Spring 2018 Statistics 210b (Theoretical Statistics) - All Lectures

Aditya Guntuboyina

16 January, 2018

Contents

1 Lecture 1

The first topic of the class will be *Empirical Process Theory*. I will give a high-level overview of what we plan to cover in the empirical processes part of the class.

1.1 Some Aspects of Empirical Process Theory

Empirical process theory usually deals with two fundamental questions.

1.1.1 Uniform Laws of Large Numbers

The first question concerns uniform strong laws of large numbers. Suppose $X_1, X_2, \ldots$ are independent and identically distributed random objects taking values in a set $\mathcal{X}$. Let $\mathcal{F}$ denote a class of real-valued functions on $\mathcal{X}$. What can one say about the random variable:

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X_1) \right| \tag{1}$$

Specifically,

1. Does the random variable (5) concentrate around its expectation?
2. Can one provide *finite-sample* (i.e., bounds that hold for every n) bounds for (5) in terms of the class of functions $\mathcal{F}$ and the common distribution P of $X_1, X_2, \ldots$?
3. Can one provide conditions on $\mathcal{F}$ such that (5) converges to zero in probability or almost surely (if this is true, we say that the uniform strong law of large numbers holds)?

Empirical process theory provides answers to these questions. Why are these questions relevant to theoretical statistics? The two examples that we shall study in detail are given below.

Example 1.1 (Classification)**.** *Consider a pair of random objects X and Y having some joint distribution where X takes values in a space $\mathcal{X}$ and Y takes only the two values: -1 or $+1$. A classifier is a function $g : \mathcal{X} \to \{-1, +1\}$. The error of the classifier is given by*

$$L(g) := \mathbb{P}\{g(X) \neq Y\}.$$

The goal of classification is to construct a classifier with small error based on n i.i.d observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ having the same distribution as (X, Y).

For a classifier g, its empirical error (i.e., its error on the observed sample) is given by

$$L_n(g) := \frac{1}{n} \sum_{i=1}^{n} I\{g(X_i) \neq Y_i\}.$$

A natural strategy for classification is to select a class of classifiers $\mathcal{C}$ and then to choose the classifier in $\mathcal{C}$ which has the smallest empirical error on the observed sample i.e.,

$$\hat{g}_n := \operatorname*{argmin}_{g \in \mathcal{C}} L_n(g).$$

How good a classifier is $\hat{g}_n$ i.e., how small is its error:

$$L(\hat{g}_n) := \mathbb{P}\left\{\hat{g}_n(X) \neq Y \middle| X_1, Y_1, \ldots, X_n, Y_n\right\}.$$

Two questions are relevant about $L(\hat{g}_n)$:

1. *Is $L(\hat{g}_n)$ comparable to $\inf_{g \in \mathcal{C}} L(g)$? i.e., is the error of $\hat{g}_n$ comparable to the best achievable error in the class $\mathcal{C}$?*
2. *is $L(\hat{g}_n)$ comparable to $L_n(\hat{g}_n)$? i.e., is the error of $\hat{g}_n$ comparable its "in-sample" empirical error?*

It is quite easy to relate these two questions to the size of $\sup_{g \in \mathcal{C}} |L_n(g) - L(g)|$. Indeed, if $g^* := \operatorname{argmin}_{g \in \mathcal{C}} L(g)$, then

$$\begin{aligned} L(\hat{g}_n) &= L(g^*) + L(\hat{g}_n) - L_n(\hat{g}_n) + L_n(\hat{g}_n) - L(g^*) \\ &\leq L(g^*) + L(\hat{g}_n) - L_n(\hat{g}_n) + L_n(g^*) - L(g^*) \leq L(g^*) + 2 \sup_{g \in \mathcal{C}} |L_n(g) - L(g)|. \end{aligned}$$

Also

$$L(\hat{g}_n) \leq L_n(\hat{g}_n) + L(\hat{g}_n) - L_n(\hat{g}_n) \leq L_n(\hat{g}_n) + \sup_{g \in \mathcal{C}} |L_n(g) - L(g)|.$$

Thus the key quantity to answering the above questions is

$$\sup_{g \in \mathcal{C}} |L_n(g) - L(g)|.$$

It is now easy to see that the above quantity is a special case of (5) when $\mathcal{F}$ is taken to be the class of all functions $I\{g(x) \neq y\}$ as g varies over $\mathcal{C}$. Also the X_is in (5) need to be replaced by (X_i, Y_i).

Sometimes, the two inequalities above can sometimes be quite loose. Later, we shall see more sharper inequalities which utilize a technique known as "localization".

Example 1.2 (Consistency and Rates of convergence of M-estimators)**.** *Many problems in statistics are concerned with estimators of the form*

$$\hat{\theta}_n := \operatorname*{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i) \tag{2}$$

for i.i.d observations $X_1, \ldots, X_n$ taking values in a space $\mathcal{X}$. Here Θ denotes the parameter space and, for each $\theta \in \Theta$, m_θ denotes a real-valued function (known as a loss or criterion function) on $\mathcal{X}$. Such an estimator $\hat{\theta}$ is called an M-estimator as it is obtained by maximizing an objective function. The most standard examples of M-estimators are:

1. ***Maximum Likelihood Estimators**: These correspond to $m_\theta(x) := \log p_\theta(x)$ for a class of densities $\{p_\theta, \theta \in \Theta\}$ on $\mathcal{X}$.*

2. ***Location Estimators**:*

   (a) ***Mean**: corresponds to $m_\theta(x) := (x-\theta)^2$.*

   (b) ***Median**: corresponds to $m_\theta(x) := |x-\theta|$.*

   (c) ***Mode**: may correspond to $m_\theta(x) := I\{|x-\theta| \leq 1\}$.*

The target quantity for the estimator $\hat{\theta}_n$ is

$$\theta_0 := \operatorname*{argmax}_{\theta \in \Theta} \mathbb{E} m_\theta(X_1).$$

The main question of interest while studying M-estimators concerns the accuracy of $\hat{\theta}_n$ for estimating θ_0. In the asymptotic framework ($n \to \infty$), the two key questions are:

1. *Is $\hat{\theta}_n$ consistent for estimating θ_0 i.e., does $d(\hat{\theta}_n, \theta_0)$ converge to zero almost surely or in probability as $n \to \infty$? Here $d(\cdot,\cdot)$ is a metric on Θ (for example, the usual Euclidean metric when Θ is a subset of $\mathbb{R}^k$ for some k).*

2. *What is the rate of convergence of $d(\hat{\theta}_n, \theta_0)$? For example, is it $O_p(n^{-1/2})$? or $O_p(n^{-1/3})$?.*

To answer these questions, it is obvious that one must investigate the closeness of $\sum_{i=1}^n m_\theta(X_i)/n$ to $\mathbb{E} m_\theta(X_1)$ in some sort of uniform sense over θ which leads to investigation of (5) for appropriate subclasses $\mathcal{F}$ of $\{m_\theta, \theta \in \Theta\}$.

Indeed, the standard argument involves first writing

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) \geq \epsilon\right\} \leq \mathbb{P}\left\{\sup_{\theta \in \Theta : d(\theta,\theta_0) \geq \epsilon} (M_n(\theta) - M_n(\theta_0)) \geq 0\right\} \tag{3}$$

where

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^n m_\theta(X_i) \quad \text{and } M(\theta) := \mathbb{E} m_\theta(X_1).$$

One then bounds the right hand side of (3) by

$$\mathbb{P}\left\{\sup_{\theta \in \Theta : d(\theta,\theta_0) \geq \epsilon} [(M_n(\theta) - M(\theta)) - (M_n(\theta_0) - M(\theta_0))] \geq -\sup_{\theta \in \Theta : d(\theta,\theta_0) \geq \epsilon} (M(\theta) - M(\theta_0))\right\}.$$

Empirical process results provide bounds for the above probability (some assumptions on the relation between M and the metric d will be needed).

Note that one can further bound the above probability by replacing the left hand side by

$$2\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|$$

but this can sometimes be too loose.

The strategy for controlling (5) is as follows (we will mostly focus on the case when $\mathcal{F}$ is a uniformly bounded class of functions):

1. The key observation is that the random variable (5) "concentrates" around its mean (or expectation).

2. Because of concentration, it is enough to control the mean of (5). The mean will be bounded by a quantity called "Rademacher Complexity" of $\mathcal{F}$ via a technique called "symmetrization".

3. The Rademacher complexity involves the expected supremum over a "sub-Gaussian process". This is further controlled via a technique known as "chaining". In the process, we shall also encounter a quantity known as the "Vapnik-Chervonenkis dimension".

The best reference for these topics is the book Boucheron et al. [3]. The viewpoint that we shall take is the nonasymptotic viewpoint where bounds are proved which hold for every n. The more classical viewpoint is the asymptotic one where statements are made that hold as $n \to \infty$. In the asymptotic viewpoint, it is said that the class $\mathcal{F}$ is "Glivenko-Cantelli" provided (5) converges almost surely as $n \to \infty$. Using our nonasymptotic bounds, it will be possible to put appropriate conditions on $\mathcal{F}$ under which $\mathcal{F}$ becomes a Glivenko-Cantelli class.

2 Lecture 2

2.1 Uniform Central Limit Theorems

Let us now describe the second fundamental question that is addressed by the theory of empirical process.

By the usual Central Limit Theorem (CLT), we have that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right)$$

converges in distribution to the normal distribution with mean zero and variance $Var(f(X_1))$ as $n \to \infty$. This statement is true for every $f \in \mathcal{F}$. Does this convergence hold uniformly over f in the class $\mathcal{F}$ in a reasonable sense? To illustrate this, let us look at the following example.

Example 2.1. *Suppose that $X_1, \ldots, X_n$ are i.i.d observations from the uniform distribution on $[0,1]$. Also suppose that $\mathcal{F}$ consists of all indicator functions $\{I_{(-\infty,t]} : t \in \mathbb{R}\}$. In this case, for $f = I_{-\infty,t]}$, the quantity*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) = F_n(t)$$

where F_n is the empirical distribution function of the observations $X_1, \ldots, X_n$. Define

$$U_n(t) := \sqrt{n}(F_n(t) - t) \qquad \text{for } t \in [0,1].$$

$U_n(t)$ represents a collection of random variables as t varies in $[0,1]$. The stochastic process $\{U_n(t), t \in [0,1]\}$ is known as the "Uniform Empirical Process". It is easy to see that every realization of $\{U_n(t), t \in [0,1]\}$, viewed as a function on $[0,1]$, is piecewise linear with jump discontinuities at the n data points $X_1, \ldots, X_n$. Also $U_n(0) = U_n(1) = 0$ for every n.

The CLT states that for each $t \in [0,1]$, the sequence of real random variables $\{U_n(t)\}$ converges in distribution to $N(0, t - t^2)$ as $n \to \infty$. Moreover, the multivariate CLT states that for every fixed $t_1, \ldots, t_k$, the sequence of random vectors $(U_n(t_1), \ldots, U_n(t_k))$ converges in distribution to the multivariate normal distribution with zero means and covariances given by $\min(t_i, t_j) - t_i t_j$.

At this point, let us introduce an object called Brownian Bridge. The Brownian Bridge on $[0,1]$ is a stochastic process $\{U(t), 0 \le t \le 1\}$ that is characterized by the following two requirements:

1. *Every realization is a continuous function on $[0,1]$ with $U(0)$ and $U(1)$ always fixed to be equal to 0.*
2. *For every fixed $t_1, \ldots, t_k$ in $[0,1]$, the random vector $(U(t_1), \ldots, U(t_k))$ has the multivariate normal distribution with zero means and covariances given by $\min(t_i, t_j) - t_i t_j$.*

We therefore see that the "finite dimensional distributions" of the process $\{U_n(t), 0 \leq t \leq 1\}$ converge to the "finite dimensional distributions" of $\{U(t), 0 \leq t \leq 1\}$. It is natural to ask here if one can claim anything beyond finite-dimensional convergence here. Does the entire process $\{U_n(t), 0 \leq t \leq 1\}$ converge to $\{U(t), 0 \leq t \leq 1\}$? This was first conjectured by Doob and rigorously proved by Donsker.

What is the meaning of the statement that the sequence of stochastic processes $\{U_n(t), t \in [0,1]\}$ converges in distribution to $\{U(t), t \in [0,1]\}$? To understand, let us first recall the usual notion of convergence in distribution for sequences of random vectors. We say that a sequence of random vectors $\{Z_n\}$ taking values in $\mathbb{R}^k$ converges in distribution to Z if and only if

$$\mathbb{E}h(Z_n) \to \mathbb{E}h(Z) \qquad \text{as } n \to \infty$$

*for every **bounded continuous real-valued function** $h : \mathbb{R}^k \to \mathbb{R}$.*

One can attempt a direct generalization of this to define convergence of $U_n(\cdot)$ to $U(\cdot)$ as a stochastic process. These processes take values not in $\mathbb{R}^k$ but in the space of all bounded functions on $[0,1]$. Let us denote this space by $\ell^\infty([0,1])$. This space can be metrized by the supremum metric: $\sup_{0 \leq t \leq 1} |g_1(t) - g_2(t)|$. We can then say that G_n converges in distribution to U as a stochastic process provided

$$\mathbb{E}h(U_n) \to \mathbb{E}h(U) \qquad \text{as } n \to \infty \tag{4}$$

*for every bounded and continuous real valued function $h : \ell^\infty[0,1] \to \mathbb{R}$. This definition almost makes sense except for one measure-theoretic issue. It turns out that there exist bounded and continuous real valued functions $h : \ell^\infty[0,1] \to \mathbb{R}$ for which the random variable $h(U_n)$ is not measurable. One therefore replaces the left hand side in (4) by its **outer** expectation $\mathbb{E}^* h(U_n)$ (formally defined later).*

In this sense, Donsker showed that U_n converges in distribution to Brownian Bridge.

Let us now return to the general case. Here we consider the stochastic process:

$$G_n(f) := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1)\right) \qquad \text{for } f \in \mathcal{F}.$$

Under a simple assumption such as $\sup_{f \in \mathcal{F}} |f(x)| < \infty$ for every $x \in \mathcal{X}$, the function $f \mapsto G_n(f)$ belongs to the space $\ell^\infty(\mathcal{F})$. We say then that the uniform central limit theorem holds over $\mathcal{F}$ if the stochastic process $G_n(f), f \in \mathcal{F}$ converges in distribution in $\ell^\infty(\mathcal{F})$ to a process $G(f), f \in \mathcal{F}$ as $n \to \infty$. The limit process $G(f), f \in \mathcal{F}$ will have the property that for every $f_1, \ldots, f_k \in \mathcal{F}$, the random vector $(G(f_1), \ldots, G(f_k))$ will have a multivariate normal distribution having the same covariance as $(G_n(f_1), \ldots, G_n(f_k))$.

We shall characterize convergence in distribution in $\ell^\infty(\mathcal{F})$ and then see some sufficient conditions on $\mathcal{F}$ that ensure that the Uniform CLT holds.

The following are some statistical applications of Uniform CLTs.

Example 2.2 (Classical Motivation: Goodness of Fit Testing)**.** *Suppose one observes i.i.d observations $X_1, \ldots, X_n$ from a distribution (cdf) F and wants to test the null hypothesis $H_0 : F = F_0$ against the alternative hypothesis $H_1 : F \neq F_0$. Here F_0 is a fixed distribution function.*

Kolmogorov recommended testing this hypothesis via the quantity

$$D_n := \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

where F_n is the empirical cdf of the data $X_1, \ldots, X_n$. The idea is to reject H_0 when D_n is large. To calculate the p-value of this test, the null distribution (i.e., the distribution of D_n under H_0) needs to be determined. An interesting property of the null distribution of D_n is that the null distribution does not depend on F_0 as long as F_0 is continuous. (I will leave this fact as an exercise; it can, for example, be proved via the quantile transformation).

Because of this fact, one can compute the null distribution of D_n assuming that F_0 is the uniform distribution on $(0,1)$. In this case, we can write

$$D_n = \sup_{0 \le t \le 1} |U_n(t)|$$

where $U_n(t)$ is the uniform empirical process from Example 180.

The fact that $\{U_n(t), t \in [0,1]\}$ converges in distribution to a Brownian bridge $\{U(t), t \in [0,1]\}$ as $n \to \infty$ actually allows one to claim that

$$\lim_{n\to\infty} \mathbb{P}\{D_n \le x\} = \mathbb{P}\left\{ \sup_{0 \le t \le 1} |U(t)| \le x \right\} \qquad \text{for every } x > 0.$$

The latter probability can be exactly computed (see, for example, Dudley [5, Proposition 12.3.4]). Thus the uniform central limit theorem gives a way of computing asymptotically valid p-values for Goodness of fit testing via the Kolmogorov Statistic.

The same argument can be used for many related goodness of fit statistics such as

1. *Cramer-Von Mises Statistic: Defined as*

$$W_n := n \int (F_n(x) - F_0(x))^2 \, dF_0(x).$$

2. *Anderson-Darling Statistic: Defined as*

$$A_n := n \int \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x).$$

3. *Smirnov statistics: Defined as*

$$D_n^+ := \sqrt{n} \sup_x \left(F_n(x) - F_0(x)\right) \qquad \text{and} \qquad D_n^+ := \sqrt{n} \sup_x \left(F_0(x) - F_n(x)\right).$$

The asymptotic null distribution of all these statistics can be computed from Brownian bridge and this will be validated by the uniform CLT.

Example 2.3 (Asymptotic Distribution of MLE)**.** *Suppose $X_1, \ldots, X_n$ are i.i.d from an unknown density p_{θ_0} belonging to a known class $\{p_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$. Let $\hat{\theta}_n$ denote the maximum likelihood estimator of θ_0 defined as the maximizer of*

$$\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i)$$

over $\theta \in \Theta$. A classical result is that, under some smoothness assumptions, $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ converges in distribution to $N_k(0, I^{-1}(\theta_0))$ where $I(\theta_0)$ denotes the $k \times k$ Fisher information matrix defined as

$$I(\theta_0) := \mathbb{E}\left(\nabla_\theta \log p_\theta(X) \left(\nabla_\theta \log p_\theta(X)\right)^T\right)$$

where the gradient ∇_θ is evaluated at $\theta = \theta_0$ and the expectation is taken with respect to the density p_{θ_0}.

What smoothness assumptions need to be imposed on $p_\theta, \theta \in \Theta$ for this result to hold? Because the result involves the information matrix $I(\theta_0)$ which involves gradients, a minimal assumption seems to be that $\theta \mapsto \log p_\theta(x)$ needs to be differentiable with respect to θ. Also because of the presence of the expectation in the definition of $I(\theta_0)$, it should be okay if the derivative with respect to θ does not exist on sets of measure zero with respect to p_{θ_0} (think about the model $p_\theta(x) := \exp(-|x - \theta|)/2$).

The classical proofs of this result assume however that this map allows two (or sometimes even three) derivatives. Using uniform central limit theorems, we shall present later a proof using a minimal differentiability assumption called Differentiability in Quadratic Mean (DQM).

Example 2.4 (Asymptotic Distribution Results for M-estimators)**.** *Uniform central limit theorems can be used to derive limiting distributions of other M-estimators as well.*

For example, consider the sample median defined as:

$$\hat{\theta}_n := \operatorname*{argmax}_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |X_i - \theta|.$$

Assuming that the distribution function F of the observations is differentiable at its median θ_0 with positive derivative $f(\theta_0)$, it can be proved that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$$

converges in distribution to $N(0, (4f^2(\theta_0))^{-1})$.

For the mode defined as $\operatorname{argmax}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} m_\theta(X_i)$ with $m_\theta(x) := I\{|x-\theta| \leq 1\}$ and $\Theta = \mathbb{R}$, the asymptotic distribution is much more complicated. The result is that

$$n^{1/3}\left(\hat{\theta}_n - \theta_0\right)$$

converges in distribution to

$$\operatorname*{argmax}_{h \in \mathbb{R}} \left(aZ(h) - bh^2\right)$$

where Z is a standard two-sided Brownian motion starting from 0,

$$a^2 := p(\theta_0 + 1) + p(\theta_0 - 1) \quad \textit{and} \quad b := \frac{1}{2}\left(p'(\theta_0 - 1) - p'(\theta_0 + 1)\right).$$

Here $p(\cdot)$ represents the density of the observations and it is assumed that p is unimodal and symmetric with mode θ_0 i.e., $p'(x) > 0$ for $x < \theta_0$ and $p'(x) < 0$ for $x > \theta_0$. This result is stated here just to illustrate that the limiting distributions of even simple-looking M-estimators can be quite complicated. We shall later see how to prove these results via Uniform CLTs.

2.2 Concentration Results

Let us now start with our discussion of uniform laws of large numbers. The key object of study is

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X_1) \right| \tag{5}$$

where $X_1, \ldots, X_n$ are i.i.d random objects taking values in a space $\mathcal{X}$ and $\mathcal{F}$ is a collection of real-valued functions on $\mathcal{X}$. We shall argue that (5) concentrates around its expectation. This is fairly easy to prove when it is assumed that the all functions in $\mathcal{F}$ are bounded by a positive constant B:

$$\sup_{x \in \mathcal{X}} |f(x)| \leq B \qquad \text{for every } f \in \mathcal{F}. \tag{6}$$

Under the above assumption, we shall prove a concentration result for (5). We shall do this as a consequence of the *bounded differences* concentration inequality.

Theorem 2.5 (Bounded Differences Concentration Inequality)**.** *Suppose $X_1, \ldots, X_n$ are independent random variables taking values in a set $\mathcal{X}$. Suppose $g : \mathcal{X} \times \cdots \times \mathcal{X} \to \mathbb{R}$ be a function that satisfies the following "bounded differences" assumption':*

$$\sup_{x_1, \ldots, x_n, x_i' \in \mathcal{X}} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i \tag{7}$$

for every $i = 1, \ldots, n$. Then for every $t \geq 0$, we have

$$\mathbb{P}\left\{g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \geq t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{8}$$

and

$$\mathbb{P}\left\{g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \leq -t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right). \tag{9}$$

Remark 2.1. *The bounded differences condition* (211) *is equivalent to the following:*

$$|g(x_1, \ldots, x_n) - g(z_1, \ldots, z_n)| \leq c_i$$

whenever $(x_1, \ldots, x_n)$ and $(z_1, \ldots, z_n)$ differ in exactly the i^{th} coordinate.

It is also equivalent to the following condition:

$$|g(x_1, \ldots, x_n) - g(z_1, \ldots, z_n)| \leq \sum_{i=1}^n c_i I\{x_i \neq z_i\} \qquad \text{for all } x_1, \ldots, x_n, z_1, \ldots, z_n \in \mathcal{X}.$$

Theorem 2.5 can be seen as a quantification of the following qualitative statement of Talagrand (see Talagrand [22, Page 2]): *A random variable that depends on the influence of many independent variables (but not too much on any of them) concentrates'*. The numbers c_i control the effect of the i^{th} variable on the function g.

We shall prove Theorem 2.5 in the next class. Let us argue here that it implies a concentration inequality for

$$Z := \sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1)|.$$

under the condition (6). Indeed, let

$$g(x_1, \ldots, x_n) := \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f(X_1)\right|.$$

We shall show below that g satisfies the bounded differences assumption (211) with $c_i := 2B/n$ for $i = 1, \ldots, n$. To see this, note that

$$\begin{aligned} g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) &= \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{j \neq i} f(x_j) + \frac{f(x_i')}{n} - \mathbb{E}f(X_1)\right| \\ &= \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{j=1}^n f(x_j) - \mathbb{E}f(X_1) + \frac{f(x_i')}{n} - \frac{f(x_i)}{n}\right|. \end{aligned}$$

For every $f \in \mathcal{F}$, by triangle inequality and the fact that $|f(x_i)| \leq B$ and $|f(x_i')| \leq B$, we have

$$\left|\frac{1}{n}\sum_{j=1}^n f(x_j) - \mathbb{E}f(X_1) + \frac{f(x_i')}{n} - \frac{f(x_i)}{n}\right| \leq \left|\frac{1}{n}\sum_{j=1}^n f(x_j) - \mathbb{E}f(X_1)\right| + \frac{2B}{n}.$$

Taking supremum over $f \in \mathcal{F}$ on both sides, we obtain

$$g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \leq g(x_1, \ldots, x_n) + \frac{2B}{n}.$$

Interchanging the roles of x_i and x_i', we can deduce that

$$|g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) - g(x_1, \ldots, x_n)| \leq \frac{2B}{n}$$

so that (211) holds with $c_i = 2B/n$. Theorem 2.5 (specifically inequality (8)) then gives

$$\mathbb{P}\left\{Z \geq \mathbb{E}Z + t\right\} \leq \exp\left(\frac{-nt^2}{2B^2}\right) \qquad \text{for every } t \geq 0.$$

Setting

$$\delta := \exp\left(\frac{-nt^2}{2B^2}\right),$$

we deduce that the following inequality:

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1) \right| \leq \mathbb{E}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1) \right| \right) + B\sqrt{\frac{2}{n} \log \frac{1}{\delta}} \tag{10}$$

holds with probability at least $1 - \delta$ for every $\delta > 0$.

This inequality implies that $\mathbb{E}(Z)$ is usually the dominating term for understanding the behavior of Z. This is because typically $\mathbb{E}(Z)$ dominates the last term on the right hand side of (10). Indeed, for every $f \in \mathcal{F}$,

$$\mathbb{E}(Z) \geq \mathbb{E}\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1) \right|. \tag{11}$$

Because

$$\mathbb{E}\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1) \right|^2 = \frac{var(f(X_1))}{n},$$

it is reasonable to believe that the right hand side of (11) will typically be of order $\sqrt{var(f(X_1))/n}$. Thus (unless $var(f(X_1))$ is much smaller compared to B^2 for every $f \in \mathcal{F}$), the first term on the right hand side of (10) usually dominates the second term and hence in order to control the random variable Z, it is enough to focus on the expectation $\mathbb{E}Z$.

3 Lecture 3

3.1 Hoeffding's Inequality and Proof of the Bounded Differences Concentration Inequality

One of the goals of this lecture is to prove the Bounded Differences Concentration Inequality. We shall prove another standard concentration inequality called Hoeffding's inequality and then tweak the proof of Hoeffding's inequality to yield the Bounded Differences Concentration Inequality.

Theorem 3.1 (Hoeffding's Inequality)**.** *Suppose $\xi_1, \ldots, \xi_n$ are independent random variables. Suppose $a_1, \ldots, a_n, b_1, \ldots, b_n$ are constants such that $a_i \leq \xi_i \leq b_i$ almost surely for each $i = 1, \ldots, n$. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left\{ \sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i) \geq t \right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right) \tag{12}$$

and

$$\mathbb{P}\left\{ \sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i) \leq -t \right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

Proof. Let $S := \sum_{i=1}^n (\xi_i = \mathbb{E}\xi_i)$ and write (for a fixed $\lambda \geq 0$)

$$\mathbb{P}\{S \geq t\} \leq \mathbb{P}\{e^{\lambda S} \geq e^{\lambda t}\} \leq e^{-\lambda t}\mathbb{E}e^{\lambda S} = \exp\left(-\lambda t + \psi_S(\lambda)\right)$$

where

$$\psi_S(\lambda) := \log \mathbb{E}e^{\lambda S}$$

is the log moment generating function of S. Now by the independence of $\xi_1, \ldots, \xi_n$,

$$\psi_S(\lambda) = \log \mathbb{E}\exp\left(\lambda \sum_{i=1}^n (\xi_i - \mathbb{E}\xi_i)\right) = \sum_{i=1}^n \log \mathbb{E}\exp\left(\lambda(\xi_i - \mathbb{E}\xi_i)\right) = \sum_{i=1}^n \psi_{\xi_i - \mathbb{E}\xi_i}(\lambda)$$

where $\psi_{\xi_i - \mathbb{E}\xi_i}(\cdot)$ denotes the log moment generating function of $\xi_i - \mathbb{E}\xi_i$. Fix $1 \leq i \leq n$ and let $U := \xi_i - \mathbb{E}\xi_i$. We shall bound $\psi_U(\lambda)$ below. We know that $\mathbb{E}U = 0$ and that $a_i - \mathbb{E}\xi_i \leq U \leq b_i - \mathbb{E}\xi_i$ almost surely. By second order Taylor expansion of $\psi_U(\lambda)$ around 0, we can write

$$\psi_U(\lambda) = \psi_U(0) + \lambda\psi_U'(0) + \frac{\lambda^2}{2}\psi_U''(\lambda')$$

for some $0 \leq \lambda' \leq \lambda$. Note now that $\psi_U(0) = \log \mathbb{E}(1) = 0$. Also

$$\psi_U'(\lambda) = \frac{1}{\mathbb{E}e^{\lambda U}} \frac{d}{d\lambda}\mathbb{E}(e^{\lambda U}) = \frac{\mathbb{E}(Ue^{\lambda U})}{\mathbb{E}e^{\lambda U}}$$

so that

$$\psi_U'(0) = \mathbb{E}U = 0.$$

And

$$\psi_U''(\lambda) = \mathbb{E}\left(U^2 \frac{e^{\lambda U}}{\mathbb{E}e^{\lambda U}}\right) - \left(\mathbb{E}\frac{Ue^{\lambda U}}{\mathbb{E}e^{\lambda U}}\right)^2.$$

Consider now a random variable V whose density with respect to the distribution of U is $e^{\lambda U}/(\mathbb{E}e^{\lambda U})$ i.e.,

$$\frac{dP_V}{dP_U} = \frac{e^{\lambda U}}{\mathbb{E}e^{\lambda U}}.$$

Based on the calculation above, it is then clear that $\psi_U''(\lambda) = var(V) \geq 0$. Note also that V is supported on the interval $[a_i - \mathbb{E}\xi_i, b_i - \mathbb{E}\xi_i]$ (because U is supported on this interval and P_V is absolutely continuous with respect to P_U). As a result,

$$\psi_U''(\lambda) = var(V) = \inf_{m \in \mathbb{R}} \mathbb{E}\left(V - m\right)^2 \leq \mathbb{E}\left(V - \eta\right)^2 \leq \left(\frac{b_i - a_i}{2}\right)^2 = \frac{(b_i - a_i)^2}{4}$$

where η is the mid-point of the interval $[a_i - \mathbb{E}\xi_i, b_i - \mathbb{E}\xi_i]$. We have thus proved that $\psi_U''(\lambda) \leq (b_i - a_i)^2/4$ for every $\lambda \geq 0$. This, along with $\psi_U(0) = 0$ and $\psi_U'(0) = 0$, gives

$$\psi_U(\lambda) \leq \frac{(b_i - a_i)^2}{8}\lambda^2.$$

As a result

$$\psi_S(\lambda) = \sum_{i=1}^n \psi_{\xi_i - \mathbb{E}\xi_i}(\lambda) \leq \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2$$

and consequently

$$\mathbb{P}\{S \geq t\} \leq \exp\left(-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right)$$

for every $\lambda \geq 0$. We can optimize this bound over $\lambda \geq 0$ by setting

$$\lambda = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

to prove (12). To prove the lower tail inequality, just apply (12) to $-\xi_1, \ldots, -\xi_n$. □

The proof given above bounds the probability $\mathbb{P}\{S \geq t\}$ in terms of the Moment Generating Function of S. This technique is known as the *Cramer-Chernoff Method*.

3.1.1 Remarks on Hoeffding's inequality

Consider the following special case of Hoeffding's inequality: Suppose $X_1, \ldots, X_n$ are i.i.d with $\mathbb{E}X_i = \mu$, $var(X_i) = \sigma^2$ and $a \leq X_i \leq b$ almost surely (a and b are constants). Suppose $\bar{X}_n := (X_1 + \cdots + X_n)/n$. Hoeffding's inequality then gives

$$\mathbb{P}\left\{\sqrt{n}(\bar{X}_n - \mu) \geq t\right\} \leq \exp\left(\frac{-2t^2}{(b-a)^2}\right) \qquad \text{for all } t \geq 0. \tag{13}$$

Is this a good bound? By "good" here, we mean if the probability on the right hand side above is close to the bound on the right or if the bound is much looser. To answer this question, we of course need a way of approximately computing the probability on the left hand side. A natural way of doing this is via invoking the Central Limit Theorem (assuming that the CLT is valid). Indeed CLT states that

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{L} N(0, \sigma^2) \qquad \text{as } n \to \infty$$

provided that the distribution of X_i (and in particular the quantities μ, σ^2, a and b) do not depend on n (note that Hoeffding's inequality needs no such assumption; in particular, (13) is valid even when μ, σ^2, a and b all depend on n). Thus we may expect

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

when n is large and when CLT holds. What is $\mathbb{P}\{N(0, \sigma^2) \geq t\}$? We can bound this again by the Cramer-Chernoff method:

$$\mathbb{P}\left\{N(0, \sigma^2) \geq t\right\} \leq \exp\left(-\lambda t + \psi_{N(0,\sigma^2)}(\lambda)\right)$$

for every $\lambda \geq 0$ where $\psi_{N(0,\sigma^2)}$ is the log moment generating function of $N(0, \sigma^2)$. By a straightforward calculation, it can be seen that $\psi_{N(0,\sigma^2)}(\lambda)$ is exactly equal to $\lambda^2\sigma^2/2$. Thus

$$\mathbb{P}\left\{N(0, \sigma^2) \geq t\right\} \leq \inf_{\lambda \geq 0} \exp\left(-\lambda t + \frac{1}{2}\lambda^2\sigma^2\right) = \exp\left(\frac{-t^2}{2\sigma^2}\right) \qquad \text{for every } t \geq 0 \tag{14}$$

Is this bound accurate? It is quite good as can be seen from the following inequality (see, for example, Feller [7, Section 7.1]):

$$\frac{1}{\sqrt{2\pi}}\left(\frac{\sigma}{t} - \frac{\sigma^3}{t^3}\right)\exp\left(\frac{-t^2}{2\sigma^2}\right) \leq \mathbb{P}\left\{N(0, \sigma^2) \geq t\right\} \leq \frac{\sigma}{t\sqrt{2\pi}}\exp\left(\frac{-t^2}{2\sigma^2}\right).$$

So $\exp\left(\frac{-t^2}{2\sigma^2}\right)$ is the correct exponential term controlling the behavior of $\mathbb{P}\{N(0,\sigma^2) \geq t\}$. Now let us compare Hoeffding with the bound (14). Hoeffding gives the bound

$$\exp\left(\frac{-2t^2}{(b-a)^2}\right)$$

while normal approximation suggests

$$\exp\left(\frac{-t^2}{2\sigma^2}\right).$$

Note now that because $a \leq X_1 \leq b$ almost surely,

$$\sigma^2 = var(X_1) \leq \mathbb{E}\left(X_1 - \frac{a+b}{2}\right)^2 \leq \frac{(b-a)^2}{4}.$$

Thus in the regime where CLT holds, Hoeffding is a looser inequality where the variance σ^2 is replaced by the upper bound $(b-a)^2/4$. This looseness can be quite pronounced when X_1 puts less mass near the end points a and b. Here is a potential statistical implication of this looseness.

Example 3.2. *Suppose $X_1, \ldots, X_n$ are i.i.d with $\mathbb{E}X_i = \mu$, $var(X_i) = \sigma^2$ and $a \leq X_i \leq b$ almost surely (a and b are constants). Suppose σ^2, a and b are known while μ is unknown and that we seek a confidence interval for μ. There are two ways of solving this problem.*

The first method uses the CLT (normal approximation). Indeed, by CLT:

$$\mathbb{P}\left\{\left|\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right| \leq t\right\} \to \mathbb{P}\{N(0,1) \leq t\}$$

as $n \to \infty$ for each t. Thus

$$\mathbb{P}\left\{\left|\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right| \leq z_{\alpha/2}\right\} \to \mathbb{P}\{N(0,1) \leq z_{\alpha/2}\} = 1 - \alpha$$

where $z_{\alpha/2}$ is defined so that the last equality above holds. This leads to the following C.I for μ:

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]. \tag{15}$$

Note that this is an "asymptotically valid" $100(1-\alpha)\%$ confidence interval for μ. Its finite sample coverage, on the other hand, may not be $100(1-\alpha)\%$.

The second method for constructing a confidence interval for μ uses the Hoeffding inequality which states that

$$\mathbb{P}\left\{|\sqrt{n}(\bar{X}_n - \mu)| \geq t\right\} \leq 2\exp\left(\frac{-2t^2}{(b-a)^2}\right) \qquad \text{for every } t \geq 0.$$

Thus, by taking,

$$t = (b-a)\sqrt{\frac{1}{2}\log\frac{2}{\alpha}},$$

one gets the following confidence interval for μ:

$$\left[\bar{X}_n - \frac{b-a}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{2}{\alpha}}, \bar{X}_n + \frac{b-a}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{2}{\alpha}}\right]. \tag{16}$$

This inequality has guaranteed finite sample coverage $100(1-\alpha)\%$. But this interval might be much too big compared to (15)*. Which of the two intervals* (15) *and* (16) *would you prefer?*

3.1.2 Hoeffding's inequality for Martingale Differences

Theorem 3.3 (Hoeffding's inequality for Martingale Differences)**.** *Suppose $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are increasing σ-fields and suppose $\xi_1, \ldots, \xi_n$ are random variables with ξ_i being $\mathcal{F}_i$-measurable. Assume that*

$$\mathbb{E}\left(\xi_i - \mathbb{E}\xi_i | \mathcal{F}_{i-1}\right) = 0 \qquad \text{almost surely} \tag{17}$$

for all $i = 1, \ldots, n$. Also assume that, for each $1 \leq i \leq n$, the conditional distribution of ξ_i given $\mathcal{F}_{i-1}$ is supported on an interval whose length is bounded from above by the deterministic quantity R_i. Then

$$\mathbb{P}\left\{\sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i) \geq t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} R_i^2}\right) \tag{18}$$

and

$$\mathbb{P}\left\{\sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i) \leq -t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} R_i^2}\right)$$

for every $t \geq 0$.

Remark 3.1. *The assumption* (17) *means that $(S_j, \mathcal{F}_j), j = 1, \ldots, n$ is a martingale where $S_j := \sum_{i=1}^{j}(\xi_i - \mathbb{E}\xi_i)$. Therefore the sequence $\{\xi_i - \mathbb{E}\xi_i, i = 1, \ldots, n\}$ is a martingale difference sequence.*

Proof. Let $S = \sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i)$. As before, for every $t \geq 0$ and $\lambda \geq 0$,

$$\mathbb{P}\left\{S \geq t\right\} \leq \exp\left(-\lambda t + \psi_S(\lambda)\right)$$

with

$$\psi_S(\lambda) := \log \mathbb{E}e^{\lambda S} = \log \mathbb{E}\exp\left(\lambda \sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i)\right).$$

Observe now that

$$\mathbb{E}\left(e^{\lambda S} | \mathcal{F}_{n-1}\right) = \exp\left(\lambda \sum_{i=1}^{n-1}(\xi_i - \mathbb{E}\xi_i)\right) \mathbb{E}\left(e^{\lambda(\xi_n - \mathbb{E}\xi_n)} | \mathcal{F}_{n-1}\right).$$

Now because $\mathbb{E}\xi_n = \mathbb{E}(\xi_n | \mathcal{F}_{n-1})$, we can use exactly the same argument as in the proof of Hoeffding's inequality in the independent case (via second order Taylor expansion of the log Moment Generating Function) to deduce that

$$\mathbb{E}\left(e^{\lambda(\xi_n - \mathbb{E}\xi_n)} | \mathcal{F}_{n-1}\right) \leq \exp\left(\frac{\lambda^2 R_n^2}{8}\right)$$

and this gives

$$\mathbb{E}e^{\lambda S} \leq \exp\left(\frac{\lambda^2 R_n^2}{8}\right) \mathbb{E}\exp\left(\lambda \sum_{i=1}^{n-1}(\xi_i - \mathbb{E}\xi_i)\right).$$

Now repeat the above argument (by conditioning on $\mathcal{F}_{n-2}$, then $\mathcal{F}_{n-3}$ and so on) to deduce that

$$\mathbb{E}e^{\lambda S} \leq \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n} R_i^2\right).$$

This gives

$$\mathbb{P}\left\{S \geq t\right\} \leq \exp\left(-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^{n} R_i^2\right).$$

Optimize over λ to deduce (18). For the proof of the lower tail inequality, argue with $-\xi_i$ in place of ξ_i. $\square$

3.1.3 Proof of the Bounded Differences Concentration Inequality

We shall now prove the bounded differences concentration inequality as a simple consequence of Theorem 3.3. Recall the statement of the Bounded Differences Concentration Inequality:

Theorem 3.4 (Bounded Differences Concentration Inequality)**.** *Suppose $X_1, \ldots, X_n$ are independent random variables taking values in a set $\mathcal{X}$. Suppose $g : \mathcal{X} \times \cdots \times \mathcal{X} \to \mathbb{R}$ be a function that satisfies the following "bounded differences" assumption':*

$$|g(x_1, \ldots, x_n) - g(z_1, \ldots, z_n)| \leq \sum_{i=1}^{n} c_i I\{x_i \neq z_i\} \tag{19}$$

for some constants $c_1, \ldots, c_n$. Then for every $t \geq 0$, we have

$$\mathbb{P}\left\{g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \geq t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{20}$$

and

$$\mathbb{P}\left\{g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \leq -t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof of Theorem 3.4. We shall apply the martingale Hoeffding inequality to

$$\xi_i := \mathbb{E}\left(g(X_1, \ldots, X_n)|X_1, \ldots, X_i\right) - \mathbb{E}\left(g(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}\right) \qquad \text{for } i = 1, \ldots, n$$

and $\mathcal{F}_i$ taken to be the sigma field generated by $X_1, \ldots, X_i$ for $i = 1, \ldots, n$. Clearly ξ_i is $\mathcal{F}_i$ measurable and $\mathbb{E}\xi_i = 0$. Also

$$\begin{aligned}
\mathbb{E}\left(\xi_i|\mathcal{F}_{i-1}\right) &= \mathbb{E}\left(\xi_i|X_1, \ldots, X_{i-1}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(g(X_1, \ldots, X_n)|X_1, \ldots, X_i\right)|X_1, \ldots, X_{i-1}\right) - \mathbb{E}\left[g(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}\right] \\
&= \mathbb{E}\left[g(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}\right] - \mathbb{E}\left[g(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}\right] = 0.
\end{aligned}$$

Thus $(\xi_i, \mathcal{F}_i)$ is a martingale difference sequence. We shall now argue that the conditional distribution of ξ_i given $\mathcal{F}_{i-1}$ is supported on an interval of length bounded from above by c_i. For this, we need to look at the condition distribution of ξ_i given $X_1, \ldots, X_{i-1}$. So let us fix $X_1, \ldots, X_{i-1}$ at $x_1, \ldots, x_{i-1}$. Then ξ_i is a function solely of X_i and we need to look at the range of values of ξ_i as $X_i = x$ varies. We therefore need to look at the values:

$$x \mapsto \mathbb{E}\left[g(X_1, \ldots, X_n)|X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x\right] - \mathbb{E}\left[g(X_1, \ldots, X_n)|X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\right]$$

as x varies and $x_1, \ldots, x_{i-1}$ are fixed. Now, by **independence** of $X_1, \ldots, X_n$, the right hand side above equals

$$\mathbb{E}g(x_1, \ldots, x_{i-1}, x, X_{i+1}, \ldots, X_n) - \textit{constant}$$

where the "constant" term only depends on $x_1, \ldots, x_{i-1}$. Thus we can take R_i to be

$$\begin{aligned}
R_i &:= \sup_{x, x' \in \mathcal{X}} \left|\mathbb{E}g(x_1, \ldots, x_{i-1}, x, X_{i+1}, \ldots, X_n) - \mathbb{E}g(x_1, \ldots, x_{i-1}, x', X_{i+1}, \ldots, X_n)\right| \\
&\leq \sup_{x, x' \in \mathcal{X}} \mathbb{E}\left|g(x_1, \ldots, x_{i-1}, x, X_{i+1}, \ldots, X_n) - g(x_1, \ldots, x_{i-1}, x', X_{i+1}, \ldots, X_n)\right|.
\end{aligned}$$

It is clear now that $R_i \leq c_i$ by the bounded differences assumption (19). We can therefore apply Theorem 3.3 with $R_i = c_i$ which finishes the proof of Theorem 3.4. □

4 Lecture 4

4.1 Bennett's Inequality

Let us recall the Hoeffding inequality from last lecture. It states that

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

for every $t \geq 0$ where $X_1, \ldots, X_n$ are independent random variables with $a_i \leq X_i \leq b_i$ almost surely. We remarked that when $\sum_{i=1}^{n} var(X_i)$ is much smaller than $\sum_{i=1}^{n}(b_i - a_i)^2/4$ and when the CLT holds, then the tail bound given by Hoeffding can be loose. Bennett's inequality attempts to given tail bounds which involve variances.

Theorem 4.1 (Bennett's inequality)**.** *Suppose $X_1, \ldots, X_n$ are independent random variables having finite variances. Suppose $X_i \leq B$ almost surely for each $i = 1, \ldots, n$ (here B is deterministic). Let $V := \sum_{i=1}^{n} \mathbb{E}X_i^2$. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(-\frac{V}{B^2}h\left(\frac{tB}{V}\right)\right) \tag{21}$$

where

$$h(u) := (1+u)\log(1+u) - u \qquad \text{for } u \geq 0. \tag{22}$$

Remark 4.1. *Bennett's inequality, as stated above, gives only the upper tail bound. To get the lower bound, one needs to impose the assumption $X_i \geq -B$. In this case, one gets*

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \leq -t\right\} \leq \exp\left(-\frac{V}{B^2}h\left(\frac{tB}{V}\right)\right)$$

Remark 4.2. *For the function h defined in (22), it is easy to see that $h(0) = 0$, $h'(0) = 0$ and $h''(0) = 1$. Therefore for u near zero, we have $h(u) \approx u^2/2$. Thus when tB/V is small, the bound given by Bennett inequality looks like:*

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \leq -t\right\} \leq \exp\left(-\frac{V}{B^2}h\left(\frac{tB}{V}\right)\right) \approx \exp\left(\frac{-t^2}{2V}\right).$$

Thus Bennett's inequality gives Gaussian like tails with $V = \sum_{i=1}^{n} \mathbb{E}X_i^2$ in some regimes.

As an example, suppose that $\mathbb{E}X_i = 0$, $var(X_i) = \sigma^2$ and $X_i \leq 1$. Then $V = n\sigma^2$ and Bennett's inequality gives

$$\mathbb{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \geq t\right\} \leq \exp\left(-Vh\left(\frac{t\sqrt{n}}{V}\right)\right) = \exp\left(-n\sigma^2 h\left(\frac{t}{\sigma^2\sqrt{n}}\right)\right).$$

When t is small compared to $\sqrt{n}\sigma^2$, we get a Gaussian type bound.

Proof of Theorem 4.1. Without loss of generality, take $B = 1$ (by working with the variables $X_1/B, \ldots, X_n/B$ instead of $X_1, \ldots, X_n$).

This proof relies on the following observation: Let $\phi : \mathbb{R} \to \mathbb{R}$ denote the function $\phi(u) := e^u - u - 1$. Then the map $u \mapsto \phi(u)/u^2$ is increasing on $\mathbb{R}$ (we take $\phi(0)/0^2 = 1/2$). I will leave as homework the verification of this fact.

