IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: A. Rakhlin — Lectures 1-26
Scribe: A. Rakhlin — Spring 2022

1. INTRODUCTION

Suppose we would like to estimate the average height μ of students at MIT. Assuming students in this course are a random sample from the overall population, we may build a confidence interval for the unknown parameter μ as

$$\left[\bar{X}_n - 1.96\sigma/\sqrt{n}, \bar{X}_n + 1.96\sigma/\sqrt{n}\right]$$

where $\bar{X}_n$ is the sample average of n student heights in this course, and σ^2 is the population variance (which we can also estimate from data). Classical statistics tells us that this random interval contains μ with probability approximately 95%. Where does the number 1.96 come from?

More formally, let $X_1, \ldots, X_n$ be i.i.d. from some distribution P on $\mathbb{R}$ such that $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathsf{var}(X_i)$ are finite. Of course, for the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ we have $\mathbb{E}\left[\bar{X}_n\right] = \mu$, i.e. $\bar{X}_n$ is an unbiased estimate of the population mean. The (weak) Law of Large Numbers provides more information: $\bar{X}_n$ converges to μ in probability as $n \to \infty$: for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\bar{X}_n - \mu\right| < \epsilon\right) = 1.$$

The strong LLN provides states that $\bar{X}_n$ converges to μ almost surely:

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

Only finiteness of μ is needed for both of these results. Assuming σ is finite as well, we have the Central Limit Theorem:

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

This means that for large enough n, tail probability

$$\mathbb{P}\left(\sqrt{n}\left|\frac{\bar{X}_n - \mu}{\sigma}\right| > u\right) \approx \mathbb{P}(|Z| > u) = 2\Phi(-u), \quad Z \sim \mathcal{N}(0,1) \tag{1.1}$$

where Φ is the cumulative distribution function (cdf) of the standard normal. This approximation gives us the number 1.96 for which the probability is close to 95%.

Note that the statement of CLT is *asymptotic*, while we apply the conclusions for finite n. The quality of the approximation for finite n should be a source of worry, but statisticians devised rules of thumb, and indeed the approximation is quite good for n above, say, 30. In practice, statisticians can perform simulations to see whether the CLT can be trusted for the given sample size.

Of course, the quality of the approximation in (1.1) also depends on P. For instance, student heights are approximately normal, and for a normal random variable we have the exact non-asymptotic result: $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ (that is, $\sqrt{n}\left|\frac{\bar{X}_n-\mu}{\sigma}\right| \sim \mathcal{N}(0,1)$). However, for highly skewed distributions, the CLT kicks in for larger n.

This course is centered on non-asymptotic results. In the context of one-dimensional mean estimation, we will show in the next lecture that, under appropriate assumptions on P,

$$\mathbb{P}\left(\sqrt{n}\left|\frac{\bar{X}_n-\mu}{\sigma}\right|>u\right)\leq 2\exp\left\{-cu^2\right\}, \tag{1.2}$$

which holds for any n. Here c is some constant that depends on properties of P and may be somewhat larger than the one suggested by the limiting distribution. Thus, confidence intervals derived with such non-asymptotic methods may be somewhat wider, yet they hold for any n. On the downside, we will have to place stronger assumptions on the distribution than those required by the CLT.

One may argue that in modern applications, n is very often large. However, it is also true that many applications of interest are characterized by high dimensionality of data, or complex structure. In such problems, as we see below, asymptotic analysis based on $n \to \infty$ may not suffice.

1.1 Covariance estimation

Suppose $X_1, \ldots, X_n, X \overset{\text{i.i.d.}}{\sim} P$ on $\mathbb{R}^d$ and $\mathbb{E}[X]=0$. Let $\Sigma = \mathbb{E}[XX^\top]$ be the covariance matrix, and $\widehat{\Sigma}=\frac{1}{n}\sum_{i=1}^n X_iX_i^\top$ sample covariance. Clearly, sample covariance is an unbiased estimate of Σ. What can we say about the quality of such an estimate? For any pair $i,j=1,\ldots,d$, it still holds that $\widehat{\Sigma}_{i,j}$ converges to $\Sigma_{i,j}$ in probability by the LLN (it's an average of independent products). Similarly,

$$\sqrt{n}(\widehat{\Sigma}_{i,j}-\Sigma_{i,j}) \xrightarrow{d} \mathcal{N}(0, \mathsf{var}(X_{1,i}X_{1,j})),$$

under the appropriate moment assumptions. In particular, $\widehat{\Sigma}$ is a consistent estimator of Σ. This positive result, however, disregards the role of dimension d. In many problems of interest, dimensionality of the data may be of the same or similar order as the number of datapoints (e.g. genomics applications). *Is it then reasonable to disregard the role of d while sending n to infinity?*

There are two approaches to address the issue of high dimensionality:

- Consider an asymptotic setting where both n and d increase.
- Develop a non-asymptotic result that exhibits explicit dependence on d and n.

We start with the first approach. To make sense of this setting, consider a sequence of problems where both n and d are increasing at a constant aspect ratio $\alpha = d/n \in (0,1]$. Such a scaling is called *high-dimensional asymptotics*.

In high-dimensional analysis we are often interested in convergence in the operator norm

$$\|A\| = \sup_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}.$$

Let's see if such convergence holds in the high-dimensional asymptotics regime. For simplicity, take $\Sigma = I$, assume that coordinates of X are i.i.d., and let $\lambda_1(\widehat{\Sigma}) \geq \ldots \geq \lambda_d(\widehat{\Sigma}) \geq 0$ be the eigenvalues of $\widehat{\Sigma}$. If $\widehat{\Sigma}$ converges in spectral norm to $\Sigma = I$, the histogram of (random) eigenvalues should be concentrated at 1. In particular, we would expect the empirical distribution of eigenvalues $\frac{1}{d}\sum_{i=1}^d \delta_{\lambda_i} \overset{n\to\infty}{\longrightarrow} \delta_1$. This indeed happens when d is kept fixed and

n taken to infinity. Yet in the proportional high-dimensional regime, the limiting distribution of the empirical spectrum is not δ_1 but follows the Marčhenko-Pastur law [23]. This distribution has density supported on $[\lambda_-, \lambda_+]$ where $\lambda_+ = (1+\sqrt{\alpha})^2$, $\lambda_- = (1-\sqrt{\alpha})^2$. The density has the form

$$p(t) \propto \frac{\sqrt{(\lambda_+ - t)(t-\lambda_-)}}{t}$$

for $\alpha \in (0,1]$ (and for $\alpha > 1$, there is an atom at 0). We see that when d, n both grow proportionally, $\widehat{\Sigma}$ does not converge to Σ in the desired sense. To conclude, if we had, say, genomic data with $d = 20K$ and $n = 30K$, we should probably not trust the sample covariance matrix as an estimate of the true covariance matrix, even though the data size appears to be large.

Analogously to the development in the previous section, we can contrast the asymptotic approach with non-asymptotic tail bounds that hold for all n, d. In particular, we will show that, under additional assumptions, the largest eigenvalue of the sample covariance matrix satisfies

$$\mathbb{P}\left(\lambda_1(\widehat{\Sigma}) \geq \left(1 + \sqrt{d/n} + u\right)^2\right) \leq \exp\left\{-nu^2/2\right\}, \quad u \geq 0 \tag{1.3}$$

1.2 Hypothesis testing in high dimension

Suppose X has distribution either $P_1 = \mathcal{N}(\mu_1, \Sigma)$ or $P_2 = \mathcal{N}(\mu_2, \Sigma)$. In this case, the Neyman-Pearson lemma says that the most powerful hypothesis test (P_1 vs P_2) is to compare the likelihood-ratio to a fixed threshold τ:

$$\log \frac{dP_1(x)}{dP_2(x)} \geq \tau.$$

Using the form of the Gaussian multivariate density yields a simple statistic for testing:[1]

$$\Psi(x) = \langle \mu_1 - \mu_2, \Sigma^{-1}(x - (\mu_1+\mu_2)/2)\rangle.$$

If Type I and II errors are weighted equally,

$$\frac{1}{2}\mathbb{P}_1(\Psi(x) \leq 0) + \frac{1}{2}\mathbb{P}_2(\Psi(x) > 0) = \Phi(-\Delta/2) \tag{1.4}$$

where $\Delta = \|\mu_1 - \mu_2\|_2$ and $\Phi(\gamma) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\gamma} e^{-t^2/2}dt$.

If μ_1 and μ_2 are unknown, we may estimate them by $\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $\bar{X}_2 = \frac{1}{n_1}\sum_{i=1}^{n_2} X_i'$ on two independent samples from the two respective distributions. Let us assume for simplicity that $\Sigma = I$. The plug-in rule is then based on the statistic

$$\widehat{\Psi}(x) = \langle \bar{X}_1 - \bar{X}_2, x - (\bar{X}_1 + \bar{X}_2)/2\rangle.$$

Kolmogorov in 1970's studied the high-dimensional asymptotics of this problem where $n_1, n_2, d \to \infty$ while $d/n_1 \to \alpha$, $d/n_2 \to \alpha$. He showed that the error of $\widehat{\Psi}$ instead converges in probability to

$$\Phi\left(-\frac{\Delta^2}{2\sqrt{\Delta^2 + 2\alpha}}\right).$$

Note that when $\alpha = 0$, the result recovers (1.4). When the asymptotics are proportional, however, the effect of dimensionality cannot be ignored: the error probability becomes skewed and failure to account for dimensionality can lead to incorrect hypothesis test acceptance/rejection.

[1] The exposition here follows [30] and [34, Chap 1].

1.3 MLE

To give a taste of some other settings where asyptotic analysis is classically used, consider the case of $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$ on $\mathbb{R}$ where $\theta \in \Theta$ is a parameter. Suppose for simplicity, the random variables are real-valued. Under some regularity conditions, the sequence of MLE solutions

$$\widehat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P_\theta(X_i)$$

satisfies $\widehat{\theta}_n \to \theta$ in probability and asymptotic normality holds:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$$

where $I(\theta)$ is the Fisher information. Once again, while the asymptotic result sheds light on the convergence of MLE for large enough n, it does not say much about finite n. In particular, for some finite n, MLE may not be the best estimator, and some biased procedure may be better.

1.4 Statistical Learning

In the problem of binary classification, we are given i.i.d. samples $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from a joint distribution $P_{X \times Y}$ on $\mathcal{X} \times \{\pm 1\}$. Based on these n data, we construct a classifier $\widehat{f}_n : \mathcal{X} \to \{\pm 1\}$. We can make the dependence on the dataset explicit by writing the prediction on $x \in \mathcal{X}$ as $\widehat{f}_n(x; \mathcal{D}_n)$.

The classification rule is said to be (weakly) consistent [8] if

$$\mathbb{P}(f_n(X; \mathcal{D}_n) \neq Y) \to L^*,$$

where L^* is the Bayes risk (lowest achievable error by any classification rule), and the probability is with respect to $\mathcal{D}_n$ and a new datum (X, Y) from the same distribution. Strong universal consistency asks for almost sure convergence.

Once again, consistency does not guarantee good performance for any finite n. Much of learning theory instead focuses on explicit rates of convergence in n, as well as on making explicit the relevant complexity parameters of the problem. Such complexity parameters are not always explicit (in contrast to dimensionality of linear models), as illustrated in the next example.

1.5 Example of Complex Structure: Neural Networks

Consider a class of feed-forward neural networks

$$f(x) = W^L(\sigma(W^{L-1} \ldots \sigma(W^1 x) \ldots))$$

of depth L and nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ applied coordinate-wise, with $W^\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$. Here fixing the neural network structure and letting the data n increase to infinity may not be too interesting. Indeed, in modern practice, the size of the network is taken to be large for large n (much like in the high-dimensional asymptotics regime). We would like to understand what plays the role of “dimension” here. With the techniques developed in the second part of the course, we will be able to develop results that hold for any particular n, particular architecture, and, say, norms of the weight matrices.

2. SUB-GAUSSIAN RANDOM VARIABLES

This lecture is based on [33, Chap 2].

Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with distribution P with mean μ and variance σ^2. Recall that CLT implies approximate results of the form (1.1), i.e. for large enough n, *tails* (that is, values of $\mathbb{P}\left(\bar{X} \geq t\right)$) of sample averages of random variables behave like those of a Gaussian. So, what are these tails?

2.1 What is it like to be normal?

It is easy to show (by simple integration) that for $Z \sim \mathcal{N}(0,1)$ and for any $t > 0$,

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}(Z \geq t) \leq \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \tag{2.5}$$

The right-hand side is especially easy to remember: for $t \geq 1$, the tail is at most the density of the standard normal itself! Further, note that the moments of a standard normal random variable have the following behavior:

$$(\mathbb{E}|Z|^p)^{1/p} = \sqrt{2} \left(\frac{\Gamma(\frac{1+p}{2})}{\Gamma(\frac{1}{2})} \right)^{1/p} \sim c\sqrt{p}, \quad p \geq 1 \tag{2.6}$$

Finally,

$$\mathbb{E}e^{\lambda Z} = e^{\lambda^2/2} \tag{2.7}$$

for any $\lambda \in \mathbb{R}$. Hence $\mathbb{E}e^{\lambda Z'} = e^{\sigma^2 \lambda^2/2}$ for $Z' \sim \mathcal{N}(0, \sigma^2)$. Since our aim is to develop CLT-like non-asymptotic tail bounds on averages of random variables, we will be checking whether approximate versions of (2.5), (2.6), (2.7) hold.

2.2 Tail bounds

Let's now discuss some of the basic tools we have at our disposal for proving non-asymptotic tail bounds. We start with some familiar probabilistic inequalities. Markov's inequality says that for any non-negative X,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}, \quad t > 0$$

As a consequence, Chebyshev's inequality[2] says that for any real-valued random variable X with mean μ,

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

Applying Markov's inequality to higher moments yields

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^p \geq t^p) \leq \min_{p \geq 1} \frac{\mathbb{E}|X - \mu|^p}{t^p},$$

and applying Markov to an exponentiated random variable gives the Cramér-Chernoff bound

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}\left(e^{\lambda(X-\mu)} \geq e^{\lambda t}\right) \leq \inf_{\lambda > 0} e^{-t\lambda} \mathbb{E}e^{\lambda(X-\mu)}. \tag{2.8}$$

[2] Chebyshev was Markov's advisor

Hence, to deduce Gaussian-like tails for the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we need to understand the behavior of its moments $\mathbb{E}|\bar{X} - \mu|^p$ or its moment generation function

$$\mathsf{M}_U(\lambda) = \mathbb{E}\exp\{\lambda U\}, \quad \lambda \in \mathbb{R}$$

(defined abstractly here for any random variable U). Since the exponential of a sum is product of exponentials, the upper bound furnished by optimizing λ in (2.8) will be easier to handle.

Before proceeding to analyze the sums and establishing tail bounds, we first discuss a family of random variables that will be useful to work with. These random variables have more restrictions than those for which CLT holds (finite second moment), and hence form a smaller family. Nevertheless, the family is rich enough to cover many applications of interest. In the next lecture we will see a larger family of random variables.

2.3 Sub-Gaussian random variables

Definition 1. *A mean-zero random variable X is sub-Gaussian with variance factor (or variance proxy) s^2 if*

$$\mathbb{E}e^{\lambda X} \le e^{s^2\lambda^2/2}$$

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

We will write $X \in \mathsf{subG}(s^2)$ to denote the fact that X belongs to the family of sub-Gaussian random variables with s^2 as the parameter.

A few remarks. First, if X is sub-Gaussian, then so is $-X$ with the same variance proxy. This will be useful for deducing bounds on $|X|$ from those of bounds on X. Second, the families of these random variables are nested in the sense that if $X \in \mathsf{subG}(s^2)$, then $X \in \mathsf{subG}(t^2)$ for all $t^2 > s^2$. Third, if $X \in \mathsf{subG}(s^2)$ then $cX \in \mathsf{subG}(c^2 s^2)$. In particular, we can often work with $\mathsf{subG}(1)$ and conclude the more general result by rescaling.

It turns out that there are several equivalent ways of defining sub-Gaussian behavior.

Lemma 1 (Prop 2.5.2 in [33])**.** *Let X be a random variable with $\mathbb{E}[X] = 0$. Then the following are equivalent, and the parameters $c_i > 0$ differ by at most absolute constant factors:*

1. *For all $\lambda \in \mathbb{R}$,*
$$\mathbb{E}\exp\{\lambda X\} \le \exp\{c_1^2\lambda^2\}$$

2. *For all $t \ge 0$,*
$$\mathbb{P}(|X| \ge t) \le 2\exp\{-t^2/c_2^2\}$$

3. *For all $p = 1, 2 \ldots$,*
$$(\mathbb{E}|X|^p)^{1/p} \le c_3\sqrt{p}$$

4. *For all λ such that $|\lambda| \le 1/c_4$,*
$$\mathbb{E}\exp\{\lambda^2 X^2\} \le \exp\{c_4^2\lambda^2\}$$

5. For some $c_5 < \infty$,

$$\mathbb{E}\exp\{X^2/c_5^2\} \le 2.$$

We will only prove a few of the implications here (please see [33] for all the proofs). Let us illustrate (1) ⇒ (2). Suppose without loss of generality that $X \in \mathsf{subG}(1)$ (and hence $c_1^2 = 1/2$). In view of (2.8),

$$\mathbb{P}(X \ge t) \le \inf_{\lambda>0} e^{-t\lambda}\mathbb{E}e^{\lambda X} \le \inf_{\lambda>0} e^{\lambda^2/2-t\lambda} = \exp\left\{-\frac{t^2}{2}\right\} \tag{2.9}$$

by plugging in the optimizing value $\lambda = t$. This is the Cramér-Chernoff method.

Let us now prove (2) ⇒ (3). By rescaling, assume $c_2 = 1$. We have

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}(|X|^p \ge u)\,du = \int_0^\infty \mathbb{P}(|X| \ge t)\,pt^{p-1}dt \le \int_0^\infty 2\exp\{-t^2\}pt^{p-1}dt. \tag{2.10}$$

By change of variables $t = \sqrt{s}$ (and hence $dt = \frac{1}{2}s^{-1/2}ds$), the last expression can be written as $p\Gamma(p/2)$ in terms of the Gamma-function. Using Stirling's approximation, $\Gamma(p/2) \le (p/2)^{p/2}$. Hence,

$$(\mathbb{E}|X|^p)^{1/p} \le p^{1/p}(p/2)^{1/2} \le c_3\sqrt{p}.$$

2.3.1 Examples

Arguably, the simplest nontrivial random variables are Bernoulli or Rademacher. The Rademacher random variable ε takes values in $\{\pm 1\}$ with equal probability. We then have

$$\mathbb{E}e^{\lambda\varepsilon} = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} = \frac{1}{2}\sum_{k=0}^\infty \frac{\lambda^k}{k!} + \frac{(-\lambda)^k}{k!} = \sum_{k=0}^\infty \frac{\lambda^{2k}}{(2k)!} \le 1 + \sum_{k=1}^\infty \frac{(\lambda^2)^k}{2^k k!} = e^{\lambda^2/2}. \tag{2.11}$$

Hence, ε is 1-subGaussian. By re-scaling, the variable $\frac{b-a}{2}\varepsilon$ has subGaussian parameter $(b-a)^2/4$ and (obviously) takes values on the endpoints of $[-\frac{b-a}{2}, \frac{b-a}{2}]$ (assuming $b \ge a$). In fact, *any* zero-mean random variable that takes on values in the interval $[a, b]$ has subGaussian parameter at most $(b-a)^2/4$. In this sense, the scaled Rademacher random variable is extremal.

Lemma 2 (Hoeffding's Lemma)**.** *For any zero-mean random variable X taking values in $[a, b]$, the moment generating function satisfies*

$$\mathbb{E}\exp\{\lambda X\} \le \exp\{\lambda^2(b-a)^2/8\}, \qquad \lambda \in \mathbb{R}.$$

Hence, $X \in \mathsf{subG}((b-a)^2/4)$.

Proof. Let $\psi(\lambda) = \log\mathbb{E}\exp\{\lambda X\}$. Then $\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}e^{\lambda X}}$. Observe that $\psi(0) = \psi'(0) = 0$. It remains to prove that $\psi''(\lambda) \le (b-a)^2/4$ since Taylor's theorem would then imply (for some $\nu \in [0, \lambda]$)

$$\psi(\lambda) = \psi(0) + \lambda\psi'(0) + \frac{\lambda^2}{2}\psi''(\nu) \le \lambda^2\frac{(b-a)^2}{8}$$

We compute the second derivative as

$$\psi''(\lambda) = \mathbb{E}\left[X^2 \frac{e^{\lambda X}}{\mathbb{E}e^{\lambda X}}\right] - \left(\mathbb{E}\left[X \frac{e^{\lambda X}}{\mathbb{E}e^{\lambda X}}\right]\right)^2 = \mathsf{var}(Y)$$

for Y with density tilted by $x \to \frac{e^{\lambda x}}{\mathbb{E}e^{\lambda X}}$. Since Y takes on values in $[a, b]$, its variance is at most $(b-a)^2/4$, concluding the proof. □

Observe now that if $X_1 \in \mathsf{subG}(\sigma_1^2)$ and $X_2 \in \mathsf{subG}(\sigma_2^2)$, then $X_1 + X_2 \in \mathsf{subG}(\sigma_1^2 + \sigma_2^2)$ whenever X_1 and X_2 are independent. As an immediate consequence,

Lemma 3. *Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ be independent Rademacher and $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}$. Then*

$$\langle \boldsymbol{\varepsilon}, \boldsymbol{a} \rangle = \sum_{i=1}^{n} \varepsilon_i a_i \in \textsf{\textit{subG}}(\|\boldsymbol{a}\|_2^2).$$

In the same vein, for any sequence of independent random variables X_i with $\mathbb{E}[X_i] = \mu_i$ and $X_i - \mu_i \in \mathsf{subG}(\sigma_i^2)$,

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right\} \tag{2.12}$$

In particular, we have

Lemma 4. *Hoeffding's inequality For independent $X_i \in [a, b]$,*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\left\{-\frac{2t^2}{n(b-a)^2}\right\} \tag{2.13}$$

We close this section with two examples that indicate that the development of sub-Gaussian tail bounds so far is lacking on several fronts.

First, we will be interested in tail bounds on norms of gaussian vectors $\|\boldsymbol{g}\|$, where coordinates are standard normal. Since $\|\boldsymbol{g}\|^2 = \sum g_i^2$, it's tempting to use the sub-Gaussian results above. However, g_i^2 is not sub-Gaussian: $\mathbb{P}\left(g^2 \geq t\right) = \mathbb{P}\left(|g| \geq \sqrt{t}\right) \leq 2\exp\{-t/2\}$, which is sub-exponential rather than sub-Gaussian. These tails are heavier (or, fatter) than those of sub-Gaussian.

The second example illustrates a larger concern with sub-Gaussian tail bounds a la Hoeffding that rely on the range of random variables but not on their variance. Consider the following variable X. Let $\mathbb{P}(X = 0) = 1 - 1/k^2$ and $\mathbb{P}(X = \pm k) = 1/2k^2$, where k is a parameter, which we think of as large. Observe that the range of this random variable is $2k$, but the mean and (importantly) variance are small: $\mathbb{E}X = 0$, $\mathsf{var}(X) = 1 - 1/k^2$. If we draw $X_1, \ldots, X_k$ i.i.d., $\mathbb{P}(X_1 = \ldots = X_k = 0) = (1 - 1/k^2)^k \approx \exp\{-1/k\}$ which is close to 1 for large k. Since Hoeffding style inequalities only depend on the range, they are not able to distinguish this small-variance distribution from one that is uniform on $[-k, k]$.

3. SUB-EXPONENTIAL RANDOM VARIABLES

In this section, we follow the notation of [34].

As mentioned at the end of last lecture, the sub-Gaussian family leaves out some interesting random variables, in particular $X = Z^2$, where $Z \sim \mathcal{N}(0,1)$. Here X is called chi-square random variable, denoted by χ^2. Let's examine its moment generating function:

$$\mathsf{M}_X(\lambda) = \mathbb{E}\exp\{\lambda(Z^2-1)\} = \frac{1}{\sqrt{2\pi}}\int e^{\lambda(z^2-1)}e^{-z^2/2}dz \tag{3.14}$$

Clearly, MGF is infinite when $\lambda \geq 1/2$, so we only consider $\lambda < 1/2$. In that case,

$$\mathsf{M}_X(\lambda) = e^{-\lambda}\frac{1}{\sqrt{2\pi}}\int e^{-z^2(1-2\lambda)/2}dz = e^{-\lambda}\frac{1}{\sqrt{1-2\lambda}}. \tag{3.15}$$

One can further check that

$$e^{-\lambda}\frac{1}{\sqrt{1-2\lambda}} \leq \exp\left\{\frac{\lambda^2}{1-2\lambda}\right\}, \quad 0 < \lambda < 1/2 \tag{3.16}$$

Moreover, for $|\lambda| < 1/4$, the expression in (3.15) is dominated by $e^{2\lambda^2}$, and thus in this range Z^2 is sub-Gaussian.

Definition 2 (p 26 in [34]). *A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-exponential if there are non-negative parameters (s^2,α) such that*

$$\mathbb{E}[\exp\{\lambda(X-\mu)\}] \leq \exp\left\{s^2\lambda^2/2\right\}, \quad \forall|\lambda| < 1/\alpha. \tag{3.17}$$

We will write $X \in \mathsf{subE}(s^2,\alpha)$.

Remarks:

- In some of the references, you will see that sub-exponential random variables are defined with only one parameter; this corresponds to insisting that $\alpha = s$, i.e. the random variable has sub-Gaussian behavior with parameter s^2 in the range $|\lambda| < 1/s$. We follow [34] and decouple these two parameters.
- If we ask that (3.17) holds for $\lambda \in (0, 1/\alpha)$, the results stated below will only hold for the upper tail of $(X-\mu)$. The behavior for the upper and lower tails can indeed be different.
- Any $X \in \mathsf{subG}(s^2)$ is also sub-exponential with parameters $(s^2, 0)$.

From the earlier calculation, $Z^2 \in \mathsf{subE}(2^2, 4)$.

Lemma 5. *Suppose $X \in \mathsf{subE}(v^2,\alpha)$. Then*

$$\mathbb{P}(X-\mu \geq t) \leq \begin{cases} \exp\left\{-t^2/2v^2\right\}, & 0 \leq t \leq v^2/\alpha \\ \exp\left\{-t/2\alpha\right\}, & t \geq v^2/\alpha \end{cases} \tag{3.18}$$

The same holds for the tail of $-(X-\mu)$.

Alternatively, we can write

$$\mathbb{P}\left(X-\mu \geq t\right) \leq \exp\left\{-\min\left\{\frac{t^2}{2v^2}, \frac{t}{2\alpha}\right\}\right\} \tag{3.19}$$

Proof. Recall that

$$\mathbb{P}\left(X-\mu \geq t\right) \leq \inf_{\lambda\in[0,1/\alpha)} e^{-t\lambda}\mathbb{E}e^{\lambda(X-\mu)} \leq \inf_{\lambda\in[0,1/\alpha)} e^{-t\lambda}e^{v^2\lambda^2/2}. \tag{3.20}$$

where the limited range of λ, as compared to (2.8), is dictated by the definition of sub-exponential random variable. By taking derivative, we see that $\lambda = t/v^2$ is the unconstrained solution; we take this whenever $t/v^2 \leq 1/\alpha$. Otherwise, the minimum is achieved at the endpoint $\lambda = 1/\alpha$, with the value of $-t/\alpha + v^2/2\alpha^2 \leq -t/2\alpha$. □

Let us discuss Lemma 5. It shows that sub-exponential random variables exhibit two behaviors: sub-Gaussian (in the range $0 \leq t \leq v^2/\alpha$) and sub-exponential (in the range $t \geq v^2/\alpha$). We remark that the two-tail behavior arises simply by asking for the sub-Gaussian behavior in an interval.

Rather than writing the tail bound with a min as in (3.19), we can relax the exponent as follows. Note that for nonnegative u, v, it holds that $\min\{1/u, 1/v\} \geq 1/(u+v)$. We can thus upper bound the right-hand side of (3.19) as

$$\exp\left\{-\min\left\{\frac{t^2/2}{v^2}, \frac{t^2/2}{t\alpha}\right\}\right\} \leq \exp\left\{-\frac{t^2/2}{v^2+t\alpha}\right\} \tag{3.21}$$

We will see this form of a tail bound later in the lecture.

As with sub-Gaussian random variables, we can easily calculate the parameters for the sum of sub-exponentials. If $X_1, \ldots, X_n$ are independent with means $\mathbb{E}[X_i] = \mu_i$ and $X_i - \mu_i \in \mathsf{subE}(v_i^2, \alpha)$, then

$$\sum_{i=1}^n (X_i - \mu_i) \in \mathsf{subE}\left(\sum v_i^2, \max \alpha_i\right)$$

Lemma 6 (Bernstein's inequality)**.** *Suppose $X_1, \ldots, X_n$ are independent zero-mean and $X_i \in \mathsf{subE}(1,1)$. Let $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^2$. Then*

$$\sum_{i=1}^n a_i X_i \in \mathsf{subE}(\|\boldsymbol{a}\|_2^2, \|\boldsymbol{a}\|_\infty)$$

and, hence,

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

In particular, if all $a_i = 1/n$, under the conditions of above lemma,

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_i| \geq t\right) \leq 2\exp\left\{-n \cdot \min\left\{\frac{t^2}{2},\ \frac{t}{2}\right\}\right\} \tag{3.22}$$

To shed some light on (3.22), consider a tail bound for a *single* sub-exponential random variable with parameters $(1, 1)$:

$$\mathbb{P}\left(|\frac{1}{n}X_i| \geq t\right) \leq 2\exp\left\{-\frac{nt}{2}\right\}, \quad t \geq 1 \tag{3.23}$$

from Lemma 5. Hence, the sub-exponential behavior of the averages in (3.22) comes not from averaging but rather from a single worst tail (e.g. that has the largest α for a general collection).

Another way to write (3.22) is

$$\mathbb{P}\left(|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i| \geq t\right) \leq \begin{cases} 2\exp\left\{-\frac{t^2}{2}\right\}, & t \leq \sqrt{n} \\ 2\exp\left\{-\frac{t\sqrt{n}}{2}\right\}, & t \geq \sqrt{n} \end{cases} \tag{3.24}$$

The CLT would say that for large enough n, the random variable $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$ should have Gaussian tails under finiteness of second moment. In contrast, (3.24) says that for the sub-exponential family (where the restriction is less strict than sub-Gaussian but more strict than finite second moment), the sub-Gaussian behavior holds until $t = \sqrt{n}$, after which is switches to heavier tails.

3.1 Bernstein's Condition

It turns out that the two tail behaviors in (3.22) play an important role in statistical applications. As we will see below, the interplay between these tails is due to the relative behavior of the variance and range of random variables. So-called "fast rates" will be derived in this course in situations with small variance or low noise. But first, we define a condition that implies that the random variable is sub-exponential.

Definition 3. *We say that a random variable X with mean $\mu = \mathbb{E}X$ satisfies the Bernstein's Condition (BC) with parameter b if*

$$|\mathbb{E}(X-\mu)^k| \leq \frac{1}{2}k!\sigma^2 b^{k-2}, \quad k = 3, 4, \ldots$$

Lemma 7. *Any bounded random variable with $|X - \mathbb{E}X| \leq B$ satisfies the Bernstein's Condition with $b = B/3$.*

Proof. For any $k = 3, \ldots,$

$$\mathbb{E}|X-\mu|^k \leq \mathbb{E}\left\{|X-\mu|^{k-2}(X-\mu)^2\right\} \leq B^{k-2}\sigma^2 \leq \frac{\sigma^2}{2}k!(B/3)^{k-2} \tag{3.25}$$

□

Lemma 8 (Bernstein's Inequality)**.** *For a random variable X satisfying the Bernstein's Condition with parameter $b > 0$, it holds that for any $|\lambda| < 1/b$,*

$$\mathbb{E}\exp\{\lambda(X-\mu)\} \leq \exp\left\{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}\right\} \tag{3.26}$$

where $\mu = \mathbb{E}X$ and $\sigma^2 = \mathsf{var}(X)$. Hence, for all $t > 0$,

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq 2\exp\left\{-\frac{t^2/2}{\sigma^2 + tb}\right\}. \tag{3.27}$$

In particular, for a bounded random variable with $|X-\mu| \leq B$ a.s.,

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq 2\exp\left\{-\frac{t^2/2}{\sigma^2 + Bt/3}\right\}. \tag{3.28}$$

It is worth comparing (3.27) to the tail in (3.21) for a $\mathsf{subE}(v^2, \alpha)$ random variable. Here, v^2 is replaced by the actual variance σ^2, and the parameter α by b.

Proof. We have

$$\begin{aligned}
\mathbb{E}\exp\{\lambda(X-\mu)\} &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}(X-\mu)^k}{k!} && (3.29)\\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}|\lambda|^{k-2}b^{k-2} && (3.30)\\
&= 1 + \frac{\lambda^2\sigma^2}{2}\left(1 + \sum_{k=1}^{\infty}|\lambda|^k b^k\right) && (3.31)\\
&\leq 1 + \frac{\lambda^2\sigma^2}{2}\left(\frac{1}{1-|\lambda|b}\right) && (3.32)
\end{aligned}$$

provided that $|\lambda| \leq 1/b$, where $\sigma^2 = \mathsf{var}(X)$. Since $1 + x \leq e^x$, we conclude

$$\mathbb{E}\exp\{\lambda(X-\mu)\} \leq \exp\left\{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}\right\} \tag{3.33}$$

Choosing $\lambda = \frac{t}{bt+\sigma^2} \in [0, 1/b)$ in the Cramér-Chernoff bound (3.20) concludes the proof. □

In particular, (3.26) implies that random variables satisfying BC are sub-exponential. Indeed, by restricting $|\lambda| \leq 1/2b$ in (3.33) we conclude that

$$\mathbb{E}\exp\{\lambda(X-\mu)\} \leq \exp\left\{\lambda^2/2 \cdot (\sqrt{2}\sigma)^2\right\}, \tag{3.34}$$

which means that $X - \mu \in \mathsf{subE}(2\sigma^2, 2b)$. This, however, does not yield the constants of (3.28) as opposed to working directly with (3.33).

Finally, we mention a one-sided tail bound that has tighter constants:

Lemma 9. *Suppose for some positive v, b it holds that*

$$\mathbb{E}\exp\{\lambda(X-\mu)\} \le \exp\left\{\frac{\lambda^2 v^2/2}{1-b\lambda}\right\}, \quad \lambda \in (0, 1/b). \tag{3.35}$$

Then

$$\mathbb{P}\left(X-\mu \ge \sqrt{2v^2 t} + bt\right) \le \exp\{-t\}. \tag{3.36}$$

See [5, p. 29] for a proof, or try to prove it yourself (Hint: solve for the optimal λ in Cramér-Chernoff).

3.2 Bernstein's inequality for sums

We now discuss the implication of Bernstein's inequality for a sum of independent random variables X_i. Let $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathsf{var}(X_i)$. If X_i satisfies BC with parameter b then $X_i - \mu \in \mathsf{subE}(2\sigma^2, 2b)$ and thus $\sum_{i=1}^n X_i - \mu \in \mathsf{subE}(2n\sigma^2, 2b)$. Crucially, this is significantly better than saying that $\sum_{i=1}^n X_i - \mu$ is a random variable with range $B \cdot n$ and variance $n\sigma^2$. However, as mentioned at the end of last section, using $\sum_{i=1}^n X_i - \mu \in \mathsf{subE}(2n\sigma^2, 2b)$ in (3.19) will lose a constant factor, so we directly repeat the proof of Lemma 8 with the sum of random variables:

Lemma 10 (Bernstein's inequality)**.** *Let $X_1, \dots, X_n$ be independent with $\mathbb{E}X_i = \mu$, $\mathsf{var}(X_i) = \sigma^2$, and range $|X_i - \mu| \le B$ almost surely. Then*

$$\mathbb{P}\left(|\sum_{i=1}^n (X_i - \mu)| \ge t\right) \le 2\exp\left\{-\frac{t^2/2}{n\sigma^2 + Bt/3}\right\}. \tag{3.37}$$

You may encounter the normalized version

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^n X_i - \mu| \ge t\right) \le 2\exp\left\{-\frac{nt^2/2}{\sigma^2 + Bt/3}\right\}. \tag{3.38}$$

from which we can read off the following transition between the two tails. If $t \le 3\sigma^2/B$, the tails are sub-Gaussian, while for $t \ge 3\sigma^2/B$ they are sub-exponential.

As already indicated by (3.36), in view of (3.33), it also holds that with probability at least $1-\delta$,

$$|\frac{1}{n}\sum_{i=1}^n X_i - \mu| \le \sqrt{\frac{2\sigma^2 \log 2/\delta}{n}} + \frac{B\log 2/\delta}{3n}. \tag{3.39}$$

We will give a short proof of this with a worse constant 2 in the last term. To this end, set

$$\delta = 2\exp\left\{-\frac{nt^2}{2\sigma^2 + 2Bt/3}\right\}$$

which is equivalent to solving quadratic equation

$$t^2 - t\frac{2B\log 2/\delta}{3n} - \frac{2\sigma^2 \log 2/\delta}{n} = 0$$

and thus

$$t \leq \sqrt{\frac{2\sigma^2 \log 2/\delta}{n}} + \frac{2B \log 2/\delta}{3n}$$

using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. For the sharper constant 1 in (3.39), see (3.36).

Let us examine (3.39). We see that for small-variance case, the last term dominates and it indicates a faster convergence rate in terms of n (though at the expense of $\log 1/\delta$ rather than $\sqrt{\log 1/\delta}$ dependence on precision).

3.3 Equivalent conditions

Just as sub-Gaussian, sub-exponential random variables have several equivalent definitions.

Lemma 11. *Let X be a random variable with $\mathbb{E}[X] = 0$. Then the following are equivalent, and the parameters $c_i > 0$ differ by at most absolute constant factors:*

1. *For all $|\lambda| < 1/c_1$,*
$$\mathbb{E}\exp\{\lambda X\} \leq \exp\{c_1^2 \lambda^2\}$$

2. *For all $t \geq 0$,*
$$\mathbb{P}\left(|X| \geq t\right) \leq 2\exp\{-t/c_2\}$$

3. *For all $p = 1, 2 \ldots$,*
$$(\mathbb{E}|X|^p)^{1/p} \leq c_3 p$$

4. *For all $\lambda \in [0, 1/c_4]$,*
$$\mathbb{E}\exp\{\lambda |X|\} \leq \exp\{c_4 \lambda\}$$

5. *For some $c_5 < \infty$,*
$$\mathbb{E}\exp\{|X|/c_5\} \leq 2.$$

In particular, from the last point we immediately conclude that X is sub-Gaussian if an only if X^2 is sub-exponential.

3.4 Application: Classification

Suppose $f : \mathcal{X} \to \{\pm 1\}$ is a classifier that we developed (e.g. by training on some data). Now, suppose we have validation data $(X_1, Y_1), \ldots, (X_n, Y_n)$ sampled i.i.d. from an unknown $P_{X\times Y}$. The indicator loss compares the output $f(X)$ of the classifier on a point X to that of the label Y, and we denote it by $\mathbf{1}\{f(X_i) \neq Y_i\}$. The validation error is then $\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{f(X_i) \neq Y_i\}$, while the true expected error is $\mathbb{E}\mathbf{1}\{f(X) \neq Y\} = \mathbb{P}\left(f(X) \neq Y\right) \triangleq p$. Note that the variance of the random variable $\mathbf{1}\{f(X) \neq Y\}$ is simply $p(1-p)$, since this is a Bernoulli random variable.

Suppose we observe that validation error is 0. What can we conclude about the actual true expected error? The CLT would suggest we are $O(1/\sqrt{n})$ away.

