\varepsilon/\|\mathcal{F}\|_\infty}^{1} \sqrt{\log N(v\|\mathcal{F}\|_\infty;\mathcal{F},L^\infty)}\,\mathrm{d}u\right\} && (u = v\|\mathcal{F}\|_\infty) \\
&= \inf_{\nu}\left\{\nu\|\mathcal{F}\|_\infty + \|\mathcal{F}\|_\infty \int_{\nu}^{1} \sqrt{\log N(v\|\mathcal{F}\|_\infty;\mathcal{F},L^\infty)}\,\mathrm{d}u\right\} && (\nu\|\mathcal{F}\|_\infty = \varepsilon) \\
&= \|\mathcal{F}\|_\infty \inf_{\varepsilon}\left\{\varepsilon + \int_{\varepsilon}^{1} \sqrt{\log N(v\|\mathcal{F}\|_\infty;\mathcal{F},L^\infty)}\,\mathrm{d}u\right\}.
\end{aligned}
$$

■

Example 40 (Parametric class)**.** *Consider the function class*

$$\mathcal{F} = \left\{f_\theta(x) = 1 - e^{-\theta x}, x \in [0,1], \theta \in [0,1]\right\}.$$

We first need to compute the covering number

$$N(\varepsilon;[0,1],|\cdot|) \le \frac{1}{2\varepsilon} + 1.$$

In a previous example we showed that $\|f_{\theta_1} - f_{\theta_2}\|_\infty \le |\theta_1 - \theta_2|$ *which implies*

$$N(\varepsilon;\mathcal{F},\|\cdot\|_\infty) \le N(\varepsilon;[0,1],|\cdot|).$$

Finally, since $f \in \mathcal{F}$ *are bounded* $D_\infty = 1$*. Thus applying the proposition above*

$$
\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &\lesssim \frac{D_\infty}{\sqrt{n}} \int_0^1 \sqrt{\log N(\|\mathcal{F}\|_\infty \cdot u;\mathcal{F},L^\infty)}\,\mathrm{d}u \\
&\lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log\left(1 + \frac{1}{2u\|\mathcal{F}\|_\infty}\right)}\,\mathrm{d}u \lesssim \frac{c}{\sqrt{n}},
\end{aligned}
$$

where the last equality follows from the fact that $\int_0^1 \sqrt{\log(1+1/x)}\,\mathrm{d}x < \infty$*. The one-step discretization bound is looser in this case and will give a bound of order* $\sqrt{\log n}\sqrt{n}$*.*

Consider now the function

$$\mathcal{F} = \left\{f_\theta : x \to \mathbb{R} : \theta \in \mathcal{B}_2^d(1)\right\}$$

and suppose that it's a class of Lipschitz functions

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \le \|\theta_1 - \theta_2\|_2.$$

Then

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{d\log\left(\frac{1}{n} + 1\right)}\,\mathrm{d}u \leqslant \sqrt{\frac{d}{n}},$$

where we used the result that $N(\varepsilon;\mathcal{B}_2^d(1),\|\cdot\|_2) \le (1 + 1/2\varepsilon)^2$*.*

Note that if we want vanishing concentration, then we need d *to grow slower than* n*.* ♣

Example 41 (Non-parametric class with smoothness/convexity)**.** *Consider now the class of functions*

$$\mathcal{F}_L = \{g : [0,1] \to \mathbb{R} \mid g(0) = 0, \quad g \text{ is } L\text{-Lipschitz } \}.$$

We showed that $N(\varepsilon; \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\varepsilon}$*. Moreover,* $\|\mathcal{F}_L^d\|_\infty = L$*.*

$$\mathcal{R}_n(\mathcal{F}_L) \lesssim L \inf_\varepsilon \left[\varepsilon + \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\varepsilon}}\right] \lesssim \frac{L}{n^{\frac{1}{3}}} \qquad \text{(One step bound)}$$

$$\mathcal{R}_n(\mathcal{F}_L) \lesssim \frac{L}{\sqrt{n}} \int_0^1 \sqrt{\frac{1}{u}}\, \mathrm{d}u \lesssim \frac{L}{\sqrt{n}}. \qquad \text{(Chaining bound)}$$

Consider now the class of functions

$$\mathcal{F}_L^d = \{g : [0,1]^d \to \mathbb{R} \mid g(0) = 0, \quad g \text{ is } L\text{-Lipschitz } \}.$$

Then $N(\varepsilon; \mathcal{F}_L^d, \|\cdot\|_\infty) \asymp \left(\frac{L}{\varepsilon}\right)^d$*. Moreover,* $\|\mathcal{F}_L^d\|_\infty = L$*.* ♣

Note. *Note that the Rademacher complexity of classes of parametric functions is typically of order* $1/\sqrt{n}$*, whereas for non-parametric functions is* $1/n^{1/d}$*.* ♦

Example 42 (Boolean function class)**.** *Consider the class of functions*

$$\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}.$$

We saw that $\mathcal{F} \in PD(\nu)$*, where* $\nu = VC(\mathcal{F})$*. We saw that this gave us*

$$\mathcal{R}_n(\mathcal{F}) \lesssim \sqrt{\frac{\nu \log(n)}{n}}.$$

We can show that for $\varepsilon < 1$

$$\sup_{\mathbb{P}} \log\left(N\left(\varepsilon; \mathcal{F}, L^2(\mathbb{P})\right)\right) \lesssim \nu \log\left(\frac{e}{\varepsilon}\right).$$

Using this result

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\nu \log(e/n)} \leq c \cdot \sqrt{\frac{\nu}{n}}$$

♣

Example 43 (Glivenko-Cantelli)**.** *Consider the class of functions*

$$\mathcal{F} = \{\mathbb{1}(\cdot \leq t) : t \in \mathbb{R}\},$$

then

$$\mathcal{R}_n(\mathcal{F}) \lesssim c \cdot \sqrt{\frac{1}{n}}$$

which also implies

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \geq \frac{c}{\sqrt{n}} + \frac{\varepsilon}{\sqrt{n}}\right) \leq 2e^{-\varepsilon^2/2}.$$

♣

Note. *There is a sharper bound than the one above and it is called* ***Dvoretzky-Kiefer-Wolfwitz-Massart*** *bound*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}|F_n(t)-F(t)|\geq\frac{\varepsilon}{\sqrt{n}}\right)\leq 2e^{-2\varepsilon^2}\quad\forall\,\varepsilon>0.$$

♦

3.7 Orlicz Processes

All the bounds in this section apply to sub-Gaussian random variables/processes. It is also useful to obtain bounds on the expected supremum and associated deviation bounds for processes that are sub-exponential in nature. The notion of **Orlicz norm** allows us to treat both sub-Gaussian and sub-exponential processes in a unified manner.

Definition 28 (Orlicz Norm)**.** *Consider a function* $\psi_q(t):=\exp(t^q)-1, q\in[1,2]$. *The* ψ_q-*Orlicz norm of a zero-mean random variable* X *is given by*

$$\|X\|_{\psi_q}:=\inf\{\lambda>0:\mathbb{E}[\psi_q(|X|/\lambda)]\leq 1]\}.$$

The Orlicz norm is infinite if there is no $\lambda\in\mathbb{R}$ *for which the given expectation is finite.*

Proposition 36. *If* $X\sim\mathsf{sG}(\sigma)$ *then* $\|X\|_{\psi_2}\leq\sigma$, *whereas if* $X\sim\mathsf{sE}(\nu,\nu)$, *then* $\|X\|_{\psi_1}\leq\nu$.

Proof. If $X\sim\mathsf{sG}(\sigma)$, then

$$\mathbb{E}[\exp(X^2/\sigma^2)]\leq 2\implies\mathbb{E}[\exp(X^2/\sigma^2)-1]\leq 1\implies\mathbb{E}[\psi_2(|X|/\sigma)]\leq 1.$$

If $X\sim\mathsf{sE}(\nu,\nu)$, then

$$\mathbb{E}[\exp(|X|/\nu)]\leq 2\implies\mathbb{E}[\exp(|X|/\nu)-1]\leq 1\implies\mathbb{E}[\psi_1(|X|/\nu)]\leq 1.$$

■

We can obtain all the results we've seen so far in terms of the Orlicz norm.

Proposition 37 (Concentration inequality)**.**

$$\mathbb{P}(|X|\geq t)=\mathbb{P}\left(\psi_q\left(|X|/\|X\|_{\psi_q}\right)\geq\psi_q\left(t/\|X\|_{\psi_q}\right)\right)\leq\frac{1}{\psi_q\left(t/\|X\|_{\psi_q}\right)}.$$

Proposition 38 (Maximal inequality)**.**

$$\mathbb{E}\left[\max_{i\in[n]}|X_i|\right]\leq\max_i\|X\|_{\psi_q}\cdot\psi^{-1}(n).$$

Proof.

$$\psi_q\left(\mathbb{E}\left[\max_{i\in[n]}|X_i|\right]/\sigma\right)\leq\mathbb{E}\left[\max_{i\in[n]}\psi_q\left(|X_i|/\sigma\right)\right]\leq\sum_{i\in[n]}\mathbb{E}\left[\psi_q\left(|X_i|/\sigma\right)\right]\leq n.$$

■

3.8 Contraction Inequalities

Suppose we have $(X_i, Y_i)_{i\in[n]} \overset{\text{iid}}{\sim} \mathbb{P}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Recall that we define the empirical risk and the population risk as

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) \qquad R(f) = \mathbb{E}[\ell(f(X), Y)].$$

We know that

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

However, we can think of the Rademacher complexity as being over the class of functions

$$\mathcal{G} = \{\ell(f(x), y), f \in \mathcal{F}\},$$

implying that

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

The question now is how does $\mathcal{R}_n(\mathcal{F})$ relate to $\mathcal{R}_n(\mathcal{G})$? If we assume that ℓ is $L-$Lipschitz in its first coordinate, then we can obtain the following result

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\hat{R}_n(f) - R(f)\right|\right] \leq 2\mathbb{E}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(x_i, y_i)\right|\right] \leq 2L \cdot \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)\right|\right] + o(1)$$

Proposition 39 (Talagrand-Ledoux Contraction)**.** *Let $\{\phi_j(\cdot)\}_{j\in[d]}$ be function such that $\phi_j : \mathbb{R} \to \mathbb{R}$, $L-$Lipschitz, and $\phi_j(0) = 0$, i.e. $\{\phi_j(\cdot)\}_{j\in[d]}$ are centered $L-$Lipschitz. Moreover, let $\mathcal{T} \subseteq \mathbb{R}^d$. Then*

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{T}} \sum_{j=1}^{d} \varepsilon_j \phi_j(\theta_j)\right] \leq L \cdot \mathbb{E}\left[\sup_{\theta\in\mathcal{T}} \sum_{j=1}^{d} \varepsilon_j \theta_j\right], \quad \mathbb{E}\left[\sup_{\theta\in\mathcal{T}} \left|\sum_{j=1}^{d} \varepsilon_j \phi_j(\theta_j)\right|\right] \leq 2L \cdot \mathbb{E}\left[\sup_{\theta\in\mathcal{T}} \sum_{j=1}^{d} \varepsilon_j \theta_j\right].$$

Note. *The centered assumption is not crucial. You can always define $\widetilde{\phi}_j(x) = \phi_j(x) - \phi_j(0)$. Then, by triangle inequality, one term will involve a centered Lipschitz function and the other would involve $\varepsilon_j\phi_j(0)$ which we can disregard by standard WLLN/CLT arguments.* ♦

Example 44. *Suppose we have $Z_i = (X_i, Y_i) \overset{iid}{\sim} \mathbb{P}$, where $Y_i \in \{\pm 1\}$, $\Theta = \mathcal{B}_2(r)$, and $\|X\|_2 \leq M$. Define the loss function as*

$$\ell(\theta; z) := \log\left(1 + \exp\left(-y\theta^\top x\right)\right).$$

Note that this is the likelihood function we would get by imposing a logistic regression model $\mathbb{P}(Y = 1 \mid X) = \Lambda(X^\top\theta)$. However, here we don't assume that. First, we already showed that

$$E := \mathbb{E}\left[\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; z) - R(\theta)\right|\right] \leq 2\mathbb{E}\left[\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \ell(\theta; z_i)\right|\right].$$

Now we want to apply the contraction inequality so we need to show that $\ell(\cdot,\cdot)$ is Lipschitz.

$$\ell(\theta; z_i) = \log(1 + \exp(-y_i \langle \theta, x_i \rangle)) = \phi_i(\widetilde{\theta}_i) + \log 2, \quad \phi_i(\widetilde{\theta}_i) = \log(1 + \exp(-y_i \tilde{\theta}_i)) - \log 2,$$

where $\widetilde{\theta} \in \widetilde{\Theta} = \{(\langle \theta, x_1 \rangle, \ldots, \langle \theta, x_n \rangle) : \theta \in \Theta, x \in \mathcal{X}\}$. Then ϕ_i is $1-$Lipschitz. Then

$$\begin{aligned}
E &\leq 2\mathbb{E}\left[\sup_{\widetilde{\theta}\in\widetilde{\Theta}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \phi_i(\widetilde{\theta}_i)\right|\right] + 2\mathbb{E}\left[\sup_{\widetilde{\theta}\in\widetilde{\Theta}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \log 2\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_{\widetilde{\theta}\in\widetilde{\Theta}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \phi_i(\widetilde{\theta}_i)\right|\right] + 2\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] \log 2 \\
&\leq 2L \cdot \mathbb{E}\left[\sup_{\widetilde{\theta}\in\widetilde{\Theta}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{\theta}_i\right|\right] + \frac{2}{\sqrt{n}} \log 2 \\
&= 2 \cdot \mathbb{E}\left[\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle x_i, \theta \rangle\right|\right] \\
&= 2 \cdot r \cdot \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i\right\|_2\right] \leq \frac{2rM}{\sqrt{n}}
\end{aligned}$$

♣

4 Random Matrix Theory

Random matrix theory seeks to study the various properties of matrices with random entries. The sample covariance matrix provides a motivating example. Given n independent random samples $(\mathbf{x}_i)_{i\in[n]} \subseteq \mathbb{R}^d, \mathbb{E}[\mathbf{x}_i] = 0$, how well does the sample covariance matrix $\widehat{\mathbf{\Sigma}} := \frac{1}{n}\sum \mathbf{x}_i\mathbf{x}_i^T$ approximate the true covariance $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T]$? This is measured in terms of

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathtt{op}} \quad \text{or} \quad |\lambda_{\mathtt{max}}(\widehat{\mathbf{\Sigma}}) - \lambda_{\mathtt{max}}(\mathbf{\Sigma})|.$$

To answer such questions, we need to understand the various properties in which these matrices can differ, and how we can control the influence of randomness on these properties.

