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

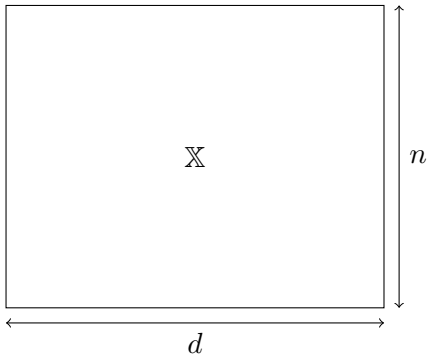
Lecturer: Philippe Rigollet
Scribe: Philippe Rigollet

Lecture 1
Feb. 4, 2020

Goals: This lecture is an introduction to the concepts covered in this class. In particular, we will discuss the difference between the *asymptotic* and *non-asymptotic* approaches to mathematical statistics.

We also give a brief overview of the some of the topics covered in class: Covariance matrix estimation, matrix estimation, empirical risk minimization, neural networks and minimax lower bounds.

A good example to keep in mind is a dataset organized as an n by d matrix $\mathbb{X}$ where, for example, the rows correspond to patients and the columns correspond to measurements on each patient (height, weight, ...). Row i is a random vector $X_i^\top \in \mathbb{R}^d$ of the measurements performed on patient i.



We now compare the *asymptotic* and *non-asymptotic* approaches to mathematical statistics.To better illustrate the difference between the two approaches, let us consider some examples.

1. ASYMPTOTIC VS. NON-ASYMPTOTIC REGIMES

1.1 Mean estimation

Here $d = 1$ and we observe $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} P$ where P is a distribution over $\mathbb{R}$ with mean μ and variance σ^2. We consider the sample mean:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \,.$$

We have the following asymptotic results:

- Law of Large Numbers (LLN): $\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu$

- Central Limit Theorem (CLT): $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$

We can also make some non-asymptotic statements:

- Quadratic risk: $\mathbb{E}[(\bar{X}_n - \mu)^2] = \dfrac{\sigma^2}{n}$
- Tail bounds: $\mathbb{P}(|\bar{X}_n - \mu| > t) \leq 2e^{-Cnt^2}$, where $C > 0$ is a constant that depends on further assumptions on P, such as having a bounded support (Hoeffding's inequality).

1.2 Covariance matrix estimation

Assume now that we observe $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} P$ where P is a distribution over $\mathbb{R}^d$ with mean μ and covariance matrix $\Sigma = \mathbb{E}[X_1 X_1^\top]$. The sample covariance matrix $\hat{\Sigma}$ is defined as

$$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top .$$

We still have some asymptotic statements:

- LLN: $\hat{\Sigma}_{i,j} \xrightarrow[n\to\infty]{\mathbb{P}} \Sigma_{i,j}$, for all $i, j = 1, \dots, d$.
- CLT: $\sqrt{n}(\hat{\Sigma}_{i,j} - \Sigma_{i,j}) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \operatorname{var}(\mathbb{X}_{1,i}\mathbb{X}_{1,j}))$.

To compute explicitly the variance $\operatorname{var}(\mathbb{X}_{1,i}\mathbb{X}_{1,j})$, one would need to make assumptions on the fourth moment P but this value does not matter for our considerations here.

In this case, letting $n \to \infty$ implicitly assumes that $n \gg d$. But what if n is of the order of d or even if $n \ll d$?

Can we make similar statements simultaneously for all the entries of $\hat{\Sigma}$? In other words, can we guarantee that the matrix $\hat{\Sigma}$ converges to the matrix Σ?

For example, let's say[1] that we are interested in understanding the random variable

$$|\hat{\Sigma} - \Sigma|_\infty := \max_{i,j} |\hat{\Sigma}_{i,j} - \Sigma_{i.j}|$$

Here is an attempt using a union bound.

$$\mathbb{P}(|\hat{\Sigma} - \Sigma|_\infty > t) = \mathbb{P}(\exists (i,j) \,:\, |\hat{\Sigma}_{i,j} - \Sigma_{i.j}| > t) \leq \sum_{1 \leq i,j \leq d} \mathbb{P}(|\hat{\Sigma}_{i,j} - \Sigma_{i.j}| > t) \leq C \frac{d^2}{nt^2} ,$$

assuming that we use Chebyshev's inequality to control

$$\mathbb{P}(|\hat{\Sigma}_{i,j} - \Sigma_{i.j}| > t) \leq \frac{\operatorname{var}(\hat{\Sigma}_{i,j})}{t^2} \leq \frac{C}{nt^2} .$$

In particular, this attempt fails if d grows faster than $\sqrt{n}$, that is, $d = \omega(\sqrt{n})$. While the above attempt uses loose arguments, we can show that convergence of $\hat{\Sigma}$ to Σ fails if $d/n \to \gamma \in (0, 1]$ (asymptotically fixed aspect ratio). This regime is sometimes refereed to as the *high-dimensional asymptotic* regime. We can see this by showing that the spectrum

[1] Later in the class, we'll be interested in understanding the operator norm of $\hat{\Sigma} - \Sigma$.

of $\hat{\Sigma}$ does not converge to that of Σ even in the simple case where $\Sigma = I_d$. To that end, we can use some tools from random matrix theory (RMT)[2].

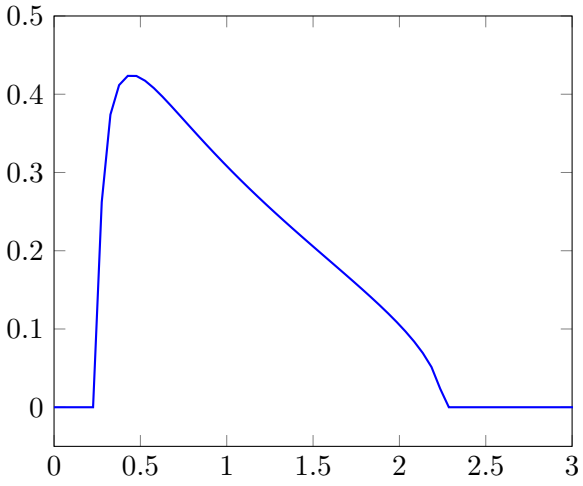
Let $\hat{\lambda}_1, \dots, \hat{\lambda}_d$ denote the eigenvalues of $\hat{\Sigma}$ and let $\lambda_1 = \dots = \lambda_d = 1$ denote those of I_d. We are going to see that the set $\{\hat{\lambda}_1, \dots, \hat{\lambda}_n\}$ does not converge to $\{1\}$. In fact it is a well known fact of RMT that

$$\frac{1}{d}\sum_{j=1}^{d} \delta_{\hat{\lambda}_j} \to T, \qquad \text{as } n \to \infty,\ d \to \infty, \frac{d}{n} \to \gamma \in (0, 1]\,,$$

where T is a random variable distributed according to the Marčenko-Pastur distribution and has density

$$f(t) = \frac{\sqrt{(\gamma_+ - t)(t - \gamma_-)}}{2\pi\gamma t} \mathbb{1}_{[\gamma_-, \gamma_+]}(t)\,, \qquad \gamma_\pm = (1 \pm \sqrt{\gamma})^2\,.$$

This density for $\gamma = 1/2$ looks like that



In particular, we can see that the eigenvalues do not concentrate at 1.

If one is interested only in the largest eigenvalue of $\hat{\Sigma}$, then it is true that $\lambda_{\max}(\hat{\Sigma}) \to \gamma_+$ and the asymptotic distribution of $\lambda_{\max}(\hat{\Sigma})$ is also known and corresponds to the Tracy-Widom distribution. In particular, one can extract the quantiles of this distribution to perform statistical inference on Σ such as hypothesis testing or confidence intervals.

Random matrix theory results are very delicate. In this class, we will see that we can get with much less effort similar qualitative results. For example, we will show that

$$\mathbb{P}(\lambda_{\max}(\hat{\Sigma}) > 1 + C(\sqrt{\frac{d}{n}} + t + \sqrt{t})) \leq e^{-nt}\,.$$

in other words, the order of the fluctuation, which is $\sqrt{d/n}$ holds for all d and n.

[2]Random matrix theory is a deep topic of probability and is not required in this class. However, knowledge of the qualitative results from this field often proves useful to get a grasp of the order of magnitude of important quantities such as the leading eigenvalue of $\hat{\Sigma}$.

1.3 Asymptotic vs non-asymptotic

In light of these examples we can see the difference between asymptotic and non-asymptotic results.

In the context of mean estimation for $d = 1$, a direct consequence of the CLT (classical asymptotic regime) is that we can build an asymptotic confidence interval for μ:

$$\mathbb{P}\big(\mu \in [\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]\big) \xrightarrow[n\to\infty]{} .95$$

It is quite useful to have exact constants (here 1.96) arising from quantiles of the standard Gaussian distribution. Similarly, sharp constants may be obtained in the high-dimensional asymptotic regime by considering the quantiles of the Tracy-Widom distribution for example.

This makes for precise confidence intervals that should be contrasted with the ones obtained in the non-asymptotic regime, using for example Hoeffding's inequality:

$$\mathbb{P}\big(\mu \in [\bar{X}_n - C\frac{\sigma}{\sqrt{n}}, \bar{X}_n + C\frac{\sigma}{\sqrt{n}}]\big) \geq .95\,,$$

where C is a constant typically larger than 1.96. Qualitatively, note that the width of the confidence interval is captured by the non-asymptotic regime: it is of order $\sigma/\sqrt{n}$.

This difference is also salient when considering hypothesis testing where it is often desirable to have sharp thresholds in order to maximize power. From this perspective, the classical asymptotic regime is more desirable.

However, note that these sharp constants are only valid as $n \to \infty$ and in particular, it requires that $n \gg d$. The high-dimensional asymptotic regime gives a partial remedy for this limitation but given n and d, it is unclear whether we are in this regime. Indeed, one may ask which of the following pairs (n, d) are in this regime:

$$(1000, 1),\ (1000, 10),\ (1000, 100),\ (1000, 1000)$$

For each of these one can estimate γ by d/n but it is unclear if asymptotic statements are valid.

Instead the non-asymptotic regime is valid for all n and d. It does not yield sharp constants but captures the performance/accuracy of the estimator $\bar{X}_n$ for all n.

In conclusion, the *classical asymptotic* or *high-dimensional asymptotic* regimes are preferred for statistical inference tasks such as confidence intervals and hypothesis testing whereas the *non-asymptotic* regime is preferred to produce a qualitative description of the performance of a possibly complicated and high-dimensional method such as the ones arising in machine learning.

In the rest of this lecture, we give an overview of the topics covered in this class from the useful perspective of the non-asymptotic regime.

2. A BRIEF OVERVIEW OF TOPICS COVERED

2.1 Empirical risk minimization

Recall our favorite statistical estimation method: maximum likelihood.

We are given a statistical model $(\mathbb{R}, \{P_\theta\}_{\theta\in\Theta})$ where the density of P_θ with respect to the Lebesgue measure is given by p_θ. Assume that we observe $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} P_{\theta^*}$ for some unknown $\theta^* \in \Theta$. The maximum likelihood estimator $\hat{\theta}$ is given by

$$\hat{\theta} = \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i)\,.$$

Wald's theorem ensures that under some technical conditions,

$$\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta^*) \to \mathcal{N}(0,1)\,.$$

where $I(\theta)$ denotes the Fisher information: $I(\theta) = -\mathbb{E}_\theta\big[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X_1)\big]$.

The maximum likelihood method belongs to a larger class of methods called *empirical risk minimization*. To see this recall that if the model is identifiable then θ^* is the unique minimizer of the expected negative log-likelihood given by

$$-\mathbb{E}_{\theta^*}[\log p_\theta(X_1)]\,.$$

Therefore, we can view the maximum likelihood method as replacing the expected value with an average and then proceeding to optimizing the resulting function. This is precisely the idea behind empirical risk minimization.

Consider a loss function $\ell(X,\theta)$ such that $\theta^* = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}_{\theta^*}\ell(X,\theta)$. The quantity $\mathbb{E}_{\theta^*}\ell(X,\theta)$ is the *risk* of θ and can can estimate it by it *empirical risk*:

$$\frac{1}{n}\sum_{i=1}^{n} \ell(X_i,\theta)\,.$$

Empirical risk minimization consists in minimizing this function over Θ:

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(X_i,\theta)\,.$$

We will consider various cases for Θ. For example

- $\Theta = \mathbb{R}^d$
- Θ is a space of smooth functions
- Θ is a combinatorial subset of the boolean hypercube $\{0,1\}^d$

While Wald's theory can apply in the first case, the other two are more tricky and asymptotic normality of $\hat{\theta}$ often does not hold. Nevertheless, we will develop some tools to quantify the accuracy of $\hat{\theta}$ in the non-asymptotic regime.

2.2 Matrix denoising

Assume that we observe a $\sqrt{d}\times\sqrt{d}$ matrix $Y = W + \Xi$ where W is a matrix of interest and Ξ is a noise matrix. Such problems arise in matrix completion and community detection for example.

Without further assumptions, estimating W is hopeless: our best guess is simply Y. In fact, we will see that we can give nontrivial guarantees to estimate W when it has a natural low-dimensional structure, for example if it has low rank.

2.3 Neural networks

The problem of fitting a neural network can be viewed as a regression problem where one observes independent copies of a predictor/response pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ that satisfy

$$Y = f(X) + \varepsilon\,,$$

where ε is a noise random variable and $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown regression function.

Neural networks have been applied successfully by fitting regression functions of the form

$$f(x) = W^L \sigma(W_{L-1}\sigma(W_{L-2} \cdots \sigma(W_1 x) \cdots)\,,$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinearity (ReLU, sigmoid, etc.) applied to each coordinate and W_j is a d_j-by-d_{j-1} matrix with $d_0 = d$ and $d_L = 1$. The number of parameters $\sum_j d_j d_{j-1}$ is typically much larger than the sample size n and cannot represent the true dimensionality of the problem if one can estimate f accurately. We will describe some recent developments on how the "true" dimensionality of such functions may be measured.

2.4 Minimax lower bounds

A typical non-asymptotic statistical guarantee is of the form

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \leq C\frac{d}{n}\,.$$

In other words, these are uniform guarantees (unlike asymptotic statements which are pointwise). For a given problem, one may ask whether we can do better, either by finding a better proof of performance for $\hat{\theta}$ or by changing the estimator $\hat{\theta}$ altogether. Indeed, since non-asymptotic results are qualitative in nature, it is important to make sure that are painting the correct picture.

A *minimax lower bound* is a result of the form

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \geq C\frac{d}{n}\,.$$

where the infimum is over all estimators (measurable functions of the data). This reads as:

> For all estimator $\hat{\theta}$, there exists $\theta \in \Theta$ that cannot be estimated (in squared norm) faster than order d/n.

Minimax lower bounds are very informative companions to non-asymptotic lower bounds. In this class we will develop a general machinery that borrows from information theory to develop minimax lower bounds systematically.

Summary: The *classical asymptotic* or *high-dimensional asymptotic* regimes are preferred for statistical inference tasks such as confidence intervals and hypothesis testing whereas the *non-asymptotic* regime is preferred to produce a qualitative description of the performance of a possibly complicated and high-dimensional method such as the ones arising in machine learning.

The object of interest often has a latent low-dimensional structure which makes seemingly impossible estimation tasks (matrix denoising, neural networks training) possible.

Minimax lower bounds allow us to assess the optimality of non-asymptotic bounds by giving the fundamental limitations of *any* estimator constructed from the data at hand.

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

Lecturer: P Rigollet
Scribes: Varkey Alumootil, Sylvia Klosin and Brice Huang

Lecture 2
Feb. 6, 2020

Goals: Last time we introduced the class and gave a brief overview of the flavor of non-asymptotic statistics as opposed to classical asymptotic and high dimensional statistics. Today we will cover probabilistic tools in this field, especially for tail bounds. In particular, we will cover subGaussian random variables, Chernoff bounds, and Hoeffding's Inequality.

A tool in classical asymptotic analysis is the central limit theorem (CLT): If $X_1, \dots, X_n \sim P$ iid with mean μ and variance σ^2, then

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

We note that if $P = \mathcal{N}(\mu, \sigma^2)$, then we actually have for all n,

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

We would like to consider a class of distributions in which we can get non-asymptotic results somewhere between the generality of the central limit theorem and the specificity of just Gaussians. This class will be the class **subGaussian** random variables. For these we will have

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

Where $\approx$ that the tails look similar, or the moment generating function are similar.

1. GAUSSIAN TAILS AND MOMENT GENERATING FUNCTIONS

Consider a standard normal random variable $Z \sim \mathcal{N}(0, 1)$. Then it has pdf given by

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

A key trait of this distribution is that the trail probability grows exponentially small.

Proposition (Mills Ratio Inequality): $\forall t > 0$,

$$\mathbb{P}[|Z| > t] \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{t^2}{2}}}{t}.$$

Proof. We have

$$\mathbb{P}[Z > t] = \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$$

However, on the interval $[t, \infty)$, $x \geq t$, so we can upper bound the right hand side by multiplying by $\frac{x}{t} \geq 1$.

$$\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x \leq \int_t^\infty \left(\frac{x}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x = \frac{1}{t\sqrt{2\pi}} \int_t^\infty x \exp(-\frac{x^2}{2}) \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2}}}{t}.$$

Since Z is symmetric (Z has the same distribution as $-Z$), we have that $\mathbb{P}[Z < -t] = \mathbb{P}[Z > t]$. Moreover,

$$\mathbb{P}[|Z| > t] = \mathbb{P}[Z > t] + \mathbb{P}[Z < -t] = 2\mathbb{P}[X > t] = 2\frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2}}}{t}.$$

□

An immediate consequence is that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then CLT says that

$$\lim_{n\to\infty} \mathbb{P}[\sqrt{n}\frac{|\bar{X}_n - \mu|}{\sigma} > t] \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{t^2}{2}}}{t}$$

To analyze the rate of convergance of the CLT, one applicable theorem is Berry-Esseen.

Theorem (Berry-Esseen): Given finite third moments of the i.i.d random variables X_i, then

$$|\mathbb{P}[\sqrt{n}\frac{|\bar{X}_n - \mu|}{\sigma} > t] - \mathbb{P}[|Z| > t)| \leq \frac{C}{\sqrt{n}}$$

for some constant C. Note that this compares the CDFs of our sample average distribution and the normal distribution.

We note that this is a 1D result, and this sort of result is still an active area of research for higher dimensions (c.f. V. Chernozhukov). Furthermore, Berry-Esseen does not give the sort of result we want. If we use triangle inequality on the Berry-Esseen result, then for a fixed n, we bound the tail probability above, but this bound is capped by a term depending on n. Restricting our random variables to only those with finite third moment gave some progress, but we will have to have more restrictions to get more useful results.

2. SUBGAUSSIAN RANDOM VARIABLES

Recall the definition of the Moment Generating Function.

Definition (Moment Generating Function (MGF)): The Moment Generating Function of a random variable X is the function $s \mapsto M(s) = \mathbb{E}[e^{sX}]$ for $s \in \mathbb{R}$.

Important Remark: Note that this does not completely identify the random variable (i.e. log normal); we need $s \in \mathbb{C}$ and the characteristic function to identify the random variable. This is good enough for our statistical purposes however.

This is called the MGF because

$$\frac{\partial^k}{\partial s^k} M(s)\Big|_{s=0} = \mathbb{E}[X^k]$$

If we can control the MGF, we can produce tail bounds using Chernoff bounds.

Definition (subGaussian random variables): A random variable X is subGaussian with variance proxy σ^2 if $\mathbb{E}X = 0$ and $\mathbb{E}[e^{sx}] \leq e^{\frac{s^2\sigma^2}{2}}$ for all $s \in \mathbb{R}$. This is written as $X \sim \mathsf{subG}(\sigma^2)$.

Note that $X \sim \mathsf{subG}(\sigma^2)$ is an abuse of notation as this is a class of distributions, not a particular one. For convenience we will assume subGaussians are centred, i.e., have mean 0.

Proposition: Let X be a random variable with mean 0 and variance 1. Then the following are equivalent:

(i) $\mathbb{E}[e^{sX}] \leq e^{c_1 s^2}$ for all $s \in \mathbb{R}$,

(ii) $\mathbb{P}[|X| \geq t] \leq 2e^{-c_2 t^2}$ for all $t \geq 0$,

(iii) $\|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}} \leq c_3\sqrt{p}$ for all $p = 1, 2 \ldots$,

(iv) $\mathbb{E}[e^{sX^2}] \leq e^{c_4 s}$ for all $s \in (0, c_4')$,

Note that (ii) is the tail bound we wanted, and (iv) says that X^2 is sub-exponential.

Proof. (i) $\Rightarrow$ (ii). We will apply a *Chernoff bound*. We note that for any positive s,

$$\mathbb{P}[X \geq t] = \mathbb{P}[e^{sX} \geq e^{st}] \leq \frac{\mathbb{E}[e^{sX}]}{e^{st}} \leq e^{c_1 s^2 - st}$$

where the first inequality is the application of Markov's inequality and the second inequality is the application of (i). This is true for any positive s, and we can minimize with respect to s:

$$\frac{\partial}{\partial s} c_1 s^2 - st = 0 \Leftrightarrow s = \frac{t}{2c_1} > 0$$

Plugging this value into the exponent yields that $\mathbb{P}[X \geq t] \leq e^{-\frac{t^2}{4c_1}}$. Similarly, we get that $\mathbb{P}[X \leq -t] \leq e^{-\frac{t^2}{4c_1}}$ and we conclude the proof of (ii) using a union bound.

(ii) $\Rightarrow$ (iii). Here we apply integration of tails which is just using Fubini's theorem to switch integrals. If $x \geq 0$ then

$$x = \int_0^x \mathrm{d}t = \int_0^\infty \mathbb{1}[t \leq x]\mathrm{d}t$$

In particular, if $X \geq 0$ a.s., then

$$\mathbb{E}X = \int_0^\infty \mathbb{P}[X \geq t]\mathrm{d}t$$

If we apply this to $|X|^p$, we get using Fubini's theorem to switch the expectation and the integral sign, we get

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}[|X|^p > t]\mathrm{d}t \leq 2\int_0^\infty e^{-c_2 t^{2/p}}\mathrm{d}t$$

We apply the change of variable $u = c_2 t^{2/p}$ to get

$$\frac{p}{c_2^{p/2}}\int_0^\infty e^{-u}u^{\frac{p}{2}-1}du$$

We recognize the integral as $\Gamma(\frac{p}{2})$, which we can upper-bound by $(\frac{p}{2})^{p/2}$. We can then produce the bound

$$\|X\|_p \leq Cp^{1/p}\sqrt{\frac{p}{2}}$$

We can however bound $p^{1/p}$ by a constant, so we get our desired result.

(iii) $\Rightarrow$ (iv).

We first employ the Taylor Expansion of the exponential function and linearity of expectation to observe that

$$\mathbb{E}[e^{sX^2}] = 1 + \sum_{p=1}^\infty \frac{s^p}{p!}\mathbb{E}X^{2p}$$

Our bound from (iii) states that for all $p > 0$,

$$\mathbb{E}X^{2p} \leq (2c_3^2 p)^p = (2c_3^2 e)^p(\frac{p}{e})^p \leq (2c_3^2 e)^p p!$$

The last inequality follows from Stirling's approximation. We substitute this to produce the bound

$$1 + \sum_{p=1}^\infty (2sc_3^2 e)^p = 1 + 2sc_3^2 e\sum_{p=0}^\infty (2sc_3^2 e)^p$$

Let us take s small enough such that $2sc_3^2 e < \frac{1}{2}$. Then this sum is bounded above by

$$1 + 2sc_3^2 e \leq e^{2sc_3^2 e}$$

as desired.

(iv) $\Rightarrow$ (i). For all $x \in \mathbb{R}$,

$$e^x \leq x + e^{x^2} \tag{2.1}$$

To verify (2.1), define the function $\psi(x) := e^{x^2} + x - e^x$. We have

$$\psi'(x) = 2xe^{x^2} + 1 - e^x\,, \qquad \psi''(x) = 2e^{x^2}(1 + 2x^2) - e^x\,,$$

Next, observe that

$$\psi''(x) \geq 2e^{x^2} - e^{|x|} = e^{|x|}\big(2e^{x^2-|x|} - 1\big) \geq \frac{e^{|x|}}{2} \geq \frac{1}{2}\,,$$

where, in the penultimate inequality, we used the fact that $e^{x^2-|x|} \geq 1 + x^2 - |x| \geq 3/4$.

In particular ψ is convex and since $\psi'(0) = 0$, it is minimized at 0 where it takes the value 0. Therefore, $\psi(x) \geq 0$ for all $x \in \mathbb{R}$. This concludes the proof of (2.1).

We now resume the proof of (iv) $\Rightarrow$ (i). Using (2.1), we get

$$\mathbb{E}[e^{sX}] \leq \mathbb{E}[sX + e^{s^2X^2}] \leq e^{c_4 s^2}$$

for $s^2 \in (0, c_4')$. Note that the expectation of sX disappears as X is centered. If $s^2 \geq c_4'$, we use a different technique. We use the fact that

$$2\lambda x \leq \delta\lambda^2 + \frac{x^2}{\delta}$$

for all $\delta > 0$. This inequality follows from $(\sqrt{\delta}\lambda - x/\sqrt{\delta})^2 \geq 0$.

Choosing of $\lambda = s/2$ and $x = X$, we get

$$\mathbb{E}[e^{sX}] \leq e^{\frac{\delta s^2}{4}}\,\mathbb{E}[e^{\frac{X^2}{\delta}}]$$

Taking $\delta = 1/c_4'$, we get from (iv) that

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2}{4c_4'}}\, e^{c_4 c_4'} \leq \exp\left(\frac{s^2}{4c_4'} + c_4 s^2\right) = e^{c_1 s^2}\,, \qquad c_1 = \frac{1}{4c_4'} + c_4$$

□

We now describe a fifth, alternative definition of subGaussianity. It relies on the general notion of Orlicz norms.

Definition (Orlicz norm): The Orlicz ψ_2-norm of a random variable X is

$$\|X\|_{\psi_2} = \inf\big\{t > 0, \quad \mathbb{E}[e^{\frac{X^2}{t^2}}] \leq 2\big\}$$

Note that not all random variables have a finite ψ_2-norm. In fact, the following proposition shows that only subGaussian ones do.

Proposition:

$$X \sim \mathsf{subG}(\sigma^2) \quad \text{iff} \quad \|X\|_{\psi_2} \leq c\sigma^2$$

Keep in mind that X is centered.

Proof. We assume that $\sigma = 1$ without loss of generality.

Assume first that $X \sim \mathsf{subG}(1)$ so that

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2}{2}}\,, \quad \forall s \in \mathbb{R}$$

Let $S \sim \mathcal{N}(0, 3/4)$ and note that the above inequality implies that

$$\mathbb{E}[e^{\frac{3X^2}{8}}] = \mathbb{E}[\mathbb{E}[e^{SX}|X]] = \mathbb{E}[\mathbb{E}[e^{SX}|S]] \leq \mathbb{E}[e^{\frac{S^2}{2}}] = 2\,.$$

Therefore, we have $\|X\|_{\psi_2}^2 \leq 4/3$.

We now prove the converse by showing part (ii) from the proposition. Set $s := 1/(2\|X\|_{\psi_2}^2)$

$$\mathbb{P}(|X| > t) \leq \mathbb{E}[e^{sX^2}]e^{-st^2} \leq 2e^{-\frac{t^2}{2\|X\|_{\psi_2}^2}}$$

□

2.1 Examples of subGaussian random variables

1. Gaussian random variables are clearly subGaussian (see, e.g., (i))
2. $X \sim \mathsf{Rad}(\frac{1}{2})$, where a Rademacher variable takes value $+1$ with probability $\frac{1}{2}$ and value -1 with probability $\frac{1}{2}$. This can be seen by checking (iv) for example:

$$\mathbb{E}[e^{sX^2}] = e^s\,.$$

3. Bounded support $|X| \leq a$ a.s. We can use (iv) here as well. In fact, sharp constants may be obtained in this case. This is the purpose of Hoeffding's inequality.

3. HOEFFDING'S INEQUALITY

We begin with a lemma that gives sharp bounds for the variance proxy of random variables with bounded support.

Lemma (Hoeffding's lemma): Let X be a centered random variable such that $x \in [a, b]$ a.s. Then

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2(b-a)^2}{8}} \quad \forall s \in \mathbb{R}$$

Proof. Define the log-MGF (a.k.a. *cumulant generating function*):

$$\psi(s) = \log \mathbb{E}[e^{sX}]\,.$$

Let us define a new probability measure $\widetilde{\mathbb{P}}$ on the real line by

$$\widetilde{\mathbb{P}}(A) = \frac{\int_A e^{sx}\mathbb{P}(\mathrm{d}x)}{\int e^{sy}\mathbb{P}(\mathrm{d}y)}\,.$$

We can compute the derivatives of the log-MGF:

$$\psi'(s) = \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]} = \widetilde{\mathbb{E}}(x)\,.$$

Moreover,

$$\begin{aligned}\psi''(s) &= \frac{\mathbb{E}[X^2 e^{sX}]\mathbb{E}[e^{sX}] - \mathbb{E}[Xe^{sX}]\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]^2} \\ &= \frac{\mathbb{E}[X^2 e^{sX}]}{\mathbb{E}[e^{sX}]} - \Big(\frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}\Big)^2 \\ &= \mathbb{E}_{\widetilde{\mathbb{P}}}[X^2] - \mathbb{E}_{\widetilde{\mathbb{P}}}[X]^2 \\ &= \widetilde{\mathrm{var}}(X)\,,\end{aligned}$$

where $\widetilde{\mathrm{var}}(X)$ denotes variance of X under the probability measure $\widetilde{\mathbb{P}}$.

Note that

$$\widetilde{\mathrm{var}}(X) = \widetilde{\mathrm{var}}\big(X - \frac{a-b}{2}\big) \leq \mathbb{E}[(X - \frac{a+b}{2})^2] \leq \frac{(b-a)^2}{4}$$

Moreover, since $\mathbb{E}[X] = 0$, we have $\psi'(0) = 0$ and it is not hard to check that $\psi(0) = 0$. By the fundamental theorem of calculus, we have

$$\psi(s) = \int_0^s \int_0^r \psi''(t)\mathrm{d}t\mathrm{d}r \leq \int_0^s \int_0^r \frac{(b-a)^2}{4}\mathrm{d}t\mathrm{d}r = \frac{s^2(b-a)^2}{8}$$

□

We are now in a position to state the main result of this section.

Theorem (Hoeffding's inequality): Let $X_1, \cdots, X_n$ be independent random variables such that $X_i \sim \mathsf{subG}(\sigma_i^2)$. Then for any $t > 0$, we have

$$\mathbb{P}[\bar{X}_n > t] \vee \mathbb{P}[\bar{X}_n < -t] \leq \exp\Big(-\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}\Big)$$

In particular, if $X_i \in [a_i, b_i]$ a.s., we have

$$\mathbb{P}[|\bar{X}_n| > t] \leq 2\exp\Big(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

Proof. We use a Chernoff bound.

$$\mathbb{P}[\bar{X}_n > t] \leq \mathbb{E}\big[e^{s\sum_{i=1}^n X_i}\big]e^{-nst} = \big[\prod_{i=1}^n \mathbb{E}\big[e^{sX_i}\big]\big]e^{-nst} \leq \big[\prod_{i=1}^n e^{\frac{\sigma_i^2 s^2}{2}}\big]e^{-nst} = e^{\frac{\sum_{i=1}^n \sigma_i^2 s^2}{2}}e^{-nst}\,.$$

We minimize the right-hand side with respect to s. To do this we want to minimize our exponent value of $\frac{\sum_{i=1}^n \sigma_i^2 s^2}{2} - nst$. Taking the derivative with respect to s and setting the derivative equal to zero gives us

$$\sum_{i=1}^n \sigma_i^2 s - nt = 0 \implies s = \frac{nt}{\sum_{i=1}^n \sigma_i^2}$$

Plugging in this value of s into our exponent gives us

$$\frac{\sum_{i=1}^n \sigma_i^2 \Big(\frac{nt}{\sum_{i=1}^n \sigma_i^2}\Big)^2}{2} - n\Big(\frac{nt}{\sum_{i=1}^n \sigma_i^2}\Big)t = \frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2} - \frac{n^2t^2}{\sum_{i=1}^n \sigma_i^2} = -\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}$$

This yields

$$\mathbb{P}[\bar{X}_n > t] \leq \exp\Big(-\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}\Big).$$

The same bound on $\mathbb{P}[\bar{X}_n < -t]$ may be obtained using the same method.

Finally, the second bound by observing that Hoeffding's lemma yields

$$X_i \sim \mathsf{subG}\big(\frac{(b_i - a_i)^2}{4}\big)$$

and applying a union bound. □

Summary: **SubGaussian random variables.** A centered r.v. X is sub-Gaussian with variance proxy 1 if either of the following equivalent definitions hold

(i) $\mathbb{E}[e^{sX}] \leq e^{c_1 s^2}$ for all $s \in \mathbb{R}$,

(ii) $\mathbb{P}[|X| \geq t] \leq 2e^{-c_2 t^2}$ for all $t \geq 0$,

(iii) $\|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}} \leq c_3\sqrt{p}$ for all $p = 1, 2\ldots$,

(iv) $\mathbb{E}[e^{sX^2}] \leq e^{c_4 s}$ for all $s \in (0, c_4')$,

(v) $\|X\|_{\psi_2} \leq c_5$

where $\|X\|_{\psi_2} = \inf\big\{t > 0, \quad \mathbb{E}[e^{\frac{X^2}{2t^2}}] \leq 2\big\}$ is the ψ_2 Orlicz norm of X.

