Theoretical Statistics

Lectures & Homeworks

Reece D. Huff
rdhuff@berkeley.edu

May 19th, 2024

STAT210B Theoretical Statistics
Nikita Zhivotovskiy
Department of Statistics
University of California, Berkeley

Contents

Lecture 1: The Non-Asymptotic Approach

Instructor: Nikita Zhivotovskiy *Scriber: Nikita Zhivotovskiy* *Proofreader: Nikita Zhivotovskiy*

1.1 Some Limitations of the Asymptotic Approach

This course focuses on non-asymptotic analysis in statistics. To illustrate the significance of the asymptotic approach, consider estimating the mean of a random variable. Suppose we observe n independent identically distributed random variables $X_1, \dots, X_n$ with mean μ and (for simplicity, known) variance σ^2. We can construct a confidence interval of the form

$$\left[-\frac{2\sigma}{\sqrt{n}} + \frac{1}{n}\sum_{i=1}^{n} X_i, \frac{2\sigma}{\sqrt{n}} + \frac{1}{n}\sum_{i=1}^{n} X_i \right],$$

which, according to the Central Limit Theorem (CLT), will asymptotically contain the true mean with a probability of at least 0.95. The constant 2 in the above bound comes from the quantiles of the Gaussian distribution, its limiting distribution. For example, if Z is a standard normal random variable and Φ its cumulative distribution function, then from the CLT

$$\lim_{n\to\infty} \Pr\left(\left| \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}} \right| \geq t \right) = \Pr(|Z| \geq t) = 2\Phi(-t).$$

And $1 - 2\Phi(-2) \approx 0.9545$. However, this approach to constructing confidence intervals has a significant flaw. It overly focuses on the scenario where the number of observations n goes to infinity. Furthermore, the expression is somewhat sensitive to the distribution; if X_i are Gaussian, then $\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma\sqrt{n}}$ is a standard normal random variable for any sample size n. For other distributions, the CLT might only apply when the sample size is quite large. One key piece of evidence for this comes from the Poisson limit theorem. Let's recall its statement.

Theorem 1.1 (Poisson Limit Theorem)**.** *Let $X_{n,i}$, $1 \leq i \leq n$, be independent random variables with $X_{n,i} \sim \mathrm{Ber}(p_n)$. Assume that for some $\lambda > 0$,*

$$\lim_{n\to\infty} np_n \to \lambda.$$

Then,

$$\sum_{i=1}^{n} X_{n,i} \xrightarrow{d} Pois(\lambda) \quad as\ n \to \infty.$$

Consider observing $n = 50$ independent Bernoulli distributions with $p = 1/50$. Which approximation would be more appropriate: Poisson or the CLT? How do we choose the right limiting theorem in each particular practical case where we only have access to a finite sample of data? The methodologies developed in this course will enable us to derive quantitative bounds that often depend on additional properties of the distribution, providing relevant quantitative results. Furthermore, we will also discuss strategies to achieve a form of universality, where minimal assumptions are made about the distribution, yet the results are akin to those obtained under some limiting distribution, such as the Gaussian distribution.

There are additional aspects where non-asymptotic bounds can be superior to asymptotic results. Consider a fundamental model, the Gaussian regression model with fixed design, which we will explore in more detail later. Take a dataset $\{(Y_i, x_i)\}_{i=1}^n$, where

$$Y_i = f(x_i) + \xi_i, \quad i = 1, \dots, n, \tag{1}$$

with $\mathcal{X}$ being some fixed set, $f : \mathcal{X} \to \mathbb{R}$ an unknown function, $x_i \in \mathcal{X}$ fixed nonrandom elements, and the random errors ξ_i independently and identically distributed (i.i.d.) Gaussian variables with mean zero and variance σ^2. The primary goal is to construct an estimator $\hat{f}$ for f using the observations $\{(Y_i, x_i)\}_{i=1}^n$. To evaluate the performance of $\widehat{f}$, we consider the squared error loss given by

$$\|\hat{f} - f\|_n^2 = \frac{1}{n} \sum_{i=1}^n \left(\widehat{f}(x_i) - f(x_i)\right)^2.$$

Moreover, we define the risk of the estimator $\hat{f}$ as $\mathbb{E}\|\hat{f} - f\|_n^2$. A straightforward, yet illustrative, case for this model is to assume that $x_i \in \mathbb{R}^d$ and $f(X_i) = \langle X_i, \theta^\star \rangle$, with $\theta^\star \in \mathbb{R}^d$ being an unknown target parameter. A common method to estimate $\theta^\star$ is through the least squares estimator:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (\langle x_i, \theta \rangle - Y_i)^2.$$

In this course, we will discuss and prove results of the form:

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left(\langle x_i, \widehat{\theta} \rangle - \langle x_i, \theta^\star \rangle\right)^2\right] \leq \frac{d\sigma^2}{n}.$$

Here d, n, σ^2 each play a specific role. The parameters d, n, and σ can grow simultaneously in any manner, a scenario somewhat problematic for an asymptotic approach. Typically, asymptotic results are either of the form where both d and σ are fixed and n goes to infinity, or σ is fixed and both d and n grow such that the ratio d/n converges to a constant. However, even in this simplified example, the asymptotic regimes described only cover a fraction of possible interactions between d, σ, and n. To elaborate further on this perspective, we refer to the quote of Pascal Massart [Mas07], one of the pioneers of non-asymptotic analysis in statistics:

"*Through our works, we have promoted a nonasymptotic approach in statistics which consists in taking the number of observations as it is and trying to evaluate the effect of all the influential parameters. At this very starting point, it seems to me that it is important to provide a first answer to the following question: why should we be interested by a nonasymptotic view for model selection at all? In my opinion, the motivation should neither be a strange interest for 'small' sets of data nor a special taste for constants and inequalities rather than for limit theorems (although since mathematics is also a matter of taste, it is a possible way for getting involved in it...). On the contrary, the nonasymptotic point of view may turn to be especially relevant when the number of observations is large. It is indeed to fit large complex sets of data that one needs to deal with possibly huge collections of models at different scales. The nonasymptotic approach for model selection precisely allows the collection of models together with their dimensions to vary freely, letting the dimensions be possibly of the same order of magnitude as the number of observations.*"

1.2 Basic tail bounds

We begin with some basic tail bounds.

Proposition 1.2 (Markov's Inequality)**.** *Let X be a non-negative random variable. Then for any $t \geq 0$, we have*

$$\Pr(X \geq t) \leq \frac{\mathbb{E}X}{t}.$$

Proposition 1.3 (Chebyshev's Inequality)**.** *Let X be a random variable with finite variance. Then for any $t \geq 0$, we have*

$$\Pr(|X - \mathbb{E}X| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proof. Chebyshev's inequality follows from applying Markov's inequality to the squared deviation $(X - \mathbb{E}X)^2$. □

Remark 1.4. *When discussing Chebyshev's inequality, it's important to give attention to the correct pronunciation of Chebyshev's name. Notably, the proper English transcription accentuates the final 'ov' in 'Chebyshov', a detail that is often missed.*

Is Chebyshev's inequality tight for the standard Gaussian random variable? The answer is no! Here is why.

Proposition 1.5 (Tails of Gaussian Random Variables)**.** *Let Z be a standard Gaussian random variable. For any $t \geq 0$, we have*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \Pr(Z \geq t) \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Proof. Consider Z, a standard Gaussian variable. For any $t \geq 0$, we have

$$\Pr(Z \geq t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx.$$

Setting $x = t + y$, we find

$$\Pr(Z \geq t) = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-t^2/2} e^{-ty} e^{-y^2/2}\, dy \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-ty}\, dy,$$

using $e^{-y^2/2} \leq 1$. The last integral equals $\frac{1}{t}$, yielding the upper bound. The lower bound follows from

$$\int_t^\infty (1 - 3x^{-4}) e^{-x^2/2}\, dx = \left(\frac{1}{t} - \frac{1}{t^3}\right) e^{-t^2/2}.$$

□

The crucial aspect is that the tail bound $\Pr(Z \geq t) \leq \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}$ is often the only necessary consideration for Gaussian variables in many applications. Chebyshev's inequality is not sufficient as it provides a bound of $\frac{1}{t^2}$. However, we can refine this using Markov's inequality, which, for any random variable X with mean μ, and any $t \geq 0$, gives

$$\Pr(|X - \mu| \geq t) = \Pr(|X - \mu|^p \geq t^p) \leq \frac{\mathbb{E}|X - \mu|^p}{t^p}.$$

By taking the infimum with respect to $p \geq 1$, we find

$$\Pr(|X - \mu| \geq t) \leq \inf_{p \geq 1} \frac{\mathbb{E}|X - \mu|^p}{t^p}.$$

This approach, while leading to the desired bound, may require more complex computations in many applications. An alternative and more practical method involves using the moment generating function (MGF). For any $\lambda > 0$,

$$\Pr(|X - \mu| \geq t) = \Pr\big(\exp(\lambda|X - \mu|) \geq \exp(\lambda t)\big) \leq \frac{\mathbb{E}\exp(\lambda|X - \mu|)}{\exp(\lambda t)}.$$

Thus,

$$\Pr(|X - \mu| \geq t) \leq \inf_{\lambda > 0} \frac{\mathbb{E}\exp(\lambda|X - \mu|)}{\exp(\lambda t)}.$$

For any random variable Z, its MGF is defined as

$$M_Z(\lambda) = \mathbb{E}\exp(\lambda Z).$$

A key advantage of the MGF over standard moments is that for independent random variables $Z_1, \ldots, Z_n$, for any $\lambda \in \mathbb{R}$,

$$M_{Z_1 + \ldots + Z_n}(\lambda) = \prod_{i=1}^{n} M_{Z_i}(\lambda),$$

a property not generally shared by moments. Under assumptions allowing the interchange of integrals and derivatives, we can show that

$$\frac{d}{d\lambda^k} M_Z(\lambda)\Big|_{\lambda=0} = \mathbb{E}\frac{d}{d\lambda^k}\exp(\lambda Z)\Big|_{\lambda=0} = \mathbb{E}Z^k \exp(\lambda Z)\Big|_{\lambda=0} = \mathbb{E}Z^k,$$

justifying the name 'moment generating function'. Finally, we calculate the MGF for a standard Gaussian. For Z, a zero-mean Gaussian random variable with variance σ^2, it holds that

$$M_Z(\lambda) = \mathbb{E}\exp(\lambda Z) = \exp\Big(\lambda^2\sigma^2/2\Big).$$

Lecture 2: Basic concentration inequalities

Instructor: Nikita Zhivotovskiy *Scriber: Nikita Zhivotovskiy* *Proofreader: Nikita Zhivotovskiy*

2.1 Sub-Gaussian random variables

We continue our discussion of the moment generating function. Recall that for Z, a zero-mean Gaussian random variable with variance σ^2, it holds that

$$M_Z(\lambda) = \mathbb{E}\exp(\lambda Z) = \exp\Big(\lambda^2\sigma^2/2\Big).$$

Let us apply Chernoff's method. Using the formula for the moment generating function of the Gaussian random variable, we have

$$\Pr(Z \geq t) \leq \inf_{\lambda>0} \exp\Big(\lambda^2\sigma^2/2 - \lambda t\Big) = \exp\Big(-t^2/2\sigma^2\Big).$$

This is quite remarkable as exponential moments alone are good enough to get the main term in the Gaussian tail. Our first definition of the sub-Gaussian random variable is the following:

Definition 2.1. *A mean-zero random variable X is sub-Gaussian with the variance parameter σ^2 if*

$$\mathbb{E}\exp(\lambda X) \leq \exp\Big(\lambda^2\sigma^2/2\Big),$$

for all $\lambda \in \mathbb{R}$.

What do we have immediately from this definition? For example, if the random variable X is sub-Gaussian, then $-X$ is also sub-Gaussian with the same variance parameter. The following result summarizes several equivalent definitions of sub-Gaussian random variables.

Proposition 2.2. *Let X be a random variable with $\mathbb{E}X = 0$. Then the following are equivalent, and the parameters $k_i > 0$ differ by at most multiplicative absolute constant factors:*

1. *For all $\lambda \in \mathbb{R}$,*
$$\mathbb{E}\exp(\lambda X) \leq \exp\Big(k_1^2\lambda^2\Big).$$
2. *For all $t \geq 0$,*
$$\Pr(|X| \geq t) \leq 2\exp\Big(-t^2/k_2^2\Big).$$
3. *For all $p \geq 1$,*
$$\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p} \leq k_3\sqrt{p}.$$
4. *For all λ such that $|\lambda| \leq 1/k_4$,*
$$\mathbb{E}\exp\Big(\lambda^2X^2\Big) \leq \exp\Big(k_4^2\lambda^2\Big).$$
5. *For some k_5,*
$$\mathbb{E}\exp\Big(X^2/k_5^2\Big) \leq 2.$$

Proof. We have already verified (1) → (2). This follows from Chernoff's method. Let us show how (2) implies (3). Without loss of generality, assume that $k_2 = 1$. We have

$$\mathbb{E}|X|^p = \int_{t=0}^{\infty} \Pr(|X|^p \geq t)dt = \int_{u=0}^{\infty} \Pr(|X| \geq u)pu^{p-1}du \leq \int_{u=0}^{\infty} 2pu^{p-1}\exp\Big(-u^2\Big)du.$$

Recall that the gamma function Γ is given by $\Gamma(x) = \int_0^{\infty} t^{x-1}\exp(-t)dt$. Using the change of variables $w = u^2$, we have

$$2\int_{u=0}^{\infty} pu^{p-1}\exp\Big(-u^2\Big)du = p\int_{w=0}^{\infty} w^{p/2-1}\exp(-w)dw = p\Gamma(p/2) \leq 3p(p/2)^{p/2},$$

where we used $\Gamma(x) \leq 3x^x$ for $x > 1/2$. The remaining proofs are left as an exercise and can be found in [Ver18, Proposition 2.5.2]. □

What other distributions are sub-Gaussian?
Consider the example of the Rademacher random sign. If $\varepsilon = \pm 1$ with equal probabilities, then we call it the Rademacher random variable. We have

$$\mathbb{E}\exp(\lambda\varepsilon) = \frac{1}{2}\exp(\lambda) + \frac{1}{2}\exp(-\lambda) \leq \exp\Big(\lambda^2/2\Big),$$

where the inequality follows from comparing Taylor series. Thus, the Rademacher random variable is sub-Gaussian with $\sigma = 1$. It appears that another basic example belongs to the class of bounded distributions.

Lemma 2.3 (Hoeffding's lemma)**.** *Assume that X is a zero-mean random variable whose values are almost surely in $[a, b]$. Then*

$$\mathbb{E}\exp(\lambda X) \leq \exp\Big(\lambda^2(b-a)^2/8\Big).$$

Proof. We show a slightly weaker inequality first using the idea of *symmetrization*. Let X' be an independent copy of X. Denote the expectation with respect to X' as $\mathbb{E}'$. Using Jensen's inequality (since the exponent function is convex) we have

$$\mathbb{E}\exp(\lambda X) = \mathbb{E}\exp(\lambda(X - \mathbb{E}'X')) \leq \mathbb{E}\mathbb{E}'\exp(\lambda(X - X')).$$

Let ε be a Rademacher random variable independent from both X and X'. We observe that by symmetry, $X - X'$ has the same distribution as $\varepsilon(X - X')$. Therefore, we have

$$\mathbb{E}\mathbb{E}'\exp(\lambda(X - X')) = \mathbb{E}\mathbb{E}'\mathbb{E}_{\varepsilon}\exp(\lambda\varepsilon(X - X')).$$

Conditioning on the values $X - X'$ we use the upper bound for the MGF of the Rademacher random variable to obtain

$$\mathbb{E}\mathbb{E}'\mathbb{E}_{\varepsilon}\exp(\lambda\varepsilon(X - X')) \leq \mathbb{E}\mathbb{E}'\exp\Big(\lambda^2(X - X')^2/2\Big).$$

The proof follows by observing that $(X - X')^2 \leq (b - a)^2$ with probability one. A slightly more refined analysis leads to the constant 8 in the bound. □

The next lemma is the standard concentration results for the sum of independent sub-Gaussian random variables. Its proof is a manifestation of why MGF is preferred over the moments: it is easy to work with MGF of independent random variables.

Proposition 2.4. *Assume that $X_1, \dots, X_n$ are independent random variables with means $\mu_1, \dots, \mu_n$ such that $X_i - \mu_i$ are sub-Gaussian with parameters σ_i for all $i = 1, \dots, n$. Then, for any $t \geq 0$,*

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\Big(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big).$$

Furthermore,

$$\Pr\left(\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right) \leq 2\exp\Big(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big).$$

Proof. Due to the independence of $X_i - \mu_i$, we have

$$\mathbb{E}\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mu_i)\right) = \prod_{i=1}^{n}\mathbb{E}\exp(\lambda(X_i - \mu_i)) \leq \exp\left(\lambda^2\sum_{i=1}^{n}\sigma_i^2/2\right).$$

Therefore, the sub-Gaussian parameter σ^2 of $\sum_{i=1}^{n}(X_i - \mu_i)$ is $\sum_{i=1}^{n}\sigma_i^2$. Thus, applying Chernoff's method and optimizing with respect to λ, we get the first inequality. Similarly, we can prove that

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mu_i) \leq -t\right) \leq \exp\Big(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big).$$

Combining both inequalities via the union bound, we obtain

$$\Pr\left(\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right) \leq 2\exp\Big(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big).$$

□

As a corollary for the bounded random variables, we have the classical Hoeffding's inequality.

Proposition 2.5 (Hoeffding's inequality)**.** *Assume that $X_1, \dots, X_n$ are random variables taking their values in $[a_i, b_i]$ respectively with means $\mu_1, \dots, \mu_n$. Then, for any $t \geq 0$, we have*

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

Furthermore,

$$\Pr\left(\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

Proof. We just replace σ_i in Proposition 2.4 with $|b_i - a_i|/2$, which is due to Hoeffding's lemma. □

Lecture 3: Sub-Gaussian and Sub-Exponential Distributions

Instructor: Nikita Zhivotovskiy *Scriber: Annie Ulichney* *Proofreader: João Vitor Romano*

In the previous lecture, we introduced sub-Gaussian distributions and equivalent characterizations of sub-Gaussian distributions. We showed that Rademacher variables and bounded random variables with mean zero are examples of sub-Gaussian variables. Finally, we introduced Hoeffding's inequality which provides a bound for the deviation of the sum of bounded independent random variables from its expectation. We pick up where we left off: with Hoeffding's inequality.

3.1 Hoeffding's inequality

Recall from the previous lecture Hoeffding's inequality, which is sometimes called the *Hoeffding bound*.

Proposition 3.1 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables such that $\mathbb{E}X_i = \mu_i$ and $X_i \in [a_i, b_i]$ almost surely. Then, for all $t \geq 0$,*

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right).$$

By symmetry and the union bound, it follows that

$$\Pr\left(\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right).$$

This result was proved in Lecture 2, Proposition 5.

Remark 3.2. *Observe that these bounds are a function of the length of the interval $[a_i, b_i]$, so they are invariant to centering X_i.*

3.1.1 Example 1: Rademacher Random Variables

Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables, i.e. $\varepsilon_i = \pm 1$ with probability $1/2$. Observe $\mathbb{E}\varepsilon_i = 0, \mathrm{Var}(\varepsilon_i) = 1$. Applying Hoeffding's inequality, we get

$$\begin{aligned}\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right| \geq t\right) &\leq 2\exp\left(\frac{-2(nt)^2}{4n}\right) \\ &= \underbrace{2\exp\left(\frac{-nt^2}{2}\right)}_{\delta}.\end{aligned}$$

We seek t such that, with high probability, $\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right| \leq t$. In other words, we want to bound δ. We do so by rearranging the relation $\delta = 2\exp\left(\frac{-nt^2}{2}\right)$ to isolate t, which yields $t = \sqrt{\frac{2\log(2/\delta)}{n}}$. After expressing t in terms of δ, we can interpret our bound as a *high-probability bound* as follows. For this value of t, with probability at least $1 - \delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right| \leq \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Remark 3.3. *Observe that this result is true for any n; it is a non-asymptotic result.*

Now, let's compare this bound to that from Chebyshev's inequality. Applying Chebyshev to the random variable $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$, for all $t > 0$,

$$\begin{aligned}\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right| \geq t\right) &\leq \frac{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)}{t^2} \\ &= \frac{1/n^2 \cdot n}{t^2} \\ &= \underbrace{\frac{1}{nt^2}}_{\delta}.\end{aligned}$$

Now, we have bounded the probability that our random variable exceeds t by δ. We can equivalently express this bound as a high-probability bound by solving for t in terms of δ to get $t = \frac{1}{\sqrt{\delta n}}$. As before, we can make the interpretation that, with probability $1-\delta$, $\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right| \leq \frac{1}{\sqrt{\delta n}}$.
We have now explored an application of Hoeffding's inequality to Rademacher random variables and compared the resulting bound to that of Chebyshev. Next, we return to the discussion of sub-Gaussian distributions in general and introduce the notion of the *sub-Gaussian norm* and its properties.

3.2 Sub-Gaussian Norm

Definition 3.4 (Sub-Gaussian Norm)**.** *Let X be a random variable (not necessarily such that $\mathbb{E}X = 0$). The Sub-Gaussian norm of X is*

$$\|X\|_{\psi_2} = \inf\left\{t \geq 0 : \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\}.$$

Remark 3.5. *Essentially, this norm is the smallest constant k_5 in relation 5 of our 5 equivalent definitions of sub-Gaussian random variables listed in Lecture 2, Proposition 2.*

Proposition 3.6 (Sub-Gaussian norm equivalent characterizations)**.** *As in Lecture 2, Proposition 2, for all $t \geq 0$, if c and C are absolute constants, for a random variable X the following are equivalent:*

1. $\Pr(|X| \geq t) \leq 2\exp\left(-ct^2/\|X\|_{\psi_2}^2\right)$;
2. $\mathbb{E}\exp\left(X^2/\|X\|_{\psi_2}^2\right) \leq 2$;
3. $\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p} \leq C\|X\|_{\psi_2}\sqrt{p} \quad \forall p \geq 1$;
4. $\mathbb{E}X = 0 \implies \mathbb{E}\exp(\lambda X) \leq \exp\left(C\lambda^2\|X\|_{\psi_2}^2\right) \quad \forall \lambda \in \mathbb{R}$.

Remark 3.7. *Note that definitions 1-3 do not require $\mathbb{E}X = 0$.*

Proof of Proposition 3.6 can be found in the proof of [Ver18, Proposition 2.5.2].

3.2.1 Properties of the Sub-Gaussian Norm

Proposition 3.8. $\|\cdot\|_{\psi_2}$ *is a valid norm.*

To prove this, we need to show that, for a random variable X and $\lambda \in \mathbb{R}$, the following are satisfied:

1. $\|\lambda X\|_{\psi_2} = |\lambda| \|X\|_{\psi_2}$;
2. $\|X + Y\|_{\psi_2} \leq \|X\|_{\psi_2} + \|Y\|_{\psi_2}$;
3. $\|X\|_{\psi_2} = 0 \iff X = 0$ almost surely.

We prove these properties in Homework 1.

Proposition 3.9. *For the normal distribution, the following properties hold:*

1. $Z \sim N(0,1) \implies \|Z\|_{\psi_2} \leq c$;
2. $Z \sim N(0,\sigma^2) \implies \|Z\|_{\psi_2} \leq c\sigma$.

Proof. Suppose $Z \sim N(0,\sigma^2)$. The MGF of the normal distribution is

$$\mathbb{E}\exp(\lambda Z) = \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

We can see that Z is sub-Gaussian with variance parameter σ^2. The desired result follows from relation 4 of Proposition 3.6. □

Proposition 3.10.

If $X \in [a,b]$, $\exp\left(X^2/t^2\right) \leq \exp\left(\max(a^2,b^2)/t^2\right)$ *which implies that* $t = \frac{1}{\sqrt{\log(2)}}\sqrt{\max(a^2,b^2)}$.

Proposition 3.11. *If* $X_1, \ldots, X_n$ *are independent random variables and* c *is an absolute constant,*

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

Proof. Without loss of generality, assume $\mathbb{E}X_i = 0$. Then, by our independence assumption and Properties 1, 4 of Proposition 3.6,

$$\begin{aligned}
\mathbb{E}\exp\left(\lambda\sum_{i=1}^n X_i\right) &= \prod_{i=1}^n \mathbb{E}\exp\left(\lambda X_i\right) \\
&\leq \prod_{i=1}^n \exp\left(c\lambda^2\|X_i\|_{\psi_2}^2\right) \\
&= \exp\left(\lambda^2 c\sum_{i=1}^n \|X_i\|_{\psi_2}^2\right).
\end{aligned}$$

□

Remark 3.12. *Note that this is an analog of the property* $\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i)$ *for independent* X_i.

Remark 3.13. *Observe that Claim 3.11 is not a result of the triangle inequality. Our claim assumes independence and is a stronger result.*

3.2.2 Example 2: Khintchine's inequality

Khintchine's inequality can be derived with Hoeffding's inequality to bound the L_p norms of the sum of independent random variables.

Proposition 3.14. *Let $X_1, \dots, X_n$ be independent sub-Gaussian random variables where $\mathbb{E}X_i = 0$ and $\mathrm{Var}(X_i) = 1$. Let $a = (a_1, \dots, a_n) \in \mathbb{R}^n$. Then, for $p \geq 2$,*

$$\left(\sum_{i=1}^n a_i^2\right)^{1/2} \leq \left\|\sum_{i=1}^n a_i X_i\right\|_p \leq cK\sqrt{p}\left(\sum_{i=1}^n a_i^2\right)^{1/2},$$

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

Proof. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables and let $a = (a_1, \dots, a_n) \in \mathbb{R}^n$. How does our vector a correlate with random coin flips? By our assumption of independence and Claim 3.11,

$$\begin{aligned}\left\|\sum_{i=1}^n a_i\varepsilon_i\right\|_{\psi_2}^2 &\leq C\sum_{i=1}^n \|a_i\varepsilon_i\|_{\psi_2}^2 \\ &= C\sum_{i=1}^n a_i^2\|\varepsilon_i\|_{\psi_2}^2 \\ &= C\sum_{i=1}^n a_i^2\frac{1}{\log(2)}.\end{aligned}$$

Taking the square root of both sides, for all $p \geq 2$ where C is an absolute constant,

$$\begin{aligned}\left\|\sum_{i=1}^n a_i\varepsilon_i\right\|_{\psi_2} &\leq C\sqrt{\sum_{i=1}^n a_i^2} \\ &\leq C\sqrt{p}\sqrt{\sum_{i=1}^n a_i^2}.\end{aligned}$$

Now, we evaluate the expectation of $\left(\sum_{i=1}^n a_i\varepsilon_i\right)^2$ by expanding the squared summation as follows:

$$\mathbb{E}\left[\left(\sum_{i=1}^n a_i\varepsilon_i\right)^2\right] = \sum_{i=1}^n a_i^2 + \mathbb{E}\sum_{i,j:i\neq j} a_i a_j \varepsilon_i\varepsilon_j.$$

The second term on the right hand side is 0 by our assumption of independence of ε_i. Therefore, we are left with:

$$\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^n a_i\varepsilon_i\right)^2\right]} = \left(\sum_{i=1}^n a_i^2\right)^{1/2}.$$

□

Next, we classify distributions with tail behavior that does not meet the definition of sub-Gaussian behavior but can be characterized analogously.

3.3 Non-Sub-Gaussian Distributions

Let $X \sim N(0,1)$, $X = Z^2 - 1$ where $\mathbb{E}X = 0$. Is X sub-Gaussian? We evaluate $\mathbb{E}(\lambda X)$ by

$$\mathbb{E}\exp\left(\lambda(Z^2-1)\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(\lambda(z^2-1)\right)\exp\left(\frac{-z^2}{2}\right)dz.$$

- If $\lambda \geq 1/2$, the exponential moment is not defined. Therefore, X is not sub-Gaussian because there is no reasonable upper bound.
- If $\lambda < 1/2$,
$$\mathbb{E}\exp\left(\lambda(Z^2-1)\right) = \exp(-\lambda)\frac{1}{\sqrt{1-2\lambda}} \leq \exp\left(\frac{\lambda^2}{1-2\lambda}\right),$$
so X is sub-Gaussian up to some moment.

3.4 Sub-Exponential Norm

Definition 3.15 (Sub-Exponential Norm)**.** *Let X be a random variable (not necessarily such that $\mathbb{E}X = 0$). The sub-exponential norm of X is*

$$\|X\|_{\psi_1} = \inf\left\{t \geq 0 : \mathbb{E}\left[\exp\left(\frac{|X|}{t}\right)\right] \leq 2\right\}.$$

3.4.1 Properties of the Sub-Exponential Norm

Proposition 3.16. *$\|\cdot\|_{\psi_1}$ is a valid norm.*

As with the sub-Gaussian norm, to prove this, we need to show that, for a random variable X and $\lambda \in \mathbb{R}$, the following are satisfied:

1. $\|\lambda X\|_{\psi_1} = |\lambda|\|X\|_{\psi_1}$;
2. $\|X + Y\|_{\psi_1} \leq \|X\|_{\psi_1} + \|Y\|_{\psi_1}$;
3. $\|X\|_{\psi_1} = 0 \iff X = 0$ almost surely.

Proposition 3.17 (Sub-Exponential Equivalent Characterizations)**.** *For a random variable X the following are equivalent for $t \geq 0$ and absolute constant c:*

1. $\Pr(|X| \geq t) \leq 2\exp\left(-ct/\|X\|_{\psi_1}\right)$;
2. $\mathbb{E}\exp\left(|X|/\|X\|_{\psi_1}\right) \leq 2$;
3. $\|X\|_{L_p} \leq cp\|X\|_{\psi_1} \quad \forall p \geq 1$;
4. $\mathbb{E}X = 0 \implies \mathbb{E}\exp(\lambda X) \leq \exp\left(C\lambda^2\|X\|_{\psi_1}^2\right), |\lambda| \leq c/\|X\|_{\psi_1}$.

Proposition 3.18. *Any sub-Gaussian random variable is also sub-exponential, but the reverse is not necessarily true.*

Proposition 3.19. $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$

Proof that a sub-exponential random variable is a sub-Gaussian squared follows from the definitions of sub-Gaussian and sub-exponential random variables. See [Ver18, Definition 2.7.3].

Proposition 3.20. *If X and Y are random variables (not necessarily independent), then $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$.*

Proof. Without loss of generality, assume $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. Recall Young's inequality, $|ab| \leq \frac{a^2}{2} + \frac{b^2}{2}$, which gives:

$$\begin{aligned}
\mathbb{E}\exp(|XY|) &\leq \mathbb{E}\left(\frac{X^2}{2} + \frac{Y^2}{2}\right) \\
&\leq \mathbb{E}\left(\frac{1}{2}\exp\left(X^2\right) + \frac{1}{2}\exp\left(Y^2\right)\right) \\
&= 2 \quad \text{by assumption } \|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1.
\end{aligned}$$

The result follows from Property 2 in Proposition 3.17.

□

3.5 Bernstein's inequality

Next, we derive bounds for sums of sub-exponential random variables analogous to our bounds for sums of sub-Gaussian random variables.

Proposition 3.21. *Let $X_1, \ldots X_n$ be independent sub-exponential random variables where $\mathbb{E}X_i = 0$. For all $t \geq 0$,*

$$\Pr\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\sum_{i=1}^{n}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_{i\in[n]}\|X_i\|_{\psi_1}}\right)\right)$$

where c is a small absolute constant. Bernstein's inequality is covered in more detail at the start of next lecture.

Lecture 4: Tail and High-Probability Bounds

Instructor: Nikita Zhivotovskiy *Scriber: João Vitor Romano* *Proofreader: Annie Ulichney*

4.1 Bernstein's inequality

In the previous lecture, we covered the sub-Gaussian norm $\|\cdot\|_{\psi_2}$, the sub-exponential norm $\|\cdot\|_{\psi_1}$, and some applications of Hoeffding's inequality. We will continue with our study of tail bounds bounds and their equivalent representation as high-probability bounds. A particular form of Bernstein's inequality (Theorem 4.9) will be compared to Hoeffding's inequality to give intuition of the trade-offs between them.

Theorem 4.1 (Bernstein's inequality (ψ_1 form))**.** *Let $X_1, \dots, X_n$ be independent, zero-mean, sub-exponential random variables. Then, for all $t \geq 0$,*

$$\Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-c \min\left(\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{i\in[n]}\|X_i\|_{\psi_1}}\right)\right),$$

and, by symmetry and the union bound,

$$\Pr\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-c \min\left(\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{i\in[n]}\|X_i\|_{\psi_1}}\right)\right),$$

where $c > 0$ is an absolute constant.

Proof. The proof follows by applying Markov's inequality to $\exp\left(\lambda \sum_{i=1}^n X_i\right)$, bounding the moment-generating function using the fact that the random variables are sub-exponential and optimizing the result. For a detailed derivation, we refer the reader to [Ver18, Theorem 2.8.1]. □

We now state an inequality that is useful in bounding the moment-generating function of random variables that assume only two values, for example, those with Bernoulli or Rademacher laws. The curious reader can consult [BK15, Appendix A] for bibliographical remarks and a proof of the following inequality.

Lemma 4.2 (Kearns-Saul inequality)**.** *For all $p \in [0,1]$ and all $\lambda \in \mathbb{R}$,*

$$p\exp(\lambda(1-p)) + (1-p)\exp(-\lambda p) \leq \exp\left(\lambda^2 \frac{1-2p}{4\log\frac{1-p}{p}}\right).$$

It is now possible to discuss a simple example of a bound that encodes information about the variance of the random variable.

Example 4.3 (Centered Bernoulli)**.** *Let $X \sim \mathrm{Ber}(p) - p$ and note that the variance $\mathrm{Var}[X] = p(n-p)$ is the same as that of a non-centered Bernoulli random variable because the variance is invariant to translation. When p is close to zero or one, the variance is small, and when p is close to 0.5, the variance is high. Using Kearns-Saul inequality (Lemma 4.2), we are able to find an upper bound that takes the variance into account:*

$$\mathbb{E}\exp(\lambda X) = p\exp(\lambda(1-p)) + (1-p)\exp(-\lambda p) \leq \exp\left(\lambda^2 \frac{1-2p}{4\log\frac{1-p}{p}}\right).$$

In Lecture 3, Proposition 15, we saw four equivalent characterizations of a sub-exponential random variable. The treatment there was in terms of the sub-exponential norm $\|\cdot\|_{\psi_1}$, in particular for the fourth characterization. More generally, a zero-mean random variable X is said to be sub-exponential with non-negative parameters ν^2 and α if, for all $|\lambda| < \frac{1}{\alpha}$,

$$\mathbb{E}\exp(\lambda X) \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right).$$

Taking $\nu \equiv \sigma\sqrt{2}$ and $\alpha \equiv 2b$ gives the equivalent characterization, for all $|\lambda| < \frac{1}{2b}$:

$$\mathbb{E}\exp(\lambda X) \leq \exp\left(\lambda^2\sigma^2\right).$$

Since $1 - b|\lambda| \geq 1/2$, we can give the following alternative definition of a sub-exponential random variable.

Definition 4.4 (Sub-exponential)**.** *A zero-mean random variable X is sub-exponential with parameters σ^2 and b if*

$$\mathbb{E}\exp(\lambda X) \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1 - b|\lambda|}\right).$$

Note that this is a generalization of the previous definitions by splitting b and σ.

Definition 4.5 (Bernstein's moment condition)**.** *A random variable X with mean $\mu = \mathbb{E}X$ and variance $\sigma^2 = \mathrm{Var}[X]$ satisfies Bernstein's moment condition with parameter b if, for all $k \in \mathbb{N}_{\geq 2}$,*

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

Proposition 4.6. *A random variable that is almost surely bounded by B when centered satisfies Bernstein's moment condition with parameter $b = B/3$.*

Proof. Let X be an almost surely bounded random variable with mean $\mu = \mathbb{E}X$ and variance $\sigma^2 = \mathrm{Var}[X]$. From boundedness, we have that its mean is finite and therefore $X - \mu$ is also bounded, that is, $|X - \mu| \leq B < \infty$. For $k \in \mathbb{N}_{\geq 2}$,

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

where $b = B\left(\frac{2}{k!}\right)^{\frac{1}{k-2}}$ for $k \geq 3$. For $k = 2$, b can assume any value since $b^{k-2} = b^0 = 1$ for all $b \in \mathbb{R}$. Note that $\left(\frac{2}{k!}\right)^{\frac{1}{k-2}}$ is decreasing in k, so choosing $b = B\left(\frac{2}{3!}\right)^{\frac{1}{3-2}} = B/3$ gives us the tightest bound that does not depend on k. □

Remark 4.7. *Although boundedness is sufficient for a random variable to satisfy Bernstein's moment condition, it is in no way necessary: some unbounded random variables, such as those with Gaussian or χ^2 law, also do so.*

Lemma 4.8. *Let X be a random variable with $\mathbb{E}X = 0$ and $\mathrm{Var}[X] = \sigma^2$. If X satisfies Bernstein's moment condition with parameter b and $|X| \leq B$ almost surely, then X is $(\sigma^2, B/3)$-sub-exponential.*

Proof. We follow by representing the moment-generating function in Taylor series form, making use of Bernstein's moment, noting that the geometric series is summable for any $|\lambda| < 1/b$, and using the fact that $1 + t \leq \exp(t)$ in the last step:

$$\mathbb{E}\exp(\lambda X) = 1 + 0 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}X^k}{k!}$$

$$
\begin{aligned}
&\leq 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{\lambda^k k! \sigma^2 b^{k-2}}{2k!} \\
&= 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} \lambda^{k-2} b^{k-2} \\
&= 1 + \frac{\lambda^2\sigma^2}{2} \left(1 + \frac{|\lambda| b}{1 - b|\lambda|}\right) \\
&= 1 + \frac{\lambda^2\sigma^2}{2} \left(\frac{1}{1 - b|\lambda|}\right) \\
&\leq \exp\left(\frac{\lambda^2\sigma^2/2}{1 - b|\lambda|}\right).
\end{aligned}
$$

The proof is complete by applying Proposition 4.6 to get $b = B/3$. □

We can now prove another version of Bernstein's inequality for random variables that satisfy Bernstein's moment condition.

Theorem 4.9 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with means $\mathbb{E}X_i = \mu_i$ and variances $\mathrm{Var}[X_i] = \sigma_i^2$ for $i = 1, \ldots, n$. If the random variables also satisfy Bernstein's moment condition with parameter b, then*

$$
\Pr\left(\sum_{i=1}^{n} X_i - \mu_i \geq t\right) \leq \exp\left(-\frac{t^2/2}{\sum_{i=1}^{n} \sigma_i^2 + bt}\right),
$$

and, by symmetry and the union bound,

$$
\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{\sum_{i=1}^{n} \sigma_i^2 + bt}\right).
$$

Proof. The result follows from a direct application of Chernoff's method and Lemma 4.8; see [Wai19, Proposition 2.14] for details. □

Let us finish this section by briefly recalling a useful technique for bridging *tail bounds* and *high-probability bounds*. Consider Hoeffding's inequality (Lecture 3, Proposition 1) and assume for simplicity that the random variables are bounded in the unit interval. For notational convenience, let $S_n := \sum_{i=1}^{n} X_i$ so that the *tail bound* is

$$
\Pr(S_n - \mathbb{E}S_n \geq t) \leq \exp\Big(-2t^2/n\Big),
$$

which can be rewritten as

$$
\Pr(S_n - \mathbb{E}S_n < t) > 1 - \exp\Big(-2t^2/n\Big) =: 1 - \delta.
$$

Expressing t in terms of δ yields the equivalent *high-probability bound*

$$
\Pr\left(S_n - \mathbb{E}S_n < \sqrt{\frac{n \log(1/\delta)}{2}}\right) > 1 - \delta.
$$

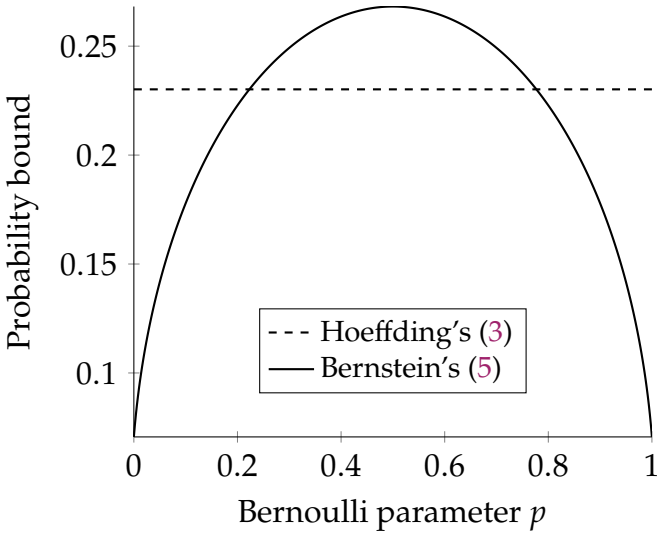


FIGURE 1: Comparison of Hoeffding's high-probability bound (Eq. 3) and Bernstein's high-probability bound (Eq. 5) for $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p) - p$ with varying $p \in [0, 1]$, $n = 50$ and $\delta = 0.01$. If $\text{Var}[X_1] = p(1-p)$ is large (p close to 0.5), Hoeffding's outperforms Bernstein's. For small variance (p close to 0 or 1), Bernstein's is preferred.

4.2 Comparing bounds

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p) - p$ be distributed as a centered Bernoulli with parameter $p \in [0, 1]$. Note that $-p \le X_1 \le 1 - p$ and $\mathbb{E}X_1 = 0$, so Hoeffding's inequality gives

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \ge t\right) \le 2\exp\left(-2nt^2\right). \tag{2}$$

Restating this result as a high-probability bound as outlined above yields that, with probability at least $1 - \delta \in (0, 1)$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \le \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{3}$$

Note that this upper bound depends on n and δ, but not on p. Given that we know $\text{Var}[X_1] = p(1-p)$, one might expect to do better under certain situations by using information about the variance. Since $|X_1| \le \max(p, 1-p) \le 1$, we can apply Bernstein's inequality to get

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2p(1-p) + 2t/3}\right). \tag{4}$$

By solving $\delta := 2\exp\left(-\frac{nt^2}{2p(1-p)+2t/3}\right)$ for t, we can restate the result as a high-probability bound. Indeed, with probability at least $1 - \delta \in (0, 1)$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \le \sqrt{\left(\frac{\log(2/\delta)}{3n}\right)^2 + \frac{2p(1-p)\log(2/\delta)}{n}} + \frac{\log(2/\delta)}{3n}. \tag{5}$$

For a cleaner but less tight bound, recall that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for all $a, b \in \mathbb{R}_{\ge 0}$, so

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \le \sqrt{\frac{2p(1-p)\log(2/\delta)}{n}} + \frac{2\log(2/\delta)}{3n}.$$

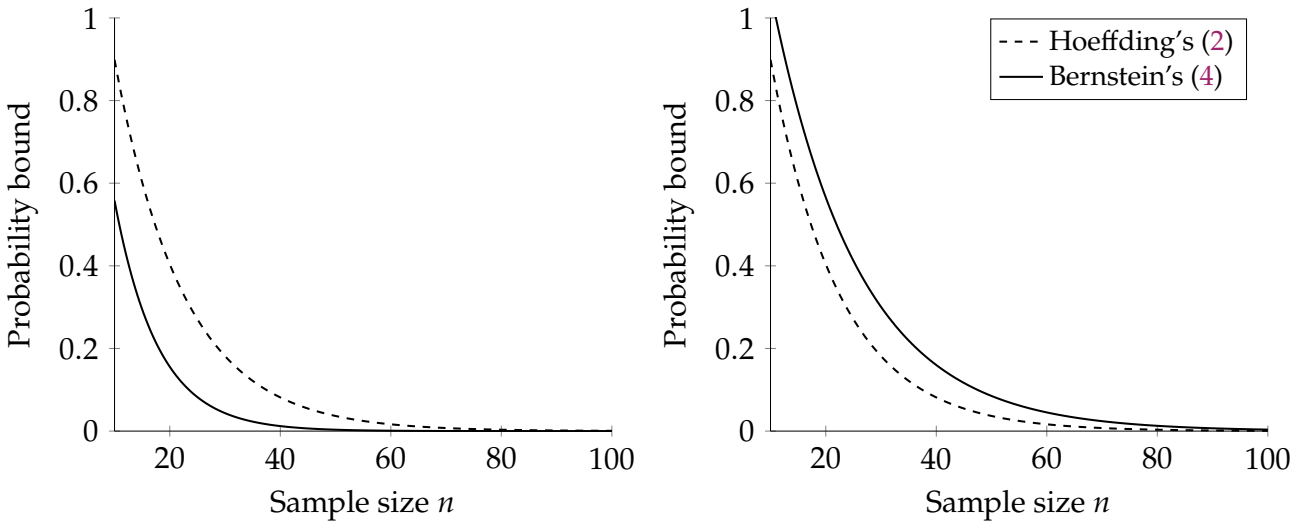


FIGURE 2: Comparison of Hoeffding's tail bound (Eq. 2) and Bernstein's tail bound (Eq. 4) for $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p) - p$ with $p = 0.9$ (left panel; low variance) and $p = 0.5$ (right panel; high variance), $t = 0.2$ and varying sample size n. Hoeffding's inequality outperforms Bernstein's when the variance $\text{Var}[X_1] = p(1-p)$ is large; Bernstein's is preferred on the low-variance regime. The gap between the bounds is especially prominent for small n and decreases as n grows.

In Figure 1, we analyze Hoeffding's high-probability bound (Equation 3) and Bernstein's high-probability bound (Equation 5) for the centered Bernoulli example with $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p) - p$ with p varying in the unit interval while $n = 50$ and $\delta = 0.01$ are kept fixed. On the one hand, when p is close to 0.5, the variance $\text{Var}[X_1] = p(1-p)$ is large and Hoeffding's gives a better bound. On the other hand, when p close to 0 or 1 and the variance therefore small, Bernstein's is preferred.
Figure 2 shows a complementary analysis of the tail bounds (Equations 2 and 4) for the same centered Bernoulli example. Now, the parameter of the distribution is fixed at $p = 0.9$ (left panel; low variance) or $p = 0.5$ (right panel; high variance), $t = 0.2$ and the sample size n varies from 10 to 100. Once again, we observe Hoeffding's advantage when the variance is large and Bernstein's when it is low. Importantly, we see that the difference between the bounds is especially prominent for small sample sizes and less relevant for large n.
For a complementary comparison, first recall from Lecture 1, Definition 1, that a zero-mean random variable X is sub-Gaussian with parameter σ^2 if, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\exp(\lambda X) \leq \exp\left(\lambda^2\sigma^2/2\right). \tag{6}$$

In the literature, the parameter σ^2 is oftentimes referred to as *proxy variance* to emphasize it is not necessarily the true variance of the random variable. One may also encounter the term *optimal proxy variance*, σ^2_{opt}, to denote the smallest proxy variance such that Equation 6 holds. Then, from [Riv12, Proposition 2.1], we have that every proxy variance at most equal to the true variance:

$$\text{Var}[X] \leq \sigma^2_{\text{opt}} \leq \sigma^2.$$

A natural question that arises is how different proxy variances compare to one another and to the true variance. Recall that the variance for the centered Bernoulli is given by

$$\text{Var}[X] = p(1-p), \tag{7}$$

Hoeffding's lemma asserts that

$$\mathbb{E}\exp(\lambda X) \leq \exp\left(\frac{\lambda^2}{2} \cdot \frac{1-p-(-p)}{8}\right) = \exp\left(\frac{\lambda^2}{2}\frac{1}{4}\right), \tag{8}$$

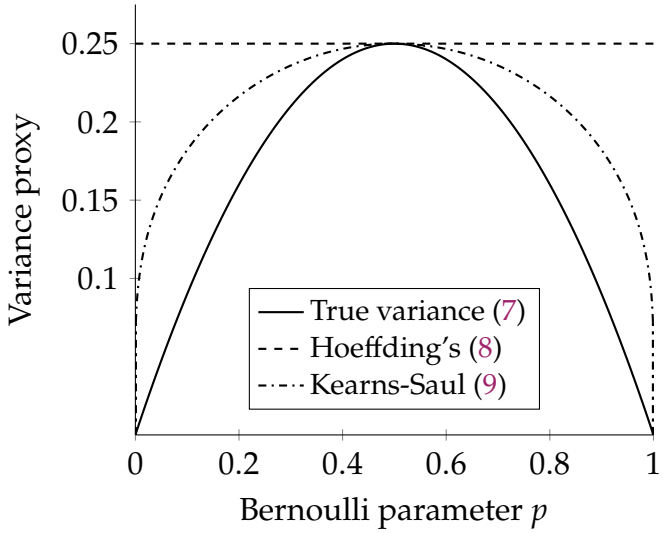


Figure 3: Comparison of the true variance (solid) of a centered Bernoulli random variable with varying parameter p, the proxy variance from Hoeffding's lemma (dashed) and the optimal proxy variance that comes from Kearns-Saul inequality (dashdotted).

and Kearns-Saul inequality yields

$$\mathbb{E}\exp(\lambda X) \le \exp\left(\frac{\lambda^2}{2}\cdot\frac{1-2p}{2\log\frac{1-p}{p}}\right). \tag{9}$$

In fact, the optimal proxy variance σ^2_{opt} for the centered Bernoulli can be shown to be that from Kearns-Saul inequality, i.e., $\frac{1-2p}{2\log\frac{1-p}{p}}$ [AMN20, Proposition 4.1].

Figure 3 compares the proxy variances coming from Equations 8 and 9 above to the true variance of the centered Bernoulli defined in Equation var-true. Note that, as expected, the true variance is a uniform lower bound and that the proxy variance from Kearns-Saul is smaller than the one from Hoeffding's.

4.3 Another application of Bernstein's inequality

Instead of bounding the probability of a random variable deviating from its mean by a given constant, one might be interested in deviations relative to the mean, that is, $\Pr[|X - \mathbb{E}X| \ge \gamma\mathbb{E}X]$ for some $\gamma \in [0,1]$. Let us consider the case $X \sim \text{Bin}(n,p)$; equivalently, $X = \sum_{i=1}^n X_i$ where $X_1,\dots,X_n \overset{\text{iid}}{\sim} \text{Ber}(p)$. From Bernstein's inequality and the triangle inequality (see Section Comparing bounds),

$$|X - \mathbb{E}X| \le \sqrt{2np(1-p)\log(2/\delta)} + \frac{2}{3}\log(2/\delta).$$

Recall that for any $a,b \ge 0$ and $\gamma > 0$, we have $\sqrt{ab} \le \frac{1}{2}(\gamma a + b/\gamma)$. Let $a \equiv 2np = 2\mathbb{E}X$ and $b \equiv (1-p)\log(2/\delta)$, such that

$$\begin{aligned}|X - \mathbb{E}X| &\le \frac{1}{2}\left(2\gamma\mathbb{E}X + \frac{(1-p)\log(2/\delta)}{\gamma}\right) + \frac{2}{3}\log(2/\delta)\\ &= \gamma\mathbb{E}X + \log(2/\delta)\left(\frac{1-p}{2\gamma} + \frac{2}{3}\right).\end{aligned}$$

By noting that $1 - p \leq 1$ and introducing the assumption that $\gamma \mathbb{E}X \geq \log(2/\delta)(\frac{1}{2\gamma} + \frac{2}{3})$, we have that, with probability at least $1 - \delta$,

$$\mathbb{E}X - 2\gamma \mathbb{E}X \leq X \leq \mathbb{E}X + 2\gamma \mathbb{E}X.$$

We can see from this that in certain circumstances a sum of Bernoulli random variables could be replaced by a proportion of its expectation. This is known as a multiplicative Chernoff bound. For $\gamma = 0.1$ for example, we have $0.8 \cdot \mathbb{E}X \leq X \leq 1.2 \cdot \mathbb{E}X$. We emphasize that this result depends on the assumption that $\gamma \mathbb{E}X \geq \log(2/\delta)(\frac{1}{2\gamma} + \frac{2}{3})$, but note that this is reasonable, especially for large n.

In the next lecture, we will start with an application of Bernstein's inequality to statistical learning theory with roots in the works of Vladimir Vapnik and Alexey Chervonenkis in the 1960s and Leslie Valiant in the 1980s.

Lecture 5: Learning Theory and Maximal Inequalities

Instructor: Nikita Zhivotovskiy *Scriber: Sohom Paul* *Proofreader: Dylan Webb*

5.1 Statistical Learning Theory

Consider the following simple model for learning a binary classifier:

- Our instances are drawn from a set $\mathcal{X}$.
- The true classifier f^* is an element of some finite family $\mathcal{F}$ of functions from $\mathcal{X}$ to $\{0,1\}$. Let M denote $|\mathcal{F}|$.
- We observe the labeled training set $(X_1, f^*(X_1)), \ldots, (X_n, f^*(X_n))$ for training points $X_1, \ldots, X_n$ drawn i.i.d. from the (unknown) probability distribution $\mathcal{P}$ over the instance space $\mathcal{X}$.
- Our goal is to output some decision rule f such that our decision rule agrees with f^* with high probability on new samples from $\mathcal{P}$. Namely, we seek to minimize $\Pr_{X\sim\mathcal{P}}(f(X) \neq f^*(X))$.

This basic model has been studied by Vapnik and Chervonenkis in [VC71] and Valiant in [Val84].

Definition 5.1. *Let the **risk** of classifier f, denoted $R(f)$, be the probability of misclassification when using classifier f for new data drawn from our distribution $\mathcal{P}$. Namely,*

$$R(f) := \Pr_{X\sim P}(f(X) \neq f^*(X)).$$

Definition 5.2. *Let the **empirical risk** of classifer f, denoted $R_n(f)$, be the proportion of errors that classifier f makes on the observed training data. Namely,*

$$R_n(f) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq f^*(X_i)\}.$$

Consider some fixed f. Note that the indicators $\mathbb{1}\{f(X_i) \neq f^*(X_i)\}$ are i.i.d. Bernoulli trials, each with probability $R(f)$ of occurring, so we conclude that $R_n(f)$ is a scaled binomial random variable. Applying Bernstein's inequality shows, with probability at least $1-\delta$,

$$\begin{aligned} R(f) - R_n(f) &\leq \sqrt{\frac{2R(f)(1-R(f))\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n} \\ &\leq \sqrt{\frac{2R(f)\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n}. \end{aligned} \tag{*}$$

In order to get the tightest bound on the risk, our strategy will be to choose $\hat{f} \in \mathcal{F}$ such that $\hat{f}(X_i) = f^*(X_i)$ for all $i \in \{1, \ldots, n\}$ (so $R_n(\hat{f}) = 0$). Such a $\hat{f}$ certainly exists because f^* itself is contained in $\mathcal{F}$. However, we cannot naively use our analysis using Bernstein's inequality above to bound the risk, as the choice of $\hat{f}$ depends on the entire observed dataset $(X_1, \ldots, X_n)$, and thus we do not necessarily have independence of the indicators $\mathbb{1}\{\hat{f}(X_i) \neq f^*(X_i)\}$.

Instead, observe that we have for any fixed f, (*) shows, with probability $1 - \delta/M$,

$$R(f) - R_n(f) \le \sqrt{\frac{2R(f)\log(M/\delta)}{n}} + \frac{2\log(M/\delta)}{3n}.$$

Thus, applying union bound to the different $f \in \mathcal{F}$ shows, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}. \quad R(f) - R_n(f) \le \sqrt{\frac{2R(f)\log(M/\delta)}{n}} + \frac{2\log(M/\delta)}{3n}.$$

Because we chose $\hat{f}$ from $\mathcal{F}$ to have $R_n(\hat{f}) = 0$, we conclude

$$R(\hat{f}) \le \sqrt{\frac{2R(\hat{f})\log(M/\delta)}{n}} + \frac{2\log(M/\delta)}{3n}.$$

One can check that for $R(\hat{f}) > 10\log(M/\delta)/n$ the inequality above is violated, so we conclude that

$$R(\hat{f}) \le \frac{10\log(M/\delta)}{n} = \frac{10(\log(M) + \log(1/\delta))}{n}.$$

For an alternative proof, fix $\varepsilon \in (0, 1)$ and notice, by union bound,

$$\begin{aligned}\Pr(\exists f \in \mathcal{F} : R(f) \ge \varepsilon, R_n(f) = 0) &\le \sum_{f \in \mathcal{F}} \Pr(R(f) \ge \varepsilon, R_n(f) = 0) \\ &\le M(1-\varepsilon)^n,\end{aligned}$$

The final inequality holds because, given that the risk of classifier f is at least ε, each X_i has an independent probability of at most $1 - \varepsilon$ of being correctly classified. Recalling $1 - x \le \exp(-x)$ holds for all x, we obtain

$$\Pr(\exists f \in \mathcal{F} : R(f) \ge \varepsilon, R_n(f) = 0) \le M\exp(-\varepsilon n),$$

so for $n \ge (\log M + \log(1/\delta))/\varepsilon$ the probability of there existing some f with empirical risk 0 but true risk at least ε can be bounded by δ. We conclude that with probability at least $1 - \delta$ that $R(\hat{f})$ is bounded by $(\log M + \log(1/\delta))/n$. This gives the same result as before, up to constants.

5.2 Maximal Inequalities

Previously, we have derived bounds on sums of random variables. It will be useful to similarly derive bounds on maxima of sets of random variables.

Theorem 5.3. *Let $X_1, \ldots, X_n$ be zero-mean, not necessarily independent, subgaussian random variables. Namely, suppose $\mathbb{E}\left[\exp(\lambda X_i)\right] \le \exp\left(\lambda^2\sigma^2/2\right)$ holds for all λ and for each $i \in [n]$. Then,*

$$\mathbb{E}\left[\max(X_1, \ldots, X_n)\right] \le \sqrt{2\sigma^2 \log n}.$$

Proof. Using Jensen's inequality, we can compute for $\lambda > 0$

$$\mathbb{E}\left[\max(X_1, \ldots, X_n)\right] = \frac{1}{\lambda}\mathbb{E}\left[\log\exp\left(\lambda \max(X_1, \ldots, X_n)\right)\right]$$

$$
\begin{aligned}
&\leq \frac{1}{\lambda} \log \mathbb{E}\left[\exp(\lambda \max(X_1, \dots, X_n))\right] \\
&= \frac{1}{\lambda} \log \mathbb{E}\left[\max_i \exp(\lambda X_i)\right] \\
&\leq \frac{1}{\lambda} \log \mathbb{E}\left[\sum_i \exp(\lambda X_i)\right] \\
&\leq \frac{1}{\lambda} \log\left(n \exp\left(\frac{\lambda^2\sigma^2}{2}\right)\right) \\
&= \frac{1}{\lambda} \log n + \frac{\lambda\sigma^2}{2}.
\end{aligned}
$$

Taking the infimum of the right-hand side over λ shows the claim. □

Theorem 5.4. *Let $X_1, \dots, X_n$ be zero-mean, not necessarily independent, subexponential random variables. Namely, suppose that for all $|\lambda| \leq 1/b$ and $i \in [n]$, we have*

$$
\mathbb{E}\left[\exp(\lambda X_i)\right] \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1 - b|\lambda|}\right).
$$

Then,

$$
\mathbb{E}\left[\max(X_1, \dots, X_n)\right] \leq \sqrt{2\sigma^2 \log n} + b \log n.
$$

In particular, there is an absolute constant C such that

$$
\mathbb{E}\left[\max(X_1, \dots, X_n)\right] \leq C \max_i \|X_i\|_{\psi_1} \log n.
$$

Proof. Following the same steps as for the previous theorem, we deduce that for any $0 < \lambda < 1/b$,

$$
\begin{aligned}
\mathbb{E}\left[\max(X_1, \dots, X_n)\right] &\leq \frac{1}{\lambda} \log \mathbb{E}\left[\sum_i \exp(\lambda X_i)\right] \\
&\leq \frac{1}{\lambda} \log\left(n \exp\left(\frac{\lambda^2\sigma^2/2}{1 - b\lambda}\right)\right) \\
&= \frac{1}{\lambda} \log n + \frac{\lambda\sigma^2}{2(1 - b\lambda)}.
\end{aligned}
$$

Solving for the infimum over λ yields the first claim. The latter holds by noting that X_i is $(C_1\|X_i\|_{\psi_1}, C_2\|X_i\|_{\psi_1})$-subexponential for some choice of absolute constants C_1, C_2, so we can collect the terms together. □

Theorem 5.5. *For any, not necessarily independent, set of random variables $X_1, \dots, X_n$ and $p \geq 1$, we have*

$$
\mathbb{E}\left[\max(X_1, \dots, X_n)\right] \leq n^{1/p} \max_i \|X_i\|_{L_p}.
$$

Proof. By Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}\left[\max(X_1, \dots, X_n)\right] &\leq \mathbb{E}\left[(\max(|X_1|^p, \dots, |X_n|^p))^{1/p}\right] \\
&\leq (\mathbb{E}\left[\max(|X_1|^p, \dots, |X_n|^p)\right])^{1/p} \\
&\leq n^{1/p} \max_i \|X_i\|_{L_p},
\end{aligned}
$$

as desired. □

Definition 5.6. *We say $g : \mathcal{X}^n \to \mathbb{R}$ satisfies the **bounded differences property** with constants $C_1, \ldots, C_n$ if, for all $i \in [n]$, we have*

$$\sup_{\substack{x_1,\ldots,x_n \in \mathcal{X} \\ x_i' \in \mathcal{X}}} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq C_i.$$

Theorem 5.7 (McDiarmid's Inequality)**.** *Let $g : \mathcal{X} \to \mathbb{R}$ satisfy the bounded differences property with constants $C_1 \ldots, C_n$, and let $X_1, \ldots, X_n$ be independent random variables over $\mathcal{X}$. Then,*

$$\Pr(g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \geq t) \leq \exp\left(\frac{-2t^2}{\sum_i C_i^2}\right).$$

Proof. Note $g := g(X_1, \ldots, X_n)$ is a random variable, and define for $i \in \{1, \ldots, n\}$

$$Y_i := \mathbb{E}\left[g \mid X_1, \ldots, X_i\right] - \mathbb{E}\left[g \mid X_1, \ldots, X_{i-1}\right].$$

(In particular, $Y_1 = \mathbb{E}\left[g \mid X_1\right] - \mathbb{E}\left[g\right]$.) Then, we have $g - \mathbb{E}g = \sum_i Y_i$. Now, for each $i \in [n]$ define $h_i(x_1, \ldots, x_i) := \mathbb{E}\left[g \mid X_1 = x_1, \ldots, X_i = x_i\right]$. Observe

$$\begin{aligned} Y_i &\leq \sup_{x \in \mathcal{X}} h_i(X_1, \ldots, X_{i-1}, x) - h_{i-1}(X_1, \ldots, X_{i-1}), \\ Y_i &\geq \inf_{x \in \mathcal{X}} h_i(X_1, \ldots, X_{i-1}, x) - h_{i-1}(X_1, \ldots, X_{i-1}), \end{aligned}$$

and thus $Y_i \mid X_1, \ldots, X_{i-1}$ belongs to an interval of size at most C_i, by the bounded differences property, and
$\mathbb{E}\left[Y_i \mid X_1, \ldots, X_{i-1}\right] = 0$ by iterated expectation. Thus, by Hoeffding's Lemma,

$$\mathbb{E}\left[\exp(\lambda Y_i) \mid X_1, \ldots, X_{i-1}\right] \leq \exp\left(\frac{\lambda^2 C_i^2}{8}\right).$$

We obtain

$$\begin{aligned} \mathbb{E}\left[\exp(\lambda(g - \mathbb{E}g))\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Y_i\right)\right] \\ &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} Y_i\right) \exp(\lambda Y_n)\right] \\ &= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\exp\left(\lambda \sum_{i=1}^{n-1} Y_i\right) \mathbb{E}_{X_n}[\exp(\lambda Y_n)]\right] \\ &\leq \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\exp\left(\lambda \sum_{i=1}^{n-1} Y_i\right) \exp\left(\frac{\lambda^2 C_i^2}{8}\right)\right]. \end{aligned}$$

Iterating this argument yields

$$\mathbb{E}\left[\exp(\lambda(g - \mathbb{E}g))\right] \leq \exp\left(\frac{\lambda^2 \sum_i C_i^2}{8}\right).$$

Finally, we use the Chernoff bound to finish the proof. □

Remark 5.8. *$g(X_1, \ldots, X_n) = \sum_i X_i$ for $X_i \in [0, 1]$ satisfies the bounded differences property. Thus, McDiarmid's inequality generalizes Hoeffding's lemma.*

5.3 Kernel Density Estimation

Consider observing $X_1, \ldots, X_n$ i.i.d. samples from some (unknown) probability density f over $\mathbb{R}$. We seek to estimate the unknown density from our data using a *kernel estimator*

$$\hat{f}_n(x; X_1, \ldots, X_n) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $K : \mathbb{R} \to \mathbb{R}$ is some kernel function satisfying $K(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int K(x)\, dx = 1$, and h is an appropriate hyperparameter representing our desired window length. We measure the quality of our estimator using expected L_1 distance:

$$\mathbb{E}_{X_1,\ldots,X_n} \int_{\mathbb{R}} |\hat{f}_n(x; X_1, \ldots, X_n) - f(x)|\, dx.$$

Next lecture, we will discuss how we can use McDiarmid's inequality to bound this objective.

Lecture 6: Kernel Density Estimation and Norm Concentration

Instructor: Nikita Zhivotovskiy *Scriber: Julian Morimoto* *Proofreader: Daniel Etaat*

6.1 Kernel Density Estimation (continued)

We begin this lecture by continuing our discussion of density estimation. Let $x_1, \dots, x_n$ be IID samples drawn from some distribution with an unknown density function f. To estimate f, we may employ kernel density estimation, a non-parametric method to estimate the density of a random variable. We estimate f by

$$\hat{f}_n := \frac{1}{nh} \sum_{j \in [n]} K\left(\frac{x - x_j}{h}\right)$$

where K is some kernel function and $h > 0$ is parameter sometimes called the bandwidth. Recall that a kernel function is a non-negative function that satisfies $\int_{\mathbb{R}} K(x)dx = 1$.

To determine whether $\hat{f}_n$ is a good estimator, we may want to consider its L^1 distance from f defined as $\|\hat{f}_n - f\|_1 = \int_{\mathbb{R}} |\hat{f}_n(x) - f(x)|dx$. This is infeasible to compute since f is unknown. Instead, we will study its expectation:

$$\mathbb{E}_{x_1,\dots,x_n}[\|\hat{f}_n - f\|_1].$$

We will attempt to bound $\Pr\Big(|\|\hat{f}_n - f\|_1 - \mathbb{E}_{x_1,\dots,x_n}[\|\hat{f}_n - f\|_1]| \geq t\Big)$ via McDiarmid's inequality (see lecture 5 notes). To do this we must first show that $g(x_1, \dots, x_n) := \|\hat{f}_n - f\|_1$ satisfies the bounded differences property (note that $x_1, \dots, x_n$ are used to construct $\hat{f}_n$). Fix some $i \in [n]$ and let $x_i' \neq x_i$. Then we have that $\left|g(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - g(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)\right|$ is equal to:

$$\left|\int_{\mathbb{R}} \left|\frac{1}{nh} \sum_j K\left(\frac{x - x_j}{h}\right) - f(x)\right| dx - \int_{\mathbb{R}} \left|\frac{1}{nh} \sum_{j \neq i} K\left(\frac{x - x_j}{h}\right) + \frac{1}{nh} K\left(\frac{x - x_i'}{h}\right) - f(x)\right| dx\right|.$$

By the triangle inequality this is less than or equal to:

$$\int_{\mathbb{R}} \frac{1}{nh} \left|K\left(\frac{x - x_i}{h}\right) - K\left(\frac{x - x_i'}{h}\right)\right| dx,$$

which is less than or equal to $2/n$ by the properties of K. Then by McDiarmid's inequality we have that,

$$\Pr\left(|\|\hat{f}_n - f\|_1 - \mathbb{E}_{x_1,\dots,x_n}[\|\hat{f}_n - f\|_1]| \geq t\right) \leq 2\exp\left(-\frac{t^2 n}{2}\right).$$

This would be tough result to prove without McDiarmid's inequality and the bounded differences property.

6.2 Concentration of Norms of Random Vectors

Suppose $X \sim \mathcal{N}(0, I_d)$ or equivalently $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i = 1, \ldots, d$. We would like to study the concentration of $\|X\|_2^2 = \sum_{i=1}^d X_i^2$. Since each X_i is trivially sub-Gaussian, we know that their sub-Gaussian norms $\|X_i\|_{\psi_2}$ are finite. Then $\|X_i^2\|_{\psi_1} \leq C$ for some finite C since $\|X_i^2\|_{\psi_1} = \|X_i\|_{\psi_2}^2$. Noting that $\mathbb{E}\left[\|X\|^2\right] = d$, we can apply Bernstein's inequality to show that:

$$\Pr\Big(\big|\|X\|_2^2 - d\big| \geq t\Big) \leq 2\exp\left(-c' \min\left\{\frac{t^2}{dC^2}, \frac{t}{C}\right\}\right).$$

Normalizing by d and restricting $t \in (0,1)$ yields

$$\Pr\left(\left|\frac{\|X\|_2^2}{d} - 1\right| \geq t\right) \leq 2\exp\left(-ct^2 d\right). \tag{1}$$

This is an interesting result that shows that the norm of some nice random vector concentrates around d as its dimension, d increases. In the next section, we use this machinery to prove another useful result.

Remark 6.1. *Given the law of large numbers, this is a reasonable thing to expect. As d increases, $\|X\|_2^2$ looks more and more like a sum of a large number of independent random variables whose expectation is 1 (since the coordinates of X are all centered at 0 and have variance 1). What this result helps us see is the rate at which this concentration happens as we increase the dimension.*

6.2.1 The Johnson–Lindenstrauss Lemma

Let $u_1, \ldots, u_n \in \mathbb{R}^d$ and $m \ll d$. We would like to find a projection map $\pi : \mathbb{R}^d \to \mathbb{R}^m$ that preserves distances between these vectors. Informally, we would like π to satisfy

$$\|\pi(u_i) - \pi(u_j)\|_2 \approx \|u_i - u_j\|_2$$

for $i, j \in [n]$.
We can construct such a projection as follows. Let Γ be an $m \times d$ random matrix with normally distributed entries $\Gamma_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Let $v \in S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2^2 = 1\}$. Then $\Gamma v \sim \mathcal{N}(0, I_m)$. Then by (1) we have that for $t \in (0,1)$:

$$\Pr\left(\left|\frac{\|\Gamma v\|_2^2}{m} - 1\right| \geq t\right) \leq 2\exp\left(-ct^2 m\right).$$

We can generalize this bound to an arbitary $v \in \mathbb{R}^d$ as:

$$\Pr\left(\left|\frac{\|\frac{1}{\sqrt{m}}\Gamma v\|_2^2}{\|v\|_2^2} - 1\right| \geq t\right) \leq 2\exp\left(-ct^2 m\right).$$

Finally, applying the union bound over all pairs (u_i, u_j) leads to the following bound:

$$\Pr\left(\frac{\|\frac{1}{\sqrt{m}}\Gamma(u_i - u_j)\|_2^2}{\|u_i - u_j\|_2^2} \notin [1-t, 1+t]\}\right) \leq 2\binom{n}{2}\exp\left(-c^2 t^2 m\right) := \delta.$$

Alternatively, we could say that with probability $1-\delta$,

$$1 - t \leq \frac{\|\frac{1}{\sqrt{m}}\Gamma(u_i - u_j)\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + t$$

for $m \geq \frac{1}{ct^2}\ln(n^2/\delta)$. Our desired projection is then $\pi(v) = \frac{1}{\sqrt{m}}\Gamma v$. What is rather remarkable about this bound is that it does not depend on the dimension d. However, m increases with n, so this projection becomes less efficient as we increase the amount of data we are projecting. If n is infinite, we will need a smarter projection scheme than the one presented here.

Remark 6.2 (Alternative view of the result)**.** *We could also formulate the bound in terms of t. One might be interested in doing this when there's a cap on the dimension of the space to which we wish to project our data, and we want to know what might be the "worst" discrepancy between the distance between the projected vectors as a multiple of the distance of the unprojected vectors. Formally, this would be the infimum of all t such that the inequality is satisfied for fixed m, n, and δ.*

Remark 6.3 (A technique for achieving dimension free bounds)**.** *One way of thinking about why this does not depend on the dimension d is that we are ensuring that whatever goes into the projection has norm 1, and this is achieved by dividing by the norm of the vector that is being projected. In other words, we are able to achieve this nice behavior because we frame our question about acceptable distances between projected and unprojected vectors as one of relative distance by way of ratios, rather than absolute differences (i.e., the distance between $\|m^{-\frac{1}{2}}\Gamma(u_i - u_j)\|_2^2$ and $\|u_i - u_j\|_2^2$ is measured relative to the size of $\|u_i - u_j\|_2^2$).*
By doing this, we ensure that whatever the projection produces is going to be some vector whose entries are sums of centered normal random variables with variances that are essentially uniformly controlled (because the projection works by taking linear combinations of entries of Γ with the constants being determined by the vector being projected, which we are essentially bounding in size through normalization of that vector). We would expect these sums to concentrate predictably even for very large d since more and more of these random variables are added together and we would be able to apply results like the Lindeberg–Feller or Lyapunov Central Limit Theorems. Without dividing by the size of the vector that we are projecting, there is no guarantee that the variances of the normal r.v.'s that we are adding up will be controlled in this way.
For example, if we applied Γ to some unnormalized vector, $v = (2^d, ..., 2^d)$, of dimension d, then the entries of $\Gamma(v)$ would be linear combinations of many normal random variables with very large variances for large d. That kind of object is not something that we would easily expect to concentrate predictably for very large d. It is through normalizing the vectors that we are projecting that we are able to ensure that whatever the projection produces is going to be something that behaves reasonably well. Thus, perhaps another lesson to take from this result is that one way to control dependence on dimensionality in some kind of problem is to normalize in some way the high dimensional objects that we are working with.

6.2.2 Concentration of $\|X\|$ instead of $\|X\|^2$ squared

Proposition 6.4. *Let X be a random vector with independent coordinates X_i for $i = 1, \ldots, n$ such that $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$. Let $K = \max_{i\in[n]} \|X_i\|_{\psi_2}$. Then,*

$$\left\| \|X\|_2 - \sqrt{d} \right\|_{\psi_2} \leq cK^2.$$

Proof. We still begin by looking at $\|X\|_2^2$. Note that $\|X_i\|_{\psi_2} \leq K$ and $\|X_i^2\|_{\psi_1} = \|X_i\|_{\psi_2}^2 \leq K^2$. Then by Bernstein's inequality we have that:

$$\Pr\left(\left|\frac{1}{d}\left(\|X\|_2^2 - d\right)\right| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2 d}{K^4}, \frac{td}{K^2}\right\}\right).$$

We can assume WLOG that $K \geq 1$. This is established by the following argument. We know that the sub-Gaussian norm $\|X_i\|_{\psi_2}^2$ is proportional to its sub-Gaussian parameter σ up to some universal constant where $\mathbb{E}\exp(\lambda X_i) \leq \exp\left(\lambda^2\sigma^2/2\right)$ for all λ. Comparing the Taylor series we have that $1+\mathbb{E}X_i^2\lambda^2/2+O(\lambda^3) \leq 1 + \sigma^2\lambda^2/2 + O(\lambda^3) \Rightarrow \mathbb{E}X_i^2 + O(\lambda) \leq \sigma^2$. Taking $\lambda \to 0$ and using the assumption that $\mathbb{E}X_i^2 = 1$ we have that $\sigma^2 \geq 1$. Then, $\|X_i\|_{\psi_2}^2 > c$ for some universal constant which we can fold into the remaining calculations. With this assumption we have that:

$$\Pr\left(\left|\frac{1}{d}(\|X\|_2^2 - d)\right| \geq t\right) \leq 2\exp\left(-cK^{-4}\min\{t^2 d, td\}\right).$$

Note the following fact: for all $z, u \in \mathbb{R}$, that $|z-1| \geq u$ implies that $|z^2-1| \geq \max\{u, u^2\}$. Combining this with the bound above gives:

$$\begin{aligned}\Pr\left(\left|\frac{\|X\|_2}{\sqrt{d}} - 1\right| \geq t\right) &\leq 2\exp\left(-cdK^{-4}\min\{\max\{t, t^2\}, \max\{t^2, t^4\}\}\right) \\ &\leq 2\exp\left(-cdt^2/K^4\right).\end{aligned}$$

By the equivalent definitions of sub-Gaussianity this implies the desired result. □

6.2.3 Concentration Without Independent Coordinates

Many random vectors do not have independent coordinates. So how can we handle situations like this? We begin with some definitions.