4.1 Linear Algebra Review

As our interest will primarily be focused on the spectral properties of these matrices, the following representations will be useful. The singular value decomposition (SVD) decomposes a matrix into a pair of orthogonal matrices and a diagonal matrix. Namely, it breaks the transformation down into a pair of rotations, one in the domain and one in the codomain, and a scaling. It also gives a way of breaking a matrix into a sum of simple rank one matrices.

Theorem 3 (Singular Value Decomposition)**.** *Any rectangular matrix $\mathbf{A} \in \mathbb{R}^{n\times m}, n \geq m$ has a decomposition:*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

$$\mathbf{U}^\top\mathbf{U} = \mathbf{I}_m, \quad \mathbf{V}\mathbf{V}^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_m, \quad \textit{and} \quad \mathbf{\Sigma} = \mathrm{diag}(\sigma_1(\mathbf{A}), \ldots, \sigma_m(A))$$

for $\mathbf{U} \in \mathbb{R}^{n\times m}$, $\mathbf{V} \in \mathbb{R}^{m\times m}$, and $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$, where $r = \mathrm{rank}(\mathbf{A})$ is the number of nonzero singular values.

In general, we will assume the singular values are in a non-increasing order. Namely:

$$\sigma_{\mathtt{max}}(\mathbf{A}) = \sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_m(\mathbf{A}) = \sigma_{\mathtt{min}}(\mathbf{A}) \geq 0.$$

The singular values also have a convenient representation in terms of the L^2 norm which is often much more tractable for use in proofs.

Theorem 4 (Variational Representation of Singular Values (Courant-Fisher-Weyl))**.**

$$\sigma_k(\mathbf{A}) = \min_{V:\dim(V)=n-k+1} \max_{x\in V, \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \max_{V:\dim(V)=k} \min_{x\in V, \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2.$$

In particular, for the maximum and minimum singular values:

$$\sigma_{\mathtt{max}}(\mathbf{A}) = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2, \quad \textit{and} \quad \sigma_{\mathtt{min}}(\mathbf{A}) = \min_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2.$$

Recall the operator norm and rank of a matrix:

$$\|\mathbf{A}\|_{\mathtt{op}} := \sigma_{\mathtt{max}}(\mathbf{A}), \quad \textit{and} \quad \operatorname{rank}(\mathbf{A}) := \max\{k : \sigma_k(\mathbf{A}) > 0\} = |\{k : \sigma_k(\mathbf{A}) \neq 0\}|.$$

Although we will not use it much, for completeness we include the Jordan decomposition. This represents A in a basis of its generalized eigenvectors, creating an upper-triangular and almost diagonal matrix.

Theorem 5 (Jordan Normal Form). *Any $\mathbf{A} \in \mathbb{R}^{d\times d}$ is similar to a block diagonal matrix $\mathbf{J} \in \mathbb{R}^{d\times d}$:*

$$\mathbf{A} = \mathbf{U}\mathbf{J}\mathbf{U}^{-1}, \quad \textit{and} \quad \mathbf{J} = \textit{diag}(\mathbf{J}_{d_1}(\lambda_1), ..., \mathbf{J}_{d_k}(\lambda_k)) \tag{4.1}$$

where the blocks J_{d_i} are of the form:

$$\mathbf{J}_{d_i} = \begin{bmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ \vdots & & \ddots & & \\ 0 & 0 & 0 & \dots & \lambda_i \end{bmatrix} \in \mathbb{C}^{d_i \times d_i}. \tag{4.2}$$

Here λ_i are the eigenvalues of $\mathbf{A}$. Moreover this representation is unique up to reordering of the $\mathbf{J}_{d_i}$.

The above lacks utility because in general a matrix does not have a nice eigenbasis, but in the special case of symmetric matrices such a basis does exist. Let $\mathcal{S}^{d\times d}$ denote the set of all $d \times d$ symmetric matrices, i.e.

$$\mathcal{S} := \{\mathbf{Q} \in \mathbb{R}^{d\times d} : \mathbf{Q} = \mathbf{Q}^\top\}.$$

The below theorem shows any symmetric matrix has an orthonormal basis of eigenvectors.

Theorem 6 (Spectral Decomposition of Symmetric Matrices). *Any $\mathbf{Q} \in \mathcal{S}^{d\times d}$ has a decomposition:*

$$\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \quad \textit{with} \quad \mathbf{U} \in \mathbb{R}^{d\times d}, \mathbf{\Lambda} \in \mathbb{R}^{d\times d}. \tag{4.3}$$

Moreover:

$$\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}_d, \quad \textit{and} \quad \mathbf{\Lambda} = \textit{diag}(\lambda_1(\mathbf{Q}), \cdots, \lambda_d(\mathbf{Q}))$$

where λ_i are the eigenvalues of $\mathbf{Q}$, all of which are real, and the columns of $\mathbf{U}$ are an orthonormal basis of the corresponding eigenvectors.

Note. *We can see that if the matrix is symmetric than its eigenvectors $\mathbf{U} = \{\mathbf{u}_1^\top, \mathbf{u}_2^\top, \dots, \mathbf{u}_d^\top\}^\top$ are orthonormal basis for the eigenspace as $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$.* ♦

Again, we assume the eigenvalues are placed in non-increasing order:

$$\lambda_{\mathtt{max}}(\mathbf{Q}) := \lambda_1(\mathbf{Q}) \geq \lambda_2(\mathbf{Q}) \geq \cdots \geq \lambda_d(\mathbf{Q}) =: \lambda_{\mathtt{min}}(\mathbf{Q}).$$

As with the singular values, the eigenvalues have a variational representation more useful for proofs.

Lemma 14 (Variational representation of extremal eigenvalues)**.** *For* $\mathbf{Q} \in \mathcal{S}^{d\times d}$*:*

$$\lambda_{\texttt{max}}(\mathbf{Q}) = \max_{||\mathbf{x}||_2=1} \langle \mathbf{x}, \mathbf{Q}\mathbf{x}\rangle, \quad \textit{and} \quad \lambda_{\texttt{min}}(\mathbf{Q}) = \min_{||\mathbf{x}||_2=1} \langle \mathbf{x}, \mathbf{Q}\mathbf{x}\rangle.$$

For $\mathbf{Q} \in \mathcal{S}^{d\times d}$ there is a simple relationship between eigenvalues and singular values:

$$\{\ |\lambda_k(\mathbf{Q})|\ \}_{k\in[d]} = \{\sigma_k(\mathbf{Q})\}_{k\in[d]}.$$

Note. *Note that the above is an equivalence between sets, but we can't say anything on how the* $k-th$ *eigenvalue relates to the* $k-th$ *singular value. This is because eigenvalues might be negative. The next statement recovers a one-to-one relationships between* λ_k *and* σ_k *by ruling out negative eigenvalues.* ♦

Let $\mathcal{S}_+^{d\times d} := \{\mathbf{Q} \in \mathcal{S}^{d\times d} : \mathbf{Q} \succeq 0\}$ denote the set of positive semi-definite matrices, then

$$\forall\, \mathbf{Q} \in \mathcal{S}_+^{d\times d}, \quad \lambda_k(\mathbf{Q}) = \sigma_k(\mathbf{Q}) \quad \forall k \in [d],$$

because all eigenvalues of positive semi-definite matrices are non-negative. In particular, for $\mathbf{A} \in \mathbb{R}^{n\times m}$ since $\mathbf{A}^\top\mathbf{A} \in \mathcal{S}_+^{m\times m}$ we have:

$$\lambda_k(\mathbf{A}^\top\mathbf{A}) = \sigma_k(\mathbf{A})^2.$$

This last equation is particularly useful because we can first take the SVD of $\mathbf{A}$ and then look at eigenvalues of $\mathbf{A}^\top\mathbf{A}$ which simplifies to

$$\mathbf{A}^\top\mathbf{A} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top,$$

where $\mathbf{V}$ is an eigenmatrix and $\mathbf{\Sigma}$ are eigenvalues.

Often we will be interested in how these spectral properties vary as we perturb our system. Namely, if we have some $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{n\times m}$ where $\mathbf{E}$ is viewed as deterministic noise, how do the eigenvalues and singular values of $\mathbf{A}$ relate to the perturbed $\overline{\mathbf{A}} = \mathbf{A} + \mathbf{E}$? The following proposition bounds the fluctuations of the singular values due to such perturbation.

Proposition 40 (Weyl's Perturbation Inequality)**.** *For any* $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{n\times m}$ *and* $k \in [m]$*:*

$$|\sigma_k(\mathbf{A}+\mathbf{E}) - \sigma_k(\mathbf{A})| \le ||\mathbf{E}||_{\texttt{op}}.$$

If furthermore $\mathbf{Q}, \mathbf{E} \in \mathcal{S}^{d\times d}$ *are symmetric,* $k \in [d]$*:*

$$|\lambda_k(\mathbf{Q}+\mathbf{E}) - \lambda_k(\mathbf{Q})| \le ||\mathbf{E}||_{\texttt{op}}.$$

Note. *The above proposition tells us that singular values and eigenvalues (interpreted as functions) are Lipschitz in the operator norm.* ♦

4.2 Sample Covariance Matrix

Let's apply our above results to the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ which is naturally a PSD matrix:

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$$

with $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{bmatrix}^\top \in \mathbb{R}^{n\times d}$. Then, by defining $\mathbf{Q} = \mathbf{\Sigma}$ and $\mathbf{E} = \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$, we see that Weyl's Inequality implies:

$$|\lambda_k(\mathbf{Q}+\mathbf{E}) - \lambda_k(\mathbf{Q})| \leq \|\mathbf{E}\|_{\mathtt{op}} \quad \iff \quad \left|\lambda_k(\widehat{\mathbf{\Sigma}}) - \lambda_k(\mathbf{\Sigma})\right| \leq ||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}||_{\mathtt{op}}. \tag{4.4}$$

Moreover, as $\widehat{\mathbf{\Sigma}} \in \mathcal{S}_+^{n\times n}$, we further know that

$$\lambda_k(\widehat{\mathbf{\Sigma}}) = \sigma_k(\mathbf{X}/\sqrt{n})^2$$

The operator norm gives us a way of bounding the spectral gap between the sample covariance and true covariance, and further motivates our interest in this norm. Using our variational representation of the eigenvalues, the upper bound in (4.4) can be re-expressed as:

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathtt{op}} := \sup_{\|\boldsymbol{\nu}\|_2=1}\left|\langle\boldsymbol{\nu}, (\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\boldsymbol{\nu}\rangle\right| = \sup_{\|\boldsymbol{\nu}\|_2=1}\left|\frac{1}{n}\sum_{i=1}^{n}\langle\mathbf{x}_i, \boldsymbol{\nu}\rangle^2 - \boldsymbol{\nu}^\top\mathbf{\Sigma}\boldsymbol{\nu}\right|.$$

Therefore, for any fixed d the above is the supremum of an empirical process. In particular, controlling the deviation $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathtt{op}}$ is equivalent to establishing a uniform law of large numbers for the class of functions $\mathbf{x} \mapsto \langle\mathbf{x}, \boldsymbol{\nu}\rangle^2$, indexed by vectors $\boldsymbol{\nu} : \|\boldsymbol{\nu}\|_2 = 1$.