Sums of independent subGaussian random variables. Let $X_1, \cdots, X_n$ be independent random variables such that $X_i \sim \mathsf{subG}(\sigma_i^2)$. Then for any $t > 0$, we have

$$\mathbb{P}[|\bar{X}_n| > t] \leq 2\exp\Big(-\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}\Big)$$

Hoeffding's lemma. If $X \in [a, b]$ a.s., then $X \sim \mathsf{subG}(\frac{(b_i - a_i)^2}{4})$

Hoeffding's inequality. Let $X_1, \cdots, X_n$ be independent copies of $X_i \sim \mathsf{subG}(\sigma^2)$. Then for any $t > 0$, we have

$$\mathbb{P}[|\bar{X}_n| > t] \leq 2\exp\Big(-\frac{nt^2}{2\sigma^2}\Big)$$

In particular, if $|X \in [a, b]$ a.s., then

$$\mathbb{P}[|\bar{X}_n| > t] \leq 2 \exp\Big(- \frac{2nt^2}{(b-a)^2} \Big)$$

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

Lecturer: P Rigollet
Scribes: Adam Block & Patrik Gerber

Lecture 3
Feb. 11, 2020

Goals: In the previous lecture, we introduced the notion of subGaussian random variables and explored some of their basic properties, including their connection to an Orlicz norm and Hoeffding's inequality. In this lecture, we apply analogous methods to consider a wider class of random variables: the subExponential distributions. We then introduce another Orlicz norm and prove basic properties about the subExponential random variables and their connection to the norm. We conclude by proving Bernstein's inequality, the analogue of Hoeffding's inequality in this more general regime.

1. SUBEXPONENTIAL RANDOM VARIABLES

Last lecture, we discussed subGaussian random variables and some of their properties. This is a nice, large class of random variables including, for example, the set of random variables with compact support (as seen by Hoeffding's lemma). Unfortunately, some random variables are not subGaussian. For instance, if we consider $Z \sim \mathcal{N}(0,1)$ then we might consider Z^2. Then we can directly compute the moment generating function. Indeed, set $s < \frac{1}{2}$ and we see

$$M(s) = \mathbb{E}\left[e^{sZ^2}\right] = \frac{1}{\sqrt{2\pi}} \int e^{sx^2 - \frac{1}{2}x^2} \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{x^2}{\frac{2}{1-2s}}\right) \mathrm{d}x \tag{1.1}$$

$$= \frac{1}{\sqrt{1-2s}} \frac{1}{\sqrt{\frac{2\pi}{1-2s}}} \int \exp\left(-\frac{x^2}{\frac{2}{1-2s}}\right) \mathrm{d}x = \frac{1}{\sqrt{1-2s}} \tag{1.2}$$

where the last equality follows from the fact that we are integrating the density of a centred Gaussian with variance $\frac{1}{1-2s}$. We note further that $M(s)$ is only finite for $s < \frac{1}{2}$ so we clearly cannot have $M(s) \leq e^{cs^2}$ for all s. In this lecture we consider a more general class of random variables, with heavier tails.

Definition (subExponential Random Variables): A random variable X is subExponential with parameter $\lambda > 0$ if $\mathbb{E}[X] = 0$ and for all $|s| \leq \frac{1}{\lambda}$, we have

$$M(s) = \mathbb{E}\left[e^{sX}\right] \leq e^{s^2\lambda^2} \tag{1.3}$$

We abbreviate this by saying that $X \sim \mathsf{subE}(\lambda)$.

We note as a side remark that this definition is a weakening of that of a subGaussian random variable, where the inequality of moment generating functions must hold for the entire real line. Much as in the case of subGaussian random variables, we have equivalent definitions providing tail bounds and moment estimates for subExponential random variables:

Proposition: Let X be a centred random variable. Then, the following are equivalent:

(i) There exists a constant $c_1 > 0$ such that for $|s| \leq \frac{1}{c_1}$, we have $\mathbb{E}\left[e^{sX}\right] \leq e^{s^2 c_1^2}$.

(ii) There is a constant $c_2 > 0$ such that $\mathbb{P}(|X| > t) \leq 2\, e^{-c_2 t}$.

(iii) There is a constant $c_3 > 0$ such that $\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq c_3 p$ for all $p \in \mathbb{N}$

Proof. We prove this result in much the same way that we proved the analogous proposition in the case of subGaussian random variables.

(i) implies (ii): We apply a Chernoff bound. Indeed, for $t > 0$, we have for $0 < s < \frac{1}{c_1}$

$$\mathbb{P}(X > t) \leq \mathbb{E}\left[e^{sX}\right] e^{-st} \leq e^{c_1^2 s^2 - st} \tag{1.4}$$

Now, we minimize $c_1^2 s^2 - st$ for $s \in \left(0, \frac{1}{c_1}\right)$. We know that this is a parabola and so it either attains its minimum at its vertex or at the right end point of this interval, depending on whether the abscissa of its vertex falls in this interval. Elementary calculus tells us that the vertex is at $s = \frac{t}{2c_1^2}$. Thus if $t > 2c_1$ then we have

$$\mathbb{P}(X > t) \leq e^{1 - \frac{t}{c_1}} \leq e^{-\frac{t}{2c_1}}$$

Using the same argument for $-X$ together with a union bound, we get that for $t > 2c_1$, we have

$$\mathbb{P}(|X| > t) \leq 2e^{-\frac{t}{2c_1}}$$

To deal with the case where $t \leq 2c_1$, observe that

$$\mathbb{P}(|X| > t) \leq 1 \leq \frac{2}{\sqrt{e}} \leq 2e^{-\frac{t}{4c_1}}\,.$$

where in the last inequality, we used $t \leq 2c_1$.

The above two displays yield that for every $t \geq 0$, it holds

$$\mathbb{P}(|X| > t) \leq 2e^{-c_2 t}\,, \qquad c_2 = \frac{1}{4c_1}\,.$$

(ii) implies (iii): We apply Fubini's theorem, in the same way that we did for the subGaussian case. Indeed, we note

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty \mathbb{P}\left(|X|^p > u\right) \mathrm{d}u = \int_0^\infty p t^{p-1} \mathbb{P}(|X| > t) \mathrm{d}t \leq 2 \int_0^\infty p t^{p-1} e^{-c_2 t} \mathrm{d}t \tag{1.5}$$

where we substituted $u = t^p$ and took advantage of the fact that $x \mapsto x^p$ is increasing on the positive half line, followed by the tail bound. Now, let $r = c_2 t$ to get that

$$\mathbb{E}[|X|^p] \leq \frac{2p}{c_2^p} \int_0^\infty r^{p-1} e^{-r} \mathrm{d}r = \frac{2p}{c_2^p} \Gamma(p) \tag{1.6}$$

Taking p^{th} roots and recalling that $\Gamma(p) \leq p^p$ gives

$$||X||_p \leq \frac{(2p)^{\frac{1}{p}}}{c_2} p \tag{1.7}$$

Finally, note that $(2p)^{\frac{1}{p}}$ converges as $p \to \infty$ so we may take a supremum to get that $||X||_p \leq c_3 p$ as desired.

(iii) implies (i): Let $s > 0$ and recall the definition of the exponential

$$e^{sX} = \sum_{p=0}^{\infty} \frac{s^p X^p}{p!}. \tag{1.8}$$

By the hypothesis we have

$$\mathbb{E}\ e^{sX} \leq \mathbb{E}e^{s|X|} = \sum_{p=0}^{\infty} s^p \frac{\mathbb{E}|X|^p}{p!} \leq \sum_{p=0}^{\infty} \frac{(c_3 sp)^p}{p!}. \tag{1.9}$$

We can easily find the radius of convergence of the above power series (for example using the ratio-test), which yields that $\mathbb{E}e^{s|X|} < \infty$ provided that $s < \frac{1}{ec_3}$. Thus by the Dominated Convergence Theorem, we have for $s < \frac{1}{ec_3}$ that

$$\mathbb{E}\left[e^{sX}\right] = \sum_{p=0}^{\infty} s^p \frac{\mathbb{E}X^p}{p!} \tag{1.10}$$

$$= 1 + \sum_{p=2}^{\infty} s^p \frac{\mathbb{E}X^p}{p!} \tag{1.11}$$

where we used that $\mathbb{E}X = 0$. Recall that Stirling's approximation says that $p! \geq \left(\frac{p}{e}\right)^p$. Using this and the hypothesis, we get

$$\mathbb{E}\left[e^{sX}\right] \leq 1 + \sum_{p=2}^{\infty} s^p \frac{(c_3 p)^p}{(p/e)^p} \tag{1.12}$$

$$= 1 + \sum_{p=2}^{\infty} s^p (ec_3)^p \tag{1.13}$$

$$= 1 + \frac{(ec_3 s)^2}{1 - ec_3 s}. \tag{1.14}$$

Let us further restrict s to be $s < \frac{1}{2ec_3}$. Using the trivial inequality $1 + x \leq e^x$ for all $x \geq 0$ we obtain

$$1 + \frac{(ec_3 s)^2}{1 - ec_3 s} \leq 1 + 2(ec_3 s)^2 \tag{1.15}$$

$$\leq e^{2(ec_3 s)^2} \tag{1.16}$$

so that $X \sim \mathsf{subE}(2ec_3)$ as required. □

Important Remark: Notice that we proved something slightly stronger: we proved that there exists a universal constant $c > 0$ such that

1. if (i) holds with c_1 then (ii) holds with $\frac{1}{c_2} \leq cc_1$
2. if (ii) holds with c_2 then (iii) holds with $c_3 \leq \frac{c}{c_2}$

3. if (iii) holds with c_3 then (i) holds with $c_1 \leq c\, c_3$.

In other words, all constants are within constant factor of each other.

Just as in the case of subGaussian random variables, we have another condition, equivalent to the above, in terms of the Orlicz norm. We first need to introduce the relevant Orlicz norm, however.

Definition (Orlicz ψ_1-norm): The Orlicz ψ_1-norm of a random variable X is

$$\|X\|_{\psi_1} = \inf\left\{t > 0 \,:\, \mathbb{E}\left[e^{\frac{|X|}{t}}\right] \leq 2\right\}. \tag{1.17}$$

Just like for subGaussian random variables, subExponentiality can be characterized using an Orlicz norm.

Proposition: There exist universal constants $0 < c_1, c_2 < \infty$ such that for any centered random variable X,

$$X \sim \mathsf{subE}(1) \quad \Longrightarrow \quad \|X\|_{\psi_1} \leq c_1. \tag{1.18}$$

and

$$\|X\|_{\psi_1} \leq 1 \quad \Longrightarrow \quad X \sim \mathsf{subE}(c_2). \tag{1.19}$$

Proof. Suppose that $X \sim \mathsf{subE}(1)$. In this case we have that for all $|s| < 1$,

$$\mathbb{E}\left[e^{sX}\right] \leq e^{s^2} \tag{1.20}$$

Suppose that $s \sim \gamma\rho$ where $\gamma \in (0,1)$ to be specified later and $\rho \sim \mathsf{Rad}\left(\frac{1}{2}\right)$. By Fubini's theorem, we have that

$$\mathbb{E}_s\left[\mathbb{E}_x\left[e^{sX}\right]\right] = \mathbb{E}_x\left[\mathbb{E}_s\left[e^{sX}\right]\right] = \mathbb{E}_x\left[\frac{e^{\gamma x} + e^{-\gamma x}}{2}\right] = \mathbb{E}[\cosh(\gamma X)] \tag{1.21}$$

We prove in the lemma below that $\cosh(\gamma x) \geq \frac{2}{3} e^{\frac{|x|}{2}}$. Thus we get that

$$\mathbb{E}[\cosh(\gamma X)] \geq \frac{2}{3}\mathbb{E}\left[e^{\frac{\gamma|X|}{2}}\right] \tag{1.22}$$

Rearranging, we have

$$\mathbb{E}\left[e^{\frac{\gamma|X|}{2}}\right] \leq \frac{3}{2}\mathbb{E}_s\left[e^{s^2}\right] = \frac{3}{2}e^{\gamma^2} \tag{1.23}$$

because $s^2 = \gamma^2$ because $|\rho| = 1$. Let $\gamma = \sqrt{\log\frac{4}{3}}$ and note that as $1 < \frac{4}{3} < e$ we have that $0 < \gamma < 1$. Then we see that

$$\mathbb{E}\left[e^{\frac{\gamma|X|}{2}}\right] \leq 2 \tag{1.24}$$

Thus by definition of the Orlicz norm, $\|X\|_{\psi_1} \leq \frac{2}{\gamma} = c_1$ as desired.

Suppose that $\|X\|_{\psi_1} \leq 1$. Then by Markov's inequality,

$$\mathbb{P}(|X| > t) = \mathbb{P}\left(e^{\frac{|X|}{\|X\|_{\psi_1}}} > e^{\frac{t}{\|X\|_{\psi_1}}}\right) \leq \mathbb{E}\left[e^{\frac{|X|}{\|X\|_{\psi_1}}}\right] e^{-\frac{t}{\|X\|_{\psi_1}}} \leq 2e^{-\frac{t}{\|X\|_{\psi_1}}} \leq 2e^{-t}. \quad (1.25)$$

By Remark 1 we know that there exists a universal constant c_2 such that the above implies that $X \sim \mathsf{subE}(c_2)$.

□

Before proving Bernstein's inequality, we wish to compare the norm introduced last lecture ($\|\cdot\|_{\psi_2}$) with that introduced this lecture ($\|\cdot\|_{\psi_2}$). We have the following proposition:

Proposition: Let X and Y be random variables. Then

(i) $\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$

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

Proof. **(i):** By definition of ψ_2, we have

$$\mathbb{E}\left[e^{\frac{X^2}{\|X\|_{\psi_2}}}\right] \leq 2 \quad (1.26)$$

and thus by the definition of the ψ_1 norm, we have $\|X\|_{\psi_2}^2 \geq \left\|X^2\right\|_{\psi_1}$. Moreover, for all $t > 0$ such that

$$\mathbb{E}\left[e^{\frac{X^2}{t^2}}\right] \leq 2 \quad (1.27)$$

we have $\|X\|_{\psi_2}^2 \leq t^2$ by definition. In particular, this holds for $t^2 = \left\|X^2\right\|_{\psi_1}$ and so $\|X\|_{\psi_2}^2 \leq \left\|X^2\right\|_{\psi_1}$, proving the other side. Thus they are equal.

(ii): Without loss of generality, by homogeneity of norms, we may assume that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. Recall that because $(X-Y)^2 \geq 0$, we have $|XY| \leq \frac{X^2}{2} + \frac{Y^2}{2}$. Thus we have by Cauchy-Schwarz,

$$\mathbb{E}\left[e^{|XY|}\right] \leq \mathbb{E}\left[e^{\frac{X^2}{2}} e^{\frac{Y^2}{2}}\right] \leq \sqrt{\mathbb{E}\left[e^{X^2}\right] \mathbb{E}\left[e^{Y^2}\right]} \leq 2 \quad (1.28)$$

because we have $\mathbb{E}\left[e^{X^2}\right] \leq 2$ and similarly for Y by the fact that their Orlicz norms are both one. But this implies that $\|XY\|_{\psi_1} \leq 1$ by definition. □

We note in passing that the second part of the above proposition together with the antecedent result and its analogy from last lecture together imply that if we centre the product of two subGaussian random variables, then we have a subExponential random variable. Thus, the example that we saw above, where $Z \sim \mathcal{N}(0,1)$ is subGaussian, tells us that $Z^2 - 1$ is subExponential. Indeed, we have

$$\left\|Z^2 - 1\right\|_{\psi_1} \leq \left\|Z^2\right\|_{\psi_1} + \|1\|_{\psi_1} \leq \|Z\|_{\psi_2}^2 + \|1\|_{\psi_1} < \infty \quad (1.29)$$

and so the result above implies that $Z^2 - 1$ is subExponential.

2. BERNSTEIN'S INEQUALITY

In this section we develop a tail bound for sums of independent subexponential random variables similar to Hoeffding's inequality.

Theorem (Bernstein's inequality): Let $X_1, ..., X_n$ be independent subExponential random variables and let

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \|X\|_{\psi_1}^2 \qquad \text{and} \qquad \sigma_{\max} = \max_{1\le i\le n} \|X\|_{\psi_1}. \tag{2.30}$$

Then for every $t \geq 0$ the following inequality holds

$$\mathbb{P}(|\overline{X}_n| > t) \leq 2\exp\left(-Cn\left[\frac{t^2}{\overline{\sigma}^2} \wedge \frac{t}{\sigma_{\max}}\right]\right) \tag{2.31}$$

for some positive constant C.

Proof. Since X_i is subExponential, there exists a universal constant $c > 0$ such that

$$X_i \sim \mathsf{subE}(c\,\|X_i\|_{\psi_1}) \tag{2.32}$$

holds for each i. In particular, for any $0 < s < \frac{1}{c\sigma_{\max}}$ we have

$$\mathbb{E}\left[e^{sX_i}\right] \leq e^{s^2c^2\|X_i\|_{\psi_1}^2}. \tag{2.33}$$

Once again we can use Markov's inequality and our control of the moment-generating function to derive the tail-bounds. For any $0 < s < \frac{1}{c\sigma_{\max}}$ we have

$$\begin{aligned}
\mathbb{P}(\overline{X}_n > t) &= \mathbb{P}\left(e^{s\sum_{i=1}^{n} X_i} > e^{nst}\right) && (2.34)\\
&\leq \mathbb{E}\left[e^{s\sum_{i=1}^{n} X_i}\right] e^{-stn} && (2.35)\\
&= e^{-stn}\prod_{i=1}^{n}\mathbb{E}\left[e^{sX_i}\right] && (2.36)\\
&\leq e^{-stn}\prod_{i=1}^{n} e^{c^2s^2\|X_i\|_{\psi_1}^2} && (2.37)\\
&= \exp\left(c^2s^2n\overline{\sigma}^2 - nst\right). && (2.38)
\end{aligned}$$

As usual, the next step is to minimize the RHS. Looking at the exponent, differentiating with respect to s and setting equal to 0 we get the minimizer

$$s^* = \frac{t}{2c^2\overline{\sigma}^2}. \tag{2.39}$$

Now, it might be that s^* falls outside the interval $(0, (c\sigma_{\max})^{-1})$ in which case the best possible bound is given by plugging in $s = (c\sigma_{\max})^{-1}$. In the latter case, the RHS becomes

$$\exp\left(\frac{n\overline{\sigma}^2}{\sigma_{\max}^2} - \frac{nt}{c\sigma_{\max}}\right) \leq \exp\left(-\frac{nt}{2c\sigma_{\max}}\right), \tag{2.40}$$

where we substituted $\frac{t}{2c} \geq \frac{\overline{\sigma}^2}{\sigma_{\max}}$. Summarising, we have

$$\mathbb{P}(\overline{X}_n > t) \leq \begin{cases} \exp\left(-\frac{nt^2}{4c^2\overline{\sigma}^2}\right) & \text{if } t \leq \frac{2c\overline{\sigma}^2}{\sigma_{\max}} \\ \exp\left(-\frac{nt}{2c\sigma_{\max}}\right) & \text{otherwise.} \end{cases} \tag{2.41}$$

This can conveniently be written as the expression

$$\mathbb{P}\left(\overline{X}_n > t\right) \leq \exp\left(-n\left[\frac{t^2}{4c^2\overline{\sigma}^2} \wedge \frac{t}{2c\sigma_{\max}}\right]\right). \tag{2.42}$$

Taking $C = \frac{1}{2c} \wedge \frac{1}{4c^2}$ together with a union bound yields (2.31). □

Let us compare the above result to what we've seen from the Central Limit Theorem (CLT). We get

$$\mathbb{P}\left(\sqrt{n}\overline{X}_n > t\right) \leq \exp\left(-\left[\frac{t^2}{4c^2\overline{\sigma}^2} \wedge \frac{\sqrt{n}t}{2c\sigma_{\max}}\right]\right). \tag{2.43}$$

We see that there is a window of width $2c\sqrt{n}$ where the rescaled average has subGaussian tails in line with the CLT while outside that growing window we only get a subExponential tail bound.

Summary:

- **subExponential random variables:** TFAE
  1. $\exists\, c_1 > 0$ such that $\mathbb{E}\left[e^{sX}\right] \leq e^{s^2c_1^2}$ for all $|s| < \frac{1}{c_1}$.
  2. $\exists\, c_2 > 0$ such that $\mathbb{P}(|X| > t) \leq 2e^{-c_2 t}$ for all $t \geq 0$.
  3. $\exists\, c_3 > 0$ such that $\|X\|_p \leq c_3 p$ for all $p \in bN$.
- **Orlicz ψ_1-norm**: There exist constants c_1, c_2 such that for any centered random variable X
  1. $X \sim \mathsf{subE}(1) \quad \Longrightarrow \quad \|X\|_{\psi_1} \leq c_1$
  2. $\|X\|_{\psi_1} \leq 1 \quad \Longrightarrow \quad X \sim (c_2)$.
- **Bernstein's Inequality**: For $X_1, ..., X_n$ independent subExponential random variables

$$\mathbb{P}\left(|\overline{X}_n| > t\right) \leq 2\exp\left(-Cn\left[\frac{t^2}{\overline{\sigma}^2} \wedge \frac{t}{\sigma_{\max}}\right]\right) \tag{2.44}$$

  for all $t \geq 0$ where

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \|X\|_{\psi_1}^2 \qquad \text{and} \qquad \sigma_{\max} = \max_{1 \leq i \leq n} \|X\|_{\psi_1}. \tag{2.45}$$

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

Lecturer: P Rigollet
Scribe: Andy Haupt, David Hughes

Lecture 4
Feb. 13, 2020

In the last lecture, we studied sub-exponential random variables. A random variable is sub-exponential if it is centered ($\mathbb{E}[X] = 0$) and its tail satisfies

$$\mathbb{P}[|X| > t] \leq 2e^{-ct}.$$

We also defined the ψ_1-norm

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

and showed that centered random variables with finite ψ_1-norm are sub-exponential. Next, we proved Bernstein's inequality, which provides a tail bound for sums of independent, sub-exponential random variables. More specifically, we proved that if $X_1, X_2, \ldots, X_n$ are independent and sub-exponential, then

$$\mathbb{P}[|\bar{X}_n| > t] \leq 2\exp\left(-Cn\left(\frac{t^2}{\bar{\sigma}^2} \wedge \frac{t}{\sigma_{\max}}\right)\right) \tag{0.1}$$

where $\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \|X_i\|_{\psi_1}^2$ and $\sigma_{\max} = \max_{i=1,2,\ldots,n} \|X_i\|_{\psi_1}$. The minimum in the exponential shows a change in the behavior of the tails for small deviations versus larger deviations (see Figure 1).

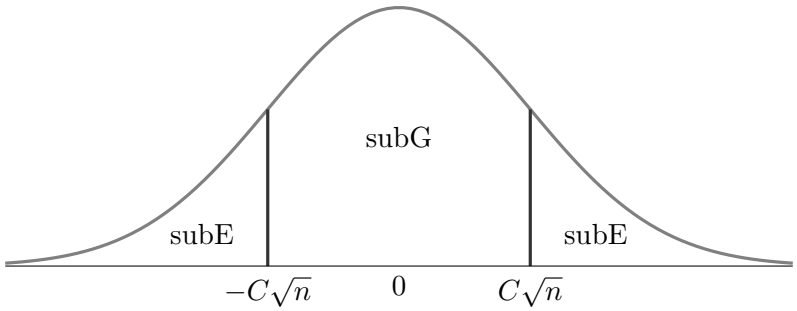


Figure 1: Probability density function of $\sqrt{n}\bar{X}_n$ for sub-exponential random variables $X_1, X_2, \ldots, X_n$. Bounds on the tails are sub-Gaussian for small deviations, but sub-exponential for $t > C\sqrt{n}$.

Alternatively, by setting the bound above equal to some δ and solving for t, Bernstein's inequality tells us that with probability at least $1 - \delta$

$$|\bar{X}_n| \lesssim \frac{\sigma_{\max}}{n}\log\left(\frac{2}{\delta}\right) + \frac{\bar{\sigma}}{\sqrt{n}}\sqrt{\log\left(\frac{2}{\delta}\right)}$$

This uses $\lesssim$ as new notation. For two sequences $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$, we denote by $a_n \lesssim b_n$ the fact that $a_n \in O(b_n)$.

Goals: Today, we will first give a more familiar formulation of the Bernstein inequality, and then apply these results to classification and mean estimation.

1. THE BERNSTEIN CONDITION

Definition: A centered random variable X satisfies the *Bernstein condition* with parameter $b > 0$ if its k-th moment satisfies the bound

$$\mathbb{E}[|X|^k] \leq \frac{\mathrm{var}(X)}{2} k! b^{k-2}. \tag{BC(b)}$$

Theorem: If $X_1, \ldots, X_n$ are independent, centered random variables that satisfy BC(b_i), for $i = 1, \ldots, n$, then

$$\mathbb{P}[|\bar{X}_n| > t] \leq 2 \exp\left(-\frac{nt^2}{2(\bar{\sigma}^2 + b_{\max} t)} \right),$$

where $\bar{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n \mathrm{var}(X_i)$ and $b_{\max} := \max_{i=1,\ldots,n}(b_i)$.

Proof. By the Chernoff bound, we have

$$\mathbb{P}[\bar{X}_n > t] \leq \prod_{i=1}^n \mathbb{E}[e^{sX_i}] e^{-nst}. \tag{1.2}$$

We can bound the moment on the right hand side using using the power series expansion for the exponential:

$$\begin{aligned}
\mathbb{E}[e^{sX_i}] &= \sum_{k=0}^{\infty} \frac{|s|^k \mathbb{E}[|X_i|^k]}{k!} \\
&\leq 1 + \frac{s^2 \mathrm{var}(X_i)}{2} + \sum_{k \geq 3} \frac{|s|^k}{k!} \frac{\mathrm{var}(X_i)}{2} k! b_i^{k-2} \\
&= 1 + \frac{s^2 \mathrm{var}(X_i)}{2} + \frac{s^2 \mathrm{var}(X_i)}{2} \sum_{k \geq 3} (|s| b_i)^{k-2} \\
&= 1 + \frac{s^2 \mathrm{var}(X_i)}{2} \sum_{k \geq 2} (|s| b_i)^{k-2},
\end{aligned}$$

where the inequality in the second line follows from the Bernstein condition. Re-indexing

and evaluating the sum gives

$$\begin{aligned}\mathbb{E}[e^{sX_i}] &\leq 1 + \frac{s^2\mathrm{var}(X_i)}{2}\sum_{k\geq 0}(|s|b_i)^k \\ &= 1 + \left(\frac{1}{1-|s|b_i}\right)\frac{s^2\mathrm{var}(X_i)}{2}, \qquad \text{for } |s| < \frac{1}{b_i} \\ &\leq \exp\left(\frac{s^2\mathrm{var}(X_i)}{2(1-|s|b_i)}\right).\end{aligned}$$

The final line uses the inequality $1 + x \leq e^x$. Combining this result with (1.2), we get

$$\mathbb{P}[\bar{X}_n > t] \leq \exp\left(\frac{s^2 n\bar{\sigma}^2}{2(1-|s|b_{\max})} - nst\right)$$

for any $|s| < \frac{1}{b_{\max}}$. Plugging in $s = \frac{t}{\bar{\sigma}^2 + b_{\max}t}$, which satisfies $|s| < \frac{1}{b_{\max}}$, the argument of the exponential reads

$$\begin{aligned}\frac{t^2}{(\bar{\sigma}^2 + b_{\max}t)^2}\frac{n\bar{\sigma}^2}{2(1-|s|b_{\max})} - \frac{nt^2}{\bar{\sigma}^2 + b_{\max}t} &= \frac{t^2}{\bar{\sigma}^2 + b_{\max}t}\left(\frac{1}{\bar{\sigma}^2 + b_{\max}t}\frac{n\bar{\sigma}^2}{2(1-\frac{tb_{\max}}{\bar{\sigma}^2+b_{\max}t})} - n\right) \\ &= \frac{t^2}{\bar{\sigma}^2 + b_{\max}t}\left(\frac{n\bar{\sigma}^2}{2\bar{\sigma}^2} - n\right) \\ &= -\frac{nt^2}{2(\bar{\sigma}^2 + b_{\max}t)},\end{aligned}$$

so that the bound simplifies to the right-hand side from the theorem. □

Note that in the proof we have shown that the moment-generating function of a random variable satisfying the Bernstein condition with parameter b is bounded by

$$\mathbb{E}[e^{sX}] \leq \exp\left(\frac{s^2\mathrm{var}(X)}{2(1-|s|b)}\right)$$

whenever $|s| < \frac{1}{b}$. This implies that these variable are sub-exponential, with parameter

$$2b \vee \sqrt{\mathrm{var}(X)}$$

by our definition of sub-exponential random variables. In other treatments of sub-exponentiality, this maximum is not necessary, as separate parameters are used to control the variance in the bound on the moment-generating function, and the range over which the bound holds. We do not cover these topics in this course.

Lemma: If $|X| \leq B$ and X is centered, then X satisfies (BC(b)) with parameter $b = B/3$.

Proof. Observe that

$$\mathbb{E}[|X|^k] \leq \mathbb{E}[|X|^2|X|^{k-2}] \leq \mathbb{E}[|X|^2 B^{k-2}] = \mathrm{var}(X)B^{k-2}$$

It remains to check that the Bernstein condition holds, that is

$$B^{k-2} \leq \frac{k!}{2}\left(\frac{B}{3}\right)^{k-2} \iff 3^{k-2} \leq \frac{k!}{2}$$

the above is clearly true for $k = 3$, and since the left-hand side increases by a factor of 3, while the right-hand increases by a factor of k, it is therefore also true for all $k \geq 3$. $\square$

Note that if the random variable is symmetric, i.e. $X \stackrel{\mathrm{d}}{=} -X$, then we can strengthen the statement by replacing the 3 with a $\sqrt{12}$. This is possible since $E[|X|^3] = 0$, so we need only control terms of order $k \geq 4$ (the proof of this is left as an exercise).

We can combine the last lemma with Bernstein's inequality to provide a tail bound for sums of bounded random variables of the form

$$\mathbb{P}\left(|\bar{X}_n| > t\right) \leq 2\exp\left(-\frac{nt^2}{2(\sigma^2 + \frac{1}{3}Bt)}\right), \quad \text{if } \mathrm{var}(X_i) \leq \sigma^2$$

The above tail bound implies that, with probability at least $1 - \delta$,

$$|\bar{X}_n| \lesssim \frac{B}{n}\log\left(\frac{1}{\delta}\right) + \frac{\sigma}{\sqrt{n}}\sqrt{\log\left(\frac{1}{\delta}\right)}. \tag{1.3}$$

The term $\frac{\sigma}{\sqrt{n}}$ usually dominates on the right-hand side; however, when σ is very small, the $\frac{B}{n}$-term will dominate.

2. ORLICZ NORM, BOUNDEDNESS AND VARIANCE

We now consider the relationship between random variables being a.s. bounded, boundedness of its ψ_1-norm and the variance of a random variable.

Boundedness implies finite Orlicz norm First, if $|X| \leq B$ holds almost surely, then

$$\mathbb{E}[e^{\frac{|X|}{t}}] \leq e^{\frac{B}{t}} \leq 2$$

holds if and only if $t \geq \frac{B}{\log(2)}$. Hence $\|X\|_{\psi_1} \leq cB$, so that bounded random variables have bounded ψ_1-norm. (This is not surprising given that we know bounded random variables are sub-Gaussian.)

Variance does not control Orlicz norm On the other hand, in general, we cannot guarantee existence of $C < \infty$ such that $\|X\|_{\psi_1}^2 \leq C\mathrm{var}(X)$. Take as an example

$$X_n = \begin{cases} \pm 1 & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{2}{n}. \end{cases}$$

Then, $\mathrm{var}(X_n) = \mathbb{E}[X_n^2] = \frac{2}{n}$. Furthermore, as

$$\mathbb{E}[e^{\frac{|X|}{t}}] = 1 - \frac{2}{n} + \frac{2}{n}e^{\frac{1}{t}} \leq 2 \iff \frac{n}{2} + 1 \geq e^{\frac{1}{t}} \iff t \geq \frac{1}{\log(\frac{n}{2}+1)} = \|X\|_{\psi_1},$$

we get that the ratio of the two sides diverges:

$$\frac{\|X_n\|_{\psi_1}^2}{\mathrm{var}(X_n)} = \frac{n}{2\log(\frac{n+2}{2})} \xrightarrow[n\to\infty]{} \infty.$$

Hence, the ψ_1-norm does not capture variance and it could be that the variance of a random variable becomes very small, even as the ψ_1-norm remains large.

3. APPLICATIONS OF BERNSTEIN'S INEQUALITY

3.1 Classification

Consider the problem of evaluating the performance of the *classifier*

$$f\colon \mathcal{X} \to \{-1, 1\},$$

which predicts labels $Y \in \{-1, 1\}$ given a set of features $X \in \mathcal{X}$. We evaluate the classifier on a set of labelled test examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ using the *indicator loss* function $Z_i = \mathbb{1}(f(X_i) \neq Y_i)$. Observe that $Z_i \sim \mathsf{Ber}(p)$, where $p = \mathbb{E}[f(X) \neq Y]$ is the expected error rate, or *classification error*.

In this case, Hoeffding's lemma applied to the random variable $Z_i - p$ (which is mean zero and is bounded in the interval $[-p, 1-p]$) implies that $(Z_i - p) \sim \mathsf{subG}\left(\frac{1}{4}\right)$. Hoeffding's inequality then gives

$$\mathbb{P}[|\bar{Z}_n - p| > t] \leq 2\exp\left(-2nt^2\right)$$

so that

$$|\bar{Z}_n - p| = \left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - p)\right| \leq \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$$

with probability $1 - \delta$.

Now imagine that $\bar{Z}_n = 0$, i.e. the classifier produces no error when tested against our sample. In this case, how can we strengthen our bound on the classification error rate p?