Bernstein's inequality tells us that with probability at least $1 - e^{-u}$,

$$\mathbb{E}\mathbf{1}\{f(X) \neq Y\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{f(X_i) \neq Y_i\} \leq \sqrt{\frac{2p(1-p)u}{n}} + \frac{u}{3n} \tag{3.40}$$

Under the event that the validation error is zero, we have

$$p \leq \sqrt{\frac{2pu}{n}} + \frac{u}{3n}$$

which means

$$p \leq \frac{4u}{n}.$$

Note that this is better than what we expected from the CLT. The effect is due to low variance (more precisely, here variance is upper bounded by expectation itself). This type of argument appears often in statistical learning. Of course, we would be interested in the case that f itself was produced by minimizing error on the same data (in which case the validation error is in fact training error). The issue of the dependence of f on the data (and hence failure of CLT due to lack of independence) will be dealt with through notions of uniform convergence in the second part of the course.

3.5 Norm concentration

We now revisit the example we considered earlier. Let $Y = \|\boldsymbol{g}\|^2 = \sum_{i=1}^d g_i^2$ be the squared norm of a random Gaussian vector with i.i.d. $\mathcal{N}(0,1)$ coordinates (Y has χ_d^2 distribution with d degrees of freedom). Recall that $g_i^2 \in \mathsf{subE}(2^2, 4)$ and thus $Y \in \mathsf{subE}(4d, 4)$. Thus,

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{i=1}^d g_i^2 - 1\right| \geq t\right) = \mathbb{P}\left(|Y - d| \geq dt\right) \leq 2\exp\{-dt^2/8\}, \quad t \in (0,1) \tag{3.41}$$

where we only took one tail of the two-tail behavior (NB: the constant 8 can be improved).

3.6 The Johnson–Lindenstrauss lemma (JL) Lemma

Let $u_1, \ldots, u_N \in \mathbb{R}^M$ be fixed vectors in M dimensions. If M is large, we may ask whether we can reduce the dimensionality while preserving the norms of these vectors (or, pairwise distances). A classical way to do this is via random projections. Let m be the target dimensionality, $m < M$. Let $\Gamma \in \mathbb{R}^{m \times M}$ be a random matrix with independent entries $\Gamma_{i,j} \sim \mathcal{N}(0,1)$. We will reduce dimensionality by mapping each $u_i \to \frac{1}{\sqrt{m}}\Gamma u_i$. It remains to analyze how norms change under the action of this matrix.

First, fix a vector $v \in \mathbb{R}^M$ with $\|v\| = 1$. Then $\langle \Gamma_i, v\rangle \sim \mathcal{N}(0,1)$ where Γ_i is the ith row of the matrix Γ. Then $\sum_{i=1}^m \langle \Gamma_i, v\rangle^2 = \|\Gamma v\|^2 \sim \chi_m^2$. As shown in the previous section,

$$\mathbb{P}\left(\left|\frac{1}{m}\|\Gamma v\|^2 - 1\right| \geq t\right) \leq 2\exp\{-mt^2/8\}, \quad t \in (0,1)$$

Hence, if we define the map $F(u) = \frac{1}{\sqrt{m}}\Gamma u$, we have proved that for any $u \neq 0$, $u \in \mathbb{R}^M$,

$$\mathbb{P}\left(\frac{\|F(u)\|^2}{\|u\|^2} \notin [1-t, 1+t]\right) \leq 2\exp\{-mt^2/8\}, \quad t \in (0,1)$$

By a union bound, for $u_1, \ldots, u_N \in \mathbb{R}^M$,

$$\mathbb{P}\left(\exists u_i \neq u_j, \quad \frac{\|F(u_i) - F(u_j)\|^2}{\|u_i - u_j\|^2} \notin [1-t, 1+t]\right) \leq 2\binom{N}{2}\exp\{-mt^2/8\}, \quad t \in (0,1)$$

since F is linear. By setting the right-hand-side to δ, we have that with probability at least $1-\delta$, all the norms are preserved up to multiplicative accuracy $1 \pm t$ as long as

$$m > \frac{16}{t^2}\log(N/\delta).$$

Interestingly, the dimension M does not enter this estimate for the target dimension.

3.7 Norm concentration: from sub-Exponential to sub-Gaussian tails

As we have seen earlier, $Y - d = \|\boldsymbol{g}\|^2 - d = \sum_{i=1}^{d}(g_i^2 - 1)$ is a sub-exponential random variable. Hence, we expect that square root of $\sqrt{Y}$ (that is, the norm of the random Gaussian vector) to be, after centering, sub-Gaussian. In fact, we will show this for a general vector with sub-Gaussian entries.

Following the exposition in [33, Chap 3.1], let $X = (X_1, \dots, X_d)$ be a vector with i.i.d. sub-Gaussian entries X_i with mean zero and variance 1. Then, $\|X\|^2 - d = \sum_{i=1}^{d}(X_i^2 - 1)$ is sub-exponential and, hence, satisfies the Bernstein's inequality

$$\mathbb{P}\left(|\frac{1}{d}\|X\|^2 - 1| \geq t\right) \leq 2\exp\left\{-Cd\min\{t^2, t\}\right\}, \tag{3.42}$$

where C depends on the sub-Gaussian constant of X_i. A simple trick will now convert the two-tailed behavior for the square into single-tail behavior for the norm itself. First, following [33], observe that $|z-1| \geq t$ implies $|z^2 - 1| \geq \max\{t, t^2\}$ for all $z \geq 0$. Now, let $u = \max\{t, t^2\}$ and observe that $\min\{u, u^2\} = \min\{\max\{t, t^2\}, (\max\{t, t^2\})^2\} = t^2$ (by considering cases). Hence,

$$\begin{aligned}
\mathbb{P}\left(|\frac{1}{\sqrt{d}}\|X\| - 1| \geq t\right) &\leq \mathbb{P}\left(|\frac{1}{d}\|X\|^2 - 1| \geq \max\{t, t^2\}\right) && (3.43)\\
&= \mathbb{P}\left(|\frac{1}{d}\|X\|^2 - 1| \geq u\right) && (3.44)\\
&\leq 2\exp\left\{-Cd\min\{u^2, u\}\right\} && (3.45)\\
&= 2\exp\left\{-Cdt^2\right\} && (3.46)
\end{aligned}$$

for all $t > 0$. Thus, norm of a random vector with sub-Gaussian entries is sub-Gaussian itself (after centering). Or, rescaling,

$$\mathbb{P}\left(|\,\|X\| - \sqrt{d}| \geq t\right) \leq 2\exp\left\{-Ct^2\right\} \tag{3.47}$$

This means that a vector with independent sub-Gaussian entries with variance 1 has norm that is tightly concentrated around the value $\sqrt{d}$. This is one of the most basic high-dimensional phenomena.

3.8 From isotropic to anisotropic vectors

Recall that in (3.41), we treated the χ_1^2 random variable g^2 (where $g \sim \mathcal{N}(0,1)$) as a $\mathsf{subE}(2^2, 4)$ random variable. However, χ_1^2 has distinct upper and lower tails, and so it may be beneficial to consider one-sided tail bounds. We will only focus on the upper tail. For this purpose, recall (3.16), which holds for $\lambda \in (0, 1/2)$. Then, using the Cramér-Chernoff bound with the choice $\lambda = \frac{t}{2t+2} \in (0, 1/2)$, we arrive at the one-sided bound

$$\mathbb{P}\left(g^2 - 1 \geq t\right) \leq \exp\left\{-\frac{t^2/2}{2 + 2t}\right\} \tag{3.48}$$

or, in view of (3.16), from (3.36),

$$\mathbb{P}\left(g^2 - 1 \geq 2\sqrt{t} + 2t\right) \leq \exp\{-t\}. \tag{3.49}$$

The rest of this subsection follows easily by checking what happens to (3.16) for the sum of random variables. For $\sum_{i=1}^d g_i^2 \sim \chi_d^2$, where $g_i \sim \mathcal{N}(0, 1)$ independently,

$$\mathbb{P}\left(\sum_{i=1}^d g_i^2 - d \geq t\right) \leq \exp\left\{-\frac{t^2/2}{2d + 2t}\right\} \tag{3.50}$$

or, from (3.36),

$$\mathbb{P}\left(\sum_{i=1}^d g_i^2 - d \geq 2\sqrt{dt} + 2t\right) \leq \exp\{-t\}. \tag{3.51}$$

For $\boldsymbol{a} = (a_1, \ldots, a_d)$, $a_i \geq 0$, the above tail bound is easily extended to

$$\mathbb{P}\left(\sum_{i=1}^d a_i g_i^2 - \|\boldsymbol{a}\|_1 \geq 2\|\boldsymbol{a}\|_2 \sqrt{t} + 2\|\boldsymbol{a}\|_\infty t\right) \leq \exp\{-t\}. \tag{3.52}$$

Let us write $\boldsymbol{g} \sim \mathcal{N}(0, I_d)$. Let $A \in \mathbb{R}^{d \times d}$ and let $\Sigma = A^\top A$. We have that the mean

$$\mathbb{E}\|A\boldsymbol{g}\|^2 = \mathbb{E}\boldsymbol{g}^\top A^\top A\boldsymbol{g} = \mathsf{tr}(A^\top A \mathbb{E}\boldsymbol{g}\boldsymbol{g}^\top) = \mathsf{tr}(\Sigma).$$

Since Σ is positive semidefinite, it has an SVD decomposition $\Sigma = U\Lambda U^\top$. Then $\Sigma^{1/2}\boldsymbol{g} \sim \mathcal{N}(0, \Sigma)$ and

$$\|A\boldsymbol{g}\|^2 = \boldsymbol{g}^\top \Sigma \boldsymbol{g} = \left\|\Sigma^{1/2}\boldsymbol{g}\right\|^2 = \left\|\Lambda^{1/2}\boldsymbol{g}\right\|^2 = \sum_{i=1}^d \lambda_i \boldsymbol{g}_i^2$$

by the rotational invariance of multivariate normal and $\Lambda = \mathsf{diag}(\lambda_1, \ldots, \lambda_d)$. Then (3.52) implies

$$\mathbb{P}\left(\|A\boldsymbol{g}\|^2 - \mathsf{tr}(\Sigma) \geq 2\sqrt{\mathsf{tr}(\Sigma^2)t} + 2\|\Sigma\| t\right) \leq \exp\{-t\} \tag{3.53}$$

because Σ and Λ share the same set of eigenvalues. Note that d does not explicitly appear, except through the trace of the eigenvalues.

The tail bound (3.53) was proved in [14] for sub-Gaussian (rather than Gaussian) mean-zero vectors, with the same constants as above. More precisely, a centered random vector $\boldsymbol{x} \in \mathbb{R}^d$ is sub-Gaussian with variance proxy v^2 if for any unit vector $\boldsymbol{u} \in \mathbb{R}^d$, $\langle \boldsymbol{x}, \boldsymbol{u} \rangle \in \mathsf{subG}(v^2)$.

We also remark that tail bounds on $\|A\boldsymbol{x}\|^2 = \boldsymbol{x}^\top A^\top A\boldsymbol{x}$ have been proved in [29] for more general quadratic forms $\boldsymbol{x}^\top B\boldsymbol{x}$, where B is not necessarily psd, but with stronger independence assumptions on coordinates of $\boldsymbol{x}$. Such bounds are known as Hanson-Wright inequalities.

4. MEAN ESTIMATION

4.1 High-dimensional mean estimation

Now, let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Gamma)$ be independent multivariate Gaussian vectors in $\mathbb{R}^d$, and recall that $X_i - \mu = \Gamma^{1/2} Z_i$ where $Z_i \sim \mathcal{N}(0, I_d)$. To estimate the mean, it is natural to take the sample average $\bar{X}_n$. Its quality can be measured by $\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right\|^2$. Without loss of generality, for the purposes of analysis we can set $\mu = 0$. Then the quality of the estimate is simply $\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|^2 = n^{-1}\left\|\Gamma^{1/2}(\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i)\right\|^2 \overset{d}{=} n^{-1}\left\|\Gamma^{1/2} Z\right\|^2$ for an independent $Z \sim \mathcal{N}(0, I_d)$. From (3.53) with $A = (\Gamma/n)^{1/2}$ (and hence $\Sigma = n^{-1}\Gamma$),

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|^2 - \frac{\mathsf{tr}(\Gamma)}{n} \geq \frac{2\sqrt{\mathsf{tr}(\Gamma^2)t}}{n} + \frac{2\left\|\Gamma\right\| t}{n}\right) \leq \exp\{-t\} \tag{4.54}$$

We conclude that when the error is measured in squared Euclidean norm, the expected error is $\mathsf{tr}(\Gamma)/n$ and the deviations above this expectation are given by the two tails in terms of the trace of Γ. Since dimension d never appears in these bounds, it can be very large or infinite, as long as the covariance matrix Γ has a fast decay of eigenvalues.

As in (3.47), if we instead consider the norm rather than squared norm, we again only have the sub-Gaussian behavior. More precisely,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\| - \sqrt{\frac{\mathsf{tr}(\Gamma)}{n}} \geq \sqrt{\frac{2\left\|\Gamma\right\| t}{n}}\right) \leq \exp\{-t\}. \tag{4.55}$$

We refer to [33, p. 135] for the proof of this fact, along the lines of our earlier conversion to sub-Gaussian tails for the norm itself.

We see that the trace of the covariance matrix, $\mathsf{tr}(\Gamma)$, serves as the *effective dimension* of the problem. Indeed, it replaces the actual dimension d that would arise if we used (3.47) with identity covariance.

4.2 Median of means and heavy-tailed distributions

Consider mean estimation in 1 dimension, but let us not assume sub-Gaussianity. We will see that sample average itself is not a good estimate of the unknown mean, and a small modification will be needed.

Let $X_1, \dots, X_n$ be i.i.d. from a distribution P with finite mean μ and variance σ^2. From Chebyshev's inequality,

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq t\right) \leq \frac{\sigma^2/n}{t^2},$$

or, with probability at least $1 - 2\delta$,

$$|\bar{X}_n - \mu| \leq \sigma\sqrt{\frac{1}{2n\delta}}.$$

Unlike the sub-Gaussian tails we had studied so far, the above tails have a polynomial dependence on t, or $1/\delta$. One may ask whether it is due to a sub-optimal choice of Chebyshev's inequality. However, the result of Catoni tells us that Chebyshev here is essentially unimprovable:

Lemma 12 (Catoni [7])**.** *For any $\delta \in (0, (2e)^{-1})$ and $\sigma^2 > 0$, there exists distribution P with mean 0 and variance σ^2 s.t.*

$$\mathbb{P}\left(|\bar{X}_n| \geq \sigma\sqrt{\frac{1}{2n\delta}}\left(1 - \frac{2e\delta}{n}\right)^{\frac{n-1}{2}}\right) \geq 2\delta.$$

Proof. Let $X_1, \dots, X_n$ i.i.d. with $\mathbb{E}X_i = 0$, and assume without loss of generality that $\sigma = 1$ (since we can divide through by σ). Fix t, to be chosen later, and define

$$\mathbb{P}\left(X_i = nt\right) = \mathbb{P}\left(X_i = -nt\right) = \frac{1}{2n^2t^2}$$

and

$$\mathbb{P}\left(X_i = 0\right) = 1 - \frac{1}{n^2t^2}.$$

We verify that

$$\mathbb{E}X_i^2 = \mathsf{var}(X_i) = n^2t^2 \cdot \frac{1}{n^2t^2} = 1$$

Then for $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$,

$$\mathbb{P}\left(\bar{X} \geq t\right) = \mathbb{P}\left(\bar{X} \leq -t\right) \geq \mathbb{P}\left(\bar{X} = t\right) \geq \frac{n}{2n^2t^2}\left(1 - \frac{1}{n^2t^2}\right)^{n-1}, \tag{4.56}$$

since $\bar{X} = t$ can be achieved whenever any of the variables is nt and the rest are 0. Now choose $t = \sqrt{\frac{1}{2n\delta}}\left(1 - \frac{2e\delta}{n}\right)^{\frac{n-1}{2}}$. If we show that the right-hand side of (4.56) with this value of t is at least δ, we will be done. This amounts to proving

$$\left(1 - \frac{1}{\frac{n}{2\delta}(1 - \frac{2e\delta}{n})^{n-1}}\right)^{n-1} \geq \left(1 - \frac{2e\delta}{n}\right)^{n-1} \tag{4.57}$$

which is true if

$$1 - \frac{2e\delta}{n} \leq 1 - \frac{2\delta}{n}\frac{1}{\left(1 - \frac{2e\delta}{n}\right)^{n-1}}.$$

The last statement is true since

$$\left(1 - \frac{2e\delta}{n}\right)^{n-1} \geq 1/e.$$

□

Since the sample mean does not exhibit sub-Gaussian tail behavior in our heavy-tailed situation, the goal is to change the estimator itself. Perhaps, this is the first "non-trivial" estimator in this course, since we only analyzed averages so far.

For simplicity of exposition, suppose $n = km$ with $k, m \geq 1$ integers. Define the median-of-means estimator as

$$\widehat{\mu} = \mathsf{median}\left(\frac{1}{m}\sum_{i=1}^{m} X_i\ ,\ \dots\ ,\ \frac{1}{m}\sum_{i=(k-1)m+1}^{km} X_i\right) \tag{4.58}$$

Lemma 13. *Let $\delta \in (0,1)$, $k = c \log 1/\delta$, $m = \frac{n}{c \log 1/\delta}$, for some absolute constant c. Then, with probability at least $1-\delta$,*

$$|\widehat{\mu} - \mu| \le \sigma \sqrt{\frac{c \log 1/\delta}{n}}.$$

Proof. For each batch of size m, we have a bound by application of Chebyshev:

$$\mathbb{P}\left(|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu| \ge \frac{2\sigma}{\sqrt{m}} \right) \le \frac{1}{4} \tag{4.59}$$

The corresponding bad event for the jth batch can be denoted by

$$Y_j = \mathbf{1}\left\{ |\frac{1}{m}\sum_{i=(j-1)m+1}^{jm} X_i - \mu| \ge \frac{2\sigma}{\sqrt{m}} \right\}$$

Note that Y_j are i.i.d. Bernoulli with bias $p \le 1/4$. Then

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \ge \frac{2\sigma}{\sqrt{m}} \right) \le \mathbb{P}\left(\sum_{j=1}^{k} Y_j \ge k/2 \right) \le \mathbb{P}\left(\sum_{j=1}^{k} Y_j - \mathbb{E}Y_j \ge k/4 \right) \le \exp\left\{ -\frac{(k/4)^2}{k} \right\}$$

□

A disadvantage of the median-of-means estimator is that k should be chosen as a function of the target accuracy δ. One may ask whether there exist estimators with sub-Gaussian tails that work for all (or at least a nontrivial range of) δ. Surprisingly, if σ is not known, this is impossible [9].

Finally, the idea of median-of-means has been extended to multivariate distributions. Here, a certain median-of-means tournament by [22] is shown to obtain the sub-Gaussian behavior (4.55). [13] achieved the first poly-time algorithm with this behavior.

4.3 Sparse mean estimation and the Gaussian Sequence Model

In Section 4.1, we considered the problem of estimating the mean of a high-dimensional vector. We saw that the trace of the covariance matrix serves as an effective dimension of the problem and determines the number n of samples needed to achieve a certain accuracy. Low trace of the covariance matrix is a coordinate-free notion of "simplicity" of the distribution of the high-dimensional random variable. In contrast, here we will make an assumption about the mean of the distribution.

First, let us extend the definition of sub-Gaussianity to random vectors.

Definition 4. *A vector-valued random variable $X \in \mathbb{R}^d$ with mean μ is v^2-sub-Gaussian if for all $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\| = 1$,*

$$\langle X - \boldsymbol{\mu}, \boldsymbol{u} \rangle \in \textsf{subG}(v^2).$$

Equivalently, we can state the definition as: $\langle X - \mu, \boldsymbol{u}\rangle \in \mathsf{subG}(v^2 \|\boldsymbol{u}\|^2)$ for any $\boldsymbol{u} \in \mathbb{R}^d$. In other words, a vector is sub-Gaussian if all its 1-dimensional marginals are sub-Gaussian. In particular, by choosing standard basis vectors, sub-Gaussianity of X implies sub-Gaussianity of its coordinates, and thus the variance of each coordinate of X is at most v^2. Note that sub-Gaussianity of the vector does not require independence of the coordinates.

Let $X_1, \ldots, X_n$ be i.i.d. from a v^2-sub-Gaussian distribution with mean $\boldsymbol{\mu}$. Let us estimate $\boldsymbol{\mu}$ by $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then it is easy to see that $\bar{X}$ is v^2/n-sub-Gaussian. Equivalently, we can think of observing a single vector from the model

$$Y = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{4.60}$$

with $\boldsymbol{\varepsilon} \in \mathsf{subG}(v^2/n)$, $\mathbb{E}\boldsymbol{\varepsilon} = 0$, and we are observing one vector realization $Y = \bar{X}$. While $\boldsymbol{\varepsilon}$ is sub-Gaussian, we will not assume independence of the coordinates. The model in (4.60)—called the *Gaussian Sequence Model*—has been studied extensively (usually under the Gaussian assumption on $\boldsymbol{\varepsilon}$).

The model in (4.60) is also a prototypical example of a statistical problem in the form "observation = signal + noise," with the goal of denoising the observation and estimating the signal, under various structural assumptions. We will study one such example now.

Suppose $\boldsymbol{\mu} \in \mathbb{R}^d$ is k-sparse:

$$\|\boldsymbol{\mu}\|_0 = \sum_{j=1}^d \mathbf{1}\{\mu_j \neq 0\} = k.$$

Returning Y (or, in our earlier example, the sample mean $\bar{X}$) as an estimate of $\boldsymbol{\mu}$ may be suboptimal if we know that $\boldsymbol{\mu}$ is sparse. Indeed, the mean squared error of the estimator $\widehat{\boldsymbol{\mu}} = Y$ is

$$\mathbb{E}\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 = \mathbb{E}\|\boldsymbol{\varepsilon}\|^2 \propto \frac{d}{n}.$$

In particular, this ignores the sparsity of the mean vector $\boldsymbol{\mu}$.

A natural modification is to threshold coordinates of Y. Given δ, consider the event $\mathcal{E} = \{|\varepsilon_i| \leq \lambda\}$, where λ will be chosen later as a function of δ, n, k such that this event holds with probability at least $1 - \delta$. Define a "kill-or-keep" estimate $\widehat{\boldsymbol{\mu}}^{\mathsf{HT}} \in \mathbb{R}^d$ by

$$\widehat{\boldsymbol{\mu}}_i^{\mathsf{HT}} = Y_i \mathbf{1}\{|Y_i| > \lambda\}$$

The superscript here stands for "hard thresholding." How close is this estimate to $\boldsymbol{\mu}$?

In the case that $\boldsymbol{\mu}_i = 0$, under the event $\mathcal{E}$ it holds that $|Y_i| \leq \lambda$, and thus $\widehat{\boldsymbol{\mu}}_i^{\mathsf{HT}} = 0$ and $|\widehat{\boldsymbol{\mu}}_i^{\mathsf{HT}} - \boldsymbol{\mu}_i| = 0$ (that is, the coordinate was zeroed out correctly). If, on the other hand, $\boldsymbol{\mu}_i \neq 0$ (i.e. i is one of the k non-zero coordinates), then

$$|\widehat{\boldsymbol{\mu}}_i^{\mathsf{HT}} - \boldsymbol{\mu}_i| \leq |\widehat{\boldsymbol{\mu}}_i^{\mathsf{HT}} - Y_i| + |Y_i - \boldsymbol{\mu}_i| \leq 2\lambda.$$

Putting these together, we have

$$|\widehat{\boldsymbol{\mu}}_i^{\mathsf{HT}} - \boldsymbol{\mu}_i| \leq 2\lambda \mathbf{1}\{\boldsymbol{\mu}_i \neq 0\}$$

or

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{HT}} - \boldsymbol{\mu}\right\|^2 \leq 4\lambda^2 \|\boldsymbol{\mu}\|_0 . \tag{4.61}$$

It remains to calculate λ such that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. To this end, note that for any $i = 1, \dots, d$, sub-Gaussianity (with parameter v^2/n) implies

$$\mathbb{P}\left(|\boldsymbol{\varepsilon}_i| > \frac{v}{\sqrt{n}}\sqrt{2\log 2/\delta}\right) \leq \delta \tag{4.62}$$

By union bound,

$$\mathbb{P}\left(\forall i \in [d], \quad |\boldsymbol{\varepsilon}_i| \leq \frac{v}{\sqrt{n}}\sqrt{2\log(2d/\delta)}\right) \leq \delta \tag{4.63}$$

Hence, from (4.61), with probability at least $1-\delta$,

$$\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{HT}} - \boldsymbol{\mu}\right\|^2 \leq \frac{8v^2 k \log(2d/\delta)}{n}. \tag{4.64}$$

A few remarks:

- The thresholding method requires the knowledge of v and δ, but not the sparsity parameter k. In this sense, the method is adaptive to the unknown sparsity, i.e. attains the rate in (4.64) that depends on k despite not knowing it.
- If the goal is to recover the correct support (i.e. non-zero entries) of $\boldsymbol{\mu}$, we need to make an assumption about "signal strength," i.e. that the minimum value of a nonzero entry of $\boldsymbol{\mu}$ is at least, say, 3λ. This ensures that signal can be separated from the noisy values outside the support. In this case, we can threshold Y at the value of 2λ.

The hard thresholding estimator is a discontinuous function. Another popular thresholding scheme is soft thresholding, defined as

$$\widehat{\boldsymbol{\mu}}_i^{\mathsf{ST}} = \begin{cases} Y_i - \lambda, & Y_i > \lambda \\ 0, & |Y_i| \leq \lambda \\ Y_i + \lambda, & Y_i < -\lambda \end{cases} \tag{4.65}$$

Finally, we mention that both hard and soft thresholding schemes can be written in the form

$$\widehat{\boldsymbol{\mu}}^{\mathsf{HT}} = \underset{\widehat{\boldsymbol{\mu}}}{\operatorname{argmin}} \|Y - \widehat{\boldsymbol{\mu}}\|^2 + \lambda^2 \|\widehat{\boldsymbol{\mu}}\|_0 \tag{4.66}$$

and

$$\widehat{\boldsymbol{\mu}}^{\mathsf{ST}} = \underset{\widehat{\boldsymbol{\mu}}}{\operatorname{argmin}} \|Y - \widehat{\boldsymbol{\mu}}\|^2 + \lambda \|\widehat{\boldsymbol{\mu}}\|_1 . \tag{4.67}$$

To see the first one, note that the objective decomposes coordinate-wise, and for each coordinate we have

$$\widehat{\boldsymbol{\mu}}_i^{\mathsf{HT}} = \underset{\widehat{\boldsymbol{\mu}}_i \in \mathbb{R}}{\operatorname{argmin}} (Y_i - \widehat{\boldsymbol{\mu}}_i)^2 + \lambda^2 \mathbf{1}\left\{\widehat{\boldsymbol{\mu}}_i \neq 0\right\}. \tag{4.68}$$

If $\lambda^2 \geq Y_i^2$, the solution is attained at 0; otherwise at Y_i.

In contrast to hard thresholding, the reformulation in (4.67) is convex. While for the Gaussian Sequence Model it may not matter computationally, such a convex reformulation helps in regression setups considered later in the course.

The model in (4.60) is also called a *direct observation* model since we are observing the signal $\boldsymbol{\mu}$ directly perturbed by noise. After a short detour into maximal inequalities, we turn to linear regression, a problem where the parameter vector is observed via linear measurements (*indirect* observations).

5. MAXIMAL INEQUALITIES: BASIC RESULTS

Before diving into linear regression, we make a brief detour and talk about maximal inequalities. This topic is a precursor to the more detailed study of the suprema of sub-Gaussian and empirical processes.

First, recall several basic notions. Given a norm $\|\cdot\|$ (say, on $\mathbb{R}^d$, although this extends to Banach spaces), the dual norm is defined as

$$\|\boldsymbol{v}\|_* = \sup_{\|\boldsymbol{u}\|\leq 1} \langle \boldsymbol{u}, \boldsymbol{v}\rangle.$$

The ℓ_p norm is dual to ℓ_q with $1/p + 1/q = 1$, $p, q \geq 1$. In particular, ℓ_1 is dual to ℓ_∞, while ℓ_2 is dual to itself.

Next, recall that the maximum of a linear function over a bounded set is achieved at the vertices. More precisely, for any $V \subset \mathbb{R}^d$, $\boldsymbol{a} \in \mathbb{R}^d$,

$$\sup_{\boldsymbol{u}\in \operatorname{conv}(V)} \langle \boldsymbol{a}, \boldsymbol{u}\rangle = \sup_{\boldsymbol{u}\in V} \langle \boldsymbol{a}, \boldsymbol{u}\rangle.$$

Next, we prove the following straightforward result.

Lemma 14. *Let $Z = (Z_1, \dots, Z_d)$ be a centered random variable with $Z \in \mathsf{subG}(\sigma^2)$. Then*

$$\mathbb{E}\,\|Z\|_2 \leq \sigma\sqrt{d}, \tag{5.69}$$

$$\mathbb{E}\max_i Z_i \leq \sigma\sqrt{2\log d}, \tag{5.70}$$

and

$$\mathbb{E}\,\|Z\|_\infty = \mathbb{E}\max_i |Z_i| \leq \sigma\sqrt{2\log(2d)} \tag{5.71}$$

Proof. First,

$$\mathbb{E}\,\|Z\|_2 = \mathbb{E}\sqrt{\|Z\|_2^2} \leq \sqrt{\mathbb{E}\,\|Z\|_2^2} = \sqrt{\sum_{i=1}^d \mathbb{E}Z_i^2} = \sigma\sqrt{d} \tag{5.72}$$

where we used the fact that variance of a random variable is at most its sub-Gaussian parameter (homework). Next, we prove (5.71). For any $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\max_i Z_i &= \frac{1}{\lambda}\mathbb{E}\max_{i=1,\ldots,d}\log\exp\{\lambda Z_i\} && (5.73)\\
&= \frac{1}{\lambda}\mathbb{E}\log\max_{i=1,\ldots,d}\exp\{\lambda Z_i\} && (5.74)\\
&\le \frac{1}{\lambda}\log\mathbb{E}\max_{i=1,\ldots,d}\exp\{\lambda Z_i\} && (5.75)\\
&\le \frac{1}{\lambda}\log\mathbb{E}\sum_{i=1}^{d}\exp\{\lambda Z_i\} && (5.76)\\
&\le \frac{1}{\lambda}\log\left[d\exp\{\lambda^2\sigma^2/2\}\right] && (5.77)
\end{aligned}$$

which is equal to

$$\frac{1}{\lambda}\log d + \frac{\lambda\sigma^2}{2} = \sigma\sqrt{2\log d} \tag{5.78}$$

upon choosing $\lambda = \sqrt{\frac{2\log d}{\sigma^2}}$. The estimate on $\|Z\|_\infty$ follows by considering $2d$ variables. □

We also leave the following as an exercise:

Lemma 15. *Let $Z_1, \ldots, Z_d$ be real-valued centered random variables satisfying*

$$\mathbb{E}\exp\{\lambda Z_i\} \le \exp\left\{\frac{\lambda^2 v^2}{2(1-b\lambda)}\right\}, \quad 0 < \lambda < 1/b.$$

Then

$$\mathbb{E}\max_i Z_i \le \sqrt{2v^2\log d} + b\log d \tag{5.79}$$

6. LINEAR REGRESSION

We now introduce the problem of linear regression, make the connection to the Gaussian Sequence Model, and motivate the need to study maximal inequalities.

Consider the model

$$Y_i = \langle x_i, \boldsymbol{\beta}^*\rangle + \varepsilon_i, \quad i = 1, \ldots, n \tag{6.80}$$

and $\boldsymbol{\beta}^*, x_i \in \mathbb{R}^d$. Assume the zero-mean noise satisfies $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathsf{subG}(\sigma^2)$. As usual, we assume that x_i's and Y_i's are observed, but the parameter vector $\boldsymbol{\beta}^*$ is unknown.

We can write the n equations together as

$$Y = X\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \tag{6.81}$$

where X is the $n \times d$ matrix with x_i as rows, and $Y = (Y_1, \ldots, Y_n)^\intercal$. For now, we will think of the matrix X as being fixed and given to us (this is called *fixed design*). In later parts of the course, we will work under the assumption that $x_1, \ldots, x_n$ are drawn i.i.d. from a distribution (i.e. *random design*).

6.1 Connection to the Gaussian Sequence Model

Multiplying both sides of (6.81) by $\frac{1}{n}X^\intercal$ yields

$$\frac{1}{n}X^\intercal Y = \frac{1}{n}X^\intercal X\boldsymbol{\beta}^* + \frac{1}{n}X^\intercal \varepsilon. \tag{6.82}$$

Consider the following assumption on the matrix X:

Definition 5. *If $\frac{1}{n}X^\intercal X = I_d$, we say that design (that is, the set $\{x_1, \ldots, x_n\}$) is orthonormal.*

In addition to orthogonality, the above definition implies that $\|x_i\|^2 = n$, which is what we would expect if coordinates of x_i were independent. In this case, (6.82) becomes

$$\tilde{Y} = \boldsymbol{\beta}^* + \tilde{\boldsymbol{\varepsilon}} \tag{6.83}$$

where $\tilde{\boldsymbol{\varepsilon}} = \frac{1}{n}X^\intercal\varepsilon$ and $\tilde{Y} = \frac{1}{n}X^\intercal Y$. If $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \sim \mathcal{N}(0, I_d)$, then $\tilde{\boldsymbol{\varepsilon}}$ is also Gaussian and mean-zero, and covariance is $\mathbb{E}\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^\intercal = \frac{1}{n^2}\mathbb{E}X^\intercal\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal X^\intercal = \frac{1}{n}I_d$. Hence, $\tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}(0, 1/n)$. We see that in orthogonal design, regression becomes the Gaussian Sequence Model. Furthermore, the problem of estimation $\boldsymbol{\beta}^*$ with respect to Euclidean norm becomes equivalent to the problem of bounding the prediction error: for an estimator $\widehat{\boldsymbol{\beta}}$

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\intercal(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\intercal(\frac{1}{n}X^\intercal X)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^n(\langle x_i, \widehat{\boldsymbol{\beta}}\rangle - \langle x_i, \boldsymbol{\beta}^*\rangle)^2 \tag{6.84}$$

6.2 Estimation, de-noising, and fixed design.

Several goals can be set for analyzing linear regression (or, any other estimator). Let us mention a few that will be central for the rest of the course.

First goal is estimation in some measure of distance on the space of parameters. For instance, a natural measure is $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2$ for the distance between the estimator $\widehat{\boldsymbol{\beta}}$ and the true parameter.

Another objective is to provide a bound on the error of the form

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

which can be called a "fixed-design error" or "de-noising objective." In other words, $x_1, \ldots, x_n$ are fixed, and we are interested in prediction (or de-noising of the y values) on these very points. Following the calculation in (6.84), but without assuming orthonormal design,

$$\frac{1}{n}\sum_{i=1}^n(\langle x_i, \widehat{\boldsymbol{\beta}}\rangle - \langle x_i, \boldsymbol{\beta}^*\rangle)^2 = \frac{1}{n}\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\|_2^2 = \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_\Sigma^2$$

where $\Sigma = \frac{1}{n}\sum_{i=1}^n x_i x_i^\intercal = \frac{1}{n}X^\intercal X$.

6.3 Unconstrained Least Squares

We now go back to the model (6.80), without the assumption on the matrix X. Our goal in this section will be to upper bound the fixed-design error.

Let $\widehat{\boldsymbol{\beta}}$ be the least-squares solution

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \|X\beta - Y\|^2 = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle x_i, \boldsymbol{\beta} \rangle)^2 \tag{6.85}$$

Setting the gradient of the objective to zero,

$$X^\intercal X \widehat{\boldsymbol{\beta}} = X^\intercal Y$$

and thus

$$\widehat{\boldsymbol{\beta}} = (X^\intercal X)^\dagger X^\intercal Y$$

where $A^\dagger$ denotes the Moore-Penrose inverse.

Rather than using the closed-form solution for the least squares, we will present analysis based on the optimality of the solution with respect to the empirical error. This analysis is more general and will hold for constrained least squares beyond linear regression.

First, observe that by optimality,

$$\left\| X\widehat{\boldsymbol{\beta}} - Y \right\|^2 \leq \|X\boldsymbol{\beta}^* - Y\|^2 = \|\boldsymbol{\varepsilon}\|^2 . \tag{6.86}$$

On the other hand,

$$\left\| X\widehat{\boldsymbol{\beta}} - Y \right\|^2 = \left\| X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^* - \boldsymbol{\varepsilon} \right\|^2 = \left\| X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^* \right\|^2 - 2\langle \boldsymbol{\varepsilon}, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^* \rangle + \|\boldsymbol{\varepsilon}\|^2 . \tag{6.87}$$

These two equations together yield the so-called *Basic Inequality*:

$$\frac{1}{n} \left\| X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^* \right\|^2 \leq \frac{2}{n} \langle \boldsymbol{\varepsilon}, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^* \rangle \tag{6.88}$$

Since generalizations of this inequality will be used many times in this course, it's worth making a few remarks. First, the inequality is deterministic. Second, on the left-hand side, we have our quantity of interest: the de-noising error. However, the right-hand side also involves our estimator $\widehat{\boldsymbol{\beta}}$. Moreover, if $X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*$ is small, we expect the product with the random $\boldsymbol{\varepsilon}$ to be "even smaller", which, in turn, leads to a smaller bound on $X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*$, which... Indeed, this argument will be formalized in terms of a certain fixed point in the second part of the course. Third, (6.86) will be the only place in our analysis where we use the fact that $\widehat{\boldsymbol{\beta}}$ minimizes empirical error, and other properties of $\widehat{\boldsymbol{\beta}}$ will be irrelevant. This observation will lead to immediate generalizations of the analysis beyond linear least squares.

Our strategy in analyzing least squares will be to "remove the hat" (i.e. the dependence of the right-hand side of the right-hand side of (6.88) on $\widehat{\boldsymbol{\beta}}$ by passing to a sufficiently localized supremum over all possible locations of $\widehat{\boldsymbol{\beta}}$.

6.4 Constrained Least Squares

Consider now a modification of the regression model (6.80), where we have the additional knowledge that $\boldsymbol{\beta}^* \in K$ for some set $K \subset \mathbb{R}^d$. It then makes sense to minimize squared error subject to being in K:

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in K}{\operatorname{argmin}} \|X\beta - Y\|^2 \tag{6.89}$$

Of course, the unconstrained model and the corresponding least squares solution in (6.85) corresponds to $K = \mathbb{R}^d$. Since the constrained case subsumes the unconstrained case, we will proceed below with the constrained analysis.

By examining (6.86) and (6.87), the Basic Inequality (6.88) still holds in constrained least squares.

6.5 Analyses of Least Squares: first strategy

For simplicity of exposition, we make the assumption that K is symmetric about the origin (i.e. $x \in K$ implies $-x \in K$). In this case, $K - K \subseteq 2K$.

We will take one of the following two paths in analyzing (6.88). The first, which leads to the so-called *fast rates*, is to divide both sides of (6.88) by $\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\|$ and then take a supremum over all unit vectors that can arise when $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon})$ ranges over K (or $\mathbb{R}^d$).

More precisely, the basic inequality leads to

$$\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\| \le 2\langle \boldsymbol{\varepsilon}, \frac{X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*}{\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\|}\rangle \le 2 \sup_{\boldsymbol{\beta} \in K} \langle \boldsymbol{\varepsilon}, \frac{X\boldsymbol{\beta} - X\boldsymbol{\beta}^*}{\|X\boldsymbol{\beta} - X\boldsymbol{\beta}^*\|}\rangle \tag{6.90}$$

Note that the right-hand side is now independent of the algorithm/estimator. We can now treat the right-hand side as a supremum of a collection of random variables indexed by $\boldsymbol{\beta}$. The smaller the collection, the smaller is the upper bound (other things kept equal). Once we have a high-probability bound on (6.90), we may square both sides to get a bound on the squared error. We should point out that it is quite surprising that such a simple strategy works for analyzing least squares. It may appear that the supremum on the right-hand side of (6.90) can be significantly larger than the middle part of that inequality.

To illustrate the strategy, we consider the unconstrained least squares (also called Ordinary Least Squares, OLS).