Let $S := \sum_{i=1}^n (X_i - \mathbb{E}X_i)$. Then for every $\lambda \geq 0$ as before

$$\mathbb{P}\{S \geq t\} \leq e^{-\lambda t}\mathbb{E}e^{\lambda \sum_{i=1}^n (X_i - \mathbb{E}X_i)} = e^{-\lambda t}e^{-\lambda \sum_{i=1}^n \mathbb{E}X_i}\prod_{i=1}^n \mathbb{E}e^{\lambda X_i} \tag{23}$$

where we have used the independence of $X_1, \ldots, X_n$. Now because $X_i \leq 1$, we have $\lambda X_i \leq \lambda$ and hence (using the fact that $\phi(u)/u^2$ is increasing), we deduce that

$$\frac{\phi(\lambda X_i)}{(\lambda X_i)^2} \leq \frac{\phi(\lambda)}{\lambda^2}$$

which implies that

$$e^{\lambda X_i} \leq \lambda X_i + 1 + X_i^2\phi(\lambda).$$

Using this bound in the right hand side of (23), we obtain

$$\mathbb{P}\{S \geq t\} \leq \exp(-\lambda t - \lambda \sum_{i=1}^n \mathbb{E}X_i)\prod_{i=1}^n \left(1 + \lambda \mathbb{E}X_i + \phi(\lambda)\mathbb{E}X_i^2\right).$$

We now use the trivial inequality $(1 + x \leq e^x)$

$$1 + \lambda \mathbb{E}X_i + \phi(\lambda)\mathbb{E}X_i^2 \leq \exp\left(\lambda \mathbb{E}X_i + \phi(\lambda)\mathbb{E}X_i^2\right)$$

to obtain

$$\mathbb{P}\{S \geq t\} \leq \exp(-\lambda t - \lambda \sum_{i=1}^n \mathbb{E}X_i)\exp\left(\lambda \sum_{i=1}^n \mathbb{E}X_i + \phi(\lambda)\sum_{i=1}^n \mathbb{E}X_i^2\right) = \exp\left(-\lambda t + \phi(\lambda)V\right)$$

for every $\lambda \geq 0$. We now optimize the above bound by taking the derivative with respect to λ and setting it equal to zero to obtain:

$$-t + V(e^\lambda - 1) = 0 \implies \lambda = \log\left(1 + \frac{t}{V}\right).$$

For this value of λ, it is straightforward to deduce (21). □

The form of the bound in Bennett's inequality can be simplified by using the following inequality (whose proof is left as exercise):

$$h(u) = (1+u)\log(1+u) - u \geq \frac{u^2}{2(1 + \frac{u}{3})} \qquad \text{for all } u \geq 0.$$

This leads to the following result which is known as Bernstein's inequality.

Theorem 4.2 (Bernstein's Inequality)**.** *Suppose $X_1, \ldots, X_n$ are independent random variables with finite variances and suppose that $|X_i| \leq B$ almost surely for each $i = 1, \ldots, n$ (B is deterministic). Let $V := \sum_{i=1}^n \mathbb{E}X_i^2$. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(\frac{-t^2}{2(V + \frac{tB}{3})}\right)$$

and

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq -t\right\} \leq \exp\left(\frac{-t^2}{2(V + \frac{tB}{3})}\right).$$

Remark 4.3. *There is a version of Bernstein's inequality that replaces the boundedness assumption by weaker moment restrictions. See Boucheron et al. [3, Theorem 2.10].*

The two bounds in Bernstein's inequality can be combined to write

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq t\right\} \leq 2\exp\left(\frac{-t^2}{2(V + \frac{tB}{3})}\right).$$

We can now attempt to find the value of t which makes the bound on the right hand side above exactly equal to α i.e., we want to solve the equation

$$2\exp\left(\frac{-t^2}{2(V + \frac{tB}{3})}\right) = \alpha.$$

This leads to the quadratic equation

$$t^2 - \frac{2tB}{3}\log\frac{2}{\alpha} - 2V\log\frac{2}{\alpha} = 0$$

whose nonnegative solution is given by

$$t = \frac{B}{3}\log\frac{2}{\alpha} + \sqrt{\frac{B^2}{9}\left(\log\frac{2}{\alpha}\right)^2 + 2V\log\frac{2}{\alpha}} \leq \sqrt{2V\log\frac{2}{\alpha}} + \frac{2B}{3}\log\frac{2}{\alpha}$$

where, in the last inequality, we used the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Thus Bernstein's inequality implies that

$$\left|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \leq \sqrt{2V\log\frac{2}{\alpha}} + \frac{2B}{3}\log\frac{2}{\alpha}$$

with probability at least $1-\alpha$. Now if $X_1, \ldots, X_n$ are i.i.d with mean zero, variance σ^2 and bounded in absolute value by B, then $V = n\sigma^2$ which gives that the inequality

$$|\bar{X}_n| \leq \frac{\sigma}{\sqrt{n}}\sqrt{2\log\frac{2}{\alpha}} + \frac{2B}{3n}\log\frac{2}{\alpha} \tag{24}$$

holds with probability at least $1-\alpha$. Note that if $\bar{X}_n$ is normal, then $|\bar{X}_n|$ will be bounded by the first term in the right hand side above with probability at least $1-\alpha$. Therefore the deviation bound (24) agrees with the normal approximation bound except for the smaller order term (which if of order $1/n$; the leading term being of order $1/\sqrt{n}$).

4.2 Back to Concentration of $\sup_{f\in\mathcal{F}}|\ldots|$

Let us now get back to the concentration behavior of

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right|. \tag{25}$$

Let us introduce some notation here. We denote the empirical measure of $X_1, \ldots, X_n$ by P_n. The common distribution of the i.i.d random observations $X_1, \ldots, X_n$ will be denoted by P. We also let

$$Pf := \mathbb{E}f(X_1) \quad \text{and} \quad P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

The quantity can therefore be written as

$$\sup_{f\in\mathcal{F}}|P_n f - Pf| \quad \text{or} \quad \sup_{f\in\mathcal{F}}|(P_n - P)f|.$$

The concentration inequality that we proved via the Bounded Differences Inequality is the following. Suppose that $\mathcal{F}$ consists of functions that are uniformly bounded by B, then

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \leq \mathbb{E}\left(\sup_{f\in\mathcal{F}} |P_n f - Pf|\right) + B\sqrt{\frac{2}{n}\log\frac{1}{\alpha}} \tag{26}$$

with probability $1-\alpha$.

We remarked previously that when $var(f(X_1))$ is small compared to B for every $f \in \mathcal{F}$, this inequality is not sharp. In such situations, it is much more helpful to use *Talagrand's concentration inequality for the suprema of empirical processes* which is stronger than (26) and also deeper and harder to prove. We shall give the statement of this inequality but not the proof (for a proof, you can refer to Boucheron et al. [3, Section 12.4]). Before stating Talagrand's inequality, let us look at a statistical application where it becomes necessary to deal with function classes $\mathcal{F}$ where the variances are small compared to the uniform bound. This application concerns the regression problem (it also applies similarly to the classification problem).

Example 4.3 (Bounded Regression)**.** *We have two random objects X and Y taking values in spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Assume that $\mathcal{Y}$ is a bounded subinterval of the real line. The problem is to predict $Y \in \mathcal{Y}$ on the basis of $X \in \mathcal{X}$. A predictor (or estimator) is any function g which maps $\mathcal{X}$ to $\mathbb{R}$. The (test) error of an estimator g is defined by*

$$L(g) := \mathbb{E}\left(Y - g(X)\right)^2.$$

The goal of regression is to construct an estimator with small error based on n i.i.d observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ having the same distribution as (X, Y). For an estimator g, its empirical error is given by

$$L_n(g) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g(X_i)\right)^2.$$

A natural strategy is to select a class of predictors $\mathcal{G}$ and then to choose the predictor in $\mathcal{G}$ which has the smallest empirical error i.e.,

$$\hat{g}_n := \operatorname*{argmin}_{g\in\mathcal{G}} L_n(g).$$

The key question now is how good a predictor is $\hat{g}_n$ in terms of test error i.e., how small is its error:

$$L(\hat{g}_n) := \mathbb{E}\left[\left(Y - \hat{g}_n(X)\right)^2 | X_1, Y_1, \ldots, X_n, Y_n\right].$$

In particular, we are interested in how small $L(\hat{g}_n)$ is compared to $\inf_{g\in\mathcal{G}} L(g)$. Suppose that this infimum is achieved at some $g^* \in \mathcal{G}$. To bound $L(\hat{g}_n) - L(g^*)$, it is natural to write:

$$\begin{aligned} L(\hat{g}_n) - L(g^*) &= \left(L_n(\hat{g}_n) - L_n(g^*)\right) + \left(L(\hat{g}_n) - L_n(\hat{g}_n)\right) + \left(L_n(g^*) - L(g^*)\right) \\ &\leq \left(L(\hat{g}_n) - L_n(\hat{g}_n)\right) + \left(L_n(g^*) - L(g^*)\right). \end{aligned}$$

We can now use Empirical Process Notation. Let P denote the joint distribution of (X, Y) and P_n denote the empirical distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let $\mathcal{F}$ denote the class of all functions $(x, y) \mapsto (y - g(x))^2$ as g varies over $\mathcal{G}$.

With this notation, the above inequality becomes

$$P(\hat{f}_n - f^*) \leq (P - P_n)(\hat{f}_n - f^*) \tag{27}$$

where $\hat{f}_n(x, y) := (y - \hat{g}_n(x))^2$ and $f^*(x, y) := (y - g^*(x))^2$. In order to proceed further, we need to bound the right hand side above. A crude bound is

$$(P - P_n)(\hat{f}_n - f^*) \leq 2\sup_{f\in\mathcal{F}} |P_n f - Pf|. \tag{28}$$

If we now assume that the class of functions $\mathcal{F}$ is uniformly bounded by B, we can use the concentration inequality (26). This will give some bound on $L(\hat{g}_n) - L(g^*)$ provided one can control the Expectation (we

shall study how to do this later). It is important now to note that this method will never give a bound better than $1/\sqrt{n}$ for $L(\hat{g}_n) - L(g^*)$. This is because there is already a term of $n^{-1/2}$ in the right hand side of (26). But in regression, at least for small classes $\mathcal{G}$ (such as finite dimensional function class), we would expect the test error to decay much faster than $n^{-1/2}$ (such as at the n^{-1} rate). Such fast rates cannot be proved by this method.

To prove faster rates, one needs to use a technique called "localization" instead of the crude bound (28). Let $\hat{\delta}$ denote the left hand side of (27) and the goal is to get bounds for $\hat{\delta}$. The inequality (27) implies that

$$\hat{\delta} \leq \sup_{f \in \mathcal{F}: P(f-f^*) \leq \hat{\delta}} (P - P_n)(f - f^*).$$

Thus we really need to understand how to bound

$$\sup_{f \in \mathcal{F}: P(f-f^*) \leq \hat{\delta}} (P - P_n)(f - f^*).$$

This is a bit complicated because the class of functions in the supremum is random and depends on $\hat{\delta}$. But let us ignore that for the moment and focus on obtaining bounds for

$$\sup_{f \in \mathcal{F}: P(f-f^*) \leq \delta} (P - P_n)(f - f^*). \tag{29}$$

for a deterministic but small δ. The key now is to realize that the functions involved here have small variances (at least in the well specified case where $g^*(x) = \mathbb{E}(Y|X = x)$). Indeed, in the well specified case, we have

$$P(f - f^*) = \mathbb{E}\left[(Y - g(X))^2 - (Y - g^*(X))^2\right] = \mathbb{E}(g(X) - g^*(X))^2.$$

Hence when $P(f - f^*) \leq \delta$, we have

$$\begin{aligned} var(f - f^*) &\leq \mathbb{E}(f(X,Y) - f^*(X,Y))^2 \\ &= \mathbb{E}\left[(Y - g(X))^2 - (Y - g^*(X))^2\right]^2 \\ &= \mathbb{E}\left[(2Y - g(X) - g^*(X))(g(X) - g^*(X))\right]^2 \leq C_B \mathbb{E}(g(X) - g^*(X))^2 \leq C_B \delta. \end{aligned}$$

If we use the concentration inequality (26) to control (29), the resulting bound will be atleast $\sqrt{B/n}$ independent of δ. This will not lead to any faster rates. However Talagrand's inequality will make sure of the small variances to give a better bound. Together wil suitable bounds for the expectation, one will obtain faster rates for regression under appropriate assumptions on $\mathcal{G}$.

Similar analysis can be done for classification but certain assumptions.

Let us now state Talagrand's concentration inequality for empirical processes. As before, assume that $\mathcal{F}$ is uniformly bounded by a constant B. Then, letting, $Z := \sup_{f \in \mathcal{F}} |P_n f - Pf|$, we have

$$Z \leq C\mathbb{E}(Z) + C\sqrt{\frac{\sup_{f \in \mathcal{F}} var(f(X_1))}{n} \log \frac{1}{\alpha}} + C\frac{B}{n} \log \frac{1}{\alpha}$$

with probability at least $1 - \alpha$. Here C is a universal constant which can be made explicit. Note that the leading terms are $\mathbb{E}Z$ and the second term which involves only the variances. The final term is of order $1/n$.

After learning how to control $\mathbb{E}Z$, we shall come back to regression and classification to provide explicit error bounds on the test error for various classes $\mathcal{G}$. We shall use Talagrand's inequality together with localization.

5 Lecture 5

This lecture was delivered by Chi Jin. He made some changes to the notes (his modified notes are in the folder).

5.1 Bounds for the Expected Suprema

The next major topic of the course involves bounding the quantity:

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf|. \tag{30}$$

The two main ideas here are *Symmetrization* and *Chaining*. We shall go over symmetrization first.

Symmetrization bounds (30) from above using the *Rademacher complexity* of the class $\mathcal{F}$. Let us first define the Rademacher complexity. A Rademacher random variable is a random variable ϵ that takes the two values $+1$ and -1 with probability $1/2$ each. For a subset $A \subseteq \mathbb{R}^n$, its Rademacher average is defined by

$$R_n(A) := \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^n \epsilon_i a_i \right|$$

where the expectation is taken with respect to i.i.d Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$. Note first that $\sum_{i=1}^n \epsilon_i a_i / n$ measures the "correlation" between the values $a_1, \ldots, a_n$ and independent Rademacher noise. This means therefore that $R_n(A)$ is large when there exists vectors $(a_1, \ldots, a_n) \in A$ that fit the Rademacher noise very well. This usually means that the set A is large. In this sense, $R_n(A)$ measures the size of the set A.

In the empirical process setup, we have i.i.d random observations $X_1, \ldots, X_n$ taking values in $\mathcal{X}$ as well as a class of real-valued functions $\mathcal{F}$ on $\mathcal{X}$. Let

$$\mathcal{F}(X_1, \ldots, X_n) := \{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}.$$

This is a random subset of $\mathbb{R}^n$ and its Rademacher average, $R_n(\mathcal{F}(X_1, \ldots, X_n))$, is a random variable. The expectation of this random variable with respect to the distirbution of $X_1, \ldots, X_n$ is called the Rademacher Complexity of $\mathcal{F}$:

$$R_n(\mathcal{F}) := \mathbb{E} R_n(\mathcal{F}(X_1, \ldots, X_n)).$$

It is easy to see that

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

where the expectation is taken with respect to $\epsilon_1, \ldots, \epsilon_n$ and $X_1, \ldots, X_n$ which are all independent (ϵ_i's are i.i.d Rademachers and X_i's are i.i.d having distribution P).

The next result shows that the expectation in (30) is bounded from above by twice the Rademacher complexity $R_n(\mathcal{F})$.

Theorem 5.1 (Symmetrization)**.** *We have*

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \leq 2R_n(\mathcal{F}) := \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \epsilon_i f(X_i) \right|$$

where the expectation on the left hand side is taken with respect to $X_1, \ldots, X_n$ being i.i.d with distribution P while the expectation on the right hand side is taken both with respect to the X's and independent Rademachers $\epsilon_1, \ldots, \epsilon_n$.

Proof. Suppose $X_1', \ldots, X_n'$ are random variables such that $X_1, \ldots, X_n, X_1', \ldots, X_n'$ are all independent having the same distribution P. We can then write

$$\mathbb{E} f(X_1) = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n f(X_i') \right).$$

As a result, we have

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}f(X_1)\right| &= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}f(X_i')\right)\right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\frac{1}{n}\sum_{i=1}^{n}f(X_i')\middle|X_1,\ldots,X_n\right)\right| \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\frac{1}{n}\sum_{i=1}^{n}f(X_i')\right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i)-f(X_i')\right)\right|.
\end{aligned}$$

The method used above is basically called symmetrization. We now introduce i.i.d Rademacher variables $\epsilon_1,\ldots,\epsilon_n$. Because X_i and X_i' are independent copies, it is clear that the distribution of $f(X_i)-f(X_i')$ is the same as that of $\epsilon_i\left(f(X_i)-f(X_i')\right)$. As a result, we have

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}f(X_1)\right| &\le \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(X_i)-f(X_i')\right)\right| \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right|+\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i f(X_i')\right|\right) \le 2R_n(\mathcal{F}).
\end{aligned}$$

□

Theorem 5.1 implies that we can control (30) by bounding from above $R_n(\mathcal{F})$. The usual strategy used for bounding $R_n(\mathcal{F})$ is the following. One first fixes points $x_1,\ldots,x_n\in\mathcal{X}$ and bounds the Rademacher average of the set

$$\mathcal{F}(x_1,\ldots,x_n):=\left\{(f(x_1),\ldots,f(x_n)):f\in\mathcal{F}\right\}. \tag{31}$$

If an upper bound is obtained for this Rademacher average that does not depend on $x_1,\ldots,x_n$, then it automatically also becomes an upper bound for $R_n(\mathcal{F})$. Note that in order to bound $R_n(\mathcal{F}(x_1,\ldots,x_n))$ for fixed points $x_1,\ldots,x_n$, we only need to deal with the simple distribution of $\epsilon_1,\ldots,\epsilon_n$ which makes this much more tractable.

The main technique for bounding $R_n(\mathcal{F}(x_1,\ldots,x_n))$ will be *chaining*. Before we get to chaining however, we shall first look at a more elementary bound that work well in certain situations for Boolean classes $\mathcal{F}$. As we shall see later, this bound will not be as accurate as the bounds given by chaining however.

5.2 Simple bounds on the Rademacher Average $R_n(\mathcal{F}(x_1,\ldots,x_n))$

These bounds are based on the following simple result.

Proposition 5.2. *Suppose A is a finite subset of $\mathbb{R}^n$ with cardinality $|A|$. Then*

$$R_n(A)=\mathbb{E}\max_{a\in A}\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i a_i\right|\le\sqrt{6}\sqrt{\frac{\log(2|A|)}{n}}\max_{a\in A}\sqrt{\frac{1}{n}\sum_{i=1}^{n}a_i^2}. \tag{32}$$

Proof of Proposition 5.2. It is trivial to see that for every nonnegative random variable X, one has

$$\mathbb{E}X=\int_0^\infty\mathbb{P}\{X>x\}dx$$

which can, for example, be proved by interchanging the integral and the probability on the right hand side. We shall use this identity below.

For every $a \in \mathcal{A}$, we have

$$\begin{aligned}\mathbb{E}\exp\left[\frac{\left(\sum_{i=1}^n a_i\epsilon_i\right)^2}{6\sum_{i=1}^n a_i^2}\right] &= \int_0^\infty \mathbb{P}\left\{\exp\left[\frac{\left(\sum_{i=1}^n a_i\epsilon_i\right)^2}{6\sum_{i=1}^n a_i^2}\right] > x\right\} dx \\ &\leq 1 + \int_1^\infty \mathbb{P}\left\{\left|\sum_{i=1}^n a_i\epsilon_i\right| > \sqrt{6\sum_{i=1}^n a_i^2}\sqrt{\log x}\right\} dx \\ &\leq 1 + 2\int_1^\infty \exp\left(-\frac{(6\sum_{i=1}^n a_i^2)(\log x)}{2\sum_{i=1}^n a_i^2}\right) dx = 1 + 2\int_1^\infty x^{-3}dx = 2.\end{aligned}$$

The probability bound above comes from Hoeffding's inequality. From the above, we have

$$\begin{aligned}\mathbb{E}\exp\left[\max_{a\in A}\frac{\left(\sum_{i=1}^n a_i\epsilon_i\right)^2}{6\sum_{i=1}^n a_i^2}\right] &= \mathbb{E}\max_{a\in A}\exp\left[\frac{\left(\sum_{i=1}^n a_i\epsilon_i\right)^2}{6\sum_{i=1}^n a_i^2}\right] \\ &\leq \mathbb{E}\sum_{a\in A}\exp\left[\frac{\left(\sum_{i=1}^n a_i\epsilon_i\right)^2}{6\sum_{i=1}^n a_i^2}\right] \leq 2|A|\end{aligned}$$

where $|A|$ is the cardinality of A. This can be rewritten as

$$\mathbb{E}\exp\left(\max_{a\in A}\left|\frac{\sum_{i=1}^n a_i\epsilon_i}{\sqrt{6\sum_{i=1}^n a_i^2}}\right|\right)^2 \leq 2|A|.$$

Now the function $x \mapsto e^{x^2}$ is convex (as can be easily checked by computing the second derivative) so that Jensen's inequality gives

$$\exp\left(\mathbb{E}\max_{a\in A}\left|\frac{\sum_{i=1}^n a_i\epsilon_i}{\sqrt{6\sum_{i=1}^n a_i^2}}\right|\right)^2 \leq \mathbb{E}\exp\left(\max_{a\in A}\left|\frac{\sum_{i=1}^n a_i\epsilon_i}{\sqrt{6\sum_{i=1}^n a_i^2}}\right|\right)^2 \leq 2|A|$$

so that

$$\mathbb{E}\max_{a\in A}\left|\frac{\sum_{i=1}^n a_i\epsilon_i}{\sqrt{6\sum_{i=1}^n a_i^2}}\right| \leq \sqrt{\log(2|A|)}.$$

From here, the inequality given in (32) follows by the trivial inequality:

$$\max_{a\in A}\left|\frac{\sum_{i=1}^n a_i\epsilon_i}{\sqrt{6\sum_{i=1}^n a_i^2}}\right| \geq \frac{\max_{a\in A}|\sum_{i=1}^n a_i\epsilon_i|}{\max_{a\in A}\sqrt{6\sum_{i=1}^n a_i^2}}.$$

□

Let us now apply Proposition 5.2 to control the Rademacher complexity of Boolean Function Classes. We say that $\mathcal{F}$ is a Boolean class if $f(x)$ takes only the two values 0 and 1 for every function f and every $x \in \mathcal{X}$. Boolean classes $\mathcal{F}$ arise in the problem of classification (where $\mathcal{F}$ can be taken to consist of all functions f of the form $I\{g(X) \neq Y\}$). They are also important for historical reasons: empirical process theory has its origins in the study of $\sup_t(F_n(t) - F(t))$ which corresponds to taking $\mathcal{F} := \{I(-\infty, t] : t \in \mathbb{R}\}$.

Let us now fix a Boolean class $\mathcal{F}$ and points $x_1, \ldots, x_n$. The set $\mathcal{F}(x_1, \ldots, x_n)$ (defined as in (31)) is obviously then finite and we can apply Proposition 5.2 to control $R_n(\mathcal{F}(x_1, \ldots, x_n))$. This gives

$$R_n(\mathcal{F}(x_1, \ldots, x_n)) \leq \sqrt{6}\sqrt{\frac{\log(2|\mathcal{F}(x_1, \ldots, x_n)|)}{n}}\max_{f\in\mathcal{F}}\sqrt{\frac{1}{n}\sum_{i=1}^n f^2(x_i)}.$$

Because $\mathcal{F}$ is Boolean, we can bound each $f^2(x_i)$ by 1 in the right hand side above to obtain

$$R_n(\mathcal{F}(x_1,\ldots,x_n)) \leq \sqrt{6}\sqrt{\frac{\log(2|\mathcal{F}(x_1,\ldots,x_n)|)}{n}}. \tag{33}$$

Now for some classes $\mathcal{F}$, the cardinality $|\mathcal{F}(x_1,\ldots,x_n)|$ can be bounded from above by a polynomial in n for every set of n points $x_1,\ldots,x_n \in \mathcal{X}$. We refer to such classes as classes having polynomial discrimination. For such classes, we can bound $\mathbb{R}_n(\mathcal{F}(x_1,\ldots,x_n))$ by a constant multiple of $\sqrt{(\log n)/n}$ for every $x_1,\ldots,x_n$. Because $R_n(\mathcal{F})$ is defined as the expectation of $\mathbb{R}_n(\mathcal{F}(X_1,\ldots,X_n))$, we would obtain that, for such Boolean classes, the Rademacher complexity is bounded by a constant multiple of $\sqrt{(\log n)/n}$.

Definition 5.3. *The class of Boolean functions $\mathcal{F}$ is said to have **polynomial discrimination** if there exists a polynomial $\rho(\cdot)$ such that for every $n \geq 1$ and every set of n points $x_1,\ldots,x_n$ in $\mathcal{X}$, the cardinality of $\mathcal{F}(x_1,\ldots,x_n)$ is atmost $\rho(n)$.*

How does one check that a given Boolean class $\mathcal{F}$ has polynomial discrimination? The most popular way is via the *Vapnik Chervonenkis dimension* (or simply the VC dimension) of the class.

Definition 5.4 (VC dimension)**.** *The VC dimension of a class of Boolean functions $\mathcal{F}$ on $\mathcal{X}$ is defined as the maximum integer D for which there exists a finite subset $\{x_1,\ldots,x_D\}$ of $\mathcal{X}$ satisfying*

$$\mathcal{F}(x_1,\ldots,x_D) = \{0,1\}^D.$$

The VC dimension is taken to be ∞ if the above condition is satisfied for every integer D.

Definition 5.5 (Shattering)**.** *A finite subset $\{x_1,\ldots,x_m\}$ of $\mathcal{X}$ is said to be **shattered** by the Boolean class $\mathcal{F}$ if*

$$\mathcal{F}(x_1,\ldots,x_m) = \{0,1\}^m.$$

By convention, we extend the definition of shattering to empty subsets as well by saying that the empty set is shattered by every nonempty class $\mathcal{F}$.

It should be clear from the above pair of definitions that an alternative definition of VC dimension is: **The maximum cardinality of a finite subset of $\mathcal{X}$ that is shattered by $\mathcal{F}$.**

The link between VC dimension and polynomial discrimination comes via the following famous result, knows as the Sauer-Shelah lemma or the VC lemma.

Lemma 5.6 (Sauer-Shelah-Vapnik-Chevronenkis)**.** *Suppose that the VC dimension of a Boolean class $\mathcal{F}$ of functions on $\mathcal{X}$ is D. Then for every $n \geq 1$ and $x_1,\ldots,x_n \in \mathcal{X}$, we have*

$$|\mathcal{F}(x_1,\ldots,x_n)| \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D}.$$

Here $\binom{n}{k}$ is taken to be 0 if $n < k$. Moreoever, if $n \geq D$, then

$$|\mathcal{F}(x_1,\ldots,x_n)| \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D} \leq \left(\frac{en}{D}\right)^D.$$

Combining (33) with Lemma 5.6, we obtain the following bound on the control of Rademacher complexity and Expected suprema for Boolean classes with finite VC dimension.

Proposition 5.7. *Suppose $\mathcal{F}$ is a Boolean function class with VC dimension D. Then, for $n \geq D$, we have*

$$R_n(\mathcal{F}) \leq C\sqrt{\frac{D}{n}\log\left(\frac{en}{D}\right)}$$

and

$$\mathbb{E}\sup_{f\in\mathcal{F}}|P_nf - Pf| \leq C\sqrt{\frac{D}{n}\log\left(\frac{en}{D}\right)}.$$

Here C is a universal positive constant.

Remark 5.1. *It turns out that the logarithmic term is not needed in the bounds given by the above proposition. We shall see later that the bounds given by chaining do not have the superfluous logarithmic factor.*

We shall provide the proof of Lemma 5.6 in the next subsection. Before that, we give two examples of Boolean classes with finite VC dimension.

Example 5.8. *Let $\mathcal{V}$ be a D-dimensional vector space of real functions on $\mathcal{X}$. Let $\mathcal{F} := \{I(f \geq 0) : f \in \mathcal{V}\}$. The VC dimension of $\mathcal{F}$ is at most D.*

Proof. For any $D+1$ points $\{x_1, ..., x_{D+1}\}$, consider the set

$$T = \{(f(x_1), ..., f(x_{D+1}) : f \in \mathcal{V}\}.$$

Since $\mathcal{V}$ is a D-dimensional vector space, T is a linear subspace of $\mathbb{R}^{D+1}$ with dimension at most D. Therefore there exists $y \in \mathbb{R}^{D+1}$ and $y \neq 0$ such that y is orthogonal to the subspace T, i.e.

$$\sum_i y_i f(x_i) = 0 \text{ for all } f \in \mathcal{V}. \tag{34}$$

Without loss of generality, we can assume that there is an index k such that $y_k > 0$. Now suppose $\mathcal{F}$ shatters $\{x_1, ..., x_{D+1}\}$. Then there is $f \in \mathcal{V}$ satisfying

$$f(x_i) < 0 \text{ for all } i \text{ such that } y_i > 0;$$
$$f(x_i) \geq 0 \text{ for all } i \text{ such that } y_i \leq 0;$$

Then we have $\sum_i y_i f(x_i) = \sum_{i:y_i \leq 0} y_i f(x_i) + \sum_{i:y_i > 0} y_i f(x_i) < 0$, which is a contradiction to (34). Thus $\mathcal{F}$ cannot shatter $\{x_1, ..., x_{D+1}\}$ and so the VC dimension is at most D. □

Example 5.9. *Let $\mathcal{H}_k$ denote the indicators of all closed half-spaces in $\mathbb{R}^k$. The VC dimension of $\mathcal{H}_k$ is exactly equal to $k+1$.*

Proof. Left as a homework problem. □

6 Lecture 6

This lecture was delivered by Max Rabinovich. He made some changes to the notes (his modified notes are in the folder).

6.1 Proof of the Sauer-Shelah-Vapnik-Chevronenkis Lemma

This section contains the proof of Lemma 5.6. The proof uses an idea called *downshifting*.

Fix a Boolean class $\mathcal{F}$ with VC dimension D and also fix $n \geq 1$ and $x_1, \ldots, x_n$. To simplify notation, let us denote the set $\mathcal{F}(x_1, \ldots, x_n)$ by Δ. Observe first that the set Δ can be represented by a Boolean $n \times N$ matrix where $N := |\Delta|$. Indeed, every element of Δ is an element of $\{0,1\}^n$; so write this element as a column in a matrix; append all the columns corresponding to the different elements to form an $n \times N$ matrix all of whose columns are distinct. In the proof of Lemma 5.6, we will use both these representations of Δ: as a subset of $\{0,1\}^n$ and also as a Boolean $n \times N$ matrix.

For a subset S of $\{x_1, \ldots, x_n\}$, let Δ_S denote the $|S| \times N$ submatrix of Δ formed by taking only the rows of Δ corresponding to S. For example, if $S := \{x_1, x_5, x_8\}$, then Δ_S is the $3 \times N$ submatrix of Δ consisting of only the first, fifth and eighth rows of Δ.

Note that a subset S of $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{F}$ if and only if every element of $\{0,1\}^{|S|}$ appears as a column of Δ_S. Because the VC dimension of $\mathcal{F}$ is D, the number of subsets of $\{x_1, \ldots, x_n\}$ that can be shattered by $\mathcal{F}$ is clearly at most

$$\binom{n}{0} + \cdots + \binom{n}{D}.$$

We therefore have to show that the number of columns of Δ is atmost the number of subsets of $\{x_1, \ldots, x_n\}$ that are shattered by $\mathcal{F}$. We can isolate this into the following result which applies only to Boolean matrices.

Result 6.1. *Let Δ denote a $n \times N$ matrix with Boolean entries all of whose columns are distinct. Say that a subset S of $\{1, \ldots, n\}$ is shattered by Δ if every element of $\{0,1\}^{|S|}$ appears as a column of Δ_S (the empty set is always shattered). Let $\mathcal{S}(\Delta)$ denote the number of subsets of $\{1, \ldots, n\}$ that are shattered by Δ. Then*

$$N \leq \mathcal{S}(\Delta).$$

Proof of Result 6.1. The proof follows an idea called *downshifting*. Pick an arbitrary row of the matrix Δ, say the first row. Change each 1 in that row of Δ to 0 unless the change would create a column already present in Δ. This will create a new Boolean matrix, call it Δ', all of whose columns are distinct. This operation is called downshifting. I claim that

$$\mathcal{S}(\Delta') \leq \mathcal{S}(\Delta). \tag{35}$$

This claim is the key component of the proof. Once this is established, the rest of the proof is immediate.

To prove (35), it is enough to show that whenever a subset S of $\{1, \ldots, n\}$ is **not** shattered by Δ, it is not shattered by Δ' as well. This means that there will be fewer subsets shattered by Δ' compared to Δ which implies (35). So let us fix a subset S that is not shattered by Δ. S is a subset of the rows of Δ. If S does not contain the first row, then we have nothing to do because Δ' and Δ are identical in all rows except the first. So assume that $1 \in S$. In fact, assume, purely for notational simplicity, that $S = \{1, 2, 3, 4\}$.

Because S is not shattered by Δ, there exists an element $u \in \{0,1\}^4$ that is not present in Δ_S. If $u_1 = 1$, then it is clear that u is not present in Δ'_S as well because the downshifting operation which created Δ' from Δ cannot create new ones. So let us assume that $u_1 = 0$ and write $u = (1, v)$. The fact that u is not in Δ_S means that an element of the form $(0, v)$ is not present as a column in Δ_S. This would then mean that $(1, v)$ is not present in Δ'_S. If not, then Δ' would include a column of the form $(1, v, x)$. But then Δ would have to include the column $(1, v, x)$ as well. But if Δ did have $(1, v, x)$, then it would have been converted to $(0, v, x)$ by the downshifting operation because $(0, v, x)$ is not alredy present as a column in Δ to prevent this shifting. This completes the proof of (35).

Now, consider Δ'. Again, pick an arbitrary row and perform downshifting on Δ'. Repeat this procedure of picking an arbitrary row and performing downshifting until we get a matrix that cannot be altered by further downshifts. Call this matrix Δ^*. Repeated application of (35) will imply that $\mathcal{S}(\Delta^*) \leq \mathcal{S}(\Delta)$. The proof will now be completed by showing that $N \leq \mathcal{S}(\Delta^*)$. To see this, consider the first column of Δ^* and let S be the indices among $\{1, \ldots, n\}$ for which there is a 1 in the first column of S^*. I claim that S is shattered by Δ^*. To see this, assume that S is non-empty because empty sets are always shattered. For notational simplicity, assume that $S = \{1, 2\}$. We need to show that all elements in $\{0,1\}^2$ appear as columns in Δ^*_S. There are only four elements in $\{0,1\}^2$: $(1,1), (1,0), (0,1)$ and $(0,0)$. Obviously $(1,1)$ appears in the first column of Δ^*_S. $(1,0)$ should also appear somewhere because otherwise, it should be possible to alter Δ^* by downshifting. Similarly for $(0,1)$ and $(0,0)$. The proof is complete. □

As mentioned previously, Result 6.1 almost gives proves Lemma 5.6. The only thing remaining is to argue that

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D} \leq \left(\frac{en}{D}\right)^D \qquad \text{when } n \geq D.$$

To see this, let B have the Binomial distribution corresponding to n tosses with probability of success $1/2$. Then the left hand side above equals

$$2^n \mathbb{P}\{B \leq D\}$$

The function $I\{B \le D\}$ is bounded from above by $(D/n)^{B-D}$ which gives

$$2^n \mathbb{P}\{B \le D\} \le 2^n \mathbb{E}(D/n)^{B-D} = (D/n)^{-D}(1 + (D/n))^n \le \Big(\frac{en}{D}\Big)^D$$

because $1 + (D/n) \le e^{D/n}$. The proof of Lemma 5.6 is now complete.

6.2 Covering and Packing Numbers

As mentioned before, chaining gives much better bounds for $R_n(\mathcal{F}(x_1, \ldots, x_n))$ compared to the simple bound of Proposition 5.2. In order to discuss chaining, we need to be familiar with the notions of covering and packing numbers.

Let T be a set equipped with a pseudometric d. A pseudometric satisfies (a) $d(x, x) = 0$ for all $x \in T$, (b) $d(x, y) = d(y, x)$, and (c) $d(x, z) \le d(x, y) + d(y, z)$ for all x, y, z. If, in addition, it also satisfies $d(x, y) > 0$ for $x \ne y$, then $d(\cdot, \cdot)$ becomes a metric. We shall need to work with pseudometrics because the function:

$$(f, g) \mapsto \sqrt{\frac{1}{n}\sum_{i=1}^{n} (f(x_i) - g(x_i))^2}$$

is usually not a metric over $\mathcal{F}$ (because it only depends on the values of functions in $\mathcal{F}$ at $x_1, \ldots, x_n$). But this is a valid pseudometric.

Definition 6.2 (Covering Numbers)**.** *For a subset F of T and $\delta > 0$, the δ-covering number of F is denoted by $N_T(\delta, F, d)$ and is defined as the smallest number of closed δ-balls needed to cover F. In other words, $N_T(\delta, F, d)$ is the smallest N for which there exist points $t_1, \ldots, t_N \in T$ with $\min_{1 \le i \le N} d(t, t_i) \le \delta$ for each $t \in F$. The set of centers $\{t_i\}$ is called a δ-net for F. The logarithm of $N_T(\delta, F, d)$ is called the δ-metric entropy of F.*

Remark 6.1. *Note that the centers $t_1, \ldots, t_N$ are not constrained to be in F. This is related to the presence of the subscript T in the definition $N_T(\delta, F, d)$ of the covering numbers of F. If we regard F as a metric space in its own right, not just as a subset of T, then the covering numbers $N_F(\delta, F, d)$ might be larger because the centers t_i would then be forced to lie in F. It is an easy exercise to prove that $N_F(2\delta, F, d) \le N_T(\delta, F, d)$ and the extra factor of 2 would usually be of little consequence.*

If $N_T(\delta, T, d) < \infty$ for every $\delta > 0$, we say that T is totally bounded.

The notion of covering numbers is closely related to that of packing numbers which are defined next.

Definition 6.3. *For $\delta > 0$, the δ-packing number of F is defined as the largest N for which there exist points $t_1, \ldots, t_N \in F$ with $d(t_i, t_j) > \delta$ for every $i \ne j$ (these points $t_1, \ldots, t_N$ are said to be δ-separated). The δ-packing number will be denoted by $M(\delta, F, d)$. x*

The following result shows that covering and packing numbers are closely related to each other.

Lemma 6.4. *For every $\delta > 0$, we have*

$$N_F(\delta, F, d) \le M(\delta, F, d) \le N_T(\delta/2, F, d) \le N_F(\delta/2, F, d). \tag{36}$$

Proof. For the first inequality in (43), let $t_1, \ldots, t_M \in F$ be maximal set of δ-separated points in F with $M = M(\delta, F, d)$. Because of the maximality, every other point of F is within δ of one of the points $t_1, \ldots, t_M$. This means that $t_1, \ldots, t_M$ is a δ-net for F so that $N_F(\delta, F, d) \le M$ and this proves the first inequality in (43).

For the second inequality, again let $t_1, \ldots, t_M \in F$ be maximal set of δ-separated points in F with $M = M(\delta, F, d)$. Now if one tries to cover F by closed balls of radius $\delta/2$, it is clear that each ball can at

most contain one of the points $t_1, \ldots, t_M$. This because the distance between any two points t_i and t_j is strictly larger than δ while the diameter of a $\delta/2$ ball is at most δ. Therefore the number of closed $\delta/2$-balls required to cover F is at least M which proves the second inequality.

The third inequality is trivial. □

Below we see some examples where explicit bounds for covering/packing numbers are possible.

Proposition 6.5. *Suppose $\|\cdot\|$ denotes any norm in $\mathbb{R}^k$. For example, it might be the usual Euclidean norm or the ℓ_n norm, $\|x\|_1 := \sum_{i=1}^k |x_i|$. Let*

$$B_R := \{x \in \mathbb{R}^n : \|x\| \leq R\}.$$

Then, for every $\epsilon > 0$, we have

$$M(\epsilon R, B_R, d) \leq \left(1 + \frac{2}{\epsilon}\right)^k \tag{37}$$

where d denotes the metric corresponding to the norm $\|\cdot\|$.

Proof. Let $x_1, \ldots, x_N$ denote any set of points in B_R that is ϵR-separated i.e., $\|x_i - x_j\| > \epsilon R$ for all $i \neq j$. Then the closed balls

$$B(x_i, \epsilon R/2) := \{x \in \mathbb{R}^n : \|x - x_i\| \leq \epsilon R/2\}$$

for $i = 1, \ldots, N$ are disjoint. Moreover, all these balls $B(x_i, \epsilon R/2)$ are contained in $B_{R+\epsilon R/2}$ (the ball of radius $R + \epsilon R/2$ centered at the origin). As a result,

$$\sum_{i=1}^N \text{Vol}(B(x_i, \epsilon R/2)) \leq \text{Vol}(B_{R+\epsilon R/2})$$

where Vol denotes volume (Lebesgue measure). If we let Λ denote the volume of the unit ball B_1, then the above inequality becomes

$$\sum_{i=1}^N \left(\frac{\epsilon R}{2}\right)^k \Lambda \leq \left(R + \frac{\epsilon R}{2}\right)^k \Lambda$$

which immediately proves (37). □

The argument used above to prove (37) is known as the **volumetric** argument because it is based on a volume comparison.

We shall consider covering/packing numbers of some function classes. Loosely, function classes can be categorized into two groups: parametric classes and nonparametric classes. The ϵ-covering numbers of parametric classes will be of the order ϵ^{-k} for some integer k while the ϵ-covering numbers of nonparametric classes will be of the form $\exp(\sim \epsilon^{-k})$ for some k. This reflects the fact that the nonparametric classes will be much larger compared to parametric classes.

The following proposition gives an example of a parametric class of functions. Note that when D and $\|\Gamma\|_Q$ below are constants, the covering number bound given by the result below is of the form ϵ^{-k}.

Proposition 6.6. *Let $\Theta \subseteq \mathbb{R}^k$ be a non-empty bounded subset with Euclidean diameter D and let $\mathcal{F} := \{f_\theta : \theta \in \Theta\}$ be a class of functions on $\mathcal{X}$ indexed by Θ such that for some nonnegative function $\Gamma : \mathcal{X} \to \mathbb{R}$, we have*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \Gamma(x) \|\theta_1 - \theta_2\| \tag{38}$$

for all $x \in \mathcal{X}$ and $\theta_1, \theta_2 \in \Theta$. Here $\|\cdot\|$ denotes the usual Euclidean norm on $\mathbb{R}^k$.

Fix a probability measure Q on $\mathcal{X}$ and let d denote the pseudometric on $\mathcal{F}$ defined by

$$d(f,g) := \sqrt{\int_{\mathcal{X}} (f(x)-g(x))^2 \, dQ(x)}.$$

Then, for every $\epsilon > 0$,

$$M(\epsilon, \mathcal{F}, d) \leq \left(1 + \frac{2D \left\|\Gamma\right\|_Q}{\epsilon}\right)^k \qquad \text{where } \left\|\Gamma\right\|_Q := \left(\int |\Gamma(x)|^2 dQ(x)\right)^{1/2}$$

Proof. The condition (46) implies that for every $\theta_1, \theta_2 \in \Theta$, we have

$$d(f_{\theta_1}, f_{\theta_2}) \leq \|\theta_1 - \theta_2\| \left\|\Gamma\right\|_Q .$$

As a result, every ϵ-separated subset $\mathcal{F}$ in the metric d is automatically an $\epsilon / \left\|\Gamma\right\|_Q$ separated subset of Θ. Consequently

$$M(\epsilon, \mathcal{F}, d) \leq M\left(\frac{\epsilon}{\left\|\Gamma\right\|_Q}, \Theta, \|\cdot\|\right).$$

To bound the Euclidean packing number, we shall use the assumption that Θ has diameter $\leq D$ so that Θ is contained in $B(a, D) := \{x \in \mathbb{R}^k : \|x - a\| \leq D\}$ for every $a \in \Theta$. As a result

$$M\left(\frac{\epsilon}{\left\|\Gamma\right\|_Q}, \Theta, \|\cdot\|\right) \leq M\left(\frac{\epsilon}{\left\|\Gamma\right\|_Q}, B(a,D), \|\cdot\|\right).$$

To bound the right hand side above, we use Proposition 7.6 (note that we can $a = 0$ above because balls of the same radius will have the same packing numbers regardless of their center). This gives

$$M\left(\frac{\epsilon}{\left\|\Gamma\right\|_Q}, B(a,D), \|\cdot\|\right) \leq \left(1 + \frac{2D}{\epsilon} \left\|\Gamma\right\|_Q\right)^k$$

which finishes the proof of Proposition 7.8. □

The most standard examples of nonparametric function classes are smoothness classes. These will have covering numbers that are exponential in $1/\epsilon$. We shall first introduce smoothness classes and describe their covering numbers in one dimension and then generalize to multiple dimensions. For proofs of the covering number results, see Dudley [6, Chapter 8].

Fix $\alpha > 0$. Let β denote the largest integer that is **strictly** smaller than α. For example, if $\alpha = 5$, then $\beta = 4$ and if $\alpha = 5.2$, then $\beta = 5$.

The class $\mathcal{S}_\alpha$ is defined to consist of functions f on $[0,1]$ that satisfy all the following properties:

1. f is continuous on $[0,1]$.
2. f is differentiable β times on $(0,1)$.
3. $|f^{(k)}(x)| \leq 1$ for all $k = 0, \ldots, \beta$ and $x \in [0,1]$ where $f^{(0)}(x) := f(x)$.
4. $|f^{(\beta)}(x) - f^{(\beta)}(y)| \leq |x - y|^{\alpha - \beta}$ for all $x, y \in (0,1)$.

Let ρ denote the supremum metric on $\mathcal{S}_\alpha$ defined by

$$\rho(f,g) := \sup_{x \in [0,1]} |f(x) - g(x)|. \tag{39}$$

Theorem 6.7. *There exist positive constants ϵ_0, C_1 and C_2 denpending on α alone such that for all $\epsilon > 0$, we have*

$$\exp\left(C_2\epsilon^{-1/\alpha}\right) \leq M(\epsilon, \mathcal{S}_\alpha, \rho) \leq \exp\left(C_1\epsilon^{-1/\alpha}\right)$$

Thus the ϵ-metric entropy (logarithm of the ϵ-covering number) of the smoothness class $\mathcal{S}_\alpha$ in one dimension grows as $\epsilon^{-1/\alpha}$. Here α denotes the degree of smoothness (the higher α is, the smoother the functions in $\mathcal{S}_\alpha$). When $\alpha = 1$, the class $\mathcal{S}_\alpha$ consists of all bounded 1-Lipschitz functions on $[0, 1]$.

This result has a direct generalization to multidimensions. As before, $\alpha > 0$ and β is the largest integer that is strictly smaller than α.

For a vector $p = (p_1, \ldots, p_d)$ consisting of nonnegative integers $p_1, \ldots, p_d$, let $\langle p \rangle := p_1 + \cdots + p_d$. Let

$$D^p := \partial^{\langle p \rangle} / \partial x_1^{p_1} \ldots \partial x_d^{p_d}$$

The class $\mathcal{S}_{\alpha,d}$ is defined to consist of all functions f on $[0, 1]^d$ that satisfy:

1. f is continuous on $[0, 1]^d$.
2. All partial derivatives D^p of f exist on $(0, 1)^d$ for $\langle p \rangle \leq \beta$.
3. $|D^p(x)| \leq 1$ for all p with $\langle p \rangle \leq \beta$ and $x \in [0, 1]^d$.
4. $|D^p f(x) - D^p f(y)| \leq |x - y|^{\alpha-\beta}$ for all p with $\langle p \rangle = \beta$ and $x, y \in (0, 1)^d$.

Once again, we consider the supremum metric defined by $\rho(f, g) := \sup_{x\in[0,1]^d} |f(x) - g(x)|$.