Definition 6.5. *Let X be a d-dimensional random vector. X is isotropic if $\mathbb{E}[XX^T] = I_d$.*

Note the following nice result. If Y is a random vector with mean $\mathbb{E}X = \mu$ and invertible covariance matrix Σ then, $X = \Sigma^{-1/2}(Y - \mu)$ is isotropic.
If X is isotropic, this does not necessarily mean that the coordinates of X are independent. Consider sampling the uniform distribution on a unit sphere $X \sim \sqrt{d}\,\text{Unif}(S^{d-1})$. X is isotropic since $\mathbb{E}X = 0$ and $\Sigma = I_d$. However, the coordinates of X are not independent since knowing any $d-1$ coordinates of X fully determines the remaining coordinate (up to a ± 1 sign).
To further handle the situation without independence, we'll introduce some different but closely related definitions of sub-Gaussianity in multiple dimensions. In the definitions below, let X be a d-dimensional random vector with $\mathbb{E}X = 0$.

Definition 6.6. *X is sub-Guassian if $\|X\|_{\psi_2} := \sup_{v \in S^{d-1}} \|\langle x, v\rangle\|_{\psi_2} < \infty$.*

Definition 6.7. *X is sub-Gaussian if for all $v \in S^{d-1}$, $\|\langle v, x\rangle\|_{\psi_2} \leq C\sqrt{v^T\Sigma v}$.*

Definition 6.8. *X is sub-Gaussian if for all $\lambda \in \mathbb{R}$, $v \in S^{d-1}$, $\mathbb{E}[\exp(\lambda\langle v, x\rangle)] \leq \exp\left(\frac{\lambda^2 v^T\Sigma v}{2}\right)$.*

Note that Definition 6.6 does not necessarily imply definitions 6.7 and 6.8 (one can construct simple examples using Bernoulli random variables illustrating why this is the case). Definitions 6.7 and 6.8 are equivalent up to multiplicative constants, and they both imply definition 6.6.
Further, note that in definition 6.8, Σ need not be a covariance matrix (the best case). It can also be any other "larger" positive semi-definite matrix, which works as a covariance proxy. For two positive semi-definite matrices, A and B, we say that A is larger than B if and only if $A - B$ is also a positive semi-definite matrix, and denote this as $B \preceq A$. Notice also that definitions 6.7 and 6.8 are more "variance-sensitive" than definition 6.6. We conclude this lecture with the following proposition.

Proposition 6.9. *Let X be a d-dimensional random vector that is sub-Gaussian in the sense of definition 6.8. Then, for all $\delta \in (0,1)$ we have*

$$\Pr\left(\|X\| \geq \sqrt{Tr(\Sigma)} + \sqrt{2\lambda_{\max}(\Sigma)\log(1/\delta)}\right) \leq \delta.$$

We will prove this proposition in the next lecture.

Lecture 7: Norm of a Sub-Gaussian Random Vector

Instructor: Nikita Zhivotovskiy *Scribe: Erez Buchweitz* *Proofreader: Zhiwei Xiao*

Notation

All random vectors are assumed to be column vectors. For $x, y \in \mathbb{R}^d$, the Euclidean inner product is denoted by $\langle x, y | x, y \rangle = x_1 y_1 + ... + x_d y_d = x^\top y$, and the Euclidean norm is denoted by $\|x\|_2 = \sqrt{\langle x, x | x, x \rangle} = \sqrt{x^\top x} = \sqrt{x_1^2 + ... + x_d^2}$. The unit sphere $S^{d-1} \subseteq \mathbb{R}^d$ is the set of all points which have Euclidean norm one, i.e. $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. For a random variable $X \in \mathbb{R}$ with finite variance, the L_2 norm is defined by $\|X\|_{L_2} = \sqrt{\mathbb{E}[X^2]}$. In denoting the normal distribution $\mathcal{N}_d(\mu, \Sigma)$ the subscript d indicates the dimension, e.g. implying the mean vector $\mu \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. The identity matrix in dimension d is denoted I_d.

7.1 Norm of a Sub-Gaussian Random Vector

Let $X \in \mathbb{R}^d$ be a mean-zero random vector, and denote $\Sigma = \mathbb{E}[XX^\top] \in \mathbb{R}^{d \times d}$. We recall that X is said to be ***sub-Gaussian*** if, additionally, for all $v \in S^{d-1}$ and all $\lambda \in \mathbb{R}$,

$$\mathbb{E} \exp(\lambda \langle X, v | X, v \rangle) \leq \exp\left(\frac{\lambda^2 v^\top \Sigma v}{2}\right). \tag{10}$$

Importantly, the coordinates $X_1, .., X_n$ need not be independent. Recall that since Σ is a symmetric and positive semidefinite matrix (it is a covariance matrix), it has d real eigenvalues. We denote by $\lambda_{\max}(\Sigma)$ its largest eigenvalue.

Theorem 7.1. *If X is a sub-Gaussian random vector, in particular, $\mathbb{E}X = 0$, then with probability at least $1 - \delta$,*

$$\|X\|_2 \leq \sqrt{\operatorname{Tr}\Sigma} + \sqrt{2\lambda_{max}(\Sigma)\log(1/\delta)}. \tag{11}$$

Example 7.2. *If $\Sigma = I_d$ (the identity $d \times d$ matrix), then $\operatorname{Tr}\Sigma = d$ and $\lambda_{max}(\Sigma) = 1$. We see that one of the summands in Inequality (11) is of order $\sqrt{d}$ and the other is dimension-free (does not depend on d). In general, it holds that*

$$\mathbb{E}\|X\|_2 \leq \sqrt{\mathbb{E}\|X\|_2^2} = \sqrt{\operatorname{Tr}\Sigma}. \tag{12}$$

The first transition in Inequality (12) is due to the Cauchy-Schwartz inequality, and the second is due the following important trick, where we treat $X^\top X$ as a 1×1 matrix and use the cyclical property of the trace;

$$\mathbb{E}\|X\|_2^2 = \mathbb{E}\left[X^\top X\right] = \mathbb{E}\left[\operatorname{Tr}\left(X^\top X\right)\right] = \mathbb{E}\left[\operatorname{Tr}\left(XX^\top\right)\right] = \operatorname{Tr}\mathbb{E}\left[XX^\top\right] = \operatorname{Tr}\Sigma. \tag{13}$$

The second transition in Equation (13) is due to the fact that the trace of a 1×1 matrix equals the matrix itself, the third due the cyclical property of the trace; $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$ as long as the products AB, BA are both well-defined square matrices. The fourth transition is due to the fact that Tr is linear in the entries of the matrix on which it operates. Note that Inequality (12) holds for any random vector $X \in \mathbb{R}^d$, and does not require sub-Gaussianity.

We will now state and prove a few lemmas toward proving Theorem 7.1.

7.2 Kullback-Leibler Divergence

Let ρ, π be probability densities supported on $\Theta \subseteq \mathbb{R}^d$. We first introduce the ***Kullback-Leibler divergence*** (abbreviated ***KL divergence***, also known as ***relative entropy***), which is defined by

$$KL(\rho\|\pi) = \int_\Theta \log \frac{\rho(\theta)}{\pi(\theta)} \rho(\theta) d\theta = \mathbb{E}_{\theta\sim\rho}\Big[\log \frac{\rho(\theta)}{\pi(\theta)}\Big]. \tag{14}$$

The expectation appearing in Equation (14) treats θ as a random variable with density ρ. Note that we require $\rho(\theta) = 0$ whenever $\pi(\theta) = 0$, in which case we define $\log(\rho(\theta)/\pi(\theta)) = 0$.

Fact 7.3. *1.* $KL(\rho\|\pi) \geq 0$ *2.* $KL(\rho\|\pi) = 0$ *if and only if* $\rho(\theta) = \pi(\theta)$ *almost everywhere.*

Proof. Since log is a concave function, by Jensen's inequality,

$$-KL(\rho\|\pi) = \mathbb{E}_{\theta\sim\rho}\Big[\log \frac{\pi(\theta)}{\rho(\theta)}\Big] \leq \log \mathbb{E}_{\theta\sim\rho}\Big[\frac{\pi(\theta)}{\rho(\theta)}\Big] = \log \int_\Theta \frac{\pi(\theta)}{\rho(\theta)} \rho(\theta) d\theta = \log \int_\Theta \pi(\theta) d\theta = 0.$$

The integral $\int_\Theta \pi(\theta) d\theta$ equals one because π is a probability density. From this it follows that $KL(\rho\|\pi) \geq 0$. Since log is a strictly concave function, Jensen's inequality is strict unless $\pi(\theta)/\rho(\theta)$ is almost-everywhere constant, which proves that $KL(\rho\|\pi) = 0$ if and only if $\rho(\theta) = \pi(\theta)$ almost everywhere. □

Due to Fact 7.3, we may think of KL divergence, informally, as a $distance^*$ between densities. It is not formally a distance, because it does not satisfy the triangle inequality.

7.3 Donsker-Varadhan Variational Formula

We state and prove the following ***Change of Measure lemma***.

Lemma 7.4 (Donsker-Varadhan variational formula)**.** *Let* π *be a probability density supported on* $\Theta \subseteq \mathbb{R}^d$*, and fix* $h : \Theta \to \mathbb{R}$ *be a bounded function. Then*

$$\log \mathbb{E}_{\theta\sim\pi} e^{h(\theta)} = \sup_\rho \Big\{\mathbb{E}_{\theta\sim\rho} h(\theta) - KL(\rho\|\pi)\Big\}.$$

where the supremum is taken over all probability densities ρ *such that* $KL(\rho\|\pi) < \infty$*.*

Proof. Define the probability density (check that it indeed integrates to one)

$$\pi'(\theta) = \frac{\pi(\theta) e^{h(\theta)}}{\mathbb{E}_{\theta\sim\pi} e^{h(\theta)}}.$$

where $\mathbb{E}_{\theta\sim\pi} e^{h(\theta)}$ acts as the normalizing constant. For any probability density ρ with $KL(\rho\|\pi) < \infty$, compute

$$\begin{aligned} KL(\rho\|\pi') &= \mathbb{E}_{\theta\sim\rho} \log \frac{\rho(\theta)}{\pi'(\theta)} \\ &= \mathbb{E}_{\theta\sim\rho} \log \Big(\frac{\rho(\theta)}{\pi(\theta)} \cdot \frac{\mathbb{E}_{\theta\sim\pi} e^{h(\theta)}}{e^{h(\theta)}}\Big) \\ &= KL(\rho\|\pi) + \log \mathbb{E}_{\theta\sim\pi} e^{h(\theta)} - \mathbb{E}_{\theta\sim\rho} h(\theta) \geq 0 \end{aligned}$$

Note that if we take $\rho = \pi'$ we obtain an equality, due to Fact 7.3. Rearranging, we obtain

$$\log \mathbb{E}_{\theta\sim\pi} e^{h(\theta)} \geq \mathbb{E}_{\theta\sim\rho} h(\theta) - KL(\rho\|\pi)$$

with equality obtained for $\rho = \pi'$. Since this holds for any such ρ, it must also hold for the supremum over all such ρ,

$$\log \mathbb{E}_{\theta\sim\pi} e^{h(\theta)} \geq \sup_{\rho} \left\{ \mathbb{E}_{\theta\sim\rho} h(\theta) - KL(\rho\|\pi) \right\}$$

and, again, equality is obtained for $\rho = \pi'$. □

7.4 Second Lemma

Lemma 7.5. *Fix some probability density π on Θ, and let $f(X, \theta)$ be a function with X being a random variable and $\theta \in \Theta \subseteq \mathbb{R}^d$. Then, with probability at least $1 - \delta$, it holds that for any probability density ρ on Θ for which $KL(\rho\|\pi) < \infty$,*

$$\mathbb{E}_{\theta\sim\rho} f(X, \theta) \leq \mathbb{E}_{\theta\sim\rho} \log \mathbb{E}_X e^{f(X,\theta)} + KL(\rho\|\pi) + \log(1/\delta). \tag{15}$$

The symbol $\mathbb{E}_X$ above means taking expectation with respect to X. The KL divergence term in (15) is the price we pay for wanting a bound that holds uniformly over all ρ.

Proof. Define a function $h(\theta)$ and a random variable Y_X, by

$$h(\theta) = f(X, \theta) - \log \mathbb{E}_X e^{f(X,\theta)} \qquad ; \qquad Y_X = \sup_\rho \{\mathbb{E}_{\theta\sim\rho} h(\theta) - KL(\rho\|\pi)\} = \log \mathbb{E}_{\theta\sim\pi} e^{h(\theta)}$$

where we have used Lemma 7.4. Notice that $\mathbb{E}_X e^{Y_x} = 1$, and this in fact directly implies the Lemma. Indeed,

$$\begin{aligned}
\mathbb{E}_X e^{Y_X} &= \mathbb{E}_X \mathbb{E}_{\theta\sim\pi} e^{h(\theta)} \\
&= \mathbb{E}_X \mathbb{E}_{\theta\sim\pi} [e^{f(X,\theta) - \log \mathbb{E}_x \exp(f(X,\theta))}] \\
&= \mathbb{E}_X \mathbb{E}_{\theta\sim\pi} \left[\frac{e^{f(X,\theta)}}{\mathbb{E}_X e^{f(X,\theta)}} \right] \\
&= \mathbb{E}_{\theta\sim\pi} \mathbb{E}_X \left[\frac{e^{f(X,\theta)}}{\mathbb{E}_X e^{f(X,\theta)}} \right] \\
&= \mathbb{E}_{\theta\sim\pi} \left[\frac{\mathbb{E}_X e^{f(X,\theta)}}{\mathbb{E}_X e^{f(X,\theta)}} \right] \\
&= 1.
\end{aligned}$$

Using Markov's inequality, we obtain the tail bound

$$\Pr(Y_X \geq t) = \Pr\left(e^{Y_X} \geq e^t\right) \leq \frac{\mathbb{E}_X e^{Y_X}}{e^t} = e^{-t}.$$

Plugging in $t = \log(1/\delta)$, for any $\delta \in (0, 1)$, we get $\Pr\left(Y_x \geq \log(1/\delta)\right) \leq \delta$. In other words, with probability at least $1 - \delta$,

$$\sup_\rho \{\mathbb{E}_{\theta\sim\rho} h(\theta) - KL(\rho\|\pi)\} = Y_X \leq \log(1/\delta).$$

In other words, with probability at least $1 - \delta$, it holds for all such probability densities ρ that

$$\log(1/\delta) \geq \mathbb{E}_{\theta\sim\rho} h(\theta) - KL(\rho\|\pi) = \mathbb{E}_{\theta\sim\rho} f(X, \theta) - \mathbb{E}_{\theta\sim\rho} \log \mathbb{E}_X e^{f(X,\theta)} - KL(\rho\|\pi).$$

Rearranging, we obtain Inequality 15. □

7.5 Useful Facts

Fact 7.6. *If $Y \sim \mathcal{N}_d(\mu, \sigma^2 I_d)$ and $A \in \mathbb{R}^{d\times d}$ then $\mathbb{E}[Y^\top A Y] = \sigma^2 \operatorname{Tr} A + \mu^\top A \mu$.*

Proof. We may write $Y = \mu + \sigma Z$ where $Z \sim \mathcal{N}_d(0, I_d)$, so

$$\mathbb{E}\left[Y^\top A Y\right] = \mathbb{E}\left[(\mu+\sigma Z)^\top A(\mu+\sigma Z)\right] = \mu^\top A\mu + \underbrace{2\sigma \mathbb{E}[Z]^\top A\mu}_{=0} + \sigma^2 \mathbb{E}\left[Z^\top A Z\right] = \mu^\top A\mu + \sigma^2 \mathbb{E}\left[Z^\top A Z\right].$$

It is left to use the cyclical trace trick to compute

$$\mathbb{E}\left[Z^\top A Z\right] = \mathbb{E}\operatorname{Tr}\left(Z^\top A Z\right) = \mathbb{E}\operatorname{Tr}\left(AZZ^T\right) = \operatorname{Tr}\left(A\underbrace{\mathbb{E}\left[ZZ^T\right]}_{=I_d}\right) = \operatorname{Tr} A.$$

□

Fact 7.7. *If ρ, π are the densities of $\mathcal{N}_d(v, I_d/\beta)$, $\mathcal{N}_d(0, I_d/\beta)$, respectively, and $\|v\|_2 = 1$, then $KL(\rho\|\pi) = \beta/2$.*

Proof. Observe that $\rho(x) = c(\beta)e^{-\beta\|x-v\|_2^2/2}$ and $\pi(x) = c(\beta)e^{-\beta\|x\|_2^2/2}$ where $c(\beta)$ is some constant that depends only on β. It follows that

$$\frac{2}{\beta}\log\frac{\rho(x)}{\pi(x)} = \|x\|_2^2 - \|x-v\|_2^2 = \|x\|_2^2 - (\|x\|_2^2 - 2\langle x, v|x, v\rangle + \|v\|_2^2) = 2\langle x, v|x, v\rangle - \|v\|_2^2.$$

Let $X \in \mathbb{R}^d$ be a random vector with density $\rho(x)$. By Equation (14) and $\mathbb{E}X = v$ we get

$$KL(\rho\|\pi) = \mathbb{E}\log\frac{\rho(X)}{\pi(X)} = \frac{\beta}{2}(2\langle \mathbb{E}X, v|\mathbb{E}X, v\rangle - \|v\|_2^2) = \frac{\beta}{2}(2\|v\|_2^2 - \|v\|_2^2) = \frac{\beta}{2}\underbrace{\|v\|_2^2}_{=1} = \frac{\beta}{2}.$$

□

Fact 7.8. *If $x \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric and positive semidefinite then $\sup_{v\in S^{d-1}}\langle x, v|x, v\rangle = \|x\|_2$ and $\sup_{v\in S^{d-1}} v^\top \Sigma v = \lambda_{max}(\Sigma)$.*

Proof. By the Cauchy-Schwartz inequality, for $v \in S^{d-1}$, $\langle x, v|x, v\rangle \le \|x\|_2$, and see that equality is obtained for $v = x/\|x\|_2$. Being a symmetric and positive semidefinite matrix, Σ has an orthogonal diagonalization $\Sigma = UD^2U^\top$ where $UU^\top = U^\top U = I_d$ and D^2 is a diagonal matrix with non-negative diagonal elements which are its eigenvalues. Define $u = U^\top v$, and note that $\|u\|_2 = 1$. Since $\lambda_{\max}(\Sigma)$ is the maximal diagonal element of D^2, we have

$$v^\top \Sigma v = u^\top D^2 u = \|Du\|_2^2 = \sum_{i=1}^n (Du)_i^2 = \sum_{i=1}^n (D_{ii})^2 u_i^2 = \sum_{i=1}^n (D^2)_{ii} u_i^2 \le \lambda_{\max}(\Sigma)\underbrace{\sum_{i=1}^n u_i^2}_{=\|u\|_2^2=1} = \lambda_{\max}(\Sigma).$$

Equality is obtained whenever v is an eigenvector with eigenvalue $\lambda_{\max}(\Sigma)$ (and still with norm one). □

Fact 7.9. *The function $f(x) = ax + \frac{b}{x}$ for $a, b, x > 0$ is minimized at $x^* = \sqrt{b/a}$ and has $f(x^*) = 2\sqrt{ab}$.*

Proof. The derivative $f'(x) = a - b/x^2$ equals zero for $x = \sqrt{b/a}$ and this is a minimum. □

7.6 Proof of Theorem 7.1

Proof of Theorem 7.1. Suppose that X is a d-dimensional sub-Gaussian random vector. Fix $\alpha, \beta > 0$ and $v \in S^{d-1}$ which will be determined later. Define the function $f(X, \theta) = \alpha \langle X, \theta | X, \theta \rangle$ for $\theta \in \mathbb{R}^d$. We will apply Lemma 7.5 to f along with π, ρ the densities of $\mathcal{N}_d(0, I_d/\beta)$ and $\mathcal{N}_d(v, I_d/\beta)$ respectively. Lemma 7.5 yields that with probability at least $1 - \delta$,

$$\mathbb{E}_{\theta\sim\rho} f(X, \theta) \leq \mathbb{E}_{\theta\sim\rho} \log \mathbb{E}_X e^{f(X,\theta)} + KL(\rho\|\pi) + \log(1/\delta). \tag{16}$$

We compute each expression in Inequality (16) in turn. First,

$$\mathbb{E}_{\theta\sim\rho} f(X, \theta) = \alpha \mathbb{E}_{\theta\sim\rho} \langle X, \theta | X, \theta \rangle = \alpha \left\langle X, \mathbb{E}_{\theta\sim\rho}\theta \middle| X, \mathbb{E}_{\theta\sim\rho}\theta \right\rangle = \alpha \langle X, v | X, v \rangle . \tag{17}$$

Second, using Inequality (10) and Facts 7.6 and 7.8

$$\begin{aligned}
\mathbb{E}_{\theta\sim\rho} \log \mathbb{E}_X e^{f(X,\theta)} &= \mathbb{E}_{\theta\sim\rho} \log \mathbb{E}_X e^{\alpha\langle X,\theta|X,\theta\rangle} \\
&\leq \mathbb{E}_{\theta\sim\rho} \log \exp\left(\frac{\alpha^2 \theta^\top \Sigma \theta}{2}\right) \\
&= \frac{\alpha^2}{2} \mathbb{E}_{\theta\sim\rho}[\theta^\top \Sigma \theta] \\
&= \frac{\alpha^2}{2} \mathbb{E}_{\theta\sim\rho}[(\theta - v + v)^\top \Sigma (\theta - v + v)] \\
&= \frac{\alpha^2}{2} \mathbb{E}_{\theta\sim\rho}[(\theta - v)^\top \Sigma (\theta - v) + v^\top \Sigma v + cross - terms] \\
&= \frac{\alpha^2}{2}\left(\frac{\operatorname{Tr}\Sigma}{\beta} + v^\top \Sigma v\right) \\
&\leq \frac{\alpha^2}{2}\left(\frac{\operatorname{Tr}\Sigma}{\beta} + \lambda_{\max}(\Sigma)\right)
\end{aligned} \tag{18}$$

Third, by Fact 7.7,

$$KL(\rho\|\pi) = \beta/2. \tag{19}$$

Plugging Equations (17), (18) and (19) into Inequality (16), we get that with probability at least $1 - \delta$,

$$\alpha \langle X, v | X, v \rangle \leq \frac{\alpha^2}{2}\left(\frac{\operatorname{Tr}\Sigma}{\beta} + \lambda_{\max}(\Sigma)\right) + \frac{\beta}{2} + \log(1/\delta). \tag{20}$$

Since this holds for all $v \in S^{d-1}$, it holds also for the supremum, i.e. with probability at least $1 - \delta$,

$$\begin{aligned}
\|X\|_2 = \sup_{v\in S^{d-1}} \langle X, v | X, v \rangle &\leq \frac{\alpha}{2}\left(\frac{\operatorname{Tr}\Sigma}{\beta} + \lambda_{\max}(\Sigma)\right) + \frac{\beta}{2\alpha} + \frac{\log(1/\delta)}{\alpha} \\
&= \frac{\gamma}{2}\operatorname{Tr}\Sigma + \frac{\alpha}{2}\lambda_{\max}(\Sigma) + \frac{1}{2\gamma} + \frac{\log(1/\delta)}{\alpha}.
\end{aligned}$$

where we have used Fact 7.8 after dividing both sides of Inequality (20) by α, then set $\gamma = \alpha/\beta$. Having set $\gamma = \alpha/\beta$, it is clear that we may optimize $\alpha, \gamma > 0$ independently from each other, because they always appear separately in the formula above. Using Fact 7.9 we may plug in optimal α, γ to obtain that with probability at least $1 - \delta$,

$$\|X\|_2 \leq \sqrt{\operatorname{Tr}\Sigma} + \sqrt{2\lambda_{\max}(\Sigma)\log(1/\delta)}.$$

□

7.7 Sub-Exponential Vectors

We say that a mean-zero random vector X is ***sub-exponential*** if, for any $v \in S^{d-1}$,

$$\| \langle X, v|X, v\rangle \|_{\psi_1} \leq C \, \|\langle X, v|X, v\rangle\|_{L_2}$$

where $C > 0$ is a universal constant that does not depend on v. Compare this definition of a sub-Gaussian random vector, from earlier in this lecture, noting the following fact;

Fact 7.10. *If X is a random vector with $\mathbb{E}[XX^\top] = \Sigma$ then, for any $v \in \mathbb{R}^n$, $\|\langle X, v|X, v\rangle\|_{L_2}^2 = v^\top \Sigma v$.*

Proof. $\|\langle X, v|X, v\rangle\|_{L_2}^2 = \mathbb{E}\left[(X^\top v)^2\right] = \mathbb{E}[v^\top XX^\top v] = v^\top \mathbb{E}[XX^\top]\, v = v^\top \Sigma v.$ □

The following theorem is a counterpart to Theorem 7.1, and the proof will appear in the homework.

Theorem 7.11. *If X is a sub-exponential random vector then with probability at least $1 - \delta$,*

$$\|X\|_2 \leq C(\sqrt{\mathrm{Tr}(\Sigma)\log(1/\delta)} + \log(1/\delta)\sqrt{\lambda_{max}(\Sigma)})$$

where $C > 0$ is a universal constant.

7.8 Log-Concave Densities

A density function $f(x)$ is said to be ***log-concave*** if $f(x) = e^{-\varphi(x)}$ where φ is a convex function.

Example 7.12.

- *The density of a multivariate Gaussian $\mathcal{N}_d(v, \Sigma)$ is log-concave, as up to additive and multiplicative constants $\varphi(x) \sim (x - v)^\top \Sigma^{-1}(x - v)$.*
- *The product of densities of independent exponential distributions is log-concave. Note that it is not sub-Gaussian.*
- *The uniform measure on a bounded convex open set $K \subseteq \mathbb{R}^d$ is log-concave, as*

$$\varphi(x) \sim \begin{cases} \log(Volume(K)) & x \in K \\ \infty & x \notin K. \end{cases}$$

Theorem 7.13 (Borell)**.** *If $X \in \mathbb{R}^d$ is a mean-zero random vector with log-concave density then, for any $v \in S^{d-1}$, $\| \langle X, v|X, v\rangle \|_{\psi_1} \leq C \, \|\langle X, v|X, v\rangle\|_{L_2}$ where C is a universal constant.*

Thus, a mean-zero random variable with log-concave density is sub-exponential. We do not prove this theorem.

7.9 Gaussian Concentration Inequality

A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be L-***Lipschitz*** if for all $x, y \in \mathbb{R}^d$, $|f(x) - f(y)| \leq L\|x - y\|_2$.

Theorem 7.14. *If $X \sim \mathcal{N}_d(0, I_d)$ and $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz then,*

$$\Pr(f(X) - \mathbb{E}f(X) \geq t) \leq \exp\Big(-\frac{t^2}{2L^2}\Big).$$

As usual, we may derive from this a tail bound on the absolute value :

$$\Pr(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\Big(-\frac{t^2}{2L^2}\Big).$$

Some facts will be useful.

Fact 7.15. *Let $\Sigma \in \mathbb{R}^{d\times d}$ be a symmetric and positive semidefinite matrix. Then there exists a unique symmetric and positive semidefinite matrix $\Sigma^{\frac{1}{2}}$ such that $(\Sigma^{\frac{1}{2}})^2 = \Sigma$.*

Proof. Being a symmetric and positive semidefinite matrix, Σ has an orthogonal diagonalization $\Sigma = UDU^\top$ where $UU^\top = U^\top U = I_d$ and D is a diagonal matrix with non-negative diagonal elements. Define $\Sigma^{\frac{1}{2}} = UD^{\frac{1}{2}}U^T$, where $D^{\frac{1}{2}}$ is a diagonal matrix with $(D^{\frac{1}{2}})_{ii} = \sqrt{D_{ii}}$, and check that it satisfies the requirements. □

Fact 7.16. *For any $x, y \in \mathbb{R}^d$, $|\|x\|_2 - \|y\|_2| \leq \|x - y\|_2$.*

Proof. Using the triangle inequality, $\|x\|_2 = \|(x - y) + y\|_2 \leq \|x - y\|_2 + \|y\|_2$. Rearranging, we get $\|x\|_2 - \|y\|_2 \leq \|x-y\|_2$. Repeating this argument with the roles of x, y reversed, the proof is concluded. □

Corollary 7.17. *If $X \sim \mathcal{N}_d(0, \Sigma)$ then with probability at least $1 - \delta$,*

$$|\|X\|_2 - \mathbb{E}\|X\|_2| \leq \sqrt{2\lambda_{max}(\Sigma)\log(2/\delta)}.$$

Proof. Using Fact 7.15, let $Z \sim \mathcal{N}_d(0, I_d)$ be such that $X = \Sigma^{\frac{1}{2}}Z$ (can take $Z = (\Sigma^{\frac{1}{2}})^{-1}X$ if Σ is full rank). Set $f(Z) = \|\Sigma^{\frac{1}{2}}Z\|_2$, and see that it is $\sqrt{\lambda_{\max}(\Sigma)}$-Lipschitz. Indeed, for $x, y \in \mathbb{R}^d$ we denote $v = (x - y)/\|x - y\|_2 \in S^{d-1}$ and compute

$$|f(x)-f(y)| = |\|\Sigma^{\frac{1}{2}}x\|_2-\|\Sigma^{\frac{1}{2}}y\|_2| \leq \|\Sigma^{\frac{1}{2}}(x-y)\|_2 = \|x-y\|_2\|\Sigma^{\frac{1}{2}}v\|_2 = \|x-y\|_2\sqrt{v^\top\Sigma v} \leq \|x-y\|_2\sqrt{\lambda_{\max}(\Sigma)}$$

with the first inequality due to Fact 7.16 and the second due to Fact 7.8. As usual, set $2\exp\Big(-\frac{t^2}{2(\sqrt{\lambda_{\max}\Sigma})^2}\Big) = \delta$. Plugging in $t = \sqrt{2\lambda_{\max}(\Sigma)\log(2/\delta)}$ into Theorem 7.14, for any $\delta \in (0, 1)$, we get that with probability at least $1 - \delta$,

$$|\|X\|_2 - \mathbb{E}\|X\|_2| \leq \sqrt{2\lambda_{\max}(\Sigma)\log(2/\delta)}.$$

□

Lecture 8: Gaussian Concentration & Fixed Design Linear Regression

Instructor: Nikita Zhivotovskiy *Scribe: Rita Lyu* *Proofreader: Erez Buchweitz*

8.1 Notation

All random vectors are assumed to be column vectors. For $x, y \in \mathbb{R}^d$, the Euclidean inner product is denoted by $\langle x, y | x, y \rangle = x_1 y_1 + ... + x_d y_d = x^\top y$, and the Euclidean norm is denoted by $\|x\|_2 = \sqrt{\langle x, x | x, x \rangle} = \sqrt{x^\top x} = \sqrt{x_1^2 + ... + x_d^2}$. The unit sphere $S^{d-1} \subseteq \mathbb{R}^d$ is the set of all points which have Euclidean norm one, i.e. $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. For a random variable $X \in \mathbb{R}$ with finite variance, the L_2 norm is defined by $\|X\|_{L_2} = \sqrt{\mathbb{E}[X^2]}$. In denoting the normal distribution $\mathcal{N}_d(\mu, \Sigma)$ the subscript d indicates the dimension, e.g. implying the mean vector $\mu \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. The identity matrix in dimension d is denoted I_d.

8.2 Gaussian Concentration

Theorem 8.1. *If $X \sim \mathcal{N}_d(0, I_d)$ and $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz then,*

$$\Pr(f(X) - \mathbb{E}f(X) \geq t) \leq \exp\Big(-\frac{t^2}{2L^2}\Big). \tag{21}$$

Remark 8.2. *We can see the right-hand side term does not contain dimension term d, which means that this inequality always holds for L-Lipschitz function regardless of the dimension. This theorem also indicates that $f(X) - \mathbb{E}f(X)$ is subgaussian.*

As usual, we may derive from this a tail bound on the absolute value :

$$\Pr(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\Big(-\frac{t^2}{2L^2}\Big). \tag{22}$$

To prove this Theorem, we need Fact 8.11 and Lemma 8.3.

Lemma 8.3. *For any convex function $\varphi : \mathbb{R} \to \mathbb{R}$ and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:*

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

Remark 8.4. *The trick used in the proof of this lemma is quite helpful. We upper bound the expectation using a mixture of Gaussians.*

Gaussian concentration can be extended to some log-concave measures. For example, strongly log-concave measures. Let us introduce the following definition.

Definition 8.5 (K-strongly convexity and K-strongly log-concavity)**.** *A differentiable function is K-strongly convex, for $K > 0$, if*

$$\left[\frac{\partial^2(\varphi(X))}{\partial X_i \partial X_j}\right]_{d \times d} \geq K I_d.$$

A measure is K-strongly log-concave if $f(X) = \exp(-\varphi(X))$, where $\varphi(X)$ is K-strongly convex.

With Definition 8.5, we can see $\exp\left(\frac{-\|X\|_2^2}{2}\right)$ is 1-strongly log-concave, while $\exp\left(\frac{-||X||_1}{2}\right)$ is concave but cannot find K to make it K-strongly log-concave. Because when $d = 1$, it can be regarded as the product of independent exponential distributions, the second derivative is 0. We have another theorem without proof that

Theorem 8.6. *For K-strongly log concave distributions (X is distributed according to $f(X) = \exp(-\varphi(X))$), Eq (21) and Eq (22) hold by replacing L to $\sim \frac{L}{\sqrt{K}}$, where $\sim$ means "in a proportion to". When K gets larger, the bound becomes tighter.*

In Lecture 7, we proved Theorem 8.7:

Theorem 8.7. *If X is a sub-Gaussian random vector, in particular, $\mathbb{E}X = 0$, then with probability at least $1 - \delta$,*

$$\|X\|_2 \leq \sqrt{\operatorname{Tr}\Sigma} + \sqrt{2\lambda_{max}(\Sigma)\log(1/\delta)}. \tag{23}$$

Example 8.8 (Multivariate Mean Estimation). *$X_1, \cdots X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \Sigma)$, we now use $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ to estimate the mean vector μ. Here we now*

$$\mathbb{C}ov(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{\Sigma}{n}.$$

With Theorem 8.1, we know $\hat{\mu} - \mu$ is subgaussian. Then, combining Theorem 8.7, we have with probability at least $1 - \delta$.

$$\|\frac{1}{n}\sum_{i=1}^n X_i - \mu\|_2 \leq \sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{2\lambda_{max}(\Sigma)\log(1/\delta)}{n}}.$$

8.3 Fixed Design Linear Regression Model

Let $x_i \in \mathbb{R}, i = 1, \cdots, n$ be fixed (we can regard it as d features for the i-th individual), $\beta^* \in \mathbb{R}^d$, $\xi_i, i = 1, \cdots n$ be independent zero mean σ-subgaussain variables. That is

$$\mathbb{E}\exp(\lambda\xi_i) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

The underlying data generating process is that

$$y_i = \langle x_i, \beta^* \rangle + \xi_i, i = 1, \cdots, n.$$

We do not know the true β^*. Instead, we can observe pairs of $\{(x_i, y_i)\}_{i=1}^n$. We use $\hat{\beta}$ to estimate β^* based on the observed data. To measure the estimation error, we have (i)Euclidean norm: $\|\hat{\beta} - \beta^*\|_2$, (ii) denoising error: (empirical sample)

$$\frac{1}{n}\sum_{i=1}^n(\langle x_i, \hat{\beta}\rangle - \langle x_i, \beta^*\rangle)^2 = \frac{1}{n}\sum_{i=1}^n(\hat{\beta} - \beta^*)^\top x_i x_i^\top(\hat{\beta} - \beta^*).$$

These two measures return the same results only when $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top = I_d$. We now define $Y = [y_1, \cdots, y_n]^\top \in \mathbb{R}^n$, $X = [x_1, \cdots, x_n]^\top \in \mathbb{R}^{n\times d}$, $\xi = [\xi_1, \cdots, \xi_n]^\top \in \mathbb{R}^n$. The linear model can be rewritten as:

$$Y = X\beta^* + \xi.$$

8.3.1 Ordinary Least Squares Estimator

The Ordinary Least Squares estimator (OLS) solves the optimization problem that

$$\hat{\beta}_{\text{OLS}} = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\langle x_i, \beta \rangle - y_i)^2 = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \|X\beta - Y\|_2^2,$$

assuming $\texttt{rank}(X^\top X) = d$ such that $X^\top X$ is invertable, then

$$\begin{aligned} \hat{\beta}_{\text{OLS}} &= (X^\top X)^{-1} X^\top Y \in \mathbb{R}^d \\ X\hat{\beta}_{\text{OLS}} &= \underbrace{X(X^\top X)^{-1} X^\top}_{\text{projector}} Y. \end{aligned}$$

The property of the orthogonal projection matrix contains

- symmetric and positive semidefinite positive $(X(X^\top X)^{-1}X^\top)^\top = (X(X^\top X)^{-1}X^\top) \in \mathbb{R}^{n \times n}$,
- idempotent $(X(X^\top X)^{-1}X^\top)^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top$,
- $\texttt{rank}(X(X^\top X)^{-1}X^\top) = d$,
- $\operatorname{Tr}\big(X(X^\top X)^{-1}X^\top\big) = d$,
- The eigenvalues of $X(X^\top X)^{-1}X^\top$ consist of d ones and $n - d$ zeros.

We now look at the population denoising error

$$\mathbb{E}_\xi \frac{1}{n} \sum_{i=1}^{n} (\langle x_i, \hat{\beta} \rangle - \langle x_i, \beta^* \rangle)^2 \tag{24}$$

$$= \frac{1}{n} \mathbb{E} \|X\hat{\beta} - X\beta^*\|_2^2 = \frac{1}{n} \mathbb{E} \|X(X^\top X)^{-1}X^\top Y - X\beta^*\|_2^2 \tag{25}$$

$$= \frac{1}{n} \mathbb{E} \|\underbrace{X(X^\top X)^{-1}X^\top}_{:=A} \xi\|_2^2 \tag{26}$$

$$= \frac{1}{n} \mathbb{E} \operatorname{Tr}\big(\xi^\top A \xi\big) \tag{27}$$

$$= \frac{1}{n} \mathbb{E} \operatorname{Tr}\big(A\xi\xi^\top\big) \quad \operatorname{Tr}(AB) = \operatorname{Tr}(BA), \text{ the dimensions are such that both } AB \text{ and } BA \text{ are well defined} \tag{28}$$

$$= \frac{1}{n} \operatorname{Tr}\big(A\mathbb{E}[\xi\xi^\top]\big), \quad A \text{ is a fixed matrix.} \tag{29}$$

$$\leq \frac{\sigma^2}{n} d, \quad \text{use Fact 8.12 and the subgaussian property } \mathbb{C}\text{ov}(\xi) \preceq \sigma^2 I_n. \tag{30}$$

8.3.2 Oracle Inequalities

Now we transfer from the unconstrained linear regression to the constrained case. We use $\mathcal{K}$ to denote a convex closed set in $\mathbb{R}^d$. The data-generating process is still

$$Y = X\beta^* + \xi,$$

we do not assume β^* belongs to $\mathcal{K}$. Now the estimated $\hat{\beta}$ within $\mathcal{K}$ is obtained by

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathcal{K}} \frac{1}{n} \|X\beta - Y\|_2^2.$$

We want to bound the denoising error as in Eq (24).

Theorem 8.9. *For the denoising error, we have*

$$\mathbb{E} \frac{1}{n} \|X\hat{\beta} - X\beta^*\|_2^2 \leq \min_{\beta \in \mathcal{K}} \frac{1}{n} \|X\beta - X\beta^*\|_2^2 + \frac{4\sigma^2 d}{n}.$$

Remark 8.10. *The term $\min_{\beta \in \mathcal{K}} \frac{1}{n} \|X\beta - X\beta^*\|_2^2$ means "the best possible solution with in $\mathcal{K}$" and can be regarded as theoretically best action.*

8.4 Useful Facts

Fact 8.11. *Assume additionally that L-Lipschitz function f is diffferentiable, then $\forall X \in \mathbb{R}^d$, $\|\nabla f(X)\|_2 \leq L$.*

Fact 8.12. *If both A, B, and C are positive semi-definite (PSD) matrix, $B \preceq C$, then*

$$\operatorname{Tr}(AB) \leq \operatorname{Tr}(AC).$$

Proof. Because A is PSD, then $A = A^{\frac{1}{2}} A^{\frac{1}{2}}$, we have

$$\operatorname{Tr}(AB) = \operatorname{Tr}\left(A^{\frac{1}{2}} B A^{\frac{1}{2}}\right), \operatorname{Tr}(AC) = \operatorname{Tr}\left(A^{\frac{1}{2}} C A^{\frac{1}{2}}\right).$$

For arbitrary vector x, because $B \preceq C$,

$$x^\top A^{\frac{1}{2}} B A^{\frac{1}{2}} x \leq x^\top A^{\frac{1}{2}} C A^{\frac{1}{2}} x.$$

Now we choose $e_i = (0, \cdots, \underbrace{1}_{i\text{-th coordinate}}, \cdots, 0), i = 1, \cdots, n$, then

$$\operatorname{Tr}(AB) = \operatorname{Tr}\left(A^{\frac{1}{2}} B A^{\frac{1}{2}}\right) = \sum_{i=1}^n e_i^\top \frac{1}{2} A^{\frac{1}{2}} B A^{\frac{1}{2}} e_i \leq \sum_{i=1}^n e_i^\top A^{\frac{1}{2}} C A^{\frac{1}{2}} e_i = \operatorname{Tr}\left(A^{\frac{1}{2}} C A^{\frac{1}{2}}\right) = \operatorname{Tr}(AC).$$

□

8.5 Proof of Theorem 8.1

Proof of Theorem 8.1. Assuming additionally that f is differentiable, combining with Lemma 8.3, we now prove the Gaussian concentration inequality. Let us use the lemma with $\varphi(\cdot) = \exp(\lambda \cdot)$.

$$\begin{aligned}
\mathbb{E}_{X,Y}\left[\exp(\lambda(f(X) - \mathbb{E}[f(X)]))\right] &\leq \mathbb{E}_{X,Y}\left[\exp\left(\frac{\lambda\pi}{2}\langle Y, \nabla f(X)\rangle\right)\right] \quad \text{(Lemma 8.3)} \\
&= \mathbb{E}_X\left[\prod_{i=1}^n \mathbb{E}_{Y_i}\left[\exp\left(\frac{\lambda\pi}{2}(\nabla f(X))_i Y_i\right)\right]\right]
\end{aligned}$$

$$= \mathbb{E}_X \left[\exp\left(\lambda^2 \frac{\pi^2}{4} ||\nabla f(x)||^2/2 \right) \right]$$

Here, we use that $\nabla f(X))_i$ is a constant and Y_i is standard normal.

$$\mathbb{E}_{X,Y} \left[\exp(\lambda(f(X) - \mathbb{E}[f(X)])) \right] \leq \exp\left(\frac{\lambda^2 \pi^2}{8} L^2 \right), \quad \text{(Fact 8.11)}$$

which shows that $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with the parameter at most $\frac{\pi L}{2}$,
The tail bound then can be

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\pi^2 L^2} \right) \text{ for all } t \geq 0.$$

□

Remark 8.13. *This is the cleanest and easiest way of proving such an inequality and results in a weaker bound (difference only occurs in constant).*

8.6 Proof of Lemma 8.3

Proof of Lemma 8.3. Because X and Y have the same distribution and $\mathbb{E}_X[f(X)]$ is a constant, replacing it with $\mathbb{E}_Y[f(Y)]$ and apply the Jensen's inequality because $\varphi(\cdot)$ is a convex function, then

$$\mathbb{E}_X \left[\varphi \left(f(X) - \mathbb{E}_X[f(X)] \right) \right] = \mathbb{E}_X \left[\varphi \left(f(X) - \mathbb{E}_Y[f(Y)] \right) \right] \leq \mathbb{E}_{X,Y}[\varphi(f(X) - f(Y))]. \tag{31}$$

Define the following random variable $Z \in \mathbb{R}^n$,

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

For each coordinate,

$$Z_k(\theta) = X_k \sin\theta + Y_k \cos\theta.$$

The variable $Z(\theta)$ can be thought of as a path between X and Y. In fact, when $\theta = 0$ we get $Z(\theta) = Y$, while if $\theta = \frac{\pi}{2}$ we get $Z(\theta) = X$. Therefore, as θ varies in the interval $[0, \frac{\pi}{2}]$ we are moving from X to Y. Use Z' to denote the derivative of Z w.r.t. θ, that is $Z' = \cos\theta X - Y \sin\theta$. The random variable Z has some nice properties

$$\forall \theta \in [0, \frac{\pi}{2}] \quad Z(\theta) \overset{d}{=} X \overset{d}{=} Y, \quad Z'(\theta) \overset{d}{=} X \overset{d}{=} Y, \quad Z(\theta) \text{ and } Z'(\theta) \text{are independent.}$$

First, for a fixed θ, $Z(\theta)$ is a linear combination of two standard normals, then $\mathbb{E}[Z(\theta)] = 0$ and $\mathbb{V}(Z(\theta)) = \mathbb{V}(X)\sin\theta^2 + \mathbb{V}(Y)\cos\theta^2 = \left(\sin\theta^2 + \cos\theta^2\right) I_n = I_n$ showing that $Z(\theta) \sim N(0, I_n)$. Consider now $Z'(\theta) = X\cos\theta - Y\sin\theta$. Using a similar reasoning we can show that $Z'(\theta) \sim N(0, I_n)$. Because both $Z(\theta)$ and $Z(\theta)'$ are normally distributed, we can show they are independent by just checking their covariance is 0. Independence comes from the fact that

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

Now since $Z_k(0) = Y_k$ and $Z_k(\pi/2) = X_k$ for all $k = 1, \ldots, n$, we have

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

where $Z'(\theta)$ in $\mathbb{R}^n$ denotes the elementwise derivative, a vector with the components $Z'_k(\theta) = X_k \cos(\theta) - Y_k \sin(\theta)$. Note that this integral may be reinterpreted as an expectation over $\theta \sim U[0, \pi/2]$,

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

Substituting the integral representation into our earlier bound Eq (31) which implies

$$\begin{aligned}
\mathbb{E}_{X,Y}[\varphi(f(X) - f(Y))] &= \mathbb{E}_{X,Y} \left[\varphi \left(\mathbb{E}_\theta \left[\frac{\pi}{2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle \right] \right) \right] \\
&\leq \mathbb{E}_{X,Y} \mathbb{E}_\theta \left[\varphi \left(\frac{\pi}{2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle \right) \right] \quad \text{(Jensen)} \\
&= \mathbb{E}_\theta \mathbb{E}_{X,Y} \left[\varphi \left(\frac{\pi}{2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle \right) \right] \quad \text{(Fubini)} \\
&= \mathbb{E}_\theta \mathbb{E}_{X,Y} \left[\varphi \left(\frac{\pi}{2} \langle \nabla f(X), Y \rangle \right) \right] \quad \text{because} \quad (Z(\theta), Z'(\theta)) \overset{d}{=} (X, Y) \\
&= \mathbb{E}_{X,Y} \left[\varphi \left(\frac{\pi}{2} \langle \nabla f(X), Y \rangle \right) \right],
\end{aligned}$$

where the equality before the last exploits the fact that θ is fixed inside the inner expectation.

□

8.7 Proof of Theorem 8.9

Proof of Theorem 8.9. We have

$$\|X\hat{\beta} - Y\|_2^2 \leq \|X\widetilde{\beta} - Y\|_2^2, \ \widetilde{\beta} \text{ is the best theoretical solution we might get within } \mathcal{K} .$$

Opening the bracket, we have

$$\begin{aligned}
&\|X\hat{\beta} - Y\|_2^2 = \|X\hat{\beta} - X\beta^* - \xi\|_2^2 = \|X\hat{\beta} - X\beta^*\|_2^2 + \|\xi\|_2^2 - 2\langle X\hat{\beta} - X\beta^*, \xi \rangle \\
&\|X\widetilde{\beta} - Y\|_2^2 = \|X\widetilde{\beta} - X\beta^* - \xi\|_2^2 = \|X\widetilde{\beta} - X\beta^*\|_2^2 + \|\xi\|_2^2 - 2\langle X\widetilde{\beta} - X\beta^*, \xi \rangle \\
&= \|X\widetilde{\beta} - X\beta^*\|_2^2 + \|\xi\|_2^2 - 2\langle X\hat{\beta} - X\beta^*, \xi \rangle + 2\langle X\hat{\beta} - X\widetilde{\beta}, \xi \rangle \\
&\Rightarrow \|X\hat{\beta} - X\beta^*\|_2^2 \leq \|X\widetilde{\beta} - X\beta^*\|_2^2 + 2\langle X\hat{\beta} - X\widetilde{\beta}, \xi \rangle.
\end{aligned}$$

Then what we need to analyze is the second term $2\langle X\hat{\beta} - X\widetilde{\beta}, \xi \rangle$. By Cauchy-Schwarz inequality,

$$2\langle X\hat{\beta} - X\widetilde{\beta}, \xi \rangle = 2\|X\hat{\beta} - X\widetilde{\beta}\|_2 \langle \frac{X\hat{\beta} - X\widetilde{\beta}}{\|X\hat{\beta} - X\widetilde{\beta}\|_2}, \xi \rangle \leq 2\|X\hat{\beta} - X\widetilde{\beta}\|_2 \|\xi\|_2.$$

Then we apply the equality we have used many times

$$ab \leq \frac{1}{2}(\frac{a^2}{\gamma} + b^2\gamma), a \geq 0, b \geq 0.$$

And here we set $\gamma = 2$, then

$$2\|X\hat{\beta} - X\widetilde{\beta}\|_2 \|\xi\|_2 \leq \frac{\|X\hat{\beta} - X\widetilde{\beta}\|_2^2}{2} + 2\|\xi\|_2^2.$$

Finally, we analyze $\|X\hat{\beta} - X\widetilde{\beta}\|_2^2$, by the convexity of the set $\mathcal{K}$, we know that

$$\begin{aligned}
&\|X\hat{\beta} - X\beta^*\|_2^2 \geq \|X\hat{\beta} - X\widetilde{\beta}\|_2^2 + \|X\beta^* - X\widetilde{\beta}\|_2^2 \\
&\Rightarrow \|X\hat{\beta} - X\widetilde{\beta}\|_2^2 \leq \|X\hat{\beta} - X\beta^*\|_2^2 - \|X\beta^* - X\widetilde{\beta}\|_2^2.
\end{aligned}$$

Plugging this into the previous inequality, we have

$$\begin{aligned}
&\|X\hat{\beta} - X\beta^*\|_2^2 \leq 2\|\xi\|_2^2 + \frac{\|X\hat{\beta} - X\beta^*\|_2^2}{2} - \frac{\|X\beta^* - X\widetilde{\beta}\|_2^2}{2} + \|X\beta^* - X\widetilde{\beta}\|_2^2 \\
&\Rightarrow \|X\hat{\beta} - X\beta^*\|_2^2 \leq 4\|\xi\|_2^2 + \|X\beta^* - X\widetilde{\beta}\|_2^2 \\
&\Rightarrow \frac{1}{n}\mathbb{E}\|X\hat{\beta} - X\beta^*\|_2^2 \leq \frac{1}{n}\mathbb{E}4\|\xi\|_2^2 + \frac{1}{n}\mathbb{E}\|X\beta^* - X\widetilde{\beta}\|_2^2 \\
&\Rightarrow \frac{1}{n}\mathbb{E}\|X\hat{\beta} - X\beta^*\|_2^2 \leq \frac{4\sigma^2 d}{n} + \frac{1}{n}\|X\beta^* - X\widetilde{\beta}\|_2^2.
\end{aligned}$$

□

Lecture 9: Fixed Design and Sparse Linear Regression

Instructor: Nikita Zhivotovskiy *Scriber: Toby Kreiman* *Proofreader: Rita Lyu*

9.1 Fixed Design Linear Regression

Recall that fixed design linear regression consists of predicting targets y_i from fixed vectors $x_i \in \mathbb{R}^d$ for $i \in [1, n]$, where $y_i = \langle x_i, \beta^* \rangle + \xi_i$. ξ_i is zero mean and $\sigma-$subgaussian random noise. We can stack the features as rows into a matrix $X \in \mathbb{R}^{n \times d}$ and write the above in matrix form:

$$Y = X\beta^* + \xi.$$

Let $K \subseteq \Delta^d$ where $\Delta^d = \{x \in \mathbb{R}^d_+ : \sum_{i=1}^d |x_i| \leq 1\}$ is the d dimensional simplex. Further, let

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in K} ||X\beta - Y||_2^2.$$

This is a constrained least squares problem. We wish to analyze:

$$\mathbb{E}\frac{1}{n}||X\hat{\beta} - X\beta^*||_2^2.$$

In order to do so, recall that last time we showed that:

$$||X\hat{\beta} - X\beta^*||_2^2 \leq ||X\beta - X\beta^*||_2^2 + 2\langle X\hat{\beta} - X\beta, \xi \rangle \tag{32}$$

for all $\beta \in K_0$ where K_0 was some arbitrary subspace. In this case, since K is the simplex, we can see that $\hat{\beta} - \beta \in B_1^d$, where B_1^d is the unit ball with respect to the l_1 distance. Thus we can bound the second term in the above inequality:

$$2\langle X\hat{\beta} - X\beta, \xi \rangle \leq 2 \sup_{v \in B_1^d} \langle v, X^\intercal \xi \rangle,$$

where we also used the fact that $\langle Xa, b \rangle = a^\intercal X^\intercal b = \langle a, X^\intercal b \rangle$. We can recognize the right hand side as the definition for the $\infty-$norm using its dual l_1 norm. Thus we get:

$$2\langle X\hat{\beta} - X\beta, \xi \rangle \leq 2||X^\intercal \xi||_\infty.$$

We now analyze $\mathbb{E}||X^\intercal \xi||_\infty$. We introduce $a = X^\intercal \xi \in \mathbb{R}\mathbb{d}$ for notational convenience. We note that $a_i = \langle X_{(i)}, \xi \rangle$, where $X_{(i)}$ represents the ith **column** of X (since we have an inner product between the transpose of X and ξ). Therefore, a_i is also subgaussian with parameter $\sigma||X_{(i)}||_2$. By definition:

$$||X^\intercal \xi||_\infty = \max\{a_1, \ldots, a_n, -a_1, \ldots, -a_n\}.$$

Using the fact that a_i is subgaussian with parameter $\sigma||X_{(i)}||_2$ and the max inequality derived in lecture 5, we can say that:

$$\mathbb{E}||X^\intercal \xi||_\infty \leq \sqrt{2\log(2d)}\sigma \max ||X_{(i)}||_2.$$

Finally, putting it all together, we can say that:

$$\mathbb{E}\frac{1}{n}||X\hat{\beta} - X\beta^*||_2^2 \leq \inf_{\beta \in K} \frac{1}{n}||X\beta - X\beta^*||_2^2 + \frac{2\sqrt{2\log(2d)}\sigma \max ||X_{(i)}||_2}{n}.$$

For reference, we note that typically we assume that $||X_{(i)}||_2 \leq \sqrt{n}$.

9.2 Sparse Linear Regression

We are interested in cases where the solution depends only on a sparse subset of the features. This time, let:

$$K = \{x : ||x||_0 \leq s\},$$

where $||x||_0$ is the number of non-zero coordinates of x. We assume that $s \ll d$ and that $\beta^* \in K$. In general this a difficult non-convex problem to compute. By equation 32,

$$||X\hat{\beta} - X\beta^*||_2^2 \leq 2\langle X\hat{\beta} - X\beta^*, \xi\rangle,$$

where we drop the infimum since $\beta^* \in K$. We can divide by $||X\hat{\beta} - X\beta^*||_2$ on both sides to obtain:

$$||X\hat{\beta} - X\beta^*||_2 \leq 2\langle \frac{X\hat{\beta} - X\beta^*}{||X\hat{\beta} - X\beta^*||_2}, \xi\rangle.$$

Note that since $\beta^*, \hat{\beta}$ are both sparse, $||\hat{\beta} - \beta^*||_0 \leq 2s$. Therefore, we can consider $\frac{X(\hat{\beta}-\beta^*)}{||X(\hat{\beta}-\beta^*)||_2}$ as an orthogonal projector onto some subset S of magnitude $|S| \leq 2s$ of the columns of the matrix X (alternatively we could use Cauchy-Schwartz to analyze this but it does not give a good upper bound since we don't take advantage of the sparseness). We call that matrix A_S. Therefore, if we take a maximum over all such sets S, we can say:

$$||X\hat{\beta} - X\beta^*||_2 \leq 2 \max_{S\subseteq[d],|S|\leq 2s} ||A_S\xi||_2,$$

implying that (by squaring both sides)

$$||X\hat{\beta} - X\beta^*||_2^2 \leq 4 \max_{S\subseteq[d],|S|\leq 2s} ||A_S\xi||_2^2.$$

We now fix $S \subseteq [d]$ with $|S| \leq 2s$ and check the subgaussianity of $A_S\xi$. Fix $v \in S^{d-1}$ and $\lambda > 0$. Then:

$$\mathbb{E}\exp(\lambda\langle A_S\xi, v\rangle) = \mathbb{E}\exp(\lambda\langle \xi, A_S v\rangle),$$

since $A_S = A_S^\intercal$ since it is an orthogonal projection matrix. We know that ξ is subgaussian, therefore:

$$\mathbb{E}\exp(\lambda\langle \xi, A_S v\rangle) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}||A_S v||_2^2\right).$$

Again using the fact that A_S is a projector and $A_S^2 = A_S$,

$$\exp\left(\frac{\lambda^2\sigma^2}{2}||A_S v||_2^2\right) = \exp\left(\frac{\lambda^2\sigma^2}{2} v^\intercal A_S v\right),$$

showing that $A_S\xi$ is a subgaussian vector. Therefore, we can use the inequality for the norm of a subgaussian vector from Lecture 7 Theorem 1 to say that with probability $1 - \delta$:

$$||A_S\xi||_2 \leq \sigma(\sqrt{\mathrm{Tr}\,(A)_S} + \sqrt{2\lambda_{max}(A_S)\log\left(\frac{1}{\delta}\right)}) \leq \sigma(\sqrt{2s} + \sqrt{2\log\left(\frac{1}{\delta}\right)}),$$

since we know that for a projection matrix of size $2s$, $\mathrm{Tr}\,(A)_S \leq 2s$ and $\lambda_{max}(A_S) = 1$.
In order to bound $\max_{S\subseteq[d],|S|\leq 2s} ||A_S\xi||_2^2$, we apply a union bound over all sets $S \in [d]$ where $|S| \leq 2s$. We call M the number of sets $S \in [d]$ where $|S| \leq 2s$:

$$M = \sum_{j=0}^{2s}\binom{d}{j} \leq \sum_{j=0}^{2s}\binom{d}{j}\left(\frac{d}{2s}\right)^{2s-j},$$

where the last inequality holds because we are multiplying by a number greater than 1 since $2s \ll d$. We can sum over more positive terms and write:

$$\sum_{j=0}^{2s} \binom{d}{j} \left(\frac{d}{2s}\right)^{2s-j} \leq \left(\frac{d}{2s}\right)^{2s} \sum_{j=0}^{d} \binom{d}{j} \left(\frac{2s}{d}\right)^{j},$$

where we also factor out $(\frac{d}{2s})^{2s}$. We can recognize this term and use the binomial theorem to get:

$$M \leq \left(\frac{d}{2s}\right)^{2s} \sum_{j=0}^{d} \binom{d}{j} \left(\frac{2s}{d}\right)^{j} \leq \left(\frac{d}{2s}\right)^{2s} \left(1 + \frac{2s}{d}\right)^{d} \leq \left(\frac{ed}{2s}\right)^{2s},$$

where we use that $1 + x \leq e^x$ in the last step. Applying the union bound over all sets S, we get that with probability $1 - \delta$,

$$\max_{S \subseteq [d], |S| \leq 2s} ||A_S \xi||_2 \leq \sigma(\sqrt{2s} + \sqrt{2(2s \log\left(\frac{ed}{2s}\right) + \log\left(\frac{1}{\delta}\right)}).$$

Putting it all together, with probability $1 - \delta$

$$\frac{1}{n}||X\hat{\beta} - X\beta^*||_2^2 \leq \frac{C\sigma^2(s \log\left(\frac{ed}{2s}\right) + \log\frac{1}{\delta})}{n},$$

for some constant C. Importantly, note the better dependence of $n \geq s \log \frac{d}{s}$ instead of the more traditional bound of $n \geq d$.

9.3 Matrices and their Concentrations

We begin by reviewing a few useful definitions. Let $A \in \mathbb{R}^{m \times n}$ be a (non-random) matrix.

Definition 9.1 (SVD)**.** *Singular Value Decomposition (SVD) for a matrix A is defined as:*

$$A = \sum_{i=1}^{rank(A)} \sigma_i u_i v_i^\intercal,$$

where $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ are the ordered singular values of the matrix A. The u_i, v_i form an orthonormal basis for $AA^\intercal, A^\intercal A$ respectively and the singular values are the square root of the eigenvalues of $AA^\intercal$ and $A^\intercal A$.

Remark 9.2. *For a square matrix $M \in \mathbb{R}^{m \times m}$, the inverse can be written as:*

$$M^{-1} = \sum_{i=1}^{m} \sigma_i^{-1} v_i u_i^\intercal.$$

Definition 9.3 (Operator Norm)**.** *The operator norm of a matrix A is defined as:*

$$||A||_{op} = \sup_{v \in S^{n-1}} ||Av||_2 = \sup_{u \in S^{m-1}, v \in S^{n-1}} u^\intercal A v.$$

Definition 9.4 (Frobenius Norm)**.** *The Frobenius (or sometimes called Hilbert–Schmidt operator) norm of a matrix A is defined as:*

$$||A||_F = \sqrt{\sum_{i,j} A_{ij}^2}$$

.

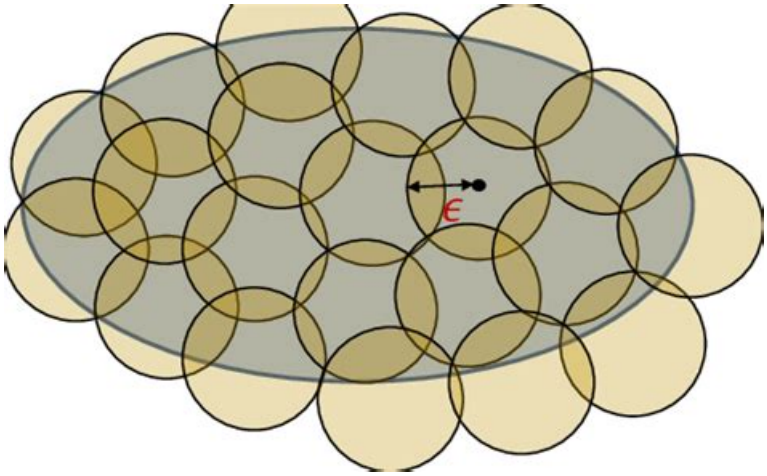


Figure 4: Each point in the grey oval is covered by one of the ε balls.

Remark 9.5. *The operator norm is the maximum singular value:*

$$||A||_{op} = \sigma_1.$$

The Frobenius norm is square root of the sum of the squared singular values (norm of singular values vector):

$$||A||_F = \sqrt{\sum_i \sigma_i^2}.$$

With these equalities in hand, we can use singular values to turn a matrix bound into a vector bound of singular values.

9.4 Covering and Packing Numbers

We review some more useful definitions.

Definition 9.6 ($\varepsilon-$Cover)**.** *Let K be a subset of $\mathbb{R}^d$. An $\varepsilon-$cover with respect to the distance ρ is the set $N_\varepsilon \subseteq K$ such that $\forall x \in K, \exists x_0 \in N_\varepsilon$ such that $\rho(x, x_o) \leq \varepsilon$. See figure 4.*

Definition 9.7 (Covering Number)**.** *The cover number $\mathcal{N}(K, \rho, \varepsilon)$ is the smallest size of an $\varepsilon-$cover N_ε of K with respect to ρ.*

Definition 9.8 (ε-separated set)**.** *Let ρ be a distance metric. A set S is ε-separated if $\forall x, y \in S$, $x \neq y$, $\rho(x, y) > \varepsilon$.*

Definition 9.9 (Packing Number)**.** *The packing number $\mathcal{P}(K, \rho, \varepsilon)$ is the size of the largest $\varepsilon-$separated subset of K with respect to ρ.*

Lemma 9.10.

$$\mathcal{P}(K, \rho, 2\varepsilon) \leq \mathcal{N}(K, \rho, \varepsilon) \leq \mathcal{P}(K, \rho, \varepsilon)$$

Proof lemma 9.10. We prove the right most inequality first. It suffices to prove that any max packing is a covering. Assume for contradiction that we have a max packing P for K and that some $x \in K$ is not covered. But then this is not a max packing since x is at least ε away from any point in P so it could have been added to P creating a larger packing set. This gives us a contradiction and thus every max packing is a cover. Therefore, $\mathcal{N}(K, \rho, \varepsilon) \leq \mathcal{P}(K, \rho, \varepsilon)$.
Now for the first inequality. Consider an $\varepsilon-$covering of K. Take two points x, y that are 2ε separated. They must be in 2 different covering balls. Therefore, there is at most one element of the 2ε packing set for each element in the ε covering set. Therefore $\mathcal{P}(K, \rho, 2\varepsilon) \leq \mathcal{N}(K, \rho, \varepsilon)$ □

Lecture 10: Upper bounds on the norms of Random Matrices

Instructor: Nikita Zhivotovskiy *Scriber: Kaihao Jing* *Proofreader: Xuelin Yang*

10.1 Preliminaries

Recall that we mentioned the concentration of random matrices, we need some preparation: For any set $K \subset \mathbb{R}^d$, recall a ϵ-net is a subset $\mathcal{N} \subset K$ such that $K \subset \cup_{x\in\mathcal{N}} B(x,\epsilon)$, where $B(x,r)$ is a Ball centered at x with radius r (with respect to some metric d), and the covering number $\mathcal{N}(K,d,\epsilon) = \min\{|\mathcal{N}| : \mathcal{N} \text{ is an } \epsilon\text{-net}\}$. A subset $\mathcal{P} \subset K$ is called $\epsilon - seperated$ if $d(x,y) > \epsilon$ for all distinct points $x, y \in \mathcal{P}$. The packing number $\mathcal{P}(K,d,\epsilon) = \max\{|\mathcal{P}| : \mathcal{P} \text{ is } \epsilon - seperated\}$. In the Lecture 9, we proved the equivalence of the covering and packing numbers:

Lemma 10.1. *For any set $K \subset \mathbb{R}^d$ and any $\epsilon > 0$, we have*

$$\mathcal{P}(K,d,2\epsilon) \le \mathcal{N}(K,d,\epsilon) \le \mathcal{P}(K,d,\epsilon).$$

For our purpose, we need an estimate of the covering number of the unit ball, which is stated as the following lemma:

Lemma 10.2. *Let B_2^d be the unit ball (with respect to $\|\cdot\|_2$ metric) in $\mathbb{R}^d$, then for any $\epsilon > 0$ we have*

$$\left(\frac{1}{\epsilon}\right)^d \le \mathcal{N}\left(B_2^d, \|\cdot\|_2, \epsilon\right) \le \left(1+\frac{2}{\epsilon}\right)^d.$$

Proof. **Lower bound:** for any ϵ-net $\mathcal{N}$ (WLOG, we assume $\mathcal{N}$ is a countable set),

$$\begin{aligned}\mathrm{Vol}\left(B_2^d\right) &\le \mathrm{Vol}\left(\cup_{x\in\mathcal{N}} B(x,\epsilon)\right) \le \sum_{x\in\mathcal{N}} \mathrm{Vol}\left(\epsilon B(x,1)\right)\\ &= |\mathcal{N}|\,\mathrm{Vol}\left(\epsilon B_2^d\right) = |\mathcal{N}|\,\epsilon^d \mathrm{Vol}\left(B_2^d\right),\end{aligned}$$

where $\mathrm{Vol}(\cdot)$ is the volume of sets in $\mathbb{R}^d$ and we use the fact that $\mathrm{Vol}\left(\epsilon B(x,1)\right) = \mathrm{Vol}\left(\epsilon B_2^d\right) = \epsilon^d \mathrm{Vol}\left(B_2^d\right)$. Thus, we have $|\mathcal{N}| \ge \left(\frac{1}{\epsilon}\right)^d$ for any $\epsilon-$net $\mathcal{N}$, which proves the lower bound.

Upper bound: Choose an $\epsilon - seperated$ set $\mathcal{P}$ such that $|\mathcal{P}| = \mathcal{P}\left(B_2^d, \|\cdot\|_2, \epsilon\right)$. Notice that for any distinct $x, y \in \mathcal{P}$, $B\left(x, \frac{\epsilon}{2}\right)$ and $B\left(y, \frac{\epsilon}{2}\right)$ are disjoint (since $d(x,y) > \epsilon$) and $B\left(x, \frac{\epsilon}{2}\right) \subset \left(1+\frac{\epsilon}{2}\right) B_2^d$, then

$$\mathcal{P}\left(B_2^d, \|\cdot\|_2, \epsilon\right) \mathrm{Vol}\left(\frac{\epsilon}{2} B_2^d\right) = \mathrm{Vol}\left(\cup_{x\in\mathcal{P}} B\left(x, \frac{\epsilon}{2}\right)\right) \le \mathrm{Vol}\left(\left(1+\frac{\epsilon}{2}\right) B_2^d\right),$$

then implies that $\mathcal{P}\left(B_2^d, \|\cdot\|_2, \epsilon\right) \le \left(1+\frac{2}{\epsilon}\right)^d$. Finally, by Lemma 10.1, $\mathcal{N}\left(B_2^d, \|\cdot\|_2, \epsilon\right) \le \left(1+\frac{2}{\epsilon}\right)^d$. □

Remark 10.3. *It's not hard to see that the upper bound for $\mathcal{N}\left(B_2^d, \|\cdot\|_2, \epsilon\right)$ in Lemma 10.2 is also an upper bound for $\mathcal{N}\left(S^{d-1}, \|\cdot\|_2, \epsilon\right)$, where S^{d-1} is the $d-1$ dimensional unit sphere.*

10.2 Upper bound for matrices with independent entries

In this section we prove a concentration inequality for the random matrices with independent entries.

Theorem 10.4. *Let $X = \left(X_{ij}\right)_{m\times n}$ be a $m \times n$ random matrix where the entries X_{ij} are independent random variables such that $\mathbb{E}X_{ij} = 0$ for any $1 \le i \le m$, $1 \le j \le n$ and $K = \max_{i,j} \left\|X_{ij}\right\|_{\psi_2} < \infty$. Recall the operator norm $\|X\| = \|X\|_{op} = \sup_{u\in S^{m-1},\, v\in S^{n-1}} u^\top Xv$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\|X\| \le CK\left(\sqrt{m} + \sqrt{n} + \sqrt{\log(1/\delta)}\right),$$

where $C > 0$ is a universal constant.

The following is a key proposition.

Proposition 10.5. *Given any matrix $A \in \mathbb{R}^{m\times n}$ and $\epsilon \in \left(0, \frac{1}{2}\right)$, let $\mathcal{M}$ be an ϵ-net for S^{m-1} and $\mathcal{N}$ be an ϵ-net for S^{n-1}, then we have*

$$\sup_{u\in\mathcal{M},\, v\in\mathcal{N}} u^\top Av \le \|A\|_{op} \le \frac{1}{1-2\epsilon} \sup_{u\in\mathcal{M},\, v\in\mathcal{N}} u^\top Av$$

Proof. The proof for lower bound is straightforward from the definition of the operator norm. For the upper bound, we choose $v_0 \in S^{n-1}$ such that $\|Av_0\|_2 = \|A\|$ (this is achievable since $\|A\| = \max_{v\in S^{n-1}} \|Av\|_2$). There exists $v_1 \in \mathcal{N}$ such that $\|v_0 - v_1\|_2 \le \epsilon$, then we have

$$\begin{aligned}\|A\| = \|Av_0\|_2 &= \|Av_0 - Av_1 + Av_1\|_2 \\ &\le \|Av_0 - Av_1\|_2 + \|Av_1\|_2 \\ &\le \epsilon\|A\| + \|Av_1\|_2,\end{aligned}$$

thus $(1-\epsilon)\|A\| \le \|Av_1\|_2 \le \sup_{v\in\mathcal{N}} \|Av\|_2$. For the same reason, for any $v \in S^{n-1}$

$$\|Av\|_2 = \sup_{u\in S^{m-1}} u^\top Av \le \frac{1}{1-\epsilon} \sup_{u\in\mathcal{M}} u^\top Av.$$

Finally,

$$\|A\| \le \frac{1}{(1-\epsilon)^2} \sup_{u\in\mathcal{M},\, v\in\mathcal{N}} u^\top Av \le \frac{1}{1-2\epsilon} \sup_{u\in\mathcal{M},\, v\in\mathcal{N}} u^\top Av,$$

where $\frac{1}{(1-\epsilon)^2} \le \frac{1}{1-2\epsilon}$ since $\epsilon \in \left(0, \frac{1}{2}\right)$. □

Proof of Theorem 10.4. Let $\mathcal{N}$ be a $\frac{1}{4}$-net for S^{n-1} and $\mathcal{M}$ be a $\frac{1}{4}$-net for S^{m-1} such that $|\mathcal{N}| \le 9^n$ and $|\mathcal{M}| \le 9^m$ (this is achievable due to Lemma 10.2 and Remark 10.3). By Proposition 10.5,

$$\|X\| \le 2 \sup_{u\in\mathcal{M},\, v\in\mathcal{N}} u^\top Av.$$

For any pair $(u, v) \in \mathcal{M} \times \mathcal{N}$, we have that

$$\begin{aligned}\left\|u^\top Xv\right\|_{\psi_2}^2 = \Big\|\sum_{i,j} u_i X_{ij} v_j\Big\|_{\psi_2}^2 &\le C_1 \sum_{i,j} \left\|u_i X_{ij} v_j\right\|_{\psi_2}^2 \\ &\le C_1 K^2 \sum_{i,j} u_i^2 v_j^2 = C_1 K^2 \left(\sum_i u_i^2\right)\left(\sum_j v_j^2\right)\end{aligned}$$

$$\leq C_1 K^2,$$

where we use C_i to represent universal constants and the last inequality is because $u \in S^{m-1}$, $v \in S^{n-1}$. Then the concentration inequality for sub-Gaussian random variables implies that for any $t > 0$

$$\Pr\left(u^\top X v \geq t\right) \leq \exp\left(-\frac{C_2 t^2}{K^2}\right),$$

then using the union bound gives us that

$$\Pr\left(\sup_{u \in \mathcal{M},\, v \in \mathcal{N}} u^\top X v \geq t\right) \leq 9^{m+n} \exp\left(-\frac{C_2 t^2}{K^2}\right)$$

For any δ, choose $t = C_3 K\left(\sqrt{m} + \sqrt{n} + \sqrt{\log(1/\delta)}\right)$, where C_3 is chosen such that $9^{m+n} \exp\left(-\frac{C_2 t^2}{K^2}\right) \leq \delta$ (this is possible because $\left(\sqrt{m} + \sqrt{n}\right)^2 \geq m + n$) we have with probability at least δ,

$$\|X\| \leq 2 \sup_{u \in \mathcal{M},\, v \in \mathcal{N}} u^\top X v \leq 2 C_3 K\left(\sqrt{m} + \sqrt{n} + \sqrt{\log(1/\delta)}\right),$$

which finishes the proof. □

Example: (Wigner matrix) Let $X = \left(X_{ij}\right)_{n \times n}$ be a $n \times n$ random matrix such that $\left\{X_{ij}\right\}_{i<j}$ are i.i.d. $\mathcal{N}(0,1)$, $\{X_{ii}\}_{1 \leq i \leq n}$ are i.i.d. $\mathcal{N}(0,2)$ (also independent from X_{ij}) and $X^\top = X$. Define $Y = \left(Y_{ij}\right)_{n \times n}$ such that

$$Y_{ij} = X_{ij} \qquad \forall i < j, \qquad Y_{ii} = \frac{X_{ii}}{2} \qquad \forall 1 \leq i \leq n, \qquad Y_{ij} = 0 \qquad \forall i > j,$$

then $X = Y + Y^\top$. By triangle inequality, $\|X\| \leq \|Y\| + \|Y^\top\| = 2\|Y\|$. Clearly $\max_{ij} \left\|Y_{ij}\right\|_{\psi_2} \leq C < \infty$ ($\|0\|_{\psi_2} = 0$), then apply Theorem 10.4 we have that for any $\delta \in (0,1)$, with probability at least $1 - \delta$

$$\|X\| \leq C\left(\sqrt{n} + \sqrt{\log(1/\delta)}\right).$$

10.3 Operator norm of sample covariance matrices

Now we consider sample covariance matrices: consider $X \in \mathbb{R}^d$, $\mathbb{E}X = 0$, and sub-Gaussian $\forall v \in S^{d-1}$, $\|\langle X, v\rangle\|_{\psi_2} \leq C\|\langle X, v\rangle\|_{L^2}$ with absolute constant C. We want to bound $\|\frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \Sigma\|_{op}$.
Define effective rank of Σ as $r(\Sigma) = \frac{tr(\Sigma)}{\|\Sigma\|_{op}}$ (the denominator is the largest eigenvalue).