Lemma 16. *Assume the regression model $Y = X\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \in \mathsf{subG}(\sigma^2)$. Let $r = rank(X^\top X)$. Then the unconstrained OLS enjoys*

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\frac{1}{n}\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\|^2\right] \le \frac{4r\sigma^2}{n}. \tag{6.91}$$

Proof. Let us write the basic inequality as

$$\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\| \le 2\langle \boldsymbol{\varepsilon}, \boldsymbol{v}(\boldsymbol{\varepsilon})\rangle \tag{6.92}$$

where $\boldsymbol{v}(\boldsymbol{\varepsilon}) = \frac{X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*}{\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\|}$. Let $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]$ be the matrix with orthonormal columns, a basis of the column space of X. Since $X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ is in the column space of X, we can write $\boldsymbol{v}(\boldsymbol{\varepsilon}) = \mathbf{U}\boldsymbol{a}$ for some $\boldsymbol{a} = \boldsymbol{a}(\boldsymbol{\varepsilon}) \in \mathbb{R}^r$ with $\|\boldsymbol{a}\| = 1$. Then

$$\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\| \leq 2 \sup_{\|\boldsymbol{a}\| \leq 1} \langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{a} \rangle = 2 \sup_{\|\boldsymbol{a}\| \leq 1} \langle \mathbf{U}^\mathsf{T} \boldsymbol{\varepsilon}, \boldsymbol{a} \rangle = 2 \left\|\mathbf{U}^\mathsf{T} \boldsymbol{\varepsilon}\right\| \tag{6.93}$$

since Euclidean norm is self-dual. Squaring both sides,

$$\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\|^2 \leq 4 \left\|\mathbf{U}^\mathsf{T} \boldsymbol{\varepsilon}\right\|^2 \tag{6.94}$$

It is easy to see that $\mathbf{U}^\mathsf{T}\boldsymbol{\varepsilon} \in \mathsf{subG}(\sigma^2)$. Indeed, for any $\boldsymbol{w} \in \mathbb{R}^r$ with $\|\boldsymbol{w}\| = 1$,

$$\mathbb{E} \exp\{\lambda \langle \boldsymbol{w}, \mathbf{U}^\mathsf{T}\boldsymbol{\varepsilon} \rangle\} = \mathbb{E} \exp\{\lambda \langle \mathbf{U}\boldsymbol{w}, \boldsymbol{\varepsilon} \rangle\} \leq \exp\{\lambda^2 \sigma^2 / 2\}.$$

Since variance of each coordinate of $\mathbf{U}^\mathsf{T}\boldsymbol{\varepsilon}$ is at most σ^2, the result follows. □

In particular, $r \leq \min\{n, d\}$, and if $r = d$, the guarantee takes on the familiar form of $\frac{\sigma^2 d}{n}$.

6.6 Analyses of Least Squares: second strategy

Consider constrained least squares with $K = \mathsf{B}_1^d = \{q \in \mathbb{R}^d : \sum |q_i| \leq 1\}$. We have that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in K - K = 2\mathsf{B}_1^d$. Note that $2\mathsf{B}_1^d$ has $2d$ vertices, and thus $X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \in X \cdot (2\mathsf{B}_1^d)$ also has at most $2d$ vertices.

Unfortunately, when we normalize $X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, we "lose" the vertices, as

$$\boldsymbol{v}(\boldsymbol{\varepsilon}) = \frac{X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{\left\|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|} \in \mathbf{S}^{n-1} \cap \mathrm{col}(X).$$

Such an approach, at least directly, would lead to rates of the previous section, without exploiting the structure of K.

We proceed by avoiding normalization and directly analyzing the basic inequality:

$$\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\|^2 \leq 2\langle \boldsymbol{\varepsilon}, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^* \rangle \leq 2 \max_{\boldsymbol{v} \in 2\mathsf{B}_1^d} \langle \boldsymbol{\varepsilon}, X\boldsymbol{v} \rangle = 4 \left\|X^\mathsf{T}\boldsymbol{\varepsilon}\right\|_\infty \tag{6.95}$$

Let x^i denote the ith column of X. From (5.71), and observing that $\langle x^i, \boldsymbol{\varepsilon} \rangle \in \mathsf{subG}(\left\|x^i\right\|^2 \sigma^2)$,

$$\mathbb{E} \left\|X^\mathsf{T}\boldsymbol{\varepsilon}\right\|_\infty \leq \sigma\sqrt{2\log(2d)} \cdot \max_i \left\|x^i\right\|$$

A natural normalization of the data is $\left\|x^i\right\| \leq \sqrt{n}$, in which case we have proved the following result:

Lemma 17. *Assume the regression model $Y = X\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \in \mathsf{subG}(\sigma^2)$ and $K = \mathsf{B}_1^d$. Suppose columns of X are normalized to be $\left\|x^i\right\| \leq \sqrt{n}$. Then the constrained OLS enjoys*

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \left[\frac{1}{n} \left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\|^2\right] \leq 4\sigma\sqrt{\frac{2\log(2d)}{n}}. \tag{6.96}$$

Since analysis of previous section applies to the constrained least squares as well, we have that

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\frac{1}{n}\left\|X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\beta}^*\right\|^2\right]\le 4\min\left\{\sigma\sqrt{\frac{2\log(2d)}{n}},\frac{r\sigma^2}{n}\right\},$$

where r is the rank of X. Disregarding the logarithmic factors, the transition between the two rates is at $r\sim\sqrt{n}$. The upper bound in (6.96) is sometimes referred to as the "slow rate," as opposed to the "fast rate" in (6.91). The fast rate kicks in for problems with small dimensionality (or rank), while the slow rate wins in high-dimensional situation. We will see how these two regimes arise more generally in parametric and nonparametric regression through the lens of covering numbers.

6.7 Sparsity

To close our discussion of linear regression, consider one more example of constrained least squares, where $\boldsymbol{\beta}^*$ is known to be sparse. In other words, assume that

$$\boldsymbol{\beta}^*\in\mathsf{B}_0^d(k)=\{\boldsymbol{\beta}\in\mathbb{R}^d:\|\boldsymbol{\beta}\|_0\le k\}.$$

Let $\widehat{\boldsymbol{\beta}}$ be the constrained least squares solution, and assume, as before, that $\boldsymbol{\varepsilon}\in\mathsf{subG}(\sigma^2)$. Constrained least squares is an inefficient method, as it requires enumeration over $\binom{d}{k}$ subsets. Note that $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\in 2K=2\mathsf{B}_0^d(k)=\mathsf{B}_0^d(2k)$, i.e. this difference vector is at most $2k$-sparse.

We shall proceed with the first approach. The Basic Inequality then gives

$$\left\|X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\beta}^*\right\|\le 2\langle\boldsymbol{\varepsilon},\frac{X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\beta}^*}{\left\|X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\beta}^*\right\|}\rangle\le 2\max_{S\subset[d],|S|\le 2k}\ \sup_{\boldsymbol{w}:\mathsf{supp}(\boldsymbol{w})\subset S}\langle\boldsymbol{\varepsilon},\frac{X\boldsymbol{w}}{\|X\boldsymbol{w}\|}\rangle \tag{6.97}$$

As before, let $\mathbf{U}_S$ be the orthonormal basis of the span of columns of X corresponding to index set S. We can then write $(X\boldsymbol{w})/\|X\boldsymbol{w}\|=\mathbf{U}_S\boldsymbol{a}$ for some $\boldsymbol{a}\in\mathsf{B}_2^{2k}$. Then we have

$$\left\|X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\beta}^*\right\|\le 2\max_{S\subset[d],|S|\le 2k}\ \sup_{\boldsymbol{a}\in\mathsf{B}_2^{2k}}\langle\mathbf{U}_S^{\mathsf{T}}\boldsymbol{\varepsilon},\boldsymbol{a}\rangle=2\max_{S\subset[d],|S|\le 2k}\|\mathbf{U}_S^{\mathsf{T}}\boldsymbol{\varepsilon}\| \tag{6.98}$$

and

$$\mathbb{E}\left\|X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\beta}^*\right\|^2\le 4\mathbb{E}\max_{S\subset[d],|S|\le 2k}\|\mathbf{U}_S^{\mathsf{T}}\boldsymbol{\varepsilon}\|^2 \tag{6.99}$$

$$\le 4\mathbb{E}\max_{S\subset[d],|S|\le 2k}\left\{\|\mathbf{U}_S^{\mathsf{T}}\boldsymbol{\varepsilon}\|^2-\mathbb{E}\|\mathbf{U}_S^{\mathsf{T}}\boldsymbol{\varepsilon}\|^2\right\}+\max_{S\subset[d],|S|\le 2k}\mathbb{E}\|\mathbf{U}_S^{\mathsf{T}}\boldsymbol{\varepsilon}\|^2. \tag{6.100}$$

where in the last expression we centered the random variables.

This first quantity in the last expression is an expected maximum of random variables $\|\mathbf{U}_S^{\mathsf{T}}\boldsymbol{\varepsilon}\|^2$, indexed by S, and centered. There are at most $\binom{d}{2k}$ such variables, and each variable is sub-exponential (indeed, each coordinate j of the vector $\mathbf{U}_S^{\mathsf{T}}\boldsymbol{\varepsilon}$ is a σ^2-sub-Gaussian random variable). Thus, this expected maximum is at most

$$C\sigma^2\log\binom{d}{2k}$$

for some absolute constant C. The second term is at most $2k\sigma^2$, a lower-order term. With the standard estimate $\binom{d}{i} \le \left(\frac{ed}{i}\right)^i$, we conclude that

$$\mathbb{E}\left\|X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*\right\|^2 \lesssim \sigma^2 \frac{k\log(d/k)}{n}. \tag{6.101}$$

7. COVERING NUMBERS: AN INTRODUCTION

7.1 ℓ_2 ball cover

In analyzing linear regression, we started with the Basic Inequality and turned its estimator-dependent upper bound into a maximum of a collection of random variables, e.g. (6.93), (6.96), and (6.97). In these examples, the maximum was expressed conveniently as either the ℓ_2 or ℓ_∞ norm of a vector-valued random variable. For constrained least squares with other sets K, we may not have such a convenient closed form, and we are seeking to understand a more general principle behind developing maximal inequalities. At the very least, we would like to see how to unify the analysis of (5.69) and (5.71), which at the moment appears to be very different: the first uses properties of the ℓ_2 norm (specifically, smoothness) while the second relies on finiteness of extremal points. Is there a unified analysis or principle for these two results? Indeed, there is.

We start with a notion of a cover for a subset of $\mathbb{R}^d$.

Definition 6. *Given $K \subset \mathbb{R}^d$ and $\varepsilon \ge 0$, a set $V \subset \mathbb{R}^d$ is an ε-net (equivalently, an ε-cover) with respect to distance measure ρ on $\mathbb{R}^d$ if for any $x \in K$ there exists a $v \in V$ such that $\rho(x, v) \le \varepsilon$. The* covering number *$\mathcal{N}(K, \rho, \varepsilon)$ is the cardinality of the smallest such cover.*

Clearly, the definition extends beyond $\mathbb{R}^d$ to any metric space (X, ρ). If $V \subset K$, the cover is called *proper*.

Lemma 18. *For any $\varepsilon \in (0, 1]$,*

$$\mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon) \le \left(1 + \frac{2}{\varepsilon}\right)^d$$

Proof. We use the following volume argument. We add centers $v_1, v_2, \ldots \in \mathsf{B}_2^d$ such that $\|v_i - v_j\| > \varepsilon$ for every $i \ne j$, until no such additional point exists. Let N be the size of this set, which is clearly an ε-net. Then

$$N \cdot \mathsf{vol}\left(\frac{\varepsilon}{2}\mathsf{B}_2^d\right) \le \mathsf{vol}\left(\mathsf{B}_2^d + \frac{\varepsilon}{2}\mathsf{B}_2^d\right) \qquad \Rightarrow \qquad N \le \frac{\left(1 + \frac{\varepsilon}{2}\right)^d}{\left(\frac{\varepsilon}{2}\right)^d} = \left(1 + \frac{2}{\varepsilon}\right)^d.$$

□

Since we can start the iterative process of placing ε-balls with $v_1 = 0$, we can assume without loss of generality that the minimal cover contains 0.

Lemma 19. *Let V be a cover of B_2^d at scale $\varepsilon \in (0,1)$ with respect to $\|\cdot\|_2$. Then for any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\max_{\boldsymbol{u}\in\mathsf{B}_2^d} \langle \boldsymbol{u}, \boldsymbol{x}\rangle \leq \frac{1}{1-\varepsilon} \max_{\boldsymbol{v}\in V} \langle \boldsymbol{v}, \boldsymbol{x}\rangle. \tag{7.102}$$

Proof. For any $\boldsymbol{u} \in \mathsf{B}_2^d$, there exists $\boldsymbol{v} \in V$ such that $\|\boldsymbol{u} - \boldsymbol{v}\| \leq \varepsilon$ (i.e. $\boldsymbol{u} - \boldsymbol{v} \in \varepsilon\mathsf{B}_2^d$). Since we have $\langle \boldsymbol{u}, \boldsymbol{x}\rangle = \langle \boldsymbol{v}, \boldsymbol{x}\rangle + \langle \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{x}\rangle$, it also holds that

$$\langle \boldsymbol{u}, \boldsymbol{x}\rangle \leq \max_{\boldsymbol{v}\in V} \langle \boldsymbol{v}, \boldsymbol{x}\rangle + \max_{\boldsymbol{w}\in\varepsilon\mathsf{B}_2^d} \langle \boldsymbol{w}, \boldsymbol{x}\rangle.$$

By linearity, the last term is $\varepsilon \max_{\boldsymbol{w}\in\mathsf{B}_2^d} \langle \boldsymbol{w}, \boldsymbol{x}\rangle$. Since the choice of $\boldsymbol{u} \in \mathsf{B}_2^d$ was arbitrary, the statement follows by rearranging the terms. □

We remark that this lemma trivially extends to norms beyond Euclidean, as long as the ball is covered in the very norm with respect to which it is defined. This situation is rather special, and we will use this comparison result only a couple of times in this course, with a constant ε.

7.2 Recovering (5.69) via covering numbers

As a sanity check, let's see if we can recover (5.69). To this end, let $Z \in \mathsf{subG}(\sigma^2)$ be a random d-dimensional vector. Then

$$\mathbb{E}\,\|Z\| = \mathbb{E} \max_{\boldsymbol{u}\in\mathsf{B}_2^d} \langle \boldsymbol{u}, Z\rangle \leq 2\mathbb{E} \max_{\boldsymbol{v}\in V} \langle \boldsymbol{v}, Z\rangle \tag{7.103}$$

where V is a minimal $1/2$-cover of B_2^d. Since $|V| \leq 5^d$ by Lemma 19, we can conclude from (5.70) that

$$\mathbb{E}\,\|Z\| = \mathbb{E} \max_{\boldsymbol{u}\in\mathsf{B}_2^d} \langle \boldsymbol{u}, Z\rangle \leq 2\sigma\sqrt{2\log(5^d)} \leq C\sigma\sqrt{d} \tag{7.104}$$

for $C = 2\sqrt{2\log 5}$. While the approach through finite discretization appears to "unify" both (5.69) and (5.71) (in the latter case, the set is already discrete), it did not recover the same constant 1 as in (5.69).

7.3 Operator norm

Recall that for a matrix $A \in \mathbb{R}^{p\times q}$, an operator norm (in the $\ell_2^q \to \ell_2^p$ sense) is defined as

$$\|A\| = \sup_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\| = \sup_{\|\boldsymbol{x}\|=1, \|\boldsymbol{y}\|=1} \boldsymbol{y}^\top A\boldsymbol{x}$$

where $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^q$. This norm is also known as the spectral norm since $\|A\| = \sqrt{\lambda_{\mathsf{max}}(A^\top A)}$, the square root of the largest eigenvalue, which is also the largest singular value of A, which we shall denote as $\sigma_{\mathsf{max}}(A)$. We emphasize our convention that the unadorned norm $\|\cdot\|$ for vectors stands for the Euclidean norm (unless stated otherwise), and the unadorned norm $\|\cdot\|$ for matrices will stand for the operator norm.

We have the following extension of Lemma 19, see [33, p. 84].

Lemma 20. *Let $\varepsilon \in (0, 1/2)$. Let V and U be proper ε-nets of B_2^p and B_2^q, respectively, with respect to Euclidean norm. Without loss of generality, assume $\mathbf{0} \in V, U$. Then for any $A \in \mathbb{R}^{p\times q}$,*

$$\max_{\boldsymbol{v}\in V, \boldsymbol{u}\in U} \boldsymbol{v}^\top A\boldsymbol{u} \le \|A\| \le \frac{1}{1-2\varepsilon} \max_{\boldsymbol{v}\in V, \boldsymbol{u}\in U} \boldsymbol{v}^\top A\boldsymbol{u}. \tag{7.105}$$

Furthermore, if $p = q$, it holds that

$$\|A\| \le \frac{1}{1-2\varepsilon} \max_{\boldsymbol{v}\in V} \boldsymbol{v}^\top A\boldsymbol{v}. \tag{7.106}$$

Proof. Let $\|A\| = \|A\boldsymbol{x}\|$ for $\|\boldsymbol{x}\| = 1$ and let $\boldsymbol{u} \in U$ be such that $\|\boldsymbol{x} - \boldsymbol{u}\| \le \varepsilon$. Then

$$\|A\| = \|A\boldsymbol{x}\| \le \|A\boldsymbol{u}\| + \|A(\boldsymbol{x} - \boldsymbol{u})\| \le \|A\boldsymbol{u}\| + \|A\|\,\varepsilon.$$

Then, combining with Lemma 19,

$$(1-\varepsilon)\,\|A\| \le \|A\boldsymbol{u}\| \le \frac{1}{1-\varepsilon} \max_{\boldsymbol{v}\in V} \langle \boldsymbol{v}, A\boldsymbol{u}\rangle \tag{7.107}$$

Taking maximum over $\boldsymbol{u}$, and noting that $(1-\varepsilon)^{-2} \le (1-2\varepsilon)^{-1}$, the upper bound follows. The lower bound is immediate since the ε-nets are proper.

For the second statement, let $\boldsymbol{x}$ be such that $\|A\| = |\boldsymbol{x}^\top A\boldsymbol{x}|$. We have for any $\boldsymbol{v}$ that is $\|\boldsymbol{x} - \boldsymbol{v}\| \le \varepsilon$,

$$\langle \boldsymbol{x}, A\boldsymbol{x}\rangle = \langle \boldsymbol{v}, A\boldsymbol{v}\rangle + \langle \boldsymbol{v}, A(\boldsymbol{x} - \boldsymbol{v})\rangle + \langle \boldsymbol{x} - \boldsymbol{v}, A\boldsymbol{x}\rangle \le \langle \boldsymbol{v}, A\boldsymbol{v}\rangle + 2\varepsilon\,\|A\|$$

and thus

$$\|A\| \le |\langle \boldsymbol{v}, A\boldsymbol{v}\rangle| + 2\varepsilon\,\|A\| \le \max_{\boldsymbol{v}\in V} |\langle \boldsymbol{v}, A\boldsymbol{v}\rangle| + 2\varepsilon\,\|A\|\,.$$

□

The following result appears, for example, in [33, Thm 4.4.5].

Lemma 21. *Suppose $A \in \mathbb{R}^{p\times q}$ be a random matrix with mean-zero independent σ^2-sub-Gaussian entries. Then for any $t > 0$,*

$$\|A\| \le C\sigma(\sqrt{p} + \sqrt{q} + t) \tag{7.108}$$

with probability at least $1 - \exp\{-t^2\}$ for some absolute constant C. Hence,

$$\mathbb{E}\,\|A\| \lesssim \sigma(\sqrt{p} + \sqrt{q}) \tag{7.109}$$

Proof. Let V, U be, respectively, $1/4$-nets for B_2^p and B_2^q, of size 9^p and 9^q as guaranteed by Lemma 18. From Lemma 20,

$$\max_{\boldsymbol{v}\in V, \boldsymbol{u}\in U} \boldsymbol{v}^\top A\boldsymbol{u} \le \|A\| \le 2 \max_{\boldsymbol{v}\in V, \boldsymbol{u}\in U} \boldsymbol{v}^\top A\boldsymbol{u}. \tag{7.110}$$

For any fixed $\boldsymbol{v} \in V, \boldsymbol{u} \in U$, the random variable $\boldsymbol{v}^\top A\boldsymbol{u} = \sum_{i=1}^p \sum_{j=1}^q A_{i,j}\boldsymbol{v}_i\boldsymbol{u}_j$ is sub-Gaussian with variance proxy $\sigma^2 \sum_{i=1}^p \sum_{j=1}^q A_{i,j}\boldsymbol{v}_i^2\boldsymbol{u}_j^2 \le \sigma^2$. Hence,

$$\mathbb{P}\left(\boldsymbol{v}^\top A\boldsymbol{u} \ge t\sigma\right) \le \exp\left\{-t^2/2\right\}.$$

Then by union bound,

$$\mathbb{P}\left(\exists \boldsymbol{u} \in U, \boldsymbol{v} \in V : \quad \boldsymbol{v}^\top A\boldsymbol{u} \ge t\sigma\right) \le 9^{p+q}\exp\left\{-t^2/2\right\}.$$

Substituting $\sqrt{C}(\sqrt{p}+\sqrt{q}+t)$ in place of t, for some absolute constant $C > 0$, and noting that $9^{p+q}\exp\{-C(\sqrt{p}+\sqrt{q}+t)^2\} \le \exp -Ct^2$ for C large enough, we conclude the proof. □

Note: as a corollary, for a symmetric (Wigner) random matrix $A \in \mathbb{R}^{p\times p}$ with independent σ^2-sub-Gaussian entries above the diagonal,

$$\mathbb{E}\,\|A\| \lesssim \sigma\sqrt{p}.$$

This holds by applying the above lemma separately to the upper and lower triangular components of A.

8. COVARIANCE ESTIMATION

Let $X_1, \dots, X_n \in \mathbb{R}^d$ be i.i.d. sub-Gaussian centered random variables with covariance Σ. Recall that sample covariance is $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_iX_i^\top$. We would like to measure the quality of the approximation of Σ by $\widehat{\Sigma}$ via the spectral norm:

$$\left\|\Sigma - \widehat{\Sigma}\right\| = \max_{\boldsymbol{v}\in \mathsf{S}^{d-1}} \left|\frac{1}{n}\sum_{i=1}^n \langle X_i, \boldsymbol{v}\rangle^2 - \boldsymbol{v}^\top\Sigma\boldsymbol{v}\right| \tag{8.111}$$

Lemma 21 does not directly apply here since entries of $\Sigma - \widehat{\Sigma}$ are not independent.

Lemma 22. *Let $X_1, \dots, X_n \in \mathbb{R}^d$ be i.i.d. with mean zero and $X_i \in \mathsf{subG}(\sigma^2)$. Let $\mathbb{E}X_iX_i^\top = \Sigma$ and let $\widehat{\Sigma}$ be the sample covariance matrix. Then*

$$\mathbb{P}\left(\left\|\Sigma - \widehat{\Sigma}\right\| \ge \sigma^2 C \max\left\{\sqrt{\frac{d}{n}+\frac{t}{n}}, \frac{d}{n}+\frac{t}{n}\right\}\right) \le 2\exp\{-t\} \tag{8.112}$$

for some absolute constant C.

Proof. Define shorthand $Q = \widehat{\Sigma} - \Sigma$. Let V be a $1/8$-cover of B_2^d of size at most 17^d. A small modification of the proof of Lemma 20 implies that

$$\|Q\|_2 \le 2\max_{\boldsymbol{v}\in V} |\langle \boldsymbol{v}, Q\boldsymbol{v}\rangle|.$$

Now, $X_i \in \mathsf{subG}(\sigma^2)$ implies that for any $\boldsymbol{u} \in \mathsf{B}_2^d$, the random variable $\langle X_i, \boldsymbol{u}\rangle^2 - \langle \boldsymbol{u}, \Sigma\boldsymbol{u}\rangle$ is sub-exponential with parameters $(c\sigma^2, c\sigma^2)$ for some absolute constant c. From (3.22), rescaling by $c\sigma^2$,

$$\mathbb{P}\left(|\langle \boldsymbol{u}, Q\boldsymbol{u}\rangle| \ge c\varepsilon\sigma^2\right) \le 2\exp\left\{-n\min\{\varepsilon, \varepsilon^2\}\right\}$$

Taking a union bound over the discretization,

$$\mathbb{P}\left(\exists \boldsymbol{v} \in V : |\langle \boldsymbol{v}, Q\boldsymbol{v}\rangle| \geq c\varepsilon\sigma^2\right) \leq 17^d \cdot 2\exp\left\{-n\min\{\varepsilon, \varepsilon^2\}\right\}$$

and thus

$$\mathbb{P}\left(\exists \boldsymbol{v} \in V : |\langle \boldsymbol{v}, Q\boldsymbol{v}\rangle| \geq \varepsilon\sigma^2\right) \leq 2\exp\left\{-c(n\min\{\varepsilon, \varepsilon^2\} - d)\right\}$$

Now for some $t > 0$, choose

$$\varepsilon = \max\left\{\sqrt{\frac{d}{n} + \frac{t}{n}}, \frac{d}{n} + \frac{t}{n}\right\}.$$

Then

$$\min\{\varepsilon, \varepsilon^2\} = \frac{d}{n} + \frac{t}{n}$$

and

$$n\min\{\varepsilon, \varepsilon^2\} - d = t.$$

This yields

$$\mathbb{P}\left(\left\|\widehat{\Sigma} - \Sigma\right\|_2 \geq 2\sigma^2 \max\left\{\sqrt{\frac{d}{n} + \frac{t}{n}}, \frac{d}{n} + \frac{t}{n}\right\}\right) \leq 2\exp\{-ct\}$$

for some absolute constant $c > 0$. □

Note that if $X_i \sim N(0, \Sigma)$, we have $\langle \boldsymbol{v}, X_i\rangle \sim N(0, \boldsymbol{v}^\top \Sigma \boldsymbol{v})$. Since for any unit vector $\boldsymbol{v}$, $\boldsymbol{v}^\top \Sigma \boldsymbol{v} \leq \|\Sigma\|$, X_i is a sub-Gaussian vector with variance proxy at most $\|\Sigma\|$. More generally, if we assume that the sub-Gaussian parameter of X_i is at most $C\|\Sigma\|$ for some constant C, then we have the following corollary:

Corollary 1. *In the setting of Lemma 22, if we additionally assume that $X_i \in \mathsf{subG}(\|\Sigma\|)$, then*

$$\mathbb{P}\left(\left\|\Sigma - \widehat{\Sigma}\right\| \geq \|\Sigma\| C \max\left\{\sqrt{\frac{d}{n} + \frac{t}{n}}, \frac{d}{n} + \frac{t}{n}\right\}\right) \leq 2\exp\{-t\} \tag{8.113}$$

Furthermore,

$$\mathbb{E}\left\|\Sigma - \widehat{\Sigma}\right\| \lesssim \|\Sigma\| \max\left\{\sqrt{\frac{d}{n}}, \frac{d}{n}\right\}$$

As discussed before, if $d = o(n)$, then sample covariance $\widehat{\Sigma}$ is a consistent estimator of Σ and we have an explicit rate. Let us mention one more result, in terms of effective rank (or, stable rank)

$$\boldsymbol{r}(\Sigma) = \frac{\mathsf{tr}(\Sigma)}{\|\Sigma\|}. \tag{8.114}$$

Note that a similar quantity arose in mean estimation in high dimension. The numerator here is the sum of eigenvalues of Σ, while the denominator is the largest eigenvalue. The ratio has the right "units" to qualify for a notion of a dimension. If $\Sigma = I_d$, we have

$\mathsf{tr}(\Sigma) = d$ and $\|\Sigma\| = 1$. More generally, effective rank can be small even though d is large, as long as the eigenvalues decay fast enough.

The more general result says that X_i are sub-Gaussian centered vectors such that the sub-Gaussian parameter of any one-dimensional projection $\langle X_i, \boldsymbol{u}\rangle$ is at most a constant multiple of its variance, it holds that

$$\mathbb{E}\left\|\Sigma - \widehat{\Sigma}\right\| \lesssim \|\Sigma\| \max\left\{\sqrt{\frac{\boldsymbol{r}(\Sigma)}{n}}, \frac{\boldsymbol{r}(\Sigma)}{n}\right\}$$

(see [17], [33, Theorem 9.2.4])

8.1 Singular values

The reason we had the two-tailed behavior of the spectral norm $\left\|\Sigma - \widehat{\Sigma}\right\|$ is that $\widehat{\Sigma}$ is an average of "squares" of a sub-Gaussian random variables. If you recall, in the earlier lecture, we deduced pure sub-Gaussian tails by taking square root of the random variable. The analogue here will be the singular values $\sigma_i(X)$ of the data matrix $X \in \mathbb{R}^{n\times d}$ which has rows X_i^{T}.

Recall that singular values of the matrix X and eigenvalues of sample covariance $\widehat{\Sigma}$ are related as

$$\sigma_j(X) = \sqrt{\lambda_j(X^{\mathsf{T}}X)}$$

or, rescaling,

$$\sigma_j(\frac{1}{\sqrt{n}}X) = \sqrt{\lambda_j(\frac{1}{n}X^{\mathsf{T}}X)} = \sqrt{\lambda_j(\widehat{\Sigma})}.$$

Weyl's Inequality then says that

$$\max_{j=1,\dots,d} |\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)| \le \left\|\widehat{\Sigma} - \Sigma\right\| \tag{8.115}$$

Suppose for the purposes of illustration that $\Sigma = I_d$ (i.e. the random variables are isotropic). Then our results tell us that

$$\left\|\widehat{\Sigma} - I_d\right\| \lesssim \sqrt{\frac{d}{n} + \frac{t}{n}} \vee \frac{d}{n} + \frac{t}{n}$$

with probability at least $2e^{-t}$, which means that for all $i = 1, \dots, d$,

$$\left|\sigma_i^2(\frac{1}{\sqrt{n}}X) - 1\right| \lesssim \sqrt{\frac{d}{n} + \frac{t}{n}} \vee \frac{d}{n} + \frac{t}{n}$$

Since $\max(|z-1|, |z-1|^2) \le |z^2 - 1|$ for $z \ge 0$, we get

$$\max\left\{|\sigma_i(\frac{1}{\sqrt{n}}X) - 1|, |\sigma_i(\frac{1}{\sqrt{n}}X) - 1|^2\right\} \le |\sigma_i^2(\frac{1}{\sqrt{n}}X) - 1|$$

which implies, after rescaling, that

$$|\sigma_i(X) - \sqrt{n}| \lesssim \sqrt{d} + \sqrt{t}$$

with probability at least $1 - 2\exp\{-t\}$. In other words, the singular values of a tall $(n > d)$ matrix X with sub-Gaussian isotropic rows can be found to be tightly concentrated in the interval $[\sqrt{n} - C\sqrt{d}, \sqrt{n} + C\sqrt{d}]$. This result holds in more generality than stated here, and we refer to [32, 33].

9. SPECTRAL METHODS

9.1 Perturbation Analysis

We now present several additional models that involve random matrices. In these applications, such as Principal Component Analysis, we will think of the matrix Y as a noisy observation of some signal matrix X, with additive noise $\mathcal{E}$, i.e.

$$Y = X + \mathcal{E}. \tag{9.116}$$

We will be mainly interested in estimating top eigenvector(s) of X from the noisy observation Y. For now, however, we think of $\mathcal{E}$ as a non-random perturbation of X. How does this perturbation affect the spectral properties? While Weil's inequality (see e.g. (8.115)) tells us that eigenvalues to not change much when the perturbation of the matrix is small in spectral norm, it does not say anything about closeness of eigenvectors. So, it is natural to ask: Are eigenvectors of X and Y close if $\|X - Y\|$ is small?

To provide some intuition, consider the following example, with some $\delta > 0$:

$$X = \begin{bmatrix} 1+\delta & 0 \\ 0 & 1-\delta \end{bmatrix}, \quad \mathcal{E} = \begin{bmatrix} -\delta & \delta \\ \delta & \delta \end{bmatrix}, \quad Y = X + \mathcal{E} = \begin{bmatrix} 1 & \delta \\ \delta & 1 \end{bmatrix} \tag{9.117}$$

The eigenvalues of X are $1+\delta$ and $1-\delta$, with the corresponding eigenvectors $\boldsymbol{u}_1 = \boldsymbol{e}_1$, $\boldsymbol{u}_2 = \boldsymbol{e}_2$. On the other hand, the eigenvalues of Y are also $1+\delta$ and $1-\delta$, with eigenvectors $\boldsymbol{v}_1 = [2^{-1/2}, 2^{-1/2}]^\intercal$ and $\boldsymbol{v}_2 = [-2^{-1/2}, 2^{-1/2}]^\intercal$. We see that $\boldsymbol{u}_1^\intercal \boldsymbol{v}_1 = 2^{-1/2}$, i.e. the eigenvectors rotated by 45 degrees because of the perturbation, even though the eigenvalues remained the same. The source of the instability of eigenvalues is the gap between the eigenvalues of X which is on the same order as the size of the perturbation $\mathcal{E}$, when measured, say, in the operator norm. One may wonder if, in general, this is the only reason that eigenvectors may move significantly. Indeed, that's the case, as shown below.

We state a simplified version of the Davis-Kahan "sin(θ)" theorem (see [36]).

Theorem 1. *Let $X, Y \in \mathbb{R}^{d\times d}$ be symmetric matrices with, respectively, eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ and $\mu_1 \geq \ldots \geq \mu_d$, as well as eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$. Then*

$$\sin(\theta) \leq \frac{2\,\|X - Y\|}{\max\{\lambda_1 - \lambda_2, \mu_1 - \mu_2\}} \tag{9.118}$$

where $\theta = \arccos(|\langle \boldsymbol{u}_1, \boldsymbol{v}_1 \rangle|)$ is the principal angle between $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$.

Proof. First, we have

$$\begin{aligned} \langle \boldsymbol{u}_1, X\boldsymbol{u}_1 \rangle - \langle \boldsymbol{v}_1, X\boldsymbol{v}_1 \rangle &= \langle \boldsymbol{u}_1, Y\boldsymbol{u}_1 \rangle - \langle \boldsymbol{v}_1, X\boldsymbol{v}_1 \rangle + \langle \boldsymbol{u}_1, (X-Y)\boldsymbol{u}_1 \rangle \\ &\leq \langle \boldsymbol{v}_1, Y\boldsymbol{v}_1 \rangle - \langle \boldsymbol{v}_1, X\boldsymbol{v}_1 \rangle + \langle \boldsymbol{u}_1, (X-Y)\boldsymbol{u}_1 \rangle \\ &= \langle X - Y, \boldsymbol{u}_1\boldsymbol{u}_1^\intercal - \boldsymbol{v}_1\boldsymbol{v}_1^\intercal \rangle \end{aligned}$$

where the last inner product is to be understood as trace. The last expression is at most

$$\|X - Y\| \cdot \|\boldsymbol{u}_1\boldsymbol{u}_1^\intercal - \boldsymbol{v}_1\boldsymbol{v}_1^\intercal\|_1 \leq \|X - Y\| \cdot \|\boldsymbol{u}_1\boldsymbol{u}_1^\intercal - \boldsymbol{v}_1\boldsymbol{v}_1^\intercal\|_{\mathsf{F}} \cdot \sqrt{2}$$

where $\|\cdot\|_1$ is the nuclear norm (ℓ_1 of eigenvalues) and $\|\cdot\|_{\mathsf{F}}$ is the Frobenius norm (ℓ_2 of eigenvalues).

On the one hand,

$$\|\boldsymbol{u}_1\boldsymbol{u}_1^{\mathsf{T}} - \boldsymbol{v}_1\boldsymbol{v}_1^{\mathsf{T}}\|_{\mathsf{F}}^2 = 2 - 2\langle \boldsymbol{u}_1, \boldsymbol{v}_1\rangle^2 = 2\sin^2(\theta).$$

On the other hand, the values $\langle \boldsymbol{u}_1, X\boldsymbol{u}_1\rangle$ and $\langle \boldsymbol{v}_1, X\boldsymbol{v}_1\rangle$ should be different if the angle is large and there is a gap in the eigenvalues λ_1 and λ_2. More precisely,

$$\langle \boldsymbol{v}_1, X\boldsymbol{v}_1\rangle = \sum_{j=1}^{d} \lambda_j \langle \boldsymbol{u}_j, \boldsymbol{v}_1\rangle^2 \le \lambda_1\langle \boldsymbol{u}_1, \boldsymbol{v}_1\rangle^2 + \lambda_2(1 - \langle \boldsymbol{u}_1, \boldsymbol{v}_1\rangle^2) = \lambda_1\cos^2(\theta) + \lambda_2\sin^2(\theta)$$

since $\sum_{j=1}^{d}\langle \boldsymbol{u}_j, \boldsymbol{v}_1\rangle^2 = \|U^{\mathsf{T}}\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_1\|^2 = 1$ for $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d]$. Hence,

$$\langle \boldsymbol{u}_1, X\boldsymbol{u}_1\rangle - \langle \boldsymbol{v}_1, X\boldsymbol{v}_1\rangle \ge \lambda_1 - \lambda_1\cos^2(\theta) - \lambda_2\sin^2(\theta) = (\lambda_1 - \lambda_2)\sin^2(\theta). \tag{9.119}$$

We conclude that

$$(\lambda_1 - \lambda_2)\sin^2(\theta) \le 2\,\|X - Y\|\sin(\theta).$$

The analogous analysis with $\langle \boldsymbol{v}_1, Y\boldsymbol{v}_1\rangle - \langle \boldsymbol{u}_1, Y\boldsymbol{u}_1\rangle$ as a starting point yields $(\mu_1 - \mu_2)$ in the denominator. We can take the best of these two bounds by introducing the maximum. □

The theorem says that the top eigenvectors of X and Y are close (i.e. the sine of the angle is small) if the gap between the top two eigenvalues of either X or Y is large compared to the spectral norm of the difference of these two matrices. Recall that the lack of this favorable comparison was exactly the reason for the instability in (9.117).

A few remarks:

- The statement of the theorem presented here is in terms of the gap between the eigenvalues of either X or Y. This form (see [36]) will be useful in statistical applications, as we often have control on the gaps of the signal matrix X. Other versions in the literature state the upper bound in terms of gaps between eigenvalues of X and the corresponding eigenvalues of Y. In this case, one can use Weyl's inequality to pass to the gap on the signal matrix X only.
- We stated the result for the top eigenvector. More general results can be found in the literature (e.g. [36]), for intermediate eigenvalues and eigenspaces.
- Wedin's theorems generalize Davis-Kahan to singular vectors rather than eigenvectors.

9.2 Principal Component Analysis

One of the most basic questions we may ask when analyzing high-dimensional data is whether there is a direction (or several directions) along which the data varies more than other directions. In many situations, this may be the case if data lives close to a low-dimensional subspace. We may determine this direction/subspace from the sample covariance matrix. When can this estimate reliably tell us about the direction of large variance in the population?

To motivate the Spiked Covariance Model below, consider two data generating scenarios. First is a model of data that lives close to a single direction $\boldsymbol{u}$. For $g \sim \mathcal{N}(0, I_d), Z \sim \mathcal{N}(0, \sigma^2 I_d)$, we have $X = \langle g, \boldsymbol{u}\rangle\boldsymbol{u} + Z$ and $\mathbb{E}XX^{\mathsf{T}} = \boldsymbol{u}\boldsymbol{u}^{\mathsf{T}} + \sigma^2 I$. The second model is a mixture model of two Gaussian populations with means at $\boldsymbol{u}$ and $-\boldsymbol{u}$: $X = \varepsilon \boldsymbol{u} + Z$, where $\varepsilon \in \{\pm 1\}$ is a Rademacher random variable. Here, again, $\mathbb{E}XX^\intercal = \boldsymbol{u}\boldsymbol{u}^\intercal + \sigma^2 I_d$. This covariance structure, present in both models, is the subject of the following investigation.