Theorem 6.8. *There exist positive constants C_1 and C_2 depending only on α and the dimension d such that for all $\epsilon > 0$, we have*

$$\exp(C_2\epsilon^{-d/\alpha}) \leq M(\epsilon, \mathcal{S}_{\alpha,d}, \rho) \leq \exp(C_1\epsilon^{-d/\alpha}).$$

Thus the metric entropy of a smoothness class of functions with smoothness α and dimension d scales as $\epsilon^{-d/\alpha}$. This grows as d increases and goes down as α increases.

7 Lecture 7

The next main topic in the class is chaining. Before we go to chaining, we shall review the topics that were covered last week by Chi and Max.

We are discussing the problem of controlling:

$$\mathbb{E} \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1) \right| = \mathbb{E} \sup_{f\in\mathcal{F}} |P_n f - Pf| .$$

The symmetrization technique introduced last week allows us to bound the above as:

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

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables which are also independent of $X_1, \ldots, X_n$. The expectation on the right hand side above is with respect to both $\epsilon_1, \ldots, \epsilon_n$ and $X_1, \ldots, X_n$. To control the expectation on the right hand side above, one usually works conditionally on $X_1, \ldots, X_n$. The conditional expectation is then of the form

$$R_n(T) := \mathbb{E} \sup_{t\in T} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i t_i \right|$$

for a subset T of $\mathbb{R}^n$. $R_n(T)$ is easier to handle because the expectation is with respect to $\epsilon_1, \ldots, \epsilon_n$ which have a particularly simple distribution (i.i.d Rademachers).

In Lecture 5, we have seen the following elementary bound on $R_n(T)$.

Proposition 7.1. *Suppose T is a finite subset of $\mathbb{R}^n$ with cardinality $|T|$. Then*

$$R_n(T) = \mathbb{E} \max_{t \in T} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i t_i \right| \leq C \sqrt{\frac{\log(2|T|)}{n}} \left(\max_{t \in T} \sqrt{\frac{1}{n} \sum_{i=1}^{n} t_i^2} \right) \tag{40}$$

for a universal constant C.

The bound given by (40) has some shortcomings. It does not give anything when T is infinite. Even when T is finite, the bound is weak in some special situations. For example, when

$$T := \left\{ (I_{(-\infty,t]}(x_1), \ldots, I_{(-\infty,t]}(x_1)) : t \in \mathbb{R} \right\}$$

for some fixed points $x_1 < \cdots < x_n$ in $\mathbb{R}$, then it is easy to check that $|T| = n + 1$ so that the bound (40) gives

$$R_n(T) \leq C \sqrt{\frac{\log(n+1)}{n}}.$$

It turns out that for this particular T, the logarithmic term $\log(n+1)$ is redundant and that $R_n(T)$ is of the order $Cn^{-1/2}$. The bound on $R_n(T)$ derived from chaining will be of the form $C/n^{-1/2}$. The extra logarithmic factor is because of the inefficiency of (40).

In spite of these drawbacks, the bound (40) is important and crucially used for deriving the chaining bound. Before proceeding further, we shall provide a proof of Proposition 7.1. This proof will be slightly different from the way it was proved last week. We shall actually prove a stronger version of (40).

Proposition 7.2. *Let T be a finite set and let $\{X_t, t \in T\}$ be a stochastic process. Suppose that for every $t \in T$ and $u \geq 0$, the inequality*

$$\mathbb{P}\left\{ |X_t| \geq u \right\} \leq 2 \exp\left(\frac{-u^2}{2\Sigma^2} \right) \tag{41}$$

holds. Here Σ is a fixed positive real number. Then, for a universal positive constant C, we have

$$\mathbb{E} \max_{t \in T} |X_t| \leq C \Sigma \sqrt{\log(2|T|)}. \tag{42}$$

Remark 7.1. *Note that Proposition 7.2 is indeed a generalization of Proposition 7.1. This is because for $X_t := \sum_{i=1}^{n} \epsilon_i t_i$ (with $t \in T$), Hoeffding's inequality assures that (260) holds with*

$$\Sigma^2 := \max_{t \in T} \left(\sum_{i=1}^{n} t_i^2 \right).$$

Proposition 7.2 holds for every set of random variables X_t satisfying (260) so in addition to $X_t = \sum_{i=1}^{n} \epsilon_i t_i$, it also holds for $X_t \sim N(0, \sigma^2)$ with $\sigma \leq \Sigma$.

Proof of Proposition 7.2. Because

$$\mathbb{E} \max_{t \in T} |X_t| = \int_0^\infty \mathbb{P}\left\{ \max_{t \in T} |X_t| \geq u \right\} du,$$

we can control $\mathbb{E} \max_{t \in T} |X_t|$ by bounding the tail probability

$$\mathbb{P}\left\{ \max_{t \in T} |X_t| \geq u \right\} du$$

for every $u \geq 0$. For this, write

$$\mathbb{P}\left\{\max_{t\in T}|X_t| \geq u\right\} du = \mathbb{P}\left\{\cup_{t\in T}\{|X_t| \geq u\}\right\} \leq \sum_{t\in T}\mathbb{P}\left\{|X_t| \geq u\right\} \leq 2|T|\exp\left(\frac{-u^2}{2\Sigma^2}\right).$$

This bound is good for large u but not so good for small u (it is quite bad for $u = 0$ for example). It is therefore good to use it only for $u \geq u_0$ for some u_0 to be specified later. This gives

$$\begin{aligned}
\mathbb{E}\max_{t\in T}|X_t| &= \int_0^\infty \mathbb{P}\left\{\max_{t\in T}|X_t| \geq u\right\} du \\
&= \int_0^{u_0} \mathbb{P}\left\{\max_{t\in T}|X_t| \geq u\right\} du + \int_{u_0}^\infty \mathbb{P}\left\{\max_{t\in T}|X_t| \geq u\right\} du \\
&\leq u_0 + \int_{u_0}^\infty 2|T|\exp\left(\frac{-u^2}{2\Sigma^2}\right) du \\
&\leq u_0 + \int_{u_0}^\infty 2|T|\frac{u}{u_0}\exp\left(\frac{-u^2}{2\Sigma^2}\right) du = u_0 + \frac{2|T|}{u_0}\Sigma^2\exp\left(\frac{-u_0^2}{2\Sigma^2}\right).
\end{aligned}$$

One can try to minimize the above term over u_0. A simpler strategy is to realize that the large term here is $2|T|$ so one can choose u_0 to kill this term by setting

$$\exp\left(\frac{u_0^2}{2\Sigma^2}\right) = 2|T| \quad \text{or} \quad u_0 = \sqrt{2}\Sigma\sqrt{\log(2|T|)}.$$

This gives

$$\mathbb{E}\max_{t\in T}|X_t| \leq \sqrt{2}\Sigma\sqrt{\log(2|T|)} + \frac{\Sigma^2}{\sqrt{2\Sigma^2\log(2|T|)}} \leq C\Sigma\sqrt{\log(2|T|)}$$

which proves the result. □

It is not hard to construct examples where the bound given by Proposition 7.2 is loose. For example, it is loose when the tail bound (260) is loose for many $t \in T$ (this can happen for instance when $X_t \sim N(0, \sigma^2)$ for some σ^2 that is much smaller than Σ^2). It can also be loose when many of the $X_t's$ are close to each other: for instance, in the extreme case when $\max_{t\in T}|X_t| \approx X_{t_0}$ for a single $t_0 \in T$, the bound in (42) is loose by a factor of $\log|T|$.

However there exist examples where the bound in (42) is tight. The simplest example is the following. Suppose $X_t, t \in T$ are independently distributed as $N(0, \Sigma^2)$. Then it can be shown that

$$\mathbb{E}\max_{t\in T}|X_t| \geq c\Sigma\sqrt{\log(2|T|)}$$

for a positive constant c. Therefore, in this case, (42) is tight up to a constant factor. I will leave the proof of the above inequality as a homework exercise. This example means that Proposition 7.2 cannot be improved without additional assumptions on the process $\{X_t, t \in T\}$. Chaining gives improved bounds for $\mathbb{E}\max_{t\in T}|X_t|$ under an assumption on $\{X_t, t \in T\}$ that is different from (260). The assumption (260) pertains to the marginal distribution of each X_t but does not say anything about how close X_t is to another X_s etc. In contrast, for chaining, one assumes the existence of a metric d on T such that

$$\mathbb{P}\left\{|X_s - X_t| \geq u\right\} \leq 2\exp\left(\frac{-u^2}{2d^2(s,t)}\right).$$

Under this assumption, chaining provides a bound on $\mathbb{E}\max_{t\in T}|X_t|$ which involves the metric properties of (T, d). Before proceeding to chaining, let us recall the notions of covering and packing numbers of a metric space.

7.1 Review of Covering and Packing numbers

Let (T, d) be a metric or pseudometric space. Covering and packing numbers are defined as follows.

Definition 7.3 (Covering Numbers)**.** *For a subset F of T and $\delta > 0$, the δ-covering number of F is denoted by $N_T(\delta, F, d)$ and is defined as the smallest number of closed δ-balls needed to cover F. In other words, $N_T(\delta, F, d)$ is the smallest N for which there exist points $t_1, \ldots, t_N \in T$ with $\min_{1 \le i \le N} d(t, t_i) \le \delta$ for each $t \in F$. The set of centers $\{t_i\}$ is called a δ-net for F. The logarithm of $N_T(\delta, F, d)$ is called the δ-metric entropy of F.*

Remark 7.2. *Note that the centers $t_1, \ldots, t_N$ are not constrained to be in F. This is related to the presence of the subscript T in the definition $N_T(\delta, F, d)$ of the covering numbers of F. If we regard F as a metric space in its own right, not just as a subset of T, then the covering numbers $N_F(\delta, F, d)$ might be larger because the centers t_i would then be forced to lie in F. It is an easy exercise to prove that $N_F(2\delta, F, d) \le N_T(\delta, F, d)$ and the extra factor of 2 would usually be of little consequence.*

If $N_T(\delta, T, d) < \infty$ for every $\delta > 0$, we say that T is totally bounded.

The notion of covering numbers is closely related to that of packing numbers which are defined next.

Definition 7.4. *For $\delta > 0$, the δ-packing number of F is defined as the largest N for which there exist points $t_1, \ldots, t_N \in F$ with $d(t_i, t_j) > \delta$ for every $i \ne j$ (these points $t_1, \ldots, t_N$ are said to be δ-separated). The δ-packing number will be denoted by $M(\delta, F, d)$.*

Because of the following result (proved in last lecture), we shall treat covering and packing numbers as roughly the same.

Lemma 7.5. *For every $\delta > 0$, we have*

$$N_F(\delta, F, d) \le M(\delta, F, d) \le N_T(\delta/2, F, d) \le N_F(\delta/2, F, d). \tag{43}$$

Below we see some examples where explicit bounds for covering/packing numbers are possible. It is useful to be aware of these results.

7.1.1 Euclidean/Parametric Covering Numbers

Proposition 7.6. *For $R > 0$, let*

$$B(a, R) := \left\{ x \in \mathbb{R}^k : \|x - a\| \le R \right\}.$$

denote the ball of radius R centered at a point a. $\|\cdot\|$ here is the usual Euclidean norm. Then, for every $\epsilon > 0$, we have

$$M(\epsilon, B(a, R), d) \le \left(1 + \frac{2R}{\epsilon}\right)^k \tag{44}$$

and

$$N_{\mathbb{R}^k}(\epsilon, B(a, R), d) \ge \left(\frac{R}{\epsilon}\right)^k \tag{45}$$

where d denotes the usual Euclidean metric.

These bounds are simple to prove (proved in last lecture) and the proofs are based on volume comparison (as a result, these bounds are often referred to as volumetric bounds). The following is an immediate corollary of (44)

Corollary 7.7. *Suppose $S \subseteq \mathbb{R}^k$ is contained in some ball of radius R. Then*

$$M(\epsilon, S, d) \leq \left(1 + \frac{2R}{\epsilon}\right)^k.$$

with d denoting the usual Euclidean metric.

The above result implies that the covering numbers of bounded sets in $\mathbb{R}^k$ grow as ϵ^{-k}. Equivalently, the **metric entropy** of sets in $\mathbb{R}^k$ grows as $k \log(1/\epsilon)$. If k is constant, then the metric entropy grows logarithmically with $1/\epsilon$. The same conclusion can often be drawn for function classes that are indexed by a bounded set in $\mathbb{R}^k$ provided the mapping between the index and the function is smooth. The following proposition provides one way of making this precise.

Proposition 7.8. *Let $\Theta \subseteq \mathbb{R}^k$ be a non-empty bounded subset with Euclidean diameter D and let $\mathcal{F} := \{f_\theta : \theta \in \Theta\}$ be a class of functions on $\mathcal{X}$ indexed by Θ such that for some nonnegative function $\Gamma : \mathcal{X} \to \mathbb{R}$, we have*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \Gamma(x) \, \|\theta_1 - \theta_2\| \tag{46}$$

for all $x \in \mathcal{X}$ and $\theta_1, \theta_2 \in \Theta$. Here $\|\cdot\|$ denotes the usual Euclidean norm on $\mathbb{R}^k$.

Fix a probability measure Q on $\mathcal{X}$ and let d denote the pseudometric on $\mathcal{F}$ defined by

$$d(f, g) := \sqrt{\int_{\mathcal{X}} (f(x) - g(x))^2 \, dQ(x)}.$$

Then, for every $\epsilon > 0$,

$$M(\epsilon, \mathcal{F}, d) \leq \left(1 + \frac{2D \, \|\Gamma\|_Q}{\epsilon}\right)^k \qquad \text{where } \|\Gamma\|_Q := \left(\int |\Gamma(x)|^2 dQ(x)\right)^{1/2}$$

Function classes whose covering numbers grow as ϵ^{-k} (or whose metric entropy grows as $\log(1/\epsilon)$) will often be referred to as *parametric* or *Euclidean* or *finite-dimensional*.

7.1.2 Nonparametric Function Classes

Nonparametric function classes are much more massive in comparison to parametric classes in the sense that their metric entropy grows as a polynomial in $(1/\epsilon)$. Some standard examples of nonparametric functions classes are provided below.

7.1.3 One-dimensional smoothness classes

Fix $\alpha > 0$. Let β denote the largest integer that is **strictly** smaller than α. For example, if $\alpha = 5$, then $\beta = 4$ and if $\alpha = 5.2$, then $\beta = 5$.

The class $\mathcal{S}_\alpha$ is defined to consist of functions f on $[0, 1]$ that satisfy all the following properties:

1. f is continuous on $[0, 1]$.
2. f is differentiable β times on $(0, 1)$.
3. $|f^{(k)}(x)| \leq 1$ for all $k = 0, \ldots, \beta$ and $x \in [0, 1]$ where $f^{(0)}(x) := f(x)$.
4. $|f^{(\beta)}(x) - f^{(\beta)}(y)| \leq |x - y|^{\alpha - \beta}$ for all $x, y \in (0, 1)$.

Let ρ denote the supremum metric on $\mathcal{S}_\alpha$ defined by

$$\rho(f,g) := \sup_{x\in[0,1]} |f(x) - g(x)|. \tag{47}$$

The following result shows that the metric entropy of $\mathcal{S}_\alpha$ grows as $\epsilon^{-1/\alpha}$. Its proof can be found in Dudley [6, Chapter 8].

Theorem 7.9. *There exist positive constants ϵ_0, C_1 and C_2 denpending on α alone such that for all $\epsilon > 0$, we have*

$$\exp\left(C_2\epsilon^{-1/\alpha}\right) \le M(\epsilon, \mathcal{S}_\alpha, \rho) \le \exp\left(C_1\epsilon^{-1/\alpha}\right)$$

Thus the ϵ-metric entropy (logarithm of the ϵ-covering number) of the smoothness class $\mathcal{S}_\alpha$ in one dimension grows as $\epsilon^{-1/\alpha}$. Here α denotes the degree of smoothness (the higher α is, the smoother the functions in $\mathcal{S}_\alpha$). When $\alpha = 1$, the class $\mathcal{S}_\alpha$ consists of all bounded 1-Lipschitz functions on $[0,1]$.

7.1.4 One-dimensional Monotone Functions

Let $\mathcal{M}$ denote the class of all functions f on $[0,1]$ such that

1. f is nondecreasing on $[0,1]$
2. $|f(x)| \le 1$ for all $x \in [0,1]$.

For a probability measure Q on $[0,1]$, let ρ_Q denote the metric on $\mathcal{M}$ given by

$$\rho_Q(f,g) := \left(\int (f(x) - g(x))^2 dQ(x)\right)^{1/2}$$

Then it can be proved that

$$M(\epsilon, \mathcal{M}, \rho_Q) \le \exp\left(\frac{C}{\epsilon}\right)$$

for every probability measure Q on $[0,1]$. There exist probability measures Q for which a lower bound of $\exp(C_2/\epsilon)$ also holds on the packing number. Comparing this result with Theorem 7.9, it is clear that the covering numbers of $\mathcal{M}$ are comparable to the smoothness class $\mathcal{S}_1$ i.e., $\mathcal{S}_\alpha$ with $\alpha = 1$. Thus bounded monotone functions have the same metric entropy as bounded Lipschitz functions even though monotone functions need not be continuous.

7.1.5 Multidimensional smoothness classes

As in the one-dimensional case, let $\alpha > 0$ and β is the largest integer that is strictly smaller than α.

For a vector $p = (p_1, \ldots, p_d)$ consisting of nonnegative integers $p_1, \ldots, p_d$, let $\langle p\rangle := p_1 + \cdots + p_d$. Let

$$D^p := \partial^{\langle p\rangle} / \partial x_1^{p_1} \ldots \partial x_d^{p_d}$$

The class $\mathcal{S}_{\alpha,d}$ is defined to consist of all functions f on $[0,1]^d$ that satisfy:

1. f is continuous on $[0,1]^d$.
2. All partial derivatives D^p of f exist on $(0,1)^d$ for $\langle p\rangle \le \beta$.
3. $|D^p(x)| \le 1$ for all p with $\langle p\rangle \le \beta$ and $x \in [0,1]^d$.

4. $|D^p f(x) - D^p f(y)| \leq |x-y|^{\alpha-\beta}$ for all p with $\langle p \rangle = \beta$ and $x, y \in (0,1)^d$.

Once again, we consider the supremum metric defined by $\rho(f,g) := \sup_{x \in [0,1]^d} |f(x) - g(x)|$.

Theorem 7.10. *There exist positive constants C_1 and C_2 depending only on α and the dimension d such that for all $\epsilon > 0$, we have*

$$\exp(C_2 \epsilon^{-d/\alpha}) \leq M(\epsilon, \mathcal{S}_{\alpha,d}, \rho) \leq \exp(C_1 \epsilon^{-d/\alpha}).$$

Thus the metric entropy of a smoothness class of functions with smoothness α and dimension d scales as $\epsilon^{-d/\alpha}$. This grows as d increases and goes down as α increases.

7.1.6 Bounded Lipschitz Convex Functions

Let $\mathcal{C}$ denote the class of all functions f on $0,1]^d$ such that

1. f is convex on $[0,1]^d$.
2. $|f(x)| \leq 1$ for all $x \in [0,1]^d$
3. $|f(x) - f(y)| \leq \|x - y\|$.

It can then be showed that (ρ is the supremum metric on $[0,1]^d$):

$$\exp\left(C_2 \epsilon^{-d/2}\right) \leq M(\epsilon, \mathcal{C}, \rho) \leq \exp\left(C_1 \epsilon^{-d/2}\right)$$

where C_1 and C_2 depend on d alone. Comparing this to Theorem 7.10, it is clear that, in terms of metric entropy, $\mathcal{C}$ is comparable to the smoothness class $\mathcal{S}_{d,2}$. This is interesting because convex functions are not necessarily twice differentiable in the usual sense. Yet, they possess the regularity of second order smoothness in terms of metric entropy.

8 Lecture 8

8.1 Dudley's Metric Entropy Bound

The main goal for today is to state and prove Dudley's entropy bound for the suprema of subgaussian processes. The proof involves an idea called chaining. Before we start with chaining, let us recall the following basic result from last class.

Proposition 8.1. *Let T be a finite set and let $\{X_t, t \in T\}$ be a stochastic process. Suppose that for every $t \in T$ and $u \geq 0$, the inequality*

$$\mathbb{P}\{|X_t| \geq u\} \leq 2 \exp\left(\frac{-u^2}{2\Sigma^2}\right) \tag{48}$$

holds. Here Σ is a fixed positive real number. Then, for a universal positive constant C, we have

$$\mathbb{E} \max_{t \in T} |X_t| \leq C\Sigma\sqrt{\log(2|T|)}. \tag{49}$$

As we remarked in the last lecture, the bound (49) can be tight (up to a multiplicative constant) in some situations. For example, this is the case when $X_t, t \in T$ are i.i.d $N(0, \Sigma^2)$. Because of this example, Proposition 8.1 cannot be improved without imposing additional conditions on the process $\{X_t, t \in T\}$. It is also easy to construct examples where (49) is quite weak. For example, if $X_t = X_0 + \eta Z_t$ for some

$X_0 \sim N(0, \Sigma^2)$ and $Z_t, t \in T \sim^{i.i.d} N(0,1)$ and η is very very small, then it is clear that $\max_{t\in T} |X_t| \approx X_0$ so that (49) will be loose by a factor of $\log(2|T|)$. In order to improve on (49), we need to make assumptions on how *close* to each other the $X_t's$ are. Dudley's entropy bound makes such an assumption explicit and provides improved upper bounds for $\mathbb{E}\max_{t\in T} |X_t|$.

We shall first state Dudley's bound when the index set T is finite and subsequently improve it to the case when T is infinite.

Theorem 8.2 (Dudley's Metric Entropy Bound for finite T)**.** *Suppose (T, d) is a finite metric space and $\{X_t, t \in T\}$ is a stochastic process such that for every $s, t \in T$ and $u \geq 0$,*

$$\mathbb{P}\left\{|X_t - X_s| \geq u\right\} \leq 2\exp\left(\frac{-u^2}{2d^2(s,t)}\right). \tag{50}$$

Then, for a universal positive constant C, the following inequality holds for every $t_0 \in T$:

$$\mathbb{E}\max_{t\in T} |X_t - X_{t_0}| \leq C\int_0^\infty \sqrt{\log M(\epsilon, T, d)}d\epsilon. \tag{51}$$

The following remarks mention some alternative forms of writing the inequality (51) and also describe some implications.

1. Let D denote the diameter of the metric space T (i.e., $D = \max_{s,t\in T} d(s,t)$). Then the packing number $M(\epsilon, T, d)$ clearly equals 1 for $\epsilon \geq D$ (it is impossible to have two points in T whose distance is strictly larger than ϵ when $\epsilon > D$). Therefore

$$\int_0^\infty \sqrt{\log M(\epsilon, T, d)}d\epsilon = \int_0^D \sqrt{\log M(\epsilon, T, d)}d\epsilon.$$

   Moreover

$$\begin{aligned}\int_0^D \sqrt{\log M(\epsilon, T, d)}d\epsilon &= \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}d\epsilon + \int_{D/2}^D \sqrt{\log M(\epsilon, T, d)}d\epsilon \\ &\leq \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}d\epsilon + \int_0^{D/2} \sqrt{\log M(\epsilon + (D/2), T, d)}d\epsilon \\ &\leq 2\int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}d\epsilon\end{aligned}$$

   because $M(\epsilon + (D/2), T, d) \leq M(\epsilon, T, d)$ for every ϵ. We can thus state Dudley's bound as

$$\mathbb{E}\max_{t\in T} |X_t - X_{t_0}| \leq C\int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}d\epsilon$$

   where the C above equals twice the constant C in (51). Similarly, again by splitting the above integral in two parts (over 0 to $D/4$ and over $D/4$ to $D/2$), we can also state Dudley's bound as

$$\mathbb{E}\max_{t\in T} |X_t - X_{t_0}| \leq C\int_0^{D/4} \sqrt{\log M(\epsilon, T, d)}d\epsilon.$$

   The constant C above now is 4 times the constant in (51).

2. The left hand side in (51) is bounded from below (by triangle inequality) by $\mathbb{E}\max_{t\in T} |X_t| - \mathbb{E}|X_{t_0}|$. Thus, (51) implies that

$$\mathbb{E}\max_{t\in T} |X_t| \leq \mathbb{E}|X_{t_0}| + C\int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}d\epsilon \qquad \text{for every } t_0 \in T.$$

3. If $X_t, t \in T$ have mean zero and are jointly Gaussian , then $X_t - X_s$ is a mean zero normal random variable for every $s, t \in T$ so that (50) holds with

$$d(s,t) := \sqrt{\mathbb{E}(X_s - X_t)^2}.$$

4. The advantages of Theorem 8.2 over Proposition 8.1 is clear from the following example. Suppose $X_t, t \in T$ are given by

$$X_t = X_0 + \eta Z_t$$

for some positive but very very small η and $X_0 \sim N(0, \Sigma^2$ and $Z_t, t \in T \sim^{i.i.d} N(0,1)$. We have seen before that in this case $\mathbb{E} \max_{t \in T} |X_t|$ should behave like $C\Sigma$ but Proposition 8.1 will give an extra factor of $\log(2|T|)$. On the other hand, because

$$d(s,t) = \sqrt{\mathbb{E}(X_t - X_s)^2} \leq \eta\sqrt{2},$$

the packing number $M(\epsilon, T, d)$ will equal 1 for all but extremely small values of ϵ (say for $\epsilon > \epsilon_0$). Thus Dudley's bound will give $\Sigma + C(\log |T|)\epsilon_0$ which is much smaller than the bound given by Proposition 8.1 (because ϵ_0 is small).

We shall now give the proof of Theorem 8.2. The proof will be based on an idea called *chaining*. Specifically, we shall split $\max_{t \in T}(X_t - X_{t_0})$ in chains and use the bound given by Proposition 8.1 within the links of each chain.

Proof of Theorem 8.2. Recall that D is the diameter of T. For $n \geq 1$, let T_n be a maximal $D2^{-n}$-separated subset of T i.e., $\min_{s,t \in T_n : s \neq t} d(s,t) > D2^{-n}$ and T_n has maximal cardinality subject to the separation restriction. The cardinality of T_n is given by the packing number $M(D2^{-n}, T, d)$. Because of the maximality,

$$\max_{t \in T} \min_{s \in T_n} d(s,t) \leq D2^{-n}. \tag{52}$$

Because T is finite and $d(s,t) > 0$ for all $s \neq t$, the set T_n will equal T when n is large. Let

$$N := \min\{n \geq 1 : T_n = T\}.$$

For each $n \geq 1$, let $\pi_n : T \mapsto T_n$ denote the function which maps each point $t \in T$ to the point in T_n that is closest to T (if there are multiple closest points to T in T_n, then choose one arbitrarily). In other words, $\pi_n(t)$ is chosen so that

$$d(t, \pi_n(t)) = \min_{s \in T_n} d(t,s).$$

As a result, from (52), we have

$$d(t, \pi_n(t)) \leq D2^{-n} \qquad \text{for all } t \in T \text{ and } n \geq 1. \tag{53}$$

Note that $\pi_N(t) = t$. Finally let $T_0 := \{t_0\}$ and $\pi_0(t) = t_0$ for all $t \in T$.

We now note that

$$X_t - X_{t_0} = \sum_{n=1}^{N} \left(X_{\pi_n(t)} - X_{\pi_{n-1}(t)} \right) \qquad \text{for every } t \in T. \tag{54}$$

The sequence

$$t_0 \to \pi_1(t) \to \pi_2(t) \to \cdots \to \pi_{N-1}(t) \to \pi_N(t) = t$$

can be viewed as a chain from t_0 to t. This is what gives the argument the name *chaining*.

By (54), we obtain

$$\max_{t \in T} |X_t - X_{t_0}| \leq \max_{t \in T} \sum_{n=1}^{N} |X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \leq \sum_{n=1}^{N} \max_{t \in T} |X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|$$

so that

$$\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| \le \sum_{n=1}^{N} \mathbb{E}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|. \tag{55}$$

Now to bound $\mathbb{E}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|$ for each $1 \le n \le N$, we shall use the elementary bound given by Proposition 8.1. For this, note first that by (50), we have

$$\mathbb{P}\left\{|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \ge u\right\} \le 2\exp\left(\frac{-u^2}{2d^2(\pi_n(t), \pi_{n-1}(t))}\right).$$

Now

$$d(\pi_n(t), \pi_{n-1}(t)) \le d(\pi_n(t), t) + d(\pi_{n-1}(t), t) \le D2^{-n} + D2^{-(n-1)} = 3D2^{-n}.$$

Thus Proposition 8.1 can be applied with $\Sigma := 3D2^{-n}$ so that we obtain (**note that the value of C might change from occurrence to occurrence**)

$$\begin{aligned}\mathbb{E}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| &\le C\frac{3D}{2^n}\sqrt{\log\left(2|T_n||T_{n-1}|\right)} \\ &\le CD2^{-n}\sqrt{\log\left(2|T_n|^2\right)} \le CD2^{-n}\sqrt{\log\left(2M(D2^{-n}, T, d)\right)}\end{aligned}$$

Plugging the above bound into (55), we deduce

$$\begin{aligned}\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| &\le C\sum_{n=1}^{N}\frac{D}{2^n}\sqrt{\log\left(2M(D2^{-n}, T, d)\right)} \\ &\le 2C\sum_{n=1}^{N}\int_{D/(2^{n+1})}^{D/(2^n)}\sqrt{\log(2M(\epsilon, T, d))}d\epsilon \\ &= C\int_{D/(2^{N+1})}^{D/2}\sqrt{\log(2M(\epsilon, T, d))}d\epsilon \\ &\le C\int_{0}^{D/2}\sqrt{\log(2M(\epsilon, T, d))}d\epsilon \\ &= C\int_{0}^{D/4}\sqrt{\log(2M(\epsilon, T, d))}d\epsilon + C\int_{0}^{D/4}\sqrt{\log(2M(\epsilon + (D/4), T, d))}d\epsilon \\ &\le 2C\int_{0}^{D/4}\sqrt{\log(2M(\epsilon, T, d))}d\epsilon.\end{aligned}$$

Note now that for $\epsilon \le D/4$, the packing number $M(\epsilon, T, d) \ge 2$ so that

$$\log(2M(\epsilon, T, d)) \le \log 2 + \log M(\epsilon, T, d) \le 2\log M(\epsilon, T, d).$$

We have thus proved that

$$\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| \le 2\sqrt{2}C\int_{0}^{D/4}\sqrt{\log M(\epsilon, T, d)}d\epsilon$$

which proves (51). □

8.2 Dudley's bound for infinite T

We shall next prove Dudley's bound for the case of infinite T. This requires a technical assumption called *separability* which will always be satisfied in our applications.

Definition 8.3 (Separable Stochastic Process)**.** *Let (T, d) be a metric space. The stochastic process $\{X_t, t \in T\}$ indexed by T is said to be separable if there exists a null set N and a countable subset $\tilde{T}$ of T such that for all $\omega \notin N$ and $t \in T$, there exists a sequence $\{t_n\}$ in $\tilde{T}$ with $\lim_{n\to\infty} d(t_n, t) = 0$ and $\lim_{n\to\infty} X_{t_n}(\omega) = X_t(\omega)$.*

Note that the definition of separability requires that $\tilde{T}$ is a dense subset of T which means that the metric space (T, d) is separable (a metric space is said to be separable if it has a countable dense subset).

The following fact is easy to check: **If (T, d) is a separable metric space and if $X_t, t \in T$ has continuous sample paths (almost surely), then $X_t, t \in T$ is separable**. The statement that $X_t, t \in T$ has continuous sample paths (almost surely) means that there exists a null set N such that for all $\omega \notin N$, the function $t \mapsto X_t(\omega)$ is continous on T.

The following fact is also easy to check: If $\{X_t, t \in T\}$ is a separable stochastic process, then

$$\sup_{t \in T} |X_t - X_{t_0}| = \sup_{t \in \tilde{T}} |X_t - X_{t_0}| \qquad \text{almost surely} \tag{56}$$

for every $t_0 \in T$. Here $\tilde{T}$ is a countable subset of T which appears in the definition of separability of $X_t, t \in T$.

In particular, the statement (56) implies that $\sup_{t \in T} |X_t - X_{t_0}|$ is measurable (note that uncountable suprema are in general not guaranteed to be measurable; but this is not an issue for separable processes).

We shall now state Dudley's theorem for separable processes. This theorem does not impose any cardinality restrictions on T (it holds for both finite and infinite T).

Theorem 8.4. *Let (T, d) be a separable metric space and let $\{X_t, t \in T\}$ be a separable stochastic process. Suppose that for every $s, t \in T$ and $u \geq 0$, we have*

$$\mathbb{P}\left\{|X_s - X_t| \geq u\right\} \leq 2 \exp\left(\frac{-u^2}{2d^2(s,t)}\right).$$

Then for every $t_0 \in T$, we have

$$\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}| \leq C \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)} d\epsilon$$

where D is the diameter of the metric space (T, d).

Proof of Theorem 8.4. Let $\tilde{T}$ be a countable subset of T such that (56) holds. We may assume that $\tilde{T}$ contains t_0 (otherwise simply add t_0 to $\tilde{T}$). For each $k \geq 1$, let $\tilde{T}_k$ be the finite set obtained by taking the first k elements of $\tilde{T}$ (in an arbitrary enumeration of the entries of $\tilde{T}$). We can ensure that $\tilde{T}_k$ contains t_0 for every $k \geq 1$.

Applying the finite index set version of Dudley's theorem (Theorem 8.2) to $\{X_t, t \in \tilde{T}_k\}$, we obtain

$$\mathbb{E} \max_{t \in \tilde{T}_k} |X_t - X_{t_0}| \leq C \int_0^{diam(\tilde{T}_k)/2} \sqrt{\log M(\epsilon, \tilde{T}_k, d)} d\epsilon \leq C \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)} d\epsilon.$$

Note that the right hand side does not depend on k. Letting $k \to \infty$ on the left hand side, we use the Monotone Convergence Theorem to obtain

$$\sup_{t \in \tilde{T}} |X_t - X_{t_0}| \leq C \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)} d\epsilon.$$

The proof is now completed by (56). □

8.3 Application of Dudley's Bound to Rademacher Averages

Suppose $T \subseteq \mathbb{R}^n$ and consider the stochastic process $X_t, t \in T$ given by

$$X_t := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i t_i$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d Rademacher random variables.

Let us define the following norm on $\mathbb{R}^n$:

$$\|t\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n t_i^2}.$$

In other words, $\|t\|_n$ is the usual Euclidean norm of t divided by $\sqrt{n}$. Also let $d_n(s,t) := \|s-t\|_n$ be the corresponding metric on $\mathbb{R}^n$.

By Hoeffding's inequality, for every $u \geq 0$,

$$\mathbb{P}\left\{|X_t - X_s| \geq u\right\} \leq 2\exp\left(\frac{-nu^2}{2\sum_{i=1}^n (s_i - t_i)^2}\right) = 2\exp\left(\frac{-u^2}{2\|s-t\|_n^2}\right) = 2\exp\left(\frac{-u^2}{2d_n^2(s,t)}\right)$$

so that $X_t, t \in T$ satisfies the assumptions in Dudley's theorems with the metric d_n. Also note that $T = \mathbb{R}^n$ is trivially separable and that the map

$$t \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i t_i$$

is linear (and hence continuous in t). This means that $X_t, t \in T$ is separable. We can therefore apply Dudley's theorem. We apply Theorem 8.4 with $t_0 = (0, \ldots, 0)$ (since this vector may not be contained in T, we shall apply Theorem 8.4 to $T \cup \{0\}$) to obtain

$$\mathbb{E}\sup_{t\in T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i t_i\right| \leq C\int_0^{diam(T\cup\{0\})/2} \sqrt{\log M(\epsilon, T\cup\{0\}, d_n)}d\epsilon$$

where the diameter and packing numbers above are with respect to the d_n metric. It is now easy to see that

$$diam(T\cup\{0\}) = \sup_{s,t\in T\cup\{0\}} \|s-t\|_n \leq 2\sigma_n \qquad \text{where } \sigma_n := \sup_{t\in T}\|t\|_n.$$

We thus obtain the following upper bound:

$$\mathbb{E}\sup_{t\in T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i t_i\right| \leq C\int_0^{\sigma_n} \sqrt{\log M(\epsilon, T\cup\{0\}, d_n)}d\epsilon$$

In the next class, we shall combine the above bound with the technique of symmetrization which will give us an important upper bound on the suprema of empirical processes.

9 Lecture 9

Let us start by recalling Dudley's entropy bound from the last class: Suppose (T,d) is a metric space and $\{X_t, t \in \mathcal{T}\}$ is a separable stochastic process satisfying

$$\mathbb{P}\left\{|X_s - X_t| \geq u\right\} \leq 2\exp\left(\frac{-u^2}{2d^2(s,t)}\right) \qquad \text{for all } u \geq 0, s \in T, t \in T.$$

Then for every $t_0 \in T$, we have

$$\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}| \leq C\int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}d\epsilon$$

where D denotes the diameter of the metric space (T,d).

We applied this bound to control the expected suprema of Rademacher averages. Suppose T is a subset of $\mathbb{R}^n$. Then

$$\mathbb{E}\sup_{t\in\mathbb{R}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i t_i\right| \leq C\int_0^{\sigma_n}\sqrt{\log M(\epsilon, T\cup\{0\}, d_n)}d\epsilon \tag{57}$$

where

$$\sigma_n := \sup_{t\in T}\|t\|_n, \quad \|t\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}t_i^2}, \quad \text{and} \quad d_n(s,t) := \|s-t\|_n.$$

Note that if T is finite, then

$$\log M(\epsilon, T\cup\{0\}, d_n) \leq 1 + \log|T|$$

and hence the bound (57) implies that

$$\mathbb{E}\max_{t\in T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i t_i\right| \leq C\sqrt{\log(2|T|)}\max_{t\in T}\sqrt{\frac{1}{n}\sum_{i=1}^{n}t_i^2}.$$

Note that we had proved the above bound earlier by the elementary bound on the expected maxima of subgaussian random variables.

We shall now apply (57) together with symmetrization to obtain our main bound for the Expected suprema of an empirical process.

9.1 Main Bound on the Expected Suprema of Empirical Processes

Consider the usual Empirical Process setup. Our goal is to obtain upper bounds on Δ where

$$\Delta := \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X_1))\right| = \mathbb{E}\sup_{f\in\mathcal{F}}\left\{\sqrt{n}|P_nf - Pf|\right\}$$

Note the presence of the $\sqrt{n}$ in the supremums above. We have seen that symmetrization gives

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

We write the expectation in two parts conditioning on $X_1,\ldots,X_n$ to get

$$\Delta \leq 2\mathbb{E}\left[\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\,\middle|\,X_1,\ldots,X_n\right)\right].$$

The inner expectation above can be controlled via the bound (57). This gives

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\,\middle|\,X_1,\ldots,X_n\right) \leq C\int_0^{\sigma_n}\sqrt{\log M(\epsilon, \mathcal{F}(X_1,\ldots,X_n)\cup\{0\}, d_n)}d\epsilon.$$

where $\mathcal{F}(X_1,\ldots,X_n) := \{(f(X_1),\ldots,f(X_n)) : f\in\mathcal{F}\}$ is a subset of $\mathbb{R}^n$,

$$\sigma_n := \sup_{f\in\mathcal{F}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}f^2(X_i)} = \sup_{f\in\mathcal{F}}\sqrt{P_nf^2}$$

and d_n is the Euclidean metric on $\mathbb{R}^n$ scaled by $\sqrt{n}$.

We shall now write

$$M(\epsilon, \mathcal{F}(X_1, \ldots, X_n) \cup \{0\}, d_n) = M(\epsilon, \mathcal{F} \cup \{0\}, L^2(P_n))$$

where $L^2(P_n)$ refers to the pseudometric on $\mathcal{F}$ given by

$$(f, g) \mapsto \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f(X_i) - g(X_i))^2}.$$

Note the trivial inequality

$$M(\epsilon, \mathcal{F} \cup \{0\}, L^2(P_n)) \leq 1 + M(\epsilon, \mathcal{F}, L^2(P_n))$$

We thus obtain

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \middle| X_1, \ldots, X_n \right) \leq C \int_0^{\sup_{f \in \mathcal{F}} \sqrt{P_n f^2}} \sqrt{1 + \log M(\epsilon, \mathcal{F}, L^2(P_n))} d\epsilon.$$

Taking expectations, we obtain

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sqrt{n} |P_n f - P f| \right) \leq C \mathbb{E} \left[\int_0^{\sup_{f \in \mathcal{F}} \sqrt{P_n f^2}} \sqrt{1 + \log M(\epsilon, \mathcal{F}, L^2(P_n))} d\epsilon \right].$$

This is our first bound on the expected supremum of an empirical process. We can simplify this bound further using *envelopes*. We say that a nonnegative valued function $F : \mathcal{X} \to [0, \infty)$ is an envelope for the class $\mathcal{F}$ if

$$\sup_{f \in \mathcal{F}} |f(x)| \leq F(x) \qquad \text{for every } x \in \mathcal{X}.$$

It is clear then that $\sup_{f \in \mathcal{F}} \sqrt{P_n f^2} \leq \sqrt{P_n F^2}$ so that

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sqrt{n} |P_n f - P f| \right) &\leq C \mathbb{E} \left[\int_0^{\sup_{f \in \mathcal{F}} \sqrt{P_n f^2}} \sqrt{1 + \log M(\epsilon, \mathcal{F}, L^2(P_n))} d\epsilon \right] \\
&\leq C \mathbb{E} \left[\int_0^{\sqrt{P_n F^2}} \sqrt{1 + \log M(\epsilon, \mathcal{F}, L^2(P_n))} d\epsilon \right] \\
&= C \mathbb{E} \left[\sqrt{P_n F^2} \int_0^1 \sqrt{1 + \log M(\epsilon \sqrt{P_n F^2}, \mathcal{F}, L^2(P_n))} d\epsilon \right] \\
&\leq C \mathbb{E} \left[\sqrt{P_n F^2} \int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \sqrt{Q F^2}, \mathcal{F}, L^2(Q))} d\epsilon \right] \\
&= C \left[\int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \sqrt{Q F^2}, \mathcal{F}, L^2(Q))} d\epsilon \right] \mathbb{E} \sqrt{P_n F^2} \\
&\leq C \left[\int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \sqrt{Q F^2}, \mathcal{F}, L^2(Q))} d\epsilon \right] \sqrt{\mathbb{E} P_n F^2} \\
&= C \left[\int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \sqrt{Q F^2}, \mathcal{F}, L^2(Q))} d\epsilon \right] \sqrt{P F^2}.
\end{aligned}$$

In the above chain of inequalities, the supremum is over all probability measures Q supported on a set of cardinality at most n in $\mathcal{X}$. Also PF^2 stands for $\mathbb{E}F^2(X_1)$.

We have therefore proved the following result.

Theorem 9.1. *Let F be an envelope for the class $\mathcal{F}$ such that $PF^2 < \infty$. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sqrt{n} |P_n f - Pf| \right) \leq C \|F\|_{L^2(P)} J(F, \mathcal{F})$$

where

$$J(F, \mathcal{F}) := \int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \sqrt{QF^2}, \mathcal{F}, L^2(Q))} d\epsilon.$$

9.2 Application to Boolean Function Classes with finite VC dimension

Let $\mathcal{F}$ be a Boolean function class with finite VC dimension and let D denote its VC dimension. Recall that the VC dimension is defined as the maximum cardinality of a set in $\mathcal{X}$ that is shattered by the class $\mathcal{F}$. An important fact about the VC dimension is the Sauer-Shelah-Vapnik-Chervonenkis lemma which states that

$$|\mathcal{F}(x_1, \ldots, x_n)| \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D}. \tag{58}$$

for every $n \geq 1$ and $x_1, \ldots, x_n \in \mathcal{X}$ where

$$\mathcal{F}(x_1, \ldots, x_n) := \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}.$$

Note that $\binom{n}{k}$ in (58) is taken to be 0 if $n < k$. The right hand of (58) equals 2^D if $n \leq D$ and is bounded from above by $(en/D)^D$ if $n \geq D$.

We have seen previously that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sqrt{n} |P_n f - Pf| \right) \leq C \sqrt{D \log \left(\frac{en}{D} \right)} \qquad \text{for } n \geq D \tag{59}$$

and this bound was proved by symmetrization and the elementary bound on the Rademacher averages. This elementary bound involved the cardinality of $\mathcal{F}(X_1, \ldots, X_n)$ which we bounded via (58).

It turns out however that the logarithmic factor is redundant in (59) and one actually has the bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sqrt{n} |P_n f - Pf| \right) \leq C \sqrt{D}. \tag{60}$$

This can be deduced as a consequence of Theorem 9.1 as we shall demonstrate in this section. Since Theorem 9.1 gives bounds in terms of packing numbers, it becomes necessary to relate the packing numbers of $\mathcal{F}$ to its VC dimension. This is done in the following important result due to Dudley.

Theorem 9.2. *Suppose $\mathcal{F}$ is a Boolean function class with VC dimension D. Then*

$$\sup_Q M(\epsilon, \mathcal{F}, L^2(Q)) \leq \left(\frac{c_1}{\epsilon} \right)^{c_2 D} \qquad \text{for all } 0 < \epsilon \leq 1. \tag{61}$$

Here c_1 and c_2 are universal positive constants and the supremum is over all probability measures Q on $\mathcal{X}$.

Note that Theorem 9.2 gives upper bounds for the ϵ-packing numbers when $\epsilon \leq 1$. Since the functions in $\mathcal{F}$ take only the two values 0 and 1, it is clear that $M(\epsilon, \mathcal{F}, L^2(Q)) = 1$ for all $\epsilon \geq 1$.

Proof of Theorem 9.2. Fix $0 < \epsilon \leq 1$ and a probability measure Q on $\mathcal{X}$. Let $N = M(\epsilon, \mathcal{F}, L^2(Q))$ and let $f_1, \ldots, f_N$ be a maximal ϵ-separated subset of $\mathcal{F}$ in the $L^2(Q)$ metric. This means therefore that for every $1 \leq i \neq j \leq N$, we have

$$\epsilon^2 < \int (f_i - f_j)^2 \, dQ = \int I\{f_i \neq f_j\} dQ = Q\{f_i \neq f_j\}.$$

Now let $Z_1, Z_2, \ldots$ be i.i.d observations from Q. By the above, we have

$$\mathbb{P}\{f_i(Z_1) = f_j(Z_1)\} = 1 - Q\{f_i \neq f_j\} < 1 - \epsilon^2 \leq e^{-\epsilon^2}.$$

By the independence of $Z_1, Z_2, \ldots$, we deduce then that for every $k \geq 1$,

$$\mathbb{P}\{f_i(Z_1) = f_j(Z_1), f_i(Z_2) = f_j(Z_2), \ldots, f_i(Z_k) = f_j(Z_k)\} \leq e^{-k\epsilon^2}.$$

In words, this means that the probability that f_i and f_j agree on every $Z_1, \ldots Z_k$ is at most $e^{-k\epsilon^2}$. By the union bound, we have

$$\mathbb{P}\{(f_i(Z_1), \ldots, f_i(Z_k)) = (f_j(Z_1), \ldots, f_j(Z_k)) \text{ for some } 1 \leq i < j \leq N\} \leq \binom{N}{2} e^{-k\epsilon^2} \leq \frac{N^2}{2} e^{-k\epsilon^2}.$$

This immediately gives

$$\mathbb{P}\{|\mathcal{F}(Z_1, \ldots, Z_k)| \geq N\} \geq 1 - \frac{N^2}{2} e^{-k\epsilon^2}.$$

Thus if we take

$$k = \left\lceil \frac{2 \log N}{\epsilon^2} \right\rceil \geq \frac{2 \log N}{\epsilon^2}, \tag{62}$$

then

$$\mathbb{P}\{|\mathcal{F}(Z_1, \ldots, Z_k)| \geq N\} \geq \frac{1}{2}.$$

Thus for the choice (62) of k, there exists a subset $\{z_1, \ldots, z_k\}$ of cardinality k such that

$$N \leq |\mathcal{F}(z_1, \ldots, z_k)|$$

We now apply the Sauer-Shelah-VC lemma and deduce that

$$N \leq \binom{k}{0} + \binom{k}{1} + \cdots + \binom{k}{D}. \tag{63}$$

We now split into two cases depending on whether $k \leq D$ or $k \geq D$.