Theorem 10.6. *For sub-Gaussian zero-mean independent sample $X_1, \ldots, X_n$,*

$$\|\frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \Sigma\|_{op} \leq C\|\Sigma\|_{op}\left(\sqrt{\frac{r(\Sigma)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

with probability at least $1 - \delta$ whenever $n \geq C_1(r(\Sigma) + \log(1/\delta))$.

Proof. Recall that from Lecture 7, we have the following lemma: Consider $f(X, \theta)$ for r.v. X, and parameter $\theta \in \Theta \in \mathbb{R}^d$. Choose the prior π on Θ. Simultaneously for all measure $\rho : KL(\rho||\pi) < \infty$. We know $\mathbb{E}_{\theta\sim\rho} f(X, \theta) \leq \mathbb{E}_{\theta\sim\rho} \log \mathbb{E}_X \exp(f(X, \theta)) + KL(\rho||\pi) + \log(1/\delta)$ with probability at least $1 - \delta$. We use this lemma to get the following corollary:

Corollary 10.7. *Assume $f(X,\theta) = \sum_{i=1}^n f(X_i,\theta)$ for random vector X with i.i.d. X_i. We plug it into the lemma and have*

$$\frac{1}{n}\sum_{i=1}^n f(X_i,\theta) \le E_{\theta\sim\rho}\log \mathbb{E}_X \exp(f(X,\theta)) + \frac{KL(\rho||\pi)+\log(1/\delta)}{n}.$$

Since we are interested in bounding

$$\|\frac{1}{n}\sum_{i=1}^n X_iX_i^\top - \Sigma\|_{op} = \sup_{u,v\in S^{d-1}}\left(\frac{1}{n}\sum_{i=1}^n \langle X_i,u\rangle\langle X_i,v\rangle - u^\top\Sigma v\right)$$

and we want to relate this to the Corollary:
We use two tricks. For the first trick, we consider $\pi(\theta)$ with $\pi \sim \mathcal{N}(0,\beta^{-1}I_d)$ where $\beta > 0$ is some parameter that we can tune. Let $\rho_v(\theta)$, $\rho > 0$ defines the density as folows:

$$\rho_v(\theta) = \frac{1}{p(2(\pi^{-1}\beta)^{d/2})}\exp\left(\frac{-\beta\|\theta-v\|^2}{2}\right)\mathbb{1}\{\|\Sigma^{1/2}(\theta-v)\|_2 \le r\}.$$

On RHS, p in the denominator is a normalization factor because we restrict to the ellipsoid. The indicator function also means putting it into the ellipsoid.
For the second trick, recall original $\theta \in \Theta \subseteq \mathbb{R}^d$. We let a new $\theta \in \Theta^{2d}$, $\theta = (\theta_1,\theta_2)$ with θ_1,θ_2 both d dimensional vectors. We have $\rho_{u,v} = \rho_u(\theta_1)\otimes\rho_v(\theta_2)$, $\pi'(\theta) = \pi(\theta_1)\otimes\pi(\theta_2)$, and pair $(\theta,\nu)\sim\rho_{u,v}$. Then,

$$\mathbb{E}_{(\theta,\nu)\sim\rho_{u,v}}\langle X,\theta\rangle\langle X,\nu\rangle = \langle X,\mathbb{E}_{\theta\sim\rho_{u,v}}\theta\rangle\langle X,\mathbb{E}_{\nu\sim\rho_{u,v}}v\rangle = \langle X,u\rangle\langle X,v\rangle.$$

Plug the following function into the corollary:

$$\sup_{u,v\in S^{d-1}}\mathbb{E}_{(\theta,\nu)\sim\rho_{u,v}}\lambda\left(\sum_{i=1}^n\langle\theta,X_i\rangle\langle X_i,\nu\rangle - \theta^\top\Sigma\nu\right) = \lambda\|\sum_{i=1}^n\left(X_iX_i^\top-\Sigma\right)\|_{op},$$

we have

$$\lambda\|\frac{1}{n}\sum_{i=1}^n\left(X_iX_i^\top-\Sigma\right)\|_{op} \le \sup_{u,v\in S^{d-1}}\Bigg(\mathbb{E}_{(\theta,\nu)\sim\rho_{u,v}}\log\mathbb{E}_X\exp\left(\lambda\left(\langle\theta,X\rangle\langle X,\nu\rangle-\theta^\top\Sigma\nu\right)\right) + \frac{KL(\rho_{u,v}||\pi')+\log(1/\delta)}{n}\Bigg),$$

Consider the first term on the RHS as if θ,ν are fixed:

$$\begin{aligned}
\mathbb{E}_{(\theta,\nu)\sim\rho_{u,v}}\log\mathbb{E}_X\exp\left(\lambda\left(\langle\theta,X\rangle\langle X,\nu\rangle-\theta^\top\Sigma\nu\right)\right) &= \|\langle\theta,X\rangle\langle X,\nu\rangle-\theta^\top\Sigma\nu\|_{\psi_1}\\
&\le C_2\|\langle\theta,X\rangle\langle X,\nu\rangle\|_{\psi_1}\\
&\le C_2\|\langle\theta,X\rangle\|_{\psi_2}\|\langle X,\nu\rangle\|_{\psi_2}\\
&\le C_3\sqrt{\theta^\top\Sigma\theta}\sqrt{\nu^\top\Sigma\nu}\\
&\le C_3(\theta^\top\Sigma\theta+\nu^\top\Sigma\nu)\\
&\le C_4\|\Sigma\|_{op}.
\end{aligned}$$

Note that in above $\theta^\top\Sigma\theta \le 2((\theta-u)^\top\Sigma(\theta-u)+(u^\top\Sigma u)) \le 2(r^2+\|\Sigma\|_{op})$ (also applies to $\nu^\top\Sigma\nu$). The last line comes from picking $r^2 = 2\|\Sigma\|_{op}$, which will also be used to bound the normalization constant p. Now what is left is to bound the second term on the RHS (i.e. $\frac{KL(\rho_{u,v}||\pi')}{n}$). □

Lecture 11: Matrix Bernstein & Gaussian Comparator Inequalities

Instructor: Nikita Zhivotovskiy *Scriber: Jessica Dai* *Proofreader: Kota Okuda*

11.1 Proof of sample covariance bound, continued.

We begin by finishing the proof of the result discussed last lecture, the bound on sample covariance for random vectors. Recall the statement below:

Theorem 11.1. *Let $X_1, \dots, X_n$ be independent random vectors in $\mathbb{R}^d$ with $\mathbb{E}[X_i] = 0$, true covariance Σ, and subgaussian, i.e. for all $v \in S^{d-1}$, $\|\langle X, v\rangle\|_{\psi_2} \le C\|\langle X, v\rangle\|_2$. Then, with probability $1-\delta$,*

$$\|\frac{1}{n}\sum_{i\in[n]} X_i X_i^\top - \Sigma\|_{\text{op}} \le C\|\Sigma\|_{\text{op}}\left(\sqrt{\frac{r(\Sigma)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where $r(\Sigma) = \frac{\text{Tr}(\Sigma)}{\|\Sigma\|_{\text{op}}}$ is the effective rank of Σ and C is some constant, as long as $n \ge C'(r(\Sigma) + \log(1/\delta))$ for some constant C'.

By the end of Lecture 10, we had shown

$$\lambda\|\frac{1}{n}\sum_{i\in[n]} X_i X_i^\top - \Sigma\|_{\text{op}} \le \sup_{u,v\in S^{d-1}}\Bigg[\mathbb{E}_{\theta\sim\rho_{u,v}}\Big[\log\Big(\mathbb{E}_X\big[\exp\big(\lambda\underbrace{(\langle\theta, X\rangle\langle X, v\rangle - \theta^\top\Sigma v)}_{(A)}\big)\big]\Big)\Big] + \underbrace{\frac{KL(\rho_{u,v}\|\pi\otimes\pi) + \log(1/\delta)}{n}}_{(B)}\Bigg],$$

where $\theta \in \mathbb{R}^d$ and $\rho_u(\theta)$ has density (with parameters p, β and r[1]) given by

$$\frac{1}{p(2\pi\beta^{-1})^{d/2}}\exp\left(\frac{-\beta\|\theta - v\|^2}{2}\right)\cdot 1[\|\Sigma^{1/2}(\theta - v)\|_2 \le r],$$

allowing us to define $\rho_{u,v}(\theta_1, \theta_2) = \rho_u(\theta_1)\otimes\rho_v(\theta_2)$, and $\pi(\theta) \sim \mathcal{N}(0, \beta^{-1}I_d)$ for $\beta > 0$.
We had begun analyzing (A) by first looking at its $\psi - 1$ norm, showing that $\|(A)\|_{\psi_1} \le C_1(\theta^\top\Sigma\theta + v^\top\Sigma v)$.

Analyzing the first term, continued. In Lecture 11, we continue with this term:

$$\begin{aligned}\|(A)\|_{\psi_1} &\le C_1(\theta^\top\Sigma\theta + v^\top\Sigma v)\\ &\le C_2(\|\Sigma\| + r^2)\\ &\le C_3\|\Sigma\|,\end{aligned}$$

where in the second transition we note that

$$\theta^\top\Sigma\theta + v^\top\Sigma v \le 2(\theta - u)^\top\Sigma(\theta - u) + 2u^\top\Sigma u \le 2(r^2 + \|\Sigma\|)$$

[1]Note this r is not the same as the $r(\Sigma)$ in the theorem statement.

and in the third transition we choose $r^2 = 2\|\Sigma\|$.
Now, consider (A) as a random variable. By the above, we know that $\|(A)\|_{\psi_2} \le C_3\|\Sigma\|$. Then, as long as $\lambda \le \frac{1}{C_3\|\Sigma\|}$, we have by subgaussianity[2] (Def. 1 from Prop. 2, Lec. 2) that

$$\sup_{u,v} \mathbb{E}_X[\exp(\lambda(A))] \le \log \exp\left(\lambda^2 C_4 \|\Sigma\|^2\right) = \lambda^2 C_4 \|\Sigma\|^2.$$

Analyzing the KL term. We now move to analyzing $n(B) := KL(\rho_{u,v}\|\pi\otimes\pi) + \log(1/\delta)$. We will proceed in three steps.
Step 1: Computations. We can explicitly compute the KL divergence between ρ_u and π as

$$\begin{aligned} KL(\rho_u\|\pi) &= \mathbb{E}_{\theta\sim\rho_u}\left[\log\frac{\rho_u(\theta)}{\pi(\theta)}\right] \\ &= \mathbb{E}_{\theta\sim\rho_u}\left[\log\left(\frac{1}{p}\exp\left(\frac{-\|\theta-u\|^2+\|\theta\|^2}{2\beta^{-1}}\right)\right)\right] \\ &= \log(1/p) + \mathbb{E}_{\theta\sim\rho_u}\left[\frac{-\|\theta\|^2-\|u\|^2+2\langle\theta,u\rangle+\|\theta\|^2}{2\beta^{-1}}\right] \\ &= \log(1/p) + \beta/2, \end{aligned}$$

where the final transition follows by noting that $\mathbb{E}_{\theta\sim\rho_u}[\theta] = u$ by symmetry and $\|u\|^2 = 1$.
Step 2: Converting to product measures. Using a property of KL for product measures, we have that

$$KL(\rho_{u,v}\|\pi\otimes\pi) = KL(\rho_u\|\pi) + KL(\rho_v\|\pi) = 2\log(1/p) + \beta.$$

Step 3: Dealing with the parameters p and β. Recall that p is a parameter to the density of ρ that can be interpreted as a normalization constant for a random variable $Z \sim \mathcal{N}(0, \beta^{-1}I_d)$. Then, we have that $p = \Pr\left[\|\Sigma^{1/2}Z\|_2 \le r\right]$; we can find a lower bound for p by upper bounding as follows:

$$\Pr\left[\|\Sigma^{1/2}Z\|_2 > r\right] \le \frac{\mathbb{E}\left[\|\Sigma^{1/2}Z\|_2^2\right]}{r^2} = \frac{\mathrm{Tr}(\Sigma)\beta^{-1}I_d}{r^2} = \frac{\mathrm{Tr}(\Sigma)}{\beta 2\|\Sigma\|} = \frac{1}{2},$$

where the final transition follows by choosing $\beta = r(\Sigma)$. Hence we have $p \ge 1/2$. Plugging this (and our choice of β) into the result from Step 2, we have

$$KL(\rho_{u,v}\|\pi\otimes\pi) = 2\log(2) + r(\Sigma) \le Cr(\Sigma).$$

Completing the proof. Optimizing over λ, we will find that

$$\lambda_{\mathrm{opt}} \approx \frac{1}{\|\Sigma\|}\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{n}};$$

plugging this in gives us the desired result. Note that this requires $n \ge C(r(\Sigma)+\log(1/\delta))$ for some constant C.

Remark 11.2. *None of these constants are any larger than approx. 20.*

[2]Note that the randomness in the below expectation is due to (A), but the only randomness in (A) is due to X because we are working inside an outer expectation over θ, which means we can take it to be fixed.

11.2 Matrix Bernstein Inequality

Fun fact: this result is only around 13-14 years old. The statement is as follows.

Theorem 11.3. *Let $X_1 \dots X_n$ be independent, zero-mean, symmetric $d \times d$ matrix such that $\|X_i\|_{\text{op}} \le K$ for all i. Then, for all $t \ge 0$, we have*

$$\Pr\left[\|\sum_{i\in[n]} X_i\|_{\text{op}} \ge t\right] \le 2d \exp\left(\frac{-t^2/2}{\sum_{i\in[n]} \mathbb{E}\|X_i^2\|_{\text{op}} + \frac{Kt}{3}}\right).$$

Rearranging, we get that with probability $1-\delta$,

$$\|\sum_{i\in[n]} X_i\|_{\text{op}} \le \sqrt{2\|\sum_{i\in[n]} \mathbb{E}\|X_i^2\|_{\text{op}} \log(2d/\delta)} + \frac{2}{3}K\log(2d/\delta).$$

When comparing to the scalar Bernstein inequality, we see that we use the operator norm instead of the exact variance; we have boundedness replaced by K; and pay an additional d term.

11.2.1 Useful facts for proof of Theorem 11.3.

We will use the following facts in the proof of Theorem 11.3.

Proposition 11.4. *Consider X symmetric and a function $f : \mathbb{R} \to \mathbb{R}$. SVD on X gives us $X = \sum_{j\in[d]} \lambda_j u_j u_j^\top$, where λ_j are the (ordered) eigenvalues of X and u_j are the corresponding eigenvectors. Define*

$$f(X) = \sum_{j\in[d]} f(\lambda_j) u_j u_j^\top.$$

Then, we have the following facts:

(a) If $f(x) \le g(x), \forall |x| \le K$, then $f(X) \succeq g(X)$ if $\|X\|_{\text{op}} \le K$, i.e. $f(X) - g(X)$ is PSD.

(b) If $0 \succeq X \succeq Y$, then $\log(X) \succeq \log(Y)$.

(c) If $X \succeq Y$, then $\text{Tr}\big(\exp(X)\big) \le Tr(\exp(Y))$.

For fact (b), note that log is monotonic in $d > 1$, but not all functions that are monotonic in one dimension preserve monotonicity in the matrix sense—for example, monotonicity is violated even for $d = 2$ for functions like $\exp(x)$ or x^2. Fact (c), on the other hand, is true for any function that is monotone in one dimension.
The fourth fact is known as Lieb's Inequality and is nontrivial to prove for $d > 1$. We will be using a corollary of Prop. 11.5, which applies the result to random matrices.

Proposition 11.5 (Lieb's Inequality.)**.** *For symmetric $H \in \mathbb{R}^{d\times d}$, the function $\psi(A) = \text{Tr}\big(\exp\big(H + \log(A)\big)\big)$ is concave for PSD A. That is, $\forall \alpha \in (0,1)$ and PSD A, B,*

$$\psi(\alpha A + (1-\alpha)B) \ge \alpha\psi(A) + (1-\alpha)\psi(B).$$

Corollary 11.6. *Let Z be a random matrix and $A = \exp(Z)$. Then, by concavity (via Lieb's) and Jensen, we have*

$$\mathbb{E}\left[\text{Tr}\big(\exp(H+Z)\big)\right] \le \text{Tr}\big(\exp\big(H + \log \mathbb{E}\left[\exp(Z)\right]\big)\big).$$

11.2.2 Proof of Theorem 11.3.

We are now ready to prove Theorem 11.3.
First, define $S := \sum_{i\in[n]} X_i$. Then, $\|S\|_{\text{op}} = \max(\lambda_{\max}(S), \lambda_{\max}(-S))$, where the second term is to handle possible negative eigenvalues. For ease of exposition, we assume all eigenvalues are nonnegative for now. Then, we have[3]:

$$\begin{aligned}
\Pr[\lambda \cdot \lambda_{\max}(S) \geq \lambda t] &\leq \frac{\mathbb{E}\left[\exp(\lambda \cdot \lambda_{\max}(S))\right]}{\exp(\lambda t)} && \text{(standard Chernoff in 1 dimension)} \\
&= \frac{\mathbb{E}\left[\lambda_{\max} \cdot \exp(\lambda S)\right]}{\exp(\lambda t)} && \text{(property of exp applied to matrices)} \\
&\leq \frac{\mathbb{E}\left[\text{Tr}\big(\exp(\lambda S)\big)\right]}{\exp(\lambda t)} && \text{(all eigenvalues are non-negative)} \\
&= \frac{\mathbb{E}\left[\text{Tr}\Big(\exp\Big(\lambda \sum_{i=1}^{n-1} X_i + \lambda X_n\Big)\Big)\right]}{\exp(\lambda t)} \\
&= \frac{\mathbb{E}_{i\in[n-1]}[\text{Tr}\Big(\exp\Big(\lambda \sum_{i=1}^{n-1} X_i\Big) \cdot \mathbb{E}_{i=n}[\exp(\lambda X_n) \mid X_{1\ldots n-1}\Big)]]}{\exp(\lambda t)}, && (*)
\end{aligned}$$

where the final transition follows by noting that all X_i are independent so we can condition on $X_{1\ldots n-1}$ to isolate the randomness in X_n. Then, with $H = \lambda \sum_{i\in[n-1]} X_i$, we can apply Cor. 11.6 to get

$$(*) \leq \frac{\mathbb{E}_{i\in[n-1]}[\text{Tr}\Big(\exp\Big(\lambda \sum_{i=1}^{n-1} X_i + \log \mathbb{E}_{i=n}[\exp(\lambda X_n) \mid X_{1\ldots n-1}]\Big)\Big)]}{\exp(\lambda t)}.$$

Applying lines from the proof of 1-d Bernstein and property *(a)* of Prop. 11.4, we have that for a single X_i,

$$\mathbb{E}\left[\exp(\lambda X_i)\right] \preceq \exp\Big(g(\lambda)\mathbb{E}\left[X_i^2\right]\Big),$$

where $g(\lambda) = \frac{\lambda^2/2}{1-\lambda K/3}$ and $|\lambda| \leq 3/K$.
Applying the conditioning trick and Lieb's repeatedly for $i \in [n-1]$, we have

$$\begin{aligned}
(*) &\leq \frac{\text{Tr}\big(\exp\big(g(\lambda)\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right]\big)\big)}{\exp(\lambda t)} \\
&\leq \frac{d\lambda_{\max}(\exp\big(g(\lambda)\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right]\big))}{\exp(\lambda t)} \\
&= \frac{d\exp\big(g(\lambda)\|\sum_{i=1}^{n} X_i^2\|_{\text{op}}\big)}{\exp(\lambda t)}.
\end{aligned}$$

We can optimize over λ; repeat these steps for $\lambda_{\max}(-S)$; and apply the union bound to complete the proof.

11.2.3 Extensions of Matrix Bernstein Inequality.

We briefly consider two extensions/applications of Theorem 11.3.

[3]Note the distinction between λ the Chernoff parameter and $\lambda_{\max}$ the max eigenvalue of S.

Controlling $\mathbb{E}\left[\|S\|_{\text{op}}\right]$. Chernoff's method would give us a bound of the form $\mathbb{E}\left[\|S\|_{\text{op}}\right] \le \frac{1}{\lambda}\log \mathbb{E}\left[\exp\left(\lambda\|S\|_{\text{op}}\right)\right]$. In the setup of Theorem 11.3, we instead have

$$\mathbb{E}\left[\|S\|_{\text{op}}\right] \le \sqrt{2\log(2d)\|\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]\|_{\text{op}}} + \frac{2}{3}K\log(2d).$$

General rectangular matrices. Consider $A \in \mathbb{R}^{d_1\times d_2}$. Then we can construct the block matrix

$$\tilde{A} = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix},$$

and can proceed with analyzing $\tilde{A}$, noting that $\lambda_{\max}(\tilde{A}) = \|A\|_{\text{op}}$.

11.3 Gaussian Comparator Inequalities

We finish with a preview of Gaussian processes.

Definition 11.7 (Gaussian process.)**.** *Consider the random process $(X_t)_{t\in\mathcal{T}}$. A Gaussian process is one where, for all finite $\mathcal{T}_0 \subseteq \mathcal{T}$, the random vector $(X_t)_{t\in\mathcal{T}_0}$ is (multivariate) Gaussian.*

Generally, we care about the behavior of $\sup_{t\in\mathcal{T}} X_t$, or $\mathbb{E}\left[\sup_{t\in\mathcal{T}} X_t\right]$. We now give a statement (to be proven in future lectures) of the Slepian Lemma.

Theorem 11.8. *Assume X_t and Y_t are zero-mean Gaussian processes such that $\mathbb{E}\left[X_t^2\right] = \mathbb{E}\left[Y_t^2\right]$ and $\forall s,t \in \mathcal{T}$, $\mathbb{E}\left[(X_t - X_s)^2\right] \le \mathbb{E}\left[(Y_t - Y_s)^2\right]$. Then, for all $t \in \mathcal{T}$:*

1. $\Pr\left[\sup_{t\in\mathcal{T}} X_t \ge \tau\right] \le \Pr\left[\sup_{t\in\mathcal{T}} Y_t \ge \tau\right]$ *for all τ, and*
2. $\mathbb{E}\left[\sup_{t\in\mathcal{T}} X_t\right] \le \mathbb{E}\left[\sup_{t\in\mathcal{T}} Y_t\right]$.

Lecture 12: Gaussian processes

Instructor: Nikita Zhivotovskiy *Scriber: Anthony Ozerov* *Proofreader: Zach Rewolinski*

12.1 Slepian's inequality

Theorem 12.1 (Slepian's inequality)**.** *Suppose $(X_t)_{t\in T}$, $(Y_t)_{t\in T}$ are zero-mean Gaussian processes such that, $\forall t, s \in T$, we have*

$$\mathbb{E}\left[X_t^2\right] = \mathbb{E}\left[Y_t^2\right] \qquad \text{and} \qquad \mathbb{E}\left[(X_t - X_s)^2\right] \le \mathbb{E}\left[(Y_t - Y_s)^2\right].$$

Then $\forall \tau \in \mathbb{R}$, we have

1. $\Pr\left(\sup_{t\in T} X_t \ge \tau\right) \le \Pr\left(\sup_{t\in T} Y_t \ge \tau\right)$
2. $\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} Y_t\right]$

Remark 12.2. *When we say $\mathbb{E}\left[\sup_{t\in T} X_t\right]$, there are some concerns with measurability depending on what exactly we mean. To avoid these, we use Talagrand's convention, which states that*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] := \sup_{\substack{T_0 \subseteq T \\ T_0 \text{ finite}}} \mathbb{E}\left[\sup_{t\in T_0} X_t\right].$$

To prove this Theorem 12.1, we will first need to establish several lemmas.

Lemma 12.3 (Stein's lemma)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. If $X \sim N(0,1)$, then we have*

$$\mathbb{E}\left[f'(X)\right] = \mathbb{E}\left[Xf(X)\right].$$

Proof of Lemma 12.3. Assume for simplicity that f has bounded support. Define

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

This is the probability density function of X, because we stated $X \sim N(0,1)$. Let's find the expectation of $f'(X)$:

$$\mathbb{E}\left[f'(X)\right] = \int_{\mathbb{R}} f'(x)g(x)dx = \left[f(x)g(x)\right]_{-\infty}^{\infty} - \int_{\mathbb{R}} f(x)g'(x)\,dx.$$

In the second equality we have simply used integration by parts. Because f has bounded support, and $g(x)$ approaches zero as $x \to \infty$ or $x \to -\infty$, the first term is zero and we get

$$\mathbb{E}\left[f'(X)\right] = \int_{\mathbb{R}} -f(x)g'(x)\,dx.$$

Now we can notice that $g'(x) = -xg(x)$, and rewrite the RHS as

$$\mathbb{E}\left[f'(X)\right] = \int_{\mathbb{R}} xf(x)g(x)\,dx = \mathbb{E}\left[Xf(X)\right].$$

□

As Slepian's inequality deals with Gaussian processes (such that any finite collection is a multivariate Gaussian), we would like to have a version of Stein's lemma which applies to multivariate Gaussians.

Lemma 12.4 (Multivariate Stein's lemma)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. If $X \sim N_n(0, \Sigma)$, then we have*

$$\mathbb{E}[Xf(X)] = \Sigma \mathbb{E}[\nabla f(X)] = \left(\sum_{j=1}^{n} \Sigma_{i,j} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(X) \right] \right)_{i=1}^{n}.$$

Note that the result is an n-vector. The proof is essentially the same as in the univariate case. Now that we have the multivariate Stein's lemma, we can get the Gaussian interpolation lemma.

Lemma 12.5 (Gaussian interpolation)**.** *Suppose $X = (X_1, \dots, X_n) \sim N(0, \Sigma^X)$ and $Y = (Y_1, \dots Y_n) \sim N(0, \Sigma^Y)$ are two independent Gaussian random vectors. Define*

$$Z(u) = \sqrt{u}X + \sqrt{1-u}Y, \qquad u \in [0, 1]. \tag{33}$$

Then if $f : \mathbb{R}^n \to \mathbb{R}$ is a twice-differentiable function (with nice properties so that we can swap integrals and derivatives, and hence expectation and derivatives), we get

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

Proof of Lemma 12.5. Under the assumptions of the Lemma, we get

$$\begin{aligned}\frac{d}{du}\mathbb{E}[f(Z(u))] &= \mathbb{E}\left[\frac{d}{du}f(Z(u))\right] = \mathbb{E}\left[\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(Z(u)) \cdot \frac{dZ_i(u)}{du}\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(u)) \cdot \frac{dZ_i(u)}{du}\right].\end{aligned}$$

Now note that, by the definition of Z in Equation 33, we have

$$\frac{dZ_i(u)}{du} = \frac{1}{2}\left(\frac{1}{\sqrt{u}}X_i - \frac{1}{\sqrt{1-u}}Y_i\right).$$

Thus we get

$$\frac{d}{du}\mathbb{E}[f(Z(u))] = \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(u)) \cdot \left(\frac{1}{\sqrt{u}}X_i - \frac{1}{\sqrt{1-u}}Y_i\right)\right].$$

Let's work on the first term. Let

$$h_i(x) = \frac{\partial f}{\partial x_i}(Z(u)), \tag{34}$$

where we think of Y (which is an additive term in Z) as fixed. We get that

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(u)) \cdot \frac{1}{\sqrt{u}}X_i\right] &= \frac{1}{2\sqrt{u}}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}[h_i(X)X_i|Y]] \\ &= \frac{1}{2\sqrt{u}}\sum_{i=1}^{n}(\mathbb{E}[\mathbb{E}[h_i(X)X|Y]])_i\end{aligned}$$

$$= \frac{1}{2\sqrt{u}} \sum_{i=1}^{n} \mathbb{E}\left[\sum_{j=1}^{n} \Sigma^X_{i,j} \mathbb{E}\left[\frac{\partial h_i}{\partial x_j}(X)|Y\right]\right]$$
$$= \frac{1}{2\sqrt{u}} \mathbb{E}\left[\sum_{i=1,j=1}^{n} \Sigma^X_{i,j} \frac{\partial h_i}{\partial x_j}(X)\right]$$
$$= \frac{1}{2} \mathbb{E}\left[\sum_{i=1,j=1}^{n} \Sigma^X_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j}(Z(u))\right].$$

The third equality is an application of Lemma 12.4. The final equality is by substituting $h_i(x)$ from Equation 34. Repeating the same computation on the other term, we get

$$\frac{d}{du}\mathbb{E}[f(Z(u)] = \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial f(Z(u))}{\partial x_i} \cdot \left(\frac{1}{\sqrt{u}} X_i - \frac{1}{\sqrt{1-u}} Y_i\right)\right] = \frac{1}{2}\sum_{i,j} (\Sigma^X_{i,j} - \Sigma^Y_{i,j}) \frac{\partial^2 f(Z(u))}{\partial x_i \partial x_j}, \tag{35}$$

which completes the proof. □

Corollary 12.6. *Under the assumptions of Lemma 12.5, we have that if additionally, for all $i \neq j$:*

1. $\Sigma^X_{i,j} \geq \Sigma^Y_{i,j}$
2. $\Sigma^X_{i,i} = \Sigma^Y_{i,i}$
3. $\dfrac{\partial^2 f}{\partial x_i \partial x_j} \geq 0$

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

Proof of Corollary 12.6. We can see this by noting that $\mathbb{E}[f(X)] = \mathbb{E}[f(Z(1))]$ and $\mathbb{E}[f(Y)] = \mathbb{E}[f(Z(0))]$, and Equation 35 with the assumed conditions tells us that the derivative of $\mathbb{E}[f(Z(u))]$ with respect to u is nonnegative for $u \in [0,1]$. □

We are now finally equipped to prove Slepian's lemma.

Proof of Result 1 of Theorem 12.1. Suppose $(X_t)_{t\in T}$, $(Y_t)_{t\in T}$ are zero-mean Gaussian processes such that, $\forall t, s \in T$, we have

$$\mathbb{E}\left[X_t^2\right] = \mathbb{E}\left[Y_t^2\right] \qquad \text{and} \qquad \mathbb{E}\left[(X_t - X_s)^2\right] \leq \mathbb{E}\left[(Y_t - Y_s)^2\right].$$

By Talagrand's convention (Remark 12.2), we can focus on a finite T, $|T| = n$, and compare the Gaussian random vectors $X = (X_1, \ldots, X_n) \sim N(0, \Sigma^X)$ and $Y = (Y_1, \ldots, Y_n) \sim N(0, \Sigma^Y)$. Now instead of thinking of a supremum over $t \in T$, we can think of a maximum over $i \in [n]$.
By the assumption that $\mathbb{E}\left[X_t^2\right] = \mathbb{E}\left[Y_t^2\right]$ and that the vectors are zero-mean, we have that $\forall i$, $\Sigma^X_{i,i} = \Sigma^Y_{i,i}$.
By the assumption that $\forall t, s\ \mathbb{E}\left[(X_t - X_s)^2\right] \leq \mathbb{E}\left[(Y_t - Y_s)^2\right]$ and that the vectors are zero-mean, we have that $\forall i, j$, $\Sigma^X_{i,j} \geq \Sigma^Y_{i,j}$ (lower squared difference but same variances means higher covariances).
Assume WLOG that $X \perp\!\!\!\perp Y$. We can make this simplifying assumption because, if X and Y are not independent, we can replace Y_t with its uncorrelated copy; all the tail bounds given in Slepian's inequality will be the same.

Let $g_\tau : \mathbb{R} \to [0,1]$ be a twice-differentiable approximation of $\mathbf{1}[x < \tau]$. This approximation can be made arbitrarily good. We then have

$$\mathbf{1}[\max_i(x_i) < \tau] \approx (g_\tau(x_1) \cdot g_\tau(x_2) \cdot \ldots \cdot g_\tau(x_n)) = f_\tau(\mathbf{x})$$

Note that f_τ is also twice-differentiable. We would like it to satisfy Condition 3 of Corollary 12.6:

$$\frac{\partial^2 f_\tau}{\partial x_i \partial x_j} = \begin{cases} g_\tau(x_1) \cdot \ldots \cdot g'_\tau(x_i) \cdot \ldots \cdot g'_\tau(x_j) \cdot \ldots \cdot g_\tau(x_n) & i \neq j \\ g_\tau(x_1) \cdot \ldots \cdot g''_\tau(x_i) \cdot \ldots \cdot g_\tau(x_n) & i = j \end{cases}$$

In the first case, the derivative is always nonpositive, hence the product of the two first derivatives will be nonnegative. The second case seems like it could be negative, but it doesn't matter to us as Condition 3 only regards $i \neq j$. Thus f_τ satisfies Condition 3 of Corollary 12.6. So this setup with X, Y, and f_τ satisfies Corollary 12.6, from which we can conclude that $\mathbb{E}[f_\tau(X)] \geq \mathbb{E}[f_\tau(Y)]$ and therefore

$$\Pr\left(\max_i(X_i) < \tau\right) = \mathbb{E}\left[\mathbf{1}[\max_i(X_i) < \tau]\right] \approx \mathbb{E}[f_\tau(X)] \geq \mathbb{E}[f_\tau(Y)] \approx \mathbb{E}\left[\mathbf{1}[\max_i(Y_i) < \tau]\right].$$

It then follows that

$$\Pr\left(\max_i(X_i) \geq \tau\right) \leq \Pr\left(\max_i(Y_i) \geq \tau\right),$$

so we have shown result 1 of Theorem 12.1.

□

We have neglected result 2 of Theorem 12.1. For this we will need a different theorem.

Theorem 12.7 (Sudakov-Fernique)**.** *If $(X_t)_{t\in T}$, $(Y_t)_{t\in T}$ are zero-mean Gaussian processes such that $\forall s, t \in T$ we have*

$$\mathbb{E}\left[(X_t - X_s)^2\right] \leq \mathbb{E}\left[(Y_t - Y_s)^2\right],$$

then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

Proof idea for Sudakov-Fernique. We can do the same trick applying Remark 12.2 to deal only with finite Gaussian random vectors. As $\lambda \to \infty$, we can approximate

$$\max_{i\in[n]}(x_i) \approx \frac{1}{\lambda} \log\left(\sum_{i=1}^n \exp(\lambda x_i)\right) = f_\lambda(x).$$

Now we can apply Theorem 12.5 to this f_λ as we did for f_τ. f_λ is twice-differentiable, defining $Z(u) = \sqrt{u}X + \sqrt{1-u}Y$ as before, we get

$$\frac{d}{du}\mathbb{E}[f_\lambda(Z(u))] = \frac{1}{2}\sum_{i,j}(\Sigma^X_{i,j} - \Sigma^Y_{i,j})\frac{\partial^2 f_\lambda}{\partial x_i \partial x_j}(Z(u)) \leq 0,$$

which implies that $\mathbb{E}[f_\lambda(Z(0))] \geq \mathbb{E}[f_\lambda(Z(1))]$ and thus

$$\mathbb{E}\left[\max_{i\in[n]}(X_i)\right] \approx \mathbb{E}[f_\lambda(X)] = \mathbb{E}[f_\lambda(Z(1))] \leq \mathbb{E}[f_\lambda(Z(0))] = \mathbb{E}[f_\lambda(Y)] \approx \mathbb{E}\left[\max_{i\in[n]}(Y_i)\right].$$

□

Proof of Result 2 of Theorem 12.1. Sudakov-Fernique directly proves the second result in Slepian's Theorem. □

12.2 Applications

Let $X \in \mathbb{R}^{m \times n}$ be a random matrix whose elements X_{ij} are iid $N(0,1)$. We can show that

$$\mathbb{E}\left[||X||_{\text{op}}\right] \leq \sqrt{m} + \sqrt{n}.$$

Proof. Recall that, by the definition of $||\cdot||_{\text{op}}$

$$||X||_{\text{op}} = \sup_{\substack{u \in S^{m-1} \\ v \in S^{n-1}}} u^\top X v,$$

where $u^\top X v$ is a Gaussian process indexed by $t = (u, v) \in T = S^{m-1} \times S^{n-1}$. Let's try to satisfy the condition of Theorem 12.7. Let u, w be in S^{m-1}, and let v, z be in S^{n-1}. We have

$$\begin{aligned}
\mathbb{E}\left[(u^\top X v - w^\top X z)^2\right] &= \mathbb{E}\left[\left(\sum_{i,j} u_i X_{ij} v_j - \sum_{i,j} w_i X_{ij} z_j\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j} u_i X_{ij} v_j\right)^2\right] - 2\mathbb{E}\left[\left(\sum_{i,j} u_i X_{ij} v_j\right)\left(\sum_{i,j} w_i X_{ij} z_j\right)\right] + \mathbb{E}\left[\left(\sum_{i,j} w_i X_{ij} z_j\right)^2\right].
\end{aligned}$$

At this point, any cross terms containing $X_{ij} X_{kl}$ where $(i,j) \neq (k,l)$ will disappear, as $\mathbb{E}\left[X_{ij} X_{kl}\right]$ will be zero (recall that $X_{ij} \sim N(0,1)$). Recall also that $\mathbb{E}\left[X_{ij}\right] = 1$. Hence we get

$$\begin{aligned}
\mathbb{E}\left[(u^\top X v - w^\top X z)^2\right] &= \sum_{i,j} u_i^2 v_j^2 \mathbb{E}\left[X_{ij}^2\right] - 2\sum_{i,j} u_i^2 w_i^2 v_j^2 z_j^2 \mathbb{E}\left[X_{ij}^2\right] + \sum_{i,j} w_i^2 z_j^2 \mathbb{E}\left[X_{ij}^2\right] \\
&= \sum_{i,j} u_i^2 v_j^2 - 2\sum_{i,j} u_i^2 w_i^2 v_j^2 z_j^2 + \sum_{i,j} w_i^2 z_j^2 \\
&= \sum_{i,j} (u_i v_j - w_i z_j)^2 \\
&= ||u v^\top - w z^\top||_{\text{F}}^2 \\
&\leq ||u - w||_2^2 + ||v - z||_2^2.
\end{aligned}$$

Note that $u v^\top - w z^\top$ is a matrix with the i,jth entry being $u_i v_j - w_i z_j$. Its squared Frobenius norm is exactly the sum of the squared elements. The final inequality is nontrivial.
Now we can consider the process $Y_{u,v} = \langle u, Z_1 \rangle + \langle v, Z_2 \rangle$, where $Z_1 \sim N(0, I_m)$ and $Z_2 \sim N(0, I_n)$. We can compute that

$$\mathbb{E}\left[(Y_{u,v} - Y_{w,z})^2\right] = ||u - w||_2^2 + ||v - z||_2^2.$$

This process $Y_{u,v}$ will act as the second, simpler, dominating process in the Sudakov-Fernique inequality, as the condition is satisfied

$$\mathbb{E}\left[(u^\top X v - w^\top X z)^2\right] \leq ||u - w||_2^2 + ||v - z||_2^2 = E[(Y_{u,v} - Y_{w,z})^2].$$

Hence by the Sudakov-Fernique inequality, we have

$$\begin{aligned}
\mathbb{E}\left[||X||_{\text{op}}\right] = \mathbb{E}\left[\sup_{\substack{u \in S^{m-1} \\ v \in S^{n-1}}} u^\top X v\right] \leq \mathbb{E}\left[\sup_{\substack{u \in S^{m-1} \\ v \in S^{n-1}}} Y_{u,v}\right] = \mathbb{E}\left[\sup_{\substack{u \in S^{m-1} \\ v \in S^{n-1}}} \langle u, Z_1 \rangle + \langle v, Z_2 \rangle\right] = \mathbb{E}\left[||Z_1||_2 + ||Z_2||_2\right] \\
\leq (\mathbb{E}\left[||Z_1||_2^2\right])^{1/2} + (\mathbb{E}\left[||Z_2||_2^2\right])^{1/2} = \sqrt{n} + \sqrt{m}.
\end{aligned}$$

The second inequality is by Jensen's inequality. □

As a corollary, if X is a square matrix, $X \in \mathbb{R}^{n\times n}$, then $\mathbb{E}\left[||X||_{\text{op}}\right] \leq 2\sqrt{n}$. But what about a high-probability bound on $||X||_{\text{op}}$? Let's think of X as a vector in $\mathbb{R}^{n^2}$, and $||X||_{\text{op}}$ as a function from $\mathbb{R}^{n^2}$ to $\mathbb{R}$. We have that

$$|\,||X||_{\text{op}} - ||Y||_{\text{op}}\,| \leq ||X-Y||_{\text{op}} \leq ||X-Y||_{\text{F}} = ||\text{vec}(X) - \text{vec}(Y)||_2,$$

where the function "vec" maps from matrices to vectors. This shows that $||\cdot||_{\text{op}}$ is 1-Lipschitz. Thus by Gaussian concentration, we get that with probability $1-\delta$,

$$||X||_{\text{op}} \leq 2\mathbb{E}\left[||X||_{\text{op}}\right] + \sqrt{2\log(1/\delta)} = 2\sqrt{n} + \sqrt{2\log(1/\delta)}.$$

Let's finish with a theorem which we will discuss more next time. Recall that $\mathcal{N}(T,d,\epsilon)$ is the covering number of ϵ-balls (under distance d) over set T.

Theorem 12.8 (Sudakov minoration)**.** *Let X_t be a zero-mean Gaussian process. Define the distance*

$$d(t,s) = \sqrt{\mathbb{E}\left[(X_t - X_s)^2\right]}.$$

Then there exists an absolute constant $c > 0$ such that $\forall \varepsilon > 0$,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq c \cdot \varepsilon \cdot \sqrt{\log(\mathcal{N}(T,d,\varepsilon))}.$$

Lecture 13: Sudakov Minoration and Gaussian Processes

Instructor: Nikita Zhivotovskiy *Scriber: Max Hirsch* *Proofreader: Michael Xiao*

1 Sudakov Minoration

Theorem 1 (Sudakov minoration)**.** *Let X_t be a zero-mean Gaussian process indexed by $t \in T$ and define for $t, s \in T$,*

$$d(t,s) = \sqrt{\mathbb{E}(X_t - X_s)^2}.$$

Then for all $\varepsilon > 0$,

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

where $c > 0$ is some absolute constant.

Proof. Let $P_\varepsilon \subseteq T$ be such that P_ε is a maximum packing. In particular, for all $t, s \in P_\varepsilon$ we have $d(t,s) > \varepsilon$. Then

$$\mathcal{N}(T,d,\varepsilon) \le |P_\varepsilon|,$$

and

$$\mathbb{E} \sup_{t\in T} X_t \ge \mathbb{E} \sup_{t\in P_\varepsilon} X_t.$$

Now define the process $Y_t = \frac{\varepsilon}{\sqrt{2}} Z_t$ for $t \in P_\varepsilon$, where $Z_t \sim \mathcal{N}(0,1)$ and $Z_1, \dots, Z_{|P_\varepsilon|}$ are independent. We have that for all $t, s \in P_\varepsilon$,

$$\mathbb{E}(X_t - X_s)^2 = d(t,s)^2 > \varepsilon^2, \quad \text{and} \quad \mathbb{E}(Y_t - Y_s)^2 = \varepsilon^2,$$

where the second equality is by construction. It follows by the Sudakov-Fernique theorem that

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

Finally, observe that

$$\mathbb{E} \sup_{t\in P_\varepsilon} Y_t = \frac{\varepsilon}{\sqrt{2}} \mathbb{E} \left(\max_{i \in \{1,\dots,|P_\varepsilon|\}} Z_t \right) \gtrsim \varepsilon \cdot C \sqrt{\log(|P_\varepsilon|)},$$

where $C > 0$ is an absolute constant, and the last inequality is an exercise used in homework. Combining these inequalities yields

$$\varepsilon \sqrt{\log \mathcal{N}(T,d,\varepsilon)} \le \varepsilon \sqrt{\log(|P_\varepsilon|)} \le C^{-1} \mathbb{E} \sup_{t\in P_\varepsilon} Y_t \le C^{-1} \mathbb{E} \sup_{t\in P_\varepsilon} X_t \le C^{-1} \mathbb{E} \sup_{t\in T} X_t.$$

It suffices to take $c = C^{-1}$. □

1.1 Canonical Gaussian Process Covering Number Examples

We now consider examples in which we use Theorem 1 to give bounds on covering numbers. The setup is as follows: Let $T \subseteq \mathbb{R}^d$ and $X_t = \langle g, t \rangle$ with $t \in T$ and $g \sim \mathcal{N}(0, I_d)$. Observe that

$$d(t,s)^2 = \mathbb{E}(X_t - X_s)^2 = \mathbb{E}(\langle g, t-s \rangle)^2 = \|t-s\|_2^2$$

so that $d(t,s) = \|t-s\|_2$ for $t, s \in T$. Now consider the following examples:

1. Let $T = \Sigma^{1/2} B_2^d$. By Theorem 1, we have

$$\varepsilon \sqrt{\log \mathcal{N}(\Sigma^{1/2} B_2^d, \|\cdot\|_2, \varepsilon)} \le c\mathbb{E} \sup_{t \in \Sigma^{1/2} B_2^d} \langle g, t \rangle = c\mathbb{E}\|g'\|_2 \le c\sqrt{\mathrm{Tr}(\Sigma)},$$

where $g' = \Sigma^{1/2} g \sim \mathcal{N}(0, \Sigma)$ and the last inequality follows from Jensen's inequality:

$$\mathbb{E}\|g'\|_2 \le \sqrt{\mathbb{E}\|g'\|_2^2} = \sqrt{\mathrm{Tr}(\Sigma)}.$$

It follows that

$$\log \mathcal{N}(\Sigma^{1/2} B_2^d, \|\cdot\|_2, \varepsilon) \le \frac{c_1 \,\mathrm{Tr}(\Sigma)}{\varepsilon^2},$$

with $c_1 = c^2$.

2. Now consider $T = B_1^d = \{x \in \mathbb{R}^d : \|x\|_1 \le 1\}$. Then by Theorem 1,

$$\varepsilon \sqrt{\log \mathcal{N}(B_1^d, \|\cdot\|_2, \varepsilon)} \le c\mathbb{E} \sup_{t \in B_1^d} \langle g, t \rangle = c\mathbb{E}\|g\|_\infty = c\mathbb{E} \max_{i \in [d]} |g_i| \le c_1 \sqrt{\log(2d)}$$

for some constants $c, c_1 > 0$. It follows that for some $c_2 > 0$,

$$\log \mathcal{N}(B_1^d, \|\cdot\|_2, \varepsilon) \le \frac{c_2 \log(2d)}{\varepsilon^2}. \tag{1}$$

Remark 2. *The same proof works for polytopes with unit diameter and d vertices.*

Now we compare this with a volumetric argument. As an exercise, it is easy to show that $B_1^d \subseteq B_2^d \subseteq \sqrt{d} B_1^d$. Then we have

$$\mathcal{N}(B_1^d, \|\cdot\|_2, \varepsilon) \le \frac{\mathrm{Vol}(B_1^d + \frac{\varepsilon}{2} B_2^d)}{\mathrm{Vol}(\frac{\varepsilon}{2} B_2^d)} \le \frac{\mathrm{Vol}(B_1^d (1 + \frac{\varepsilon\sqrt{d}}{2}))}{\mathrm{Vol}(\frac{\varepsilon}{2} B_2^d)} = \frac{(1 + \frac{\varepsilon\sqrt{d}}{2})^d}{(\varepsilon/2)^d} \left(\frac{\mathrm{Vol}(B_1^d)}{\mathrm{Vol}(B_2^d)} \right) \le \left(c \left(\frac{2}{\varepsilon\sqrt{d}} + 1 \right) \right)^d,$$

where we used the fact from Wikipedia that

$$\frac{\mathrm{Vol}(B_1^d)}{\mathrm{Vol}(B_2^d)} \le \left(\frac{c}{\sqrt{d}} \right)^d.$$

It follows that we have

$$\log \mathcal{N}(B_1^d, \|\cdot\|_2, \varepsilon) \le d \log \left(c \left(1 + \frac{2}{\varepsilon\sqrt{d}} \right) \right).$$

Combining this with the first bound (1) gives

$$\log \mathcal{N}(B_1^d, \|\cdot\|_2, \varepsilon) \le \min \left(\underbrace{d \log \left(c \left(1 + \frac{2}{\varepsilon\sqrt{d}} \right) \right)}_{I}, \underbrace{\frac{c_2 \log(2d)}{\varepsilon^2}}_{II} \right).$$

Note that when $\varepsilon = 1/\sqrt{d}$, we have $I \approx d$ and $II \approx d \log d$. When $\varepsilon \lesssim 1/\sqrt{d}$, the bound I is better, while II is better for $\varepsilon \gtrsim 1/\sqrt{d}$.

2 Dual Sudakov Minoration

Definition 3. *Take T a convex and symmetric (meaning $T = -T$) set in $\mathbb{R}^d$. Then the polar set T° is*

$$T^\circ = \left\{ y \in \mathbb{R}^d : \sup_{x \in T} \langle x, y \rangle \leq 1 \right\}.$$

We list a few examples of polar sets:

1. $(B_2^d)^\circ = B_2^d$.
2. $(B_p^d)^\circ = B_q^d$, where B_p^d is the ℓ_p ball with $p \geq 1$ and $p^{-1} + q^{-1} = 1$.
3. If T is an ellipsoid with semi-axes $a_1, \ldots, a_d > 0$ then T° is an ellipsoid with semi-axes $a_1^{-1}, \ldots, a_d^{-1}$.

In what follows, we will use the notation $\mathcal{N}(T, \|\cdot\|_2, \varepsilon) := \mathcal{N}(T, \varepsilon B_2^d)$, the minimum number of εB_2^d required to cover T.

Theorem 4 (Dual Sudakov minoration)**.** *If T is a symmetric convex body, then for all $\varepsilon > 0$,*

$$\varepsilon \sqrt{\log \mathcal{N}(B_2^d, \varepsilon T^\circ)} \leq c \mathbb{E} \sup_{t \in T} \langle g, t \rangle.$$

We will not prove this result. We further have the following conjecture:

Conjecture 5. *For any T, K convex, symmetric bodies, there are $c, C > 0$ universal constants such that*

$$c \log \mathcal{N}(T, K) \leq \log \mathcal{N}(K^\circ, T^\circ) \leq C \log \mathcal{N}(T, K).$$

2.1 Euclidean Ball Covering Number

Let $T = \Sigma^{1/2} B_2^d$ and note that

$$\mathbb{E} \sup_{t \in T} \langle g, t \rangle \leq \sqrt{\mathrm{Tr}(\Sigma)}.$$

We have that

$$T^\circ = \{ y \in \mathbb{R}^d : \sup_{x \in T} \langle x, y \rangle \leq 1 \} = \{ y \in \mathbb{R}^d : \|\Sigma^{1/2} y\|_2 \leq 1 \}.$$

Thus, we are covering B_2^d with the sets $\{ y \in \mathbb{R}^d : \|\Sigma^{1/2} y\|_2 \leq 1 \}$. Defining

$$d_\Sigma(t, s)^2 = (t - s)^\top \Sigma (t - s),$$

we then obtain by Theorem 4 that

$$\log \mathcal{N}(B_2^d, d_\Sigma, \varepsilon) \leq \frac{c \, \mathrm{Tr}(\Sigma)}{\varepsilon^2}.$$

3 Gaussian Width

Definition 6. *Let $T \subseteq \mathbb{R}^d$ and $g \sim \mathcal{N}(0, I_d)$. Then the Gaussian width of T is*

$$W(T) = \mathbb{E} \sup_{t \in T} \langle t, g \rangle.$$

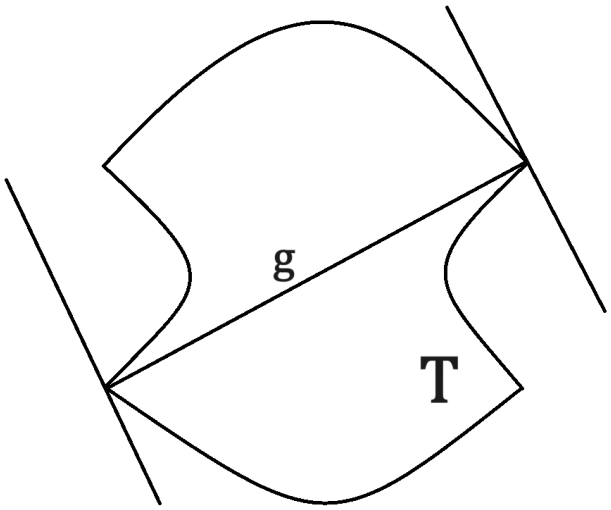


Figure 1: Gaussian width measures the width of T in the direction g and averages over $g \sim \mathcal{N}(0, I_d)$.

3.1 Properties of Gaussian Width

We have the following properties of the Gaussian width:

1. $W(T)$ is finite if and only if T is bounded.
2. If Q is an orthogonal matrix and y a fixed vector, then $W(QT + y) = W(T)$.
3. $W(T+K) = W(T)+W(K)$ and $W(\alpha T) = |\alpha|W(T)$, where $\alpha \in \mathbb{R}$ and we recall that $T+K = \{t+k : t \in T, k \in K\}$.
4. $W(T) = \frac{1}{2}W(T-T) = \frac{1}{2}\mathbb{E}\sup_{x,y\in T}\langle g, x-y\rangle$.
5. If T is a finite set then $W(T) \leq c\sqrt{\log(|T|)} \cdot \operatorname{diam}(T)$.

3.2 Gaussian Width Examples

1. $W(B_2^d) = \mathbb{E}\|g\|_2 \leq \sqrt{d}$
2. $W(\Sigma^{1/2}B_2^d) \leq \sqrt{\operatorname{Tr}(\Sigma)}$
3. $W(B_1^d) = \mathbb{E}\|g\|_\infty \leq \sqrt{2\log(2d)}$
4. $W(B_\infty^d) = \mathbb{E}\|g\|_1 = d\sqrt{\frac{2}{\pi}}$

3.3 Gaussian Concentration Inequality

We conclude this section with the following result:

Theorem 7 (Gaussian concentration inequality)**.** *Let $\varphi_1, \ldots, \varphi_d : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz. Then*

$$\mathbb{E}\sup_{t\in T}\sum_{i=1}^{d} g_i\varphi_i(t_i) \leq \mathbb{E}\sup_{t\in T}\sum_{i=1}^{d} g_i t_i = W(T).$$

Proof. For $t, s \in T$, we have that

$$\mathbb{E}\left(\sum_{i=1}^{d}(g_i\varphi_i(t_i) - g_i\varphi_i(s_i))\right)^2 = \sum_{i=1}^{d}(\varphi_i(t_i) - \varphi_i(s_i))^2 \leq \sum_{i=1}^{d}(t_i - s_i)^2 = \mathbb{E}\left(\sum_{i=1}^{d} g_i(t_i - s_i)\right)^2,$$

so applying the Sudakov-Fernique theorem yields the result. □

4 Next Time

Next lecture, we will begin discussing empirical processes. As an example, consider $X_1, \ldots, X_n$ i.i.d. random variables and the CDF and empirical CDF

$$F(t) = \Pr(X \leq t), \quad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Ind}[X_i \leq t].$$

To test whether this empirical distribution came from the distribution corresponding to the CDF F, Kolmogorov suggested the test statistic

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)|.$$

As $n \to \infty$, if the data are sampled from the distribution F, then this statistic converges almost surely to 0. Our question is this: for a finite sample size n, what should we expect from the above test statistic?

Lecture 14: Empirical Process Theory

Instructor: Nikita Zhivotovskiy *Scriber: Xueda Shen* *Proofreader: Xueda Shen*

1 Motivation

Last class we introduced the KS-test statistics to motivate the study of empirical process theory. Suppose we observe $X_i, i = 1, ..., n$ sampled i.i.d. from a distribution. We would like to test whether $X_i \sim P$ a given probability measure. Let $F(t) := P(X \leq t)$ the population cdf function, $F_n(t) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \leq t\}$. In fact, the KS statistics is known as an instantiation of a wider class of process called empirical process.

Definition 1 (Empirical Process)**.** *Given $\mathcal{F}$ a class of functions with $X_{1:n} \sim_{i.i.d.} P$. The process $\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)$ is called empirical process indexed by $\mathcal{F}$.*

Definition 2 (Gilvenko-Cantelli)**.** *The function class $\mathcal{F}$ is called Gilvenko-Cantelli with respect to measure P_X if $\sup_{f\in\mathcal{F}} \left|\mathbb{E}f(X) - \frac{1}{n}f(X_i)\right| \to 0$ almost surely.*

2 Symmetrization

One of the central techniques used to analyze empirical process is to establish an expectation upper bound via symmetrization.

Lemma 3 (Symmetrization upperbound)**.** *Let $\varepsilon_1, ..., \varepsilon_n$ be i.i.d. Rademacher random variables, $X_1, ..., X_n \sim_{i.i.d.} P$ We have:*

$$\mathbb{E}_X \sup_{f\in\mathcal{F}} \left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] \leq 2\,\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right]$$

$$\mathbb{E}_X \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right] \leq 2\,\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right]$$

$$\mathbb{E}_X \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right| \leq 2\,\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i)\right|$$

Proof. We only prove the first one, as the rest are identical arguments. Introduce ghost samples $X_1', ..., X_n'$ i.i.d copies of X_i. We have that

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}} \left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] &= \mathbb{E}_X \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}f(X_i') - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] \\ &\leq \mathbb{E}_X\mathbb{E}_{X'} \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^n f(X_i') - f(X_i)\right]\end{aligned}$$

The first inequality essentially follows from the observation that for $f = f(X, \beta)$, $\sup_\beta \mathbb{E}f(X,\beta) \leq \mathbb{E}\sup_\beta f(X,\beta)$ Now observe that $f(X_i) - f(X_i') \overset{D}{=} \varepsilon(f(X_i) - f(X_i'))$ where ε is a Rademacher variable. This could be seen

via conditioning on the value of ε. This leaves us with

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_{X'}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n f(X_i^{'})-f(X_i)\right] &= \mathbb{E}_X\mathbb{E}_{X'}\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(f(X_i^{'})-f(X_i)\right)\right]\\ &\leq \mathbb{E}_X\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right]+\mathbb{E}_{X'}\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i^{'})\right]\\ &= 2\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right]\end{aligned}$$

where in the first inequality we observed that $\varepsilon f(X_i) = -\varepsilon f(X_i)$. □

3 Desymmetrization

Alternatively, given quantity of the form $\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)$, we can upperbound it via Desymmetrization, essentially unwinding what we have done before.

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right] &= \mathbb{E}_X\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \left(\varepsilon_i f(X_i)+\mathbb{E}f(X_i)-\mathbb{E}f(X_i)\right)\right]\\ \leq \mathbb{E}_X\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(f(X_i)-f(X_i^{'})\right)\right] &+ \mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{E}f(X_i)\right]\end{aligned}$$

We first analyze the first term on RHS.

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(f(X_i)-f(X_i^{'})\right)\right] &= \mathbb{E}_{X,X'}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(f(X_i)-\mathbb{E}f(X_i)+\mathbb{E}f(X_i^{'})-f(X_i^{'})\right)\right]\\ &= \mathbb{E}_{X,X'}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \left(f(X_i)-\mathbb{E}f(X_i)+\mathbb{E}f(X_i^{'})-f(X_i^{'})\right)\right]\\ &\leq \mathbb{E}_{X,X'}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \left(f(X_i)-\mathbb{E}f(X_i)+\mathbb{E}f(X_i^{'})-f(X_i^{'})\right)\right|\right]\\ &\leq \mathbb{E}_X\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}f(X_i)\right|\right]+\mathbb{E}_{X'}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i^{'})-\right.\right.\\ &= 2\mathbb{E}_X\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}f(X_i)\right|\right]\end{aligned}$$

Recall our remainder term is yet analyzed. This is a great place to exercise Hölder's inequality.

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

This allows us to immediately recognize $|\sum_{i=1}^n \varepsilon_i\mathbb{E}f(X_i)|$ as a 1-norm and apply Hölder's inequality with $1,\infty$ norm.

$$\left|\sum_{i=1}^n \varepsilon_i\mathbb{E}f(X_i)\right| \leq \left|\sum_{i=1}^n \varepsilon_i\right| \,|\mathbb{E}f(X)|$$

The last term to analyze is $\mathbb{E}|\sum_{i=1}^n \varepsilon_i|$. However, we have

$$\mathbb{E}\left|\sum_{i=1}^n \varepsilon_i\right| = \mathbb{E}\sqrt{\left|\sum_{i=1}^n \varepsilon_i\right|^2} \leq \sqrt{\mathbb{E}\left|\sum_{i=1}^n \varepsilon_i\right|^2} = \sqrt{n}$$

Recollecting the pieces, we are able to bound the remainder

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}f(X_i)\right] \leq n^{-\frac{1}{2}} \sup_{f\in\mathcal{F}} |\mathbb{E}f(X)|$$

Remark 4. *The same conclusion applies to* $\mathbb{E}\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)|$, *whose proof we essentially established.*

4 Analysis of KS statistics

We analyze the KS test statistics by establishing expectation and high probability bound in turn. Via symmetrization, we immediately have

$$\mathbb{E}_X \sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \leq t\} - P(X\leq t)\right| \leq 2\,\mathbb{E}_X\mathbb{E}_\varepsilon\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}\{X_i\leq t\}\right|$$

A naive union bound would not work since there are uncountably many events involved. However, a closer inspection tells us that conditional on $X_{1:n}$, at most $n+1$ values of $\mathbf{1}\{X_i \leq t\}$ is realizable. Hence if we are able to find a random variable is sub-Gaussian, then this expectation could be controlled. Let's find sG constant for $\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i\leq t\}$ with a fixed t. A quick calculation suggests the variance proxy $\sigma \leq n^{-\frac{1}{2}}$. Hence by maximal inequality, we have

$$\mathbb{E}_X\left\{\mathbb{E}_\varepsilon \sup_{t\in T, |T|\leq n+1}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i\leq t\}\right]\right\} \leq 2\sqrt{\frac{2\log(2(n+1))}{n}}$$

Next we establish the high probability bound by relating $\sup_{f\in\mathcal{F}}\left|\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right|$ to its expectation. We can readily establish such bounds if the function class in question is uniformly bounded: i.e., $||f||_\infty \leq l$. This leads to the following proposition

Proposition 5. *Consider a uniformly bounded function class* $\mathcal{F}$, *where* $||f||_\infty \leq l$. *Then with probability at least* $1-\delta$ *we have*

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

Proof. The main task is to ascertain the bounded difference constant for a suitably defined function. Afterwhich we can just apply the bounded difference inequality. Consider

$$g(X_{1:n}) = \sup_{f\in\mathcal{F}}\left|\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right|$$

Let $X_{1,,i',n}$ denote the sequence of $X_1, ..., X_{i-1}, X_i', X_{i+1}, ..., X_n$ i.e. where $i-$th element is replaced. We now work out the bounded difference constant of g. Without loss of generality, suppose $f^\star \in \mathcal{F}$ be the maximizer

of $\left|\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right|$. We have that

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right| - \sup_{f'\in\mathcal{F}}\left|\mathbb{E}f'(X) - \frac{1}{n}\left(\sum_{j\neq i} f'(X_j) + f'(X_i')\right)\right| \\
&\leq \left|\mathbb{E}f^\star(X) - \frac{1}{n}\sum_{i=1}^{n} f^\star(X_i)\right| - \left|\mathbb{E}f^\star(X) - \frac{1}{n}\left(\sum_{j\neq i} f^\star(X_j) + f^\star(X_i')\right)\right| \\
&\leq \left|\frac{1}{n}\left(f^\star(X_i) - f^\star(X_i')\right)\right| \leq \frac{2b}{n}.
\end{aligned}$$

The first inequality follows from

$$\left|\mathbb{E}f^\star(X) - \frac{1}{n}\left(\sum_{j\neq i} f^\star(X_j) + f^\star(X_i')\right)\right| \leq \sup_{f'\in\mathcal{F}}\left|\mathbb{E}f(X) - \frac{1}{n}\left(\sum_{j\neq i} f(X_j) + f(X_i')\right)\right|$$

since the maximizer with X_i' instead of X_i is not necessary $f^\star$. The second inequality is an application of $|a| - |b| \leq |a - b|$, and the final inequality applies boundedness assumption. Apply bounded difference inequality. □

5 Vapnik-Chervonenkis Theory

Before we shift gear to discuss VC theory, we introduce some empirical process theory notations, and motivate why we shift to discuss such theory. Let $\mathcal{A}$ denote a collection of events, and $X_{1:n}$ i.i.d samples on $\mathcal{X}$. We define $Pf := \mathbb{E}f(X), P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. We are interested in the following quantity, also called Uniform Law of Large Numbers:

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|$$

where $P(A) = \Pr(X \in A); P_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[X_i \in A]$

Definition 6 (Growth/Shattering for a set of events)**.** *Let $n \in \mathbb{N}$. The shattering number of a set of events $\mathcal{A}$ is*

$$S_{\mathcal{A}}(n) := \max_{x_1,\ldots,x_n\in\mathcal{X}} |\{(\mathbf{1}\{x_1 \in A\}, ..., \mathbf{1}\{x_n \in A\}), A \in \mathcal{A}\}|$$

In words, the shattering number is the maximum cardinality of set of binary vectors.

We finally introduce some basic properties of shattering function.

- $S_{\mathcal{A}}(n) \leq 2^n$
- If $|\mathcal{A}| \leq \infty, S_{\mathcal{A}}(n) \leq |\mathcal{A}|$
- If $\mathcal{A}$ is induced by cylinder sets $(-\infty, t)$, then $S_{\mathcal{A}}(n) = n + 1$

Lecture 15: Shattering Function Bound & VC dimension

Instructor: Nikita Zhivotovskiy *Scribe: Rita Lyu* *Proofreader: Xueda Shen*

1 Notations

Let $\mathcal{A}$ denote a collection of events, and $X_1, \cdots, X_n$ denote i.i.d. samples on $\mathcal{X}$.

2 Growth function

Definition 1 (Growth function/Shattering function for a set of events)**.** *Let $n \in \mathbb{N}$. The shattering number of a set of events $\mathcal{A}$ is*

$$S_{\mathcal{A}}(n) := \max_{x_1,\cdots,x_n \in \mathcal{X}} |\{(\mathbf{1}\{x_1 \in A\}, \ldots, \mathbf{1}\{x_n \in A\}), A \in \mathcal{A}\}|$$

In words, the shattering number is the maximum number of different values of n indicator functions that can take on a set of events $\mathcal{A}$.

For $S_{\mathcal{A}}$, we know it has following properties:

- $S_{\mathcal{A}}(n) \leq 2^n$ (this is because we have n binary elements and this inequality always holds, but is not a good bound.)
- If $|\mathcal{A}| < \infty, S_{\mathcal{A}}(n) \leq |\mathcal{A}|$ (this is because of each A in $\mathcal{A}$, we only have one vector with binary outcome.)
- If $\mathcal{A}$ is induced by cylinder sets $(-\infty, t)$, then $S_{\mathcal{A}}(n) = n + 1$.

Proposition 2. *If the family of events $\mathcal{A}$ has the shatter function $S_{\mathcal{A}}(n)$, then with probability at least $1-\delta$,*

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \leq 2\sqrt{\frac{2\log(2S_{\mathcal{A}}(n))}{n}} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}},$$

where $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \in A\}$ is the empirical measure and $P(A) = \Pr(x \in A)$.

Remark 3. *We emphasize (i) this bound only holds when $X_1, \cdots, X_n$ are i.i.d. samples; (ii) $\mathcal{A}$ is the collection of events and can be infinite. This proposition indicates that even though $|\mathcal{A}|$ is infinite, the total number can be bounded by $S_{\mathcal{A}}(n)$, which is the projection to indicators; (iii) This bound holds simultaneously for all events in $\mathcal{A}$. (iv) For the bound, the first 2 (outside of square root) comes from the symmetrization, the other 2s come from the bound; (v) The typical example is that when $S_{\mathcal{A}}(n) = n + 1$, then the bound becomes $O\left(\sqrt{\frac{\log(n)}{n}}\right)$ (see Big O notation for more detailed explanation for this notation), and we can see as $n \to \infty$, the error converges to 0.*

The general idea is that the "uniform law of large numbers" holds if $S_{\mathcal{A}}(n) \ll 2^n$, because otherwise $\sqrt{\frac{\log(S_{\mathcal{A}(n)})}{n}}$ will not converges to 0, as $n \to \infty$.

Example 4. Let $\mathcal{A}$ be the collection of all subsets of $\mathbb{R}$, then $S_{\mathcal{A}}(n) = 2^n$, and the first term in error bound turns to be $2\sqrt{\frac{2(n+1)}{n}}$, which is the constant bound and does not converge to 0.

Now the question comes when can we have $S_{\mathcal{A}}(n) \ll 2^n$? The answer is that we can infer whether $S_{\mathcal{A}}(n) \ll 2^n$ by looking at the Vapnik-Chervonenkis dimension (VC dim) of the set $\mathcal{A}$.

3 VC dimension

Definition 5 (VC dimension). *The VC dimension of $\mathcal{A}$ is the largest integer d, such that $S_{\mathcal{A}}(d) = 2^d$.*

Remark 6. *We can regard $VC(\mathcal{A}) = \max\{i | S_{\mathcal{A}}(i) = 2^i\}$ and it characterizes the richness of the class $\mathcal{A}$.*

We introduce the following examples to characterize the interplay between the shattering function and the VC dimension.

Example 7. Let $\mathcal{A}$ be the set of all closed intervals in $\mathbb{R}$. We determine the VC dimension by working through the shattering function with $d = 2$ and $d = 3$. We claim VC dimension is 2, so we need to show when d = 3, we have less than 8 patterns. From Figure 7, we can see

1. When there are just two points, $S_{\mathcal{A}}(2) = 2$.
2. When there are three points, $S_{\mathcal{A}}(3)$ is at most 2^3 patterns. However, the pattern in the Figure cannot be realized. Thus, $VC(\mathcal{A}) = 2$.

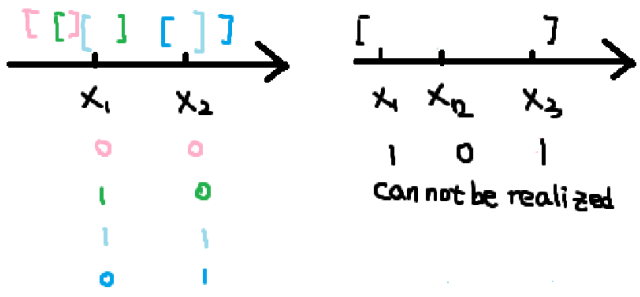


Figure 1: VC dimension for Example 7.

Remark 8. *For $\mathcal{A}$, $VC(\mathcal{A}) = d$ means (i) $\forall n \leq d$, we are able to have 2^n patterns by the definition 1, (ii) when $n > d$, the 2^n patterns cannot be realized.*

Example 9. $\mathcal{A}$ is induced by half spaces in $\mathbb{R}^2$, then $VC(\mathcal{A}) = 3$. From the left panel of Figure 2, we can find a shattered set of size 3, satisfying $S_{\mathcal{A}}(3) = 2^3$. However, we cannot find a half space such that the right panel is realized, because the convex hulls of between points with label 1 and points with label 0 intersect. This would work for any 4 points, not only for those on the picture.

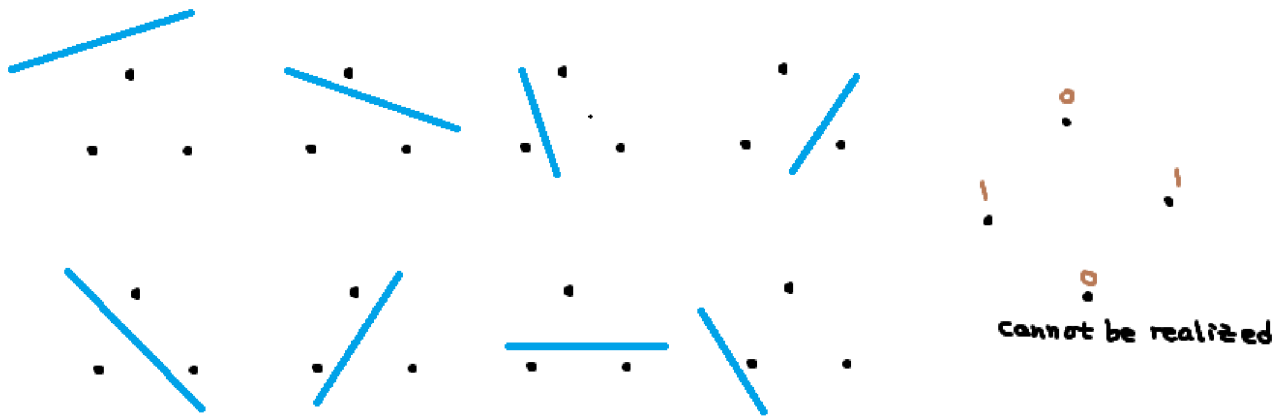


Figure 2: VC dimension for Example 9.

More rigorously, we provide the following theorem without proof (see more discussion in Radon Theorem).

Theorem 10 (Radon). *If we have $p+2$ points in $\mathbb{R}^p$, then we can split these points into two groups $A \sqcup B$ (no intersect), such that their convex hulls intersect.*

Thus, in this example, with $p = 2$, we can always separate 4 points into two groups with their convex hulls intersecting. Then, we cannot make $S_{\mathcal{A}}(4) = 2^4$. Based on Theorem 10, we have the following corollary.

Corollary 11. *For $\mathcal{A}$ induced by half spaces in $\mathbb{R}^p$, $VC(\mathcal{A}) = p + 1$.*

For example, a simplex in $\mathbb{R}^3$, we can find 2^{3+1} binary vectors, thus $VC(\mathcal{A}) \geq 3 + 1$. But because of Theorem 10, if we take any $3 + 2$ points, there are 2 groups with intersecting convex hulls. Thus, $VC(\mathcal{A}) < 3 + 2$. Finally, $VC(\mathcal{A}) = 4$.

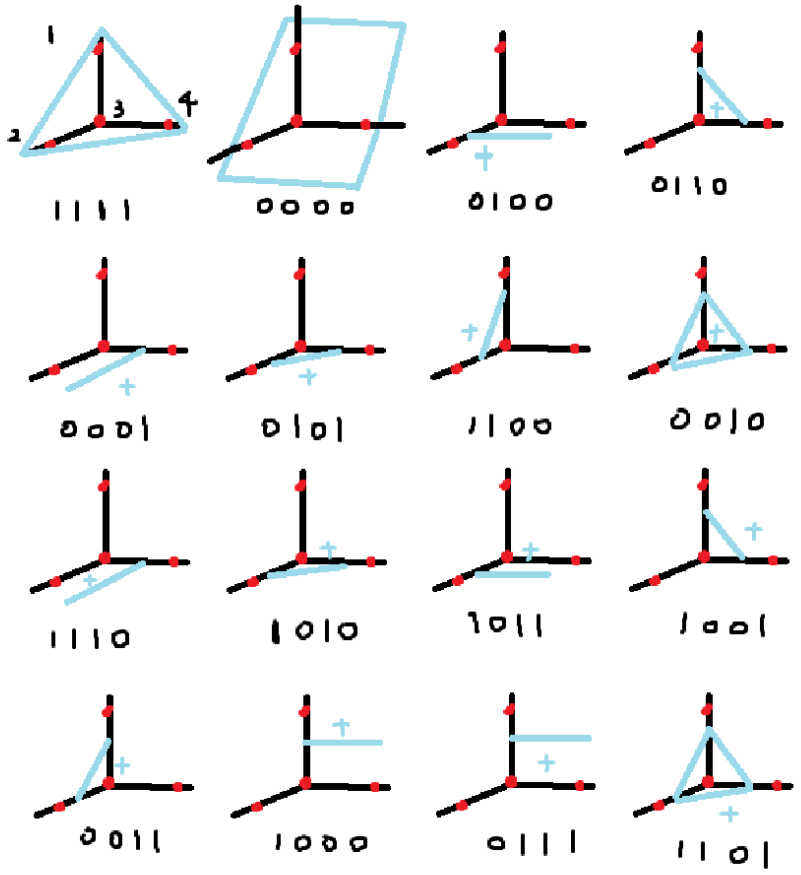


Figure 3: VC dimension for simplex in $\mathbb{R}^3$.