Note. *Sometimes we may want to work with other matrix norms beyond $\|\cdot\|_{\mathtt{op}}$. A standard result in linear algebra is that all such norms are equivalent, differing only up to a universal constant. However, this constant depends on the dimension of the matrices, so we must be careful to take this into account if we want to make claims about arbitrary dimensions.* ♦

4.2.1 Eigenvalues of Sample Covariance of Gaussian Ensembles

Suppose we have a sample where each $\mathbf{x}_i \overset{\mathtt{iid}}{\sim} N(0, \mathbf{\Sigma})$. In this case we say that the associated matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, with $\mathbf{x}_i^\top$ as its $i-$th row, is draw from the $\mathbf{\Sigma}-$Gaussian ensemble. The associated sample covariance $\widehat{\mathbf{\Sigma}}$ is said to follow a multivariate Wishart distribution.

Theorem 7. *Consider the setting:*

$$(\mathbf{x}_i)_{i\in[n]} \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top\end{bmatrix} \in \mathbb{R}^{n\times d}, \quad \textit{and} \quad \widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top = \frac{1}{n}\mathbf{X}^\top\mathbf{X}.$$

Then

$$\mathbb{P}\left(\frac{\sigma_{\mathtt{max}}(\mathbf{X})}{\sqrt{n}} \geq \lambda_{\mathtt{max}}(\sqrt{\mathbf{\Sigma}})(1+t) + \sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})}{n}}\right) \leq e^{-nt^2/2}.$$

Moreover, if $n \geq d$:

$$\mathbb{P}\left(\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{n}} \leq \lambda_{\min}(\sqrt{\boldsymbol{\Sigma}})(1-t) - \sqrt{\frac{\operatorname{tr}(\boldsymbol{\Sigma})}{n}}\right) \leq e^{-nt^2/2}$$

and

$$\mathbb{E}[\sigma_{\max}(\mathbf{X})] \leq \sqrt{n} + \sqrt{d}, \qquad \mathbb{E}[\sigma_{\min}(\mathbf{X})] \geq \sqrt{n} - \sqrt{d}.$$

Corollary 5. *When $\boldsymbol{\Sigma} = \mathbf{I}_d$, the following event happens with probability at least $1-\delta$:*

$$\left\{\begin{array}{l} \dfrac{\sigma_{\max}(\mathbf{X})}{\sqrt{n}} \leq 1 + \sqrt{\dfrac{2\log(2/\delta)}{n}} + \sqrt{\dfrac{d}{n}} \\ \dfrac{\sigma_{\min}(\mathbf{X})}{\sqrt{n}} \geq 1 - \sqrt{\dfrac{2\log(2/\delta)}{n}} - \sqrt{\dfrac{d}{n}} \end{array}\right\}.$$

Note. *By standard properties of the multivariate Gaussian, we can always write $\mathbf{X} = \mathbf{W}\sqrt{\boldsymbol{\Sigma}}$, where $\mathbf{W}$ is a standard Gaussian random matrix and $\boldsymbol{\Sigma} = \sqrt{\boldsymbol{\Sigma}}\sqrt{\boldsymbol{\Sigma}}$. Therefore,*

$$\begin{aligned} \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\text{op}} &= \left\|\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \boldsymbol{\Sigma}\right\|_{\text{op}} \\ &= \left\|\sqrt{\boldsymbol{\Sigma}}\left(\frac{1}{n}\mathbf{W}^\top\mathbf{W}\right)\sqrt{\boldsymbol{\Sigma}} - \sqrt{\boldsymbol{\Sigma}}\sqrt{\boldsymbol{\Sigma}}\right\|_{\text{op}} \\ &= \left\|\sqrt{\boldsymbol{\Sigma}}\left(\frac{1}{n}\mathbf{W}^\top\mathbf{W} - \mathbf{I}_d\right)\sqrt{\boldsymbol{\Sigma}}\right\|_{\text{op}} \\ &= \|\boldsymbol{\Sigma}\|_{\text{op}} \cdot \left\|\frac{1}{n}\mathbf{W}^\top\mathbf{W} - \mathbf{I}_d\right\|_{\text{op}} \end{aligned}$$

which reduces the problem to the case $\boldsymbol{\Sigma} = \mathbf{I}$.[2] ♦

Before proving the expectation bounds, we introduce two useful Gaussian comparison results that we will use to prove the upper and lower bounds of the singular values, respectively.

Proposition 41 (Sudakov-Fernique inequality)**.** *Let $(Z_\theta)_{\theta\in\mathcal{T}}$ and $(Y_\theta)_{\theta\in\mathcal{T}}$ be two continuous Gaussian processes on a separable space $\mathcal{T}$ such that $\mathbb{E}[Z_\theta] = \mathbb{E}[Y_\theta]$ (not necessarily 0) for all $\theta \in \mathcal{T}$. If $\mathbb{E}[(Z_\theta - Z_{\theta'})^2] \leq \mathbb{E}[(Y_\theta - Y_{\theta'})^2]$ for all $\theta, \theta' \in \mathcal{T}$, then*

$$\mathbb{E}\left[\max_{\theta\in\mathcal{T}} Z_\theta\right] \leq \mathbb{E}\left[\max_{\theta\in\mathcal{T}} Y_\theta\right].$$

Note. *There is another way to interpret the condition of the second moment of the difference*

$$\mathbb{E}[(Z_\theta - Z_{\theta'})^2] \leq \mathbb{E}[(Y_\theta - Y_{\theta'})^2].$$

[2] This trick of splitting $\boldsymbol{\Sigma}$ is called "unwhitening" in machine learning

If $\mathbb{E}[Z_\theta^2] = \mathbb{E}[Y_\theta^2]$, *then*

$$\mathbb{E}[(Z_\theta - Z_{\theta'})^2] \leq \mathbb{E}[(Y_\theta - Y_{\theta'})^2] \iff \mathbb{E}[Z_\theta Z_{\theta'}] \geq \mathbb{E}[Y_\theta Y_{\theta'}].$$

In words, we can think of this condition as imposing some structure on the covariance of the Gaussian processes. In particular, it's telling us that the elements in the Z *process covary more than the elements in* Y*. Intuitively, if the elements of a stochastic process covary a lot, then their expected squared differences will be small.* ♦

Example 45. *Let's see an extreme example of two Gaussian processes to fix ideas. In particular, we will consider the case of two processes: one that is perfectly correlated, another that is completely independent. Let* $\mathcal{T} = \{1, 2, \ldots, N\}$, $Y_\theta \overset{iid}{\sim} N(0,1)$ *and* $Z_\theta = Z \sim N(0,1), \forall\, \theta \in \mathcal{T}$*. Then*

$$\mathbb{E}[(Z_\theta - Z_{\theta'})^2] = 0 \leq 2 = \mathbb{E}[(Y_\theta - Y_{\theta'})^2].$$

By the Sudakov-Fernique inequality we get that

$$\mathbb{E}\left[\max_{\theta\in\mathcal{T}} Z_\theta\right] \leq \mathbb{E}\left[\max_{\theta\in\mathcal{T}} Y_\theta\right],$$

but we can actually calculate it! By perfect correlation

$$\mathbb{E}\left[\max_{\theta\in\mathcal{T}} Z_\theta\right] = \mathbb{E}[Z] = 0,$$

whilst by the maximal inequality

$$\mathbb{E}\left[\max_{\theta\in\mathcal{T}} Y_\theta\right] \asymp \sqrt{2\log n}.$$

♣

Note (Smoothed Max Operator)**.** *Before going through the proof, we introduce the "smoothed max" operator with parameter* $\beta > 0$*, which is defined as*

$$F_\beta : \mathbb{R}^n \to \mathbb{R}, \quad F_\beta(v) := \frac{1}{\beta}\log\left(\sum_{i=1}^n e^{\beta v_i}\right).$$

Its first nice property is that

$$\max_{i\in[n]} v_i \leq F_\beta(v) \leq \max_{i\in[n]} v_i + \frac{\log n}{\beta}.$$

To see this note that

$$\max_{i\in[n]} e^{\beta v_i} \leq \sum_{i=1}^n e^{\beta v_i} \leq n \max_{i\in[n]} e^{\beta v_i},$$

because non-negativity of the exponential function and β*. Since the logarithm is an increasing function and* $\beta > 0$

$$\frac{1}{\beta}\log\max_{i\in[n]} e^{\beta v_i} \leq \frac{1}{\beta}\log\sum_{i=1}^n e^{\beta v_i} \leq \frac{1}{\beta}\log n \max_{i\in[n]} e^{\beta v_i}$$

$$\frac{1}{\beta}\max_{i\in[n]}\log e^{\beta v_i} \le F_\beta(v) \le \frac{\log n}{\beta} + \frac{1}{\beta}\max_{i\in[n]}\log e^{\beta v_i}$$
$$\max_{i\in[n]} v_i \le F_\beta(v) \le \frac{\log n}{\beta} + \max_{i\in[n]} v_i.$$

In words, the smoothed max operator is sandwiched between the true maximum and the maximum shifted by a term which is $o(\beta^{-1})$. From this fact we also know that

$$\lim_{\beta\to\infty} F_\beta(x) = \max_{i\in[n]} x_i.$$

The next property is the most important one, which justifies both the name of the operator and why it is extremely useful in practice:

$$\nabla F_\beta : \mathbb{R}^n \to [0,1]^n, \quad \nabla F_\beta(v) = \left(\frac{e^{v_i\beta}}{\sum_{j=1}^n e^{v_j\beta}}\right)_{i\in[n]}, \quad \lim_{\beta\to\infty}\nabla F_\beta(v) = \left(\mathbb{1}(X_i = \max_{j\in[n]]} X_j)\right)_{i\in[n]}.$$

We see that the smoothed max operator is differentiable and its gradient converges to the indicator denoting the maximum. Typically

$$p_i(x) := \frac{e^{v_i\beta}}{\sum_{j=1}^n e^{v_j\beta}}$$

is called the "softmax". Finally

$$\frac{\partial^2}{\partial v_i \partial v_j} F_\beta(v) = \beta(p_i(v)\mathbb{1}(i=j) - p_i(v)p_j(v))$$

♦

Proof of Proposition 41. We prove an equivalent version of the statement which say, suppose that Z and Y are two Gaussian vectors in $\mathbb{R}^n$ with $\mathbb{E}[Z] = \mathbb{E}[Y]$ and

$$\mathbb{E}\left[(Z_i - Z_j)^2\right] \le \mathbb{E}\left[(Y_i - Y_j)^2\right] \quad \forall i,j \in [n].$$

Then

$$\mathbb{E}\left[\max_{i\in[n]} Z_i\right] \le \mathbb{E}\left[\max_{i\in[n]} Y_i\right].$$

The two versions are equivalent despite n being finite because we required $\mathcal{T}$ to be a separable space.

We will rely on the **interpolation method** and the **smoothed max operator**. Let $\mu = \mathbb{E}[Z] = \mathbb{E}[Y]$ and define $\tilde{Z} = Z - \mu$ and $\tilde{Y} = Y - \mu$, so that $\tilde{Z}$ and $\tilde{Y}$ are both Gaussian with zero-mean (but not necessarily independent). Define an auxiliary random vector $W(\theta) := \tilde{Z}\sin\theta + \tilde{Y}\cos\theta \cdot + \mu$ for $\theta \in [0, \pi/2]$ to interpolate between Z and Y. In addition, define $\varphi(\theta) := \mathbb{E}[F_\beta(W(\theta))]$ for $\theta \in [0,\pi/2]$ and $\beta > 0$, where F_β is the smoothed max operator.

Note that the endpoints of the path from 0 to $\pi/2$ are the objects we want to compare

$$\varphi(0) = \mathbb{E}[F_\beta(Z)], \qquad \varphi(\pi/2) = \mathbb{E}[F_\beta(Y)].$$

Therefore we want to show that φ is increasing on $[0, \pi/2]$ for any positive β. To show this, we evaluate the derivative of ϕ using Stein's lemma (essentially integration by part), which states for a standard Gaussian random variable $G \sim N(0,1)$ and any differentiable function $f : \mathbb{R} \to \mathbb{R}$ that $\mathbb{E}[Gf(G)] = \mathbb{E}[f'(G)]$, granted that the two expectations both exist. Using Fubini's theorem to justify changing order of limits and applying Stein's lemma, some algebra yields

$$\begin{aligned}\varphi'(\theta) &= \mathbb{E}\left[\sum_{i=1}^{n} \partial_i F_\beta(W(\theta))(-\sin\theta \cdot \tilde{Z}_i + \cos\theta \cdot \tilde{Y}_i + \mu)\right] \\ &= \cos\theta \sin\theta \sum_{i,j=1}^{n} \mathbb{E}[\partial_{i,j}^2 F_\beta(W(\theta))] \cdot \big(\mathbb{E}[\tilde{Y}_i \tilde{Y}_j] - \mathbb{E}[\tilde{Z}_i \tilde{Z}_j]\big).\end{aligned}$$

Denote $p_i(v) := \partial_i F_\beta(v) = e^{\beta v_i} / \big(\sum_{j=1}^{n} e^{\beta v_j}\big)$, which defines a probability distribution on $[n]$. We can express second derivatives of F_β as

$$\partial_{i,g}^2 F_\beta(v) = \begin{cases} \beta\big(p_i(v) - p_i(v)^2\big) & \text{when } i = j, \\ -\beta p_i(v) p_j(v) & \text{when } i \neq j. \end{cases}$$

The remaining work is just computation:

$$\begin{aligned}&\sum_{i,j=1}^{n} \mathbb{E}[\partial_{i,j}^2 F_\beta(W(\theta))] \cdot \big(\mathbb{E}[\tilde{Y}_i \tilde{Y}_j] - \mathbb{E}[\tilde{Z}_i \tilde{Z}_j]\big) \\ &= \beta \sum_{i=1}^{n} p_i(x)\big(\mathbb{E}[\tilde{Y}_i^2] - \mathbb{E}[\tilde{Z}_i^2]\big) - \beta \sum_{i,j=1}^{n} p_i(x) p_j(x)\big(\mathbb{E}[\tilde{Y}_i \tilde{Y}_j] - \mathbb{E}[\tilde{Z}_i \tilde{Z}_j]\big) \\ &= \frac{\beta}{2} \sum_{i,j=1}^{n} p_i(x) p_j(x)\big(\mathbb{E}[\tilde{Y}_i^2 + \tilde{Y}_j^2] - \mathbb{E}[\tilde{Z}_i^2 + \tilde{Z}_j^2]\big) - \beta \sum_{i,j=1}^{n} p_i(x) p_j(x)\big(\mathbb{E}[\tilde{Y}_i \tilde{y}_j] - \mathbb{E}[\tilde{Z}_i \tilde{Z}_j]\big) \\ &= \frac{\beta}{2} \sum_{i,j=1}^{n} p_i(x) p_j(x)\big(\mathbb{E}[(\tilde{Y}_i - \tilde{Y}_j)^2] - \mathbb{E}[(\tilde{Z}_i - \tilde{Z}_j)^2]\big).\end{aligned}$$

Therefore, $\mathbb{E}[F_\beta(x)] \leq \mathbb{E}[F_\beta(y)]$, and sending $\beta \to \infty$ gives the desired comparison result. ■

The following is an extension of the Sudakov-Fernique inequality.