Case 1: $k \leq D$: Here (63) gives

$$M(\epsilon, \mathcal{F}, L^2(Q)) = N \leq 2^D \leq \left(\frac{2}{\epsilon}\right)^D$$

which proves (61).

Case 2: $k \geq D$: Here (63) gives

$$N \leq \left(\frac{ek}{D}\right)^D$$

so that (using (62))

$$N^{1/D} \leq \frac{ek}{D} \leq \frac{4e}{D\epsilon^2} \log N = \frac{8e}{\epsilon^2} \log N^{1/(2D)} \leq \frac{8e}{\epsilon^2} N^{1/(2D)}$$

where we have used $\log x \leq x$. This immediately gives

$$M(\epsilon, \mathcal{F}, L^2(Q)) = N \leq \left(\frac{8e}{\epsilon^2}\right)^{2D}.$$

The proof of Theorem 9.2 is complete. □

The bound (60) immediately follows from Theorem 9.1 and Theorem 9.2 as shown below.

Theorem 9.3. *Suppose $\mathcal{F}$ is a Boolean class of functions with VC dimension D, then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}|P_nf - Pf| \leq C\sqrt{\frac{D}{n}}. \tag{64}$$

Proof. Since $\mathcal{F}$ is a Boolean class, we can apply Theorem 9.1 with $F(x) = 1$ for all x. This gives

$$\mathbb{E}\sup_{f\in\mathcal{F}}|P_nf - Pf| \leq \frac{C}{\sqrt{n}}J(1,\mathcal{F}) \qquad \text{with } J(1,\mathcal{F}) := \int_0^1 \sqrt{1+\log\sup_Q M(\epsilon,\mathcal{F},L^2(Q))}d\epsilon.$$

The packing numbers above can be bounded by Theorem 9.2 which gives

$$\begin{aligned}
J(1,\mathcal{F}) &:= \int_0^1 \sqrt{1+\log\sup_Q M(\epsilon,\mathcal{F},L^2(Q))}d\epsilon \\
&\leq \int_0^1 \sqrt{1+2D\log\frac{8e}{\epsilon^2}}d\epsilon \\
&\leq 1+\int_0^1 \sqrt{2D\log\frac{8e}{\epsilon^2}}d\epsilon \;\; \text{because } \sqrt{a+b}\leq\sqrt{a}+\sqrt{b} \\
&= 1+\sqrt{D}\int_0^1 \sqrt{2\log\frac{8e}{\epsilon^2}}d\epsilon \leq C\sqrt{D}.
\end{aligned}$$

This completes the proof of Theorem 9.3. □

The following are immediate applications of Theorem 9.3.

Example 9.4. *Suppose $X_1,\ldots,X_n$ are i.i.d real valued observations having a common cdf F. Let F_n denote the empirical cdf of the data $X_1,\ldots,X_n$. Then Theorem 9.3 immediately gives*

$$\mathbb{E}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \leq \frac{C}{\sqrt{n}}. \tag{65}$$

This is because the Boolean class $\mathcal{F} := \{I_{(-\infty,x]} : x\in\mathbb{R}\}$ has VC dimension 1.

One can also obtain a high probability upper bound on $\sup_x|F_n(x)-F(x)|$ using the Bounded Differences concentration inequality that we discussed previously. This (together with (65)) gives

$$\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \leq \frac{C}{\sqrt{n}} + \sqrt{\frac{2}{n}\log\frac{1}{\alpha}} \qquad \textit{with probability} \geq 1-\alpha.$$

Example 9.5 (Classification with VC Classes)**.** *Consider the classification problem where we observe i.i.d data $(X_1,Y_1),\ldots,(X_n,Y_n)$ with $X_i\in\mathcal{X}$ and $Y_i\in\{0,1\}$. Let $\mathcal{C}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ (these are classifiers). For a classifier g, we define its test error and training error by*

$$L(g) = \mathbb{P}\{g(X_1)\neq Y_1\} \quad \textit{and} \quad L_n(g) := \frac{1}{n}\sum_{i=1}^n I\{g(X_i)\neq Y_i\}$$

respectively. The ERM (Empirical Risk Minimizer) classifier is given by

$$\hat{g}_n := \operatorname*{argmin}_{g\in\mathcal{C}} L_n(g).$$

It is usually of interest to understand the test error of $\hat{g}_n$ relative to the best test error in the class $\mathcal{C}$ i.e.,

$$L(\hat{g}_n) - \inf_{g\in\mathcal{C}} L(g).$$

If g^* minimizes $L(g)$ over $g \in \mathcal{C}$, then we can bound the discrepancy above as

$$\begin{aligned} L(\hat{g}_n) - L(g^*) &= L_n(\hat{g}_n) - L_n(g^*) + L(\hat{g}_n) - L_n(\hat{g}_n) + L(g^*) - L_n(g^*) \\ &\leq L(\hat{g}_n) - L_n(\hat{g}_n) + L(g^*) - L_n(g^*) \\ &\leq 2 \sup_{g \in \mathcal{C}} |L_n(g) - L(g)|. \end{aligned}$$

The last inequality above can be quite loose (we shall look at improved bounds later). The term above can be written as $\sup_{f \in \mathcal{F}} |P_n f - Pf|$ where

$$\mathcal{F} := \{(x, y) \mapsto I\{g(x) \neq y\} : g \in \mathcal{C}\},$$

P_n is the empirical distribution of $(X_i, Y_i), i = 1, \ldots, n$ and P is the distribution of (X_1, Y_1).

Using the bounded differences concentration inequality and the bound given by Theorem 9.3, we obtain (for every $\alpha \in (0, 1)$)

$$L(\hat{g}_n) - L(g^*) \leq C\sqrt{\frac{VC(\mathcal{F})}{n}} + \sqrt{\frac{8}{n} \log \frac{1}{\alpha}} \tag{66}$$

with probability $\geq 1 - \alpha$.

It can now be shown that $VC(\mathcal{F}) \leq VC(\mathcal{C})$. To see this, it is enough to argue that if $\mathcal{F}$ can shatter $(x_1, y_1), \ldots, (x_n, y_n)$, then $\mathcal{C}$ can shatter $x_1, \ldots, x_n$. For this, let $\eta_1, \ldots, \eta_n$ be arbitary in $\{0, 1\}$. We need to obtain a function $g \in \mathcal{C}$ for which $g(x_i) = \eta_i$. Define $\delta_1, \ldots, \delta_n$ by

$$\delta_i = \eta_i I\{y_i = 0\} + (1 - \eta_i) I\{y_i = 1\}.$$

Because $\mathcal{F}$ can shatter $(x_1, y_1), \ldots, (x_n, y_n)$, there exists a function $f \in \mathcal{F}$ with $f(x_i, y_i) = \delta_i$ for $i = 1, \ldots, n$. If $f(x, y) = I\{g(x) \neq y\}$ for some $g \in \mathcal{C}$, then it is now easy to verify that $g(x_i) = \eta_i$. This proves that $\mathcal{C}$ shatters $x_1, \ldots, x_n$. The proof of $VC(\mathcal{F}) \leq VC(\mathcal{C})$ is complete.

We thus obtain from (66),

$$L(\hat{g}_n) - \inf_{g \in \mathcal{C}} L(g) \leq C\sqrt{\frac{VC(\mathcal{C})}{n}} + \sqrt{\frac{8}{n} \log \frac{1}{\alpha}} \qquad \text{with prob} \geq 1 - \alpha.$$

Thus, as long as $VC(\mathcal{C}) = o(n)$, the test error of $\hat{g}_n$ relative to the best test error in $\mathcal{C}$ converges to zero as $n \to \infty$.

10 Lecture 10

10.1 Recap of the Main Empirical Process Bound from last class

Let us first recall our main bound on the expected suprema of empirical processes. If $\mathcal{F}$ is a class of real-valued functions on $\mathcal{X}$ with envelope F, then (assuming that $PF^2 < \infty$), we have

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \leq \frac{C}{\sqrt{n}} \|F\|_{L^2(P)} J(F, \mathcal{F}) \tag{67}$$

where

$$J(F, \mathcal{F}) := \int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))} d\epsilon.$$

An important implication of this bound is that when $\mathcal{F}$ is a Boolean function class with finite VC dimension D,

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \leq C\sqrt{\frac{D}{n}}. \tag{68}$$

Today, we shall discuss an application of the bounds (67) and (68) to a problem in M-estimation. My treatment here will be a mix of rigor and heuristics. We shall come back to M-estimation next week when all the heuristic arguments will be rigorized.

10.2 Application to an M-estimation problem

This can be seen as a mode estimation problem. Suppose $X_1, \ldots, X_n$ are i.i.d observations from a univariate density p. We shall assume that p has a single mode θ_0 and that it is symmetric about $\theta_0 \in \mathbb{R}$. In addition, we shall assume that p is smooth and bounded and that $p'(x) > 0$ for $x < \theta_0$ and that $p'(x) < 0$ for $x > \theta_0$. You can think of p as the normal density with mean θ_0 or the Cauchy density centered at θ_0.

Consider now the problem of estimating θ_0. For this, let us define

$$M(\theta) := \int_{\theta-1}^{\theta+1} p(x)dx = \mathbb{P}\left\{|X_1 - \theta| \leq 1\right\} = PI_{\{\theta-1 \leq X_1 \leq \theta+1\}}. \tag{69}$$

Note that

$$M'(\theta) = p(\theta + 1) - p(\theta - 1).$$

Because of the assumptions on p, it is clear that $M'(\theta_0) = 0$ and for $\theta \neq \theta_0$, we have $M'(\theta) < 0$ for $\theta > \theta_0$ and $M'(\theta) > 0$ for $\theta < \theta_0$. This implies that $\theta \mapsto M(\theta)$ has a unique maximum at θ_0. Also $M''(\theta_0) = p'(\theta_0 + 1) - p'(\theta_0 - 1) < 0$.

Because θ_0 uniquely maximizes $M(\theta)$ over $\theta \in \mathbb{R}$, a reasonable method of estimating θ_0 is to estimate it by $\hat{\theta}_n$ where $\hat{\theta}_n$ is any maximizer of $M_n(\theta)$ over $\theta \in \mathbb{R}$ with

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} I\left\{X_i \in [\theta - 1, \theta + 1]\right\}$$

It is now natural to ask the following questions:

1. Is $\hat{\theta}_n$ consistent as an estimator for θ_0 i.e., is it true that $|\hat{\theta}_n - \theta_0|$ converges in probability to zero.
2. One does have consistency in this example as we shall show. One can then ask: what is the rate of convergence r_n of $|\hat{\theta}_n - \theta_0|$ to zero?
3. What is the limiting distribution of $r_n(\hat{\theta}_n - \theta_0)$?

Below we shall prove consistency of $\hat{\theta}_n$ rigorously. I will also provide a heuristic argument for finding the rate r_n which we shall make rigorous next week. The limiting distribution will be addressed in a few weeks after the discussion on uniform central limit theorems.

The fundamental first step for analyzing M-estimators is the following inequality:

$$\begin{aligned} 0 \leq M(\theta_0) - M(\hat{\theta}_n) &= M_n(\theta_0) - M_n(\hat{\theta}_n) + M(\theta_0) - M_n(\theta_0) - M(\hat{\theta}_n) + M_n(\hat{\theta}_n) \\ &\leq M(\theta_0) - M_n(\theta_0) - M(\hat{\theta}_n) + M_n(\hat{\theta}_n) \end{aligned}$$

where we have used the inequality $M_n(\theta_0) \leq M_n(\hat{\theta}_n)$ which holds because $\hat{\theta}_n$ maximizes $M_n(\cdot)$. We can rewrite the above inequality in Empirical Process notation. For $\theta \in \mathbb{R}$, let us define the function $m_\theta : \mathbb{R} \to \mathbb{R}$ by

$$m_\theta(x) := I\left\{\theta - 1 \leq x \leq \theta + 1\right\}.$$

With this notation, the inequality becomes

$$0 \leq P\left(m_{\theta_0} - m_{\hat{\theta}_n}\right) \leq (P_n - P)\left(m_{\hat{\theta}_n} - m_{\theta_0}\right). \tag{70}$$

This inequality is so fundamental that is has been referred to as the *basic inequality*.

To derive the consistency of $\hat{\theta}_n$ from (70), we can crudely bound the right hand side of (70) as

$$(P_n - P)\left(m_{\hat{\theta}_n} - m_{\theta_0}\right) \leq 2 \sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta| .$$

It is now easy to check that $\{m_\theta, \theta \in \mathbb{R}\}$ is a Boolean class of functions with VC dimension 2 and hence inequality (68) implies that

$$\sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta| \xrightarrow{P} 0.$$

Combining this with (70), we obtain

$$M(\theta_0) - M(\hat{\theta}_n) = P\left(m_{\theta_0} - m_{\hat{\theta}_n}\right) \xrightarrow{P} 0. \tag{71}$$

Now because of our assumptions on p, the following is true:

$$\sup_{\theta \in \mathbb{R}: |\theta - \theta_0| \geq \epsilon} M(\theta) < M(\theta_0) \qquad \text{for every } \epsilon > 0. \tag{72}$$

Indeed, for $M(\theta)$ as in (69), under our assumptions on p, we have

$$\sup_{\theta \in \mathbb{R}: |\theta - \theta_0| \geq \epsilon} M(\theta) = M(\theta_0 \pm \epsilon) < M(\theta_0)$$

because $M(\cdot)$ has a unique maximum at θ_0. The two assumptions (71) and (72) imply together that $|\hat{\theta}_n - \theta_0| \xrightarrow{P} 0$. To see this, first fix $\epsilon > 0$ and use (72) to obtain $\eta > 0$ such that

$$\sup_{\theta \in \mathbb{R}: |\theta - \theta_0| \geq \epsilon} M(\theta) < M(\theta_0) - \eta.$$

It follows then that

$$\mathbb{P}\left\{|\theta_0 - \hat{\theta}_n| \geq \epsilon\right\} \leq \mathbb{P}\left\{M(\hat{\theta}_n) < M(\theta_0) - \eta\right\} = \mathbb{P}\left\{M(\theta_0) - M(\hat{\theta}_n) > \eta\right\} \to 0$$

where the last (converging to zero) assertion follows from (71). This proves that $|\hat{\theta}_n - \theta_0| \xrightarrow{P} 0$.

This argument for proving consistency of an M-estimator is quite general and can be isolated in the following theorem (which can be found, for example, in Van der Vaart [24, Theorem 5.7]).

Theorem 10.1 (Consistency)**.** *Let $\{M_n\}$ be a sequence of random functions of $\theta \in \Theta$ and let $\{M\}$ be a fixed deterministic function of $\theta \in \Theta$. Let $\hat{\theta}_n$ be any maximizer of $\{M_n(\theta), \theta \in \Theta\}$ and let θ_0 be the unique maximizer of $\{M(\theta), \theta \in \Theta\}$. Suppose the following two conditions hold*

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$.
2. *For every $\epsilon > 0$, the inequality $\sup_{\theta \in \Theta: d(\theta, \theta_0) \geq \epsilon} M(\theta) < M(\theta_0)$. Here d is a metric on Θ.*

Then $d(\hat{\theta}_n, \theta_0) \xrightarrow{P} 0$ as $n \to \infty$.

The assumption $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$ is often too strong for consistency (and also not always easy to check) but there exist results with weaker conditions.

Now that the consistency of $\hat{\theta}_n$ is established, the next natural question is about the rate of convergence. We can first try to go over the consistency argument again to see it gives an explicit rate of convergence for $|\hat{\theta}_n - \theta_0|$. I shall argue heuristically.

The consistency argument given above was based on the inequality:

$$0 \leq M(\theta_0) - M(\hat{\theta}_n) \leq (P_n - P)\left(m_{\hat{\theta}_n} - m_{\theta_0}\right) \leq 2 \sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta|. \tag{73}$$

For consistency, we used that the right hand side above converges in probability to zero. But the inequality (68) actually implies that

$$\mathbb{E} \sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta| \leq \frac{C}{\sqrt{n}}$$

which gives

$$\sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta| = O_P(n^{-1/2}).$$

Inequality (73) then gives

$$0 \leq M(\theta_0) - M(\hat{\theta}_n) = O_P(n^{-1/2}). \tag{74}$$

From here, to obtain an explicit rate for $|\hat{\theta}_n - \theta_0|$ we need to use some structure of the function $M(\cdot)$. Note that the second derivative of M at θ_0 equals

$$M''(\theta_0) = p'(\theta_0 + 1) - p'(\theta_0 - 1)$$

which is strictly negative because, by assumption, $p'(\theta_0 + 1) < 0$ and $p'(\theta_0 - 1) > 0$. As a result, there exists a constant C and aneighbourhood of θ_0 such that for all θ in that neighbourhood, we can write

$$M(\theta_0) - M(\theta) \geq C(\theta - \theta_0)^2. \tag{75}$$

The value of C is related to $M''(\theta_0)$. Using this, we can heuristically write

$$M(\theta_0) - M(\hat{\theta}_n) \geq C(\hat{\theta}_n - \theta_0)^2. \tag{76}$$

In other words, I am assuming that $\hat{\theta}_n$ belongs to the neighborhood of θ_0 where (75) holds. Because of the consistency of $\hat{\theta}_n$ (which we have rigorously proved), the inequality (76) can be made rigorous. Combining (76) with (74), we deduce that

$$\left(\hat{\theta}_n - \theta_0\right)^2 = O_P(n^{-1/2})$$

which gives

$$\left|\hat{\theta}_n - \theta_0\right| = O_P(n^{-1/4}).$$

We have therefore obtained $n^{-1/4}$ as a rate of convergence for $|\hat{\theta}_n - \theta_0|$. It turns out however that

$$\left|\hat{\theta}_n - \theta_0\right| = O_P(n^{-1/3}).$$

In other words, $n^{-1/4}$ is slower than the actual rate of convergence and is a reflection of some loosness in our proof technique. The main source of looseness is in the inequality:

$$(P_n - P)\left(m_{\hat{\theta}_n} - m_{\theta_0}\right) \leq 2 \sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta| = O_P(n^{-1/2}) \tag{77}$$

It turns out that the left hand side above is much smaller than the right hand side. To get a heuristic understanding of the size of the left hand side above, let us first compute bounds for

$$\mathbb{E}\left|(P_n - P)\left(m_\theta - m_{\theta_0}\right)\right|$$

for a fixed θ which is close to θ_0. Clearly

$$\begin{aligned}\mathbb{E}\left|(P_n - P)\left(m_\theta - m_{\theta_0}\right)\right| &\leq \sqrt{\operatorname{Var}\left(P_n(m_\theta - m_{\theta_0})\right)} \\ &= \frac{1}{\sqrt{n}}\sqrt{\operatorname{Var}\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)} \leq \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)^2}\end{aligned} \tag{78}$$

Now for θ close to θ_0 (and $\theta < \theta_0$), we have

$$m_\theta(x) - m_{\theta_0}(x) = I\{\theta - 1 \le x \le \theta_0 - 1\} + I\{\theta + 1 \le X_1 \le \theta_0 + 1\}.$$

Thus

$$\mathbb{E}\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)^2 \le \mathbb{P}\{\theta - 1 \le X_1 \le \theta_0 + 1\} + \mathbb{P}\{\theta + 1 \le X_1 \le \theta_0 + 1\} \le 2p(\theta_0)|\theta - \theta_0|.$$

where, in the last inequality, we used the fact that the density of X_1 has a mode at θ_0 (so that the density at every other point is bounded by $p(\theta_0)$). Combining this with (78), we obtain

$$\mathbb{E}\left|(P_n - P)\left(m_\theta - m_{\theta_0}\right)\right| \le \sqrt{\frac{2p(\theta_0)}{n}}\sqrt{|\theta - \theta_0|}$$

so that

$$(P_n - P)\left(m_\theta - m_{\theta_0}\right) = O_P\left(\sqrt{\frac{|\theta - \theta_0|}{n}}\right).$$

This is true for a fixed θ that is close to θ_0. Heuristically, this suggests that

$$(P_n - P)\left(m_{\hat{\theta}_n} - m_{\theta_0}\right) = O_P\left(\sqrt{\frac{|\hat{\theta}_n - \theta_0|}{n}}\right).$$

We shall formally justify this later. Note that this bound is an stronger compared to our earlier bound (77). Plugging this in the right hand side of the basic inequality (70) and using the quadratic bound (76) on the left hand side of (70), we deduce that

$$C\left(\hat{\theta}_n - \theta_0\right)^2 \le O_P\left(\sqrt{\frac{|\hat{\theta}_n - \theta_0|}{n}}\right)$$

"Cancelling" $|\hat{\theta}_n - \theta_0|^{1/2}$ from both sides, we deduce that

$$\left|\hat{\theta}_n - \theta_0\right| = O_P(n^{-1/3}). \tag{79}$$

As mentioned earlier, this is the correct rate for $\hat{\theta}_n - \theta_0$. Indeed, it turns out that

$$n^{1/3}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} \operatorname*{argmax}_{h \in \mathbb{R}}\left\{aZ_h - bh^2\right\}$$

as $n \to \infty$ where $Z_h, h \in \mathbb{R}$ is two-sided Brownian motion starting from 0 and a, b are two constants depending on p and θ_0. We shall prove this limiting result later but it tells us now that $n^{-1/3}$ is the correct rate of convergence.

We shall make the rate result (79) rigorous next week. A key ingredient in the rigorous argument will involve establishing the inequality

$$\mathbb{E}\sup_{\theta:|\theta-\theta_0|\le\delta}|(P_n - P)(m_\theta - m_{\theta_0})| \le C\sqrt{\frac{\delta}{n}} \tag{80}$$

for δ sufficiently small. The above inequality can be derived as a consequence of (67). Indeed, to apply (67), we first need to obtain an envelope for the class $\{m_\theta - m_{\theta_0} : |\theta - \theta_0| \le \delta\}$. It is not hard to see that

$$F(x) := I_{[\theta_0 - 1 - \delta, \theta_0 - 1 + \delta]}(x) + I_{[\theta_0 + 1 - \delta, \theta_0 + 1 + \delta]}(x)$$

is an envelope function. Further

$$PF^2 \leq 2\mathbb{P}\left\{\theta_0 - 1 - \delta \leq X_1 \leq \theta_0 - 1 + \delta\right\} + 2\mathbb{P}\left\{\theta_0 + 1 - \delta \leq X_1 \leq \theta_0 + 1 + \delta\right\} \leq Cp(\theta_0)\delta \leq C\delta.$$

Thus (67) gives

$$\mathbb{E} \sup_{\theta:|\theta-\theta_0|\leq\delta} |(P_n - P)(m_\theta - m_{\theta_0})| \leq C\sqrt{\frac{\delta}{n}} J(F, \mathcal{F})$$

where $\mathcal{F} := \{m_\theta - m_{\theta_0} : |\theta - \theta_0| \leq \delta\}$. This will prove (80) provided $J(F, \mathcal{F}) < \infty$. This will follow from the fact that the class $\{m_\theta - m_{\theta_0}\}$ has finite VC subgraph dimension (to be defined shortly). Note that this is not a Boolean class of functions so we need VC subgraph dimension as opposed to VC dimension.

Let us now summarize this discussion of a heuristic argument for the rate of M-estimators. Although we did for a special M-estimator which corresponded to $m_\theta(x) := I\{\theta - 1 \leq x \leq \theta + 1\}$, the ideas are actually fairly general. The most important ingredient is the Basic inequality (70). The left hand side $P(m_{\theta_0} - m_{\hat{\theta}_n})$ is bounded from below by an assumption on the second derivative of $\theta \mapsto Pm_\theta$ at $\theta = \theta_0$. The right hand side can be understood by calculating

$$\mathbb{E}\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)^2.$$

For the specific choice of $m_\theta(x) = I\{\theta - 1 \leq x \leq \theta + 1\}$, it turned out that

$$\mathbb{E}\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)^2 \leq C|\theta - \theta_0|.$$

For other m_θ, the right hand side might be different (for example, it is common to have $C|\theta - \theta_0|^2$ on the right hand side). Plugging these bounds in the basic inequality will yield an inequality involving $|\hat{\theta}_n - \theta_0|$ and n which can be solved to get explicit rates (such as $n^{-1/3}$) in this problem. This heuristic will be justified next week. Before concluding this section, let us state a result from Van der Vaart and Wellner [25, Page 294] on the rate of convergence of M-estimators. We shall formally prove this result later but, based on the above heuristics, its conclusion should be quite obvious.

Theorem 10.2 (Van der Vaart and Wellner, Page 294)**.** *Let $X_1, X_2, \ldots$ be i.i.d observations from a distribution P. Suppose $\Theta \subseteq \mathbb{R}$ is an open set and let $m_\theta, \theta \in \Theta$ be a collection of real-valued functions on $\mathcal{X}$ that are indexed by Θ. Suppose there exist $\alpha > 0$ and a function M on $\mathcal{X}$ with $PM^2 < \infty$ for which*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq M(x)|\theta_1 - \theta_2|^\alpha \qquad \text{for all } \theta_1, \theta_2 \in \Theta. \tag{81}$$

Let $\hat{\theta}_n$ and θ_0 denote maximizers of $P_n m_\theta$ and Pm_θ over $\theta \in \Theta$. If $\theta \mapsto Pm_\theta$ has two derivatives at θ_0 with the second derivative strictly negative, then

$$|\hat{\theta}_n - \theta_0| = O_P(n^{-1/(4-2\alpha)}). \tag{82}$$

Heuristic Argument. We use the heuristics summarized above to justify (82) based on the assumptions made in the theorem. Note that assumption (81) implies that

$$\mathbb{E}\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)^2 \leq \mathbb{E}\left(M^2(X_1)|\theta - \theta_0|^{2\alpha}\right) \leq |\theta - \theta_0|^{2\alpha} PM^2 \leq C|\theta - \theta_0|^{2\alpha}.$$

This suggests the heuristic

$$(P_n - P)\left(m_{\hat{\theta}_n} - m_{\theta_0}\right) \leq n^{-1/2}|\hat{\theta}_n - \theta_0|^\alpha.$$

Combining with the basic inequality (and the lower bound $C(\hat{\theta}_n - \theta_0)^2$ on $P(m_{\theta_0} - m_{\hat{\theta}_n})$, we obtain the inequality

$$(\hat{\theta}_n - \theta_0)^2 \lesssim n^{-1/2}|\hat{\theta}_n - \theta_0|^\alpha$$

which gives

$$|\hat{\theta}_n - \theta_0| \lesssim n^{-1/(4-2\alpha)}.$$

□

Note that when $\alpha = 1$, Theorem 10.2 gives the usual $n^{-1/2}$ rate for $|\hat{\theta}_n - \theta_0|$.

11 Lecture 11

For Boolean function classes $\mathcal{F}$, we have seen that the VC dimension gives useful upper bounds on covering numbers:

$$\sup_Q \log M(\epsilon, \mathcal{F}, L^2(Q)) \leq \left(\frac{c_1}{\epsilon}\right)^{c_2 VC(\mathcal{F})} \qquad \text{for all } 0 < \epsilon \leq 1. \tag{83}$$

What happens for function classes that are not Boolean? We shall see today that there exist two notions of combinatorial dimension for general function classes which allow control of covering numbers via bounds similar to (83). These are the notions of *VC subgraph dimension* and *fat shattering dimension* which we shall go over today.

11.1 VC Subgraph Dimension

The VC subgraph dimension of $\mathcal{F}$ is simply the VC dimension of the Boolean class obtained by taking the indicators of the subgraphs of functions in $\mathcal{F}$. To formally define this, let us first define the notion of subgraph of a function.

Definition 11.1 (Subgraph)**.** *For a function $f : \mathcal{X} \to \mathbb{R}$, its subgraph $sg(f)$ is a subset of $\mathcal{X} \times \mathbb{R}$ that is defined as*

$$sg(f) := \{(x, t) \in \mathcal{X} \times \mathbb{R} : t < f(x)\}.$$

In other words, $sg(f)$ consists of all points that lie below the graph of the function f.

We can now define the VC subgraph dimension of $\mathcal{F}$ as:

Definition 11.2 (VC Subgraph Dimension)**.** *The VC subgraph dimension of $\mathcal{F}$ is defined as the VC dimension of the Boolean class $\{I_{sg(f)} : f \in \mathcal{F}\}$. We shall denote this by just $VC(\mathcal{F})$.*

The VC subgraph dimension can be related to covering numbers in the same way as (83). This is done in the following result.

Theorem 11.3. *Suppose $\mathcal{F}$ is a class of functions with envelope F. Let the VC-subgraph dimension of $\mathcal{F}$ be equal to D. Then*

$$\sup_Q M(\epsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q)) \leq \left(\frac{c_1}{\epsilon}\right)^{c_2 D} \qquad \textit{for all } 0 < \epsilon \leq 1 \tag{84}$$

where c_1 and c_2 are universal positive constants.

Proof. The idea is to relate the $L^2(Q)$ norm between two functions in $\mathcal{F}$ to an L^2 norm between their subgraphs. Fix $f, g \in \mathcal{F}$ and write

$$\int |f - g|^2 dQ \leq \int 2F(x)|f(x) - g(x)| dQ(x)$$

where we used the fact that $|f(x) - g(x)| \leq 2F(x)$ (this is true because F is the envelope of $\mathcal{F}$). We now use the fact that for every two real numbers a and b, we have the identity

$$|a - b| = \int |I\{t < a\} - I\{t < b\}| dt$$

This gives

$$\begin{aligned}
\int |f-g|^2 dQ &\leq \int 2F(x)|f(x)-g(x)|dQ(x) \\
&= \int 2F(x)\left(\int |I\{t<f(x)\}-I\{t<g(x)\}|\,dt\right)dQ(x) \\
&= \int\int \left|I_{sg(f)}(x,t)-I_{sg(g)}(x,t)\right| 2F(x)dQ(x)dt \\
&= \int\int\limits_{(x,t):|t|\leq F(x)} \left|I_{sg(f)}(x,t)-I_{sg(g)}(x,t)\right| 2F(x)dQ(x)dt \quad \text{as } I_{sg(f)}(x,t)=I_{sg(g)}(x,t) \text{ for } |t|>F(x) \\
&= \left(\int\int\limits_{(x,t):|t|\leq F(x)} 2F(x)dQ(x)dt\right) \int\int\limits_{(x,t):|t|\leq F(x)} \left|I_{sg(f)}(x,t)-I_{sg(g)}(x,t)\right| \frac{2F(x)dQ(x)dt}{\int\int\limits_{(x,t):|t|\leq F(x)} 2F(x)dQ(x)dt} \\
&= \left(4\int F^2(x)dQ(x)\right) \int\int\limits_{(x,t):|t|\leq F(x)} \left|I_{sg(f)}(x,t)-I_{sg(g)}(x,t)\right|^2 \frac{2F(x)dQ(x)dt}{\int\int\limits_{(x,t):|t|\leq F(x)} 2F(x)dQ(x)dt}.
\end{aligned}$$

We have thus proved that

$$\|f-g\|_{L^2(Q)} \leq 2\|F\|_{L^2(Q)} \left\|I_{sg(f)}-I_{sg(g)}\right\|_{L^2(Q')} \tag{85}$$

where Q' is the probability measure on $\mathcal{X}\times\mathbb{R}$ whose density with respect to $Q\times Leb$ is proportional to

$$2F(x)I\{(x,t):|t|\leq F(x)\}.$$

It is routine to deduce from (85) that

$$M(\epsilon\|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q)) \leq M(\epsilon/2, \{I_{sg(f)}, f\in\mathcal{F}\}, L^2(Q'))$$

To bound the right hand side, we simply use the earlier result for Boolean classes (see (83)). This completes the proof of Theorem 11.3. □

The following is an immediate corollary of Theorem 11.3 and our main bound on the Expected suprema of empirical processes.

Corollary 11.4. *If $\mathcal{F}$ has envelope F and VC subgraph dimnension D, then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}|P_nf-Pf| \leq C\|F\|_{L^2(P)}\sqrt{\frac{D}{n}}.$$

Proof. Our main bound on the expected suprema of empirical processes gives

$$\mathbb{E}\sup_{f\in\mathcal{F}}|P_nf-Pf| \leq C\|F\|_{L^2(P)}\frac{J(F,\mathcal{F})}{\sqrt{n}}$$

and we bound $J(F,\mathcal{F})$ (using Theorem 11.3) as

$$\begin{aligned}
J(F,\mathcal{F}) &= \int_0^1 \sqrt{1+\log\sup_Q M(\epsilon\|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))}d\epsilon \\
&\leq \int_0^1 \sqrt{1+c_2D\log\frac{c_1}{\epsilon}}d\epsilon \leq 1+\sqrt{D}\int_0^1\sqrt{c_2\log\frac{c_1}{\epsilon}}d\epsilon \leq C\sqrt{D}
\end{aligned}$$

which completes the proof of Corollary 11.4. □

Example 11.5. *In the last lecture, I remarked that*

$$\mathbb{E} \sup_{\theta \in \mathbb{R}: |\theta - \theta_0| \leq \delta} |(P_n - P)(m_\theta - m_{\theta_0})| \leq C\sqrt{\frac{\delta}{n}}$$

where $m_\theta(x) := I\{\theta - 1 \leq x \leq \theta + 1\}$*. I gave a partial proof of this fact in the last class. Complete this proof by proving that the function-class*

$$\left\{ I_{[\theta-1,\theta+1]} - I_{[\theta_0-1,\theta_0+1]} : \theta \in \mathbb{R} \right\}$$

has finite VC subgraph dimension (≤ 3*??).*

Let us now look at a reformulation of the VC subgraph dimension. This defines the dimension directly in terms of the class $\mathcal{F}$ without going to subgraphs. This reformulation will also make clear the connection to fat shattering dimension.

We say that a subset $\{x_1, \ldots, x_n\}$ of $\mathcal{X}$ is subgraph-shattered by $\mathcal{F}$ if there exist real numbers $t_1, \ldots, t_n$ such that for every subset $S \subseteq \{1, \ldots, n\}$, there exists $f \in \mathcal{F}$ with

$$(x_s, t_s) \notin sg(f) \text{ for } s \in S \quad \text{and} \quad (x_s, t_s) \in sg(f) \text{ for } s \notin S. \tag{86}$$

Note that (86) is equivalent to

$$f(x_s) \leq t_s \text{ for } s \in S \quad \text{and} \quad f(x_s) > t_s \text{ for } s \notin S. \tag{87}$$

In words, we say that $\{x_1, \ldots, x_n\}$ is subgraph-shattered by $\mathcal{F}$ if there exist *levels* $t_1, \ldots, t_n$ such that for every subset $S \subseteq \{1, \ldots, n\}$, there exists a function f which goes **under** the level for each $x_s, s \in S$ and **strictly over** the level for $x_s, s \notin S$.

The VC subgraph dimension $VC(\mathcal{F})$ is defined as the maximum cardinality of a finite subset of $\mathcal{X}$ that is subgraph shattered by $\mathcal{F}$.

Let us now describe a potential problem with using the VC subgraph dimension to control covering numbers. Let $\mathcal{M}$ denote the class of all nondecreasing functions $f : \mathbb{R} \to [-1, 1]$ i.e., $\mathcal{M}$ consists of all nondecreasing functions on $\mathbb{R}$ that are bounded by 1. It turns out then that

$$\sup_Q M(\epsilon, \mathcal{M}, L^2(Q)) \leq \exp\left(\frac{C}{\epsilon}\right) \qquad \text{for all } \epsilon > 0. \tag{88}$$

It is also easy to see that the VC-subgraph dimension of $\mathcal{M}$ equals ∞ (i.e., for every $n \geq 1$, there exists a finite subset of $\mathbb{R}$ that is subgraph shattered by $\mathcal{M}$). Therefore, the notion of VC-subgraph dimension is not useful here and Theorem 11.3 does not give anything meaningful for this class $\mathcal{M}$. It is actually possible to prove (88) using the notion of fat shattering dimension which is discussed next.

11.2 Fat Shattering Dimension

Fat Shattering is a scale sensitive notion of dimension. Specifically, fat shattering dimension is actually a function on $(0, \infty)$ i.e., it is defined for each $\epsilon > 0$. We shall denote this by $\text{fat}_{\mathcal{F}}(\epsilon)$ and is defined in the following way.

Definition 11.6 (ϵ-shattering)**.** *We say that a subset* $\{x_1, \ldots, x_n\}$ *of* $\mathcal{X}$ *is* ϵ*-shattered by* $\mathcal{F}$ *if there exist real numbers* $t_1, \ldots, t_n$ *such that for every* $S \subseteq \{1, \ldots, n\}$*, there exists a function* $f \in \mathcal{F}$ *such that*

$$f(x_s) \leq t_s \text{ for } s \in S \quad \text{and} \quad f(x_s) \geq t_s + \epsilon \text{ for } s \notin S. \tag{89}$$

In words, we say that $\{x_1, \ldots, x_n\}$ is subgraph-shattered by $\mathcal{F}$ if there exist *levels* $t_1, \ldots, t_n$ such that for every subset $S \subseteq \{1, \ldots, n\}$, there exists a function f which goes **under** the level for each $x_s, s \in S$ and **exceeds by** ϵ the level for $x_s, s \notin S$. Note that the only difference between the notion of ϵ-shattering and the notion of subgraph-shattering from the previous subsection is that the words "strictly over" are replaced by "exceeds by ϵ".

Definition 11.7 (Fat Shattering Dimension)**.** *For $\epsilon > 0$, the fat shattering dimension $\text{fat}_{\mathcal{F}}(\epsilon)$ is defined as the maximum cardinality of a finite subset of $\mathcal{X}$ that is ϵ-shattered by $\mathcal{F}$.*

It is clear that $\text{fat}_{\mathcal{F}}(\epsilon)$ is a decreasing function of ϵ. In fact, it can be shown (exercise) that

$$VC(\mathcal{F}) = \sup_{\epsilon > 0} \text{fat}_{\mathcal{F}}(\epsilon)$$

where $VC(\mathcal{F})$ above refers to VC subgraph dimension. This inequality means, in particular, that $\text{fat}_{\mathcal{F}}(\epsilon)$ is always bounded from above by $VC(\mathcal{F})$. Another easy fact is that when $\mathcal{F}$ is Boolean, then $\text{fat}_{\mathcal{F}}(\epsilon)$ equals $VC(\mathcal{F})$ for every $0 < \epsilon \leq 1$.

Recall now the class $\mathcal{M}$ from the last subsection consisting of all nondecreasing functions $f : \mathbb{R} \to [-1, 1]$. It was mentioned before that the VC subgraph dimension of $\mathcal{M}$ is infinity. We shall show now that $\text{fat}_{\mathcal{M}}(\epsilon)$ is finite for every $\epsilon > 0$ and in fact

$$\text{fat}_{\mathcal{M}}(\epsilon) \leq 1 + \frac{2}{\epsilon} \qquad \text{for every } \epsilon > 0.$$

In fact, the above fat shattering dimension bound holds for the larger class of all functions $f : \mathbb{R} \to \mathbb{R}$ whose variation is bounded by 2. This is proved in the following result. Recall that the variation of a function $f : \mathbb{R} \to \mathbb{R}$ is defined by

$$\|f\|_{TV} := \sup_{n \geq 1} \sup_{x_1 < x_2 \cdots < x_n} \sum_{i=1}^{n-1} |f(x_i) - f(x_{i+1})|.$$

Lemma 11.8. *Fix $V > 0$ and let $\mathcal{F}$ denote the class of all functions $f : \mathbb{R} \to \mathbb{R}$ with $\|f\|_{TV} \leq V$. Then*

$$\text{fat}_{\mathcal{F}}(\epsilon) = 1 + \left\lfloor \frac{V}{\epsilon} \right\rfloor \qquad \textit{for every } \epsilon > 0.$$

Proof. Fix $\epsilon > 0$. Let us first prove that

$$\text{fat}_{\mathcal{F}}(\epsilon) \leq 1 + \left\lfloor \frac{V}{\epsilon} \right\rfloor \tag{90}$$

For this, $\{x_1, \ldots, x_n\}$ be ϵ-shattered by $\mathcal{F}$. We shall show then that n cannot be larger than the right hand side of (90) which will prove (90). Note first that because $\{x_1, \ldots, x_n\}$ are ϵ-shattered, there exist real numbers $t_1, \ldots, t_n$ which satisfy the condition in the definition of ϵ-shattering. This means, in particular, that there exist two functions f_1 and f_2 in $\mathcal{F}$ which satisfy the following:

$$f_1(x_i) \text{ is } \begin{cases} \leq t_i & : \text{ for odd } i \\ \geq t_i + \epsilon & : \text{ for even } i \end{cases}$$

and

$$f_2(x_i) \text{ is } \begin{cases} \leq t_i & : \text{ for even } i \\ \geq t_i + \epsilon & : \text{ for odd } i \end{cases}$$

Now let $f = (f_1 - f_2)/2$. The conditions above imply together that

$$f(x_i) \text{ is } \begin{cases} \leq -\epsilon/2 & : \text{ for odd } i \\ \geq \epsilon/2 & : \text{ for even } i \end{cases}$$

which immediately implies that

$$\|f\|_{TV} = \sum_{i=1}^{n-1} |f(x_i) - f(x_{i+1})| \geq (n-1)\epsilon.$$

On the other hand,

$$\|f\|_{TV} = \|(f_1 - f_2)/2\|_{TV} \leq \frac{\|f_1\|_{TV} + \|f_2\|_{TV}}{2} \leq V.$$

Combining the above two inequalities, we obtain $(n-1)\epsilon \leq V$ which implies (90) (note that n has to be an integer which allows us to put integer part around V/ϵ).

We now show that $\text{fat}_{\mathcal{F}}(\epsilon)$ is larger than or equal to the right hand side of (90). For this, let $d = \lfloor V/\epsilon \rfloor$. Consider any set of d points $y_1 < \cdots < y_d$. These form $d+1$ intervals $I_j := [y_j, y_{j+1})$ for $j = 1, \ldots, d-1$ and $I_0 := (-\infty, y_1]$ and $I_d := [y_d, \infty)$. Let G consist of the 2^{d+1} functions from $\mathbb{R}$ to $\{0, \epsilon\}$ that are piecewise constant on each of the intervals $I_0, \ldots, I_d$. Let $\{x_1, \ldots, x_{d+1}\}$ be any finite set obtained by picking one point from each of the $d+1$ intervals $I_0, \ldots, I_d$. It is then clear that $\{x_1, \ldots, x_{d+1}\}$ is ϵ-shattered by G. Further, the variation of every function in G is atmost $d\epsilon = \epsilon\lfloor V/\epsilon \rfloor \leq V$. Thus $\mathcal{F}$ shatters $\{x_1, \ldots, x_{d+1}\}$ which means

$$\text{fat}_{\mathcal{F}}(\epsilon) \geq 1 + d = 1 + \left\lfloor \frac{V}{\epsilon} \right\rfloor.$$

This completes the proof of Lemma 11.8. □

Let us now describe a result which bound covering numbers in terms of the fat-shattering dimension $\text{fat}_{\mathcal{F}}(\epsilon), \epsilon > 0$. This is the following theorem due to Mendelson and Vershynin [16].

Theorem 11.9 (Mendelson-Vershynin)**.** *Suppose $\mathcal{F}$ is a class of functions that are uniformly bounded by 1. Then there exist a universal positive constant $C \geq 1$ such that*

$$\sup_Q M(\epsilon, \mathcal{F}, L^2(Q)) \leq \left(\frac{2}{\epsilon}\right)^{C\text{fat}_{\mathcal{F}}(\epsilon/C)} \qquad \text{for all } 0 < \epsilon \leq 1. \tag{91}$$

Let us see what this result gives for class $\mathcal{M}$ of all nondecreasing functions $f : \mathbb{R} \to [-1, 1]$. Every function in this class has variation at most 2 and thus Lemma 11.8 implies that

$$\text{fat}_{\mathcal{M}}(\epsilon) \leq 1 + \frac{2}{\epsilon}. \tag{92}$$

This, together with Theorem 11.9, allows us to deduce that

$$\sup_Q M(\epsilon, \mathcal{M}, L^2(Q)) \leq \exp\left(\frac{C}{\epsilon} \log \frac{2}{\epsilon}\right) \qquad \text{for } 0 < \epsilon \leq 1.$$

Note that this result is weaker compared to (88) by a factor of $\log(2/\epsilon)$ in the exponent. We can now ask if it is possible to derive (88) via the fat shattering dimension. This is possible if the bound (91) can be improved to $\exp(C\text{fat}_{\mathcal{F}}(\epsilon/C))$. Note that this cannot be done in general. For example, when $\mathcal{F}$ is Boolean with finite VC dimension, then $\text{fat}_{\mathcal{F}}(\epsilon) = VC(\mathcal{F})$ for every $0 < \epsilon \leq 1$ and in this case, one obviously cannot replace $2/\epsilon$ by a constant in (91). However, Rudelson and Vershynin [21] have showed that under a technical regularity assumption on $\text{fat}_{\mathcal{F}}(\epsilon)$ (which rules out the case when $\mathcal{F}$ is Boolean with finite VC dimension), it is indeed possible to improve Theorem 11.9. This result is given below.

Theorem 11.10 (Rudelson-Vershynin)**.** *Suppose $\mathcal{F}$ is a class of functions. Suppose $a > 2$ and a decreasing function $v : (0, \infty) \to (0, \infty)$ are such that*

$$\text{fat}_{\mathcal{F}}(\epsilon) \leq v(\epsilon) \quad \text{and} \quad v(a\epsilon) \leq \frac{1}{2} v(\epsilon) \tag{93}$$

for all $\epsilon > 0$. Then

$$\sup_Q M(\epsilon, \mathcal{F}, L^2(Q)) \leq \exp\left(C(\log a) v\left(\frac{\epsilon}{C}\right)\right) \tag{94}$$

for every $\epsilon > 0$. C as usual is a universal constant.

Note that the regularity condition (93) rules out situations such as the case when $v(\epsilon)$ is constant. Also notice that there is no explicit boundedness assumption on the functions in $\mathcal{F}$; this is hidden in the regularity condition.

Let us now show that Theorem 11.10 does indeed imply the result (88) for the class $\mathcal{M}$ of nondecreasing functions that are constrained to take values in $[-1, 1]$. Indeed, for this class we first have (92). Also because the functions in $\mathcal{M}$ are constrained to take values in $[-1, 1]$, the fat shattering dimension will be zero for large ϵ. In fact, it is easy to see that $\text{fat}_{\mathcal{M}}(\epsilon) = 0$ for $\epsilon \geq 3$. This, along with (92), implies that

$$\text{fat}_{\mathcal{M}}(\epsilon) \leq \frac{5}{\epsilon} \qquad \text{for all } \epsilon > 0.$$

We can therefore apply Theorem 11.10 with $v(\epsilon) := 5/\epsilon$. The condition $v(a\epsilon) \leq v(\epsilon)/2$ is easily seen to be satisfied with $a = 3$. It is straightforward then to show that inequality (88) is a consequence of (94).