Since Z_i is Bernoulli, $\mathrm{var}(Z_i) = p(1-p) \leq p$, and $|Z_i| \leq B$ for $B = \max\{p, 1-p\} \leq 1$. Using Bernstein's inequality for bounded random variables (1.3) and applying $\bar{Z}_n = 0$, we see that the bound on the classification error:

$$p \lesssim \frac{\log(\frac{1}{\delta})}{n} + \sqrt{\frac{p\log(\frac{1}{\delta})}{n}}$$

holds with probability $1 - \delta$. Plugging in $\delta = \frac{1}{100}$ (any other constant would work as well), we get with 99% accuracy that

$$p \lesssim \sqrt{\frac{p}{n}} + \frac{1}{n}. \tag{3.4}$$

We can strengthen this result using a recursive argument: If $p \lesssim \frac{1}{n}$, then (3.4) can be strengthened to $p \lesssim \frac{1}{n}$. Otherwise, for large enough n, we get $p \geq \frac{c}{n}$ for a $c \gg 0$ to be chosen later. The latter is equivalent to $\frac{1}{n} \leq \frac{p}{c}$. Substituting this into (3.4), we get with high probability that $p \leq C_1 p + \frac{C_2}{n}$, where we can choose c large enough to have $C_1 < 1$. Rearranging yields also in this case $p \lesssim \frac{1}{n}$. So when $\bar{Z}_n = 0$ we only need on the order of $\frac{1}{n}$ observations to test the classification error rate.

3.2 Mean Estimation

We first consider the isotropic case $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2 I_d)$. Then, by standard properties of the multivariate normal distribution, we get for $Z \sim \mathcal{N}(0, I_d)$

$$\frac{n\|\bar{X}_n - \mu\|_2^2}{\sigma^2} \overset{\text{d}}{=} \|Z\|_2^2$$

which is χ^2_d-distributed. This implies that

$$\mathbb{E}[\|\bar{X}_n - \mu\|^2] = \frac{d\sigma^2}{n}.$$

Furthermore, $\|Z_i^2\|_{\psi_1} = \|Z_i\|_{\psi_2}^2 \leq C$, as Z_i is sub-Gaussian. Using Bernstein's inequality, we get the sub-exponential tail bound

$$\mathbb{P}\left[\left|\frac{1}{d}\sum_{i=1}^{d}(Z_i^2 - 1)\right| > t\right] \leq 2\exp(-cd(t^2 \wedge t)) =: \delta,$$

where $c > 0$ is another constant. Hence, with probability $1 - \delta$,

$$\left|\frac{1}{d}\sum_{i=1}^{d}(Z_i^2 - 1)\right| \lesssim \sqrt{\frac{\log(\frac{2}{\delta})}{d}} + \frac{\log(\frac{2}{\delta})}{d}.$$

Re-transforming to the X_i, we get

$$\left|\|\bar{X}_n - \mu\|^2 - \frac{\sigma^2 d}{n}\right| \lesssim \frac{\sigma^2}{n}\log\left(\frac{2}{\delta}\right) + \frac{\sigma^2}{n}\sqrt{d\log\left(\frac{2}{\delta}\right)}.$$

As a second application, consider the non-isotropic case $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$, where we assume for simplicity that $\mu = 0$. Then we have for a spectral decomposition $\Sigma = U\Lambda U^\top$ that

$$n\|\bar{X}_n\|^2 \overset{\text{d}}{=} n\|\Sigma^{\frac{1}{2}}\bar{Z}_n\|^2 \overset{\text{d}}{=} \|U\Lambda^{\frac{1}{2}}U^\top Z\|^2 \overset{\text{d}}{=} \|\Lambda^{\frac{1}{2}}Z\|^2 \overset{\text{d}}{=} \sum_{i=1}^{d}\lambda_i Z_i^2$$

From the definition of the ψ_1-norm, $\|\lambda_i(Z_i^2 - 1)\|_{\psi_1} = \lambda_i\|(Z_i^2 - 1)\|_{\psi_1} \leq C\lambda_i$ since $(Z_i^2 - 1)$ is sub-exponential and hence has finite norm.

Applying the version of Bernstein's inequality introduced last lecture, we get

$$\mathbb{P}\left[\frac{1}{d}\left|\sum_{i=1}^{d}\lambda_i(Z_i^2 - 1)\right| > t\right] \leq 2\exp\left(-Cd\left(\frac{t^2}{\frac{1}{d}\sum_{i=1}^{d}\lambda_i^2} \wedge \frac{t}{\lambda_{\max}}\right)\right)$$

Re-transforming to X_i, we get

$$\left|\|\bar{X}_n - \mu\|_2^2 - \frac{\operatorname{Tr}(\Sigma)}{n}\right| \lesssim \frac{\|\Sigma\|_{\text{op}}}{n}\log\left(\frac{2}{\delta}\right) + \frac{\|\Sigma\|_{\text{F}}}{n}\sqrt{\log\left(\frac{2}{\delta}\right)}.$$

where we recall that $\|\Sigma\|_{\text{op}} = \lambda_{\max}$ (operator norm) and $\|\Sigma\|_{\text{F}}^2 = \sum_{i=1}^{d}\lambda_i^2$ (Frobenius norm), and hence, the bound depends non-trivially on the spectrum of the covariance matrix.

Summary: In this lecture, we gave a different formulation of Bernstein's inequality based on the so-called *Bernstein condition*. In addition, we discovered applications of Hoeffding's and Bernstein's inequality to classification and mean estimation.

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

Lecturer: P Rigollet
Scribe: Hussein Mozannar and Arnab Sarker

Lecture 5
Feb. 20, 2020

Goals: Develop tools to characterize the behavior of the maximum of a set of random variables.

In the last two lectures we covered Bernstein's and Hoeffding's inequalities, which provide concentration inequalities on the average of independent random variables $\bar{X}_n$, and can be generally extended to linear combinations of random variables. However, we may often be concerned with the maximum of a set of random variables, for example in the study of empirical risk minimization. In this lecture, we will turn our focus on maximal inequalities to upper bound the maximum of a collection of random variables which may not necessarily be independent.

1. MAXIMUM OVER A FINITE SET

Our first problem will be to consider the maximum over a finite collection of random variables $X_1, \cdots, X_N$ which may *not necessarily be independent*.

As a first attempt, we may introduce the following inequalities:

$$\max_{1\le 1\le N} X_i \le \max_{1\le i\le N} |X_i| \le \sum_{i=1}^{N} |X_i| \,.$$

Taking expectation on both sides of the above inequality we obtain

$$\mathbb{E}\left[\max_{1\le i\le N} X_i\right] \le N \max_{1\le i\le N} \mathbb{E}\left[|X_i|\right] \,.$$

The bound above has a linear dependence on N, the size of our collection, but our analysis can be refined by considering the p-norm of the random variables.

$$\begin{aligned}
\mathbb{E}\left[\max_{1\le i\le N} X_i\right] &\le \mathbb{E}\left[\left(\max_{1\le i\le N} |X_i|^p\right)^{\frac{1}{p}}\right] && (\forall\ p \ge 1)\\
&\le \left(\mathbb{E}\left[\max_{1\le i\le N} |X_i|^p\right]\right)^{\frac{1}{p}} && \text{(Jensen's Inequality)}\\
&\le \left(\sum_{i=1}^{N} \mathbb{E}\left[|X_i|^p\right]\right)^{\frac{1}{p}}\\
&\le N^{\frac{1}{p}} \max_{1\le i\le N} \|X_i\|_p
\end{aligned}$$

We now have a polynomial dependence on N; however, we might suffer through the p-norm of X_i. As a specific example when the random variables have finite p-norms for all $p \ge 1$,

consider the case where $X_i \sim \mathsf{subG}(\sigma^2)$ for all i. Then, setting $p = \log(N)$ we obtain

$$\begin{aligned}\mathbb{E}\left[\max_{1\le i\le N} X_i\right] &\le N^{\frac{1}{\log(N)}} \max_{1\le i\le N} \|X_i\|_{\log(N)} \\ &\le ec\sigma\sqrt{\log(N)}\,, \qquad \left(N^{\frac{1}{\log(N)}} = e^{\frac{\log N}{\log N}} = e\right)\end{aligned}$$

For some constant c as discussed in Lecture 2. The following theorem makes the above result more precise, and refines the constant factor.

Theorem: Let $X_1, \dots, X_N$ be N random variables (not necessarily independent) such that $X_i \sim \mathsf{subG}(\sigma^2)$.

Then

$$\mathbb{E}\left[\max_{1\le i\le N} X_i\right] \le \sigma\sqrt{2\log(N)}\,, \quad \text{and} \quad \mathbb{E}\left[\max_{1\le i\le N} |X_i|\right] \le \sigma\sqrt{2\log(2N)}\,.$$

Moreover, for any $t > 0$,

$$\mathbb{P}\left(\max_{1\le i\le N} X_i > t\right) \le Ne^{-\frac{t^2}{2\sigma^2}}\,, \quad \text{and} \quad \mathbb{P}\left(\max_{1\le i\le N} |X_i| > t\right) \le 2Ne^{-\frac{t^2}{2\sigma^2}}\,.$$

Proof. For any $s > 0$,

$$\begin{aligned}\mathbb{E}\left[\max_{1\le i\le N} X_i\right] &\le \frac{1}{s}\mathbb{E}\left[\log e^{s\max_{1\le i\le N} X_i}\right] \\ &\le \frac{1}{s}\log\mathbb{E}\left[e^{s\max_{1\le i\le N} X_i}\right] && \text{(By Jensen's Inequality)} \\ &= \frac{1}{s}\log\mathbb{E}\left[\max_{1\le i\le N} e^{sX_i}\right] \\ &\le \frac{1}{s}\log\sum_{i=1}^{N}\mathbb{E}\left[e^{sX_i}\right] \\ &\le \frac{1}{s}\log\sum_{i=1}^{N} e^{\frac{\sigma^2 s^2}{2}} && (X_i \sim \mathrm{subG}(\sigma^2)) \\ &= \frac{\log(N)}{s} + \frac{\sigma^2 s}{2}\,.\end{aligned}$$

To obtain the tightest bound we minimize over $s > 0$ the RHS of the above bound, since it is convex we can set the derivative with respect to s to 0 and get $s = \sqrt{2\log(N)/\sigma^2}$.

For the high probability bound,

$$\begin{aligned}\mathbb{P}\left(\max_{1\le i\le N} X_i > t\right) &= \mathbb{P}\left(\bigcup_{i=1}^{N}\{X_i > t\}\right) \\ &\le \sum_{i=1}^{N}\mathbb{P}(X_i > t) && \text{(Union Bound)} \\ &\le Ne^{-\frac{t^2}{2\sigma^2}}\,.\end{aligned}$$

To prove the inequalities involving the absolute values of $|X_i|$, define $\tilde{X}_i = X_i$ and $\tilde{X}_{N+i} = -X_i$ for $i = 1, \cdots, N$, then note that:

$$\max_{1 \le i \le 2N} \tilde{X}_i = \max_{1 \le i \le N} |X_i|$$

This new collection is of course *not independent* but since we did not require independence in our proof then we can apply the results we just proved above to a collection of $2N$ random variables which amounts to replacing N by $2N$ in all the bounds.

□

One might be interested in studying the maximum or supremum over an infinite collection of random variables; however, the following simple example shows that such results may not always generalize to an infinite collection of random variables.

Let $X_1, \cdots, X_N \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. Then, for any $N \ge 1$, and any $t > 0$,

$$\begin{aligned} \mathbb{P}\left(\max_{1 \le i \le N} X_i > t \right) &= 1 - \mathbb{P}\left(\max_{1 \le i \le N} X_i \le t \right) \\ &= 1 - \mathbb{P}(X_1 \le t)^N \to 1 \quad (N \to \infty)\,. \end{aligned}$$

Therefore, in this setting, the maximum of an infinite set of random variables is unbounded almost surely. On the other hand, if $X_1 = X_2 = \cdots = X_N$, i.e., all the random variables are equal to the same random variable X_1, then we have for any $t > 0$,

$$\mathbb{P}\left(\max_{1 \le i \le N} X_i > t \right) = \mathbb{P}(X_1 > t) < 1\,, \quad \forall N \ge 1\,.$$

This simple example illustrates that if an infinite collection of random variables has a certain structure, then their maximum may in fact be finite. The following sections review other examples where we can exploit the structure of the set of random variables to analyze the behavior of the maximum of the set.

2. MAXIMUM OVER A CONVEX POLYTOPE

We now turn our attention to an uncountably infinite set of random variables. Specifically, given a random vector X taking values in $\mathbb{R}^d$, we seek to understand the set of random variables $\{\theta^\top X \mid \theta \in P\}$ where $P \subset \mathbb{R}^d$. We first consider the case in which P is a *convex polytope*, and provide a general result regarding maximization over a convex polytope.

Definition (Convex Polytope): A convex polytope P is a compact set with a finite number of vertices, $\mathcal{V}(P)$ (also called extreme points), such that $P = \mathsf{conv}\left(\mathcal{V}(P)\right)$

In the definition above, conv refers to the *convex hull* of a set of points. Formally,

$$\mathsf{conv}\left(\{v_1, \ldots, v_k\}\right) = \left\{ \sum_{i=1}^k \lambda_i v_i \;\middle|\; \lambda_i \ge 0 \;\forall i,\; \sum_{i=1}^k \lambda_i = 1 \right\}$$

We next show the following lemma regarding maximization of a linear function over a convex polytope.

Lemma: For any $c \in \mathbb{R}^d$, and any convex polytope P with extreme points $\mathcal{V}(P)$, it holds

$$\max_{x \in P} c^\top x = \max_{x \in \mathcal{V}(P)} c^\top x$$

Proof. First, note that since $\mathcal{V}(P) \subseteq P$, we must have:

$$\max_{x \in P} c^\top x \geq \max_{x \in \mathcal{V}(P)} c^\top x \,.$$

Next, fix some $x \in P$, and let $\mathcal{V}(P) = \{v_1, \ldots, v_N\}$. By definition of a convex polytope, we may write $x = \sum_{i=1}^N \lambda_i v_i$ for some non-negative values λ_i such that $\sum_{i=1}^N \lambda_i = 1$. Therefore, we may write

$$c^\top x = \sum_{i=1}^N \lambda_i c^\top v_i \leq \left(\max_{1 \leq i \leq N} c^\top v_i \right) \sum_{i=1}^N \lambda_i = \max_{1 \leq i \leq N} c^\top v_i = \max_{v \in \mathcal{V}(P)} c^\top v$$

Since the above holds for any $x \in P$, we can take the maximum over x in the left-hand side to get

$$\max_{x \in P} c^\top x \leq \max_{x \in \mathcal{V}(P)} c^\top x \,.$$

This completes the proof of the lemma. □

The above lemma leads to the following maximal inequality, as we may consider maximization over a convex polytope to have similar properties as maximization over a finite set.

Corollary: Consider a convex polytope P with N vertices $\mathcal{V}(P) = \{v_1, \ldots, v_N\}$. If for each v_i, we have $v_i^\top X \sim \mathsf{subG}(\sigma^2)$, then

$$\mathbb{E}\left[\max_{\theta \in P} \theta^\top X\right] \leq \sigma\sqrt{2\log(N)}\,, \quad \text{and} \quad \mathbb{E}\left[\max_{\theta \in P} |\theta^\top X|\right] \leq \sigma\sqrt{2\log(2N)}\,.$$

Further, for any $t > 0$,

$$\mathbb{P}\left[\max_{\theta \in P} \theta^\top X > t\right] \leq N e^{-\frac{t^2}{2\sigma^2}}\,, \quad \text{and} \quad \mathbb{P}\left[\max_{\theta \in P} |\theta^\top X| > t\right] \leq 2N e^{-\frac{t^2}{2\sigma^2}}\,.$$

Note that the condition $v^\top X \sim \mathsf{subG}(\sigma^2)$ for all $v \in \mathcal{V}(P)$ is equivalent to $v^\top X \sim \mathsf{subG}(\sigma^2)$ for all $v \in P$.

Proof. From the lemma above, we see that the following two random variables are equivalent:

$$\max_{\theta \in P} \theta^\top X = \max_{\theta \in \mathcal{V}(P)} \theta^\top X$$

Hence, we may apply the theorem from section 2 to the N random variables $v_1^\top X, \ldots, v_N^\top X$. □

The theorem above is most useful when considering polytopes with a small number of vertices. One particular convex polytope of interest is the ℓ_1 ball,

$$\mathcal{B}_1 = \left\{ x = (x^{(1)}, \ldots, x^{(d)})^\top \in \mathbb{R}^d,\ |x|_1 := \sum_{i=1}^d |x^{(i)}| \leq 1 \right\}.$$

Here, $x^{(i)}$ denotes the ith element of the vector x. The ℓ_1 ball has vertices at each of the standard basis vectors e_i (defined as the vector with a 1 in the ith position and 0's elsewhere) and their negations $-e_i$, for a total of $2d$ vertices.

3. MAXIMUM OVER THE EUCLIDEAN BALL

We now consider the case in which θ belongs to the set of vectors in the Euclidean ball,

$$\mathcal{B}_2 = \left\{ x \in \mathbb{R}^d,\ |x|_2^2 := \sum_{i=1}^d |x^{(i)}|^2 \leq 1 \right\}.$$

The Euclidean ball $\mathcal{B}_2$ is not a polytope, as has an infinite number of extreme points. However, we have that[1] $\mathcal{B}_2 \subset \sqrt{d}\mathcal{B}_1$.

Next, observe that for any $x \in \mathcal{B}_2$,

$$\begin{aligned}
\sum_{i=1}^d |x_i| &\leq \sqrt{\left(\sum_{i=1}^d |x^{(i)}|^2\right)\left(\sum_{i=1}^d 1^2\right)} && \text{(Cauchy-Schwarz inequality)} \\
&= |x|_2\sqrt{d} \\
&\leq \sqrt{d} && (x \in \mathcal{B}_2)
\end{aligned}$$

Therefore $x \in \sqrt{d}\mathcal{B}_1$ and this completes the proof that $\mathcal{B}_2 \subset \sqrt{d}\mathcal{B}_1$. Hence,

$$\max_{\theta \in \mathcal{B}_2} \theta^\top X \leq \sqrt{d} \max_{\theta \in \mathcal{B}_1} \theta^\top X = \sqrt{d} \max_{1 \leq i \leq d} |X^{(i)}|.$$

Therefore, if $X^{(i)} \sim \mathsf{subG}(\sigma^2)$ for all i, then

$$\mathbb{E}\left[\max_{\theta \in \mathcal{B}_2} \theta^\top X\right] \leq \sqrt{d}\mathbb{E}\left[\max_{1 \leq i \leq d} |X^{(i)}|\right] \leq \sigma\sqrt{2d\log(2d)}.$$

However, we can refine our analysis and remove the dependence on $\log d$. This will require the notion of a ε-net (covering).

Definition: Fix $K \subset \mathbb{R}^d$ and $\varepsilon > 0$. A set $\mathcal{N}$ is called an ε-net of K with respect to a distance $d(\cdot, \cdot)$ on $\mathbb{R}^d$, if $\mathcal{N} \subset K$ and for any $z \in K$, there exists $x \in \mathcal{N}$ such that $d(x, z) \leq \varepsilon$

If $\mathcal{N}$ is an ε-net of K with respect to a distance $d(\cdot, \cdot)$, then every point of K is at distance at most ε from a point in $\mathcal{N}$. Note that K is trivially an ε-net of K with respect to any distance. Moreover, every compact set admits a finite ε-net by definition.

We will be interested in ε-nets of small size and the following lemma gives an upper bound on the size of the smallest ε-net of $\mathcal{B}_2$.

[1] we use the notation that if $A \subset R^d$, then for $a \in \mathbb{R}$, and $b \in \mathbb{R}^d$, $aA + b$ is defined as the set $\{ax + b \mid x \in A\}$.

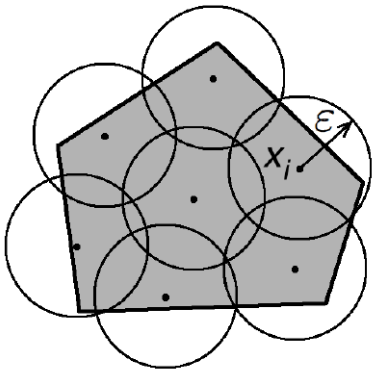


Figure 1: An example of an ε-net of size 7 with respect to the Euclidean norm for a pentagon. Figure from [Ver18]

Lemma: Fix $\varepsilon \in (0,1)$. The unit Euclidean ball $\mathcal{B}_2$ admits an ε-net $\mathcal{N}$ with respect to the Euclidean distance such that:

$$|\mathcal{N}| \leq \left(1 + \frac{2}{\varepsilon}\right)^d \leq \left(\frac{3}{\varepsilon}\right)^d$$

Proof. The following proof is an example of a "Volume argument" where we consider the ratio of the volume of the covering to the size of the set we are trying to cover. We show the proof by constructing an ε-net of $\mathcal{B}_2$ with bounded size.
Consider the following iterative construction of the ε-net:

1. Set $x_1 = 0$
2. For $i \geq 2$, take x_i to be any $x \in \mathcal{B}_2 \setminus \bigcup_{j=1}^{i-1}\{\varepsilon\mathcal{B}_2 + x_j\}$.
   If no such x_i exists, then output $\mathcal{N} = \{x_1, \cdots, x_{i-1}\}$ as the ε-net.

Clearly this procedure will create an ε-net. We now compute its size.

Observe that for any $x, y \in \mathcal{N}$, $|x - y|_2 > \varepsilon$. Hence, the set of Euclidean balls centered at x_j with radius $\varepsilon/2$ for $j = 1, \cdots, |\mathcal{N}|$ are disjoint. Their total volume is

$$\text{vol}\left(\bigcup_{j=1}^{|\mathcal{N}|}\left\{\frac{\varepsilon}{2}\mathcal{B}_2 + x_j\right\}\right) = \sum_{j=1}^{|\mathcal{N}|}\text{vol}\left(\left\{\frac{\varepsilon}{2}\mathcal{B}_2\right\}\right) = |\mathcal{N}|\left(\frac{\varepsilon}{2}\right)^d \text{vol}(\mathcal{B}_2)\,.$$

Moreover,

$$\bigcup_{j=1}^{|\mathcal{N}|}\{\frac{\varepsilon}{2}\mathcal{B}_2 + x_j\} \subset (1 + \frac{\varepsilon}{2})\mathcal{B}_2\,.$$

This is justified as the farthest point $x \in \mathcal{N}$ could lie from the origin is on the surface of $\mathcal{B}_2$, and hence the collection of balls with center $x \in \mathcal{N}$ with radius $\epsilon/2$ lie inside the enlarged

$(1+\frac{\varepsilon}{2})\mathcal{B}_2$. Translating the above set relation in terms of volumes we obtain:

$$|\mathcal{N}|\left(\frac{\varepsilon}{2}\right)^d \mathrm{vol}(\mathcal{B}_2) \leq \left(1+\frac{\varepsilon}{2}\right)^d \mathrm{vol}(\mathcal{B}_2)$$
$$\iff |\mathcal{N}| \leq \left(1+\frac{2}{\varepsilon}\right)^d .$$

Then, since $\varepsilon \in (0,1)$, $\left(1+\frac{2}{\varepsilon}\right)^d \leq \left(\frac{3}{\varepsilon}\right)^d$, completing the proof. □

Theorem: Assume that for any $u \in \mathbb{R}^d$, we have $u^\top X \sim \mathsf{subG}(\sigma^2|u|_2^2)$. Then,

$$\mathbb{E}\left[\max_{\theta\in\mathcal{B}_2} \theta^\top X\right] = \mathbb{E}\left[\max_{\theta\in\mathcal{B}_2} |\theta^\top X|\right] \leq 4\sigma\sqrt{d}$$

Moreover, for any $\delta > 0$, with probability $1-\delta$, it holds

$$\max_{\theta\in\mathcal{B}_2} \theta^\top X = \max_{\theta\in\mathcal{B}_2} \left|\theta^\top X\right| \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$$

Proof. Let $\mathcal{N}$ be a 1/2-net of $\mathcal{B}_2$ with respect to the Euclidean norm obtained from the construction in the previous Lemma. We have $|\mathcal{N}| \leq 5^d$. For every $\theta \in \mathcal{B}_2$, there exists $z \in \mathcal{N}$ and x such that $|x|_2 \leq 1/2$ and $\theta = z + x$. Therefore,

$$\max_{\theta\in\mathcal{B}_2} \theta^\top X \leq \max_{z\in\mathcal{N}} z^\top X + \max_{x\in\frac{1}{2}B_2} x^\top X$$

the rightmost term on the RHS is nothing but

$$\max_{x\in\frac{1}{2}B_2} x^\top X = \frac{1}{2}\max_{\theta\in B_2} \theta^\top X$$

Combining the inequalities above and referring to the theorem in section 2 on the maximum of a finite collection of random variables, we obtain:

$$\mathbb{E}\left[\max_{\theta\in\mathcal{B}_2} \theta^\top X\right] \leq 2\mathbb{E}\left[\max_{z\in\mathcal{N}} z^\top X\right] \leq 2\sigma\sqrt{2\log(|\mathcal{N}|)} \leq 2\sigma\sqrt{2\log(5)d} \leq 4\sigma\sqrt{d}\,.$$

For the high probability bound,

$$\mathbb{P}\left(\max_{\theta\in\mathcal{B}_2} \theta^\top X > t\right) \leq \mathbb{P}\left(2\max_{z\in\mathcal{N}} z^\top X > t\right) \leq |\mathcal{N}| e^{-\frac{t^2}{8\sigma^2}} \leq 5^d e^{-\frac{t^2}{8\sigma^2}}\,.$$

Setting the RHS equal to δ yields

$$t \geq 2\sqrt{2\log(5)}\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$$

which is sufficient thus completing the proof. □

3.1 Application: Operator Norm of a Random Matrix

Consider a random matrix $A = (A_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}}$, where each entry $A_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{subG}(\sigma^2)$. We wish to analyze the *operator norm* of the random matrix,

$$\|A\|_{\text{op}} = \max_{x \in \mathcal{B}_2(\mathbb{R}^n)} |Ax|_2 = \max_{x \in \mathcal{B}_2(\mathbb{R}^n)} \max_{y \in \mathcal{B}_2(\mathbb{R}^m)} y^\top A x \,.$$

Proposition: For the random matrix A defined above,

$$\mathbb{E}\left[\|A\|_{\text{op}}\right] \le 2\sigma\sqrt{2(n+m)\log 9}\,.$$

Proof. Let $\mathcal{N}_n$ and $\mathcal{N}_m$ be $\frac{1}{4}$-nets of $\mathcal{B}_2(\mathbb{R}^n)$ and $\mathcal{B}_2(\mathbb{R}^m)$, respectively. We may select $\mathcal{N}_n$ and $\mathcal{N}_m$ such that $|\mathcal{N}_n| \le 9^n$, and $|\mathcal{N}_m| \le 9^m$, from the arguments above. Then, since any $x \in \mathcal{B}_2(\mathbb{R}^n)$ may be written as $z + \delta$ for some $z \in \mathcal{N}_n$ and $\delta \in \frac{1}{4}\mathcal{B}_2(\mathbb{R}^n)$,

$$\begin{aligned}
\|A\|_{\text{op}} &= \max_{x \in \mathcal{B}_2(\mathbb{R}^n)} |Ax|_2 \\
&\le \max_{z \in \mathcal{N}_n} |Az|_2 + \max_{\delta \in \frac{1}{4}\mathcal{B}_2(\mathbb{R}^n)} |A\delta|_2 \\
&= \max_{z \in \mathcal{N}_n} |Az|_2 + \frac{1}{4}\|A\|_{\text{op}} \\
&= \max_{z \in \mathcal{N}_n} \max_{y \in \mathcal{B}_2(\mathbb{R}^m)} y^\top A z + \frac{1}{4}\|A\|_{\text{op}} \,. \qquad (3.1)
\end{aligned}$$

Next, for fixed $z \in \mathcal{N}_n$, we similarly note

$$\begin{aligned}
\max_{y \in \mathcal{B}_2(\mathbb{R}^m)} y^\top A z &\le \max_{w \in \mathcal{N}_m} w^\top A z + \max_{\delta \in \frac{1}{4}\mathcal{B}_2(\mathbb{R}^m)} \delta^\top A z \\
&\le \max_{w \in \mathcal{N}_m} w^\top A z + \frac{1}{4}\|A\|_{\text{op}} \,.
\end{aligned}$$

Combining the above inequality with (3.1) and rearranging, we get

$$\|A\|_{\text{op}} \le 2 \max_{z \in \mathcal{N}_n} \max_{w \in \mathcal{N}_m} w^\top A z \,.$$

Further, we note that each random variable $w^\top A z$ can be written as $\sum_{i,j} w^{(i)} A_{ij} z^{(j)}$, which means that each random variable $w^\top A z \sim \mathsf{subG}(\sigma^2 |w|_2^2 |z|_2^2)$ as each A_{ij} is independent. Since w and z lie within the Euclidean ball, their norms are at most one, and we see that $\|A\|_{\text{op}}$ is bounded by the maximum of 9^{n+m} random variables which are sub-Gaussian with variance proxy σ^2. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|A\|_{\text{op}}\right] &\le 2\mathbb{E}\left[\max_{z \in \mathcal{N}_n} \max_{w \in \mathcal{N}_m} w^\top A z\right] \\
&\le 2\sigma\sqrt{2\log(9^{n+m})} \\
&= 2\sigma\sqrt{2(n+m)\log 9}\,,
\end{aligned}$$

completing the proof. □

Exercise. Provide a high probability bound on the size of $\|A\|_{\text{op}}$.

Summary: In this lecture, we developed tools for characterizing the behavior of the maximum of a set of variables. We began by considering the case in which the set of random variables is finite, and showed that the expectation of the maximum of N random variables is bounded by a term on the order of $\sqrt{\log(N)}$.

We then considered the case in which the set of random variables is characterized by a convex polytope, and found a similar bound– when the convex polytope has N vertices, the expected maximum of associated random variables again is bounded by a term on the order of $\sqrt{\log(N)}$.

Finally, we considered the case when the set of random variables is characterized by the Euclidean ball in $\mathbb{R}^d$, and found the expected maximum of associated random variables to be bounded by a term on the order of $\sqrt{d}$ – we then extended the reasoning to characterize the operator norm of a random matrix with independent sub-Gaussian entries.

References

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

Lecturer: P Rigollet
Scribe: Alex Gu, Chandler Squires

Lecture 6
Feb. 25, 2020

Goals: In this lecture, we introduce the linear regression model. We present the least squares estimator for this model, and develop results on the finite sample performance of this estimator. Then, we introduce constrained least squares estimation, and develop results for estimator when the regression coefficients are constrained to the ℓ_1-ball.

1. LINEAR REGRESSION SETUP

In this section, we setup the linear regression model. We observe n pairs (X_i, Y_i), such that

$$Y_i = f(X_i) + \varepsilon_i, 1 \le i \le n$$

with $\mathbb{E}\varepsilon_i = 0$. We consider the fixed design model, in which the X_i's are deterministic. In a fixed design setting, we wish to estimate some $\mu = (f(X_1), \cdots, f(X_n))^\top \in \mathbb{R}^n$, given the vector $\mu + \varepsilon$.

In other words, we wish to reconstruct a function $\hat{f}_n$ from our n given samples such that $\hat{f}_n(X_i)$ are as close to the original $f(X_i)$ as possible. We measure the performance of $\hat{f}_n$ using the *mean squared error*, given by

$$\mathsf{MSE}(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(x_i) - f(x_i))^2 = \frac{1}{n}|\hat{\mu_n} - \mu|_2^2$$

where $\hat{\mu_n} = (\hat{f}(X_1), \cdots, \hat{f}(X_n))^\top$.

In linear regression, we consider the class of linear functions f given by $f(x) = x^\top\theta^*$, where θ^* is unknown. This can be rewritten as follows:

$$\begin{bmatrix} | \\ \mu_i \\ | \end{bmatrix} = \begin{bmatrix} | \\ x_i^\top\theta^* \\ | \end{bmatrix} \Rightarrow \mu = \mathbb{X}\theta^*, \qquad \mathbb{X} = \begin{bmatrix} — & x_1^\top & — \\ — & x_2^\top & — \\ & \vdots & \end{bmatrix}$$

We also consider the set of linear candidate estimators $\hat{\mu} = \mathbb{X}\hat{\theta}$. The mean squared error can then be written as

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) = \frac{1}{n}|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 = (\hat{\theta} - \theta^*)^\top\frac{\mathbb{X}^\top\mathbb{X}}{n}(\hat{\theta} - \theta^*)$$

2. THE LEAST SQUARES ESTIMATOR

The least squares estimator is defined as any estimator that minimizes the mean squared error, that is,

$$\hat{\theta}^{\mathrm{LS}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} |Y - \mathbb{X}\theta|_2^2$$

Theorem: Let $\mathbb{X}$, Y be defined as in the previous section. The least-squares estimator is given by

$$\hat{\theta}^{\text{LS}} = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y$$

where $A^\dagger$ is the Moore-Penrose pseudoinverse of a matrix A.

Proof. We have

$$|Y - \mathbb{X}\theta|_2^2 = |Y|^2 + \theta^\top \mathbb{X}^\top \mathbb{X}\theta - 2\theta^\top \mathbb{X}^\top Y$$

Since the MSE function is convex, the estimator that minimizes the MSE must satisfy

$$\nabla_\theta |Y - \mathbb{X}\theta|_2^2 \Big|_{\theta=\hat{\theta}^{\text{LS}}} = 0 \Rightarrow 2\mathbb{X}^\top \mathbb{X}\hat{\theta}^{\text{LS}} - 2\mathbb{X}^\top Y = 0$$

or

$$\mathbb{X}^\top \mathbb{X}\hat{\theta}^{\text{LS}} = \mathbb{X}^\top Y$$

A solution must exist, since the column space of $\mathbb{X}^\top \mathbb{X}$ is equal to the column space of $\mathbb{X}^\top$. Moreover,

$$\hat{\theta}^{\text{LS}} = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y$$

is a solution since $\mathbb{X}^\top \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top = \mathbb{X}^\top$.