We claim that for $\mathcal{A}$, if VC dimension is small, then the shattering function is also small.

Theorem 12 (Sauer-Shelah-Vapnik-Chervonenkis). *If $VC(\mathcal{A}) = d$, then*

$$S_{\mathcal{A}}(n) \leq \sum_{i=0}^{n} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d (n \geq d).$$

Proof for the second inequality can be found in Lecture 9 Section 2 of Sparse Linear Regression.

Remark 13. *From Theorem 12, we can see if $d = \infty$, then $S_{\mathcal{A}}(n) = 2^n$, which is the naive bound. If $d < \infty$, $S_{\mathcal{A}}(n) = O(n^d)$, which is the polynomial bound.*

Corollary 14. *If $VC(\mathcal{A}) = d$, by applying Proposition 2 and Theorem 12, then with probability $1 - \delta$,*

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \leq 4\sqrt{\frac{d \log\left(\frac{en}{d}\right)}{n}} + \sqrt{\frac{2 \log\left(\frac{1}{\delta}\right)}{n}}.$$

If d is finite, we can see

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| = O\left(\sqrt{\frac{d\log(n)}{n}}\right) \to 0, \ \text{as } n \to \infty.$$

Remark 15. *This corollary result indicates the relationship between the VC dimension, the shattering function, and the uniform law of large numbers. We can bound the distance between the empirical measure and the true probability measure using the shattering function. Moreover, the shattering function can have an upper bound induced by the VC dimension. Then, we can characterize whether the uniform law of large numbers holds by looking at the VC dimension of the class $\mathcal{A}$.*

Corollary 16. *From Corollary 14, we can see that set system with finite VC dimension satisfy the uniform law of large numbers.*

4 Proof of Proposition 2

Proof. Because our $\mathbf{1}\{x \in A\}$ is an indicator function, it is bounded by 1. Apply results in Lecture 14 Section 4, we observe that conditional on $X_{1:n}$, at most $S_{\mathcal{A}}(n)$ values of $\sum_{i=1}^n \mathbf{1}\{X_i \in A\}, \forall A \in \mathcal{A}$ is realizable. Then by symmetrization and the maximal inequality, we have

$$\mathbb{E}\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \le 2\sqrt{\frac{2\log(2S_{\mathcal{A}}(n))}{n}}.$$

Because the function is bounded by 1, we apply the Proposition 5 in Lecture 14 with $l = 1$, then we have, with probability at least $1-\delta$,

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \le \mathbb{E}\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \le 2\sqrt{\frac{2\log(2S_{\mathcal{A}}(n))}{n}} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}.$$

□

5 Proof of Theorem 12

Proof. Our main idea is that in the definition of shattering function, we just care about the binary function (fix $x_1, \ldots, x_n, (\mathbf{1}\{x_1 \in A\}, \ldots, \mathbf{1}\{x_n \in A\}) \in \{0,1\}^n$)[1]. Thus, we reduce our problem to counting the size of V, where V is the matrix whose row vectors are realized values of indicator vector with n columns, (from set system to the matrix), such that, by VC dimension definition, (i) we can find d columns in V, such that all 2^d vectors are realized, (ii) $\forall d+1$ columns, we have smaller than 2^{d+1} different vectors.

$$V = \underbrace{\begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 1 & 1 \\ & & \cdots & & \end{bmatrix}}_{n}.$$

E.g. For VC dimension 2, for the first two columns, we find 2^2 distinct binary row vectors, but for the first three columns, we cannot see full 2^3 row vectors.

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

[1] We do not fix arbitrary $x_1, \ldots, x_n$. We fix $x_1, \ldots, x_n$ that satisfy the VC dimension.

We use shifting-based proof here. Shifting allows us to change the original set of vectors without changing its size. Induction-based proof can be found in [Vershynin, 2018]. Before moving on, we first introduce the shifting operator S. For $v \in V$, take $i \in [n]$, we change 1 to 0 iff this does not cause copies of the same vector.

$$S_i(v) = \begin{cases} (v_1, \ldots, v_{i-1}, 0, v_{i+1}, \ldots, v_n), \text{ if } (v_1, \ldots, v_{i-1}, 0, v_{i+1}, \ldots, v_n) \notin V \\ v = (v_1, \ldots, v_{i-1}, v_i, v_{i+1}, \ldots, v_n), \text{ Otherwise} \end{cases}.$$

If we apply S_i to every vector in V, then by construction

$$|S_i(V)| = |V|,$$

meaning that the size will not change after shifting. More importantly, if a subset of columns is shattered by $S_i(V)$, then it is also shattered by the original V ("shattered by " here means if we project these columns to indicator and we can see all patterns of 2 taking power to the number of columns). E.g.

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \rightarrow \text{shifting on the first column } \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \text{ all two combinations are possible.}$$

We then prove this claim.

Proof. W.L.G, take $i = 1$, columns are first K columns, $1, \ldots, K$, by definition of shattering, $\forall u \in \{0,1\}^K$ (for all possible patterns in K columns), $\exists v \in S_i(V)$, such that $u_i = v_i, i = 1, \cdots, K$. The previous claim holds because our condition is that the first K columns are shattered by $S_i(V)$. We next want to show there is $v' \in V$, such that $u_i = v'_i, i = 1, \cdots, K$. Assume $u = [1, u_2, \cdots, u_K, \cdots], \exists v \in S_i(V), u_i = v_i, i \in [K]$

1. one pattern $u = [1, u_2, \cdots, u_K, \cdots]$, is a vector starting with 1 after shifting means $[0, u_2, \cdots, u_K] \in V$ (because shifting will lead to $[0, u_2, \cdots, u_K, \cdots]$ when the vector does not in V, but now the vector starts with 1, meaning $[0, u_2, \cdots, u_K, \cdots]$ is already in V.)
2. Since 1 does not be changed, meaning that $[1, u_2, \cdots, u_K, \cdots]$ is also in V.

$\Rightarrow, \forall u \in \{0,1\}^K, \exists v' \in V$, such that $v'_i = u'_i, i \in [K]$. □

Then, we apply $S_i(V)$ in a loop for all i, in any order until we reach such a set V^*, so that for all i,

$$S_i(V^*) = V^*, \forall i,$$

meaning that shifting will not change the element anymore. We know that by construction, $|V^*| = |V|$. Finally, we need to compute $|V^*|$, for $v \in V^*$. W.L.G., we assume all ones are stacked at the front, that is,

$$1, 1, 1, \cdots, 1, 0, \cdots, 0, 0.$$

If we assume there are more than $(d+1)-$ones at the beginning, as shifting will not change the size (it is already stable), then

$$\begin{gathered} 0, 1, 1, \cdots, 1, 0, \cdots, 0, 0 \\ 1, 0, 1, \cdots, 1, 0, \cdots, 0, 0 \\ 0, 1, 1, \cdots, 1, 0, \cdots, 0, 0 \\ \cdots \end{gathered}$$

should also be in V^*. Finally, the more than $(d+1)-$ones by shifting will result in more than 2^{d+1} patterns. However, this contradicts the VC dimension. This is because if the subset is shattered by $S_i(V)$, then it should also be shattered by the original V. We conclude that no vector $v \in V^*$ contains more than d-ones.

So the vector in V^* contains at most d-ones. Therefore, by taking the number of ones to be 0 to d, we have

$$|V^*| = |V| \leq \sum_{i=0}^{d} \begin{pmatrix} n \\ i \end{pmatrix}.$$

□

6 Proof of Corollary 14

Proof. Based on Proposition 2 and Theorem 12, we know that

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \leq 2\sqrt{\frac{2\log(2S_{\mathcal{A}}(n))}{n}} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}},$$

$$\log(2S_{\mathcal{A}}(n)) \leq \log\left(2\left(\frac{en}{d}\right)^d\right) \leq \log\left(\left(\frac{en}{d}\right)^{2d}\right) = 2d\log\left(\frac{en}{d}\right)$$

$$\Rightarrow \sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \leq 4\sqrt{\frac{d\log\left(\frac{en}{d}\right)}{n}} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}.$$

□

References

[Vershynin, 2018] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Lecture 16: Empirical Risk Minimization & Dudley Integral

Instructor: Nikita Zhivotovskiy *Scriber: Jingxi Wang* *Proofreader: Jinglin Yang*

16.1 Example: Statistical learning (classification)

16.1.1 Definitions

We will start by introducing some definitions. A set of classifiers is as follows $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$. $(X_i, Y_i)_{i=1}^n$ is an i.i.d sample from some unknown distribution $P_{X,Y}$, where $X \in \mathcal{X}$, $Y \in \{0,1\}$. Given $f \in \mathcal{F}$, ***the empirical error*** is defined by

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \text{Ind}\{f(X_i) \neq Y_i\},$$

and ***the population risk*** is defined by

$$R(f) = \Pr_{X,Y\sim P_{X,Y}}(f(X) \neq Y).$$

Next we introduce ***VC-dimension of*** $\mathcal{F}$. The following two definitions are equivalent: (a) We say $\mathcal{F}$ has the VC-dimension d, if d is the size of the largest shattered set; (b) Define $S_{\mathcal{F}}(n) := \max_{X_1,\dots,X_n} \#\{(f(X_1),\dots,f(X_n)) \mid f \in \mathcal{F}\}$, then the VC-dimension of $\mathcal{F}$ is the largest d such that $S_{\mathcal{F}}(d) = 2^d$.
For a classifier $\hat{f}$ constructed based on the sample, we define ***the excess risk*** is $R(\hat{f}) - \inf_{f\in\mathcal{F}} R(f)$. We denote $f^* = \text{argmin}_{f\in\mathcal{F}} R(f)$, and $\hat{f} = \arg\min_{f\in\mathcal{F}} R_n(f)$, i.e. f^* is the minimizer of $R(f)$ and $\hat{f}$ is the minimizer of $R_n(f)$. Then we define ***the generalization error*** to be $|R_n(\hat{f}) - R(\hat{f})|$.

16.1.2 Empirical Risk Minimization

Utilizing results about VC-dimension from previous lectures, we can bound $R(\hat{f}) - R(f^*)$.

Proposition 16.1.

$$R(\hat{f}) - R(f^*) \leq C\left(\sqrt{\frac{d\log(en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{n}}\right).$$

Proof. By definition of $\hat{f}$, we have $R_n(\hat{f}) - R_n(f^*) \leq 0$. It follows that

$$R(\hat{f}) - R(f^*) = R(\hat{f}) - R_n(\hat{f}) + R_n(\hat{f}) - R_n(f^*) + R_n(f^*) - R(f^*) \leq 2\sup_{f\in\mathcal{F}} |R(f) - R_n(f)|.$$

Now our goal is to bound $2\sup_{f\in\mathcal{F}} |R(f) - R_n(f)|$.
Step 1: Let $A_f := \{f(X) \neq Y\}$, $\mathcal{A}_{\mathcal{F}} := \{A_f | f \in \mathcal{F}\}$. We want to show $S_{\mathcal{F}}(n) = S_{\mathcal{A}_{\mathcal{F}}}(n)$. Verify that the value of

$$\sup_{\substack{X_1,\dots,X_n\in\mathcal{X}\\ Y_1,\dots,Y_n\in\{0,1\}}} \#\{(\text{Ind}[f(X_1) \neq Y_1],\dots,\text{Ind}[f(X_n) \neq Y_n]) : f \in \mathcal{F}\}$$

is invariant of the choice of $Y_1, \ldots, Y_n$! (Since after changing a single Y_i, it either remains unchanged or flips the bit in a column) Therefore we can choose $Y_1 = Y_2 = \cdots = Y_n = 0$. Observe that $\text{Ind}[t \neq 0] = t$ for $t \in \{0, 1\}$. Since f is $\{0, 1\}$-valued, then $(\text{Ind}[f(X_1) \neq 0], \ldots, \text{Ind}[f(X_n) \neq 0]) = (f(X_i), \ldots, f(X_n))$, indicating that $S_{\mathcal{F}}(n) = S_{\mathcal{A}_{\mathcal{F}}}(n)$. It follows that $\text{VC}(\mathcal{A}_{\mathcal{F}}) = \text{VC}(\mathcal{F})$.
Step 2: By Sauer–Shelah lemma from previous lecture, with probability at least $1 - \delta$, we have

$$\begin{aligned} R(f) - R_n(f) &= \Pr(f(X) \neq Y) - \frac{1}{n}\sum_{i=1}^{n} \text{Ind}\{f(X_i) \neq Y_i\} \\ &\leq C\left(\sqrt{\frac{\log\left(S_{\mathcal{A}_{\mathcal{F}}}(n)\right)}{n}} + \sqrt{\frac{\log(2/\delta)}{n}}\right) \\ &\leq C\left(\sqrt{\frac{d\log(en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{n}}\right). \quad (C \leq 10) \end{aligned}$$

□

16.2 Sub-Gaussian Process

Definition 16.2. *(Sub-Gaussian process) The following two definitions are equivalent: (a) We say $(X_t)_{t\in T}$ is a **sub-Gaussian process** wrt. the metric $d(t, s)$ on T, if it is 0-mean, and for all $t, s \in T$,*

$$\mathbb{E}\exp(\lambda(X_t - X_s)) \leq \exp\left(\frac{\lambda^2 d^2(t, s)}{2}\right);$$

*(b) Say $(X_t)_{t\in T}$ is a **sub-Gaussian process** wrt. the metric $d(t, s)$ on T, if it is 0-mean, and $\exists C > 0$ (absolute constant), such that $\forall t, s \in T$,*

$$\|X_t - X_s\|_{\psi_2} \leq C d(t, s).$$

16.2.1 Examples

Example 1 Suppose $T \subseteq \mathbb{R}^d$, $g \sim \mathcal{N}(0, I_d)$, $X_t = \langle g, t\rangle$, the distance $d(t, s) = \|t - s\|_2$. Then

$$\mathbb{E}\exp\left(\lambda\langle g, t - s\rangle\right) \leq \exp\left(\frac{\lambda^2\|t - s\|_2^2}{2}\right).$$

Example 2 (Rademacher Process)
Suppose $T \subseteq \mathbb{R}^d$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, where ε_i's are i.i.d Rademacher random variables, $X_t = \langle \varepsilon, t\rangle$, the distance $d(t, s) = \|t - s\|_2$. Then

$$\mathbb{E}\exp\left(\lambda(X_t - X_s)\right) \leq \exp\left(\frac{\lambda^2\|t - s\|_2^2}{2}\right).$$

16.2.2 Definitions

Definition 16.3. *We introduce some definitions.*

- ***Gaussian width** is $\mathbb{E}\sup_{t\in T}\langle g, t\rangle$;*

- ***The Rademacher average*** *is* $\mathbb{E}\sup_{t\in T}\langle \varepsilon, t\rangle$;
- ***Gaussian complexity*** *is* $\mathbb{E}\sup_{t\in T}|\langle g, t\rangle|$;
- ***Rademacher complexity*** *is* $\mathbb{E}\sup_{t\in T}|\langle \varepsilon, t\rangle|$;
- ***(Empirical) Rademacher complexity*** *of class of functions* $\mathcal{F}: \{f: \mathcal{X}\to\mathbb{R}\}$ *is*

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

- ***Rademacher complexity*** *is*

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_\varepsilon\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right\}.$$

By symmetrization, we have

$$\mathbb{E}_X \sup_{f\in\mathcal{F}}\left|\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right| \le 2\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| = 2\mathcal{R}(\mathcal{F}).$$

16.3 Dudley Integral

16.3.1 Statement

Theorem 16.4. *Assume that* $(X_t)_{t\in T}$ *is a sub-Gaussian process with metric* d, $diam(T) := \sup_{t,s\in T} d(t,s)$, *then for any* $\delta > 0$,

$$\mathbb{E}\sup_{t\in T} X_t \le 2\mathbb{E}\sup_{t,s\in T, d(t,s)\le\delta}(X_t - X_s) + 16\int_{\frac{\delta}{4}}^{\frac{diam(T)}{2}}\sqrt{\log\mathcal{N}(T,d,\varepsilon)}d\varepsilon,$$

where $\int_{\frac{\delta}{4}}^{\frac{diam(T)}{2}}\sqrt{\log\mathcal{N}(T,d,\varepsilon)}d\varepsilon$ *is called* ***Dudley integral****.*

Proof. This is left for next lecture. □

Remark 16.5. *It can also apply to absolute value.*

16.3.2 Application

Consider $\mathbb{E}\sup_{t\in T}\langle g, t\rangle$, where $T = B_2^d$, $d(t,s) = \|t-s\|_2$. Then $\mathrm{diam}(T) = 2$, and $\varepsilon \le \mathrm{diam}(T)/2 = 1$. We can let $\delta = 0$. Recall that $\mathcal{N}(B_2^d, d, \varepsilon) \le (1+2/\varepsilon)^d \le (3/\varepsilon)^d$, then

$$\begin{aligned}\mathbb{E}\sup_{t\in B_2^d}\langle g, t\rangle &\le 16\int_0^1\sqrt{d\log\left(\frac{3}{\varepsilon}\right)}d\varepsilon \\ &= 16\sqrt{d}\int_0^1\sqrt{\log\left(\frac{3}{\varepsilon}\right)}d\varepsilon \le C\sqrt{d},\end{aligned}$$

Since $\int_0^1\sqrt{\log(3/\varepsilon)}d\varepsilon$ is a finite absolute constant.

Comparison: Recall $\mathbb{E}\sup_{t\in B_2^d}\langle g, t\rangle = \mathbb{E}\|g\| \le \sqrt{d}$.

From theorem 16.4, we can bound $\mathbb{E}\sup_{t\in T} X_t$ using covering number. We know that for different norms, the covering number may not be the same. Next, we introduce two useful norms.

Definition 16.6. *Given the i.i.d sample $X_1, \dots, X_n \sim \mathbb{P}$, define $L_2(\mathbb{P})$-norm is $\|f\|^2_{L_2(\mathbb{P})} = \mathbb{E}_{X\sim\mathbb{P}} f^2(X)$, and $L_2(\mathbb{P}_n)$-norm is $\|f\|^2_{L_2(\mathbb{P}_n)} = \frac{1}{n}\sum_{i=1}^n f^2(X_i)$, then the **covering number** wrt. $L_2(\mathbb{P}_n)$ is*

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_n)}, \varepsilon).$$

Corollary 16.7. *Given $x_1, \dots x_n$, the $L_2(\mathbb{P}_n)$-distance is defined by*

$$\|f-g\|^2_{L_2(\mathbb{P}_n)} = \frac{1}{n}\sum_{i=1}^n (f(x_i) - g(x_i))^2.$$

Define the zero-mean random variable $Z_f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(x_i)$, and the sub-Gaussian process $\{Z_f\}_{f\in\mathcal{F}}$. Substituting the $L_2(\mathbb{P}_n)$-distance into the covering number in Dudley integral, we can obtain following upper bound

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right) \le 2\cdot\frac{1}{n}\delta\cdot n + \frac{16}{\sqrt{n}}\int_{\frac{\delta}{4}}^{\frac{diam(\mathcal{F})}{2}} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(\mathbb{P}_n), \varepsilon)}d\varepsilon.$$

Proof. For conciseness, we denote condition $M = \{f, g \in \mathcal{F}\}$, and condition $N = \{\|f-g\|_{L_2(\mathbb{P}_n)} \le \delta\}$. By theorem 16.4, we only need to show that $\mathbb{E}\sup_{M,N}\sum_{i=1}^n \varepsilon_i(f(X_i)-g(X_i)) \le n\delta$. In fact,

$$\begin{aligned}
&\mathbb{E}\sup_{M,N}\sum_{i=1}^n \varepsilon_i(f(X_i)-g(X_i)) \\
\text{(Cauchy-Schwarz)}\quad &\le \mathbb{E}\sup_{M,N}\|\varepsilon\|\sqrt{\sum_{i=1}^n (f(X_i)-g(X_i))^2} \\
&= \sqrt{n}\,\mathbb{E}\sup_{M,N}\|\varepsilon\|\sqrt{\frac{1}{n}\sum_{i=1}^n (f(X_i)-g(X_i))^2} \\
(\|f-g\|_{L_2(\mathbb{P}_n)} \le \delta)\quad &\le \delta\sqrt{n}\,\mathbb{E}\|\varepsilon\| \\
\text{(Cauchy-Schwarz)}\quad &\le \delta\sqrt{n}\left[\mathbb{E}\|\varepsilon\|^2\right]^{\frac{1}{2}} \\
&= \delta\cdot n.
\end{aligned}$$

□

Definition 16.8. *We say that $\mathcal{F}$ is a **parametric class of functions** if*

$$\sup_{\mathbb{P}_n}\mathcal{N}(\mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_n)}, \varepsilon) \le \left(\frac{C}{\varepsilon}\right)^p,$$

where p plays the role of dimension.

For parametric classes $\mathcal{F}$ such that $\|f\|_\infty \le 1$, applying theorem 16.4, we can derive

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right) &\le \frac{16}{\sqrt{n}}\int_0^1 \sqrt{\log\mathcal{N}(\mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_n)}, \varepsilon)}d\varepsilon \\
&= \frac{16}{\sqrt{n}}\int_0^1 \sqrt{p\log\left(\frac{C}{\varepsilon}\right)}d\varepsilon \\
&= \frac{C_1\sqrt{p}}{\sqrt{n}},
\end{aligned}$$

where C, C_1 are absolute constants.

Proposition 16.9. *(Dudley) If $\mathcal{F}$ is a class of $\{0,1\}$-valued functions with VC-dimension d, then*

$$\sup_{\mathbb{P}_n} \mathcal{N}(\mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_n)}, \varepsilon) \leq \left(\frac{C}{\varepsilon}\right)^{4d}.$$

Proof. This is left for next lecture. □

Lecture 17: Proof of Dudley's integral

Instructor: Nikita Zhivotovskiy *Scriber: Zora Tung* *Proofreader: Weijie Zhao*

1 Dudley's Lemma

Lemma 1. *Assume that $\mathcal{F}$ is a class of binary-valued functions/classifiers with the VC dimension d. Then $\exists c > 0$ (an absolute constant) such that for any empirical measure P_n,*

$$\boxed{\sup_{P_n} \mathcal{N}\left(\mathcal{F}, L_2\left(P_n\right), \epsilon\right) \leq \left(\frac{c}{\epsilon}\right)^{4d}}$$

As per Lecture 14 (VC Theory) P_n is understood to be parameterized by $x_1, ...x_n \in \mathcal{X}$ (we are, more precisely, taking the supremum over these x_i), and more exactly defined as $P_n\left(A \subseteq \mathcal{X}\right) = \frac{1}{n}\sum_{i=1}^n 1\left\{x_i \in A\right\}$. For a real-valued function f, we can $P_n f = \frac{1}{n}\sum_{i=1}^n f\left(X_i\right)$.

Proof. We think of a subset $V \subseteq \{0,1\}^n$ where V has VC-dimension d; i.e. there is a subset of d "columns" which are shattered (see Lecture 16), and there is no subset of $d+1$ columns which are shattered.

We will bound the packing numbers instead of covering numbers, which works because the packing number is always greater than the covering number (see Lecture 9),

$$\mathcal{N}\left(K, \rho, \epsilon\right) \leq \mathcal{P}\left(K, \rho, \epsilon\right).$$

Our distance measure will be,

$$\|f - g\|_{L_2(P_n)} = \sqrt{\frac{1}{n}\sum_{i=1}^n \underbrace{\left(g\left(x_i\right) - f\left(x_i\right)\right)^2}_{\text{for binary fcn's, 0 or 1}}}.$$

If we write these functions as vectors in V, such that $u_i = g\left(x_i\right)$, then we get an equivalent distance metric on V,

$$\rho\left(u, v\right) = \sqrt{\frac{1}{n}\sum_{i=1}^n 1\left\{u_i \neq v_i\right\}} \quad \left(u, v \in V\right).$$

Recall V^0 is a packing of V if $\forall u, v \in V^0$ $(u \neq v)$,

$$\rho\left(u, v\right) > \epsilon \iff \sum_{i=1}^n 1\left\{u_i \neq v_j\right\} > n\epsilon^2. \tag{1}$$

Fix a $u, v \in V^0$ and consider a fixed

$$A_{u,v} = \left\{i \in [n]\,; u_i \neq v_i\right\}.$$

Consider $Y_1, ..., Y_k$ iid random variables distributed uniformly on $[n] = \{1, ..., n\}$. Then

$$\Pr\left(Y_1 \notin A_{u,v}\right) \leq 1 - \epsilon^2 \tag{2}$$

since (1) implies that *on average* the ith coordinate is $\geq \epsilon^2$ apart for $u \neq v$. We want to bound,

$$\Pr\left(\forall u, v \in V^0, u \neq v, \exists j \in [k] \text{ such that } Y_j \in A_{u,v}\right), \tag{3}$$

so we can apply the union bound to all pairs u, v (there are $\leq \left|V^0\right|^2$ pairs) and from (2) that $\Pr\left(\forall j \in [k], Y_j \notin A_{u,v}\right) = \left(1-\epsilon^2\right)^k$, this is

$$\begin{aligned}
&\geq 1 - \left|V^0\right|^2 \left(1-\epsilon^2\right)^k \\
&\geq 1 - \left|V^0\right|^2 \exp\left(-\epsilon^2 k\right).
\end{aligned}$$

We can choose

$$k = \left\lceil \frac{2\log\left|V^0\right|}{\epsilon^2} \right\rceil + 1,$$

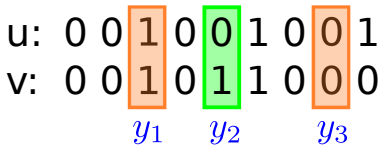
then (3) becomes

$$\geq 1 - \left|V^0\right|^2 \exp\left(\log\left(\left|V^0\right|^{-2}\right) - \underbrace{2\epsilon^2}_{\text{from ceil}}\right) = 1 - \exp\left(-2\epsilon^2\right) \geq 0$$

(replace $2\epsilon^2$ with ϵ^2 when the ceiling doesn't change the expression in k). Then because (3) is greater than 0, there is a realization of the random variables such that the event is true, i.e.

$$\exists y_1, ..., y_k \in [n] \text{ such that } \forall u, v \in V^0, u \neq v, \exists j \in [k] \text{ such that } y_j \in A_{u,v}.$$

As an illustration, we can pick a set of indices where for any u and v, at least one of the indices picks up on one of the the difference between the vectors; y_2 in the figure below,



Therefore, if we project V^0 on only columns corresponding to indices $y_1, ..., y_k$, then the size of the newly-obtained set $V' \subseteq \{0,1\}^k$ is the same as V^0: we cannot project any $u \neq v$ to the same point, because one of the y_j's will distinguish them.

Then the Sauer-Shelah-Vapnik-Chervonenkis Lemma tells us

$$\left|V^0\right| = |V'| \leq \left(\frac{ek}{d}\right)^d,$$

and since $4x \geq \lceil 2x \rceil + 1$ for $x \geq 1$, assuming $\log\left|V^0\right| \geq \epsilon^2$,

$$k = \left\lceil \frac{2\log\left|V^0\right|}{\epsilon^2} \right\rceil + 1 \leq \frac{4\log\left(\left|V^0\right|\right)}{\epsilon^2}$$

then we can plug this in to get,

$$\begin{aligned}
\left|V^0\right| = \left|V^1\right| &\leq \left(\frac{4e\log\left(\left|V^0\right|\right)}{d\epsilon^2}\right)^d \\
\log\left(\left|V^0\right|\right) &\leq d\log\left(\frac{4e}{\epsilon^2}\right) + \cancel{d}\log\left(\frac{\log\left(\left|V^0\right|\right)}{\cancel{d}}\right),
\end{aligned}$$

use the fact $\log(x) \leq \frac{x}{e}$, and get,

$$\begin{aligned}
\log\left(\left|V^0\right|\right) &\leq d\log\left(\frac{4e}{\epsilon^2}\right) + \frac{\log\left(\left|V^0\right|\right)}{e} \\
\log\left(\left|V^0\right|\right) &\leq \left(1-\frac{1}{e}\right)^{-1} d\log\left(\frac{4e}{\epsilon^2}\right).
\end{aligned}$$

Now taking the exponential again

$$\begin{aligned} |V^0| &\leq \left(\frac{4e}{\epsilon^2}\right)^{2d} \\ &= \left(\frac{\sqrt{4e}}{\epsilon}\right)^{4d}. \end{aligned}$$

□

2 Applications

If $\mathcal{F}$ is a class of $\{0,1\}$-valued functions with VC-dimension d, then

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right)\right] \leq C\sqrt{\frac{d}{n}}. \tag{4}$$

We showed this last time, deriving the bound

$$\frac{16}{\sqrt{n}}\int_0^1 \sqrt{ud\log\left(\frac{c'}{\epsilon}\right)}d\epsilon,$$

which is $\leq C\sqrt{\frac{d}{n}}$.

Corollary (Dvoretzky–Kiefer–Wolfowitz inequality) Let $F(x)$ be the true CDF, $F_n(x)$ be an empirical CDF. Then we are interested in achieving a tail bound for,

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)|,$$

where F_n is the random variable. We just apply

- symmetrization
- bounded differences
- the bound for Rademacher complexity of a $\{0,1\}$-valued function with VC-dimension d (4)
- the fact that the VC dimension of the class of intervals $\{(-\infty,t)\,;t\in\mathbb{R}\}$ used to define the CDF is 1

to get with probability $1-\delta$,

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \underbrace{C}_{\text{absolute const}}\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}\right).$$

More specifically, this is

$$\leq C'\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}.$$

Remark 2. *This is not the sharpest possible constant; Massart's version of DKW tells us that*

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}$$

but this is optimal (we cannot make it sharper for all CDFs).

3 Dudley integral proof

Let's return to the Dudley integral bound proof.

We want to show: If X_t is a sub-Gaussian process indexed by $\mathcal{T}$ (with canonical distance d), then

$$\boxed{\mathbb{E}\left[\sup_{t\in\mathcal{T}} X_t\right] \leq 2\mathbb{E}\left[\sup_{\substack{t,s\in\mathcal{T}\\ d(t,s)\leq\delta}} (X_t - X_s)\right] + 16\int_{\delta/4}^{\text{Diam}/2} \sqrt{\log\left(\mathcal{N}\left(\mathcal{T}, d, \epsilon\right)\right)}\mathrm{d}\epsilon}$$

where $\text{Diam} = \sup_{t,s\in\mathcal{T}} d\left(t,s\right)$.

We generally analyze the first term case-by-case (see previous lecture for Rademacher process, for example), and we can always take $\delta = 0$ to just have the integral.

3.1 Part 1: Finite net bounds

In this part, we work with a "single-shot" approximation of $\mathcal{T}$. Assume that $T^0 \subseteq \mathcal{T}$, where T^0 is finite. Then we can say

$$\mathbb{E}\left[\sup_{t,s\in T^0} (X_t - X_s)\right] \leq 2 \max_{t,s\in T^0} d\left(t,s\right)\sqrt{\log\left(|T^0|\right)}.$$

We know that since X_t is a sub-Gaussian process, then by definition / equivalent properties,

$$\mathbb{E}\left[\exp\left(\lambda\left(X_t - X_s\right)\right)\right] \leq \exp\left(\frac{\lambda^2\left(d\left(t,s\right)\right)^2}{2}\right),$$

and $\max_{t,s\in T^0} d\left(t,s\right)$ is the upper bound on the sub-Gaussian process. Then

$$\mathbb{E}\left[\sup_{t,s\in T^0} (X_t - X_s)\right] \leq \max_{t,s\in T^0} d\left(t,s\right)\sqrt{2\log\left(\left|T^0\right|^2\right)}.$$

because if $Y_1, ..., Y_n$ are SG with parameter $\sigma_1, ..., \sigma_n$, then

$$\mathbb{E}\left[\max_i Y_i\right] \leq \max \sigma_i \sqrt{2\log\left(n\right)},$$

where we are using T^0 being finite. Here we can think of i indexing pairs in $T^0 \times T^0$, so $Y_{i\leftrightarrow s,t} = X_s - X_t$ which is mean-zero and sub-Gaussian.

3.2 Step 2

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\mathcal{T}} X_t\right] &= \mathbb{E}\left[\sup_{t\in\mathcal{T}} \left(X_t - X_s\right)\right] \text{ (for any } s \text{ since } \mathbb{E}\left[X_s\right] = 0) \\
&\leq \mathbb{E}\left[\sup_{t,s\in\mathcal{T}} \left(X_t - X_s\right)\right].
\end{aligned}$$

Let N be a covering of $\mathcal{T}$ at scale δ. Given t, and let $\hat{t}$ be the closest $\hat{t} \in N$ to t, the same for $\hat{s}$ (just choosing the closest points in the net). Then this is

$$\begin{aligned}
&\leq \mathbb{E}\left[\sup_{t,s\in\mathcal{T}} (X_t - X_s + X_{\hat{t}} - X_{\hat{t}} + X_{\hat{s}} - X_{\hat{s}})\right] \\
&\underbrace{=}_{\text{reorder}} \mathbb{E}\left[\sup_{t,s\in\mathcal{T}} (X_t - X_{\hat{t}} + X_{\hat{s}} - X_s + X_{\hat{t}} - X_{\hat{s}})\right] \\
&\underbrace{\leq}_{\text{property of sup}} \mathbb{E}\left[\sup_{t\in\mathcal{T}} (X_t - X_{\hat{t}}) + \sup_{s\in\mathcal{T}} (X_{\hat{s}} - X_s) + \sup_{\hat{t},\hat{s}\in N} (X_{\hat{t}} - X_{\hat{s}})\right] \\
&= \mathbb{E}\left[\sup_{\hat{t},\hat{s}\in N} (X_{\hat{t}} - X_{\hat{s}})\right] + 2\mathbb{E}\left[\sup_{t\in\mathcal{T}} \left(X_t - \underbrace{X_{\hat{t}}}_{\text{closest to } X_t}\right)\right] \\
&\leq \mathbb{E}\left[\sup_{\hat{t},\hat{s}\in N} (X_{\hat{t}} - X_{\hat{s}})\right] + 2\mathbb{E}\left[\sup_{\substack{t,s\in\mathcal{T} \\ d(t,s)\leq\delta}} (X_t - X_s)\right].
\end{aligned}$$

we've basically used the net to make it so we are taking the supremum over a finite set.

3.3 Step 3

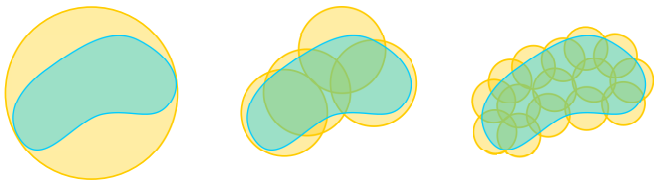
We now want to use a trick of "chaining", where we look at the sequence where we "zoom in", and improve the granularity. We are analyzing the first term from Step 2,

$$\mathbb{E}\left[\sup_{\hat{t},\hat{s}\in N} (X_{\hat{t}} - X_{\hat{s}})\right].$$

For $j = 0, 1, ...$, consider the cover/covering number

$$N_j \subseteq \mathcal{T} \text{ at scale / using balls of radius } 2^{-j} \cdot \text{Diam}(\mathcal{T}).$$

Visually,



$j = 0$ $j = 1$ $j = 2$

Let m be the first integer such that

$$2^{-m} \cdot \text{Diam}(\mathcal{T}) \leq \delta,$$

so we only have to bound

$$\mathbb{E}\left[\sup_{t_m,s_m\in N_m} (X_{t_m} - X_{s_m})\right] \tag{5}$$

Chaining We can write

$$X_{t_m} = \sum_{i=1}^{m} \left(X_{t_i} - X_{\pi_{i-1}(t_i)}\right) + X_{t_0},$$

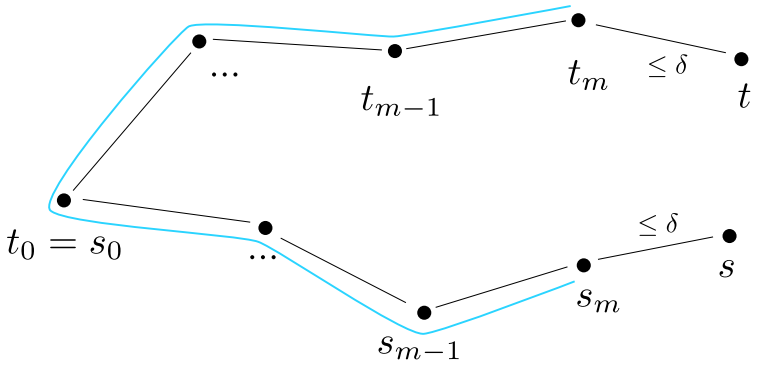
where we recursively define

$$t_{i-1} := \pi_{i-1}(t_i) \text{ is the closest element in} N_{i-1} \text{ to } t_i \in N_i.$$

Relating this back to our supremum (5), we write

$$X_{t_m} - X_{s_m} = \sum_{i=1}^{m} \left(X_{t_i} - X_{\pi_{i-1}(t_i)}\right) + \cancel{X_{t_0}} - \sum_{i=1}^{m} \left(X_{s_i} - X_{\pi_{i-1}(s_i)}\right) - \cancel{X_{s_0}}.$$

Pictorially, we follow the path to the coarsest cover, which is a single point $t_0 = s_0$,



. So (5) is

$$\begin{aligned}
&2\mathbb{E}\left[\sup_{t_m \in N_m} \sum_{i=1}^{m} \left(X_{t_i} - X_{\pi_{i-1}(t_i)}\right)\right] \\
\leq &2\mathbb{E}\left[\sum_{i=1}^{m} \sup_{t_i \in N_i} \left(X_{t_i} - X_{\pi_{i-1}(t_i)}\right)\right] \\
\leq &2\sum_{i=1}^{m} \text{Diam} \cdot \left(2^{-(i-1)}\right) \sqrt{2\log\left(\mathcal{N}\left(\mathcal{T}, d, 2^{-i} \cdot \text{Diam}\right)\right)} \\
\leq &16 \int_{\delta/4}^{\text{Diam}/2} \sqrt{\log\left(\mathcal{T}, d, \epsilon\right)} d\epsilon.
\end{aligned}$$

where the third step uses the expectation bound from the Step 1 (for a single shot). The last step comes from using the integral as a smooth upper bound of the summation; it is not hard to derive.

Remark 3. *This is a very useful bound for empirical processes; you can see papers on Arxiv using it every day.*

4 Remarks about covering numbers

There is a L_p norm

$$L_p(P) \text{ defined as } \|X\|_{L_p} = \mathbb{E}_{X \sim P}\left[|X|^p\right]^{1/p},$$

where for vectors

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}.$$

The "order" of p for these L_p norms is reversed from the geometric sense. For $1 \leq p \leq q \leq \infty$,

- $\mathcal{N}\left(\mathcal{F}, L_p(P), \epsilon\right) \leq \mathcal{N}\left(\mathcal{F}, L_q(P), \epsilon\right)$
- $\mathcal{N}\left(\mathcal{T}, \|\cdot\|_p, \epsilon\right) \geq \mathcal{N}\left(\mathcal{T}, \|\cdot\|_q, \epsilon\right)$ for $1 \leq p \leq q \leq \infty$. i.e. it is harder to cover a space with ℓ^1 balls than ℓ^∞ balls.

Lecture 18: Nonparametric classes, Contraction, Bracketing

Instructor: Nikita Zhivotovskiy *Scriber: Vilas Winstein* *Proofreader: Zekai Wang*

18.1 Nonparametric classes

We start with a motivating example.

18.1.1 Example: Lipschitz functions

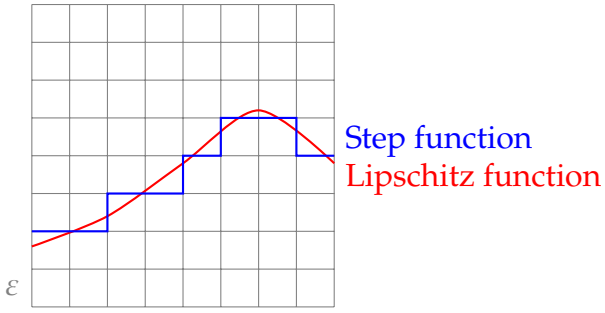
Let $\mathcal{F}$ denote the class of 1-Lipschitz functions $f : [0,1] \to [0,1]$. We are interested in

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

for some i.i.d. random variables $X_i \in [0,1]$. We wish to apply the chaining argument (Dudley's integral) to bound $(*)$, but for that we need to get a bound on the covering numbers. First notice that since the L_∞ norm of a random variable is bigger than the L_2 norm, the L_∞ balls are smaller and so the L_∞ covering numbers are bigger. In other words, we have

$$\mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \le \mathcal{N}(\mathcal{F}, L_\infty(P_n), \varepsilon).$$

In order to get a bound on the L_∞ covering numbers, we need to be able to find a net of functions which approximate 1-Lipschitz functions uniformly on the interval. For this, draw an ε-spaced grid on the unit square $[0,1]^2$ as in the figure below.



Since a 1-Lipschitz function (shown in red) cannot cross two different horizontal grid lines in between two of the vertical grid lines, there is always a horizontal segment of the grid which stays within ε of the 1-Liptschitz function between any two vertical grid lines. Taking the step function consisting of these horizontal segments (shown in blue), we see that the L_∞ distance between the step function and the Liptschitz function is at most ε.
So, the set of all of these step functions is an ε-net of $\mathcal{F}$. Note that this is an example of a net which consists of functions that are *not* in the set to be covered, since the step functions are not 1-Lipschitz. It remains to count the number of possible step functions. A generous overcounting, noticing that there are at most $\frac{2}{\varepsilon}$ horizontal intervals and at most $\frac{2}{\varepsilon}$ segments to choose from yields

$$\mathcal{N}(\mathcal{F}, L_\infty(P_n), \varepsilon) \le \left(\frac{2}{\varepsilon}\right)^{\left(\frac{2}{\varepsilon}\right)}.$$

In fact, since the next segment is constrained to be not too far away from the previous one, we could probably get away with an upper bound of the formalism $C^{1/\varepsilon}$, but on a logarithmic scale this only differs by a logarithmic factor from the previous bound.
Now we apply the Dudley integral to obtain

$$(*) \leq \frac{16}{\sqrt{n}} \int_0^1 \sqrt{\frac{2}{\varepsilon} \log \frac{2}{\varepsilon}}\, d\varepsilon.$$

Since $\int_0^1 \frac{1}{\varepsilon^q}\, d\varepsilon$ converges when $q < 1$, the integral above is a finite constant, and we obtain $(*) \leq \frac{C}{\sqrt{n}}$.
Note that this is the *same* rate as for a single function in $\mathcal{F}$; by Höfding's inequality, for any fixed $f \in \mathcal{F}$, we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right] \leq \frac{C}{\sqrt{n}}.$$

But we have gotten the same rate *uniformly* in $\mathcal{F}$.

18.1.2 General nonparametric classes

Recall that for parametric classes $\mathcal{F}$ we had

$$\sup_{P_n} \log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \leq p \log\left(\frac{C}{\varepsilon}\right)$$

for some p, and thus that we got the order $\frac{1}{\sqrt{n}}$ bound uniformly in such classes.
The covering number bound for parametric classis is much better than the bound we obtained for the 1-Lipschitz example above, which is polynomial in $\frac{1}{\varepsilon}$. Nonetheless, we were able to obtain the same order of uniform bound on $(*)$. We would like to be able to understand when we can get this kind of bound for somewhat worse classes than parametric classes, like the 1-Lipschitz functions, in a more systematic way. This motivates the following definition:

Definition 18.1. *A* nonparametric *class $\mathcal{F}$ is one for which*

$$\sup_{P_n} \log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \lesssim \left(\frac{C}{\varepsilon}\right)^p$$

for some p.

Let's calculate the Dudley integral for nonparametric classes. First, as long as $p < 2$, we have

$$\begin{aligned}
\frac{16}{\sqrt{n}} \int_0^1 \sqrt{\left(\frac{C}{\varepsilon}\right)^p}\, d\varepsilon &= \frac{16C^{p/2}}{\sqrt{n}} \int_0^1 \varepsilon^{-p/2}\, d\varepsilon \\
&= \frac{16C^{p/2}}{\sqrt{n}} \left[\frac{\varepsilon^{1-p/2}}{1-p/2}\right]_{\varepsilon=0}^1 \\
&= \frac{16}{\sqrt{n}} \frac{C^{p/2}}{1-p/2} \\
&= O_{\varepsilon,p}\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

In other words, as long as $p < 2$, we get the same bound uniformly as for a single function. For $p \geq 2$ the above integral does not converge, and so we need to use the full form of the Dudley integral bound. Let's also assume for simplicity that $\|f\|_\infty \leq 1$ for every $f \in \mathcal{F}$. The bound is then

$$\begin{aligned} 2\delta + \frac{16C^{p/2}}{\sqrt{n}} \int_{\delta/4}^{1} \varepsilon^{-p/2}\, d\varepsilon &= 2\delta + \frac{16C^{p/2}}{\sqrt{n}} \left[\frac{\varepsilon^{1-p/2}}{1-p/2}\right]_{\varepsilon=\delta/4}^{1} \\ &= 2\delta + \frac{C}{\sqrt{n}}\delta^{1-p/2}, \end{aligned}$$

possibly changing C and using the fact that $\delta < 1$. Taking $\delta = n^{-\frac{1}{p}}$ we then obtain

$$2n^{-\frac{1}{p}} + Cn^{-\frac{1}{2}-\frac{1}{p}+\frac{p}{2}\frac{1}{p}} = Cn^{-\frac{1}{p}}$$

again changing C. In conclusion, for nonparametric classes, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right)\right] \leq C \cdot \begin{cases} \frac{1}{\sqrt{n}} & \text{when } p < 2, \\ \frac{1}{\sqrt[p]{n}} & \text{when } p > 2. \end{cases}$$

Finally, we did not prove it in class, but for $p = 2$ there is an extra logarithmic factor.

18.2 Contraction for Rademacher averages/processes

We start by stating a theorem which states that we can "erase" Lipschitz functions when taking Rademacher averages to get an upper bound.

18.2.1 Theorem statement and proof

Theorem 18.2 (Ledoux-Talagrand)**.** *Let $\varphi_1, \ldots, \varphi_n$ be L-Liptschitz functions $\mathbb{R} \to \mathbb{R}$ such that $\varphi_i(0) = 0$. Then for any $T \subseteq \mathbb{R}^n$, we have*

$$\mathbb{E}\left[\sup_{t\in T}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varphi_i(t_i)\right|\right] \leq 2L\mathbb{E}\left[\sup_{t\in T}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i t_i\right|\right].$$

For this theorem, unlike many of the results previously discussed in the course, the absolute values make a big difference, in terms of the difficulty of the proof. We will only prove a much easier version without the absolute values (and without the assumption that $\varphi_i(0) = 0$).

Theorem 18.3. *Let $\varphi_1, \ldots, \varphi_n$ be L-Liptschitz functions $\mathbb{R} \to \mathbb{R}$. Then for any $T \subseteq \mathbb{R}^n$, we have*

$$\mathbb{E}\left[\sup_{t\in T}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varphi_i(t_i)\right] \leq L\mathbb{E}\left[\sup_{t\in T}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i t_i\right].$$

Proof. Without loss of generality, assume that $L = 1$. Then, expanding the expectation over ε_n only,

$$\begin{aligned} \mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^{n}\varepsilon_i\varphi_i(t_i)\right] &= \frac{1}{2}\mathbb{E}\left[\sup_{t\in T}\left(\sum_{i=1}^{n-1}\varepsilon_i\varphi_i(t_i) + \varphi_n(t_n)\right) + \sup_{s\in T}\left(\sum_{i=1}^{n-1}\varepsilon_i\varphi_i(s_i) - \varphi_n(s_n)\right)\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[\sup_{t,s\in T}\left(\sum_{i=1}^{n-1}\varepsilon_i(\varphi_i(t_i) + \varphi_i(s_i)) + |t_n - s_n|\right)\right] \end{aligned}$$

$$= \frac{1}{2}\mathbb{E}\left[\sup_{t,s\in T}\left(\sum_{i=1}^{n-1}\varepsilon_i(\varphi_i(t_i)+\varphi_i(s_i))+t_n-s_n\right)\right]$$
$$= \mathbb{E}\left[\sup_{t\in T}\left(\sum_{i=1}^{n-1}\varepsilon_i\varphi_i(t_i)+\varepsilon_n t_n\right)\right],$$

where in the second equality we used the symmetry between t and s to remove the absolute value. The result follows by induction. □

18.2.2 Application: excess risk in general

We will apply the contraction theorem to excess risk. First, we define a few different loss functions to which the next computation applies. In general, we may have an i.i.d. sample of (X_i, Y_i), and hypothesize that $Y = f(X)$ describes the data. We can measure the loss of this hypothesis in a variety of ways, using a loss function $\ell(f(X), Y)$ which represents the "price" for predicting $Y = f(X)$. For example, we could take

$$\ell(f(X),Y) = \begin{cases} \mathrm{Ind}(f(X)=Y) & \text{(binary loss)} \\ (Y-f(X))^2 & \text{(squared loss)} \\ |Y-f(X)| & \text{(absolute loss)} \\ \max\{0, 1-Y\cdot f(X)\} & \text{(hinge loss).} \end{cases}$$

Note that hinge loss generalizes binary loss, in the case where $Y \in \{\pm 1\}$. From a loss function we can define the *population risk* R and the *empirical risk* R_n of predicting $Y = f(X)$ as

$$R(f) = \mathbb{E}\left[\ell(f(X),Y)\right],$$
$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i),Y_i).$$

We may prefer to choose the estimator $\hat{f}$ which minimizes the empirical risk. Then the *excess risk* $\hat{\mathcal{E}}$ is

$$\hat{\mathcal{E}} = R(\hat{f}) - \inf_{f\in\mathcal{F}} R(f), \qquad \text{where} \quad \hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} R_n(f).$$

Now we calculate, using the version of the theorem which we have proved, that

$$\mathbb{E}\left[\hat{\mathcal{E}}\right] \le \mathbb{E}\left[\sup_{f\in\mathcal{F}}(R(f)-R_n(f))\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}}(R_n(f)-R(f))\right]$$
$$\le 4\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell(f(X_i),Y_i)\right]$$
$$\le 4L\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right].$$

18.2.3 Application: hinge loss

Suppose that $Y_i \in \{-1, +1\}$, and define $\psi(t) = \max\{0, 1-t\}$. It iis easy to check that $\psi(0) = 0$ and ψ is 1-Lipschitz. So we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\psi(Y_i f(X_i))\right] \le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i Y_i f(X_i)\right]$$

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

where in the last step we used the fact that (ε_i) has the same distribution as $(\varepsilon_i Y_i)$, since the Y_i are independent and ±1-valued.

18.2.4 Application: Rademacher complexity of a linear class

Let B_2^d be the Euclidean ball in $\mathbb{R}^d$, let $b > 0$, and let

$$\mathcal{F} = \{X \mapsto \langle X, w\rangle : w \in b \cdot B_2^d\}.$$

Further assume that $\|X_i\|_2 \le r$ almost surely for all i. Then we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{w\in b\cdot B_2^d} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\langle X_i, w\rangle\right] &= \frac{b}{n}\mathbb{E}\left[\left\|\sum_{i=1}^{n} \varepsilon_i X_i\right\|\right] \\
&\le \frac{b}{n}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{n} \varepsilon_i X_i\right\|^2\right]} \\
&\le \frac{b}{n}\sqrt{nr^2} \\
&= \frac{br}{\sqrt{n}}.
\end{aligned}$$

Note that we could have used the Dudley integral to bound the left-hand side, but this would have resulted in an upper bound which depends on dimension. Here we have used the linear structure present in the family, and avoided the dependence on dimension.

18.3 Bracketing entropy

This section just contains some definitions that we will use in the next lecture.

Definition 18.4. *For a class $\mathcal{F}$ of functions on $\mathcal{X}$, the* bracket *between two functions $u, l : \mathcal{X} \to \mathbb{R}$ is defined to be the set*

$$[u, l] = \{f \in \mathcal{F} : l(x) \le f(x) \le u(x) \text{ for all } x \in \mathcal{X}\}.$$

Definition 18.5. *For a fixed distribution P on $\mathcal{X}$, and a fixed number q, a bracket $[u, l]$ is an* ε-bracket *with respect to $L_q(P)$ if $\|u - l\|_{L_q(P)} \le \varepsilon$.*

Definition 18.6. *The* ε-bracketing number *of $\mathcal{F}$ with respect to $L_q(P)$ is*

$$\mathcal{N}_{[\,]}(\mathcal{F}, L_q(P), \varepsilon) = \text{minimal number of } \varepsilon\text{-brackets covering } \mathcal{F}.$$

Exercise 18.7. *Show that $\mathcal{N}(\mathcal{F}, L_q(P), \varepsilon) \le \mathcal{N}_{[]}(\mathcal{F}, L_q(P), \varepsilon/2)$.*

In light of the above exercise, it may seem pointless to introduce this notion. Indeed, we will see that many results about $\mathcal{N}$ can be extended to $\mathcal{N}_{[\,]}$. However, importantly, $\mathcal{N}_{[\,]}$ depends on P, whereas $\mathcal{N}$ in principle does not (since we typically take the supremum over P_n).

Lecture 19: Bracketing, Sub-Gaussian Mean Estimators

Instructor: Nikita Zhivotovskiy *Scriber: Xiyuan Zhang* *Proofreader: Xiyuan Zhang*

1 Bracketing

1.1 Recap from last lecture

The ***bracket*** $[u, l]$ is formed by $f \in \mathcal{F} : l(x) \leq f(x) \leq u(x)$ for all $x \in \mathcal{X}$.
$[u, l]$ is an ε***-bracket*** with respect to $L_q(P)$ if $[u, l]$ is a bracket and $\|u - l\|_{L_q(P)} \leq \varepsilon$.
For a fixed distribution P on $\mathcal{X}$, $\mathcal{N}_{[\,]}(\mathcal{F}, L_q(P), \varepsilon)$ is a ***bracketing entropy***.
The power of bracketing is that we do not work with the *sup* with respect to empirical measures.

1.2 Theorem

Theorem 1. *Assume that $\mathcal{F}$ is a class of functions such that $\|f\|_{L_\infty(P)\leq m}$. Then if $X_1 \cdots X_n$ is an i.i.d. sample of copies of the random variable X which distributes to P, there exists an absolute constant C such that*

$$\mathbb{E} \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)| \leq \frac{C}{\sqrt{n}} \int_0^m \sqrt{\log N_{[]}(\mathcal{F}, L_2(P), \varepsilon)} d\varepsilon.$$

To sum up, if we want to bound $\mathbb{E}\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)|$, one method is to bound it as above. Another method is to bound it through standard symmetrization and chaining. Then we can get $\sup_{P_n} N(\mathcal{F}, L_2(P_n), \varepsilon) \leq \sup_P N(\mathcal{F}, L_2(P), \varepsilon)$, where P denotes all measures.

Remark 2 (Koltchinskii-Pollard entropy)**.** $\sup_{P_n} N(\mathcal{F}, L_2(P_n), \varepsilon) \leq \sup_P N(\mathcal{F}, L_2(P), \varepsilon)$*, where P denotes all measures.*

2 Sub-Gaussian mean estimators

Let's start with an example. Let $X_1, \cdots, X_n$ be i.i.d. Sub-Gaussian random variables with parameter σ and mean μ. For non-asymptotic regime,

$$|\frac{1}{n}\sum_{i=1}^{n} -\mu| \leq \sigma\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}.$$

Compare this with CLT with $\mathrm{Var}(X_i) = 1$,

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu) \xrightarrow{d} \mathcal{N}(0, 1).$$

The difference is that the non-asymptotic method requires all moments while CLT only requires 2 moments. There is a fix in the non-asymptotic regime.

2.1 Median-of-means estimator (Mean estimator in $\mathbb{R}$)

Let $X_1, \cdots, X_n$ be i.i.d. random variables with mean μ and $\mathrm{Var}(X_i) = \sigma^2$. Split n points in to K non-intersecting blocks $B_1, \cdots, B_K$ where $|B_j| = \frac{n}{K} = m$, and $\bar{X_j} = \frac{1}{m}\sum_{i \in B_j} X_i$. Then the median of means estimator is $\hat{X_j} = Med(\bar{X_1}, \bar{X_2}, \cdots, \bar{X_K})$, where the median is defined as $Med(y_1, \cdots, y_n) = y_i$ such that

$$|j \in [K] : y_j \leq y_i| \geq \frac{K}{2} \text{ and } |j \in [K] : y_j \geq y_i| \geq \frac{K}{2}.$$

Theorem 3. *Fix the number of blocks $K = 8\log(\frac{1}{\delta})$. Assume without loss of generality that K is an integer and $\frac{n}{K}$ is an integer. $X_1, \cdots, X_n$ are i.i.d. copies of the random variable X with mean μ and $Var(X_i) = \sigma^2$. Then with probability $1 - \delta$,*

$$|\hat{\mu} - \mu| \leq \sigma\sqrt{\frac{32\log(\frac{1}{\delta})}{n}}$$

Proof. First fix the j-th block B_j, where $|B_j| = m$ and $\bar{X_j} = \frac{1}{|B_j|}\sum_{i \in B_j} X_i$. By Chebyshev's inequality,

$$\Pr(|\bar{X_j} - \mu| \geq t) \leq \frac{\mathrm{Var}(X_i)}{t^2 m} = \frac{\sigma^2}{mt^2}.$$

Choose $t = \frac{2\sigma}{\sqrt{m}}$. Then the good event happens with probability at least $\frac{3}{4}$,

$$|\bar{X_j} - \mu| \leq \frac{2\sigma}{\sqrt{m}}.$$

Since the good event for a block has the probability $\geq \frac{3}{4}$, and these events are independent, the probability that there are more than $\frac{K}{2}$ blocks corresponding to a good event can be interpreted as

$$\Pr(\mathrm{Binom}(\frac{3}{4}, K) \geq \frac{K}{2}) \Leftrightarrow \Pr(\mathrm{Binom}(\frac{1}{4}, K) - \frac{1}{4}K < \frac{K}{4}) \geq 1 - \exp(-\frac{2 \times \frac{K^2}{16}}{K}) = 1 - \delta$$

with $K = 8\log(\frac{1}{\delta})$.

Thus, with probability at least $1 - \delta$, there are more than $K/2$ blocks satisfying $|\bar{X_j} - \mu| \leq \frac{2\sigma}{\sqrt{m}}$, and the median should pick one of those good $\bar{X_j} = \hat{\mu}$. Therefore, with probability $1 - \delta$,

$$|\hat{\mu} - \mu| \leq \frac{2\sigma}{\sqrt{m}} = \frac{2\sigma}{\sqrt{\frac{n}{8\log(\frac{1}{\delta})}}} = \sigma\sqrt{\frac{32\log(\frac{1}{\delta})}{n}}.$$

□

Remark 4. *The form of the bound as if $\hat{\mu}$ is a sample mean and the distribution is Sub-Gaussian.*

2.2 Multivariate extensions of Median-of-means estimator

Let $B_1, \cdots B_K$ be the blocks, for a function f,

$$\mathrm{MOM}(f) := Med(\bar{f_1}, \cdots, \bar{f_K}) = Med(\frac{1}{m}\sum_{i \in B_1} f(X_i), \cdots, \frac{1}{m}\sum_{i \in B_K} f(X_i)).$$

Proposition 5 (uniform bound for Median-of-means estimator)**.** *Fix $K = 8\log(\frac{1}{\delta})$, and let ε_i be Rademacher random sign. Then with probability $1 - \delta$,*

$$\sup_{f \in \mathcal{F}}(\mathrm{MOM}(f) - \mathbb{E}f) \leq 32\mathbb{E}\sup_{f \in \mathcal{F}}(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)) + \sqrt{\frac{128\sup_{f \in \mathcal{F}} Var(f(X))\log(\frac{1}{\delta})}{n}}$$

$$\sup_{f\in\mathcal{F}}(\mathbb{E}f - \mathrm{MOM}(f)) \le 32\mathbb{E}\sup_{f\in\mathcal{F}}(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)) + \sqrt{\frac{128\sup_{f\in\mathcal{F}} Var(f(X))\log(\frac{1}{\delta})}{n}}$$

Remark 6. *The nice thing about this uniform bound is that it only requires 2 moments as in CLT.*

Proof. We want to control

$$\sup_{f\in\mathcal{F}}(\mathrm{MOM}(f) - \mathbb{E}f) = \sup_{f\in\mathcal{F}}(\mathbb{E}f - \mathrm{MOM}(f)) = \sup_{f\in\mathcal{F}}(\mathrm{MOM}(\mathbb{E}f - f)) < t.$$

It is enough to show that $\forall f \in \mathcal{F}$, there are more than $1/2$ of blocks satisfying $\mathbb{E}f - \bar{f}_j < t$,

$$\frac{1}{K}\sup_{f\in\mathcal{F}}\sum_{j=1}^{K}\mathrm{Ind}[\mathbb{E}f - \bar{f}_j \ge t] < \frac{1}{2},$$

where $\bar{f}_j = \frac{1}{|B_j|}\sum_{i\in B_j} f(X_i)$.
Let $\varphi(t) = (t-1)\mathrm{Ind}[1 \le t \le 2] + \mathrm{Ind}[t > 2]$, and by construction, we have

$$\varphi(t) \ge \mathrm{Ind}(t \ge 2),$$

and

$$\varphi(t) \le \mathrm{Ind}(t \ge 1).$$

Now we bound the probability of the bad event.

$$\begin{aligned}
\frac{1}{K}\sup_{f\in\mathcal{F}}\sum_{j=1}^{K}\mathrm{Ind}[\mathbb{E}f - \bar{f}_j \ge t] &= \frac{1}{K}\sup_{f\in\mathcal{F}}\sum_{j=1}^{K}\mathrm{Ind}[\frac{2(\mathbb{E}f - \bar{f}_j)}{t} \ge 2] \\
&\le \sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\varphi(\frac{2(\mathbb{E}f - \bar{f}_j)}{t}) \\
&= \sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\varphi(\frac{2(\mathbb{E}f - \bar{f}_j)}{t}) \qquad (*) \\
&+ \sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\varphi(\frac{2(\mathbb{E}f - \bar{f}_j)}{t}) - \sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\varphi(\frac{2(\mathbb{E}f - \bar{f}_j)}{t}) \qquad (**)
\end{aligned}$$

For the first part,

$$\begin{aligned}
(*) &\le \sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\mathrm{Ind}(\frac{2(\mathbb{E}f - \bar{f}_j)}{t} \ge 1) \\
&= \sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\Pr(\mathbb{E}f - \bar{f}_j \ge \frac{t}{2}) \\
&= \sup_{f\in\mathcal{F}}\Pr(\mathbb{E}f - \bar{f}_1 \ge \frac{t}{2}) \\
&\le \sup_{f\in\mathcal{F}}\frac{4\mathrm{Var}(f(X))}{mt^2}. \qquad \text{(by Chebyshev's inequality)}
\end{aligned}$$

For the second part,

$$\sup_{h\in\mathcal{H}}\frac{1}{K}\sum_{j=1}^{K}h(Y_j)-\sup_{h\in\mathcal{H}}\left(\frac{1}{K}\left(\sum_{\substack{j=1\\j\neq i}}^{K}h(Y_j)+h(Y_i')\right)\right)\leq\frac{1}{K}\sum_{j=1}^{K}h^*(Y_j)-\frac{1}{K}\left(\sum_{\substack{j=1\\j\neq i}}^{K}h^*(Y_j)+h^*(Y_i')\right)$$
$$=\frac{1}{K}(h^*(Y_i)-h^*(Y_i'))$$
$$\leq\frac{1}{K}|h^*(Y_i)-h^*(Y_i')|$$
$$\leq\frac{1}{K}.$$

Then by bounded diffenrence inequality,

$$\Pr((**)-\mathbb{E}(**)\geq y)\leq\exp(-\frac{2y^2}{K/K^2})=\exp(-2y^2K).$$

With probability at least $1-\exp(-2Ky^2)$,

$$(**)\leq\mathbb{E}(\sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\varphi(\frac{2(\mathbb{E}f-\bar{f}_j)}{t})-\sup_{f\in\mathcal{F}}\sum_{j=1}^{K}\mathbb{E}\varphi(\frac{2(\mathbb{E}f-\bar{f}_j)}{t}))\qquad(***)$$
$$+y$$

By symmetrization,

$$(***)\leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\varepsilon_j\varphi(\frac{2(\mathbb{E}f-\bar{f}_j)}{t})\qquad(\sup(a-b)\leq\sup(a)+\sup(b))$$
$$\leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\varepsilon_j(\frac{2(\mathbb{E}f-\bar{f}_j)}{t})\qquad(\varphi(t)\text{ is 1-Lipschitz})$$
$$=\frac{4}{t}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\varepsilon_j(\mathbb{E}f-\frac{1}{m}\sum_{i\in B_j}f(X_i))$$
$$=\frac{4}{t}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{K}\sum_{j=1}^{K}\varepsilon_j(\mathbb{E}'\frac{1}{m}\sum_{i\in B_j}f(X_i')-\frac{1}{m}\sum_{i\in B_j}f(X_i))\qquad(X_i'\text{ is an i.i.d. copy of }X_i)$$
$$\leq\frac{4}{tK}\mathbb{E}\mathbb{E}'\sup_{f\in\mathcal{F}}\sum_{j=1}^{K}\varepsilon_j(\frac{1}{m}\sum_{i\in B_j}(f(X_i')-f(X_i)))\qquad(\text{Jensen's Inequality})$$
$$=\frac{4}{tKm}\mathbb{E}\mathbb{E}'\sup_{f\in\mathcal{F}}\sum_{j=1}^{K}\varepsilon_j\sum_{i\in B_j}\varepsilon'(f(X_i')-f(X_i))$$
$$\leq\frac{8}{tn}\mathbb{E}\mathbb{E}'\sup_{f\in\mathcal{F}}\sum_{j=1}^{K}\sum_{i\in B_j}\varepsilon_j\varepsilon_i'f(X_i)$$
$$=\frac{8}{tn}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\varepsilon_if(X_i).\qquad(\varepsilon_j\varepsilon_i'\sim\varepsilon_i)$$

Then with probability at least $1 - \exp(-2Ky^2)$,

$$\sup_{f\in\mathcal{F}} \frac{1}{K}\sum_{j=1}^{K} \mathrm{Ind}[\mathbb{E}f - \bar{f}_j \geq t] \leq \frac{8}{tn}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\varepsilon_i f(X_i) + \frac{4\sup_{f\in\mathcal{F}}\mathrm{Var}(f(X))K}{nt^2} \qquad (****)$$
$$+ y$$

To ensure this bound is less than $1/2$, set $y = \frac{1}{4}$, and let $K = 8\log(\frac{1}{\delta})$, yielding $1 - \exp(-2Ky^2) = 1 - \delta$. When

$$t = 32\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right) + \sqrt{\frac{128\sup_{f\in\mathcal{F}}\mathrm{Var}(f(X))\log\left(\frac{1}{\delta}\right)}{n}},$$

it follows that $(****) < \frac{1}{4}$.

□

Lecture 20: Applications of the Median-Of-Means Estimator

Instructor: Nikita Zhivotovskiy *Scribe: Jianzhi Wang*

In this lecture, we apply the median-of-means (MOM) estimator to two examples: estimating the mean of a random vector and estimating the higher moments of a random variable. In both examples, we only require the first two moments. We conclude by proving a one-sided lower tail bound under the same conditions (having only the first two moments).

1 Estimating the mean of a random vector

1.1 Motivation

Let $X \sim N(\mu, \Sigma)$. From both Gaussian concentration and bound for the concentration of norm, we have $\left\| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right\|_2 \leq \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + 2\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}} \log \frac{1}{\delta}}{n}}$. However, can we get an estimator which has a similar bound given that we only know the first two moments $\mathbb{E}[X] = \mu$ and $\mathbb{E}\left[(X-\mu)(X-\mu)^T\right] = \Sigma$? After all, Central Limit Theorem, which only requires the first two moments, seems to suggest it is possible.

The idea is to construct an estimator using median-of-means. Last time, we proved that given $k \geq 8 \log\left(\frac{1}{\delta}\right)$ where k is the number of blocks, then:

$$\sup_{f \in \mathcal{F}} \{|\mathbb{E}[f] - \mathrm{MOM}(f)|\} \leq 64 \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right\}\right] + 2\sqrt{\frac{128 \sup_{f \in \mathcal{F}} \{\mathrm{Var}[f(x)]\} \log\left(\frac{2}{\delta}\right)}{n}}$$

1.2 Set-Up and Derivations

Consider $\mathcal{F} = \left\{ f_v : v \in B_2^d \right\}$ where $f_v(x) = \langle x, v \rangle$ for $x \in \mathbb{R}^d$. Let $\hat{\mu}$ be an estimator for the mean and μ be the true mean (i.e. $\mu = \mathbb{E}[X]$). Then $\|\hat{\mu} - \mu\|_2 = \sup_{v \in B_2^d} \{|\langle \hat{\mu} - \mu, v \rangle|\}$.

We construct our chosen estimator $\hat{\mu} := \arg\min_{\nu \in \mathbb{R}^d} \left\{ \sup_{v \in B_2^d} \{|\langle \nu, v \rangle - \mathrm{MOM}(\langle X, v \rangle)|\} \right\}$. Intuitively, we are finding the vector ν^* that best approximates the median-of-means estimator, as measured by the worst difference in the projection along any direction $v \in B_2^d$.

The L^2 difference can be bounded as follows:

$$\begin{aligned}
\|\hat{\mu} - \mu\|_2 &= \sup_{v \in B_2^d} |\langle \hat{\mu} - \mu, v \rangle| \\
&\leq \sup_{v \in B_2^d} |\langle \hat{\mu}, v \rangle - \mathrm{MOM}(\langle x, v \rangle)| + \sup_{v \in B_2^d} |\langle \mu, v \rangle - \mathrm{MOM}(\langle x, v \rangle)| \\
&\leq 2 \sup_{v \in B_2^d} |\langle \mu, v \rangle - \mathrm{MOM}(\langle x, v \rangle)| \\
&= 2 \sup_{v \in B_2^d} |\mathrm{MOM}(\langle x - \mu, v \rangle)| \\
&= 2 \sup_{f \in \mathcal{F}} |\mathrm{MOM}(f(x - \mu))|
\end{aligned}$$

The first inequality is due to the triangle inequality and $\sup\{a+b\} \leq \sup a + \sup b$. The second inequality is because $\hat{\mu}$ is the minimiser of the objective. The second-to-last equality is due to the translation equivariant property of $\mathrm{MOM}(\cdot)$ estimator.

Now, note that $\mathbb{E}\left[f_v\left(X-\mu\right)\right] = \mathbb{E}\left[\langle X-\mu, v\rangle\right] = 0 \; \forall v \in B_2^d$. Applying the uniform bound for median-of-means inequality, we get:

$$\begin{aligned}
2\sup_{f\in\mathcal{F}} \left|\mathrm{MOM}\left(f\left(x-\mu\right)\right)\right| &= 2\sup_{f\in\mathcal{F}} \left|\mathrm{MOM}\left(f\left(x-\mu\right)\right) - \mathbb{E}\left[f(x-\mu)\right]\right| \\
&\leq 128\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right\}\right] + 4\sqrt{\frac{128\sup_{f\in\mathcal{F}}\left\{\mathrm{Var}\left[f(x)\right]\right\}\log\left(\frac{2}{\delta}\right)}{n}}
\end{aligned}$$

The first term can be bounded directly by optimisation.

$$\begin{aligned}
128\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right\}\right] &= \frac{128}{n}\mathbb{E}\left[\sup_{v\in B_2^d}\left\{\sum_{i=1}^{n}\epsilon_i\langle x_i, v\rangle\right\}\right] \\
&= \frac{128}{n}\mathbb{E}\left[\sup_{v\in B_2^d}\left\{\langle\sum_{i=1}^{n}\epsilon_i x_i, v\rangle\right\}\right] \\
&= \frac{128}{n}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\epsilon_i\left(x_i-\mu\right)\right\|_2\right] \\
&\leq \frac{128}{n}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{n}\epsilon_i\left(x_i-\mu\right)\right\|_2^2\right]} \\
&= \frac{128}{n}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{n}\left(x_i-\mu\right)\right\|_2^2\right]} \\
&\leq 128\sqrt{\frac{\mathrm{tr}\left(\Sigma\right)}{n}}
\end{aligned}$$

The second term can be bounded by direct optimisation as well:

$$\begin{aligned}
4\sqrt{\frac{128\sup_{f\in\mathcal{F}}\left\{\mathrm{Var}\left[f(x)\right]\right\}\log\left(\frac{2}{\delta}\right)}{n}} &= 4\sqrt{\frac{128\sup_{v\in B_2^d}\left\{\mathrm{Var}\left[\langle x, v\rangle\right]\right\}\log\left(\frac{2}{\delta}\right)}{n}} \\
&= 4\sqrt{\frac{128\sup_{v\in B_2^d}\left\{\mathbb{E}\left[\langle x, v\rangle^2\right]\right\}\log\left(\frac{2}{\delta}\right)}{n}} \\
&= 4\sqrt{\frac{128\left\|\Sigma\right\|_{\mathrm{op}}\log\left(\frac{2}{\delta}\right)}{n}}
\end{aligned}$$

Hence,

$$\left\|\hat{\mu}-\mu\right\|_2 \leq C\left(\sqrt{\frac{\mathrm{tr}\left(\Sigma\right)}{n}} + \sqrt{\frac{\left\|\Sigma\right\|_{\mathrm{op}}\log\left(\frac{2}{\delta}\right)}{n}}\right)$$

Thus, we recover a similar bound for $\hat{\mu}$ compared to the sample mean $\frac{1}{n}\sum_{i=1}^{n}X_i$. The upshot is that this estimator works for a larger class of distributions, including the heavy-tailed Student's t-distribution.

Remark 1. *We are assuming that there is an efficient algorithm to calculate the median-of-means estimator. There is an algorithm that does so in polynomial time.*

2 Estimating higher moments of a random variable

2.1 Definitions

Definition 2. *((p,q) Hypercontractive) Let X be a random variable. Then, X is (p,q)-**hypercontractive** if there exists a "nice" function $L_{p,q}$ such that $q \geq p$ implies $\|X\|_q \leq L_{p,q} \|X\|_p$.*

For example, take $X \sim N(0, \sigma^2)$. Then, for $p \geq 2$, $\|X\|_p \leq C\sqrt{p}\,\|X\|_2$, where C is an absolute constant. Hence, the higher L^p norms are controlled by the lower L^2 norm, implying that X is $(2,p)$-hypercontractive.

We can extend the definition of hypercontractivity to the high dimensional case. If $X \in \mathbb{R}^d$ is a vector, then X is hypercontractive if it is hypercontractive in all directions i.e. $\langle X, v\rangle$ is hypercontractive $\forall v \in S^{d-1}$.