To this end, consider the following simple model. We assume that $X_1, \dots, X_n \in \mathbb{R}^d$ are centered i.i.d. random variables with $\mathbb{E}X_i X_i^\intercal = \Sigma$ and $X_i \in \mathsf{subG}(\|\Sigma\|)$. Assume that the population covariance matrix has the following structure, called the Spiked Covariance Model:

$$\Sigma = \lambda \boldsymbol{u}\boldsymbol{u}^\intercal + I_d \tag{9.120}$$

for some fixed $\boldsymbol{u} \in \mathbb{R}^d$, $\|\boldsymbol{u}\| = 1$. The parameter $\lambda \geq 1$ here determines the signal-to-noise ratio, the strength of the "spike." Clearly, the top eigenvector of Σ is $\boldsymbol{u}_1 = \boldsymbol{u}$, corresponding to the eigenvalue $1 + \lambda$. The question is whether this spike persists in the sample covariance matrix $\widehat{\Sigma}$. To this end, we view $Y = \widehat{\Sigma}$ as a randomly perturbed observation of a signal matrix $X = \Sigma$ with $\mathcal{E} = \widehat{\Sigma} - \Sigma$, as in (9.116). Let $\boldsymbol{v}_1$ be the leading eigenvector of $\widehat{\Sigma}$. Since $-\boldsymbol{v}_1$ is also an eigenvector of $\widehat{\Sigma}$, we can only determine closeness to $\boldsymbol{u}_1$ up to a sign. Observe that

$$\min_{\varepsilon \in \{\pm 1\}} \|\varepsilon \boldsymbol{v}_1 - \boldsymbol{u}_1\|^2 = 2 - 2|\langle \boldsymbol{u}_1, \boldsymbol{v}_1 \rangle| \leq 2 - 2\langle \boldsymbol{u}_1, \boldsymbol{v}_1 \rangle^2 = 2\sin^2(\angle(\boldsymbol{u}_1, \boldsymbol{v}_1)) \tag{9.121}$$

We also have that the gap $\lambda_1 - \lambda_2$ of the top two eigenvalues of Σ is λ, while $\|\Sigma\| = 1 + \lambda$. Together with results of the previous lecture,

$$\min_{\varepsilon \in \{\pm 1\}} \|\varepsilon \boldsymbol{v}_1 - \boldsymbol{u}_1\| \lesssim \frac{1+\lambda}{\lambda} \max\left\{ \sqrt{\frac{d}{n}} + \sqrt{\frac{t}{n}}, \frac{d}{n} + \frac{t}{n} \right\} \tag{9.122}$$

with probability at least $1 - 2\exp\{-t\}$. When d is large compared to n, we may employ the corresponding results with low effective rank, or sparsity.

9.3 Spectral Clustering and Stochastic Block Model

Suppose we have d vertices, subdivided into two equal-sized groups. For concreteness, the first $d/2$ vertices belong to the first cluster. A random graph is constructed as follows: independently, each pair within the community has an edge with probability p, and any pair across the two communities has an edge with probability $q < p$. The resulting distribution of the random graph is denoted by $G(n, p, q)$. One of the questions we may ask is: If we observe the random graph, but not identity of the vertices, can we recover the communities?

Note that the adjacency matrix A of the random graph is a random matrix with entries 0 and 1. We can also calculate its expected value as

$$\mathbb{E}A = \begin{bmatrix} pJ_{d/2} & qJ_{d/2} \\ qJ_{d/2} & pJ_{d/2} \end{bmatrix},$$

a matrix made up of four blocks, where $J_{d/2}$ is a $d/2 \times d/2$ matrix of all 1's.

We now view the observation $Y = A$ as a noisy value of the signal matrix $X = \mathbb{E}A$, as in (9.116), with $\mathcal{E} = A - \mathbb{E}A$. Let us examine the eigenstructure of $\mathbb{E}A$. The first (normalized) eigenvector is $\boldsymbol{u}_1 = \frac{1}{\sqrt{d}}\mathbf{1}$ and the corresponding eigenvalue is $\lambda_1 = d(p+q)/2$. This vector is not informative. The second eigenvector is $\boldsymbol{u}_2 = \frac{1}{\sqrt{d}}[1, \dots, 1, -1, \dots, -1]^\intercal$, with the corresponding eigenvalue $\lambda_2 = d(p-q)/2$ (the rest of the eigenvalues are 0). Interestingly, this second eigenvector contains community memberships. Note that not knowing identity of the vertices means that the rows/columns of A are renamed, or permuted. This only permutes the corresponding coordinates of the eigenvectors. Hence, we have the hope that the second eigenvector $\boldsymbol{v}_2$ of A also contains the necessary information about the community memberships. In what follows, we will show that clustering vertices into two communities according to the sign of $\boldsymbol{v}_2(i)$ for each vertex i leads to only a constant number of errors. This algorithm is known as *spectral clustering.*

Since we are aiming to recover the second rather than first eigenvector, we need to appeal to a more general version of Davis-Kahan, which has $\min\{\lambda_1 - \lambda_2, \lambda_2 - \lambda_3\}$ instead of $\lambda_1 - \lambda_2$ in the denominator of (9.118):

$$\min_{\varepsilon \in \{\pm 1\}} \|\varepsilon \boldsymbol{v}_2 - \boldsymbol{u}_2\| \leq \sqrt{2}\sin(\angle(\boldsymbol{u}_2, \boldsymbol{v}_2)) \leq \frac{2\sqrt{2}\,\|A - \mathbb{E}A\|}{\min\{\lambda_1 - \lambda_2, \lambda_2 - \lambda_3\}} \tag{9.123}$$

Now recall that from (21), with high probability, $\|A - \mathbb{E}A\| \lesssim \sqrt{d}$ where we use the fact that each entry is sub-Gaussian (recall that we need to apply the lemma separately to upper and lower triangular parts of the matrix). On the other hand,

$$\min\{\lambda_1 - \lambda_2, \lambda_2 - \lambda_3\} = d\min\{q, (p-q)/2\}.$$

We now recall that $\sqrt{d}\boldsymbol{u}_2$ is a vector of ±1's, while $\sqrt{d}\boldsymbol{v}_2(i) \in [-1, 1]$ for any coordinate i. Then we have for any $\varepsilon \in \{\pm 1\}$,

$$\|\varepsilon \boldsymbol{v}_2 - \boldsymbol{u}_2\|^2 = \frac{1}{d}\sum_{i=1}^{d}(\varepsilon\sqrt{d}\boldsymbol{v}_2(i) - \sqrt{d}\boldsymbol{u}_2(i))^2 \geq \frac{1}{d}\sum_{i=1}^{d}\mathbf{1}\left\{\varepsilon\,\mathrm{sign}(\boldsymbol{v}_2(i)) \neq \mathrm{sign}(\boldsymbol{u}_2(i))\right\} \tag{9.124}$$

Thus, if we think of p, q as constants (that is, for d large enough), we have that the right-hand side of (9.123) is of order $1/\sqrt{d}$, and by squaring both sides we get that for a constant $C_{p,q}$,

$$\sum_{i=1}^{d}\mathbf{1}\left\{\varepsilon\,\mathrm{sign}(\boldsymbol{v}_2(i)) \neq \mathrm{sign}(\boldsymbol{u}_2(i))\right\} \lesssim C_{p,q}, \tag{9.125}$$

i.e. only a constant number (out of d) vertices are misclassified by the *spectral clustering algorithm* which separates the nodes into two clusters according to the sign of the second eigenvector of A.

10. UNIFORM LAWS OF LARGE NUMBERS: MOTIVATION

By now you have seen a number of finite-sample guarantees: estimation of a mean vector, matrix estimation, constrained and unconstrained linear regression. In all the examples, the key technical step was a control of the maximum of some collection of random variables. Over the next few lectures, we will extend the toolkit to arbitrary classes of functions and then apply it to questions of parametric and nonparametric estimation and statistical learning.

10.1 Kolmogorov's Goodness-of-Fit test

Given n independent draws of a real-valued random variable X, you may want to ask whether it has a hypothesized distribution with cdf F_0. For instance, can you test the hypothesis that heights of people are $N(63, 3^2)$ (in inches)? Of course, we can try to see if the sample mean is "close" to the mean of the hypothesized distribution. We can also try the median, or some quantiles. In fact, we can try to compare all the quantiles at once and see if they match the quantiles of F_0. It turns out that comparing "all quantiles" is again a question about control of a maximum of a collection of correlated random variables. We will make this connection precise.

If you have taken a course on statistics, you might have seen several approaches to the hypothesis testing problem of whether X has a given distribution. One classical approach is the Kolmogorov-Smirnov test. Let

$$F(\theta) = P(X \le \theta)$$

be the cdf of X, and let

$$F_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le \theta\}$$

be the empirical cdf obtained from n examples. While for a single θ, the random variable $|F(\theta) - F_n(\theta)|$ converges to zero almost surely by the Laws of Large Numbers, the analogous convergence of

$$D_n = \sup_\theta |F(\theta) - F_n(\theta)|$$

to zero (that is, convergence *uniform* in θ) is less clear since we have a maximum of an uncountable collection of correlated random variables.

Nevertheless, the Glivenko-Cantelli Theorem (1933) states that

$$D_n \to 0 \quad a.s.$$

Hence, given a candidate F, one can test whether X has distribution with cdf F, but for this we need to know the (asymptotic) distribution of D_n. Assuming continuity of F, Kolmogorov (1933) showed that the distribution of D_n does not depend on the law of X, and he calculated the asymptotic distribution (now known as the Kolmogorov distribution). Without going into details, we can observe that $F(X)$ has cdf of a uniform random variable supported on $[0, 1]$, and this transformation does not change the supremum. Hence, it is enough to calculate D_n for the uniform distribution on $[0, 1]$. D_n fluctuates on the order of $1/\sqrt{n}$ and

$$\sqrt{n}D_n \longrightarrow \sup_{\theta\in\mathbb{R}} |B(F(\theta))|.$$

Here $B(x)$ is a Brownian bridge on $[0, 1]$ (a continuous-time stochastic process with distribution being Wiener process conditioned on being pinned to 0 at the endpoints).

In particular, Kolmogorov in his 1933 paper calculates the asymptotic distribution, as well a table of a few values. For instance, he states that

$$P(D_n \le 2.4/\sqrt{n}) \longrightarrow \quad \text{approx} \quad 0.999973.$$

In the spirit of this course, we will take a non-asymptotic approach to this problem. While we might not obtain such sharp constants, the deviation inequalities will be valid for finite n.

We will now come to the same question of uniform deviations from a different angle – Statistical Learning Theory.

10.2 Statistical Learning and Empirical Risk Minimization

Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be n i.i.d. copies of a random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with distribution $P = P_X \times P_{Y|X}$, where the X variable lives in some abstract space $\mathcal{X}$ and $\mathcal{Y} \subseteq \mathbb{R}$. Fix a loss function $\boldsymbol{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a class of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$. We may think of $\mathcal{F}$ as a set of neural networks, or decision trees, or whatever model you may have. Given the dataset S, the empirical risk minimization (ERM) method is defined as

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}(f(X_i), Y_i)$$

Examples:

- Linear regression: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, $\mathcal{F} = \{x \mapsto \langle w, x \rangle : w \in \mathbb{R}^d\}$, $\boldsymbol{\ell}(a, b) = (a - b)^2$
- Linear classification: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{F} = \{x \mapsto (\operatorname{sign}(\langle w, x \rangle) + 1)/2 : w \in \mathsf{B}_2\}$, $\boldsymbol{\ell}(a, b) = \mathbf{1}\{a \neq b\}$

We now define expected loss (error) as

$$\mathbf{L}(f) = \mathbb{E}_{(X,Y)} \boldsymbol{\ell}(f(X), Y) \tag{10.126}$$

and empirical loss (error) as

$$\widehat{\mathbf{L}}(f) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}(f(X_i), Y_i). \tag{10.127}$$

A central question in Statistical Learning is: what is an upper bound on the expected error of ERM?

Lemma 23. *For any class $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, the ERM $\widehat{f}$ satisfies*

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \inf_{f \in \mathcal{F}} \mathbf{L}(f) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right], \tag{10.128}$$

where the expectations are with respect to S.

Proof. Suppose without loss of generality that $f^* = \inf_{f \in \mathcal{F}} \mathbf{L}(f)$. The decomposition holds:

$$\mathbf{L}(\widehat{f}) - \mathbf{L}(f^*) = \left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] + \left[\widehat{\mathbf{L}}(\widehat{f}) - \widehat{\mathbf{L}}(f^*)\right] + \left[\widehat{\mathbf{L}}(f^*) - \mathbf{L}(f^*)\right].$$

By definition of ERM, the second term is nonpositive. Since f^* is independent of the random sample, the third term is a difference between an average of random variables $\boldsymbol{\ell}(f^*(X_i), Y_i)$ and their expectation. Hence, this term is zero-mean, and its fluctuations can be controlled with the tail bounds we have seen in class. The first term, however, is generally not zero in expectation, i.e. $\mathbb{E}_S \widehat{\mathbf{L}}(\widehat{f}) \neq \mathbb{E}_S \mathbf{L}(\widehat{f})$ (why?). Let us proceed by taking expectation (with respect to S) of both sides:

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \mathbf{L}(f^*) \leq \mathbb{E}\left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] \leq \mathbb{E} \sup_{f \in \mathcal{F}} \left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right]. \tag{10.129}$$

□

Here we "removed the hat" on $\widehat{f}$ by "supping out" this data-dependent choice. We are only using the knowledge that $f \in \mathcal{F}$, and nothing else about the method. We will see later that for "curved" loss functions, such as square loss, the supremum can be further localized within $\mathcal{F}$. Note that (10.128) can lead to a vacuous (e.g. infinite) upper bound: one such example is linear unconstrained regression.

10.3 Example: Classification with thesholds.

We now specialize to the classification scenario with indicator loss $\boldsymbol{\ell}(a,b) = \mathbf{1}\{a \neq b\}$. Observe that $\mathbf{1}\{a \neq b\} = a + (1-2a)b$ for $a, b \in \{0,1\}$. Hence, by taking $a = Y$ and $b = f(X)$,

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] &= \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}(Y + (1-2Y)f(X)) - \frac{1}{n}\sum_{i=1}^{n}(Y_i + (1-2Y_i)f(X_i))\right] \\ &= \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}((1-2Y)f(X)) - \frac{1}{n}\sum_{i=1}^{n}(1-2Y_i)f(X_i)\right]\end{aligned}$$

Observe that $(1-2Y)$ is a random sign that is jointly distributed with X. Let us omit this random sign for a moment, and consider

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n}f(X_i)\right]. \tag{10.130}$$

Over the next few lectures, we will develop upper bounds on the above expected supremum for any class $\mathcal{F}$. For now, let us gain a bit more intuition about this object by looking at a particular class of 1D thresholds:

$$\mathcal{F} = \{x \mapsto \mathbf{1}\{x \leq \theta\} : \theta \in \mathbb{R}\}.$$

Substituting this choice, (10.130) becomes

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[P(X \leq \theta) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq \theta\}\right] = \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[F(\theta) - F_n(\theta)\right]. \tag{10.131}$$

which is precisely the quantity from the beginning of the lecture (albeit without absolute values and in expectation). Again, (10.131) is the expected largest pointwise (and one-sided) distance between the CDF and empirical CDF. Does it go to zero as $n \to \infty$? How fast?

Let's introduce the shorthand

$$U_\theta = \mathbb{E}\mathbf{1}\{X \leq \theta\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq \theta\}.$$

$\{U_\theta\}_{\theta\in\mathbb{R}}$ is an uncountable collection of *correlated* random variables, so how does the maximum behave? We have already encountered the question in the context of linear forms $\langle X, \theta\rangle$, indexed by $\theta \in \mathsf{B}_2$ and we were able to use a covering argument to control the expected supremum. Recall the key step in that proof: we can introduce a cover $\theta_1, \ldots, \theta_N$ such that control of $\sup U_\theta$ can be reduced to control of $\max_{j=1,\ldots,N} U_{\theta_i}$. Does this idea work here? Problems with this approach start appearing immediately: how do we cover $\mathbb{R}$ by a finite collection?

In the next two sections, we present two approaches for upper-bounding (10.131); both extend to the general case of (10.130).

10.4 Approach 1: Bracketing

While we cannot provide a finite ϵ-grid of $\mathbb{R}$ directly, we observe that we should be placing the covering elements according to the underlying measure P. Informally, U_θ is likely to be constant over regions of θ with small mass.

For simplicity assume that P does not have atoms, and let $\theta_1, \theta_1, \ldots, \theta_N$ (with $\theta_0 = -\infty, \theta_{N+1} = +\infty$) correspond to the quantiles: $P(\theta_i \le X \le \theta_{i+1}) = \frac{1}{N+1}$. For a given θ, let $u(\theta)$ and $\ell(\theta)$ denote, respectively, the upper and lower elements corresponding to the discrete collection $\theta_0, \ldots, \theta_{N+1}$. Then, trivially,

$$\begin{aligned}\mathbb{E}\mathbf{1}\{X \le \theta\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le \theta\} &\le \mathbb{E}\mathbf{1}\{X \le u(\theta)\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le \ell(\theta)\} \\ &\le \mathbb{E}\mathbf{1}\{X \le \ell(\theta)\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le \ell(\theta)\} + \frac{1}{N+1}\end{aligned}$$

and thus

$$\begin{aligned}&\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\{X \le \theta\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le \theta\}\right] \\ &\le \frac{1}{N+1} + \mathbb{E}\max_{j\in\{0,\ldots,N\}} \mathbb{E}\mathbf{1}\{X \le \theta_j\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le \theta_j\}\end{aligned}$$

Now, each random variable $\mathbb{E}\mathbf{1}\{X \le \theta\} - \mathbf{1}\{X_i \le \theta\}$ is centered and 1-subGaussian. Hence, for each j, U_{θ_j} is $\frac{1}{\sqrt{n}}$-subGaussian, and the expected maximum is at most $\sqrt{\frac{2\log(N+1)}{n}}$. The overall upper bound is then

$$\frac{1}{N+1} + \sqrt{\frac{\log(N+1)}{n}} = O\left(\sqrt{\frac{\log n}{n}}\right)$$

if we choose, for instance, $N+1 = n$.

Before presenting an alternative to this approach, we state a general lemma.

10.5 The Symmetrization Lemma

An alternative is a powerful technique that replaces the expected value by a ghost sample. To motivate the technique, recall the following inequality for variance:

$$\mathbb{E}(X - \mathbb{E}X)^2 \le \mathbb{E}(X - X')^2 = 2\mathbb{E}(X - \mathbb{E}X)^2$$

where X' is an independent copy of X.

Lemma 24. *Let $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ be a class of real-valued functions. Let $X, X_1, \ldots, X_n$ be i.i.d. random variables with values in $\mathcal{X}$, and let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] \le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right].$$

We also have

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)-\mathbb{E}f(X)\right]\le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]$$

and

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i)-\mathbb{E}f(X)\right|\le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|.$$

Furthermore, the opposite direction holds:

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\mathbb{E}f(X)-\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right|+\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}|\mathbb{E}f|$$

Proof. For the first statement, we introduce an i.i.d. sample $X_1',\ldots,X_n'$ with the same distribution as X. Observe that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i')\right]=\mathbb{E}f(X)$. Hence,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}f(X)-\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]=\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i')\right]-\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]. \tag{10.132}$$

By Jensen's inequality, the last expression is at most

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i')-f(X_i)\right] \tag{10.133}$$

where the expectation is over $X_{1:n}, X_{1:n}'$. Now, since distribution of $f(X_i')-f(X_i)$ is the same as the distribution of $-\left(f(X_i')-f(X_i)\right)$, we can insert arbitrary signs ϵ_i without changing the expected value:

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i')-f(X_i))\right]. \tag{10.134}$$

Since the quantity is constant for all the choices of $\epsilon_1,\ldots,\epsilon_n$, we have the same value by taking an expectation. We have

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}f(X)-\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]\le\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i')-f(X_i))\right], \tag{10.135}$$

where ϵ_i's are now Rademacher random variables. Breaking up the supremum into two terms leads to an upper bound

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i')\right]+\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}-\epsilon_i f(X_i)\right] \tag{10.136}$$

$$=2\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right] \tag{10.137}$$

by the symmetry of Rademacher random variables. The second and third statement follow from the same argument. For the last part,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\le\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i)-\mathbb{E}f)\right|+\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbb{E}f\right|.$$

Consider the first term on the RHS:

$$
\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i)-\mathbb{E}f)\right| &\leq \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i)-f(X_i'))\right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f+\mathbb{E}f-f(X_i'))\right| \\
&\leq \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}f-f(X_i))\right| + \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f)\right|.
\end{aligned}
$$

As for the second term,

$$
\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbb{E}f\right| \leq \sup_{f\in\mathcal{F}}|\mathbb{E}f|\cdot\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \tag{10.138}
$$

□

10.6 Approach 2: Symmetrization

We now illustrate the power of the symmetrization lemma for the case of thresholds. Recall, that our goal is to upper bound

$$
\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\left\{X\leq\theta\right\}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i\leq\theta\right\}\right].
$$

From Lemma 24, this expected supremum is upper bounded by

$$
2\mathbb{E}\sup_{\theta\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i\leq\theta\right\}.
$$

Let us condition on $X_1,\ldots,X_n$ and think of the random variables

$$
V_\theta=\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i\leq\theta\right\}
$$

as a function of the Rademacher random variables. How many truly distinct V_θ's do we have? Since $X_1,\ldots,X_n$ are now fixed, there are only at most $n+1$ choices (say, midpoints between datapoints), and so the last expression is

$$
2\mathbb{E}\left[\mathbb{E}\left[\sup_{\theta\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i\leq\theta\right\}\middle|X_{1:n}\right]\right]=2\mathbb{E}\mathbb{E}\left[\max_{\theta\in\{\theta_1,\ldots,\theta_{n+1}\}}V_\theta\middle|X_{1:n}\right] \tag{10.139}
$$

Since each V_θ is $1/\sqrt{n}$-subGaussian, and we get an overall upper bound of

$$
2\sqrt{\frac{2\log(n+1)}{n}}
$$

which, up to constants, matches the bound with the bracketing approach.

10.7 Discussion

The bracketing and symmetrization approaches produced similar upper bounds for the case of thresholds. We will see, however, that for more complex classes of functions, the two approaches can give different results.

Of course, the symmetrization lemma can also be applied to the class of functions

$$\{(x,y) \mapsto (1-2y)f(x)\}.$$

Since $(1-2y)$ is $\{\pm 1\}$-valued, the distribution of $(1-2Y_i)\epsilon_i$ is also Rademacher. Hence,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(1-2Y_i)f(X_i)\right] = \mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]. \tag{10.140}$$

This justifies omitting $(1-2Y)$ for binary classification in our earlier exposition. Hence, in view of (10.129), the upper bounds we derived guarantee that for empirical risk minimization,

$$\mathbb{E}\mathbf{L}(\widehat{f}) - \min_{f^*\in\mathcal{F}}\mathbf{L}(f^*) \lesssim \sqrt{\frac{\log(n+1)}{n}}$$

The power of symmetrization for studying the suprema of empirical processes has been described in [11], who, in turn, attribute the technique to [15].

10.8 Empirical Processes

Let us also define an empirical process:

Definition 7. *Let* $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ *and* $X, X_1, \ldots, X_n$ *are i.i.d. The stochastic process*

$$\nu_f = \mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

is called the empirical process indexed by $\mathcal{F}$.

We note that it is also customary to scale the empirical process as

$$\nu_f = \sqrt{n}\left(\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right)$$

Second, empirical process theory often employs the notation

$$\nu_f = \sqrt{n}(\mathbb{P} - \mathbb{P}_n)f$$

where $\mathbb{P}$ is the distribution of X and $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ is the empirical measure. You may also see the notation

$$\sup_{f\in\mathcal{F}}|\nu_f| = \|\mathbb{P} - \mathbb{P}_n\|_{\mathcal{F}}.$$

We can view supremum of the empirical process as the difference between the true and empirical distributions when viewed through the lens of $\mathcal{F}$.

Definition 8. *A class $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ is (weak) Glivenko-Cantelli with respect to P if*

$$\sup_{f\in\mathcal{F}} \left| \mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i) \right| \to 0 \tag{10.141}$$

in probability (and strong Glivenko-Cantelli for almost sure convergence; these are equivalent under certain boundedness assumptions).

Note: there do exist classes that are not Glivenko-Cantelli. These classes are, in a certain sense, very rich, and both learning and uniform GC property fail. One trivial example is

$$\mathcal{F} = \{\mathbf{1}\{S\} : |S| = m, m \geq 1\},$$

indicators of discrete sets of arbitrary size, and P is absolutely continuous with respect to Lebesgue. Another example is a class is bounded continuous functions on $[0, 1]$ with respect to, say, Lebesgue measure.

11. SUPREMA OF GAUSSIAN AND SUBGAUSSIAN PROCESSES

Definition 9. *Stochastic process $(U_\theta)_{\theta\in\Theta}$, indexed by $\theta \in \Theta$, is a collection of random variables on a common probability space.*

The index θ can be "time," but we will be primarily interested in cases where Θ has some metric structure.

We will be interested in the behavior of the supremum of the stochastic process, and in particular its expected value:

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta.$$

To understand this object, we need to have a sense of the dependence structure of U_θ and $U_{\theta'}$ for a pair of parameters, but also about the metric structure of Θ.

Gaussian process is a collection of random variables such that any finite collection $U_{\theta_1}, \ldots, U_{\theta_n}$, for any $n \geq 1$, is zero-mean and jointly Gaussian. In this case

$$\mathbb{E}\exp\{\lambda(U_\theta - U_{\theta'})\} = \exp\{\lambda^2 d(\theta, \theta')^2/2\}$$

with $d(\theta, \theta')^2 = \mathbb{E}(U_\theta - U'_\theta)^2$. Hence, there is a natural metric for Gaussian process.

11.1 SubGaussian Processes

Definition 10. *Stochastic process $(U_\theta)_{\theta\in\Theta}$ is sub-Gaussian with respect to a metric d on Θ if U_θ is zero-mean and*

$$\forall \theta, \theta' \in \Theta, \lambda \in \mathbb{R}, \qquad \mathbb{E}\exp\{\lambda(U_\theta - U_{\theta'})\} \leq \exp\{\lambda^2 d(\theta, \theta')^2/2\}$$

The main examples we will be studying have a particular linearly parametrized form:

Gaussian process: Let $G_\theta = \langle g, \theta \rangle$, $g = (g_1, \ldots, g_n)$, $g_i \sim N(0,1)$ i.i.d. Take $d(\theta, \theta') = \|\theta - \theta'\|$. Then

$$G_\theta - G'_\theta = \langle g, \theta - \theta' \rangle \sim N(0, \left\|\theta - \theta'\right\|^2)$$

In particular, this Gaussian process is also, trivially, sub-Gaussian with respect to the Euclidean distance on Θ.

Rademacher process: Let $R_\theta = \langle \boldsymbol{\varepsilon}, \theta \rangle$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$, ε_i i.i.d. Rademacher. Again, take $d(\theta, \theta') = \|\theta - \theta'\|$. Then

$$R_\theta - R'_\theta = \langle \boldsymbol{\varepsilon}, \theta - \theta' \rangle$$

is subGaussian with parameter $\|\theta - \theta'\|^2$.

Note that in this linear parametrization of U_θ, the expected supremum can be seen as a kind of average 'width' of the set Θ.

Definition 11. *We will call*

$$\widehat{\mathcal{R}}(\Theta) = \mathbb{E} \sup_{\theta \in \Theta} \langle \boldsymbol{\varepsilon}, \theta \rangle$$

the (empirical) Rademacher averages of Θ. The corresponding expected supremum of the Gaussian process will be called the Gaussian averages or the Gaussian width of Θ and denoted by $\widehat{\mathcal{G}}(\Theta)$.

11.1.1 A few examples

Let $U_\theta = \langle \boldsymbol{\varepsilon}, \theta \rangle$, $\Theta \subset \mathbb{R}^n$, and take Euclidean distance as the metric. Let B_p^n denote the unit ℓ_p ball in $\mathbb{R}^n$. We have

$$\widehat{\mathcal{R}}(\mathsf{B}_\infty^n) = \mathbb{E} \sup_{\theta \in \mathsf{B}_\infty^n} U_\theta = \mathbb{E} \sup_{\theta \in \mathsf{B}_\infty^n} \langle \boldsymbol{\varepsilon}, \theta \rangle = n.$$

To get a sublinear growth in n, we have to make sure Θ is significantly smaller than B_∞^n.

A few other sets:

$$\widehat{\mathcal{R}}(\mathsf{B}_2^n) = \mathbb{E} \sup_{\theta \in \mathsf{B}_2^n} \langle \boldsymbol{\varepsilon}, \theta \rangle = \mathbb{E} \left\|\boldsymbol{\varepsilon}\right\|_2 = \sqrt{n}$$

and

$$\widehat{\mathcal{G}}(\mathsf{B}_2^n) \leq \sqrt{n}.$$

However, we observe that

$$\widehat{\mathcal{R}}(\mathsf{B}_1^n) = \mathbb{E} \sup_{\theta \in \mathsf{B}_1^n} \langle \boldsymbol{\varepsilon}, \theta \rangle = \mathbb{E} \left\|\boldsymbol{\varepsilon}\right\|_\infty = 1$$

while for the Gaussian process,

$$\widehat{\mathcal{G}}(\mathsf{B}_1^n) = \mathbb{E} \sup_{\theta \in \mathsf{B}_1^n} \langle g, \theta \rangle = \mathbb{E} \max_{i \in [n]} |g_i| \leq \sqrt{2 \log(2n)}.$$

In fact, this discrepancy between the Rademacher and Gaussian averages for B_1^n is the worst that can happen and for any Θ

$$\widehat{\mathcal{R}}(\Theta) \lesssim \widehat{\mathcal{G}}(\Theta) \lesssim \sqrt{\log n} \cdot \widehat{\mathcal{R}}(\Theta). \tag{11.142}$$

Furthermore, the discrepancy is only there because B_1^n has a small ℓ_1 diameter, and for many of the applications in statistics, we will work with a function class that will not have such a small ℓ_1 diameter.

For a singleton,

$$\widehat{\mathcal{R}}(\{\theta\}) = 0$$

while for the vector $\mathbf{1}_n = (1, \ldots, 1)$,

$$\widehat{\mathcal{R}}(\{-\mathbf{1}_n, \mathbf{1}_n\}) = \mathbb{E} \max\{\langle \epsilon, \mathbf{1}_n \rangle, -\langle \epsilon, \mathbf{1}_n \rangle\} = \mathbb{E} \left| \sum_{i=1}^n \epsilon_i \right| \leq \sqrt{n}.$$

Some further properties of both Rademacher and Gaussian averages:

$$\widehat{\mathcal{R}}(\Theta) \lesssim \operatorname{diam}(\Theta) \sqrt{\log \operatorname{card}(\Theta)},$$
$$\widehat{\mathcal{R}}(\operatorname{conv}(\Theta)) = \widehat{\mathcal{R}}(\Theta),$$
$$\widehat{\mathcal{R}}(c\Theta) = |c| \widehat{\mathcal{R}}(\Theta) \quad \text{for constant} \;\; c$$

11.2 Finite-class lemma and a single-scale covering argument

Lemma 25. *Let d be a metric on Θ and assume (U_θ) is a subGaussian process. Then for any finite subset $A \subseteq \Theta \times \Theta$,*

$$\mathbb{E} \max_{(\theta,\theta') \in A} U_\theta - U_{\theta'} \leq \max_{(\theta,\theta') \in A} d(\theta, \theta') \cdot \sqrt{2 \log \operatorname{card}(A)} \tag{11.143}$$

How do we go beyond finite cover?

Definition 12. *Let (Θ, d) be a metric space. A set $\theta_1, \ldots, \theta_N \in \Theta$ is a (proper) cover of Θ at scale ϵ if for any θ there exists $j \in [N]$ such that $d(\theta, \theta_j) \leq \epsilon$. The covering number of Θ at scale ϵ is the size of the smallest cover, denoted by $\mathcal{N}(\Theta, d, \epsilon)$.*

As a simple consequence,

Lemma 26. *If $(U_\theta)_{\theta \in \Theta}$ is subGaussian with respect to d on Θ, then for any $\delta > 0$,*

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 2\mathbb{E} \sup_{d(\theta,\theta') \leq \delta} (U_\theta - U_{\theta'}) + 2\, diam(\Theta) \sqrt{\log \mathcal{N}(\Theta, d, \delta)}$$

Proof. Observe that

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta = \mathbb{E} \sup_{\theta \in \Theta} U_\theta - U_{\theta'} \leq \mathbb{E} \sup_{\theta, \theta' \in \Theta} U_\theta - U_{\theta'}$$

Let $\widehat{\Theta}$ be a δ-cover of Θ. Then for $\hat{\theta}, \hat{\theta}' \in \widehat{\Theta}$ with $d(\theta, \hat{\theta}), d(\theta', \hat{\theta}') \leq \delta$,

$$\begin{aligned} U_\theta - U_{\theta'} &= U_\theta - U_{\hat{\theta}} + U_{\hat{\theta}} - U_{\hat{\theta}'} + U_{\hat{\theta}'} - U_{\theta'} && (11.144) \\ &\leq 2 \sup_{d(\theta,\theta') \leq \delta} (U_\theta - U_{\theta'}) + \sup_{\hat{\theta}, \hat{\theta}' \in \widehat{\Theta}} (U_{\hat{\theta}} - U_{\hat{\theta}'}) && (11.145) \end{aligned}$$

The last term is

$$\mathbb{E} \sup_{\hat{\theta},\hat{\theta}'\in\widehat{\Theta}} U_{\hat{\theta}} - U_{\hat{\theta}'} \leq \mathrm{diam}(\Theta)\sqrt{2\log(\mathrm{card}(\widehat{\Theta})^2)}$$

□

11.3 Example: Rademacher/Gaussian processes

Let $U_\theta = \langle g, \theta\rangle$ or $\langle \boldsymbol{\varepsilon}, \theta\rangle$, $\Theta \subset \mathbb{R}^n$, and take Euclidean distance as the metric. Then

$$\mathbb{E} \sup_{d(\theta,\theta')\leq\delta} U_\theta - U_{\theta'} \leq \mathbb{E} \sup_{\|\theta\|\leq\delta} \langle g, \theta\rangle \leq \delta\mathbb{E}\,\|g\| \leq \delta\sqrt{n}$$

Hence,

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta \leq 2\delta\sqrt{n} + 2\mathrm{diam}(\Theta)\sqrt{\log\mathcal{N}(\Theta, \|\cdot\|_2, \delta)} \tag{11.146}$$

Roughly speaking, the supremum over Θ can be upper bounded by the supremum within a ball of radius δ ("local complexity") and the maximum over a finite collection of centers of δ-balls. We will see this decomposition/idea again within the context of optimal estimators with general (possibly nonparametric) classes of functions.

Is (11.146) a tight upper bound? To investigate this question, consider two examples. First is the example of $\Theta = \mathsf{B}_2^n$. In this case, (5.69) gives an upper bound of $\sqrt{n}$ and a multiplicative-cover approach of (7.104) recovers this up to constant factors (here n is the dimensionality rather than d). We see that the same guarantee can be achieved by (11.146) by taking δ a constant.

The next example, however, brings bad news: (11.146) is not necessarily tight. Consider

$$\Theta = \{(0,\ldots,0), (1,0,\ldots,0), (1,1,0,\ldots,0), \ldots, (1,\ldots,1)\}. \tag{11.147}$$

In this case, the expected supremum in (11.146) is $O(\sqrt{n})$, as we shall see soon, but this cannot be recovered from the upper bound. To establish a guarantee, we need to take δ to be constant, yet the diameter of Θ is $\sqrt{n}$ while the covering number of Θ at a constant scale must grow with n. We will soon see that (11.146) can lead to suboptimal rates.

Notice that we have seen the set Θ in (11.147) earlier: it corresponds to the $n+1$ effective "signatures" of threshold functions when $x_1, \ldots, x_n$ are fixed (see (10.139)).

11.4 Chaining

Theorem 2. *Let $(U_\theta)_{\theta\in\Theta}$ be a sub-Gaussian stochastic process with respect to a metric d. Let $D = diam(\Theta)$. Then for any $\delta \in [0, D]$,*

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta \leq 2\mathbb{E} \sup_{d(\theta,\theta')\leq\delta} (U_\theta - U'_\theta) + 8\sqrt{2}\int_{\delta/4}^{D/2} \sqrt{\log\mathcal{N}(\Theta, d, \varepsilon)}d\varepsilon \tag{11.148}$$

Proof. Let Θ_j be a cover of Θ at scale $2^{-j}D$. We have $\mathrm{card}(\Theta_0) = 1$. Let

$$N = \min\left\{j : 2^{-j}D \leq \delta\right\}$$

(which means $2^{-N}D \le \delta \le 2^{-(N-1)}D$) and $\mathrm{card}(\Theta_N) = \mathcal{N}(\Theta, d, 2^{-N}D) \ge \mathcal{N}(\Theta, d, \delta)$. As before, we start with a single (finest-scale) cover:

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta \le 2\mathbb{E}\sup_{d(\theta,\theta')\le\delta}(U_\theta - U_{\theta'}) + \mathbb{E}\sup_{\theta_N,\theta'_N\in\Theta_N}(U_{\theta_N} - U_{\theta'_N}).$$

For $\theta_N \in \Theta_N$,

$$U_{\theta_N} = \sum_{i=1}^{N} U_{\theta_i} - U_{\pi_{i-1}(\theta_i)} + U_{\theta_0} \tag{11.149}$$

where, recursively, we define $\theta_{i-1} = \pi_{i-1}(\theta_i)$ to be the element of Θ_{i-1} closest to θ_i. The sequence $\theta_0, \theta_1, \ldots, \theta_N$ is a "chain" linking an element of the covering to the corresponding closest element at the coarser scale.

Let the corresponding chain for $\theta'_N \in \Theta_N$ be denoted by $\theta'_0, \theta'_1, \ldots, \theta'_N$. Then

$$U_{\theta_N} - U_{\theta'_N} = \left(\sum_{i=1}^{N} U_{\theta_i} - U_{\pi_{i-1}(\theta_i)}\right) - \left(\sum_{i=1}^{N} U_{\theta'_i} - U_{\pi_{i-1}(\theta'_i)}\right)$$

and

$$\begin{aligned}
\mathbb{E}\max_{\theta,\theta'\in\Theta_N} U_\theta - U_{\theta'} &\le \sum_{i=1}^{N}\mathbb{E}\max_{\theta_i\in\Theta_i}(U_{\theta_i} - U_{\pi_{i-1}(\theta_i)}) + \sum_{i=1}^{N}\mathbb{E}\max_{\theta'_i\in\Theta_i}(U_{\pi_{i-1}(\theta'_i)} - U_{\theta'_i}) && (11.150)\\
&\le 2\sum_{i=1}^{N} D2^{-(i-1)}\sqrt{2\log\mathcal{N}(\Theta, d, 2^{-i}D)} && (11.151)\\
&= 8\sum_{i=1}^{N} D2^{-(i+1)}\sqrt{2\log\mathcal{N}(\Theta, d, 2^{-i}D)} && (11.152)\\
&\le 8\sum_{i=1}^{N}\int_{2^{-(i+1)}D}^{2^{-i}D}\sqrt{2\log\mathcal{N}(\Theta, d, \varepsilon)}d\varepsilon && (11.153)
\end{aligned}$$

Observe that $2^{-(N+1)}D \ge \delta/4$, which concludes the proof. □

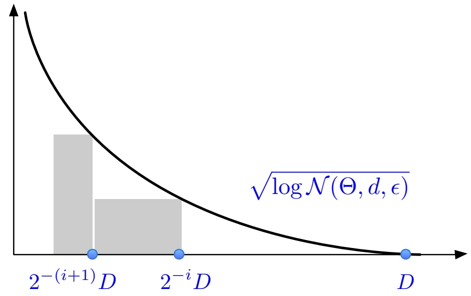


Figure 1: Illustration of the Dudley integral upper bound

Sudakov's theorem gives a single-scale lower bound:

Theorem 3. *For a Gaussian process $(U_\theta)_{\theta\in\Theta}$,*

$$C \sup_{\alpha\geq 0} \alpha\sqrt{\log \mathcal{N}(\Theta, d, \alpha)} \leq \mathbb{E}\sup_{\theta\in\Theta} U_\theta$$

for some constant C.

We can interpret this lower bound as the largest rectangle under the curve in Figure 1. This lower bound can be tight in the applications we consider (whenever the sum of the areas of rectangles Figure 1 is of the same order as the largest one).