Remark. For a matrix $\mathbb{X}$, if $\mathbb{X}^\top \mathbb{X}$ is not invertible (i.e., the x_i's belong to a subspace of $\mathbb{R}^d$), then $\ker(\mathbb{X}^\top \mathbb{X}) \neq \{0\}$ where $\forall a \in \ker(\mathbb{X}^\top \mathbb{X})$, we have $\mathbb{X}^\top \mathbb{X} a = 0$. Each θ can be divided into two parts, $\theta_1 \in \ker^\perp(\mathbb{X}^\top \mathbb{X})$, and $\theta_2 \in \ker(\mathbb{X}^\top \mathbb{X})$, such that $\theta = \theta_1 + \theta_2$. The Moore-Penrose pseudoinverse picks the solution with $\theta_2 = 0$, which is the solution that minimizes $|\theta|_2$.

If the noise ε is centered and subGaussian, we can bound the expectation of the MSE of the least-squares solution, as shown in the following theorem:

Theorem: Assume the linear regression model $Y = \mathbb{X}\theta^* + \varepsilon_i$, where ε_i are independent and in $\mathsf{subG}(\sigma^2)$. Then,

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}})] \lesssim \frac{\sigma^2 r}{n}$$

where $r = \text{rank}(\mathbb{X}^\top \mathbb{X}) \leq d \wedge n$.

Moreover, for $\delta > 0$, we have with probability $1 - \delta$,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}) \lesssim \frac{\sigma^2}{n} \log(\frac{1}{\delta}) + \frac{\sigma^2 r}{n}$$

Pre-proof. We first provide some intuition around this result. In the proof, we write $\hat{\theta}$ to mean $\hat{\theta}^{\text{LS}}$. First, consider $\hat{\mu} = \mathbb{X}\hat{\theta} = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y$. Let $P = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top$. Observe that P is a projection matrix, because

$$P^2 = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top = P$$

In fact, it is the projection of matrix Y onto the column span of $\mathbb{X}$. Therefore, if $Y = \mathbb{X}\theta^* + \varepsilon$, then $PY = P\mathbb{X}\theta^* + P\varepsilon = \mathbb{X}\theta^* + P\varepsilon$, because we know $\mathbb{X}\theta^*$ is in the column span of $\mathbb{X}$. Therefore, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) = \frac{1}{n}|\mathbb{X}\theta^* - \mathbb{X}\hat{\theta}|_2^2 = \frac{1}{n}|\mathbb{X}\theta^* - PY|_2^2 = \frac{1}{n}|\mathbb{X}\theta^* - \mathbb{X}\theta^* - P\varepsilon|_2^2 = \frac{1}{n}|P\varepsilon|_2^2$$

Observe that if $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then $|P\varepsilon|_2 \sim |\mathcal{N}(0, \sigma^2 I_r)|_2$, which makes $|P\varepsilon|_2^2 \sim \sigma^2 \chi_r^2$. Since $\mathbb{E}[\sigma^2 \chi_r^2] = \sigma^2 \mathbb{E}[\chi_r^2] = \sigma^2 r$, we get that

$$\mathbb{E}[\mathsf{MSE}(X\hat{\theta})] = \frac{\sigma^2 r}{n}$$

Now, we show how to prove the statement for general ε.

Proof. We first use the basic inequality:

$$|Y - \mathbb{X}\hat{\theta}|_2^2 \leq |Y - \mathbb{X}\theta|_2^2, \quad \forall \theta \in \mathbb{R}^d$$

In particular, we can take $\theta = \theta^*$ in the right-hand side, so that

$$|Y - \mathbb{X}\hat{\theta}|_2^2 \leq |Y - \mathbb{X}\theta^*|_2^2$$

Since $Y = \mathbb{X}\theta^* + \varepsilon$, the inequality can be written as

$$|\mathbb{X}\theta^* + \varepsilon - \mathbb{X}\hat{\theta}|_2^2 \leq |\mathbb{X}\theta^* + \varepsilon - \mathbb{X}\theta^*|_2^2$$

Cancelling and expanding both sides:

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 + |\varepsilon|_2^2 - 2\langle \varepsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^* \rangle \leq |\varepsilon|_2^2$$

Then, rearranging:

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 2\langle \varepsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^* \rangle$$

And dividing by a factor of $|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2$,

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2 \leq 2\langle \varepsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2} \rangle$$

Then squaring both sides:

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 4\langle \varepsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2} \rangle^2$$

Normally, we would think of applying Cauchy-Schwarz, which would give us a bound of

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 4\langle \varepsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2} \rangle^2 \leq 4|\varepsilon|^2 \left| \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2} \right|^2 = 4|\varepsilon|^2$$

Looking at what we want to prove, we notice that there is a dependence on the rank of the matrix r. However, after applying Cauchy-Schwarz, we completely got rid of this dependence. Hence, this is a dead end, and we need to try a different approach.

Now, let $\Phi \in \mathbb{R}^{n\times r} = [\phi_1, \phi_2, \cdots, \phi_r]$, where $\phi_1, \cdots, \phi_r$ comprise an orthonormal basis of the column span of $\mathbb{X}$. This means that there exists $\nu \in \mathbb{R}^r$ such that $\mathbb{X}(\hat{\theta} - \theta^*) = \Phi\nu$, or

$$|\mathbb{X}(\hat{\theta} - \theta^*)|_2 = |\Phi\nu|_2 = |\nu|_2$$

Therefore, we can write the MSE as

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 4\langle \varepsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2}\rangle^2 = 4\langle \varepsilon, \frac{\Phi\nu}{|\nu|_2}\rangle^2 = 4\langle \Phi^\top\varepsilon, \frac{\nu}{|\nu|_2}\rangle^2 \leq 4|\tilde{\varepsilon}|_2^2$$

where $\tilde{\varepsilon} = \Phi^\top\varepsilon \in \mathbb{R}^r$ and the last inequality follows by Cauchy-Schwarz. We claim that $\tilde{\varepsilon}^\top u \sim \mathsf{subG}(\sigma^2|u|_2^2)$ for $u \in \mathbb{R}^r$. To see this, notice that

$$\begin{aligned}
\mathbb{E}[\exp\{s\tilde{\varepsilon}^\top u\}] &= \mathbb{E}[\exp\{s\varepsilon^\top \Phi u\}] \\
&= \prod_{i=1}^r \mathbb{E}[\exp\{s\varepsilon_i(\Phi u)_i\}] \\
&\leq \prod_{i=1}^r \exp\left\{\frac{\sigma^2 s^2 (\Phi u)_i^2}{2}\right\} \\
&= \exp\left\{\frac{\sigma^2 s^2 |\Phi u|_2^2}{2}\right\} \\
&= \exp\left\{\frac{\sigma^2 s^2 |u|_2^2}{2}\right\}
\end{aligned}$$

where the inequality comes from the fact that all the ε_i are $\mathsf{subG}(\sigma^2)$. This shows that $\tilde{\varepsilon}^\top u \sim \mathsf{subG}(\sigma^2|u|_2^2)$.

Thus, $\varepsilon_j = \langle \tilde{\varepsilon}, e_j\rangle$ is $\mathsf{subG}(\sigma^2)$, and thus has variance at most σ^2. Therefore,

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \frac{1}{n} \cdot 4\mathbb{E}[|\tilde{\varepsilon}|_2^2] = \frac{4}{n}\sum_{j=1}^r \mathbb{E}[\tilde{\varepsilon}_j^2] \leq \frac{4r\sigma^2}{n}$$

Using the high-probability bounds developed in the last lecture for maximizing over the unit ball, we have

$$\mathbb{P}[|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 > t] \leq \mathbb{P}[4\langle \tilde{\varepsilon}, \frac{\nu}{|\nu|_2}\rangle^2 > t] \leq \mathbb{P}\left[\sup_{u\in\mathcal{B}_2(\mathbb{R}^r)} |\langle \tilde{\varepsilon}, u\rangle| > \frac{\sqrt{t}}{2}\right] \leq \exp\left\{\frac{-t}{32\sigma^2} + r\log 5\right\}$$

Therefore, we get that with probability $1-\delta$, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) \lesssim \frac{\sigma^2}{n}\log(\frac{1}{\delta}) + \frac{\sigma^2 r}{n}$$

3. CONSTRAINED LEAST SQUARES

We may wish to consider cases where $\mu = X\theta^*$ for $\theta^* \in K \subsetneq \mathbb{R}^d$, and constrain our minimization to K, i.e., solve the constrained least squares problem

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in K} |Y - \mathbb{X}\theta|_2^2$$

In this lecture, we consider the case $K = B_1 = \{x \in \mathbb{R}^d : |x|_1 \leq 1\}$. Recall that B_1 has $2d$ vertices at $\pm e_j$ for $j = 1, \ldots, d$. We will establish the following result:

Theorem: Assume the linear regression model $Y_i = X_i^\top \theta^* + \varepsilon_i$, with ε_i independent and $\mathsf{subG}(\sigma^2)$. Further, let $\theta^* \in B_1$, and $\max_j |\mathbb{X}_j|_2 \leq \sqrt{n}$, where $\mathbb{X}_j$ is the j-th column of $\mathbb{X}$. Denote the constrained least square solution $\hat{\theta} \in \operatorname{argmin}_{\theta \in B_1} |Y - \mathbb{X}\theta|_2$. Then

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \lesssim \sigma \sqrt{\frac{\log 2d}{n}}$$

and with probability $1 - \delta$, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) \lesssim \sigma \sqrt{\frac{\log \frac{2d}{\delta}}{n}}$$

Proof. We start with the basic inequality,

$$|Y - \mathbb{X}\hat{\theta}|^2 \leq |Y - \mathbb{X}\theta^*|^2$$

Once again substituting $Y = \mathbb{X}\theta^* + \varepsilon$, cancelling, and re-arranging to obtain

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|^2 \leq 2\langle \varepsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^* \rangle$$

Instead of using the fixed-point argument as in the unconstrained case, we will replace the right-hand side with a worst-case bound over the whole image of B_1 under the linear transformation $\mathbb{X}$

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 4 \sup_{v \in \mathbb{X}B_1} \langle \varepsilon, v \rangle$$

As established in the last lecture, we need only maximize over the vertices of $\mathbb{X}B_1$, i.e. over the columns of $\mathbb{X}$. By our assumptions on ε and $\mathbb{X}_j$, we have $\mathbb{X}_j^\top \varepsilon \sim \mathsf{subG}(\sigma^2 n)$. Thus, we meet the conditions for the maximal inequality from the last lecture and have

$$\mathbb{E}[\max_j |\mathbb{X}_j^\top \varepsilon|] \lesssim 2\sigma\sqrt{n}\sqrt{\log(2d)}$$

Hence,

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \frac{4}{n}\mathbb{E}[\max_j |\mathbb{X}_j^\top \varepsilon|] \lesssim \sigma \sqrt{\frac{\log(2d)}{n}}$$

Similarly, using the high-probability bounds developed in the last lecture for maximizing over a convex polytope, we have

$$\mathbb{P}(\mathsf{MSE}(\mathbb{X}\hat{\theta}) > t) \leq \mathbb{P}(\max_j \mathbb{X}_j^\top \varepsilon > \frac{nt}{4}) \leq 2d \exp\{-\frac{nt^2}{32\sigma^2}\}$$

Thus, picking

$$t = \sigma \sqrt{32 \frac{\log(\frac{2d}{\delta})}{n}}$$

bounds the right-hand side by δ.

□

Summary: In this lecture, we developed finite-sample results for linear regression with a fixed design matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ and $\mathsf{subG}(\sigma^2)$ noise, showing that the expectation of the mean squared error is bounded up to constant factor by $\frac{\sigma^2 r}{n}$, where $r = \text{rank}(\mathbb{X}^\top \mathbb{X})$.

We then considered constrained linear regression over the ℓ_1 ball, with additional constraints on the columns of $\mathbb{X}$.

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

Lecturer: P Rigollet
Scribe: Kayhan Behdin, Wei Fang

Lecture 7
Feb. 27, 2020

Goals: In the last lecture we introduced the linear regression model with fixed design, and provided solutions using least-squares and constrained least-squares estimators. In this lecture, we continue to analyze linear regression. Specifically, we assume that the underlying model in the regression is sparse (i.e. has many zeros). We consider two scenarios where we know the true sparsity or we do not, and provide analysis for each case.

1. SPARSITY IN LINEAR REGRESSION

Recall that the linear regression model with subGaussian noise is

$$Y = \mathbb{X}\theta^* + \varepsilon,$$

where $\varepsilon^\top u \sim \mathsf{subG}(\sigma^2|u|_2^2)\ \forall u \in \mathbb{R}^d$. In this section, we consider the case that θ^* is s-sparse, meaning that it has s non-zero coordinates where $s \ll d$. Additionally, for this first section we assume that s is known.

In the last lecture, we introduced constrained least-squares estimators by restricting θ^* inside the ℓ_1 ball. In this lecture, rather than the ℓ_1 ball we consider the ℓ_0 "ball". Formally we define the ℓ_0 ball as

$$\mathcal{B}_0(s) \coloneqq \{\theta \in \mathbb{R}^d : \sum_{j=1}^{d} \mathbb{I}(\theta_j \neq 0) \leq s\},$$

and if $\theta^* \in \mathcal{B}_0(s)$ we say θ^* is s-sparse. $\mathcal{B}_0(s)$ includes all vectors in $\mathbb{R}^d$ with up to s non-zero coordinates. Another way to describe this space is to view it as an union of subspaces of dimension s that are aligned with the coordinate axes. As an example, we can see how this relates to linear regression $y = \sum_{j=1}^{d} \theta_j^* X_j + \varepsilon$ by observing that in this model, $\theta_j^* = 0 \iff$ "X_j does not enter the regression".

With the definition of $\mathcal{B}_0(s)$, we consider the estimator $\hat{\theta}_{\mathcal{B}_0(s)}$ where

$$\hat{\theta}_{\mathcal{B}_0(s)} \in \operatorname*{argmin}_{\theta \in \mathcal{B}_0(s)} |Y - \mathbb{X}\theta|_2^2.$$

To analyze this estimator, we again utilize the basic inequality

$$|Y - \mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)}|_2^2 \leq |Y - \mathbb{X}\theta^*|_2^2.$$

Rearranging we get

$$|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2^2 \leq 2\langle \varepsilon, \mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^* \rangle.$$

Then we use the fixed point trick, as we did in the previous lecture, giving us

$$|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2 \leq 2\langle \varepsilon, \frac{\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2} \rangle.$$

Next, we control the supremum in the ℓ_0 ball:

$$|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2^2 \leq 4\big(\max_{\substack{S\subset[d]\\ |S|\leq 2s}} \sup_{v:\ \mathrm{supp}(v)\subset S} \langle \varepsilon, \frac{\mathbb{X}v}{|\mathbb{X}v|_2}\rangle^2\big).$$

Now, similar to the previous lecture, we introduce $\Phi_S \in \mathbb{R}^{n\times 2s}$, consisting of an orthonormal basis for the span of the columns of $\mathbb{X}$, $\{\mathbb{X}_j : j \in S\}$. Note that when the span is rank-deficient, we pad the basis up to $2s$. We can now write $\mathbb{X}v = \Phi_S\nu$, where ν are the coordinates in this new coordinate system, and in particular, we know that $|\Phi_S\nu|_2 = |\nu|_2$, thus

$$\sup_{v:\ \mathrm{supp}(v)\subset S} \langle \varepsilon, \frac{\mathbb{X}v}{|\mathbb{X}v|_2}\rangle^2 \leq \sup_{\nu\in\mathcal{B}_2(\mathbb{R}^{2s})} \langle \Phi_S^\top\varepsilon, \nu\rangle^2.$$

With this inequality, we can now bound the MSE using the union bound:

$$\begin{aligned}
\mathbb{P}[|X\hat{\theta}_{\mathcal{B}_0(s)} - X\theta^*|_2^2 > t] &\leq \mathbb{P}[4\max_{|S|=2s}\sup_{\nu\in\mathcal{B}_2(\mathbb{R}^{2s})}\langle\Phi_s^\top\varepsilon,\nu\rangle^2 > t] \\
&\leq \sum_{|S|=2s}\mathbb{P}[\sup_{\nu\in\mathcal{B}_2(\mathbb{R}^{2s})}\langle\Phi_s^\top\varepsilon,\nu\rangle > \frac{\sqrt{t}}{2}] \\
&\leq \binom{d}{2s}\exp(-\frac{t}{32\sigma^2} + 2s\log 5) \\
&= \exp\big(-\frac{t}{32\sigma^2} + 2s\log 5 + \log\binom{d}{2s}\big).
\end{aligned}$$

In the third inequality, notice that $\tilde{\varepsilon} = \Phi_s^\top\varepsilon$, $\tilde{\varepsilon}^\top u \sim \mathsf{subG}(\sigma^2|u|_2^2)\ \forall u \in \mathbb{R}^{2s}$, so the tail $\mathbb{P}[\sup_{\nu\in\mathcal{B}_2(\mathbb{R}^{2s})}\langle\Phi_s^\top\varepsilon,\nu\rangle > \frac{\sqrt{t}}{2}]$ is bounded by $5^{2s}\exp(-\frac{(\frac{\sqrt{t}}{4})^2}{2\sigma^2})$ using the theorem found in Section 3, Lecture 5, with the term 5^{2s} corresponding to the half-net of $\mathcal{B}_2(\mathbb{R}^{2s})$. Additionally, note that in the first inequality we set $|S| = 2s$ and not $|S| < 2s$ since we padded the orthonormal basis.

Before bounding the log term, notice that without the log term this is exactly the least-squares bound that we would get if we were told which of the coordinates of θ^*, up to $2s$ dimensions, are nonzero. The price we are paying for not knowing which of those coordinates are nonzero is exactly the log of the choices that we have.

In order to bound the term $\log\binom{d}{2s}$, we claim that $\binom{d}{k} \leq (\frac{ed}{k})^k$ and prove by induction. For $k = 1$, $d \leq (ed)$ holds. Suppose true for k, so for $k+1$ we have

$$\begin{aligned}
\binom{d}{k+1} &= \frac{d!}{(k+1)!(d-k-1)!} = \binom{d}{k}\frac{(d-k)}{(k+1)} \\
&\leq \frac{e^k d^k}{k^k}\cdot\frac{(d-k)}{(k+1)} \leq \frac{e^k d^{k+1}}{k^k(k+1)} \\
&\leq \frac{e^k d^{k+1}}{(k+1)^{k+1}}\cdot \underbrace{\frac{(k+1)^{k+1}}{k^k(k+1)}}_{=(1+\frac{1}{k})^k = e^{k\log(1+\frac{1}{k})} \leq e^{\frac{k}{k}} = e} \\
&\leq (\frac{ed}{k+1})^{k+1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}[|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2^2 > t] &\leq \exp\big(-\frac{t}{32\sigma^2} + 2s\log 5 + \log\binom{d}{2s}\big) \\ &\leq \exp\big(-\frac{t}{32\sigma^2} + 2s\log 5 + 2s\log(\frac{ed}{2s})\big).\end{aligned}$$

Setting $\delta \coloneqq \exp\big(-\frac{t}{8\sigma^2} + 2s\log 5 + 2s\log(\frac{ed}{2s})\big)$ results in

$$t \lesssim \sigma^2 s + \sigma^2 s\log(\frac{d}{s}) + \sigma^2\log(1/\delta) \lesssim \sigma^2 s\log(\frac{d}{s\delta}).$$

Notice that the first term comes from the cardinality of the ℓ_2 ball of size $2s$, and the second term comes from the number of subsets of size s. Thus we are not paying only a log factor for not knowing where the sparsity is, but instead a full additional term that is a log factor larger than the original.

Finally, with probability $1-\delta$, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)}) \leq \frac{\sigma^2 s}{n}\log(\frac{d}{s\delta}).$$

2. GAUSSIAN SEQUENCE MODEL

For this section, we assume $\frac{\mathbb{X}^\top}{\sqrt{n}}\frac{\mathbb{X}}{\sqrt{n}} = I_d$ which we call orthogonal design. Let's consider the linear regression model with Gaussian noise as

$$Y = \mathbb{X}\theta^* + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$. By multiplying both sides by $\mathbb{X}^\top/n$, we have

$$\frac{\mathbb{X}^\top Y}{n} = \frac{\mathbb{X}^\top\mathbb{X}}{n}\theta^* + \frac{\mathbb{X}^\top\varepsilon}{n}$$

or equivalently,

$$\tilde{Y} = \theta^* + \frac{\mathbb{X}^\top\varepsilon}{n}$$

where $\tilde{Y} = \frac{\mathbb{X}^\top Y}{n} \in \mathbb{R}^d$ is the new observations vector and $\frac{\mathbb{X}^\top\varepsilon}{n} \sim \mathcal{N}(0, \frac{\mathbb{X}^\top\mathbb{X}}{n^2}\sigma^2) = \mathcal{N}(0, \frac{\sigma^2}{n}I_d)$. Therefore, under the orthogonal design assumption, we can consider the following equivalent model:

$$Y = \theta^* + \varepsilon \in \mathbb{R}^d$$

where $\varepsilon \sim \mathcal{N}(0, \frac{\sigma^2}{n}I_d)$. Note that this is estimating the mean of a Gaussian random variable and the variance of noise is the only location n or number of observations appears. As n goes to infinity, the variance of noise goes to zero and we can observe θ^* exactly. Such a model is called Gaussian Sequence Model (GSM). The literature of GSM is quite rich, e.g. see Tsybakov (09) chapter 3 or Johnstone ('20+) which is a 467-page long draft about GSMs. In addition, this approach to linear regression is also called the direct (observation) model, in contrast to inverse problem where the goal is to estimate the inverse of an operator A in the model $Y = A\theta^* + \varepsilon$.

We also consider a slight generalization of GSM, namely SubGaussian Sequence Models. We assume

$$Y = \theta^* + \varepsilon$$

where $\varepsilon^\top u \sim \mathsf{subG}(\frac{\sigma^2}{n}|u|_2^2)$ for any $u \in \mathbb{R}^d$. Under this model,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) = (\hat{\theta} - \theta^*)^\top \frac{\mathbb{X}^\top \mathbb{X}}{n}(\hat{\theta} - \theta^*) = |\hat{\theta} - \theta^*|_2^2$$

which we denote by $\mathsf{MSE}(\hat{\theta})$. Under the direct model, we have $\hat{\theta}^{\mathrm{LS}} = Y$ and for any $j \in [d]$, $\hat{\theta}^{(j)}_{\mathcal{B}_0(s)}$ is equal to $Y^{(j)}$ if $Y^{(j)}$ is among the s largest elements of Y and otherwise, $\hat{\theta}^{(j)}_{\mathcal{B}_0(s)} = 0$.

3. SPARSITY ADAPTIVE THRESHOLDING ESTIMATION

In this part of lecture, we try to solve the direct model linear regression with sparse underlying model. However, we no longer assume that the sparsity s of the solution is known. The algorithm we use here is a hard thresholding algorithm. To be more specific, if an observation is smaller than the threshold, we decide that the observation is just noise and we set it to zero. Otherwise, we decide to keep the observation as probably the additive noise in the observation is not too high. Mathematically,

$$\hat{\theta}_j^{\mathrm{HRD}} = \begin{cases} Y_j & \text{if } |Y_j| > 2\tau \\ 0 & \text{if } |Y_j| \leq 2\tau \end{cases}$$

where τ is the threshold.

Theorem: If $\tau = \sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$ and $\theta^* \in B_0(s)$, then with probability at least $1 - \delta$,

$$\mathsf{MSE}(\hat{\theta}^{\mathrm{HRD}}) = |\hat{\theta}^{\mathrm{HRD}} - \theta^*|_2^2 \lesssim \frac{\sigma^2 s}{n}\log(\frac{2d}{\delta}).$$

Proof. For the sake of simplicity, we use θ and $\hat{\theta}$ instead of θ^* and $\hat{\theta}^{\mathrm{HRD}}$, respectively. Let

$$\mathcal{A} = \{\max_{j\in[d]} |\varepsilon_j| \leq \tau\}.$$

From maximal inequalities, we have $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$. On this event, we know

1. $|Y_j| > 2\tau \Rightarrow |\theta_j| \geq |Y_j| - |\varepsilon_j| > \tau$.
2. $|Y_j| \leq 2\tau \Rightarrow |\theta_j| \leq |Y_j| + |\varepsilon_j| \leq 3\tau$.

Therefore, one can write

$$
\begin{aligned}
|\theta-\hat{\theta}|_2^2 = \sum_{j\in[d]}(\hat{\theta}_j-\theta_j)^2 &= \sum_{\substack{j\in[d]\\ |Y_j|>2\tau}}(Y_j-\theta_j)^2 + \sum_{\substack{j\in[d]\\ |Y_j|\le 2\tau}}\theta_j^2 \\
&= \sum_{\substack{j\in[d]\\ |Y_j|>2\tau}}\varepsilon_j^2 + \sum_{\substack{j\in[d]\\ |Y_j|\le 2\tau}}\theta_j^2 \\
&\le \tau^2\sum_{j\in[d]}\mathbb{1}(|Y_j|>2\tau) + \sum_{j\in[d]}\theta_j^2\mathbb{1}(|Y_j|\le 2\tau) \\
&\le \tau^2\sum_{j\in[d]}\mathbb{1}(|\theta_j|>\tau) + \sum_{j\in[d]}\theta_j^2\mathbb{1}(|\theta_j|\le 3\tau) \\
&\le \sum_{j\in[d]}\min(\tau,|\theta_j|)^2 + \sum_{j\in[d]}(3\min(\tau,|\theta_j|))^2 \\
&= 10\sum_{j\in[d]}\min(\tau^2,|\theta_j|^2) \\
&\le 10s\tau^2 = 20\frac{s\sigma^2}{n}\log(\frac{2d}{\delta})
\end{aligned}
$$

where the last inequality results from the fact that $\theta\in\mathcal{B}_0(s)$. □

Note that τ introduced above does not depend on s which is what we require. In addition, comparing this result to the known s result, we lose $1/s$ in the logarithm here which considering $s\ll d$, does not change the final result much.

Summary: We considered sparse linear regression with $\mathsf{subG}(\sigma^2)$ noise in this lecture. First, we assumed that the underlying sparsity s is known, and define s-sparsity:

$$\mathcal{B}_0(s) := \{\theta\in\mathbb{R}^d : \sum_{j=1}^d \mathbb{1}(\theta_j\neq 0)\le s\},$$

By defining the estimator

$$\hat{\theta}_{\mathcal{B}_0(s)}\in\arg\min_{\theta\in\mathcal{B}_0(s)}|Y-\mathbb{X}\theta|_2^2,$$

we showed with probability $1-\delta$,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)})\lesssim\frac{\sigma^2 s}{n}\log(\frac{d}{s}).$$

In the next part, we assumed $\frac{\mathbb{X}^\top\mathbb{X}}{n}=I_d$ and under orthogonality assumption, we showed our problem is equivalent to a direct model which can be solved by hard thresholding as

$$\hat{\theta}_j^{\text{HRD}} = \begin{cases} Y_j & \text{if } |(\mathbb{X}^\top Y/n)_j|>2\tau \\ 0 & \text{if } |(\mathbb{X}^\top Y/n)_j|\le 2\tau \end{cases}$$

with the guarantee

$$\mathsf{MSE}(\hat{\theta}^{\mathrm{HRD}}) \lesssim \frac{\sigma^2 s}{n} \log(\frac{2d}{\delta})$$

with probability $1 - \delta$ for $\tau = \sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$.

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

Lecturer: P Rigollet
Scribe: Chin-Chia Hsu abd Guang-He Lee

Lecture 8
Mar. 3, 2020

Goals: In the last lecture we study the linear regression when the parameter is sparse, considering the sparsity either known or unknown. In this lecture, a variational representation is introduced and can be generalized from HRD estimator to BIC and Lasso estimators. Later we investigate the case when the misspecified linear model in which the regression function is not in the linear form. Finally we move on to Chapter 3– matrix estimation.

1. LINEAR REGRESSION: VARIATIONAL FORM AND GENERALIZATION

One can represent the hard thresholding (HRD) estimator as a minimizer to the following variational formulation:

$$\hat{\theta}^{\mathrm{HRD}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{ |Y-\theta|_2^2 + 4\tau^2|\theta|_0 \right\} \tag{1.1}$$

or equivalently,

$$\hat{\theta}^{\mathrm{HRD}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \sum_{i=1}^{d} \left\{ (Y_i-\theta_i)^2 + 4\tau^2 \mathbb{1}(\theta_i \neq 0) \right\} \tag{1.2}$$

To verify (1.2), first given an index i we have

$$(Y_i-\hat{\theta}_i^{\mathrm{HRD}})^2 + 4\tau^2\mathbb{1}(\hat{\theta}_i^{\mathrm{HRD}}\neq 0) = \begin{cases} Y_i^2 & \text{, if } Y_i^2 \le 4\tau^2 (\hat{\theta}_i^{\mathrm{HRD}} = 0) \\ 4\tau^2 & \text{, if } Y_i^2 > 4\tau^2 \end{cases} = \min(Y_i^2, 4\tau^2)$$

Moreover, for any $\theta \in \mathbb{R}^d$,

$$(Y_i-\theta_i)^2 + 4\tau^2\mathbb{1}(\theta_i\neq 0) = \begin{cases} Y_i^2 & \text{, if } \theta_i = 0 \\ (Y_i-\theta_i)^2 + 4\tau^2 & \text{, if } \theta_i \neq 0 \end{cases} \ge \min(Y_i^2, 4\tau^2)$$

which shows that $\hat{\theta}^{\mathrm{HRD}}$ minimizes (1.2).

The formulation (1.1) can be generalized to any design matrix. Here we is the example of BIC estimator.

$$\hat{\theta}^{\mathrm{BIC}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{ |Y-\mathbb{X}\theta|_2^2 + 4\tau^2|\theta|_0 \right\} \tag{1.3}$$

The rate is the same as hard thresholding linear estimator or ℓ_0-constrained estimator. However, it is NP-hard to compute the BIC estimator in the worst case. In particular, one needs to use the brute force and search among all 2^d sparsity patterns.

By contrast, we can change the problem and replace the ℓ_0-norm by ℓ_1 norm to make it a convex optimization problem.

$$\hat{\theta}^{\mathcal{L}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{ |Y-\mathbb{X}\theta|_2^2 + 4\tau|\theta|_1 \right\} \tag{1.4}$$

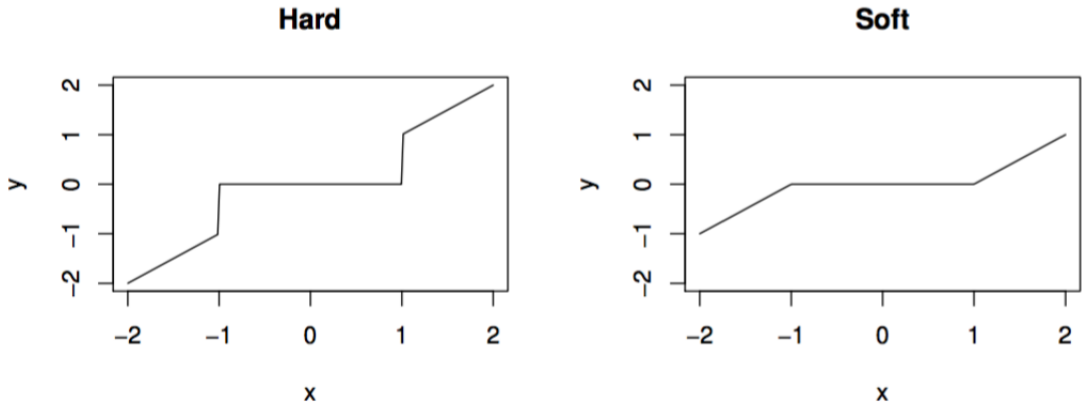


Figure 1: Transformation applied to Y_j with $2\tau = 1$ to obtain the hard (left) and soft (right) thresholding estimators [RH].

It is the Lasso estimator. There exist many efficient algorithms to solve it fast, even on large scale (coordinate decent is one of the popular ways to solve it). We derive "almost" the same rate but pay a cost: we need to assume that $\frac{\mathbb{X}^\top \mathbb{X}}{n} \approx I_d$. Recall that in Pset 1 we define "incoherence" as one measure on the closeness between two matrices. There are many ways for two matrices to be close. For more details, see notes [RH].

What if $\frac{\mathbb{X}^\top \mathbb{X}}{n} = I_d$ in (1.4), does the lead to an intuitive estimator? If $\mathbb{X} = I_d$, in this case $\hat{\theta}^{\mathcal{L}} = \hat{\theta}^{\text{SFT}}$, which is the "soft thresholding estimator," defined as

$$\hat{\theta}_j^{\text{SFT}} = (1 - \frac{2\tau}{|Y_i|})_+ Y_j, \forall j \tag{1.5}$$

Figure 1 illustrates how the soft thresholding function makes the hard thresholding function continuous at $x = |2\tau|$, softening the sharp transition. Basically, this soft thresholding function has the same property. Since we are looking for finite type of results in this course, no constants actually matter, and from this perspective $\hat{\theta}^{\mathcal{L}}$ and $\hat{\theta}^{\text{SFT}}$ are the same estimators. One can check that $\hat{\theta}^{\text{SFT}}$ is indeed the solution to (1.4) by writing the first order condition and using sub-gradient (ℓ_1-norm is not differentiable).

2. MISSPECIFIED LINEAR MODEL

We start from the regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, ...n$$

and so far we make the assumption that $f(x) = x^\top \theta^*$ for some $\theta^* \in \mathbb{R}$. What if this assumption is violated but in an approximate way $f(x) \approx x^\top \theta^*$? The technique to solve this scenario is different from what we did in linear regression since $|Y - \mathbb{X}\theta^*|_2$ is no longer equal to ε in the basic inequality $|Y - \mathbb{X}\hat{\theta}|_2 \leq |Y - \mathbb{X}\theta^*|_2$.