2.2 Set-Up and Derivations

Theorem 3. *Let p be an even integer. Assume that X is a zero-mean random vector in $\mathbb{R}^d$ such that $\forall v \in S^{d-1}$, $\mathbb{E}\left[\langle X, v\rangle^{2p}\right]^{\frac{1}{2p}} \leq L\mathbb{E}\left[\langle X, v\rangle^{p}\right]^{\frac{1}{p}}$ where L is some "nice" function (i.e. X is $(p, 2p)$-hypercontractive). Then, with probability $1-\delta$, $\forall v \in S^{d-1}$:*

$$|MOM(\langle X, v\rangle^p) - \mathbb{E}\left[\langle X, v\rangle^p\right]| \leq C2\sqrt{2}L^p\mathbb{E}\left[\langle X, v\rangle^p\right]\sqrt{\frac{d\log p + \log\left(\frac{1}{\delta}\right)}{n}}$$

This works simultaneously for all v, with C being an absolute constant. Before proceeding with the proof, we first require two lemmas.

Lemma 4. *(Warren) The VC dimension of a binary class induced by polynomials of d variables and power at most p is less than $2d\log(12p)$.*

In other words, let $X \in \mathbb{R}^d$ be a vector of d variables. Then, the binary class induced by polynomials of d variables and power at most p consists functions of the form $\mathbb{1}\{P(X) > 0\}$ where P is a polynomial of degree at most p.

The proof of the lemma, which requires machinery from Algebraic Geometry, is omitted. As an example, this lemma says that the VC dimension of $\left\{\langle X, v\rangle^{2p} : v \in S^{d-1}\right\}$ is small, since it is a polynomial of d variables and degree at most $2p$ when expanded. Furthermore, note that this lemma concerns only the degree and the number of variables in the family of polynomials, giving no regards to their coefficients.

Lemma 5. *Assume that $Y_1, ..., Y_k$ are independent random vectors and $\mathcal{F}$ is a class of $\{0,1\}$ valued functions with VC dimension d. Assume also that $\forall f \in \mathcal{F}$, $\mathbb{P}[f(Y_i) = 0] \geq \frac{7}{8}$ $\forall i$. Then, if k is chosen as $C'\left(d + \log\frac{1}{\delta}\right)$, then with probability $1-\delta$, $\forall f \in \mathcal{F}$, $\frac{1}{k}\sum_{i=1}^{k} f(Y_i) \leq \frac{1}{4}$.*

The constant C' is adjusted for the other constant $\frac{7}{8}$; otherwise, it is absolute. This lemma allows us to conduct block-level analysis first, then merge them into an overall statement.

Proof. By the VC bound for empirical processes, with probability $1-\delta$, $\forall f \in \mathcal{F}$:

$$\frac{1}{k}\sum_{i=1}^{k} f(Y_i) \leq \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} f(Y_i)\right] + C''\sqrt{\frac{d + \log\frac{1}{\delta}}{k}} \leq \frac{1}{8} + \frac{1}{8} \leq \frac{1}{4}$$

The second inequality is due to $\mathbb{E}[f(Y_i)] = \mathbb{P}[f(Y_i) = 1] \leq \frac{1}{8}$ and our choice of k. □

Proof. (of Theorem 3) To set up the problem, let there be n observations, k blocks for median-of-means and $m = \frac{n}{k}$ elements per block. Consider the mapping $(x_1, ..., x_m) \mapsto \mathbb{1}\left\{\frac{\left|\frac{1}{m}\sum_{i=1}^{m}\langle x_i, v\rangle^p - \mathbb{E}[\langle x, v\rangle^p]\right|}{\mathbb{E}[\langle x, v\rangle^p]} \geq 2\sqrt{2}L^p\frac{1}{\sqrt{m}}\right\}$. Intuitively, it maps elements of a block to an indicator, denoting whether the block is "bad". Denote the mapping by f_v, where v is just a particular direction.

By Chebyshev's inequality, we bound the probability of a "bad" block:

$$\mathbb{P}\left[\frac{\left|\frac{1}{m}\sum_{i=1}^{m}\langle x_i, v\rangle^p - \mathbb{E}\left[\langle x, v\rangle^p\right]\right|}{\mathbb{E}\left[\langle x, v\rangle^p\right]} \geq 2\sqrt{2}L^p\frac{1}{\sqrt{m}}\right] \leq \frac{\mathbb{E}\left[\langle x, v\rangle^{2p}\right]}{m\left(\mathbb{E}\left[\langle x, v\rangle^p\right]\right)^2 8L^{2p}\frac{1}{m}} \leq \frac{1}{8}$$

where the last inequality holds by hypercontractivity, which gives $\mathbb{E}\left[\langle x, v\rangle^{2p}\right] \leq L^{2p}\mathbb{E}\left[\langle x, v\rangle^p\right]^2$.

We repeat the above process for each direction $v \in S^{d-1}$, obtaining a class of $\{0,1\}$-valued functions defined on each block $Y_i = \left(X_{im-(m-1)}, ..., X_{im}\right)$ for $i \in \{1, ..., k\}$. Mathematically, our function class is $\mathcal{F} = \left\{f_v : v \in S^{d-1}\right\}$, each satisfying $\mathbb{P}\left[f_v(Y_i) = 0\right] \geq \frac{7}{8}\ \forall i \in \{1, ..., k\}$. It suffices to check that $\mathcal{F}$ has a small VC dimension, which is guaranteed by Warren's lemma. To see this, note that $\frac{\frac{1}{m}\sum_{i=1}^{m}\langle x_i, v\rangle^p - \mathbb{E}[\langle x, v\rangle^p]}{\mathbb{E}[\langle x, v\rangle^p]} - 2\sqrt{2}L^p\frac{1}{\sqrt{m}}$ is a polynomial of degree p and has d variables. By Warren's lemma, $\mathcal{F}$ has VC dimension less than $C'''d\log p$.

Thus, by Lemma 5, we choose the number of blocks $k = C_1\left(d\log p + \log\left(\frac{1}{\delta}\right)\right)$, it implies that the condition $\frac{\left|\frac{1}{m}\sum_{i=1}^{m}\langle x_i, v\rangle^p - \mathbb{E}[\langle x, v\rangle^p]\right|}{\mathbb{E}[\langle x, v\rangle^p]} \leq 2\sqrt{2}L^p\frac{1}{\sqrt{m}}\ \forall v \in S^{d-1}$ is violated in less than $\frac{1}{4}$ of the blocks with high probability. In the case where the condition is not violated, the median block satisfies the condition (since the condition must be violated consecutively starting from the tails). Hence, the median block satisfies the condition with high probability.

In conclusion, with $\frac{1}{\sqrt{m}} = \frac{d\log p + \log\left(\frac{1}{\delta}\right)}{n}$, we have:

$$\left|\mathrm{MOM}\left(\langle X, v\rangle^p\right) - \mathbb{E}\left[\langle X, v\rangle^p\right]\right| \leq C2\sqrt{2}L^p\mathbb{E}\left[\langle X, v\rangle^p\right]\sqrt{\frac{d\log p + \log\left(\frac{1}{\delta}\right)}{n}}$$

□

The upshot is that the median-of-means estimator allows you to estimate the moments of hypercontractive distributions while only knowing the first two moments.

3 One-sided Lower Tail Bound Under Few Moments

3.1 Motivation

Many statistics, such as variances and singular values, are always nonnegative. For example, we care about the smallest singular value because it appears as we invert a covariance matrix in regression problems. In those scenarios, we can still give a non-asymptotic, high probability, one-sided lower tail bound with only the first two moments.

Lemma 6. *Let $X_1, ..., X_n$ be i.i.d. random variables such that $\mathbb{E}\left[X_i\right] = \mu$ and $\mathbb{E}\left[X_i^2\right] = \sigma^2$ and $X_i \geq 0$ $\forall i \in \{1, ..., n\}$. Then $\forall t \geq 0$, $\mathbb{P}\left[\mu - \frac{1}{n}\sum_{i=1}^{n}X_i > t\right] \leq e^{-\frac{t^2 n}{2\sigma^2}}$*

Proof. Take $\lambda > 0$. Consider $\mathbb{E}\left[e^{-\lambda X_i}\right]$.

$$\begin{aligned}\mathbb{E}\left[e^{-\lambda X_i}\right] &\leq \mathbb{E}\left[1 - \lambda X_i + \frac{1}{2}\lambda^2 X_i^2\right] \\ &= 1 - \lambda\mu + \frac{1}{2}\lambda^2\sigma^2 \\ &\leq e^{-\lambda\mu + \frac{1}{2}\lambda^2\sigma^2}\end{aligned}$$

Thus, $\mathbb{E}\left[e^{\lambda(\mu-X_i)}\right] \leq e^{\frac{1}{2}\lambda^2\sigma^2}$.

$$\mathbb{P}\left[\mu - \frac{1}{n}\sum_{i=1}^{n} X_i > t\right] \leq \frac{\mathbb{E}\left[e^{\lambda\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i\right)}\right]}{e^{\lambda t}} = \frac{\Pi_{i=1}^{n}\mathbb{E}\left[e^{\frac{\lambda}{n}(\mu - X_i)}\right]}{e^{\lambda t}} \leq \frac{e^{\frac{\lambda^2\sigma^2}{2n}}}{e^{\lambda t}} = e^{\frac{\lambda^2\sigma^2}{2n} - \lambda t}$$

Optimising over λ yields $\lambda^* = \frac{tn}{\sigma^2}$ and $\mathbb{P}\left[\mu - \frac{1}{n}\sum_{i=1}^{n} X_i > t\right] \leq e^{\frac{-t^2 n}{2\sigma^2}}$ as desired. □

Lecture 21: Scribe template

Instructor: Nikita Zhivotovskiy *Scriber: Jimmy Chin*

1 One-sided Lower Tail Bound Under Few Moments (cont.)

From the last lecture, we can derive a one-sided tail bound for the sample mean.

Proposition 1. *Suppose $X_1, ..., X_n$ are iid with $X_i \geq 0$, $EX_i^2 = \sigma^2$, and $EX_i = \mu$. Then $\forall t$,*

$$\Pr(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq t) \leq \exp\{-\frac{t^2 n}{2\sigma^2}\}.$$

Note that the right-hand side of the above looks sub-Gaussian. The "magic" of this bound is that we only require two moments.

2 Least Singular Value of the Sample Covariance

Let X be a zero-mean random vector in $\mathbb{R}^d$.

Assumption 2. *Assume there exists $c \in (0,1)$, $\beta \in (0,1)$, and for all $v \in S^{d-1}$,*

$$\Pr(|\langle X, v\rangle| > c\sqrt{\mathbb{E}\langle X, v\rangle^2}) \geq \beta$$

Note $\lambda_{\min}(\Sigma) = \inf_{v \in S^{d-1}} v^T \Sigma v$, where Σ is psd. Then for all $v \in S^{d-1}$,

$$\begin{aligned} v^T(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T)v &= \frac{1}{n}\sum_{i=1}^{n}\langle X_i, v\rangle^2 \\ &\geq \frac{c^2\mathbb{E}\langle X, v\rangle^2}{n} \mid \{i \in [n] : |\langle X_i, v\rangle| \geq c\sqrt{\mathbb{E}\langle X, v\rangle^2}\} \mid, \end{aligned} \tag{1}$$

where we have used Assumption 2 to obtain the lower bound. We also have

$$\sum_{i=1}^{n} \mathrm{Ind}[|\langle X_i, v\rangle| \geq c\sqrt{\mathbb{E}\langle X, v\rangle^2}] - \underbrace{\Pr(|\langle X_i, v\rangle| \geq c\sqrt{\mathbb{E}\langle X, v\rangle^2})}_{\beta} \leq c_2\sqrt{n(d + \log(1/\delta))}, \tag{2}$$

which follows by either applying properties of the VC dimension for half spaces or Warren's lemma (which implies the VC dimension $\leq c_1 d$).

Inequalities (1) and (2) imply

$$\begin{aligned} v^T(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T)v &\geq \frac{c^2\mathbb{E}\langle X, v\rangle^2}{n}[\beta n - c_2\sqrt{n(d + \log(1/\delta))}] \\ &= c^2\mathbb{E}\langle X, v\rangle^2 \underbrace{[\beta - c_2\sqrt{\frac{d + \log(1/\delta)}{n}}]}_{(\star)}. \end{aligned}$$

Assume that n is such that $(\star) \geq \frac{\beta}{2} \Leftrightarrow c_2\sqrt{\frac{d+\log(1/\delta)}{n}} \leq \frac{\beta}{2} \Leftrightarrow n \geq (\frac{2c_2}{\beta})^2(d+\log(1/\delta))$. Then it follows that

$$\lambda_{\min}(\frac{1}{n}\sum_{i=1}^{n} X_i X_I^T) \geq \frac{c^2\beta}{2}\lambda_{\min}(\Sigma),$$

where $\Sigma = \mathbb{E}XX^T$.

When does Assumption 2 hold?

Lemma 3. *Paley-Zygmund Inequality. If $Z \geq 0$ is a random variable and $c \in (0,1)$, then*

$$\Pr(Z \geq c\mathbb{E}Z) \geq (1-c)^2\frac{(\mathbb{E}Z)^2}{\mathbb{E}Z^2}.$$

Proof of Lemma 3.

$$\begin{aligned}\mathbb{E}Z &= \mathbb{E}[Z \cdot \mathrm{Ind}[Z \geq c\mathbb{E}Z]] + \mathbb{E}[Z \cdot \mathrm{Ind}[Z < c\mathbb{E}Z]] \\ &\leq \sqrt{\mathbb{E}Z^2}\sqrt{\Pr(Z \geq c\mathbb{E}Z)} + c\mathbb{E}Z \\ \Rightarrow \Pr(Z \geq c\mathbb{E}Z) &\geq (1-c)^2\frac{(\mathbb{E}Z)^2}{\mathbb{E}Z^2},\end{aligned}$$

where we have used Cauchy-Schwarz in the first inequality. □

Now we can apply this to $(\star)$. Assume that

$$\forall \upsilon \in S^{d-1}, (\mathbb{E}\langle X, \upsilon\rangle^4)^{1/4} \leq L(\mathbb{E}\langle X, \upsilon\rangle^2)^{1/2}, \text{(hypercontractivity)}.$$

Apply the Paley-Zygmund inequality to obtain

$$\begin{aligned}\Pr(|\langle X, \upsilon\rangle| \geq c\sqrt{\mathbb{E}\langle X, \upsilon\rangle^2}) &= \Pr(\langle X, \upsilon\rangle^2 \geq c^2\mathbb{E}\langle X, \upsilon\rangle^2) \\ &= (1-c^2)^2\frac{(\mathbb{E}\langle X, \upsilon\rangle^2)^2}{\mathbb{E}\langle X, \upsilon\rangle^4} \\ &\geq (1-c^2)^2\frac{1}{L^4} \\ &= \beta.\end{aligned}$$

3 Nonparametric Least Squares

The setup is

$$Y_i = f^*(X_i) + \xi_i,$$

where $X_1, ..., X_n$ are fixed design vectors, $\xi \sim \mathcal{N}(0,1)$, and $\xi_1, ..., \xi_n$ are independent. Previously, we studied linear regression where $f^*(X_i) = \langle \beta^*, X_i\rangle$. Here, we only know that f^* belongs to some class $\mathcal{F}$, which we could estimate via least squares

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

We define the following norm

Definition 4.

$$|| f - Y ||_n^2 := \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

Our goal is to bound $|| \hat{f} - f^* ||_n^2$. By definition we have

$$\begin{aligned}
|| f^* - Y ||_n^2 \geq & || \hat{f} - Y ||_n^2 \\
= & || \hat{f} - f^* ||_n^2 + || f^* - Y ||_n^2 + \frac{2}{n}\langle \hat{f} - f^*, \underbrace{f^* - Y}_{-\xi} \rangle \\
\Rightarrow & || \hat{f} - f^* ||_n^2 \leq \frac{2}{n}\langle \xi, \hat{f} - f^* \rangle \\
\Leftrightarrow & || \hat{f} - f^* ||_n \leq \frac{2}{n}\langle \xi, \frac{\hat{f} - f^*}{|| \hat{f} - f^* ||_n} \rangle,
\end{aligned}$$

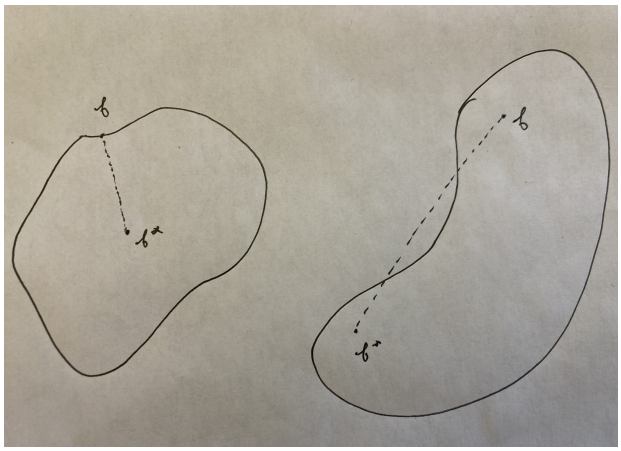
where in the second line $\langle \hat{f} - f^*, f^* - Y \rangle = \sum_{i=1}^{n}(\hat{f}(X_i) - f^*(X_i))(f^*(X_i) - Y_i)$.

For the next trick fix $t > 0$. Then

$$\begin{aligned}
|| \hat{f} - f^* ||_n = & || \hat{f} - f^* ||_n \, \mathrm{Ind}\{|| \hat{f} - f^* ||_n < t\} + || \hat{f} - f^* ||_n \, \mathrm{Ind}\{|| \hat{f} - f^* ||_n > t\} \\
\leq & \, t + \underbrace{\sup_{f \in \mathcal{F}, ||f - f^*|| \geq t} \frac{2}{n}\langle \xi, \frac{f - f^*}{|| f - f^* ||_n} \rangle}_{(\star\star)}.
\end{aligned}$$

In what follows, we will make use of the following definition.

Definition 5. *$\mathcal{F}$ is star-shaped around f^* if $\forall \alpha \in [0,1]$ and $f \in \mathcal{F}$, $\alpha(f - f^*) \in \mathcal{F} - f^*$.*



The left side of the above is an example of a star-shaped class. The right side is an example of a class that is not star-shaped.

Remark 6. *If $\mathcal{F}$ is convex, then it is star-shaped. A star-shaped set is not necessarily convex.*

We claim that if $\mathcal{F}$ is star-shaped around f^*. Then the supremum in $(\star\star)$ is achieved at some f such that $||f - f^*|| = t$. Let $f - f^*$ be a maximizer with $||f - f^*||_n > t$. Then

$$t\frac{f - f^*}{|| f - f^* ||_n} \in \mathcal{F} - f^*.$$

Moreover,

$$(\star\star) \leq t + \frac{2}{nt} \sup_{f \in \mathcal{F}, ||f-f^*||_n} \langle \xi, f - f^* \rangle = G(\xi),$$

where G is the Gaussian width.

Consider $\frac{1}{n} \sup_{f \in \mathcal{F}, ||f-f^*|| \leq t} \langle \xi, f - f^* \rangle = G(\xi)$. Then

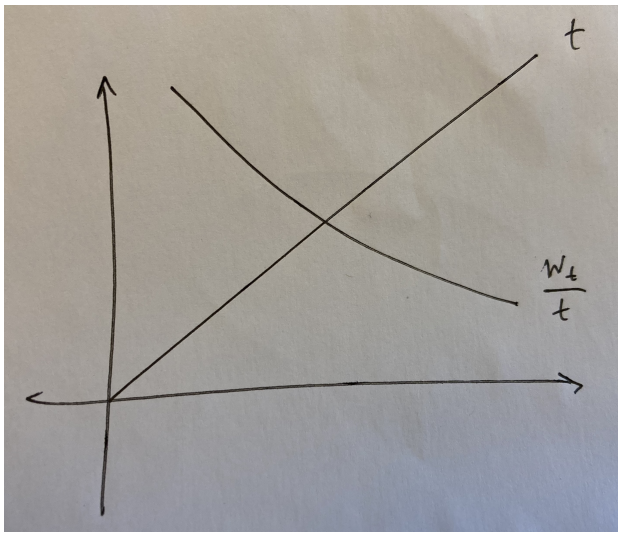
$$\begin{aligned} G(\xi) - G(\upsilon) &\leq \frac{1}{n} [\sup_{f \in \mathcal{F}} \langle \xi, f - f^* \rangle - \sup_{f \in \mathcal{F}} \langle \upsilon, f - f^* \rangle] \\ &\leq \frac{1}{n} \sup_{f \in \mathcal{F}} \langle \xi - \upsilon, f - f^* \rangle \\ &= G(\xi - \upsilon) \\ &\leq \frac{1}{n} \, || \, \xi - \upsilon \, ||_2 || \, f - f^* \, ||_2 \\ &\leq \frac{t}{\sqrt{n}} \, || \, \xi - \upsilon \, ||_2 \, . \end{aligned}$$

By Gaussian concentration, we have

$$\Pr(\frac{1}{n} \sup_{f \in \mathcal{F}, ||f-f^*||_n \leq t} \langle \xi, f - f^* \rangle \geq \mathbb{E}G(\xi) + \mu) \leq \exp(-\frac{\mu^2 n}{2t^2}),$$

for all $\mu > 0$.

Let $W_t = EG_t(\xi)$. Before we had $(\star\star) \leq t + \frac{2}{t}(W_t + \mu)$ with high probability.



Check that since $\mathcal{F}$ is star-shaped around f^*, the function W_t/t is non-increasing. We want to find the fixed point t_n such that $t_n = \frac{2}{t_n} W_{t_n}$ (see figure above). Moreover, by the non-increasing property, we have for all $t \geq t_n$ that $\frac{2}{t} W_t \leq t$. Choose $t \geq t_n$ such that $(\star\star) \leq \inf_{t \geq t_n} 2t + \frac{2\mu}{t}$. To sum up, we have

Proposition 7. *If $\mathcal{F}$ is star-shaped around f^* and ξ is Gaussian noise, then for all $t \geq t_n$*

$$|| \, \hat{f} - f^* \, ||_n^2 \leq (2t + \frac{2\mu}{t})^2,$$

with probability at least $1 - \exp(-\frac{\mu^2 n}{2t^2})$.

Lecture 22: Applications of Localization

Instructor: Nikita Zhivotovskiy *Scriber: Yichen Xu* *Proofreader: Shunan Jiang*

1 Proposition revisit

In last the lecture, we investigated the Proposition related to the localization methods, approaches in statistics working with Gaussian width/Rademacher averages of localized sets of functions.

Proposition 1. *Suppose $\mathcal{F}$ is a star-shaped function shape. The model is $y = f^*(x_i)+\xi_i$, where $x_1, \cdots, x_n \in \mathbb{R}^d$, $\xi_i \sim N(0,1)$, $\xi = (\xi_1, \cdots, \xi_n)^T$. Let $f^* \in F$ and $\mathcal{F}$ star shaped around f^*. Let $\hat{f}$ be the least squares estimator, i.e. $\hat{f} = argmin_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(X_i) - Y_i)^2$. Let t_n be the solution of the equation*

$$t = \frac{2}{tn}\mathbb{E}sup_{f\in\mathcal{F}, \|f-f^*\|_n \le t}\langle \xi, f - f^*\rangle$$

Then for any $u > 0$ with probability $1 - exp(-\frac{u^2 n}{2t^2})$, we have

$$\|\hat{f} - f^*\|_n^2 \le \left(2t + \frac{2}{t}u\right)^2$$

Remark 2. *In the proposition, we do not care about the complexity of the whole class, but those f close to f^*. The uniform convergence tells us that for all f, $\|f - Y\|_n \approx \mathbb{E}\|f - Y\|_n$, while in the localization, we do not care about f faraway since they do not affect the empirical minimizer a lot.*

2 Example 1.

We explore the application of Proposition 1 by examples. Suppose we are interested in the case $u = t^2$. Then $\|\hat{f} - f^*\|_n^2 \le (2t + 2t)^2 = 16t^2$, with probability $1 - \exp\left(-\frac{t^2 n}{2}\right)$. First we bound the Gaussian width of $\{f : \|f - f^*\|_n^2 \le t\}$, by Dudley integral.

$$\frac{1}{n}\mathbb{E}\sup_{f\in F, \|f-f^*\|\le t}\langle \xi, f - f^*\rangle \le c(\alpha + \underbrace{\frac{1}{\sqrt{n}}\int_{\frac{\alpha}{4}}^t \sqrt{\log N(F - f^*, L_2(P_n), \epsilon)}d\epsilon}_{\text{Denoted as } (*)})$$

For nonparametric classes, we have $\sup_{P_n} \log N(\mathcal{F}, L_2(P_n), \epsilon) \le c^{'}\epsilon^{-p}$. Here, we assume $p \in (0, 2)$. Choose $\alpha = 0$, we have

$$(*) \le \frac{c^{''}\int_0^t \epsilon^{-\frac{p}{2}}d\epsilon}{\sqrt{n}} \le \frac{c^{'''}t^{-\frac{p}{2}+1}}{\sqrt{n}}$$

Then we reduce everything to solving

$$t^2 = c_1 \frac{t^{-\frac{p}{2}+1}}{\sqrt{n}}$$
$$t^{\frac{p}{2}+1} = \frac{c_1}{n^{\frac{1}{2}}}$$
$$t = \frac{c_2}{n^{\frac{1}{2}\cdot\frac{1}{\frac{p}{2}+1}}} = \frac{c_2}{n^{\frac{1}{p+2}}}$$

Remark 3. *When we want to apply Proposition 1, we can solve the equation by substituting the upper bound of the "Gaussian width term" since the we are guaranteed that $t \geq t_n$ and the inequality holds for $t \geq t_n$. When $t = t_n$, the bound is the tightest.*

Plug in the solved t and continue with Proposition 1, we have

$$\|\hat{f} - f^*\| \leq c_3 n^{-\frac{2}{p+2}}$$

Note that in uniform convergence, $\|f - f^*\|_n^2 - \mathbb{E}\|f - f^*\|_n^2 \sim n^{-\frac{1}{2}}$. When $p \in (0, 2)$, $\frac{1}{n^{\frac{2}{p+2}}} \ll \frac{1}{\sqrt{n}}$. More concretely, we can choose $t = c_2 n^{-\frac{1}{p+2}} + \frac{2\sqrt{\log(\frac{1}{\delta})}}{\sqrt{n}} > t_n$ since $\frac{2\sqrt{\log(\frac{1}{\delta})}}{\sqrt{n}}$ is a positive term. Dropping $c_2 n^{-\frac{1}{p+2}}$ from t and plug $\frac{2\sqrt{\log(\frac{1}{\delta})}}{\sqrt{n}}$ into the probability estimator $1 - \exp(-\frac{nt^2}{2})$, we obtain the conservative high probability $1 - \delta$. On the other hand, observing that

$$t^2 = \left(c_2 n^{-\frac{1}{p+2}} + \frac{2\sqrt{\log(\frac{1}{\delta})}}{\sqrt{n}}\right)^2 \leq 2\left(c_2^2 n^{-\frac{2}{p+2}} + \frac{4\log(\frac{1}{\delta})}{n}\right) \ll c\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right)$$

we make the bound tighter than uniform convergence.

3 Example 2.

Consider the parametric cases. Let $H_t = \{f : \|f - f^*\|_n \leq t\}$, $f^*(x) = \langle x, \beta\rangle$, $N(H_t, L_2(P_n), \epsilon) \leq c_4(1 + \frac{2t}{\epsilon})^p$, where p is the dimension. It is expected that the covering number of the localized set scales with the radius t. Plug this upper bound to Dudley integral, we obtain obtain the upper bound for the Gaussian width

$$\frac{1}{\sqrt{n}}\int_0^t \sqrt{p}\sqrt{\log(1 + \frac{2t}{\epsilon})}d\epsilon = \frac{1}{\sqrt{n}} t \int_0^1 \sqrt{p}\sqrt{\log(1 + \frac{2}{\epsilon})}d\epsilon \leq \frac{c_5\sqrt{p}t}{\sqrt{n}}$$

where we use change of variable in the first equality. Again, by solving $t^2 = \frac{2c_5 t\sqrt{p}}{\sqrt{n}}$, we obtain t_n. We choose $t = c_6\sqrt{\frac{p+\log(\frac{1}{\delta})}{n}}$ so that $t \geq t_n$, and we have

$$\|\hat{f} - f\|_n^2 \leq c_7\left(\frac{p + \log(\frac{1}{\delta})}{n}\right) \text{ with probability } 1 - \delta$$

4 Example 3. (Non-parametric regression with random design)

Now, we study the non-parametric regression with random design. Let $\mathcal{F}$ be the class of function and is convex. Suppose $|Y| \leq m$, $\forall f \in \mathcal{F}$, $|f| \leq m$ bounded. We use notations as follows

$$R(f) = \mathbb{E}(f(X) - Y)^2$$
$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2$$

The least squares estimator in $\mathcal{F}$, $\hat{f} = \text{argmin}_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$

The goal is to bound $R(\hat{f}) - \inf_{f\in\mathcal{F}}R(f)$.

Remark 4. *Here we do not assume that the model is $Y = f^*(X) + \xi$, with ξ independent with X (called mispecified model), opposite to assuming that $f^*_{Bayes}(X) = \mathbb{E}, [Y|X = x]$, $f^*_B \in \mathcal{F}$ and ξ independent noise.*

Previously, we used the trick $R_n(\hat{f}) \leq R_n(f^*)$. Here, we use the claim that

$$R_n(f) - R_n(\hat{f}) \geq \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \hat{f}(X_i))^2$$

This claim is due the convex nature of $\mathcal{F}$ as can be demonstrated by Figure 1.

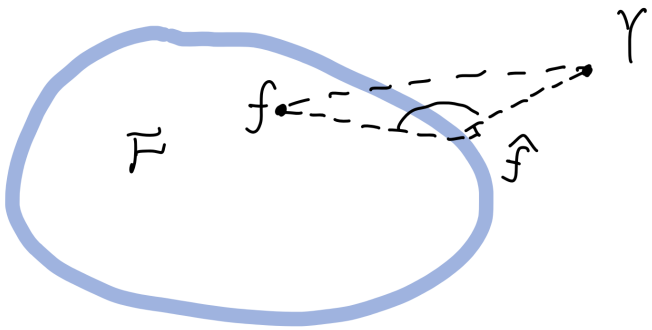


Figure 1: Visualization of the claim.

Recall the notation of $\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \hat{f}(X_i))^2 = P_n(f - \hat{f})^2$ and $\mathbb{E}[Y - f(X)]^2 = P(Y - f)^2$. By substituting the above inequality,

$$R(f) - R(f^*) \leq P(f - Y)^2 - P(f^* - Y)^2 + P_n(f^* - Y)^2 - P_n(\hat{f} - Y)^2 - P_n(\hat{f} - f^*)^2$$

Write $P(f - Y)^2 = P(f - f^* + f^* - Y)$ and $P_n(\hat{f} - Y)^2 = P_n(\hat{f} - f^* + f^* - Y)^2$ and expand them, the above becomes

$$(P - P_n)(f - f^*)^2 + 2(P - P_n)(f - f^*)(f^* - Y) - P_n(\hat{f} - f^*)^2 \qquad (*)$$

Let $\xi = Y - f^*(X)$, we further write equation (*) as

$$
\begin{aligned}
(*) &= 2(P_n - P)\xi(\hat{f} - f^*) + P(\hat{f} - f^*)^2 - 2P_n(\hat{f} - f^*)^2 \\
&= \underbrace{2(P_n - P)\xi(\hat{f} - f^*)}_{\text{I}} + \underbrace{\frac{5}{4}P(\hat{f} - f^*)^2}_{\text{II}} - \underbrace{\frac{1}{4}P(\hat{f} - f^*)^2}_{\text{III}} - \underbrace{\frac{7}{4}P_n(\hat{f} - f^*)^2}_{IV} - \underbrace{\frac{1}{4}P_n(\hat{f} - f^*)^2}_{V}
\end{aligned}
$$

Group I, III, V together and II, IV together and take sup, we have

$$
\begin{aligned}
\mathbb{E}[R(\hat{f}) - R(f^*)] \leq & \underbrace{\mathbb{E}\sup_{f\in\mathcal{F}}\left(2(P_n - P)\xi - \frac{1}{4}P(f - f^*) - \frac{1}{4}P_n(f - f^*)^2\right)}_{\text{(I)}} \\
& + \underbrace{\mathbb{E}\sup_{f\in F}\left(\frac{5}{4}P(f - f^*)^2 - \frac{7}{4}P_n(f^* - f)^2\right)}_{\text{(II)}}
\end{aligned}
$$

We first analyze (I). Let $(X_i^{'}, Y_i^{'})$ be independent copies of $(X_i, Y_i)_{i=1}^n$ and $P_n^{'}$ be corresponding empirical measure. Note that $P = \mathbb{E}^{'} P_n^{'}$. By Jensen's inequality

$$
\text{(I)} \leq \mathbb{E}\sup_{f\in\mathcal{F}}\left(2(P_n - P_n^{'})\xi(f - f^*) - \frac{1}{4}P_n^{'}(f - f^*)^2 - \frac{1}{4}P_n(f - f^*)^2\right)
$$

Applying symmetrization on

$$
(P_n - P_n^{'})\xi(f - f^*) = \frac{1}{n}\sum_{i=1}^n\left(\xi_i(f(X_i) - f^*(X_i)) - \xi^{'}(f(X_i^{'}) - f^*(X_i^{'}))\right)
$$

We obtain

$$
\begin{aligned}
\text{(I)} &= \mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{2}{n}\sum_{i=1}^n\epsilon_i(\xi_i(f(X_i) - f^*(X_i)) - \xi_i^{'}(f(X_i^{'}) - f^*(X_i^{'}))) - \frac{1}{4}P_n^{'}(f - f^*)^2 - \frac{1}{4}P_n(f - f^*)^2\right) \\
&\leq 2\mathbb{E}_{(X_i,Y_i),\epsilon_i}\sup_{f\in\mathcal{F}}\left(\frac{2}{n}\sum_{i=1}^n\epsilon_i\xi_i(f(X_i) - f^*(X_i)) - \frac{1}{4}P_n(f - f^*)^2\right) \qquad (**)
\end{aligned}
$$

Remark 5. *Here we can apply symmetrization safely because both terms in $-\frac{1}{4}P_n^{'}(f - f^*)^2 - \frac{1}{4}P_n(f - f^*)^2$ have the same negative sign.*

In (**), the term $\frac{1}{4}P_n(f - f^*)^2$ can be bounded using Rademacher average with offset shown in the next lecture. For the another term, given $|\xi| = |Y - f^*(X)| \leq 2m$, the contraction inequality tells (repeating the contraction proof), with Lipschitz constant $2m$

$$
(**) \leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{2}{n}\sum_{i=1}^n(2m)\epsilon_i(f(X_i) - f^*(X_i)) - \frac{1}{4}P_n(f - f^*)^2\right)
$$

Lecture 23: Random Design Regression

Instructor: Nikita Zhivotovskiy *Scriber: Anish Muthali* *Proofreader: Yanbo Feng*

23.1 Random Design Regression

Recall the setup of random design regression from last lecture. Suppose the pair $(X, Y) \sim P_{X,Y}$, where $P_{X,Y}$ is an unknown distribution. We observe n IID copies of (X, Y), i.e., we observe $(X_i, Y_i)_{i=1}^n \sim_{\text{IID}} (X, Y)$. We would like to estimate a function that maps X to Y, by searching over a convex class of functions $\mathcal{F}$. Assume that $|Y| \le m$ and $|f(X)| \le m, \forall f \in \mathcal{F}$. The least squares estimator is given by

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \tag{36}$$

We would like to analyze $\mathbb{E}\left[R(\hat{f})\right] - \inf_{f \in \mathcal{F}} R(f)$, where $R(f) := \mathbb{E}\left[(Y - f(X))^2\right]$.

Proposition 23.1. *For the random design regression above, for any $\alpha, \gamma \ge 0$ such that $\alpha < \gamma$, we have*

$$\mathbb{E}\left[R(\hat{f})\right] - \inf_{f \in \mathcal{F}} R(f) \le \mathbb{E}\left[Cm\left(\alpha + \frac{1}{\sqrt{n}} \int_{\alpha}^{\gamma} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)}\, \mathrm{d}\varepsilon + \frac{m \log \mathcal{N}(\mathcal{F}, L_2(P_n), \gamma)}{n}\right)\right] \tag{37}$$

where C is an absolute constant.

Proof of Proposition 23.1. Define $P_n(f)^2 := \frac{1}{n}\sum_{i=1}^n (f(X_i))^2$ and $P(f)^2 := \mathbb{E}\left[f(X)^2\right]$. From last lecture, we were able to write

$$\mathbb{E}\left[R(\hat{f})\right] - \inf_{f \in \mathcal{F}} R(f) \le \underbrace{\mathbb{E}\left[\sup_{f \in \mathcal{F}}\left(2(P_n - P)(\xi(f - f^\star)) - \frac{1}{4}P(f - f^\star)^2 - \frac{1}{4}P_n(f - f^\star)^2\right)\right]}_{\text{Term (I)}} \tag{38}$$

$$+ \underbrace{\mathbb{E}\left[\sup_{f \in \mathcal{F}}\left(\frac{5}{4}P(f - f^\star)^2 - \frac{7}{4}P_n(f - f^\star)^2\right)\right]}_{\text{Term (II)}} \tag{39}$$

We can upper bound Term (I) using symmetrization. Define $(\varepsilon_i)_{i=1}^n$ to be IID Rademacher random variables. Hence, we upper bound Term (I) following the steps from last lecture:

$$2\mathbb{E}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} 4m\varepsilon_i(f(X_i) - f^\star(X_i)) - \frac{1}{4}P_n(f - f^\star)^2\right)\right] \tag{40}$$

$$= 8m\mathbb{E}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(f(X_i) - f^\star(X_i)) - \frac{1}{16m}P_n(f - f^\star)^2\right)\right] \tag{41}$$

Term (II) can be simplified as

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}}\left(\frac{5}{4}P(f - f^\star)^2 - \frac{7}{4}P_n(f - f^\star)^2\right)\right] \tag{42}$$

$$= \frac{5}{4}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(P(f-f^\star)^2 - \frac{7}{5}P_n(f-f^\star)^2\right)\right] \tag{43}$$

$$= \frac{5}{4}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\left(1+\frac{1}{10}\right)P(f-f^\star)^2 - \frac{6}{5}P_n(f-f^\star)^2 - \frac{1}{5}P_n(f-f^\star)^2 - \frac{1}{10}P(f-f^\star)^2\right)\right] \tag{44}$$

We can bound eq. 44 by defining an IID copy of the initial data $(X_i', Y_i')_{i=1}^n \sim_{\text{IID}} (X, Y)$. Define $Z \coloneqq (X_i, Y_i)_{i=1}^n$ and $Z' \coloneqq (X_i', Y_i')_{i=1}^n$. Using Jensen's inequality, we obtain

$$\begin{aligned}
&\frac{5}{4}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\left(1+\frac{1}{10}\right)P(f-f^\star)^2 - \frac{6}{5}P_n(f-f^\star)^2 - \frac{1}{5}P_n(f-f^\star)^2 - \frac{1}{10}P(f-f^\star)^2\right)\right] \\
&\le \frac{5}{4}\mathbb{E}_Z\left[\mathbb{E}_{Z'}\left[\sup_{f\in\mathcal{F}}\left(\frac{11}{10}P_n'(f-f^\star)^2 - \frac{11}{10}P_n(f-f^\star)^2 - \frac{1}{10}P_n(f-f^\star)^2 - \frac{1}{10}P_n'(f-f^\star)^2\right)\right]\right] \\
&\le \frac{5}{4}\mathbb{E}_Z\left[\mathbb{E}_{Z'}\left[\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left(\frac{11}{10}\sum_{i=1}^n\frac{1}{n}\varepsilon_i\left((f(X_i)-f^\star(X_i))^2 - (f(X_i')-f^\star(X_i'))^2\right)\right.\right.\right.\right. \\
&\qquad\left.\left.\left.\left. - \frac{1}{10}P_n(f-f^\star)^2 - \frac{1}{10}P_n'(f-f^\star)^2\right)\right]\right]\right] \qquad (45) \\
&\le \frac{10}{4}\mathbb{E}_{Z,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left(\frac{11}{10}\sum_{i=1}^n\frac{1}{n}\varepsilon_i(f(X_i)-f^\star(X_i))^2 - \frac{1}{10}P_n(f-f^\star)^2\right)\right] \qquad (46)
\end{aligned}$$

where $P_n'(f)^2 \coloneqq \frac{1}{n}\sum_{i=1}^n (f(X_i'))^2$. We obtain eq. 45 using symmetrization. Notice that $|f(X) - f^\star(X)| \le 2m$, so eq. 46 can be bounded as

$$\begin{aligned}
&\frac{10}{4}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{11}{10}\sum_{i=1}^n\frac{1}{n}\varepsilon_i(f(X_i)-f^\star(X_i))^2 - \frac{1}{10}P_n(f-f^\star)^2\right)\right] \qquad (47) \\
&\le \frac{10}{4}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{11}{10}\sum_{i=1}^n\frac{1}{n}\varepsilon_i\cdot 4m(f(X_i)-f^\star(X_i)) - \frac{1}{10}P_n(f-f^\star)^2\right)\right] \qquad (48) \\
&= \frac{10}{4}\cdot\frac{11}{10}\cdot 4m\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f(X_i)-f^\star(X_i)) - \frac{1}{10}\cdot\frac{10}{11}\cdot\frac{1}{4m}P_n(f-f^\star)^2\right)\right] \qquad (49)
\end{aligned}$$

Combining the bounds for terms (I) and (II), we have

$$\mathbb{E}\left[R(\hat{f})\right] - \inf_{f\in\mathcal{F}} R(f) \le 20m\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f(X_i)-f^\star(X_i)) - \frac{1}{50m}P_n(f-f^\star)^2\right)\right] \tag{50}$$

To complete the rest of the proof, we require an additional proposition.
We leverage an additional fact to complete the proof above.

Proposition 23.2. *Assume that some class of functions $\mathcal{G}$ has a covering number $\mathcal{N}(\mathcal{G}, L_2(P_n), \gamma)$ for some $\gamma > 0$, such that the zero function belongs to this cover. Then, for any $\alpha \ge 0$,*

$$\mathbb{E}_\varepsilon\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i g(X_i) - c'g^2(X_i)\right)\right] \le C\left(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^\gamma\sqrt{\log\mathcal{N}(\mathcal{G}, L_2(P_n), \varepsilon)}\,\mathrm{d}\varepsilon\right) \tag{51}$$

$$+ \frac{\frac{1}{c'} \log \mathcal{N}(\mathcal{G}, L_2(P_n), \gamma)}{n} \tag{52}$$

where C is an absolute constant

Proof of Proposition 23.2. For any $g \in \mathcal{G}$, let $\pi[g]$ denote the closest element to g in the covering net at scale γ. Hence, we have

$$\mathbb{E}_\varepsilon \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X_i) - c' g^2(X_i) \right) \right] \leq \underbrace{\mathbb{E}_\varepsilon \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i (g(X_i) - \pi[g](X_i)) \right) \right]}_{\text{Term (I)}} \tag{53}$$

$$+ \underbrace{\mathbb{E}_\varepsilon \left[\sup_{g \in \mathcal{G}} \left(\sum_{i=1}^n \frac{c'}{4} (\pi[g](X_i))^2 - c' g^2(X_i) \right) \right]}_{\text{Term (II)}} \tag{54}$$

$$+ \underbrace{\mathbb{E}_\varepsilon \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \pi[g](X_i) - \frac{c'}{4} (\pi[g](X_i))^2 \right) \right]}_{\text{Term (III)}} \tag{55}$$

First, we can approach term (II). If $\|g\|_{L_2(P_n)} \leq \gamma$, then $\pi[g] = \varnothing$, the zero function. That is, among all functions in the covering net, g is closest to the zero function. Hence, in this case, term (II) is at most 0. Otherwise, if $\|g\|_{L_2(P_n)} > \gamma$, $\|g - \pi[g]\|_{L_2(P_n)} \leq \gamma$, so we can apply the triangle inequality to again show that term (II) is at most 0.
Next, we can approach term (I) by repeating the Dudley integral proof, this time replacing the $L_2(P_n)$ diameter in the upper limit of the integral with γ, since $\|g - \pi[g]\|_{L_2(P_n)} \leq \gamma$.
Lastly, we can approach term (III) using the MGF bound on Rademacher random variables. Define the covering net at scale γ by $\mathcal{N}_\gamma$, and fix $\lambda > 0$. We have

$$\mathbb{E}_\varepsilon \left[\sup_{h \in \mathcal{N}_\gamma} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i h(X_i) - \frac{c'}{4} h^2(X_i) \right) \right] \leq \frac{1}{n} \frac{1}{\lambda} \log \mathbb{E}_\varepsilon \left[\exp \left\{ \max_{h \in \mathcal{N}_\gamma} \lambda \left(\sum_{i=1}^n \varepsilon_i h(X_i) - \frac{c'}{4} h^2(X_i) \right) \right\} \right] \tag{56}$$

$$\leq \frac{1}{n} \frac{1}{\lambda} \log \mathbb{E}_\varepsilon \left[\sum_{h \in \mathcal{N}_\gamma} \exp \left\{ \lambda \left(\sum_{i=1}^n \varepsilon_i h(X_i) - \frac{c'}{4} h^2(X_i) \right) \right\} \right] \tag{57}$$

$$\leq \frac{1}{n} \frac{1}{\lambda} \log \left| \mathcal{N}_\gamma \right| \cdot \max_{h \in \mathcal{N}_\gamma} \left(\exp \left\{ \frac{\lambda^2}{2} \sum_{i=1}^n h^2(X_i) - \frac{c' \lambda}{4} \sum_{i=1}^n h^2(X_i) \right\} \right) \tag{58}$$

$$\leq \frac{2}{n c'} \log \left| \mathcal{N}_\gamma \right| \tag{59}$$

$$= \frac{2}{n c'} \log \mathcal{N}(\mathcal{G}, L_2(P_n), \gamma) \tag{60}$$

where in eq. 59 we set $\lambda = \frac{c'}{2}$ to make the argument of the exponent zero. Combining the bounds for terms (I), (II), and (III) provide the desired result. □

Now, we can resume the proof of the random design regression bound from before.

Proof of Proposition 23.1 (continued). We resume with the previously determined bound.

$$\mathbb{E}\left[R(\hat{f})\right] - \inf_{f\in\mathcal{F}} R(f) \le 20m\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i)-f^\star(X_i)) - \frac{1}{50m}P_n(f-f^\star)^2\right)\right] \tag{61}$$

$$= 20m\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i)-f^\star(X_i)) - \frac{1}{50m}\cdot\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-f^\star(X_i))^2\right)\right] \tag{62}$$

$$= 20m\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i(f(X_i)-f^\star(X_i)) - \frac{1}{50m}(f(X_i)-f^\star(X_i))^2\right)\right)\right] \tag{63}$$

We can use the result from Proposition 23.2 by setting $g(X) = f(X) - f^\star(X)$ and $c' = \frac{1}{50m}$. Notice that $\mathcal{N}(\mathcal{F}\setminus\{f^\star\}, L_2(P_n), \varepsilon) \le \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)$. Hence,

$$\mathbb{E}\left[R(\hat{f})\right] - \inf_{f\in\mathcal{F}} R(f) \le \mathbb{E}_Z\left[Cm\left(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^\gamma \sqrt{\log\mathcal{N}(\mathcal{F}\setminus\{f^\star\}, L_2(P_n), \varepsilon)}\,\mathrm{d}\varepsilon + \frac{m\log\mathcal{N}(\mathcal{F}\setminus\{f^\star\}, L_2(P_n), \gamma)}{n}\right)\right]$$

$$\le \mathbb{E}_Z\left[Cm\left(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^\gamma \sqrt{\log\mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)}\,\mathrm{d}\varepsilon + \frac{m\log\mathcal{N}(\mathcal{F}, L_2(P_n), \gamma)}{n}\right)\right]$$

as desired. □

We can use this result in an example, Consider a non-parametric function class $\mathcal{F}$ such that, for all P_n, $\log\mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \sim \varepsilon^{-p}$, where $p \in (0,2)$. Further, assume WLOG that $m \le 1$. Also, recall that

$$\int_0^\gamma \sqrt{\log\mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)}\,\mathrm{d}\varepsilon \le \gamma^{-\frac{p}{2}+1} \tag{64}$$

To optimize over γ, we can set the two terms of the bound to be equal, i.e.,

$$\frac{1}{\sqrt{n}}\gamma^{-\frac{p}{2}+1} = \frac{\gamma^{-p}}{n} \tag{65}$$

$$\implies \gamma^{\frac{p}{2}+1} = \frac{1}{\sqrt{n}} \tag{66}$$

$$\implies \gamma = n^{-\frac{1}{p+2}} \tag{67}$$

Hence, if we set $\alpha = 0$, the final bound simplifies to

$$\mathbb{E}\left[R(\hat{f})\right] - \inf_{f\in\mathcal{F}} R(f) \lesssim n^{-\frac{2}{p+2}} \tag{68}$$

23.2 Online Learning

The next main topic we will cover is online learning. Many online learning algorithms involve some kind of dynamic update step, such as stochastic gradient descent. We would like to devise bounds on the total number of mistakes that our algorithm makes, despite these dynamic update steps.

As an example, consider a classification problem where we utilize online learning. Given a finite function class $\mathcal{F}$ of $\{0,1\}$ valued functions, let $f^\star \in \mathcal{F}$ be a target function. That is, you observe X_i, and, shortly after, $f^\star(X_i)$ will be revealed. In the coming lecture, we will show that there is a way to make at most $\log_2(|\mathcal{F}|)$ mistakes on sequences of any length.

Lecture 24: Online Learning

Instructor: Nikita Zhivotovskiy *Scribe: Zhiwei Xiao* *Proofreader: Zhiwei Xiao*

Online Learning Background

Online (machine) learning refers to the process of continuously updating and improving the best predictor (machine learning model) at each step as new data becomes available in a sequential order. It's applied in many areas including recommendation systems, natural language processing, neural networks, etc.

Online Learning Protocol/Example

Consider a classification problem, where $\mathcal{F}$ is an unknown finite class of $\{0,1\}$ classifiers (functions), and $f^\star \in \mathcal{F}$ is the target function. We also have a sequence of data points $(x_1, x_2, ..., x_n)$ with true labels. For y_1, a classifier f will give us the predicted label $\hat{y_1}$, and then the nature will reveal the true label $f^\star(x_1)$. The same process is repeated for $i = 2, ..., n$. The aim is for $\sum_{i=1}^{n} \text{Ind}[\hat{y_i} \neq f^\star(x_i)]$ to be small.

1. Naive Strategy

We call this "Follow the leader". Let $\mathcal{F}_1 = \mathcal{F}$, for $i = 1, 2, ..., n$, pick any $f \in \mathcal{F}_i$, predict $\hat{y_i} = f(x_i)$, then check the true label $f^\star(x_i)$ revealed by the nature. Then we update by setting $\mathcal{F}_{i+1} = \{f \in \mathcal{F}_i : f(x_i) = f^\star(x_i)\}$, which means that we remove all functions that don't agree with the data seen so far. This strategy makes at most $|\mathcal{F}|$ mistakes.

2. Halving Algorithm

Let $\mathcal{F}_1 = \mathcal{F}$, for $i = 1, 2, ..., n$, we predict $\hat{y_i}$ as the majority vote of $\{f(x_i) : f \in \mathcal{F}_i\}$, then check the true label $f^\star(x_i)$ revealed by the nature and do the same update as in the naive strategy. This strategy makes at most $\log_2 |\mathcal{F}|$ mistakes. (Intuition is that if the majority is right, it's right; if the majority is wrong, we'll drop at least half of the functions.)

More General Setup of Online Learning / Regret Definition

We have a loss function $L(f(X), Y)$ which is non-negative, and $\mathcal{F}$ which is a class of functions. We have a sequence of predictors $\hat{f}_1, \hat{f}_2, ..., \hat{f}_n$ such that $\hat{f}_1$ has seen no data on previous rounds, $\hat{f}_2$ has seen one data point, etc. Notice that $\hat{f}_i$ does not know $y_i|x_i$ by definition. Hence, more and more information have been seen and trained on as we move along the sequence. And we define the regret as $\sum_{i=1}^{n} L(\hat{f}_i(x_i), y_i) - \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} L(f(x_i), y_i)$.

Proposition (Online to Batch Conversion)

Assume that $L(f(X), Y)$ is convex with respect to the first argument (definition of $\hat{f}$ above) and that $(x_i, y_i)_{i=1}^n$ is an i.i.d sample. Assume also that the sequential estimators $\hat{f}_1, \hat{f}_2, ..., \hat{f}_n$ satisfy that the regret

$\sum_{i=1}^{n} L(\hat{f}_i(x_i), y_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^{n} L(f(x_i), y_i) \le \mathsf{R}^{(n)}$ almost surely, where $\mathsf{R}^{(n)}$ is a constant. Then we have that

$$\mathop{\mathbb{E}}_{(x_i,y_i)_{i=1}^n} \mathop{\mathbb{E}}_{(X,Y)\sim P_{X,Y}} L(\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(X), Y) - \inf_{f\in\mathcal{F}} \mathbb{E}\, L(f(X), Y) \le \frac{\mathsf{R}^{(n)}}{n}.$$

Proof:

Define $S = (x_i, y_i)_{i=1}^n$. We know the regret satisfies $\sum_{i=1}^{n} L(\hat{f}_i(x_i), y_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^{n} L(f(x_i), y_i) \le \mathsf{R}^{(n)}$ ♦ for any sequence. And remember that $\hat{f}_i$ does not know $y_i|x_i$ by definition, and that the pair (x_i, y_i) is independent of $(x_1, y_1), ..., (x_{i-1}, y_{i-1})$. Hence, we can take the expectation on the both sides of ♦, and we first analyze its left-hand side after taking the expectation:

$$\mathop{\mathbb{E}}_{S} \sum_{i=1}^{n} L(\hat{f}_i(x_i), y_i) - \mathop{\mathbb{E}}_{S} \inf_{f\in\mathcal{F}} \sum_{i=1}^{n} L(f(x_i), y_i) \ge \sum_{i=1}^{n} \mathop{\mathbb{E}}_{S} R(\hat{f}_i) - \inf_{f\in\mathcal{F}} nR(f) \ge n\left(\mathop{\mathbb{E}}_{S} R(\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i) - \inf_{f\in\mathcal{F}} R(f)\right),$$

where the risk function $R(f) = \mathbb{E}_{(X,Y)\sim P_{X,Y}}[L(f(X), Y)]$, we used Jensen's inequality and inf's concavity in the first $\ge$, and used first argument and the convexity in the second $\ge$. We finish the proof by dividing both sides by n.

Exponential Weights (Hedge Algorithm)

Consider $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ - a family of functions parameterized by $\Theta \subseteq \mathbb{R}^d$. The prior distribution $\pi(\theta)$ over Θ does not depend on any data. $\eta > 0$ is the fixed learning rate. After seeing i data points $(x_1, y_1), ..., (x_i, y_i)$, denote

$$\hat{\rho}_i(\theta) = \frac{\exp\left(-\eta \sum_{j=1}^{i} L(f_\theta(x_j), y_j)\right)\pi(\theta)}{\mathbb{E}_{\theta\sim\pi} \exp\left(-\eta \sum_{j=1}^{i} L(f_\theta(x_j), y_j)\right)}.$$

And denote $g_\theta(z_i) = L(f_\theta(x_i), y_i)$ for simplicity. See that we can work with

$$\hat{\rho}_0(\theta) = \pi(\theta), \quad \hat{\rho}_i(\theta) = \frac{\exp(-\eta g_\theta(z_i)) \cdot \hat{\rho}_{i-1}(\theta)}{\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}} \exp(-\eta g_\theta(z_i))}.$$

The claim is that $-\frac{1}{\eta}\log \mathbb{E}_{\theta\sim\pi} \exp(-\eta \sum_{i=1}^{n} g_\theta(z_i)) = -\frac{1}{\eta}\sum_{i=1}^{n} \log \mathbb{E}_{\theta\sim\hat{\rho}_{i-1}} \exp(-\eta g_\theta(z_i))$, where we denote the left hand side as $H(n)$ and the right hand side as the so-called "mix-loss".
See that

$$\begin{aligned} H(n) - H(n-1) &= -\frac{1}{\eta}\log\left(\frac{\mathbb{E}_{\theta\sim\pi} \exp\left(-\eta g_\theta(z_n) - \eta \sum_{i=1}^{n-1} g_\theta(z_i)\right)}{\mathbb{E}_{\theta\sim\pi} \exp(-\eta \sum_{i=1}^{n-1} g_\theta(z_i))}\right) \\ &= -\frac{1}{\eta}\log\left(\mathop{\mathbb{E}}_{\theta\sim\hat{\rho}_{n-1}} \exp(-\eta g_\theta(z_n))\right). \end{aligned}$$

The idea is that we can relate the mix-loss with the true loss of our prediction. We can build $\hat{f}_1, ..., \hat{f}_n$ such that $L(\hat{f}_i(x_i), y_i) \leq -\frac{1}{\eta} \log\left(\mathbb{E}_{\theta \sim \hat{\rho}_{i-1}} \exp(-\eta L(f_\theta(x_i), y_i))\right)$. Then by our claim before, we have that

$$\begin{aligned}\sum_{i=1}^{n} L(\hat{f}_i(x_i), y_i) &\leq -\frac{1}{\eta} \sum_{i=1}^{n} \log\left(\mathop{\mathbb{E}}_{\theta \sim \hat{\rho}_{i-1}} \exp\left(-\eta L(f_\theta(x_i), y_i)\right)\right) \\ &= -\frac{1}{\eta} \log\left(\mathop{\mathbb{E}}_{\theta \sim \pi} \exp(-\eta \sum_{i=1}^{n} L(f_\theta(x_i), y_i))\right).\end{aligned}$$

Ways to Interpret the Logarithmic Loss

1: $\mathcal{F} = \left\{\forall f \in \mathcal{F} : \int f(x)dx = 1, f(x) \geq 0\right\}$ is a family of densities in $\mathbb{R}$.
If $X \sim f(X)$, $\mathbb{E}_X - \log(f(x)) = \int -\log(f(x))f(x)dx$: entropy.

2: X follows a distribution according to $f(X)$.

$$\mathop{\mathbb{E}}_{X}\left[-\log(\hat{f}(x)) - (-\log(f(x)))\right] = \mathop{\mathbb{E}}_{X} \log(\frac{f(x)}{\hat{f}(x)}) = \int \log(\frac{f(x)}{\hat{f}(x)})f(x)dx = \mathrm{KL}(f||\hat{f}).$$

3: Cross-Entropy Loss.

Lecture 25: Prediction with Logarithmic Loss

Instructor: Nikita Zhivotovskiy *Scriber: Zach Rewolinski* *Proofreader: Reece Huff*

25.1 Reminder: Why We Use Logarithmic Loss

The following neat properties encourage us to use logarithmic loss.

1. Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a family of densities. If $\int f_\theta(x)dx = 1$ with $\mathcal{L}_\theta(x) \geq 0$, then $\mathbb{E}_{x\sim f_\theta}[-\log(f_\theta(x))]$ is the entropy.
2. $\mathbb{E}_{x\sim f_\theta}[-\log(g(x)) - (-\log(f_\theta(x)))] = KL(f_\theta||g)$
3. Consider a classification task, where $y \in \{0,1\}$ and we predict the probability of a 'success' $\hat{p} \in (0,1)$. Note that $-(y\log(\hat{p}) + (1-y)\log(1-\hat{p}))$ is equivalent to the cross-entropy loss.
4. Consider data points $z_1, \ldots, z_n$ and density f_θ. The maximum likelihood procedure $\log\left(\prod_{i=1}^n f_\theta(z_i)\right) = \sum_{i=1}^n \log(f_\theta(z_i))$. Maximizing this quantity over $\theta \in \Theta$ is equivalent to minimizing $-\sum_{i=1}^n \log(f_\theta(z_i))$ over $\theta \in \Theta$.

25.2 Density Estimation

Let us focus on the density estimation problem with data $z_1, \ldots, z_n$. We want to predict $\hat{f}_i$ such that $\ell(\hat{f}_i(z_i)) \leq -\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\exp(-\eta\ell(f_\theta(z_i)))\right)$

Claim 25.1. *If ℓ is logarithmic loss, we choose $\eta = 1$ and $\hat{f}_i = \mathbb{E}_{\theta\sim\hat{\rho}_{i-1}} f_\theta(z_i)$.*

Proof of Claim 25.1. Then $-\log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\exp(-\eta\ell(f_\theta(z_i)))\right) = -\log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}} f_\theta(z_i)\right)$. □

From the formula for sum of mixed losses (previous lecture) where π is a prior, we have that

$$\begin{aligned}\sum_{i=1}^n -\log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}} f_\theta(z_i)\right) &= \sum_{i=1}^n -\log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\exp\left(-(-\log(f_\theta(z_i)))\right)\right) \\ &= -\log\left(\mathbb{E}_{\theta\sim\pi}\exp\left(-\sum_{i=1}^n(-\log(f_\theta(z_i)))\right)\right). \qquad (69)\end{aligned}$$

We want to find an upper bound on this quantity. This is when we apply the Donsker-Varadhan formula, which tells us that

$$-\log\mathbb{E}_{\theta\sim\pi}\exp(h(\theta)) = \sup_\rho(\mathbb{E}[h(\theta)] - KL(\rho||\pi)).$$

Thus, (69) above is less than or equal to

$$-\log\mathbb{E}_{\theta\sim\pi}\exp\left(-\sum_{i=1}^n(-\log(f_\theta(z_i)))\right) \leq \inf_\rho\left(-\sum_{i=1}^n\log(f_\theta(z_i)) + KL(\rho||\pi)\right). \qquad (70)$$

Example 1

Let $\Theta = \{f_i : i \in [m]\}$ be a finite family of distributions with $|\Theta| = m$. Let π be a uniform prior on Θ. Let ρ be the distribution centered on f^*, where

$$f^* = \underset{j\in[m]}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log\big(f_j(z_i)\big) \right\}. \tag{71}$$

Plugging (70) into (71) gives

$$-\sum_{i=1}^{n} \log\left(\mathbb{E}_{j\sim\hat{\rho}_{i-1}} f_j(z_i)\right) - \left(-\sum_{i=1}^{n} \log(f^*(z_i))\right) \le \log(m).$$

What is the exponential weights prediction?

$$\hat{\rho}_i(j) = \frac{\exp\left(-\left(-\sum_{k=1}^{i}\log\big(f_j(z_i)\big)\right)\right)^{\frac{1}{m}}}{\sum_{j=1}^{m}\exp\left(-\left(-\sum_{k=1}^{i}\log\big(f_j(z_i)\big)\right)\right)^{\frac{1}{m}}} = \frac{\prod_{k=1}^{i} f_j(z_k)}{\sum_{j=1}^{m}\prod_{k=1}^{i} f_j(z_k)}.$$

Thus,

$$\mathbb{E}_{j\sim\hat{\rho}_{i-1}} f_j(z_i) = \sum_{j=1}^{m} f_j(z_i)\cdot\frac{\prod_{k=1}^{i-1} f_j(z_k)}{\sum_{j=1}^{m}\prod_{k=1}^{i-1} f_j(z_k)}.$$

Example 2

Let $\mathcal{F} = \{f_1, ..., f_m\}$ be densities, and assume that $z_1, ..., z_n$ are sampled from $f^* \in \mathcal{F}$.

Claim 25.2. *$\exists \hat{f}$, an estimator of f^* based on $z_1, \dots, z_n$ such that $\mathbb{E}_{z_1,\dots,z_n} KL(f^*||\hat{f}) \le \frac{\log m}{n}$.*

Proof of Claim 25.2. Indeed, recall from "Online to Batch" that we choose the progressive mixture

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{j\sim\hat{\rho}_{i-1}} f_j(z).$$

□

25.3 Working with Infinite Θ (Yang-Barron Construction)

Let $\mathcal{F}$ be a collection of densities. We then have that

$$\mathcal{N}(\mathcal{F}, KL, \varepsilon) = \min\{N \in \mathbb{N} : \exists q_1, \dots, q_N \text{ such that for all } \theta \in \Theta, \exists i \in [N] \text{ such that } KL(f_\theta, q_i) \le \varepsilon^2\}.$$

We note that this definition is special, since the covering number with the KL divergence distance is defined with ε^2.
Idea: Fix $\gamma > 0$ and let N_ε be the net corresponding to $\mathcal{N}(\mathcal{F}, KL, \varepsilon)$.
Note: $\hat{f}$ is just a progressive mixture on $q_1, \dots, q_{|N_\varepsilon|}$ with the uniform prior on this set.

Proposition 25.3. *Asume $z_1, \dots, z_n \sim f^*$, with $f^* \in \mathcal{F}$. Then there exists a $\hat{f}$ which satisfies*

$$\mathbb{E}_{z_1,\dots,z_n} KL(f^*||\hat{f}) \le \inf_{\varepsilon>0}\left\{\varepsilon^2 + \frac{\log \mathcal{N}(\mathcal{F}, KL, \varepsilon)}{n}\right\}.$$

Proof of Proposition 25.3. Fix $\varepsilon > 0$. We have that

$$\mathop{\mathbb{E}}_{z_1,\dots,z_n}[KL(f^*||\hat{f})] = \mathop{\mathbb{E}}_{z_1,\dots,z_n}\left[\mathop{\mathbb{E}}_{z\sim f^*}\log\left(\frac{f^*(z)}{\hat{f}(z)}\right)\right]. \tag{72}$$

Let q^* be a density in the net such that $KL(f^*||q^*) \leq \varepsilon^2$. Then we have that (72) is equivalent to

$$\mathop{\mathbb{E}}_{z_1,\dots,z_n}\left[\mathop{\mathbb{E}}_{z\sim f^*}\left[\log\left(\frac{f^*(z)}{q^*(z)}\right) + \log\left(\frac{q^*(z)}{\hat{f}(z)}\right)\right]\right] \leq \varepsilon^2 + \mathop{\mathbb{E}}_{z_1,\dots,z_n}\left[\mathop{\mathbb{E}}_{z\sim f^*}\left[-\log\left(\hat{f}(z)\right) - (-\log(q^*(z))\right]\right]. \tag{73}$$

We will use the regret to bound the right-hand side of (73). Note that the weights on $\hat{f}_i$ only depend on $z_1, \dots, z_{i-1}$ (in other words, $\hat{p}_{i-1}$). By convexity, we know that

$$\begin{aligned}\mathop{\mathbb{E}}_{z_1,\dots,z_n}\left[\mathop{\mathbb{E}}_{z\sim f^*}\left[-\frac{1}{n}\sum_{i=1}^{n}\log\left(\hat{f}_i(z)\right) - (-\log(q^*(z))\right]\right] &= \frac{1}{n}\mathop{\mathbb{E}}_{z_1,\dots,z_n}\left[\mathop{\mathbb{E}}_{z\sim f^*}\left[-\sum_{i=1}^{n}\log\left(\hat{f}_i(z_i)\right) - (-\log(q^*(z_i))\right]\right] \\ &\leq \frac{\log \mathcal{N}(\mathcal{F}, KL, \varepsilon)}{n}.\end{aligned}$$

□

Example 3

Let $\mathcal{F} = \{N(\theta, I_d) : \theta \in \Theta\}$, where $\Theta = B_2^d$. We then observe $z_1, \dots, z_n \overset{iid}{\sim} N(\theta^*, I_d)$, with $\theta^* \in \Theta$. We want $\hat{f}$ such that $\mathop{\mathbb{E}}_{z_1,\dots,z_n} KL(N(\theta^*, I_d)||\hat{f})$ is small. Observe that

$$KL\left(N(\theta_1, I_d)||N(\theta_2, I_d)\right) = \frac{1}{2}||\theta_1 - \theta_2||_2^2,$$

thus giving us that

$$\mathcal{N}(\mathcal{F}, KL, \varepsilon) \leq (c/\varepsilon)^d$$

by a volumetric argument. From the proposition above, the progressive mixture $\hat{f}$ gives

$$\mathop{\mathbb{E}}_{z_1,\dots,z_n} KL(N(\theta^*, I_d)||\hat{f}) \leq \inf_{\varepsilon>0}\left\{\varepsilon^2 + \frac{d\log(c/\varepsilon)}{n}\right\} \leq \frac{cd\log n}{n}.$$

Lecture 25: Exponential Weights Estimator for Bounded Losses

Instructor: Nikita Zhivotovskiy *Scriber: Yanbo Feng*

1 Exponential Weights Estimator for Bounded Losses

In this lecture we will discuss exponential weights estimator under bounded loss. Bounded losses are common in classification. But we should note that the log loss isn't bounded. And the log loss is usually preferred over bounded losses due to its convexity.

Theorem 1. *Assuming $0 \le \ell(f_\theta(x), y) \le m$. Considering the mix loss, we have:*

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \ell(f_\theta(x_i), y_i)) \le \frac{nm^2\eta}{8} + \inf_{\gamma}(\mathop{\mathbb{E}}_{\theta\sim\gamma} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i) + \frac{KL(\gamma||\boldsymbol{\pi})}{\eta}).$$

Proof. Since

$$\left(-\eta\ell(f_\theta(X), Y) - \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} - \eta\ell(f_\theta(X), Y)\right) \in [-m\eta + \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \eta\ell(f_\theta(X), Y), \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \eta\ell(f_\theta(X), Y)].$$

By Hoeffding's Inequality, we have:

$$\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \exp(-\eta\ell(f_\theta(x_i), y_i)) \le \exp(-\eta \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \ell(f_\theta(x_i), y_i)) + \frac{(\eta m)^2}{8}).$$

Applying $(-\frac{1}{\eta}\log)(\cdot)$ to both side we have:

$$-\frac{1}{\eta}\log(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \exp(-\eta\ell(f_\theta(x_i), y_i))) \ge \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \ell(f_\theta(x_i), y_i)) - \frac{\eta m^2}{8}.$$

Summing them up, we get:

$$\sum_{i=1}^{n} -\frac{1}{\eta}\log(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \exp(-\eta\ell(f_\theta(x_i), y_i))) \ge \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \ell(f_\theta(x_i), y_i)) - \frac{n\eta m^2}{8}.$$

Using the formula for the sum of mis losses:

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \ell(f_\theta(x_i), y_i)) \le \frac{n\eta m^2}{8} - \frac{1}{\eta}\log(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \exp(-\eta\ell(f_\theta(x_i), y_i))).$$

By Donsker-Varadhan Variational Formula, we have:

$$\begin{aligned}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \ell(f_\theta(x_i), y_i)) &\le \frac{n\eta m^2}{8} - \frac{1}{\eta}\log(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}} \exp(-\eta\ell(f_\theta(x_i), y_i))) \\ &\le \frac{nm^2\eta}{8} + \inf_{\gamma}(\mathop{\mathbb{E}}_{\theta\sim\gamma} \sum_{i=1}^{n} \ell(f_\theta(x_i), yi) + \frac{KL(\gamma||\boldsymbol{\pi})}{\eta}).\end{aligned}$$

□

2 Example 1: Binary Classification

Now show an example for how exponential weights estimator used binary. classification.
Suppose $\mathcal{F} = \{f_1 = 1, f_2 = 0\}$:

Fact 2. *Suppose $\mathcal{F} = \{f_1 = 1, f_2 = 0\}$. Then the deterministic strategy can't guaratee regret $\ll O(n)$.*

Proof. If choosing a deterministic strategy. Then the regret will be

$$\sum_{i=1}^{n} \mathbf{1}_{\hat{y_i} \neq y_i} - \min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbf{1}_{f \neq y_i}.$$

Consider the worst case. The first term could be n since y_i can be always opposite to our prediction. And the second one is less than $\frac{n}{2}$, thus the regret $\geq \frac{n}{2}$. □

Fact 3. *Suppose $\mathcal{F} = \{f_1 = 1, f_2 = 0\}$. Exponential weights estimator can guaratee regret $\ll O(n)$.*

Proof. We use indicator loss here, so $m = 1$. Let π be the uniform distribution on $\mathcal{F}$ ($\pi(f_1) = \pi(f_2) = \frac{1}{2}$). Let γ pick the best estimator in $\mathcal{F}$. Then we have:

$$KL(\gamma || \boldsymbol{\pi}) = -\log(\frac{1}{2}) = \log 2.$$

Plugging in Theorem 1, we have:

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_{i-1}} \mathbf{1}_{(f_\theta(x_i)_i)} \leq \frac{n\eta}{8} + \frac{\log 2}{\eta} + \min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbf{1}_{(f(x_i) \neq y_i)}.$$

And

$$\frac{\log 2}{\eta} = \frac{n\eta}{8} \Rightarrow$$
$$\eta = \sqrt{\frac{8 \log 2}{n}}.$$

So,

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_{i-1}} \mathbf{1}_{(f_\theta(x_i) \neq y_i)} - \min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbf{1}_{(f(x_i) \neq y_i)} \leq \sqrt{\frac{n \log(2)}{2}}.$$

□

Remark 4. *If $\mathcal{F} = \{f_1, f_2 \cdots, f_k\}$, we still let π be the uniform distribution on $\mathcal{F}$. Then $KL(\gamma || \boldsymbol{\pi}) = \log(k)$. So similarly, we can deduce $\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_{i-1}} \mathbf{1}_{(f_\theta(x_i) \neq y_i)} - \min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbf{1}_{(f(x_i) \neq y_i)} \leq \sqrt{\frac{n \log(k)}{2}}$.*

3 Logistic Regression

Definition 5. *(Sigmoid function) $\sigma(z) = \frac{1}{1+\exp(-z)}$.*

Definition 6. *(Loss function and predictor functional class) In this example, we choose $\mathcal{F} = \{\sigma(\langle x, \theta \rangle) : \theta \in R^d\}$ as predictor functional class. And let $\ell(f_\theta(x), y) = -\log(\sigma(y\langle x, \theta \rangle))$.*

Remark 7. *It is noted that when the $\langle x, \theta \rangle$ approximate the true value y, the prediction $y\langle x, \theta \rangle$ made by the model tends to be closer to y^2. Conversely, when the parameters $\langle x, \theta \rangle$ are far from the true value y, the prediction $y\langle x, \theta \rangle$ diverges towards $-y^2$. Additionally, the term $-\log(\sigma(\cdot))$ is decreasing, implying that as the model predictions approach the actual values, the overall loss decreases. This reduction in loss is desirable as it indicates that the predictions are becoming more accurate.*

Now, using exponential weights predictor, we want to bound

$$-\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \log(\sigma(y_i\langle x_i,\theta\rangle) - \inf_{\theta\in R^d}(-\sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta\rangle))).$$

Assumption 8. *Let $\theta^* = \arg\inf_{\theta\in R^d}(-\sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta\rangle)))$. In this example, we assume $\|\theta^*\|_2 \le b$, where b is a constant.*

Lemma 9. *Let $\tilde{\sigma}(z) = -\log(\frac{1}{1+\exp(-z)})$, then $\tilde{\sigma}(z)$ is convex, $\tilde{\sigma}''(z) \le \frac{1}{4}$.*

Proposition 10.

$$-\sum_{i=1}^{n}\log(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}}\sigma(y_i\langle x_i,\theta\rangle) \le -\sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta^*\rangle) - \log(\mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}}\exp(\frac{1}{8}(\theta-\theta^*)^T(-\sum_{i=1}^{n}y_i^2x_ix_i^T)(\theta-\theta^*))).$$

Proof. Since $\tilde{\sigma}$ is convex, by Taylor's expansion, we have:

$$-\sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta\rangle) \le -\sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta^*\rangle) + 0 + \frac{1}{2!}\cdot\frac{1}{4}(\theta-\theta^*)^T(\sum_{i=1}^{n}y_i^2x_ix_i^T)(\theta-\theta^*).$$

$$\sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta\rangle) \ge \sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta^*\rangle) + 0 + \frac{1}{2!}\cdot\frac{1}{4}(\theta-\theta^*)^T(-\sum_{i=1}^{n}y_i^2x_ix_i^T)(\theta-\theta^*).$$

The first derivative is 0 since the function is convex and θ^* is the minimizer. The second derivative is less than $\frac{1}{4}$, which has been stated in above lemma.
So we have:

$$\begin{aligned}
-\sum_{i=1}^{n}\log(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}}\sigma(y_i\langle x_i,\theta\rangle) &= -\ \log(\mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}}\exp(\sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta\rangle)))) \\
&\le -\log(\mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}}\exp\left(\sum_{i=1}^{n}\log(\sigma(y_i\langle x_i,\theta^*\rangle) + 0 + \frac{1}{2!}\cdot\frac{1}{4}(\theta-\theta^*)^T(-\sum_{i=1}^{n}y_i^2x_ix_i^T)(\theta-\theta^*)\right) \\
&= -\sum_{i=1}^{n}\log(\sigma(y_i\langle x,\theta^*\rangle) - \log\mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}}\exp(\frac{1}{8}(\theta-\theta^*)^T(-\sum_{i=1}^{n}y_i^2x_ix_i^T)(\theta-\theta^*)) \\
&= -\sum_{i=1}^{n}\log(\sigma(y_i\langle x,\theta^*\rangle) - \log\mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}}\exp(\frac{1}{8}(\theta-\theta^*)^T(-\sum_{i=1}^{n}y_i^2x_ix_i^T)(\theta-\theta^*)).
\end{aligned}$$

□

Now, our problem is how to bound $-\log(\mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}}\exp((\theta-\theta^*)^T(-\sum_{i=1}^{n}y_i^2x_ix_i^T)(\theta-\theta^*)))$ in Proposition 9.