12. RADEMACHER/GAUSSIAN AVERAGES FOR FUNCTION CLASSES

We have developed general machinery for upper- and lower-bounding the expected suprema of sub-Gaussian processes, including Rademacher and Gaussian processes linearly parametrized by a $\Theta \subset \mathbb{R}^n$. How are these results relevant to the problem of learning or estimation with a class of functions $\mathcal{F}$?

The symmetrization lemma (Lemma 24) tells us that for a class of real-valued functions $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ we can upper bound the expected supremum of the empirical process indexed by $\mathcal{F}$ in terms of the expected supremum of the Rademacher processes:

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] \leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right].$$

The key is that we can now condition on $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, and

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right]$$

precisely corresponds to Rademacher averages of the following indexing set Θ. To see this correspondence, let

$$\Theta = \frac{1}{\sqrt{n}}\mathcal{F}|_{x_1,\ldots,x_n} = \left\{\frac{1}{\sqrt{n}}(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\right\} \subseteq \mathbb{R}^n \tag{12.154}$$

a (scaled by $1/\sqrt{n}$) *projection* (or, restriction) of $\mathcal{F}$ onto $x_1, \ldots, x_n$. Take d to be

$$d(\theta, \theta')^2 = \left\|\theta - \theta'\right\|^2 = \frac{1}{n}\sum_{i=1}^n (f(x_i) - f'(x_i))^2 \triangleq \left\|f - f'\right\|_n^2 \tag{12.155}$$

where $\theta = (f(x_1), \ldots, f(x_n))$ and $\theta' = (f'(x_1), \ldots, f'(x_n))$, $f, f' \in \mathcal{F}$. Note that $\|\cdot\|_n$ is a pseudo-metric, as it can be zero for functions that differ outside the given data. With these definitions, we write

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\sup_{f\in\mathcal{F}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(X_i)\right] = \mathbb{E}_{\boldsymbol{\varepsilon}}\sup_{\theta\in\Theta}\langle\boldsymbol{\varepsilon}, \theta\rangle.$$

Furthermore,

$$\mathcal{N}(\Theta, \|\cdot\|_2, \alpha) = \mathcal{N}(\mathcal{F}, \|\cdot\|_n, \alpha).$$

Then Theorem 2 tell us that for any $\delta \in [0, D]$,

$$\mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(x_i) \leq 2\delta\sqrt{n} + 8\sqrt{2} \int_{\delta/4}^{D/2} \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_n, \alpha)} d\alpha$$

Let us formalize this as the following corollary.

Corollary 2. *Let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$. For any $x_1, \ldots, x_n$,*

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \leq \inf_{\delta\geq 0} \left\{ 8\delta + \frac{12}{\sqrt{n}} \int_{\delta}^{D/2} \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_n, \alpha)} d\alpha \right\}$$

where $D = \sup_{f,g\in\mathcal{F}} \|f - g\|_n \leq 2 \sup_{f\in\mathcal{F}} \|f\|_n \leq 2 \sup_{f\in\mathcal{F}} \|f\|_\infty$ and $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$.

Putting together the symmetrization lemma and above Corollary, we have

Corollary 3. *Let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ be a class of functions and let $X_1, \ldots, X_n \sim P$ be independent. Then*

$$\mathbb{E} \sup_{f\in\mathcal{F}} \left\{ \mathbb{E}f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\} \leq \mathbb{E} \inf_{\delta\geq 0} \left\{ 16\delta + \frac{24}{\sqrt{n}} \int_{\delta}^{D} \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_n, \alpha)} d\alpha \right\} \tag{12.156}$$

where $D = \sup_{f\in\mathcal{F}} \|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f(X_i)^2}$.

Expectations on both sides are with respect to $X_1, \ldots, X_n$. Note that the above results hold for the absolute value of the empirical process if we replace $\log \mathcal{N}$ by $\log 2\mathcal{N}$, and the $\log 2$ can be further absorbed into the multiplicative constant.

The Sudakov lower bound for the Gaussian process implies (together with the relationship between Rademacher and Gaussian processes) the following lower bound for the Rademacher averages:

Corollary 4. *For any $X_1, \ldots, X_n$,*

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \geq \frac{c}{\sqrt{\log n}} \cdot \sup_{\alpha\geq 0} \alpha \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_n, \alpha)}{n}}$$

for some absolute constant c.

We note that a version of the lower bound (for a particular choice of α) without the logarithmic factor is available, under some conditions, and it often matches the upper bound (see a few pages below).

Definition 13. *Given $x_1, \ldots, x_n$ and a class $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$,*

$$\widehat{\mathcal{R}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(x_i) \tag{12.157}$$

are called the (empirical) Rademacher averages of $\mathcal{F}$.

Note that we will occasionally adopt the $1/n$ scaling to follow the literature.

13. COVERING AND PACKING

Given a probability measure P on $\mathcal{X}$, we define

$$\|f\|^2_{L^2(P)} = \mathbb{E} f(X)^2 = \int f(x)^2 P(dx).$$

Similarly, for a given $X_1, \ldots, X_n$ we define a random pseudometric

$$\|f\|^2_{L^2(P_n)} = \frac{1}{n} \sum_{i=1}^{n} f(X_i)^2 = \|f\|^2_n \, .$$

Of course, the second definition is just a special case of the first for empirical measure $\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$.

Definition 14. *An α-net (or, α-cover) of $\mathcal{F}$ with respect to $L^2(P)$ is a set of functions $f_1, \ldots, f_N$ such that*

$$\forall f \in \mathcal{F}, \quad \exists j \in [N] \quad s.t. \quad \|f - f_j\|_{L^2(P)} \le \alpha.$$

The size of the smallest α-net is denoted by $\mathcal{N}(\mathcal{F}, L^2(P), \alpha)$.

The above definition can be also generalized to $L^r(P)$. Next, we spell out the above definition specifically for the empirical measure P_n:

Definition 15. *Let $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ be the empirical measure supported on $x_1, \ldots, x_n$. A set $V = \{v_1, \ldots, v_N\}$ of vectors in $\mathbb{R}^n$ forms an α-net (or, α-cover) of $\mathcal{F}$ with respect to $L^r(P_n)$ if*

$$\forall f \in \mathcal{F}, \quad \exists j \in [N] \quad s.t. \quad \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - v_j(i)|^r \le \alpha^r$$

The size of the smallest α-net is denoted by $\mathcal{N}(\mathcal{F}, L^r(P_n), \alpha)$. Similarly, an α-net (or, α-cover) with respect to $L^\infty(P_n)$ requires

$$\forall f \in \mathcal{F}, \quad \exists j \in [N] \quad s.t. \quad \max_{i \in [n]} |f(x_i) - v_j(i)| \le \alpha$$

The size of the smallest α-net is denoted by $\mathcal{N}(\mathcal{F}, L^\infty(P_n), \alpha)$.

Observe that the elements of the cover V can be "improper," i.e. they do not need to correspond to values of some function on the data. However, one can go between proper and improper covers at a cost of a constant (check!).

Second, observe that

$$\mathcal{N}(\mathcal{F}, L^r P_n), \alpha) \le \mathcal{N}(\mathcal{F}, L^q(P_n), \alpha)$$

for $r \le q$ since $\|f\|_{L^r(P_n)}$ is nondecreasing with r. Note that this is different for unweighted metrics: e.g. $\|x\|_r$ is nonincreasing in r, and hence $\mathcal{N}(\Theta, \|\cdot\|_r, \alpha)$ is also nonincreasing in r.

Definition 16. *An α-packing of $\mathcal{F}$ with respect to $L^r(P_n)$ is a set $f_1, \ldots, f_N \in \mathcal{F}$ such that*

$$\frac{1}{n}\sum_{i=1}^{n} |f_j(x_i) - f_k(x_i)|^r \ge \alpha^r$$

for any $j \neq k$. The size of the largest α-packing is denoted by $\mathcal{D}(\mathcal{F}, L^r(P_n), \alpha)$.

A standard relationship between covering and packing holds for any P:

$$\mathcal{D}(\mathcal{F}, L^r(P), 2\alpha) \le \mathcal{N}(\mathcal{F}, L^r(P), \alpha) \le \mathcal{D}(\mathcal{F}, L^r(P), \alpha).$$

In fact, this relationship is true for any metric.

14. PARAMETRIC AND NONPARAMETRIC CLASSES OF FUNCTIONS

There is no clear definition of what constitutes a "nonparametric class," especially since the same class of functions (e.g. neural networks) can be treated as either parametric or nonparametric (e.g. if neural network complexity is measured by matrix norms rather than number of parameters).

Consider the following (slightly vague) definition as a possibility:

Definition 17. *We will say that a class $\mathcal{F}$ is* parametric *if there is a constant C and a notion of dimension dim such that*

$$\sup_{P_n} \mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \lesssim \left(\frac{C}{\epsilon}\right)^{dim}.$$

We will say that $\mathcal{F}$ is nonparametric *if there is a $p > 0$ and C such that*

$$\sup_{P_n} \log \mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \asymp \left(\frac{C}{\epsilon}\right)^{p}. \qquad (14.158)$$

The requirement that (14.158) holds for all measures P_n and values of n is quite strong. Yet, we will show that as an upper bound, it is true for a variety of function classes.

However, one should keep in mind that there are also cases where dependence of the upper bound on n can lead to better overall estimates. The quantity

$$\sup_Q \log \mathcal{N}(\mathcal{F}, L^2(Q), \epsilon),$$

where supremum is taken over all discrete measures, is called *Koltchinskii-Pollard entropy.*

Let's consider a "parametric" class $\mathcal{F}$ such that functions in $\mathcal{F}$ are uniformly bounded: $|f|_\infty \leq 1$. Uniform boundedness implies an upper bound on the diameter: $D/2 \leq 1$. Then, taking $\delta = 0$ in Corollary 2, conditionally on $X_1, \ldots, X_n$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) &\leq \frac{12}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \alpha)} d\alpha \\
&\leq \frac{12}{\sqrt{n}} \int_0^1 \sqrt{d \log(1/\alpha)} d\alpha \\
&\leq c\sqrt{\frac{d}{n}}
\end{aligned}$$

Here it's useful to note that

$$\int_0^a \sqrt{\log(1/\alpha)} d\alpha \leq \begin{cases} 2a\sqrt{\log(1/a)} & a \leq 1/e \\ 2a & a > 1/e \end{cases}$$

The following theorem is due to D. Haussler (an earlier version with exponent $O(d)$ is due to Dudley '78):

Theorem 4. *Let $\mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$ be a class of binary-valued functions with VC dimension $vc(\mathcal{F}) = d$. Then for any n and any P_n,*

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \leq Cd(4e)^d \left(\frac{1}{\epsilon}\right)^{2d}.$$

We will explain what "VC dimension" means a bit later, and let's just say here that the class of thresholds has dimension 1 and the class of homogenous linear classifiers in $\mathbb{R}^d$ has dimension d. In particular, this removes the extraneous $\log(n+1)$ factor we had in Lecture 14 when analyzing thresholds.

14.1 A phase transition

Let us inspect the Dudley integral upper bound. Note that when we plug in

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \lesssim \left(\frac{1}{\epsilon}\right)^p,$$

the integral becomes

$$\int_\delta^{D/2} \varepsilon^{-p/2} d\varepsilon$$

If $p < 2$, the integral converges, and we can take $\delta = 0$. However, when $p > 2$, the lower limit of the integral matters and we get an overall bound of the order

$$\delta + n^{-1/2} \left[\varepsilon^{1-p/2}\right]_{\delta}^{D/2} \leq \delta + n^{-1/2}\delta^{1-p/2}$$

By choosing δ to balance the two terms (and thus minimize the upper bound) we obtain $\delta = n^{-1/p}$. Hence, for $p > 2$, the estimate on Rademacher averages provided by the Dudley bound is

$$\frac{1}{\sqrt{n}}\widehat{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/p}.$$

On the other hand, for $p < 2$, the Dudley entropy integral upper bound becomes (by setting $\delta = 0$) on the order of

$$n^{-1/2}D^{1-p/2} = O(n^{-1/2}),$$

yielding

$$\frac{1}{\sqrt{n}}\widehat{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/2}.$$

We see that there is a transition at $p = 2$ in terms of the growth of Rademacher averages (“elbow” behavior). The phase transition will be important in the rest of the course when we study optimality of nonparametric least squares.

Remark that in the $p < 2$ regime, the rate $n^{-1/2}$ is the same rate CLT rate we would have if we simply considered $\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f\right|$ (or the average with random signs) with a single function. Hence, the payment for the supremum over class $\mathcal{F}$ is only in a constant that may depend on $\mathcal{F}$ but does not depend on n.

14.2 Single scale vs chaining

It is also worthwhile to compare the single-scale upper bound we obtained earlier to the tighter upper bound given by chaining. In other words, we are comparing

$$\delta + \sqrt{\frac{\log \mathcal{N}(\delta)}{n}}$$

versus

$$\delta + \int_{\delta}^{D/2} \sqrt{\frac{\log \mathcal{N}(\varepsilon)}{n}} d\varepsilon,$$

simplifying the notation for brevity.

In the parametric case, the single-scale bound becomes (with the choice of $\delta = 1/n$)

$$\sqrt{\frac{\dim\ \log n}{n}}$$

while chaining gives

$$\sqrt{\frac{\dim}{n}}.$$

In the nonparametric case, the difference is more stark:

$$\delta + \sqrt{\frac{\delta^{-p}}{n}} \asymp n^{-\frac{1}{2+p}}$$

vs

$$n^{-1/2}$$

for $p < 2$, and

$$\delta + \frac{\delta^{1-p/2}}{\sqrt{n}} \asymp n^{-1/p}$$

for $p > 2$.

14.3 Linear class: Parametric or Nonparametric?

Let's take a closer look at the function class

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle : w \in \mathsf{B}_2^d\}$$

and take $\mathcal{X} = \mathsf{B}_2^d$. Recall that for a given $x_1, \ldots, x_n$,

$$\mathcal{F}|_{x_1,\ldots,x_n} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\} = \left\{Xw : w \in \mathsf{B}_2^d\right\}$$

where X is the $n \times d$ data matrix. As we have seen, the key quantity we need to compute is

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon).$$

What is a good upper bound for this quantity? What we had done earlier in the course was to discretize the set B_2^d to create a ε-net $w_1, \ldots, w_N$ of size $\mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon)$. Observe that for any $w, w' \in \mathsf{B}_2^d$,

$$\begin{aligned}\left(\frac{1}{n}\sum_{i=1}^n (\langle w, x_i \rangle - \langle w', x_i \rangle)^2\right)^{1/2} &\le \max_{i \in [n]} |\langle w - w', x_i \rangle| \\ &\le \max_{x \in \mathsf{B}_2^d} |\langle w - w', x \rangle| \\ &\le \left\|w - w'\right\|^2.\end{aligned}$$

This sequence of inequalities corresponds to

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \le \mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \le \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon). \tag{14.159}$$

where the sup-norm (or, *pointwise* over the domain) metric is $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)|$. Recall that the covering number of B_2^d is

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

This gives a "parametric" growth of entropy

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2/\varepsilon).$$

However, if d is large or infinite, this bound is loose. We will show that it also holds that

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \lesssim \varepsilon^{-2},$$

which is a nonparametric behavior. Hence, *the same class can be viewed as either parametric or nonparametric.* In fact, in the parametric behavior, it is not important that the domain of w is B_2^d since we would expect a similar estimate for other sets (including B_∞^d). In contrast, it will be crucial in nonparametric estimates that the norm of w is ℓ_2-bounded.

Jumping ahead, we will study neural networks and show a similar phenomenon: we can either count the number of neurons or connections (parameters) or we can calculate nonparametric "norm-based" estimates by looking at the norms of the layers in the network.

It's worth emphasizing again that (**??**) can lead to very loose bounds in high-dimensional situations. *A cover of function values on finite set of data can be significantly smaller than a cover with respect to sup norm.*

14.4 A more general result (Optional)

We have that for any fixed function

$$\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X))\right| \leq \mathrm{var}(f)^{1/2} = \|f-\mathbb{E}f\|_{L^2(P)}\,.$$

Obviously this implies

$$\sup_{f\in\mathcal{F}}\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X))\right| \leq \sup_{f\in\mathcal{F}}\mathrm{var}(f)^{1/2} =: \sigma$$

If we could ever prove

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X))\right| \leq C(\mathcal{F})\cdot\sigma,$$

it would imply that we only paid $C(\mathcal{F})$ for having a statement uniform in $f\in\mathcal{F}$.

Next, rather than assuming that functions in $\mathcal{F}$ are uniformly bounded, it will be enough to assume that they have an $L_2(P)$-integrable envelope F:

$$F(x) = \sup_{f\in\mathcal{F}}|f(x)|.$$

Rather than assuming that $F(x)\leq 1$, we shall assume that $\|F\|_{L^2(P)}^2 = \mathbb{E}F(X)^2 \leq \infty$ and everything will be phrased in terms of $\|F\|_{L^2(P)}^2$.

Now, let $H:[0,\infty)\mapsto[0,\infty)$ is such that $H(z)$ is non-decreasing for $z>0$ and $z\sqrt{H(1/z)}$ is non-decreasing for $z\in(0,1]$. Assume

$$\int_0^D \sqrt{H(1/x)}dx \leq C_H D\sqrt{H(1/D)}$$

for all $D\in(0,1]$, and suppose that

$$\sup_Q \log 2\mathcal{N}(\mathcal{F},L^2(Q),\tau\|F\|_{L^2(Q)}) \leq H(1/\tau)$$

for all $\tau>0$. With this control on Koltchinskii-Pollard entropy, it follows that

$$\mathbb{E}\sup\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X))\right| \lesssim \sigma\sqrt{H\left(\frac{2\|F\|_{L^2(P)}}{\sigma}\right)} \tag{14.160}$$

if n is large enough. We refer to [10] for more details, in particular Theorem 3.5.6 and the following corollaries.

Remarkably, under additional mild conditions on size of n, the inequality (14.160) can be reversed for a given P as soon as the entropy with respect to $L^2(P)$ indeed grows at least as $H\left(\frac{\|F\|_{L^2(P)}}{\sigma}\right)$.

Hence, the price we pay for uniformity in $f \in \mathcal{F}$ is truly

$$C(\mathcal{F}) \asymp \sqrt{H\left(\frac{\|F\|_{L^2(P)}}{\sigma}\right)}.$$

Of course, this expression is even simpler if $\sigma^2 = \sup_{f\in\mathcal{F}} \mathbb{E}(f(X) - \mathbb{E}f)^2$ is on the same order as $\|F\|^2_{L^2(P)} = \mathbb{E}\sup_f |f(X)|^2$.

15. COMBINATORIAL PARAMETERS

Let us gain some intuition for what can make $\widehat{\mathcal{R}}(\Theta)$ large. First, recall that

$$\widehat{\mathcal{R}}(\{\pm1\}^n) = \mathbb{E} \sup_{\theta\in\{\pm1\}^n} \langle \theta, \epsilon\rangle = n.$$

Next, suppose that for $\alpha > 0$ and $v \in \mathbb{R}^n$,

$$\alpha\{\pm1\}^n + v \subseteq \Theta.$$

Then

$$\widehat{\mathcal{R}}(\Theta) \geq \widehat{\mathcal{R}}(\alpha\{\pm1\}^n + v) = \widehat{\mathcal{R}}(\alpha\{\pm1\}^n) = \alpha\widehat{\mathcal{R}}(\{\pm1\}^n) \geq \alpha n$$

Hence, “large cubes” inside Θ make Rademacher averages large. It turns out, this is the only reason $\widehat{\mathcal{R}}(\mathcal{F}|_{x_1,\ldots,x_n})$ can be large!

The key question is whether $\mathcal{F}|_{x_1,\ldots,x_n}$ contains large cubes for a given class $\mathcal{F}$.

15.1 Binary-Valued Functions

Let's start with function classes of $\{0,1\}$-valued functions. In this case, $\mathcal{F}_{x_1,\ldots,x_n}$ is either a full $\{0,1\}^n$ cube or not. Consider the particular example of threshold functions on the real line. Take any point x_1. Clearly, $\mathcal{F}|_{x_1} = \{0,1\}$, which is a one-dimensional cube. Take two points x_1, x_2. We can only realize sign patters $(0,0), (0,1), (1,1)$, but not $(1,0)$. Hence, for no two points can we get a cube.

Definition 18. *Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$. We say that $\mathcal{F}$ shatters $x_1,\ldots,x_n \in \mathcal{X}$ if $\mathcal{F}|_{x_1,\ldots,x_n} = \{0,1\}^n$. The Vapnik-Chervonenkis dimension of $\mathcal{F}$ is*

$$vc(\mathcal{F}) = \max\{n : \mathcal{F} \text{ shatters some } x_1,\ldots,x_n\}$$

Lemma 27 (Sauer-Shelah-Vapnik-Chervonenkis)**.** *If $vc(\mathcal{F}) = d < \infty$,*

$$\operatorname{card}\left(\mathcal{F}|_{x_1,\ldots,x_n}\right) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$$

This result is quite remarkable. It says that as soon as $n > \mathrm{vc}(\mathcal{F})$, the proportion of the cube that can be realized by $\mathcal{F}$ becomes very small (n^d vs 2^n). This combinatorial result is at the heart of empirical process theory and the early developments in pattern recognition.

In particular, the lemma can be interpreted as a covering number upper bound:

$$\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \leq \left(\frac{en}{d}\right)^d$$

for any $\epsilon > 0$. Observe that these numbers are with respect to $L^\infty(P_n)$ rather than $L^2(P_n)$, and hence can be an overkill. Indeed, $L^\infty(P_n)$ covering numbers are necessarily n-dependent while we can hope to get dimension-independent $L^2(P_n)$ covering numbers. Indeed, this result (Dudley, Haussler) was already mentioned: for a binary-valued class with finite $\mathrm{vc}(\mathcal{F}) = d$,

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \lesssim \left(\frac{C}{\epsilon}\right)^{Cd}.$$

Hence, a class with finite VC dimension is "parametric". On the other hand, if $\mathrm{vc}(\mathcal{F})$ is infinite, then $\mathcal{F}|_{x_1,\ldots,x_n}$ is a full cube for arbitrarily large n (for some appropriately chosen points). Hence, Rademacher averages of this set are too large and there is no uniform convergence for all P (to see this, consider P supported on the shattered set). Hence, finiteness of VC dimension is a characterization (of both distribution-free learnability and uniform convergence).

A word of caution: VC dimension does not always correspond to "number of parameters." For instance, the one-parameter family $\mathcal{F} = \{x \mapsto \mathbf{1}\{\mathsf{sin}(\alpha x) \geq 0\} : \alpha \in \mathbb{R}\}$ over $\mathcal{X} = \mathbb{R}$ has infinite VC dimension.

15.2 Real-Valued Functions

For binary-valued functions, the size of the cube contained in $\mathcal{F}|_{x_1,\ldots,x_n}$ was trivially 1, and we only varied n to see where the phase transition occurs. In contrast, for a general real-valued function class, it is feasible that $\mathcal{F}|_{x_1,\ldots,x_n}$ contains a cube of size α, but not larger than α; this extra parameter is in addition to the dimensionality of the cube. To deal with this extra degree of freedom, we fix the scale α and ask for the largest size n such that $\mathcal{F}|_{x_1,\ldots,x_n}$ contains a (translate of a) cube of size α. A true containment statement would read $s + (\alpha/2)\{-1, 1\}^n \subseteq \mathcal{F}|_{x_1,\ldots,x_n}$. However, it is enough to ask that the equalities for the vertices are replaced with inequalities:

Definition 19. *We say that $\mathcal{F}$* shatters *a set of points $x_1, \ldots, x_n$ at scale α if there exists $s \in \mathbb{R}^n$ such that*

$$\forall \epsilon \in \{\pm 1\}^n, \exists f \in \mathcal{F} \quad s.t. \quad \begin{cases} f(x_t) \geq s_t + \alpha/2 & \text{if } \epsilon_t = +1 \\ f(x_t) \leq s_t - \alpha/2 & \text{if } \epsilon_t = -1 \end{cases}$$

The combinatorial dimension $vc(\mathcal{F}, \alpha)$ of $\mathcal{F}$ (on domain $\mathcal{X}$) at scale α is defined as the size n of the largest shattered set.

This scale-sensitive dimension (or 'fat-shattering' dimension) was introduced by [16].

15.2.1 Example: non-decreasing functions

Consider the class of nondecreasing functions $f : \mathbb{R} \to [0, 1]$. First, observe that a pointwise cover of this class does not exist ($\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) = \infty$ for any $\epsilon < 1/2$). However, $\mathcal{N}(\mathcal{F}, L^\infty(P_n), \epsilon)$ is necessarily finite. Let's calculate the scale-sensitive dimension of this class.

Claim: $\mathrm{vc}(\mathcal{F}, \epsilon) \lesssim \epsilon^{-1}$. Indeed, fix any $x_1, \ldots, x_n$ and assume these are arranged in an increasing order. Suppose $\mathcal{F}$ shatters this set. Take the alternating sequence $\epsilon = (+1, -1, \ldots)$. We then must have a nondecreasing function that is at least $s_1 + \alpha/2$ at x_1 but then no greater than $s_2 - \alpha/2$ at x_2. The nondecreasing constraint implies that $s_2 \geq s_1 + \alpha$. A similar argument then holds for the next point and so forth. Since functions are bounded, $n\alpha \leq 1$, which concludes the proof.

15.2.2 Control of covering numbers

The following generalization of the earlier result for binary-valued functions is due to Mendelson and Vershynin:

Theorem 5. *Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to [-1, 1]$. Then for any distribution P,*

$$\mathcal{N}(\mathcal{F}, L_2(P), \varepsilon) \leq \left(\frac{c}{\varepsilon}\right)^{c \cdot vc(\mathcal{F}, \varepsilon/c)}$$

for all $\epsilon > 0$. Here c is an absolute constant.

In particular, plugging into the entropy integral yields

$$\int \sqrt{\mathrm{vc}(\mathcal{F}, \varepsilon) \log(1/\varepsilon)} d\varepsilon$$

Rudelson-Vershynin: $\log(1/\epsilon)$ can be removed.

Back to the class of non-decreasing functions, we immediately get

$$\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \lesssim \varepsilon^{-1} \cdot \log\left(\frac{c}{\varepsilon}\right).$$

In particular, Rademacher averages of this class scale as $n^{-1/2}$ since this is a nonparametric class with entropy exponent $p < 2$.

15.3 Scale-sensitive dimension of linear class via Perceptron

In this section, we will prove that

Proposition 1. *For*

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathsf{B}_2^d\}$$

and $\mathcal{X} \subseteq \mathsf{B}_2^d$, it holds that

$$vc(\mathcal{F}, \alpha) \le 16\alpha^{-2}.$$

We turn to the Perceptron algorithm, defined as follows. We start with $\widehat{w}_0 = 0$. At time $t = 1, \ldots, n$, we observe $x_t \in \mathcal{X}$ and predict $\widehat{y}_t = \text{sign}(\langle \widehat{w}_t, x_t\rangle)$, a *deterministic* guess of the label of x_t given the hypothesis $\widehat{w}_t$. We then observe the true label of the example $y_t \in \{\pm 1\}$. If $\widehat{y}_t \neq y_t$, we update

$$\widehat{w}_{t+1} = \widehat{w}_t + y_t x_t,$$

and otherwise $\widehat{w}_{t+1} = \widehat{w}_t$.

Lemma 28 (Novikoff'62)**.** *For any sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathsf{B}_2^d \times \{\pm 1\}$ the Perceptron algorithm makes at most γ^{-2} mistakes, where γ is the margin of the sequence, defined as*

$$\gamma = \max_{w^* \in \mathsf{B}_2^d} \min_t y_t \langle w^*, x_t\rangle \vee 0$$

Proof. If a mistake is made on round t,

$$\|\widehat{w}_{t+1}\|^2 = \|\widehat{w}_t + y_t x_t\|^2 \le \|\widehat{w}_t\|^2 + 2y_t\langle \widehat{w}_t, x_t\rangle + 1 \le \|\widehat{w}_t\|^2 + 1$$

Denote the number of mistakes at the end as m. Then $\|\widehat{w}_{n+1}\|^2 \le m$. Next, for w^*,

$$\gamma \le \langle w^*, y_t x_t\rangle = \langle w^*, \widehat{w}_{t+1} - \widehat{w}_t\rangle,$$

and so by summing and telescoping, $m\gamma \le \langle w^*, \widehat{w}_{n+1}\rangle \le \sqrt{m}$. This concludes the proof. □

Remarkably, the number of mistakes does not depend on the dimension d. We will now show that the mistake bound translates into a bound on the scale-sensitive dimension.

Proof of Proposition. Suppose there exist a shattered set $x_1, \ldots, x_m \in \mathsf{B}_2^d$: there exists $s_1, \ldots, s_m \in [-1, 1]$ such that for any sequence of signs $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ there exists a $w_\epsilon \in \mathsf{B}_2^d$ such that

$$\epsilon_i(\langle w_\epsilon, x_i\rangle - s_i) \ge \alpha/2.$$

Claim: we can reparametrize the problem so that $s_i = 0$. Indeed, take

$$\tilde{w}_\epsilon = [w_\epsilon, 1], \quad \tilde{x}_i = [x_i, -s_i].$$

Then we have

$$\epsilon_i \langle \tilde{w}_\epsilon, \tilde{x}_i\rangle \ge \alpha/2.$$

while the norms are at most $\sqrt{2}$:

$$\|\tilde{w}_\epsilon\|^2 = \|w_\epsilon\|^2 + 1 \le 2, \qquad \|\tilde{x}_i\|^2 \le 2$$

Now comes the key step. We run Perceptron on the sequence $\tilde{x}_1/\sqrt{2}, \ldots, \tilde{x}_m/\sqrt{2}$ and $y_i = -\widehat{y}_i$. That is, we force Perceptron to make mistakes on every round, no matter what the predictions are. It is important that Perceptron makes deterministic predictions for this argument to work. Note that the sequence of predictions of Perceptron defines the sequence $y = (y_1, \ldots, y_m)$ with

$$y_i \langle \tilde{w}_y/\sqrt{2}, \tilde{x}_i/\sqrt{2} \rangle \geq \alpha/4.$$

Hence, by Novikoff's result,

$$m \leq 16/\alpha^2.$$

□

Interestingly, both Perceptron and VC theory were developed in the 60's as distinct approaches (online vs batch), yet the connection between them runs deeper than was recognized, until recently. In particular, the above proof in fact shows that a stronger *sequential* version of $\mathrm{vc}(\mathcal{F}, \alpha)$ is also bounded by $16\alpha^{-2}$, where (roughly speaking) sequential analogues allow the sequence to evolve as a predictable process with respect to a dyadic filtration. It turns out that there are sequential analogues of Rademacher averages, covering numbers, Dudley chaining, and combinatorial dimensions, and these govern *online* (rather than i.i.d.) learning. We will mention these towards the end of the course.

16. PREDICTION WITH LIPSCHITZ LOSS FUNCTIONS

In the past few lectures, we have developed tools for analyzing the expected suprema of empirical processes. We have already seen that such quantities can be used to derive sample complexity bounds for empirical risk minimization algorithms. Let us recall the setup. *Excess loss* with respect to a class of functions $\mathcal{F}$ is defined as

$$\mathbb{E}\boldsymbol{\ell}(f(X), Y) - \min_{f \in \mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) \tag{16.161}$$

for some $\boldsymbol{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Earlier, we have shown that ERM

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}(f(X_i), Y_i)$$

enjoys

$$\mathbb{E}\boldsymbol{\ell}(\widehat{f}(X), Y) - \min_{f \in \mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}(f(X_i), Y_i).$$

The latter is at most

$$2\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{\ell}(f(X_i), Y_i) \tag{16.162}$$

by symmetrization, which is Rademacher averages of the loss class

$$\boldsymbol{\ell} \circ \mathcal{F}|_{(X_1, Y_1), \ldots, (X_n, Y_n)}$$

We would like to further upper bound this with Rademacher averages of the function class itself. This can be done if $\boldsymbol{\ell}$ is Lipschitz in the first argument.

Lemma 29 (Contraction)**.** *Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz, $i = 1, \dots, n$. Let $\Theta \subset \mathbb{R}^n$ and $\phi \circ \theta = (\phi_1(\theta_1), \dots, \phi_n(\theta_n))$ for $\theta \in \Theta$. Denote $\phi \circ \Theta = \{\phi \circ \theta : \theta \in \Theta\}$. Then*

$$\widehat{\mathcal{R}}(\phi \circ \Theta) \leq \widehat{\mathcal{R}}(\Theta).$$

Proof. Conditionally on $\epsilon_1, \dots, \epsilon_{n-1}$,

$$\begin{aligned}
\mathbb{E}_{\epsilon_n} \sup_{\theta \in \Theta} \langle \phi \circ \theta, \epsilon \rangle &= \frac{1}{2} \left(\sup_{\theta \in \Theta} \left\{ \langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \phi_n(\theta_n) \right\} + \sup_{\theta' \in \Theta} \left\{ \langle \phi \circ \theta'_{1:n-1}, \epsilon_{1:n-1} \rangle - \phi_n(\theta'_n) \right\} \right) \\
&\leq \frac{1}{2} \sup_{\theta, \theta' \in \Theta} \langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \langle \phi \circ \theta'_{1:n-1}, \epsilon_{1:n-1} \rangle + |\theta_n - \theta'_n| \\
&= \frac{1}{2} \sup_{\theta, \theta' \in \Theta} \langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \langle \phi \circ \theta'_{1:n-1}, \epsilon_{1:n-1} \rangle + \theta_n - \theta'_n \\
&= \frac{1}{2} \left(\sup_{\theta \in \Theta} \left\{ \langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \theta_n \right\} + \sup_{\theta' \in \Theta} \left\{ \langle \phi \circ \theta'_{1:n-1}, \epsilon_{1:n-1} \rangle - \theta'_n \right\} \right) \\
&= \mathbb{E}_{\epsilon_n} \sup_{\theta \in \Theta} \langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \epsilon_n \theta_n
\end{aligned}$$

The inequality follows from the Lipschitz condition and the following equality is justified because of the symmetry of the other two terms with respect to renaming θ and θ'. Proceeding to remove the other signs concludes the proof. □

We now apply this lemma to functions $\phi_i(\cdot) = \boldsymbol{\ell}(\cdot, Y_i)$. As long as these functions are L-Lipschitz, contraction lemma gives

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{\ell}(f(X_i), Y_i) \leq L \cdot \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \tag{16.163}$$

the (expected) Rademacher averages of $\mathcal{F}$. The argument can be seen as a generalization of the argument in (10.140) for classification where we "erased" multipliers $(1 - 2Y_i)$.

The simple analysis we just performed applies to any Lipschitz loss function. For uniformly bounded $\mathcal{F}$ and $\mathcal{Y}$, square loss is Lipschitz, but that is no longer true for unbounded $\mathcal{Y}$ (e.g. for real-value prediction with Gaussian noise). Hence, such an analysis only goes so far.

Second, observe that one would only obtain rates $n^{-1/2}$ or worse with such an analysis, while we might hope to have faster decrease. For instance, in finite-dimensional regression, one can recall the classical $d \cdot n^{-1}$ rates for Least Squares.

A quick inspection tells us that the second step in the sequence of inequalities

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \mathbf{L}(f^*) \leq \mathbb{E}\left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] \leq \mathbb{E} \sup_{f \in \mathcal{F}} \left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] \tag{16.164}$$

for ERM $\widehat{f}$ may be too loose. The second step only used the fact that $\widehat{f}$ belongs to $\mathcal{F}$. It turns out one can localize its place in $\mathcal{F}$ better than that. Before turning to this question of localization, let us point out a relationship between the problems of estimation and prediction with square loss.

17. REGRESSION. PREDICTION VS ESTIMATION

As before, let $\mathcal{S} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be a set of i.i.d. pairs with distribution $P = P_X \times P_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$. Let $f^*(x) = \mathbb{E}[Y|X = x]$ be the *regression function.* One can show that

$$f^* \in \operatorname*{argmin}_f \mathbb{E}(f(X) - Y)^2$$

where minimization is over all measurable functions. Given a class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$, we also define

$$f_{\mathcal{F}} \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2$$

to be the best predictor within the class $\mathcal{F}$.

Note that for any function $f : \mathcal{X} \to \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}(f(X) - Y)^2 - \inf_{h \in \mathcal{F}} \mathbb{E}(h(X) - Y)^2 && (17.165)\\
&= \mathbb{E}(f(X) - f^*(X) + f^*(X) - Y)^2 - \inf_{h \in \mathcal{F}} \mathbb{E}(h(X) - f^*(X) + f^*(X) - Y)^2 \\
&= \mathbb{E}(f(X) - f^*(X))^2 - \inf_{h \in \mathcal{F}} \mathbb{E}(h(X) - f^*(X))^2 \\
&= \|f - f^*\|^2_{L^2(P)} - \inf_{h \in \mathcal{F}} \|h - f^*\|^2_{L^2(P)} \,. && (17.166)
\end{aligned}$$

The penultimate equality holds because the cross term

$$\mathbb{E}\left[(f(X) - f^*(X))(f^*(X) - Y)\right] = 0,$$

as follows by conditioning on X and using the definition of f^*.

On the left-hand side of (17.165), we have the object of interest in Statistical Learning: predicting well relative to a given class $\mathcal{F}$ (e.g. agnostic PAC learning in the realm of computational learning theory). On the other hand, (17.166) measures the quality of estimation of an unknown regression function in the $L^2(P)$ norm. This object is within the purview of Statistics. We see that the problem of prediction and the problem of estimation naturally coincide for square loss.

Of course, we will be interested in analyzing estimators $\widehat{f}$ constructed on the basis of n datapoints. The hat on $\widehat{f}$ reminds us about the dependence on $\mathcal{S}$. Then (17.165) will be evaluated with f replaced by $\widehat{f}$.

Two standard scenarios:

- Well-specified case: given some class $\mathcal{F}$, assume $f^* \in \mathcal{F}$. More precisely, P is such that the regression function is in the class $\mathcal{F}$. In this case, (17.166) becomes $\|f - f^*\|^2_{L^2(P)}$.
- Misspecified case: do not insist that $f^* \in \mathcal{F}$. Upper bounds on (17.166) are called Oracle Inequalities in statistics, while the prediction form has been studied in statistical learning theory (sometimes under the name of Agnostic PAC).

The misspecified problem arises naturally as a relaxation of an assumption on the form of the distribution.

18. NONPARAMETRIC REGRESSION: WELL-SPECIFIED CASE

We will start with "fixed design": $x_1, \ldots, x_n \in \mathcal{X}$ are fixed. Let

$$Y_i = f^*(x_i) + \eta_i$$

where η_i are zero-mean independent subGaussian. Suppose $f^* \in \mathcal{F}$. Goal: estimate f^* on the points $x_1, \ldots, x_n$ (denoise the observed values). That is, the goal is to provide nonasymptotic bounds on

$$\mathbb{E}_\eta \left\| \widehat{f} - f^* \right\|^2_{L^2(P_n)},$$

where $\widehat{f}$ is the least squares (ERM) constrained to $\mathcal{F}$. In constrast, in random design the goal is w.r.t. $L^2(P)$ with P unknown, while here P_n is known. We write the $L^2(P_n)$ norm more succinctly as $\mathbb{E} \left\| \widehat{f} - f^* \right\|^2_n$.

Since

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - Y_i)^2 = \|f - Y\|_n^2$$

we have

$$\|f^* - Y\|_n^2 \geq \left\| \widehat{f} - Y \right\|^2_n = \left\| \widehat{f} - f^* + f^* - Y \right\|^2_n = \left\| \widehat{f} - f^* \right\|^2_n + \|f^* - Y\|_n^2 + 2\langle \widehat{f} - f^*, f^* - Y \rangle_n$$

where $\langle a, b \rangle_n = \frac{1}{n}\langle a, b \rangle$. Thus,

$$\left\| \widehat{f} - f^* \right\|^2_n \leq 2\langle \eta, \widehat{f} - f^* \rangle_n \tag{18.167}$$

which is the *Basic Inequality* developed earlier in (6.88) for linear regression.

18.1 Informal intuition for localization

Before developing the localization approach, we provide some intuition. The first intuition comes from viewing (18.167) as a fixed point.