Theorem 8 (Gordon's inequality)**.** *Let $(Z_{s,t})_{s\in S, t\in T}$ and $(Y_{s,t})_{s\in S, t\in T}$ be two Gaussian processes such that $\mathbb{E}[Z_{s,t}] = \mathbb{E}[Y_{s,t}]$ (not necessarily 0) for all $s \in S$ and $t \in T$. If we have*

$$\mathbb{E}\left[(Z_{s,t_1} - Z_{s,t_2})^2\right] \geq \mathbb{E}\left[(Y_{s,t_1} - Y_{s,t_2})^2\right] \quad \forall t_1, t_2 \in T, s \in S$$

and moreover

$$\mathbb{E}\left[(Z_{s_1,t_1} - Z_{s_2,t_2})^2\right] \leq \mathbb{E}\left[(Y_{s_1,t_1} - Y_{s_2,t_2})^2\right], \forall t_1, t_2 \in T, s_1 \neq s_2 \in S.$$

Then we have

$$\mathbb{E}\left[\max_{s\in S} \min_{t\in T} Z_{s,t}\right] \leq \mathbb{E}\left[\max_{s\in S} \min_{t\in T} Y_{s,t}\right].$$

Proof of Theorem 7. We do the proof for the case $\mathbf{\Sigma} = \mathbf{I}$. By the whitening trick we know that this is without loss of generality. We will only prove the first tail bound (7) involving $\sigma_{\texttt{max}}(\mathbf{X})$ and the proof of the other tail bound is similar. We use the standard two-part proof technique of first proving a concentration bound of $\sigma_{\texttt{max}}(\mathbf{X})$ around its mean, and then bound the mean.

First we prove the concentration result. Namely we aim to show:

$$\mathbb{P}\Big(\Big|\sigma_k(\mathbf{X}) - \mathbb{E}[\sigma_k(\mathbf{X})]\Big| \geq t\Big) \leq 2e^{-t^2/2}. \tag{4.5}$$

By Weyl's Inequality:

$$|\sigma_k(\mathbf{X}) - \sigma_k(\mathbf{Y})| \leq ||\mathbf{X}-\mathbf{Y}||_{\mathtt{op}} \leq ||\mathbf{X}-\mathbf{Y}||_F = ||vec(\mathbf{X}) - vec(\mathbf{Y})||_2. \tag{4.6}$$

Hence viewing the matrices $\mathbf{X}, \mathbf{Y}$ as gaussian vectors, we see the function $\sigma_k(\cdot)$ is a 1-Lipschitz function of i.i.d. $N(0,1)$ random variables. Then we can use Gaussian concetration to get (4.5).

The second step is to bound the expectation of $\sigma_{\mathtt{max}}(\mathbf{X})$ and $\sigma_{\mathtt{min}}(\mathbf{X})$. We first establish the expectation bound for $\sigma_{\mathtt{max}}(\mathbf{X})$. By the variational representation

$$\sigma_{\mathtt{max}}(\mathbf{X}) = \sup_{(\mathbf{u},\mathbf{v})\in S^{n-1}\times S^{d-1}} \langle \mathbf{u}, \mathbf{X}\mathbf{v}\rangle.$$

Note that for a fixed $\mathbf{u} \in S^{n-1}, \mathbf{v} \in S^{d-1}$ (with $S^{m-1} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 = 1\}$) we have that $Z_{\mathbf{u},\mathbf{v}} := \langle \mathbf{u}, \mathbf{X}\mathbf{v}\rangle$ is a Gaussian random variable. Moreover, if we let $\theta = (\mathbf{u}, \mathbf{v})$, we can see that $\{Z_\theta\}_\theta$ is a Gaussian process.[3] This follows from linearity of the inner product and closure of the Gaussian family to linear transformations. Define $Y_{\mathbf{u},\mathbf{v}} = \langle \mathbf{u}, \mathbf{g}\rangle + \langle \mathbf{v}, \mathbf{h}\rangle, (g_j)_{j\in[n]} \overset{\mathtt{iid}}{\sim} N(0,1), (h_j)_{j\in[n]} \overset{\mathtt{iid}}{\sim} N(0,1)$. We now verify the conditions to apply the Sudakov-Fernique inequality to $Z_{\mathbf{u},\mathbf{v}}$ and $Y_{\mathbf{u},\mathbf{v}}$.

$$\begin{aligned}
\mathbb{E}[Z_{\mathbf{u},\mathbf{v}} Z_{\mathbf{u}',\mathbf{v}'}] &= \mathbb{E}[\langle \mathbf{X}, \mathbf{u}\mathbf{v}^\top\rangle \langle \mathbf{X}, \mathbf{u}'\mathbf{v}'^\top\rangle] \\
&= \mathbb{E}\left[\sum_{i,i'=1}^{n} \sum_{j,j'=1}^{d} X_{ij} u_i v_j \cdot X_{i'j'} u'_{i'} v'_{j'} \right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{d} u_i u'_i \cdot v_j v'_j \\
&= \langle \mathbf{u}\mathbf{v}^\top, \mathbf{u}'\mathbf{v}'^\top\rangle \\
&= \langle \mathbf{u}, \mathbf{u}'\rangle \cdot \langle \mathbf{v}, \mathbf{v}'\rangle,
\end{aligned}$$

where in the third equality the cross terms vanish by the independence assumption among the entries and ve also used that each X_{ij} has unit variance. Therefore,

$$\mathbb{E}[(Z_{\mathbf{u},\mathbf{v}} - Z_{\mathbf{u}',\mathbf{v}'})^2] = \mathbb{E}[Z^2_{\mathbf{u},\mathbf{v}}] - 2\,\mathbb{E}[Z_{\mathbf{u},\mathbf{v}} Z_{\mathbf{u}',\mathbf{v}'}] + \mathbb{E}[Z^2_{\mathbf{u},\mathbf{v}}] = 2 - 2\cdot\langle \mathbf{u}, \mathbf{u}'\rangle\langle \mathbf{v}, \mathbf{v}'\rangle$$

since $\|u\|_2 = \|v\|_2 = 1$. Similarly, we have

$$\mathbb{E}[Y_{\mathbf{u},\mathbf{v}} Y_{\mathbf{u}',\mathbf{v}'}] = \mathbb{E}[(\langle \mathbf{u}, \mathbf{g}\rangle + \langle \mathbf{v}, \mathbf{h}\rangle)(\langle \mathbf{u}', \mathbf{g}\rangle + \langle \mathbf{v}', \mathbf{h}\rangle)] = \langle \mathbf{u}, \mathbf{u}'\rangle + \langle \mathbf{v}, \mathbf{v}'\rangle,$$

and thus

$$\mathbb{E}[(Y_{\mathbf{u},\mathbf{v}} - Y_{\mathbf{u}',\mathbf{v}'})^2] = \mathbb{E}[Y^2_{\mathbf{u},\mathbf{v}}] - 2\,\mathbb{E}[Y_{\mathbf{u},\mathbf{v}} Y_{\mathbf{u}',\mathbf{v}'}] + \mathbb{E}[Y^2_{\mathbf{u}',\mathbf{v}'}] = 4 - 2(\langle \mathbf{u}, \mathbf{u}'\rangle + \langle \mathbf{v}, \mathbf{v}'\rangle).$$

The last step to verify the condition for Sudakov-Fernique inequality

$$\mathbb{E}[(Y_{\mathbf{u},\mathbf{v}} - Y_{u',v'})^2] - \mathbb{E}[(Z_{\mathbf{u},\mathbf{v}} - Z_{\mathbf{u}',\mathbf{v}'})^2] = 2\cdot(1 - \langle \mathbf{u}, \mathbf{u}'\rangle)(1 - \langle \mathbf{v}, \mathbf{v}'\rangle) \geq 0,$$

where we used the Cauchy-Schwartz inequality to deduce $\langle \mathbf{u}, \mathbf{u}'\rangle, \langle \mathbf{v}, \mathbf{v}'\rangle \leq 1$ in the last inequality.

Therefore

$$\mathbb{E}\left[\max_{(\mathbf{u},\mathbf{v})\in S^{n-1}\times S^{d-1}} \langle \mathbf{u}, \mathbf{X}\mathbf{v}\rangle \right] \leq \mathbb{E}\left[\max_{(\mathbf{u},\mathbf{v})\in S^{n-1}\times S^{d-1}} (\langle \mathbf{u}, \mathbf{g}\rangle + \langle \mathbf{v}, \mathbf{h}\rangle) \right] \tag{Sudakov-Fernique}$$

[3] A stochastic process $\{Z_\theta\}_{\theta\in\Theta}$ is a Gaussian process if $\forall\, n, (\theta_1, \ldots, \theta_n) \subseteq \Theta$ the distribution of $(Z_{\theta_1}, Z_{\theta_2}, \ldots, Z_{\theta_n})$ is multivariate Gaussian.

$$
\begin{aligned}
&= \mathbb{E}[\max_{\mathbf{u}\in S^{n-1}} \langle \mathbf{u}, \mathbf{g}\rangle] + \mathbb{E}[\max_{\mathbf{v}\in S^{d-1}} \langle \mathbf{v}, \mathbf{h}\rangle] && \text{(independence)}\\
&= \mathbb{E}[\|\mathbf{g}\|_2] + \mathbb{E}[\|\mathbf{h}\|_2] && \text{(definition)}\\
&\le \sqrt{n} + \sqrt{d}. && \text{(Hölder's)}
\end{aligned}
$$

The definition step follows from the fact that we defined the unit sphere S with respect to the ℓ_2-norm and the fact that the dual norm of the ℓ_2-norm is the ℓ_2-norm.

Now we establish the lower bound of $\mathbb{E}[\sigma_{\texttt{min}}(\mathbf{X})]$. Note that we require $n \ge d$, since the inequality is trivial otherwise. Define $Z_{\mathbf{u},\mathbf{v}}, Y_{\mathbf{u},\mathbf{v}}$ as before. Some straightforward calculation verifies that $Z_{\mathbf{u},\mathbf{v}}$ and $Y_{\mathbf{u},\mathbf{v}}$ indeed satisfy the conditions for Gordon's inequality (Theorem 8). Recall the following variational definition of smallest singular value

$$\sigma_{\texttt{min}}(\mathbf{X}) = \min_{v\in S^{d-1}} \max_{u\in S^{n-1}} Z_{\mathbf{u},\mathbf{v}}.$$

Using the variational characterization, we have

$$
\begin{aligned}
-\mathbb{E}[\sigma_{\texttt{min}}(X)] &\overset{(1)}{=} -\mathbb{E}\left[\min_{\mathbf{v}\in S^{d-1}} \max_{\mathbf{u}\in S^{n-1}} Z_{\mathbf{u},\mathbf{v}}\right]\\
&\overset{(2)}{=} -\mathbb{E}\left[\min_{\mathbf{v}\in S^{d-1}} \max_{\mathbf{u}\in S^{n-1}} Y_{\mathbf{u},\mathbf{v}}\right]\\
&= \mathbb{E}\left[\max_{\mathbf{v}\in S^{d-1}} \min_{\mathbf{u}\in S^{n-1}} \left(\langle \mathbf{g}, \mathbf{u}\rangle + \langle \mathbf{h}, \mathbf{v}\rangle\right)\right]\\
&\overset{(3)}{=} \mathbb{E}\left[\max_{\mathbf{v}\in S^{d-1}} \langle \mathbf{h}, \mathbf{v}\rangle\right] + \mathbb{E}\left[\min_{\mathbf{u}\in S^{n-1}} \langle \mathbf{g}, \mathbf{u}\rangle\right]\\
&\overset{(4)}{=} \mathbb{E}[\|\mathbf{h}\|_2] - \mathbb{E}[\|\mathbf{g}\|_2] \lesssim \sqrt{d} - \sqrt{n},
\end{aligned}
$$

where we have used the variational definition of singular value in (1), Gordon's inequality in (2), linearity of expectation and separability of the objective in (3), and finally the definition of dual norm in (4).