Lemma 11. *If $Q(\theta) = \theta^TA\theta + b\theta + c$, A is positive definite, then*

$$\int\exp(-Q(\theta))d\theta = \exp(-\inf_{\theta\in R^d}(Q(\theta))\cdot\frac{\pi^{\frac{d}{2}}}{\sqrt{\det(A)}}.$$

Remark 12. *This lemma offers a tool to deal with the expectation of exponential quadratic term. The proof comes from multivariate Gaussian distribution density. I skip it here.*

Applying Lemma 11 we can give a cleaner expression to Proposition 10.

Proposition 13. *Let $\boldsymbol{\pi} = N(0, \frac{1}{2}(a\pi_1)^{-1}\mathcal{I}_d)$, then*

$$-\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \log(\sigma(y_i\langle x_i,\theta\rangle) \le -\sum_{i=1}^{n} \log(\sigma(y_i\langle x_i,\theta^*\rangle) + ab^2\pi + \frac{1}{2}\log(\det(A)) - \frac{d}{2}\log(\pi) - \frac{d}{2}\log(a),$$

where $A = \frac{1}{8}\sum_{i=1}^{n} y_i^2 x_i x_i^T + a\pi\mathcal{I}_d$.

Proof. By proposition 9, we have

$$-\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \log(\sigma(y_i\langle x_i,\theta\rangle) \le -\sum_{i=1}^{n} \log(\sigma(y_i\langle x_i,\theta^*\rangle) - \log(\mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}} \exp(\frac{1}{8}(\theta-\theta^*)^T(-\sum_{i=1}^{n} y_i^2 x_i x_i^T)(\theta-\theta^*))).$$

Let $\Sigma = (\frac{1}{2}a\pi)^{-1}\mathcal{I}_d$, then $det(\Sigma) = (2a\pi)^{-d}$.

So the density function $\boldsymbol{\pi}(\theta) = \frac{1}{\sqrt{(2\pi_1)^d \det(\Sigma)}} \exp(-\frac{\theta^T\Sigma^{-1}\theta}{2}) = a^{\frac{d}{2}} \exp(-a\pi \left\|\theta\right\|_2^2)$.
Let $Q(\theta) = \frac{1}{8}(\theta-\theta^*)^T(\sum_{i=1}^{n} y_i^2 x_i x_i^T)(\theta-\theta^*) + a\pi \left\|\theta\right\|_2^2$, we first transform it into the form described in Lemma 10.

$$\begin{aligned}
Q(\theta) &= \frac{1}{8}(\theta-\theta^*)^T(\sum_{i=1}^{n} y_i^2 x_i x_i^T)(\theta-\theta^*) + a\pi \left\|\theta\right\|_2^2 \\
&= (\theta-\theta^*)^T(\frac{1}{8}\sum_{i=1}^{n} y_i^2 x_i x_i^T + a\pi\mathcal{I}_d)(\theta-\theta^*) + (\theta^*)^T(a\pi\mathcal{I}_d)(\theta^*) \\
&\ge (\theta^*)^T(a\pi\mathcal{I}_d)(\theta^*).
\end{aligned}$$

Let $A = \frac{1}{8}\sum_{i=1}^{n} y_i^2 x_i x_i^T + a\pi\mathcal{I}_d$, we have:

$$\begin{aligned}
\log(\mathop{E}_{\theta\sim\boldsymbol{\pi}} \exp(\frac{1}{8}(\theta-\theta^*)^T(-\sum_{i=1}^{n} y_i^2 x_i x_i^T)(\theta-\theta^*))) &= \log(\int \exp(\frac{1}{8}(\theta-\theta^*)^T(-\sum_{i=1}^{n} y_i^2 x_i x_i^T)(\theta-\theta^*))a^{\frac{d}{2}} \exp(-a\pi_1 \left\|\theta\right. \\
&= \log\left(a^{\frac{d}{2}} \int \exp\left(-Q(\theta)\right)d\theta\right) \\
&= \log\left(\exp(-\inf_{\theta\in R^d} Q(\theta)) \cdot \frac{\pi^{\frac{d}{2}}}{\sqrt{\det(A)}}\right) + \frac{d}{2}\log(a).
\end{aligned}$$

We know $\inf_{\theta\in R^d} Q(\theta) = a\pi \left\|\theta^*\right\|_2^2 \le ab^2\pi$. So

$$\begin{aligned}
\log\left(\exp(-\inf_{\theta\in R^d} Q(\theta)) \cdot \frac{\pi^{\frac{d}{2}}}{\sqrt{\det(A)}}\right) + \frac{d}{2}\log(a) &= -\inf_{\theta\in R^d} Q(\theta) + \log(\pi^{\frac{d}{2}}) - \log(\sqrt{\det(A)}) + \frac{d}{2}\log(a) \\
&\ge -ab^2\pi - \frac{1}{2}\log(\det(A)) + \frac{d}{2}\log(\pi) + \frac{d}{2}\log(a).
\end{aligned}$$

So

$$\begin{aligned}
-\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_{i-1}} \log(\sigma(y_i\langle x_i,\theta\rangle) &\le -\sum_{i=1}^{n} \log(\sigma(y_i\langle x_i,\theta^*\rangle) - \log(\mathop{\mathbb{E}}_{\theta\sim\boldsymbol{\pi}} \exp(\frac{1}{8}(\theta-\theta^*)^T(-\sum_{i=1}^{n} y_i^2 x_i x_i^T)(\theta-\theta^*))) \\
&\le -\sum_{i=1}^{n} \log(\sigma(y_i\langle x_i,\theta^*\rangle) + ab^2\pi + \frac{1}{2}\log(\det(A)) - \frac{d}{2}\log(\pi) - \frac{d}{2}\log(a).
\end{aligned}$$

□

Lecture 27: Logistic Regression, Exponential-Concavity

Instructor: Nikita Zhivotovskiy *Scriber: Michael Xiao* *Proofreader: Dylan Webb*

In the last lecture, we analyzed the logistic regression model. Recall the setup: We have $x_t \in \mathbb{R}^d$, with $\|x_t\|_2 \leqslant r$, and $y_t \in \{-1, 1\}$. We further assume the predictor function class of the form $\sigma(\langle x, \theta \rangle)$, where the sigmoid function $\sigma(z) = \frac{1}{1+\exp(-z)}$. Lastly, we assume that the MLE solution θ^* is bounded: $\|\theta^*\|_2 \leqslant b$. To produce the estimated coefficient, we define the loss function $l(f_\theta(x), y) = -\log(\sigma(y \langle x, \theta \rangle))$.

1 Bounds on total loss in logistic regression

Following lecture 26, we want to further develop the bound of the total log loss in logistic regression.

Proposition 1. *Under the logistic regression setting,*

$$-\sum_{i=1}^{n} \log \mathbb{E}_{\theta \sim \hat{\rho}_{i-1}} \sigma(y_i \langle x_i, \theta \rangle) \leqslant -\sum_{i=1}^{n} \log \sigma(y_i \langle x_i, \theta^* \rangle) + d + \frac{d}{2} \log\left(1 + \frac{nb^2 r^2}{8d^2}\right)$$

Proof of Proposition 1. From lecture 26, we showed that

$$\begin{aligned} -\sum_{i=1}^{n} \log \mathbb{E}_{\theta \sim \hat{\rho}_{i-1}} \sigma(y_i \langle x_i, \theta \rangle) \leqslant & -\sum_{i=1}^{n} \log \sigma(y_i \langle x_i, \theta^* \rangle) \\ & + \inf_{\theta \in \mathbb{R}^d} \left(a\pi \|\theta\|_2^2 + \frac{1}{8} (\theta - \theta^*)^\top \sum_{i=1}^{n} x_i x_i^\top (\theta - \theta^*) \right) \qquad (1) \\ & + \frac{1}{2} \log\det\left(\frac{1}{8} \sum_{i=1}^{n} x_i x_i^\top + a\pi I_d \right) - \frac{d}{2} \log(a) - \frac{d}{2} \log(\pi) \qquad (2) \end{aligned}$$

First notice that (1) $\leqslant a\pi \|\theta^*\|_2^2 \leqslant a\pi b^2$. To bound (2), we employ the manipulation $-\frac{d}{2}\log(a\pi) = -\frac{1}{2}\log\det(a\pi I_d)$. Therefore,

$$\begin{aligned} (2) &= \frac{1}{2} \log\left(\frac{\det(\frac{1}{8}\sum_{i=1}^{n} x_i x_i^\top + a\pi I_d)}{\det(a\pi I_d)} \right) \\ &= \frac{1}{2} \log \prod_{j=1}^{d} \frac{\frac{1}{8}\lambda_j + a\pi}{a\pi} \\ &= \frac{1}{2} \sum_{j=1}^{d} \log \frac{\frac{1}{8}\lambda_j + a\pi}{a\pi} \end{aligned}$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of the sample covariance matrix, $\frac{1}{8}\sum_{i=1}^{n} x_i x_i^\top \in \mathbb{R}^{d \times d}$. To proceed, we introduce the *Gram matrix* in $\mathbb{R}^{n \times n}$, whose (i, j)-th element is $\langle x_i, x_j \rangle$. The useful fact here is that, up to zero eigenvalues, the spectrum of the sample covariance matrix is the same as that of the Gram matrix. That is

$$\sum_{j=1}^{d} \lambda_j = \mathrm{Tr}(\text{Gram matrix}) = \sum_{j=1}^{n} \langle x_j, x_j \rangle = \sum_{j=1}^{n} \|x_j\|_2^2 \leqslant nr^2$$

With this fact, we have

$$(2) = \frac{1}{2}\sum_{j=1}^{d}\log\frac{\frac{1}{8}\lambda_j + a\pi}{a\pi} \leqslant \frac{1}{2}d\log\left(1+\frac{nr^2}{8da\pi}\right)$$

Combining the above bounds, we conclude

$$\begin{aligned}-\sum_{i=1}^{n}\log\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\sigma(y_i\langle x_i,\theta\rangle) &\leqslant -\sum_{i=1}^{n}\log\sigma(y_i\langle x_i,\theta^*\rangle) + a\pi b^2 + \frac{1}{2}d\log\left(1+\frac{nr^2}{8da\pi}\right)\\ &= -\sum_{i=1}^{n}\log\sigma(y_i\langle x_i,\theta^*\rangle) + d + \frac{d}{2}\log\left(1+\frac{nb^2r^2}{8d^2}\right)\end{aligned}$$

The second line follows from choosing $a = \frac{d}{\pi b^2}$. □

2 Square loss and its exponential-concavity

By construction of the loss function in logistic regression, it is somewhat natural to consider a log-exponential type bound for the loss. Namely, we would like to produce a bound in the forms of

$$\ell(\hat{f}_i(x_i), y_i) \leqslant -\frac{1}{\eta}\log\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\exp\left(-\eta\ell(f_\theta(x_i), y_i)\right) \tag{3}$$

If we assume that

$$\ell(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}f_\theta(x_i), y_i) = -\frac{1}{\eta}\log\exp\left(-\eta\ell(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}f_\theta(x_i), y_i)\right) \leqslant -\frac{1}{\eta}\log\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\exp\left(-\eta\ell(f_\theta(x_i), y_i)\right)$$

and hence let $\hat{f} = \mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}f_\theta$, we have the equivalence

$$(3) \Leftrightarrow \mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\exp\left(-\eta\ell(f_\theta(x_i), y_i)\right) \leqslant \exp\left(-\eta\ell(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}f_\theta(x_i), y_i)\right) \tag{4}$$

This gives us a framework to analyze the loss function with the following definition:

Definition 2. *A loss function ℓ is* **exponentially-concave** *with respect to $\eta > 0$ if (4) holds for all i. From another perspective, if we define γ as a distribution over θ, ℓ is exponentially-concave w.r.t. η if (4) holds for all such distributions γ.*

The function $\log(\cdot)$ is naturally exponentially-concave with $\eta = 1$. The following proposition demonstrates that the squared loss is also exponentially-concave.

Proposition 3. *Define the quadratic loss $\ell(f_\theta(x), y) = (f_\theta(x) - y)^2$. Assuming $|y| \leqslant m$ and $|f_\theta(x)| \leqslant m$, then the quadratic loss is $\frac{1}{8m^2}$-exponentially-concave.*

Proof. The main idea is that if for any y such that $|y| \leqslant m$, the function $f_y(z) = \exp(-\eta(z-y)^2)$ is concave, we have the desired result by Jensen's inequality. To verify this, we take the derivatives of f:

$$\begin{aligned}\frac{\partial f_y(z)}{\partial z} &= -2\eta(z-y)\exp\left(-\eta(z-y)^2\right)\\ \frac{\partial^2 f_y(z)}{\partial z^2} &= \underbrace{\exp\left(-\eta(z-y)^2\right)}_{(*)}\underbrace{\left(-2\eta + (2\eta(z-y))^2\right)}_{(**)}\end{aligned}$$

Notice that we always have $(*) \geqslant 0$. And $(**) < 0$ if $2\eta(z-y)^2 \leqslant 1$, with $(z-y)^2 \leqslant 4m^2$. So we can choose $\eta = \frac{1}{8m^2}$ and the result follows. □

We now show an application of exponential concavity in the context of model selection. Define the class of functions $\mathcal{F} = \{f_1, \dots, f_M\}$ as our candidate models. Assume that $|y| \leqslant m$ and $|f| \leqslant m$. We want to bound the quadratic loss of the optimal model. That is, we would like

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \leqslant \min_{f \in \mathcal{F}} \sum_{i=1}^n (f(x_i) - y_i)^2 + \text{small term}$$

This is made possible by the exponential-concavity of the quadratic loss. Define the parameter space $\Theta = \{\theta_1, \dots, \theta_M\}$, and let π be the uniform prior on Θ. If we take $\hat{y}_i = \mathbb{E}_{\theta \sim \hat{\rho}_{i-1}} f_\theta(x_i)$ and $\eta = \frac{1}{8m^2}$, then by the Donsker-Varadhan formula, we have

$$-\frac{1}{\eta} \log \mathbb{E}_{\theta \sim \pi} \exp\left(-\eta \sum_{i=1}^n (y_i - f_\theta(x_i))^2\right) \leqslant \min_{f \in \mathcal{F}} \sum_{i=1}^n (y_i - f(x_i))^2 + 8m^2 \log(M)$$

This allows us to conclude about the risk:

Corollary 4. *If (x_i, y_i) are i.i.d. samples, and $|y| \leqslant m$, $|f| \leqslant m$, then*

$$\mathbb{E}_{(x,y)} \mathcal{R}\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\theta \sim \hat{\rho}_{i-1}} f_\theta\right) \leqslant \min_{f \in \mathcal{F}} \mathcal{R}(f) + \frac{8m^2 \log(M)}{n}$$

The rough idea of the proof relies on the observation

$$\sum_{i=1}^n -\frac{1}{\eta} \log \mathbb{E}_{\theta \sim \hat{\rho}_{i-1}} \exp(-\eta l(f_\theta(x_i), y_i)) = -\frac{1}{\eta} \log \mathbb{E}_{\theta \sim \pi} \exp(-\eta \sum_{i=1}^n l(f_\theta(x_i), y_i))$$

The RHS of which we can bound by the Donsker-Varadhan formula or direct computation.

Lastly, we turn to linear regression. Consider the setup: $|y_i| \leqslant m$, $\|x_i\|_2 \leqslant r$, $\|\theta^*\| \leqslant b$, where $\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n (y_i - \langle \theta, x_i \rangle)^2$. Our goal is to make a sequence of predictions $\hat{y}_1, \dots, \hat{y}_n$ such that

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \leqslant \sum_{i=1}^n (y_i - \langle \theta^*, x_i \rangle)^2 + \text{small term}$$

Before next lecture, we provide a bit motivation for the forthcoming method. If we select the distribution

$$\hat{\rho}_{i-1} \sim \exp\left(-\eta \sum_{j=1}^{i-1} (y_i - \langle x_j, \theta \rangle)^2\right) \cdot \pi$$

where $\pi(\theta) \sim \exp(-\eta\lambda\|\theta\|^2)$, then $\hat{\rho}_{i-1}$ is Gaussian with its mean exactly equal to $\hat{\theta}_\lambda$, the *ridge regression* prediction.

Lecture 28: Sequential Linear Regression

Instructor: Nikita Zhivotovskiy *Scriber: Zora Tung* *Proofreader: Daniel Etaat*

1 Sequential Linear Regression

Given a deterministic sequence of pairs $(y_i, x_i)_{i=1}^n$, $x_i \in \mathbb{R}^d$, and $y_i \in \mathbb{R}$, we define

$$\theta^\star = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \left(y_i - \langle \theta, x_i \rangle\right)^2.$$

Further assume that, $\|x_i\| \le r$, $|y_i| \le m$, and $\|\theta^\star\|_2 \le b$, for some fixed constants $m, r, b \in \mathbb{R}$. Our aim is to create a sequence $\hat{y}_1, ..., \hat{y}_n$, where $\hat{y}_i$ depends only on the data seen before the ith round, such that

$$\sum_{i=1}^n \left(\hat{y}_i - y_i\right)^2 - \sum_{i=1}^n \left(y_i - \langle x_i, \theta^\star \rangle\right)^2,$$

i.e. such that the total regret is small. We also define the clip function:

$$\text{clip}_m(x) = \min\{m, \max\{-m, x\}\} = \begin{cases} x & |x| \le m \\ m \cdot \text{sign}(x) & \text{otherwise} \end{cases}.$$

Theorem 1 (**Vovk-Azoury-Warmuth**). *In the sequential regression setup set* $\hat{y}_i = clip_m\left(\left\langle \hat{\theta}_{i-1}, x_i \right\rangle\right)$, *where*

$$\hat{\theta}_{i-1} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{j=1}^{i-1} \left(y_j - \langle x_j, \theta \rangle\right)^2 + \lambda \|\theta\|_2^2$$

is the ridge regression predictor. Then there is a choice of λ *such that*

$$\sum_{i=1}^n \left(y_i - \hat{y}_i\right)^2 \le \sum_{i=1}^n \left(y_i - \langle x_i, \theta^\star \rangle\right)^2 + m^2 \left(d + 4d \log\left(1 + \frac{nr^2b^2}{d^2m^2}\right)\right).$$

Remark 2. *As homework, we derived a bound that scales as* $m^2 + r^2b^2$. *In general, this is less favorable than the* $m^2 \log\left(\frac{r^2b^2n}{d^2m^2}\right)$ *bound in the theorem above.*

The reason we require clipping for this estimator is to force exp-concavity of the squared loss. Recall from the previous lecture that the squared loss is $\frac{1}{8m^2}$-exp-concave if $|y| \le m$ and $|f_\theta(x)| \le m$. However, in our setup $|f_\theta(x)| = |\langle x, \theta \rangle| \le rb$.

Lemma 3. *Let* $y \in [-m, m]$ *and let* $Z \sim \mathcal{N}\left(a, \sigma^2\right)$. *Then if* $\eta = \frac{1}{8m^2}$,

$$\left(y - clip_m(a)\right)^2 \le -\frac{1}{\eta} \log\left(\mathop{\mathbb{E}}_{Z \sim \mathcal{N}(a, \sigma^2)} \left[\exp\left(-\eta\left(Z - y\right)^2\right)\right]\right).$$

Proof. Let's compute the RHS. It equals,

$$-\frac{1}{\eta} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\eta\left(z - y\right)^2 - \underbrace{\frac{\left(z - a\right)^2}{2\sigma^2}}_{\text{Gaussian pdf}}\right) dz\right) \tag{1}$$

Recall that if $Q(\theta) = \theta^\top A\theta + b\theta + c$ is a quadratic form, where A is a positive definite matrix, then

$$\int_{\mathbb{R}^d} \exp\left(-Q(\theta)\right) d\theta = \exp\left(-\inf_{\theta\in\mathbb{R}^d} Q(\theta)\right) \frac{\pi^{\frac{d}{2}}}{\sqrt{\det(A)}}. \tag{2}$$

In this case $A = \eta + \frac{1}{2\sigma^2}$ so (1) becomes,

$$-\frac{1}{\eta}\log\left(\frac{\sqrt{\pi}}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{\eta+1/2\sigma^2}}\exp\left(-\inf_z\left(\eta(z-y)^2 + \frac{(z-a)^2}{2\sigma^2}\right)\right)\right) \tag{3}$$

Solving for the inf by setting the derivative to zero, we have

$$\begin{aligned}
2\eta(z-y) + \frac{2(z-a)}{2\sigma^2} &= 0\\
z\left(\eta + \frac{1}{2\sigma^2}\right) &= \eta y + \frac{a}{2\sigma^2}\\
z &= \frac{\eta y + \frac{a}{2\sigma^2}}{\eta + \frac{1}{2\sigma^2}}\\
z &= \frac{2\sigma^2\eta y + a}{2\sigma^2\eta + 1}.
\end{aligned}$$

Plugging this back into the inf we have that,

$$\begin{aligned}
\inf_z\left(\eta(z-y)^2 + \frac{(z-a)^2}{2\sigma^2}\right) &= \eta\left(\frac{2\sigma^2\eta y + a}{2\sigma^2\eta+1} - y\right)^2 + \frac{1}{2\sigma^2}\left(\frac{2\sigma^2\eta y + a}{2\sigma^2\eta+1} - a\right)^2\\
&= \eta\left(\frac{2\sigma^2\eta y + a - 2\sigma^2\eta y - y}{2\sigma^2\eta+1}\right)^2 + \frac{1}{2\sigma^2}\left(\frac{2\sigma^2\eta y + a - 2\sigma^2\eta a - a}{2\sigma^2\eta+1}\right)^2\\
&= \eta\left(\frac{y-a}{2\sigma^2\eta+1}\right)^2 + 2\sigma^2\eta^2\left(\frac{y-a}{2\sigma^2\eta+1}\right)^2\\
&= \eta\left(2\sigma^2\eta+1\right)\left(\frac{y-a}{2\sigma^2\eta+1}\right)^2 = \frac{\eta(y-a)^2}{2\sigma^2\eta+1}.
\end{aligned}$$

Finally plugging this into (3) we get,

$$-\frac{1}{\eta}\log\left(\frac{1}{\sqrt{2\sigma^2 n+1}}\exp\left(-\frac{\eta(y-a)^2}{2\sigma^2\eta+1}\right)\right) = \frac{1}{2\eta}\log\left(2\sigma^2\eta+1\right) + \frac{(y-a)^2}{2\sigma^2\eta+1}$$

We will now compare $(y-\text{clip}_m(a))^2$ with $\frac{1}{2\eta}\log\left(2\sigma^2\eta+1\right) + \frac{(y-a)^2}{2\sigma^2\eta+1}$. We start by assuming that $a\in[-m,m]$ and we will show that,

$$(y-a)^2 \le \frac{1}{2\eta}\log\left(2\sigma^2\eta+1\right) + \frac{(y-a)^2}{2\sigma^2\eta+1}$$

This will immediately imply the desired result for $(y-\text{clip}_m(a))^2$ since allowing larger values of a will only increase the RHS of the inequality while leaving the LHS unchanged. Factoring out the $(y-a)^2$ terms, we have

$$(y-a)^2\left(1 - \frac{1}{1+2\sigma^2\eta}\right) \le \frac{1}{2\eta}\log\left(2\sigma^2\eta+1\right)$$

and by our assumptions, $y - a \in [-2m, 2m] \implies (y-a)^2 \in [0, 4m^2]$ and $\eta = \frac{1}{8m^2}$. Then we can reparameterize by $w = 2\sigma^2\eta$ and it is enough to show that,

$$\left(1 - \frac{1}{1+w}\right) \leq \log(1+w).$$

This is true analytically, see https://www.desmos.com/calculator/f5eyqdz2ze. □

Proof of Theorem 1.

Proof. We would like to use the exponential weights algorithm with the Gaussian prior:

$$\boldsymbol{\pi}(\theta) = (a\eta)^{\frac{d}{2}} \exp\left(-a\eta\pi \|\theta\|_2^2\right)$$

where a is a tuning parameter. Observe that the weights take on the form,

$$\hat{\rho}_{i-1} \propto \exp\left(-\eta \sum_{j=1}^{i-1} (y_j - \langle x_j, \theta\rangle)^2 - a\eta\pi\|\theta\|_2^2\right)$$

By construction, $\hat{\rho}_{i-1}$ is multi-variate Gaussian since $\hat{\rho}_{i-1} \propto \exp(-Q(\theta))$. Since the Gaussian density is maximized at its mean, the mean of $\hat{\rho}_{i-1}$ must be,

$$\hat{\theta}_{i-1} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left(\sum_{j=1}^{i-1} (y_j - \langle x_j, \theta\rangle)^2 + \underbrace{a\pi}_{=:\lambda} \|\theta\|_2^2\right)$$

Then by recalling (2) we must have that,

$$\hat{\rho}_{i-1} = c \exp\left(Q(\hat{\theta}_{i-1}) - Q(\theta)\right).$$

where c is a normalizing constant. Now, by our lemma we only need to show that our prediction is the clip_m of a Gaussian mean. Observe that,

$$\mathop{\mathbb{E}}_{\theta\sim\hat{\rho}_{i-1}} [f_\theta(x_i)] = \mathop{\mathbb{E}}_{\theta\sim\hat{\rho}_{i-1}} \langle\theta, x_i\rangle = \langle\hat{\theta}_{i-1}, x_i\rangle$$

Concretely, we are training a ridge regressor on the first $i-1$ points and using it to predict the label for x_i. Moreover,

$$\langle\theta, x_i\rangle \sim \mathcal{N}\left(\left\langle\hat{\theta}_{i-1}, x_i\right\rangle, \sigma_i^2\right)$$

Together with our lemma this implies that,

$$\left(y_i - \text{clip}_m\left(\langle x_i, \hat{\theta}_{i-1}\rangle\right)\right)^2 \leq -\frac{1}{\eta}\log\left(\mathop{\mathrm{E}}_{\theta\sim\hat{\rho}_{i-1}}\left[\exp\left(-\eta(y_i - \langle x_i, \theta\rangle)^2\right)\right]\right)$$

Summing both sides we get,

$$\begin{aligned}\sum_{i=1}^{n}\left(y_i - \text{clip}_m\left(\langle x_i, \hat{\theta}_{i-1}\rangle\right)\right)^2 &\leq -\frac{1}{\eta}\log\left(\mathop{\mathrm{E}}_{\theta\sim\pi}\left[\exp\left(-\eta\sum_{i=1}^{n}(y_i - \langle x_i, \theta\rangle)^2\right)\right]\right) \\ &= -\frac{1}{\eta}\log\left(\int_{\mathbb{R}^d}\exp\left(-\eta\sum_{i=1}^{n}(y_i - \langle x_i, \theta\rangle)^2 - a\eta\|\theta\|_2^2\right)(a\eta)^{\frac{d}{2}}\,d\theta\right).\end{aligned}$$

which is the now familiar integral of a exponentiated quadratic form. Once again recalling (2) and noting that $\inf_{\theta \in \mathbb{R}^d} Q(\theta) \leq Q'(\theta^\star)$ we can bound the integral above by,

$$\sum_{i=1}^{n} (y_i - \langle x_i, \theta^\star \rangle)^2 - \frac{1}{\eta} \log\left((a\eta)^{\frac{d}{2}}\right) - \frac{1}{\eta} \log\left(\pi^{\frac{d}{2}}\right) + \lambda \|\theta^\star\|_2^2 + \frac{1}{2\eta} \log\left(\det\left(\left(\sum_{i=1}^{n} x_i x_i^\top + \lambda I_d\right)\eta\right)\right)$$

Repeating the the computations for our analysis of logistic regression and fixing $\eta = \frac{1}{8m^2}$ and $\lambda = \frac{m^2 d}{b^2}$ gives us,

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leq \sum_{i=1}^{n} (y_i - \langle x_i, \theta^\star \rangle)^2 + m^2 \left(d + 4d \log\left(1 + \frac{nr^2}{d\lambda}\right)\right).$$

Therefore, since

$$m^2 \left(d + 4d \log\left(1 + \frac{n^2 r^2}{d\lambda}\right)\right) = m^2 \left(d + 4d \log\left(1 + \frac{nr^2 b^2}{d^2 m^2}\right)\right),$$

this completes the proof. □

Remark 4. *The same regret bound holds the estimator* $\hat{y}_i = \langle \hat{\theta}'_{i-1}, x_i \rangle$ *where*

$$\hat{\theta}'_{i-1} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{j=1}^{i-1} (y_j - \langle x_j, \theta \rangle)^2 + \lambda \|\theta\|_2^2 + \langle x_i, \theta \rangle^2,$$

essentially adds a point with label 0 to the training data.

STAT210B Theoretical Statistics February 12th, 2024

Homework # 1: Concentration Inequalities

Reece D. Huff

Regrades

When regrading, I only attach problems in which I did not receive 100%. If the mistake is minor, I highlight my changes in purple. If the mistake is major, I highlight the entire problem in purple. I provide the regrade justification in the gray box below the problem statement.

Notation

Let **c** and **C** represent a small and large constant, respectively (e.g., $\mathbf{c} = 10^{-5}$ and $\mathbf{C} = 10^{5}$).

Problem 0

> Familiarize yourself with the proof of the result showing the equivalent definitions for sub-exponential and sub-Gaussian random variables.
>
> *Note*: This exercise is not checked and does not have to be typed, though we expect you are familiar with the derivations.

Recall the definition of $\|\cdot\|_{\psi_2}$

$$\|X\|_{\psi_2} = \inf_t \left\{ t > 0 : \mathbb{E}\left[\exp\left\{ \left(\frac{X^2}{t^2} \right) \right\} \right] \leq 2 \right\}$$

and its properties

Property 1 $\|\cdot\|_{\psi_2}$ is a norm. Let $X, Y, \lambda \in \mathbb{R}$, then

$$\begin{aligned} \|\lambda X\|_{\psi_2} &= |\lambda| \, \|X\|_{\psi_2} \\ \|X + Y\|_{\psi_2} &\leq \|X\|_{\psi_2} + \|Y\|_{\psi_2} \\ \|X\|_{\psi_2} = 0 &\Leftrightarrow X = 0 \quad \text{(a.s.)} \end{aligned}$$

Property 2 (a.) $Z \sim \mathcal{N}(0,1) \implies \|Z\|_{\psi_2} \leq \mathbf{C}$

(b.) $Z \sim \mathcal{N}(0,\sigma^2) \implies \|Z\|_{\psi_2} \leq \sigma\mathbf{C}$

Property 3 If $X_1, \ldots, X_n$ are independent, sub-Gaussian random variables, then

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

Problem 1 ($\|\cdot\|_{\psi_2}$ is a norm)

Prove that the sub-Gaussian norm $\|\cdot\|_{\psi_2}$ indeed satisfies the properties of a norm. In particular, show that it satisfies the triangle inequality and that $\|X\|_{\psi_2} = 0$ if and only if $X = 0$ almost surely.

Recall the definition of $\|\cdot\|_{\psi_2}$

$$\|X\|_{\psi_2} := \inf_t \left\{ t > 0 : \mathbb{E}\left[\psi_2\left(\frac{|X|}{t}\right)\right] \leq 1 \right\}$$

where $\psi_2(X) = e^{X^2} - 1$.
To begin, we prove the triangle inequality

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

Proof. We define $\psi_2(X) = e^{X^2} - 1$ and consider

$$\begin{aligned}\psi_2\left(\frac{|X+Y|}{a+b}\right) &\leq \psi_2\left(\frac{|X|+|Y|}{a+b}\right) = \psi_2\left(\frac{a|X|}{a(a+b)} + \frac{b|Y|}{b(a+b)}\right) \\ &\leq \frac{a}{a+b}\psi_2\left(\frac{|X|}{a}\right) + \frac{b}{a+b}\psi_2\left(\frac{|Y|}{b}\right)\end{aligned}$$

by convexity of $\psi_2(X)$ (i.e., $f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$ for all x, y, and $t \in [0,1]$). Now, we set $a = \|X\|_{\psi_2}$ and $b = \|Y\|_{\psi_2}$ and take the expectation of both sides

$$\mathbb{E}\left[\psi_2\left(\frac{|X+Y|}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}}\right)\right] \leq \frac{\|X\|_{\psi_2}}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} \mathbb{E}\left[\psi_2\left(\frac{|X|}{\|X\|_{\psi_2}}\right)\right] + \frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} \mathbb{E}\left[\psi_2\left(\frac{|Y|}{\|Y\|_{\psi_2}}\right)\right]$$

and take the inf of both sides

$$\begin{aligned}\inf\left\{\mathbb{E}\left[\psi_2\left(\frac{|X+Y|}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}}\right)\right]\right\} \leq{}& \frac{\|X\|_{\psi_2}}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} \inf\left\{\mathbb{E}\left[\psi_2\left(\frac{|X|}{\|X\|_{\psi_2}}\right)\right]\right\} \\ &+ \frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} \inf\left\{\mathbb{E}\left[\psi_2\left(\frac{|Y|}{\|Y\|_{\psi_2}}\right)\right]\right\}\end{aligned}$$

$$\begin{aligned}\|X+Y\|_{\psi_2} &\leq \frac{\|X\|_{\psi_2}}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} \|X\|_{\psi_2} + \frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} \|Y\|_{\psi_2} \\ &= \frac{\|X\|_{\psi_2}^2 + \|Y\|_{\psi_2}^2}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} \\ &\leq \frac{\|X\|_{\psi_2}^2 + 2\|X\|_{\psi_2}\|Y\|_{\psi_2} + \|Y\|_{\psi_2}^2}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} \\ &= \frac{(\|X\|_{\psi_2} + \|Y\|_{\psi_2})^2}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}} = \|X\|_{\psi_2} + \|Y\|_{\psi_2}.\end{aligned}$$

Thus we arrive at the desired result

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

□

Next, we will prove that $\|X\|_{\psi_2} = 0$ if and only if $X = 0$ almost surely.

Proof. Starting with the definition of $\|\cdot\|_{\psi_2}$

$$\|X\|_{\psi_2} = \inf_t \left\{ t > 0 : \mathbb{E}\left[\exp\left\{ \left(\frac{X^2}{t^2} \right) \right\} \right] \leq 2 \right\}.$$

Let $Y = \exp\left\{ \left(\frac{X^2}{t^2} \right) \right\}$, and let us assume that $\mathbb{E}[Y] = 0$ and $\Pr[Y > 0] = k > 0$. Then, we have

$$0 = \mathbb{E}[Y] \geq \Pr(Y > 0)\, \mathbb{E}[Y \mid Y > 0] = k\, \mathbb{E}[Y \mid Y > 0] > 0.$$

□

Finally, we show that $\|\lambda X\|_{\psi_2} = |\lambda| \, \|X\|_{\psi_2}$.

Proof. This proof follows directly from the properties of inf, e.g., $\inf\{|\gamma| x\} = |\gamma| \inf\{x\}$ for $\gamma \in \mathbb{R}$.

$$\|\lambda X\|_{\psi_2} = \inf_t \left\{ t > 0 : \mathbb{E}\left[\exp\left\{ \left(\frac{\lambda^2 X^2}{t^2} \right) \right\} \right] \leq 2 \right\}$$

Let $k = \frac{t}{|\lambda|}$, then

$$\begin{aligned} \|\lambda X\|_{\psi_2} &= \inf_t \left\{ |\lambda| k > 0 : \mathbb{E}\left[\exp\left\{ \left(\frac{X^2}{k^2} \right) \right\} \right] \leq 2 \right\} \\ &= |\lambda| \inf_t \left\{ k > 0 : \mathbb{E}\left[\exp\left\{ \left(\frac{X^2}{k^2} \right) \right\} \right] \leq 2 \right\} \end{aligned}$$

Thus we achieve the desired result,

$$\|\lambda X\|_{\psi_2} = |\lambda| \, \|X\|_{\psi_2}.$$

□

Problem 2 (Moments are sharper than MGF)

Show that moment bounds for tail probabilities are always sharper than Chernoff-method bounds. Specifically, let X be a non-negative random variable and let $t > 0$. The best moment bound for the tail probability $\Pr(X \geq t)$ is $\inf_{q \geq 0} t^{-q} \mathbb{E} X^q$. The best Chernoff bound is $\inf_{\lambda > 0} \mathbb{E} \exp(\lambda(X - t))$. Prove that

$$\inf_{q \geq 0} \frac{\mathbb{E} X^q}{t^q} \leq \inf_{\lambda > 0} \frac{\mathbb{E} \exp(\lambda X)}{\exp(\lambda t)}.$$

Hint: Consider two positive summable sequences $\{a_i\}_{i=1}^{\infty}$ and $\{b_i\}_{i=1}^{\infty}$. Suppose $c \leq \frac{a_i}{b_i}$ for $i \in \mathbb{N}$. Then $c \leq \frac{\sum_{i=1}^{\infty} a_i}{\sum_{i=1}^{\infty} b_i}$. Moreover, if $\frac{a_i}{b_i} \neq \frac{a_j}{b_j}$ for some i, j, then $c < \frac{\sum_{i=1}^{\infty} a_i}{\sum_{i=1}^{\infty} b_i}$. Proving this might be helpful.

Recall that the moment bounds for tail probability are defined as

$$\Pr(X \geq t) \leq \inf_{q \geq 0} \frac{\mathbb{E} X^q}{t^q} \tag{moment bound}$$

and the Chernoff-method bounds are defined as

$$\Pr(X \geq t) \leq \inf_{\lambda > 0} \frac{\mathbb{E} \exp(\lambda X)}{\exp(\lambda t)}. \tag{Chernoff}$$

In order to prove that

$$\inf_{q \geq 0} \frac{\mathbb{E} X^q}{t^q} \leq \inf_{\lambda > 0} \frac{\mathbb{E} \exp(\lambda X)}{\exp(\lambda t)},$$

we first establish the following lemma.

Lemma HW1.P2.1. *Consider two positive summable sequences $\{a_i\}_{i=0}^{\infty}$ and $\{b_i\}_{i=0}^{\infty}$. Suppose $c \leq \frac{a_i}{b_i}$ for all $i \in \mathbb{N}$. Then*

$$c \leq \frac{\sum_{i=0}^{\infty} a_i}{\sum_{i=0}^{\infty} b_i}. \tag{74}$$

Proof. By assumption $c b_i \leq a_i$ for all $i \in \mathbb{N}$. Summing both sides from 1 to ∞ results in Equation (74), i.e.,

$$\sum_{i=0}^{\infty} c b_i \leq \sum_{i=0}^{\infty} a_i \implies c \leq \frac{\sum_{i=0}^{\infty} a_i}{\sum_{i=0}^{\infty} b_i}.$$

□

Now we may proceed with proving the following theorem.

Theorem HW1.P2.2 ([PN95])**.** *For a non-negative random variable X and for all $t > 0$,*

$$\inf_{q \geq 0} \frac{\mathbb{E} X^q}{t^q} \leq \inf_{\lambda > 0} \frac{\mathbb{E} \exp(\lambda X)}{\exp(\lambda t)}.$$

Proof. To begin, recall the Taylor series expansion of the exponential function (i.e., Maclaurin series)

$$\exp(kx) = \sum_{i=0}^{\infty} \frac{k^i x^i}{i!}.$$

Taking a Taylor series expansion w.r.t. to λ of the objective function of the Chernoff-method bound results in

$$\frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda t)} = \sum_{i=0}^{\infty} \frac{\lambda^i \mathbb{E}X^i}{i!} \Bigg/ \sum_{i=0}^{\infty} \frac{\lambda^i t^i}{i!}$$

Let $a_i = \sum_{i=0}^{\infty} \frac{\lambda^i \mathbb{E}X^i}{i!}$ and let $b_i = \sum_{i=0}^{\infty} \frac{\lambda^i t^i}{i!}$, then we have

$$\frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda t)} = \frac{\sum_{i=0}^{\infty} a_i}{\sum_{i=0}^{\infty} b_i}.$$

We note that

$$\frac{a_i}{b_i} = \frac{\lambda^i \mathbb{E}X^i}{i!} \frac{i!}{\lambda^i t^i} = \frac{\mathbb{E}X^i}{t^i} \geq \inf_{q \geq 0} \frac{\mathbb{E}X^q}{t^q} \quad \forall i \in \mathbb{N}.$$

We can then set c to $\inf_{q \geq 0} \frac{\mathbb{E}X^q}{t^q}$ and leverage Lemma HW1.P2.1 such that

$$c = \inf_{q \geq 0} \frac{\mathbb{E}X^q}{t^q} \leq \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda t)} = \frac{\sum_{i=0}^{\infty} a_i}{\sum_{i=0}^{\infty} b_i}$$

Note that the above inequality holds for all $\lambda > 0$ including the λ that minimizes $\mathbb{E}\exp(\lambda(X - t))$. Thus, we have

$$\inf_{q \geq 0} \frac{\mathbb{E}X^q}{t^q} \leq \inf_{\lambda > 0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda t)}$$

and achieve the desired result of the proof. □

Problem 3 (Hoeffding's lemma with the correct constant)

Let X be a zero-mean random variable within the interval $[a, b]$. Then, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\exp(\lambda X) \leq \exp\frac{\lambda^2(b-a)^2}{8}.$$

Hint: Define $L_X(\lambda) = \log\big(\mathbb{E}\exp(\lambda X)\big)$. Use the Taylor expansion of L_X with Lagrange's error bound. Then, create a random variable $Y_\lambda \in [a, b]$ whose variance equals $L_X''(\lambda)$. Use the boundedness of Y_λ and the upper bound on the variance of any bounded random variable to control $L_X''(\lambda)$.

In order to prove that

$$\mathbb{E}\exp(\lambda X) \leq \exp\frac{\lambda^2(b-a)^2}{8}.$$

we will leverage the two following lemmas.

Lemma HW1.P3.3 (Taylor's theorem)**.** *Suppose that $\mathcal{I} \subseteq \mathbb{R}$ is an closed interval and that $f : \mathcal{I} \to \mathbb{R}$. For $a \in \mathcal{I}$ and $h \in \mathbb{R}$ such that $a + h \in \mathcal{I}$, there exists some $\theta \in (0, 1)$ such that*

$$f(a+h) = f(a) + hf'(a) + \frac{h^2}{2}f''(a+\theta h).$$

Lemma HW1.P3.4 (Maximum variance)**.** *Let Z represent a random variable $Z \in [a, b]$. Then* $\mathrm{Var}[Z]$ *is upper bounded by*

$$\mathrm{Var}[Z] \leq \frac{(b-a)^2}{4}$$

Now we may proceed with proving the following theorem.

Theorem HW1.P3.5. *Let X be a zero-mean random variable within the interval $[a, b]$. Then, for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\exp(\lambda X) \leq \exp\frac{\lambda^2(b-a)^2}{8}.$$

Proof. To begin, let us define

$$L_X(\lambda) = \log\big(\mathbb{E}\exp(\lambda X)\big).$$

To begin, let us calculate the derivatives of $L_X(\lambda)$. The first derivative of $L_X(\lambda)$ with respect to λ is

$$L_X'(\lambda) = \frac{\mathbb{E}X\exp(\lambda X)}{\mathbb{E}\exp(\lambda X)}$$

and the second derivative is

$$\begin{aligned} L_X''(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\mathbb{E}X\exp(\lambda X)}{\mathbb{E}\exp(\lambda X)} &= \frac{\mathbb{E}[\exp(\lambda X)]\,\mathbb{E}[X^2\exp(\lambda X)] - \mathbb{E}[X\exp(\lambda X)]\,\mathbb{E}[X\exp(\lambda X)]}{\Big(\mathbb{E}\exp(\lambda X)\Big)^2} \\ &= \frac{\mathbb{E}[\exp(\lambda X)]\,\mathbb{E}[X^2\exp(\lambda X)] - \mathbb{E}[X\exp(\lambda X)]^2}{\Big(\mathbb{E}\exp(\lambda X)\Big)^2} \end{aligned}$$

$$L_X''(\lambda) = \frac{\mathbb{E}\left[X^2 \exp(\lambda X)\right]}{\mathbb{E}\exp(\lambda X)} - \left(\frac{\mathbb{E}\left[X \exp(\lambda X)\right]}{\mathbb{E}\exp(\lambda X)}\right)^2$$

We now define a random variable Y_λ such that $\mathrm{Var}[Y_\lambda] = L_X''(\lambda)$, i.e.,

$$Y_\lambda := \frac{X \exp(\lambda X)}{\mathbb{E}\exp(\lambda X)} \quad \text{and} \quad Y_\lambda^2 := \frac{X^2 \exp(\lambda X)}{\mathbb{E}\exp(\lambda X)} \implies \mathrm{Var}\left[Y_\lambda\right] = L_X''(\lambda).$$

Note that Y_λ is in the interval $[a, b]$

$$\Pr(Y_\lambda \in [a,b]) = \mathbb{E}\left[\mathbb{1}_{Y_\lambda \in [a,b]}\right] = \frac{\mathbb{E}\left[\mathbb{1}_{X \in [a,b]} \exp(\lambda X)\right]}{\mathbb{E}\exp(\lambda X)} = 1.$$

and therefore we may use Lemma HW1.P3.4 to bound $L_X''(\lambda)$

$$\mathrm{Var}\left[Y_\lambda\right] \le \frac{(b-a)^2}{4} \implies L_X''(\lambda) \le \frac{(b-a)^2}{4}.$$

We have by Lemma HW1.P3.3 with $f = L_X$, $a = 0$, $h = \lambda$, there is a $\theta \in (0,1)$ such that

$$\begin{aligned} L_X(\lambda) &= L_X(0) + \lambda L_X'(0) + \frac{\lambda^2}{2} L_X''(\theta\lambda) \\ &= \log(1) + \lambda \frac{\mathbb{E}X}{1} + \frac{\lambda^2}{2} L_X''(\theta\lambda) \le \frac{\lambda^2 (b-a)^2}{8} \end{aligned}$$

since $L_X''(\lambda) \le \frac{(b-a)^2}{4}$ for all λ. Thus, we conclude that

$$\begin{aligned} L_X(\lambda) = \log\big(\mathbb{E}\exp(\lambda X)\big) &\le \frac{\lambda^2 (b-a)^2}{8} \\ \mathbb{E}\exp(\lambda X) &\le \exp \frac{\lambda^2 (b-a)^2}{8}. \end{aligned}$$

□

Problem 4 (Binomial concentration with sharp constants)

In this problem, we aim to obtain sharper bounds for Chernoff bounds than those derived from Hoeffding and Bernstein inequalities for the binomial distribution. First, for two probability distributions on k elements $P = (p_1, p_2, ..., p_k)$ and $Q = (q_1, q_2, ..., q_k)$ (i.e., $p_i, q_i \in [0, 1]$ for all $i \in [k]$ and $\sum_{i=1}^k p_i = \sum_{i=1}^k q_i = 1$), define the *Kullback-Leibler* divergence as follows (log is the logarithm with base e):

$$\mathrm{KL}(P, Q) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i}.$$

Consider X to be a binomial random variable with parameters n and $p \in [0, 1]$. Prove the following:

1. For any $t \in [0, 1 - p]$, it holds that

$$\Pr(X \geq \mathbb{E}X + tn) \leq \exp(-n\mathrm{KL}((p + t, 1 - p - t), (p, 1 - p))),$$

   and for any $t \in [0, p]$,

$$\Pr(X \leq \mathbb{E}X - tn) \leq \exp(-n\mathrm{KL}((p - t, 1 - p + t), (p, 1 - p))),$$

2. For any $\delta \geq 0$, we have

$$\Pr(X \geq (1 + \delta)\mathbb{E}X) \leq \left(\frac{e^{\delta}}{(1 + \delta)^{1+\delta}}\right)^{\mathbb{E}X},$$

   and for $\delta \in (0, 1)$,

$$\Pr(X \leq (1 - \delta)\mathbb{E}X) \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}}\right)^{\mathbb{E}X}.$$

   Demonstrate that the bounds can be further simplified for the same values of δ. That is, show that

$$\left(\frac{e^{\delta}}{(1 + \delta)^{1+\delta}}\right)^{\mathbb{E}X} \leq \exp\left(\frac{-\delta^2 \mathbb{E}X}{2 + \delta}\right), \quad \text{and} \quad \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}}\right)^{\mathbb{E}X} \leq \exp\left(\frac{-\delta^2 \mathbb{E}X}{2}\right).$$

   These inequalities, when δ is a small but fixed constant, are often referred to as the *multiplicative Chernoff bounds*, as they compare the random variable to multiple times its expectation.

Hint: Apply Chernoff's method and a precise computation of the MGF for Bernoulli random variables. In 2.), simplify the formulas in 1.) by combining the exact computation for the KL divergence with the inequality $(1 - x) \leq \exp(-x)$ for $x \geq 0$. For the last inequalities, it may be useful to first show that $\frac{\delta}{2+\delta} \leq \log(1 + \delta)$ for any $\delta \geq 0$ and that $-\delta + \delta^2/2 \leq (1 - \delta)\log(1 - \delta)$ for $\delta \in (0, 1)$.

Recall the Chernoff's method

$$\Pr(X \geq k) \leq \inf_{\lambda > 0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda k)},$$

various properties of the Bernoulli random variables

$$\mathbb{E}X = p \quad \text{and} \quad \text{Var}[X] = pq = p(1-p) \quad \text{and} \quad \text{MGF}_{\text{Ber}}(\lambda) = \mathbb{E}\exp(\lambda X) = q + pe^{\lambda} = 1 - p + pe^{\lambda},$$

and various properties of random variables following the binomial distribution

$$\begin{aligned}
\mathbb{E}X = np, \qquad \text{Var}[X] &= npq = np(1-p) \\
\text{MGF}_{\text{bin}}(\lambda) = \mathbb{E}\exp(\lambda X) &= (q + pe^{\lambda})^n = (1 - p + pe^{\lambda})^n, \\
\frac{\partial}{\partial \lambda}\text{MGF}_{\text{bin}}(\lambda) &= npe^{\lambda}(q + pe^{\lambda})^{n-1} = npe^{\lambda}(1 - p + pe^{\lambda})^{n-1}
\end{aligned}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

Theorem HW1.P4.6. *For any $t \in [0, 1-p]$, it holds that*

$$\Pr(X \geq \mathbb{E}X + tn) \leq \exp(-n\text{KL}((p+t, 1-p-t), (p, 1-p))).$$

Proof. We begin by utilizing Chernoff's method with $k = \mathbb{E}X + tn$. Note that the mean of the binomial distribution is np, so $k = np + nt = n(p+t)$. Additionally, we utilize the MGF_{bin} to arrive at

$$\Pr(X \geq \mathbb{E}X + tn) \leq \inf_{\lambda > 0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p+t))} = \inf_{\lambda > 0} \frac{(1 - p + pe^{\lambda})^n}{\exp(\lambda n(p+t))}$$

for $t \in [0, 1-p]$. Next, we aim to solve for $\lambda^\star$ that minimizes the objective function. That is,

$$\lambda^\star := \operatorname*{arginf}_{\lambda > 0} \frac{(1 - p + pe^{\lambda})^n}{\exp(\lambda n(p+t))}$$

Therefore, we calculate the partial derivative of the objective with respect to λ,

$$\begin{aligned}
\frac{\partial}{\partial \lambda}\left\{\frac{(1 - p + pe^{\lambda})^n}{\exp(\lambda n(p+t))}\right\} &= \frac{(1 - p + pe^{\lambda})^n n(p+t)\exp(\lambda n(p+t)) - npe^{\lambda}(1 - p + pe^{\lambda})^{n-1}\exp(\lambda n(p+t))}{\left(\exp(\lambda n(p+t))\right)^2} \\
&= \frac{n(p+t)(1 - p + pe^{\lambda})^n - npe^{\lambda}(1 - p + pe^{\lambda})^{n-1}}{\exp(\lambda n(p+t))}
\end{aligned}$$

and set the resulting partial derivative to 0 to solve for $\lambda^\star$

$$\begin{aligned}
n(p+t)(1 - p + pe^{\lambda})^n &= npe^{\lambda}(1 - p + pe^{\lambda})^{n-1} \\
(p+t)(1 - p + pe^{\lambda})^n &= pe^{\lambda}(1 - p + pe^{\lambda})^{n-1} \\
\frac{pe^{\lambda}}{p+t} &= \frac{(1 - p + pe^{\lambda})^n}{(1 - p + pe^{\lambda})^{n-1}} \\
\frac{pe^{\lambda}}{p+t} &= (1 - p + pe^{\lambda})^{n-(n-1)} \\
\frac{pe^{\lambda}}{p+t} &= 1 - p + pe^{\lambda} \\
pe^{\lambda} &= (p+t)(1 - p + pe^{\lambda}) \\
pe^{\lambda} &= (p+t)(1-p) + p(p+t)e^{\lambda}
\end{aligned}$$

$$
\begin{aligned}
pe^{\lambda} - p(p+t)e^{\lambda} &= (p+t)(1-p) \\
[p - p(p+t)]e^{\lambda} &= (p+t)(1-p) \\
e^{\lambda^\star} &= \frac{(p+t)(1-p)}{p(1-p-t)} = \frac{t+p-p^2-pt}{p-p^2-pt} = \frac{t}{p(1-p-t)} + 1 \\
\lambda^\star &= \log\left(\frac{(p+t)(1-p)}{p(1-p-t)}\right)
\end{aligned}
$$

Substituting $\lambda^\star$ back into the objective function results in

$$
\begin{aligned}
\inf_{\lambda>0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p+t))} &= \left.\frac{(1-p+pe^{\lambda})^n}{\exp(\lambda n(p+t))}\right|_{\lambda=\lambda^\star} \\
&= \left(1-p+pe^{\lambda^\star}\right)^n \Big/ \exp\left(\lambda^\star n(p+t)\right) \\
&= \left(1-p+p\left[\frac{t}{p(1-p-t)}+1\right]\right)^n \Big/ \exp\left(n(p+t)\log\left(\frac{(p+t)(1-p)}{p(1-p-t)}\right)\right) \\
&= \left(1+\frac{t}{1-p-t}\right)^n \Big/ \left(\frac{(p+t)(1-p)}{p(1-p-t)}\right)^{n(p+t)} \\
&= \left(\frac{1-p-t+t}{1-p-t}\right)^n \times \Big((p+t)(1-p)\Big)^{-n(p+t)} \times \Big(p(1-p-t)\Big)^{n(p+t)} \\
&= \left(\frac{1-p}{1-p-t}\right)^n \times (p+t)^{-n(p+t)} \times (1-p)^{-n(p+t)} \times p^{n(p+t)} \times (1-p-t)^{n(p+t)} \\
&= (1-p)^n \times (1-p-t)^{-n} \times (p+t)^{-n(p+t)} \times (1-p)^{-n(p+t)} \times p^{n(p+t)} \times (1-p-t)^{n(p+t)} \\
&= (1-p)^{n-n(p+t)} \times (1-p-t)^{-n+n(p+t)} \times (p+t)^{-n(p+t)} \times p^{n(p+t)} \\
&= (1-p)^{n(1-p-t)} \times (1-p-t)^{-n(1-p-t)} \times (p+t)^{-n(p+t)} \times p^{n(p+t)} \\
&= \left(\frac{1-p}{1-p-t}\right)^{n(1-p-t)} \times \left(\frac{p}{p+t}\right)^{n(p+t)} \\
&= \left(\frac{1-p-t}{1-p}\right)^{-n(1-p-t)} \left(\frac{p+t}{p}\right)^{-n(p+t)}
\end{aligned}
$$

Taking the exp log

$$
\begin{aligned}
&= \exp\left[-n(p+t)\log\left(\frac{p+t}{p}\right) - n(1-p-t)\log\left(\frac{1-p-t}{1-p}\right)\right] \\
&= \exp\left[-n\left[(p+t)\log\left(\frac{p+t}{p}\right) + (1-p-t)\log\left(\frac{1-p-t}{1-p}\right)\right]\right] \\
\inf_{\lambda>0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p+t))} &= \exp\Big(-n\mathrm{KL}((p+t, 1-p-t), (p, 1-p))\Big).
\end{aligned}
$$

Thus, we have arrived at the desired result,

$$
\Pr(X \geq \mathbb{E}X + tn) \leq \inf_{\lambda>0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p+t))} = \exp\Big(-n\mathrm{KL}((p+t, 1-p-t), (p, 1-p))\Big)
$$

for $t \in [0, 1-p]$. □

Recall the Chernoff's method for the left tail

$$\Pr(X \le k) \le \inf_{\lambda<0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda k)}.$$

Theorem HW1.P4.7. *For any $t \in [0, p]$, it holds that*

$$\Pr(X \le \mathbb{E}X - tn) \le \exp(-n\mathrm{KL}((p-t, 1-p+t), (p, 1-p))).$$

Proof. We begin our proof by utilizing Chernoff's method with $k = \mathbb{E}X - tn$. Note that the mean of the binomial distribution is np, so $k = np - nt = n(p-t)$. Additionally, we utilize the $\mathrm{MGF}_{\mathrm{bin}}$ to arrive at

$$\Pr(X \ge \mathbb{E}X + tn) \le \inf_{\lambda<0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p-t))} = \inf_{\lambda<0} \frac{(1-p+pe^{\lambda})^n}{\exp(\lambda n(p-t))}$$

for $t \in [0, p]$. Next, we aim to solve for $\lambda^\star$ that minimizes the objective function. That is,

$$\lambda^\star := \operatorname*{arginf}_{\lambda<0} \frac{(1-p+pe^{\lambda})^n}{\exp(\lambda n(p-t))}$$

Therefore, we calculate the partial derivative of the objective with respect to λ,

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\left\{\frac{(1-p+pe^{\lambda})^n}{\exp(\lambda n(p-t))}\right\} &= \frac{(1-p+pe^{\lambda})^n n(p-t)\exp(\lambda n(p-t)) - npe^{\lambda}(1-p+pe^{\lambda})^{n-1}\exp(\lambda n(p-t))}{\left(\exp(\lambda n(p-t))\right)^2} \\
&= \frac{n(p-t)(1-p+pe^{\lambda})^n - npe^{\lambda}(1-p+pe^{\lambda})^{n-1}}{\exp(\lambda n(p-t))}
\end{aligned}$$

and set the resulting partial derivative to 0 to solve for $\lambda^\star$

$$\begin{aligned}
n(p-t)(1-p+pe^{\lambda})^n &= npe^{\lambda}(1-p+pe^{\lambda})^{n-1} \\
&\vdots \\
e^{\lambda^\star} &= \frac{(p-t)(1-p)}{p(1-p+t)} = \frac{-t+p-p^2+pt}{p-p^2+pt} = \frac{-t}{p(1-p+t)} + 1 \\
\lambda^\star &= \log\left(\frac{(p-t)(1-p)}{p(1-p+t)}\right)
\end{aligned}$$

Substituting $\lambda^\star$ back into the objective function results in

$$\begin{aligned}
\inf_{\lambda<0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p-t))} &= \left.\frac{(1-p+pe^{\lambda})^n}{\exp(\lambda n(p-t))}\right|_{\lambda=\lambda^\star} \\
&= \left(1-p+pe^{\lambda^\star}\right)^n \Big/ \exp\left(\lambda^\star n(p-t)\right) \\
&= \left(1-p+p\left[\frac{-t}{p(1-p+t)}+1\right]\right)^n \Big/ \exp\left(n(p-t)\log\left(\frac{(p-t)(1-p)}{p(1-p+t)}\right)\right) \\
&\vdots \\
&= \left(\frac{1-p+t}{1-p}\right)^{-n(1-p+t)} \left(\frac{p-t}{p}\right)^{-n(p-t)}
\end{aligned}$$

Taking the exp log

$$= \exp\left[-n\left[(p-t)\log\left(\frac{p-t}{p}\right) + (1-p+t)\log\left(\frac{1-p+t}{1-p}\right)\right]\right]$$

$$\inf_{\lambda<0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p-t))} = \exp\Big(-n\mathrm{KL}((p-t, 1-p+t), (p, 1-p))\Big).$$

Thus, we have arrived at the desired result,

$$\Pr(X \leq \mathbb{E}X - tn) \leq \inf_{\lambda<0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p-t))} = \exp\Big(-n\mathrm{KL}((p-t, 1-p+t), (p, 1-p))\Big)$$

for $t \in [0, p]$. □

Theorem HW1.P4.8. *For any $\delta \geq 0$, we have*

$$\Pr(X \geq (1+\delta)\mathbb{E}X) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}X}.$$

Proof. Recall our earlier result,

$$\begin{aligned}\Pr(X \geq \mathbb{E}X + tn) \leq \inf_{\lambda>0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p+t))} &= \exp\Big(-n\mathrm{KL}((p+t, 1-p-t), (p, 1-p))\Big) \\ &= \exp\left[-n\left[(p+t)\log\left(\frac{p+t}{p}\right) + (1-p-t)\log\left(\frac{1-p-t}{1-p}\right)\right]\right]\end{aligned}$$

for $t \in [0, 1-p]$. Note that

$$\mathbb{E}X + tn = np + nt = (1 + t/p)np = (1+t/p)\mathbb{E}X$$

therefore we can set $\delta = t/p$. In other words, $t = \delta p$ and we have

$$\begin{aligned}\Pr(X \geq (1+\delta)\mathbb{E}X) &\leq \exp\left[-n\left[(p+\delta p)\log\left(\frac{p+\delta p}{p}\right) + (1-p-\delta p)\log\left(\frac{1-p-\delta p}{1-p}\right)\right]\right] \\ &= (1+\delta)^{-n(p+\delta p)} \times \left(\frac{1-p}{1-p-\delta p}\right)^{n(1-p-\delta p)} \\ &= \frac{1}{(1+\delta)^{(1+\delta)np}} \times \left(\frac{1-p-\delta p+\delta p}{1-p-\delta p}\right)^{n(1-p-\delta p)} \\ &= \frac{1}{(1+\delta)^{(1+\delta)np}} \times \left(1 + \frac{\delta p}{1-p-\delta p}\right)^{n(1-p-\delta p)} \\ &\leq \frac{1}{(1+\delta)^{(1+\delta)np}} \times \left(\exp\left(\frac{\delta p}{1-p-\delta p}\right)\right)^{n(1-p-\delta p)} \\ &= \frac{1}{(1+\delta)^{(1+\delta)np}} \times \exp(n\delta p)\end{aligned}$$

$$\boxed{\Pr(X \geq (1+\delta)\mathbb{E}X) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}X}.}$$

□

Theorem HW1.P4.9. *For $\delta \in (0,1)$,*

$$\Pr(X \le (1-\delta)\mathbb{E}X) \le \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mathbb{E}X}.$$

Proof. Recall our previous result

$$\begin{aligned}\inf_{\lambda<0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p-t))} &= \exp\Big(-n\mathrm{KL}((p-t,1-p+t),(p,1-p))\Big)\\ &= \exp\left[-n\left[(p-t)\log\left(\frac{p-t}{p}\right)+(1-p+t)\log\left(\frac{1-p+t}{1-p}\right)\right]\right]\end{aligned}$$

Thus, we have arrived at the desired result,

$$\Pr(X \le \mathbb{E}X - tn) \le \inf_{\lambda<0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda n(p-t))} = \exp\Big(-n\mathrm{KL}((p-t,1-p+t),(p,1-p))\Big)$$

for $t \in [0,p]$. Note that

$$\mathbb{E}X - tn = np - nt = (1-t/p)np = (1-t/p)\mathbb{E}X$$

therefore we can set $\delta = t/p$. In other words, $t = \delta p$ and we have

$$\begin{aligned}\Pr(X \le (1-\delta)\mathbb{E}X) &\le \exp\left[-n\left[(p-\delta p)\log\left(\frac{p-\delta p}{p}\right)+(1-p+\delta p)\log\left(\frac{1-p+\delta p}{1-p}\right)\right]\right]\\ &= (1-\delta)^{-n(p-\delta p)} \times \left(\frac{1-p}{1-p+\delta p}\right)^{n(1-p+\delta p)}\\ &= \frac{1}{(1-\delta)^{(1-\delta)np}} \times \left(\frac{1-p+\delta p-\delta p}{1-p+\delta p}\right)^{n(1-p+\delta p)}\\ &= \frac{1}{(1-\delta)^{(1-\delta)np}} \times \left(1-\frac{-\delta p}{1-p+\delta p}\right)^{n(1-p+\delta p)}\\ &\le \frac{1}{(1-\delta)^{(1-\delta)np}} \times \left(\exp\left(\frac{-\delta p}{1-p+\delta p}\right)\right)^{n(1-p+\delta p)}\\ &= \frac{1}{(1-\delta)^{(1-\delta)np}} \times \exp(-n\delta p)\end{aligned}$$

$$\boxed{\Pr(X \le (1-\delta)\mathbb{E}X) \le \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mathbb{E}X}.}$$

□

In order to prove that for any $\delta \geq 0$, we have

$$\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}X} \leq \exp\left(\frac{-\delta^2 \mathbb{E}X}{2+\delta}\right)$$

we must first prove the following Lemma.

Lemma HW1.P4.10. *For any $\delta \geq 0$,*

$$\frac{2\delta}{2+\delta} \leq \log(1+\delta).$$

Proof. Starting with

$$f(\delta) := \log(1+\delta) - \frac{2\delta}{2+\delta}$$

$$f'(\delta) = \frac{1}{1+\delta} - \frac{2(2+\delta)-2\delta}{(2+\delta)^2} = \frac{1}{1+\delta} - \frac{4}{(2+\delta)^2} = \frac{(2+\delta)^2 - 4(1+\delta)}{(1+\delta)(2+\delta)^2} = \frac{\delta^2}{(1+\delta)(2+\delta)^2}$$

We see that $f'(\delta) \geq 0$, meaning that $f(\lambda)$ is *strictly increasing* for all $\delta \geq 0$. Taken with $f(0) = 0$, we can conclude that $f(\lambda) \geq 0$ for all $\delta \geq 0$. Thus,

$$f(\delta) = \log(1+\delta) - \frac{2\delta}{2+\delta} \geq 0 \implies \frac{2\delta}{2+\delta} \leq \log(1+\delta) \quad \forall \delta \geq 0.$$

proving our lemma. □

Now we can proceed to the main Theorem.

Theorem HW1.P4.11. *For any $\delta \geq 0$, we have*

$$\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}X} \leq \exp\left(\frac{-\delta^2 \mathbb{E}X}{2+\delta}\right).$$

Proof. Starting with

$$\begin{aligned}
\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}X} &= \exp\left(\mathbb{E}X \log\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)\right) \\
&= \exp\left[\mathbb{E}X\left[\delta - (1+\delta)\log(1+\delta)\right]\right] \\
&= \frac{\exp[\mathbb{E}X\delta]}{\exp\left[\mathbb{E}X(1+\delta)\log(1+\delta)\right]}
\end{aligned}$$

Note that $\exp\big(-k\log(1+\delta)\big) \leq \exp(-k\delta/(2+\delta))$

$$\leq \exp\left[\mathbb{E}X\left[\delta - (1+\delta)\frac{\delta}{2+\delta}\right]\right]$$

Focusing on term $\delta - (1+\delta)\frac{2\delta}{2+\delta}$,

$$\delta - (1+\delta)\frac{2\delta}{2+\delta} = \delta - \frac{2\delta^2+2\delta}{2+\delta} = \frac{\delta(2+\delta) - 2\delta^2 - 2\delta}{2+\delta} = \frac{-\delta^2}{2+\delta}$$

Thus we arrive at the desired result:

$$\boxed{\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}X} \leq \exp\left(\frac{-\delta^2 \mathbb{E}X}{2+\delta}\right)}$$

□

In order to prove that for $\delta \in (0,1)$, we have

$$\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mathbb{E}X} \leq \exp\left(\frac{-\delta^2 \mathbb{E}X}{2}\right).$$

we must first prove the following Lemma.

Lemma HW1.P4.12. *For $\delta \in (0,1)$,*

$$-\delta + \delta^2/2 \leq (1-\delta)\log(1-\delta).$$

Proof. Starting with

$$f(\delta) := (1-\delta)\log(1-\delta) + \delta - \frac{\delta^2}{2}$$
$$f'(\delta) = -1\log(1-\delta) + (1-\delta)\frac{-1}{1-\delta} + 1 - \delta = -\log(1-\delta) - \delta$$

We note that $f'(\delta) \geq 0$, meaning that $f(\lambda)$ is *strictly increasing* for all $\delta \in (0,1)$. Taken with $f(0) = 0$, we can conclude that $f(\lambda) \geq 0$ for all $\delta \in (0,1)$. Thus,

$$f(\delta) = (1-\delta)\log(1-\delta) + \delta - \frac{\delta^2}{2} \geq 0 \implies -\delta + \delta^2/2 \leq (1-\delta)\log(1-\delta) \quad \forall \delta \in (0,1).$$

proving our lemma. □

Now we can proceed to the main Theorem.

Theorem HW1.P4.13. *For $\delta \in (0,1)$, we have*

$$\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mathbb{E}X} \leq \exp\left(\frac{-\delta^2 \mathbb{E}X}{2}\right).$$

Proof. Starting with

$$\begin{aligned}\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mathbb{E}X} &= \exp\left(\mathbb{E}X \log\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)\right) \\ &= \exp\left[\mathbb{E}X\left[-\delta - (1-\delta)\log(1-\delta)\right]\right]\end{aligned}$$

Note that $-(1-\delta)\log(1+\delta)) \leq \delta - \delta^2/2$

$$\leq \exp\left[\mathbb{E}X\left[-\delta + \delta - \delta^2/2\right]\right]$$

Thus we arrive at the desired result:

$$\boxed{\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mathbb{E}X} \leq \exp\left(\frac{-\delta^2 \mathbb{E}X}{2}\right).}$$

□

Problem 5 (Sample mean of heavy-tailed random variables)

Assume that X is a random variable for which $\mathbb{E}|X - \mathbb{E}X|^{1+\epsilon} \leq \sigma$, where $\epsilon \in (0,1)$. Note that we do not assume the existence of the variance for this random variable. Let $X_1, ..., X_n$ be independent copies of X. Show that for any $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \leq \mathbb{E}X + \left(\frac{3\sigma}{\delta n^{\epsilon}}\right)^{\frac{1}{1+\epsilon}}.$$

In your proof, a constant different from 3 might emerge.

Hint: Utilize the following decomposition. For any $a, t \geq 0$,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq t\right) \leq \Pr\left(\exists i \in [n] : |X_i - \mathbb{E}X| > a\right) + \Pr\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X)\mathbb{1}_{|X_i - \mathbb{E}X| \leq a} \geq t\right).$$

Apply the union bound to analyze the first term and Chebyshev's inequality for the second term.

Regrade justification: The original solution was incorrect because I made the incorrect assumption that $\mathbb{E}\left[\sum(|Xi - \mathbb{E}[X]|)^{1+\epsilon}\right] \leq \sum(\mathbb{E}\left[|Xi - \mathbb{E}[X]|^{1+\epsilon}\right])$. This led to an incorrect bound for the first term in the decomposition. The solution that follows is a corrected version of the original solution.

Proof. To begin, we apply the decomposition provided in the hint. For any $a, t \geq 0$, we have

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq t\right) \leq \underbrace{\Pr\left(\exists i \in [n] : |X_i - \mathbb{E}X| > a\right)}_{\text{Term 1}} + \underbrace{\Pr\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X)\mathbb{1}_{|X_i - \mathbb{E}X| \leq a} \geq t\right)}_{\text{Term 2}}.$$

We begin by analyzing Term 1. Applying the union bound, we have

$$\Pr\left(\exists i \in [n] : |X_i - \mathbb{E}X| > a\right) \leq \sum_{i=1}^{n} \Pr\left(|X_i - \mathbb{E}X| > a\right).$$

We will bound this probability by using a more general form of Chebyshev's inequality, i.e.,

$$\Pr(|X - \mathbb{E}X| \geq t) = \Pr(|X - \mathbb{E}X|^p \geq t^p) \leq \frac{\mathbb{E}|X - \mathbb{E}X|^p}{t^p}$$

which holds for any $t > 0$ and $p > 0$. We apply this inequality with $p = 1 + \epsilon$ and $t = a$ to get

$$\sum_{i=1}^{n} \Pr\left(|X_i - \mathbb{E}X| > a\right) \leq \sum_{i=1}^{n} \frac{\mathbb{E}|X_i - \mathbb{E}X|^{1+\epsilon}}{a^{1+\epsilon}} \leq \frac{n\sigma}{a^{1+\epsilon}}. \tag{75}$$

Next, we analyze Term 2. First, we note that for any independent and identically distributed random variables $A_1, A_2, \ldots, A_n$, we have[4],

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} A_i\right)^2\right] \leq \frac{\mathbb{E}\left[A_1^2\right]}{n} + \mathbb{E}\left[A_1\right]^2.$$

Then we define the random variables $Y_i = (X_i - \mathbb{E}X)$ and apply the above inequality to get

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{|Y_i|\leq a}\right)^2\right] \leq \underbrace{\frac{\mathbb{E}\left[(Y_1 \mathbb{1}_{|Y_1|\leq a})^2\right]}{n}}_{\text{Term A}} + \underbrace{\mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|\leq a}\right]^2}_{\text{Term B}}$$

We will analyze Term A and Term B separately. Starting with the numerator of Term A, we have

$$\mathbb{E}\left[(Y_1 \mathbb{1}_{|Y_1|\leq a})^2\right] = \mathbb{E}\left[(Y_1 \mathbb{1}_{|Y_1|\leq a})^{1+\epsilon+1-\epsilon}\right] = \mathbb{E}\left[(Y_1 \mathbb{1}_{|Y_1|\leq a})^{1+\epsilon}(Y_1 \mathbb{1}_{|Y_1|\leq a})^{1-\epsilon}\right].$$

Since $(Y_1 \mathbb{1}_{|Y_1|\leq a})^{1+\epsilon} \leq \left|Y_1 \mathbb{1}_{|Y_1|\leq a}\right|^{1+\epsilon}$ and $Y_1 \mathbb{1}_{|Y_1|\leq a}$ is bounded by a, we have that

$$\mathbb{E}\left[(Y_1 \mathbb{1}_{|Y_1|\leq a})^2\right] \leq \mathbb{E}\left[\left|Y_1 \mathbb{1}_{|Y_1|\leq a}\right|^{1+\epsilon} a^{1-\epsilon}\right] = a^{1-\epsilon}\mathbb{E}\left[\left|Y_1 \mathbb{1}_{|Y_1|\leq a}\right|^{1+\epsilon}\right] \leq \sigma a^{1-\epsilon}$$

implying that Term A is bounded by $\frac{\sigma a^{1-\epsilon}}{n}$. Next, we analyze Term B. To begin we note that

$$\mathbb{E}\left[Y_1\right] = \mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|\leq a}\right] + \mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|> a}\right] = 0 \implies \mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|\leq a}\right] = -\mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|> a}\right].$$

Thus, Term B may be written as

$$\mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|\leq a}\right]^2 = \mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|> a}\right]^2.$$

Bounding this term is more complicated and require Hölder's inequality. We have

$$\mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|> a}\right]^2 \leq \mathbb{E}\left[\left|Y_1 \mathbb{1}_{|Y_1|> a}\right|\right]^2 \leq \mathbb{E}\left[|Y_1|^p\right]^{2/p} \mathbb{E}\left[\left|\mathbb{1}_{|Y_1|> a}\right|^q\right]^{2/q}.$$

By setting $p = {}^{\epsilon+1}/_{1}$ and $q = {}^{\epsilon+1}/_{\epsilon}$, we have

$$\mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|> a}\right]^2 \leq \mathbb{E}\left[|Y_1|^{1+\epsilon}\right]^{\frac{2}{1+\epsilon}} \mathbb{E}\left[\mathbb{1}_{|Y_1|> a}\right]^{\frac{2\epsilon}{\epsilon+1}} \leq \sigma^{\frac{2}{1+\epsilon}} \Pr\left(|Y_1| > a\right)^{\frac{2\epsilon}{\epsilon+1}}.$$

By Equation (75), we have that $\Pr\left(|Y_1| > a\right) \leq \frac{\sigma}{a^{1+\epsilon}}$. Thus, we have that

$$\mathbb{E}\left[Y_1 \mathbb{1}_{|Y_1|\leq a}\right]^2 \leq \sigma^{\frac{2}{1+\epsilon}} \left(\frac{\sigma}{a^{1+\epsilon}}\right)^{\frac{2\epsilon}{\epsilon+1}} = \frac{\sigma^{\frac{2}{1+\epsilon}}\sigma^{\frac{2\epsilon}{1+\epsilon}}}{a^{2\epsilon}} = \frac{\sigma^2}{a^{2\epsilon}}.$$

[4] *Proof.* Let $A_1, A_2, \ldots, A_n$ be iid random variables. Then

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} A_i\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n} A_i^2 + \sum_{i\neq j} A_i A_j\right] = \frac{1}{n^2}\left(n\mathbb{E}\left[A_1^2\right] + n(n-1)\mathbb{E}\left[A_1 A_2\right]\right).$$

Since A_i's are iid, we have that $\mathbb{E}\left[A_1 A_2\right] = \mathbb{E}\left[A_1\right]\mathbb{E}\left[A_2\right] = \mathbb{E}\left[A_1\right]^2$. Thus, we have that

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} A_i\right)^2\right] = \frac{1}{n^2}\left(n\mathbb{E}\left[A_1^2\right] + n(n-1)\mathbb{E}\left[A_1\right]^2\right) \leq \frac{\mathbb{E}\left[A_1^2\right]}{n} + \mathbb{E}\left[A_1\right]^2.$$

Combining the bounds for Term A and Term B, we have

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{|Y_i|\le a}\right)^2\right] \le \frac{\sigma a^{1-\epsilon}}{n} + \frac{\sigma^2}{a^{2\epsilon}}.$$

Thus the overall bound is

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \ge t\right) \le \frac{n\sigma}{a^{1+\epsilon}} + \frac{\sigma a^{1-\epsilon}}{n} + \frac{\sigma^2}{a^{2\epsilon}}$$

for any $a, t \ge 0$. We will now choose $a = nt$, which gives

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \ge t\right) \le \frac{n\sigma}{(nt)^{1+\epsilon}} + \frac{\sigma(nt)^{1-\epsilon}}{n} + \frac{\sigma^2}{(nt)^{2\epsilon}} = \frac{\sigma}{n^\epsilon t^{1+\epsilon}} + \frac{\sigma}{n^\epsilon t^{1+\epsilon}} + \left(\frac{\sigma}{n^\epsilon t^{1+\epsilon}}\right)^2.$$

In the case when $\sigma/n^\epsilon t^{1+\epsilon} \le 1$, the bound is no more than $3\sigma/n^\epsilon t^{1+\epsilon}$. In the case when $\sigma/n^\epsilon t^{1+\epsilon} > 1$, the bound is no more than $3\sigma/n^\epsilon t^{1+\epsilon}$. Thus, we have that

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \ge t\right) \le \frac{3\sigma}{n^\epsilon t^{1+\epsilon}}.$$

Setting the right-hand side to δ and solving for t, we have

$$t = \left(\frac{3\sigma}{\delta n^\epsilon}\right)^{\frac{1}{1+\epsilon}}$$

This implies that with probability at least $1 - \delta$, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i \le \mathbb{E}X + \left(\frac{3\sigma}{\delta n^\epsilon}\right)^{\frac{1}{1+\epsilon}}$$

as desired. □

Problem 6 (Maximum degree of a random graph)

> Consider $G = (V, E)$, a random graph with n vertices. The graph is constructed such that for each pair of distinct vertices, an edge is added with probability $\frac{1}{2}$, with all pairs sampled independently (this is commonly denoted by $G \sim \mathcal{G}(n, \frac{1}{2})$). Recall that the degree of a vertex v, $d(v)$, is the number of neighbors of v, and the maximum degree of the graph G is $\max_{v \in V} d(v)$. Use the concentration inequalities to derive both the high-probability upper bound on $\max_{v \in V} d(v)$ and the upper bound on $\mathbb{E} \max_{v \in V} d(v)$.