Let's assume for simplicity that η_i are 1-subGaussian. For fixed $a \in \mathbb{R}^n$, we have that with high probability

$$\langle \eta, a \rangle \lesssim \|a\| \tag{18.168}$$

Hence, if it holds that

$$\|a\|^2 \leq \langle \eta, a \rangle,$$

then $\|a\| \lesssim 1$, or, dividing by n, $\|a\|_n^2 \lesssim 1/n$.

We can try to repeat this argument with a being the values of $\widehat{f} - f^*$ on the data. However, since $\widehat{f}$ depends on η, we do not have the averaging in (18.168) that we need. Still, we can do the mental experiment of assuming that the dependence is "weak" (e.g. we fit linear regression in small d and large n). Then a bound on the size of $\left\| \widehat{f} - f^* \right\|_n$ would lead to an improved bound on the RHS of the basic inequality, which would in turn tighten the bound on the LHS of the basic inequality, suggesting some kind of a fixed point. It also seems intuitive that this fixed point likely depends on $\mathcal{F}$ and its richness.

18.2 1st approach to localization: ratio-type inequalities

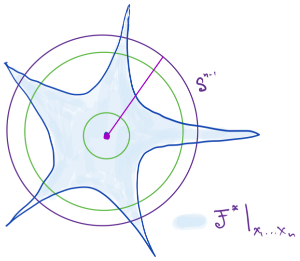
To simplify the proof somewhat, we will assume that $\eta_1, \ldots, \eta_n$ are independent standard normal $N(0,1)$.

We proceed as in the linear case earlier in the course. First, we divide both sides of the Basic Inequality (18.167) by $\left\|\widehat{f} - f^*\right\|_n$ and further upper bound the right-hand side by a supremum over f, removing the dependence of the algorithm on the data:

$$\left\|\widehat{f} - f^*\right\|_n \leq 2 \sup_{f \in \mathcal{F}} \langle \eta, \frac{f - f^*}{\|f - f^*\|_n} \rangle_n \tag{18.169}$$

By squaring both sides, we would get an upper bound on the estimation error (in probability or in expectation).

Let us use the shorthand $\mathcal{F}^* = \mathcal{F} - f^*$. The rest of the discussion will be about complexity of the neighborhood around f^* in $\mathcal{F}$, or, equivalently, complexity of the neighborhood of 0 in $\mathcal{F}^*$. Observe that we only care about values of functions on the data $x_1, \ldots, x_n$, so the discussion is really about the set $\mathcal{F}^*|_{x_1,\ldots,x_n}$, drawn in blue below.



At this point, one can say that there is no difference from the linear case, and we should just go ahead and analyze

$$\sup_{g \in \mathcal{F}^*} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

After all, this is just the Gaussian width (normalized by $\sqrt{n}$) of the subset of the sphere obtained by rescaling all the functions:

$$K = \{v \in \mathsf{S}^{n-1} : \exists g \in \mathcal{F}^* \ \text{ s.t. } \ v = (g(x_1), \ldots, g(x_n))/(\sqrt{n}\,\|g\|_n)\}.$$

(here the normalization is because $\|g\|_n$ is scaled as $1/\sqrt{n}$ times the ℓ_2 norm.) How big is this subset of the sphere? Note: if the set is all of S^{n-1}, we are doomed since in that case

$$\sup_{g \in \mathcal{F}^*} \langle \eta, \frac{g}{\|g\|_n} \rangle_n = \sup_{v \in \mathsf{S}^{n-1}} \frac{1}{\sqrt{n}} \langle \eta, v \rangle = \frac{1}{\sqrt{n}} \|\eta\| \sim 1$$

and does not converge to zero. What we would need is that K is a *significantly smaller* subset of the sphere. In the linear case, this was easy: we simply used the fact that the subset is d-dimensional. However, for nonlinear functions, it is not easy to see what the set is.

There is a bigger problem, however. Upon rescaling every vector to the sphere, all the functions are treated equally even if their unscaled versions are very close to being zero (that is, close to f^* in the original class $\mathcal{F}$). In other words, the quantity

$$\sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

can be potentially much smaller than the unrestricted supremum. This is depicted in the above figure. If we look at functions within the smaller green sphere, its rescaled version is the whole sphere. However, at larger scales (e.g. the larger green sphere), the set can be much smaller. Understanding the map

$$u \mapsto \sup_{g \in \mathcal{F}^*: \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

will be key. In particular, we can break up the balance at scale u and instead have a better upper bound

$$\left\| \widehat{f} - f^* \right\|_n \leq u + 2 \sup_{g \in \mathcal{F}^*: \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n \tag{18.170}$$

Indeed, to show (18.170), write

$$\begin{aligned}
\left\| \widehat{f} - f^* \right\|_n &= \left\| \widehat{f} - f^* \right\|_n \mathbf{1}\left\{ \left\| \widehat{f} - f^* \right\|_n < u \right\} + \left\| \widehat{f} - f^* \right\|_n \mathbf{1}\left\{ \left\| \widehat{f} - f^* \right\|_n \geq u \right\} \\
&\leq u + \left\| \widehat{f} - f^* \right\|_n \mathbf{1}\left\{ \left\| \widehat{f} - f^* \right\|_n \geq u \right\} \\
&\leq u + 2\langle \eta, \frac{\widehat{f} - f^*}{\left\| \widehat{f} - f^* \right\|_n} \rangle_n \times \mathbf{1}\left\{ \left\| \widehat{f} - f^* \right\|_n \geq u \right\} \\
&\leq u + 2 \sup_{g \in \mathcal{F}^*: \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n
\end{aligned}$$

Consider the following assumption:

Definition 20. *A class $\mathcal{H}$ is* star-shaped *(around 0) if $h \in \mathcal{H}$ implies $\lambda h \in \mathcal{H}$ for $\lambda \in [0,1]$. In particular, if $\mathcal{H}$ is convex and contains 0, it is star-shaped.*

We will assume that $\mathcal{F}^*$ is star-shaped. In particular, if $\mathcal{F}$ is convex, then $\mathcal{F}^*$ is star-shaped. The key property of a star-shaped class is that by increasing the radius, the sets cannot become more complex, as for any function there is a scaled copy of it at a smaller magnitude.

In light of this last remark, we claim that the inequality $\|g\|_n \geq u$ in the supremum in (18.170) can be replaced with an *equality* if the class is star-shaped. Indeed, for any $g \in \mathcal{F}^*$ with $\|g\|_n \geq u$, there is a corresponding function $h = u\frac{g}{\|g\|_n}$ with norm $\|h\|_n = u$ and

$$\langle \eta, \frac{g}{\|g\|_n} \rangle_n = \langle \eta, \frac{h}{u} \rangle_n$$

Hence,

$$\langle \eta, \frac{g}{\|g\|_n} \rangle_n \leq \frac{1}{u} \sup_{h \in \mathcal{F}^*: \|h\|_n = u} \langle \eta, h \rangle_n$$

Taking a supremum on the LHS over g with $\|g\|_n \geq u$ gives an upper bound on (18.170) as

$$\begin{aligned}
\left\| \widehat{f} - f^* \right\|_n &\leq u + \frac{2}{u} \sup_{g \in \mathcal{F}^*: \|g\|_n = u} \langle \eta, g \rangle_n \\
&\leq u + \frac{2}{u} \sup_{g \in \mathcal{F}^*: \|g\|_n \leq u} \langle \eta, g \rangle_n
\end{aligned} \tag{18.171}$$

where in the last step we included all the functions below level u. We will use concentration to replace the second term with its expectation. In particular, define

$$Z(u) = \sup_{g \in \mathcal{F}^* : \|g\|_n \leq u} \langle \eta, g \rangle_n$$

and

$$G(u) = \mathbb{E}Z(u).$$

If we were to replace $Z(u)$ on the RHS of (18.171) with $G(u)$, the natural balance between the two terms would be

$$u = \frac{2}{u} G(u)$$

Definition 21. *The* critical radius δ_n *will be the minimum* δ *satisfying*

$$G(\delta) \leq \delta^2/2$$

One can ask if this critical radius is actually well-defined. This follows from the following:

Lemma 30. *If* $\mathcal{F}^*$ *is star-shaped, the function* $u \mapsto G(u)/u$ *is non-increasing.*

Proof. Let $\delta' < \delta$. Take any $h \in \mathcal{F}^*$ with $\delta' < \|h\|_n \leq \delta$. By star-shapedness,

$$h' = \left(\frac{\delta'}{\delta}\right) h \in \mathcal{F}^*$$

and $\|h'\|_n = \frac{\delta'}{\delta} \|h\|_n \leq \delta'$. Hence,

$$\langle \eta, h \rangle_n = \frac{\delta}{\delta'} \langle \eta, h' \rangle_n \leq \frac{\delta}{\delta'} Z(\delta')$$

Taking supremum on the left-hand side over h with $\|h\|_n \leq \delta$, as well as expectation on both sides, finishes the proof. □

In particular, for any $u \geq \delta_n$,

$$G(u) \leq u^2/2$$

Indeed,

$$G(u) = u\frac{G(u)}{u} \leq u\frac{G(\delta_n)}{\delta_n} \leq u\delta_n/2 \leq u^2/2. \tag{18.172}$$

To formally replace $Z(u)$ with $G(u)$ in the balancing equation, we need a concentration result.

Lemma 31 (Gaussian Concentration)**.** *Let $\eta = (\eta_1, \dots, \eta_n)$ be a vector of independent standard normals. Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz (w.r.t. Euclidean norm). Then for all $t > 0$*

$$\mathbb{P}\left(\phi(\eta) - \mathbb{E}\phi \geq t\right) \leq \exp\left\{-\frac{t^2}{2L^2}\right\}$$

First, observe that $Z(u)$ is $(u/\sqrt{n})$-Lipschitz function of η. Omitting the argument u,

$$Z[\eta] - Z[\eta'] \leq \sup_{g \in \mathcal{F}^*, \|g\|_n \leq u} \langle \eta, g\rangle_n - \langle \eta', g\rangle_n \leq \left\|\eta - \eta'\right\|_n \sup_{g \in \mathcal{F}^*, \|g\|_n \leq u} \|g\|_n \leq \frac{u}{\sqrt{n}} \left\|\eta - \eta'\right\|$$

Hence, for any $u > 0$,

$$\mathbb{P}\left(Z(u) - \mathbb{E}Z(u) \geq t\right) \leq \exp\left\{-\frac{nt^2}{2u^2}\right\} \tag{18.173}$$

In particular, by setting $t = u^2$,

$$\mathbb{P}\left(Z(u) \geq G(u) + u^2\right) \leq \exp\left\{-\frac{nu^2}{2}\right\} \tag{18.174}$$

In light of (18.172), we have proved

Lemma 32. *Assuming $\mathcal{F}^*$ is star-shaped, with probability at least $1 - \exp\left\{-\frac{nu^2}{2}\right\}$,*

$$Z(u) \leq 1.5u^2 \tag{18.175}$$

for any $u \geq \delta_n$.

Thus, from (18.171), we have

$$\left\|\widehat{f} - f^*\right\|_n \leq 4u \tag{18.176}$$

with probability at least $1 - \exp\left\{-\frac{nu^2}{2}\right\}$, for any $u \geq \delta_n$. Squaring both sides, yields

Theorem 6. *Assume $x_1, \dots, x_n$ are fixed, $\eta_1, \dots, \eta_n$ are i.i.d. standard normal, and $Y_i = f^*(x_i) + \eta_i$ with $f^* \in \mathcal{F}$. Assume $\mathcal{F} - f^*$ is star-shaped and δ_n the corresponding critical radius. Then constrained least squares $\widehat{f}$ satisfies*

$$\mathbb{P}\left(\left\|\widehat{f} - f^*\right\|_n^2 \geq 16s\delta_n^2\right) \leq \exp\left\{-\frac{ns\delta_n^2}{2}\right\} \tag{18.177}$$

for any $s \geq 1$. In particular, this implies

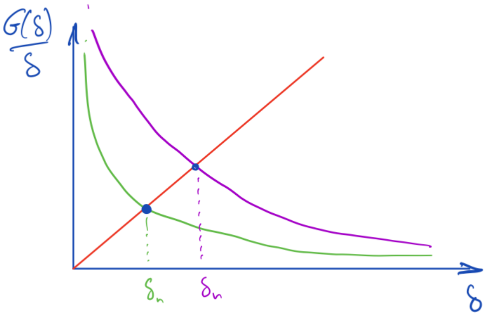
$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 \lesssim \delta_n^2 + \frac{1}{n}.$$

Note: in the literature, you will find a slightly different parametrization. Write $\psi(r) = \mathbb{E}Z(\sqrt{r})$. In other words, $\psi(u^2) = G(u)$. Then the property $G(u)/u$ non-increasing translates into ψ having the *subroot* property:

$$\psi(ra) \le \sqrt{a}\psi(r)$$

using the same type of proof as above. The fixed point then reads as the smallest r such that $\psi(r) \le r$ (ignoring the constant).

Let's quickly discuss the behavior of $G(\delta)/\delta$.



The above sketch shows the function $\delta \mapsto G(\delta)/\delta$ for two classes of functions. The purple curve corresponds to a more complex class, since the Gaussian width (normalized by δ) grows faster as $\delta \to 0$. The corresponding fixed point is larger for a more rich class.

18.3 2nd approach to localization: offset

We start again with the basic inequality

$$\left\|\widehat{f} - f^*\right\|_n^2 \le 2\langle \eta, \widehat{f} - f^* \rangle_n$$

and trivially write it as

$$\left\|\widehat{f} - f^*\right\|_n^2 \le 4\langle \eta, \widehat{f} - f^* \rangle_n - \left\|\widehat{f} - f^*\right\|_n^2$$

Now take the supremum and expectation on both sides:

$$\begin{aligned}\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 &\le \mathbb{E}\sup_{f\in\mathcal{F}} 4\langle \eta, f - f^* \rangle_n - \|f - f^*\|_n^2 \\ &= \mathbb{E}\sup_{g\in\mathcal{F}-f^*} \frac{1}{n}\sum_{i=1}^n 4\eta_i g(x_i) - g(x_i)^2\end{aligned}$$

which we shall call *the offset Rademacher (or Gaussian) averages.*

Contrast this approach with the first approach where we divided both sides by the norm $\left\|\widehat{f} - f^*\right\|_n$ and then upper bounded by supremum over an appropriately localized subset, then squared both sides.

Surprisingly, this somewhat simpler approach yields correct upper bounds. Note that the negative quadratic term annihilates the fluctuations of the term $\eta_i g(x_i)$ when the magnitude of g becomes large enough (beyond some critical radius). Hence, the supremum is achieved in a finite radius, no larger than the critical radius:

Lemma 33. *Let δ_n be the critical radius. Then for any $c \geq 1$,*

$$\mathbb{P}\left(\sup_{g\in\mathcal{F}^*} 2c\langle\eta, g\rangle_n - \|g\|_n^2 > 2c^2\delta_n^2 + \frac{2c^2u}{n}\right) \leq \exp\{-u/2\} \tag{18.178}$$

In particular,

$$\mathbb{E}\sup_{g\in\mathcal{F}^*} 2\langle\eta, g\rangle_n - \|g\|_n^2 \lesssim \delta_n^2 + \frac{1}{n}.$$

Proof. By Gaussian concentration,

$$\mathbb{P}\left(Z(\delta_n) \geq \mathbb{E}Z(\delta_n) + t\delta_n\right) \leq \exp\left\{-\frac{nt^2}{2}\right\}. \tag{18.179}$$

We now condition on the complement of the above event. Take $g \in \mathcal{F}^*$. Consider two cases. First, if $\|g\|_n \leq \delta_n$ then

$$2c\langle\eta, g\rangle_n - \|g\|_n^2 \leq 2cZ(\delta_n) \leq 2c\left(\mathbb{E}Z(\delta_n) + t\delta_n\right) \leq 2c\left(\frac{\delta_n^2}{2} + t\delta_n\right) \leq c(t+\delta_n)^2 \tag{18.180}$$

Second, if $\|g\|_n \geq \delta_n$, we set $r = \delta_n / \|g\|_n \leq 1$. Then

$$2c\langle\eta, g\rangle_n - \|g\|_n^2 = \frac{2c}{r}\langle\eta, \frac{\delta_n}{\|g\|_n}g\rangle_n - \frac{\delta_n^2}{r^2} \leq \frac{2c}{r}Z(\delta_n) - \frac{\delta_n^2}{r^2} = \frac{2\delta_n}{r}\frac{cZ(\delta_n)}{\delta_n} - \frac{\delta_n^2}{r^2}. \tag{18.181}$$

Using $2ab - b^2 \leq a^2$, we get a further upper bound of

$$c^2\left(\frac{Z(\delta_n)}{\delta_n}\right)^2 \leq c^2\left(\frac{\delta_n^2/2 + t\delta_n}{\delta_n}\right)^2 = c^2(\delta_n/2 + t)^2 \tag{18.182}$$

□

18.3.1 Example: linear regression

To get a sense of the behavior of the offset process, consider the linear class $\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathbb{R}^d\}$. First, $\mathcal{F} - f^* = \mathcal{F}$. Second, note that functions are unbounded, and so Rademacher/Gaussian averages are unbounded too. However, offset Gaussian/Rademacher averages are

$$\sup_{w\in\mathbb{R}^d}\sum_{i=1}^n \eta_i\langle w, x_i\rangle - c\langle w, x_i\rangle^2 = \sup_{w\in\mathbb{R}^d}\langle w, \sum_{i=1}^n \eta_i x_i\rangle - c\|w\|_\Sigma^2 \tag{18.183}$$

$$= \frac{1}{4c}\left\|\sum_{i=1}^n \eta_i x_i\right\|_{\Sigma^\dagger}^2 \tag{18.184}$$

where $\Sigma = \sum_{i=1}^n x_i x_i^\mathsf{T}$ and $\Sigma^\dagger$ is the pseudoinverse. Assuming $\mathbb{E}\eta_i^2 \leq 1$,

$$\mathbb{E}\left\|\sum_{i=1}^n \eta_i x_i\right\|_{\Sigma^\dagger}^2 \leq \sum_{i=1}^n x_i^\mathsf{T}\Sigma^\dagger x_i = \mathrm{tr}(\Sigma\Sigma^\dagger) = \mathrm{rank}(\Sigma)$$

We see that, these offset Rademacher/Gaussian averages have the right behavior: we already saw in the first part of the course that the fast rate for linear regression is $O\left(\frac{\text{rank}(\Sigma)}{n}\right)$ without further assumptions.

We can view the negative term that extinguishes the fluctuations of the zero-mean process as coming from the curvature of the square loss. Without the curvature, the negative term is not there and we are left with the usual Rademacher/Gaussian averages.

18.4 Minimax Optimal Rates

It can be shown that minimax optimal rates of estimation (for any estimator) for fixed design are given by the fixed point (see [35])

$$\frac{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \delta_*)}{n} \asymp \delta_*^2 \tag{18.185}$$

For instance, if $\log \mathcal{N}(\mathcal{F}, L^2(P_n), \delta) \asymp \delta^{-p}$, the balance is

$$\delta_*^{-p} n^{-1} \asymp \delta_*^2$$

which gives the rate of $\delta_*^2 = n^{-\frac{2}{2+p}}$.

Do we recover this rate for least squares?

18.5 Rates for Fixed Design Least Squares

We would like to calculate the critical radius δ_n for some function clases of interest. Recall that δ_n is defined as the smallest number such that

$$\mathbb{E} \sup_{g \in \mathcal{F}^*: \|g\|_n \leq \delta} \langle \eta, g \rangle_n \leq \delta^2/2.$$

The strategy is to find upper bounds on the left-hand-side in terms of δ and then solve for the minimal δ. In particular, we know that for any $\alpha \geq 0$,

$$\mathbb{E} \sup_{g \in \mathcal{F}^*: \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \alpha + \frac{1}{\sqrt{n}} \int_{\alpha/4}^{\delta} \sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon)} d\varepsilon \tag{18.186}$$

If the Dudley integral in (18.186) is of the order of the single-scale value (think area under the curve)

$$\delta \times \frac{1}{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \delta)}$$

then an upper bound on the critical radius is obtained by the balance

$$\delta \times \frac{1}{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \delta)} \asymp \delta^2 \tag{18.187}$$

which matches the optimal rate in (18.185). In this case, least squares is an optimal procedure. Below we compute the fixed point under entropy growth conditions.

18.5.1 Nonparametric

Suppose we have

$$\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon) \lesssim \varepsilon^{-p}$$

for $p \in (0, 2)$. Then, taking $\alpha = 0$,

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim n^{-1/2} [\varepsilon^{1-p/2}]_0^\delta = n^{-1/2} \delta^{1-p/2}$$

Setting

$$n^{-1/2} \delta^{1-p/2} = \delta^2$$

yields

$$\delta_n \lesssim n^{-\frac{1}{2+p}}$$

and thus the rate of the least squares estimator is

$$\mathbb{E} \left\| \widehat{f} - f^* \right\|_n^2 \lesssim n^{-\frac{2}{2+p}}$$

Hence, least squares are optimal in this minimax sense for $p \in (0, 2)$.

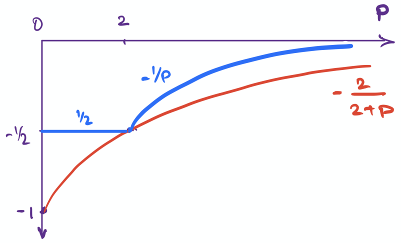


Figure 2: Optimal (in general) rates $n^{-\frac{2}{2+p}}$ (obtained with localization for $p \in (0, 2)$ by ERM) vs without localization (e.g. via global Rademacher averages)

Example: Convex L-Lipschitz functions on a compact domain in $\mathbb{R}^d$:

$$\log \mathcal{N}(\mathcal{F}_{\text{cvx,lip}}, L^2(P_n), \varepsilon) \leq (L/\varepsilon)^{d/2}$$

Example: L-Lipschitz functions on a compact domain in $\mathbb{R}^d$:

$$\log \mathcal{N}(\mathcal{F}_{\text{lip}}, L^2(P_n), \varepsilon) \leq (L/\varepsilon)^{d}$$

18.5.2 Parametric

Consider the parametric case,

$$\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2/\varepsilon)$$

Then

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{d \log(1 + 2/\varepsilon)} d\varepsilon \tag{18.188}$$

Change of variables gives an upper bound

$$\sqrt{\frac{d}{n}}\delta \cdot \int_0^1 \sqrt{\log(1+2/(u\delta))}du \tag{18.189}$$

Unfortunately, this gives a pesky logarithmic factor that should not always be there. For some parametric cases one can, in fact, prove that *local covering numbers* behave as

$$\log \mathcal{N}(\mathcal{F}^* \cap \{g : \|g\|_n \leq \delta\}, L^2(P_n), \varepsilon) \lesssim d \log(1+2\delta/\varepsilon) \tag{18.190}$$

In this case, the change-of-variables leads to

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \sqrt{\frac{d}{n}}\delta \cdot \int_0^1 \sqrt{\log(1+2/\varepsilon)}d\varepsilon \lesssim \sqrt{\frac{d}{n}}\delta \tag{18.191}$$

Equating

$$\delta\sqrt{\frac{d}{n}} \asymp \delta^2$$

yields

$$\delta_n^2 \asymp \frac{d}{n}$$

Note that local covering numbers (18.190) are available in some parametric cases (e.g. when we discretize the parameter space of linear functions) but may not be available for some other classes (e.g. for VC classes, except under additional conditions).

18.6 Remarks

- to bound metric entropy of $\mathcal{F}^* = \mathcal{F} - f^*$, instead consider $\mathcal{F} - \mathcal{F}$. This often leads to only mild increase in a constant. For instance, if $\mathcal{F}$ is a class of L-Lipschitz functions, then $\mathcal{F} - \mathcal{F}$ is a subset of $2L$-Lipschitz functions.

- Note that the rate δ_n^2 depends on local covering numbers (or, local complexity) around f^*. This gives a path to proving adaptivity results (e.g. if f^* is convex but has only k linear pieces, the rate of estimation is parametric because its neighborhood is "simple").

- A simple counting argument (see Yang & Barron 1999, Section 7) shows that for rich enough classes (e.g. nonparametric) worst-case local entropy (worst-case location in the class) and global entropies behave similarly. This implies, in particular, that instead of constructing a local packing for a lower bound (via hypothesis testing), one can instead use global entropy with Fano inequality, justifying the LHS of (18.185) as the lower bound for estimation. See also Mendelson's "local vs global parameters" paper for an in-depth discussion.

19. ORACLE INEQUALITIES

What if we do not assume the regression function f^* is in $\mathcal{F}$? How can we prove an oracle inequality

$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 - \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 \leq \phi(\mathcal{F}, n)$$

Again, we will focus on fixed design.

19.1 Convex $\mathcal{F}$

Suppose $\mathcal{F}$ is convex (or, rather, $\mathcal{F}|_{x_1,\ldots,x_n}$ is convex). Let $\widehat{f}$ be the constrained least squares:

$$\widehat{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-Y_i)^2 = \operatorname*{argmin}_{f\in\mathcal{F}} \|f-Y\|_n^2$$

For the basic inequality we used

$$\left\|\widehat{f}-Y\right\|_n^2 \le \|f^*-Y\|_n^2$$

but in the misspecified case this is no longer true. However, what is true is that

$$\left\|\widehat{f}-Y\right\|_n^2 \le \|f_{\mathcal{F}}-Y\|_n^2$$

Unfortunately, this inequality is not strong enough to get us the desired result. Fortunately, we can do better. Since $\widehat{f}$ is a projection of Y onto $F=\mathcal{F}|_{x_1,\ldots,x_n}$, it holds that

$$\left\|\widehat{f}-Y\right\|_n^2 \le \|f-Y\|_n^2 - \left\|\widehat{f}-f\right\|_n^2 \tag{19.192}$$

for any $f\in\mathcal{F}$, and in particular for $f_{\mathcal{F}}$. This is a simple consequence of convexity and pythagorean theorem. The negative quadratic will give us the extra juice we need.

Adding and subtracting f^* on both sides and expanding,

$$\left\|\widehat{f}-f^*\right\|_n^2+\|f^*-Y\|_n^2+2\langle\widehat{f}-f^*,-\eta\rangle_n \le \|f_{\mathcal{F}}-f^*\|_n^2+\|f^*-Y\|_n^2+2\langle f_{\mathcal{F}}-f^*,-\eta\rangle_n-\left\|f_{\mathcal{F}}-\widehat{f}\right\|_n^2$$

which leads to

$$\begin{aligned}\left\|\widehat{f}-f^*\right\|_n^2-\|f_{\mathcal{F}}-f^*\|_n^2 &\le 2\langle\eta,\widehat{f}-f_{\mathcal{F}}\rangle_n-\left\|\widehat{f}-f_{\mathcal{F}}\right\|_n^2 &(19.193)\\ &\le \sup_{h\in\mathcal{F}-f_{\mathcal{F}}} 2\langle\eta,h\rangle_n-\|h\|_n^2 &(19.194)\end{aligned}$$

We conclude that for convex $\mathcal{F}$ and fixed design, the *upper bounds* we find for well-specified and misspecified cases match. Moreover, since the misspecified case is strictly more general, and since *lower bounds* for the well-specified case and polynomial entropy growth (in the $p<2$ regime) match the upper bounds, we conclude that constrained least squares are also minimax optimal for fixed design misspecified case.

Note: a crucial observation is that offset complexity would arise even if (19.192) had a different constant multiplier in front of $-\left\|f-\widehat{f}\right\|_n^2$. We will exploit this observation in a bit.

19.2 General $\mathcal{F}$

What if $\mathcal{F}$ is not convex? It turns out that least squares (ERM) can be suboptimal even if $\mathcal{F}$ is a finite class!

19.2.1 A lower bound for ERM (or any proper procedure)

The suboptimality can be illustrated on a very simple example. Suppose $\mathcal{X} = \{x\}$, Y is $\{0,1\}$-valued, and $\mathcal{F} = \{f_0, f_1\}$ such that $f_0(x) = 0$ and $f_1(x) = 1$. The marginal distribution is the trivial $P_X = \delta_x$ and suppose we have two conditional distributions $P_0(Y = 1) = 1/2 - \alpha$ and $P_1(Y = 1) = 1/2 + \alpha$. Clearly, the population minimizer for P_j is f_j. Also, under P_0 the regression function is $f_0^* = 1/2 - \alpha$ while under P_1 it is $f_1^* = 1/2 + \alpha$. Finally, ERM is a method that goes after the most frequent observation in the data $Y_1, \ldots, Y_n$.

However, if $\alpha \propto 1/\sqrt{n}$, there is a constant probability of error in determining whether P_0 or P_1 generated the data. Note that the oracle risk is $\min_{f \in \{f_0, f_1\}} \|f - f_i^*\|^2 = (1/2-\alpha)^2$ while the risk of the estimator $p(1/2+\alpha)^2 + (1-p)(1/2-\alpha)^2$ where p is the probability of making a mistake and not selecting f_i under the distribution P_i. Hence, the overall comparison to the oracle is at least $p((1/2+\alpha)^2 - (1/2-\alpha)^2) = \Omega(\alpha)$ when p is constant.

Hence, ERM (or any "proper" method that selects from $\mathcal{F}$) cannot achieve excess loss smaller than $\Omega(n^{-1/2})$:

$$\max_{P_i \in \{P_0, P_1\}} \left\{ \mathbb{E} \left\| \widehat{f} - f_i^* \right\|^2 - \min_{f \in \{f_0, f_1\}} \|f - f_i^*\|^2 \right\} = \Omega(n^{-1/2})$$

Yet, an improper method that selects $\widehat{f}$ outside $\mathcal{F}$ can achieve an $O(n^{-1})$ rate.

A similar simple lower bound can be constructed for ERM with random design.[3]

19.2.2 How about ERM over Convex Hull?

Given that the procedure has to be "improper" (select from outside of $\mathcal{F}$), one can hypothesize that doing ERM over $\mathrm{conv}(\mathcal{F})$ may work. Interestingly, this procedure is also rate-suboptimal for a finite $\mathcal{F}$ since $\mathrm{conv}(\mathcal{F})$ is too expressive.[4]

19.2.3 An improper procedure

Somewhat surprisingly, only a small modification of ERM is required to make it optimal for general classes. Consider the following two-step procedure[5] (*Star Estimator*):

$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \|f - Y\|_n^2 \tag{19.195}$$

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathrm{star}(\mathcal{F}, \widehat{g})} \|f - Y\|_n^2 \tag{19.196}$$

where

$$\mathrm{star}(\mathcal{F}, g) = \{\alpha f + (1-\alpha) g : f \in \mathcal{F}, \alpha \in [0,1]\}.$$

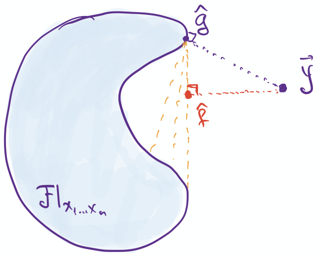
Note that $\widehat{f}$ need not be in $\mathcal{F}$ but is an average of two elements of $\mathcal{F}$.

Note: the method is, in general, different from single ERM over a convex hull of $\mathcal{F}$, and so it is not clear that a version of (19.192) holds [21]:

[3] For more detailed discussion, we refer to [20].

[4] Proof can be found in Lecué & Mendelson

[5] For a finite class, the above estimator was analyzed by J-Y. Audibert [1].



Lemma 34. *For any $f \in \mathcal{F}$,*

$$\|f - Y\|_n^2 - \left\|\widehat{f} - Y\right\|_n^2 \geq \frac{1}{18}\left\|\widehat{f} - f\right\|_n^2. \tag{19.197}$$

The above inequality is an approximate version of (19.192), a generalization of the pythagorean relationship for convex sets.

As a consequence,

$$\left\|\widehat{f} - f^*\right\|_n^2 - \|f_{\mathcal{F}} - f^*\|_n^2 \leq 2\langle \eta, \widehat{f} - f_{\mathcal{F}}\rangle_n - \frac{1}{18}\left\|f_{\mathcal{F}} - \widehat{f}\right\|_n^2$$

and the same upper bounds hold as in the convex case, up to constants. The difference is that the supremum is now in $\mathrm{star}(\mathcal{F}, \widehat{g}) \subseteq \mathcal{F} - f^* + \mathrm{star}(\mathcal{F} - \mathcal{F})$ which is not significantly larger than $\mathcal{F}$ in terms of entropy (unless $\mathcal{F}$ is finite, which can be handled separately).

Remarks:

1. if the set is convex, $\widehat{f} = \widehat{g}$.
2. the Star Estimator can be viewed as one step of Frank-Wolfe. More steps can improve the constant.

Exercise: for any $\alpha > 0$ and a set $F \subset \mathbb{R}^n$, the covering numbers satisfy

$$\log \mathcal{N}(F, \|\cdot\|, 2\alpha) \leq \log \mathcal{N}(\mathrm{star}(F, 0), \|\cdot\|, 2\alpha) \leq \log(\mathrm{diam}(F)/\alpha) + \log \mathcal{N}(F, \|\cdot\|, \alpha)$$

19.3 Offset Rademacher averages

We have seen that offset Gaussian or Rademacher averages are a convenient way of proving rates of convergence for least squares in well-specified, as well as misspecified (both convex and general function class) cases. We finish this section by estimating offset complexities via covering numbers. We start with a finite set.

For a set $V \subset \mathbb{R}^n$, the offset process indexed by V is defined as a stochastic process

$$v \mapsto \sum_{i=1}^{n} \varepsilon_i v_i - c v_i^2 = \langle \varepsilon, v \rangle - c\|v\|^2.$$

Here ε_i are independent Rademacher, but the same results hold for any sub-Gaussian random variables.

Lemma 35. *Let $V \subset \mathbb{R}^n$ be a finite set of vectors,* $\mathrm{card}(V) = N$*. Then for any $c > 0$,*

$$\mathbb{E}_\varepsilon \max_{v \in V} \langle \varepsilon, v \rangle - c \|v\|^2 \leq \frac{\log N}{2c}.$$

Furthermore,

$$\mathbb{P}\left(\max_{v \in V} \langle \varepsilon, v \rangle - c \|v\|^2 \geq \frac{1}{2c}(\log N + \log(1/\delta)) \right) \leq \delta$$

Proof. Assuming the random variables are 1-subGaussian,

$$\begin{aligned}
\mathbb{E} \max_{v \in V} \langle \varepsilon, v \rangle - c \|v\|^2 &= \frac{1}{\lambda} \mathbb{E} \log \exp \max \lambda \langle \varepsilon, v \rangle - \lambda c \|v\|^2 \\
&\leq \frac{1}{\lambda} \log \sum_{v \in V} \mathbb{E} \exp\{\lambda \langle \varepsilon, v \rangle - \lambda c \|v\|^2\} \\
&\leq \frac{1}{\lambda} \log \left(\sum_{v \in V} \exp\{\lambda^2 \|v\|^2 / 2 - \lambda c \|v\|^2\} \right) \\
&= \frac{1}{2c} \log N
\end{aligned}$$

where we chose $\lambda = 2c$. □

Finally, we state a general result.

Theorem 7. *Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$. Then for any $x_1, \ldots, x_n \in \mathcal{X}$ and the corresponding empirical measure P_n,*

$$\begin{aligned}
&\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) - c f(x_i)^2 \qquad (19.198) \\
&\leq \inf_{\gamma \geq 0, \alpha \in [0, \gamma]} \left\{ \frac{(2/c) \log \mathcal{N}(\mathcal{F}, L^2(P_n), \gamma)}{n} + 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \delta)} d\delta \right\}
\end{aligned}$$

We have shown in Lemma 33 that the offset Gaussian process cannot be more than a constant multiple of the critical radius. But what if the bound of the above theorem is too loose to be useful? To see that it attains the optimal balance of (18.185) in some cases, consider the situation where, as in the discussion preceding (18.187), the Dudley entropy integral is of the order of the single scale estimate

$$\gamma \times \frac{1}{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \gamma)}.$$

In this case, the optimal balance in (19.198) is

$$\gamma \times \frac{1}{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \gamma)} \asymp \frac{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \gamma)}{n}. \qquad (19.199)$$

Dividing and squaring, we do recover (18.185), an optimal rate. We conclude that the upper bound of Theorem 7 recovers optimal rates for regression in the $p \in (0, 2)$ regime and, more generally, under the above-stated condition on the Dudley integral.

20. TALAGRAND'S INEQUALITY AND APPLICATIONS

For the last half of the course, we have only considered the expected suprema of empirical, Rademacher, or Gaussian processes. We mentioned that high-probability statements follow from different arguments. In this lecture, we provide the tools to study deviations of random suprema above (or below) their expected values.

The following version of Talagrand's inequality is due to Bousquet:

Theorem 8. *Let $X_1, \dots, X_n$ be i.i.d., and let $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$. Suppose*

$$\sup_{f \in \mathcal{F}} \mathsf{var}(f(X)) \le \sigma^2$$

for some $\sigma > 0$. Let either

$$Z = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(X_i) \right\} \quad or \quad Z = \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f \right\}$$

and set $v = \sigma^2 + 2\mathbb{E}Z$. Then for any $t \ge 0$,

$$\mathbb{P}\left(Z \ge \mathbb{E}Z + \sqrt{\frac{2tv}{n}} + \frac{t}{3n} \right) \le e^{-t}.$$

Consider a particular case of a singleton $\mathcal{F} = \{f\}$. Then $Z = \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f$ (or the other form), $v = \sigma^2 = \mathsf{var}(f(X))$ because $\mathbb{E}Z = 0$. Then Theorem 8 says that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f \ge \sigma\sqrt{\frac{2t}{n}} + \frac{t}{3n} \right) \le e^{-t}$$

which is Bernstein's inequality. Moreover, the constants match those in Bernstein's inequality, which is remarkable.

Now, recall the definition of empirical Rademacher averages. In this lecture we will scale these averages by $1/n$:

$$\widehat{\mathcal{R}}(\mathcal{F}) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \;\middle|\; X_1, \dots, X_n \right].$$

We have $\widehat{\mathcal{R}}(\mathcal{F}) \ge 0$ by Jensen's inequality. Moreover, this function satisfies a self-bounding property [5, Ch. 6], which implies the following

Theorem 9. *Let $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$. Then for any $t > 0$,*

$$\mathbb{P}\left(\widehat{\mathcal{R}}(\mathcal{F}) \ge \mathbb{E}\widehat{\mathcal{R}}(\mathcal{F}) + \sqrt{\frac{2t\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F})}{n}} + \frac{t}{3n} \right) \le e^{-t}$$

and

$$\mathbb{P}\left(\widehat{\mathcal{R}}(\mathcal{F}) \leq \mathbb{E}\widehat{\mathcal{R}}(\mathcal{F}) - \sqrt{\frac{2t\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F})}{n}}\right) \leq e^{-t}$$

This first statement has a similar form to that of Theorem 8 (after normalizing Z by n) with v replaced by $\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F})$, a consequence of the self-bounding property of $\widehat{\mathcal{R}}(\mathcal{F})$.

In particular, by using the inequality

$$\forall x, y, \lambda > 0, \qquad \sqrt{xy} \leq \frac{\lambda}{2}x + \frac{1}{2\lambda}y,$$

we have

$$\mathbb{P}\left(\widehat{\mathcal{R}}(\mathcal{F}) \geq 2\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F}) + \frac{5t}{6n}\right) \leq e^{-t}$$

and

$$\mathbb{P}\left(\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F}) \geq 2\widehat{\mathcal{R}}(\mathcal{F}) + \frac{2t}{n}\right) \leq e^{-t}.$$

Finally, recall that symmetrization lemma states that for the supremum of the empirical process Z in Theorem 8,

$$\mathbb{E}Z \leq 2\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F}).$$

Together with Theorem 8 and Theorem 9, this yields (see e.g. [3, Thm 2.1])

Theorem 10. *Let $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$. Let $\sup_{f\in\mathcal{F}} \mathsf{var}(f(X)) \leq \sigma^2$. Then for any $t > 0$, with probability at least $1 - 2e^{-t}$, for any $f \in \mathcal{F}$,*

$$\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1} f(X_i) \leq 6\widehat{\mathcal{R}}(\mathcal{F}) + \sigma\sqrt{\frac{2t}{n}} + \frac{11t}{n} \tag{20.200}$$

A few remarks. First, the constants here can be balanced differently (see [3, Thm 2.1]). Second, the same result holds with $\frac{1}{n}\sum_{i=1} f(X_i) - \mathbb{E}f(X)$ on the left-hand-side of (20.200). Third, Theorem 8 can be replaced with McDiarmid's inequality if one does not aim to take advantage of small variance σ; however, some of the key results on fast rates in learning theory do take advantage of this Bernstein-style bound.