12 Lecture 12

12.1 Bracketing Control

Our main empirical process bound so far is the following. Under the usual notation:

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \leq \frac{C}{\sqrt{n}} \|F\|_{L^2(P)} J(F, \mathcal{F}) \tag{95}$$

where

$$J(F, \mathcal{F}) := \int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))} d\epsilon.$$

Bracketing methods provide another upper bound for $\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf|$ which we shall describe next. This bound will be very similar to (95) except that $\sup_Q M(\epsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))$ will be replaced by the ϵ-**bracketing** number of $\mathcal{F}$ in $L^2(P)$. Before we state this result, let us first define the notion of bracketing numbers:

1. Given two real-valued functions ℓ and u on $\mathcal{X}$, the bracket $[\ell, u]$ is defined as the collection of all functions $f : \mathcal{X} \to \mathbb{R}$ for which $\ell(x) \leq f(x) \leq u(x)$ for all $x \in \mathcal{X}$.
2. Given a probability measure P on $\mathcal{X}$, the $L^2(P)$-size of a bracket $[\ell, u]$ is defined as $\|u - \ell\|_{L^2(P)}$.
3. Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. For $\epsilon > 0$, the bracketing number $N_{[]}(\epsilon, \mathcal{F}, L^2(P))$ is defined as the smallest number of brackets each having $L^2(P)$-size at most ϵ such that every $f \in \mathcal{F}$ belongs to one of the brackets.

It is important to notice that the bracketing numbers are larger than covering numbers as shown below.

Lemma 12.1. *For every $\epsilon > 0$,*

$$N_{\mathcal{F}}(\epsilon, \mathcal{F}, L^2(P)) \leq N_{\mathcal{F}_{all}}(\epsilon/2, \mathcal{F}, L^2(P)) \leq N_{[]}(\epsilon, \mathcal{F}, L^2(P)).$$

Here $\mathcal{F}_{all}$ denotes the class of all real-valued functions on $\mathcal{X}$.

Proof. The first inequality is something we have already seen when discussing covering numbers. The second inequality is proved as follows. First get brackets $[\ell_i, u_i], i = 1, \ldots, N$ each of $L^2(P)$-size ϵ which cover $\mathcal{F}$. Then it is obvious to see that the mid-point functions $(\ell_i + u_i)/2, i = 1, \ldots, N$ form an $\epsilon/2$-net for $\mathcal{F}$ in the $L^2(P)$ metric. □

Next, we provide an example where the bracketing numbers can be explicitly computed.

Example 12.2. *Let $\mathcal{F} := \{I_{(\infty,t]} : t \in \mathbb{R}\}$ and let P be a fixed probability measure on $\mathbb{R}$. Then we shall argue that*

$$N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \leq 1 + \frac{1}{\epsilon^2} \qquad \text{for every } \epsilon > 0. \tag{96}$$

Here is an argument for (96). Let $t_0 := -\infty$ and recursively define

$$t_i := \sup \{x > t_{i-1} : P(t_{i-1}, x] \leq \epsilon\}.$$

Then, for every $\delta > 0$ sufficiently small, it is clear that $P(t_{i-1}, t_i - \delta] \leq \epsilon$ (because otherwise $t_i - \epsilon$ would be the supremum) and hence (by letting $\delta \to 0$), we deduce that $P(t_{i-1}, t_i) \leq \epsilon$. Also if $t_i < \infty$, then for every $\delta > 0$, we have $P(t_{i-1}, t_i + \delta] > \epsilon$ so that (by letting $\delta \downarrow 0$), $P(t_{i-1}, t_i] \geq \epsilon$.

Let $k \geq 1$ be the smallest integer for which $t_k = \infty$. Then, by the above, we have $P(t_{i-1}, t_i] \geq \epsilon$ for $i = 1, \ldots, k-1$ so that

$$1 = P(-\infty, \infty) = \sum_{i=1}^{k} P(t_{i-1}, t_i] \geq (k-1)\epsilon$$

which gives $k \leq 1 + \epsilon^{-1}$. Now consider the brackets $[I_{(-\infty,t_{i-1}]}, I_{(-\infty,t_i)}]$ for $i = 1, \ldots, k$. These obviously cover $\mathcal{F}$ (i.e., each function in $\mathcal{F}$ belongs to one of these brackets) and their $L^2(P)$-size is

$$\sqrt{P(t_{i-1}, t_i)} \leq \sqrt{\epsilon}.$$

We have thus proved that

$$N_{[]}(\sqrt{\epsilon}, \mathcal{F}, L^2(P)) \leq 1 + \frac{1}{\epsilon}$$

This, being true for all $\epsilon > 0$, is the same as (96).

Before stating the analogue of (95) involving bracketing numbers, let us first state and prove a simple classical asymptotic result which shows that bracketing number bounds can be used to control $\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf|$.

Proposition 12.3. *Suppose $\mathcal{F}$ is a function class such that $N_{[]}(\epsilon, \mathcal{F}, L^2(P)) < \infty$ for every $\epsilon > 0$. Then*

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{P} 0 \qquad \text{as } n \to \infty. \tag{97}$$

Proof. Fix $\epsilon > 0$. Let $[\ell_i, u_i], i = 1, \ldots, N$ denote brackets of $L^2(P)$-size $\leq \epsilon$ which cover $\mathcal{F}$. We shall first argue that

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \leq \max_{1 \leq i \leq N} \max \left(|P_n u_i - P u_i|, |P_n \ell_i - P \ell_i| \right) + \epsilon \tag{98}$$

Let us first complete the proof of (97) assuming that (98) is true. To see this, note that the right hand side above converges to 0 almost surely as $n \to \infty$. This is because, by the strong law of large numbers, $|P_n u_i - P u_i|$ and $|P_n \ell_i - P \ell_i|$ converge to zero almost surely as $n \to \infty$ for each i (note that the functions u_i and ℓ_i do not change with n) and hence the finite maximum of these over $i = 1, \ldots, N$ also converges to zero. Thus, from (98), we deduce that

$$\limsup_{n \to \infty} \sup_{f \in \mathcal{F}} |P_n f - Pf| \leq \epsilon \qquad \text{almost surely for every } \epsilon > 0.$$

Applying this for each $\epsilon = 1/m$ and letting $m \to \infty$, it is possible to deduce (97).

It remains therefore to prove (98). Fix $f \in \mathcal{F}$ and get a bracket $[\ell_i, u_i]$ which contains f. This means that $\ell_i(x) \leq f(x) \leq u_i(x)$ for every $x \in \mathcal{X}$. Write

$$\begin{aligned} P_n f - Pf &\leq P_n u_i - P u_i + P u_i - Pf \\ &\leq P_n u_i - P u_i + P u_i - P \ell_i \\ &\leq P_n u_i - P u_i + \|u_i - \ell_i\|_{L^2(P)} \leq P_n u_i - P u_i + \epsilon. \end{aligned}$$

It can similarly be proved that $P_n f - Pf \geq P_n \ell_i - P\ell_i - \epsilon$. Both these inequalities together imply (98) which completes the proof of Proposition 12.3. □

We shall now state the analogue of (95) involving bracketing numbers. This will be our second main result for bounding the expected suprema of empirical processes (the first main result being (95)).

Theorem 12.4. *Let F be an envelope for the class $\mathcal{F}$ such that $PF^2 < \infty$. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sqrt{n} |P_n f - Pf| \right) \leq C \left\| F \right\|_{L^2(P)} J_{[]}(F, \mathcal{F}) \tag{99}$$

where

$$J_{[]}(F, \mathcal{F}) := \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon \left\| F \right\|_{L^2(P)}, \mathcal{F}, L^2(P))} d\epsilon.$$

The bound (99) is very similar to (95) the only difference being that the "uniform" covering numbers $\sup_Q M(\epsilon \left\| F \right\|_{L^2(Q)}, \mathcal{F}, L^2(Q))$ are replaced by the bracketing numbers $N_{[]}(\epsilon \left\| F \right\|_{L^2(P)}, \mathcal{F}, L^2(P))$ with respect to $L^2(P)$. Importantly, note that there is supremum over Q in (99) and the bracketing numbers involving only the measure P. In contrast, the bound (95) would be false if $\sup_Q M(\epsilon \left\| F \right\|_{L^2(Q)}, \mathcal{F}, L^2(Q))$ is replaced by $M(\epsilon \left\| F \right\|_{L^2(P)}, \mathcal{F}, L^2(P))$.

Example 12.5. *Suppose $X_1, \ldots, X_n$ are i.i.d observations having cdf F and let F_n be the empirical cdf. We have seen previously that*

$$\mathbb{E} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \frac{C}{\sqrt{n}}$$

for every $n \geq 1$. This was deduce as a consequence of (95). We shall show here that this can also be deduced via (99). Indeed for $\mathcal{F} := \{I_{(-\infty, t]} : t \in \mathbb{R}\}$, we have obtained bounds for $N_{[]}(\epsilon, \mathcal{F}, L^2(P))$ in (96). We deduce from these and (99) that

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \leq \frac{C}{\sqrt{n}} \int_0^1 \sqrt{1 + \log \left(1 + \frac{1}{\epsilon^2} \right)} d\epsilon \leq \frac{C}{\sqrt{n}}.$$

The following presents a situation where bounding the bracketing numbers is much more tractable compared to bounding the uniform covering numbers.

Proposition 12.6. *Let $\Theta \subseteq \mathbb{R}^d$ be contained in a ball of radius R. Let $\mathcal{F} := \{m_\theta : \theta \in \Theta\}$ be a function class indexed by Θ. Suppose there exists a function M with $\left\| M \right\|_{L^2(P)} < \infty$ such that*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq M(x) \left\| \theta_1 - \theta_2 \right\| \tag{100}$$

for all $x \in \mathcal{X}$ and $\theta_1, \theta_2 \in \Theta$ (here $\left\| \cdot \right\|$ denotes the usual Euclidean norm). Then for every $\epsilon > 0$,

$$N_{[]}(\epsilon \left\| M \right\|_{L^2(P)}, \mathcal{F}, L^2(P)) \leq \left(1 + \frac{4R}{\epsilon} \right)^d. \tag{101}$$

Proof. Let $\theta_1, \ldots, \theta_N$ be a maximal $\epsilon/2$-packing subset of Θ in the Euclidean metric. Consider the brackets $[m_{\theta_i} - \epsilon M/2, m_{\theta_i} + \epsilon M/2]$ for $i = 1, \ldots, N$. Note that

1. These brackets cover $\mathcal{F}$. Indeed, for every $\theta \in \Theta$, there exists $1 \leq i \leq N$ with $\left\| \theta - \theta_i \right\| \leq \epsilon/2$. Then by the condition (100),
$$|m_\theta(x) - m_{\theta_i}(x)| \leq M(x) \left\| \theta - \theta_i \right\| \leq \frac{\epsilon M(x)}{2}$$
which implies that m_θ lies in the bracket $[m_{\theta_i} - \epsilon M/2, m_{\theta_i} + \epsilon M/2]$.
2. The $L^2(P)$-size of these brackets is at most ϵ. This is obvious.

Because of these two observations, $N_{[]}(\epsilon \left\| M \right\|_{L^2(P)}, \mathcal{F}, L^2(P))$ is bounded from above the Euclidean $\epsilon/2$-packing number of Θ which we bounded previously. This completes the proof of Proposition 12.6. □

12.2 M-estimation

We shall now come to the first statistics topic of the course: M-estimation. The basic abstract setting is the following.

Let Θ be an abstract parameter space. Usually, it is a subset of $\mathbb{R}^d$ for parametric estimation problems or it is a function class for nonparametric estimation problems. We have two processes (one stochastic and one deterministic) that are indexed by $\theta \in \Theta$. The stochastic process will usually depend on a "sample" size n and will be denoted by $M_n(\theta), \theta \in \Theta$. The deterministic process will usually not depend on n and will simply be denoted by $M(\theta), \theta \in \Theta$. We expect M_n to be close to M for large n.

Let $\hat{\theta}_n$ denote a maximizer of $M_n(\theta)$ over $\theta \in \Theta$ and let θ_0 denote a maximizer of $M(\theta)$ over $\theta \in \Theta$. The goal in M-estimation is to study the behavior of $\hat{\theta}_n$ in relation to θ_0.

Some concrete M-estimators are described below.

1. **Classical Parametric Estimation**: The most classical M-estimator is the Maximum Likelihood Estimator (MLE). Here one typically has data $X_1, \ldots, X_n$ in $\mathcal{X}$ that are i.i.d having distribution P. One also has a class $\{p_\theta, \theta \in \Theta\}$ of densities over the space. The MLE maximizes $M_n(\theta) := P_n \log p_\theta$ over $\theta \in \Theta$. The process $M(\theta)$ here is $M(\theta) := P \log p_\theta$ and θ_0 can then be taken to the parameter value in Θ for which p_θ is closest to P in terms of the Kullback-Leibler divergence.

   More generally, one can take $M_n(\theta) = P_n m_\theta$ and $M(\theta) = P m_\theta$ for other functions m_θ. For example, $m_\theta(x) := |x-\theta|$ corresponds to median estimation (here $\hat{\theta}_n$ is the sample median and θ_0 is the population median) and $m_\theta(x) := I\{|x-\theta| \leq 1\}$ can be taken to correspond to mode estimation.

2. **Least Squares Estimators in Regression**: In regression problems, one observes data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$ which can be modeled as i.i.d observations having some distribution P. Let Θ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The least squares estimator over the class Θ corresponds to the maximizer of

$$M_n(\theta) := -P_n(y - \theta(x))^2$$

   over $\theta \in \Theta$. It is natural to compare this $\hat{\theta}_n$ to θ_0 which is the maximizer of

$$M(\theta) := -P(y - \theta(x))^2.$$

3. **Empirical Risk Minimization Procedures in Classification**: Here one observes data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i \in \mathcal{X}$ and $Y_i \in \{-1, +1\}$. We model the data as i.i.d having a distribution P. Let Θ denote a class of functions from $\mathcal{X}$ to $\mathbb{R}$; we are thinking of the sign of $\theta(x)$ as the output of the classifier. It is natural to consider

$$M_n(\theta) := -P_n I\{y \neq sign(\theta(x))\} \quad \text{and} \quad M(\theta) := -P\{y \neq sign(\theta(x))\}.$$

   In this case, $\hat{\theta}_n$ will be the empirical minimizer of the misclassification rate and θ_0 will be the minimizer of the test error, both in the class Θ. It is therefore natural to compare the performance of $\hat{\theta}_n$ to that of θ_0.

   Note that it is difficult to compute $\hat{\theta}_n$ as the minimization of $M_n(\theta)$ is a combinatorial problem. For this, one also studies other choices of $M_n(\theta)$ in classification. To motivate these other choices, let us first rewrite the above $M_n(\theta)$ as

$$M_n(\theta) = -P_n I\{y \neq sign(\theta(x))\} = -P_n \phi_0(-y\theta(x)) \qquad \text{where } \phi_0(t) := I\{t \geq 0\}.$$

   For computational considerations, one often replaces ϕ_0 by another loss function ϕ that is *convex* and *similar* to ϕ_0. Common choices of ϕ include (a) Hinge loss: $\phi(t) := (1+t)_+$, (b) Exponential loss: $\phi(t) := \exp(t)$, and (c) Logistic loss: $\phi(t) := \log(1 + e^t)$. Note that these three functions are convex

on $\mathbb{R}$ and they are similar to ϕ_0 (note that they also satisfy $\phi(t) \geq \phi_0(t)$ for al t). We shall study procedures $\hat{\theta}_n$ which minimize

$$M_n(\theta) := -P_n\phi(-y\theta(x)) \qquad \text{over } \theta \in \Theta$$

and compare their performance to θ_0.

The theory of M-estimation concerns itself usually with three questions: (a) Consistency, (b) Rate of Convergence, and (c) Limiting Behavior. Consistency asserts that the discrepancy between $\hat{\theta}_n$ and θ_0 converges to zero as $n \to \infty$. Rate of convergence aims to characterize the precise rate of this convergence. The goal of the third question will be to give a precise characterization of the limiting distribution of the discrepancy in the asymptotic setting where $n \to \infty$.

Consistency usually always holds and we have already seen a theorem last week on consistency. We shall mainly concentrate on the problem of rates of convergence. In many cases, a rate of convergence result automatically implies consistency. In other cases, one needs a preliminary consistency result so that attention can be focused in a local neighbourhood of θ_0 in order to determine the rate of convergence. In cases where preliminary consistency is required and our consistency theorem last week is not sufficient, we shall provide a different argument for consistency. Let us ignore consistency for the time being and proceed directly to the rates. For studying limiting behavior, we need theory on uniform central limit theorems which we are yet to cover; we shall come back to these in a few weeks.

12.3 Rates of Convergence of M-estimators

It is cleanest to work in the abstract setting where $\hat{\theta}_n$ maximizes a stochastic process $M_n(\theta)$ over $\theta \in \Theta$ and θ_0 maximizes a deterministic process $M(\theta)$ over $\theta \in \Theta$. The argument for deriving rates starts from the following basic inequaility:

$$M(\theta_0) - M(\hat{\theta}_n) \leq \left[M_n(\hat{\theta}_n) - M(\hat{\theta}_n)\right] - [M_n(\theta_0) - M(\theta_0)]$$

We have already seen this inequality multiple times and it is a consequence of the simple inequality $M_n(\hat{\theta}_n) \geq M_n(\theta_0)$. For convenience, we shall denote the right hand side above by $(M_n - M)(\hat{\theta}_n - \theta_0)$ so that

$$M(\theta_0) - M(\hat{\theta}_n) \leq (M_n - M)(\hat{\theta}_n - \theta_0). \tag{102}$$

We shall use this inequality to study rates of convergence of $\hat{\theta}_n$ to θ_0. We need to first fix a measure of discrepancy between $\hat{\theta}_n$ and θ_0. Let this be given by $d(\hat{\theta}_n, \theta_0)$. In cases where Θ is a subset of $\mathbb{R}^d$, it is natural to take $d(\cdot, \cdot)$ as the usual Euclidean metric. In general, we shall not require that $d(\cdot, \cdot)$ is a metric; at this stage, we only require it to be nonnegative.

Note that the discrepancy measure $d(\cdot, \cdot)$ is somewhat external to the problem and, therefore, to understand the behavior of $d(\hat{\theta}_n, \theta_0)$, we need to connect it to $M(\theta)$ or $M_n(\theta)$. The usual assumption for this is to assume that:

$$M(\theta_0) - M(\theta) \gtrsim d^2(\theta, \theta_0). \tag{103}$$

Here the notation $a \gtrsim b$ means that $a \geq Cb$ for a universal constant C (the notation $a \lesssim b$ is defined analogously).

Let us assume that (103) is true for all $\theta \in \Theta$. In some situations, it is only true in a neighborhood of θ_0 (we can come back to this later). Note that (103) is automatically true if we define d as

$$d^2(\theta, \theta_0) := M(\theta_0) - M(\theta).$$

This is the most natural choice for studying rates of M-estimators. In parametric estimation problems, this usually does not correspond to the Euclidean metric so this choice is not usually used. However in function estimation problems, this is a very common choice.

Combining (102) and (103), we obtain

$$d^2(\hat{\theta}_n, \theta_0) \lesssim (M_n - M)(\hat{\theta}_n - \theta_0).$$

Let $\hat{\delta}_n := d(\hat{\theta}_n, \theta_0)$. Then the above inequality clearly implies

$$\hat{\delta}_n^2 \lesssim \sup_{\theta \in \Theta : d(\theta, \theta_0) \leq \hat{\delta}_n} (M_n - M)(\theta - \theta_0)$$

This suggests that $\hat{\delta}_n \lesssim \delta_n$ for any rate δ_n that satisfies

$$\delta_n^2 \lesssim \mathbb{E} \sup_{\theta \in \Theta : d(\theta, \theta_0) \leq \delta_n} (M_n - M)(\theta - \theta_0). \tag{104}$$

We shall rigorize this intuition in the next class. The critical inequality (104) gives a nice way to determine the rate of convergence of M-estimators in a variety of problems. The expectation on the right hand side can be controlled via the empirical process methods that we have studied in the past many weeks.

13 Lecture 13

13.1 Rigorous Derivation of Rates of Convergence of M-estimators

Let us recall the setup. Θ is an abstract parameter space. We have two processes (one stochastic and one deterministic) that are indexed by $\theta \in \Theta$. The stochastic process will usually depend on a "sample" size n and will be denoted by $M_n(\theta), \theta \in \Theta$. The deterministic process will usually not depend on n and will simply be denoted by $M(\theta), \theta \in \Theta$. We expect M_n to be close to M for large n.

Let $\hat{\theta}_n$ denote a maximizer of $M_n(\theta)$ over $\theta \in \Theta$ and let θ_0 denote a maximizer of $M(\theta)$ over $\theta \in \Theta$. Let $d(\hat{\theta}_n, \theta_0)$ be a nonnegative discrepancy measure gauging the gap between $\hat{\theta}_n$ and θ_0. We shall assume that

$$M(\theta_0) - M(\theta) \gtrsim d^2(\theta, \theta_0) \tag{105}$$

for every $\theta \in \Theta$. Here the notation $a \gtrsim b$ means that $a \geq Cb$ for a universal positive constant C (the notation $a \lesssim b$ is defined analogously). In light of (105), the canonical choice for d will be

$$d(\theta, \theta_0) := \sqrt{M(\theta_0) - M(\theta)}. \tag{106}$$

When d is not the canonical choice above, it usually happens that (105) holds only in a neighbourhood of θ_0. We shall come back to this situation later.

We shall now rigorously find upper bounds for the rate of convergence of $d(\hat{\theta}_n, \theta_0)$. Formally, we say that δ_n is a rate of convergence of $d(\hat{\theta}_n, \theta_0)$ to zero if for every $\epsilon > 0$, there exists a constant C_ϵ such that

$$d(\hat{\theta}_n, \theta_0) \leq C_\epsilon \delta_n \qquad \text{with probability } \geq 1 - \epsilon. \tag{107}$$

Note that this is equivalent to

$$\mathbb{P}\left\{ d(\hat{\theta}_n, \theta_0) > 2^M \delta_n \right\} \to 0 \qquad \text{as } M \to \infty. \tag{108}$$

It should be noted that (107) and (108) are nonasymptotic statements (they hold for each finite n). They imply, in particular, the asymptotic rate statement: $d(\hat{\theta}_n, \theta_0) = O_P(\delta_n)$ which means the following: For every $\epsilon > 0$, there exists C_ϵ and an integer N_ϵ such that

$$\mathbb{P}\{d(\hat{\theta}_n, \theta_0) \leq C_\epsilon \delta_n\} \geq 1 - \epsilon \qquad \text{for all } n \geq N_\epsilon. \tag{109}$$

The difference between the (109) and (107) is that (109) holds for all $n \geq N_\epsilon$ while (107) holds for all n.

Let us now study the probability

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\right\}$$

for fixed δ_n and large M. We need to understand for which δ_n does this probability become small as $M \to \infty$.

Write

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\right\} = \sum_{j>M} \mathbb{P}\left\{2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n\right\}.$$

We shall now use the basic inequality (together with the condition (105)):

$$d^2(\hat{\theta}_n, \theta_0) \lesssim M(\theta_0) - M(\hat{\theta}_n) \leq (M_n - M)(\hat{\theta}_n - \theta_0)$$

This clearly gives

$$\begin{aligned}
\mathbb{P}\left\{2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n\right\} &\leq \mathbb{P}\left\{(M_n - M)(\hat{\theta}_n - \theta_0) \gtrsim 2^{2j-2}\delta_n^2, d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n\right\} \\
&\leq \mathbb{P}\left\{\sup_{\theta: d(\theta,\theta_0) \leq 2^j \delta_n} (M_n - M)(\theta - \theta_0) \gtrsim 2^{2j-2}\delta^2\right\} \\
&\lesssim \frac{1}{2^{2j-2}\delta^2} \mathbb{E} \sup_{\theta: d(\theta,\theta_0) \leq 2^j \delta} (M_n - M)(\theta - \theta_0).
\end{aligned}$$

Suppose that the function $\phi_n(\cdot)$ is such that

$$\mathbb{E} \sup_{\theta: d(\theta,\theta_0) \leq u} (M_n - M)(\theta - \theta_0) \lesssim \phi_n(u) \qquad \text{for every } u. \tag{110}$$

We thus get

$$\mathbb{P}\left\{2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n\right\} \lesssim \frac{\phi_n(2^j \delta_n)}{2^{2j}\delta_n^2}$$

for every j. As a consequence,

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\right\} \lesssim \sum_{j>M} \frac{\phi_n(2^j \delta_n)}{2^{2j}\delta_n^2}.$$

The following assumption on $\phi_n(\cdot)$ is usually made to simplify the expression above: There exists $\alpha < 2$ such that

$$\phi_n(cx) \leq c^\alpha \phi_n(x) \qquad \text{for all } c > 1 \text{ and } x > 0. \tag{111}$$

Under this assumption, we get

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\right\} \lesssim \frac{\phi_n(\delta_n)}{\delta_n^2} \sum_{j>M} 2^{j(\alpha-2)}.$$

The quantity $\sum_{j>M} 2^{j(\alpha-2)}$ converges to zero as $M \to \infty$. Therefore if δ_n is such that

$$\phi_n(\delta_n) \lesssim \delta_n^2,$$

then

$$d(\hat{\theta}_n, \theta_0) \leq 2^M \delta_n \qquad \text{with probability at least } 1 - u_M$$

where $u_M \to 0$ as $M \to \infty$.

This gives us the following nonasymptotic rate of convergence theorem:

Theorem 13.1. *Assume the condition* (105) *and that the function* $\phi_n(\cdot)$ *satisfies* (110) *and* (111). *Then for every* $M > 0$, *we get* $d(\hat{\theta}_n, \theta_0) \leq 2^M \delta_n$ *with probability at least* $1 - u_M$ *provided* $\phi_n(\delta_n) \lesssim \delta_n^2$. *Here* $u_M = \sum_{j>M} 2^{j(\alpha-2)} \to 0$ *as* $M \to \infty$.

13.2 Application to Bounded Lipschitz Regression

Suppose f_0 is an unknown function on $[0,1]$. We observe data $Y_1, \dots, Y_n$ on f_0 that are generated according to the model:

$$Y_i = f_0(i/n) + \epsilon_i \qquad \text{for } i = 1, \dots, n$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d $N(0,1)$ random variables.

Suppose that we assume that f_0 is 1-Lipschitz on and that it is bounded by 1 on $[0,1]$. In other words we assume that $f_0 \in \mathcal{F}$ where $\mathcal{F}$ is the collection of all functions on $[0,1]$ that are bounded in absolute value by 1 and that are 1-Lipschitz. Under this assumption, it is reasonable to estimate f_0 by

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(i/n)\right)^2.$$

$\hat{f}_n$ is clearly an M-estimator and we can use Theorem 13.1 to study some aspects of its behavior. For this, we first write (using $Y_i = f_0(i/n) + \epsilon_i$):

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(i/n)\right)^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(\epsilon_i + f_0(i/n) - f(i/n)\right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(f_0(i/n) - f(i/n)\right)^2 - \frac{2}{n} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - f_0(i/n)\right) + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2.\end{aligned}$$

As a result

$$\hat{f}_n = \operatorname*{argmax}_{f \in \mathcal{F}} \left(\frac{2}{n} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - f_0(i/n)\right) - \frac{1}{n} \sum_{i=1}^{n} \left(f_0(i/n) - f(i/n)\right)^2 \right).$$

In order to use Theorem 13.1, we can thus take

$$M_n(f) := \frac{2}{n} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - f_0(i/n)\right) - \frac{1}{n} \sum_{i=1}^{n} \left(f_0(i/n) - f(i/n)\right)^2.$$

It is then natural to take

$$M(f) := \mathbb{E} M_n(f) = -\frac{1}{n} \sum_{i=1}^{n} \left(f_0(i/n) - f(i/n)\right)^2.$$

For the discrepancy d, we can use the canonical choice (106):

$$d(f, f_0) := \sqrt{M(f_0) - M(f)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(f_0(i/n) - f(i/n)\right)^2}.$$

To find the rate, we need to control

$$\mathbb{E} \sup_{f \in \mathcal{F} : d(f, f_0) \le \delta} (M_n - M)(f - f_0) = 2\mathbb{E} \sup_{f \in \mathcal{F} : d(f, f_0) \le \delta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - f_0(i/n)\right).$$

For this, we can use Dudley's entropy bound. Let

$$X_f := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - f_0(i/n)\right)$$

and note that

$$X_f - X_g = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - g(i/n) \right) \sim N(0, d^2(f,g)).$$

This implies that

$$\mathbb{P}\left\{ |X_f - X_g| \geq u \right\} \leq 2 \exp\left(\frac{-u^2}{2d^2(f,g)} \right) \qquad \text{for all } u \geq 0.$$

Dudley's entropy bound then immediately gives

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}: d(f,f_0) \leq \delta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - f_0(i/n) \right) &\leq \frac{1}{\sqrt{n}} \mathbb{E} \sup_{f \in \mathcal{F}: d(f,f_0) \leq \delta} |X_f - X_{f_0}| \\
&\leq \frac{C}{\sqrt{n}} \int_0^{\delta} \sqrt{\log M(\epsilon, \{f \in \mathcal{F} : d(f,f_0) \leq \delta\}, d)} d\epsilon \\
&\leq \frac{C}{\sqrt{n}} \int_0^{\delta} \sqrt{\log M(\epsilon, \mathcal{F}, d)} d\epsilon.
\end{aligned}$$

We have previously noted that

$$M(\epsilon, \mathcal{F}, d) \leq M(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \exp\left(\frac{C}{\epsilon} \right).$$

This gives

$$\mathbb{E} \sup_{f \in \mathcal{F}: d(f,f_0) \leq \delta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - f_0(i/n) \right) \leq \frac{C}{\sqrt{n}} \int_0^{\delta} \sqrt{\log M(\epsilon, \mathcal{F}, d)} d\epsilon \leq \frac{C}{\sqrt{n}} \int_0^{\delta} \sqrt{\frac{C}{\epsilon}} d\epsilon \lesssim \sqrt{\frac{\delta}{n}}.$$

We can thus take $\phi_n(\delta) := \sqrt{\delta/n}$ in Theorem 13.1. Note that this clearly satisfies the condition $\phi_n(cx) \leq c^\alpha \phi_n(x)$ with $\alpha = 1/2 < 2$. The critical rate determining equation then becomes:

$$\phi_n(\delta) = \sqrt{\frac{\delta}{n}} \lesssim \delta^2$$

which gives $\delta_n \gtrsim n^{-1/3}$. Thus Theorem 13.1 is valid here with $\delta_n = n^{-1/3}$ which allows us to deduce the following:

$$\frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_n(i/n) - f_0(i/n) \right)^2 = O_P(n^{-2/3})$$

or, more specifically,

$$\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_n(i/n) - f_0(i/n) \right)^2 > 2^{2M} n^{-2/3} \right\} \lesssim 2^{-M}.$$

We have therefore proved that the rate of convergence in Lipschitz regression is $n^{-2/3}$. We shall prove later that, in a minimax or worst case sense (worst case over all functions $f_0 \in \mathcal{F}$), this $n^{-2/3}$ rate cannot be improved by any other estimator. In other words, $n^{-2/3}$ is the minimax optimal rate of convergence for Lipschitz functions on $[0,1]$.

More generally, suppose we now assume that f_0 is in the smoothness class $\mathcal{S}_\alpha$ that we previously defined. In that case, the same argument as above leads to the inequality:

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{S}_\alpha: d(f,f_0) \leq \delta} (M_n - M)(f - f_0) &= 2\mathbb{E} \sup_{f \in \mathcal{S}_\alpha: d(f,f_0) \leq \delta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left(f(i/n) - f_0(i/n) \right) \\
&\leq \frac{C}{\sqrt{n}} \int_0^{\delta} \sqrt{\log M(\epsilon, \mathcal{S}_\alpha, d)} d\epsilon \\
&\leq \frac{C}{\sqrt{n}} \int_0^{\delta} \left(\frac{1}{\epsilon} \right)^{1/(2\alpha)}. \qquad (112)
\end{aligned}$$

Suppose now that $\alpha > 1/2$. In that case the integral above is finite and we get

$$\mathbb{E} \sup_{f \in \mathcal{S}_\alpha : d(f, f_0) \leq \delta} (M_n - M)(f - f_0) \lesssim \frac{1}{\sqrt{n}} \delta^{1-(1/(2\alpha))}$$

We can thus take

$$\phi_n(\delta) := \frac{1}{\sqrt{n}} \delta^{1-(1/(2\alpha))}$$

so that the critical inequality for determining the rate becomes

$$\frac{1}{\sqrt{n}} \delta_n^{1-(1/(2\alpha))} \lesssim \delta_n^2$$

which gives

$$\delta_n \gtrsim n^{-\alpha/(2\alpha+1)}$$

leading us to the conclusion

$$\frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_n(i/n) - f_0(i/n) \right)^2 = O_P(n^{-2\alpha/2\alpha+1})$$

It turns out that $n^{-2\alpha/(2\alpha+1)}$ is indeed the minimax optimal rate of estimation of functions in $\mathcal{S}_\alpha$.

Suppose now that $\alpha \leq 1/2$. In this case, the integral in (112) is infinite so Dudley's bound in the form that we used does not give us anything useful. In this case, one can use a modification of Dudley's bound where the lower limit in the integral is not *zero* but strictly positive (this is Problem 6 Homework 3). However the resulting rate for $d^2(\hat{f}_n, f_0)$ will be slower than $n^{-2\alpha/(2\alpha+1)}$. It is not known if the least squares estimator over $\mathcal{S}_\alpha$ is minimax optimal for $\alpha \leq 1/2$.

13.3 Back to the rate theorem

Now let us get back to Theorem 13.1. For proving the theorem, we assumed that

$$M(\theta) - M(\theta_0) \lesssim -d^2(\theta, \theta_0) \tag{113}$$

for all $\theta \in \Theta$. We also assumed that

$$\mathbb{E} \sup_{\theta \in \Theta : d(\theta, \theta_0) \leq u} (M_n - M)(\theta - \theta_0) \lesssim \phi_n(u) \tag{114}$$

for all $u > 0$. Here $\phi_n(u)$ is some function on $(0, \infty)$ which satisfies $\phi_n(cx) \leq c^\alpha \phi_n(x)$ for some $\alpha < 2$.

Under these two assumptions, Theorem 13.1 asserted that $d(\hat{\theta}_n, \theta_0) = O_P(\delta_n)$ for every δ_n that satisfies $\phi_n(\delta_n) \lesssim \delta_n^2$.

In some situations, it is not possible to ensure that (113) holds for all $\theta \in \Theta$. It is also not possible to ensure that (114) holds for all $u > 0$. On the contrary, it is usually possible to ensure the existence of a positive real number u^* (not depending on n) such that (113) holds for all $\theta \in \Theta$ with $d(\theta, \theta_0) \leq u^*$ and such that (114) holds for all $u \leq u^*$. In that case, it is still possible to assert that $d(\hat{\theta}_n, \theta_0) = O_P(\delta_n)$ under the additional assumption that $d(\hat{\theta}_n, \theta_0)$ converges in probability to 0. This is the content of the following theorem (which is Theorem 3.2.5 in Van der Vaart and Wellner [25]).

Theorem 13.2. *Let u^* be a strictly positive real number (not depending on n) such that (113) holds for all $\theta \in \Theta$ with $d(\theta, \theta^*) \leq u^*$. Let ϕ_n be a function on $(0, \infty)$ which satisfies the condition for some $\alpha < 2$: $\phi_n(cx) \leq c^\alpha \phi_n(x)$ for all $c > 1$ and $x > 0$. Suppose that (114) holds for all $0 < u \leq u^*$. Assume that $d(\hat{\theta}_n, \theta_0) = O_P(1)$. Then $d(\hat{\theta}_n, \theta_0) = O_P(\delta_n)$ for every δ_n satisfying $\phi_n(\delta_n) \lesssim \delta_n^2$.*

Proof. Write

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\right\} \leq \sum_{j > M : 2^j \delta_n \leq u^*} \mathbb{P}\left\{2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n\right\} + P\left\{2d(\hat{\theta}_n, \theta_0) > u^*\right\}.$$

The first term can be bounded in exactly the same way as in the proof of Theorem 13.1. This gives

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\right\} \lesssim \frac{\phi_n(\delta_n)}{\delta_n^2} \sum_{j > M} 2^{j(\alpha-2)} + \mathbb{P}\left\{2d(\hat{\theta}_n, \theta_0) > u^*\right\}.$$

If δ_n is chosen such that $\phi_n(\delta_n) \lesssim \delta_n^2$, the first term above converges to zero as $M \to \infty$. The second term, on the other hand, converges to 0 as $n \to \infty$ by the assumption that $d(\hat{\theta}_n, \theta_0)$ converges to zero in probability. This concludes the proof of Theorem 13.2. □

14 Lecture 14

14.1 Recap of the main rate theorem

$\hat{\theta}_n$ maximizes a stochastic process $M_n(\theta)$ over $\theta \in \Theta$ while θ_0 maximizes a deterministic process $M(\theta)$ over $\theta \in \Theta$. Assume that $d(\cdot, \cdot)$ is such that

$$M(\theta) - M(\theta_0) \lesssim -d^2(\theta, \theta_0)$$

for all $\theta \in \Theta$. Let $\phi_n(\cdot)$ be a function satisfying a mild condition (there exists $\alpha < 2$ such that $\phi_n(cx) \leq c^\alpha \phi_n(x)$) such that

$$\mathbb{E} \sup_{\theta \in \Theta : d(\theta, \theta_0) \leq u} (M_n - M)(\theta - \theta_0) \lesssim \phi_n(u)$$

for all $u > 0$. Then the random quantity $d(\hat{\theta}_n, \theta_0)$ will be controlled by δ_n for every δ_n satisfying

$$\phi_n(\delta_n) \lesssim \delta_n^2.$$

Formally, we have

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\right\} \lesssim \sum_{j > M} 2^{j(\alpha-2)}. \tag{115}$$

Here are some strengths and weaknesses of this theorem. Let us start with the strengths:

1. It rigorizes the heuristic argument well. Indeed, the heuristic argument starts with the basic inequality:

   $$d^2(\hat{\theta}_n, \theta_0) \lesssim M(\theta_0) - M(\hat{\theta}_n) \leq (M_n - M)(\hat{\theta}_n - \theta_0).$$

   From here, it is easy to derive that $\hat{\delta}_n := d(\hat{\theta}_n, \theta_0)$ satisfies

   $$\left(\hat{\delta}_n\right)^2 \lesssim \sup_{\theta \in \Theta : d(\hat{\theta}_n, \theta_0) \leq \hat{\delta}_n} (M_n - M)(\theta - \theta_0).$$

   From here, it is reasonable to conjecture that the $d(\hat{\theta}_n, \theta_0)$ will be controlled by any δ_n satisfying

   $$\delta_n^2 \lesssim \mathbb{E} \sup_{\theta \in \Theta : d(\theta, \theta_0) \leq \delta_n} (M_n - M)(\theta - \theta_0).$$

   The theorem basically proves this if the $\lesssim$ inequality above is changed to $\gtrsim$. This also suggests that the rate obtained by solving $\phi_n(\delta_n) \sim \delta_n^2$ should be the correct rate for $d(\hat{\theta}_n, \theta_0)$.

2. It is very general. The theorem is quite general and applies to a variety of problems. We have already seen examples of this and we shall some more examples in the near future.

3. It is very simple to prove. The proof is only a few lines long and does not use any complicated machinery.

Here are some important weaknesses of the rate theorem.

1. The most important weakness is that, although the rate obtained is usually correct, the deviation inequaiity (115) is usually quite weak. To see this, observe that when $\alpha = 1$ (for example), inequality (115) becomes

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\right\} \lesssim 2^{-M}$$

from which it follows that

$$\mathbb{P}\left\{d(\hat{\theta}_n, \theta_0) > t\delta_n\right\} \lesssim \frac{1}{t}$$

for all t. This inequality is quite weak in the sense that it does not even imply that

$$\mathbb{E}d^2(\hat{\theta}_n, \theta_0) \lesssim \delta_n^2.$$

The main reason for the looseness comes from the use of Markov's inequality in the proof:

$$\mathbb{P}\left\{\sup_{\theta\in\Theta: d(\theta,\theta_0)\leq 2^j\delta_n} (M_n - M)(\theta - \theta_0) \gtrsim 2^{2j-2}\delta_n^2\right\} \leq \frac{1}{2^{2j-2}\delta_n^2}\mathbb{E}\sup_{\theta\in\Theta: d(\theta,\theta_0)\leq 2^j\delta_n} (M_n - M)(\theta - \theta_0).$$

This inequality is quite loose. More sophisticated arguments (under more specialized settings) can be used in place of this and these give improved bounds for $\mathbb{P}\{d(\hat{\theta}_n, \theta_0) > t\delta_n\}$. We shall see some examples of such improved results later.

2. Calculating the rate via the theorem requires one to bound

$$\mathbb{E}\sup_{\theta\in\Theta: d(\theta,\theta_0)\leq u} (M_n - M)(\theta - \theta_0).$$

Although there exist general techniques for this, getting good bounds in specific situations can still be quite hard.

We shall next apply the rate theorem for understanding the loss behavior of convex penalized least squares estimators in the Gaussian sequence model. Before that, let us first introduce the Gaussian sequence model.

14.2 The Gaussian Sequence Model

The (finite) Gaussian Sequence Model is an important model for studying the theoretical performance of many commonly used statistical procedures. For a comprehensive treatment of estimation theory under the Gaussian sequence model, see Johnstone [11].

Suppose we observe real-valued observations $Y_1, \ldots, Y_n$. Under the Gaussian sequence model, we model the observed data as

$$Y_i = \theta_i^* + \epsilon_i \qquad \text{for } i = 1, \ldots, n$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d $N(0, 1)$. In other words, $\epsilon \sim N(0, I_n)$ and $Y \sim N(\theta^*, I_n)$. The goal is to estimate $\theta_1^*, \ldots, \theta_n^*$ from data $Y_1, \ldots, Y_n$ under the loss function:

$$\frac{1}{n}\sum_{i=1}^n \left(\hat{\theta}_i - \theta_i^*\right)^2 = \frac{1}{n}\left\|\hat{\theta} - \theta^*\right\|^2.$$

The *risk* of an estimator $\hat{\theta}$ will be defined as

$$R(\hat{\theta}, \theta^*) := \frac{1}{n}\mathbb{E}\left\|\hat{\theta} - \theta^*\right\|^2.$$

Usually one imposes some additional structure on the unknown parameters $\theta_1^*, \ldots, \theta_n^*$. Here are some standard examples of such additional structure:

1. **No Structure**: In this case, no information is available on $\theta_1^*, \ldots, \theta_n^*$. It is intuitively clear then that one cannot do much better than the simple estimator $\hat{\theta}_i = Y_i, i = 1, \ldots, n$. The risk of this estimator clearly equals 1. It turns out that

   $$\inf_{\tilde{\theta}} \sup_{\theta \in \mathbb{R}^n} R(\tilde{\theta}, \theta) = 1.$$

   This means that the simple estimator Y is minimax optimal over $\mathbb{R}^n$ (or equivalently, the worst case risk over $\theta \in \mathbb{R}^n$ of every other estimator is at least 1).

2. **Fixed Linear Subspace**: Here it is assumed that $\theta^* \in S$ for a known linear subspace S. In this case, the most natural estimator is the projection of Y onto S. This estimator has risk k/n where k is the dimension of S. The minimax risk over S equals k/n i.e.,

   $$\inf_{\tilde{\theta}} \sup_{\theta \in S} R(\tilde{\theta}, \theta) = \frac{k}{n}.$$

3. **Smoothness**: Suppose $\theta^* = (f^*(1/n), \ldots, f^*(1))$ for some function $f^* : [0,1] \to [-1,1]$ which is 1-Lipschitz. In the last class, we saw that if we consider the estimator $\hat{\theta} = (\hat{f}(1/n), \ldots, \hat{f}(1))$ where $\hat{f}$ is any least squares estimator over the class of all 1-Lipschitz functions that are bounded by 1, then

   $$\frac{1}{n}\left\|\hat{\theta} - \theta^*\right\| = O_P(n^{-2/3}).$$

   It is possible to also prove that the above bound also holds in expectation i.e.,

   $$R(\hat{\theta}, \theta^*) \lesssim n^{-2/3}.$$

   It turns out that $n^{-2/3}$ is actually the minimax risk over the class of all vectors $(f(1/n), \ldots, f(1))$ where $f : [0,1] \to [-1,1]$ is 1-Lipschitz. We shall prove this later. The estimator $\hat{\theta}$ is actually non-linear. However it is possible to achieve the risk $n^{-2/3}$ also with a linear estimator based on local averaging of $Y_1, \ldots, Y_n$.

4. **Sparsity**: Suppose the vector θ^* is sparse in the sense that only a few of its entries are non-zero. More precisely assume that $\theta^* \in \Theta_k$ where Θ_k is the class of all vectors in $\mathbb{R}^n$ which have at most k non-zero entries. Assume that k is small compared to n (specifically, assume that $k/n \to 0$ as $n \to \infty$). In this case, there exist estimators $\hat{\theta}$ which satisfy

   $$\sup_{\theta \in \Theta_k} R(\hat{\theta}, \theta^*) = \frac{2k}{n} \log \frac{n}{k} \left(1 + o(1)\right). \tag{116}$$

   It can be proved that the minimax risk over Θ_k also equals the right hand side above (we shall prove this later). The estimator $\hat{\theta}$ achieving (116) can be taken to be LASSO or soft-thresholding with an appropriate choice of the tuning parameter. This is an M-estimator that can be studied via the rate theorem. We shall do this later.

5. **Low Rank Structure**: Suppose θ^* is the vectorization of a $d \times d$ matrix A^* with $n = d^2$. Suppose it is assumed that A^* is rank at most r. Then a penalized estimator based on penalizing the nuclear norm (sum of singular values) can be shown to be minimax optimal.

14.3 Convex Penalized Estimators in the Gausssian Sequence Model

In the Gaussian sequence model $Y = \theta^* + \epsilon$ with $\epsilon \sim N(0, I_n)$, a very standard class of estimators is given by estimators of the form:

$$\hat{\theta}_{\lambda,f} := \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left(\frac{1}{2} \|Y - \theta\|^2 + \lambda f(\theta) \right) \tag{117}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\lambda > 0$ is an appropriate tuning parameter. The function f will be the L^1 norm of θ when we believe that the true signal is sparse and will be the nuclear norm of the matrix corresponding to θ under a low rank assumption. Other functions f are also used (such as $f(\theta) := \sum_{i=1}^{n-1} |\theta_i - \theta_{i-1}|$ for piecewise constant structure).