First we denote that $Y = \mu + \varepsilon$ and $\mu \approx \mathbb{X}\theta^*$ in the sense that

$$\frac{1}{n}|\mu - \mathbb{X}\theta^*|_2^2 \tag{2.6}$$

is small for some θ^*. How small? We will make this quantity appear in our bound: If it is small, the bound is good. Then denote $K \subset \mathbb{R}^d$ ($K = \mathbb{R}^d, K = B_1$ are two cases to which we paid lots of attention). Formulate the problem

$$\hat{\theta} \in \underset{\theta \in K}{\operatorname{argmin}} |Y - \mathbb{X}\theta|_2^2 \tag{2.7}$$

We are still searching for a solution in the column space of $\mathbb{X}$. What we can do is to compete with the best estimate. The best estimate of μ of this column space is the projection of vector μ on this span. We find the parameters over the set K

$$\theta^* \in \underset{\theta \in K}{\operatorname{argmin}} |\mu - \mathbb{X}\theta|_2^2 \tag{2.8}$$

which are sometimes called the Oracle. It is something that one cannot compute and only ORACLE can. Oracle tells us something about the truth not in a perfect manner: it knows μ but can only answer the closest object to μ in the column space. We are competing with the Oracle, that is, we want $\hat{\theta}$ to achieve as good as θ^* in terms of mean squared error.

Beginning from the basic inequality

$$|Y - \mathbb{X}\hat{\theta}|_2 \leq |Y - \mathbb{X}\theta|_2, \quad \forall \theta \in K. \tag{2.9}$$

$$\Rightarrow |\mu + \varepsilon - \mathbb{X}\hat{\theta}|_2 \leq |\mu + \varepsilon - \mathbb{X}\theta^*|_2 \tag{2.10}$$

$$\Rightarrow |\mu - \mathbb{X}\hat{\theta}|_2^2 + 2\langle \mu - \mathbb{X}\hat{\theta}, \varepsilon \rangle + |\varepsilon|_2^2 \leq |\mu - \mathbb{X}\theta^*|_2^2 + 2\langle \mu - \mathbb{X}\theta^*, \varepsilon \rangle + |\varepsilon|_2^2 \tag{2.11}$$

$$\Rightarrow |\mu - \mathbb{X}\hat{\theta}|_2^2 - |\mu - \mathbb{X}\theta^*|_2^2 \leq 2\langle \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*, \varepsilon \rangle \tag{2.12}$$

$$\Rightarrow |\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 2\langle \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*, \varepsilon \rangle \tag{2.13}$$

The last derivation comes from that if the projection of μ can be represented by some $\theta^* \in K$ and Pythagoras theorem. This is the same formula as we have seen before despite the different meanings. We can still apply our tricks in previous lectures. For example, as for least square estimator $\hat{\theta}^{\mathrm{LS}}$,

$$\frac{1}{n} |\mathbb{X}\hat{\theta}^{\mathrm{LS}} - \mathbb{X}\theta^*|_2^2 \lesssim \sigma^2 \frac{\operatorname{rank}(\mathbb{X}^\top \mathbb{X})}{n} \tag{2.14}$$

We add μ and subtract it on the left hand side of (2.14) and expand to obtain

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \mathsf{MSE}(\mathbb{X}\theta^*) + C\sigma^2 \frac{\operatorname{rank}(\mathbb{X}^\top \mathbb{X})}{n} \text{ for some constant } C \tag{2.15}$$

$$= \inf_{\theta \in \mathbb{R}^d} \mathsf{MSE}(\mathbb{X}\theta) + C\sigma^2 \frac{\operatorname{rank}(\mathbb{X}^\top \mathbb{X})}{n} \tag{2.16}$$

and we call it Oracle inequality. The term $\inf_{\theta \in \mathbb{R}^d} \mathsf{MSE}(\mathbb{X}\theta)$ is the misspecified error that will goes away if linear model is used. From a statistical perspective, Oracle inequalities are used as devices to guarantee some adaptations to such as smoothness or sparsity. Later when we touch the topics about machine learning, we will see many inequalities that look like the Oracle inequality to bound the risk in a hypothesis class.

Let's consider $K = B_1; |\mathbb{X}_j|_2 \leq \sqrt{n}$. Pythagoras theorem is not valid here. However, with the particular structure of B_1, using Hölder's inequality,

$$|\mathbb{X}\hat{\theta} - \mu|_2^2 \leq |\mathbb{X}\theta^* - \mu|_2^2 + 2\langle \mathbb{X}^\top \varepsilon, \hat{\theta} - \theta^* \rangle \tag{2.17}$$

$$\leq |\mathbb{X}\theta^* - \mu|_2^2 + 2|\mathbb{X}^\top \varepsilon|_\infty |\hat{\theta} - \theta^*|_1 \tag{2.18}$$

where $|\hat{\theta} - \theta^*| \leq 2$ and $|\mathbb{X}^\top \varepsilon|_\infty \lesssim \sigma\sqrt{n \log d}$. Therefore we obtain

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \inf_{\theta \in B_1} \mathsf{MSE}(\mathbb{X}\theta) + C\sigma\sqrt{\frac{\log d}{n}} \tag{2.19}$$

3. MATRIX ESTIMATION: BASICS

Let's first remind ourselves of the (sub)Gaussian sequence model:

$$Y = \theta^* + \varepsilon \in \mathbb{R}^d. \tag{3.20}$$

We can always reorganize these vectors to matrices. For example, we can divide the vector into 3 chunks, put the 3 chunks as 3 columns in a matrix, and then we have a matrix estimation problem.

$$Y = \theta^* + \varepsilon \in \mathbb{R}^{(d/3)\times 3}. \tag{3.21}$$

Honestly, if we don't impose any structure on the data other than being a matrix, then the estimation problem is exactly the same as the original (sub)Gaussian sequence model. For example, if θ^* is sparse, we can use $\hat{\theta}^{\mathrm{HRD}}$ as the estimator: we look at the matrix, keep the entries with high magnitudes and kill the entries with low magnitudes. Then we immediately obtain a matrix estimator, assuming that it is sparse. Essentially, even though we have a matrix here, the estimator only concerns the sparsity without considering other properties of the matrix. We can also impose some nice structures on the matrix. For example, we will talk about covariance matrix estimation. We could assume that these are covariances of things that observed in different points in time. Therefore, as things are spread in time, it is natural to assume that the covariance matrix is concentrated on the diagonal. The quintessential low dimensional structure that we can impose on matrix is governed by the rank of the matrix. We could ask, for example, what is the rate of estimation for a low-rank matrix with additive noises.

The matrix estimation problem is motivated by Netflix prize (2006-2011)[1], a 1 million grand prize for estimating a matrix for Netflix. Concretely, the researchers are given a sparse matrix, where the rows correspond to users and the columns correspond to movies. The matrix M contains sparse rating observations $M_{i,j} \in \{1, 2, 3, 4, 5\}$ for the movie j from the user i. There are at least two characteristics of such matrix. First, the observations are "noisy", as an integer value clearly does not well-calibrate the rating in the user's mind. Second, lots of the entries are missing, since each user only watch and rate a small portion of entries. The second problem is termed as deletion noise by the lecturer.

The simplest thing that we can do is to assume each rating $M_{i,j}$ can be

$$M_{i,j} = u_i v_j, \tag{3.22}$$

where $u_i, v_j \in \mathbb{R}$. This entails that the resulting estimation Θ^* is a rank 1 matrix:

$$\Theta^* = uv^\top. \tag{3.23}$$

We can generalize the rank 1 matrix to a rank r matrix as:

$$\Theta^* = \sum_{j=1}^{r} a_j u_j v_j^\top, \tag{3.24}$$

[1] https://www.netflixprize.com

where a_j is a scalar, and u_j, v_j are vectors. To ensure that everyone is on the same page, below we review some basic facts about matrices.

- Eigenvalue and eigenvectors for a square matrix A satisfy the folllowing equation:

$$Au = \lambda u, \tag{3.25}$$

  where u is an eigenvector and λ is the corresponding eigenvalue. If $A = A^\top$, then we have n real eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. In this course, we always assume that the eigenvectors have unit ℓ_2 norm $|u|_2 = 1$. Moreover the eigenvectors of A for an orthonormal basis of the linear span of the columns of A.

- Singular value decomposition (SVD): given $A \in \mathbb{R}^{m\times n}$, it can be factorized by SVD as:

$$A = UDV^\top, U \in \mathbb{R}^{m\times r}, D \in \mathbb{R}^{r\times r}, V \in \mathbb{R}^{n\times r} \tag{3.26}$$

  where r is the rank of A, $U^\top U = I_r, V^\top V = I_r$, and D is a diagonal matrix with singular values. The SVD can also be written in vector from:

$$A = \sum_{j=1}^{r} \lambda_j u_j v_j^\top, \lambda_j \in \mathbb{R}, u_j \in \mathbb{R}^m, v_j \in \mathbb{R}^n. \tag{3.27}$$

  The vector form can be extended to the largest possible rank $\min(m, n)$ by letting $\lambda_j = 0, \forall j > r$ as:

$$A = \sum_{j=1}^{\min(m,n)} \lambda_j u_j v_j^\top, \lambda_j \in \mathbb{R}, u_j \in \mathbb{R}^m, v_j \in \mathbb{R}^n. \tag{3.28}$$

  Note that we have

$$AA^\top u_j = \lambda_j^2 u_j, A^\top A v_j = \lambda_j^2 v_j. \tag{3.29}$$

  If A is positive semi-definite (PSD; $A = A^\top$ and $u^\top Au \geq 0, \forall u \in \mathbb{R}^n$), the eigenvalues are equal to the singular values.

  We use the largest singular value to define the matrix operator norm:

$$\|A\|_{\text{op}} = \lambda_{\max}(A) = \max_{x\in\mathbb{R}^n : x\neq 0} \frac{|Ax|_2}{|x|_2} = \max_{y\in B_2(\mathbb{R}^m), x\in B_2(\mathbb{R}^n)} y^\top Ax. \tag{3.30}$$

  This is called the operator norm as A is a linear operator from $(\mathbb{R}^n, |\cdot|_2)$ to $(\mathbb{R}^m, |\cdot|_2)$.

  If A is PSD,

$$\|A\|_{\text{op}} = \lambda_{\max}(A) = \max_{x\in B_2(\mathbb{R}^m)} x^\top Ax. \tag{3.31}$$

- vector norms and inner product. Let $A = a_{ij}, B = b_{ij}$.
  - $|A|_q = (\sum_{i,j} |a_{ij}|^q)^{1/q}, q > 0$.
  - $|A|_\infty = \max_{i,j} |a_{ij}|$.

- $|A|_0 = \sum_{i,j} \mathbf{1}(a_{ij} \neq 0)$.
- (Frobenius norm) $\|A\|_F = |A|_2 = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\mathrm{Tr}(A^\top A)}$.
- (inner product) $\langle A, B\rangle = \mathrm{Tr}(A^\top B) = \mathrm{Tr}(AB^T)$.

- Spectral norms. Let $\lambda = (\lambda_1, \ldots, \lambda_r)$ be singular values of A.
  - (Schatten q-norm) $\|A\|_q = |\lambda|_q$.
  - If $q = 2$, $\|A\|_2^2 = \|A\|_F^2 = \mathrm{Tr}(A^\top A) = \mathrm{Tr}(VDU^\top UDV^\top) = \mathrm{Tr}(D^2) = \sum_{j=1}^r \lambda_j^2$.
  - If $q = 1$, $\|A\|_1 = \|A\|_*$ (called nuclear norm or trace norm).
  - If $q = \infty$, $\|A\|_\infty = \lambda_{\max}(A) = \|A\|_{\mathrm{op}}$

- Useful matrix inequalities. Let $A, B \in \mathbb{R}^{n\times m}, n \leq m$. Let $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A) \geq 0$ denote the singular values of A, and $\lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_n(B) \geq 0$ the singular values of B.
  - (Weyl' 12): $\max_j |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_{\mathrm{op}}$.
  - (Hoffman-Wielandt '53): $\sum_k |\lambda_k(A) - \lambda_k(B)|^2 \leq \|A - B\|_F^2$.
  - (Hölder): $\langle A, B\rangle \leq \|A\|_p \|A\|_q$, where $p, q \geq 1$ and $1/p + 1/q = 1$. Note that we also have the vector version $\langle A, B\rangle \leq |A|_p |A|_q$.
  - Lemma (Eckart–Young). Given $A = \sum_{j=1}^r \lambda_j u_j v_j^\top, \forall k < r$, we let $A_k = \sum_{j=1}^k \lambda_j u_j v_j^\top$ be the truncated SVD. Then, $\|A - A_k\|_F^2 = \inf_{B:\mathrm{rank}(B)\leq k} \|A - B\|_F^2 = \sum_{j=k+1}^r \lambda_j^2$.

    Proof.

$$\|A - A_k\|_F^2 = \|\sum_{j=k+1}^r \lambda_j u_j v_j^\top\|_F^2 = \sum_{i,j=k+1}^r \lambda_i \lambda_j \mathrm{Tr}(u_i v_i^\top v_j u_j^\top) \tag{3.32}$$

$$= \sum_{i,j=k+1}^r \lambda_i \lambda_j \mathrm{Tr}(v_i^\top v_j u_j^\top u_i) = \sum_{i=1}^r \lambda_i^2, \tag{3.33}$$

    where the last equality is due to $u_j^\top u_i = v_i^\top v_j = \delta_{ij}$ Now we take B with singular values $\sigma_1 \geq \cdots \geq \sigma_k \geq 0 \geq \cdots \geq 0$.

$$\|A - B\|_F^2 \geq \sum_{i=1}^r (\lambda_j - \sigma_j)^2 = \sum_{i=1}^k (\lambda_j - \sigma_j)^2 + \sum_{i=1}^k \lambda_j^2, \tag{3.34}$$

    where the inequality is due to Hoffman-Wielandt. □

Summary:

- Linear regression
  1. Hard thresholding and variational form

$$\hat{\theta}^{\mathrm{HRD}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{|Y - \theta|_2^2 + 4\tau^2 |\theta|_0\right\}$$

2. BIC estimator

$$\hat{\theta}^{\mathrm{BIC}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ |Y - \mathbb{X}\theta|_2^2 + 4\tau^2 |\theta|_0 \right\}$$

3. Lasso and soft thresholding estimator

$$\hat{\theta}^{\mathcal{L}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ |Y - \mathbb{X}\theta|_2^2 + 4\tau |\theta|_1 \right\}$$

If $\mathbb{X} = I_d$, $\hat{\theta}^{\mathcal{L}} = \hat{\theta}^{\mathrm{SFT}}$ where

$$\hat{\theta}_j^{\mathrm{SFT}} = (1 - \frac{2\tau}{|Y_i|})_+ Y_j, \forall j$$

- Misspecified linear model and oracle inequality: When $\mu \approx \mathbb{X}\theta^*$ for some $\theta* \in K$, we derived the oracle inequality for expected MSE of two estimators

1. $K = \mathbb{R}^d$:

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \inf_{\theta \in \mathbb{R}^d} \mathsf{MSE}(\mathbb{X}\theta) + C\sigma^2 \frac{\mathrm{rank}(\mathbb{X}^\top \mathbb{X})}{n}$$

2. $K = B_1; |\mathbb{X}_j|_2 \leq \sqrt{n}$:

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \inf_{\theta \in B_1} \mathsf{MSE}(\mathbb{X}\theta) + C\sigma \sqrt{\frac{\log d}{n}}$$

- Matrix basics:

1. Eigenvalues and eigenvectors: $Au = \lambda u$.
2. SVD: $A = UDV^\top$, where $D = \mathrm{diag}(\lambda)$.
3. Operator norm: $\|A\|_{\mathrm{op}} = \lambda_{\max(A)}$.
4. Vector norms $|A|_q = (\sum_{i,j} |a_{ij}^q|)^{1/q}, q > 0$.
5. Spectral norms: $\|A\|_q = |\lambda|_q$.
6. (Weyl' 12): $\max_j |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_{\mathrm{op}}$.
7. (Hoffman-Wielandt '53): $\sum_k |\lambda_k(A) - \lambda_k(B)|^2 \leq \|A - B\|_F^2$.
8. (Hölder): $\langle A, B \rangle \leq \|A\|_p \|A\|_q$, where $p, q \geq 1$ and $1/p + 1/q = 1$.
9. Truncated SVD is the best rank k approximation in Frobenius norm.

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

Lecturer: P Rigollet
Scribe: Abraham Shalom, Ziad Mansour

Lecture 9
Mar. 5, 2020

Goals: In the last lecture, we introduced the fundamental matrix estimation model and some basic matrix properties and inequalities. In this lecture, we introduce the singular value thresholding estimator for matrix denoising. We then show that, with a good threshold choice, this estimator's MSE will decay as the matrix dimensions grow. Finally, we introduce the Davis-Kahan-Sin(Theta) Theorem for use in matrix perturbation analysis.

1. MATRIX DENOISING

We consider the subGaussian matrix model:

$$Y = \Theta^* + E$$

where $Y \in \mathbb{R}^{m \times n}$ is the matrix of observed responses, and E is an $m \times n$ noise matrix such that

$$u^\top E v \sim \mathsf{subG}(\sigma^2)\,, \quad \forall u \in \mathbb{R}^m, v \in \mathbb{R}^n\,.$$

This is the case for example if $\forall\, , i, j \in [m] \times [n]$, each of entry of E denoted by $E_{i,j}$ is an independent $\mathsf{subG}(\sigma^2)$. Indeed for any $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$, it holds

$$\mathbb{E}[e^{su^\top Ev}] = \mathbb{E}[e^{\sum_{ij} u_i E_{ij} v_j}] = \prod_{i,j} \mathbb{E}[e^{su_i E_{ij} v_j}] \le \prod_{i,j} e^{\frac{s^2\sigma^2 u_i^2 v_j^2}{2}} = e^{\sum_{i,j} \frac{s^2\sigma^2 u_i^2 v_j^2}{2}} = e^{\frac{s^2\sigma^2 |u|_2^2 |v|_2^2}{2}}$$

Similar to the sub-Gaussian sequence model, we have a direct observation model where we observe the parameter of interest with additive noise. This enables us to use thresholding methods for estimating Θ^*. The analysis becomes interesting when we assume that Θ^* is low rank which is equivalent to sparsity in its unknown eigenbasis. Hence, we can consider the SVD of Θ^* and of Y:

$$\Theta^* = \sum_{j=1}^{m \wedge n} \lambda_j u_j v_j^\top$$

$$Y = \sum_{j=1}^{m \wedge n} \hat{\lambda}_j \hat{u}_j \hat{v}_j^\top$$

Hence, if we knew the u_j and v_j, we can estimate the λ_j's by hard thresholding. We claim that it is sufficient to estimate the eigenvectors of Θ^* by the eigenvectors of Y.

We define the Singular Value Thresholding Estimator, SVT, with threshold $2\tau \ge 0$ as:

$$\hat{\Theta}^{\text{SVT}} = \sum_j \hat{\lambda}_j \mathbb{I}(\hat{\lambda}_j > 2\tau) \hat{u}_j \hat{v}_j^\top$$

How to select τ? Recall in the Gaussian sequence model, we select τ so that it is larger than the maximum magnitude of the noise with probability $1-\delta$. We take a similar approach here except that the norm in which the magnitude of the noise is measured is adapted to the matrix case. Basically, we are trying to control the operator norm of the matrix E.

$$\text{Our Candidate}: \quad \tau \geq \max_j \lambda_j(E) = \|E\|_{\text{op}}$$

Recall from Lecture 5, the result on operator norms of $m \times n$ matrices: If $E_{i,j} \sim \mathsf{subG}(\sigma^2)$ are independent, then $\|E\|_{\text{op}} \leq \sigma(\sqrt{m}+\sqrt{n})$ with high probability. We generalize this result to the following lemma which will allow us to control the operator norm of the matrix E.

Lemma: Let E be an $m \times n$ random matrix defined as above. Then

$$\|E\|_{\text{op}} \leq 2\sigma\sqrt{5(m+n)} + 2\sigma\sqrt{2\log(1/\delta)}$$

with probability $1-\delta$.

Proof. Let $\mathcal{N}_m$ be a 1/4-net for the euclidean ball $\mathcal{B}_2(R^m)$ and $\mathcal{N}_n$ be a 1/4-net for the euclidean ball $\mathcal{B}_2(R^n)$. Then it follows from the results of lecture 5 that for $\varepsilon = 1/4$:

$$|\mathcal{N}_m| \leq (1+\frac{2}{\varepsilon})^m = 9^m$$

$$|\mathcal{N}_n| \leq (1+\frac{2}{\varepsilon})^n = 9^n$$

Now we can write the operator norm on E as:

$$\|E\|_{\text{op}} = \max_{x \in \mathcal{S}^{n-1}} |Ex|_2 = \max_{\substack{x \in \mathcal{S}^{n-1} \\ y \in \mathcal{S}^{m-1}}} y^\top E x$$

On the operator norm, we use the triangular inequality and decompose each point x on the sphere $\mathcal{S}^{n-1}$ into a point z on the epsilon net $\mathcal{N}_n$ and a remainder term that can be upper bounded by 1/4. By applying the described decomposition on the maximum $x \in \mathcal{S}^{n-1}$ we can upper bound the operator norm of E by :

$$\|E\|_{\text{op}} \leq \max_{z \in \mathcal{N}_n} |Ez|_2 + \frac{1}{4}\|E\|_{\text{op}}$$

Now we examine the $|Ez|_2$ term:

$$\begin{aligned} |Ez|_2 &= \max_{y \in \mathcal{S}^{m-1}} y^\top E z \\ &\leq \max_{w \in \mathcal{N}_m} w^\top E z + \frac{1}{4}\|E\|_{\text{op}} \end{aligned}$$

Hence applying this bound to our $\|E\|_{\text{op}}$

$$\|E\|_{\text{op}} \leq \max_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} w^\top E z + \frac{1}{4}\|E\|_{\text{op}} + \frac{1}{4}\|E\|_{\text{op}}$$

Rearranging the above inequality

$$\|E\|_{\text{op}} \leq 2 \max_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} w^\top E z$$

where $\forall w, z;\ w^\top E z \sim \mathsf{subG}(\sigma^2 |w|_2^2 |z|_2^2)$. Using the fact that $|w|_2^2 \leq 1$ and $|z|_2^2 \leq 1$, then $w^\top E z \sim \mathsf{subG}(\sigma^2)$. Using union bounds:

$$\mathbb{P}(\max_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} w^\top E z > \frac{t}{2}) \leq \sum_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} \mathbb{P}(w^\top E z > \frac{t}{2}) \leq \sum_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} e^{\frac{-t^2}{8\sigma^2}} \leq 9^{n+m} e^{\frac{-t^2}{8\sigma^2}} =: \delta$$

Solving for t we get

$$t \leq 2\sqrt{2}\sigma\sqrt{\log(9)(m+n)} + 2\sqrt{2}\sigma\sqrt{-\log(\delta)} \leq 2\sigma\sqrt{5(n+m)} + 2\sigma\sqrt{-2\log(\delta)}$$

Thus, we choose $\tau = 2\sigma\sqrt{5(n+m)} + 2\sigma\sqrt{-2\log(\delta)}$. □

Using the above lemma, we propose the following theorem:

Theorem: The SVT estimator $\hat{\Theta}^{\text{SVT}}$ with τ as above, satisfies (with probability $1-\delta$):

$$\frac{1}{m \cdot n} \left\| \hat{\Theta}^{\text{SVT}} - \Theta^* \right\|_F^2 \lesssim \frac{\sigma^2 \cdot \text{rank}(\Theta^*)}{m \cdot n}(m + n + \log\frac{1}{\delta})$$

Proof. To prove this, we first define a random set of indices S:

$$S = \{j \ :\ |\hat{\lambda}_j| > 2\tau\}$$

We then define the event $\mathcal{A} = \{\|E\|_{\text{op}} \leq \tau\}$ and recall that $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$. Now, we will make purely deterministic statements assuming that $\mathcal{A}$ holds.

By Weyl's inequality,

$$|\hat{\lambda}_j - \lambda_j| \leq \|Y - \Theta^*\|_{\text{op}} = \|E\|_{\text{op}} \leq \tau$$

Based off of the above in conjunction with triangle inequality, we can make statements about $j \in S$ and $j \in S^c$ as follows:

$$\begin{aligned} j \in S: &\quad |\hat{\lambda}_j| > 2\tau \implies |\hat{\lambda}_j| \geq |\hat{\lambda}_j| - |\hat{\lambda}_j - \lambda_j| > \tau \\ j \in S^c: &\quad |\hat{\lambda}_j| \leq 2\tau \implies |\lambda_j| \leq |\hat{\lambda}_j| + |\lambda_j - \hat{\lambda}_j| \leq 3\tau \end{aligned}$$

We introduce an oracle $\bar{\Theta}$ that knows the singular values and vectors, but not the support.

$$\bar{\Theta} = \sum_{j \in S} \lambda_j u_j v_j^\top$$

Next,

$$\left\| \hat{\Theta}^{\text{SVT}} - \bar{\Theta} \right\|_F^2 = \sum_j \lambda_j^2 (\hat{\Theta}^{\text{SVT}} - \bar{\Theta}) \leq \left\| \hat{\Theta}^{\text{SVT}} - \bar{\Theta} \right\|_{\text{op}}^2 \text{rank}(\hat{\Theta}^{\text{SVT}} - \bar{\Theta}),$$

where we used the fact that $\lambda_j^2(\hat{\Theta}^{\text{SVT}} - \bar{\Theta}) \leq \left\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\right\|_{\text{op}}^2$ for all j.

Note that the rank of the sum of two matrices is at most the sum of the two ranks. Since $\text{rank}(\hat{\Theta}^{\text{SVT}}) \vee \text{rank}(\bar{\Theta}) \leq |S|$, then $\text{rank}(\hat{\Theta}^{\text{SVT}} - \bar{\Theta}) \leq 2|S|$.

Now, we need to bound the operator norm in the above inequality. We first apply the triangle inequality, bringing in Θ^* and Y:

$$\left\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\right\|_{\text{op}} \leq \left\|\hat{\Theta}^{\text{SVT}} - Y\right\|_{\text{op}} + \|Y - \Theta^*\|_{\text{op}} + \left\|\Theta^* - \bar{\Theta}\right\|_{\text{op}}$$

Now, we bound each of the individual terms that appear above. First, we have that, by construction,

$$\|Y - \Theta^*\|_{\text{op}} = \|E\|_{\text{op}} \leq \tau$$

We now work to bound $\left\|\hat{\Theta}^{\text{SVT}} - Y\right\|_{\text{op}}$, using the fact that $\forall j \in S^c : \ \hat{\lambda}_j \leq 2\tau$.

$$Y - \hat{\Theta}^{\text{SVT}} = \sum_{j \in S^c} \hat{\lambda}_j \hat{u}_j \hat{v}_j^\top$$
$$\left\|\hat{\Theta}^{\text{SVT}} - Y\right\|_{\text{op}} = \max_{j \in S^c} \hat{\lambda}_j \leq 2\tau$$

We use similar process to bound $\left\|\Theta^* - \bar{\Theta}\right\|_{\text{op}}$, using $\forall j \in S^c : \ \lambda_j \leq 3\tau$.

$$\Theta^* - \bar{\Theta} = \sum_{j \in S^c} \lambda_j u_j v_j^\top$$
$$\left\|\Theta^* - \bar{\Theta}\right\|_{\text{op}} = \max_{j \in S^c} \lambda_j \leq 3\tau$$

Putting the above together, we get

$$\left\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\right\|_{\text{op}} \leq 2\tau + \tau + 3\tau = 6\tau$$
$$\left\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\right\|_F^2 \leq 72\tau^2|S| = 72\sum_{j \in S} \tau^2$$

Now, we note that we are not interested in $\bar{\Theta}$ (from the oracle); instead, we want to compare our estimate to Θ^*. We therefore use the triangle inequality (again) to squeeze in the $\bar{\Theta}$.

$$\begin{aligned}\left\|\hat{\Theta}^{\text{SVT}} - \Theta^*\right\|_F^2 &\leq 2\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\|_F^2 + 2\left\|\bar{\Theta} - \Theta^*\right\|_F^2 \\ &\leq 144\sum_{j \in S} \tau^2 + 2\sum_{j \in S^c} \lambda_j^2\end{aligned}$$

Notice that the term in the left-hand sum, τ^2, is in fact equal to $\min(\tau^2, \lambda_j^2)$. In the

right-hand sum, the term $\lambda_j^2 \leq \min(9\lambda_j^2, 9\tau^2)$. Thus, we add this to the above inequality:

$$\begin{aligned}
\left\|\hat{\Theta}^{\text{SVT}} - \Theta^*\right\|_F^2 &\leq 144 \sum_{j \in S} (\tau^2 \wedge \lambda_j^2) + 18 \sum_{j \in S^c} (\tau^2 \wedge \lambda_j^2) \\
&\leq 144 \sum_{j \in S} (\tau^2 \wedge \lambda_j^2) + 144 \sum_{j \in S^c} (\tau^2 \wedge \lambda_j^2) \\
&= 144 \sum_{j} (\tau^2 \wedge \lambda_j^2) \\
&\lesssim \tau^2 \text{rank}(\Theta^*) \\
&\lesssim (\sigma^2(m+n) + \sigma^2 \log \frac{1}{\delta}) \text{rank}(\Theta^*)
\end{aligned}$$

To finish the proof, we can simply multiply both sides of our inequality by $\frac{1}{m \cdot n}$ in order to get mean-squared error. This gives us a final bound of:

$$\frac{\left\|\hat{\Theta}^{\text{SVT}} - \Theta^*\right\|_F^2}{m \cdot n} \lesssim \frac{(\sigma^2(m+n) + \sigma^2 \log \frac{1}{\delta})}{m \cdot n} \text{rank}(\Theta^*)$$

□

Using this, we see that without knowing anything about $\hat{u}$, $\hat{v}$ and their closeness to u, v respectively we can get an MSE that is vanishing as $m, n \to \infty$. The question now becomes: can we get greedy and find any results indicating closeness of our estimates to the true singular vectors?

2. PERTURBATION THEORY

We give a brief introduction to Perturbation Theory, with the subject to be continued in the next lecture. The focus of perturbation theory is to understand how a matrix's spectrum changes if its entries are perturbed. We introduce an important result in perturbation theory.

Theorem (Davis-Kahan $\sin(\theta)$ Theorem): Let A, $\hat{A}$ by symmetric $n \times n$ matrices such that:

$$\begin{aligned}
A &= \sum_{j=1}^{n} \lambda_j u_j u_j^\top, \quad \lambda_1 \geq \lambda_2 \geq \dots \\
\hat{A} &= \sum_{j=1}^{n} \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top, \quad \hat{\lambda}_1 \geq \hat{\lambda_2} \geq \dots \\
|u_j|_2 &= |\hat{u}_j|_2 = 1
\end{aligned}$$

Then:

$$|\sin(\measuredangle(\hat{u}_1, u_1))| \leq \frac{2}{\max(\lambda_1 - \lambda_2, \hat{\lambda}_1 - \hat{\lambda}_2)} \left\|A - \hat{A}\right\|_{\text{op}}$$

Moreover:

$$\min_{\varepsilon\in\{\pm 1\}} |u_1 - \varepsilon\hat{u}_1|_2 \leq \sqrt{2}|\sin(\measuredangle(u_1, \hat{u}_1))|$$

The proof to this Theorem will be given in the following lecture.

Summary: In this lecture, we develop tools for matrix denoising. We defined the Singular Value Thresholding Estimator (SVT) with threshold 2τ as

$$\hat{\Theta}^{\text{SVT}} = \sum_j \hat{\lambda}_j \mathbb{1}(\hat{\lambda}_j > 2\tau)\hat{u}_j\hat{v}_j^\top$$

We then established the following lemma that helps us control the operator norm of the noise matrix:
Let E be an $m \times n$ random matrix such that $u^\top Ev \sim \mathsf{subG}(\sigma^2)$ for all $u \in \mathbb{R}^m, v \in \mathbb{R}^n$, then

$$\|E\|_{\text{op}} \lesssim \sigma\sqrt{m+n} + \sigma\sqrt{\log(1/\delta)}$$

with probability $1-\delta$.

This, in turn, allows us to bound the MSE of the SVT estimator using the following theorem, and show that the average error goes to 0 as $m, n \to \infty$ (for $\text{rank}(\Theta^*) \ll m, n$).

The SVT estimator $\hat{\Theta}^{\text{SVT}}$ with τ as above, satisfies (with probability $1-\delta$):

$$\frac{1}{m\cdot n}\left\|\hat{\Theta}^{\text{SVT}} - \Theta^*\right\|_F^2 \lesssim \frac{\sigma^2\cdot\text{rank}(\Theta^*)}{m\cdot n}(m+n+\log\frac{1}{\delta})$$

Finally, we introduce the Davis-Kahan $\sin(\theta)$ theorem, which is an important result in determining the closeness of eigenspaces. This serves as the beginning of our studies into perturbation theory, which we will continue in the next lecture.

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

Lecturer: P Rigollet
Scribe: Abhimanyu Dubey, Zachary Markovich

Lecture 10
Mar. 10, 2020

Goals: In the previous lecture, we introduced the Davis-Kahan $\sin(\theta)$ theorem, a result in perturbation theory that allows us to provide a bound on the 2-norm of the difference between the top eigenvector of a matrix and its perturbed version. In this lecture we will first provide a proof for the Davis-Kahan $\sin(\theta)$ theorem and subsequently we will provide two of its applications in matrix denoising and community detection.

1. DAVIS-KAHAN $\sin(\theta)$ THEOREM

The Davis-Kahan $\sin(\theta)$ theorem provides an important result in perturbation theory of matrices. We begin by restating the theorem.