Theorem 10 can be applied in a variety of situations. To start, since $\widehat{\mathcal{R}}$ does not depend on the unknown distribution of X, all the terms (except for σ) on the right-hand-side of (20.200) can be computed from the data. While the supremum of the empirical process on the left-hand-side of (20.200) cannot be computed in general since the distribution of X is not known, the expression provides a data-dependent estimate of this quantity.

In the setting of prediction and model selection, we could consider a model $\mathcal{G}_\lambda$ such that $\mathcal{G}_\lambda \subseteq \mathcal{G}_{\lambda'}$ for $\lambda \leq \lambda'$, i.e. λ is a tunable parameter that controls complexity of the model (e.g. width of a neural network). Inequality (20.200) can then be viewed as an upper bound on the expected loss of any function in $\mathcal{G}_\lambda$ in terms of its empirical fit to data plus a penalty term for model complexity, as given by the Rademacher averages. Moreover, this penalty is data-driven.

The above theorems are also at the heart of proving localization results for random design, both in the well-specified and misspecified settings. Let us only mention one consequence (see [3, Thm 4.1]).

Theorem 11. *Let $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$. Suppose for every $f \in \mathcal{F}$, it holds that $\mathbb{E}f(X)^2 \leq B\mathbb{E}f(X)$. Then with probability at least $1 - 3e^{-t}$, for all $f \in \mathcal{F}$,*

$$\mathbb{E}f(X) \leq 2\frac{1}{n}\sum_{i=1}^{n} f(X_i) + c\delta_n^2 + \frac{c't}{n} \tag{20.201}$$

where δ_n be the critical radius[a] of $\textsf{star}(\mathcal{F}, 0)$ and c, c' are constants that depend on B.

[a] To be precise, [3, Thm 4.1] includes a confidence term t/n in the computation of the critical radius.

20.1 Application: Learning and Low-Noise

Consider the setting of statistical learning with a class $\mathcal{F}$ and a 1-Lipschitz loss function $\boldsymbol{\ell}$. Let $\widehat{f}$ be a minimizer of empirical risk $\widehat{\mathbf{L}}(f)$ as defined in (10.127). Let $f_{\mathcal{F}}$ be a minimizer of expected risk $\mathbf{L}(f)$ over $\mathcal{F}$. We apply Theorem 11 to the class $\boldsymbol{\ell} \circ \mathcal{F} - \boldsymbol{\ell} \circ f_{\mathcal{F}}$. Under the high-probability event of the Theorem, the inequality holds for all functions, so we can apply it to $\widehat{f} \in \mathcal{F}$. Since $\widehat{\mathbf{L}}(\widehat{f}) - \widehat{\mathbf{L}}(f_{\mathcal{F}}) \leq 0$, we have that, with probability at least $1 - 3e^{-t}$,

$$\mathbf{L}(\widehat{f}) - \mathbf{L}(f_{\mathcal{F}}) \leq c\delta_n^2 + \frac{c't}{n}. \tag{20.202}$$

This conclusion holds under the assumption

$$\mathbb{E}(f(X) - f^*(X))^2 \leq B(\mathbf{L}(f) - \mathbf{L}(f_{\mathcal{F}})),$$

which, together with the Lipschitz condition on the loss implies the so-called Bernstein condition

$$\mathbb{E}(\boldsymbol{\ell} \circ f - \boldsymbol{\ell} \circ f_{\mathcal{F}})^2 \leq B\mathbb{E}(\boldsymbol{\ell} \circ f - \boldsymbol{\ell} \circ f_{\mathcal{F}}) = B(\mathbf{L}(f) - \mathbf{L}(f_{\mathcal{F}}))$$

Such a condition (or closely-related variants) are implied by, for instance, convexity of $\mathcal{F}$ and uniform convexity of $\boldsymbol{\ell}$, or by low-noise assumptions in classification settings. For the case of square loss, (20.202) implies a random design oracle inequality in the misspecified case. We now provide more details for the case of well-specified random design regression and develop general tools for passing from fixed to random design.

21. FROM FIXED TO RANDOM DESIGN

Recall that in fixed design regression we aim to prove that for a given set of points $x_1, \ldots, x_n$, an estimator (such as constrained least squares) attains

$$\left\|\widehat{f} - f^*\right\|_{L^2(P_n)}^2 \leq \cdots$$

where on the right-hand side we have either a quantity that goes to zero with n or oracle risk as in the misspecified case. We would like to analyze random design regression where $X_1, \ldots, X_n$ are i.i.d from P. Importantly, we also measure the risk through the $L^2(P)$ norm. However,

$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_{L^2(P_n)}^2 \neq \mathbb{E}\left\|\widehat{f} - f^*\right\|_{L^2(P)}^2$$

since the algorithm $\widehat{f}$ depends on $X_1, \ldots, X_n$, and so lifting the results from the fixed design case is not straightforward.

Imagine, however, we could prove that with high probability, for all functions $f \in \mathcal{F}$,

$$\|f - f^*\|_{L^2(P)}^2 \leq 2 \|f - f^*\|_{L^2(P_n)}^2 + \psi(n, \mathcal{F}). \tag{21.203}$$

In that case, a guarantee for fixed-design regression *would* translate into a guarantee for random design regression as long as $\widehat{f} \in \mathcal{F}$ (for the Star Algorithm, just enlarge $\mathcal{F}$ appropriately). Furthermore, as long as $\psi(n, \mathcal{F})$ decays with n at least as fast as the rate of fixed design regression, we would be able to conclude that random design is not harder than fixed design. Let's see if this can be shown.

21.1 Uniformly Bounded Functions

Our plan of action for proving results of the form (21.203) is to view the inequality as an instance of a more general uniform comparison

$$\forall g \in \mathcal{G}, \quad \mathbb{E} g(X) \leq \frac{2}{n} \sum_{i=1}^{n} g(X_i) + \psi(n, \mathcal{G})$$

for a class $\mathcal{G}$ of *nonnegative* functions. In this part of the lecture, we sketch analysis for uniformly bounded functions. This requirement is necessitated by the use of Theorems 8 and 10.

Let $\bar{\delta}$ be such that for all $\delta \geq \bar{\delta}$,

$$\mathbb{E}_{\varepsilon} \sup_{g \in \mathcal{G}: \frac{1}{n}\sum_{i=1}^{n} g(X_i) \leq \delta^2} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(X_i) \leq \delta^2/2 \tag{21.204}$$

conditionally on $X_1, \ldots, X_n$.

Alternatively, we can write (21.204) as

$$\mathbb{E}_{\varepsilon} \sup_{g \in \mathcal{G}: \|\sqrt{g}\|_n \leq \delta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(X_i) \leq \delta^2/2 \tag{21.205}$$

The following result can be proved using Theorem 10 (see [6, Theorem 6.1]):

Lemma 36. *Let $\mathcal{G}$ be a class of functions with values in $[0,1]$. Then with probability at least $1 - e^{-t}$ for all $g \in \mathcal{G}$*

$$\mathbb{E} g(X) \leq \frac{2}{n} \sum_{i=1}^{n} g(X_i) + c \cdot \bar{\delta}^2 + \frac{c' \cdot (t + \log\log n)}{n} \tag{21.206}$$

where $\bar{\delta} = \bar{\delta}(\mathcal{G})$ is any upper bound on the fixed point in (21.204).

Applying this inequality for the class $\mathcal{G} = \{(f - f')^2 : f, f' \in \mathcal{F}\}$, assuming $\mathcal{F}$ is a class of $[0,1]$-valued functions, yields

$$\|f - f'\|_{L^2(P)}^2 \leq 2 \|f - f'\|_{L^2(P_n)}^2 + c \cdot \bar{\delta}^2 + \frac{c' \cdot (t + \log\log n)}{n}. \tag{21.207}$$

A few remarks. First, $\mathcal{G} = (\mathcal{F} - \mathcal{F})^2$ can be replaced by $(\mathcal{F} - f^*)^2$, even if $f^* \notin \mathcal{F}$, as long as the resulting class is uniformly bounded. Second, we observe that (21.204) is defined with a localization restriction $\frac{1}{n}\sum_{i=1}^n g(X_i) \leq \delta^2$ rather than $\frac{1}{n}\sum_{i=1}^n g(X_i)^2 \leq \delta^2$ in the previous lecture. Since functions are bounded by 1, the set

$$\widehat{\mathcal{M}} := \left\{ g : \frac{1}{n}\sum_{i=1}^n g(X_i) \leq \delta^2 \right\} \subseteq \{\|g\|_n^2 \leq \delta^2\}$$

and hence the set in (21.204) is smaller. Thus the fixed point (21.204) is potentially smaller than the one defined in the previous lecture.

21.1.1 Evaluating the new critical radius

Now, one can ask how to compute a suitable upper bound on the critical radius in (21.204) for particular classes of interest. As in the earlier lectures, the strategy is to upper bound the left-hand side of (21.204) in terms of some more tangible measures of complexity and δ, and then balance with $\delta^2/2$.

In particular, we are interested in the case when $\mathcal{G} = \mathcal{F}^2$ (same analysis works for $(\mathcal{F} - \mathcal{F})^2$ or $(\mathcal{F} - f^*)^2$) for some class $\mathcal{F}$ of $[-1, 1]$-valued functions. In this case, it is tempting to proceed with the help of contraction inequality and upper bound

$$\mathbb{E}_\varepsilon \sup_{g \in \mathcal{F}^2 \cap \widehat{\mathcal{M}}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i) \leq 2\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}: \|f\|_n^2 \leq \delta^2} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \tag{21.208}$$

since square is 2-Lipschitz on $[-1, 1]$. Balancing this with δ^2 gives, up to constants, precisely the critical radius of $\mathcal{F}$, as in Definition 21 (modulo the use of Gaussian vs Rademacher random variables). Interestingly, one can significantly improve upon this argument and show that the localization radius for $\mathcal{F}^2$ with the left-hand-side of (21.208) can be smaller than that of $\mathcal{F}$. In particular, a useful result is the following:

Lemma 37. *For any class $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$ of bounded functions, the critical radius in* (21.204) *for the class $\mathcal{G} = \mathcal{F}^2$ can be upper bounded by a solution to*

$$\frac{12}{\sqrt{n}} \int_{\delta/16}^{1} \sqrt{\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), u/2))} du \leq \delta/4. \tag{21.209}$$

Proof. We start upper bounding the left-hand side of (21.204). Observe that functions in $\mathcal{G}$ are nonnegative and bounded uniformly in $[0, 1]$. As discussed earlier, the restriction $\frac{1}{n}\sum_{i=1}^n g(X_i) \leq \delta^2$ implies $\|g\|_n \leq \delta$, and hence the left-hand-side of (21.204) is upper bounded by

$$\inf_\alpha \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\delta \sqrt{\log \mathcal{N}(\mathcal{G} \cap \widehat{\mathcal{M}}, L^2(P_n), \varepsilon)} d\varepsilon \right\}. \tag{21.210}$$

Let $V = \{\tilde{f}_1, \ldots, \tilde{f}_N\}$ be a proper $L^\infty(P_n)$-cover of $\mathcal{F} \cap \{\|f\|_n \leq \delta\}$ at scale $\tau \leq \delta$ (proper implies $\left\|\tilde{f}\right\|_n \leq \delta$). Fix any $g = f^2 \in \mathcal{G} \cap \widehat{\mathcal{M}}$. Let $\tilde{f}$ be an element of V that is τ-close to

f. Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(f(x_i)^2-\tilde{f}(x_i)^2)^2 &= \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\tilde{f}(x_i))^2(f(x_i)+\tilde{f}(x_i))^2 \\ &\le \max_i(f(x_i)-\tilde{f}(x_i))^2\cdot\frac{1}{n}\sum_{i=1}^{n}(f(x_i)+\tilde{f}(x_i))^2 \\ &\le \tau^2(2\|f\|_n^2+2\left\|\tilde{f}\right\|_n^2) \\ &\le 4\tau^2\delta^2 := \varepsilon^2\end{aligned}$$

We conclude that

$$\begin{aligned}\mathcal{N}(\mathcal{G}\cap\widehat{\mathcal{M}},L^2(P_n),\varepsilon) &\le \mathcal{N}(\mathcal{F}\cap\{\|f\|_n\le\delta\},L^\infty(P_n),\varepsilon/(2\delta)) \\ &\le \mathcal{N}(\mathcal{F},L^\infty(P_n),\varepsilon/(2\delta))\end{aligned}$$

Substituting into (21.210), the upper bound on the right-hand side becomes

$$\begin{aligned}&\inf_{\alpha\ge 0}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^\delta\sqrt{\log\mathcal{N}(\mathcal{F},L^\infty(P_n),\varepsilon/(2\delta))}d\varepsilon\right\} \\ &\le \delta^2/4+\delta\times\frac{12}{\sqrt{n}}\int_{\delta/16}^{1}\sqrt{\log\mathcal{N}(\mathcal{F},L^\infty(P_n),u/2))}du\end{aligned}$$

where we performed change-of-variables $u=\varepsilon/\delta$ and chose $\alpha=\delta^2/16$. Using this in (21.204) and balancing with $\delta^2/2$ yields (21.209). □

A key outcome of the above lemma is that the critical radius of $\mathcal{G}=\mathcal{F}^2$ (or $(\mathcal{F}-\mathcal{F})^2$) given by (21.204) is much smaller than that of $\mathcal{F}$. Note that whenever the Dudley integral in (21.209) converges with $\delta=0$, the solution is $\delta\propto n^{-1/2}$ (up to $\log n$ factors) and hence the remainder in (21.207) is of the order $1/n$, a smaller order term as compared to the rate of estimation for fixed design. The fact that the remainder term is or a lower order can be shown, for instance, more generally under the polynomial growth of entropy, or in the parametric cases. For instance, for

$$\mathcal{N}(\mathcal{F},L^\infty(P_n),\varepsilon)\le\left(\frac{cn}{\varepsilon}\right)^d,$$

the localization radius of $\mathcal{G}=\mathcal{F}^2$ can be upper bounded as

$$\bar{\delta}(\mathcal{G})=C\sqrt{\frac{d}{n}\log\left(\frac{cn}{d}\right)}$$

and for a finite class we immediately have

$$\bar{\delta}(\mathcal{G})\le C\sqrt{\frac{\log|\mathcal{F}|}{n}}.$$

We can also prove a general and useful result, albeit with extra log factors (due to its generality). Following [31], we have

Lemma 38. *For any class* $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$*, the critical radius in* (21.209) *is at most*

$$\bar{\delta}(\mathcal{F}^2) \leq C \log^2 n \cdot \bar{\mathcal{R}}(\mathcal{F}),$$

where

$$\bar{\mathcal{R}}(\mathcal{F}) = \sup_{x_1,\ldots,x_n} \widehat{\mathcal{R}}(\mathcal{F}).$$

Proof. Substitute the following estimate for L^∞ covering numbers in terms of the scale-sensitive dimension (see e.g. [28]):

$$\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), \alpha) \leq 2\text{vc}(\mathcal{F}, c\alpha) \cdot \log n \cdot \left(\frac{cn}{\text{vc}(\mathcal{F}, c\alpha)\cdot \alpha}\right) \tag{21.211}$$

and then use the following fact: for any $\alpha > \bar{\mathcal{R}}(\mathcal{F})$,

$$\text{vc}(\mathcal{F}, \alpha) \leq \frac{4n\bar{\mathcal{R}}(\mathcal{F})^2}{\alpha^2}. \tag{21.212}$$

This last inequality can be written in the more familiar form

$$\sup_{\alpha > \bar{\mathcal{R}}(\mathcal{F})} \alpha \sqrt{\frac{\text{vc}(\mathcal{F}, \alpha)}{4n}} \leq \bar{\mathcal{R}}(\mathcal{F}), \tag{21.213}$$

which bears similarity to Sudakov's minoration. This inequality is proved by taking the α-shattered set, replicating it $\lceil n/\text{vc}(\mathcal{F}, \alpha)\rceil$ times, and using our previous argument about Rademacher averages being large when there is a cube inside the set. We leave it as an exercise.

Back to the estimate, we have

$$\frac{1}{\sqrt{n}} \int_{\delta/64}^{1/4} \sqrt{\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), \alpha))} d\alpha \lesssim \frac{\sqrt{\log n}}{\sqrt{n}} \int_{\delta/64}^{1/4} \sqrt{\text{vc}(\mathcal{F}, c\alpha) \log\left(\frac{cn}{\alpha}\right)} d\alpha \tag{21.214}$$

$$\lesssim \sqrt{\log n} \bar{\mathcal{R}}(\mathcal{F}) \int_{\delta/64}^{1/4} \frac{1}{\alpha} \sqrt{\log\left(\frac{cn}{\alpha}\right)} d\alpha \tag{21.215}$$

To finish the proof, choose $\delta = 64\bar{\mathcal{R}}(\mathcal{F})$ and observe that

$$\int_{\bar{\mathcal{R}}(\mathcal{F})}^{1} \frac{1}{\alpha}\sqrt{\log\left(\frac{cn}{\alpha}\right)} d\alpha \lesssim \log^2(cn/\bar{\mathcal{R}}(\mathcal{F})).$$

□

Hence, for $\mathcal{G} = \mathcal{F}^2$, ignoring logarithmic factors, $\bar{\delta}(\mathcal{G}) \leq \tilde{O}(n^{-1})$ when $\bar{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/2}$ and $\bar{\delta}(\mathcal{G}) \leq \tilde{O}(n^{-2/p})$ when $\bar{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/p}$, which is *smaller* than the rate of estimation for least squares, ignoring logarithmic factors.

We conclude that rates of estimation for fixed design translate into rates for estimation with random design, at least for bounded functions. It is worth emphasizing that the extra factors one gains from comparing $\|f - f^*\|^2_{L^2(P)}$ to $2\,\|f - f^*\|^2_{L^2(P_n)}$ are typically of smaller order than what one gets from denoising for fixed design. The next section provides further motivation for why this happens, and presents an approach that does not rely on uniform boundedness of functions.

21.2 Beyond boundedness: the small-ball method

This approach was pioneered by [18] and then developed by Mendelson in a series of papers starting with [24]. Importantly, this approach does not rely on uniform boundedness of functions as in the application of Talagrand's inequality.

Roughly speaking, the realization is that whenever the population norm $\|f\|_{L^2(P)}$ is large enough, it is highly unlikely that the random empirical norm $\|f\|_{L^2(P_n)}$ can be smaller than a fraction of the population norm. Moreover, conditions for such a statement to be true are rather weak and do not require uniform boundedness.

We first recall the Paley-Zygmund inequality (1932) stating that for a nonnegative random variable Z with finite variance,

$$\mathbb{P}\left(Z \geq t\mathbb{E}Z\right) \geq (1-t)^2 \frac{(\mathbb{E}Z)^2}{\mathbb{E}Z^2}$$

for any $0 \leq t \leq 1$.

Let us use the following shorthand. We will write $\|f\|_2 = \|f\|_{L^2(P)} = (\mathbb{E}f(X)^2)^{1/2}$ and $\|f\|_4 = \|f\|_{L^4(P)} = (\mathbb{E}f(X)^4)^{1/4}$. Then

$$\mathbb{P}\left(|f(X)| \geq t\|f\|_2\right) = \mathbb{P}\left(f(X)^2 \geq t^2\|f\|_2^2\right) \geq (1-t^2)^2 \frac{\|f\|_2^4}{\|f\|_4^4}$$

Now, we make an assumption that for every $f \in \mathcal{F}$,

$$\mathbb{E}f(X)^4 \leq c(\mathbb{E}f(X)^2)^2$$

for some c.

Under this $L^4 - L^2$ norm comparison, it holds that

$$\mathbb{P}\left(|f(X)| \geq t\|f\|_2\right) \geq (1-t^2)^2 c,$$

an "anti-concentration" inequality. More generally, the condition that there exists c and c' such that for all $f \in \mathcal{F}$,

$$\mathbb{P}\left(|f(X)| \geq c\|f\|_2\right) \geq c' \tag{21.216}$$

is called the small-ball property.

Lemma 39. *Assume* (21.216). *Let $\mathcal{F}$ be star-shaped around* 0. *Then with probability at least $e^{-c_1 n}$,*

$$\inf_{f\in\mathcal{F}:\|f\|_2\geq\tilde{\delta}} \frac{\|f\|_n}{\|f\|_2} \geq c_2$$

for some constants c_1, c_2, i.e. for all $f \in \mathcal{F}$,

$$\|f\|_2^2 \lesssim \|f\|_n^2 + \tilde{\delta}^2,$$

where $\tilde{\delta}$ is a critical radius defined as the smallest δ such that

$$\mathbb{E} \sup_{f\in\mathcal{F},\|f\|_2\leq\delta} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \leq c_3\delta. \tag{21.217}$$

Proof. Let's see how we can compare the empirical and population norms, uniformly over $\mathcal{F}$, given such a condition. First, let's consider any function with norm $\|f\|_2 = 1$. Observe that if we could show with high probability

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{|f(X_i)| \geq c_1\right\} \geq c_2 \tag{21.218}$$

for some constants c_1, c_2, we would be done since such a lower bound implies a constant lower bound on $\frac{1}{n}\sum_{i=1}^{n} f(X_i)^2 \geq c\,\|f\|_2^2 = c$). By rescaling and assuming star-shapedness, we would extend the result to all functions in $\mathcal{F}$ (above some critical level for which we can prove (21.218)).

For a given $c > 0$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{|f(X_i)| \geq c\right\} &= \mathbb{E}\mathbf{1}\left\{|f(X)| \geq 2c\right\} - \left(\mathbb{E}\mathbf{1}\left\{|f(X)| \geq 2c\right\} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{|f(X_i)| \geq c\right\}\right)\\ &\geq \mathbb{E}\mathbf{1}\left\{|f(X)| \geq 2c\right\} - \left(\mathbb{E}\phi(|f(X)|) - \frac{1}{n}\sum_{i=1}^{n} \phi(|f(X_i)|)\right)\end{aligned}$$

for $\phi(u) = 0$ on $(-\infty, c]$, $\phi(u) = u/c - 1$ on $[c, 2c]$, and $\phi(u) = 1$ on $[2c, \infty)$.

$$\geq \inf_{f\in\mathcal{F}} \mathbb{P}\left(|f(X)| \geq 2c\,\|f\|_2\right) - \sup_{f\in\mathcal{F},\|f\|_2=1} \left(\mathbb{E}\phi(|f|) - \frac{1}{n}\sum_{i=1}^{n} \phi(|f(X_i)|)\right)$$

Now, using concentration (since $\phi(|f|)$ are in $[0, 1]$), the random supremum

$$\sup_{f\in\mathcal{F},\|f\|_2=1} \left(\mathbb{E}\phi(|f|) - \frac{1}{n}\sum_{i=1}^{n} \phi(|f(X_i)|)\right)$$

can be upper bounded with probability at least $1 - e^{-2u^2}$ by its expectation

$$\mathbb{E} \sup_{f\in\mathcal{F},\|f\|_2=1} \left(\mathbb{E}\phi(|f|) - \frac{1}{n}\sum_{i=1}^{n} \phi(|f(X_i)|)\right) + \frac{u}{\sqrt{n}}$$

which, in turn, can be upper bounded via symmetrization and contraction inequality (since ϕ is $1/c$-Lipschitz) by

$$\frac{4}{c}\mathbb{E} \sup_{f\in\mathcal{F},\|f\|_2=1} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i) + \frac{u}{\sqrt{n}}$$

By choosing $u = \sqrt{n}\cdot c''$, we can make the additive term an arbitrarily small constant c''. Now, we see that (21.218) will hold with a non-zero constant c_2 as long as

$$\mathbb{E} \sup_{f\in\mathcal{F},\|f\|_2=1} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i) \leq c''$$

for an appropriately small constant c''. We now need to extend this control to all $\|f\|_2$ above some critical radius. Assuming that $\mathcal{F}$ is star-shaped around 0, the control extends for all f such that $\|f\| \geq \tilde{\delta}$. □

Observe that $\tilde{\delta}$ can be significantly smaller than if (21.217) were defined with δ^2 on the right-hand side, as before.

21.3 Example: Random Projections and Johnson-Lindenstrauss lemma

The development here can be seen as a nonlinear generalization of the random projection method and the Johnson–Lindenstrauss lemma. Let $\Gamma \in \mathbb{R}^{n \times d}$ be an appropriately scaled random matrix. We then prove that for any fixed $v \in \mathbb{R}^d$, with high probability

$$(1-\varepsilon)^2 \left\|v\right\|_2^2 \leq \left\|\Gamma v\right\|_2^2 \leq (1+\varepsilon)^2 \left\|v\right\|_2^2 .$$

Of particular interest in applications is the lower side of this inequality:

$$\frac{\left\|\Gamma v\right\|_2^2}{\left\|v\right\|_2^2} \geq 1-\alpha$$

where $\alpha \in (0,1)$. A corresponding *uniform* statement over a set $V \subset \mathbb{R}^d$ asks that with high probability,

$$\inf_{v \in V} \frac{\left\|\Gamma v\right\|_2^2}{\left\|v\right\|_2^2} \geq 1-\alpha.$$

Statements of this form are very useful in statistics, signal processing, etc. The lower isometry says that the energy of the signal is preserved under random measurement. Or, the null space of the random matrix Γ is likely to miss (in a quantitative way) the set V. Of course, if V is too large, it's not possible to miss it, and so complexity of V (as quantified by the measures we have studied) enters the picture.

The connection to today's lecture can be seen by taking

$$\Gamma = \frac{1}{\sqrt{n}} \begin{pmatrix} -X_1- \\ \cdots \\ -X_n- \end{pmatrix}$$

with $X_1, \ldots, X_n$ i.i.d. from an isotropic distribution. Then

$$\left\|\Gamma v\right\|_2^2 = \frac{1}{n} \sum_{i=1}^n \langle v, X_i \rangle^2$$

while $\|v\| = \mathbb{E}_x \langle v, X \rangle^2$. Each $v \in V$ then corresponds to $f \in \mathcal{F}$ in our earlier notation.

21.4 Example: Interpolation

Suppose we observe *noiseless* values $y_i = f^*(X_i)$ at i.i.d. locations $X_1, \ldots, X_n$. Let $\widehat{f}$ be an ERM with respect to square loss over $\mathcal{F}$ and assume $f^* \in \mathcal{F}$. Clearly, $\widehat{f}$ achieves zero error, and the question is what the expected deviation from f^* is. This is a question of a "version space size" – what is the $L^2(P)$ diameter of the random subset of $\mathcal{F}$ that matches f^* on a set of data points. More precisely, define the interpolation set

$$\mathcal{I}_{X_1,\ldots,X_n} = \{f \in \mathcal{F} : f(X_i) = f^*(X_i)\},$$

a random subset of the class $\mathcal{F}$, and its diameter as

$$\mathrm{diam}_2(\mathcal{I}_{X_1,\ldots,X_n}) = \sup_{f,f' \in \mathcal{I}_{X_1,\ldots,X_n}} \left\|f - f'\right\|_{L^2(P)} .$$

Of course, from the earlier calculations, we have that with high probability

$$\left\|f - f'\right\|^2_{L^2(P)} \lesssim \bar{\delta}^2$$

where $\bar{\delta}$ is the localization radius for $(\mathcal{F} - \mathcal{F})^2$ and can be upper bounded by $\sup_{x_{1:n}} \widehat{\mathcal{R}}(\mathcal{F})^2$, up to polylog factors. Alternatively, we can use the fixed point $\tilde{\delta}^2$ under the small ball property.

22. LARGE MARGIN THEORY

We now switch gears and discuss the problem of classification with margin. Recall that for a class of binary functions $\mathcal{G} = \{g : \mathcal{X} \to \{\pm 1\}\}$, we established learning and uniform convergence results in terms of the ratio $\mathrm{vc}(\mathcal{G})/n$. Yet, the VC dimension can be easily larger than the sample size for neural networks (where it is related to the number of parameters) and high-dimensional linear separators (e.g. kernels). Perhaps more importantly, we do not usually work with a class $\mathcal{G}$ directly but rather with a real-valued class $\mathcal{F}\{f : \mathcal{X} \to \mathbb{R}\}$, with sign of the function determining the class label. That is, in applications of interest, we work with $\mathrm{sign}(\mathcal{F}) = \{\mathrm{sign}(f) : f \in \mathcal{F}\}$. It is important to realize that nearly-constant f can generate very complex $\mathrm{sign}(f)$, a situation we would like to avoid. Large-margin approach below allows us to replace complexity of $\mathrm{sign}(\mathcal{F})$ with that of $\mathcal{F}$ itself, as a class of real-valued functions.

The proof below utilizes the same technique as that in Lemma 39 (not surprisingly, when we look at the authors of [19] and [18]).

Let $\mathcal{F}$ be a class of $\mathbb{R}$-valued functions. Consider a classification problem with binary $Y \in \{\pm 1\}$. Fix $\gamma > 0$ as a margin parameter.

Let $\phi : \mathbb{R} \to \mathbb{R}$ as

$$\phi(s) = \begin{cases} 1 & \text{if } s \le 0 \\ 1 - s/\gamma & \text{if } 0 < s < \gamma \\ 0 & \text{if } s \ge \gamma \end{cases}$$

Then

$$\mathbf{1}\{yf(x) \le 0\} \le \phi(yf(x)) \le \mathbf{1}\{yf(x) \le \gamma\}.$$

Hence, with probability at least $1 - e^{-2u^2}$, for any $f \in \mathcal{F}$,

$$\begin{aligned} \mathbb{E}\mathbf{1}\{Yf(X) \le 0\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Y_i f(X_i) \le \gamma\} &\le \sup_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)) - \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i)) \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)) - \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i)) + \frac{u}{\sqrt{n}} \end{aligned}$$

since ϕ is in $[0, 1]$. By symmetrization, the above expectation is at most

$$2\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \phi(Y_i f(X_i)) \le \frac{2}{\gamma}\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i Y_i f(X_i) = \frac{2}{\gamma}\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) = \frac{2}{\gamma}\mathcal{R}(\mathcal{F})$$

Hence, with probability at least $1 - e^{-2u^2}$, for any $f \in \mathcal{F}$,

$$\mathbb{E}\mathbf{1}\{Yf(X) \le 0\} \le \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Y_i f(X_i) \le \gamma\} + \frac{2}{\gamma}\mathcal{R}(\mathcal{F}) + \frac{u}{\sqrt{n}}.$$

By a union bound over a discretization of $(0, B]$, we can prove the following result [19, Thm 2]:

Theorem 12. *Let $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$. For all $u > 0$, with probability at least $1 - 2e^{-2u^2}$, for all $f \in \mathcal{F}, \gamma \in (0, 1]$,*

$$\mathbb{E}\mathbf{1}\{Yf(X) \le 0\} \le \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{Y_i f(X_i) \le \gamma\} + \frac{8}{\gamma}\mathcal{R}(\mathcal{F}) + \sqrt{\frac{\log\log(2/\gamma)}{n}} + \frac{u}{\sqrt{n}} \quad (22.219)$$

The key message of this theorem is that upper bound on the expected error is in terms of the complexity of $\mathcal{F}$ as a class of real-valued functions, rather than complexity of $\text{sign}(\mathcal{F})$. The price for this is the margin parameter γ which sets the resolution at which we view predictions as being incorrect (or not confident enough). Given that we would like to have small left-hand-side (for some estimator), the above bound suggest maximize the margin (i.e. minimize the number of margin mistakes) while balancing this goal with complexity of the class. Methods such as support-vector-machines or boosting can be seen as directly or indirectly having this goal.

Finally, suppose we apply the above theorem to some class $\mathcal{F}_B = \{f_{\boldsymbol{\theta}} : \textsc{compl}(\boldsymbol{\theta}) \le B\}$, where $\textsc{compl}(\boldsymbol{\theta})$ is some notion of complexity of the parameter. Suppose $\mathcal{F}_B \subseteq \mathcal{F}_{B'}$ for $B \le B'$ and that $\sup_{f \in \mathcal{F}_B} |f|_\infty \le \psi(B)$. In other words, we allow the function range to increase (linearly or otherwise) with increasing B. We can then apply a union bound to obtain a statement for any $f \in \cup_{B>0}\mathcal{F}_B$ in terms of the complexity of f, defined as the smallest radius B such that $f \in \mathcal{F}_B$. We leave this as an exercise.

22.1 Linear example and comparison to perceptron

As an example, consider the class of linear functions

$$\mathcal{F} = \{x \mapsto \langle x, w\rangle : w \in \mathsf{B}_2^d\}$$

and $\mathcal{X} \in \mathsf{B}_2^d$. We saw earlier that

$$\mathcal{R}(\mathcal{F}) \le \frac{1}{\sqrt{n}}$$

(recall that here we normalized Rademacher averages by $1/n$). Thus, one can derive an upper bound on classification out-of-sample performance that does not depend on the dimensionality of the space despite the fact that the VC dimension of the set of hyperplanes in $\mathbb{R}^d$ is d and covering numbers of $\text{sign}(\mathcal{F})$ necessarily grow with d. Similarly, one can prove margin bounds for neural networks in terms of norms of the weight matrices and without any dependence on the number of neurons.

Observe that the rate of $\frac{1}{\gamma\sqrt{n}}$ is the "slow rate analogue" of the $\frac{1}{n\gamma^2}$ rate we can prove under the assumption that the distribution of the data has a hard margin γ. We show this argument in Section 24.1.

23. COMPLEXITIES OF NEURAL NETWORKS

Neural networks are a class of functions built in a hierarchical manner. Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be a fixed 1-Lipschitz function. Given parameters $\boldsymbol{\theta} = (\mathbf{W}_1, \ldots, \mathbf{W}_L)$, we define

$$f_{\boldsymbol{\theta}}(x) = \sigma(\mathbf{W}_L\sigma(\mathbf{W}_{L-1}\cdots\sigma(\mathbf{W}_1 x)\cdots)), \quad (23.220)$$

where $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and $d_0 = d$ and σ is applied coordinate-wise.

In our setting, the architecture of a neural network corresponds to the choices of input and intermediate dimensions. For the fixed architecture, the set of neural networks we consider is

$$\mathcal{F} = \{f_{\boldsymbol{\theta}} : \textsc{compl}(\boldsymbol{\theta}) \leq B\}$$

where $\textsc{compl}(\boldsymbol{\theta})$ is some notion of complexity of the weight matrices. That is, just as in the case of a class of linear functions $\mathcal{F}_{lin} = \{x \mapsto \langle w, x\rangle : \|w\| \leq B\}$, we would like to define a "ball" in the space of neural networks.

Note that many tuples $(\mathbf{W}_1, \ldots, \mathbf{W}_L)$ lead to the same function $f_{\boldsymbol{\theta}}$. For example, take ReLU activation, scale one layer up by 100, another down by 100. The function does not change under this transformation. There are many transformations that leave the function intact, and we would like to make sure $\textsc{compl}$ does not assign different values of complexity to different sets of parameters if they lead to same function.

As an example, take Frobenius norm of all the layers:

$$\textsc{compl}(\boldsymbol{\theta}) = \sum_{j=1}^{L} \|\mathbf{W}_j\|_F$$

since this is a natural "generalization" of the corresponding Euclidean norm for $\mathcal{F}_{lin}$. Unfortunately, this measure does not capture the scaling invariance of the layers. However, a product of Frobenius norms would reflect the invariance (though it may not reflect many other invariances)

$$\textsc{compl}(\boldsymbol{\theta}) = \prod_{j=1}^{L} \|\mathbf{W}_j\|_F$$

Of course, it is not at all clear that the Rademacher averages of a unit ball defined with respect to this complexity is non-vacuous. Remember that we relied heavily on linearity of functions to analyze $\widehat{\mathcal{R}}(\mathcal{F}_{lin})$.

23.1 Short primer on matrix norms

Before we start, we briefly describe some other norms of a $d_1 \times d_2$ matrix A. We have already seen the operator norm (or, spectral norm, or 2-norm) of a matrix A:

$$\|A\| = \sigma_{\mathsf{max}}(A) = \sqrt{\lambda_{\mathsf{max}}(A^*A)}$$

which can also be written as

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

General Schatten norms are

$$\|A\|_p = \left(\sum_{i=1}^{\min(d_1,d_2)} \sigma_i^p\right)^{1/p},$$

and the $p = 2$ case coincides with the Frobenius norm. The $p = 1$ case is termed nuclear norm, or trace norm, or Ky Fan norm:

$$\|A\|_{nuc} = \sum_{i=1}^{\min(d_1,d_2)} \sigma_i = \operatorname{trace}(\sqrt{A^*A}).$$

Next, we describe entry-wise norms. We start with the sum of ℓ_2 norms of columns:

$$\|A\|_{2,1} = \sum_{j=1}^{d_2} \|A_{\cdot,j}\| = \sum_{j=1}^{d_2} \left(\sum_{i=1}^{d_1} A_{i,j}^2 \right)^{1/2}$$

whereas the maximum ℓ_2 norm of columns is

$$\|A\|_{2,\infty} = \max_{j=1\ldots d_2} \|A_{\cdot,j}\|$$

For general $p, q \geq 1$,

$$\|A\|_{p,q} = \left(\sum_{j=1}^{d_2} \left(\sum_{i=1}^{d_1} |A_{i,j}|^p \right)^{q/p} \right)^{1/q}$$

23.2 Neural networks with bounded $(1, \infty)$ and Frobenius norms

Let us generalize the definition in (23.220) and write it down recursively as follows. Take a base class

$$\mathcal{F}_1 = \{f : \mathcal{X} \to \mathbb{R}\}, \ \mathcal{X} \subset \mathbb{R}^d,$$

and assume (in order to simplify the proof) that $0 \in \mathcal{F}_1$. We now define

$$\mathcal{F}_i = \{x \mapsto \sum_{j=1}^{d_{i-1}} w_j \sigma\left(f_j(x)\right) \ : \ f_j \in \mathcal{F}_{i-1}, \ \|w\|_1 \leq B_i\} \tag{23.221}$$

The following was proved in [2]:

Lemma 40. *Let $\mathcal{F}_i$ be defined recursively as in* (23.221)*, with a base function class $\mathcal{F}_1$ that contains the zero function. Assuming σ is* 1*-Lipschitz and $\sigma(0) = 0$. Then*

$$\widehat{\mathcal{R}}(\mathcal{F}_i) \leq 2B_i \widehat{\mathcal{R}}(\mathcal{F}_{i-1}).$$

Proof.

$$\widehat{\mathcal{R}}(\mathcal{F}_i) = \mathbb{E}_\epsilon \sup_{\substack{\|w\|_1 \leq B_i \\ f_j \in \mathcal{F}_{i-1}}} \sum_{t=1}^n \epsilon_t \left(\sum_j w_j \sigma\left(f_j(x_t)\right) \right) = \mathbb{E}_\epsilon \sup_{\substack{\|w\|_1 \leq B_i \\ f_j \in \mathcal{F}_{i-1}}} \sum_j w_j \sum_{t=1}^n \epsilon_t \sigma\left(f_j(x_t)\right)$$

which is upper bounded via Hölder's inequality by

$$\mathbb{E}_\epsilon \sup_{\substack{w : \|w\|_1 \leq B_i \\ f_j \in \mathcal{F}_{i-1}}} \|w\|_1 \max_j \left| \sum_{t=1}^n \epsilon_t \sigma\left(f_j(x_t)\right) \right| \leq B_i \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^n \epsilon_t \sigma\left(f(x_t)\right) \right|$$

Next, we remove the absolute values and pay a factor of 2:

$$\begin{aligned} \sup_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^n \epsilon_t \sigma\left(f(x_t)\right) \right| &= \sup_{f \in \mathcal{F}_{i-1}} \max \left\{ \sum_{t=1}^n \epsilon_t \sigma\left(f(x_t)\right), -\sum_{t=1}^n \epsilon_t \sigma\left(f(x_t)\right) \right\} \\ &\leq \max \left\{ \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n \epsilon_t \sigma\left(f(x_t)\right), \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n -\epsilon_t \sigma\left(f(x_t)\right) \right\} \end{aligned}$$

Since $0 \in \mathcal{F}$ and $\sigma(0) = 0$, it also holds that $0 \in \mathcal{F}_{i-1}$. Hence both terms in the above max are nonnegative and we can further upper bound the maximum by the sum of two terms, which are equal in expectation:

$$\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^{n} \epsilon_t \sigma\left(f(x_t)\right) \right| \leq 2\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{n} \epsilon_t \sigma\left(f(x_t)\right) \leq 2\widehat{\mathcal{R}}(\mathcal{F}_{i-1}).$$

□

Observe that the restriction that the norm of incoming weights for every neuron is bounded as $\|w\|_1 \leq B_i$ is equivalent to constraining the rows of $\mathbf{W}_i$, which can be written as

$$\|\mathbf{W}_i^\intercal\|_{1,\infty} \leq B_i.$$

Hence, we have the following corollary:

Corollary 5. *Under the assumptions of Lemma 40, the Rademacher averages of $\mathcal{F}_L$ with weight matrices $\|\mathbf{W}_i^\intercal\|_{1,\infty} \leq B_i$ is*

$$\widehat{\mathcal{R}}(\mathcal{F}_L) \lesssim 2^L \prod_{i=1}^{L} B_i \cdot \sqrt{\frac{\log d}{n}},$$

where we also assumed $\mathcal{F}_1 = \{x \mapsto \langle w, x\rangle : \|w\|_1 \leq B_1\}$ and $\mathcal{X} \subseteq B_\infty^d$.