The estimator (117) is obviously an M-estimator so we can use our general rate theorem to study:

$$\frac{1}{n} \sum_{i=1}^{n} \left\| \theta^* - \hat{\theta}_{\lambda,f} \right\|^2.$$

For this, we first write (using $Y = \theta^* + \epsilon$),

$$\frac{1}{2} \|Y - \theta\|^2 + \lambda f(\theta) = \frac{1}{2} \|\theta^* - \theta\|^2 - \langle \epsilon, \theta - \theta^* \rangle + \lambda f(\theta) + \|\epsilon\|^2.$$

Therefore, we can write

$$\hat{\theta}_{\lambda,f} = \operatorname*{argmax}_{\theta \in \mathbb{R}^n} \left(\langle \epsilon, \theta - \theta^* \rangle - \frac{1}{2} \|\theta - \theta^*\|^2 - \lambda f(\theta) \right).$$

We can therefore apply the rate theorem with $\Theta = \mathbb{R}^n$ and

$$M_n(\theta) := \langle \epsilon, \theta - \theta^* \rangle - \frac{1}{2} \|\theta - \theta^*\|^2 - \lambda f(\theta) \quad \text{and} \quad M(\theta) := \frac{-1}{2} \|\theta - \theta^*\|^2.$$

Note that $M(\theta)$ is maximized at θ^* and the condition

$$M(\theta) - M(\theta^*) \lesssim -d^2(\theta, \theta^*)$$

is trivially satisfied when

$$d(\theta, \theta^*) := \|\theta - \theta^*\|.$$

We can therefore apply the rate theorem which will require us to bound the expectation of

$$\sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \le u} (M_n - M)(\theta - \theta^*) = \sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \le u} (\langle \epsilon, \theta - \theta^* \rangle - \lambda f(\theta) + \lambda f(\theta^*)).$$

We now bound this term in the following way. Because f is convex, for every subgradient s of f at θ^*, we have

$$f(\theta) \ge f(\theta^*) + \langle s, \theta - \theta^* \rangle.$$

As a result,

$$\begin{aligned}
\sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \le u} (M_n - M)(\theta - \theta^*) &= \sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \le u} (\langle \epsilon, \theta - \theta^* \rangle - \lambda f(\theta) + \lambda f(\theta^*)) \\
&\le \sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \le u} (\langle \epsilon, \theta - \theta^* \rangle - \lambda \langle s, \theta - \theta^* \rangle) \\
&= \sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \le u} \langle \epsilon - \lambda s, \theta - \theta^* \rangle \\
&\le \sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \le u} \|\epsilon - \lambda s\| \, \|\theta - \theta^*\| \le u \|\epsilon - \lambda s\|.
\end{aligned}$$

where, in the last inequality above, we used the Cauchy-Schwarz inequality. Note that the above is true for every subgradient s of f at θ^*. The set of all subgradients of f at θ^* is called the subdifferential of f at θ^* and is denoted by $\partial f(\theta^*)$. Because the above chain of inequalities is true for every $s \in \partial f(\theta^*)$, we can take an infimum over $s \in \partial f(\theta^*)$ which allows us to deduce that

$$\sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \leq u} (M_n - M)(\theta - \theta^*) \leq u \inf_{s \in \partial f(\theta^*)} \|\epsilon - \lambda s\| =: u \ \text{dist}(\epsilon, \lambda \partial f(\theta^*))$$

and thus

$$\mathbb{E} \sup_{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| \leq u} (M_n - M)(\theta - \theta^*) \leq u \ \mathbb{E}\text{dist}(\epsilon, \lambda \partial f(\theta^*)).$$

The rate theorem therefore implies that $\left\| \hat{\theta}_{\lambda,f} - \theta^* \right\|$ is controlled by δ_n which solves

$$\delta_n \ \mathbb{E}\text{dist}(\epsilon, \lambda \partial f(\theta^*)) = \delta_n^2$$

which means that we can take

$$\delta_n := \mathbb{E}\text{dist}(\epsilon, \lambda \partial f(\theta^*)).$$

The precise inequality given by the rate theorem for bounding $\left\| \hat{\theta}_{\lambda,f} - \theta^* \right\|$ via δ_n is slightly weak (as mentioned earlier, this happens in quite a few situations). However, in this context, it can be proved (see Oymak and Hassibi [17]) that

$$\mathbb{E} \left\| \hat{\theta}_{\lambda,f} - \theta^* \right\|^2 \leq \mathbb{E}\text{dist}^2(\epsilon, \lambda \partial f(\theta^*)). \tag{118}$$

I will include a sketch of the proof of this inequality in the homework. Inequality (118) implies that if $\partial f(\theta^*)$ is a large set in $\mathbb{R}^n$, then the risk of $\hat{\theta}_{\lambda,f}$ will be small. Note that inequality (118) holds for every convex function f so it is applicable in a variety of situations. In the next section, we shall demonstrate its use for studying risks in sparse signal estimation.

14.3.1 Application of inequality (118) for sparse signal estimation

We shall now apply the inequality (118) to the case when $f(\theta) = |\theta_1| + \cdots + |\theta_n|$ is the L^1 norm of θ. For a fixed λ, we need to bound the squared expected distance of a standard Gaussian vector ϵ to $\lambda \partial f(\theta^*)$. The first step is to obtain a characterization of $\partial f(\theta^*)$. It is straightforward to see that $\partial f(\theta^*)$ consists of all vectors $(v_1, \ldots, v_n) \in \mathbb{R}^n$ such that

$$v_i \begin{cases} = \{1\} & \text{if } \theta_i^* > 0 \\ = \{-1\} & \text{if } \theta_i^* < 0 \\ \in [-1, 1] & \text{if } \theta_i^* = 0. \end{cases}$$

As a result, $\lambda \partial f(\theta^*)$ consists of all vectors $(v_1, \ldots, v_n) \in \mathbb{R}^n$ such that

$$v_i \begin{cases} = \{\lambda\} & \text{if } \theta_i^* > 0 \\ = \{-\lambda\} & \text{if } \theta_i^* < 0 \\ \in [-\lambda, \lambda] & \text{if } \theta_i^* = 0. \end{cases}$$

As a result,

$$\text{dist}^2(\epsilon, \lambda \partial f(\theta^*)) = \sum_{i : \theta_i^* \neq 0} (\epsilon_i - \text{sign}(\theta_i^*)\lambda)^2 + \sum_{i : \theta_i^* = 0} (\epsilon_i - p_\lambda(\epsilon_i))^2$$

where $p_\lambda(\epsilon_i)$ is the point in the interval $[-\lambda, \lambda]$ that is closest to ϵ_i. It is now easy to see that

$$\epsilon_i - p_\lambda(\epsilon_i) = \begin{cases} \epsilon_i - \lambda & \text{if } \epsilon_i > \lambda \\ 0 & \text{if } -\lambda \leq \epsilon_i \leq \lambda \\ \epsilon_i + \lambda & \text{if } \epsilon_i < -\lambda. \end{cases}$$

This function above has a name and it is called the soft thresholding of ϵ_i with level λ. We thus have

$$\epsilon_i - p_\lambda(\epsilon_i) = \text{soft}_\lambda(\epsilon_i).$$

We have thus obtained:

$$\begin{aligned} \mathbb{E}\text{dist}^2(\epsilon, \lambda\partial f(\theta^*)) &= \sum_{i:\theta_i^* \neq 0} \mathbb{E}\left(\epsilon_i - \text{sign}(\theta_i^*)\lambda\right)^2 + \sum_{i:\theta_i^*=0} \mathbb{E}\left[\text{soft}_\lambda(\epsilon_i)\right]^2 \\ &= k(1+\lambda^2) + (n-k)\mathbb{E}\left[\text{soft}_\lambda(\epsilon_1)\right]^2 \end{aligned}$$

where k is the number of non-zero entries in θ^*. To proceed further, we need to compute $\mathbb{E}\left[\text{soft}_\lambda(\epsilon_1)\right]^2$ which we shall do in the next class.

15 Lecture 15

In the last lecture, we were studying the performance of estimators of the form:

$$\hat{\theta}_{\lambda,f} := \underset{\theta \in \mathbb{R}^n}{\text{argmin}} \left(\frac{1}{2} \|Y - \theta\|^2 + \lambda f(\theta) \right)$$

under the model $Y = \theta^* + \epsilon$ with $\epsilon \sim N(0, I_n)$. Here $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function.

For this estimator, we remarked last time that the following inequality is true:

$$\mathbb{E}\left\|\hat{\theta}_{\lambda,f} - \theta^*\right\|^2 \leq \mathbb{E}n\text{dist}^2(\epsilon, \lambda\partial f(\theta^*)). \tag{119}$$

This inequality is due to Oymak and Hassibi [17]; it has a simple proof which is given below.

15.1 Proof of Inequality (119)

Let $g(\theta) := \|y - \theta\|^2/2 + \lambda f(\theta)$ for $\theta \in \mathbb{R}^n$. The statement that $\hat{\theta}_{\lambda,f}$ minimizes g is equivalent to the statement that $0 \in \partial g(\hat{\theta}_{\lambda,f})$ (this is trivial because $\hat{\theta}_{\lambda,f}$ minimizing g is equivalent to $g(\theta) \geq g(\hat{\theta}_{\lambda,f}) + \left\langle 0, \theta - \hat{\theta}_{\lambda,f} \right\rangle$). It is now easy to check that

$$\partial g(\hat{\theta}_{\lambda,f}) = \hat{\theta}_{\lambda,f} - Y + \lambda\partial f(\hat{\theta}_{\lambda,f})n$$

so that we have

$$0 \in \hat{\theta}_{\lambda,f} - Y + \lambda\partial f(\hat{\theta}_{\lambda,f})$$

or equivalently

$$Y - \hat{\theta}_{\lambda,f} \in \lambda\partial f(\hat{\theta}_{\lambda,f}).$$

We now use Lemma 15.1 below to deduce that for every $s \in \partial f(\theta^*)$, we have

$$\left\langle Y - \hat{\theta}_{\lambda,f} - \lambda s, \hat{\theta}_{\lambda,f} - \theta^* \right\rangle \geq 0.$$

Writing $Y = \theta^* + \epsilon$, we obtain

$$\left\|\hat{\theta}_{\lambda,f} - \theta^*\right\|^2 \leq \left\langle \epsilon - \lambda s, \hat{\theta}_{\lambda,f} - \theta^* \right\rangle.$$

The Cauchy-Schwarz inequality can be used on the right hand side above which will give:

$$\left\|\hat{\theta}_{\lambda,f} - \theta^*\right\| \leq \|\epsilon - \lambda s\|.$$

This is true for every $s \in \partial f(\theta^*)$ so we can take an infimum over all such s and deduce

$$\left\|\hat{\theta}_{\lambda,f} - \theta^*\right\| \leq \text{dist}(\epsilon, \lambda\partial f(\theta^*))$$

which completes the proof of inequality (119).

Lemma 15.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then for every $\theta_1, \theta_2 \in \mathbb{R}^n$ and $s_1 \in \partial f(\theta_1)$, $s_2 \in \partial f(\theta_2)$, we have*

$$\langle \theta_1 - \theta_2, s_1 - s_2 \rangle \geq 0. \tag{120}$$

Proof. By the definition of subgradients, we have

$$f(\theta_1) \geq f(\theta_2) + \langle s_2, \theta_1 - \theta_2 \rangle \quad \text{and} \quad f(\theta_2) \geq f(\theta_1) + \langle s_1, \theta_2 - \theta_1 \rangle.$$

Adding these two inequalities results in (120). □

15.2 Application of (119) to $f(x) = \|x\|_1$

In the last lecture, we applied (119) to the situation when $f(x) = \|x\|_1 = |x_1| + \cdots + |x_n|$ and argued that

$$\begin{aligned}\mathbb{E}\text{dist}^2(\epsilon, \lambda\partial f(\theta^*)) &= \sum_{i:\theta_i^* \neq 0} \mathbb{E}\left(\epsilon_i - \text{sign}(\theta_i^*)\lambda\right)^2 + \sum_{i:\theta_i^*=0} \mathbb{E}\left[\text{soft}_\lambda(\epsilon_i)\right]^2 \\ &= k(1+\lambda^2) + (n-k)\mathbb{E}\left[\text{soft}_\lambda(\epsilon_1)\right]^2 \end{aligned} \tag{121}$$

where

$$\text{soft}_\lambda(y) = \begin{cases} y - \lambda & \text{if } y > \lambda \\ 0 & \text{if } -\lambda \leq y \leq \lambda \\ y + \lambda & \text{if } y < -\lambda. \end{cases}$$

We now proceed via (below $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ is the standard Gaussian density)

$$\begin{aligned}\mathbb{E}\left[\text{soft}_\lambda(\epsilon_1)\right]^2 &= \int_{-\infty}^{\infty} \left\{\text{soft}_\lambda(x)\right\}^2 \phi(x)dx \\ &= 2\int_0^{\infty} \left\{\text{soft}_\lambda(x)\right\}^2 \phi(x)dx \\ &= 2\int_\lambda^{\infty} (x-\lambda)^2\phi(x)dx \\ &= 2\left[\int_\lambda^{\infty} x^2\phi(x)dx - 2\lambda\int_\lambda^{\infty} x\phi(x)dx + \lambda^2\int_\lambda^{\infty}\phi(x)dx\right].\end{aligned}$$

We now apply integration by parts in the first integral above (with $u = x$ and $dv = x\phi(x)dx$), evaluate the second integral in closed form and leave the third integral as is to obtain

$$\mathbb{E}\left[\text{soft}_\lambda(\epsilon_1)\right]^2 = 2(1+\lambda^2)\left(1 - \Phi(\lambda)\right) - 2\lambda\phi(\lambda)$$

where $\Phi(\lambda) = \int_{-\infty}^{\lambda}\phi(x)dx$ is the Gaussian cdf. We now use the standard Mill's ratio Gaussian bound:

$$1 - \Phi(\lambda) \leq \frac{\phi(\lambda)}{\lambda} \qquad \text{for } \lambda > 0$$

to obtain

$$\mathbb{E}\left[\text{soft}_\lambda(\epsilon_1)\right]^2 \le 2(1+\lambda^2)\frac{\phi(\lambda)}{\lambda} - 2\lambda\phi(\lambda) = \frac{2\phi(\lambda)}{\lambda}. \tag{122}$$

Using this in (121), we obtain

$$\mathbb{E}\text{dist}^2(\epsilon, \lambda\partial f(\theta^*)) \le k(1+\lambda^2) + 2(n-k)\frac{\phi(\lambda)}{\lambda}$$

which implies, via inequality (119),

$$\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 \le k(1+\lambda^2) + (n-k)\frac{e^{-\lambda^2/2}}{\lambda}\sqrt{\frac{2}{\pi}}. \tag{123}$$

If we now make the choice

$$\lambda = \sqrt{2\log\frac{n}{k}}, \tag{124}$$

we obtain the risk bound

$$\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 \le k\left(1 + 2\log\frac{n}{k}\right) + (n-k)\frac{k}{n}\sqrt{\frac{2}{\pi}}\left(2\log\frac{n}{k}\right)^{-1/2} = \left(2k\log\frac{n}{k}\right)(1+o(1)) \tag{125}$$

as $n \to \infty$ provided $k/n \to 0$. Thus the LASSO with the penality (124) achieves the risk $2k\log(n/k)$.

Note that if the locations of the non-zero entries in θ^* are known, then the naive estimator which estimates the non-zero entries by Y_i and the zero entries by 0 with achieve risk equal to k. This, in relation to (125), means that the LASSO with tuning (124) is paying a price of $2\log(n/k)$ for not knowing the non-zero locations. We shall later prove that every estimator will have to pay this price in a minimax sense. This is not too hard to see intuitively. For example, if $k = 1$ and the magnitude of the non-zero signal is $\sqrt{c\log n}$ for some $c < 2$, then the noise in the data will drown the signal so every estimator will most likely miss the signal and incur a loss of $c\log n$. We shall make this precise later.

In order to use the choice (124) for the tuning parameter λ, we need knowledge of k. One can instead use

$$\lambda = \sqrt{2\log n} \tag{126}$$

which does not depend on k. With this choice, the bound (123) gives

$$\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 \le (2k\log n)(1+o(1)) \tag{127}$$

which is only slightly worse compared to (125). If k is of constant order, then there is not much difference between (125) and (127) but for $k = n/(\log n)$, there is a difference.

The bound (123) for the LASSO can actually be derived by a more direct method without relying on the inequality (119). This is because the estimator can be written in closed form via the soft thresholding operator. This is done next.

15.3 Soft Thresholding

The estimator

$$\hat{\theta}_\lambda := \underset{\theta\in\mathbb{R}^n}{\text{argmin}}\left(\frac{1}{2}\left\|Y-\theta\right\|^2 + \lambda\sum_{i=1}^n |\theta_i|\right)$$

can be written in closed form. Indeed observe first that if $\hat{\theta}_\lambda = (\hat{\theta}_\lambda(1), \ldots, \hat{\theta}_\lambda(n))$, then

$$\hat{\theta}_\lambda(i) = \underset{\theta_i \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2}(Y_i - \theta_i)^2 + \lambda|\theta_i|.$$

The function $Q(\theta_i) := (Y_i - \theta_i)^2/2 + \lambda|\theta_i|$ is convex and

$$Q'(\theta_i) = \begin{cases} \theta_i - Y_i - \lambda & \text{if } \theta_i < 0 \\ \theta_i - Y_i + \lambda & \text{if } \theta_i > 0 \end{cases}$$

The derivative $Q'(\theta_i)$ is therefore piecewise linear with positive slope except for an upward jump of 2λ at $\theta_i = 0$. Thus, $Q'(\theta_i)$ has exactly one sign change from negative to positive at a single point which must therefore be the minimizing value of $Q(\theta_i)$. Depending on the value of Y_i, this crossing point is positive, zero or negative, and we can then check that

$$\hat{\theta}_\lambda(i) = \begin{cases} Y_i - \lambda & \text{if } Y_i > \lambda \\ 0 & \text{if } -\lambda \le Y_i \le \lambda \\ Y_i + \lambda & \text{if } Y_i < -\lambda \end{cases}$$

In other words,

$$\hat{\theta}_\lambda(i) = \text{soft}_\lambda(Y_i) \qquad \text{for } i = 1, \ldots, n.$$

Using this, we can directly study the risk of $\hat{\theta}_\lambda$ as follows:

$$\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 = \sum_{i=1}^n \mathbb{E}\left(\text{soft}_\lambda(Y_i) - \theta_i^*\right)^2 = \sum_{i=1}^n r_S(\lambda, \theta_i^*)$$

where

$$r_S(\lambda, \mu) := \mathbb{E}\left(\text{soft}_\lambda(y) - \mu\right)^2 \qquad \text{with } y \sim N(\mu, 1).$$

This quantity $r_S(\lambda, \mu)$ is the risk of the soft thresholding estimator (at threshold λ) in the univariate problem with data $y \sim N(\mu, 1)$. We can explicitly write this as

$$\begin{aligned} r_S(\lambda, \mu) &:= \mathbb{E}\left(\text{soft}_\lambda(y) - \mu\right)^2 \\ &= \int_{-\lambda}^{\lambda} \mu^2 \phi(x - \mu)dx + \int_{-\infty}^{-\lambda} (x + \lambda - \mu)^2 \phi(x - \mu)dx + \int_{\lambda}^{\infty} (x - \lambda - \mu)^2 \phi(x - \mu)dx \\ &= \mu^2 \int_{-\lambda-\mu}^{\lambda-\mu} \phi(z)dz + \int_{-\infty}^{-\lambda-\mu} (z + \lambda)^2 \phi(z)dz + \int_{\lambda-\mu}^{\infty} (z - \lambda)^2 \phi(z)dz. \end{aligned}$$

The following are some basic properties of $r_S(\lambda, \mu)$:

1. The function $\mu \mapsto r_S(\lambda, \mu)$ is increasing on $[0, \infty)$. This is intuitive and easy to check via the calculation:

$$\frac{\partial}{\partial \mu} r_S(\lambda, \mu) = 2\mu\left(\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)\right) \tag{128}$$

   which is positive when $\mu > 0$.

2. When $\mu = 0$, we have

$$r_S(\lambda, 0) \le \frac{2\phi(\lambda)}{\lambda} = \sqrt{\frac{2}{\pi}} \frac{1}{\lambda} e^{-\lambda^2/2} \qquad \text{for all } \lambda > 0.$$

   We proved this in (122).

3. When μ approaches $\pm\infty$, the risk $r_S(\lambda, \mu)$ behaves like $1 + \lambda^2$:

$$\lim_{\mu \to \infty} r_S(\lambda, \mu) = 1 + \lambda^2.$$

I will leave this as an exercise to prove this. Intuitively, this is obvious since for large μ, most likely $\text{soft}_\lambda(y) = y - \lambda$ and $\mathbb{E}(y - \lambda - \mu)^2 = 1 + \lambda^2$. Combined with the fact that $\mu \mapsto r_S(\lambda, \mu)$ is increasing on $[0, \infty)$, we can deduce that

$$\sup_{\mu \in \mathbb{R}} r_S(\lambda, \mu) = 1 + \lambda^2. \tag{129}$$

These facts imply that, compared to the naive estimator y, the risk of the soft thresholding estimator is much smaller at $\mu = 0$ while its worst case risk is larger. Therefore, it makes sense to use it only when it is believed that μ is zero or small.

Using the above observations, we can give an alternative proof of the risk bound (123) for LASSO. Indeed, we can write

$$\begin{aligned}
\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 &= \sum_{i=1}^n \mathbb{E}\left(\text{soft}_\lambda(Y_i) - \theta_i^*\right)^2 \\
&= \sum_{i=1}^n r_S(\lambda, \theta_i^*) = \sum_{i: \theta_i^* \neq 0} r_S(\lambda, \theta_i^*) + \sum_{i: \theta_i^* = 0} r_S(\lambda, \theta_i^*) \leq k \sup_\mu r_S(\lambda, \mu) + (n-k) r_S(\lambda, 0).
\end{aligned}$$

Using the second and third facts above, we obtain

$$\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 \leq k(1 + \lambda^2) + 2(n-k)\frac{\phi(\lambda)}{\lambda}.$$

This proves (123) which, we have seen in the last subsection, allows us to deduce the rate results (125) and (127) under the choices (124) and (126) for the tuning parameter λ respectively.

Note that in the above bound, we used

$$r_S(\lambda, \theta_i^*) \leq \sup_{\mu \in \mathbb{R}} r_S(\lambda, \mu) \tag{130}$$

whenever $\theta_i^* \neq 0$. If more information is provided about the θ^*, then this might not be a very good bound. For example, it is common to also study the performance of LASSO under the assumption:

$$\|\theta^*\|_1 \leq C_n \tag{131}$$

for some $C_n > 0$. Under this assumption, potentiall all of the θ_i^*'s can be non-zero so that use of (130) will give very poor bounds. Note that, even though under (131), all entries of θ^* can be non-zero, they have to satisfy the property that the j^{th} largest entry in absolute value (to be denoted by $|\theta^*|_{(j)}$) should be bounded by C_n/j. This means that the entries of θ^* have to satisfy a certain decay. The assumption (131) can therefore be considered to be some form of *weak sparsity* assumption on θ^*.

Let us now study the risk of $\hat{\theta}_\lambda$ under the assumption (131). As mentioned earlier, we need some bound for $r_S(\lambda, \mu)$ for $\mu \neq 0$ that is better than $1 + \lambda^2$. For this, we use inequality (128) to write

$$\begin{aligned}
r_S(\lambda, \mu) - r_S(\lambda, 0) &= \int_0^\mu \frac{\partial}{\partial \mu} r_S(\lambda, d\mu) d\mu \\
&= \int_0^\mu 2\nu \left(\Phi(\lambda - \nu) - \Phi(-\lambda - \nu)\right) d\nu \leq \int_0^\mu 2\nu d\nu = \mu^2
\end{aligned}$$

which gives

$$r_S(\lambda, \mu) \leq r_S(\lambda, 0) + \mu^2.$$

Combining this with (129), we deduce

$$r_S(\lambda, \mu) \leq r_S(\lambda, 0) + \min(\mu^2, 1 + \lambda^2).$$

This implies that the risk of LASSO is bounded by

$$\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 \leq n r_S(\lambda, 0) + \sum_{i=1}^{n} \min\left((\theta_i^*)^2, 1 + \lambda^2\right).$$

We shall further bound this under the assumption (131). Because $\min(a^2, b^2) \leq ab$ for $a, b \geq 0$, we obtain

$$\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 \leq n r_S(\lambda, 0) + \sqrt{1+\lambda^2}\sum_{i=1}^{n} |\theta_i^*| \leq n r_S(\lambda, 0) + \sqrt{1+\lambda^2} C_n \leq 2n\frac{\phi(\lambda)}{\lambda} + \sqrt{1+\lambda^2} C_n.$$

The choice $\lambda = \sqrt{2 \log n}$ will now lead to

$$\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 \leq \sqrt{\frac{2}{\log n}} + C_n\sqrt{1 + 2\log n} \lesssim C_n\sqrt{\log n}$$

This further gives

$$\frac{1}{n}\mathbb{E}\left\|\hat{\theta}_\lambda - \theta^*\right\|^2 \lesssim \frac{C_n\sqrt{\log n}}{n}.$$

If, for example, $C_n \lesssim \sqrt{n}$, then the bound above becomes $\sqrt{(\log n)/n}$. We shall see later that this rate is minimax under the assumption (131).

16 Lecture 16

In the last lecture, we studied the behavior of the soft thresholding estimator in the Gaussian sequence model $Y \sim N_n(\theta^*, I_n)$. We noted that this estimator is just the same as LASSO and looked at bounds on its risk in the case where θ^* has exact sparsity and in the case of weak sparsity. We start this lecture with hard thresholding which has many similar properties to the soft thresholding estimator.

16.1 Hard Thresholding Estimator

The hard thresholding function is defined as:

$$\text{hard}_\lambda(y) := yI\{|y| > \lambda\} \quad \text{or} \quad \text{hard}_\lambda(y) := yI\{|y| \geq \lambda\}$$

Let us fix on the first definition above for concreteness. In words, $\text{hard}_\lambda(y)$ equals 0 when $-\lambda \leq y \leq \lambda$ and equals y otherwise. It is similar to $\text{soft}_\lambda(y)$ in that both equal 0 when $|y| \leq \lambda$. However, it is different in that it is discontinuous in y while $\text{soft}_\lambda(y)$ is continuous.

The Hard Thresholding estimator $\hat{\theta}_\lambda^H$ for θ^* in the Gaussian sequence model $Y \sim N(\theta^*, I_n)$ is given by

$$\hat{\theta}_\lambda^H = (\text{hard}_\lambda(Y_1), \ldots, \text{hard}_\lambda(Y_n)).$$

It is easy to see that $\hat{\theta}_\lambda^H$ is the solution to the optimization problem:

$$\hat{\theta}_\lambda^H = \underset{\theta \in \mathbb{R}^n}{\text{argmin}} \left(\|Y - \theta\|^2 + \lambda^2 \|\theta\|_0\right).$$

where $\|\theta\|_0 := \sum_{i=1}^n I\{\theta_i \neq 0\}$.

Note the similarity and dissimilarity of this optimization with the soft-thresholding (or LASSO) which is

$$\hat{\theta}^S_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left(\frac{1}{2} \|Y - \theta\|^2 + \lambda \|\theta\|_1 \right).$$

Notice that the tuning parameter in $\hat{\theta}^H_\lambda$ is λ^2 while it is λ in $\hat{\theta}^S_\lambda$. Also there is no factor of $1/2$ for the sum of squares term in $\hat{\theta}^H_\lambda$.

The hard thresholding estimator $\hat{\theta}^H_\lambda$ has very similar properties to the soft thresholding estimator in sparse settings. For example, we shall show below that when θ^* is k-sparse (i.e., $\|\theta^*\|_0 = k$) with $k/n \to 0$,

$$\mathbb{E} \left\| \hat{\theta}^H_\lambda - \theta^* \right\|^2 \leq (2k \log(n/k)) (1 + o(1)) \qquad \text{provided } \lambda = \sqrt{2 \log(n/k)}. \tag{132}$$

To see this, write

$$\mathbb{E} \left\| \hat{\theta}^H_\lambda - \theta^* \right\|^2 = \sum_{i=1}^n r_H(\lambda, \theta^*_i)$$

where

$$r_H(\lambda, \mu) := \mathbb{E} \left(\text{hard}_\lambda(y) - \mu \right)^2 \qquad \text{with } y \sim N(\mu, 1).$$

Therefore

$$\mathbb{E} \left\| \hat{\theta}^H_\lambda - \theta^* \right\|^2 = \sum_{i=1}^n r_H(\lambda, \theta^*_i) \leq (n-k) r_H(\lambda, 0) + k \sup_{\mu \in \mathbb{R}} r_H(\lambda, \mu).$$

It is now easy to see that

$$r_H(\lambda, 0) = 2 \int_\lambda^\infty x^2 \phi(x) dx = 2\lambda\phi(\lambda) + 2 (1 - \Phi(\lambda))$$

where the integral above was computed by integration by parts. The standard Mills ratio bound now gives

$$r_H(\lambda, 0) \leq 2 \left(\lambda + \frac{1}{\lambda} \right) \phi(\lambda) \qquad \text{for all } \lambda > 0.$$

It can also be shown (homework) that

$$\sup_{\mu \in \mathbb{R}} r_H(\lambda, \mu) \leq 1 + \lambda^2 \qquad \text{for all } \lambda > 0.$$

Unlike soft thresholding, the function $\mu \mapsto r_H(\lambda, \mu)$ is not monotonically increasing in $\mu > 0$. Indeed, it is easy to see that $\lim_{\mu \to \infty} r_H(\lambda, \mu) = 1$ but the maximum is achieved at some finite μ (near λ).

Because of the above facts, it follows that

$$\mathbb{E} \left\| \hat{\theta}^H_\lambda - \theta^* \right\|^2 \leq (n-k) 2(\lambda + \lambda^{-1}) \phi(\lambda) + k(1 + \lambda^2) = k \left(1 + \lambda^2 + \left(\frac{n}{k} - 1 \right) \sqrt{\frac{2}{\pi}} e^{-\lambda^2/2} \left(\lambda + \lambda^{-1} \right) \right).$$

The choice $\lambda = \sqrt{2 \log(n/k)}$ now easily gives (132). It also follows that

$$\mathbb{E} \left\| \hat{\theta}^H_\lambda - \theta^* \right\|^2 \leq (2k \log n) (1 + o(1)) \qquad \text{for } \lambda = \sqrt{2 \log n}.$$

It is also true that $\hat{\theta}^H_\lambda$ works similarly to soft thresholding under the assumption $\|\theta^*\|_1 \leq C_n$ (this is homework).

16.2 Linear Regression

We shall now study the linear regression model. Here we observe again a data vector $Y \in \mathbb{R}^n$ that we shall model as

$$Y = X\theta^* + \epsilon \qquad \text{with } \epsilon \sim N(0, I_n)$$

for some known deterministic $n \times p$ matrix X. The $p \times 1$ parameter θ^* is the unknown parameter of interest.

Given an estimator $\hat{\theta}$ of θ^*, we will be interested in the *prediction risk*:

$$\mathbb{E}\frac{1}{n}\left\|X\hat{\theta} - X\theta^*\right\|^2. \tag{133}$$

Depending on the particular context, it might be more or less natural to study $\mathbb{E}\left\|\hat{\theta} - \theta^*\right\|^2/n$ but this will usually require more assumptions on X and we shall refrain from studying this loss function.

Both the hard and soft thresholding estimators can be extended in a straightforward manner to the case of Linear Regression. The extension of the hard thresholding estimator will be:

$$\hat{\theta}_\lambda := \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left(\|Y - X\theta\|^2 + \lambda^2 \|\theta\|_0 \right). \tag{134}$$

Note that when $X = I$, we used this estimator under sparse settings with $\lambda = \sqrt{2\log(n/k)}$ or $\lambda = \sqrt{2\log n}$. Therefore the term multiplying $\|\theta\|_0$ in (134) will involve $\log n$. For this reason, we shall refer to this estimator as the BIC estimator (see, for example, `https://en.wikipedia.org/wiki/Bayesian_information_criterion`) and denote this by $\hat{\theta}_\lambda^{\text{BIC}}$.

The extension of the soft thresholding estimator directly gives the LASSO estimator:

$$\hat{\theta}_\lambda^{\text{LASSO}} := \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{2}\|Y - \theta\|^2 + \lambda \|\theta\|_1 \right). \tag{135}$$

We shall analyze both these estimators in terms of the prediction risk (133). We shall focus mainly on the exact sparsity setting where $k := \|\theta^*\|_0$ is small compared to p and n. From the computational perspective, it is easy to see that (135) can be computed via convex optimization while (134) can be very hard to compute depending on X (in the worst case, computing (134) is NP hard).

Let us start with the analysis for the BIC estimator (134).

16.3 The Prediction Risk of $\hat{\theta}_\lambda^{\text{BIC}}$

The key question is: when θ^* is k-sparse, does the BIC estimator, properly regularized, have prediction risk bounded by a constant multiple of $(k/n)(\log(ep/k))$? We shall see that this will be true **without any assumptions** on the design matrix X.

Before answering this question, let us first analyze a simple estimator that is given by

$$\hat{\theta} := \underset{\theta : \|\theta\|_0 \le k}{\operatorname{argmin}} \left(\|Y - X\theta\|^2 \right).$$

This estimator simply minimizes the sum of squares over all θ having at most k entries. Remember that k is the L_0 norm of the true vector θ^* so that $\|\theta^*\|_0$ needs to be known for using this estimator. We shall prove that for this estimator,

$$\frac{1}{n}\mathbb{E}\left\|X\hat{\theta} - X\theta^*\right\|^2 \le C\frac{k}{n}\log\left(\frac{ep}{k}\right)$$

for a universal constant C. Let us first use the rate theorem to see why this should be true. We can write $\hat{\theta}$ as

$$\hat{\theta} = \operatorname*{argmax}_{\theta:\|\theta\|_0 \leq k} \left(2\langle \epsilon, X\theta - X\theta^* \rangle - \|X\theta - X\theta^*\|^2 \right) = \operatorname*{argmax}_{\theta \in \Theta} M_n(\theta)$$

where $\Theta := \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq k\}$ and

$$M_n(\theta) := 2\langle \epsilon, X\theta - X\theta^* \rangle - \|X\theta - X\theta^*\|^2 .$$

We will use the rate theorem with this M_n and

$$M(\theta) := -\|X\theta - X\theta^*\|^2 \quad \text{and} \quad d(\theta, \theta^*) := \|X\theta - X\theta^*\| .$$

Thus to obtain the rate via the rate theorem, we need to bound

$$\mathbb{E} \sup_{\theta:\|\theta\|_0 \leq k, d(\theta,\theta^*) \leq u} (M_n - M)(\theta - \theta^*) \leq 2\mathbb{E} \sup_{\theta:\|\theta\|_0 \leq k, \|X\theta - X\theta^*\| \leq u} \langle \epsilon, X\theta - X\theta^* \rangle = \mathbb{E} \sup_{v \in V} \langle \epsilon, v \rangle$$

where V consists of all vectors $v \in \mathbb{R}^n$ with $\|v\| \leq u$ and which satisfy $v = X(\theta - \theta^*)$ for some θ with $\|\theta\|_0 \leq k$. We shall use Dudley's entropy bound to control the expected supremum above:

$$\mathbb{E} \sup_{v \in V} \langle \epsilon, v \rangle \leq C \int_0^{\text{diam}(V)/2} \sqrt{\log M(\epsilon, V)} d\epsilon \tag{136}$$

where $M(\epsilon, V)$ and $\text{diam}(V)$ are the packing number and diameter in the usual Euclidean metric on $\mathbb{R}^n$.

Note now that for every θ with $\|\theta\|_0 \leq k$, we have $\|\theta - \theta^*\|_0 \leq 2k$ (because $\|\theta^*\|_0 \leq k$). As a result, we can write

$$V \subseteq \cup \{V_S : S \subseteq \{1, \ldots, p\}, |S| \leq 2k\}$$

(where $|S|$ denotes the cardinality of S) where V_S denotes the set of all vectors $v \in \mathbb{R}^n$ for which $\|v\| \leq u$ and $v = X\beta$ for some β that is supported on S (i.e., $\{i : \beta_i \neq 0\} \subseteq S$). Therefore, we deduce that

$$M(\epsilon, V) \leq \sum_{S \subseteq \{1,\ldots,p\}: |S| \leq 2k} M(\epsilon, V_S).$$

Because,

$$V_S := \{v \in \mathbb{R}^n : \|v\| \leq u, v \in \mathcal{C}(X_S)\}$$

where X_S is the matrix formed by including only those columns of X whose indices belong to S and $\mathcal{C}(X_S)$ denotes the column space of X_S. This means that V_S is a ball in a linear subspace of dimension at most $2k$ so that (by an earlier result on packing numbers of balls in linear spaces)

$$M(\epsilon, V_S) \leq \left(1 + \frac{2u}{\epsilon}\right)^{2k}.$$

Consequently,

$$M(\epsilon, V) \leq \left(1 + \frac{2u}{\epsilon}\right)^{2k} |\{S \subseteq \{1, \ldots, p\} : |S| \leq 2k\}| \leq \left(1 + \frac{2u}{\epsilon}\right)^{2k} \left[\binom{p}{0} + \cdots + \binom{p}{2k}\right] \leq \left(1 + \frac{2u}{\epsilon}\right)^{2k} \left(\frac{ep}{2k}\right)^{2k}.$$

Plugging this in (136), we obtain

$$\begin{aligned}
\mathbb{E} \sup_{v \in V} \langle \epsilon, v \rangle &\leq C \int_0^u \sqrt{2k \log\left(1 + \frac{2u}{\epsilon}\right) + 2k \log \frac{ep}{2k}} d\epsilon \\
&\leq \int_0^u \sqrt{2k \log\left(1 + \frac{2u}{\epsilon}\right)} d\epsilon + \int_0^u \sqrt{2k \log \frac{ep}{2k}} d\epsilon \lesssim u\sqrt{k}\sqrt{\log \frac{ep}{2k}}.
\end{aligned}$$

Equating this to u^2, we would obtain

$$u = \sqrt{k \log \frac{ep}{2k}}.$$

This suggests therefore that

$$\mathbb{E}\left[\frac{1}{n}\left\|X\hat{\theta} - X\theta^*\right\|^2\right] \leq \frac{Ck}{n} \log \frac{ep}{2k}. \tag{137}$$

As usual, our rate theorem is not strong enough to give this expectation control. To prove the above bound, we can argue as follows. The basic inequality corresponding to the M-estimator $\hat{\theta}$ is: $M(\theta^*) - M(\hat{\theta}) \leq (M_n - M)(\hat{\theta} - \theta^*)$ which is the same as

$$\left\|X\hat{\theta}_n - X\theta^*\right\|^2 \leq 2\left\langle \epsilon, X\hat{\theta} - X\theta^* \right\rangle \leq 2\left\langle \epsilon, \frac{X\hat{\theta} - X\theta^*}{\left\|X\hat{\theta} - X\theta^*\right\|} \right\rangle \left\|X\hat{\theta} - X\theta^*\right\|$$

so that

$$\left\|X\hat{\theta}_n - X\theta^*\right\| \leq 2\left\langle \epsilon, \frac{X\hat{\theta} - X\theta^*}{\left\|X\hat{\theta} - X\theta^*\right\|} \right\rangle \leq 2 \sup_{v \in V^*} \langle \epsilon, v \rangle .$$

where

$$V^* := \{v : \|v\| \leq 1, v = X\beta \text{ with } \|\beta\|_0 \leq 2k\}.$$

This gives

$$E\left\|X\hat{\theta}_n - X\theta^*\right\|^2 \leq 4\mathbb{E}\left\{\left(\sup_{v \in V^*} \langle \epsilon, v \rangle\right)^2\right\}$$

Note that we have just rigorously proved that

$$\mathbb{E} \sup_{v \in V^*} \langle \epsilon, v \rangle \lesssim \sqrt{k \log \frac{ep}{2k}}. \tag{138}$$

From here and the fact that

$$\epsilon \mapsto \sup_{v \in V^*} \langle \epsilon, v \rangle$$

is a Lipschitz function (with Lipschitz constant 2), one can prove that

$$\mathbb{E}\left\{\left(\sup_{v \in V^*} \langle \epsilon, v \rangle\right)^2\right\} \leq k \log \frac{ep}{2k} \tag{139}$$

which proves (137).

One way to see how (138) implies (139) is via the following important result on the concentration of Lipschitz functions of Gaussian random vectors.

Theorem 16.1. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is an L-Lipschitz function i.e., $|f(x) - f(y)| \leq L\|x - y\|$ and suppose $Z \sim N(0, I_n)$. Then for all $t \geq 0$,*

$$\mathbb{P}\{f(Z) \geq \mathbb{E}f(Z) + t\} \leq \exp\left(\frac{-t^2}{2L^2}\right) \quad \textit{and} \quad \mathbb{P}\{f(Z) \leq \mathbb{E}f(Z) - t\} \leq \exp\left(\frac{-t^2}{2L^2}\right) \tag{140}$$

Theorem 16.1 can be used to derive (139) from (138). Indeed, let

$$f(\epsilon) := \sup_{v \in V^*} \langle \epsilon, v \rangle .$$

It is easy to see then that f is 1-Lipschitz. Indeed,

$$\begin{aligned} f(\epsilon_1) &= \sup_{v \in V^*} \langle \epsilon_1, v \rangle \\ &= \sup_{v \in V^*} \{ \langle \epsilon_2, v \rangle + \langle \epsilon_1 - \epsilon_2, v \rangle \} \leq f(\epsilon_2) + \|\epsilon_1 - \epsilon_2\| \sup_{v \in V^*} \|v\| \end{aligned}$$

Because every vector in V^* has norm at most 1, we conclude that f is 1-Lipschitz. Thus, by (140),

$$\mathbb{P}\left\{ |f(\epsilon) - \mathbb{E}f(\epsilon)| \geq u \right\} \leq 2 \exp\left(\frac{-u^2}{2} \right).$$

so

$$\mathbb{P}\left\{ |f(\epsilon) - \mathbb{E}f(\epsilon)|^2 \geq t \right\} \leq 2 \exp\left(\frac{-t}{2} \right).$$

As a result

$$\mathbb{E}\left| f(\epsilon) - \mathbb{E}f(\epsilon) \right|^2 \leq \int_0^\infty 2e^{-t/2} dt \leq 4.$$

This gives

$$\mathbb{E}f^2(\epsilon) \leq 2\left(\mathbb{E}f(\epsilon)\right)^2 + 2\mathbb{E}\left| f(\epsilon) - \mathbb{E}f(\epsilon) \right|^2 \leq 2\left(\mathbb{E}f(\epsilon)\right)^2 + 8.$$

This, combined with (138), allows us to deduce

$$\mathbb{E}\left\{ \left(\sup_{v \in V^*} \langle \epsilon, v \rangle \right)^2 \right\} \leq 2\left(\mathbb{E} \sup_{v \in V^*} \langle \epsilon, v \rangle \right)^2 + 8 \lesssim k \log \frac{ep}{2k}$$

which proves (139).

17 Lecture 17

17.1 The Prediction Risk of $\hat{\theta}_\lambda^{\mathrm{BIC}}$ (continued)

We will study the prediction risk of the estimator:

$$\hat{\theta}_\lambda := \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \left(\|Y - X\theta\|^2 + \lambda^2 \|\theta\|_0 \right). \tag{141}$$

in the linear regression model $Y = X\theta^* + \epsilon$ where $\epsilon \sim N(0, I_n)$. Recall that this is the Hard Thresholding estimator when $X = I_n$. The following result will be true for $\hat{\theta}_\lambda^{\mathrm{BIC}}$: If

$$\lambda := c_1 \sqrt{\log(ep)}$$

for a sufficiently large c_1, then

$$\frac{1}{n}\mathbb{E}\left\| X\hat{\theta}_\lambda^{\mathrm{BIC}} - X\theta^* \right\|^2 \leq C(c_1)\frac{k}{n}\log(ep) \qquad \text{where } k := \|\theta^*\|_0. \tag{142}$$

We shall not prove the expectation bound given above but we shall use the rate theorem which will imply that the loss $\left\|X\hat{\theta}^{\text{BIC}}_{\lambda} - X\theta^*\right\|$ will be bounded by a constant multiple of $\sqrt{k\log(ep)}$ with high probability. Before we proceed to this argument, let us recall first that in the last lecture, we proved

$$\frac{1}{n}\mathbb{E}\left\|X\hat{\theta}^{(k)} - X\theta^*\right\|^2 \leq \frac{Ck}{n}\log\left(\frac{ep}{k}\right) \qquad \text{for } \hat{\theta}^{(k)} := \operatorname*{argmin}_{\theta:\|\theta\|_0 \leq k} \|Y - X\theta\|^2. \tag{143}$$

Note that this estimator $\hat{\theta}^{(k)}$ can be viewed as a constrained version of $\hat{\theta}^{\text{BIC}}_{\lambda}$. In the process of proving (143), we derived that

$$\mathbb{E}\sup_{\theta:\|\theta\|_0 \leq s, \|X\theta - X\theta^*\| \leq u} \langle \epsilon, X\theta - X\theta^*\rangle \leq Cu\sqrt{(s+k)\log\frac{ep}{s+k}} \qquad \text{where } k := \|\theta^*\|_0. \tag{144}$$

We shall use this inequality in the sequel. Now let us proceed to analyze $\left\|X\hat{\theta}^{\text{BIC}}_{\lambda} - X\theta^*\right\|$ via the rate theorem. Note first that $\hat{\theta}^{\text{BIC}}_{\lambda}$ maximizes

$$M_n(\theta) := 2\langle \epsilon, X\theta - X\theta^*\rangle - \|X\theta - X\theta^*\|^2 - \lambda^2 \|\theta\|_0 .$$

So we shall apply the rate theorem with this $M_n(\theta)$ and $M(\theta)$ taken to be

$$M(\theta) := -\|X\theta - X\theta^*\|^2 .$$

Also $d(\theta, \theta^*) := \|X\theta - X\theta^*\|$. For applying the rate theorem, we would need to bound

$$\Gamma := \mathbb{E}\sup_{\theta: d(\theta,\theta^*) \leq u} (M_n - M)(\theta - \theta^*)$$

from above and equate the resulting bound to u^2. For this, note that

$$\begin{aligned}
\Gamma &= \mathbb{E}\sup_{\theta:\|X\theta - X\theta^*\| \leq u} \left(2\langle \epsilon, X\theta - X\theta^*\rangle - \lambda^2\|\theta\|_0 + \lambda^2\|\theta^*\|_0\right) \\
&= \mathbb{E}\max_{1\leq s\leq p}\left\{\sup_{\theta:\|\theta\|_0 = s, \|X\theta - X\theta^*\| \leq u} 2\langle \epsilon, X\theta - X\theta^*\rangle - \lambda^2 s + \lambda^2 k\right\} \\
&\leq \mathbb{E}\max_{1\leq s\leq p}\left\{\sup_{\theta:\|\theta\|_0 = s, \|X\theta - X\theta^*\| \leq u} 2\langle \epsilon, X\theta - X\theta^*\rangle - \mathbb{E}\sup_{\theta:\|\theta\|_0 = s, \|X\theta - X\theta^*\| \leq u} 2\langle \epsilon, X\theta - X\theta^*\rangle\right\} \\
&+ \max_{1\leq s\leq p}\left\{\mathbb{E}\sup_{\theta:\|\theta\|_0 = s, \|X\theta - X\theta^*\| \leq u} 2\langle \epsilon, X\theta - X\theta^*\rangle - \lambda^2 s + \lambda^2 k\right\}.
\end{aligned}$$

Let us denote the two terms above by Γ_1 and Γ_2 respectively. To bound Γ_1, note that for every $1 \leq s \leq p$, the map

$$\epsilon \mapsto \sup_{\theta:\|\theta\|_0 = s, \|X\theta - X\theta^*\| \leq u} 2\langle \epsilon, X\theta - X\theta^*\rangle$$

is $2u$-Lipschitz so by the Gaussian concentration inequality from last time, we have

$$\mathbb{P}\left\{\sup_{\theta:\|\theta\|_0 = s, \|X\theta - X\theta^*\| \leq u} 2\langle \epsilon, X\theta - X\theta^*\rangle - \mathbb{E}\sup_{\theta:\|\theta\|_0 = s, \|X\theta - X\theta^*\| \leq u} 2\langle \epsilon, X\theta - X\theta^*\rangle \geq a\right\} \leq \exp\left(\frac{-a^2}{8u^2}\right)$$

for every $a > 0$. As a result, we have

$$\Gamma_1 \leq Cu\sqrt{\log(ep)}.$$

For bounding Γ_2, we simply use inequality (144) to obtain

$$\Gamma_2 \leq \max_{1\leq s\leq p}\left\{Cu\sqrt{(s+k)\log\frac{ep}{s+k}} - \lambda^2 s + \lambda^2 k\right\} \leq \max_{1\leq s\leq p}\left\{Cu\sqrt{(s+k)\log(ep)} - \lambda^2 s + \lambda^2 k\right\}.$$

Thus by enlarging the constant C appropriately, we have

$$\Gamma = \Gamma_1 + \Gamma_2 \leq \max_{1 \leq s \leq p} \left\{ Cu\sqrt{(s+k)\log(ep)} - \lambda^2 s + \lambda^2 k \right\}.$$

We just maximize the above function in terms of s by calculus. Setting s to be such that

$$s + k = \frac{C^2 u^2 \log(ep)}{4\lambda^4},$$

we obtain

$$\Gamma \leq \frac{C^2 u^2 \log(ep)}{4\lambda^2} + k\lambda^2.$$

Thus, by the rate theorem, $\left\| X\hat{\theta}^{\mathrm{BIC}}_{\lambda} - X\theta^* \right\|$ will be controlled by the solution to

$$\frac{C^2 u^2 \log(ep)}{4\lambda^2} + k\lambda^2 = u^2.$$

Choosing $\lambda = c_1\sqrt{\log(ep)}$, we obtain

$$\frac{C^2}{4c_1^2}u^2 + k\lambda^2 = u^2.$$

Now if $c_1 = C^2/2$, then we obtain the equation

$$\frac{1}{2}u^2 = k\lambda^2 = kc_1^2 \log(ep)$$

which gives

$$u = c_1\sqrt{2k\log(ep)}.$$

This, therefore, proves that $\left\| X\hat{\theta}^{\mathrm{BIC}}_{\lambda} - X\theta^* \right\|$ will be controlled by $\sqrt{k\log(ep)}$ with high probability. This argument does not yield the expecation bound (142) which can be proved a modification of the above argument.