Theorem: Let $A, \hat{A}$ be two symmetric $n \times n$ matrices, with eigen-decompositions given by:

$$A = \sum_{j=1}^{n} \lambda_j u_j u_j^\top \text{ and } \hat{A} = \sum_{j=1}^{n} \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top,$$

where, $\lambda_1 \geq \lambda_2 ... \geq \lambda_n$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 ... \geq \hat{\lambda}_n$ without loss of generality. Then,

$$|\sin(\measuredangle(u_1, \hat{u}_1))| \leq 2 \frac{\|A - \hat{A}\|_{\text{op}}}{\max\left(\lambda_1 - \lambda_2, \hat{\lambda}_1 - \hat{\lambda}_2\right)}.$$

Moreover,

$$\min_{\varepsilon \in \{\pm 1\}} |u_1 - \varepsilon \hat{u}_1|_2 \leq \sqrt{2} \, |\sin(\measuredangle(u_1, \hat{u}_1))| \,.$$

Proof. The proof proceeds by considering the eigendecomposition of A. For any $x \in \mathbb{R}^n$, such that $|x|_2 = 1, \sum_{j=1}^{n} \left(x^\top u_j\right)^2 = 1$. Therefore, since $\lambda_1 \geq \lambda_2 \geq ... \lambda_n$,

$$\begin{aligned}
x^\top A x &= \sum_{j=1}^{n} \lambda_j (x^\top u_j)^2 = \lambda_1 (x^\top u_1)^2 + \sum_{j \geq 1}^{n} \lambda_j (x^\top u_j)^2 \\
&\leq \lambda_1 (x^\top u_1)^2 + \lambda_2 \sum_{j \geq 2} (x^\top u_j)^2 && (\lambda_j \leq \lambda_2 \; \forall \, j \geq 2) \\
&= (\lambda_1 - \lambda_2)(x^\top u_1)^2 + \lambda_2 && (\textstyle\sum_{j=1}^{n} \left(x^\top u_j\right)^2 = 1) \\
&= \lambda_1 - (\lambda_1 - \lambda_2) \sin^2(\measuredangle(x, u_1)) \,. && (\cos(\measuredangle(x, u_1)) = x^\top u_1) \\
&= u_1^\top A u_1 - (\lambda_1 - \lambda_2) \sin^2(\measuredangle(x, u_1)) \,. && (\lambda_1 = u_1^\top A u_1)
\end{aligned}$$

Now, by taking $x = \hat{u}_1$,

$$\begin{aligned}
(\lambda_1 - \lambda_2)\sin^2\left(\measuredangle(\hat{u}_1, u_1)\right) &\leq u_1^\top A u_1 - \hat{u}_1^\top A \hat{u}_1 \\
&= u_1^\top \hat{A} u_1 - \hat{u}_1^\top A \hat{u}_1 + u_1^\top (A - \hat{A}) u_1. \\
&\leq \hat{u}_1^\top \hat{A} \hat{u}_1 - \hat{u}_1^\top A \hat{u}_1 + u_1^\top (A - \hat{A}) u_1. \quad (x^\top A x \leq \hat{u}_1^\top \hat{A} \hat{u}_1 \ \forall |x|_2 = 1) \\
&= \langle \hat{A} - A, \hat{u}_1 \hat{u}_1^\top - u_1 u_t^\top \rangle. \\
&\leq \|\hat{A} - A\|_{\text{op}} \|\hat{u}_1 \hat{u}_1^\top - u_1 u_t^\top\|_1 \qquad \text{(Hölder)} \\
&\leq \sqrt{2} \|\hat{A} - A\|_{\text{op}} \|\hat{u}_1 \hat{u}_1^\top - u_1 u_t^\top\|_F \qquad \text{(Cauchy-Schwarz)}
\end{aligned}$$

We can simplify the second term as follows.

$$\begin{aligned}
\|\hat{u}_1 \hat{u}_1^\top - u_1 u_t^\top\|_F^2 &= \underbrace{\text{Tr}\left(u_1 u_1^\top u_1 u_1^\top\right)}_{=1} + \underbrace{\text{Tr}\left(\hat{u}_1 \hat{u}_1^\top \hat{u}_1 \hat{u}_1^\top\right)}_{=1} - 2\,\text{Tr}\left(\hat{u}_1 \hat{u}_1^\top u_1 u_1^\top\right) \\
&= 2 - 2\left(\hat{u}_1^\top u_1\right)^2 \\
&= 2\sin^2(\measuredangle(\hat{u}_1, u_1)). \qquad (\cos(\measuredangle(\hat{u}_1, u_1)) = \hat{u}_1^\top u_1)
\end{aligned}$$

Replacing this in the original statement gives us:

$$(\lambda_1 - \lambda_2)\sin^2\left(\measuredangle(\hat{u}_1, u_1)\right) \leq 2\|A - \hat{A}\|_{\text{op}} \left|\sin(\measuredangle(\hat{u}_1, u_1))\right|.$$

An identical analysis can be done by considering the eigendecomposition of $\hat{A}$ (and replacing $x = u_1$ subsequently), which will provide:

$$(\hat{\lambda}_1 - \hat{\lambda}_2)\sin^2\left(\measuredangle(\hat{u}_1, u_1)\right) \leq 2\|A - \hat{A}\|_{\text{op}} \left|\sin(\measuredangle(\hat{u}_1, u_1))\right|.$$

Combining the two results,

$$\left|\sin\left(\measuredangle\left(u_1, \hat{u}_1\right)\right)\right| \leq 2\frac{\|A - \hat{A}\|_{\text{op}}}{\max\left(\lambda_1 - \lambda_2, \hat{\lambda}_1 - \hat{\lambda}_2\right)}.$$

Typically, we would not like to use the eigenvalues of the perturbed matrix $\hat{A}$, since it is a random quantity in our applications and non-trivial to control. However, domain-specific assumptions can be made about the spectrum of A, leading to useful results. Now, for the second part of the proof.

$$\min_{\varepsilon \in \{\pm 1\}} |u_1 - \varepsilon \hat{u}_1|_2^2 = 2 - 2|u_1^\top \hat{u}_1| \overset{(a)}{\leq} 2 - 2(u_1^\top \hat{u}_1)^2 \overset{(b)}{\leq} 2\sin^2(\measuredangle(u_1, \hat{u}_1)).$$

Here, (a) follows from the fact that $x \leq x^2 \ \forall x \leq 1$, and (b) follows from the relationship of the cosine and dot product. □

2. APPLICATIONS

2.1 Matrix Denoising

We will consider a sub-Gaussian matrix denoising problem similar to the previous lecture.

$$Y = \Theta^* + E$$

We will assume that all of $Y, \Theta^* \in \mathbb{R}^{n\times n}$ and $E = (e_{ij})_{i\geq j}^n \sim \mathsf{subG}(\sigma^2)$ are symmetric for simplicity. We can represent the matrices by their SVD:

$$\Theta^* = \sum \lambda_j u_j u_j^\top \quad \text{and} \quad Y = \sum \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top.$$

From the Davis-Kahan $\sin(\theta)$ theorem, we have:

$$\min_{\varepsilon\in\{\pm 1\}} |\hat{u}_1 - \varepsilon u_1|_2^2 \leq 2\frac{\|E\|_{\text{op}}}{\lambda_1 - \lambda_2}$$

We have construction, $u^\top E v \sim \mathsf{subG}(\sigma^2|u|_2^2|v|_2^2)$. We can therefore conclude that there exists a constant such that with probability at least 0.99,

$$\min_{\varepsilon\in\{\pm 1\}} |\hat{u}_1 - \varepsilon u_1|_2^2 \leq 2\frac{\|E\|_{\text{op}}}{\lambda_1 - \lambda_2} \lesssim \frac{\sigma\sqrt{n}}{\lambda_1 - \lambda_2}.$$

We see that the crucial quantity controlling the quality of approximation is the signal to noise ratio (SNR) given by $\frac{\lambda_1 - \lambda_2}{\sigma}$.

2.2 Community Detection

Community detection is an important problem in social network analysis. We consider the stochastic block model in this example. The social network is determined by an undirected graph represented by an adjacency matrix $\tilde{A}$, where the n nodes represent people and an edge (i, j) denotes a friendship between persons i and j.

We assume persons belong to one of two groups. If two persons i, j belong to the same group then there is an edge (i, j) in $\tilde{A}$ with probability p. Alternatively, if they belong to different groups, there is an edge (i, j) with probability q. Therefore, $\tilde{A}_{ij} \sim \mathsf{Ber}(p)$ if i and j belong to the same community, and $\tilde{A}_{ij} \sim \mathsf{Ber}(q)$ otherwise. Self-edges, i.e., A_{ii}, can be modeled based on the application. In this setting, we will assume they are random variables, i.e., $A_{ii} \sim \mathsf{Ber}(p)$.

We assume that p and q are known *a priori*, however, relaxations can be made in case they are not. Each edge is assumed independent from the others (i.e., no network effects). The goal of the problem is to recover the community structure, given a realization $\tilde{A}$ of the network. From our formulation, we have:

$$\mathbb{E}[\tilde{A}_{ij}] = \begin{cases} p \text{ if } i \text{ and } j \text{ belong to the same community,} \\ q \text{ otherwise.} \end{cases} \tag{2.1}$$

To represent the matrices, assume that the first group is the first $n/2$ nodes. Then, we can write $\mathbb{E}[\tilde{A}]$ as:

$$\mathbb{E}[\tilde{A}] = \left[\begin{array}{ccc|ccc} p & \dots & p & q & \dots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \dots & p & q & \dots & q \\ \hline q & \dots & q & p & \dots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q & \dots & q & p & \dots & p \end{array}\right].$$

Let $E = \tilde{A} - \mathbb{E}[\tilde{A}]$. Then,

$$\tilde{A} = \mathbb{E}[\tilde{A}] + E.$$

We see that $\mathbb{E}[E_{ij}] = 0$ and $|E_{ij}| \leq 1 \ \forall i, j$. From Hoeffding's lemma, for vectors u and v,

$$u^\top E v \sim \mathsf{subG}(\|u\|_2^2 \cdot \|v\|_2^2).$$

For the community detection problem, we will analyse the centered matrix A:

$$A = \tilde{A} - \left(\frac{p-q}{2}\right) \mathbb{1}_n \mathbb{1}_n^\top.$$

Similar to $\mathbb{E}[\tilde{A}]$, we see that $\mathbb{E}[A]$ can be written as:

$$\mathbb{E}[A] = \left[\begin{array}{ccc|ccc} \frac{p-q}{2} & \dots & \frac{p-q}{2} & \frac{q-p}{2} & \dots & \frac{q-p}{2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{p-q}{2} & \dots & \frac{p-q}{2} & \frac{q-p}{2} & \dots & \frac{q-p}{2} \\ \hline \frac{q-p}{2} & \dots & \frac{q-p}{2} & \frac{p-q}{2} & \dots & \frac{p-q}{2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{q-p}{2} & \dots & \frac{q-p}{2} & \frac{p-q}{2} & \dots & \frac{p-q}{2} \end{array}\right] = \frac{p-q}{2} \left[\begin{array}{ccc|ccc} 1 & \dots & 1 & -1 & \dots & -1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \dots & 1 & -1 & \dots & -1 \\ \hline -1 & \dots & -1 & 1 & \dots & 1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -1 & \dots & -1 & 1 & \dots & 1 \end{array}\right].$$

We see that $\mathbb{E}[A]$ is of rank 1 and it can be written as:

$$\mathbb{E}[A] = \frac{n(p-q)}{2} \left(\frac{u_1}{\sqrt{n}}\right) \left(\frac{u_1}{\sqrt{n}}\right)^\top, \text{ where } u_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}.$$

To relax the ordered-ness assumption of A, we can consider that we are provided with an alternate matrix $\Pi A \Pi^\top$, where Π is an $n \times n$ permutation of A. In that case, we can represent the resulting permutation of $\mathbb{E}[A]$ as follows:

$$\Pi \mathbb{E}[A] \Pi^\top = \frac{n(p-q)}{2} \left(\frac{\Pi u_1}{\sqrt{n}}\right) \left(\frac{\Pi u_1}{\sqrt{n}}\right)^\top.$$

To estimate u from the (centered) observed matrix, we consider the eigendecomposition of A, i.e.,

$$A = \tilde{A} - \left(\frac{p-q}{2}\right) \mathbf{1}_n \mathbf{1}_n^\top = \sum_{j=1}^{n} \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top.$$

From the Davis-Kahan $\sin(\theta)$ theorem, we can bound the difference in the first eigenvectors, in a fashion similar to the last example:

$$\min_{\varepsilon \in \{\pm 1\}} \left\| \frac{\hat{u}_1}{\sqrt{n}} - \frac{\varepsilon u_1}{\sqrt{n}} \right\|_2^2 \overset{(a)}{\leq} 2 \frac{\|A - \mathbb{E}[A]\|_{\text{op}}}{\lambda_1 - 0} \overset{(b)}{=} 4 \frac{\|E\|_{\text{op}}}{n(p-q)} \overset{(c)}{\lesssim} \frac{1}{(p-q)\sqrt{n}}.$$

Here, (a) follows from Davis-Kahan, (b) follows from the fact that $A - \mathbb{E}[A] = \tilde{A} - \mathbb{E}[\tilde{A}] = E$ and that $\lambda_1 = \frac{n(p-q)}{2}$ and (c) holds with probability at least 0.99, obtained from the bound on the operator norm for sub-Gaussian random matrices. Therefore, we can say that if $(p-q) \gg \frac{1}{\sqrt{n}}$,

$$\min_{\varepsilon \in \{\pm 1\}} \left| \frac{\hat{u}_1}{\sqrt{n}} - \frac{\varepsilon u_1}{\sqrt{n}} \right|_2^2 \ll 1.$$

A popular measure of performance is the *classification error*, i.e., the average number of times our prediction disagrees with the true assignment. This can be given by:

$$L(u_1, \hat{u}_1) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1} \left\{ \text{sign}(\hat{u}_1^{(j)}) \neq \text{sign}(u_1^{(j)}) \right\}.$$

$L(u_1, \hat{u}_1)$ is non-convex. However, we can bound it by a convex relaxation that we can provide guarantees for, as follows.

$$\begin{aligned}
L(u_1, \hat{u}_1) &= \frac{1}{n} \sum_{j=1}^{n} \mathbb{1} \left\{ \text{sign}(\hat{u}_1^{(j)}) \neq \text{sign}(u_1^{(j)}) \right\} \\
&= \frac{1}{n} \sum_{j=1}^{n} \mathbb{1} \left\{ \hat{u}_1^{(j)} \cdot u_1^{(j)} < 0 \right\} \\
&\leq \frac{1}{n} \sum_{j=1}^{n} \left(\hat{u}_1^{(j)} \cdot u_1^{(j)} - 1 \right)^2 \\
&= \frac{1}{n} \sum_{j=1}^{n} \left(\hat{u}_1^{(j)} - u_1^{(j)} \right)^2 \\
&= \left\| \frac{\hat{u}_1}{\sqrt{n}} - \frac{u_1}{\sqrt{n}} \right\|_2^2.
\end{aligned}$$

From the previous result, we can claim that $L(\hat{u}_1, u_1) \lesssim \frac{1}{(p-q)\sqrt{n}}$. Therefore, we can additionally conclude that as $\sqrt{n}(p-q) \to \infty$, the probability of error, $L(u_1, \hat{u}_1) \to 0$.

Additional Remarks. Stronger statements can be made regarding the recovery problem than the ones described in this lecture. [ABH16] prove lower bounds for *exact* recovery. Specifically, they state that, for constants α and β such that

$$p = \frac{\alpha \log(n)}{n} \text{ and } q = \frac{\beta \log(n)}{n},$$

they demonstrate that exact recovery of u_1 as $n \to \infty$ is possible when $\alpha + \beta - 2\sqrt{\alpha\beta} > 2$, and impossible otherwise. Exact recovery of u_1 implies that there exists an estimator $\hat{u}_1$ such that $\forall j$ simultaneously,

$$u_1^{(j)} \cdot \hat{u}_1^{(j)} > 0, \text{ w.p. } \to 1 \text{ as } n \to \infty.$$

Additionally, the authors provide an efficient algorithm for exact recovery. They consider the following optimization problem:

$$\max_{\|u\|_2 = 1, u^{(j)} \in \{\pm \frac{1}{\sqrt{n}}\}, \sum_j u^{(j)} = 0} u^\top \left(A - \frac{p+q}{2} \mathbf{1}_n \mathbf{1}_n^\top \right) u.$$

Since this problem is NP-Hard, the authors solve an SDP relaxation, the solution of which is demonstrated to also be close to exact.

Alternately, if graph structure is very sparse (i.e., many isolated pairs) then subgroup recovery will be impossible. [MNS13] and [Mas14] provide a series of results for the sparse case. When, for constants a and b such that $p = \frac{a}{n}$ and $q = \frac{b}{n}$, recovery is impossible in general, but it is possible to do better than chance in certain cases. Specifically, when $(a-b)^2 \geq 2(a+b)$, there exists an estimator $\hat{u}_1$ and constant $\alpha > 0$ such that,

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left\{\hat{u}_1^{(j)} - u_1^{(j)}\right\} \to \frac{1}{2} - \alpha, \text{ as } n \to \infty.$$

Even using the techniques in our analysis, tighter bounds can be obtained by carefully analysing the concentration of the noise matrix. In our analysis, we assumed that the noise is $\mathsf{subG}(1)$, which can be tightened by considering a careful Bernstein concentration.

Summary: In this lecture, we first provided a proof for Davis-Kahan $\sin(\theta)$ theorem, that provides the following bound on the perturbations of top eigenvectors:

$$\min_{\varepsilon\in\{\pm 1\}} |u_1 - \varepsilon\hat{u}_1|_2 \leq \sqrt{2}\,|\sin\left(\measuredangle\left(u_1, \hat{u}_1\right)\right)| \leq 2\sqrt{2}\frac{\|A - \hat{A}\|_{\text{op}}}{\max\left(\lambda_1 - \lambda_2, \hat{\lambda}_1 - \hat{\lambda}_2\right)}.$$

Next, for the matrix denoising problem, and demonstrate that with high probability,

$$\min_{\varepsilon\in\{\pm 1\}} |\hat{u}_1 - \varepsilon u_1|_2^2 \leq 2\frac{\|E\|_{\text{op}}}{\lambda_1 - \lambda_2} \lesssim \frac{\sigma\sqrt{n}}{\lambda_1 - \lambda_2}.$$

Finally, we considered the community detection problem (stochastic block model), for which we bounded the misclassification error as follows.

$$L(u_1, \hat{u}_1) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left\{\text{sign}(\hat{u}_1^{(j)}) \neq \text{sign}(u_1^{(j)})\right\} \lesssim \frac{1}{(p-q)\sqrt{n}}.$$

We concluded the lecture with a short discussion on state-of-the-art results in the community detection problem.

References

[MNS13] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[ABH16] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, Jan 2016.

[Mas14] Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

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

Lecturer: P Rigollet
Scribe: Maggie Makar, Ziao Lin

Lecture 11
Mar. 12, 2020

Goals: In this lecture we will focus on *Principal component analysis* (PCA), where the task is to project a high dimensional vector X onto a low dimensional space. At the crux of PCA is studying Σ, the covariance matrix of X.

We assume that the true Σ follows a spiked covariance model. We consider the empirical estimator $\hat{\Sigma}$, and quantify how close it is to the true Σ in terms of Σ's eigenspace and dimension as well as number of samples.Our analysis will rely on the Davis-Kahan theorem from the previous 2 lectures.

1. SPIKED COVARIANCE MODEL

Consider the following problem. Suppose we observe some data $X_i, \ldots, X_n \sim \mathcal{N}_d(0, \Sigma)$. We want to consider some model that allows us to uncover a low dimensional space in which X lies (e.g., for visualization purposes). Specifically, we will consider a linear structure where we take a vector $v \in R^d$. The expectation of the observed matrix $X = [X_1, X_2, ..., X_n]^\top \in R^{n \times d}$ would be represented as $E[X] = Yv$, where $Y = [Y_1, Y_2, ..., Y_n]^\top \in R^{n \times 1}$ and $y_i \in R$.

Realistically, we would not observe perfectly aligned points. Instead, data is typically corrupted by some noise in the full d dimension. We denote the noise by Z and assume that $Z_1, \cdots, Z_n \sim \mathcal{N}(0, I_d)$, with $Z \perp\!\!\!\perp Y$. So we can represent the observed $X_i = Y_i v + Z_i$. Because Y_i, and Z_i might not be on the same scale, we introduce a tuning parameter $\sqrt{\theta}$ for some $\theta > 0$, and we say that $X_i = \sqrt{\theta} Y_i v + Z_i$. We also assume that v has been normalized, i.e. $|v|_2 = 1$. Since $Z \perp\!\!\!\perp Y$, we have that $X \sim \mathcal{N}(0, \Sigma)$ based on a linear transformation of a multivariate random vector also has a multivariate normal distribution, with

$$\begin{aligned}
\Sigma &= \mathbb{E}[X_i X_i^\top] \\
&= \mathbb{E}[(\sqrt{\theta} Y_i v + Z_i)(\sqrt{\theta} Y_i v + Z_i)^\top] \\
&= \theta \mathbb{E}[Y_i^2] v v^\top + \mathbb{E}[Z_i Z_i^\top] \\
&= \theta v v^\top + I_d
\end{aligned}$$

where the last equality follows from the fact that $\mathbb{E}[Y_i^2] = 1$, and $\mathbb{E}[Z_i Z_i^\top] = I_d$. When $|v|_2$ is fixed to be $= 1$, this model is referred to as the *spiked covariance model*. Under the spiked covariance model, we can claim the following:

Claim: v is an eigenvector of Σ.

This is because $\Sigma v = \theta(v^\top v)v + I_d v = (1 + \theta)v$. We also have that:

$$\begin{aligned}
&\max_{|u|_2 = 1} u^\top \Sigma u \\
&= \theta (u^\top v)^2 + 1 \\
&= v^\top v,
\end{aligned}$$

where the last equality follows from the fact that this quantity is maximized when u, and v are aligned. Knowing that $\forall u \perp\!\!\!\perp v$, $u^\top \Sigma u = 1 < 1 + \theta$. This identifies all our eignvalues:

$$\lambda_1 = (1+\theta) \geq \lambda_2 = 1 \geq \lambda_3 = 1 \ldots \lambda_d = 1$$

2. ESTIMATING Σ

We will take the empirical covariance estimate,

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

to be an estimator for Σ. By LLN, we have that this is a consistent estimator. We know that the largest eigenvector is v and the associated eigenvalue is λ_1. So if we want to identify what v is, we can apply Davis-Kahan:

$$|\sin(\angle(\hat{v}, v))| \leq \frac{2||\hat{\Sigma} - \Sigma||_{op}}{\lambda_1 - \lambda_2} = \frac{2||\hat{\Sigma} - \Sigma||_{op}}{\theta}$$

where $\hat{v}$ is the leading eigenvector of $\hat{\Sigma}$. This tells us that the norm we need to control in order to do PCA is the operator norm. Note that even if $\hat{\Sigma}$ and Σ is positive semidefinite since they are real symmetric matrices, the difference $E = \hat{\Sigma} - \Sigma$ in general is not guaranteed to be positive semidefinite. Thus we cannot directly apply the leading eigenvector u_1 into $u_1^T E u_1$ to get operator norm. We will instead move on to control this operator norm using ε-Nets.

Let $E := ||\hat{\Sigma} - \Sigma||_{op}$. We have that:

$$E_{jk} = \frac{1}{n} \sum_{i=1}^{n} X_i^{(j)} X_i^{(k)} - \mathbb{E}[X_i^{(j)} X_i^{(k)}]$$

Using the definition of the operator norm (see lecture 8, expression 3.30) and a previous result (see lecture 9, proof of Lemma), we have that:

$$||E||_{op} \leq 2 \max_{x \in \mathcal{N}_d, y \in \mathcal{N}_d} x^\top (\hat{\Sigma} - \Sigma) y,$$

where $\mathcal{N}_d$ is the $\frac{1}{4}$-net of $B_2(\mathbb{R}^d)$, and we can control $|\mathcal{N}_d|$ to get $|\mathcal{N}_d| \leq 9^d$. We have that:

$$x^\top (\hat{\Sigma} - \Sigma) y = \frac{1}{n} \sum_{i=1}^{n} (x^\top X_i)(y^\top X_i) - \mathbb{E}[(x^\top X_i)(y^\top X_i)]. \tag{2.1}$$

It turns out that the distribution of this variable is subexponential. To see that note that:

$$x^\top X_i \sim \mathcal{N}(0, x^\top \Sigma x).$$

If we take $|x|_2 \leq 1$, we have that $x^\top \Sigma x \leq ||\Sigma||_{op}$, we have that

$$x^\top X_i \sim subG(||\Sigma||_{op}).$$

Since the term 2.1 includes a product of 2 subGaussian variables, it is subExponential, which means that we will likely use Brenstien's inequality. To use Brenstien's inequality:

$$\begin{aligned}
&||(x^\top X_i)(y^\top X_i) - \mathbb{E}[(x^\top X_i)(y^\top X_i)]||_{\varphi_1} \\
&\leq ||(x^\top X_i)(y^\top X_i)||_{\varphi_1} + ||\mathbb{E}[(x^\top X_i)(y^\top X_i)]||_{\varphi_1} \\
&\leq ||(x^\top X_i)||_{\varphi_2}||(y^\top X_i)||_{\varphi_2} + ||(x^\top X_i)||_{\varphi_2}||(y^\top X_i)||_{\varphi_2} \\
&\leq 2\sqrt{||\Sigma||_{op}}\sqrt{||\Sigma||_{op}} \\
&\leq 2||\Sigma||_{op},
\end{aligned}$$

where the first inequality follows from triangle inequality, and the second inequality is an application of Jensen's inequality due to the convexity of φ_1-norm plus the inequality s.t. $||xy||_{\varphi_1} \leq ||x||_{\varphi_2}||y||_{\varphi_2}$. The third inequality follows from the property of subGaussian variables. We can now apply Bernstein:

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^{n}(x^\top X_i)(y^\top X_i) - \mathbb{E}[(x^\top X_i)(y^\top X_i)] > t\bigg) \\
&\leq \sum_{x,y} \exp\bigg(-Cn\bigg(\frac{t^2}{||\Sigma||_{op}^2} \wedge \frac{t}{||\Sigma||_{op}}\bigg)\bigg) \\
&\leq 9^{2d} \exp\bigg(-Cn\bigg(\frac{t^2}{||\Sigma||_{op}^2} \wedge \frac{t}{||\Sigma||_{op}}\bigg)\bigg),
\end{aligned}$$

for some constant C. And the second inequality follows from the fact that the terms in the sum do not depend on x, y.
Now let's denote the desired threshold to be δ, then resolve the above inequality we will get: $t \leq C||\Sigma||_{op}[\sqrt{\frac{d+lg(1/\delta)}{n}} + \frac{d+lg(1/\delta)}{n}]$ for some constant C. Then we can hopefully control $||E||_{op} \leq C||\Sigma||_{op}\sqrt{\frac{d}{n}}$ for some constant C. Plug in the results back to Davis-Kahan, we eventually get a bound on the difference in angle between the two leading eigenvenctors $\hat{v}$ and v:

$$|\sin(\angle(\hat{v}, v))| \leq C\frac{1+\theta}{\theta}\sqrt{\frac{d}{n}}$$

for some constant C.
The result can be generalized to multiple spiked model with some scaling factor proportional to the square root of number of spikes.

3. SPARSE PCA

A slightly different model that could have generated Σ is known as the *sparse spiked model*. In this model v is assumed to be sparse. Consider the example where $v \in \mathbb{R}^2$. The spiked

covariance model assumes that v_1, v_2 are a linear combination of possibly all the dimensions in the original space. Instead, the sparse spiked covariance matrix assumes that v_1, v_2 are a linear combination of a small subset of cardinality s contribute to the principle directions v_1, v_2. In that case, we would want to include a sparsity constraint when estimating $\hat{v}_1, \hat{v}_2$. The estimator becomes:

$$\hat{v} = \max_{|u|_2=1, u \in B_0(s)} u^\top \hat{\Sigma} u.$$

Because we're considering B_0 in the constraint, this problem is computationally very expensive. Significant research has been done to find efficient ways to solve this problem (e.g., convex relaxations, ScoTLASS)

Summary: By applying Davis-Kahan theorem, we derive a upper bound on the difference in angle between the two leading eigenvectors in sample covariance estimator $\hat{\Sigma}$ and the truth covariance matrix Σ in *Principal component analysis* (PCA). The results and methods used here are generalizable to multiple spiked model.

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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lectures 14 & 15
Apr. 1, 2020

Goals: In this lecture, we motivate the study of the maxima of certain stochastic processes. The first motivation comes from the Kolmogorov-Smirnov test, and the second — from Statistical Learning. We present two approaches to analyzing the supremum of an empirical process: bracketing and symmetrization.

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.

First, we present a couple of motivating examples.

1. KOLMOGOROV'S GOODNESS-OF-FIT TEST

Given n indepenent 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. The Glivenko-Cantelli Theorem (1933) states that

$$D_n = \sup_{\theta \in \mathbb{R}} |F_n(\theta) - F(\theta)| \to 0 \quad a.s.$$

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.

2. STATISTICAL LEARNING

2.1 Empirical Risk Minimization

Let $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be n i.i.d. copies of a random variable (X, Y) with distribution $P = P_X \times P_{Y|X}$, where the X variable lives in some abstract space $\mathcal{X}$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$. Fix a loss function $\boldsymbol{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Fix a class of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$. 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: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0,1\}$, $\mathcal{F} = \{x \mapsto (\text{sign}(\langle w, x\rangle) + 1)/2 : w \in \mathsf{B}_2\}$, $\boldsymbol{\ell}(a,b) = \mathbf{1}\{a \ne b\}$

We now define expected loss (error) as

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

and empirical loss (error) as

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

For any $f^* \in \mathcal{F}$, The decomposition

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

holds true. By definition of ERM, the second term is nonpositive. If 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 not zero in expectation (why?).

Let us proceed by taking expectation (with respect to S) of both sides:

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \mathbf{L}(f^*) \leq \mathbb{E}\left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] \leq \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] \tag{2.1}$$

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 $\mathcal{F}$.

2.2 Classification

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

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]. \tag{2.2}$$

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, (2.2) becomes

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[P(X \leq \theta) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq \theta\}\right] = \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[F(\theta) - F_n(\theta)\right]. \tag{2.3}$$

which is precisely the quantity from the beginning of the lecture (albeit without absolute values and in expectation). Again, (2.3) 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 (e.g. Lecture 5) 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?

We will now present two approaches for upper-bounding (2.3); both extend to the general case of (2.2).

2.2.1 The bracketing approach

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,

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

and thus

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

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

2.2.2 The symmetrization approach

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.

Observe that

$$\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i' \le \theta\right\}\right]$$

where $X_1', \ldots, X_n'$ are n independent copies of X. We have the following upper bound on (2.3):

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.4}$$

$$\le \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.5}$$

by convexity of the sup. Now, since distribution of $\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}$ is the same as the distribution of $-\left(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}\right)$, we can insert arbitrary signs ϵ_i without changing the expected value:

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\})\right]. \tag{2.6}$$

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

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.7}$$

$$\le \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\})\right], \tag{2.8}$$

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

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i' \le \theta\right\}\right] + \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}-\epsilon_i\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.9}$$

$$= 2\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.10}$$

by symmetry of Rademacher random variables.

Now comes the key step. Let us condition on $X_1, \ldots, X_n$ and think of the random variables

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

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

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

Since each V_θ is 1-subGaussian, and we get an overall upper bound

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

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

2.3 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.

In view of (2.1), the upper bounds we derived guarantee (modulo the fact that we omitted "$1-2Y$") that for empirical risk minimization,

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

It is worth stating the symmetrization lemma more formally:

Lemma: 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

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

Furthermore,

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

Proof. We only prove the second part since the first statement was proved earlier (for indicators). Write

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

Consider the first term on the RHS:

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

As for the second term,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbb{E}f\right] \leq \sup_{f\in\mathcal{F}}|\mathbb{E}f|\cdot\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \tag{2.11}$$

□

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,

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

This justifies omitting $(1-2Y)$ for binary classification, at least with the symmetrization approach.

2.4 Empirical Process

Let us also define an empirical process:

Definition: 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 $\mathbb{P}_n$ is the empirical measure. You may also see the notation

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

Summary: We presented two approaches for analyzing the supremum of the difference of expected and empirical values: bracketing and symmetrization. We stated the symmetrization lemma in full generality.

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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 16 & 17
Apr. 7 & 9, 2020

Goals: We will study suprema of stochastic processes with a certain metric structure. We will develop a single-scale covering argument and then improve it through a chaining technique.

1. SUPREMA OF GAUSSIAN AND SUBGAUSSIAN PROCESSES

Definition: 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

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

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 $U_{\theta_1}, \dots, U_{\theta_n}$, 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.

1.1 SubGaussian Processes

Definition: 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 $G_\theta = \langle g, \theta\rangle$, $g = (g_1, \dots, g_n)$, $g_i \sim N(0,1)$ 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 $R_\theta = \langle \epsilon, \theta \rangle$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, ϵ i.i.d. Rademacher. Again, take $d(\theta, \theta') = \|\theta - \theta'\|$. Then

$$R_\theta - R'_\theta = \langle \epsilon, \theta - \theta' \rangle$$

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: We will call $\widehat{\mathcal{R}}(\Theta) = \mathbb{E} \sup_{\theta \in \Theta} \langle \epsilon, \theta \rangle$ 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 $\widehat{\mathcal{G}}(\Theta)$.