Putting together the pieces gives the claim. ■

The Gaussian comparison inequalities also lead to the following lower bound on the supremum of Gaussian processes, known as the Sudakov minoration.

Theorem 9 (Sudakov minoration)**.** *Let $(X_\theta)_{\theta\in T}$ be a zero-mean Gaussian process on $T \neq \emptyset$. Then*

$$\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right] \ge \sup_{\epsilon>0} \frac{\epsilon}{2}\sqrt{\log M(\epsilon; T, \rho_X)},$$

where $M(\epsilon; T, \rho_X)$ is the ϵ-packing number of T in the metric $\rho_X(\theta, \theta') := \sqrt{V(X_\theta - X_{\theta'})}$.

Heuristically, the Sudakov lower bound implies that the one-step discretization bound for Gaussian complexity is nearly tight.

4.3 Concentration of sub-Gaussian sample covariance

Let's first quickly remind ourselves with the definition of variance and sample covariance. Given any set of i.i.d. sampled random variables $\{X_i\}_{i\in[n]} \overset{\texttt{iid}}{\sim} \mathbb{P} \in \mathcal{P}(\mathbb{R}^d), \mathbb{E}[X_i] = 0$, its covariance (denoted as Σ) and sample covariance (denoted as $\widehat{\Sigma}$) are defined respectively as:

$$\Sigma = \mathbb{E}[X_i X_i^\top] \in \mathcal{S}_+^{d\times d}, \text{ and } \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top = \frac{1}{n} X^\top X \in \mathcal{S}_+^{d\times d}.$$

So far we proved that if $(X_i)_{i\in[n]} \overset{\texttt{iid}}{\sim} N(0,\Sigma)$, then

$$\|\widehat{\Sigma}-\Sigma\|_{\texttt{op}} \lesssim \|\Sigma\|_{\texttt{op}} \cdot \left(\sqrt{\frac{d}{n}}+\frac{d}{n}\right),$$

with probability at least $1-\delta$.

Now we introduced the notion of sub-Gaussian random vector, which is a natural extension of sub-Gaussian random variables:

Definition 29 (Sub-Gaussian random vector)**.** *We say a r.v. $x \in \mathbb{R}^d$ with mean 0 is a sub-Gaussian random vector, denoted as $\mathsf{sG}(\sigma)$, if:*

$$\mathbb{E}[e^{\lambda\langle v,x\rangle}] \le e^{\lambda^2\|v\|_2^2\sigma^2/2}, \qquad \forall\,\lambda\in\mathbb{R}, \forall\, v\in\mathbb{R}^d.$$

A sufficient condition for $X\in\mathbb{R}^d$ to be $\mathsf{sG}(\sigma)$ is that each X_i is independent and $X_i\sim\mathsf{sG}(\sigma)$ since:

$$\mathbb{E}[e^{\lambda\langle v,X\rangle}] = \prod_{i=1}^d \mathbb{E}[e^{\lambda v_i X_i}] \le \prod_{i=1}^d e^{\lambda^2 v_i^2\sigma^2/2} = e^{\lambda^2\|v\|_2^2\sigma^2/2}.$$

Theorem 10. *Let $\{X_i\}_{i\in[n]}\in\mathbb{R}^d$ be independent mean-0 $\mathsf{sG}(\sigma)$. Then with probability at least $1-\delta$, we have*

$$\|\widehat{\Sigma}-\Sigma\|_{\texttt{op}} \le C\sigma^2\left(\sqrt{\frac{d+\log(2/\delta)}{n}}+\frac{d+\log(2/\delta)}{n}\right),$$

for some universal constant C.

Up to a constant, this result matches the one obtained assuming Gaussianity.

We need to introduce one more lemma before we can tackle the proof of Theorem 10:

Lemma 15. *Let $\Omega_\epsilon=\{v^1,\dots,v^{N_\epsilon}\}$ be the ϵ-covering of $\mathcal{S}^{d-1}$ in $\|\cdot\|_2$ norm. For all $A\in\mathbb{R}^{d\times d}$,*

$$\|A\|_{\texttt{op}} \le \frac{1}{1-2\epsilon-\epsilon^2}\sup_{v\in\Omega_\epsilon}|\langle v,Av\rangle|.$$

Proof. Given the definition of ϵ-covering, we know that for all $v\in\mathcal{S}^{d-1}$, there exists $v^j\in\Omega_\epsilon$ such that $\|v-v^j\|\le\epsilon$. This implies that:

$$\begin{aligned}
&\langle v,Av\rangle = \langle v^j,Av^j\rangle + 2\langle v-v^j,Av^j\rangle + \langle v-v^j,A(v-v^j)\rangle\\
\implies &\sup_{v\in\mathcal{S}^{d-1}}|\langle v,Av\rangle| \le \sup_{v^j\in\Omega_\epsilon}|\langle v^j,Av^j\rangle| + (2\epsilon+\epsilon^2)\|A\|_{\texttt{op}}\\
\implies &\|A\|_{\texttt{op}} \le \frac{1}{1-2\epsilon-\epsilon^2}\sup_{v\in\Omega_\epsilon}|\langle v,Av\rangle|.
\end{aligned}$$

■

Proof of Theorem 10. Let $\epsilon = 1/4$. By previous results on cover number, we have $|\Omega_\epsilon| \leq (1 + \frac{2}{\epsilon})^d = 17^d$. Also by the previous Lemma, it follows that:

$$\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \leq 2 \sup_{v \in \Omega_{1/4}} |\langle v, (\hat{\Sigma} - \Sigma)v\rangle|.$$

Considering the sub-Exponential tail bound of $|\langle v, (\hat{\Sigma} - \Sigma)v\rangle|$ for fixed v, we have

$$\left|\langle v, (\hat{\Sigma} - \Sigma)v\rangle\right| = \left|\frac{1}{n}\sum_{i=1}^{n}(\langle v, x_i\rangle^2 - \mathbb{E}[\langle v, X_i\rangle^2])\right|.$$

Then, since we know that $\langle v, X_i\rangle/\sigma$ is $\mathsf{sG}(1)$, previous lectures gives:

$$\begin{aligned}
&\left[\langle v, X_i\rangle^2 - \mathbb{E}[\langle v, X_i\rangle^2]\right]/\sigma^2 \in \mathsf{sE}(1, 1) \\
\implies &\frac{1}{n}\sum_{i=1}^{n}(\langle v, X_i\rangle^2 - \mathbb{E}[\langle v, X_i\rangle^2])/\sigma^2 \in \mathsf{sE}(\frac{1}{\sqrt{n}}, \frac{1}{n}) \\
\implies &\mathbb{P}(|\langle v, (\widehat{\Sigma} - \Sigma)v\rangle| \geq \sigma^2 t) \leq 2\exp(-n\min\{t^2, t\})
\end{aligned}$$

Then by a union bound:

$$\mathbb{P}(\sup_{v \in \Omega_{1/8}} |\langle v, (\widehat{\Sigma} - \Sigma)v\rangle| \geq \sigma^2 t) \leq |\Omega_\epsilon| 2\exp(-n\min\{t^2, t\}) = 2\exp(-n\min\{t^2, t\} + d\log 17).$$

Finally, setting

$$\delta = 2\exp(-n\min\{t^2, t\} + d\log 17)$$

yields

$$t = C\max\{\sqrt{\frac{d + \log(1/\delta)}{n}}, \frac{d + \log(1/\delta)}{n}\}$$

for some constant C, thereby proving the original statement. ■

Note. *With respect to the Gaussian case we lost the possibility of saying something about the behavior of the singular values.* ♦

4.3.1 Concentration of sample covariance of bounded random vector

Theorem 11. *Given a set of independent random vectors $\{X_i\}_{i\in[n]} \in \mathbb{R}^d$ with $\mathbb{E}[X_i] = 0$ and covariance $\Sigma = \mathbb{E}[X_i X_i^\top]$, and $\|X_i\|_2 \leq b$ almost surely for all i, then with probability $1 - \delta$,*

$$\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \lesssim \sqrt{\frac{b\|\Sigma\|_{\mathsf{op}}\log(2d/\delta)}{n}} + \frac{b}{n}\log(2d/\delta).$$

In some sense, $\|X\|_2 \leq b$ is stronger than sub-Gaussianity, but if we apply the result we saw in the previous section we will lose a factor of d.

Example 46. *Consider $X \sim \mathtt{Unif}(\{\sqrt{d}\cdot \mathbf{e}_i\}_{i\in[n]})$. Then $\Sigma = I_d$, and $b = \sqrt{d}$. Consequently,*

$$\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \lesssim \sqrt{\frac{d\log(2d/\delta)}{n}} + \frac{d}{n}\log(2d/\delta),$$

which is the sharp bound. Note that Theorem 10 would get an additional factor of d. ♣

The proof of this Theorem is a direct corollary of the matrix Bernstein theorem, which we will introduce and prove in later sections.

4.4 Matrix Hoeffding/Bernstein inequality

We already showed, in the scalar case, that if $X_i \overset{\texttt{iid}}{\sim} \mathsf{sG}(\sigma)$ and $\mathbb{E}[X_i] = 0$, then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2 \cdot \exp\left\{-\frac{nt^2}{2\sigma}\right\}.$$

In what follows we will show that the same results carries over to matrices, by accounting for the appropriate modifications. Namely, if $Q_i \overset{\texttt{iid}}{\sim} \mathsf{sG}(V),]E[Q_i] = 0, Q_i \in \mathcal{S}^{d\times d}$, then

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} Q_i\right\|_{\texttt{op}} \geq t\right) \leq 2 \cdot d \cdot \exp\left\{-\frac{nt^2}{2\|V\|_{\texttt{op}}}\right\}.$$

Note. *We haven't defined what it means that $Q_i \overset{iid}{\sim} \mathsf{sG}(V)$ when Q_i is a matrix, but we see the similarity between the two cases.* ♦

To show the result above, we start with defining new quantities.

Definition 30. *For any symmetric matrix $Q \in \mathcal{S}^{d\times d}$, and an analytic function f, with a slight abuse of notation we will define*

$$f(Q) \triangleq U\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_d))U^\top.$$

In words, when an analytic function f is applied to a symmetric matrix Q, we need consider it's impact on its eigenvalue, it doesn't act on Q pointwise.

Example 47. *If $f(x) = e^x$, then $e^Q = U\mathrm{diag}(e^{\lambda_1}, \ldots, e^{\lambda_d})U^\top$. If f admits a Taylor expansion, then*

$$f(x) = \sum_{l=0}^{\infty} \frac{f^{(l)}(0)}{l!} x^l,$$

similarly

$$f(Q) = \sum_{l=0}^{\infty} \frac{f^{(l)}(0)}{l!} Q^l, \qquad e^Q = \sum_{l=0}^{\infty} \frac{1}{l!} Q^l,$$

where $Q^l = Q \times \cdots \times Q$. ♣

Let's now refresh the proof for the Hoeffding's bound in the scalar case

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \geq t\right) &\leq \inf_{\lambda\geq 0} \frac{\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])}\right]}{e^{\lambda t}} && \text{(Chernoff)} \\
&= \inf_{\lambda\geq 0} \frac{\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right]}{e^{\lambda t}} && \text{(Scalar Ternsorization)} \\
&\leq \inf_{\lambda\geq 0} \left(\prod_{i=1}^{n} e^{\frac{\lambda^2}{2}\sigma^2}\right) e^{-\lambda t} && \text{(scalar } \mathsf{sG}\text{)}
\end{aligned}$$

$$= e^{-\frac{nt^2}{2\varepsilon^2}} \qquad \text{(optimization)}$$

Now we will substitute each of those steps with the appropriate matrix version.

Matrix Chernoff. Let $Q \in \mathcal{S}^{d\times d}$, e.g., $Q = \frac{1}{n}\sum_{i=1}^n Q_i$ where $Q_i = X_i X_i^\top$ so that $\mathbb{E}[Q_i] = 0$. Then for any $\lambda \geq 0$,

$$\begin{aligned}
\mathbb{P}(\lambda_{\max}(Q) \geq t) &= \mathbb{P}(e^{\lambda\cdot\lambda_{\max}(Q)} \geq e^{\lambda t}) \\
&= \mathbb{P}(\lambda_{\max}(e^{\lambda Q}) \geq e^{\lambda t}) \\
&\leq \frac{\mathbb{E}[\lambda_{\max}(e^{\lambda Q})]}{e^{\lambda t}} && \text{(Markov)} \\
&\leq \frac{\mathbb{E}[\operatorname{tr}(e^{\lambda Q})]}{e^{\lambda t}} \\
&= \frac{\operatorname{tr}(\mathbb{E}[e^{\lambda Q}])}{e^{\lambda t}} && \text{(linearity of trace)}
\end{aligned}$$

where the first equality is needed to make the random variable non-negative and apply Markov's inequality; the second equality follows form the fact that the index of the maximum eigenvalue of Q is the same as the index of the maximum eigenvalue of $e^{\lambda Q}$ because λ is positive and the exponential function is increasing, thus $\lambda_{\max}(e^{\lambda Q}) = e^{\lambda\cdot\lambda_{\max}(Q)}$; the last inequality follows from the fact that the trace of a square matrix is the sum of the eigenvalues.