It is easy to see that, in general, the factor 2^L is superfluous in the above bound. Indeed, consider a *thin* neural network $f(x) = w_L\sigma\left(\ldots\sigma(w_1 x)\ldots\right)$ with $w_1 \in \mathbb{R}^{1\times d}$ and all $w_j \in \mathbb{R}_{\geq 0}$ for $j > 1$ be nonnegative numbers. Take σ to be ReLU. Then by positive homogeneity of ReLU,

$$f(x) = \prod_{j>1} w_j \cdot \langle w_1, x\rangle$$

Clearly, in this trivial case there is no exponential dependence on depth.

We mention a result that nearly removes dependence on the depth [12, Theorem 1]:

Theorem 13. *Let σ be 1-homogenous (that is, $\sigma(\alpha x) = \alpha\sigma(x)$ for all $x \in \mathbb{R}, \alpha \geq 0$). Suppose $\mathcal{F}$ is a class of functions of the form* (23.220) *with $\|\mathbf{W}_i\|_F \leq B_i$. Then*

$$\mathcal{R}(\mathcal{F}) \lesssim \sqrt{L} \prod_{i=1}^{L} B_i \cdot \frac{1}{\sqrt{n}}.$$

Under additional mild assumptions, [12, Corollary 1] also shows a depth-independent upper bound of order (up to log factors)

$$\prod_{j=1}^{L} B_i \cdot \frac{1}{n^{1/4}}.$$

24. BEYOND UNIFORM CONVERGENCE?

24.1 Perceptron

Recall Perceptron and its mistake bound in Lemma 28. Perceptron is an online method that, given the next x_t, predicts the label $\widehat{y}_t$ and corrects the hyperplane only in case of a mistake. Given any sequence, the number of mistakes is at most γ^{-2}, where γ is the margin of the sequence.

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. from a distribution on $\mathcal{X} \times \{\pm 1\}$, and suppose $\mathcal{X} \subseteq \mathsf{B}_2^d$. Consider the following procedure. Cycle through the data multiple times until until there is a pass with no more mistakes. The length T of the resulting sequence (T is a multiple of n) is at most $n\gamma^{-2}$, corresponding to the case of one mistake per pass. Let $\boldsymbol{w}_T$ be the final hyperplane output by this procedure. Clearly, it separates the data perfectly, i.e. $\widehat{\mathbf{L}}_{01}(\boldsymbol{w}_T) = 0$ where

$$\widehat{\mathbf{L}}_{01}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{Y_i \langle \boldsymbol{w}, X_i \rangle \le 0\right\}.$$

Therefore, the function $\widehat{f}(x) = \operatorname{sign}(\langle \boldsymbol{w}_T, x \rangle)$ is a particular ERM solution (one of many). Can we say anything about future performance of $\boldsymbol{w}_T$ on data from the same distribution?

Lemma 41. *Let $\boldsymbol{w}_T$ be the output of Perceptron after no mistakes are made in a pass over the i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{\pm 1\}$, and let $\mathcal{X} \subseteq B_2^d$. Let γ be a (random) margin of $n+1$ data points drawn i.i.d. from the distribution. Then*

$$\mathbf{L}_{01}(\boldsymbol{w}_T) = \mathbb{E}\mathbf{1}\left\{Y \langle \boldsymbol{w}_T, X \rangle \le 0\right\} \le \frac{1}{n+1} \times \mathbb{E}[\gamma^{-2}]$$

Proof. Let us use the notation $\mathcal{S} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, $Z_i = (X_i, Y_i)$, $Z = (X, Y)$ and $\boldsymbol{\ell}(\boldsymbol{w}, Z) = \mathbf{1}\left\{Y \langle \boldsymbol{w}, X \rangle \le 0\right\}$. First,

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_Z \boldsymbol{\ell}(\boldsymbol{w}_T, Z) = \mathbb{E}_{\mathcal{S}, Z_{n+1}} \left[\frac{1}{n+1} \sum_{t=1}^{n+1} \boldsymbol{\ell}(\boldsymbol{w}^{(-t)}, Z_t) \right] \tag{24.222}$$

where $\boldsymbol{w}^{(-t)}$ is Perceptron's final hyperplane after (hypothetically) cycling through data $Z_1, \ldots, Z_{t-1}, Z_{t+1}, \ldots, Z_{n+1}$. That is, *leave-one-out* is unbiased estimate of expected loss.

Now consider cycling Perceptron on $Z_1, \ldots, Z_{n+1}$ until no more errors, and call the output $\bar{\boldsymbol{w}}$. Let $i_1, \ldots, i_m$ be indices on which Perceptron errs in *any* of the cycles. We know $m \le \gamma^{-2}$. However, if index $t \notin \{i_1, \ldots, i_m\}$, then whether or not Z_t was included in the computation of $\bar{\boldsymbol{w}}$ does not matter, and so $\bar{\boldsymbol{w}} = \boldsymbol{w}^{(-t)}$. Furthermore, Z_t is correctly classified by $\boldsymbol{w}^{(-t)}$. Thus, at most γ^{-2} terms in (24.222) can be nonzero. □

If we assume hard margin in the distribution (otherwise, expected γ^{-2} will be infinite), Bayes error $\mathbf{L}_{01}(f^*) = 0$. Such an assumption on P is not about its parametric or nonparametric form, but rather on what happens at the boundary. As in Section 3.4, here we beat the CLT rate of $1/\sqrt{n}$.

More importantly, we improved upon the $\frac{1}{\gamma\sqrt{n}}$ rate in Section 22.1 (which was an application of Theorem 12) by a square factor! Recall that Theorem 12 was proved with relatively heavy machinery of uniform convergence, while here we used a trivial argument to obtain a better result. This observation motivates two questions: (1) is there a version of Theorem 12 that achieves the correct rate? and (2) does the Perceptron-based argument magically avoid uniform convergence altogether? The answer to the first question is yes, and it involves developing an L^*-style bound, beyond the scope of this course. But the answer to the second question is more subtle, and should be morally taken as a 'no'. The mechanism employed in the proof of the above lemma is a version of the so-called online-to-batch conversion, where one first proves an online mistake or regret bound for an arbitrary sequence and then uses the i.i.d. nature of the sequence to derive a result on expected loss. However, the very fact that one can show an online mistake bound or an online regret bound for an arbitrary sequence implies a stronger version of uniform convergence – uniform convergence for martingales. We will describe this in detail in the last lecture.

25. BIAS-VARIANCE DECOMPOSITION

For a large part of the course, we studied risk bounds for $\widehat{f}$ defined implicitly as an empirical minimizer over some class of functions. Yet, in certain situations, an estimator of interest is defined explicitly. This was the case, for instance, with linear unconstrained regression. In that case, however, we opted for not using the closed-form solution since such an approach would not be generalizable to nonlinear cases (or even to linear constrained regression). Here we describe a classical approach that is convenient for analyzing closed-form estimators.

We consider random design regression. To this end, let P be the law of (X, Y), and $f^*(x) = \mathbb{E}[Y|X = x]$ be the regression function. We write $Y_i = f^*(X_i) + \xi_i$ for zero-mean ξ_i. Let $\mathbf{X} = [X_1, \ldots, X_n]^\intercal \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} = [Y_1, \ldots, Y_n]^\intercal$. Given an estimator $\widehat{f}(\cdot) = \widehat{f}(\cdot; \mathbf{X}, \boldsymbol{y})$, define

$$\boldsymbol{B}^2 = \mathbb{E}_X \left(f^*(X) - \mathbb{E}_{\boldsymbol{y}}\widehat{f}(X)\right)^2, \qquad \boldsymbol{V} = \mathbb{E}_{X,\boldsymbol{y}} \left(\widehat{f}(X) - \mathbb{E}_{\boldsymbol{y}}\widehat{f}(X)\right)^2. \tag{25.223}$$

Both $\boldsymbol{B}^2$ and $\boldsymbol{V}$ are random variables (in $\mathbf{X}$), and it is easy to check that

$$\mathbb{E}\|\widehat{f} - f^*\|^2_{L^2(P)} = \mathbb{E}_{\mathbf{X}}\left[\boldsymbol{B}^2\right] + \mathbb{E}_{\mathbf{X}}\left[\boldsymbol{V}\right]. \tag{25.224}$$

Consider estimators that are linear in $\boldsymbol{y}$:

$$\widehat{f}(x) = \sum_{i=1}^n Y_i \omega_i(x). \tag{25.225}$$

We then have

$$\boldsymbol{B}^2 = \mathbb{E}_X \left(f^*(X) - \sum_{i=1}^n f^*(X_i)\omega_i(X)\right)^2 \tag{25.226}$$

and

$$\boldsymbol{V} = \mathbb{E}_{X,\boldsymbol{\xi}} \left(\sum_{i=1}^n \xi_i \omega_i(X)\right)^2 \leq \sigma_\xi^2 \sum_{i=1}^n \mathbb{E}_X \left(\omega_i(X)\right)^2. \tag{25.227}$$

The form of (25.226) and (25.227) is particularly useful for analyzing "local methods". Indeed, let's think of $\omega_i(x)$ as the "relevance" of example (X_i, Y_i) to the given point x. In this case, $\widehat{f}(x)$ in (25.225) aggregates the responses Y_i according to these weights. The bias term then asks whether the problem is easy if there is no noise ξ_i, and (25.226) has the interpretation of the expected difference between the value of the true regression function at

X and its "reconstruction" from datapoints, assuming no noise. Smoothness of f^* helps to upper bound this term. The variance term increases with the noise level σ_ξ^2 and the sum of $L^2(P)$-norms of the weight functions. If these weight functions ω_i are sufficiently localized around X_i, one can often compute simple upper bounds on the variance term.

25.1 Example: Local Smoothing

In local smoothing,

$$\omega_i(x) = \frac{K((x - x_i)/h)}{\sum_{j=1}^n K((x - x_j)/h)} \tag{25.228}$$

where $K(u) : \mathbb{R}^d \mapsto [0, \infty)$ is a kernel function and $h > 0$ is a bandwidth parameter. Example: $K(u) = \exp\{-\|u\|\}$ or its truncated version $K(u) = \exp\{-\|u\|\}\mathbf{1}\{\|u\|_2 \le 1\}$. Another example is $K(u) = \|u\|^{-a}\mathbf{1}\{\|u\|_2 \le 1\}$ for $0 < a < d/2$, which is singular at 0 and leads to an interpolant of the data.

25.2 Example: Least Squares

Consider unconstrained Least Squares, which has a closed-form solution

$$\widehat{f}(x) = \langle \widehat{\boldsymbol{\theta}}, x \rangle = \langle \mathbf{X}^\dagger \boldsymbol{y}, x \rangle = (\mathbf{X}x)^\intercal (\mathbf{X}\mathbf{X}^\intercal)^{-1} \boldsymbol{y}, \tag{25.229}$$

This solution can be written as $\widehat{f}(x) = \sum_{i=1}^n Y_i \omega_i(x)$, where

$$\omega_i(x) = (x^\intercal \mathbf{X}^\dagger)_i = (\mathbf{X}x)^\intercal (\mathbf{X}\mathbf{X}^\intercal)^{-1} \boldsymbol{e}_i. \tag{25.230}$$

To avoid confusion, we will use the lower-case x for a random $x \sim P$. The bias is then

$$\boldsymbol{B}^2 = \mathbb{E}_x \langle P^\perp x, \boldsymbol{\theta}^* \rangle^2 = \left\| \boldsymbol{\Sigma}^{1/2} P^\perp \boldsymbol{\theta}^* \right\|_2^2, \tag{25.231}$$

where $P^\perp = \mathbf{I}_d - \mathbf{X}^\intercal (\mathbf{X}\mathbf{X}^\intercal)^{-1} \mathbf{X}$. On the other hand, the variance term is

$$\boldsymbol{V} \le \sigma_\xi^2 \cdot \mathbb{E}_x \left\| (\mathbf{X}\mathbf{X}^\intercal)^{-1} (\mathbf{X}x) \right\|_2^2 = \sigma_\xi^2 \cdot \mathsf{trace}\left((\mathbf{X}\mathbf{X}^\intercal)^{-2} \mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\intercal \right). \tag{25.232}$$

25.3 Example: Regularized Least Squares

Regularized Least Squares (or, Ridge Regression) is a classical method employed for high-dimensional data,

$$\widehat{\boldsymbol{\theta}}_\lambda = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{\theta}\|^2 \tag{25.233}$$

and it has a closed-form expression

$$\widehat{\boldsymbol{\theta}}_\lambda = \mathbf{X}^\intercal (\mathbf{X}\mathbf{X}^\intercal + \lambda I_n)^{-1} \boldsymbol{y}. \tag{25.234}$$

It is easy to extend (25.231) and (25.232) to this case (exercise).

25.4 Example: Kernel Ridge/Ridgeless Regression

Observe that the solution for Least Squares and Regularized Least Squares only depends on inner products between data points X_i and X_j, $i, j \in [n]$. There are several ways to motivate kernel methods, but the one we take here just replaces x with some feature map $\phi(x) : \mathcal{X} \to \mathbb{R}^D$ with D large or infinite. Let $\Phi \in \mathbb{R}^{n\times D}$ be the matrix with rows $\phi(X_i)^\intercal$. From the earlier discussion, the least squares solution in this high- or infinite-dimensional space is simply

$$\widehat{f}(x) = \langle \widehat{\boldsymbol{\theta}}, \phi(x)\rangle = (\Phi\phi(x))^\intercal(\Phi\Phi^\intercal)^{-1}\boldsymbol{y}. \tag{25.235}$$

It is useful to write $k(x, x') = \langle\phi(x), \phi(x')\rangle$ (k is called a *kernel*), as well as write $\Phi\Phi^\intercal$ as $K(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n\times n}$ with (i, j) entry $k(X_i, X_j)$. Furthermore, we write $\Phi\phi(x)$ as $K(x, \mathbf{X}) \in \mathbb{R}^n$ with $[K(x, \mathbf{X})]_i = k(X_i, x)$. With this notation, (25.235) becomes

$$\widehat{f}(x) = K(x, \mathbf{X})^\intercal K(\mathbf{X}, \mathbf{X})^{-1}\boldsymbol{y}. \tag{25.236}$$

and the Kernel Ridge Regression solution becomes

$$\widehat{f}(x) = K(x, \mathbf{X})^\intercal(K(\mathbf{X}, \mathbf{X}) + \lambda I_n)^{-1}\boldsymbol{y}. \tag{25.237}$$

25.5 Example: Linear Regime in Nonlinear Models

Let $f(x, \boldsymbol{\theta})$ be a function $\mathcal{X} \to \mathbb{R}$ parametrized (potentially non-linearly) by $\boldsymbol{\theta} \in \mathbb{R}^p$. A running example here is a neural network with 1 hidden layer

$$f(x, \boldsymbol{\theta}) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m} b_j\sigma(\langle \boldsymbol{w}_j, x\rangle), \quad \boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m) \tag{25.238}$$

and $b_1, \ldots, b_m$ fixed (for simplicity, we do not include them in $\boldsymbol{\theta}$). Here $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linearity such as $\sigma(a) = \max\{a, 0\}$.

Suppose our estimator is a solution of (potentially non-convex) least-squares problem

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{n}\|\boldsymbol{y} - f_n(\boldsymbol{\theta})\|^2 \tag{25.239}$$

where $f_n : \boldsymbol{\theta} \mapsto (f(X_1, \boldsymbol{\theta}), \ldots, f(X_n, \boldsymbol{\theta}))$ is the evaluation of the function parametrized by $\boldsymbol{\theta}$ on the data. Despite potential non-convexity of the problem, we can aim to minimize the squared loss by gradient flow (or gradient descent). Taking $\boldsymbol{\theta}_0$ as a starting point, the evolution is given by

$$\frac{d\boldsymbol{\theta}_t}{dt} = \frac{1}{n}Df_n(\boldsymbol{\theta})^\intercal(\boldsymbol{y} - f_n(\boldsymbol{\theta}_t)) \tag{25.240}$$

where $Df_n(\boldsymbol{\theta}) \in \mathbb{R}^{n\times p}$ is the Jacobian of f_n. Let us linearize

$$f_n(\boldsymbol{\theta}_t) \approx f_n(\boldsymbol{\theta}_0) + Df_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)$$

around $\boldsymbol{\theta}_0$. This linearization can be a good approximation if $\boldsymbol{\theta}$ does not move too far from $\boldsymbol{\theta}_0$ and f_n is "regular" enough. Since the linearization introduces different dynamics, we use $\bar{\boldsymbol{\theta}}_t$ to denote it. We have

$$\frac{d\bar{\boldsymbol{\theta}}_t}{dt} = \frac{1}{n}Df_n(\boldsymbol{\theta})^\intercal(\boldsymbol{y} - f_n(\boldsymbol{\theta}_0) - Df_n(\boldsymbol{\theta}_0)(\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)) \tag{25.241}$$

Under certain conditions (see e.g. [4, Thm 5.1]), parameters $\boldsymbol{\theta}_t$ stay close to $\bar{\boldsymbol{\theta}}_t$, square loss (25.239) decays exponentially fast to 0 under the dynamics of $\boldsymbol{\theta}_t$, and function values $f(x, \boldsymbol{\theta}_t)$ are close to those of the linear model

$$f^{\mathrm{lin}}(x, \boldsymbol{\theta}) = f(x, \boldsymbol{\theta}_0) + Df(x, \boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \tag{25.242}$$

in $L^2(P)$ (here, D is extended to be a linear operator from the parameter space $\mathbb{R}^p$ to the space of functions, $L^2(P)$). See [4] for references to prior work on this.

When we examine the linear model (25.242) on the data (note the subscript n),

$$f_n^{\mathrm{lin}}(\boldsymbol{\theta}) = f_n(\boldsymbol{\theta}_0) + Df_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{25.243}$$

we see that the data $X_1, \dots, X_n$ are mapped to a feature space, with the feature matrix being

$$\Phi = Df_n(\theta_0),$$

and then we find the best parameter $\boldsymbol{\theta}$ in this feature space.

Suppose we initialize the model in such a way that $f_n(\boldsymbol{\theta}) = 0$ (or approximately 0) so that we can drop it from the above expression. In this case, $\Phi\Phi^\intercal$ is the kernel matrix

$$K(\mathbf{X}, \mathbf{X}) = Df_n(\boldsymbol{\theta}_0)Df_n(\boldsymbol{\theta}_0)^\intercal \in \mathbb{R}^{n \times n}.$$

If this kernel matrix is full-rank, the linearized gradient flow can be shown to converge to the minimum interpolant of the data:

$$\bar{\boldsymbol{\theta}}_\infty = \operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ \left\| \bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right\|_2 : Df_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \boldsymbol{y} - f_n(\boldsymbol{\theta}_0) \right\}$$

(again, we can ensure $f_n(\boldsymbol{\theta}_0) = 0$)

25.5.1 Feature map and kernels for (25.238)

We can calculate the Jacobian $Df_n(\boldsymbol{\theta})$ for the model in (25.238). Here $p = md$, where $x \in \mathbb{R}^d$ with d input dimension. Then, trivially,

$$[Df_n(\boldsymbol{\theta})]_{i,(j,a)} = \frac{1}{\sqrt{m}} b_j \sigma'(\langle \boldsymbol{w}_j, x_i \rangle) x_{i,a}$$

for $i \in [n], (j, a) \in [m] \times [d]$. This corresponds to the feature matrix $\Phi \in \mathbb{R}^{n \times md}$ given by

$$\Phi = \begin{bmatrix} b_1\sigma'(\langle x_1, \boldsymbol{w}_1\rangle)x_1^\intercal & b_2\sigma'(\langle x_1, \boldsymbol{w}_2\rangle)x_1^\intercal & \dots & b_m\sigma'(\langle x_1, \boldsymbol{w}_m\rangle)x_1^\intercal \\ b_1\sigma'(\langle x_2, \boldsymbol{w}_1\rangle)x_2^\intercal & b_2\sigma'(\langle x_2, \boldsymbol{w}_2\rangle)x_2^\intercal & \dots & b_m\sigma'(\langle x_2, \boldsymbol{w}_m\rangle)x_2^\intercal \\ \dots & \dots & \dots & \dots \\ b_1\sigma'(\langle x_n, \boldsymbol{w}_1\rangle)x_n^\intercal & b_2\sigma'(\langle x_n, \boldsymbol{w}_2\rangle)x_n^\intercal & \dots & b_m\sigma'(\langle x_n, \boldsymbol{w}_m\rangle)x_n^\intercal \end{bmatrix} \tag{25.244}$$

The corresponding kernel, termed the Neural Tangent Kernel (NTK), is

$$K_m(x_1, x_2) = \frac{1}{m} \sum_{i=1}^{m} \langle x_1, x_2 \rangle \sigma'(\langle \boldsymbol{w}_i, x_1 \rangle)\sigma'(\langle \boldsymbol{w}_i, x_2 \rangle).$$

where we assumed $b_i = \pm 1$. Now, suppose the weights $\boldsymbol{w}_i \sim \mathcal{N}(0, I_d/d)$, independently. As the number of neurons m increases, the finite-width kernel K_m converges (under conditions) to an infinite-width NTK given by

$$K(x_1, x_2) = \mathbb{E}_{\boldsymbol{w}} \langle x_1, x_2 \rangle \sigma'(\langle \boldsymbol{w}, x_1 \rangle)\sigma'(\langle \boldsymbol{w}, x_2 \rangle)$$

Under the conditions which ensure that the nonlinear gradient flow (25.240) stays close to the linearized gradient flow, $\boldsymbol{\theta}_t$ converges to a minimum-norm interpolant of the data with respect to the NTK kernel (see e.g. [4]), and can be analyzed with the bias-variance decomposition. These cases are among the few where we can provably analyze *both* optimization and statistical properties of neural network models. Arguably, however, the linear regime is not very interesting in practice.

Another aspect we have not discussed here is that of interpolation. See [4].

26. TIME SERIES

Suppose we observe a sequence

$$\boldsymbol{x}_{t+1} = f^*(\boldsymbol{x}_t) + \eta_t, \qquad t = 1, \ldots, n$$

where $\boldsymbol{x}_t \in \mathbb{R}^d$ and η_t are independent zero mean vectors. The function f^* is unknown, but we assume it is a member of a known class $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}^d\}$. Let us treat this problem as a fixed-design regression problem, except that the outcomes are now vectors rather than reals, and the sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is a sequence of *dependent* random variables.

Consider the least squares solution:

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \|\boldsymbol{x}_{t+1} - f(\boldsymbol{x}_t)\|_2^2 ,$$

where the norm is the Euclidean norm. This is a natural generalization of least squares to vector-valued regression. As before, we denote

$$\|f - g\|_n^2 = \frac{1}{n} \sum_{t=1}^{n} \|f(\boldsymbol{x}_t) - g(\boldsymbol{x}_t)\|_2^2$$

The basic inequality can now be written as (exercise):

$$\left\|\widehat{f} - f^*\right\|_n^2 \leq 2 \frac{1}{n} \sum_{t=1}^{n} \langle \eta_t, \widehat{f}(\boldsymbol{x}_t) - f^*(\boldsymbol{x}_t) \rangle.$$

Choosing the offset-style approach covered in previous lectures, we have

$$\left\|\widehat{f} - f^*\right\|_n^2 \leq \sup_{g \in \mathcal{F} - f^*} \frac{1}{n} \sum_{t=1}^{n} 4\langle \eta_t, g(\boldsymbol{x}_t) \rangle - \|g(\boldsymbol{x}_t)\|^2 .$$

Up until now, the statement is conditional on $\{\eta_1, \ldots, \eta_n\}$. What happens if we take expectations on both sides? On the left-hand side we have a denoising guarantee on the sequence. On the right-hand side, we have a "dependent version" of offset Gaussian/Rademacher complexity where $\boldsymbol{x}_t$ is measurable with respect to $\sigma(\eta_1, \ldots, \eta_{t-1})$. To analyze this object, we first need to understand the simpler $\mathbb{R}$-valued version without the offset: what is the behavior of

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t)$$

where $\boldsymbol{x}_t$ is $\sigma(\epsilon_1, \ldots, \epsilon_{t-1})$-measurable, $\mathcal{F}$ is a class of real-valued functions $\mathcal{X} \to \mathbb{R}$, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables.

27. SEQUENTIAL COMPLEXITIES

We choose to study the random process generated by Rademacher random variables for several reasons. First, just as in the classical case, conditioning on the data will lead to a simpler object (binary tree) and, second, other noise processes can be reduced to the Rademacher case, under moment assumptions on the noise. The development here is based on [27], and we refer also to [26] for an introduction.

Let us elaborate on the first point. Note that $\boldsymbol{x}_t$ being measurable with respect to $\sigma(\epsilon_1, \ldots, \epsilon_{t-1})$ simply means $\boldsymbol{x}_t$ is a function of $\epsilon_1, \ldots, \epsilon_{t-1}$ (in other words, it's a predictable process). Note that the collection $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ can be "summarized" as a depth-n binary tree decorated with elements of $\mathcal{X}$ at the nodes. Indeed, $\boldsymbol{x}_1 \in \mathcal{X}$ is a constant (root), $\boldsymbol{x}_2 = \boldsymbol{x}_2(\epsilon_1)$ takes on two possible values depending on the sign of ϵ_1 (left or right), and so forth. It is useful to think of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ as a tree, even though it doesn't bring any more information into the picture. We shall denote the collection of n functions $\boldsymbol{x}_i : \{\pm 1\}^{i-1} \to \mathcal{X}$ as $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and call it simply as an $\mathcal{X}$-valued *tree*. We shall refer to $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ as a *path* in the tree. We will also talk about $\mathbb{R}$-valued trees, such as $f \circ \boldsymbol{x}$ for $f : \mathcal{X} \to \mathbb{R}$.

Given a tree $\boldsymbol{x}$, we shall call

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}, \boldsymbol{x}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1}))$$

the *sequential Rademacher complexity* of $\mathcal{F}$ on the tree $\boldsymbol{x}$.

Comparing to the classical version,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(x_t)$$

where $x_1, \ldots, x_n$ are constant values, we see that it is a special case of a tree with constant levels $\boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1}) = x_t$. Hence, sequential Rademacher complexity is a generalization of the classical notion.

To ease the notation, we will write $\boldsymbol{x}_t$ without explicit dependence on ϵ, or for brevity write $\boldsymbol{x}_t(\epsilon)$ even though $\boldsymbol{x}_t$ only depends on the prefix $\epsilon_{1:t-1}$.

Observe that for any $f \in \mathcal{F}$, the variable

$$\nu_f = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t)$$

is zero mean. Moreover, it is an average of martingale differences $\epsilon_t f(\boldsymbol{x}_t)$, and so we expect $1/\sqrt{n}$ behavior from Azuma-Hoeffding's inequality. It should be clear that, say, for $\mathcal{F}$ consisting of a finite collection of $[-1, 1]$-valued functions on $\mathcal{X}$, we have

$$\mathbb{E} \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t) \leq \sqrt{\frac{2 \log \operatorname{card}(\mathcal{F})}{n}}$$

Given that there is no difference with the classical case, one may wonder if we can just reduce everything to the classical Rademacher averages. The answer is no, and the differences already start to appear when we attempt to define covering numbers.

More precisely, since any tree $\boldsymbol{x}$ is defined by $2^n - 1$ values, one might wonder if we could define a notion of pseudo-distance between f and f' as an ℓ_2 distance on these $2^n - 1$ values.

It is easy to see that this is a huge overkill. Perhaps one of the key points to understand here is: what is the equivalent of the projection $\mathcal{F}|_{x_1,\ldots,x_n}$ for the tree case? Spoiler: it's not $\mathcal{F}|_{\boldsymbol{x}}$. The following turns out to be the right definition:

Definition 22. *A set V of $\mathbb{R}$-valued trees is an 0-cover of $\mathcal{F}$ on a tree $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ if*

$$\forall f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad s.t. \quad f(\boldsymbol{x}_t(\epsilon_{1:t-1})) = \boldsymbol{v}_t(\epsilon_{1:t-1}) \quad \forall t \in [n]$$

The size of the smallest 0-cover of $\mathcal{F}$ on a tree $\boldsymbol{x}$ will be denoted by $\mathcal{N}(\mathcal{F}, \boldsymbol{x}, 0)$.

The key aspect of this definition is that $\boldsymbol{v} \in V$ can be chosen based on the sequence $\epsilon \in \{\pm 1\}^n$. In other words, in contrast with the classical definition, for the same function f different elements $\boldsymbol{v} \in V$ can provide a cover on different paths. This results in the needed reduction in the size of V.

As an example, take a set of 2^{n-1} functions that take a value of 1 on one of the 2^{n-1} leaves of $\boldsymbol{x}$ and zero everywhere else. Then the projection $\mathcal{F}|_{\boldsymbol{x}}$ is of size 2^{n-1} but the size of the 0-cover is only 2 (exercise!), corresponding to our intuition that the class is simple (as it only varies on the last example). Indeed, the size of the 0-cover is the analogue of the size of $\mathcal{F}|_{x_1,\ldots,x_n}$ in the binary-valued case.

For real-valued functions, consider the following definition.

Definition 23. *A set V of $\mathbb{R}$-valued trees is an α-cover of $\mathcal{F}$ on a tree $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ with respect to ℓ_2 if*

$$\forall f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad s.t. \quad \frac{1}{n}\sum_{i=1}^{n} (f(\boldsymbol{x}_t(\epsilon_{1:t-1})) - \boldsymbol{v}_t(\epsilon_{1:t-1}))^2 \leq \alpha^2$$

The size of the smallest α-cover of $\mathcal{F}$ on a tree $\boldsymbol{x}$ with respect to ℓ_2 will be denoted by $\mathcal{N}_2(\mathcal{F}, \boldsymbol{x}, \alpha)$.

A similar definition can be stated for cover with respect to ℓ_p.

The following is an analogue of the chaining bound:

Theorem 14. *For any class of $[-1,1]$-valued functions $\mathcal{F}$,*

$$\widehat{\mathcal{R}}^{seq}(\mathcal{F}, \boldsymbol{x}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \boldsymbol{x}, \varepsilon)} d\varepsilon \right\}$$

Recall the definition of VC dimension and a shattered set. Here is the right sequential analogue:

Definition 24. *Function class $\mathcal{F}$ of $\{\pm 1\}$-valued functions shatters a tree $\boldsymbol{x}$ of depth d if*

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists f \in \mathcal{F}, \quad s.t. \quad \forall t \in [d], \quad f(\boldsymbol{x}_t(\epsilon)) = \epsilon_t$$

The largest depth d for which there exists a shattered $\mathcal{X}$-valued tree is called the Littlestone dimension *and denoted by ldim($\mathcal{F}$).*

To contrast with the classical definition, the path on which the signs should be realized is given by the path itself. But it's clear that the definition serves the same purpose: if $\boldsymbol{x}$ is shattered by $\mathcal{F}$ then $\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}, \boldsymbol{x}) = 1$. It is also easy to see that $\text{vc}(\mathcal{F}) \leq \text{ldim}(\mathcal{F})$, and the gap can be infinite.

The following is an analogue of the Sauer-Shelah-Vapnik-Chervonenkis lemma.

Theorem 15. *For a class of binary-valued functions $\mathcal{F}$ with Littlestone dimension ldim($\mathcal{F}$),*

$$\mathcal{N}(\mathcal{F}, \boldsymbol{x}, 0) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$$

Scale-sensitive sequential versions are defined as follows:

Definition 25. *Function class $\mathcal{F}$ of $\mathbb{R}$-valued functions shatters a tree $\boldsymbol{x}$ of depth d at scale α if there exists a witness $\mathbb{R}$-valued tree $\boldsymbol{s}$ such that*

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists f \in \mathcal{F}, \quad s.t. \quad \forall t \in [d], \qquad \epsilon_t(f(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{s}_t(\epsilon)) \geq \alpha/2$$

The largest depth d for which there exists an α-shattered $\mathcal{X}$-valued tree is called sequential scale-sensitive dimension and denoted ldim($\mathcal{F}, \alpha$).

We note that the above definitions reduce to the classical ones if we consider only trees $\boldsymbol{x}$ with constant levels.

Theorem 16. *For any class of $[-1,1]$-valued functions $\mathcal{F}$ and $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n*

$$\mathcal{N}_\infty(\mathcal{F}, \boldsymbol{x}, \alpha) \leq \left(\frac{2en}{\alpha}\right)^{ldim(\mathcal{F},\alpha)}$$

Finally, it is possible to show an analogue of symmetrization lemma: for any joint distribution of $(X_1, \ldots, X_n)$,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[f(X_t)|X_{1:t-1}] - f(X_t) \leq 2 \sup_{\boldsymbol{x}} \widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}, \boldsymbol{x})$$

If the sequence $(X_1, \ldots, X_n)$ is i.i.d., the left-hand side is the expected supremum of the empirical process. The present version provides a martingale generalization. Furthermore, if we take supremum over all joint distributions on the left-hand-side, then the lower bound is also matching the upper bound, up to a constant.

The offset Rademacher complexity has been analyzed in [25].

28. ONLINE LEARNING

Consider the following online classification problem. On each of n rounds $t = 1, \ldots, n$, the learner observes $x_t \in \mathcal{X}$, makes a prediction $\widehat{y}_t \in \{\pm 1\}$, and observes the outcome $y_t \in \{\pm 1\}$. The learner models the problem by fixing a class $\mathcal{F}$ of possible models $f : \mathcal{X} \to \{\pm 1\}$, and aims to predict nearly as well as the best model in $\mathcal{F}$ in the sense of keeping *regret*

$$\operatorname{Reg}(\mathcal{F}) = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\left\{\widehat{y}_t \neq y_t\right\}\right] - \inf_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\left\{f(x_t) \neq y_t\right\}\right] \tag{28.245}$$

small for any sequence $(x_1, y_1), \ldots, (x_n, y_n)$. At least visually, this looks like oracle inequalities for misspecified models. The distinguishing feature of this online framework is that (a) data arrives sequentially, and (b) we aim to have low regret for any sequence without assuming any generative process.

It is also worth noting that in the above protocol there is no separation of training and test data: the online nature of the problem allows us to first test our current hypothesis by making a prediction, then observe the outcome and incorporate the datum in to our dataset.

The expectation on the first term in (28.245) is with respect to learner's internal randomization. More specifically, let Q_t be the distribution on $\{\pm 1\}$ that the learner uses to predict $\widehat{y}_t \sim Q_t$. Let $q_t = \mathbb{E}\widehat{y}_t$ be the (conditional) mean of this distribution. In other words, $q_t = 0$ would correspond to the learner tossing a fair coin.

A note about the protocol. The results below hold even if the sequence is chosen based on learner's past predictions. However, in this case, y_t may only depend on q_t but not on the realization $\widehat{y}_t$. To simplify the presentation, let us just assume that the sequence $(x_1, y_1), \ldots, (x_n, y_n)$ is fixed in advance (this turns out not to matter).

We will answer the following question: what is the best achievable $\operatorname{Reg}(\mathcal{F})$ for a given $\mathcal{F}$ by any prediction strategy?

Let us first rewrite $\mathbf{1}\left\{\widehat{y}_t \neq y_t\right\} = (1 - \widehat{y}_t y_t)/2$ and do the same for the oracle term. Cancelling $1/2$, we have

$$2\operatorname{Reg}(\mathcal{F}) = \frac{1}{n}\sum_{t=1}^{n} -q_t y_t - \inf_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} -y_t f(x_t)\right] \tag{28.246}$$

$$= \sup_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} y_t f(x_t)\right] - \frac{1}{n}\sum_{t=1}^{n} q_t y_t \tag{28.247}$$

Now, consider a particular stochastic process for generating the data sequence: fix any $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n, and on round t let $x_t = \boldsymbol{x}_t(y_1, \ldots, y_{t-1})$ and $y_t = \epsilon_t$ be an independent Rademacher random variable. This defines a stochastic process with 2^n possible sequences $(x_1, y_1), \ldots, (x_n, y_n)$. Now, clearly

$$2\operatorname{Reg}(\mathcal{F}) \geq 2\mathbb{E}_\epsilon \operatorname{Reg}(\mathcal{F}).$$

Observe that $q_t = q_t(\epsilon_1, \ldots, \epsilon_{t-1})$ and thus

$$\mathbb{E}_\epsilon \left[\frac{1}{n}\sum_{t=1}^{n} q_t \epsilon_t\right] = 0.$$

Hence,

$$\mathbb{E}_\epsilon \mathrm{Reg}(\mathcal{F}) = \mathbb{E} \sup_{f\in\mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t) \right]. \tag{28.248}$$

Since the argument holds for any $\boldsymbol{x}$, we have proved that the optimal value of $\mathrm{Reg}(\mathcal{F})$ is lower bounded by half of

$$\bar{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}) = \sup_{\boldsymbol{x}} \widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}, \boldsymbol{x}).$$

It turns out that this lower bound is within a factor of 2 from optimal. Define the minimax value

$$\mathcal{V} = \min_{\text{Algo}} \max_{\{(x_t,y_t)\}_{t=1}^n} \mathrm{Reg}(\mathcal{F})$$

Theorem 17. *For a binary-valued class $\mathcal{F}$,*

$$\frac{1}{2}\bar{\mathcal{R}}^{seq}(\mathcal{F}) \leq \mathcal{V} \leq \bar{\mathcal{R}}^{seq}(\mathcal{F})$$

Similar results also holds for absolute value and other Lipschitz loss functions. For square loss, the sequential Rademacher averages are replaced by offset sequential Rademacher averages (again, as both upper and lower bounds).

In short, sequential complexities in online learning play a role similar to the role played by i.i.d. complexities as studied in this course. However, quite a large number of questions still remains open. But that's a topic for a different course.

References

[1] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2008.

[2] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[3] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[4] P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[6] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms.* PhD thesis, Biologische Kybernetik, 2002.

[7] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer, 1996.

[9] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[10] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

[11] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

[12] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[13] S. B. Hopkins. Mean estimation with sub-gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020.

[14] D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

[15] J.-P. Kahane. *Some random series of functions.* D. C. Heath, 1968.

[16] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[17] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

[18] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23): 12991–13008, 2015.

[19] V. Koltchinskii, D. Panchenko, et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[20] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

[21] T. Liang, A. Rakhlin, and K. Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

[22] G. Lugosi and S. Mendelson. Sub-gaussian estimators of the mean of a random vector. *The annals of statistics*, 47(2):783–794, 2019.

[23] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[24] S. Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[25] A. Rakhlin and K. Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

[26] A. Rakhlin and K. Sridharan. On martingale extensions of Vapnik–Chervonenkis theory with applications to online learning. In *Measures of Complexity*, pages 197–215. Springer, 2015.

[27] A. Rakhlin, K. Sridharan, and A. Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

[28] M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

[29] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[30] V. I. Serdobolskii. *Multivariate statistical analysis: A high-dimensional approach*, volume 41. Springer Science & Business Media, 2000.

[31] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

[32] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

[34] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[35] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

[36] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians, 2014.