1.1.1 A few examples

Let $U_\theta = \langle \epsilon, \theta \rangle$, $\Theta \subset \mathbb{R}^n$, and take Euclidean distance as the metric. We have

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

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

A few other sets:

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

and

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

However, we observe that

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

and yet 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

$$\widehat{\mathcal{R}}(\Theta) \lesssim \widehat{\mathcal{G}}(\Theta) \lesssim \sqrt{\log n} \cdot \widehat{\mathcal{R}}(\Theta). \tag{1.1}$$

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:

$$\widehat{\mathcal{R}}(\Theta) \lesssim \operatorname{diam}(\Theta)\sqrt{\log \operatorname{card}(\Theta)},$$

$$\widehat{\mathcal{R}}(\operatorname{conv}(\Theta)) = \widehat{\mathcal{R}}(\Theta),$$

$$\widehat{\mathcal{R}}(c\Theta) = |c|\widehat{\mathcal{R}}(\Theta) \quad \text{for constant } \ c$$

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

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

$$\mathbb{E} \max_{(\theta,\theta')\in A} U_\theta - U_{\theta'} \leq \max_{(\theta,\theta')\in A} d(\theta,\theta') \cdot \sqrt{2\log \operatorname{card}(A)} \tag{1.2}$$

How do we go beyond finite cover?

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

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

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

$$\begin{aligned} U_\theta - U_{\theta'} &= U_\theta - U_{\hat{\theta}} + U_{\hat{\theta}} - U_{\hat{\theta}'} + U_{\hat{\theta}'} - U_{\theta'} && (1.3)\\ &\leq 2 \sup_{d(\theta,\theta')\leq\delta} (U_\theta - U_{\theta'}) + \sup_{\hat{\theta},\hat{\theta}'\in\widehat{\Theta}} (U_{\hat{\theta}} - U_{\hat{\theta}'}) && (1.4)\end{aligned}$$

The last term is

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

□

1.3 Example: Rademacher/Gaussian processes

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

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

Hence,

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \le 2\delta\sqrt{n} + 2\mathrm{diam}(\Theta) \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|_2, \delta)} \tag{1.5}$$

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.

Let's step back and ask what kind of generic statement we can say about a d-dimensional subset of a Euclidean ball. Suppose that $\Theta \subseteq \mathsf{B}_2^n$ and assume that Θ lives in a d-dimensional subspace. Then

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

and by taking $\delta = \sqrt{d/n}$ the estimate in (1.5) becomes

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \le 2\sqrt{d} + 4\sqrt{d \log\left(1 + 2\sqrt{n/d}\right)} \lesssim \sqrt{d \log(n/d)}. \tag{1.6}$$

Here we tacitly assumed $d < n$. Recall that in Lecture 5 we obtained an upper bound of $O(\sqrt{d})$ in this setup by having a cover at scale $1/2$ and comparing the supremum to the maximum *multiplicatively*. Another way to see it is

$$\mathbb{E} \sup_{\theta \in \mathsf{B}_2^d} \langle \epsilon, \theta \rangle = \mathbb{E} \|\epsilon\| = \sqrt{d}$$

and similarly

$$\mathbb{E} \sup_{\theta \in \mathsf{B}_2^d} \langle g, \theta \rangle = \mathbb{E} \|g\| \le \sqrt{\sum_{i=1}^n \mathbb{E} g_i^2} \le \sqrt{d}$$

Hence, we lost a logarithmic factor by appealing to the general machinery of the previous section. We will also see that we can remove the extraneous logarithm by looking at a cover at multiple scales.

1.4 Function class

In particular, we will be interested in the following indexing set Θ. Let $x_1, \ldots, x_n$ be fixed, and let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$. We call

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

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

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

where $\theta = (f(x_1), \ldots, f(x_n))$ and $\theta' = (f'(x_1), \ldots, f'(x_n))$, $f, f' \in \mathcal{F}$. With these definitions, we can define a Gaussian or Rademacher process with respect to Θ and d.

Important point: the symmetrization lemma allows us to relate supremum of the empirical process to supremum of a Rademacher process.

1.4.1 Example: Linear Function Class

We now focus on a specific example of linear functions

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle : w \in \mathsf{B}_2^d\}.$$

Then for fixed $x_1, \ldots, x_n \in \mathsf{B}_2^d$, a direct calculation yields

$$\mathbb{E} \sup_{w \in \mathsf{B}_2^d} \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i \langle w, x_i \rangle = \frac{1}{\sqrt{n}} \mathbb{E} \left\| \sum_{i=1}^n \epsilon_i x_i \right\| \leq \frac{1}{\sqrt{n}} \sqrt{\mathbb{E} \left\| \sum_{i=1}^n \epsilon_i x_i \right\|^2} \leq 1. \tag{1.7}$$

Let's see if we can recover this via our machinery. After all, the above object is precisely a supremum of a subgaussian process. Observe that

$$\Theta = \frac{1}{\sqrt{n}} \mathcal{F}|_{x_1, \ldots, x_n} \subseteq \frac{1}{\sqrt{n}} \mathsf{B}_\infty^n \subset \mathsf{B}_2^n \tag{1.8}$$

and that

$$\mathcal{F}|_{x_1, \ldots, x_n} = \left\{ (\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle) \in \mathbb{R}^n : w \in \mathsf{B}_2^d \right\} = \{Xw : w \in \mathsf{B}_2^d\}$$

is a subset of a d-dimensional subspace. Hence, appealing to the previous example (1.6), we get an upper bound of $O(\sqrt{d \log(n/d)})$.

Looking back at (1.7), however, we see that we also gained an extra $\sqrt{d}$ factor, which can be a big loss in high-dimensional situations. Where did we gain it? We can see that the set $\frac{1}{\sqrt{n}} \mathcal{F}|_{x_1, \ldots, x_n}$ in (1.8) is, in fact, much smaller than a d-dimensional Euclidean ball.

2. CHAINING

Theorem: Let $(U_\theta)_{\theta \in \Theta}$ be a (mean-zero) subGaussian stochastic process with respect to a metric d. Let $D = \operatorname{diam}(\Theta)$. Then for any $\delta \in [0, D]$,

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 2\mathbb{E} \sup_{d(\theta, \theta') \leq \delta} (U_\theta - U'_\theta) + 8\sqrt{2} \int_{\delta/4}^{D/2} \sqrt{\log \mathcal{N}(\Theta, d, \varepsilon)} d\varepsilon \tag{2.9}$$

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

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

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

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

For $\theta_N \in \Theta_N$,

$$U_{\theta_N} = \sum_{i=1}^{N} U_{\theta_i} - U_{\pi_{i-1}(\theta_i)} + U_{\theta_0} \tag{2.10}$$

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

$$\mathbb{E} \max_{\theta,\theta' \in \Theta_N} U_\theta - U_{\theta'} \leq \sum_{i=1}^{N} \mathbb{E} \max_{\theta_i \in \Theta_i} (U_{\theta_i} - U_{\pi_{i-1}(\theta_i)}) + \sum_{i=1}^{N} \mathbb{E} \max_{\theta'_i \in \Theta_i} (U_{\pi_{i-1}(\theta'_i)} - U_{\theta'_i}) \tag{2.11}$$

$$\leq 2 \sum_{i=1}^{N} D 2^{-(i-1)} \sqrt{2 \log \mathcal{N}(\Theta, d, 2^{-i} D)} \tag{2.12}$$

$$= 8 \sum_{i=1}^{N} D 2^{-(i+1)} \sqrt{2 \log \mathcal{N}(\Theta, d, 2^{-i} D)} \tag{2.13}$$

$$\leq 8 \sum_{i=1}^{N} \int_{2^{-(i+1)} D}^{2^{-i} D} \sqrt{2 \log \mathcal{N}(\Theta, d, \varepsilon)} d\varepsilon \tag{2.14}$$

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

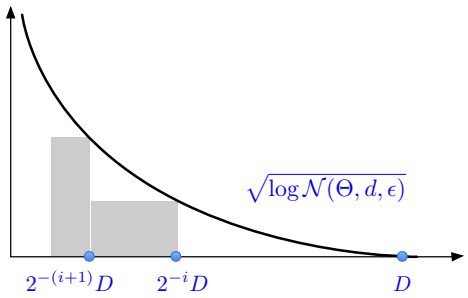


Figure 1: Illustration of the Dudley integral upper bound

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

Theorem: 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).

Summary: We now have tools to analyze suprema of subGaussian processes in terms of the geometric descriptions (e.g. covering numbers) of the indexing set. These techniques will be applied to a number of parametric and nonparametric regression and classification problems in the subsequent lectures, after we introduce a few more tools such as combinatorial parameters.

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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 18 & 19
Apr. 14 & 16, 2020

Goals: We continue the investigation of model complexity through the lens of covering/packing numbers and combinatorial dimensions. These are convenient tools for upper/lower bounding the supremum of the Rademacher or empirical process.

1. COVERING AND PACKING

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

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

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

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

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: An ε-net (or, ε-cover) of $\mathcal{F}$ with respect to $L^2(P)$ is a set of functions $f_1, \dots, f_N$ such that

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

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

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

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

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

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

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

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

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

$$\mathcal{N}(\mathcal{F}, L^p(P_n), \varepsilon) \leq \mathcal{N}(\mathcal{F}, L^q(P_n), \varepsilon)$$

for $p \leq q$ since $\|f\|_{L^p(P_n)}$ increases with p. Note that this is different for unweighted metrics: e.g. $\|x\|_p$ is nonincreasing in p, and hence $\mathcal{N}(\Theta, \|\cdot\|_p, \varepsilon)$ is also nonincreasing in p.

Definition: An ε-packing of $\mathcal{F}$ with respect to $L^p(P_n)$ is a set $f_1, \ldots, f_N \in \mathcal{F}$ such that

$$\frac{1}{n}\sum_{i=1}^n |f_j(x_i) - f_k(x_i)|^p \geq \varepsilon^p$$

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

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

$$\mathcal{D}(\mathcal{F}, L^p(P), 2\varepsilon) \leq \mathcal{N}(\mathcal{F}, L^p(P), \varepsilon) \leq \mathcal{D}(\mathcal{F}, L^p(P), \varepsilon)$$

2. UPPER AND LOWER BOUNDS FOR RADEMACHER AVERAGES

As before, we let $U_\theta = \langle \epsilon, \theta \rangle$, $\Theta = \frac{1}{\sqrt{n}}\mathcal{F}|_{x_1,\ldots,x_n}$, and d Euclidean distance. Then from last lecture

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(x_i) &= \mathbb{E}\sup_{\theta\in\Theta} U_\theta \\ &\leq 2\delta\sqrt{n} + 8\sqrt{2}\int_{\delta/4}^{D/2} \sqrt{\log \mathcal{N}(\Theta, d, \varepsilon)}d\varepsilon\end{aligned}$$

Trivially,

$$\mathcal{N}(\Theta, d, \varepsilon) = \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon).$$

Corollary: For any $X_1, \ldots, X_n$,

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

with $D = \sup_{f,g\in\mathcal{F}} \|f - g\|_n \leq 2\sup_{f\in\mathcal{F}} \|f\|_n \leq 2\sup_{f\in\mathcal{F}} \|f\|_\infty$.

Putting together the symmetrization lemma and above Corollary, we have

Corollary: 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

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\} \leq \mathbb{E} \inf_{\delta \geq 0} \left\{ 16\delta + \frac{24}{\sqrt{n}} \int_{\delta}^{D} \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon)} d\varepsilon \right\} \tag{2.1}$$

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

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: For any $X_1, \ldots, X_n$,

$$\mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \geq \frac{c}{\sqrt{\log n}} \cdot \sup_{\alpha \geq 0} \alpha \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \alpha)}{n}}$$

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).

3. 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: We will say that a class $\mathcal{F}$ is *parametric* if for any empirical measure P_n,

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

We will say that $\mathcal{F}$ is *nonparametric* if for any empirical measure P_n,

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \asymp \left(\frac{1}{\epsilon}\right)^{p}. \tag{3.2}$$

The requirement that (3.2) 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: $|f|_\infty \leq 1$. This provides an upper bound on the diameter: $D/2 \leq 1$. Then, taking $\delta = 0$, conditionally on $X_1, \ldots, X_n$,

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

Here it's useful to note that

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

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

Theorem: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$ be a class of binary-valued functions with VC dimension $\text{vc}(\mathcal{F}) = d$. 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.

3.1 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

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

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

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

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 doesn't depend on n.

3.2 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 $p > 2$.

3.3 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 in Lecture 16 was to discretize the set B_2^d to create a ε-net $w_1, \ldots, w_N$ of size $\mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon)$. Clearly, for any w and the corresponding ε-close element w_j of the cover,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (\langle w, x_i\rangle - \langle w_j, x_i\rangle)^2 &\le \max_{i\in[n]} \langle w - w_j, x_i\rangle^2 \\ &\le \max_{i\in[n]} \|w - w_j\|^2 \cdot \|x_i\|^2 \\ &\le \varepsilon^2.\end{aligned}$$

Hence,

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \le \mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon). \tag{3.3}$$

In fact, a much stronger statement can be made: Since for any $x \in \mathcal{X}$

$$|\langle w, x\rangle - \langle w_j, x\rangle| \le \|w - w_j\| \, \|x\| \le \varepsilon,$$

the cover of the parameter space induces a cover of the function class *pointwise* (in the sup-norm $\|f - g\|_\infty = \sup_{x\in\mathcal{X}} |f(x) - g(x)|$) over the domain:

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \le \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon). \tag{3.4}$$

Recall that the covering number of B_2^d is

$$\left(1 + \frac{2}{\varepsilon}\right)^d.$$

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 (3.4) 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.*

3.4 A more general result (Optional)

We have that for any fixed function

$$\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X))\right| \le \operatorname{var}(f)^{1/2} = \|f - \mathbb{E}f\|_{L^2(P)}.$$

Obviously this implies

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

If we could ever prove

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

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 $F(x) \le 1$, we shall assume that $\|F\|_{L^2(P)}^2 = \mathbb{E}F(X)^2 \le \infty$ 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

$$\int_0^D \sqrt{H(1/x)}dx \le C_H D\sqrt{H(1/D)}$$

for all $D \in (0,1]$, and suppose that

$$\sup_Q \log 2\mathcal{N}(\mathcal{F}, L^2(Q), \tau\|F\|_{L^2(Q)}) \le H(1/\tau)$$

for all $\tau > 0$. With this control on Koltchinskii-Pollard entropy, it follows that

$$\mathbb{E}\sup\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X))\right| \lesssim \sigma\sqrt{H\left(\frac{2\|F\|_{L^2(P)}}{\sigma}\right)} \tag{3.5}$$

if n is large enough. We refer to Giné & Nickl "Mathematical Foundations of Infinite-Dimensional Statistical Models" for more details, in particular Theorem 3.5.6 and the following corollaries.

Remarkably, under additional mild conditions on size of n, the inequality (3.5) 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 $\|F\|^2_{L^2(P)} = \mathbb{E}\sup_f |f(X)|^2$.

4. COMBINATORIAL PARAMETERS

Let us gain some intuition for what can make $\widehat{\mathcal{R}}(\Theta)$ large. First, recall that

$$\widehat{\mathcal{R}}(\{\pm 1\}^n) = \mathbb{E}\sup_{\theta\in\{\pm1\}^n}\langle\theta,\epsilon\rangle = n.$$

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 $\mathcal{F}$.

4.1 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: 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

$$\mathrm{vc}(\mathcal{F}) = \max\{n : \mathcal{F} \text{ shatters some } \ x_1, \ldots, x_n\}$$

Lemma (Sauer-Shelah-Vapnik-Chervonenkis): If $\mathrm{vc}(\mathcal{F}) = d < \infty$,

$$\mathrm{card}\left(\mathcal{F}|_{x_1,\ldots,x_n}\right) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$$

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).

4.2 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: 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

$$\forall \epsilon \in \{\pm 1\}^n, \exists f \in \mathcal{F} \text{ s.t. } \begin{cases} f(x_t) \geq s_t + \alpha/2 & \text{if } \epsilon = +1 \\ f(x_t) \leq s_t - \alpha/2 & \text{if } \epsilon = -1 \end{cases}$$

The combinatorial dimension $\mathrm{vc}(\mathcal{F}, \alpha)$ of $\mathcal{F}$ (on domain $\mathcal{X}$) at scale α is defined as the size n of the largest shattered set.

4.2.1 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: $\mathrm{vc}(\mathcal{F}, \epsilon) \leq \epsilon^{-1}$. 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.

4.2.2 Control of covering numbers

The following generalization of the earlier result for binary-valued functions is due to Mendelson and Vershynin:

Theorem: Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to [-1, 1]$. Then for any distribution P,

$$\mathcal{N}(\mathcal{F}, L_2(P), \varepsilon) \leq \left(\frac{c}{\varepsilon}\right)^{c \cdot \mathrm{vc}(\mathcal{F}, \varepsilon/c)}$$

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 $p < 2$.

4.3 Scale-sensitive dimension of linear class via Perceptron

In this section, we will prove that

Proposition: 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

$$\mathrm{vc}(\mathcal{F}, \alpha) \lesssim 16\alpha^{-2}.$$

We turn to the Perceptron algorithm, defined as follows. We start with $\widehat{w}_0 = 0$. At time $t = 1, \ldots, T$, we observe $x_t \in \mathcal{X}$ and predict $\widehat{y}_t = \mathrm{sign}(\langle \widehat{w}_t, x_t\rangle)$, 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 (Novikoff'62): For any sequence $(x_1, y_1), \ldots, (x_T, y_T) \in \mathsf{B}_2^d \times \{\pm 1\}$ the Perceptron algorithm makes at most γ^{-2} mistakes, where γ is the margin of the sequence, defined as

$$\gamma = \max_{w^* \in \mathsf{B}_2^d} \min_t y_t \langle w^*, x_t\rangle$$

Proof. If a mistake is made on round t,

$$\|\widehat{w}_{t+1}\|^2 = \|\widehat{w}_t + y_t x_t\|^2 \leq \|\widehat{w}_t\|^2 + 2y_t\langle \widehat{w}_t, x_t\rangle + 1 \leq \|\widehat{w}_t\|^2 + 1$$

Denote the number of mistakes at the end as m. Then $\|\widehat{w}_T\|^2 \leq m$. Next, for w^*,

$$\gamma \leq \langle w^*, y_t x_t\rangle = \langle w^*, \widehat{w}_{t+1} - \widehat{w}_t\rangle,$$

and so by summing and telescoping, $m\gamma \leq \langle w^*, \widehat{w}_T\rangle \leq \sqrt{m}$. 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

$$\epsilon_i(\langle w_\epsilon, x_i\rangle - s_i) \geq \alpha/2.$$

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

$$\epsilon_i\langle \tilde{w}_\epsilon, \tilde{x}_i\rangle \geq \alpha/2.$$

while the norms are at most $\sqrt{2}$:

$$\|\tilde{w}_\epsilon\|^2 = \|w_\epsilon\|^2 + 1 \leq 2, \qquad \|\tilde{x}_i\|^2 \leq 2$$

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 $y = (y_1, \ldots, y_n)$ 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. If there is time, we will mention these towards the end of the course.

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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 20 & 21
Apr. 21 & 23, 2020

Goals:

1. 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 $\mathcal{F}$.

Risk of a function f is defined as

$$\mathbb{E}(f(X) - f^*(X))^2 = \|f - f^*\|^2_{L^2(P)} = \|f - f^*\|^2$$

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 $\mathcal{S}$.

Note that for any function f,

$$\begin{aligned}
\mathbb{E}(f(X) - Y)^2 - \min_h \mathbb{E}(h(X) - Y)^2 &= \mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\
&= \mathbb{E}(f(X) - f^*(X) + f^*(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\
&= \mathbb{E}(f(X) - f^*(X))^2
\end{aligned}$$

Question: given i.i.d. data $\mathcal{S}$, can we select estimator $\widehat{f}$ such that risk

$$\left\| \widehat{f} - f^* \right\|^2$$

is small in expectation or high-probability (with respect to the draw of $\mathcal{S}$)? Without further assumptions this is not possible.

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 $\mathcal{F}$.

- Misspecified case (agnostic learning in CS community): Redefine goal as

$$\begin{aligned} &\left\|\widehat{f} - f^*\right\|^2 - \min_{f\in\mathcal{F}} \|f - f^*\|^2 \\ &= \mathbb{E}(\widehat{f}(X) - Y)^2 - \min_{f\in\mathcal{F}} \mathbb{E}(f(X) - Y)^2 \end{aligned} \tag{1.1}$$

  but do not insist that $f^* \in \mathcal{F}$. Upper bounds on (1.1) are called Oracle Inequalities in statistics, while the prediction form has been also studied in statistical learning theory.

We see that the problem of prediction and the problem of estimation naturally coincide for square loss. Moreover, the misspecified problem arises naturally as a relaxation of an assumption on the form of the distribution.

Here, the road naturally forks into at least several paths: analyze the well-specified case, analyze the misspecified case, or change the loss function altogether. Let us briefly consider the last generalization.

2. PREDICTION WITH OTHER LOSS FUNCTIONS

This will be a brief but useful detour. Consider changing the loss function in the prediction problem (1.1) on the previous page:

$$\mathbb{E}\boldsymbol{\ell}(f(X), Y) - \min_{f\in\mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) \tag{2.2}$$

for some $\boldsymbol{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. In Lecture 14 we already showed 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

$$\mathbb{E}\boldsymbol{\ell}(\widehat{f}(X), Y) - \min_{f\in\mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) \le \mathbb{E} \sup_{f\in\mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}(f(X_i), Y_i).$$

The latter is at most

$$2\mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{\ell}(f(X_i), Y_i) \tag{2.3}$$

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 (Contraction): Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz, $i = 1, \ldots, n$. Let $\Theta \subset \mathbb{R}^n$ and $\phi \circ \theta = (\phi_1(\theta_1), \ldots, \phi_n(\theta_n))$ for $\theta \in \Theta$. Denote $\phi \circ \Theta = \{\phi \circ \theta : \theta \in \Theta\}$. Then

$$\widehat{\mathcal{R}}(\phi \circ \Theta) \le \widehat{\mathcal{R}}(\Theta).$$

Proof. Conditionally on $\epsilon_1, \dots, \epsilon_{n-1}$,

$$\begin{aligned}
\mathbb{E}_{\epsilon_n} \sup_{\theta\in\Theta} \langle \phi\circ\theta, \epsilon\rangle &= \frac{1}{2}\left(\sup_{\theta\in\Theta}\left\{\langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \phi_n(\theta_n)\right\} + \sup_{\theta'\in\Theta}\left\{\langle \phi\circ\theta'_{1:n-1}, \epsilon_{1:n-1}\rangle - \phi_n(\theta'_n)\right\}\right) \\
&\le \frac{1}{2}\sup_{\theta,\theta'\in\Theta} \langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \langle \phi\circ\theta'_{1:n-1}, \epsilon_{1:n-1}\rangle + |\theta_n - \theta'_n| \\
&= \frac{1}{2}\sup_{\theta,\theta'\in\Theta} \langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \langle \phi\circ\theta'_{1:n-1}, \epsilon_{1:n-1}\rangle + \theta_n - \theta'_n \\
&= \frac{1}{2}\left(\sup_{\theta\in\Theta}\left\{\langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \theta_n\right\} + \sup_{\theta'\in\Theta}\left\{\langle \phi\circ\theta'_{1:n-1}, \epsilon_{1:n-1}\rangle - \theta'_n\right\}\right) \\
&= \mathbb{E}_{\epsilon_n} \sup_{\theta\in\Theta} \langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \epsilon_n\theta_n
\end{aligned}$$

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 $\phi_i(\cdot) = \ell(\cdot, Y_i)$. As long as these functions are L-Lipschitz, contraction lemma gives

$$\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(f(X_i), Y_i) \le L\cdot \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) = L\cdot \frac{1}{n}\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F}|_{X_1,\dots,X_n}), \tag{2.4}$$

the (expected) Rademacher averages of $\mathcal{F}$. The argument can be seen as a generalization of the argument we did in Lecture 14 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 (see Lecture 14) in the sequence of inequalities

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \mathbf{L}(f^*) \le \mathbb{E}\left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] \le \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] \tag{2.5}$$

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.

Next few lectures will be on nonparametric regression. We will start with well-specified models.

3. NONPARAMETRIC REGRESSION: WELL-SPECIFIED CASE

We will start with "fixed design": $x_1, \dots, x_n \in \mathcal{X}$ 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

$$\mathbb{E}_\eta \left\| \widehat{f} - f^* \right\|_{L^2(P_n)}^2,$$

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 $\mathbb{E} \left\| \widehat{f} - f^* \right\|_n^2$.

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

$$\|f^* - Y\|_n^2 \geq \left\| \widehat{f} - Y \right\|_n^2 = \left\| \widehat{f} - f^* + f^* - Y \right\|_n^2 = \left\| \widehat{f} - f^* \right\|_n^2 + \|f^* - Y\|_n^2 + 2\langle \widehat{f} - f^*, f^* - Y \rangle_n$$

where $\langle a, b \rangle_n = \frac{1}{n} \langle a, b \rangle$. Thus,

$$\left\| \widehat{f} - f^* \right\|_n^2 \leq 2\langle \eta, \widehat{f} - f^* \rangle_n \tag{3.6}$$

which we will call *the basic inequality*.

3.1 Informal intuition for localization

Before developing the localization approach, we provide some intuition. The first intuition comes from viewing (3.6) as a fixed point.

Let's assume for simplicity that η_i are 1-subGaussian. For $a \in \mathbb{R}^n$, we have that with high probability

$$\langle \eta, a \rangle \lesssim \|a\|$$

Hence, if it holds that

$$\|a\|^2 \leq \langle \eta, a \rangle,$$

then $\|a\| \lesssim 1$.

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 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.

3.2 1st approach to localization: ratio-type inequalities

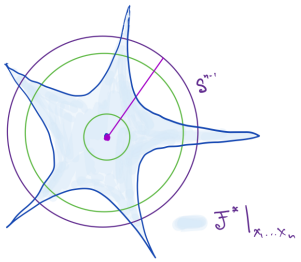
To simplify the proof somewhat, we will assume that $\eta_1, \ldots, \eta_n$ are independent standard normal $N(0, 1)$.

We proceed as in the linear case earlier in the course. First, we divide both sides of the Basic Inequality (3.6) 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:

$$\left\|\widehat{f} - f^*\right\|_n \leq 2 \sup_{f \in \mathcal{F}} \langle \eta, \frac{f - f^*}{\|f - f^*\|_n} \rangle_n \tag{3.7}$$

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

$$\left\| \widehat{f} - f^* \right\|_n \leq u + 2 \sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n \tag{3.8}$$

Consider the following assumption:

Definition: A class $\mathcal{H}$ is *star-shaped* (around 0) if $h \in \mathcal{H}$ implies $\lambda h \in \mathcal{H}$ for $h \in [0, 1]$. 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 (3.8) 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 (3.8) as

$$\begin{aligned} \left\| \widehat{f} - f^* \right\|_n &\leq u + \frac{2}{u} \sup_{g \in \mathcal{F}^* : \|g\|_n = u} \langle \eta, g \rangle_n \\ &\leq u + \frac{2}{u} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq u} \langle \eta, g \rangle_n \end{aligned} \tag{3.9}$$

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 (3.9) with $G(u)$, the natural balance between the two terms would be

$$u = \frac{2}{u} G(u)$$

Definition: 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: 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,

$$G(u) = u\frac{G(u)}{u} \leq u\frac{G(\delta_n)}{\delta_n} \leq u\delta_n/2 \leq u^2/2. \tag{3.10}$$

To formally replace $Z(u)$ with $G(u)$ in the balancing equation, we need a concentration result.

Lemma (Gaussian Concentration): Let $\eta = (\eta_1, \ldots, \eta_n)$ 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 $u > 0$,

$$\mathbb{P}\left(Z(u) - \mathbb{E}Z(u) \geq t\right) \leq \exp\left\{-\frac{nt^2}{2u^2}\right\} \tag{3.11}$$

In particular, by setting $t = u^2$,

$$\mathbb{P}\left(Z(u) \geq G(u) + u^2\right) \leq \exp\left\{-\frac{nu^2}{2}\right\} \tag{3.12}$$

In light of (3.10), we have proved

Lemma: Assuming $\mathcal{F}^*$ is star-shaped, with probability at least $1 - \exp\left\{-\frac{nu^2}{2}\right\}$,

$$Z(u) \leq 1.5u^2 \tag{3.13}$$

for any $u \geq \delta_n$.

Thus, from (3.9), we have

$$\left\|\widehat{f} - f^*\right\|_n \leq 4u \tag{3.14}$$

with probability at least $1 - \exp\left\{-\frac{nu^2}{2}\right\}$, for any $u \geq \delta_n$. Squaring both sides, yields

Theorem: Assume $x_1, \ldots, x_n$ are fixed, $\eta_1, \ldots, \eta_n$ 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

$$\mathbb{P}\left(\left\|\widehat{f} - f^*\right\|_n^2 \geq 16s\delta_n^2\right) \leq \exp\left\{-\frac{ns\delta_n^2}{2}\right\} \tag{3.15}$$

for any $s \geq 1$. In particular, this implies

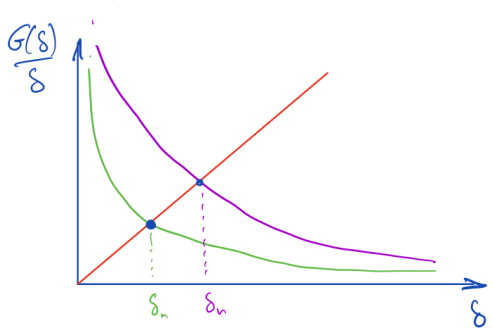
$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 \lesssim \delta_n^2 + \frac{1}{n}.$$

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 ψ has the *subroot* property:

$$\psi(ra) \leq \sqrt{a}\psi(r)$$

using the same type of proof as above. The fixed point then reads as the smallest r such that $\psi(r) \leq r$ (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.

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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 22 & 23
Apr. 28 & 30, 2020

1. NONPARAMETRIC REGRESSION, CONTINUED

1.1 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 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: Let δ_n be the critical radius. Then for any $c \ge 1$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{F}^*} 2c\langle \eta, g\rangle_n - \|g\|_n^2 > 2c^2\delta_n^2 + \frac{2c^2 u}{n}\right) \le \exp\{-u/2\} \tag{1.1}$$

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,

$$\mathbb{P}\left(Z(\delta_n) \ge \mathbb{E}Z(\delta_n) + t\delta_n\right) \le \exp\left\{-\frac{nt^2}{2}\right\}. \tag{1.2}$$

We now condition on the complement of the above event. Take $g \in \mathcal{F}^*$. Consider two cases. First, if $\|g\|_n \le \delta_n$ then

$$2c\langle \eta, g\rangle_n - \|g\|_n^2 \le 2cZ(\delta_n) \le 2c\left(\mathbb{E}Z(\delta_n) + t\delta_n\right) \le 2c\left(\frac{\delta_n^2}{2} + t\delta_n\right) \le c(t+\delta_n)^2 \tag{1.3}$$

Second, if $\|g\|_n \ge \delta_n$, we set $r = \delta_n / \|g\|_n \le 1$. Then

$$2c\langle \eta, g\rangle_n - \|g\|_n^2 = \frac{2c}{r}\langle \eta, \frac{\delta_n}{\|g\|_n} g\rangle - \frac{\delta_n^2}{r^2} \le \frac{2c}{r} Z(\delta_n) - \frac{\delta_n^2}{r^2} = \frac{2\delta_n}{r}\frac{cZ(\delta_n)}{\delta_n} - \frac{\delta_n^2}{r^2}. \tag{1.4}$$

Using $2ab - b^2 \le a^2$, we get a further upper bound of

$$c^2\left(\frac{Z(\delta_n)}{\delta_n}\right)^2 \le c^2\left(\frac{\delta_n^2/2 + t\delta_n}{\delta_n}\right)^2 = c^2(\delta_n/2 + t)^2 \tag{1.5}$$

□

1.1.1 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 averages are unbounded too. However, offset averages are

$$\sup_{w\in\mathbb{R}^d} \sum_{i=1}^n \eta_i\langle w, x_i\rangle - c\langle w, x_i\rangle^2 = \sup_{w\in\mathbb{R}^d} \langle w, \sum_{i=1}^n \eta_i x_i\rangle - c\|w\|_\Sigma^2 \tag{1.6}$$

$$= \frac{1}{4c}\left\|\sum_{i=1}^n \eta_i x_i\right\|_{\Sigma^\dagger}^2 \tag{1.7}$$

where $\Sigma = \sum_{i=1}^n x_i x_i^\mathsf{T}$ and $\Sigma^\dagger$ is the pseudoinverse. Assuming $\mathbb{E}\eta_i^2 \le 1$,

$$\mathbb{E}\left\|\sum_{i=1}^n \eta_i x_i\right\|_{\Sigma^{-1}}^2 \le \sum_{i=1}^n x_i^\mathsf{T}\Sigma^\dagger x_i = \mathrm{tr}(\Sigma\Sigma^\dagger) = \mathrm{rank}(\Sigma)$$

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 $O\left(\frac{\mathrm{rank}(\Sigma)}{n}\right)$ 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.

2. LEAST SQUARES

2.0.1 Nonparametric

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

$$\mathbb{E}\sup_{g\in\mathcal{F}^*:\|g\|_n\le\delta} \langle \eta, g\rangle_n \le \delta^2/2.$$

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 $\alpha \geq 0$,

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \alpha + \frac{1}{\sqrt{n}} \int_{\alpha/4}^{\delta} \sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon)} d\varepsilon$$

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

$$\delta_n = n^{-\frac{1}{2+p}}$$

and thus the rate of the least squares estimator is

$$\mathbb{E} \left\| \widehat{f} - f^* \right\|_n^2 \lesssim n^{-\frac{2}{2+p}}$$

It can be shown that minimax optimal rates of estimation (for any estimator) for fixed design are given by[1] the fixed point

$$\frac{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \delta_*)}{n} \asymp \delta_*^2 \tag{2.8}$$

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 same rate of $\delta_*^2 = n^{-\frac{2}{2+p}}$. Hence, least squares are optimal in this minimax sense for $p \in (0, 2)$.