Thus this directly gives the Matrix Chernoff Lemma:

Lemma 16 (Matrix Chernoff)**.** *For any random symmetric matrix $Q \in \mathcal{S}^{d\times d}$, we have*

$$\mathbb{P}(\lambda_{\max}(Q) \geq t) \leq \inf_{\lambda\geq 0}\left[\frac{\operatorname{tr}(\mathbb{E}(e^{\lambda Q}))}{e^{\lambda t}}\right].$$

Sub-Gaussian/Exponential Matrix. Now we introduce the notions of sub-Gaussian or sub-Exponential random matrices, where are matrix counterparts of sub-Guassuan or sub-Exponential random variables. We will use the notation $A \preceq B$ to denote that the matrix $B - A$ is positive semi-definite.

Definition 31 (Sub-Gaussian random matrix)**.** *For any random symmetric matrix $Q \in \mathcal{S}^{d\times d}$ with $\mathbb{E}[Q] = 0$, if*

$$\mathbb{E}[e^{\lambda Q}] \preceq e^{\lambda^2 V/2}, \quad \forall\lambda \in \mathbb{R},$$

for some $V \in \mathcal{S}_+^{d\times d}$, then Q is called a sub-Gaussian matrix denoted as $\mathsf{sG}(V)$.

Note. *Note that if V is a scalar we have to pay attention because if $V = \sigma^2$, then we would conclude that $X \sim \mathsf{sG}(\sigma^2)$ when we instead know that $X \sim \mathsf{sG}(\sigma)$. To make everything coherent, we need to take the square root of the sub-Gaussian parameter in the scalar case.* ♦

Example 48. *As an example, if $Q = \epsilon B$, where $B \in \mathcal{S}^{d\times d}$ is a deterministic matrix and $\varepsilon \sim \text{Unif}(\{\pm 1\})$. Then $\mathbb{E}[Q^\ell] = 0$ for ℓ odd. Therefore,*

$$\mathbb{E}[e^{\lambda Q}] = \sum_{\ell=0}^{\infty} \frac{\lambda^\ell}{\ell!}\mathbb{E}[Q^\ell] = \sum_{l=0}^{\infty} \frac{\lambda^{2\ell}}{(2\ell)!}\mathbb{E}[Q^{2\ell}] = \sum_{\ell=0}^{\infty} \frac{\lambda^{2\ell}}{(2\ell)!}B^{2\ell} \preceq \sum_{\ell=0}^{\infty} \frac{1}{\ell!}\left(\frac{\lambda^2 B^2}{2}\right)^\ell = e^{\lambda^2 B^2/2},$$

where the inequality step follows from the factorial/bi-factorial relationship which says that $(2\ell)! \geq (2\ell)!! = \ell! 2^\ell$. Therefore, $Q \sim \mathsf{sG}(B^2)$. ♣

Definition 32 (Sub-Exponential random matrix)**.** *For any random symmetric matrix $Q \in \mathcal{S}^{d\times d}$ with $\mathbb{E}[Q] = 0$, if*

$$\mathbb{E}[e^{\lambda Q}] \preceq e^{\lambda^2 V/2}, \quad \forall |\lambda| \leq 1/\alpha,$$

for some $V \in \mathcal{S}_+^{d\times d}, \alpha \in \mathbb{R} \geq 0$, then Q is called a sub-Exponential matrix denoted as $\mathsf{sE}(V, \alpha)$.

Note. *A sufficient condition of Q being $\mathsf{sE}(\mathbb{V}(Q), b)$ is that $\mathbb{E}[Q] = 0$ and $\|Q\|_{\mathsf{op}} \leq b$ a.s. Note $\mathbb{V}(Q) = \mathbb{E}[Q^2] - \mathbb{E}[Q]^2$. As an example, if $\|X_i\|_2 \leq \sqrt{b}$ a.s. and $\mathbb{E}[X_i X_i^\top] = \Sigma$, then defining $Q = X_i X_i^\top - \Sigma$ means $\|Q\|_{\mathsf{op}} \leq b$ and $\mathbb{V}(Q) \preceq b\Sigma$, therefore $Q \sim \mathsf{sE}(b\Sigma, b)$.* ♦

Tensorization of matrix MGF Note that unlike the scalar case, even for independent random matrices $\{Q_i\}_{i\in[n]}$, in general we have

$$\mathbb{E}[e^{\lambda \sum_i Q_i}] \neq \prod_i \mathbb{E}[e^{\lambda Q_i}]$$

since $e^{A+B} \neq e^A \cdot e^B$ for matrices A, B, i.e., the exponential function is non-commutative in the field of matrices. Thus, we use the following lemma.

Lemma 17. *Assuming $\{Q_i\}_{i\in[n]}$ independent,*

$$\text{tr}(\mathbb{E}[e^{\lambda \sum_{i=1}^n Q_i}]) \leq \text{tr}(e^{\sum_{i=1}^n \log \mathbb{E}[\exp(\lambda Q_i)]})$$

This can be easily proved by the sequential application of Jensen's inequality, given this following Lemma:

Lemma 18 (Lieb's inequality, Lieb 1973)**.** *For $H \in \mathcal{S}^{d\times d}$, if $f : \mathcal{S}^{d\times d} \mapsto \mathbb{R}$ is a function such that:*

$$f(A) = \text{tr}(\exp(H + \log A)),$$

then f is concave.

Then we are prepared to introduce the Matrix Hoeffding and Matrix Bernstein results:

Theorem 12 (Matrix Hoeffding)**.** *Given $\{Q_i\} \overset{iid}{\sim} \mathsf{sG}(V_i)$, $\mathbb{E}[Q_i] = 0$, then:*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_{\mathsf{op}} \geq t\right) \leq 2d\exp(-nt^2/(2\sigma^2)),$$

where $\sigma^2 = \|1/n \sum_{i=1}^n V_i\|_{\mathsf{op}}$.

Proof of Theorem 12.

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n Q_i\right) \geq t\right) &\leq \inf_{\lambda\geq 0} \mathbb{E}\left[\operatorname{tr}(e^{\lambda\sum_i Q_i}) \geq t\right] e^{-\lambda nt} && \text{(by Matrix Chernoff)}\\
&\leq \inf_{\lambda\geq 0} \operatorname{tr}(e^{\sum_{i=1}^n \log \mathbb{E}[\exp(\lambda Q_i)]})e^{-\lambda nt} && \text{(by Lemma 17)}\\
&\leq \inf_{\lambda\geq 0} \operatorname{tr}(e^{\sum_{i=1}^n (\lambda^2/2)V_i})e^{-\lambda nt}) && \text{(by sG property)}\\
&\leq d \inf_{\lambda\geq 0} e^{(\lambda^2/2)n\|V\|_{\mathsf{op}}} e^{-\lambda nt}\\
&= d e^{-(nt^2)/(2\|V\|_{\mathsf{op}})}.
\end{aligned}$$

The third step is not so trivial:

$$\begin{aligned}
\mathbb{E}[\exp(\lambda Q_i)] \preceq e^{\lambda^2/2V_i} &\iff \log \mathbb{E}[\exp(\lambda Q_i)] \preceq \log e^{\lambda^2/2V_i} && \text{(log is matrix monotone)}\\
&\iff \sum_{i=1}^n \log \mathbb{E}[\exp(\lambda Q_i)] \preceq \frac{\lambda^2}{2}\sum_{i=1}^n V_i && \text{(sum of PSD is PSD)}\\
&\iff \operatorname{tr}\left(e^{\sum_{i=1}^n \log \mathbb{E}[\exp(\lambda Q_i)]}\right) \preceq \operatorname{tr}\left(e^{\frac{\lambda^2}{2}\sum_{i=1}^n V_i}\right). && \text{(property of trace)}
\end{aligned}$$

In general, if $A \preceq B$ it is not true that $e^A \preceq e^B$. Therefore, the last row above uses the fact that if $f : \mathbb{R} \to \mathbb{R}$ is continuous and non-decreasing, then

$$\operatorname{tr} f(A) \leq \operatorname{tr} f(B).$$

■

Theorem 13 (Matrix Bernstein)**.** *Given* $\{Q_i\} \overset{iid}{\sim} \mathsf{sE}(V_i, \alpha_i)$*, then:*

$$\mathbb{P}(\|\frac{1}{n}\sum_{i=1}^n Q_i\|_{\mathsf{op}} \geq t) \leq 2d\exp(-n\min\{\frac{t^2}{2\sigma^2}, \frac{t}{2\alpha_*}\})$$

where $\sigma^2 = \|1/n\sum_{i=1}^n V_i\|_{\mathsf{op}}$ *and* $\alpha_* = \max_i \alpha_i$

Proof of Theorem 11. We know that $x_i \sim \mathsf{sE}(b\Sigma, b)$, so then applying Matrix Bernstein directly gives:

$$\mathbb{P}(\|\hat{\Sigma} - \Sigma\|_{\mathsf{op}} \geq t) \leq 2d\exp(-n\min\{\frac{t^2}{2b\|\Sigma\|_{\mathsf{op}}}, \frac{t}{2b}\})$$

implying

$$\|\hat{\Sigma} - \Sigma\|_{\mathsf{op}} \lesssim \sqrt{\frac{b\|\Sigma\|_{\mathsf{op}}\log(d/\delta)}{n}} + \frac{b}{n}\log(d/\delta).$$

■

5 Sparse Linear Models

In this section, we aim to develop a theory that can be applied to high-dimensional regimes of linear models where the dimension scales with the sample size (i.e. $d \asymp n$ or $d \gg n$). To this end, we introduce sparsity constraints to ensure consistent estimation and study efficient solutions to the estimation problem.

We recall the basic setting of a linear model and briefly introduce the high-dimensional linear model. First, we define some notation:

$$\begin{aligned} \mathbf{y} \in \mathbb{R}^n &: \text{ response vector} \\ \mathbf{X} \in \mathbb{R}^{n \times d} &: \text{ design matrix} \\ \theta^\star \in \mathbb{R}^d &: \text{ ground truth} \\ \mathbf{w} \in \mathbb{R}^n &: \text{ noise vector} \end{aligned}$$

where

$$\mathbf{y} = (y_1, y_2, \dots, y_n)^\top = \mathbf{X}\theta^\star + \mathbf{w},$$
$$\mathbf{X} = (\mathbf{x}_1^\top, \mathbf{x}_2^\top, \dots, \mathbf{x}_n^\top)^\top, \mathbf{x}_i \in \mathbb{R}^d, i = 1, 2, \dots n, \quad \theta^\star = (\theta_1^\star, \theta_2^\star, \dots, \theta_d^\star)^\top, \quad \mathbf{w} = (w_1, w_2, \dots, w_n)^\top.$$

The model can also be expressed in scalar form as follows:

$$y_i = \langle \mathbf{x}_i, \theta^\star \rangle + w_i \quad \text{for } i = 1, 2, \dots, n.$$

Note. *At this level of generality, we have not imposed some structure on* $(\mathbf{y}, \mathbf{w}, \mathbf{X})$*. In what follows we will prove two sets of results: deterministic results and stochastic results.* ♦

Our objective is to estimate $\theta^\star$ given $(\mathbf{X}, \mathbf{y})$. In the classical asymptotic setting, the dimension d is fixed and we consider the regime where $n \to \infty$. In the high dimensional regime, n and d can be both large and in particular, d can be much larger than n. The **least squares estimator does not admit unique solution** in the regime where $d \gg n$, therefore, **additional structural assumptions** on $\theta^\star$ and $\mathbf{X}$ are required for consistent estimation of $\theta^\star$ in high-dimensional linear models. One simple assumption that can be imposed on a linear model is a *sparsity* assumption.