To begin, we recall a few things about the degree distribution of a random graph. Let $G \sim \mathcal{G}(n, p)$ represent a random graph with n vertices and with edges added with probability p. We denote v as a vertex with a degree $d(v)$. The distribution of degree of the vertices then has a binomial distribution of degrees k:

$$\mathcal{P}_{\mathcal{G}(n,p)} = \Pr(d(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

The expected degree in the random graph is then

$$\mathbb{E}[d(v)] = (n-1)p.$$

In our case, we consider $p = \frac{1}{2}$, simplifying the above expressions to

$$\mathcal{P}_G(k) = \mathcal{P}_{\mathcal{G}(n,1/2)} = \binom{n-1}{k}\left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-1-k} = \left(\frac{1}{2}\right)^{n-1}\left(\frac{(n-1)!}{k!(n-1-k)!}\right) \quad \text{and} \quad \mathbb{E}[d(v)] = \frac{n-1}{2}.$$

Our goal is to show

$$\Pr\left(\max_{v \in V} d(v) \geq t\right) \leq \delta.$$

In order to do so, we will utilize Hoeffding's inequality with $X_i = d(v) \in [0, n-1]$ and $\mu_i = {}^{n-1}/_2$ for all $i \in [n]$. Then we have that

$$\Pr\left(\sum_{v \in V}\left(d(v) - \frac{n-1}{2}\right) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(n-1)^2}\right)$$

$$\Pr\left(n \max_{v \in V} d(v) - \frac{n(n-1)}{2} \geq t\right) \leq \underbrace{\exp\left(-\frac{2t^2}{n(n-1)^2}\right)}_{\delta}.$$

Solving for t in terms of δ results in

$$\log(\delta) = -\frac{2t^2}{n(n-1)^2} \implies t = \sqrt{\frac{-\log(\delta)n(n-1)^2}{2}} \implies t = (n-1)\sqrt{\frac{\log({}^1/_\delta)n}{2}}.$$

Then we have that with probability at least $1 - \delta$,

$$\boxed{\max_{v \in V} d(v) \leq \mathbb{E}[d(v)]\left(1 + \sqrt{\frac{2\log({}^1/_\delta)}{n}}\right).}$$

Problem 7 (Uniform distribution on the ball is sub-Gaussian)

Let X be a random vector uniformly distributed on the unit Euclidean ball in $\mathbb{R}^d$ centered at the origin. Show that for any $v \in S^{d-1}$,

$$\|\langle X, v\rangle\|_{\psi_2} \leq \mathbf{C}\, \|\langle X, v\rangle\|_{L_2},$$

where $C \geq 0$ is some absolute constant.

Hint: Recall that $\|Y\|_{L_p} = (\mathbb{E}|Y|^p)^{1/p}$, where $p \geq 1$. To control $\|\langle X, v\rangle\|_{L_2}$, compute the covariance matrix of X. Consider $Z \sim \mathcal{N}(0, I_d)$ and U, a uniform random variable independent of Z on $[0, 1]$. Utilize the fact that $U^{1/d} \cdot Z/\|Z\|_2$ is uniformly distributed in the unit ball (prove this) and that $\|Z\|_2$ concentrates strongly around its expectation.

We begin by proving the following lemma.

Lemma HW1.P6.14. *Let $Z \sim \mathcal{N}(0, I_d)$ and U, a uniform random variable independent of Z on $[0, 1]$. Then (i.) $U^{1/d} \cdot Z/\|Z\|_2$ is uniformly distributed in the unit ball and (ii.) $\|Z\|_2$ concentrates strongly around its expectation.*

Proof of (i.) in Lemma HW1.P6.14. We begin by showing that $U^{1/d} \cdot Z/\|Z\|_2$ is uniformly distributed in the unit ball, $\mathcal{B}_1$. A random vector $Y \in \mathbb{R}^d$ is uniformly distributed in the unit ball if its cumulative distribution function (CDF) is equal to r^n where $r \in [0, 1]$ is the radius of the ball. To see this, consider the volume of the d-dimensional ball with radius r,

$$V_d(r) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)} r^d$$

where Γ is Euler's gamma function. Then probability that a random vector Y is in the ball is

$$\Pr(Y \in \mathcal{B}_1) = \Pr(\|Y\|_2 \leq r) = F_Y(r) = \frac{V_d(r)}{V_1(r)} = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)} \frac{r^d}{\pi^{d/2}/\Gamma(\frac{d}{2}+1)} = r^d \quad \text{for all } r \in [0, 1].$$

We define Y as a random vector equal to $U^{1/d} \cdot Z/\|Z\|_2$. Then, the ℓ_2-norm of Y is

$$\|Y\|_2 = \left\|U^{1/d} \cdot Z/\|Z\|_2\right\|_2 = \frac{U^{1/d}}{\|Z\|_2}\|Z\|_2 = U^{1/d}.$$

Thus, the CDF of Y is equal to

$$F_Y(r) = \Pr(\|Y\|_2 \leq r) = \Pr\left(U^{1/d} \leq r\right) = \Pr\left(U \leq r^d\right) = F_U(r^d) = \frac{r^d - 0}{1 - 0} = r^d$$

and we achieve the desired result that $Y = U^{1/d} \cdot Z/\|Z\|_2$ is uniformly distributed in the unit ball, $\mathcal{B}_1$. □

Proof of (ii.) in Lemma HW1.P6.14. We now prove the later statement. Let Z be a random vector in $\mathbb{R}^d$ with independent components $Z_i \sim \mathcal{N}(0, 1)$ for all $i \in [d]$. Then we have

$$\mathbb{E}\|Z\|_2^2 = \mathbb{E}\sum_{i=1}^d Z_i^2 = \sum_{i=1}^d \mathbb{E}Z_i^2 = \sum_{i=1}^d \mathrm{Var}(Z_i) = d.$$

Thus the expected length of Z is $\mathbb{E}\|Z\|_2 = \sqrt{d}$. Additionally, recall in Lecture 6 that we prove that for a random vector $Z \in \mathbb{R}^d$ with independent coordinates Z_i such that $\mathbb{E}[Z_i] = 0$ and $\mathbb{E}[Z_i^2] = 1$ for all $i \in [d]$,

$$\Pr\left(\left|\frac{\|Z\|_2}{\sqrt{d}} - 1\right| \geq t\right) \leq 2\exp\left(-\mathbf{c}\frac{dt^2}{K^4}\right) \quad \text{where} \quad K = \max_{i\in[d]} \|Z_i\|_{\psi_2} \geq 1.$$

Changing variables to $\delta = t\sqrt{d}$, we obtain the desired tail

$$\Pr\left(\left|\|Z\|_2 - \sqrt{d}\right| \geq \delta\right) \leq 2\exp\left(-\frac{\mathbf{c}\delta^2}{K^4}\right) \leq 2\exp\left(-\mathbf{c}\delta^2\right) \quad \text{for all } \delta \geq 0.$$

Hence, we have that with probability at least $1 - \delta$

$$\left|\|Z\|_2 - \mathbb{E}\|Z\|_2\right| \leq 2\exp\left(-\mathbf{c}\delta^2\right).$$

and we can conclude that $\|Z\|_2$ concentrates strongly around its expectation. □

Now we proceed by proving the following theorem.

Theorem HW1.P6.15. *Let X be a random vector uniformly distributed on the unit Euclidean ball in $\mathbb{R}^d$ centered at the origin. Then for any $v \in S^{d-1}$,*

$$\|\langle X, v\rangle\|_{\psi_2} \leq \mathbf{C}\,\|\langle X, v\rangle\|_{L_2},$$

where $\mathbf{C} \geq 0$ is some absolute constant.

Proof. Let $X = U^{1/d} \cdot Z/\|Z\|_2$ where $Z \sim \mathcal{N}(0, I_d)$ and U, a uniform random variable independent of Z is on $[0, 1]$. We begin by calculating the covariance matrix of X.

$$\mathbb{E}_X XX^\top = \mathop{\mathbb{E}}_U\left[\mathop{\mathbb{E}}_Z\left[\frac{U^{2/d}}{\|Z\|_2^2}ZZ^\top\right]\right] \approx \frac{1}{d}\mathop{\mathbb{E}}_U\left[U^{2/d}\right]\mathop{\mathbb{E}}_Z\left[ZZ^\top\right] = \frac{1}{d}\left(\frac{u^{(2/d)+1}}{(2/d)+1}\right)\Bigg|_0^1 I_d = \left(\frac{1}{d}\right)\left(\frac{d}{d+2}\right)I_d = \left(\frac{1}{d+2}\right)I_d.$$

Next, we calculate the L_2 norm of $\langle X, v\rangle$, To begin note that

$$\|\langle X, v\rangle\|_{L_2} = (\mathbb{E}\left[|\langle X, v\rangle|^2\right])^{1/2} = (\mathbb{E}\left[(X^\top v)^2\right])^{1/2} = (\mathbb{E}\left[v^\top XX^\top v\right])^{1/2} = (v^\top \underbrace{\mathbb{E}\left[XX^\top\right]}_{\mathrm{Cov}(X,X)} v)^{1/2}.$$

Thus, the L_2 norm of $\langle X, v\rangle$ is simply,

$$\|\langle X, v\rangle\|_{L_2} = \frac{1}{\sqrt{d+2}}.$$

We now recall our definition of a random sub-Gaussian vector from class, where we let X be a d-dimensional random vector with $\mathbb{E}X = 0$.

Definition HW1.P6.16. *X is sub-Gaussian if for all $v \in S^{d-1}$, $\|\langle v, x\rangle\|_{\psi_2} \leq C\sqrt{v^T\Sigma v}$.*

In our case, $\sqrt{v^T\Sigma v}$ is precisely equal to the L_2 norm of $\langle X, v\rangle$. Finally, we proceed by calculating the $\|\cdot\|_{\psi_2}$ of $\langle X, v\rangle$. We have that

$$\begin{aligned}
\|\langle X, v\rangle\|_{\psi_2} = \|X^\top v\|_{\psi_2} &= \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{v^\top XX^\top v}{t^2}\right)\right] \leq 2\right\} \\
\text{(By Jensen's inequality)} \quad &= \inf\left\{t > 0 : \exp\left(\mathbb{E}\left[\frac{v^\top XX^\top v}{t^2}\right]\right) \leq 2\right\} \\
&= \inf\left\{t > 0 : \exp\left(\frac{1}{(d+2)t^2}\right) \leq 2\right\} \\
&= \inf\left\{t > 0 : \frac{1}{(d+2)t^2} \leq \log(2)\right\}
\end{aligned}$$

$$\|\langle X, v\rangle\|_{\psi_2} = \inf\left\{t > 0 : \frac{1}{\sqrt{(d+2)\log(2)}} \leq t\right\}$$

Thus, we will pick the t the minimizes the objective while still being at least as large as $\frac{1}{\sqrt{(d+2)\log(2)}}$. Therefore,

$$\|\langle X, v\rangle\|_{\psi_2} = \frac{1}{\sqrt{(d+2)\log(2)}} = \frac{1}{\sqrt{\log(2)}} \|\langle X, v\rangle\|_{L_2} \implies \boxed{\|\langle X, v\rangle\|_{\psi_2} \leq \mathbf{C} \|\langle X, v\rangle\|_{L_2}.}$$

□

Problem 8 (Non-asymptotic analysis of fixed design linear regression)

Consider a fixed design linear regression model. Let $x_1, \ldots, x_n$ be fixed design vectors in $\mathbb{R}^d$. Assume the response variables $Y_1, \ldots, Y_n$ are independent, with $Y_i - \mathbb{E}Y_i$ being sub-Gaussian with parameter σ for all $i \in [n]$. That is, for all λ and $i \in [n]$,

$$\mathbb{E}\exp(\lambda(Y_i - \mathbb{E}Y_i)) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

We are interested in a high-probability (with respect to the realization of $Y_1, \ldots, Y_n$) upper bound on the *excess risk* (which measures the statistical performance of some proposed estimator compared to the population best estimator):

$$\mathcal{E}(\beta) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^\top\beta - Y_i)^2\right] - \inf_{\beta\in\mathbb{R}^d}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^\top\beta - Y_i)^2\right].$$

Let $\hat{\beta}$ be the ordinary least squares estimator in $\mathbb{R}^d$. Show that, with probability at least $1-\delta$,

$$\mathcal{E}(\hat{\beta}) \leq \frac{\sigma^2\left(d + 2\sqrt{2d\log(1/\delta)} + 2\log(1/\delta)\right)}{n}.$$

What conclusions can be drawn about $\mathbb{E}\mathcal{E}(\beta)$?

Hint: Assume without loss of generality that the sample covariance matrix is invertible. Then, simplify the expression for the excess risk and apply one of the concentration inequalities discussed in the lectures.

Let us rewrite the excess risk in matrix form. Let $X \in \mathbb{R}^{n\times d}$ be a matrix with $x_i^\top$ as its rows. Additionally, let $\beta^\star$ represent the optimal $\beta \in \mathbb{R}^d$. Then, we have

$$\mathcal{E}(\beta) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^\top\beta - Y_i)^2\right] - \inf_{\beta\in\mathbb{R}^d}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^\top\beta - Y_i)^2\right] = \frac{1}{n}\mathbb{E}\left[\|X\beta - Y\|_2^2 - \|X\beta^\star - Y\|_2^2\right].$$

Recall that the ordinary least squares solution $\hat{\beta}$ is $(X^\top X)^{-1}X^\top Y$. Note that the optimal solution is in expectation, i.e., $\mathbb{E}[Y - X\beta^\star] = 0$. Then we have

$$\mathcal{E}(\hat{\beta}) = \frac{1}{n}\mathbb{E}\left[\|X\hat{\beta} - Y\|_2^2 - \|X\beta^\star - Y\|_2^2\right] = \frac{1}{n}\mathbb{E}\left[\|\underbrace{X(X^\top X)^{-1}X^\top}_{P}Y - Y\|_2^2 - \|X\beta^\star - Y\|_2^2\right]$$

We note that $P = X(X^\top X)^{-1}X^\top$ is a symmetric, projection matrix such that $P^\top = P$ and $P^2 = P$,

$$P^\top = (X(X^\top X)^{-1}X^\top)^\top = X(X^\top X)^{-1}X^\top = P \quad \text{and}$$

$$P^2 = (X(X^\top X)^{-1}X^\top)(X(X^\top X)^{-1}X^\top) = (X(X^\top X)^{-1}X^\top) = P.$$

We also observe that $PX\beta = X\beta$,

$$(X(X^\top X)^{-1}X^\top)X\beta = X\beta.$$

Then we have

$$\mathcal{E}(\hat{\beta}) = \frac{1}{n}\mathbb{E}\left[\|PY - Y\|_2^2 - \|PX\beta^\star - Y\|_2^2\right]$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbb{E}\left[Y^\top P^\top PY + Y^\top Y - 2Y^\top P^\top Y - \left(\beta^{\star\top} X^\top P^\top PX\beta^\star + Y^\top Y - 2\beta^{\star\top} X^\top P^\top Y\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[Y^\top P^\top PY - 2Y^\top P^\top Y - \beta^{\star\top} X^\top P^\top PX\beta^\star + 2\beta^{\star\top} X^\top P^\top Y\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[Y^\top PY + \mathbb{E}\left[Y\right]^\top P\mathbb{E}\left[Y\right] - 2\mathbb{E}\left[Y\right]^\top PY\right] \\
\mathcal{E}(\hat{\beta}) &= \frac{1}{n}\mathbb{E}\left[\|P(Y - \mathbb{E}Y)\|_2^2\right].
\end{aligned}
$$

Additionally, we introduce $Z = Y - \mathbb{E}Y$ and $\widetilde{P} = P/\sqrt{n}$. The excess risk then becomes

$$
\mathcal{E}(\hat{\beta}) = \mathbb{E}\left[\|\widetilde{P}Z\|_2^2\right]
$$

Recall the following proposition which we proved in class.

Proposition HW1.P7.17. *Let X be a d-dimensional sub-Gaussian random vector such that for all $\lambda \in \mathbb{R}$, $v \in S^{d-1}$,*

$$
\mathbb{E}[\exp(\lambda\langle v, X\rangle)] \leq \exp\left(\frac{\lambda^2 v^T \Sigma v}{2}\right).
$$

where $\Sigma = \mathbb{E}XX^\top$. Then, for all $\delta \in (0,1)$ we have

$$
\Pr\left(\|X\| \geq \sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{2\lambda_{\max}(\Sigma)\log(1/\delta)}\right) \leq \delta.
$$

Remark HW1.P7.18. *Note that squaring both sides of the inequality in Proposition HW1.P7.17 results in*

$$
\Pr\left(\|X\|_2^2 \geq \left(\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{2\lambda_{\max}(\Sigma)\log(1/\delta)}\right)^2\right) \leq \delta
$$

$$
\Pr\left(\|X\|_2^2 \geq \mathrm{Tr}(\Sigma) + 2\sqrt{\mathrm{Tr}(\Sigma)}\sqrt{2\lambda_{\max}(\Sigma)\log(1/\delta)} + 2\lambda_{\max}(\Sigma)\log(1/\delta)\right) \leq \delta.
$$

Then by Proposition HW1.P7.17 with $X = \widetilde{P}Z$ and $\Sigma = \mathbb{E}Z^\top\widetilde{P}^\top\widetilde{P}Z$, we have that

$$
\Pr\left(\|\widetilde{P}Z\|_2^2 < \mathrm{Tr}(\Sigma) + 2\sqrt{\mathrm{Tr}(\Sigma)}\sqrt{2\lambda_{\max}(\Sigma)\log(1/\delta)} + 2\lambda_{\max}(\Sigma)\log(1/\delta)\right) > 1 - \delta.
$$

Let us analyze $\mathrm{Tr}(\Sigma)$ and $\lambda_{\max}(\Sigma)$. We note that

$$
\Sigma = \mathbb{E}\widetilde{P}ZZ^\top\widetilde{P}^\top = \widetilde{P}\mathbb{E}\left[ZZ^\top\right]\widetilde{P}^\top,
$$

and then since each entry in Z is independent (e.g., $Z_i = Y_i - \mathbb{E}Y_i$), $\mathbb{E}\left[ZZ^\top\right] \leq \sigma^2 I_n$ by the sub-Gaussianity of Y_i. Then we have $\Sigma \leq \sigma^2\widetilde{P}I_n\widetilde{P}^\top = \sigma^2 P/n$ and since P is a projection matrix, $\mathrm{Tr}(P) = \mathrm{rank}(X) \leq d$ and $\lambda_{\max} = 1$. Thus we are left with

$$
\Pr\left(\|\widetilde{P}Z\|_2^2 \leq \frac{d + 2\sqrt{d}\sqrt{2(1)\log(1/\delta)} + 2(1)\log(1/\delta)}{n}\right) > 1 - \delta
$$

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

$$
\boxed{\mathcal{E}(\hat{\beta}) \leq \frac{d + 2\sqrt{2d\log(1/\delta)} + 2\log(1/\delta)}{n}.}
$$

STAT210B Theoretical Statistics March 6th, 2024

Homework # 2: Bounds for Random Matrices

Reece D. Huff

Regrades

When regrading, I only attach problems in which I did not receive 100%. If the mistake is minor, I highlight my changes in purple. If the mistake is major, I highlight the entire problem in purple. I provide the regrade justification in the gray box below the problem statement.

Notation

Let $\mathbf{c}$ and $\mathbf{C}$ represent a small and large positive constant, respectively (e.g., $\mathbf{c} = 10^{-5}$ and $\mathbf{C} = 10^5$). Unless otherwise specified, we use the notation $[n]$ to represent the set of integers $\{1, ..., n\}$. Given we matrix $A \in \mathbb{R}^{m \times n}$, we use $||A||$ to denote the operator norm,

$$||A||_{\text{op}} = \sup_{v \in S^{n-1}} ||Av||_2 = \sup_{u \in S^{m-1}, v \in S^{n-1}} u^\intercal A v = \lambda_{\max}(A).$$

Problem 1 (Covering the unit cube in ℓ_∞)

Consider the cube $[-1, 1]^d$ in $\mathbb{R}^d$, equipped with the distance

$$\rho(\theta, \theta') = \|\theta - \theta'\|_\infty = \max_{i \in [d]} |\theta_i - \theta_i'|.$$

Show that the covering numbers of this set at scale ε are bounded by $\left(1 + \frac{1}{\varepsilon}\right)^d$.

Theorem HW2.P1.

Consider the cube $[-1, 1]^d$ in $\mathbb{R}^d$, equipped with the distance

$$\rho(\theta, \theta') = \|\theta - \theta'\|_\infty = \max_{i \in [d]} |\theta_i - \theta_i'|.$$

Then the ε-covering number is upper-bounded by,

$$\mathcal{N}\left([-1, 1]^d, \rho, \varepsilon\right) \leq \left(1 + \frac{1}{\varepsilon}\right)^d.$$

Proof. To begin we consider dividing each coordinate θ_i for all $i \in [d]$ into $M := \lfloor 1/\varepsilon \rfloor$ [5] $+ 1$ sub-intervals. We define the centers of these sub-intervals as $\theta_i^{(j)} = -1 + 2(j-1)\varepsilon$ for all $i \in [d]$ and $j \in [M]$ and note

[5] For a scalar $\alpha \in \mathbb{R}$, we use $\lfloor \alpha \rfloor$ to represent the "floor" or the greatest integer less than or equal to α.

that the length of each sub-interval is at most 2ε. It then follows that for any $\widetilde{\theta}_i \in [0,1]$, there exists some $j \in [M]$ such that $\left|\theta_i^{(j)} - \widetilde{\theta}_i\right| \leq \varepsilon$ for all $i \in [d]$, which implies that

$$\mathcal{P}([-1,1],\rho,2\varepsilon) \leq \mathcal{N}([-1,1],\rho,\varepsilon) \leq 1 + \frac{1}{\varepsilon} \text{ for all coordinates implying that}$$

$$\boxed{\mathcal{N}\left([-1,1]^d,\rho,\varepsilon\right) \leq \left(1+\frac{1}{\varepsilon}\right)^d.}$$

□

Problem 2 (Sample covariance of bounded distributions)

Assume that $X_1,\ldots,X_n$ are independent zero mean random vectors in $\mathbb{R}^d$ with covariance matrix Σ such that $\|X_i\| \leq r$ almost surely. Show that there is an absolute constant $c > 0$ such that, with probability at least $1-\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i X_i^T - \Sigma\right\| \leq c\left(\sqrt{\frac{r^2\|\Sigma\|(\log(d)+\log(1/\delta))}{n}} + \frac{r^2(\log(d)+\log(1/\delta))}{n}\right).$$

Hint: You might need the following computation. Recall that the variance of a symmetric random matrix A is given by $\mathrm{Var}(A) = \mathbb{E}A^2 - (\mathbb{E}A)^2$. Show that $\mathrm{Var}(A) \succeq 0$.

Theorem HW2.P1.

Let $X_1,\ldots,X_n$ be independent zero mean random vectors in $\mathbb{R}^d$ with covariance matrix Σ such that $\|X_i\| \leq r$ almost surely. Then we have that with probability at least $1-\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i X_i^T - \Sigma\right\| \leq \mathbf{c}\left(\sqrt{\frac{r^2\|\Sigma\|(\log(d)+\log(1/\delta))}{n}} + \frac{r^2(\log(d)+\log(1/\delta))}{n}\right)$$

where $\mathbf{c} > 0$ is an absolute constant.

Regrade justification: Tiny issue with bounding the operator norm with the triangle inequality. My original solution was correct, just a tad loose.

Proof. To begin we show that the variance of a symmetric random matrix A is positive semi-definite. Given a symmetric random matrix $A \in \mathbb{R}^{d\times d}$, we note that

$$\mathrm{Var}(A) = \mathbb{E}\left[A - \mathbb{E}A\right]^2 = \mathbb{E}\left[A^2\right] - 2\mathbb{E}\left[A\mathbb{E}A\right] + \mathbb{E}\left[(\mathbb{E}A)^2\right] = \mathbb{E}A^2 - (\mathbb{E}A)^2.$$

It that follows that for any v in the unit sphere S^{d-1},

$$v^\top \mathrm{Var}(A) v = v^\top\left(\mathbb{E}A^2 - (\mathbb{E}A)^2\right)v = v^\top \mathbb{E}A^2 v - v^\top(\mathbb{E}A)^2 v = \mathbb{E}\|Av\|_2^2 - \|\mathbb{E}Av\|_2^2.$$

By Jensen's inequality (i.e., $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$), we have that $\mathbb{E}\|Av\|_2^2 \geq \|\mathbb{E}Av\|_2^2$ and thus

$$v^\top \mathrm{Var}(A) v \geq 0 \text{ for all } v \in S^{d-1} \text{ implying that } \mathrm{Var}(A) \succeq 0.$$

We will leverage the Matrix Bernstein's inequality to complete this proof (Theorem HW2.P1.1).

Theorem HW2.P1.1 (Bernstein bound for random matrices [Wai19]). *Let $\{Q_i\}_{i=1}^n$ be a sequence of independent, zero-mean, symmetric random $d \times d$-matrices that satisfy $\|Q_i\|_{\text{op}} \leq K$ almost surely for all i. Then, for every $t \geq 0$, we have*

$$\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\| \geq t\right) \leq 2d\exp\left(-\frac{nt^2}{2(\sigma^2+Kt)}\right) \leq 2d\exp\left(-\min\left\{\frac{nt^2}{2\sigma^2}, \frac{nt}{2K}\right\}\right).$$

where $\sigma^2 = \frac{1}{n}\left\|\sum_{i=1}^n \mathbb{E}Q_i^2\right\|$ is the norm of the matrix variance of the sum.

We use Theorem HW2.P1.1 with $Q_i = X_iX_i^\top - \Sigma$. First we show that Q_i is in fact zero-mean

$$\mathbb{E}[Q_i] = \mathbb{E}\left[X_iX_i^\top - \Sigma\right] = \mathbb{E}\left[X_iX_i^\top\right] - \Sigma = 0 \quad \text{for all } i \in [n].$$

with variance

$$\mathbb{E}\left[Q_i^2\right] = \mathbb{E}\left[(X_iX_i^\top - \Sigma)^2\right] = \mathbb{E}\left[X_iX_i^\top X_iX_i^\top + \Sigma\Sigma - 2\Sigma X_iX_i^\top\right] = \mathbb{E}\left[\|X_i\|_2^2 X_iX_i^\top\right] - \Sigma\Sigma \preceq \|X_i\|_2^2\mathbb{E}\left[X_iX_i^\top\right]$$
$$\mathbb{E}\left[Q_i^2\right] \preceq r^2\Sigma.$$

Next we calculate the norm of the matrix variance of the sum

$$\sigma^2 = \frac{1}{n}\left\|\sum_{i=1}^n \mathbb{E}Q_i^2\right\| \leq \frac{1}{n}\sum_{i=1}^n \left\|\mathbb{E}Q_i^2\right\| \leq \frac{1}{n}\sum_{i=1}^n \left\|r^2\Sigma\right\| \leq r^2\|\Sigma\|.$$

Finally we note that

$$\|XX^\top\|_{\text{op}} = \|X\|_2^2 \quad \text{which follows from the def.} \quad \|XX^\top\|_{\text{op}} = \sup_{v\in S_{d-1}} v^\top XX^\top v \quad = \sup_{v\in S_{d-1}} (X^\top v)^2.$$

Clearly, $X^\top v$ is maximized when v points in the same direction as X. Then we have $v^\star = {}^X/_{\|X\|_2}$

$$\|XX^\top\|_{\text{op}} = \frac{X^\top}{\|X\|_2}XX^\top\frac{X}{\|X\|_2} = \left(\frac{\|X\|_2^2}{\|X\|_2}\right)\left(\frac{\|X\|_2^2}{\|X\|_2}\right) = \|X\|_2^2$$

This implies that

$$\|Q_i\|_{\text{op}} = \|X_iX_i^\top - \Sigma\|_{\text{op}} \leq \|X_iX_i^\top\|_{\text{op}} + \|\Sigma\|_{\text{op}} \leq 2r^2 \quad \text{for all } i \in [n]$$

as $\left\|X_iX_i^\top\right\|_{\text{op}} = \|X_i\|_2^2 \leq r^2$.

Then, we use Theorem HW2.P1.1 to solve for t in terms of δ for the two different regimes

$$\delta = 2d\exp\left(-\frac{nt^2}{2\sigma^2}\right) \implies \log\left(\frac{\delta}{2d}\right) = -\frac{nt^2}{2\sigma^2} \implies t = \sqrt{\frac{2\sigma^2}{n}\log\left(\frac{2d}{\delta}\right)}$$
$$\delta = 2d\exp\left(-\frac{nt}{2K}\right) \implies \log\left(\frac{\delta}{2d}\right) = -\frac{nt}{2K} \implies t = \frac{2K}{n}\log\left(\frac{2d}{\delta}\right)$$

Taken together this implies that with at least probability $1-\delta$

$$\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_{\text{op}} \leq \sqrt{\frac{2\sigma^2}{n}\log\left(\frac{2d}{\delta}\right)} + \frac{2K}{n}\log\left(\frac{2d}{\delta}\right) \leq \sqrt{\frac{2r^2\|\Sigma\|}{n}\log\left(\frac{2d}{\delta}\right)} + \frac{4r^2}{n}\log\left(\frac{2d}{\delta}\right)$$

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

$$\boxed{\left\|\frac{1}{n}\sum_{i=1}^n X_iX_i^T - \Sigma\right\| \leq \mathbf{c}\left(\sqrt{\frac{r^2\|\Sigma\|(\log(d)+\log(1/\delta))}{n}} + \frac{r^2(\log(d)+\log(1/\delta))}{n}\right).}$$

where $\mathbf{c} > 0$ is an absolute constant. □

Problem 3 (Norm of sub-exponential random vectors)

Let X be a zero-mean random vector in $\mathbb{R}^d$ with covariance Σ, which satisfies that for all $v \in S^{d-1}$,

$$\|\langle X, v\rangle\|_{\psi_1} \le \mathbf{c}\|\langle X, v\rangle\|_{L_2},$$

where $\mathbf{c} > 0$ is an absolute constant.

1. Show that there is an absolute constant $\mathbf{C} > 0$ such that for all $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,

$$\|X\| \le \mathbf{C}\left(\sqrt{\mathrm{Tr}(\Sigma)}\log(1/\delta) + \log(1/\delta)\sqrt{\lambda_{\max}(\Sigma)}\right).$$

2. Compare your result with what one can get if the ε-net argument is applied together with the union bound to upper bound the norm of $\|X\|$.

Hint: Adapt the analysis of the sub-Gaussian sample covariance matrix.

Recall the following Lemmas and Facts we proved in class:

Lemma HW2.P2.2. *Fix some probability density π on Θ, and let $f(X, \theta)$ be a function with X being a random variable and $\theta \in \Theta \subseteq \mathbb{R}^d$. Then, with probability at least $1 - \delta$, it holds that for any probability density ρ on Θ for which $KL(\rho\|\pi) < \infty$,*

$$\mathbb{E}_{\theta\sim\rho} f(X, \theta) \le \mathbb{E}_{\theta\sim\rho} \log \mathbb{E}_X e^{f(X,\theta)} + KL(\rho\|\pi) + \log(1/\delta). \tag{76}$$

Fact HW2.P2.3. *If $Y \sim \mathcal{N}_d(\mu, \sigma^2 I_d)$ and $A \in \mathbb{R}^{d\times d}$ then $\mathbb{E}[Y^\top A Y] = \sigma^2 \operatorname{Tr} A + \mu^\top A \mu$.*

Fact HW2.P2.4. *If ρ, π are the densities of $\mathcal{N}_d(v, I_d/\beta)$, $\mathcal{N}_d(0, I_d/\beta)$, respectively, and $\|v\|_2 = 1$, then $KL(\rho\|\pi) = \beta/2$.*

Fact HW2.P2.5. *If $x \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric and positive semidefinite then $\sup_{v\in S^{d-1}} \langle x, v\rangle = \|x\|_2$ and $\sup_{v\in S^{d-1}} v^\top \Sigma v = \lambda_{max}(\Sigma)$.*

Fact HW2.P2.6. *The function $f(x) = ax + \frac{b}{x}$ for $a, b, x > 0$ is minimized at $x^* = \sqrt{b/a}$ and has $f(x^*) = 2\sqrt{ab}$.*

Fact HW2.P2.7. *If X is a random vector with $\mathbb{E}[XX^\top] = \Sigma$ then, for any $v \in \mathbb{R}^n$, $\|\langle X, v\rangle\|_{L_2}^2 = v^\top \Sigma v$.*

Finally, beyond the Lemmas and Facts we proved in class, we will also need the following inequality for sub-exponential random vectors.

Lemma HW2.P2.8 (Zero mean sub-exponential inequality)**.** *Let Z be a zero-mean sub-exponential random variable. Then, we have that*

$$\mathbb{E}\left[\exp(\lambda X)\right] \le \exp\left(\mathbf{C}^2 \|Z\|_{\psi_1}^2 \lambda^2\right) \quad \text{for all } \lambda \le \frac{1}{\mathbf{C}\|Z\|_{\psi_1}},$$

where $\mathbf{C} > 0$ is an absolute constant.

Proof. Proving this inequality is a standard application of the sub-exponential properties.
By Proposition 2.7.1 Property (d.) in [Ver18, Vershynin's High-dimensional probability], we have that for any sub-exponential random variable Z, we have that

$$\mathbb{E}\exp(|Z|/K_4) \le 2$$

where $K_4 > 0$ is an absolute constant.
By Proposition 2.7.1 Property (e.) in [Ver18, Vershynin's High-dimensional probability], we have that for a zero-mean ($\mathbb{E}Z = 0$) sub-exponential random variable Z, we have that

$$\mathbb{E}\exp(\lambda Z) \le \exp\left(K_5^2\lambda^2\right) \quad \text{for all } \lambda \text{ such that } |\lambda| \le \frac{1}{K_5}$$

where $K_5 > 0$ is an absolute constant.
In the properties in Proposition 2.7.1 in [Ver18, Vershynin's High-dimensional probability], there exists an absolute constant $\mathbf{C}$ such that property i implies property j with parameter $K_j \le \mathbf{C}K_i$ for any two properties.
We set $K_4 = \|Z\|_{\psi_1}$ and $K_5 = \mathbf{C}K_4$ to get the desired result, i.e.,

$$\mathbb{E}\exp(\lambda Z) \le \exp\left(\mathbf{C}^2\|Z\|_{\psi_1}^2\lambda^2\right) \quad \text{for all } \lambda \le \frac{1}{\mathbf{C}\|Z\|_{\psi_1}}.$$

□

Now we are ready to prove the following theorem.

Theorem HW2.P2.1a

Show that there is an absolute constant $\mathbf{C} > 0$ such that for all $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,

$$\|X\|_2 \le \mathbf{C}\left(\sqrt{\mathrm{Tr}(\Sigma)}\log(1/\delta) + \log(1/\delta)\sqrt{\lambda_{\max}(\Sigma)}\right).$$

Proof. To begin, we use Lemma HW2.P2.2 with $f(X, \theta) = \eta\langle X, \theta\rangle$ and ρ and π being the densities of $\mathcal{N}_d(v, \Sigma/\beta)$ and $\mathcal{N}_d(0, \Sigma/\beta)$, respectively, where $v \in S^{d-1}$ and $\beta > 0$ is a parameter to be chosen later. We have that

$$\underbrace{\mathbb{E}_{\theta\sim\rho}[\eta\langle X, \theta\rangle]}_{\text{Term 1}} \le \underbrace{\mathbb{E}_{\theta\sim\rho}\log\mathbb{E}_X e^{\eta\langle X,\theta\rangle}}_{\text{Term 2}} + \underbrace{KL(\rho\|\pi)}_{\text{Term 3}} + \log(1/\delta).$$

We will bound each term separately. Starting with Term 1, we have

$$\mathbb{E}_{\theta\sim\rho}[\eta\langle X, \theta\rangle] = \eta\mathbb{E}_{\theta\sim\rho}[\langle X, \theta\rangle] = \eta\left\langle X, \mathbb{E}_{\theta\sim\rho}[\theta]\right\rangle = \eta\langle X, v\rangle.$$

For Term 2, we leverage Lemma HW2.P2.8 with the sub-exponential random variable $Z = \langle X, \theta\rangle$ with parameter η. We have that

$$\mathbb{E}_{\theta\sim\rho}\log\mathbb{E}_X e^{\eta\langle X,\theta\rangle} \le \mathbb{E}_{\theta\sim\rho}\log\exp\left(\mathbf{C}^2\|\langle X, \theta\rangle\|_{\psi_1}^2\eta^2\right)$$
$$\mathbb{E}_{\theta\sim\rho}\log\mathbb{E}_X e^{\eta\langle X,\theta\rangle} \le \mathbf{C}^2\eta^2\mathbb{E}_{\theta\sim\rho}[\|\langle X, \theta\rangle\|_{\psi_1}^2] \quad \text{for all } |\eta| \le \frac{1}{\mathbf{C}\|\langle X, \theta\rangle\|_{\psi_1}}.$$

Now we analyze $\mathbb{E}_{\theta\sim\rho}[\|\langle X, \theta\rangle\|_{\psi_1}^2]$. By assumption and Fact HW2.P2.7, we have that

$$\mathbb{E}_{\theta\sim\rho}[\|\langle X, \theta\rangle\|_{\psi_1}^2] \le \mathbb{E}_{\theta\sim\rho}\left[\mathbf{c}^2\|\langle X, \theta\rangle\|_{L_2}^2\right] \le \mathbb{E}_{\theta\sim\rho}\left[\|\langle X, \theta\rangle\|_{L_2}^2\right] = \mathbb{E}_{\theta\sim\rho}[\theta^\top\Sigma\theta].$$

We can then expand $\theta^\top\Sigma\theta$ as

$$\mathbb{E}_{\theta\sim\rho}\left[\theta^\top\Sigma\theta\right] = \mathbb{E}_{\theta\sim\rho}\left[(\theta - v + v)^\top\Sigma(\theta - v + v)\right] = \mathbb{E}_{\theta\sim\rho}\left[(\theta - v)^\top\Sigma(\theta - v) + v^\top\Sigma v + 2(\theta - v)^\top\Sigma v\right]$$

noting that the cross term $2(\theta - v)^\top \Sigma v$ vanishes due to the expectation, i.e., $\mathbb{E}_{\theta\sim\rho}\left[(\theta - v)^\top \Sigma v\right] = (\mathbb{E}_{\theta\sim\rho}[\theta] - v)^\top \Sigma v = 0$. We then have that

$$\mathbb{E}_{\theta\sim\rho}\left[\theta^\top \Sigma \theta\right] = \mathbb{E}_{\theta\sim\rho}\left[(\theta - v)^\top \Sigma(\theta - v)\right] + v^\top \Sigma v.$$

Finally, we note that the random vector $\theta - v$ follows a Gaussian distribution with mean zero and covariance Σ/β. We can then apply Fact HW2.P2.3 to get that

$$\mathbb{E}_{\theta\sim\rho}\left[\theta^\top \Sigma \theta\right] = \mathbb{E}_{\theta\sim\rho}\left[(\theta - v)^\top \Sigma(\theta - v)\right] + v^\top \Sigma v = \frac{\mathrm{Tr}(\Sigma)}{\beta} + v^\top \Sigma v.$$

Thus, the overall bound for Term 2 is

$$\mathbb{E}_{\theta\sim\rho} \log \mathbb{E}_X \exp(\eta \langle X, \theta\rangle) \le \mathbf{C}^2\eta^2\left(\frac{\mathrm{Tr}(\Sigma)}{\beta} + v^\top \Sigma v\right) \quad \text{for all } |\eta| \le \frac{1}{\mathbf{C}\,\|\langle X, \theta\rangle\|_{\psi_1}}.$$

For Term 3, we use Fact HW2.P2.4 to get that

$$KL(\rho\|\pi) = \frac{\beta}{2}.$$

Combining the bounds for each term, we have that

$$\eta \langle X, v\rangle \le \mathbf{C}^2\eta^2\left(\frac{\mathrm{Tr}(\Sigma)}{\beta} + v^\top \Sigma v\right) + \frac{\beta}{2} + \log(1/\delta) \quad \text{for all } |\eta| \le \frac{1}{\mathbf{C}\,\|\langle X, \theta\rangle\|_{\psi_1}}.$$

Taking the supremum over all $v \in S^{d-1}$, we are able to leverage Fact HW2.P2.5 to get that

$$\begin{aligned}
\eta\|X\|_2 = \eta \sup_{v\in S^{d-1}} \langle X, v\rangle &\le \mathbf{C}^2\eta^2\left(\frac{\mathrm{Tr}(\Sigma)}{\beta} + \sup_{v\in S^{d-1}}\left\{v^\top \Sigma v\right\}\right) + \frac{\beta}{2} + \log(1/\delta) \\
\implies \|X\|_2 &\le \mathbf{C}^2\eta\left(\frac{\mathrm{Tr}(\Sigma)}{\beta} + \lambda_{\max}(\Sigma)\right) + \frac{\beta}{2\eta} + \frac{\log(1/\delta)}{\eta} \quad \text{for all } |\eta| \le \frac{1}{\mathbf{C}\,\|\langle X, \theta\rangle\|_{\psi_1}}.
\end{aligned}$$

Unfortunately, we are not able to directly optimize over η in the above inequality. For example, if we used Fact HW2.P2.6 to optimize for η, we would get that $\eta = 2\sqrt{\log(1/\delta)/\mathbf{C}^2\lambda_{\max}(\Sigma)}$ which may be much larger than $1/\mathbf{C}\|\langle X,\theta\rangle\|_{\psi_1}$ as δ approaches zero. Thus, we seek to set η to meet the sub-exponential condition, i.e., $|\eta| \le \frac{1}{\mathbf{C}\|\langle X,\theta\rangle\|_{\psi_1}}$, and then optimize over β to get the best bound. To begin, we note that

$$\begin{aligned}
\|\langle X, \theta\rangle\|_{\psi_1}^2 = \mathbb{E}_{\theta\sim\rho}[\|\langle X, \theta\rangle\|_{\psi_1}^2] &\le \frac{\mathrm{Tr}(\Sigma)}{\beta} + v^\top \Sigma v \le \frac{\mathrm{Tr}(\Sigma)}{\beta} + \lambda_{\max}(\Sigma) \\
\implies \|\langle X, \theta\rangle\|_{\psi_1} &\le \sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta} + \lambda_{\max}(\Sigma)}.
\end{aligned}$$

It then follows that

$$\|\langle X, \theta\rangle\|_{\psi_1} \le \sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta} + \lambda_{\max}(\Sigma)} \implies |\eta| \le \frac{1}{\mathbf{C}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta} + \lambda_{\max}(\Sigma)}} \le \frac{1}{\mathbf{C}\,\|\langle X, \theta\rangle\|_{\psi_1}}$$

thus, we set $\eta = 1/\left(\mathbf{C}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta}+\lambda_{\max}(\Sigma)}\right)$ to meet the sub-exponential condition. The norm of X is then bounded by

$$\|X\|_2 \leq \frac{\mathbf{C}^2\left(\frac{\mathrm{Tr}(\Sigma)}{\beta}+\lambda_{\max}(\Sigma)\right)}{\mathbf{C}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta}+\lambda_{\max}(\Sigma)}} + \frac{\beta\mathbf{C}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta}+\lambda_{\max}(\Sigma)}}{2} + \log(1/\delta)\mathbf{C}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta}+\lambda_{\max}(\Sigma)}$$

$$\|X\|_2 \leq \mathbf{C}\sqrt{\mathrm{Tr}(\Sigma)/\beta+\lambda_{\max}(\Sigma)}\left(1+\frac{\beta}{2}+\log(1/\delta)\right) \leq \mathbf{C}\sqrt{\mathrm{Tr}(\Sigma)/\beta+\lambda_{\max}(\Sigma)}\left(\frac{\beta}{2}+\log(1/\delta)\right).$$

To finish the proof, we optimize over β to get the best bound. We start with a bit of algebra to get that

$$\begin{aligned}
\|X\|_2 &\leq \mathbf{C}\sqrt{\mathrm{Tr}(\Sigma)/\beta+\lambda_{\max}(\Sigma)}\left(\frac{\beta}{2}+\log(1/\delta)\right) \leq \mathbf{C}\left(\sqrt{\mathrm{Tr}(\Sigma)/\beta}+\sqrt{\lambda_{\max}(\Sigma)}\right)\left(\frac{\beta}{2}+\log(1/\delta)\right) \\
&\leq \mathbf{C}\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta}}\beta+\sqrt{\lambda_{\max}(\Sigma)}\beta+\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta}}\log(1/\delta)+\sqrt{\lambda_{\max}(\Sigma)}\log(1/\delta)\right) \\
&= \mathbf{C}\left(\sqrt{\mathrm{Tr}(\Sigma)\beta}+\sqrt{\lambda_{\max}(\Sigma)}\beta+\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\beta}}\log(1/\delta)+\sqrt{\lambda_{\max}(\Sigma)}\log(1/\delta)\right)
\end{aligned}$$

We set $\beta = \log(1/\delta)$ to get that

$$\begin{aligned}
\|X\|_2 &\leq \mathbf{C}\left(\sqrt{\mathrm{Tr}(\Sigma)\log(1/\delta)}+\sqrt{\lambda_{\max}(\Sigma)}\log(1/\delta)+\sqrt{\frac{\mathrm{Tr}(\Sigma)}{\log(1/\delta)}}\log(1/\delta)+\sqrt{\lambda_{\max}(\Sigma)}\log(1/\delta)\right) \\
&= \mathbf{C}\left(2\sqrt{\mathrm{Tr}(\Sigma)\log(1/\delta)}+2\sqrt{\lambda_{\max}(\Sigma)}\log(1/\delta)\right) \\
\|X\|_2 &\leq \mathbf{C}\left(\sqrt{\mathrm{Tr}(\Sigma)\log(1/\delta)}+\sqrt{\lambda_{\max}(\Sigma)}\log(1/\delta)\right)
\end{aligned}$$

as desired. □

ε-net Argument

We aim to bound the norm of a zero-mean random vector X in $\mathbb{R}^d$ with covariance Σ. We will use the ε-net argument and compare it with the variational approach. This follows directly from the proof of Theorem 4.4.5. in [Ver18, Vershynin's High-dimensional probability].

Approximation via an ε-net: Consider an arbitrary $\varepsilon \in (0, 1/2)$ and denote by $\mathcal{N}_\varepsilon$ an ε-net of the unit sphere S^{d-1}, characterized by its cardinality $N_\varepsilon := |\mathcal{N}_\varepsilon|$. Invoking Exercise 4.4.3 from Vershynin, the following inequality is obtained:

$$\|X\| \leq \frac{1}{1-2\varepsilon}\max_{v\in N_\varepsilon}\langle X, v\rangle.$$

The factor $\frac{1}{1-2\varepsilon}$ is regarded as an absolute constant, as the specific choice of ε does not influence the resulting bound.

Concentration: Considering a fixed vector $v \in S^{d-1}$, then the inner product $\langle X, v\rangle$ is a zero-mean sub-exponential random variable—coupled with the equivalence of Proposition 2.7.1 (a) and (d) from Vershynin—yields the following probability bound for any positive u:

$$\Pr(\langle X, v\rangle \geq u) \leq \exp\left(\frac{-cu}{\|\langle X, v\rangle\|_{\psi_1}}\right) \leq \exp\left(\frac{-cu}{\|\langle X, v\rangle\|_{L_2}}\right) \leq \exp\left(\frac{-cu}{\sqrt{v^\top\Sigma v}}\right)$$

$$\Pr(\langle X, v\rangle \geq u) \leq \exp\left(\frac{-cu}{\sqrt{\lambda_{\max}(\Sigma)}}\right)$$

where c is an absolute constant.

Union Bound: By applying the union bound across the elements of the ε-net, we derive the following

$$\Pr\left(\max_{v\in\mathcal{N}_\varepsilon}\langle X, v\rangle > u\right) \leq \sum_{v\in\mathcal{N}_\varepsilon}\Pr(\langle X, v\rangle > u) \leq N_\varepsilon \exp\left(\frac{-cu}{\sqrt{\lambda_{\max}(\Sigma)}}\right).$$

Parameter Selection: Corollary 4.2.13 from Vershynin ensures the existence of an ε-net satisfying $N_\varepsilon \leq (1 + 2/\varepsilon)^d$. Setting $u = C\sqrt{\lambda_{\max}(\Sigma)}(d + t)$ for a suitably large absolute constant C such that $cC \leq C$, we obtain the following bound:

$$\Pr\left(\max_{v\in\mathcal{N}_\varepsilon}\langle X, v\rangle > u\right) \leq N_\varepsilon \exp\left(\frac{-cu}{\sqrt{\lambda_{\max}(\Sigma)}}\right) \leq (1 + 2/\varepsilon)^d \exp\left(\frac{-cC\sqrt{\lambda_{\max}(\Sigma)}(d + t)}{\sqrt{\lambda_{\max}(\Sigma)}}\right)$$

$$\Pr\left(\max_{v\in\mathcal{N}_\varepsilon}\langle X, v\rangle > u\right) \leq \exp\big(d\log(1 + 2/\varepsilon)\big)\exp\left(-C(d + t)\right) = \exp\big(d\log(1 + 2/\varepsilon) - C(d + t)\big)$$

$$\Pr\left(\max_{v\in\mathcal{N}_\varepsilon}\langle X, v\rangle > u\right) \leq \exp(-C(d + t)) \leq \exp(-t).$$

Now we set $\delta = \exp(-t)$ to solve for t in terms of δ. As such, we have that $t = \log(1/\delta)$. Thus, we have that with probability at least $1 - \delta$,

$$\Pr\left(\max_{v\in\mathcal{N}_\varepsilon}\langle X, v\rangle > u\right) \leq \delta \implies \max_{v\in\mathcal{N}_\varepsilon}\langle X, v\rangle \leq \max_{v\in S^{d-1}}\langle X, v\rangle = \|X\| \leq u = C\sqrt{\lambda_{\max}(\Sigma)}(d + \log(1/\delta)).$$

Comparison with the Variational Approach:

The bound reveals that the ε-net approach exhibits a dependency on the dimension d. Defining $r(\Sigma)$ as the effective rank of Σ, i.e., $r(\Sigma) = \mathrm{Tr}(\Sigma)/\lambda_{\max}(\Sigma)$, we can compare the two approaches. The variational approach can be rewritten as

$$\|X\| \leq C\sqrt{\lambda_{\max}(\Sigma)}\left(\sqrt{r(\Sigma)\log(1/\delta)} + \log(1/\delta)\right).$$

However it is still difficult to compare the two approaches directly. If we return to our derivation of the variational approach, we can set $\beta = r(\Sigma)$ to get that

$$\|X\| \leq C\sqrt{\lambda_{\max}(\Sigma)}\left(r(\Sigma) + \log(1/\delta)\right).$$

This shows that the variational approach is more efficient than the ε-net approach, as $r(\Sigma) \leq d$. Importantly, sometimes the effective rank $r(\Sigma)$ can be much smaller than the dimension d, which would make the variational approach significantly more efficient.

Problem 4 (Gaussian matrix series)

Assume that $g_1, \dots, g_n$ are independent standard Gaussian random variables and let $A_1, \dots, A_n$ be a sequence of deterministic symmetric d by d matrices.

1. Using the matrix Chernoff bound, show that
$$\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\| \le \sqrt{2\log(2d)}\left\|\sum_{i=1}^n A_i^2\right\|.$$
2. Show that
$$\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|^2 \ge \left\|\sum_{i=1}^n A_i^2\right\|.$$
3. Use Gaussian concentration to prove a high probability upper bound on $\left\|\sum_{i=1}^n g_i A_i\right\|$.
4. Compare the tails you achieve with the tails that follow from the matrix Chernoff bound.
5. Assume that A_i are rank one matrices of the form $A_i = u_i u_i^T$, where $u_i \in S^{d-1}$. Show that for some absolute constant $c > 0$ we can bound
$$\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\| \le c\sqrt{n}.$$
Note that this bound does not depend on $\log(d)$.
6. * Provide a collection of matrices $A_1, \dots, A_n$, showing that the multiplicative $\log(d)$-factor cannot be improved in general.

Hint: You might find it helpful to upper bound the operator norm in terms of the Frobenius norm. In the last bullet point you might find it useful to provide a lower bound on the maximum of n independent standard Gaussian random variables.

Theorem HW2.P3.1

Let $g_1, \dots, g_n$ represent n independent standard Gaussian random variables and let $A_1, \dots, A_n$ be a sequence of n deterministic symmetric d by d matrices. Then we have that
$$\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\| \le \sqrt{2\log(2d)}\left\|\sum_{i=1}^n A_i^2\right\|.$$

Proof. We begin by recalling the matrix Chernoff bound.

Lemma HW2.P3.9 (Matrix Chernoff bound)**.** *Consider a finite sequence $\mathbf{A}_k$ of fixed square matrices in $\mathbb{R}^{d\times d}$ and let ξ_k be a finite sequence of independent standard normal random variables. Then, for all*
$$\Pr\left\{\left\|\sum_k \xi_k \mathbf{A}_k\right\|_{\text{op}} \ge t\right\} \le d \cdot e^{-t^2/2\sigma^2} \quad \text{where} \quad \sigma^2 = \left\|\sum_k \mathbf{A}_k^2\right\|_{\text{op}}.$$

It then follows that

$$\Pr\left\{\left\|\left\|\sum_i^n g_i A_i\right\|_{\text{op}}\right| \geq t\right\} \leq \underbrace{2d \cdot e^{-t^2/2\sigma^2}}_{\delta} \implies \delta = 2d \cdot e^{-t^2/2\sigma^2}$$

$$\log(\delta/2d) = -t^2/2\sigma^2$$

$$t = \sqrt{2\sigma^2 \log(2d/\delta)}$$

Therefore with non-zero probability, we have that

$$\boxed{\mathbb{E}\left\|\sum_i^n g_i A_i\right\|_{\text{op}} \leq \sqrt{2\log(2d)\left\|\sum_{i=1}^n A_i^2\right\|_{\text{op}}}.}$$

□

Theorem HW2.P3.2

Let $g_1, \ldots, g_n$ represent n independent standard Gaussian random variables and let $A_1, \ldots, A_n$ be a sequence of n deterministic symmetric d by d matrices. Then we have that

$$\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}}^2 \geq \left\|\sum_{i=1}^n A_i^2\right\|_{\text{op}}.$$

Proof. To begin, by applying Jensen's inequality twice we have that

$$\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}}^2 \geq \mathbb{E}\left\|\left(\sum_{i=1}^n g_i A_i\right)^2\right\|_{\text{op}} \geq \left\|\mathbb{E}\left(\sum_{i=1}^n g_i A_i\right)^2\right\|_{\text{op}} = \left\|\mathbb{E}\sum_{i=1}^n\sum_{j=1}^n g_i g_j A_i A_j\right\|_{\text{op}}.$$

We note that all of the cross-terms cancel due to the independence of g_i for all $i \in [n]$, which implies that

$$\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}}^2 \geq \left\|\sum_{i=1}^n \mathbb{E}g_i^2 A_i^2\right\|_{\text{op}} = \left\|\sum_{i=1}^n A_i^2\right\|_{\text{op}} \implies \boxed{\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}}^2 \geq \left\|\sum_{i=1}^n A_i^2\right\|_{\text{op}}.}$$

□

Theorem HW2.P3.3

Let $g_1, \ldots, g_n$ represent n independent standard Gaussian random variables and let $A_1, \ldots, A_n$ be a sequence of n deterministic symmetric d by d matrices. Then we have that

$$\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}} \leq \mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}} + \sqrt{2\log(2/\delta)\sum_{i=1}^n \|A_i\|_{\text{op}}^2}$$

with at least probability $1 - \delta$.

We begin by recalling the Lemma we proved in class regarding Gaussian concentration.

Lemma HW2.P3.10. *If $X \sim \mathcal{N}_d(0, I_d)$ and $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz then,*

$$\Pr(f(X) - \mathbb{E}f(X) \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right) \quad \Leftrightarrow \quad \Pr(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\left(-\frac{t^2}{2L^2}\right). \tag{77}$$

We know utilize the above lemma with $X = g$ where $g := [g_1, ..., g_n]$ and $f(g) = \left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}}$. Then we have

$$\Pr\left(\left|\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}} - \mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}}\right| \geq t\right) \leq \underbrace{2\exp\left(-\frac{t^2}{2L^2}\right)}_{\delta}$$

where

$$L \geq \frac{|f(g^{(1)}) - f(g^{(2)})|}{\|g^{(1)} - g^{(2)}\|_2} \quad \text{for all} \quad g^{(1)}, g^{(2)} \sim \mathcal{N}(0, I_d).$$

Now we solve for t in terms of δ and arrive at

$$\delta = 2\exp\left(-t^2/2L^2\right) \implies \log(\delta/2) = -t^2/2L^2 \implies t = \sqrt{2L^2\log(2/\delta)}.$$

Then we have with probability at least $1 - \delta$

$$\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}} - \mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}} \leq \sqrt{2L^2\log(2/\delta)}$$

Finally we solve for L with

$$\begin{aligned}
\left|f(g^{(1)}) - f(g^{(2)})\right| = \left|\left\|\sum_{i=1}^n g_i^{(1)} A_i\right\|_{\text{op}} - \left\|\sum_{i=1}^n g_i^{(2)} A_i\right\|_{\text{op}}\right| &\leq \left\|\sum_{i=1}^n g_i^{(1)} A_i - \sum_{i=1}^n g_i^{(2)} A_i\right\|_{\text{op}} \\
&= \left\|\sum_{i=1}^n g_i^{(1)} A_i - g_i^{(2)} A_i\right\|_{\text{op}} \\
&= \left\|\sum_{i=1}^n \left(g_i^{(1)} - g_i^{(2)}\right) A_i\right\|_{\text{op}} \\
&\leq \sqrt{\sum_{i=1}^n \left(g_i^{(1)} - g_i^{(2)}\right)^2}\left\|\sum_{i=1}^n A_i^2\right\|_{\text{op}}
\end{aligned}$$

Therefore we are left with

$$\left|f(g^{(1)}) - f(g^{(2)})\right| \leq \left\|g_i^{(1)} - g_i^{(2)}\right\|_2 \sqrt{\sum_{i=1}^n \|A_i\|_{\text{op}}^2} \implies L = \sqrt{\sum_{i=1}^n \|A_i\|_{\text{op}}^2} \geq \frac{\left|f(g^{(1)}) - f(g^{(2)})\right|}{\left\|g_i^{(1)} - g_i^{(2)}\right\|_2}$$

Substititing L^2 into our high probability bound results in that with probability at least $1 - \delta$

$$\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}} - \mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_{\text{op}} \leq \sqrt{2L^2\log(2/\delta)} \leq \sqrt{2\log(2/\delta)\sum_{i=1}^n \|A_i\|_{\text{op}}^2}$$

Theorem HW2.P3.5

Let $g_1, \ldots, g_n$ represent n independent standard Gaussian random variables and let $A_1, \ldots, A_n$ be a sequence of n one matrices of the form $A_i = u_i u_i^T$, where $u_i \in S^{d-1}$. Then we have that

$$\mathbb{E}\left\|\sum_{i=1}^{n} g_i A_i\right\|_{\text{op}} \leq \mathbf{c}\sqrt{n}$$

for some absolute constant $\mathbf{c} > 0$.

Proof. We begin with square of the left hand side and apply Jensen's inequality twice

$$\mathbb{E}\left\|\sum_{i=1}^{n} g_i A_i\right\|_{\text{op}}^2 = \mathbb{E}\left\|\sum_{i=1}^{n} g_i u_i u_i^\top\right\|_{\text{op}}^2 \geq \left\|\mathbb{E}\left(\sum_{i=1}^{n} g_i u_i u_i^\top\right)^2\right\|_{\text{op}} = \left\|\mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{n} g_i g_j u_i u_i^\top u_j u_j^\top\right\|_{\text{op}}$$

$$\text{(again by independence of } g_i \text{ for all } i \in [n]) \quad = \left\|\sum_{i=1}^{n} \mathbb{E} g_i^2 u_i u_i^\top u_i u_i^\top\right\|_{\text{op}}$$

which implies that

$$\mathbb{E}\left\|\sum_{i=1}^{n} g_i A_i\right\|_{\text{op}}^2 \geq \left\|\sum_{i=1}^{n} \mathbb{E} g_i^2 u_i u_i^\top\right\|_{\text{op}} = \left\|\sum_{i=1}^{n} u_i u_i^\top\right\|_{\text{op}} = \sup_{v \in S^{d-1}} v^\top \left(\sum_{i=1}^{n} u_i u_i^\top\right) v = \sup_{v \in S^{d-1}} \sum_{i=1}^{n} v^\top u_i u_i^\top v$$

$$= \sup_{v \in S^{d-1}} \sum_{i=1}^{n} (u_i^\top v)^2$$

$$\mathbb{E}\left\|\sum_{i=1}^{n} g_i A_i\right\|_{\text{op}}^2 \geq \sum_{i=1}^{n} (u_i^\top v)^2 \quad \text{for all } v \in S^{d-1}.$$

Since the last inequality holds for $v \in S^{d-1}$, we set v to $-u_i$ for all $i \in [n]$ such that

$$\mathbb{E}\left\|\sum_{i=1}^{n} g_i A_i\right\|_{\text{op}}^2 \geq -n \implies \boxed{\mathbb{E}\left\|\sum_{i=1}^{n} g_i A_i\right\|_{\text{op}} \leq \mathbf{c}\sqrt{n}}$$

where $\mathbf{c} = 1 > 0$ is an absolute constant. □

Problem 5 (Non-asymptotic analysis of ridge regression)

Consider the random design linear model

$$Y = \langle X, \beta^* \rangle + \xi,$$

where β^* and X are in $\mathbb{R}^d$, ξ is a zero mean random noise with variance σ^2 independent of X. Let $\Sigma = \mathbb{E}XX^T$ and assume that it is invertible. Assume that $\|X\| \le r$ with probability one, where $r > 0$ is some constant. We observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent copies of the random pair (X, Y). Fix $\lambda > 0$ and consider the ridge regression estimator

$$\hat{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^d} \left(\frac{1}{n} \sum_{i=1}^n (Y_i - \langle \beta, X_i \rangle)^2 + \lambda \|\beta\|^2 \right) = (\hat{\Sigma}_n + \lambda I_d)^{-1} \cdot \frac{1}{n} \sum_{i=1}^n Y_i X_i,$$

where $\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$ is the empirical matrix of second moments and I_d is the d by d identity matrix. We are interested in upper bounding

$$\mathbb{E} \left\| \Sigma^{1/2} (\hat{\beta}_\lambda - \beta^*) \right\|_2^2,$$

where the expectation is taken with respect to the random observations (X_i, Y_i), $i = 1, \ldots, n$.

Part 1. Show the following decomposition

$$\mathbb{E} \left\| \Sigma^{1/2} \left(\hat{\beta}_\lambda - \beta^* \right) \right\|_2^2 \le \lambda^2 \mathbb{E} \left(\beta^{*\top} (\hat{\Sigma}_n + \lambda I_d)^{-1} \Sigma (\hat{\Sigma}_n + \lambda I_d)^{-1} \beta^* \right) + \frac{\sigma^2}{n} \mathbb{E} \operatorname{Tr} \left((\hat{\Sigma}_n + \lambda I_d)^{-1} \Sigma \right). \quad (78)$$

Part 2. We want to understand how much we lose if we replace the sample covariance matrix $\hat{\Sigma}_n$ by the population matrix Σ in the above formulas. Our final goal will be to show that we lose at most a small multiplicative factor. Returning to the population level quantities, show that

$$\lambda^2 \left(\beta^{*\top} (\Sigma + \lambda I_d)^{-1} \Sigma (\Sigma + \lambda I_d)^{-1} \beta^* \right) \le \lambda \left(\beta^{*\top} (\Sigma + \lambda I_d)^{-1} \Sigma \beta^* \right) = \inf_{\beta \in \mathbb{R}^d} \left(R(\beta) + \lambda \|\beta\|^2 - R(\beta^*) \right).$$

where $R(\beta) = \mathbb{E}(Y - \langle X, \beta \rangle)^2$ is the prediction risk of β. This would imply that the population analog of Equation (78) is

$$\inf_{\beta \in \mathbb{R}^d} \left(R(\beta) + \lambda \|\beta\|^2 - R(\beta^*) \right) + \frac{\sigma^2}{n} \operatorname{Tr} \left((\Sigma + \lambda I_d)^{-1} \Sigma \right).$$

Part 3. Quantify the error from replacing $\hat{\Sigma}_n$ with Σ. Start with the second term. We want to show that for all $v \in S^{d-1}$,

$$\mathbb{E} v^\top \Sigma^{1/2} (\hat{\Sigma}_n + \lambda I_d)^{-1} \Sigma^{1/2} v \le (1 + \Delta)(v^\top \Sigma^{1/2} (\Sigma + \lambda I_d)^{-1} \Sigma^{1/2} v),$$

where $\Delta = \Delta(r, n, \lambda)$ is a "small" term. To do so, apply the matrix Bernstein inequality to analyze the matrix $(\Sigma + \lambda I_d)^{-1/2} (\hat{\Sigma}_n + \lambda I_d)(\Sigma + \lambda I_d)^{-1/2}$, and obtain control over its inverse. This will imply uniform control over $v^\top \Sigma^{1/2} (\hat{\Sigma}_n + \lambda I_d)^{-1} \Sigma^{1/2} v$ and the entire second term.

Part 4. * Repeat a similar analysis with a possibly different error term Δ to replace the sample covariance by the population covariance in the first term of the upper bound Equation (78).

Part 5. Try to interpret (informally) the final bound by discussing the performance of ridge regression depending on λ.

Hints: For the first inequality we need to exploit the explicit formulas for Y and $\hat{\beta}_\lambda$ together with the fact that ξ is independent of X. You might also need to use that $(\hat{\Sigma}_n+\lambda I_d)^{-1}\hat{\Sigma}_n(\hat{\Sigma}_n+\lambda I_d)^{-1} \preceq (\hat{\Sigma}_n + \lambda I_d)^{-1}$ together with $\mathrm{Tr}(AB) \leq \mathrm{Tr}(AC)$ for PSD matrices A, B, C with $B \preceq C$.

Regrade Justification: I did not have time to complete this problem. Below is my solution that closely follows the provided solution.

STAT210B Theoretical Statistics April 5th, 2024

Homework # 3: Empirical Processes and Applications

Reece D. Huff

Regrade

I did not submit this homework originally, so I am submitting it now. I have completed most of the problems.

Notation

Let **c** and **C** represent a small and large positive constant, respectively (e.g., $\mathbf{c} = 10^{-5}$ and $\mathbf{C} = 10^{5}$). Unless otherwise specified, we use the notation $[n]$ to represent the set of integers $\{1, \ldots, n\}$.

Problem 1 (VC dimension)

Part 1. Show that the VC dimension of the set induced by axis aligned rectangles in $\mathbb{R}^p$ is equal to $2p$.

Part 2. Show that the VC dimension of the family of sets induced by all convex polygons on the real plane, without any restriction on the number of vertices, has infinite VC dimension.

Part 3. For a scalar $t \in \mathbb{R}$, consider the class of functions $\mathcal{F} = \{x \mapsto \text{sign}(\sin(tx)) : t \in \mathbb{R}\}$. Prove that $\mathcal{F}$ has infinite VC dimension. (This shows that VC dimension is not always equal to the number of parameters of a function class.)

Hint: For 2, you may start with a collection of points on the unit circle.

Part 1. Axis-aligned rectangles.

In this analysis, we delve into the VC dimension of various function classes:

- **Axis-aligned rectangles in $\mathbb{R}^p$.** The VC dimension of this family is deduced to be precisely $2p$. Consdier the positive point set $P = \{e_i\}_{i=1}^p$ in $\mathbb{R}^p$ and the negative point set $P' = \{-e_i\}_{i=1}^p$. We define a set T as the union of positive and negative point sets, $T = P \cup P'$ (containing $2p$ points). Any subset S of T can be uniquely selected by an axis-aligned rectangle within the subspace defined by S. Now consider the next largest set that contains all of these elements, but one more arbitruary point $T' = P \cup P' \cup \{x\}$. Any subset S' of T' can be shattered by an axis-aligned rectangle in $\mathbb{R}^{p+1}$, as the additional dimension allows for the inclusion of the extra point. However, the set T' cannot be shattered by any axis-aligned rectangle in $\mathbb{R}^p$, as the extra point will always be excluded. This implies that the VC dimension of axis-aligned rectangles in $\mathbb{R}^p$ is $2p$.

- **Families of all convex polygons.** For any integer n, consider the set $P = \{(\sin(2\pi k/n), \cos(2\pi k/n)) : k \in [n]\}$, representing n equidistant points on the unit circle. Any subset S of these points can be shattered by a convex polygon crafted by connecting the points sequentially, thereby including S and excluding its complement S^c.

Formally, let $S = \{p_1, \ldots, p_k\}$ be a subset of P. Let $x = \sum_{i=1}^{k} \alpha_i p_i$ be a convex combination of the points in S with $\sum_i^k \alpha_k = 1$. Let y be any point on the unit circle not in S. Then y is not in the convex hull of S, because

$$x \cdot y = \left(\sum_{i=1}^{k} \alpha_i p_i \right) \cdot y \leq \sum_{i=1}^{k} \alpha_i |p_i \cdot y| < 1.$$

Therefore the convex hull of S can shatter any subset of P, implying that the VC dimension of the family of sets induced by all convex polygons on the real plane is infinite.

- TODO

Problem 2 (Classification and population risk bounds)

Consider the binary classification problem with feature space $\mathcal{X} \subseteq \mathbb{R}^p$ and two classes $\mathcal{Y} = \{1, -1\}$. Assume that there is some unknown probability distribution $P_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$. Let $\eta(x) = \mathbb{E}[Y|X = x]$. The Bayes optimal rule f_B^* is given by $f_B^*(x) = \text{sign}(\eta(x))$ (assume that $\eta(X) \neq 0$ with probability one). Show the following:

Part 1. The Bayes optimal classifier indeed minimizes the population risk $R(f) = \Pr(f(X) \neq Y)$ among all (measurable) functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. Here, the probability is computed with respect to $P_{X,Y}$.

Part 2. Show that for any classifier $f : \mathcal{X} \to \mathcal{Y}$ it holds that

$$R(f) - R(f_B^*) = \mathbb{E}[|\eta(X)| \cdot \mathbb{I}\{f(X) \neq f_B^*(X)\}].$$

Part 3. Assume that there is a finite set of classifiers $\mathcal{F}$ such that $f_B^* \in \mathcal{F}$. Assume also that there is some $h > 0$ such that $|\eta(X)| \geq h$ with probability one. Show that there is an absolute constant c such that, with probability $1 - \delta$, (where $\delta \in (0, 1/2)$), it holds that

$$R(\hat{f}) - R(f_B^*) \leq c \frac{\log(|\mathcal{F}|) + \log(1/\delta)}{nh},$$

where

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{I}\{f(X_i) \neq Y_i\}.$$

is any empirical risk minimizer constructed via the i.i.d sample $(X_i, Y_i)_{i=1}^n$ sampled from $P_{X,Y}$. Compare this result with the bound we obtained earlier in the class in the special case where $Y = f_B^*(X)$ with probability one.

Hint: In the proof of 3, apply the Bernstein inequality together with the union bound to the set of functions $\{(x, y) \mapsto \mathbb{I}[f(x) \neq y] - \mathbb{I}[f_B^*(x) \neq y] : f \in \mathcal{F}\}$.

Part 1.