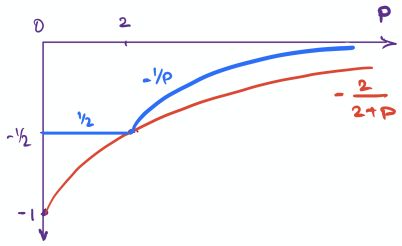


Figure 1: Optimal (in general) rates $n^{-\frac{2}{2+p}}$ (obtained with localization for $p \in (0, 2)$ by ERM) vs without localization (e.g. via global Rademacher averages)

[1] See Yang and Barron "Information-theoretic determination of minimax rates of convergence," 1999

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 $\mathbb{R}^d$:

$$\log \mathcal{N}(\mathcal{F}_{\text{lip}}, L^2(P_n), \varepsilon) \leq (L/\varepsilon)^{d}$$

2.0.2 Parametric

Consider the parametric case,

$$\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2/\varepsilon)$$

Then

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{d \log(1 + 2/\varepsilon)} d\varepsilon \tag{2.9}$$

Change of variables gives an upper bound

$$\sqrt{\frac{d}{n}} \delta \cdot \int_0^1 \sqrt{\log(1 + 2/(u\delta))} du \tag{2.10}$$

Unfortunately, this gives a pesky logarithmic factor that should not be there. However, for some parametric cases one can, in fact, prove that *local covering numbers* behave as

$$\log \mathcal{N}(\mathcal{F}^* \cap \{g : \|g\|_n \leq \delta\}, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2\delta/\varepsilon) \tag{2.11}$$

In this case, the change-of-variables leads to

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \sqrt{\frac{d}{n}} \delta \cdot \int_0^1 \sqrt{\log(1 + 2/\varepsilon)} d\varepsilon \lesssim \sqrt{\frac{d}{n}} \delta \tag{2.12}$$

Equating

$$\delta \sqrt{\frac{d}{n}} \asymp \delta^2$$

yields

$$\delta_n^2 \asymp \frac{d}{n}$$

Note that local covering numbers (2.11) 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).

2.1 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 (2.8) as the lower bound for estimation. See also Mendelson's "local vs global parameters" paper for an in-depth discussion.

3. ORACLE INEQUALITIES

What if we do not assume the regression function f^* is in $\mathcal{F}$? How can we prove an oracle inequality

$$\mathbb{E}\left\|\widehat{f}-f^*\right\|_n^2 - \inf_{f\in\mathcal{F}}\|f-f^*\|_n^2 \le \phi(\mathcal{F},n)$$

Again, we will focus on fixed design.

3.1 Convex $\mathcal{F}$

Suppose $\mathcal{F}$ is convex (or, rather, $\mathcal{F}|_{x_1,\dots,x_n}$ 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 $F=\mathcal{F}_{x_1,\dots,x_n}$, it holds that

$$\left\|\widehat{f}-Y\right\|_n^2 \le \|f-Y\|_n^2 - \left\|\widehat{f}-f\right\|_n^2 \tag{3.13}$$

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,

$$\left\|\widehat{f}-f^*\right\|_n^2 + \|f^*-Y\|_n^2 + 2\langle \widehat{f}-f^*, -\eta\rangle_n + \left\|f_{\mathcal{F}}-\widehat{f}\right\|_n^2 \le \|f_{\mathcal{F}}-f^*\|_n^2 + \|f^*-Y\|_n^2 + 2\langle f_{\mathcal{F}}-f^*, -\eta\rangle_n$$

which leads to

$$\left\|\widehat{f} - f^*\right\|_n^2 - \|f_{\mathcal{F}} - f^*\|_n^2 \le 2\langle \eta, \widehat{f} - f_{\mathcal{F}} \rangle_n - \left\|\widehat{f} - f_{\mathcal{F}}\right\|_n^2 \tag{3.14}$$

$$\le \sup_{h \in \mathcal{F} - f_{\mathcal{F}}} 2\langle \eta, h \rangle_n - \|h\|_n^2 \tag{3.15}$$

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 lower bounds for the well-specified case and polynomial entropy growth 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 (3.13) had a different constant multiplier in front of $-\left\|f - \widehat{f}\right\|_n^2$. We will exploit this observation in a bit.

3.2 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!

3.2.1 A lower bound for ERM

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.[2]

3.2.2 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.[3]

[2] For more detailed discussion, we refer to "The importance of convexity in learning with squared loss" by Lee, Bartlett, Williamson, 1996.

[3] Proof can be found in Lecué & Mendelson

3.2.3 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[4] (*Star Estimator*):

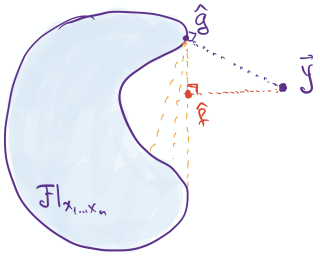
$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \|f - Y\|_n^2 \tag{3.16}$$

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathrm{star}(\mathcal{F}, \widehat{g})} \|f - Y\|_n^2 \tag{3.17}$$

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 (3.13) holds.[5]

Lemma: For any $f \in \mathcal{F}$,

$$\|f - Y\|_n^2 - \left\|\widehat{f} - Y\right\|_n^2 \geq \frac{1}{18} \left\|\widehat{f} - f\right\|_n^2. \tag{3.18}$$

The above inequality is an approximate version of (3.13), 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 $\mathrm{star}(\mathcal{F}, \widehat{f}) \subseteq \mathcal{F} - f^* + \mathrm{star}(\mathcal{F} - \mathcal{F})$ 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, $\widehat{f} = \widehat{g}$.

[4] For a finite class, the above estimator was analyzed by J-Y. Audibert in 2007 in "Progressive mixture rules are deviation suboptimal".

[5] See Liang-R-Sridharan 2015 for a proof.

2. the Star Estimator can be viewed as one step of Frank-Wolfe. More steps can improve the constant.

Exercise: for any $\varepsilon > 0$ and a set $F \subset \mathbb{R}^n$, the covering numbers satisfy

$$\log \mathcal{N}(F, \|\cdot\|, 2\varepsilon) \le \log \mathcal{N}(\mathrm{star}(F), \|\cdot\|, 2\varepsilon) \le \log(\mathrm{diam}(F)/\epsilon) + \log \mathcal{N}(F, \|\cdot\|, \varepsilon)$$

3.3 Offset Rademacher averages

For a set $V \subset \mathbb{R}^n$, the offset process indexed by V is defined as a stochastic process

$$v \mapsto \sum_{i=1}^n \epsilon_i v_i - c v_i^2 = \langle \epsilon, v \rangle - c \|v\|^2 .$$

Here ϵ_i are independent Rademacher, but the same results hold for any subGaussian random variables.

Lemma: Let $V \subset \mathbb{R}^n$ be a finite set of vectors, $\mathrm{card}(V) = N$. Then for any $c > 0$,

$$\mathbb{E}_\epsilon \max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 \le \frac{\log N}{2c}.$$

Furthermore,

$$\mathbb{P}\left(\max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 \ge \frac{1}{2c}(\log N + \log(1/\delta)) \right) \le \delta$$

Proof. Assuming the random variables are 1-subGaussian,

$$\begin{aligned}
\mathbb{E} \max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 &= \frac{1}{\lambda} \mathbb{E} \log \exp \max \langle \epsilon, v \rangle - c \|v\|^2 \\
&\le \frac{1}{\lambda} \log \sum_{v \in V} \mathbb{E} \exp\{\lambda \langle \epsilon, v \rangle - \lambda c \|v\|^2\} \\
&\le \frac{1}{\lambda} \log \left(N \exp\{\lambda^2 \|v\|^2 / 2 - \lambda c \|v\|^2\} \right) \\
&= \frac{1}{2c} \log N
\end{aligned}$$

where we chose $\lambda = 2c$. □

Theorem: 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,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) - c f(x_i)^2 \tag{3.19}$$

$$\le \inf_{\gamma \ge 0, \alpha \in [0, \gamma]} \left\{ \frac{(2/c) \log \mathcal{N}(\mathcal{F}, L^2(P_n), \gamma)}{n} + 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \delta)} d\delta \right\} \tag{3.20}$$

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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 24 & 25
May 5 & 7, 2020

1. TALAGRAND'S INEQUALITY AND APPLICATIONS

The following version of Talagrand's inequality is due to Bousquet:

Theorem: Let $X_1, \dots, X_n$ be i.i.d., and let $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$. Suppose $\mathbb{E}f(X) = 0$ and let

$$\sup_{f \in \mathcal{F}} \mathbb{E}f^2(X) \le \sigma^2$$

for some $\sigma > 0$. Let

$$Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i), \qquad v = n\sigma^2 + 2\mathbb{E}Z$$

Then for any $t \ge 0$,

$$Z \le \mathbb{E}Z + \sqrt{2tv} + \frac{t}{3}$$

with probability at least $1 - e^{-t}$.

Consider a particular case of a singleton $\mathcal{F} = \{f\}$. Then $Z = \sum_{i=1}^n f(X_i)$, $\sigma^2 = \mathbb{E}f^2$ and $v = n\mathbb{E}f^2$ because $\mathbb{E}Z = \mathbb{E}f = 0$. Then the theorem says that

$$\mathbb{P}\left(\sum_{i=1}^n f(X_i) \ge \sigma\sqrt{2tn} + \frac{t}{3} \right) \le e^{-t}$$

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 $1/n$:

$$\widehat{\mathcal{R}}(\mathcal{F}) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i),$$

conditionally on $X_1, \dots, X_n$ and its expectation

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}\widehat{\mathcal{R}}(\mathcal{F})$$

where the expectation is over the data.

The following holds for Rademacher averages (proof via self-bounding, see [2]):

Theorem: Let $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$. Then

$$\mathbb{P}\left(\widehat{\mathcal{R}}(\mathcal{F}) \geq \mathcal{R}(\mathcal{F}) + \sqrt{\frac{2t\mathcal{R}(\mathcal{F})}{n}} + \frac{t}{3n}\right) \leq e^{-t}$$

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

$$\mathbb{P}\left(\widehat{\mathcal{R}}(\mathcal{F}) \geq 2\mathcal{R}(\mathcal{F}) + \frac{5t}{6n}\right) \leq e^{-t}.$$

This and other deviation inequalities for empirical Rademacher averages around their expected value immediately result in data-dependent measures of complexity whenever one can derive a bound in terms of expected (over data) Rademacher averages. Specifically, Talagrand's inequality can be used to relate the random supremum of the empirical process to its expectation; then symmetrization can relate the expected supremum of the empirical process to the expected supremum of the Rademacher process; then above theorem can be employed to relate the latter to the random data-dependent Rademacher averages.

For this lecture, we will note that above theorems are at the heart of proving localization results for random design, both in the well-specified and misspecified settings. We will not flesh out all the details and instead refer to [1]. In particular, in the remainder of this lecture, we would like to develop tools for comparing random and population norms. This will allow us to go from fixed to random design. The tools are also useful more generally.

2. 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

$$\left\|\widehat{f} - f^*\right\|^2_{L^2(P_n)} \leq \ldots$$

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,

$$\mathbb{E}\left\|\widehat{f} - f^*\right\|^2_{L^2(P_n)} \neq \mathbb{E}\left\|\widehat{f} - f^*\right\|^2_{L^2(P)}$$

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 $f \in \mathcal{F}$,

$$\|f - f^*\|^2_{L^2(P)} \leq 2\|f - f^*\|^2_{L^2(P_n)} + \psi(n, \mathcal{F}). \tag{2.1}$$

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.

Our plan of action for proving results of the form (2.1) 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 uniformly bounded and *nonnegative* functions.

Let $\hat{\delta}$ satisfy

$$\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}:\frac{1}{n}\sum_{i=1}^{n} g(X_i)\leq\delta^2} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i) \leq \delta^2/2 \tag{2.2}$$

conditionally on $X_1, \ldots, X_n$. Then the following result can be proved from the theorems in the previous section (see e.g. [3]):

Lemma: Let $\mathcal{G}$ be a class of functions with values in $[0, 1]$. Then with probability at least $1 - e^{-t}$ for all $g \in \mathcal{G}$

$$\mathbb{E}g(X) \leq \frac{2}{n}\sum_{i=1}^{n} g(X_i) + c \cdot \hat{\delta}^2 + \frac{c' \cdot (t + \log\log n)}{n} \tag{2.3}$$

where $\hat{\delta} = \hat{\delta}(\mathcal{G})$ is any upper bound on the fixed point in (2.2).

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

$$\left\|f - f'\right\|^2_{L^2(P)} \leq 2\left\|f - f'\right\|^2_{L^2(P_n)} + c \cdot \hat{\delta}^2 + \frac{c' \cdot (t + \log\log n)}{n}. \tag{2.4}$$

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 (2.2) 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 (2.2) is smaller. Thus the fixed point (2.2) is potentially smaller than the one defined in the previous lecture.

Now, one can ask how to compute a suitable upper bound on the critical radius in (2.2) for particular classes of interest. As in the earlier lectures, the strategy is to upper bound the left-hand side of (2.2) 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

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{F}^2 \cap \widehat{\mathcal{M}}} \frac{1}{n} \sum_{i=1}^n \epsilon_i g(X_i) \leq 2\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}: \|f\|_n^2 \leq \delta^2} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \tag{2.5}$$

since square is 2-Lipschitz on $[-1, 1]$. Balancing this with δ^2 gives, up to constants, the critical radius of $\mathcal{F}$ as defined in previous lectures. Interestingly, one can significantly improve upon this argument and show that the localization radius for $\mathcal{F}^2$ can be smaller than that of $\mathcal{F}$. In particular, a useful result is the following:

Lemma: For any class $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$ of bounded functions, the critical radius in (2.2) for the class $\mathcal{G} = \mathcal{F}^2$ can be upper bounded by a solution to

$$\frac{12}{\sqrt{n}} \int_{\delta/16}^1 \sqrt{\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), u/2)} du \leq \delta/4. \tag{2.6}$$

Proof. We start upper bounding the left-hand side of (2.2), aiming to get an upper bound proportional to the scale δ. 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 (2.2) is upper bounded by

$$\inf_\alpha \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\delta \sqrt{\log \mathcal{N}(\mathcal{G} \cap \widehat{\mathcal{M}}, L^2(P_n), \varepsilon)} d\varepsilon \right\}. \tag{2.7}$$

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

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n (f(x_i)^2 - \tilde{f}(x_i)^2)^2 &= \frac{1}{n} \sum_{i=1}^n (f(x_i) - \tilde{f}(x_i))^2 (f(x_i) + \tilde{f}(x_i))^2 \\
&\leq \max_i (f(x_i) - \tilde{f}(x_i))^2 \cdot \frac{1}{n} \sum_{i=1}^n (f(x_i) + \tilde{f}(x_i))^2 \\
&\leq \tau^2 (2 \|f\|_n^2 + 2 \left\|\tilde{f}\right\|_n^2) \\
&\leq 4\tau^2 \delta^2 := \varepsilon^2
\end{aligned}$$

We conclude that

$$\begin{aligned}
\mathcal{N}(\mathcal{G} \cap \widehat{\mathcal{M}}, L^2(P_n), \varepsilon) &\leq \mathcal{N}(\mathcal{F} \cap \{\|f\|_n \leq \delta\}, L^\infty(P_n), \varepsilon/(2\delta)) \\
&\leq \mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon/(2\delta))
\end{aligned}$$

Substituting into (2.7), the upper bound on the right-hand side becomes

$$\begin{aligned}
&\inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\delta \sqrt{\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon/(2\delta))} d\varepsilon \right\} \\
&\leq \delta^2/4 + \delta \times \frac{12}{\sqrt{n}} \int_{\delta/16}^1 \sqrt{\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), u/2)} du
\end{aligned}$$

where we performed change-of-variables $u = \varepsilon/\delta$ and chose $\alpha = \delta^2/16$. Using this in (2.2) and balancing with $\delta^2/2$ yields (2.6). □

A key outcome of the above lemma is that the critical radius of $\mathcal{F}^2$ (or $(\mathcal{F}-\mathcal{F})^2$) is much smaller than that of $\mathcal{F}$. The latter would have δ^2 rather than δ on the right-hand side of (2.6). In particular, if the left-hand side of (2.6) is of order $1/\sqrt{n}$, the solution is $\delta \propto 1/\sqrt{n}$ and hence the remainder in (2.4) is of the order $1/n$, a smaller order term as compared to the rate of estimation for fixed design. For instance, for a class that exhibits polynomial growth of entropy

$$\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \leq \left(\frac{cn}{\varepsilon}\right)^d,$$

the localization radius of $\mathcal{G}$ can be upper bounded as

$$\hat{\delta}(\mathcal{G}) = C\sqrt{\frac{d}{n}\log\left(\frac{cn}{d}\right)}$$

and for a finite class we immediately have

$$\hat{\delta}(\mathcal{G}) \leq C\sqrt{\frac{\log|\mathcal{F}|}{n}}.$$

We can also prove a general and useful result, albeit with extra log factors (due to its generality). Following [8], we have

Lemma: For any class $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$, the critical radius in (2.6) is at most

$$C\log^2 n \cdot \bar{\mathcal{R}}(\mathcal{F}),$$

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. [7]):

$$\log\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \leq 2\mathrm{vc}(\mathcal{F}, c\varepsilon)\cdot\log n\cdot\left(\frac{cn}{\mathrm{vc}(\mathcal{F}, c\varepsilon)\cdot\varepsilon}\right) \tag{2.8}$$

and then use the following fact: for any $\varepsilon > \bar{\mathcal{R}}(\mathcal{F})$,

$$\mathrm{vc}(\mathcal{F}, \varepsilon) \leq \frac{4n\bar{\mathcal{R}}(\mathcal{F})^2}{\varepsilon^2}. \tag{2.9}$$

This last inequality can be written in the more familiar form

$$\sup_{\varepsilon > \bar{\mathcal{R}}(\mathcal{F})} \varepsilon\sqrt{\frac{\mathrm{vc}(\mathcal{F},\varepsilon)}{4n}} \leq \bar{\mathcal{R}}(\mathcal{F}), \tag{2.10}$$

which bears similarity to Sudakov's minoration. This inequality is proved by taking the ε-shattered set, replicating it $\lceil n/\mathrm{vc}(\mathcal{F},\varepsilon)\rceil$ 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

$$\frac{1}{\sqrt{n}}\int_{\delta/64}^{1/4}\sqrt{\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon))}d\varepsilon \lesssim \frac{\sqrt{\log n}}{\sqrt{n}}\int_{\delta/64}^{1/4}\sqrt{\mathrm{vc}(\mathcal{F}, c\varepsilon)\log\Big(\frac{cn}{\varepsilon}\Big)}d\varepsilon \tag{2.11}$$

$$\lesssim \sqrt{\log n}\bar{\mathcal{R}}(\mathcal{F})\int_{\delta/64}^{1/4}\frac{1}{\varepsilon}\sqrt{\log\Big(\frac{cn}{\varepsilon}\Big)}d\varepsilon \tag{2.12}$$

To finish the proof, choose $\delta = 64\bar{\mathcal{R}}(\mathcal{F})$ and observe that

$$\int_{\bar{\mathcal{R}}(\mathcal{F})}^{1}\frac{1}{\varepsilon}\sqrt{\log\Big(\frac{cn}{\varepsilon}\Big)}d\varepsilon \lesssim \log^2(cn/\bar{\mathcal{R}}(\mathcal{F})).$$

□

Hence, ignoring logarithmic factors, $\hat{\delta}(\mathcal{G}) \le \tilde{O}(n^{-1})$ when $\bar{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/2}$ and $\hat{\delta}(\mathcal{G}) \le \tilde{O}(n^{-2/p})$ 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 $2\|f - f^*\|^2_{L^2(P_n)}$ is typically of smaller order than what one gets from denoising for fixed design. The next section explains why this happens.

3. BEYOND BOUNDEDNESS: THE SMALL-BALL METHOD

This approach was pioneered by [4] and then developed by Mendelson in a series of papers starting with [6].

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 definitely do not require boundedness.

We first recall the Paley-Zygmund inequality (1932) stating that for a nonnegative random variable Z with finite variance,

$$\mathbb{P}\left(Z \ge t\mathbb{E}Z\right) \ge (1-t)^2\frac{(\mathbb{E}Z)^2}{\mathbb{E}Z^2}$$

for any $0 \le t \le 1$.

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

$$\mathbb{P}\left(|f(X)| \ge t\|f\|_2\right) = \mathbb{P}\left(f(X)^2 \ge t^2\|f\|_2^2\right) \ge (1-t^2)^2\frac{\|f\|_2^4}{\|f\|_4^4}$$

Now, we make an assumption that for every $f \in \mathcal{F}$,

$$\mathbb{E}f(X)^4 \le c(\mathbb{E}f(X)^2)^2$$

for some c.

Under this $L^4 - L^2$ norm comparison, it holds that

$$\mathbb{P}\left(|f(X)| \geq t \|f\|_2\right) \geq (1-t^2)^2 c$$

More generally, the condition

$$\mathbb{P}\left(|f(X)| \geq c \|f\|_2\right) \geq c' \tag{3.13}$$

for some c, c' is called the small-ball property.

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

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{|f(X_i)| \geq c_1\right\} \geq c_2 \tag{3.14}$$

for some constants c_1, c_2, we would be done since such a lower bound implies a constant lower bound on $\frac{1}{n}\sum_{i=1}^n f(X_i)^2 \geq c_3 \|f\|_2 = c_3$). 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 (3.14)).

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 (3.14) 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. The key observation is that the critical radius β^* can be defined as the smallest β such that

$$\mathbb{E} \sup_{f \in \mathcal{F}, \|f\|_2 \leq \beta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \leq c''\beta \tag{3.15}$$

Assuming that $\mathcal{F}$ is star-shaped around 0, the control extends for all $\beta \geq \beta^*$.

To summarize, with probability at least e^{-cn},

$$\inf_{f \in \mathcal{F}: \|f\|_2 \geq \beta^*} \frac{\|f\|_n}{\|f\|_2} \geq c'$$

for some constants c, c'. Alternatively, we have with probability at least e^{-cn}, for all $f \in \mathcal{F}$,

$$\|f\|_2^2 \leq C \|f\|_n^2 + (\beta^*)^2.$$

Observe that β^* can be significantly smaller than if (3.15) were defined with β^2 on the right-hand side, as before.

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

$$\operatorname{diam}_2(\mathcal{I}_{X_1,\ldots,X_n}) = \sup_{f, f' \in \mathcal{I}_{X_1,\ldots,X_n}} \left\|f - f'\right\|_{L^2(P)}.$$

Of course, from the earlier calculations, we have that with high probability

$$\left\|f - f'\right\|_{L^2(P)} \lesssim \hat{\delta}^2$$

where $\hat{\delta}$ is the localization radius for $(\mathcal{F} - \mathcal{F})^2$ and can be upper bounded by $\sup_{x_{1:n}} \widehat{\mathcal{R}}(\mathcal{F})^2$. Alternatively, we can use the fixed point $(\beta^*)^2$ under the small ball property.

5. 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

$$(1-\varepsilon)^2 \|v\|_2^2 \le \|\Gamma v\|_2^2 \le (1+\varepsilon)^2 \|v\|_2^2 .$$

Of particular interest in applications is the lower side of this inequality:

$$\frac{\|\Gamma v\|_2^2}{\|v\|_2^2} \ge 1-\alpha$$

where $\alpha \in (0,1)$. A corresponding *uniform* statement over a set $V \subset \mathbb{R}^d$ asks that with high probability,

$$\inf_{v\in V} \frac{\|\Gamma v\|_2^2}{\|v\|_2^2} \ge 1-\alpha.$$

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

$$\Gamma = \frac{1}{\sqrt{n}} \begin{pmatrix} -X_1- \\ \dots \\ -X_n- \end{pmatrix}$$

with $X_1, \dots, X_n$ i.i.d. from an isotropic distribution. Then

$$\|\Gamma v\|_2^2 = \frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle^2$$

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.

6. LARGE MARGIN THEORY

We end this lecture with a result from large margin classification, because its proof utilizes the same technique (not surprisingly, the authors of [5] and [4] have a nonzero intersection).

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 $\phi : \mathbb{R} \to \mathbb{R}$ be defined by $\phi(a) = 0$ on $(-\infty, 0]$, $\phi(a) = a/\gamma$ on $[0,\gamma]$, and $\phi(a) = 1$ on $[\gamma, \infty)$. Then with probability at least $1 - e^{-2u^2}$, for any $f \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}\mathbf{1}\{Yf(X) \ge 0\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Y_i f(X_i) \ge \gamma\} &\le \sup_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)) - \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i)) \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)) - \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i)) + \frac{u}{\sqrt{n}}\end{aligned}$$

since ϕ is in $[0,1]$. By symmetrization, the above expectation is at most

$$2\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi(Y_if(X_i)) \le \frac{2}{\gamma}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_iY_if(X_i) = \frac{2}{\gamma}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_if(X_i) \le \frac{2}{\gamma}\mathcal{R}(\mathcal{F})$$

Hence, with probability at least $1-e^{-2u^2}$, for any $f\in\mathcal{F}$,

$$\mathbb{E}\mathbf{1}\left\{Yf(X)\ge 0\right\} \le \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{Y_if(X_i)\ge\gamma\right\} + \frac{2}{\gamma}\mathcal{R}(\mathcal{F}) + \frac{u}{\sqrt{n}}$$

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 $\mathrm{sign}(\mathcal{F})$ 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.

References

[1] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[2] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[3] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Biologische Kybernetik, 2002.

[4] 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.

[5] 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.

[6] S. Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[7] M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

[8] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 26
May 12, 2020

1. TIME SERIES

Suppose we observe a sequence

$$\boldsymbol{x}_{t+1} = f^*(\boldsymbol{x}_t) + \eta_t, \quad t = 1, \dots, n$$

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 $\mathcal{F}$. Let us treat this problem as a fixed-design regression problem, except that the outcomes are now vectors rather than reals, and the sequence $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ 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):

$$\left\| \widehat{f} - f^* \right\|_n^2 \le 2 \frac{1}{n} \sum_{t=1}^{n} \langle \eta_t, \widehat{f}(\boldsymbol{x}_t) - f^*(\boldsymbol{x}_t) \rangle .$$

Choosing the offset-style approach covered in previous lectures, we have

$$\left\| \widehat{f} - f^* \right\|_n^2 \le \sup_{g \in \mathcal{F} - f^*} \frac{1}{n} \sum_{t=1}^{n} 4\langle \eta_t, g(\boldsymbol{x}_t) \rangle - \|g(\boldsymbol{x}_t)\|^2 .$$

Up until now, the statement is conditional on $\{\eta_1, \dots, \eta_n\}$. 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 $\sigma(\eta_1, \dots, \eta_{t-1})$. 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 $\sigma(\epsilon_1, \dots, \epsilon_{t-1})$-measurable, $\mathcal{F}$ is a class of real-valued functions $\mathcal{X} \to \mathbb{R}$, and $\epsilon_1, \dots, \epsilon_n$ are i.i.d. Rademacher random variables.

2. 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 [3], and we refer also to [2] for an introduction.

Let us elaborate on the first point. Note that $\boldsymbol{x}_t$ being measurable with respect to $\sigma(\epsilon_1, \dots, \epsilon_{t-1})$ simply means $\boldsymbol{x}_t$ is a function of $\epsilon_1, \dots, \epsilon_{t-1}$ (in other words, it's a predictable process). Note that the collection $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ 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 $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ 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 $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ and call it simply as an $\mathcal{X}$-valued *tree*. We shall refer to $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ 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

$$\widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}, \boldsymbol{x}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t(\epsilon_1, \dots, \epsilon_{t-1}))$$

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 $x_1, \dots, x_n$ are constant values, we see that it is a special case of a tree with constant levels $\boldsymbol{x}_t(\epsilon_1, \dots, \epsilon_{t-1}) = x_t$. 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

$$\mathbb{E} \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t) \leq \sqrt{\frac{2 \log \mathrm{card}(\mathcal{F})}{n}}$$

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 $\mathcal{F}|_{x_1,\dots,x_n}$ for the tree case? Spoiler: it's not $\mathcal{F}|_{\boldsymbol{x}}$. The following turns out to be the right definition:

Definition: A set V of $\mathbb{R}$-valued trees is an 0-cover of $\mathcal{F}$ on a tree $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ if

$$\forall f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad \text{s.t.} \quad f(\boldsymbol{x}_t(\epsilon_{1:t-1})) = \boldsymbol{v}_t(\epsilon_{1:t-1}) \quad \forall t \in [n]$$

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, 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 $\mathcal{F}|_{x_1,\dots,x_n}$ in the binary-valued case.

For real-valued functions, consider the following definition.

Definition: A set V of $\mathbb{R}$-valued trees is an α-cover of $\mathcal{F}$ on a tree $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ with respect to ℓ_2 if

$$\forall f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad \text{s.t.} \quad \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_t(\epsilon_{1:t-1})) - \boldsymbol{v}_t(\epsilon_{1:t-1}))^2 \leq \alpha^2$$

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: For any class of $[-1, 1]$-valued functions $\mathcal{F}$,

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}, \boldsymbol{x}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \boldsymbol{x}, \varepsilon)} d\varepsilon \right\}$$

Recall the definition of VC dimension and a shattered set. Here is the right sequential analogue:

Definition: Function class $\mathcal{F}$ of $\{\pm 1\}$-valued functions shatters a tree $\boldsymbol{x}$ of depth d if

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists f \in \mathcal{F}, \quad \text{s.t.} \quad \forall t \in [d], \quad f(\boldsymbol{x}_t(\epsilon)) = \epsilon_t$$

The largest depth d for which there exists a shattered $\mathcal{X}$-valued tree is called the *Littlestone dimension* and denoted by $\mathrm{ldim}(\mathcal{F})$.

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 $\widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}, \boldsymbol{x}) = 1$. It is also easy to see that $\mathrm{vc}(\mathcal{F}) \le \mathrm{ldim}(\mathcal{F})$, and the gap can be infinite.

The following is an analogue of the Sauer-Shelah-Vapnik-Chervonenkis lemma.

Theorem: For a class of binary-valued functions $\mathcal{F}$ with Littlestone dimension $\mathrm{ldim}(\mathcal{F})$,

$$\mathcal{N}(\mathcal{F}, \boldsymbol{x}, 0) \le \sum_{i=0}^{d} \binom{n}{i} \le \left(\frac{en}{d}\right)^d$$

Scale-sensitive sequential versions are defined as follows:

Definition: 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

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F}, \quad \text{s.t.} \quad \forall t \in [d], \quad \epsilon_t (f(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{s}_t(\epsilon)) \ge \alpha/2$$

The largest depth d for which there exists an α-shattered $\mathcal{X}$-valued tree is called sequential scale-sensitive dimension and denoted $\mathrm{ldim}(\mathcal{F}, \alpha)$.

We note that the above definitions reduce to the classical ones if we consider only trees $\boldsymbol{x}$ with constant levels.

Theorem: For any class of $[-1, 1]$-valued functions $\mathcal{F}$ and $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n

$$\mathcal{N}_\infty(\mathcal{F}, \boldsymbol{x}, \alpha) \le \left(\frac{2en}{\alpha}\right)^{\mathrm{ldim}(\mathcal{F}, \alpha)}$$

Finally, it is possible to show an analogue of symmetrization lemma: for any joint distribution of $(X_1, \dots, X_n)$,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[f(X_t) | X_{1:t-1}] - f(X_t) \le 2 \sup_{\boldsymbol{x}} \widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}, \boldsymbol{x})$$

If the sequence $(X_1, \dots, X_n)$ 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 [1].

3. ONLINE LEARNING

Consider the following online classification problem. On each of n rounds $t = 1, \dots, n$, 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*

$$\mathrm{Reg}(\mathcal{F}) = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\left\{\widehat{y}_t \neq y_t\right\}\right] - \inf_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\left\{f(x_t) \neq y_t\right\}\right] \tag{3.1}$$

small for any sequence $(x_1, y_1), \dots, (x_n, y_n)$. 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 (3.1) 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 $(x_1, y_1), \dots, (x_n, y_n)$ is fixed in advanced (this turns out not to matter).

We will answer the following question: what is the best achievable $\mathrm{Reg}(\mathcal{F})$ 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

$$2\mathrm{Reg}(\mathcal{F}) = \frac{1}{n}\sum_{t=1}^{n} -q_t y_t - \inf_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{t=1}^{n} -y_t f(x_t)\right] \tag{3.2}$$

$$= \sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{t=1}^{n} y_t f(x_t)\right] - \frac{1}{n}\sum_{t=1}^{n} q_t y_t \tag{3.3}$$

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 $x_t = \boldsymbol{x}_t(y_1, \dots, y_{t-1})$ and $y_t = \epsilon_t$ be an independent Rademacher random variable. This defines a stochastic process with 2^n possible sequences $(x_1, y_1), \dots, (x_n, y_n)$. Now, clearly

$$\sup_{(x_1,y_1),\dots,(x_n,y_n)} 2\mathrm{Reg}(\mathcal{F}) \geq 2\mathbb{E}_\epsilon \mathrm{Reg}(\mathcal{F}).$$

Observe that $q_t = q_t(\epsilon_1, \dots, \epsilon_{t-1})$ and thus

$$\mathbb{E}_\epsilon\left[\frac{1}{n}\sum_{t=1}^{n} q_t \epsilon_t\right] = 0.$$

Hence,

$$\mathbb{E}_\epsilon \mathrm{Reg}(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t) \right]. \tag{3.4}$$

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

$$\mathcal{V} = \min_{\mathrm{Algo}} \max_{\{(x_t, y_t)\}_{t=1}^n} \mathrm{Reg}(\mathcal{F})$$

Theorem: For a binary-valued class $\mathcal{F}$,

$$\frac{1}{2} \bar{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}) \leq \mathcal{V} \leq \bar{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F})$$

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] A. Rakhlin and K. Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

[2] 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.

[3] 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.