Definition 33 (Support of a vector)**.** *For a given* $\theta \in \mathbb{R}^d$*, the support of* θ *is defined as follows:*

$$\mathcal{S}(\theta) := \big\{ j \subseteq [d] : \theta_j \neq 0 \big\}.$$

Furthermore, define the set of vectors whose support is smaller than s *by*

$$\mathcal{B}_0(s) := \big\{ \theta \in \mathbb{R}^d : |\mathcal{S}(\theta)| \leq s \big\}.$$

Note. *In words, the ball* $\mathcal{B}_0(s)$ *contains all the vectors with no more than* s *non-zero elements. Note that this is not a ball in the traditional sense as the* ℓ_0 *norm is not a norm. The* ℓ_0-*norm is defined as*

$$\|\mathbf{x}\|_0 := \sum_{i=1}^{d} \mathbb{1}(x_i \neq 0)$$

and it is not a norm because it does not satisfy the triangle inequality. To see this take $d = 2$ *and* $\mathbf{x} = (1, 1)^\top$ *and* $\mathbf{y} = -\mathbf{x}$. *Then,* $\|\mathbf{x} + \mathbf{y}\|_0 = \|0\|_0 = 2$ *but* $\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 = 0$. *Indeed, because the* ℓ_0-*norm is not a norm,* $\mathcal{B}_0(\cdot)$ *is not compact, indeed it is an unbounded set (e.g., if* $d = 2$ *it contains the whole* x *and* y *axes).* ♦

Note. *Typically,* $\mathcal{S}(\theta^\star)$ *is unknown, but often assume that* $|\mathcal{S}(\theta^\star)| \leq s$, *where* s *is the sparsity level. Note that for known* $\mathcal{S}(\theta^\star)$, *the condition* $n \geq s$ *is sufficient for obtaining a unique solution to*

$$\arg\min_{\theta_s \in \mathcal{S}(\theta^\star)} \|\mathbf{y} - \mathbf{X}_s \theta_s\|_2^2$$

where $\mathbf{X}_s = (\mathbf{x}_{1,\mathcal{S}}^\top, \mathbf{x}_{2,\mathcal{S}}^\top, \cdots, \mathbf{x}_{n,\mathcal{S}}^\top)^\top \in \mathbb{R}^{n \times s}$, *and* $\mathbf{x}_{i,\mathcal{S}}$ *denotes the* s*-length vector that contains* $x_{i,j}$ *where* $j \in \mathcal{S}$. *However, in the case where* $\mathcal{S}(\theta^\star)$ *is unknown, more than* s *samples are required to ensure consistent estimation. In later lectures we will show that in fact, under certain conditions,* $O(s \log(d/s))$ *samples are required to achieve consistent estimation. These samples are more than* s, *but still way less than* d. *The regime we will study can be summarized as follows:*

$$s \text{ (sparsity level)} \ll n \text{ (sample size)} \ll d \text{ (ambient dimension)}.$$

♦

Under a sparsity assumption, we assume that $\theta^\star \in \mathcal{B}_0(s)$ for some s. In words, we are assuming that the ground truth is sparse.

5.1 Convex Relaxation of Sparsity Constraint and Basis Pursuit

Before we delve into the setting with noise (i.e. $\mathbf{y} = \mathbf{X}\theta^\star + \mathbf{w}$), we first consider the noiseless setting. In this setting we assume the true model to be

$$\mathbf{y} = \mathbf{X}\theta^\star \in \mathbb{R}^n,$$

where $\theta^\star \in \mathbb{R}^d$ is s-sparse ($|\mathcal{S}(\theta^\star)| \leq s$).

Notice that when $d > n$, there are infinitely many solutions that satisfy $\mathbf{y} = \mathbf{X}\theta$. In particular, all $\theta \in \theta^\star + \text{Null}(\mathbf{X})$ are valid solutions, where $\text{Null}(\mathbf{X}) := \left\{\Delta \in \mathbb{R}^d : \mathbf{X}\Delta = 0\right\}$ is the null space of $\mathbf{X}$. The set of feasible solutions would be $\theta^\star + \text{Null}(\mathbf{X})$ and, typically, $\text{Null}(\mathbf{X})$ will be a dense set.[4]

We can first consider the problem of minimizing the ℓ_0-norm

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0, \qquad \text{s.t.} \quad \mathbf{y} = \mathbf{X}\theta \tag{5.1}$$

Unfortunately, it is computationally challenging to solve this optimization problem due to the fact that the **objective function is non-convex**. For example, let $\hat{\theta}$ be the minimizer of problem (5.1). Now, one can consider searching over all possible subsets $S \subseteq [d]$, and checking

[4] Let (X, d) be a metric space. A set $Y \subseteq X$ is called dense in X if for every $x \in X$ and every $\varepsilon > 0$, there exists $y \in Y$ such that $d(x, y) < \varepsilon$.

whether $\mathbf{y} = \mathbf{X}_S\theta_S$. However, if $\|\hat{\theta}\|_0 = s$, this requires checking at least $\sum_{k=1}^{s}\binom{d}{k} \approx d^s$ subsets before finding the optimal solution $\hat{\theta}$, which is computationally expensive. To address this issue, we relax the sparsity constraint to convert this problem into a convex optimization problem. We can consider the following ℓ_1-norm optimization problem instead.

$$\min_{\theta\in\mathbb{R}^d} \|\theta\|_1, \qquad \text{s.t.} \quad \mathbf{y} = \mathbf{X}\theta \tag{5.2}$$

This problem is called *basis pursuit* and was introduced in Chen, Donoho and Saunders (1998, 2001), Tibshirani (1996).[5] Solutions to this problem have been studied by numerous researchers over the decades; the basis pursuit problem is a **convex problem**, thus it can be re-formulated as a linear programming problem, which can be solved efficiently.

Note. *The use of ℓ_1-norm is justified by the fact that ℓ_p-norms are convex as long as $p \geq 1$ and the ℓ_1-norm is the closest one to the ℓ_0-norm.* ♦

5.1.1 A Sufficient Condition for Exact Recovery in the Noiseless Setting

Now that we have relaxed our sparsity constraint, a natural question that arises is the following: under what conditions can we obtain $\theta^\star$ from solving problem (5.2) instead of problem (5.1). In particular, we aim to address the following question: what are sufficient (or necessary) conditions such that we have exact recovery, that is

$$\arg\min_{\theta\in\mathbb{R}^d}\{\|\theta\|_1 : \mathbf{y} = \mathbf{X}\theta\} = \theta^\star.$$

Formally, fix $\theta^\star, \mathcal{S}(\theta^\star) \equiv \mathcal{S}$. What are the conditions on $\mathbf{X} \in \mathbb{R}^{n\times d}$ under which

$$\widehat{\theta} := \arg\min_{\theta}\{\|\theta\|_1 : \mathbf{X}\theta^\star = \mathbf{X}\theta\} = \theta^\star, \tag{5.3}$$

or, equivalently,

$$\forall\ \theta \in \theta^\star + \text{Null}(\mathbf{X}) \setminus \{0\}, \quad \|\theta\|_1 > \|\theta^\star\|_1, \tag{5.4}$$

where the ℓ_1-norms of θ and $\theta^\star$ are only equal if $\theta = \theta^\star$. Note that condition (5.4) is ruling out all the other possible solutions by requiring them to have a higher loss than the ground truth $\theta^\star$. To analyze this problem more concretely, we can study the tangent cone of the ℓ_1-ball at $\theta^\star$. The *tangent cone* of the ℓ_1-ball at $\theta^\star$ is given by

$$\mathcal{T}(\theta^\star) := \{\Delta \in \mathbb{R}^d : \|\theta^\star + t\Delta\|_1 \leq \|\theta^\star\|_1 \text{ for some } t > 0\}, \tag{5.5}$$

where Δ characterizes the direction and t the scale of the vectors in the set. We can see that the tangent cone characterizes all the "tilts" to $\theta^\star$ in the ambient space ($\mathbb{R}^d$) that would yield a loss not larger than $\|\theta^\star\|_1$. Intuitively, we would like none of such "titls" to belong to the nullspace of $\mathbf{X}$ and, thus, be a solution of (5.2). In other words, we would like $\theta^\star + \mathcal{T}(\theta^\star)$

[5] The pre-print version of these papers appeared in 1994. Lasso is the acronym for *least absolute shrinkage and selection operator*.

to intersect with $\theta^\star + \text{Null}(\mathbf{X})$ only at $\theta^\star$. Then, condition (5.4) is equivalent to both the following conditions

$$\theta^\star + \text{Null}(\mathbf{X}) \cap \theta^\star + \mathcal{T}(\theta^\star) = \{\theta^\star\} \tag{5.6}$$

$$\text{Null}(\mathbf{X}) \cap \mathcal{T}(\theta^\star) = \{0\}. \tag{5.7}$$

Notice that conditions (5.6) and (5.7) are equivalent since $\theta^\star + \text{Null}(\mathbf{X})$ is an affine space passing through $\theta^\star$. The following diagram in 2-dimensions aims to provide some intuition.

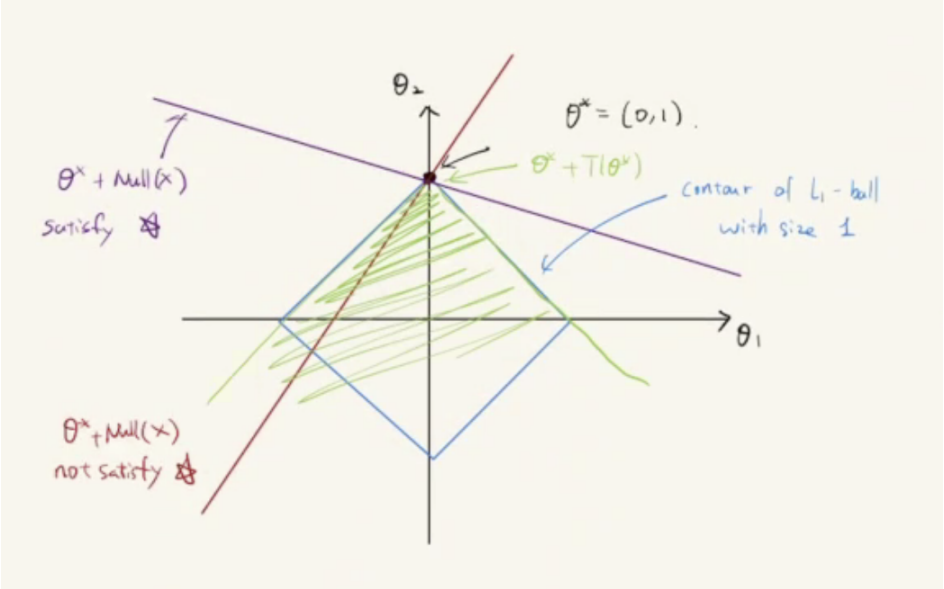


Figure 2: *In this figure,* $d = 2$, $n = 1$, *and* $\dim(\text{Null}(\mathbf{X})) = 1$. *The purple line gives the favorable case in which the set* $\theta^\star + \text{Null}(\mathbf{X})$ *only intersects the tangent cone (green lines) at* $\theta^\star$. *The red line gives the unfavorable case in which the set* $\theta^\star + \text{Null}(\mathbf{X})$ *passes through the tangent cone.*

The cone $\mathcal{T}(\theta^\star)$ allows us to characterize that portion of the space that should not be in the feasible space of solutions. Indeed, whenever the feasible space intersects with the interior of $\mathcal{T}(\theta^\star)$ the minimizer of $\|\mathbf{y} - \mathbf{X}\theta\|_1 \neq \theta^\star$.

Let $\Delta_{\mathcal{S}} = (\Delta_i)_{i\in\mathcal{S}}$ and $\Delta_{\mathcal{S}^c}$ is defined similarly. In general, if $\mathcal{S}(\theta^\star) = \mathcal{S} \subseteq [d]$, then

$$\mathcal{T}(\theta^\star) = \{\Delta \in \mathbb{R}^d : \|\Delta_{\mathcal{S}^c}\|_1 \leq \|\Delta_{\mathcal{S}}\|_1,\ \Delta_i\theta_i^\star \leq 0,\ \forall\, i \in \mathcal{S}\} \tag{5.8}$$

is equivalent to (5.5). The following example shows this in a particular case. We won't give a formal proof of this equivalence.

Example 49. *In Figure 2, we have* $d = 2, \mathcal{S} = \{2\}$, *and* $\theta^\star = (0,1)^\top$. *The equation of the cone is then given by*

$$\begin{aligned}\mathcal{T}(\theta^\star) &= \{\Delta \in \mathbb{R}^2 : \|(0,1) + t\Delta\|_1 \leq \|(0,1)\|_1,\ \textit{for some } t > 0\} \\ &= \{(\Delta_1, \Delta_2) : |t\Delta_1| + |1 + t\Delta_2| \leq 1,\ \textit{for some } t > 0\} \\ &= \{(\Delta_1, \Delta_2) : |\Delta_1| \leq |\Delta_2|,\ \Delta_2 \leq 0\}.\end{aligned}$$

♣

Notice that $\mathcal{T}(\theta^\star)$ depends on both the support $\mathcal{S}(\theta^\star)$ and $\text{sign}(\theta_i^\star)$ for all $i \in \mathcal{S}(\theta^\star)$.