Here are two comments on (142):

1. It requires no assumptions on the design matrix X.
2. If λ is allowed to depend on k, then it is possible to derive the bound $(k/n)\log(ep/k)$ (i.e., $\log(ep)$ in (142) can be replaced by $\log(ep/k)$) (see, Johnstone [11, Chapter 11]).

17.2 Prediction Risk of $\hat{\theta}^{\mathrm{LASSO}}_{\lambda}$

We now study the LASSO estimator:

$$\hat{\theta}^{\mathrm{LASSO}}_{\lambda} := \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \left(\frac{1}{2} \left\| Y - \theta \right\|^2 + \lambda \left\| \theta \right\|_1 \right). \tag{145}$$

We shall prove two bounds on the prediction error

$$\frac{1}{n}\mathbb{E}\left\| X\hat{\theta}^{\mathrm{LASSO}}_{\lambda} - X\theta^* \right\|^2.$$

The first bound involves $\left\|\theta^*\right\|_1$ (weak or approximate sparsity regime) and the second bound involves $\left\|\theta^*\right\|_0$ (strong or exact sparsity regime)

17.2.1 Weak Sparsity Bound

Here we shall prove the following bound for the prediction risk of $\hat{\theta}_\lambda^{\text{LASSO}}$. Let $X_1, \ldots, X_p \in \mathbb{R}^n$ denote the p columns of the $n \times p$ design matrix X. If

$$\lambda = c_1 \sqrt{\log(ep)} \max_{1 \le i \le p} \|X_i\| \tag{146}$$

for a sufficiently large c_1, then

$$\frac{1}{n}\mathbb{E}\left\|X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^*\right\|^2 \le C(c_1)\frac{\|\theta^*\|_1 \sqrt{\log(ep)}}{n} \max_{1 \le i \le p} \|X_i\| \tag{147}$$

for a constant $C(c_1)$ depending only on c_1.

Note that this bound specializes to our earlier bound for soft thresholding if we take $X = I_n$. Usually, the scaling for the design matrix is chosen so that $\max_{1 \le i \le p} \|X_i\| \le \sqrt{n}$. Under this assumption, the choice of the tuning parameter becomes

$$\lambda = c_1 \sqrt{n \log(ep)}$$

and the prediction risk bound (147) becomes

$$\frac{1}{n}\mathbb{E}\left\|X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^*\right\|^2 \le C(c_1)\|\theta^*\|_1 \sqrt{\frac{\log(ep)}{n}}.$$

We shall now prove (147). Note first that $\hat{\theta}_\lambda^{\text{LASSO}}$ maximizes $M_n(\theta)$ over $\theta \in \mathbb{R}^p$ where

$$M_n(\theta) := \langle \epsilon, X\theta - X\theta^* \rangle - \frac{1}{2}\|X\theta - X\theta^*\|^2 - \lambda \|\theta\|_1 .$$

It is also easy to see that θ^* maximizes $M(\theta), \theta \in \mathbb{R}^p$ where

$$M(\theta) := \frac{-1}{2}\|X\theta - X\theta^*\|^2 .$$

The basic inequality therefore is

$$M(\theta^*) - M(\hat{\theta}_\lambda^{\text{LASSO}}) \le (M_n - M)(\hat{\theta}_\lambda^{\text{LASSO}} - \theta^*)$$

which becomes

$$\frac{1}{2}\left\|X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^*\right\|^2 \le \left\langle \epsilon, X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^* \right\rangle - \lambda \left\|\hat{\theta}_\lambda^{\text{LASSO}}\right\|_1 + \lambda \|\theta^*\|_1 .$$

We shall bound the term involving ϵ on the right hand side above via

$$\left\langle \epsilon, X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^* \right\rangle = \left\langle X^T\epsilon, \hat{\theta}_\lambda^{\text{LASSO}} - \theta^* \right\rangle \le \left\|X^T\epsilon\right\|_\infty \left\|\hat{\theta}_\lambda^{\text{LASSO}} - \theta^*\right\|_1$$

which gives

$$\frac{1}{2}\left\|X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^*\right\|^2 \le \left\|X^T\epsilon\right\|_\infty \left\|\hat{\theta}_\lambda^{\text{LASSO}} - \theta^*\right\|_1 - \lambda \left\|\hat{\theta}_\lambda^{\text{LASSO}}\right\|_1 + \lambda \|\theta^*\|_1 . \tag{148}$$

Triangle inequality:

$$\left\|\hat{\theta}_\lambda^{\text{LASSO}} - \theta^*\right\|_1 \le \left\|\hat{\theta}_\lambda^{\text{LASSO}}\right\|_1 + \|\theta^*\|_1$$

now gives

$$\frac{1}{2}\left\|X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^*\right\|^2 \le \left(\left\|X^T\epsilon\right\|_\infty - \lambda\right)\left\|\hat{\theta}_\lambda^{\text{LASSO}}\right\|_1 + \left(\left\|X^T\epsilon\right\|_\infty + \lambda\right)\|\theta^*\|_1 .$$

Therefore if λ is chosen so that

$$\lambda \ge \left\|X^T\epsilon\right\|_\infty , \tag{149}$$

then we would obtain

$$\left\|X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^*\right\|^2 \le 4\lambda \|\theta^*\|_1 .$$

When the tuning parameter λ is chosen as in (146), it is easy to see that (149) holds with high probability and then inequality (147) follows from the above inequality (rigorize this argument and complete the proof of (147)).

17.2.2 Strong Sparsity Bound

We shall now attempt to bound the prediction error in terms of $\|\theta^*\|_0 = k$. Before proceeding further, let us introduce the following notation. Let S denote the set of indices i among $1, \ldots, p$ where $\theta_i^* \neq 0$. Then $|S| = k$. For a vector $\theta \in \mathbb{R}^n$, let $\theta_S \in \mathbb{R}^n$ denote the vector whose i^{th} entry equals θ_i if $i \in S$ and equals 0 if $i \notin S$. We analogously define θ_{S^c}. Let us start with inequality (148) (where, for simplicity of notation, we write $\hat{\theta}$ for $\hat{\theta}_\lambda^{\text{LASSO}}$)

$$\frac{1}{2}\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq \left\|X^T\epsilon\right\|_\infty \left\|\hat{\theta} - \theta^*\right\|_1 - \lambda\left\|\hat{\theta}\right\|_1 + \lambda\left\|\theta^*\right\|_1.$$

We rewrite the right hand side above as

$$\frac{1}{2}\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq \left\|X^T\epsilon\right\|_\infty \left(\left\|\hat{\theta}_S - \theta_S^*\right\|_1 + \left\|\hat{\theta}_{S^c} - \theta_{S^c}^*\right\|_1\right) - \lambda\left(\left\|\hat{\theta}_S\right\|_1 + \left\|\hat{\theta}_{S^c}\right\|_1\right) + \lambda\left(\|\theta_S^*\|_1 + \|\theta_{S^c}^*\|_1\right).$$

Because $\theta_{S^c}^* = 0$, we can simplify this as

$$\frac{1}{2}\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq \left\|X^T\epsilon\right\|_\infty \left(\left\|\hat{\theta}_S - \theta_S^*\right\|_1 + \left\|\hat{\theta}_{S^c}\right\|_1\right) - \lambda\left(\left\|\hat{\theta}_S\right\|_1 + \left\|\hat{\theta}_{S^c}\right\|_1\right) + \lambda\left\|\theta_S^*\right\|_1.$$

Rearranging terms, we deduce

$$\frac{1}{2}\left\|X\hat{\theta} - X\theta^*\right\|^2 + \lambda\left\|\hat{\theta}_{S^c}\right\|_1 \leq \left\|X^T\epsilon\right\|_\infty \left\|\hat{\theta}_S - \theta_S^*\right\|_1 + \left\|X^T\epsilon\right\|_\infty \left\|\hat{\theta}_{S^c}\right\|_1 + \lambda\left\|\theta_S^*\right\|_1 - \lambda\left\|\hat{\theta}_S\right\|_1.$$

The last two terms can be bounded by triangle inequality as

$$\lambda\left\|\theta_S^*\right\|_1 - \lambda\left\|\hat{\theta}_S\right\|_1 \leq \lambda\left\|\hat{\theta}_S - \theta_S^*\right\|_1$$

and so we obtain

$$\frac{1}{2}\left\|X\hat{\theta} - X\theta^*\right\|^2 + \lambda\left\|\hat{\theta}_{S^c}\right\|_1 \leq \left(\left\|X^T\epsilon\right\|_\infty + \lambda\right)\left\|\hat{\theta}_S - \theta_S^*\right\|_1 + \left\|X^T\epsilon\right\|_\infty \left\|\hat{\theta}_{S^c}\right\|_1. \tag{150}$$

Suppose now that we assume that the regularization parameter λ satisfies (149) as before. We can then cancel the terms $\lambda\left\|\hat{\theta}_{S^c}\right\|_1$ and $\left\|X^T\epsilon\right\|_\infty \left\|\hat{\theta}_{S^c}\right\|_1$ from both sides of the inequality above to obtain

$$\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq 4\lambda\left\|\hat{\theta}_S - \theta_S^*\right\|_1.$$

Applying the Cauchy-Schwarz inequality on the right hand side, we get

$$\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq 4\lambda\left\|\hat{\theta}_S - \theta_S^*\right\|_1 \leq 4\lambda\sqrt{k}\left\|\hat{\theta}_S - \theta_S^*\right\| \leq 4\lambda\sqrt{k}\left\|\hat{\theta} - \theta^*\right\|. \tag{151}$$

Further

$$\begin{aligned}
\left\|\hat{\theta} - \theta^*\right\| &= n^{-1/2}\left\|X(\hat{\theta} - \theta^*)\right\| \frac{\left\|\hat{\theta} - \theta^*\right\|}{n^{-1/2}\left\|X(\hat{\theta} - \theta^*)\right\|} \\
&\leq n^{-1/2}\left\|X(\hat{\theta} - \theta^*)\right\| \sup_{\Delta \in \mathbb{R}^p : \Delta \neq 0} \frac{\|\Delta\|}{n^{-1/2}\|X\Delta\|} \\
&= n^{-1/2}\left\|X(\hat{\theta} - \theta^*)\right\| \frac{1}{\inf_{\Delta \in \mathbb{R}^p : \Delta \neq 0} n^{-1/2}\|X\Delta\| / \|\Delta\|} = n^{-1/2}\left\|X(\hat{\theta} - \theta^*)\right\| \frac{1}{\sqrt{\lambda_{\min}(X^TX/n)}}.
\end{aligned}$$

This gives (from (151))

$$\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq 4\lambda\sqrt{k}\left\|\hat{\theta} - \theta^*\right\| \leq 4\lambda\sqrt{\frac{k}{n}}\left\|X\hat{\theta} - X\theta^*\right\| \frac{1}{\sqrt{\lambda_{\min}(X^TX/n)}}.$$

Cancelling one $\left\|X\hat{\theta} - X\theta^*\right\|$, we obtain

$$\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq \frac{16\lambda^2 k}{n} \frac{1}{\lambda_{\min}(X^T X/n)}.$$

We can now take λ to be as in (146) which ensures (149) with high probability (if c_1 is sufficiently large) which implies that

$$\frac{1}{n}\left\|X\hat{\theta}^{\text{LASSO}}_{\lambda} - X\theta^*\right\|^2 \leq C(c_1)\frac{k}{n^2}\log(ep)\frac{\max_i \|X_i\|^2}{\lambda_{\min}(X^T X/n)}$$

with high probability (expecation bound can be derived as well). When $X = I_n$, this is a weaker form of our earlier soft-thresholding risk bound for exact sparsity. However, a major problem of this bound is that it depends on $\lambda_{\min}(X^T X/n)$. When $p > n$, we necessarily have $\lambda_{\min}(X^T X/n) = 0$ so that the above bound is vacuous. It is however to replace $\lambda_{\min}(X^T X/n)$ by a smaller quantity by a slight tweak on the above argument with λ chosen to be $2\left\|X^T\epsilon\right\|_\infty$.

Indeed, if $\lambda \geq 2\left\|X^T\epsilon\right\|_\infty$, then one can first observe from (150) the inequality:

$$\frac{\lambda}{2}\left\|\hat{\theta}_{S^c}\right\|_1 \leq \frac{3\lambda}{2}\left\|\hat{\theta}_S - \theta^*_S\right\|_1$$

which is identical to

$$\left\|\hat{\theta}_{S^c} - \theta^*_{S^c}\right\|_1 \leq 3\left\|\hat{\theta}_S - \theta^*_S\right\|_1.$$

In other words, the vector $\hat{\theta} - \theta^*$ belongs to the cone:

$$\mathcal{C} := \left\{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\right\}.$$

Using this observation, one can replace the bound

$$\frac{\left\|\hat{\theta} - \theta^*\right\|}{n^{-1/2}\left\|X(\hat{\theta} - \theta^*)\right\|} \leq \sup_{\Delta \in \mathbb{R}^p : \Delta \neq 0} \frac{\|\Delta\|}{n^{-1/2}\|X\Delta\|}$$

by

$$\frac{\left\|\hat{\theta} - \theta^*\right\|}{n^{-1/2}\left\|X(\hat{\theta} - \theta^*)\right\|} \leq \sup_{\Delta \in \mathcal{C} : \Delta \neq 0} \frac{\|\Delta\|}{n^{-1/2}\|X\Delta\|}.$$

Note that the only difference between the above two bounds is that in the latter the supremum is over $\Delta \in \mathcal{C}$ while, in the former, the supremum is over all $\Delta \in \mathbb{R}^p$. As we shall see in the next lecture, this improvement gives results that potentially work even when $p > n$.

18 Lecture 18

The main topic for study is to complete the discussion of the performance of the LASSO, in terms of prediction risk, for the case of exact sparsity. Let us first recap the ideas from the previous couple of lectures and present the main problem of interest.

18.1 Recap: linear regression with exact sparsity

We observe a data vector $Y \in \mathbb{R}^n$ that we model as

$$Y = X\theta^* + \epsilon \qquad \text{with } \epsilon \sim N(0, I_n).$$

X is a deterministic $n \times p$ matrix and θ^* is a vector in $\mathbb{R}^p$. The dimension p can be larger than the sample size n. We shall work with the assumption of exact sparsity where θ^* is supported on a subset $S \subset \{1, \ldots, p\}$ with $|S| = k$ and k is assumed to be smaller than both p and n. The prediction error of an estimator $\hat{\theta}$ is defined as

$$\frac{1}{n}\left\|X\hat{\theta} - X\theta^*\right\|^2.$$

We shall study the prediction error of the LASSO defined as

$$\hat{\theta}_\lambda^{\text{LASSO}} := \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left(\frac{1}{2}\|Y - X\theta\|^2 + \lambda \|\theta\|_1\right). \tag{152}$$

Before proceeding, let us first recall the following observations:

1. If we know the support S of θ^*, then one can simply estimate θ^* by linear regression of Y on X_S (where X_S is the matrix obtained from X by dropping columns not present in S). It is elementary to check that this *Oracle* estimator will satisfy the prediction error bound:

   $$\frac{1}{n}\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq C\frac{k}{n}$$

   in expectation and in high probability. It is important to note that there are no assumptions on the X matrix here and the bound above is independent of scaling. If I change X by multiplying each column by a constant, then the bound will not change.

2. We have seen in the last class that the BIC estimator defined by

   $$\hat{\theta}_\lambda^{\text{BIC}} := \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left(\|Y - X\theta\|^2 + \lambda^2 \|\theta\|_0\right) \tag{153}$$

   achieves the prediction error bound

   $$\frac{1}{n}\left\|X\hat{\theta}_\lambda^{\text{BIC}} - X\theta^*\right\|^2 \leq C\frac{k}{n}\log(ep)$$

   with high probability and in expectation provided that the tuning parameter λ is chosen as $c_1\sqrt{\log(ep)}$ for a large enough c_1. Therefore compared to the Oracle estimator described above, the BIC estimator only pays a price that is logarithmic in p. However, even though efficient computation of $\hat{\theta}_\lambda^{\text{BIC}}$ is possible for certain special design matrices X (e.g., when $X = I_n$ or when $X(i,j) = I\{i \geq j\}$ for $1 \leq i, j \leq n$), in most cases, it is computationally intractable.

In light of the above two observations, it is most interesting to see if $\hat{\theta}_\lambda^{\text{LASSO}}$ satisfies

$$\frac{1}{n}\left\|X\hat{\theta}_\lambda^{\text{LASSO}} - X\theta^*\right\|^2 \leq C\frac{k}{n}\log(ep). \tag{154}$$

Indeed, unlike $\hat{\theta}_\lambda^{\text{BIC}}$, the lasso estimator is computationally tractable and can be obtained efficiently by convex optimization for fairly large values of n and p. It is therefore of interest to see if any price is to be paid in terms of prediction risk performance (compared to $\hat{\theta}_\lambda^{\text{BIC}}$) for this computational tractability.

We shall see below that (154) will be true **under some assumptions on** X. These assumptions are unfortunately quite restrictive and cannot usually be checked in practice. At a high level, these assumptions can be understood to be saying that X behaves like an identity matrix in a certain sense. Note that we already know that (154) is true when X is the identity matrix (in this case, $\hat{\theta}_\lambda^{\text{LASSO}}$ is the soft thresholding estimator).

18.2 Prediction Error of the LASSO under Exact Sparsity

We shall complete the argument that we started in the last class to establish (154) under some assumptions on X.

As a simple consequence of the basic inequality for the LASSO, we proved, in the last class, the following inequality (where we write $\hat{\theta}$ for $\hat{\theta}_\lambda^{\text{LASSO}}$ for simplicity):

$$\frac{1}{2}\left\|X\hat{\theta} - X\theta^*\right\|^2 + \lambda\left\|\hat{\theta}_{S^c}\right\|_1 \leq \left(\left\|X^T\epsilon\right\|_\infty + \lambda\right)\left\|\hat{\theta}_S - \theta^*_S\right\|_1 + \left\|X^T\epsilon\right\|_\infty \left\|\hat{\theta}_{S^c}\right\|_1. \tag{155}$$

We now take the tuning parameter λ to be such that

$$\lambda \geq 2\left\|X^T\epsilon\right\|_\infty. \tag{156}$$

The following analysis can be done with $\lambda \geq (1+\eta)\left\|X^T\epsilon\right\|_\infty$ for any $\eta > 0$ but it is customary to take $\eta = 1$. The choice (156) for λ immediately implies that $\left\|X^T\epsilon\right\|_\infty \leq \lambda/2$. Using this on the right hand side in (155), we obtain

$$\frac{1}{2}\left\|X\hat{\theta} - X\theta^*\right\|^2 + \lambda\left\|\hat{\theta}_{S^c}\right\|_1 \leq \frac{3\lambda}{2}\left\|\hat{\theta}_S - \theta^*_S\right\|_1 + \frac{\lambda}{2}\left\|\hat{\theta}_{S^c}\right\|_1.$$

which readily simplifies to

$$\left\|X\hat{\theta} - X\theta^*\right\|^2 + \lambda\left\|\hat{\theta}_{S^c}\right\|_1 \leq 3\lambda\left\|\hat{\theta}_S - \theta^*_S\right\|_1.$$

Note that this inequality simultaneously implies the following two inequalities:

$$\left\|\hat{\theta}_{S^c}\right\|_1 \leq 3\left\|\hat{\theta}_S - \theta^*_S\right\|_1 \quad \text{and} \quad \left\|X\hat{\theta} - X\theta^*\right\|^2 \leq 3\lambda\left\|\hat{\theta}_S - \theta^*_S\right\|_1. \tag{157}$$

The first inequality in (157) above can be rewritten as

$$\left\|\hat{\theta}_{S^c} - \theta^*_{S^c}\right\|_1 \leq 3\left\|\hat{\theta}_S - \theta^*_S\right\|_1 \tag{158}$$

because $\theta^*_{S^c} = 0$ (note that S is the support of θ^*). This means that $\hat{\theta} - \theta^* \in \mathcal{C}_S$ where

$$\mathcal{C}_S := \left\{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\right\}.$$

Note that the set $\mathcal{C}_s$ is a convex cone.

The second inequality in (157) is obviously more relevant for proving the prediction risk bound (154) of $\hat{\theta}_\lambda^{\text{LASSO}}$. Indeed, dividing both sides by n, we obtain

$$\frac{1}{n}\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq \frac{3\lambda}{n}\left\|\hat{\theta}_S - \theta^*_S\right\|_1.$$

We shall now plug in an explicit value for λ. Indeed by the assumption that $\epsilon \sim N(0, I_n)$, the assumption (156) will be satisfied with high probability for

$$\lambda = c_1 \max_i \|X_i\| \sqrt{\log(ep)} \tag{159}$$

for a sufficiently large value of c_1. Here $X_1, \ldots, X_p$ denote the columns of X. Plugging this value of λ in the bound above, we obtain

$$\frac{1}{n}\left\|X\hat{\theta} - X\theta^*\right\|^2 \leq \frac{C}{n}\left(\max_i \|X_i\|\right)\sqrt{\log(ep)}\left\|\hat{\theta}_S - \theta^*_S\right\|_1$$

for a constant C depending on c_1. We shall now impose a particular scaling on the columns of X. Specifically, we assume that

$$\|X_i\| = \sqrt{n} \qquad \text{for each } i = 1, \ldots, p. \tag{160}$$

For this scaling, the above bound becomes

$$\frac{1}{n}\left\|X\hat{\theta}-X\theta^*\right\|^2 \leq C\sqrt{\frac{\log(ep)}{n}}\left\|\hat{\theta}_S-\theta^*_S\right\|_1$$

which we can rewrite as

$$\frac{1}{\sqrt{n}}\left\|X\hat{\theta}-X\theta^*\right\| \leq C\sqrt{\frac{k\log(ep)}{n}}\left(\frac{\left\|\hat{\theta}_S-\theta^*_S\right\|_1/\sqrt{k}}{\left\|X\hat{\theta}-X\theta^*\right\|/\sqrt{n}}\right).$$

From here, it is obvious that the required bound (154) holds provided

$$\frac{\left\|\hat{\theta}_S-\theta^*_S\right\|_1/\sqrt{k}}{\left\|X\hat{\theta}-X\theta^*\right\|/\sqrt{n}}$$

is bounded from above by a constant. This is exactly the standard assumption under which (154) is proved. To make the assumption seem less blatant, one usually bounds the above quantity as:

$$\frac{\left\|\hat{\theta}_S-\theta^*_S\right\|_1/\sqrt{k}}{\left\|X\hat{\theta}-X\theta^*\right\|/\sqrt{n}} \leq \left(\inf\frac{\|X\Delta\|/\sqrt{n}}{\|\Delta_S\|_1/\sqrt{k}}\right)^{-1}$$

where the infimum can be taken over any set in $\mathbb{R}^p$ which contains $\hat{\theta}-\theta^*$. Because we know that $\hat{\theta}-\theta^* \in \mathcal{C}_S$ when λ satisfies (156) (see (158)), we can take the infimum above over $\Delta \in \mathcal{C}_S$ which gives

$$\frac{1}{\sqrt{n}}\left\|X\hat{\theta}-X\theta^*\right\| \leq C\sqrt{\frac{k\log(ep)}{n}}\left(\inf_{\Delta\in\mathcal{C}_S:\Delta_S\neq 0}\frac{\|X\Delta\|/\sqrt{n}}{\|\Delta_S\|_1/\sqrt{k}}\right)^{-1}.$$

The quantity

$$\phi(S) := \inf_{\Delta\in\mathcal{C}_S:\Delta_S\neq 0}\frac{\|X\Delta\|/\sqrt{n}}{\|\Delta_S\|_1/\sqrt{k}}$$

is called the *compatibility factor*. We thus have proved that

$$\frac{1}{n}\left\|X\hat{\theta}-X\theta^*\right\|^2 \leq \frac{C}{\phi^2(S)}\frac{k\log(ep)}{n} \tag{161}$$

with high probability and expectation when λ is chosen as in (159). If we now make the assumption that

$$\phi(S) \geq \phi_0 > 0 \tag{162}$$

for a constant ϕ_0, then we have proved that (154) holds. The assumption (162) above is called the *compatibility condition*. It is a property of the design matrix X and the set S (which is the support of θ^*). Because it depends on the unknown θ^*, it cannot be verified in practice. One therefore replaces it by the assumption that

$$\inf_{S\subseteq\{1,\ldots,p\}:|S|\leq k}\phi(S) \geq \phi_0 > 0. \tag{163}$$

This is, in principle, verifiable because it only depends on the design matrix X and the sparsity level k of the unknown θ^*. But, unfortunately, even if k is known, verifying (163) requires going over all subsets of $\{1,\ldots,p\}$ of size $\leq k$ which is computationally intractable.

Let us now quickly go over *restricted eigenvalues* which are closely related to compatibility factors. By the Cauchy-Schwarz inequality, we have

$$\frac{\|\Delta_S\|_1}{\sqrt{k}} \leq \|\Delta_S\| \qquad \text{for every } \Delta \in \mathbb{R}^p.$$

As a result

$$\phi(S) \geq \gamma(S) := \inf_{\Delta \in \mathcal{C}_S : \Delta_S \neq 0} \frac{\|X\Delta\| / \sqrt{n}}{\|\Delta_S\|}.$$

This quantity $\gamma(S)$ is called a restricted eigenvalue of X. Clearly because $\gamma(S)$ is smaller than $\phi(S)$, inequality (161) also holds if $\phi(S)$ is replaced by $\gamma(S)$ which means that (154) holds under the assumption that

$$\inf_{S \subseteq \{1,\ldots,p\} : |S| \leq k} \gamma(S) \geq \gamma_0 > 0 \tag{164}$$

for a constant γ_0. This is called the restricted eigenvalue (RE) conditon which implies (and hence weaker than) the compatibility condition. Unfortunately, checking (164) is also computationally intractable and hopeless in practice.

18.3 A simple sufficient condition for checking the RE and compatibility conditons

The following lemma presents a simple sufficient condition for checking the assumptions (163) and (164). The condition (165) on X that appears in this lemma is sometimes referred to by the phrase: "X is ρ-incoherent".

Lemma 18.1. *Suppose X is an $n \times p$ matrix such that*

$$\left(\frac{X^T X}{n}\right)(i,i) = 1 \text{ for all } 1 \leq i \leq n \quad \text{and} \quad \max_{i \neq j} \left|\frac{X^T X}{n}(i,j)\right| = \rho. \tag{165}$$

Then for every $S \subseteq \{1,\ldots,p\}$ with cardinality k, we have

$$\gamma(S) \geq \sqrt{(1 - 16\rho k)_+} \tag{166}$$

Note that the first assumption in (165) just means that each column of X is normalized to have norm equal to $\sqrt{n}$. Also the conclusion (166) is non-trivial only when $\rho < 1/(16k)$. For example, when $\rho \leq 1/(32k)$, then (166) says that $\gamma(S) \geq 1/\sqrt{2}$.

Proof of Lemma 18.1. We need to prove that

$$\frac{1}{n}\|X\Delta\|^2 \geq (1 - 16\rho k)\|\Delta_S\|^2$$

for every $\Delta \in \mathbb{R}^p$ satisfying $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$. Fix such a Δ and write

$$\begin{aligned}
\frac{1}{n}\|X\Delta\|^2 &= \frac{1}{n}\sum_{i,j} \Delta_i \Delta_j \frac{X^T X}{n}(i,j) \\
&= \sum_{i=1}^n \Delta_i^2 + \sum_{i \neq j} \Delta_i \Delta_j \frac{X^T X}{n}(i,j) \\
&= \|\Delta\|^2 - \rho \sum_{i \neq j} |\Delta_i||\Delta_j| \geq \|\Delta\|^2 - \rho \sum_{i,j} |\Delta_i||\Delta_j| \geq \|\Delta\|^2 - \rho \|\Delta\|_1^2 .
\end{aligned}$$

Note now that $\|\Delta\|_1 = \|\Delta_{S^c}\|_1 + \|\Delta_S\|_1 \leq 4\|\Delta_S\|_1$ under the assumption that $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$. We therefore obtain

$$\frac{1}{n}\|X\Delta\|^2 \geq \|\Delta\|^2 - 16\rho\|\Delta_S\|_1^2 \geq \|\Delta\|^2 - 16\rho k\|\Delta_S\|^2 \geq (1 - 16\rho k)\|\Delta_S\|^2$$

which completes the proof. □

18.4 The Restricted Isometry Property

The Restricted Isometry Property (RIP) is related to the RE and compatibility conditions. It is defined as follows.

Definition 18.2 (RIP)**.** *Let X be an $n \times p$ matrix. For $\delta \in (0,1)$ and $k \leq p$, we say that X has the $RIP(\delta, k)$ property if*

$$(1-\delta)^2 \|\Delta\|^2 \leq \Delta^T \left(\frac{X^T X}{n}\right) \Delta \leq (1+\delta)^2 \|\Delta\|^2 \tag{167}$$

for all $\Delta \in \mathbb{R}^p$ with $\|\Delta\|_0 \leq k$.

For $S \subseteq \{1, \ldots, p\}$, let X_S denote the $n \times |S|$ submatrix of X formed by dropping all the columns X_i of X for $i \notin S$. With this notation, it is easy to see that the above definition of RIP is equivalent to the following definition.

Definition 18.3 (Alternative Definition of RIP)**.** *Let X be an $n \times p$ matrix. For $\delta \in (0,1)$ and $k \leq p$, we say that X has the $RIP(\delta, k)$ property if*

$$(1-\delta)^2 \leq \lambda_{\min}(X_S^T X_S/n) \leq \lambda_{\max}(X_S^T X_S/n) \leq (1+\delta)^2 \tag{168}$$

for every subset $S \subseteq \{1, \ldots, p\}$ with $|S| \leq k$. Here $\lambda_{\min}$ and $\lambda_{\max}$ refer to the smallest and largest eigenvalue respectively.

From (168), it is clear that for $X_{n \times p}$ to satisfy the $RIP(\delta, k)$ property, it is necessary that $n \geq k$. The following result shows that the RIP property implies the RE condition.

Lemma 18.4. *Suppose X satisfies $RIP(\delta, k+m)$. Then for every $S \subseteq \{1, \ldots, p\}$ with $|S| \leq k$, we have*

$$\gamma(S) \geq \left(1 - \delta - 3(1+\delta)\sqrt{\frac{k}{m}}\right)_+. \tag{169}$$

From (169), it is trivial to deduce the following:

$$\gamma(S) \geq (1-\delta)\frac{u}{1+u} \qquad \text{provided } m \geq (1+u)^2 k \left(\frac{3(1+\delta)}{1-\delta}\right)^2.$$

This means

$$RIP\left(\delta, k\left\{1 + (1+u)^2\left(\frac{3(1+\delta)}{1-\delta}\right)^2\right\}\right) \implies \inf_{S \subseteq \{1,\ldots,p\}: |S| \leq k} \gamma(S) \geq (1-\delta)\frac{u}{1+u}.$$

For example, by taking $\delta = 1/4$ and $u = 1/5$, we obtain

$$RIP\left(\frac{1}{4}, 37k\right) \implies \inf_{S \subseteq \{1,\ldots,p\}: |S| \leq k} \gamma(S) \geq \frac{1}{8}.$$

Thus, using Lemma 18.4, we can deduce a positive lower bound on $\inf_{S \subseteq \{1,\ldots,p\}: |S| \leq k} \gamma(S)$ provided $RIP(\delta, m)$ holds for a small constant δ and m equal to a constant (depending on δ) multiple of k.

Proof of Lemma 18.4. We need to prove that

$$\frac{1}{\sqrt{n}} \|X\Delta\| \geq \left(1 - \delta - 3\sqrt{\frac{k}{m}}(1+\delta)\right)_+ \|\Delta_S\|$$

for every $\Delta \in \mathbb{R}^p$ satisfying $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$. Fix such a vector Δ. Let I_1 consist of the indices in S^c corresponding to the m largest (in absolute value) entries of Δ. Also, let I_2 consist of the indices in $(S \cup I_1)^c$ corresponding to the m largest (in absolute value) entries of Δ. Continue this way to define a partition $I_1, \ldots, I_l$ of S^c with

$$|I_1| = \cdots = |I_{l-1}| = m \quad \text{and} \quad |I_l| \leq m.$$

We now write

$$\Delta = \Delta_S + \Delta_{I_1} + \cdots + \Delta_{I_l} = \Delta_{S \cup I_1} + \sum_{i=2}^{l} \Delta_{I_i}.$$

so that

$$\frac{1}{\sqrt{n}}\|X\Delta\| = \frac{1}{\sqrt{n}}\left\|X\left(\Delta_{S\cup I_1} + \sum_{i=2}^{l}\Delta_i\right)\right\| \geq \frac{1}{\sqrt{n}}\|X\Delta_{S\cup I_1}\| - \frac{1}{\sqrt{n}}\sum_{i=2}^{l}\|X\Delta_{I_i}\|.$$

We now use the fact that X satisfies $RIP(\delta, k+m)$ which gives (note that $|S \cup I_1| \leq k+m$ and $|I_i| \leq m$ for all i)

$$\frac{1}{\sqrt{n}}\|X\Delta\| \geq (1-\delta)\|\Delta_{S\cup I_1}\| - (1+\delta)\sum_{i=2}^{l}\|\Delta_{I_i}\| \geq (1-\delta)\|\Delta_S\| - (1+\delta)\sum_{i=2}^{l}\|\Delta_{I_i}\|.$$

Now, by construction, the absolute value of every entry in Δ_{I_i} is smaller than the absolute value of every entry in $\Delta_{I_{i-1}}$. This implies, in particular, that the absolute value of every entry in Δ_{I_i} is smaller than the average of the absolute values of entries in $\Delta_{I_{i-1}}$. This gives

$$\|\Delta_{I_i}\|^2 \leq \left(\frac{1}{m}\|\Delta_{i-1}\|_1\right)^2 + \cdots + \left(\frac{1}{m}\|\Delta_{i-1}\|_1\right)^2 = \frac{1}{m}\left\|\Delta_{I_{i-1}}\right\|_1^2 \qquad \text{for every } 2 \leq i \leq l-1.$$

As a result, we obtain

$$\frac{1}{\sqrt{n}}\|X\Delta\| \geq (1-\delta)\|\Delta_S\| - \frac{1+\delta}{\sqrt{m}}\sum_{i=2}^{l}\left\|\Delta_{I_{i-1}}\right\|_1 \geq (1-\delta)\|\Delta_S\| - \frac{1+\delta}{\sqrt{m}}\|\Delta_{S^c}\|_1.$$

Using the cone condition $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$, we further deduce

$$\frac{1}{\sqrt{n}}\|X\Delta\| \geq (1-\delta)\|\Delta_S\| - 3\frac{1+\delta}{\sqrt{m}}\|\Delta_S\|_1 \geq (1-\delta)\|\Delta_S\| - 3\frac{1+\delta}{\sqrt{m}}\sqrt{k}\|\Delta_S\| \geq \left(1 - \delta - 3(1+\delta)\sqrt{\frac{k}{m}}\right)\|\Delta_S\|$$

which completes the proof. □

An important fact is that the RIP will be satisfied by certain kinds of random matrices X with high probability. It then means (by the above lemma) that for such matrices X, the RE condition will hold with high probability. The simplest example of this is when the entries of X are i.i.d standard Gaussian. This is proved below (the proof is taken from Baraniuk et al. [1]; see the paper for extensions to some other random ensembles).

Theorem 18.5. *Suppose that the entries of the $n \times p$ matrix X are independent and identically distributed as $N(0,1)$. Then X satisfies $RIP(\delta, k)$ with probability at least $1 - \exp(-n\delta^2/64)$ provided*

$$n \geq \frac{64}{\delta^2}\log\left(\frac{9e}{\delta}\right)k\log\left(\frac{ep}{k}\right).$$

Proof. It is easy to see that because the entries of X are i.i.d $N(0,1)$, we have

$$\frac{\|X\Delta\|^2}{\|\Delta\|^2} \sim \chi^2_n.$$

Now χ^2_n random variables satisfy the standard concentration inequality (whose proof is left as exercise):

$$\mathbb{P}\left\{\left|\frac{\chi^2_n}{n}-1\right|\geq t\right\}\leq 2\exp\left(-nt^2/8\right)\qquad\text{for all } 0\leq t\leq 1.$$

This immediately gives that for every $\Delta\in\mathbb{R}^p$ and $\delta\in(0,1)$

$$\mathbb{P}\left\{\left(1-\frac{\delta}{2}\right)\|\Delta\|^2\leq\Delta^T\frac{X^TX}{n}\Delta\leq\left(1+\frac{\delta}{2}\right)\|\Delta\|^2\right\}\geq 1-2\exp\left(\frac{-n\delta^2}{32}\right).$$

Because $1+\delta/2\leq(1+\delta/2)^2$ and $1-\delta/2\geq(1-\delta/2)^2$, we also have

$$\mathbb{P}\left\{\left(1-\frac{\delta}{2}\right)^2\|\Delta\|^2\leq\Delta^T\frac{X^TX}{n}\Delta\leq\left(1+\frac{\delta}{2}\right)^2\|\Delta\|^2\right\}\geq 1-2\exp\left(\frac{-n\delta^2}{32}\right).\tag{170}$$

Now let $\Delta_1,\ldots,\Delta_M$ be a maximal $\delta/4$-packing subset (in the usual Euclidean metric) of the set

$$\{\Delta:\|\Delta\|=1\text{ and }\|\Delta\|_0\leq k\}.$$

By a standard volumetric argument, it can be shown that

$$M\leq\left(1+\frac{2}{\delta/4}\right)^k\left(\binom{p}{0}+\binom{p}{1}+\cdots+\binom{p}{k}\right)\leq\left(\frac{9}{\delta}\right)^k\left(\frac{ep}{k}\right)^k.$$

By the union bound, it follows from (170) that the probability

$$\Upsilon:=\mathbb{P}\left\{\left(1-\frac{\delta}{2}\right)^2\|\Delta_j\|^2\leq\Delta_j^T\frac{X^TX}{n}\Delta_j\leq\left(1+\frac{\delta}{2}\right)^2\|\Delta_j\|^2\text{ for all }1\leq j\leq M\right\}$$

satisfies the bound

$$\begin{aligned}\Upsilon&\geq 1-2M\exp\left(\frac{-n\delta^2}{32}\right)\\&\geq 1-2\left(\frac{9ep}{k\delta}\right)^k\exp\left(\frac{-n\delta^2}{32}\right)=1-2\exp\left(-n\left\{\frac{\delta^2}{32}-\frac{k}{n}\log\left(\frac{9ep}{k\delta}\right)\right\}\right).\end{aligned}$$

Suppose now that $n\geq c_1k\log(ep/k)$ for some constant c_1. Then

$$\frac{\delta^2}{32}-\frac{k}{n}\log\left(\frac{9ep}{k\delta}\right)\geq\frac{\delta^2}{32}-\frac{1}{c_1}\left(1+\frac{\log(9/\delta)}{\log(ep/k)}\right)\geq\frac{\delta^2}{32}-\frac{1}{c_1}\log\left(\frac{9e}{\delta}\right).$$

Thus when

$$c_1=\frac{64}{\delta^2}\log\left(\frac{9e}{\delta}\right)$$

we have

$$\frac{\delta^2}{32}-\frac{k}{n}\log\left(\frac{9ep}{k\delta}\right)\geq\frac{\delta^2}{64}$$

so that

$$\Upsilon\geq 1-2\exp\left(\frac{-n\delta^2}{64}\right).$$

To complete the proof therefore, we only need to argue that

$$\left(1-\frac{\delta}{2}\right)^2\|\Delta_j\|^2\leq\Delta_j^T\frac{X^TX}{n}\Delta_j\leq\left(1+\frac{\delta}{2}\right)^2\|\Delta_j\|^2\qquad\text{for all } j=1,\ldots,M\tag{171}$$

implies that

$$(1-\delta)^2 \|\Delta\|^2 \leq \Delta^T \frac{X^T X}{n} \Delta \leq (1+\delta)^2 \|\Delta\|^2 \qquad \text{for all } \Delta \text{ with } \|\Delta\| = 1 \text{ and } \|\Delta\|_0 \leq k. \tag{172}$$

The argument for proving this implication is the following. Let A be the smallest number for which

$$\Delta^T \frac{X^T X}{n} \Delta \leq (1+A)^2 \|\Delta\|^2 \qquad \text{for all } \Delta \text{ with } \|\Delta\| = 1 \text{ and } \|\Delta\|_0 \leq k. \tag{173}$$

We shall show that $A \leq \delta$. Note first that $A < \infty$ because $\lambda_{\max}(X^T X/n) < \infty$. Now fix Δ such that $\|\Delta\| = 1$ and $\|\Delta\|_0 \leq k$. By the packing property and construction of $\{\Delta_1, \ldots, \Delta_M\}$, there will exist $1 \leq j \leq M$ such that $\|\Delta - \Delta_j\| \leq \delta/4$ and $\|\Delta - \Delta_j\|_0 \leq k$. Write

$$\Delta^T \frac{X^T X}{n} \Delta = \left\| \frac{1}{\sqrt{n}} X\Delta \right\| \leq \left\| \frac{1}{\sqrt{n}} X\Delta_j \right\| + \left\| \frac{1}{\sqrt{n}} X(\Delta - \Delta_j) \right\| \leq \left(1 + \frac{\delta}{2}\right) \|\Delta_j\| + (1+A) \|\Delta - \Delta_j\|$$

where to get the final inequality we used (171) and (173) with Δ replaced by $\Delta - \Delta_j$. Because $\|\Delta_j\| = 1$ and $\|\Delta - \Delta_j\| \leq \delta/4$, we obtain

$$\Delta^T \frac{X^T X}{n} \Delta \leq 1 + \frac{\delta}{2} + (1+A)\frac{\delta}{4}.$$

Comparing this with (173), we deduce that (by the definition of A)

$$A \leq \frac{\delta}{2} + (1+A)\frac{\delta}{4}$$

which gives $A \leq \delta$. This proves the upper inequality in (172). To prove the lower inequality, write

$$\left\| \frac{1}{\sqrt{n}} X\Delta \right\| \geq \left\| \frac{1}{\sqrt{n}} X\Delta_j \right\| - \left\| \frac{1}{\sqrt{n}} X\left(\Delta - \Delta_j\right) \right\|$$

Using (171) and (173) with $A = \delta$ (note that we can choose Δ_j so that $\|\Delta - \Delta_j\|_0 \leq k$), we get

$$\left\| \frac{1}{\sqrt{n}} X\Delta \right\| \geq 1 - \frac{\delta}{2} - (1+\delta)\|\Delta - \Delta_j\| \geq 1 - \frac{\delta}{2} - (1+\delta)\frac{\delta}{4} = 1 - \frac{3\delta}{4} - \frac{\delta^2}{4} \geq 1 - \delta.$$

This proves the lower bound in (172) and completes the proof of the theorem. □

19 Lecture 19

The next topic of the class is convergence of stochastic processes. Our main motivation for studying this is to prove limiting distribution results for M-estimators. We shall start with two classical examples.

19.1 Limiting Distribution of Sample Median

Suppose $X_1, \ldots, X_n$ are i.i.d observations from the normal density f with mean θ_0 and variance 1. Actually, it will be clear that results below hardly require normality and hold more generally but let us assume that f is $N(\theta_0, 1)$ for simplicity. Let $\hat{\theta}_n$ denote a sample median based on $X_1, \ldots, X_n$ defined as any minimizer of

$$M_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} |X_i - \theta|$$

over $\theta \in \mathbb{R}$. Also let $M(\theta) := \mathbb{E}|X_1 - \theta|$ and note that θ_0 uniquely minimizes $M(\theta)$ over $\theta \in \mathbb{R}$.

We have seen in one of the homeworks that $\hat{\theta}_n$ converges to θ_0 in probability i.e., $\hat{\theta}_n$ is a consistent estimator of θ_0. Our general rate theorem can also be applied directly here to deduce that $\hat{\theta}_n - \theta_0 = O_P(n^{-1/2})$ i.e., the rate of convergence of $\hat{\theta}_n$ to θ_0 is $n^{-1/2}$. We shall now address the question of finding the limiting or asymptotic distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$. There are many approaches for finding this limiting distribution but we shall follow the standard empirical processes approach which easily generalizes to other M-estimators. This approach also highlights the need to study convergence of stochastic processes.