Proof. Let $f : \mathcal{X} \to \mathcal{Y}$ be any measurable function. The population risk $R(f)$ is defined as the probability that $f(X)$ does not equal Y, that is,

$$R(f) = \Pr(f(X) \neq Y) = \mathbb{E}\left[\mathbb{1}_{f(X) \neq Y}\right],$$

where $\mathbb{1}_{\{f(X) \neq Y\}}$ is the indicator function that is 1 if $f(X) \neq Y$ and 0 otherwise.
By the law of total probability, we have

$$\begin{aligned}\Pr(f(X) \neq Y|X = x) &= \Pr(f(x) \neq Y|X = x, Y = 1) \Pr(Y = 1|X = x) \\ &\quad + \Pr(f(x) \neq Y|X = x, Y = -1) \Pr(Y = -1|X = x) \\ \Pr(f(X) \neq Y|X = x) &= \mathbb{1}_{f(x) \neq 1} \Pr(Y = 1|X = x) + \mathbb{1}_{f(x) \neq -1} \Pr(Y = -1|X = x)\end{aligned}$$

which can be simplified to

$$\Pr(f(x) \neq Y|X = x) = \begin{cases} \Pr(Y = -1|X = x) & \text{if } f(x) = 1, \\ \Pr(Y = 1|X = x) & \text{if } f(x) = -1. \end{cases}$$

In order to minimize the above quantity, we should choose $f(x)$ to be 1 if $\eta(x) > 0$ and -1 if $\eta(x) < 0$. This is the Bayes optimal rule $f_B^*(x) = \text{sign}(\eta(x))$. □

Part 2.

Proof. To establish the relationship between the risk of any classifier f and the Bayes optimal classifier f_B^*, we start by expressing the risk difference:

$$R(f) - R(f_B^*) = \Pr(f(X) \neq Y) - \Pr\big(f_B^*(X) \neq Y\big) \quad = \mathbb{E}\left[\mathbb{1}_{f(X)\neq Y}\right] - \mathbb{E}\left[\mathbb{1}_{f_B^*(X)\neq Y}\right] = \mathbb{E}\left[\mathbb{1}_{f(X)\neq Y} - \mathbb{1}_{f_B^*(X)\neq Y}\right].$$

Now, we note that $\mathbb{1}_{f(X)\neq Y} - \mathbb{1}_{f_B^*(X)\neq Y}$ is nonzero only when $f(X) \neq f_B^*(X)$, and in such cases, it equals $|\eta(X)|$ since $\eta(X)$ is nonzero with probability one. Therefore, we have:

$$R(f) - R(f_B^*) = \mathbb{E}\left[|\eta(X)| \cdot \mathbb{1}_{f(X)\neq f_B^*(X)}\right].$$

This completes the proof. □

Part 3.

Proof. Given the optimality of $\hat{f}$, we have $R_n(\hat{f}) \leq R_n(f_B^*)$. Let us define the excess risk process as $E(f) = R(f) - R(f_B^*)$ and its empirical counterpart as $E_n(f) = R_n(f) - R_n(f_B^*)$. By the optimality of $\hat{f}$, we note that the excess risk of $\hat{f}$ as

$$E(\hat{f}) \leq E(\hat{f}) - E_n(\hat{f}).$$

Fix a classifier $f \in \mathcal{F}$ and denote $Z_i(f) := \mathbb{1}_{f(X_i)\neq Y_i} - \mathbb{1}_{f_B^*(X_i)\neq Y_i}$. The difference $E(f) - E_n(f)$ can be written as $\mathbb{E}[Z(f)] - \frac{1}{n}\sum_{i=1}^n Z_i(f)$, where $Z(f) := Z_1(f)$.
Applying Bernstein's inequality, we find that with probability at least $1 - \delta$,

$$E(f) - E_n(f) \leq \sqrt{\frac{2\sigma_f^2 \log(1/\delta)}{n}} + \frac{2}{3}B_f \frac{\log(1/\delta)}{n},$$

where $\mathrm{Var}(Z(f)) \leq \sigma_f^2$ and $|Z(f) - \mathbb{E}[Z(f)]| \leq B_f$. We can take $B_f = 2$, but σ_f^2 should depend on f.
Since $Z(f)^2 = \mathbb{1}_{f(X)\neq f_B^*(X)}$, and using the Massart noise condition $|\eta(X)| \geq h$, we have

$$\mathrm{Var}(Z(f)) \leq \mathbb{E}\left[Z(f)^2\right] = \Pr\big(f(X) \neq f_B^*(X)\big) \leq \frac{1}{h}E[|\eta(X)| \cdot \mathbb{1}_{f(X)\neq f_B^*(X)}] = \frac{1}{h}E(f) := \sigma_f^2.$$

Substituting these values into the Bernstein bound and using the union bound over all $f \in \mathcal{F}$, we get with probability at least $1 - \delta$,

$$E(f) - E_n(f) \leq \sqrt{\frac{2E(f)\log(|\mathcal{F}|/\delta)}{nh}} + \frac{2\log(|\mathcal{F}|/\delta)}{n}.$$

For the empirical risk minimizer $\hat{f}$, this implies

$$E(\hat{f}) \leq E(\hat{f}) - E_n(\hat{f}) \leq \sqrt{\frac{2E(f)\log(|\mathcal{F}|/\delta)}{nh}} + \frac{2\log(|\mathcal{F}|/\delta)}{n} \leq 2\max\left\{\sqrt{\frac{2E(\hat{f})\log(|\mathcal{F}|/\delta)}{nh}}, \frac{2\log(|\mathcal{F}|/\delta)}{n}\right\}$$

Solving for $E(\hat{f})$ when the first term achieves the maximum gives $E(\hat{f}) \leq \frac{8\log(|\mathcal{F}|/\delta)}{nh}$. When the second term achieves the maximum, we have $E(\hat{f}) \leq \frac{4\log(|\mathcal{F}|/\delta)}{n} \leq \frac{4\log(|\mathcal{F}|/\delta)}{nh}$, since $h \leq 1$.

Overall, with probability at least $1 - \delta$, we have

$$E(\hat{f}) = R(\hat{f}) - R(f_B^*) \leq 8\frac{\log(|\mathcal{F}|) + \log(1/\delta)}{nh}.$$

In the "noiseless" setting where $Y = f_B^*(X)$, we have $R(f_B^*) = 0$ and $|\eta(X)| = 1$ with probability 1, allowing us to take $h = 1$. In this case, the result simplifies to

$$R(\hat{f}) \leq 8\frac{\log(|\mathcal{F}|) + \log(1/\delta)}{n},$$

recovering the bound from Lecture 5 up to constants. Thus, our result is more general as it accounts for noise through the parameter h and reduces to the noiseless setting as $h \to 1$. □

Problem 3 (Empirical processes and random design linear regression)

Consider the random design linear regression model. That is, assume we observe an i.i.d. sample of $(X_i, Y_i)_{i=1}^n$ sampled according to some unknown distribution $P_{X,Y}$ over $\mathbb{R}^d \times \mathbb{R}$. For any $w \in \mathbb{R}^d$ define its population risk $R(w) = \mathbb{E}(Y - \langle X, w\rangle)^2$. Fix $b > 0$ and consider the constrained least squares regression problem to the Euclidean ball of radius b,

$$\hat{w} = \arg \min_{w \in \mathbb{R}^d : \|w\|_2 \le b} \frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, w\rangle)^2.$$

(Note, that we do not assume the model $Y = \langle X, w^*\rangle + \xi$, where ξ is an independent zero mean noise.)

Assume that there are absolute constants $m, r > 0$ such that with probability one, we have $|Y| \le m$ and $\|X\|_2 \le r$. Using the Dudley integral method, show that there is some absolute constant $c > 0$ such that, with probability at least $1 - \delta$,

$$R(\hat{w}) - \inf_{w \in \mathbb{R}^d : \|w\|_2 \le b} R(w) \le c \left(m^2 + r^2 b^2\right) \sqrt{\frac{d + \log(1/\delta)}{n}}.$$

Is there a way to improve this bound with other methods we studied? Can we remove the explicit dependence on the dimension?

Hint: You might require several tools we developed so far including symmetrization, contraction, and bounded differences inequality.

Proof. Let $w^* \in \arg\min_{w \in \mathbb{R}^d : \|w\|_2 \le b} R(w)$ be an optimal solution in the population risk sense. We want to bound the excess risk $R(\hat{w}) - R(w^*)$.
To begin, we recall in class (Proposition 16.1.) that we proved that

$$R(\hat{w}) - R(w^*) \le 2 \sup_{w \in \mathbb{R}^d : \|w\|_2 \le b} |R_n(w) - R(w)|$$

Then we define the function $G(Z) = G(Z_1, \dots, Z_n) := \sup_{w \in B_2^d(b)} |R_n(w) - R(w)|$. We will verify that G satisfies the bounded differences condition with some parameter L. First, using the inequality $(a+b)^2 \le 2a^2 + 2b^2$ and the Cauchy-Schwarz inequality, we observe that

$$(Y - \langle X, w\rangle)^2 \le 2Y^2 + 2\langle X, w\rangle^2 \le 2m^2 + 2r^2 b^2.$$

Next, we examine the difference $G(Z) - G(Z^{-i})$, where Z^{-i} denotes the sequence Z with the i-th element replaced:

$$\begin{aligned}
G(Z) - G(Z^{-i}) &= \sup_{w \in B_2^d(b)} |R_n(w; Z) - R(w)| - \sup_{v \in B_2^d(b)} |R_n(v; Z^{-i}) - R(v)| \\
&\le \sup_{w \in B_2^d(b)} |R_n(w; Z) - R(w) - (R_n(w; Z^{-i}) - R(w))| \\
&= \sup_{w \in B_2^d(b)} \left| \frac{1}{n}(Y_i - \langle X_i, w\rangle)^2 - \frac{1}{n}(Y_i' - \langle X_i', w\rangle)^2 \right| \\
&\le \frac{4(m^2 + r^2 b^2)}{n},
\end{aligned}$$

which shows that G satisfies the bounded differences condition with parameter $L = \frac{4(m^2+r^2b^2)}{n}$ uniformly. Applying the bounded differences inequality, we obtain that with probability at least $1-\delta$,

$$G(Z) - \mathbb{E}\left[G(Z)\right] \leq \sqrt{\frac{nL^2 \log(1/\delta)}{2}},$$

which can be simplified to

$$\sup_{w \in B_2^d(b)} |R_n(w) - R(w)| \leq \mathbb{E}\left[\sup_{w \in B_2^d(b)} |R_n(w) - R(w)|\right] + 4(m^2 + r^2b^2)\sqrt{\frac{\log(1/\delta)}{n}}.$$

Next, we apply symmetrization to the first term on the RHS. By introducing Rademacher variables σ_i (Lecture 14, Lemma 3), we have

$$\mathbb{E}_X\left[\sup_{\|w\|_2 \leq b}\left|\frac{1}{n}\sum_{i=1}^n (Y_i - \langle X_i, w\rangle)^2 - R(w)\right|\right] \leq 2\mathbb{E}_X\mathbb{E}_\sigma\left[\sup_{\|w\|_2 \leq b}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i (Y_i - \langle X_i, w\rangle)^2\right|\right].$$

We derive this bound via

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left[\sup_{w \in B_2^d(b)} \sum_{i=1}^n \sigma_i (Y_i - \langle X_i, w\rangle)^2\right] &\leq \frac{1}{n} \cdot 4M\mathbb{E}\left[\sup_{w \in B_2^d(b)} \sum_{i=1}^n \sigma_i (Y_i - \langle X_i, w\rangle)\right] \\
&\leq \frac{1}{n} \cdot 4M\left(\mathbb{E}\left[\sum_{i=1}^n \sigma_i Y_i\right] + \mathbb{E}\left[\sup_{w \in B_2^d(b)} \sum_{i=1}^n \sigma_i \langle X_i, w\rangle\right]\right) \\
&\leq \frac{1}{n} \cdot 4M\left(\sqrt{n}m + \sqrt{n}br\right) \\
&\leq 8\sqrt{2}\frac{m^2 + r^2b^2}{\sqrt{n}}.
\end{aligned}$$

The initial inequality leverages the contraction principle, Ledoux-Talagrand's Theorem 2 (Lecture 18), to eliminate the quadratic term. This is achieved by applying a $2M$-Lipschitz continuous function $\varphi : [-M, M] \to \mathbb{R}$, defined by $\varphi(x) = x^2$, where M is set to $\sqrt{2(m^2 + r^2b^2)}$.
The second step employs the triangle inequality to separate the empirical risk into two components.
The third step utilizes Jensen's inequality to the first term, and a combination of the Cauchy-Schwarz inequality with Jensen's inequality for the second term.
The final step is derived by noting that the quantities $\sqrt{m^2 + r^2b^2}$ and $br\sqrt{m^2 + r^2b^2}$ are both bounded above by $m^2 + r^2b^2$.
Combining the above steps, we obtain with probability at least $1-\delta$,

$$R(\hat{w}) - R(w^*) \leq c\left(m^2 + r^2b^2\right)\sqrt{\frac{d + \log(1/\delta)}{n}},$$

where c is an absolute constant that absorbs other constants.
To address the question of improving the bound or removing the explicit dependence on the dimension, we could consider other methods such as localized Rademacher complexities or covering numbers that adapt to the intrinsic dimensionality of the problem. For example, if the data lies in a lower-dimensional subspace or exhibits certain sparsity, we might be able to obtain tighter bounds that reflect these properties. □

Problem 4 (Gaussian width, Rademacher averages and Dudley integral)

Assume that $T \subseteq \mathbb{R}^d$. Let

$$R(T) = \mathbb{E} \sup_{t \in T} \sum_{i=1}^{d} \varepsilon_i t_i, \quad W(T) = \mathbb{E} \sup_{t \in T} \sum_{i=1}^{d} g_i t_i,$$

be the Rademacher averages and the Gaussian width of T respectively.

Part 1. Show that $R(T) \leq \sqrt{\frac{\pi}{2}} W(T)$.

Part 2. Compare the values of $R(T)$ and $W(T)$ when $T = B_1^d$ (i.e., it is the ℓ_1 unit ball).

Part 3. Compute the Gaussian width of the set of s-sparse vectors. That is, show that there is some absolute constant $c > 0$ such that

$$W(T) \leq c \sqrt{s \log\left(\frac{ed}{s}\right)},$$

where

$$T = \{x \in \mathbb{R}^d : \|x\|_0 \leq s, \|x\|_2 \leq 1\},$$

Part 4. Let T be a convex hull of the set $\left\{ \frac{e_i}{\sqrt{1+\log(i)}} : i = 1, \ldots, d \right\}$, where e_i is the i-th standard basis vector. Show that the Dudley integral upper bound is not sharp in this case. That is, as d grows, the Dudley integral

$$\int_0^\infty \sqrt{\log(N(T, \|\cdot\|_2, \varepsilon))}\, d\varepsilon$$

goes to infinity, while $W(T)$ is bounded by an absolute constant for all d.

Hint: For 1, it might be helpful to use that $\mathbb{E}|g_i| = \sqrt{\frac{2}{\pi}}$. For 4, one can lower bound the size of the covering numbers using appropriate packing numbers, which are easier to estimate (from below).

Part 1.

Proof. Consider independent standard Gaussian variables $g = (g_1, \ldots, g_d)$ and independent Rademacher variables $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$. The Gaussian width $W(T)$ can be expressed as

$$W(T) = \mathbb{E}\left[\sup_{t \in T} \sum_{i=1}^{d} g_i t_i \right].$$

Since g_i can be decomposed as $g_i \overset{d}{=} \varepsilon_i |g_i|$, we have

$$W(T) = \mathbb{E}\left[\sup_{t \in T} \sum_{i=1}^{d} \varepsilon_i |g_i| t_i \right].$$

Applying Jensen's inequality to the convex supremum function, we obtain

$$W(T) \geq \mathbb{E}\left[\mathbb{E}\left[\sup_{t \in T} \sum_{i=1}^{d} \varepsilon_i |g_i| t_i \,\middle|\, \varepsilon \right] \right] = \sqrt{\frac{\pi}{2}} \mathbb{E}\left[\sup_{t \in T} \sum_{i=1}^{d} \varepsilon_i t_i \right] = \sqrt{\frac{\pi}{2}} R(T),$$

where the last equality follows from the definition of Rademacher averages $R(T)$. □

Part 2.

Proof. For any vector $v \in \mathbb{R}^d$ and $\theta \in B_1^d$, the ℓ_1 unit ball, Hölder's inequality gives $\langle \theta, v \rangle \leq \|\theta\|_1 \|v\|_\infty = \|v\|_\infty$. There exists $\theta^* \in B_1^d$ such that $\langle \theta^*, v \rangle = \|v\|_\infty$, specifically the vector that selects the largest absolute value entry of v. Thus, we have $\sup_{\theta \in B_1^d} \langle \theta, v \rangle = \|v\|_\infty$.

Consequently, the Rademacher averages for B_1^d are

$$R(B_1^d) = \mathbb{E}\left[\sup_{t \in B_1^d} \langle \varepsilon, t \rangle\right] = \mathbb{E}\|\varepsilon\|_\infty = 1,$$

since $\|\varepsilon\|_\infty = 1$ for Rademacher sequences.

For the Gaussian width, we have

$$W(B_1^d) = \mathbb{E}\left[\sup_{t \in B_1^d} \langle g, t \rangle\right] = \mathbb{E}\|g\|_\infty,$$

where $\|g\|_\infty$ is the maximum absolute value of the Gaussian entries. Using bounds on the expected maximum of Gaussian random variables, we obtain

$$\sqrt{\log d} \leq \mathbb{E}\|g\|_\infty \leq \sqrt{2\log(2d)} \leq 2\sqrt{\log d},$$

for $d \geq 12$. This implies $W(B_1^d) \leq \sqrt{\log d} R(B_1^d)$ so long as $d \geq 12$. □

Part 3.

Proof. Let $S \subseteq \{1, \ldots, d\}$ and define $A_d^S(1) = \{v \in \mathbb{R}^d : v = v_S, \|v\|_2 = 1\}$, where v_S is the projection of v onto the coordinates in S. The set T is the union of $A_d^S(1)$ over all subsets S of size s.

The Gaussian width of T is given by

$$W(T) = \mathbb{E}\left[\sup_{t \in T} \langle g, t \rangle\right] = \mathbb{E}\left[\max_{|S|=s} \sup_{t \in A_d^S(1)} \langle g_S, t_S \rangle\right] = \mathbb{E}\left[\max_{|S|=s} \|g_S\|_2\right].$$

Using Jensen's inequality and the fact that $\|\cdot\|_2$ is 1-Lipschitz, we have

$$\mathbb{E}\left[\max_{|S|=s} \|g_S\|_2\right] \leq \sqrt{\mathbb{E}\left[\max_{|S|=s} \|g_S\|_2^2\right]}.$$

The expectation inside the square root can be bounded by considering the subgaussian property of $\|g_S\|_2 - \mathbb{E}\|g_S\|_2$ and the number of subsets of size s. We obtain

$$\mathbb{E}\left[\max_{|S|=s} \|g_S\|_2^2\right] \leq \mathbb{E}\|g_S\|_2^2 + 2\log\binom{d}{s} \leq s + 2s\log\left(\frac{ed}{s}\right),$$

where the last inequality uses the bound $\binom{d}{s} \leq \left(\frac{ed}{s}\right)^s$. Thus, we have

$$W(T) \leq \sqrt{s + 2s\log\left(\frac{ed}{s}\right)} \leq c\sqrt{s\log\left(\frac{ed}{s}\right)},$$

for some absolute constant $c > 0$. □

Part 4.

TODO

Problem 5 (Catoni's mean estimator)

Consider the problem of estimating the mean of a random variable that is assumed to have a finite second moment. The objective is to construct a non-asymptotic mean estimator that exhibits sub-Gaussian tail behavior, ideally mirroring the classical Central Limit Theorem (CLT) asymptotic rate. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a continuous, non-decreasing function satisfying

$$-\log\left(1 - x + \frac{x^2}{2}\right) \le \psi(x) \le \log\left(1 + x + \frac{x^2}{2}\right).$$

Given i.i.d. random variables $X_1, \dots, X_n$ with mean μ and variance σ^2, we define the estimator $\hat{\mu}$ as the root of the equation

$$\sum_{i=1}^{n} \psi(\lambda(X_i - \hat{\mu})) = 0, \tag{3.1}$$

where λ is a positive tuning parameter. For a fixed $\delta \in (0, 1/2)$, assume $n > 2\log(1/\delta)$ and set

$$\eta = \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n\left(1 - 2\log(1/\delta)/n\right)}}, \quad \lambda = \frac{2\log(1/\delta)}{n(\sigma^2 + \eta^2)}.$$

Note that λ depends on both δ and σ. Show that with probability at least $1 - 2\delta$,

$$|\mu - \hat{\mu}| \le \eta.$$

The leading term in this bound, $\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}$, closely mirrors the tail bound for estimating the mean of i.i.d. Gaussian variables using the sample mean, with the leading constant $\sqrt{2}$ precisely matching the optimal rate. Note that in your proof, the residual term $\sqrt{\frac{1}{1-2\log(1/\delta)/n}}$ might be slightly larger.

Hint: Begin by examining the exponential moments of (for any ν)

$$\frac{1}{n\lambda} \sum_{i=1}^{n} \psi(\lambda(X_i - \nu)).$$

Then, argue that $\hat{\mu}$ is bracketed by solutions to two equations similar to (3.1), though obtained using high-probability bounds from the exponential moments analysis. You may also employ the inequality $1 + x \le \exp(x)$.

The proof leverages the following lemma:

Lemma HW3.P5.1. *Consider a continuous stochastic process $Z(t)$, monotonic decreasing in $t \in \mathbb{R}$. Define functions $U(t)$ and $L(t)$ such that for each t,*

$$\Pr(Z(t) \le U(t)) \ge 1 - \delta \quad \textit{and} \quad \Pr(Z(t) \ge L(t)) \ge 1 - \delta.$$

For zeros t^+ of U and t^- of L, and a well-defined zero $\hat{t}$ of Z, the following holds:

$$\Pr\left(\hat{t} \le t^+\right) \ge 1 - \delta \quad \textit{and} \quad \Pr\left(\hat{t} \ge t^-\right) \ge 1 - \delta.$$

Proof: The analysis begins with the case of t^+, leveraging the implications:

$$Z(t^*) \leq U(t^*) \implies Z(t^*) \leq 0 \implies Z(t^*) \leq Z(\hat{t}) \implies \hat{t} \leq t^*.$$

Hence, it follows:

$$1 - \delta \leq \Pr(Z(t^*) \leq U(t^*)) \leq \Pr\big(\hat{t} \leq t^*\big).$$

Start by determining functions $U(\nu)$ and $L(\nu)$ that ensure:

$$L(\nu) \leq \frac{1}{n} \sum_{i=1}^{n} \psi(\lambda(X_i - \nu)) \leq U(\nu)$$

holds with probability at least $1 - 2\delta$. Establish roots μ^+ and μ^- where $U(\mu^+) = L(\mu^-) = 0$. By the Lemma, $\mu^- \leq \hat{\mu} \leq \mu^+$ holds with probability $1 - 2\delta$.
Calculations for $U(\nu)$: Examine the exponential moment for $X_d = X_1$. Given ψ's upper limit,

$$\mathbb{E}\left[\exp(\lambda\psi(X - \nu))\right] \leq \exp\left(\lambda(\mu - \nu) + \frac{\lambda^2}{2}(\sigma^2 + (\mu - \nu)^2)\right).$$

Chernoff bounds imply:

$$\Pr\left(\sum_{i=1}^{n} \lambda\psi(X_i - \nu) \geq n\lambda u\right) \leq \exp\left(n\lambda(\mu - \nu) + \frac{n\lambda^2}{2}(\sigma^2 + (\mu - \nu)^2) - n\lambda u\right) = \delta.$$

Solving gives the bound for $U(\nu)$:

$$\frac{1}{n} \sum_{i=1}^{n} \psi(\lambda(X_i - \nu)) \leq \lambda(\mu - \nu) + \frac{\lambda^2}{2}(\sigma^2 + (\mu - \nu)^2) + \frac{\log(1/\delta)}{n} := U(\nu).$$

Taking $\nu = \mu + \eta$ where η is as defined, ensures $\nu = \mu + \eta$ satisfies $U(\nu) = 0$. The monotonicity in ν implies by the Lemma that $\Pr(\mu \leq \hat{\mu} + \eta) \geq 1 - \delta$.
By a similar argument for $L(\nu)$ and applying a union bound, we conclude:

$$|\hat{\mu} - \mu| \leq \eta$$

holds with probability at least $1 - 2\delta$.

STAT210B Theoretical Statistics May 1st, 2024

Homework # 4: Empirical Processes and Applications

Anonymous Author

Notation

Let **c** and **C** represent a small and large positive constant, respectively (e.g., $\mathbf{c} = 10^{-5}$ and $\mathbf{C} = 10^{5}$). Unless otherwise specified, we use the notation $[n]$ to represent the set of integers $\{1, ..., n\}$.

Problem 1 (Covering numbers for star-shaped hulls)

Let $\mathcal{F}$ be a class of functions absolutely bounded by 1. That is, for any $f \in \mathcal{F}$ and $x \in \mathcal{X}$, we have $|f(x)| \leq 1$. Let star$(\mathcal{F})$ denote the star-shaped hull of $\mathcal{F}$ around zero (i.e., the set $\{\alpha f : f \in \mathcal{F}, \alpha \in [0,1]\}$). Show that for any $\varepsilon > 0$,

$$\log \mathcal{N}(\text{star}(\mathcal{F}), L_2(P), 2\varepsilon) \leq \log \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon) + \log\left(\frac{2}{\varepsilon}\right).$$

That is, the covering numbers for star-shaped hulls are approximately the same as for the original class.

Proof. Let $\mathcal{N}(\mathcal{F}, L_2(P), \varepsilon)$ denote the smallest number of balls of radius ε in the $L_2(P)$ metric required to cover the class $\mathcal{F}$. We want to show that

$$\log \mathcal{N}(\text{star}(\mathcal{F}), L_2(P), 2\varepsilon) \leq \log \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon) + \log\left(\frac{2}{\varepsilon}\right).$$

Let $\{f_1, f_2, \ldots, f_N\}$ be an ε-cover for $\mathcal{F}$ in the $L_2(P)$ metric, where $N = \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon)$. This means that for any $f \in \mathcal{F}$, there exists some f_i such that $\|f - f_i\|_{L_2(P)} < \varepsilon$.
Consider the star-shaped hull of $\mathcal{F}$, denoted as star$(\mathcal{F})$. For any function $g \in \text{star}(\mathcal{F})$, there exists an $\alpha \in [0,1]$ and $f \in \mathcal{F}$ such that $g = \alpha f$. Since $\{f_1, f_2, \ldots, f_N\}$ is an ε-cover for $\mathcal{F}$, there exists an f_i such that $\|f - f_i\|_{L_2(P)} < \varepsilon$.
Now, consider the function $g_i = \alpha f_i$. We have

$$\|g - g_i\|_{L_2(P)} = \|\alpha f - \alpha f_i\|_{L_2(P)} = \alpha \|f - f_i\|_{L_2(P)} < \alpha\varepsilon \leq \varepsilon.$$

To cover star$(\mathcal{F})$ with balls of radius 2ε, we can use the functions $\{g_1, g_2, \ldots, g_N\}$ along with a discretization of the interval $[0,1]$ into points $\{\alpha_j\}$ such that the distance between consecutive points is at most ε. The number of such points is at most $\lceil 1/\varepsilon \rceil \leq 1/\varepsilon + 1$.
Therefore, the covering number for star$(\mathcal{F})$ can be bounded by the product of the covering number for $\mathcal{F}$ and the number of points in the discretization of $[0,1]$, which gives us

$$\mathcal{N}(\text{star}(\mathcal{F}), L_2(P), 2\varepsilon) \leq N\left(\frac{1}{\varepsilon} + 1\right) \leq \frac{2N}{\varepsilon}.$$

Taking the logarithm of both sides, we obtain

$$\log \mathcal{N}(\text{star}(\mathcal{F}), L_2(P), 2\varepsilon) \leq \log N + \log\left(\frac{2}{\varepsilon}\right) = \log \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon) + \log\left(\frac{2}{\varepsilon}\right).$$

This completes the proof. □

Problem 2 (d/n rate for random design linear regression)

We are going to improve the linear regression bounds from the previous homework assignment.

Consider the random design linear regression model. That is, assume we observe an i.i.d. sample of $(X_i, Y_i)_{i=1}^n$ sampled according to some unknown distribution $P_{X,Y}$ over $\mathbb{R}^d \times \mathbb{R}$. For any $w \in \mathbb{R}^d$, define its population risk $R(w) = \mathbb{E}\left[(Y - \langle X, w\rangle)^2\right]$. Fix $b > 0$ and consider the constrained (to the Euclidean ball of radius b) least squares estimator

$$\hat{w} = \operatorname*{argmin}_{\substack{w \in \mathbb{R}^d \\ \|w\|_2 \le b}} \frac{1}{n} \sum_{i=1}^n \big(Y_i - \langle X_i, w\rangle\big)^2.$$

Assume that there are absolute constants $m, r > 0$ such that with probability one, we have $|Y| \le m$ and $\|X\|_2 \le r$. Using the offset term for Rademacher averages as in the lectures, show that for some absolute constant $c > 0$,

$$\mathbb{E}\left[R(\hat{w})\right] - \inf_{\substack{w \in \mathbb{R}^d \\ \|w\|_2 \le b}} R(w) \le \frac{cd\big(m^2 + r^2 b^2\big)}{n}.$$

Here, the expectation is taken with respect to the training sample.

Hint: You might need to directly bound the process without using the Dudley integral. If needed, you may assume without loss of generality that the sample covariance matrix (empirical matrix of second moments) is invertible.

Proof. We consider the random design linear regression model where we observe an i.i.d. sample of $(X_i, Y_i)_{i=1}^n$ sampled according to some unknown distribution $P_{X,Y}$ over $\mathbb{R}^d \times \mathbb{R}$. We are given that $\|X_i\|_2 \le r$ and $|Y_i| \le m$ with probability one.
We respectively define the *population risk* and *emperical risk* for any $w \in \mathbb{R}^d$ as:

$$R(w) := \mathbb{E}[(Y - \langle X, w\rangle)^2], \quad \text{and} \quad R_n(w) := \frac{1}{n}\sum_{i=1}^n (Y_i - \langle X_i, w\rangle)^2.$$

Additionally, we consider the constrained least squares estimator:

$$\hat{w} = \operatorname*{argmin}_{\substack{w \in \mathbb{R}^d \\ \|w\|_2 \le b}} \frac{1}{n}\sum_{i=1}^n (Y_i - \langle X_i, w\rangle)^2.$$

Our goal is to bound the expected excess risk,

$$\mathbb{E}\left[\mathcal{E}(w)\right] := \mathbb{E}[R(\hat{w})] - \inf_{\substack{w \in \mathbb{R}^d \\ \|w\|_2 \le b}} R(w).$$

To bound the excess risk $\mathbb{E}\left[\mathcal{E}(w)\right]$, we recall our Proposition from Lecture 23. Specifically, recall the equation right before Proposition 23.2.

$$\mathbb{E}\left[\mathcal{E}(w)\right] \le 20m\mathbb{E}\sup_{f \in \mathcal{F}} \left\{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i) - f^\star(X_i)) - \frac{1}{50m} P_n(f - f^\star)^2\right\}$$

where $P_n(f - f^\star)^2 := \frac{1}{n}\big(f(X_i) - f^\star(X_i)\big)^2$ is the empirical measure, $f^\star$ is the minimizer of the population risk, and $\mathcal{F}$ is the class of functions we are considering, and $m > 0$ is a constant such that $|Y| \le m$ and $|f(X)| \le m$ for all $f \in \mathcal{F}$. Then we have the following lemma:

Lemma HW4.P2.1. *Consider the random design linear regression model. Let $(X_i, Y_i)_{i=1}^n$ be an i.i.d. sample from some unknown distribution $P_{X,Y}$ over $\mathbb{R}^d \times \mathbb{R}$. Assume that with probability one, we have $\max\{|Y|, |f(X)|\} \le K$ for all $f \in \mathcal{F}$. Then, for some absolute constant $C = 20$ and $c = 1/50$, we have*

$$\mathbb{E}[R(\hat{f})] - R(f^\star) \le CK\mathbb{E}_{P_{X,Y}}\mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i) - f^\star(X_i)) - \frac{c}{K}\left\|f - f^\star\right\|_{L_2(P_n)}^2\right\}\right].$$

Here σ_i are Rademacher random variables. The optimal function $f^\star$ is the minimizer of the population risk $R(f)$ and the optimal function $\hat{f}$ is the minimizer of the empirical risk $R_n(f)$.

Now we may apply this lemma to our problem. To begin we note that $f(X_i) - f^\star(X_i) = (Y_i - \langle X_i, w\rangle) - (Y_i - \langle X_i, w^\star\rangle) = \langle X_i, w^\star - w\rangle$. Then we have

$$\begin{aligned}
\mathbb{E}[R(\hat{f})] - R(f^\star) &\le CK\mathbb{E}_{P_{X,Y}}\mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i) - f^\star(X_i)) - \frac{c}{K}\left\|f - f^\star\right\|_{L_2(P_n)}^2\right\}\right] \\
&= CK\mathbb{E}_{P_{X,Y}}\mathbb{E}_{\sigma_i}\left[\sup_{\substack{w\in\mathbb{R}^d \\ \|w\|_2\le b}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i\left\langle X_i, w^\star - w\right\rangle - \frac{c}{K}\left(\left\langle X_i, w^\star - w\right\rangle\right)^2\right\}\right]
\end{aligned}$$

We note that we can pull out the $1/n$, and that $\langle X_i, w^\star - w\rangle = -\langle X_i, w - w^\star\rangle$ while $(\langle X_i, w^\star - w\rangle)^2 = (\langle X_i, w - w^\star\rangle)^2$. Then we have the following:

$$= \frac{CK}{n}\mathbb{E}_{P_{X,Y}}\mathbb{E}_{\sigma_i}\left[\inf_{\substack{w\in\mathbb{R}^d \\ \|w\|_2\le b}}\left\{\sum_{i=1}^n \underbrace{\sigma_i\left\langle X_i, w - w^\star\right\rangle + \frac{c}{K}\left(\left\langle X_i, w - w^\star\right\rangle\right)^2}_{f_i(w)}\right\}\right]$$

We now minimize $f_i(w)$ with respect to w. We note that $f_i(w)$ is a quadratic function in w and thus we can find the minimum by taking the derivative and setting it equal to zero. We have

$$\begin{aligned}
\sum_{i=1}^n \nabla_w f_i(w) &= \sum_{i=1}^n \left(\sigma_i X_i + \frac{2c}{K}X_i X_i^\top w - \frac{2c}{K}X_i X_i^\top w^\star\right) = 0 \\
\sum_{i=1}^n X_i X_i^\top (w - w^\star) &= \sum_{i=1}^n \frac{K}{2c}\sigma_i X_i
\end{aligned}$$

Now we multiply both sides by $\left(X_j X_j^\top\right)^{-1}$ for an arbitruary $j \in [n]$. We have

$$w - w^\star = \frac{K}{2c}\left(X_j X_j^\top\right)^{-1}\sigma_j X_j + \underbrace{\sum_{\substack{i=1 \\ i\ne j}}^n \left(\frac{K}{2c}\left(X_j X_j^\top\right)^{-1}\sigma_i X_i - \left(X_j X_j^\top\right)^{-1} X_i X_i^\top (w - w^\star)\right)}_{\text{cross terms}}$$

We note that the cross terms will eventually cancel out by the independence of X_i and X_j for $i \ne j$ when we take the expectation $\mathbb{E}_{P_{X,Y}}[\,\cdot\,]$. Thus we have the following:

$$\inf_{\substack{w\in\mathbb{R}^d \\ \|w\|_2\le b}}\left\{\sum_{i=1}^n f_i(w)\right\} = \sigma_j\left\langle X_j, \frac{K}{2c}\left(X_j X_j^\top\right)^{-1}\sigma_j X_j\right\rangle + \frac{c}{K}\left(\left\langle X_j, \frac{K}{2c}\left(X_j X_j^\top\right)^{-1}\sigma_j X_j\right\rangle\right)^2$$

Again, note that the sum essentially vanishes as we are only interested in the j-th term and all other terms will cancel out. It then follows that

$$\inf_{\substack{w\in\mathbb{R}^d \\ \|w\|_2\le b}} \left\{\sum_{i=1}^{n} f_i(w)\right\} = \frac{K\sigma_j^2}{2c} X_j^\top \left(X_j X_j^\top\right)^{-1} X_j + \frac{K\sigma_j^2}{4c}\left(X_j^\top \left(X_j X_j^\top\right)^{-1} X_j\right)^2$$

We note that $\Pi = X_j^\top \left(X_j X_j^\top\right)^{-1} X_j$ is defined such that $\Pi^2 = X_j^\top \left(X_j X_j^\top\right)^{-1} X_j X_j^\top \left(X_j X_j^\top\right)^{-1} X_j = \Pi$. Then we have the following:

$$\inf_{\substack{w\in\mathbb{R}^d \\ \|w\|_2\le b}} \left\{\sum_{i=1}^{n} f_i(w)\right\} = \frac{K\sigma_i^2}{2c}\Pi + \frac{K\sigma_i^2}{4c}\Pi = \frac{3K\sigma_i^2}{4c}\Pi = \frac{K\sigma_i^2}{c'} X_j^\top \left(X_j X_j^\top\right)^{-1} X_j \quad \text{where} \quad c' = \frac{4}{3}c$$

Finally, we note that $X_j^\top \left(X_j X_j^\top\right)^{-1} X_j = \mathrm{Tr}\left(X_j^\top \left(X_j X_j^\top\right)^{-1} X_j\right) = \mathrm{Tr}\left(X_j X_j^\top \left(X_j X_j^\top\right)^{-1}\right) = \mathrm{Tr}(I_d) = d$. Thus we have the following:

$$\mathbb{E}[R(\hat{f})] - R(f^\star) \le \frac{CK}{n}\mathbb{E}_{P_{X,Y}}\mathbb{E}_{\sigma_i}\left[\frac{dK\sigma_j^2}{c'}\right] = \frac{dCK^2}{c'n}$$

Recalling $c' = 4/3c$, $c = 1/50$, and $C = 20$, we have the following:

$$\mathbb{E}[R(\hat{f})] - R(f^\star) \le \frac{60K^2}{200n} = \frac{c^\star K^2}{n} \quad \text{where} \quad c^\star = \frac{3}{10}.$$

To finish, we simply need to show what K is. We are given that $|Y_i| \le m$, $\|X_i\|_2 \le r$, and $\|w\|_2 \le b$. Then we have

$$\max\{|Y|, |f(X)|\} \le m + rb \le K \implies m^2 + r^2b^2 \le K^2.$$

Thus we have the following:

$$\mathbb{E}[R(\hat{f})] - R(f^\star) \le \frac{c^\star K^2}{n} \le \frac{c^\star\left(m^2 + r^2b^2\right)}{n} = \frac{3}{10}\frac{m^2 + r^2b^2}{n}.$$

□

Problem 3 (Regression with expressive non-parametric classes)

Consider the non-parametric linear regression problem with random design given by the model

$$Y = f^*(X) + \xi,$$

where f^* belongs to some known convex class $\mathcal{F}$. Assume that $\max\{|\xi|, |Y|\} \leq m$ and $|f(X)| \leq m$ for all $f \in \mathcal{F}$, and that ξ is independent of X and ξ is zero mean. Finally, we assume that the class $\mathcal{F}$ is non-parametric in the sense that

$$\log \mathcal{N}(\mathcal{F}, L_2(P_n), \epsilon) \leq C\epsilon^{-p},$$

where C is some constant and $p > 2$. The assumption $p > 2$ corresponds to expressive classes of functions for which the uniform convergence at rate $\frac{1}{\sqrt{n}}$ is not possible.

As before, we are given an i.i.d. sample of $(X_i, Y_i^n)_{i=1}^n$ sampled according to some unknown distribution $P_{X,Y}$ over $\mathbb{R}^d \times \mathbb{R}$.

Part 1. Using the offset term for Rademacher averages, show the upper bound (the best upper bound you can get with this technique up to multiplicative constant factors) on $\mathbb{E}[R(\hat{f})] - R(f^*)$, where $\hat{f}$ is an empirical risk minimizer in $\mathcal{F}$ (i.e., $\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2$.)

Part 2. Our goal is now to improve the above Dudley integral-based bound. First, show that for any $f \in \mathcal{F}$, it holds that

$$R(f) - R(f^*) = \|f - f^*\|^2_{L_2(P_X)}.$$

Part 3. Assume that when building your estimator, you have access to the distribution P_X (but not $P_{Y|X}$), and for any $\epsilon > 0$ you can build the smallest ϵ-net with respect to the $L_2(P_X)$ distance (denote this set by $N(\mathcal{F}, \epsilon)$; we assume $N(\mathcal{F}, \epsilon) \subseteq \mathcal{F}$). Show that there is a choice of the value of ϵ such that the predictor

$$\hat{f}_\epsilon = \operatorname*{argmin}_{f \in N(\mathcal{F}, \epsilon)} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2$$

leads to the classical rate of convergence $n^{-\frac{2}{p+2}}$ for $R(\hat{f}_\epsilon) - R(f^*)$ despite being in the regime $p > 2$ (recall that this rate was shown in the lecture for the standard least squares in $\mathcal{F}$ but only for $p \in (0, 2)$).

PROBLEM 3

$(X_i, Y_i)_{i=1}^n$ is i.i.d. over $P_{X,Y} \in \mathbb{R}^d \times \mathbb{R}$

$R(w) = \mathbb{E}\left[(Y - X^T w)^2\right]$ is the population risk.

$R_n(w) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^T w)^2$ is the empirical risk.

$\hat{w} = \underset{\substack{w \in \mathbb{R}^d \\ \|w\|_2 \le b}}{\arg\min} \{ R_n(w) \}$ least squares estimator

$w^* = \underset{\substack{w \in \mathbb{R}^d \\ \|w\|_2 \le b}}{\arg\inf} \{ R(w) \}$ optimal solution

$Y = f^*(X) + \xi$ ξ is zero mean

$\max\{|Y|, |\xi|\} \le m \quad |f(X)| \le m$

$\log N(\mathcal{F}, L_2(P_n), \varepsilon) \le C \varepsilon^{-p}$ for some constant $C > 0$ and $p > 2$.

- Using the offset term for Rademacher averages, show the upper bound (the best upper bound you can get with this technique up to multiplicative constant factors) on $\mathbb{E}[R(\widehat{f})] - R(f^\star)$, where $\widehat{f}$ is an empirical risk minimizer in $\mathcal{F}$ (i.e., $\widehat{f} = \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2$.)
- Our goal is now to improve the above Dudley integral-based bound. First, show that for any $f \in \mathcal{F}$, it holds that
$$R(f) - R(f^\star) = \|f - f^\star\|^2_{L_2(P_X)}.$$
- Assume that when building your estimator, you have access to the distribution P_X (but not $P_{Y|X}$), and for any $\varepsilon > 0$ you can build the smallest ε-net with respect to the $L_2(P_X)$ distance (denote this set by $\mathcal{F}_\varepsilon$, and its cardinality by $\mathcal{N}(\mathcal{F}, \varepsilon)$; we assume $\mathcal{F}_\varepsilon \subseteq \mathcal{F}$). Show that there is a choice of the value of ε such that the predictor
$$\widehat{f}_\varepsilon = \text{argmin}_{f \in \mathcal{F}_\varepsilon} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2$$
leads to the classical rate of convergence $n^{-\frac{2}{p+2}}$ for $R(\widehat{f}_\varepsilon) - R(f^\star)$ despite being in the regime $p > 2$ (recall that this rate was shown in the lecture for the standard least squares in $\mathcal{F}$ but only for $p \in (0, 2)$).

Part 1. Recall

Proposition 23.1. *For the random design regression above, for any $\alpha, \gamma \geq 0$ such that $\alpha < \gamma$, we have*

$$\mathbb{E}\left[R(\hat{f})\right] - \inf_{f \in \mathcal{F}} R(f) \leq \mathbb{E}\left[Cm\left(\alpha + \frac{1}{\sqrt{n}}\int_{\alpha}^{\gamma}\sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)}\,d\varepsilon + \frac{m \log \mathcal{N}(\mathcal{F}, L_2(P_n), \gamma)}{n}\right)\right] \quad (37)$$

where C is an absolute constant.

where $L_2(P_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bullet_i)^2}$ and $L_2(P) = \sqrt{\mathbb{E}(\bullet)^2}$.

We have

$$\mathbb{E}\left[R(\hat{w})\right] - R(w^*) \leq Cm\,\mathbb{E}\left[\alpha + \frac{1}{\sqrt{n}}\int_{\alpha}^{\gamma}\sqrt{\log e^{C\varepsilon^{-p}}}\,d\varepsilon + \frac{m}{n}\log e^{C\gamma^{-p}}\right]$$

$$= Cm\alpha + \frac{C'm}{\sqrt{n}}\int_{\alpha}^{\gamma}\varepsilon^{-p/2}\,d\varepsilon + \frac{Cm}{n}\gamma^{-p}$$

$$= C''m\underbrace{\left[\alpha + \frac{1}{\sqrt{n}}\int_{\alpha}^{\gamma}\varepsilon^{-p/2}\,d\varepsilon + \frac{1}{n}\gamma^{-p}\right]}_{\text{select } \alpha \text{ and } \gamma \text{ to minimize this}}$$

$$\alpha + \frac{1}{\sqrt{n}}\int_{\alpha}^{\gamma}\varepsilon^{-p/2}\,d\varepsilon + \frac{1}{n}\gamma^{-p} =$$

$$\alpha + \frac{1}{\sqrt{n}}\int_{\alpha}^{u}\varepsilon^{-p/2}\,d\varepsilon + \frac{1}{\sqrt{n}}\int_{u}^{\gamma}\varepsilon^{-p/2}\,d\varepsilon + \frac{1}{n}\gamma^{-p}$$

$$\frac{\partial}{\partial\alpha}\left\{\alpha + \frac{1}{\sqrt{n}}\int_{\alpha}^{u}\varepsilon^{-p/2}\,d\varepsilon\right\} = 0 \Rightarrow 1 - \frac{\alpha^{-p/2}}{\sqrt{n}} = 0 \Rightarrow \alpha^* = n^{-1/p}$$

$$\frac{\partial}{\partial\gamma}\left\{\frac{1}{\sqrt{n}}\int_{u}^{\gamma}\varepsilon^{-p/2}\,d\varepsilon + \frac{1}{n}\gamma^{-p}\right\} = 0 \Rightarrow \frac{\gamma^{-p/2}}{\sqrt{n}} - \frac{p}{n}\gamma^{-p-1} = 0 \Rightarrow \gamma^* = n^{-\frac{1}{2+p}}$$

$$\mathbb{E}\left[R(\hat{w})\right] - R(w^*) \leq C''m\left(n^{-1/p} + \frac{1}{\sqrt{n}}\left[(-p/2+1)\varepsilon^{p/2+1}\right]_{\varepsilon=n^{-1/p}}^{\varepsilon=n^{-\frac{1}{p}+2}} + \frac{1}{n}n^{-\frac{p}{2+p}}\right)$$

$$= C''m\left(n^{-1/\rho} + \frac{1}{\sqrt{n}}\left[(-\rho/2+1)\left(n^{\frac{(\frac{-\rho+2}{2})(\frac{-1}{\rho+2})}{}} - n^{\frac{(\frac{-\rho+2}{2})(\frac{-1}{\rho})}{}}\right)^{\frac{-\rho}{2+\rho}-1} + n\right]\right)$$

$$= C''m\left(n^{-1/\rho} + \frac{1}{\sqrt{n}}\left[(-\rho/2+1)\left(n^{1/2} - n^{\frac{\rho-2}{2\rho}}\right) + n^{\frac{-\rho-\rho-2}{\rho+2}}\right]\right)$$

$$= C''m\left(n^{-1/\rho} + \frac{1}{\sqrt{n}}\left[(-\rho/2+1)\left(n^{1/2} - n^{1/2} n^{-1/\rho}\right) + n^{\frac{-(\rho+2)-\rho}{\rho+2}}\right]\right)$$

$$= C''m\left(-\rho/2\, n^{-1/\rho} + n^{-2/(\rho+2)}\right] \leq C''m\, n^{-1/\rho}$$

Part 2. $R(f) - R(f^*) \leq \|f - f^*\|_{L_2(P_X)}$

$$\mathbb{E}_{XY}\left[(Y - f(X))^2\right] - \mathbb{E}_{XY}\left[(Y - f^*(X))^2\right]$$

$$= \mathbb{E}_{XY}\left[Y^2 + f(X)^2 - 2Yf(X)\right] - \mathbb{E}_{XY}\left[Y^2 + f^*(X)^2 - 2Yf^*(X)\right]$$

$$= \mathbb{E}_X\left[f(X)^2 - f^*(X)^2 - 2f^*(X)f(X) + 2f^*(X)^2\right] \quad \text{since } \xi \text{ is zero mean}$$

$$= \mathbb{E}_X\left[f(X)^2 + f^*(X)^2 - 2f(X)f^*(X)\right] = \mathbb{E}_X\left[(f(X) - f^*(X))^2\right]$$

$$R(f) - R(f^*) \leq \|f(X) - f^*(X)\|^2_{L_2(P_X)}$$

Part 3. $\hat{f}_\varepsilon = \underset{f \in \mathcal{F}_\varepsilon}{\operatorname{argmin}} \{R_n(f)\}$ and $f^*_\varepsilon = \underset{f \in \mathcal{F}_\varepsilon}{\operatorname{argmin}} \{R(f)\}$

$R(\hat{f}_\varepsilon) - R(f^*) \leq ?$

Note that $R(\hat{f}_\varepsilon) - R(f^*) = R(\hat{f}_\varepsilon) - R(f_\varepsilon^*) + R(f_\varepsilon^*) - R(f^*)$

$\leq R(\hat{f}_\varepsilon) - R(f_\varepsilon^*) + \| f_\varepsilon^* - f^* \|_{L_2(P_X)}$

$\leq R(\hat{f}_\varepsilon) - R(f_\varepsilon^*) + \varepsilon^2$

Then

$$\mathbb{E}\left[R(\hat{f}_\varepsilon)\right] - R(f_\varepsilon^*) \leq \frac{C_m}{\sqrt{n}} \mathbb{E} \int_0^\infty \sqrt{\log N(\mathcal{F}_\varepsilon, L_2(P_n), t)}\, dt$$

$$\leq \frac{C_m}{\sqrt{n}} \mathbb{E} \int_0^\infty \sqrt{\min\{t^{-p}, \varepsilon^{-p}\}}\, dt$$

$$\leq \frac{C_m}{\sqrt{n}} \left[\mathbb{E} \int_\varepsilon^\infty t^{-p/2} dt + \int_0^\varepsilon \varepsilon^{-p/2} dt \right]$$

$$= \frac{C}{\sqrt{n}} \varepsilon^{1-p/2}$$

Thus we have

$$R(\hat{f}_\varepsilon) - R(f^*) \leq R(\hat{f}_\varepsilon) - R(f_\varepsilon^*) + \varepsilon^2$$

$$\leq \underbrace{\frac{C}{\sqrt{n}} \varepsilon^{1-p/2} + \varepsilon^2}_{:= g(\varepsilon)}$$

minimize w.r.t ε.

$$\frac{\partial g}{\partial \varepsilon} = \frac{C}{\sqrt{n}} (1 - p/2) \varepsilon^{-p/2} + 2\varepsilon = 0 \quad \Rightarrow \quad \varepsilon = \frac{C}{2\sqrt{n}} (p/2 - 1) \varepsilon^{-p/2}$$

$$\varepsilon^{\frac{p+2}{2}} = \varepsilon^{1+p/2} = \frac{C}{2\sqrt{n}} (p/2 - 1) \asymp C n^{-1/2}$$

$$\varepsilon^* = C n^{-\frac{1}{p+2}}$$

$$R(\hat{f}_\varepsilon) - R(f^*) \leq \frac{C}{\sqrt{n}} n^{\frac{2-p}{2}\left(-\frac{1}{p+2}\right)} + C n^{-\frac{2}{p+2}} \leq C n^{-2/p+2}$$

Problem 4 (Estimation of Bernoulli mean in KL-distance)

Assume that we observe n independent Bernoulli random variables $X_1, ..., X_n$ with an unknown parameter $p \in [0,1]$. Instead of the absolute or quadratic loss, our aim is to construct $\hat{p}_n$ such that

$$\mathbb{E}KL(p, \hat{p}_n) = \mathbb{E}\left[(1-p)\log\left(\frac{1-p}{1-\hat{p}_n}\right) + p\log\left(\frac{p}{\hat{p}_n}\right)\right]$$

is as small as possible. Here, the expectation is taken with respect to the realization of the sample $X_1, ..., X_n$.

Part 1. Prove that the standard sample mean $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$ can lead to arbitrarily large values of $\mathbb{E}KL(p, \hat{p}_n)$.

Part 2. Using the exponential weights algorithm with logarithmic loss and using the uniform distribution over $[0,1]$ as a prior, construct an estimator $\hat{p}_n$, satisfying for some absolute constant $c > 0$,

$$\mathbb{E}KL(p, \hat{p}_n) \leq \frac{c\log(n)}{n}.$$

Part 3. Explain how your estimator is different from the sample mean.

Hint: You might need to use the following. For any integers n_1, n_2 such that $n = n_1 + n_2$, it holds

$$\int_0^1 p^{n_1}(1-p)^{n_2}dp = \frac{1}{(n+1)\binom{n}{n_1}}.$$

One way to prove this is through backward induction over n_1.

Part 1.

Proof. Consider the standard sample mean estimator $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$. We will show that $\mathbb{E}KL(p, \hat{p}_n)$ can be arbitrarily large for certain values of p.
First, note that the Kullback-Leibler divergence from p to $\hat{p}_n$ is given by

$$KL(p, \hat{p}_n) = (1-p)\log\left(\frac{1-p}{1-\hat{p}_n}\right) + p\log\left(\frac{p}{\hat{p}_n}\right).$$

This divergence becomes infinite if $\hat{p}_n = 0$ and $p > 0$ or if $\hat{p}_n = 1$ and $p < 1$, due to the logarithmic terms. Now, consider the case when p is very close to 0 (but not 0). The probability that all X_i are 0 (and hence $\hat{p}_n = 0$) is $(1-p)^n$, which is close to 1 for small p. Therefore, with high probability, $KL(p, \hat{p}_n)$ will approach infinity as n grows.
Similarly, when p is very close to 1, the probability that all X_i are 1 (and hence $\hat{p}_n = 1$) is p^n, which is close to 1 for p near 1. Again, with high probability, $KL(p, \hat{p}_n)$ will approach infinity as n grows.
In both cases, the expected value $\mathbb{E}KL(p, \hat{p}_n)$ can be arbitrarily large, which proves the statement. □

PROBLEM 4

$X_1, \ldots, X_n$ is i.i.d Bernoulli with parameter $p \in [0,1]$.

Aim is to construct $\hat{p}_n$ that minimizes

$$\mathbb{E}\, KL(p, \hat{p}_n) = \mathbb{E}\left[(1-p)\log\left(\frac{1-p}{1-\hat{p}_n}\right) + p\log\left(\frac{p}{\hat{p}_n}\right)\right]$$

Part 2 Goal: use exponential weights algorithm to show

$$\mathbb{E}\, KL(p, \hat{p}_n) \leq \frac{c\log n}{n}. \quad (1)$$

Define log loss as $l(f_\theta, x) = -\left(x\log\theta + (1-x)\log(1-\theta)\right)$ (2)

exponential weights algorithm

$$\hat{p}_k(\theta) = \frac{\exp\left(-\eta\sum_{i=1}^{k} l(f_\theta, x_i)\right)\pi(\theta)}{\mathbb{E}_{\theta\sim\pi}\left[\exp\left(-\eta\sum_{i=1}^{k} l(f_\theta, x_i)\right)\pi(\theta)\right]} \quad (3)$$

where η is learning rate
π prior
k time step

Vovk stated that a sequence of online predictors $\hat{f}_1, \ldots, \hat{f}_n$ satisfies the <u>mixability</u> condition for some $\eta > 0$ if

$$l(\hat{f}_k, x_k) \leq \frac{-1}{\eta}\log \mathop{\mathbb{E}}_{\theta\sim\hat{p}_{k-1}} \exp\left(-\eta\, l(f, x_k)\right) \quad (4)$$

Then we can bound the accumulated loss (Lecture 24) as

$$\sum_{k=1}^{n} l\left(\hat{f}_k, x_k\right) \leq -\frac{1}{\eta} \log \mathop{\mathbb{E}}_{\theta \sim \pi} \exp\left(-\eta \sum_{i=1}^{n} l(f_\theta, x_k)\right). \tag{5}$$

Set $\eta = 1$ and assume the mixability condition holds for $\hat{f}_1, \ldots, \hat{f}_n$ (Eq. (4)). Denote prior as $\pi = \text{Unif}[0,1]$ and the number of 1's as $m = \sum_{i=1}^{n} x_i$, then the r.h.s of Eq. (5) yields

$$-\log \mathop{\mathbb{E}}_{\theta \sim \pi} \exp\left(-\sum_{k=1}^{n} l(f_\theta, x_k)\right) = -\log \mathop{\mathbb{E}}_{\theta \sim \pi} \exp\left(m \log \theta - (n-m) \log(1-\theta)\right)$$

$$= -\log \mathop{\mathbb{E}}_{\theta \sim \pi} \left[\theta^m (1-\theta)^{n-m}\right]$$

$$= -\log \int_0^1 \theta^m (1-\theta)^{n-m} \, d\theta$$

$$= -\log\left(\frac{1}{(n+1)\binom{n}{m}}\right)$$

$$-\log \mathop{\mathbb{E}}_{\theta \sim \pi} \exp\left(-\sum_{k=1}^{n} l(f_\theta, x_k)\right) = \log(n+1) + \log\binom{n}{m}.$$

Now we solve for $\hat{\theta}_{MLE}$, i.e.,

$$\inf_f \left\{\sum_{k=1}^{n} l(f, x_i)\right\} = \inf_f \left\{-\sum_{k=1}^{n} \left(x_k \log\theta - (1-x_k)\log(1-\theta)\right)\right\}$$

$$= \sup_f \left\{\sum_{k=1}^{n} \left(x_k \log\theta - (1-x_k)\log(1-\theta)\right)\right\}$$

$$-\frac{\partial l}{\partial \theta} = \frac{x_k}{\theta} - \frac{1-x_k}{1-\theta} = 0 \Rightarrow \frac{m}{\theta} - \frac{n-m}{1-\theta} \Rightarrow (1-\theta)m = (n-m)\theta \Rightarrow \theta^* = \frac{m}{n} = \hat{\theta}_{MLE}$$

$$= -\sum_{k=1}^{n} \left(x_k \log\frac{m}{n} + (1-x_k)\log\left(1-\frac{m}{n}\right)\right)$$

$$= -m \log\frac{m}{n} - (n-m)\log\left(\frac{n-m}{n}\right)$$

$$= \log\left(\frac{n}{m}\right)^m + \log\left(\frac{n}{n-m}\right)^{n-m}$$

$$= \log\left(\left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m}\right) \geq \log\left(\frac{n}{m}\right)$$

Note: $\binom{n}{m} \leq \left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m}$ from $\frac{m^m}{m!}\frac{(n-m)^{n-m}}{(n-m)!} \leq \frac{n^n}{n!}$

Thus we have

$$R_n = \sum_{k=1}^{n} \ell(\hat{f}_k, x_k) - \inf_f \left\{\sum_{k=1}^{n} \ell(f, x_k)\right\}$$

$$\overset{\text{Eq. (5)}}{\leq} -\log \mathop{\mathbb{E}}_{\theta\sim\pi} \exp\left(-\sum_{i=1}^{n} \ell(f_\theta, x_k)\right) - \sum_{k=1}^{n} \ell(f_{MLE}, x_k)$$

$$\leq \log(n+1) + \log\binom{n}{m} - \log\left(\frac{n}{m}\right) = \log(n+1)$$

$R_n \leq c \log n$

Define $R(\theta) = \mathop{\mathbb{E}}_{X\sim \text{Bern}(p)} \ell(f_\theta, X) = -(p\log\theta + (1-p)\log(1-\theta))$

We note that

$$KL(p,\theta) = p\log(p/\theta) + (1-p)\log\left(\frac{1-p}{1-\theta}\right)$$

$$= \left(p\log p + (1-p)\log(1-p)\right) - \left(p\log\theta + (1-p)\log(1-\theta)\right)$$

$$= -R(p) + R(\theta)$$

because $p = \arg\inf_\theta R(\theta)$

$$= R(\theta) - \inf_\theta R(\theta)$$

Finally, we define $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_k$

Then $R(\hat{p}_n) - \inf_\theta R(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta, x_k) - \inf_\theta R(\theta)$

$$\leq \frac{R_n}{n} = \frac{c\log n}{n}$$

Thus $\boxed{\mathbb{E}kL(p, \hat{p}_n) \leq \frac{c\log n}{n}.}$

Part 3. we define $\hat{f}_k(k) = \mathbb{E}_{\theta \sim \hat{p}_{k-1}}[f_\theta(x)]$.

Now we note

$$\hat{p}_k(\theta) \propto \exp\left(-\sum_{i=1}^{k} l(f_\theta, x_i)\right)$$

$$= \exp\left(\sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta)\right)$$

$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

$$= \theta^{\alpha-1}(1-\theta)^{\beta}$$

$$= \text{Beta}(\alpha, \beta)$$

where $\alpha = 1 + \sum_{i=1}^{n} x_i$

$\beta = 1 + (n - \Sigma x_i)$

Thus the mean is $\alpha/(\alpha+\beta)$ i.e. $\hat{\theta}_k = \mathbb{E}_{\theta \sim \hat{p}_{k-1}}[\theta] = \frac{1+\Sigma_{i=1}^{k} x_i}{2+n}$.

Problem 5 (Hypercontractivity, quadratic forms, and linear regression)

Our final problem aims to achieve an error bound similar to the one in Problem 3, but focusing on the hypercontractivity of the distributions (and only the existence of four moments) instead of boundedness. As before, consider the random design linear regression model. We observe an i.i.d. sample of $(X_i, Y_i)_{i=1}^n$ sampled according to some unknown distribution $P_{X,Y}$ over $\mathbb{R}^d \times \mathbb{R}$. Let $w^* = \operatorname{argmin}_{w \in \mathbb{R}^d} R(w)$. Prove the following:

Part 1. (Hypercontractivity relations) Let $\xi \in \mathbb{R}$ be an additional variable. Show that if there are $L_1, L_2 > 1$ such that for all w,

$$\begin{aligned} \mathbb{E}\left[(Y - \langle X, w^* \rangle)^4\right]^{1/4} &\le L_1 \mathbb{E}\left[(Y - \langle X, w^* \rangle)^2\right]^{1/2} \quad \text{and} \\ \mathbb{E}\left[(\langle X, w \rangle)^4\right]^{1/4} &\le L_2 \mathbb{E}\left[(\langle X, w \rangle)^2\right]^{1/2}, \end{aligned} \tag{79}$$

then for some $L \le c(L_1 + L_2)$, where $c > 0$ is an absolute constant, it holds that for all $w \in \mathbb{R}^d$, $\xi \in \mathbb{R}$,

$$\mathbb{E}\left[(\xi Y - \langle X, w \rangle)^4\right]^{1/4} \le L \mathbb{E}\left[(\xi Y - \langle X, w \rangle)^2\right]^{1/2}.$$

Part 2. Denote $N(\xi, w) = \mathbb{E}\left[(\xi Y - \langle X, w \rangle)^2\right]$. Under the hypercontractivity assumptions Equation (79), use the median-of-means based estimator from the lectures to provide a quadratic form $\widehat{N}(\xi, w)$ such that, with probability at least $1 - \delta$,

$$\left| \frac{\widehat{N}(\xi, w)}{N(\xi, w)} - 1 \right| \le c_1 (L_1 + L_2)^2 \sqrt{\frac{d + \log(1/\delta)}{n}},$$

where $c_1 > 0$ is some absolute constant. Note that when constructing $\widehat{N}(\xi, w)$ based on the training sample, you can completely ignore the computational efficiency restrictions.

Part 3. Let

$$\hat{w} = \operatorname*{argmin}_{w \in \mathbb{R}^d} \widehat{N}(1, w).$$

Show that, with probability at least $1 - \delta$,

$$R(\hat{w}) - R(w^*) \le c_2 (L_1 + L_2)^4 \cdot R(w^*) \cdot \left(\frac{d + \log\left(\frac{1}{\delta}\right)}{n} \right),$$

provided that $n \ge c_3 (L_1 + L_2)^4 (d + \log(1/\delta))$, where $c_2, c_3 > 0$ are absolute constants.

Part 1.

Proof. We are given that for all $w \in \mathbb{R}^d$,

$$\mathbb{E}\left[(Y - \langle X, w^* \rangle)^4\right]^{1/4} \le L_1 \mathbb{E}\left[(Y - \langle X, w^* \rangle)^2\right]^{1/2} \quad \text{and} \quad \mathbb{E}\left[(\langle X, w \rangle)^4\right]^{1/4} \le L_2 \mathbb{E}\left[(\langle X, w \rangle)^2\right]^{1/2}.$$

We want to show that for some $L \le c(L_1 + L_2)$, where $c > 0$ is an absolute constant, it holds that for all $w \in \mathbb{R}^d$, $\xi \in \mathbb{R}$,

$$\mathbb{E}\left[(\xi Y - \langle X, w \rangle)^4\right]^{1/4} \le L \mathbb{E}\left[(\xi Y - \langle X, w \rangle)^2\right]^{1/2}.$$

Let us denote $Z = \xi Y - \langle X, w \rangle$. We can write Z as $Z = \xi(Y - \langle X, w^* \rangle) + \xi \langle X, w^* - w \rangle$. By Minkowski's inequality for the L^4 norm, we have

$$\mathbb{E}\left[Z^4\right]^{1/4} \leq \mathbb{E}\left[\xi^4 (Y - \langle X, w^* \rangle)^4\right]^{1/4} + \mathbb{E}\left[\xi^4 (\langle X, w^* - w \rangle)^4\right]^{1/4}.$$

Applying the hypercontractivity relations to each term, we get

$$\mathbb{E}\left[Z^4\right]^{1/4} \leq |\xi| L_1 \mathbb{E}\left[(Y - \langle X, w^* \rangle)^2\right]^{1/2} + |\xi| L_2 \mathbb{E}\left[(\langle X, w^* - w \rangle)^2\right]^{1/2}.$$

Now we note that the excess risk can be written as

$$\begin{aligned}
\mathcal{E}(w) &= R(w) - R(w^*) \\
&= \mathbb{E}\left[(Y - \langle X, w \rangle)^2\right] - \mathbb{E}\left[(Y - \langle X, w^* \rangle)^2\right] \\
&= \mathbb{E}\left[Y^2\right] + \mathbb{E}\left[(\langle X, w \rangle)^2\right] - 2\mathbb{E}\left[Y \langle X, w \rangle\right] - \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[(\langle X, w^* \rangle)^2\right] + 2\mathbb{E}\left[Y \langle X, w^* \rangle\right] \\
&= \mathbb{E}\left[(\langle X, w \rangle)^2\right] - \mathbb{E}\left[(\langle X, w^* \rangle)^2\right] - 2\mathbb{E}\left[Y \langle X, w - w^* \rangle\right] \\
&= \mathbb{E}\left[(\langle X, w - w^* \rangle)^2\right] - 2\mathbb{E}\left[(\langle X, w^* \rangle)^2\right] + 2\mathbb{E}\left[(\langle X, w \rangle)(\langle X, w^* \rangle)\right] - 2\mathbb{E}\left[(\langle X, w - w^* \rangle) Y\right] \\
\mathcal{E}(w) &= \mathbb{E}\left[(\langle X, w - w^* \rangle)^2\right] - 2w^{*\top}\mathbb{E}\left[XX^\top\right] w^* + 2w^{*\top}\mathbb{E}\left[XX^\top\right] w + 2w^\top \mathbb{E}\left[XY\right] - 2w^{*\top}\mathbb{E}\left[XY\right]
\end{aligned}$$

By the optimality of w^*, we have $\nabla R(w^*) = 0$, which implies that $\mathbb{E}\left[XX^\top\right] w^* = \mathbb{E}\left[XY\right]$. Therefore, we can simplify the above expression to

$$\mathcal{E}(w) = R(w) - R(w^*) = \mathbb{E}\left[(\langle X, w - w^* \rangle)^2\right].$$

Now we can return to our expression from the hypercontractivity relations and write

$$\begin{aligned}
\mathbb{E}\left[Z^4\right]^{1/4} &\leq |\xi| L_1 \mathbb{E}\left[(Y - \langle X, w^* \rangle)^2\right]^{1/2} + |\xi| L_2 \mathbb{E}\left[(\langle X, w^* - w \rangle)^2\right]^{1/2} = |\xi| \Big(L_1 R(w^*)^{1/2} + L_2 \mathcal{E}(w)^{1/2} \Big) \\
\mathbb{E}\left[Z^4\right]^{1/2} &\leq \xi^2 \left(L_1^2 R(w^*) + L_2^2 \mathcal{E}(w) + 2L_1 L_2 R(w^*)^{1/2} \mathcal{E}(w)^{1/2} \right) \\
&= \xi^2 \left(L_1^2 R(w^*) + L_2^2 \big(R(w) - R(w^*)\big) + 2L_1 L_2 \Big(R(w^*) \big(R(w) - R(w^*)\big) \Big)^{1/2} \right)
\end{aligned}$$

By optimality, we have that $R(w^*) \leq R(w)$ and $R(w^*) - R(w) \leq 0 \leq R(w)$ for all w. Thus we can simplify the above expression to

$$\begin{aligned}
\mathbb{E}\left[Z^4\right]^{1/2} &\leq \xi^2 \Big(L_1^2 R(w) + L_2^2 R(w) + 2L_1 L_2 R(w) \Big) = \xi^2 \Big((L_1 + L_2)^2 R(w) \Big) = (L_1 + L_2)^2 \mathbb{E}\left[\xi^2 (Y - \langle X, w \rangle)^2\right] \\
\mathbb{E}\left[Z^4\right]^{1/4} &\leq (L_1 + L_2) E[(\xi Y - \langle X, w \rangle)^2]^{1/2}
\end{aligned}$$

This completes the proof, i.e., we have shown that for some $L \leq c(L_1 + L_2)$, where $c > 0$ is an absolute constant, it holds that for all $w \in \mathbb{R}^d$, $\xi \in \mathbb{R}$,

$$\mathbb{E}\left[Z^4\right]^{1/4} \leq L \mathbb{E}\left[Z^2\right]^{1/2}, \tag{80}$$

where $L \leq c(L_1 + L_2)$ for some absolute constant $c > 0$. □

Part 2.

Proof. We begin by recalling the theorem on the median-of-means estimator for hypercontractive distributions we proved in class.

Theorem HW4.P5.2 (Lecture 20 – Median-of-means for hypercontractive distributions)**.** *Let p be an even integer. Assume that X is a zero-mean random vector in $\mathbb{R}^d$ such that for all $v \in S^{d-1}$, $\mathbb{E}\left[\langle X, v\rangle^{2p}\right]^{1/2p} \leq L\mathbb{E}\left[\langle X, v\rangle^{p}\right]^{1/p}$, where L is some "nice" function (i.e., X is $(p, 2p)$-hypercontractive). Then, with probability $1-\delta$, for all $v \in S^{d-1}$,*

$$\left|MOM(\langle X, v\rangle^p) - \mathbb{E}\left[\langle X, v\rangle^p\right]\right| \leq C2\sqrt{2}L^p\mathbb{E}\left[\langle X, v\rangle^p\right]\sqrt{\frac{d\log p + \log(1/\delta)}{n}},$$

where $C > 0$ is an absolute constant.

In order to apply this theorem to our problem, we consider define the vector $v = \xi w - w^*$ and note that $N(\xi, w) = \mathbb{E}\left[(\xi Y - \langle X, w\rangle)^2\right]$ can be written as $\mathbb{E}\left[\langle X, \xi w - w^*\rangle^2\right]$. We will apply the theorem with $p = 2$ to the random variable $\langle X, v\rangle$ for $v = \xi w - w^*$.
By the hypercontractivity assumptions given in Part 1, we have that for all $v \in S^{d-1}$,

$$\mathbb{E}\left[\langle X_i, v\rangle^4\right]^{1/4} \leq L\mathbb{E}\left[\langle X, v\rangle^2\right]^{1/2},$$

where $L \leq c(L_1 + L_2)$ for some absolute constant $c > 0$. This implies that X is $(2, 4)$-hypercontractive.
Applying the median-of-means theorem, we obtain that with probability at least $1-\delta$, for all $v \in S^{d-1}$,

$$\left|\mathrm{MOM}(\langle X, v\rangle^2) - \mathbb{E}\left[\langle X, v\rangle^2\right]\right| \leq C'L^2\mathbb{E}\left[\langle X, v\rangle^2\right]\sqrt{\frac{d + \log(1/\delta)}{n}},$$

where $C' > 0$ is an absolute constant.
We define the median-of-means estimator $\widehat{N}(\xi, w)$ as $\mathrm{MOM}(\langle X, \xi w - w^*\rangle^2)$, which is a quadratic form in ξ and w and can be written as

$$\widehat{N}(\xi, w) = \mathrm{MOM}((\xi Y - \langle X, w\rangle)^2) = \frac{1}{m}\sum_{i=1}^{m}(\xi Y_i - \langle X_i, w\rangle)^2 = \frac{1}{m}\sum_{i=1}^{m}(\langle X_i, \xi w - w^*\rangle)^2.$$

Then, the above inequality implies that with probability at least $1-\delta$,

$$\left|\frac{\widehat{N}(\xi, w)}{N(\xi, w)} - 1\right| \leq C'L^2\sqrt{\frac{d + \log(1/\delta)}{n}}.$$

We complete the proof with leveraging the hypercontractivity assumptions, $L \leq c(L_1 + L_2)$, and the absolute constants C' and c to arrive at

$$\left|\frac{\widehat{N}(\xi, w)}{N(\xi, w)} - 1\right| \leq c_1(L_1 + L_2)^2\sqrt{\frac{d + \log(1/\delta)}{n}},$$

where $c_1 > 0$ is an absolute constant. □

Part 3.

Proof. Let $\hat{w} = \mathrm{argmin}_{w\in\mathbb{R}^d}\,\widehat{N}(1, w)$. By the result from Part 2, with probability at least $1-\delta$, we have

$$\left|\frac{\widehat{N}(1, \hat{w})}{N(1, \hat{w})} - 1\right| \leq c_1(L_1 + L_2)^2\sqrt{\frac{d + \log(1/\delta)}{n}}.$$

We note that $R_n(w) = \widehat{N}(1, w)$. Since $\hat{w}$ minimizes $\widehat{N}(1, w)$, we have $\widehat{N}(1, \hat{w}) \leq \widehat{N}(1, w^*)$, and therefore

$$N(1, \hat{w}) \leq \frac{\widehat{N}(1, w^*)}{1 - c_1(L_1 + L_2)^2\sqrt{\frac{d+\log(1/\delta)}{n}}}.$$

Using the fact that $N(1, w^*) = R(w^*)$ and rearranging terms, we get

$$R(\hat{w}) - R(w^*) \leq c_2(L_1 + L_2)^4 \cdot R(w^*) \cdot \left(\frac{d + \log\left(\frac{1}{\delta}\right)}{n}\right),$$

provided that $n \geq c_3(L_1 + L_2)^4(d + \log(1/\delta))$, where $c_2, c_3 > 0$ are absolute constants.
This shows that the estimator $\hat{w}$ achieves a bound on the excess risk that is proportional to the hypercontractivity constants L_1 and L_2, the dimension d, and the logarithm of the inverse of the confidence level δ, scaled by the number of samples n. □

References

[AMN20] Julyan Arbel, Olivier Marchal, and Hien D. Nguyen, *On strict sub-gaussianity, optimal proxy variance and symmetry for bounded random variables*, ESAIM: PS **24** (2020), 39–55.

[BK15] Daniel Berend and Aryeh Kontorovich, *A finite sample analysis of the naive bayes classifier*, Journal of Machine Learning Research **16** (2015), no. 44, 1519–1545.

[Mas07] Pascal Massart, *Concentration inequalities and model selection*, Springer, 2007.

[PN95] Thomas K Philips and Randolph Nelson, *The moment bound is tighter than chernoff's bound for positive tail probabilities*, The American Statistician **49** (1995), no. 2, 175–178.

[Riv12] Omar Rivasplata, *Subgaussian random variables: An expository note*, 2012.

[Val84] L. G. Valiant, *A theory of the learnable*, Commun. ACM **27** (1984), no. 11, 1134–1142.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability & Its Applications **16** (1971), no. 2, 264–280.

[Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge University Press, 2018.

[Wai19] Martin J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.