The intuition from Figure 2 leads us to a sufficient condition for exact recovery. Define

$$\mathcal{C}(\mathcal{S}) := \{\Delta \in \mathbb{R}^d : \| \underbrace{\Delta_{\mathcal{S}^c}}_{\in \mathbb{R}^{|\mathcal{S}^c|}} \|_1 \leq \| \underbrace{\Delta_{\mathcal{S}}}_{\in \mathbb{R}^{|\mathcal{S}|}} \|_1\} \tag{5.9}$$

$\mathcal{C}(\mathcal{S})$ is a superset of $\mathcal{T}(\theta^\star)$ for $\mathcal{S}(\theta^\star) = \mathcal{S}$, so a sufficient condition for exact recovery is

$$\mathcal{C}(\mathcal{S}) \cap \text{Null}(\mathbf{X}) = \{0\},$$

since this condition implies conditions (5.6) and (5.7). Therefore,

$$\begin{aligned} \mathcal{T}(\theta^\star) \cap \text{Null}(\mathbf{X}) = \{0\}., & \qquad \text{(necessary condition)} \\ \mathcal{C}(\mathcal{S}) \cap \text{Null}(\mathbf{X}) = \{0\}. & \qquad \text{(sufficient condition)} \end{aligned}$$

Now, we can define a condition on $\mathbf{X}$ known as the *restricted null space property*, which will lead us to a theorem giving a precise statement for when condition (5.3) is satisfied.

Definition 34. *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and let $\mathcal{S} \subseteq [d]$. Then, $\mathbf{X}$ satisfies the **Restricted Nullspace Property** with respect to $\mathcal{S}$ (abbreviated $\text{RN}(\mathcal{S})$), if*

$$\mathcal{C}(\mathcal{S}) \cap \text{Null}(\mathbf{X}) = \{0\}.$$

Theorem 14. *The following two statements are equivalent:*

1. *For all $\theta^\star \in \mathbb{R}^d$ with $\mathcal{S}(\theta^\star) = \mathcal{S}$,*

$$\theta^\star = \widehat{\theta} := \arg\min_\theta \{\|\theta\|_1 : \mathbf{y} = \mathbf{X}\theta\}$$

   where $\theta^\star$ is the unique minimizer.

2. *$\mathbf{X}$ satisfies the $\text{RN}(\mathcal{S})$, i.e.*

$$\text{Null}(\mathbf{X}) \cap \mathcal{C}(\mathcal{S}) = \{0\}.$$

Proof. First, we show that (2) implies (1). Assume that $\mathbf{X}$ satisfies the restricted null space property. Let $\widehat{\theta} \in \arg\min_\theta\{\|\theta\|_1 : \mathbf{y} = \mathbf{X}\theta\}$. Define $\widehat{\Delta} := \widehat{\theta} - \theta^\star$ to be the error vector. Note that $\widehat{\Delta} \in \text{Null}(\mathbf{X})$ because both $\widehat{\theta}$ and $\theta^\star$ are in the feasible space. Since $\widehat{\Delta} \in \text{Null}(\mathbf{X})$, it suffices to show that $\widehat{\Delta} \in \mathcal{C}(\mathcal{S})$. We have the following sequence of equations:

$$\begin{aligned} \|\theta^\star_{\mathcal{S}}\|_1 &= \|\theta^\star\|_1 && \text{(sparsity of } \theta^\star\text{)} \\ &\geq \|\widehat{\theta}\|_1 && (\widehat{\theta} \text{ is optimal}) \\ &= \|\theta^\star + \widehat{\Delta}\|_1 && \text{(definition of } \widehat{\Delta}\text{)} \\ &= \|\theta^\star_{\mathcal{S}} + \widehat{\Delta}_{\mathcal{S}}\|_1 + \|\widehat{\theta}^\star_{\mathcal{S}_c} + \widehat{\Delta}_{\mathcal{S}^c}\|_1 \\ &= \|\theta^\star_{\mathcal{S}} + \widehat{\Delta}_{\mathcal{S}}\|_1 + \|\widehat{\Delta}_{\mathcal{S}^c}\|_1 && \text{(sparsity of } \theta^\star\text{)} \end{aligned}$$

$$\geq \|\theta^\star_{\mathcal{S}}\|_1 - \|\widehat{\Delta}_{\mathcal{S}}\|_1 + \|\widehat{\Delta}_{\mathcal{S}^c}\|_1. \qquad \text{(reverse triangle inequality)}$$

Rearranging the last inequality implies that $\|\widehat{\Delta}_{\mathcal{S}^c}\|_1 \leq \|\widehat{\Delta}_{\mathcal{S}}\|_1$ and thus we have shown that $\widehat{\Delta} \in \mathcal{C}(\mathcal{S}) \cap \text{Null}(\mathbf{X})$. By our assumption, this implies that $\widehat{\Delta} = 0$ and $\widehat{\theta} = \theta^\star$.

Now, assume (1) and we will show (2). Let $\tilde{\theta} \in \text{Null}(\mathbf{X}) \setminus \{0\}$. It suffices to show that $\tilde{\theta} \notin \mathcal{C}(\mathcal{S})$. Let $\theta^\star = \begin{pmatrix}\tilde{\theta}_{\mathcal{S}} & 0\end{pmatrix}^\top$ be $\mathcal{S}$-sparse. By (a), we know that

$$\arg\min_{\beta} \left\{ \|\beta\|_1 : \mathbf{X}\beta = \mathbf{X}\begin{pmatrix}\tilde{\theta}_{\mathcal{S}} \\ 0\end{pmatrix} \right\} = \begin{pmatrix}\tilde{\theta}_{\mathcal{S}} \\ 0\end{pmatrix}.$$

and this is a unique minimizer. Since $\tilde{\theta} \in \text{Null}(\mathbf{X})$ so

$$\mathbf{X}\begin{pmatrix}0 \\ -\tilde{\theta}_{\mathcal{S}^c}\end{pmatrix} = \mathbf{X}\begin{pmatrix}\tilde{\theta}_{\mathcal{S}} \\ 0\end{pmatrix},$$

this implies that

$$\left\| \begin{pmatrix}\tilde{\theta}_{\mathcal{S}} \\ 0\end{pmatrix} \right\|_1 < \left\| \begin{pmatrix}0 \\ -\tilde{\theta}_{\mathcal{S}^c}\end{pmatrix} \right\|_1,$$

which in turn implies that $\|\tilde{\theta}_{\mathcal{S}}\|_1 < \|\tilde{\theta}_{\mathcal{S}^c}\|_1$. Thus, $\tilde{\theta} \notin \mathcal{C}(\mathcal{S})$. ■

In summary, a matrix $\mathbf{X}$ satisfying the restricted nullspace property with respect to $\mathcal{S}(\theta^\star)$ is a sufficient condition for the exact recovery of $\hat{\theta}$, which solves problem (5.1) (the basis pursuit problem). In later lectures, we will show that random matrices with certain distributions satisfy the restricted null space property as long as $n \gtrsim s\log(d/s)$.

5.2 Sufficient Conditions for RN($\mathcal{S}$)

We now ask ourselves what are the sufficient conditions on our design matrix $\mathbf{X}$ to ensure that it is RN($\mathcal{S}$). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$. If the quantity

$$\sup_{v\in\mathcal{C}(\mathcal{S}),\|v\|_2=1} \left| \left\langle v, \left(\frac{\mathbf{X}^\top\mathbf{X}}{n} - \mathbf{I}_d \right) v \right\rangle \right|,$$

is small, then

$$\frac{1}{n}\|\mathbf{X}v\|_2^2 = \frac{1}{n}\langle v, \mathbf{X}^\top\mathbf{X}v\rangle \approx \|v\|_2^2 > 0, \quad \forall\, v \in \mathcal{C}(\mathcal{S})$$

which implies

$$\text{Null}(\mathbf{X}) \cap \mathcal{C}(\mathcal{S}) = \{0\}.$$

So we want $\mathbf{\Gamma} = (\mathbf{X}^\top\mathbf{X} - \mathbf{I}_d)$ to be small in some sense. We are going to propose two alternative ways of looking at $\mathbf{\Gamma}$. As such, we will have two different sufficient conditions to guarantee that $\mathbf{X}$ is RN($\mathcal{S}$). let's first define two measures of how small $\mathbf{\Gamma}$ is:

1. **Pairwise incoherence**

$$\delta_{pw}(\mathbf{X}) := \sup_{i\neq j} |\mathbf{\Gamma}| = \sup_{i\neq j} \left| \frac{1}{n}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_d \right|.$$

2. **Restricted isometry constant**

$$\delta_s(\mathbf{X}) := \max_{|\mathcal{S}| \leq s} \left\| \frac{1}{n} \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} - \mathbf{I}_d \right\|_{op}.$$

In general, it is true that

$$\delta_{pw}(\mathbf{X}) \leq \delta_s(\mathbf{X}) \leq s \cdot \delta_{pw}(\mathbf{X}), \quad \forall\, s \in [d].$$

Pairwise incoherence is harder to prove.

Proposition 42 (Sufficient Condition for RN($\mathcal{S}$))**.** $\mathbf{X}$ *satisfies* $\mathrm{RN}(\mathcal{S}), \forall\, |\mathcal{S}| \leq s, s \in [d]$ *if any of the following holds*

1. $\delta_{pw}(\mathbf{X}) < \frac{1}{3s}$,
2. $\delta_{2s} < \frac{1}{3}$.

5.3 Noisy Linear Models

We now consider a linear model of the form

$$\mathbf{y} = \mathbf{X}\theta^\star + \mathbf{w},$$

with $\mathbf{w} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times d}$. The typical goal is to fin $\widehat{\theta}$ such that $\|\widehat{\theta} - \theta^\star\|_2$ is small. When $n > d$ we can simply use OLS, but when $n < d$, LASSO is a natural choice:

$$\widehat{\theta} := \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_1. \qquad (\lambda \text{ form})$$

The problem above can be expressed in two equivalent ways

$$\widehat{\theta} := \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\theta\|_2^2, \quad \text{s.t. } \|\theta\|_1 \leq R \qquad (\ell_1\text{−norm})$$

and

$$\widehat{\theta} := \arg\min_{\theta \in \mathbb{R}^d} \|\theta\|_1, \quad \text{s.t. } \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 \leq b^2. \qquad (\text{error form})$$

Definition 35 (Restricted Eigenvalue Property)**.** *We say that* $\mathbf{X}$ *satisfies* $\mathrm{RE}(\mathcal{S}, (k, \alpha))$ *if*

$$\frac{1}{n} \|\mathbf{X}\Delta\|_2^2 \geq k \|\Delta\|_2^2, \quad \forall\, \Delta \in \mathcal{C}_\alpha(\mathcal{S}),$$

where

$$\mathcal{C}_\alpha(\mathcal{S}) = \{\Delta \in \mathbb{R}^d : \alpha \|\Delta_{\mathcal{S}}\|_1 \geq \|\Delta_{\mathcal{S}^c}\|_1\}.$$

Note. *The name of* RE *comes from the fact that it implies that all the eigenvalues of* $\mathbf{X}$ *are larger than* k. ♦

Corollary 6. *If* $\alpha = 1, k > 0$, *then* RE *implies* RN.

Proposition 43. *If* $|\mathcal{S}(\theta^\star)| \equiv \mathcal{S}$ *and* $|\mathcal{S}| = s$ *and* $\mathbf{X}$ *satisfies* $\mathrm{RE}(\mathcal{S}, (k, 3))$, *then:*

1. $\lambda-$*form. If* $\lambda_n \geq 2\|n^{-1}\mathbf{X}^\top \mathbf{w}\|_\infty$ *then*

$$\|\widehat{\theta} - \theta^\star\|_2 \leq \frac{3}{k}\sqrt{s}\lambda_n.$$

2. ℓ_1-*norm form. Take* $R = \|\theta^\star\|_1$. *then*

$$\|\widehat{\theta} - \theta^\star\|_2 \leq \frac{4}{k}\sqrt{s}\left\|\frac{\mathbf{X}^\top \mathbf{w}}{n}\right\|_\infty.$$

3. *Error form. If* $b^2 \geq \|\mathbf{w}\|_2^2/2n$, *then*

$$\|\widehat{\theta} - \theta^\star\|_2 \leq \frac{4}{k}\sqrt{s}\left\|\frac{\mathbf{X}^\top \mathbf{w}}{n}\right\|_\infty + \frac{2}{\sqrt{k}}\sqrt{b^2 - \frac{\|\mathbf{w}\|_2^2}{n}}.$$

References

Cantelli, F. P. (1933), 'Sulla determinazione empirica delle leggi di probabilita', Giorn. Ist. Ital. Attuari **4**(421-424).

Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001), 'Atomic decomposition by basis pursuit', SIAM review **43**(1), 129–159.

Chen, S. S., Donoho, L. D. and Saunders, M. A. (1998), 'Atomic decomposition by basis pursuit', SIAM journal on scientific computing **20**(1), 33–61.

Efron, B. and Stein, C. (1981), 'The jackknife estimate of variance', The Annals of Statistics pp. 586–596.

Glivenko, V. (1933), 'Sulla determinazione empirica delle leggi di probabilita', Gion. Ist. Ital. Attauri. **4**, 92–99.

Sauer, N. (1972), 'On the density of families of sets', Journal of Combinatorial Theory, Series A **13**(1), 145–147.

Shelah, S. (1972), 'A combinatorial problem; stability and order for models and theories in infinitary languages', Pacific Journal of Mathematics **41**(1), 247–261.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288.

Vapnik, V. N. and Chervonenkis, A. Y. (1971), 'On the uniform convergence of relative frequencies of events to their probabilities', Theory of Probability and its Applications **16**, 264–280.

Wainwright, M. J. (2019), High-dimensional statistics: A non-asymptotic viewpoint, Vol. 48, Cambridge University Press.