Our approach for finding the limiting distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ is based on the following *localized, centered* and *rescaled* stochastic process:

$$\tilde{M}_n(h) := n\left(M_n(\theta_0 + n^{-1/2}h) - M_n(\theta_0)\right) \qquad \text{for } h \in \mathbb{R}.$$

This is a stochastic process that is indexed by $h \in \mathbb{R}$. Its important property (easy to see) is that $\hat{h}_n := \sqrt{n}(\hat{\theta}_n - \theta_0)$ minimizes $\tilde{M}_n(h), h \in \mathbb{R}$ i.e.,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \operatorname*{argmin}_{h \in \mathbb{R}} \tilde{M}_n(h).$$

This suggests the following approach to find the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. We study the process $\tilde{M}_n(h), h \in \mathbb{R}$ and argue that it converges as $n \to \infty$ to some limit process $\tilde{M}(h), h \in \mathbb{R}$ in an appropriate sense. If this process convergence is strong enough, then we can hopefully argue that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \operatorname*{argmin}_{h \in \mathbb{R}} \tilde{M}_n(h) \xrightarrow{L} \operatorname*{argmin}_{h \in \mathbb{R}} \tilde{M}(h).$$

It is actually not too hard to understand the behavior of $\tilde{M}_n(h)$ as $n \to \infty$ for each fixed $h \in \mathbb{R}$. For this, we can write

$$\begin{aligned}
\tilde{M}_n(h) &= n\left(M_n(\theta_0 + n^{-1/2}h) - M_n(\theta_0)\right) \\
&= n\left(M_n(\theta_0 + n^{-1/2}h) - M(\theta_0 + n^{-1/2}h) - M_n(\theta_0) + M(\theta_0)\right) + n\left(M(\theta_0 + n^{-1/2}h) - M(\theta_0)\right) \\
&=: A_n + B_n. \qquad (174)
\end{aligned}$$

Let us now analyze A_n and B_n separately. Clearly, B_n is a deterministic sequence. To understand this, we shall use a second order Taylor explansion for $M(\theta_0 + n^{-1/2}h)$ around θ_0. Note that $M(\theta) := \mathbb{E}|X_1 - \theta|$ is a smooth function. Also note that $M'(\theta_0) = 0$ because θ_0 maximizes $M(\theta), \theta \in \mathbb{R}$. We thus get

$$B_n = n\left(M(\theta_0 + n^{-1/2}h) - M(\theta_0)\right) = \frac{1}{2}M''(\tilde{\theta}_n)h^2$$

where $\tilde{\theta}_n$ is some number between θ_0 and $\theta_0 + n^{-1/2}h$. Clearly $\tilde{\theta}_n \to \theta_0$ as $n \to \infty$ so that

$$B_n \to \frac{1}{2}M''(\theta_0)h^2 \qquad \text{as } n \to \infty.$$

Let us now come to the mean zero random variable A_n. To understand it, let us first compute its variance:

$$\begin{aligned}
\operatorname{var}(A_n) &= n^2 \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n}\left\{|X_i - \theta_0 - n^{-1/2}h| - |X_i - \theta_0|\right\}\right) \\
&= n \operatorname{var}\left(|X_1 - \theta_0 - n^{-1/2}h| - |X_1 - \theta_0|\right) \\
&\approx n \operatorname{var}\left(I\{X_1 < \theta_0\}n^{-1/2}h - I\{X_1 > \theta_0\}n^{-1/2}h\right)
\end{aligned}$$

where I have ignored the contribution from X_1 lying between θ_0 and $\theta_0 + n^{-1/2}h$ (should not matter for large n; verify this). This gives

$$\operatorname{var} A_n \approx h^2 \operatorname{var}\left(I\{X_1 < \theta_0\} - I\{X_1 > \theta_0\}\right).$$

Now because $\mathbb{P}\{X_1 < \theta_0\} = \mathbb{P}\{X_1 > \theta_0\}$ (θ_0 is a population median), it is easy to check that the variance of $I\{X_1 < \theta_0\} - I\{X_1 > \theta_0\}$ appearing above equals 1. We have therefore obtained

$$\text{var}(A_n) \to h^2 \qquad \text{as } n \to \infty.$$

It is actually possible to prove that

$$A_n \xrightarrow{L} N(0, h^2) = hN(0,1) \qquad \text{as } n \to \infty.$$

For this, we can use the Lindeberg-Feller Central Limit Theorem (stated next).

19.2 Lindeberg-Feller Central Limit Theorem

Theorem 19.1. *For each n, let $Y_{n1}, \ldots, Y_{nk_n}$ be k_n independent random vectors with $\mathbb{E}\|Y_{ni}\|^2 < \infty$ for each $i = 1, \ldots, k_n$. Suppose the following two conditions hold:*

$$\sum_{i=1}^{k_n} \text{Cov}(Y_{ni}) \to \Sigma \qquad \text{as } n \to \infty \tag{175}$$

where $\text{Cov}(Y_{ni})$ denotes the covariance matrix of the random vector Y_{ni} and

$$\sum_{i=1}^{k_n} \mathbb{E}\left(\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right) \to 0 \qquad \text{as } n \to \infty \text{ for every } \epsilon > 0. \tag{176}$$

Then

$$\sum_{i=1}^{k_n} (Y_{ni} - \mathbb{E}Y_{ni}) \xrightarrow{L} N(0, \Sigma) \qquad \text{as } n \to \infty. \tag{177}$$

For a proof of this result, see, for example, Pollard [20, Page 181]. It is easy to see that this result generalizes the usual CLT. Indeed, the usual CLT states that for i.i.d random variables $X_1, X_2, \ldots$ with $\mathbb{E}X_i = \mu$, $\mathbb{E}\|X_i\|^2 < \infty$ and $\text{Cov}(X_i) = \Sigma$, we have

$$\sum_{i=1}^{n} \left(\frac{X_i}{\sqrt{n}} - \frac{\mu}{\sqrt{n}}\right) \xrightarrow{L} N(0, \Sigma) \qquad \text{as } n \to \infty.$$

Indeed this can be proved by applying Theorem 19.1 to

$$Y_{ni} = \frac{X_i}{\sqrt{n}}.$$

The condition (175) is obvious while for (176) note that

$$\sum_{i=1}^{n} \mathbb{E}\left(\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\|X_i\|^2 I\{\|X_i\| > \sqrt{n}\epsilon\} = \mathbb{E}\left(\|X_1\|^2 I\{\|X_1\| > \sqrt{n}\epsilon\}\right)\right)$$

which clearly converges to zero by the Dominated Convergence Theorem (under the assumption $\mathbb{E}\|X_1\|^2 < \infty$).

19.3 Back to the Limiting Distribution of Sample Median

Recall the random variables A_n from (174). The Lindeberg-Feller CLT can be used to prove that $A_n \xrightarrow{L} N(0, h^2)$. Note first that

$$A_n = \sum_{i=1}^{n} (Y_{ni} - \mathbb{E}Y_{ni})$$

where

$$Y_{ni} := \left|X_i - \theta_0 - n^{-1/2}h\right| - |X_i - \theta_0|\,.$$

We have already checked that

$$\sum_{i=1}^{n} \mathrm{var}(Y_{ni}) = \mathrm{var}(A_n) \to h^2 \qquad \text{as } n \to \infty.$$

To check (176), note that

$$\begin{aligned}\sum_{i=1}^{k_n} \mathbb{E}\left(\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right) &= \sum_{i=1}^{n} \mathbb{E}\left(\left||X_i - \theta_0 - n^{-1/2}h| - |X_i - \theta_0|\right|^2 I\{\left||X_i - \theta_0 - n^{-1/2}h| - |X_i - \theta_0|\right| > \epsilon\}\right)\\ &= n\mathbb{E}\left(\left||X_1 - \theta_0 - n^{-1/2}h| - |X_1 - \theta_0|\right|^2 I\{\left||X_1 - \theta_0 - n^{-1/2}h| - |X_1 - \theta_0|\right| > \epsilon\}\right)\end{aligned}$$

Using the trivial inequality

$$\left||X_1 - \theta_0 - n^{-1/2}h| - |X_1 - \theta_0|\right| \le n^{-1/2}|h|,$$

we obtain

$$\sum_{i=1}^{k_n} \mathbb{E}\left(\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right) \le h^2 I\{n^{-1/2}|h| > \epsilon\} \to 0 \qquad \text{as } n \to \infty.$$

The conditions of Theorem 19.1 therefore hold and we obtain

$$A_n \xrightarrow{L} N(0, h^2) \qquad \text{as } n \to \infty.$$

Thus if we define

$$\tilde{M}(h) := hZ + \frac{1}{2}h^2 M''(\theta_0) \qquad \text{for } h \in \mathbb{R}$$

where $Z \sim N(0,1)$, then we have shown that

$$\tilde{M}_n(h) \xrightarrow{L} \tilde{M}(h) \qquad \text{as } n \to \infty \text{ for every } h \in \mathbb{R}.$$

It turns out that the process $\tilde{M}_n$ converges to $\tilde{M}$ in a stronger sense than convergence in distirbution for each fixed $h \in \mathbb{R}$. We shall see this later. This stronger convergence allows us to deduce that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} \operatorname*{argmin}_{h \in \mathbb{R}} \left(hZ + \frac{1}{2}h^2 M''(\theta_0)\right).$$

The argmax above can be written in closed form (note that we have a quadratic in h) so that we get

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} \operatorname*{argmin}_{h \in \mathbb{R}} \left(hZ + \frac{1}{2}h^2 M''(\theta_0)\right) = \frac{-Z}{M''(\theta_0)} \sim N\left(0, \frac{1}{(M''(\theta_0))^2}\right).$$

We can simplify this slightly by writing $M''(\theta_0)$ in terms of $f(\theta_0)$. Indeed, first write

$$M(\theta) = \mathbb{E}|X_1 - \theta| = \int_{-\infty}^{\theta} (\theta - x) f(x)dx + \int_{\theta}^{\infty} (x - \theta) f(x)dx = \theta\,(2F(\theta) - 1) + \int_{\theta}^{\infty} x f(x)dx - \int_{-\infty}^{\theta} x f(x)dx$$

where F is the cdf corresponding to f. This gives

$$M'(\theta) = 2\theta f(\theta) + 2(F(\theta) - 1) - 2\theta f(\theta) = 2(F(\theta) - 1)$$

and $M''(\theta) = 2f(\theta)$. We thus have

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N\left(0, \frac{1}{4f^2(\theta_0)}\right)$$

To make this argument rigorous, we have to prove that the stochastic process $\tilde{M}_n$ converges to $\tilde{M}$ in a strong enough sense so that their argmins also converge.

19.4 Limiting Distribution of Sample Mode

The general method given in the preceding section to derive the limiting distribution of sample median is quite broad and can be used for other M-estimators as well. To illustrate this, let us apply this to determine the limiting distribution of sample model. Let $X_1, \ldots, X_n$ de i.i.d observations from the normal density f with mean θ_0 and variance 1. Again the results do not require normality (and also hold if, for example, f is the Cauchy density centered at θ_0) but let us assume f is $N(\theta_0, 1)$ for simplicity.

Let $\hat{\theta}_n$ denote any sample mode which is defined as any maximizer of

$$M_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} I\left\{X_i \in [\theta - 1, \theta + 1]\right\}$$

over $\theta \in \mathbb{R}$. Also let

$$M(\theta) := \mathbb{P}\left\{X_1 \in [\theta - 1, \theta + 1\right\} = \int_{\theta-1}^{\theta+1} f(x)dx. \tag{178}$$

We have previously seen that $\hat{\theta}_n$ is a consistent estimator of θ_0 and that

$$\hat{\theta}_n - \theta_0 = O_P(n^{-1/3}).$$

We shall now heuristically determine the limiting distribution of $n^{1/3}(\hat{\theta}_n - \theta_0)$. The necessary process convergence results needed to rigorize the argument will be given later. To study $\hat{h}_n := n^{1/3}(\hat{\theta}_n - \theta_0)$, it is natural to define the process

$$\tilde{M}_n(h) := n^{2/3}\left(M_n(\theta_0 + n^{-1/3}h) - M_n(\theta_0)\right) \qquad \text{for } h \in \mathbb{R}$$

and note that $\hat{h}_n$ maximizes $\tilde{M}_n(h)$ over $h \in \mathbb{R}$. Let us try to understand the behavior of $\tilde{M}_n(h)$ as $n \to \infty$ for each fixed $h \in \mathbb{R}$. First write

$$\begin{aligned}
\tilde{M}_n(h) &= n^{2/3}\left(M_n(\theta_0 + n^{-1/3}h) - M_n(\theta_0)\right) \\
&= n^{2/3}\left(M_n(\theta_0 + n^{-1/3}h) - M(\theta_0 + n^{-1/3}h) - M_n(\theta_0) + M(\theta_0)\right) + n^{2/3}\left(M(\theta_0 + n^{-1/3}h) - M(\theta_0)\right) \\
&=: A_n + B_n.
\end{aligned}$$

The expectation term B_n is handled exactly as in the median case by a second order Taylor expansion of the smooth function $M(\theta_0 + n^{-1/3}h)$ at θ_0 (note that $M'(\theta_0) = 0$) to obtain

$$B_n = n^{2/3}\left(M(\theta_0 + n^{-1/3}h) - M(\theta_0)\right) \to \frac{1}{2}M''(\theta_0)h^2 \qquad \text{as } n \to \infty. \tag{179}$$

For the stochastic term A_n, let us, as before, start by computing its variance:

$$\text{var}(A_n) = n^{1/3}\text{var}\left(I\{\theta_0 + n^{-1/3}h - 1 \le X_1 \le \theta_0 + n^{-1/3} + 1\} - I\{\theta_0 - 1 \le X_1 \le \theta_0 + 1\}\right).$$

If $h > 0$ and n is large, it is easy to see that

$$\text{var}(A_n) \approx n^{1/3}\text{var}(I_1 - I_2)$$

where

$$I_1 := I\{\theta_0 + 1 \le X_1 \le \theta_0 + 1 + n^{-1/3}h\} \quad \text{and} \quad I_2 := I\{\theta_0 - 1 \le X_1 < \theta_0 + n^{-1/3}h - 1\}.$$

As a result

$$\begin{aligned}\text{var}(A_n) \approx n^{1/3}\text{var}(I_1 - I_2) &= n^{1/3}\left[\mathbb{E}(I_1 - I_2)^2 - (\mathbb{E}I_1 - \mathbb{E}I_2)^2\right] \\ &= n^{1/3}\left[\mathbb{E}\left(I_1 + I_2\right) - (\mathbb{E}I_1 - \mathbb{E}I_2)^2\right] \\ &= n^{1/3}\left[\int_{\theta_0+1}^{\theta_0+1+n^{-1/3}h} f(x)dx + \int_{\theta_0-1}^{\theta_0-1+n^{-1/3}h} f(x)dx + o(n^{-1/3})\right] \\ &\to h\left(f(\theta_0+1) + f(\theta_0-1)\right).\end{aligned}$$

For $h < 0$, one would have to replace h by $-h$ above. Therefore, for every $h \in \mathbb{R}$, we have

$$\text{var}(A_n) \to |h|\left(f(\theta_0+1) + f(\theta_0-1)\right) \qquad \text{as } n \to \infty.$$

In fact, by the Lindeberg-Feller CLT (as in the case of the median), it can be shown that (this is left as homework)

$$A_n \xrightarrow{L} N\left(0, |h|\left\{f(\theta_0+1) + f(\theta_0-1)\right\}\right) \qquad \text{as } n \to \infty.$$

Combining this with (179), we obtain

$$\tilde{M}_n(h) \xrightarrow{L} \sqrt{f(\theta_0+1) + f(\theta_0-1)}N(0,|h|) + \frac{h^2}{2}M''(\theta_0).$$

as $n \to \infty$ for every $h \in \mathbb{R}$. Suppose now that $\{B_h, h \in \mathbb{R}\}$ is a two-sided Brownian motion starting at zero i.e. $B_0 = 0$ and $\{B_h, h \geq 0\}$ is a standard Brownian motion and $\{B_{-h}, h \geq 0\{$ is another standard Brownian motion that is independent of $\{B_h, h \geq 0\}$. Note that $B_h \sim N(0,|h|)$ for every $h \in \mathbb{R}$. If we now define

$$\tilde{M}(h) := \sqrt{f(\theta_0+1) + f(\theta_0-1)}B_h + \frac{h^2}{2}M''(\theta_0),$$

we have

$$\tilde{M}_n(h) \xrightarrow{L} \tilde{M}(h) \qquad \text{as } n \to \infty.$$

Our arguments above can be strengthened to argue that $(\tilde{M}_n(h_1), \ldots, \tilde{M}_n(h_k))$ converge in distribution to $(\tilde{M}(h_1), \ldots, \tilde{M}(h_k))$ for every fixed points $h_1, \ldots, h_k \in \mathbb{R}$. This is a consequence of the Lindeberg-Feller CLT and left as an exercise. We shall see later that $\tilde{M}_n$ converges to $\tilde{M}$ in a much stronger sense than just for every fixed $k \geq 1$ and points $h_1, \ldots, h_k \in \mathbb{R}$. This stronger convergence allows us to conclude that

$$n^{1/3}\left(\hat{\theta}_n - \theta_0\right) = \operatorname*{argmax}_{h\in\mathbb{R}} \tilde{M}_n(h) \xrightarrow{L} \operatorname*{argmax}_{h\in\mathbb{R}} \tilde{M}(h).$$

Because of (178), it is easy to see that $M''(\theta_0) = f'(\theta_0+1) - f'(\theta_0-1)$. We have therefore deduced (non-rigorously) that the limiting distribution of $n^{1/3}\left(\hat{\theta}_n - \theta_0\right)$ is given by

$$\operatorname*{argmax}_{h\in\mathbb{R}}\left\{\sqrt{f(\theta_0+1) + f(\theta_0-1)}B_h - \frac{h^2}{2}\left(f'(\theta_0-1) - f'(\theta_0+1)\right)\right\}.$$

Note that $f'(\theta_0-1) - f'(\theta_0+1) > 0$. The distribution of the random variable above is related to the Chernoff distribution (see `https://en.wikipedia.org/wiki/Chernoff%27s_distribution`).

In the next lecture, we shall see more examples of process convergence and move toward understanding and formalizing this notion more rigorously.

20 Lecture 20

We shall start our formal study of the theory of convergence of stochastic processes in this lecture. To understand the general ideas, it is helpful to look at the special case of the uniform empirical process.

20.1 The Uniform Empirical Process

Suppose $X_1, \ldots, X_n$ are i.i.d random variables that are uniformly distributed on $[0,1]$. For each $t \in [0,1]$, let

$$U_n(t) := \sqrt{n}\,(F_n(t) - t) = \sqrt{n}\left(\frac{1}{n} I\{X_i \leq t\} - t\right). \tag{180}$$

The collection of random variables $\{U_n(t) : 0 \leq t \leq 1\}$ represents a stochastic process indexed by $[0,1]$. Every realization of this process (which corresponds to every realization of $X_1, \ldots, X_n$) is a function on $[0,1]$ that is bounded (and also right continuous having left limits at every point in $(0,1]$). Note that realizations of $\{U_n(t) : t \in [0,1]\}$ are not continuous functions.

By the usual Multivariate Central Limit Theorem, for every $k \geq 1$ and $t_1, \ldots, t_k \in [0,1]$,

$$(U_n(t_1), \ldots, U_n(t_k)) \xrightarrow{L} N_k(0, \Sigma) \qquad \text{as } n \to \infty$$

where Σ is given by $\Sigma(i,j) := \min(t_i, t_j) - t_i t_j$.

The Brownian Bridge is a stochastic process $\{U(t) : 0 \leq t \leq 1\}$ is a stochastic process indexed by $[0,1]$ that is defined by the following two properties:

1. Every realization of $U(t), t \in [0,1]$ is continuous function on $[0,1]$ with $U(0) = U(1) = 0$.
2. For every fixed $t_1, \ldots, t_k \in [0,1]$, the random vector $(U(t_1), \ldots, U(t_k))$ has the multivariate normal distribution with mean vector 0 and covariance matrix Σ given by $\Sigma(i,j) := \min(t_i, t_j) - t_i t_j$.

Based on the above, it is clear that for every $k \geq 1$ and $t_1, \ldots, t_k \in [0,1]$, we have

$$(U_n(t_1), \ldots, U_n(t_k)) \xrightarrow{L} (U(t_1), \ldots, U(t_k)) \qquad \text{as } n \to \infty \tag{181}$$

and this is a consequence of the usual CLT. By definition of convergence in distribution, the statement (181) means that

$$\mathbb{E}g(U_n(t_1), \ldots, U_n(t_k)) \to \mathbb{E}g(U(t_1), \ldots, U(t_k)) \qquad \text{as } n \to \infty \tag{182}$$

for every bounded, continuous function $g : \mathbb{R}^k \to \mathbb{R}$. It is also true that (182) holds for all bounded continuous functions $g : \mathbb{R}^k \to \mathbb{R}$ if and only if (182) holds for all bounded Lipschitz functions $g : \mathbb{R}^k \to \mathbb{R}$. For a proof of this equivalence, see, for example, Pollard [20].

The result (181) can therefore be rephrased in the following manner: the expectation of any bounded continuous function of the stochastic process U_n that depends on U_n only through its values at a finite set of points in $[0,1]$ converges to the corresponding expectation of the Brownian Bridge U. For example, this implies that

$$\mathbb{E}h(\max_{1 \leq i \leq k} |U_n(t_i)|) \to \mathbb{E}h(\max_{1 \leq i \leq k} |U(t_i)|) \qquad \text{as } n \to \infty. \tag{183}$$

for every bounded continuous function $h : \mathbb{R} \to \mathbb{R}$ which is equivalent to

$$\max_{1 \leq i \leq k} |U_n(t_i)| \xrightarrow{L} \max_{1 \leq i \leq k} |U(t_i)| \qquad \text{as } n \to \infty.$$

While this is useful, one often needs to deal with functions of U_n that depend on the entire process U_n and not just at its values at a finite set of points. For example, it is of interest (for example, for the statistical application of testing goodness of fit via the Kolmogorov-Smirnov test) to ask if

$$\sup_{0 \leq t \leq 1} |U_n(t)| \xrightarrow{L} \sup_{0 \leq t \leq 1} |U(t)| \qquad \text{as } n \to \infty$$

which is equivalent to

$$\mathbb{E}h(\sup_{0 \leq t \leq 1} |U_n(t)|) \to \mathbb{E}h(\sup_{0 \leq t \leq 1} |U(t)|) \tag{184}$$

for all bounded continuous functions $h : \mathbb{R} \to \mathbb{R}$. Obviously these functions depend on U_n through all its values on $[0,1]$ and not just at finitely many values. Here is a reasonable strategy to prove (184). Take a large finite grid of points $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ in $[0,1]$. By right-continuity of $U_n(t)$, it would seem possible to choose a large enough grid so that

$$h\left(\sup_{0\le t\le 1}|U_n(t)|\right) \approx h\left(\max_{t\in F}|U_n(t)|\right) \qquad \text{where } F := \{t_0, t_1, \ldots, t_k\}. \tag{185}$$

Also because Brownian Bridge $\{U(t), 0 \le t \le 1\}$ has continuous sample paths, it seems reasonable that

$$h\left(\sup_{0\le t\le 1}|U(t)|\right) \approx h\left(\max_{t\in F}|U(t)|\right).$$

Now by (183),

$$\mathbb{E}h\left(\max_{t\in F}|U_n(t)|\right) \to \mathbb{E}h\left(\max_{t\in F}|U(t)|\right) \qquad \text{as } n \to \infty.$$

Putting the above three displayed equations together, it would seem to be possible to deduce (184). For this strategy to work, it is important that the approximation (185) holds "uniformly" in n for all large n. Indeed, if the grid F has to change considerably as n changes to maintain approximation, then this strategy cannot work. It seems clear from this discussion that move from finite-dimensional convergence of stochastic processes to infinite-dimensional convergence should be possible under an assumption which guarantees a grid approximation to the process uniformly at all large values of n. This is the so-called assumption of asymptotic equicontinuity (also known as stochastic equicontinuity) which is formulated in the abstract result stated next.

20.2 An Abstract Result

For the next result, we use the following notation. $\ell^\infty[0,1]$ denotes the class of all bounded functions on $[0,1]$ (i.e., all functions f for which $\sup_{0\le t\le 1}|f(t)| < \infty$). We shall view $\ell^\infty[0,1]$ as a metric space under the following metric:

$$(f,g) \mapsto \|f-g\|_\infty := \sup_{0\le t\le 1}|f(t)-g(t)|. \tag{186}$$

When we refer to a continuous function $h : \ell^\infty[0,1] \to \mathbb{R}$, we mean that h is continuous in the metric defined above.

Also, $C[0,1]$ denotes the class of all continuous functions on $[0,1]$.

Theorem 20.1. *Suppose for each $n \ge 1$, $\{X_n(t), t \in [0,1]\}$ is a stochastic process whose realizations are functions in $\ell^\infty[0,1]$. Suppose $\{X_t, t \in [0,1]\}$ is another stochastic process whose realizations are functions in $C[0,1]$. Assume that the following two conditions hold:*

1. *For every $k \ge 1$ and $t_1, \ldots, t_k \in [0,1]$,*

$$(X_n(t_1), \ldots, X_n(t_k)) \xrightarrow{L} (X(t_1), \ldots, X(t_k)) \qquad \text{as } n \to \infty. \tag{187}$$

   *This assumption will be referred to as **Finite Dimensional Convergence**.*

2. *For every $\epsilon > 0$ and $\delta > 0$, there exists an integer $N_{\epsilon,\delta}$ and a finite grid $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ such that*

$$\mathbb{P}\left\{\max_{0\le i\le k-1}\sup_{t\in[t_i,t_{i+1})}|X_n(t)-X_n(t_i)| > \delta\right\} < \epsilon \qquad \text{for all } n \ge N_{\epsilon,\delta}. \tag{188}$$

   *This assumption will be referred to as **Stochastic Equicontinuity** or **Asymptotic Equicontinuity**.*

Then for every bounded continuous function $h : \ell^\infty[0,1] \to \mathbb{R}$, we have

$$\mathbb{E}h(X_n) \to \mathbb{E}h(X) \qquad \text{as } n \to \infty. \tag{189}$$

Remark 20.1. *We can simplify assumption* (188) *slightly by taking $\epsilon = \delta$ i.e., we change it to: for every $\eta > 0$, there exists an integer N_η and a finite grid $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ such that*

$$\mathbb{P}\left\{\max_{0\le i\le k-1} \sup_{t\in[t_i,t_{i+1})} |X_n(t) - X_n(t_i)| > \eta\right\} < \eta \qquad \text{for all } n \ge N_\eta. \tag{190}$$

It is easy to see that (188) *and* (190) *are equivalent. Indeed,* (188) *obviously implies* (190)*. Also,* (190) *for $\eta = \min(\epsilon, \delta)$ implies* (188)*.*

The Stochastic Equicontinuity assumption(188) essentially says that $X_n(t), 0 \le t \le 1$ can be aprproximated by $X_n(t), t \in \{t_0, t_1, \ldots, t_k\}$ for all large n i.e., the approximation holds uniformly in n as long as n is large.

Proof of Theorem 20.1. We shall prove (189) for all functions $h : \ell^\infty[0,1] \to \mathbb{R}$ that are bounded and Lipschitz. It turns out that if (189) holds for all bounded Lipschitz h, then it also holds for all bounded continuous h but we shall skip the proof of this.

Let us therefore assume that h is bounded in absolute value by B and is L-Lipschitz i.e.,

$$\sup_{u\in\ell^\infty[0,1]} |h(u)| \le B \quad \text{and} \quad |h(u) - h(v)| \le L\|u - v\|_\infty = L \sup_{0\le t\le 1} |u(t) - v(t)|. \tag{191}$$

Fix $\epsilon > 0$ and invoke the stochastic equicontinuity assumption with $\epsilon > 0$ and $\delta = \epsilon$ to get an integer $N = N_\epsilon$ and a grid $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ such that (188) holds. Let $F := \{t_0, t_1, \ldots, t_k\}$ and let $A_F : \ell^\infty[0,1] \to \ell^\infty[0,1]$ defined by

$$(A_F x)(t) := \sum_{i=0}^{k-1} x(t_i) I\{t_i \le t < t_{i+1}\} \qquad \text{for } x \in \ell^\infty[0,1] \text{ and } t \in [0,1)$$

and $(A_F x)(1) = x(1)$. It is easy to check that for every $x \in \ell^\infty[0,1]$,

$$\max_i \sup_{t\in[t_i,t_{i+1})} |x(t) - x(t_i)| = \|x - A_F x\|_\infty = \sup_{0\le t\le 1} |x(t) - (A_F x)(t)|$$

Therefore

$$\mathbb{P}\{\|X_n - A_F X_n\| > \epsilon\} < \epsilon \qquad \text{for all } n \ge N_\epsilon.$$

We now change the grid F so that the above inequality also holds for the process $X(t), 0 \le t \le 1$ as well (note that X has continuous sample paths). For this, let $S = \{s_0, s_1, s_2, \ldots\}$ be a countable dense subset of $[0,1]$ with $s_0 = 0$ and $s_1 = 1$. Then, for every $x \in C[0,1]$,

$$\lim_{m\to\infty} \left\|A_{\{s_0,s_1,\ldots,s_m\}} x - x\right\|_\infty = 0.$$

Because X has continuous sample paths, we have

$$\left\|A_{\{s_0,s_1,\ldots,s_m\}} X - X\right\|_\infty \to 0 \qquad \text{almost surely as } m \to \infty.$$

Thus for all large m, we have

$$\mathbb{P}\left\{\left\|A_{\{s_0,s_1,\ldots,s_m\}} X - X\right\|_\infty > \epsilon\right\} < \epsilon.$$

Take such a large m "merge" the two grids $\{s_0, s_1, \ldots, s_m\}$ and $\{t_0, t_1, \ldots, t_k\}$. For the resulting merged grid, say T, we have

$$\mathbb{P}\{\|X_n - A_T X_n\|_\infty > 2\epsilon\} < \epsilon \quad \text{and} \quad \mathbb{P}\{\|X - A_T X\|_\infty > 2\epsilon\} < \epsilon.$$

Note that ϵ changed to 2ϵ inside the probability (this is because when $t_i \le s_j \le t < t_i + \epsilon$, we used the bound $|X_n(t) - X_n(s_j)| \le |X_n(t) - X_n(t_i)| + |X_n(t_i) - X_n(s_j)|$ and similarly for X).

Now for the function h satisfying (191), we can write

$$|\mathbb{E}h(X_n) - \mathbb{E}h(X)| \le \mathbb{E}\,|h(X_n) - h(A_T X_n)| + \mathbb{E}\,|h(X) - h(A_T X)| + |\mathbb{E}h(A_T X_n) - \mathbb{E}h(A_T X)|\,. \tag{192}$$

For the first term on the right hand side above, we argue as

$$\begin{aligned}\mathbb{E}\,|h(X_n) - h(A_T X_n)| &\le \mathbb{E}\,|h(X_n) - h(A_T X_n)|\, I\{\|X_n - A_T X_n\|_\infty \le 2\epsilon\} + \mathbb{E}\,|h(X_n) - h(A_T X_n)|\, I\{\|X_n - A_T X_n\|_\infty > 2\epsilon\} \\ &\le L(2\epsilon) + 2B\left(\mathbb{P}\left\{I\{\|X_n - A_T X_n\|_\infty > 2\epsilon\}\right) \le 2L\epsilon + 2B\epsilon.\end{aligned}$$

The same upper bound also holds for the second term in (192). For the third term in (192), use the finite-dimensional convergence assumption (note that the grid T does not depend on n) to claim that

$$|\mathbb{E}h(A_T X_n) - \mathbb{E}h(A_T X)| \to 0 \qquad \text{as } n \to \infty.$$

We have thus proved that

$$\limsup_{n\to\infty} |\mathbb{E}h(X_n) - \mathbb{E}h(X)| \le 4L\epsilon + 4B\epsilon.$$

Since $\epsilon > 0$ is arbitrary, we have proved (189). □

20.3 Back to the Uniform Empirical Process

Recall the uniform empirical process U_n in (180) and the Brownian Bridge U. Then, as we have seen, the multivariate CLT implies finite dimensional convergence. We shall argue here that the U_n satisfies stochastic equicontinuity as well. For this, let us first note that stochastic equicontinuity follows from

$$\mathbb{E} \sup_{s,t\in[0,1]:|s-t|\le\eta} |U_n(s) - U_n(t)| \to 0 \qquad \text{as } n \to \infty \text{ and } \eta \to 0. \tag{193}$$

Indeed, if (193) holds, then given $\epsilon > 0$ and $\delta > 0$, there exists $\eta > 0$ and an integer $N_{\epsilon\delta}$ such that for every $n \ge N_{\epsilon,\delta}$, we have

$$\mathbb{E} \sup_{s,t\in[0,1]:|s-t|\le\eta} |U_n(s) - U_n(t)| < \epsilon\delta \qquad \text{for all } n \ge N_{\epsilon\delta}.$$

Let $0 = t_0 < t_1 < \cdots < t_k = 1$ be a uniform grid in $[0,1]$ with spacing η. Then clearly

$$\max_i \sup_{t\in[t_i,t_{i+1})} |U_n(t) - U_n(t_i)| \le \sup_{s,t\in[0,1]:|s-t|\le\eta} |U_n(t) - U_n(s)|$$

and thus, by Markov's inequality, we have

$$\mathbb{P}\left\{\max_i \sup_{t\in[t_i,t_{i+1})} |U_n(t) - U_n(t_i)| > \delta\right\} \le \frac{1}{\delta}\mathbb{E} \sup_{s,t\in[0,1]:|s-t|<\eta} |U_n(t) - U_n(s)| \le \epsilon \qquad \text{for all } n \ge N_{\epsilon\delta}$$

which gives stochastic equicontinuity.

We shall now verify (193). This will be done via a bound for the expected suprema of empirical processes that we studied way back in Lecture 9. Let P_n denote the empirical measure of $X_1, \ldots, X_n$ and let P denote the uniform measure on $[0,1]$. Then

$$\mathbb{E} \sup_{s,t\in[0,1]:|s-t|\le\eta} |U_n(s) - U_n(t)| = \mathbb{E} \sup_{s,t\in[0,1]:|s-t|\le\eta} \left|\sqrt{n}(P_n - P)(I_{[0,s]} - I_{[0,t]})\right| = \mathbb{E} \sup_{f\in\mathcal{F}} \left|\sqrt{n}(P_n - P)f\right|$$

where

$$\mathcal{F} := \left\{ I_{[0,s]} - I_{[0,t]} : s, t \in [0,1], |s-t| \leq \eta \right\}.$$

We then use the following inequality (proved in Lecture 9)

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sqrt{n}(P_n - P)f \right| &\leq C \mathbb{E} \left[\int_0^{\sup_{f \in \mathcal{F}} \sqrt{P_n f^2}} \sqrt{1 + \log M(\epsilon, \mathcal{F}, L^2(P_n))} d\epsilon \right] \\
&\leq C \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sqrt{P_n f^2} \int_0^1 \sqrt{1 + \log M(\epsilon \sup_{f \in \mathcal{F}} \sqrt{P_n f^2}, \mathcal{F}, L^2(P_n))} d\epsilon \right].
\end{aligned}$$

The class $\mathcal{F}$ has the trivial envelope $F \equiv 1$ so we get (**There is a mistake here. We cannot deduce from $\sup_{f \in \mathcal{F}} \sqrt{P_n f^2} \leq 1$ that $M(\epsilon \sup_{f \in \mathcal{F}} \sqrt{P_n f^2}, \mathcal{F}, L^2(P_n)) \leq M(\epsilon, \mathcal{F}, L^2(P_n))$; the inequality will actually go the other way because ϵ-packing numbers increase as ϵ decreases ; see next lecture for the correct argument. I am leaving this incorrect argument here so we know that it does not work.**)

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sqrt{n}(P_n - P)f \right| &\leq C \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sqrt{P_n f^2} \int_0^1 \sqrt{1 + \log M(\epsilon \sqrt{P_n F^2}, \mathcal{F}, L^2(P_n))} d\epsilon \right] \\
&\leq C \left(\int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \sqrt{QF^2}, \mathcal{F}, L^2(Q))} \right) \mathbb{E} \sup_{f \in \mathcal{F}} \sqrt{P_n f^2}.
\end{aligned}$$

Bound the VC subgraph dimension of $\mathcal{F}$ and use the relation between packing numbers and the VC subgraph dimension to show that

$$\int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \sqrt{QF^2}, \mathcal{F}, L^2(Q))} \leq C$$

which gives

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sqrt{n}(P_n - P)f \right| &\leq C \mathbb{E} \sup_{f \in \mathcal{F}} \sqrt{P_n f^2} \\
&= C \mathbb{E} \sup_{f \in \mathcal{F}} \sqrt{P f^2 + (P_n - P) f^2} \\
&\leq C \left(\mathbb{E} \sup_{f \in \mathcal{F}} \sqrt{P f^2} + \mathbb{E} \sup_{f \in \mathcal{F}} \sqrt{|(P_n - P) f^2|} \right)
\end{aligned}$$

where we used the trivial inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. For every $f = I_{[0,s]} - I_{[0,t]} \in \mathcal{F}$, we have

$$P f^2 = P \left(I_{[0,s]} - I_{[0,t]} \right)^2 = s + t - 2 \min(s,t) = |s - t| \leq \eta$$

so we get

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sqrt{n}(P_n - P)f \right| \leq C \left(\sqrt{\eta} + \mathbb{E} \sup_{f \in \mathcal{F}} \sqrt{|(P_n - P) f^2|} \right) \leq C \left(\sqrt{\eta} + \sqrt{\mathbb{E} \sup_{f \in \mathcal{F}} |(P_n - P) f^2|} \right).$$

Argue now that $\{f^2 : f \in \mathcal{F}\}$ is a Boolean class of VC dimension at most 2 so that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sqrt{n}(P_n - P)f \right| \leq C \sqrt{\eta} + C n^{-1/4}$$

which goes to zero as $\eta \to 0$ and $n \to \infty$ thereby proving (193).

The finite dimensional convergence of U_n to U along with stochastic equicontinuity implies that (by Theorem 20.1)

$$\mathbb{E} h(U_n) \to \mathbb{E} h(U) \qquad \text{as } n \to \infty \tag{194}$$

for every bounded continuous function $h : \ell^\infty[0,1] \to \mathbb{R}$.

20.4 An Issue with Measurability

There are some measurability issues with the assertion (194). It turns out that it cannot happen that $h(U_n)$ is measurable for every bounded continuous function $h : \ell^\infty[0,1] \to \mathbb{R}$. Let us illustrate this below for the case when $n = 1$ (the argument can be extended for higher values of n as well; see, for example, Pollard [18, Problem 1, Page 86]).

Note first that the stochastic process U_1 depends on X_1 alone. In fact, the function U_1 in $\ell^\infty[0,1]$ precisely equals

$$U_1 = I_{[X_1,1]} - Id$$

where Id is the function $Id(t) = t$. We shall assume, if possible, that

$$h(U_1) = h(I_{[X_1,1]} - Id) \text{ is measurable for every } h : \ell^\infty[0,1] \to \mathbb{R} \text{ that is bounded and continuous.}$$

and arrive at a contradiction. It is easy to see that the above assertion is equivalent to

$$h(I_{[X_1,1]}) \text{ is measurable for every } h : \ell^\infty[0,1] \to \mathbb{R} \text{ that is bounded and continuous.} \tag{195}$$

By a standard connection between continuous functions and closed sets (see for example, Billingsley [2, Chapter 1]), it can be shown that (195) implies that

$$I\left\{I_{[X_1,1]} \in O\right\} \text{ is measurable for every open set } O \subseteq \ell^\infty[0,1].$$

Here open sets in $\ell^\infty[0,1]$ are defined with respect to the metric (186). It can now be verified that for every subset $A \subseteq [0,1]$, the following is true:

$$I\{X_1 \in A\} = I\left\{I_{[X_1,1]} \in O\right\} \qquad \text{where } O := \cup_{s \in A} B(I_{[s,1]}, 1/2)$$

where $B(I_{[s,1]}, 1/2)$ refers to the open ball in $\ell^\infty[0,1]$ (with respect to the metric (186)) centered at $I_{[s,1]}$ and of radius $1/2$. Because an arbitrary union of open sets is open, the set O defined above is open. We have therefore obtained, as a consequence of (195), that $I\{X_1 \in A\}$ is measurable for every subset A of $[0,1]$. Because X_1 is distributed according to the uniform distributoin on $[0,1]$, this means that it would be possible to define a probability measure on the set of all subsets of $[0,1]$ such that the probability of every interval equals the length of the interval. This cannot happen under the axiom of choice.

Due to the contradiction above, it follows that $h(U_1)$ cannot be measurable for all bounded continuous functions $h : \ell^\infty[0,1] \to \mathbb{R}$. One can also show that the same holds for $h(U_n)$ for every $n \geq 1$. This therefore means that we cannot really talk about $\mathbb{E}h(U_n)$. As a fix, one considers outer Expectations here (denoted by $\mathbb{E}^* h(U_n)$). Fortunately, the theory goes through with this fix. More details will be provided in the next lecture.

21 Lecture 21

We proved the following process convergence theorem in the last class.

Theorem 21.1. *Suppose for each $n \geq 1$, $\{X_n(t), t \in [0,1]\}$ is a stochastic process whose realizations are functions in $\ell^\infty[0,1]$. Suppose $\{X_t, t \in [0,1]\}$ is another stochastic process whose realizations are functions in $C[0,1]$. Assume that the following two conditions hold:*

1. *For every $k \geq 1$ and $t_1, \ldots, t_k \in [0,1]$,*

$$(X_n(t_1), \ldots, X_n(t_k)) \xrightarrow{L} (X(t_1), \ldots, X(t_k)) \qquad \text{as } n \to \infty. \tag{196}$$

*This assumption will be referred to as **Finite Dimensional Convergence**.*

2. *For every $\epsilon > 0$ and $\delta > 0$, there exists an integer $N_{\epsilon,\delta}$ and a finite grid $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ such that*

$$\mathbb{P}\left\{\max_{0\le i\le k-1}\sup_{t\in[t_i,t_{i+1})}|X_n(t)-X_n(t_i)|>\delta\right\}<\epsilon \qquad \text{for all } n\ge N_{\epsilon,\delta}. \tag{197}$$

*This assumption will be referred to as **Stochastic Equicontinuity** or **Asymptotic Equicontinuity**.*

Then for every bounded continuous function $h : \ell^\infty[0,1] \to \mathbb{R}$ (here we are viewing $\ell^\infty[0,1]$ as a metric space under the uniform metric), we have

$$\mathbb{E}h(X_n) \to \mathbb{E}h(X) \qquad \text{as } n \to \infty. \tag{198}$$

Here are some remarks on this theorem.

1. If the sample paths of X_n have jumps (such as when $X_n(t) = \sqrt{n}(F_n(t) - t)$), then (as mentioned in the previous class) $h(X_n)$ need not be measurable for every bounded continuous funciton $h : \ell^\infty[0,1] \to \mathbb{R}$. In this case, $\mathbb{E}h(X_n)$ may not be properly defined. This can be fixed by replacing $\mathbb{E}h(X_n)$ by its outer expectation $\mathbb{E}^*h(X_n)$ defined as

$$\mathbb{E}^*h(X_n) := \inf\left\{\mathbb{E}B : B \text{ is measurable }, B \ge h(X_n), \mathbb{E}B \text{ exists}\right\}.$$

   The result of the theorem will be true if $\mathbb{E}h(X_n)$ is replaced by $\mathbb{E}^*h(X_n)$. We shall ignore these measurability issues in our treatment. For a careful analysis, see Kato [12].

2. We take (198) to be the definition of the convergence of the sequence of stochastic processes $\{X_n\}$ to X in $\ell^\infty[0,1]$. We shall write this as $X_n \xrightarrow{L} X$ as $n \to \infty$. Like in the case of convergence in distribution on Euclidean spaces, the following are equivalent definitions of $X_n \xrightarrow{L} X$:
   (a) $\mathbb{E}^*h(X_n) \to \mathbb{E}h(X)$ for every bounded Lipschitz function $h : \ell^\infty[0,1] \to \mathbb{R}$.
   (b) For every open set G in $\ell^\infty[0,1]$, we have $\liminf_{n\to\infty}\mathbb{P}^*\{X_n \in G\} \ge \mathbb{P}\{X \in G\}$.
   (c) For every closed set F in $\ell^\infty[0,1]$, we have $\limsup_{n\to\infty}\mathbb{P}^*\{X_n \in F\} \le \mathbb{P}\{X \in F\}$.

   An important consequence of process convergence is the *continuous mapping theorem*: Suppose $X_n \xrightarrow{L} X$ and $g : \ell^\infty[0,1] \to \mathbb{R}^k$ is continuous, then $g(X_n) \xrightarrow{L} g(X)$. This is a trivial consequence of the definition of process convergence.

3. Finite dimensional convergence is usually a consequence of the Lindeberg-Feller Central Limit Theorem.

4. Stochastic equicontinuity is implied by

$$\lim_{\eta\downarrow 0}\limsup_{n\to\infty}\mathbb{P}\left\{\sup_{0\le s,t\le 1:|s-t|\le\eta}|X_n(s)-X_n(t)|>\delta\right\}=0 \qquad \text{for every } \delta>0$$

   which is further implied by

$$\lim_{\eta\downarrow 0}\limsup_{n\to\infty}\mathbb{E}\sup_{0\le s,t\le 1:|s-t|\le\eta}|X_n(s)-X_n(t)|=0. \tag{199}$$

   For example, to see that (199) implies (197), note that for every $\epsilon > 0$ and $\delta > 0$, there exists $\eta > 0$ such that

$$\limsup_{n\to\infty}\mathbb{E}\sup_{0\le s,t\le 1:|s-t|\le\eta}|X_n(s)-X_n(t)|<\epsilon\delta$$

   This further implies the existence of an integer $N_{\epsilon,\delta}$ such that for all $n \ge N_{\epsilon,\delta}$,

$$\mathbb{E}\sup_{0\le s,t\le 1:|s-t|\le\eta}|X_n(s)-X_n(t)|<\epsilon\delta.$$

Let $0 = t_0 < t_1 < \cdots < t_k = 1$ be a uniform grid in $[0,1]$ with spacing η so that

$$\max_i \sup_{t \in [t_i, t_{i+1})} |X_n(t) - X_n(t_i)| \leq \sup_{0 \leq s,t, \leq 1 : |s-t| \leq \eta} |X_n(s) - X_n(t)|$$

so that by Markov inequality, we have

$$\mathbb{P}\left\{ \max_i \sup_{t \in [t_i, t_{i+1})} |X_n(t) - X_n(t_i)| > \delta \right\} \leq \frac{1}{\delta} \mathbb{E} \sup_{0 \leq s,t \leq 1 : |s-t| \leq \eta} |X_n(s) - X_n(t)| < \epsilon \qquad \text{for all } n \geq N_{\epsilon,\delta}$$

which proves (197).

In Theorem 21.1, the interval $[0,1]$ can be replaced by any other compact subinterval $[a,b]$ of $\mathbb{R}$. In fact, it can be replaced by any abstract set T. In this case, one gets the following theorem whose proof we will skip (the proof can be found, for example, in Kato [12, Theorem 11]).

Let $\ell^\infty(T)$ denote the space of all bounded functions on T viewed as a metric space with the metric $(f_1, f_2) \mapsto \sup_{t \in T} |f_1(t) - f_2(t)|$.

Theorem 21.2. *For each $n \geq 1$, let $X_n(t), t \in T$ be a stochastic process with realizations in $\ell^\infty(T)$. Suppose*

1. *For every $k \geq 1$ and $t_1, \ldots, t_k \in T$, the random vector sequence $(X_n(t_1), \ldots, X_n(t_k))$ converges in distribution to some limit.*
2. *There exists a semi-metric d on T for which (T,d) is totally bounded and such that*

$$\lim_{\eta \downarrow 0} \limsup_{n \to \infty} \mathbb{E} \sup_{0 \leq s,t \leq 1 : d(s,t) \leq \eta} |X_n(s) - X_n(t)| = 0.$$

Then there exists a stochastic process $X(t), t \in T$ whose realizations are continuous functions on T (with respect to the metric d) such that $X_n \xrightarrow{L} X$ in $\ell^\infty(T)$ or, equivalently, $\mathbb{E}^* h(X_n) \to \mathbb{E} h(X)$ as $n \to \infty$ for every bounded continuous function $h : \ell^\infty(T) \to \mathbb{R}$.

Note that Theorem 21.2 does not start with a limit object X but it rather asserts the existence of a process $X(t), t \in T$ with continuous sample paths (with respect to the metric d with respect to which X_n satisfies stochastic equicontinuity). Note that the limit object necessarily satisfies the property that $(X_n(t_1), \ldots, X_n(t_k))$ converges in distribution to $(X(t_1), \ldots, X(t_k))$ for every $k \geq 1$ and $t_1, \ldots, t_k \in T$.

21.1 Maximal Inequalities and Stochastic Equicontinuity

As we have seen, the key condition for process convergence is stochastic equicontinuity. To prove it, we obviously need bounds on

$$\mathbb{E} \sup_{0 \leq s,t \leq 1 : d(s,t) \leq \eta} |X_n(s) - X_n(t)|.$$

In most of the applications X_n will be related to an empirical process and hence we are led to bounds on the expected suprema of empirical process.

Let us illustrate the general ideas with an example first. Suppose that the index set $T = [0,1]$ and X_n is the uniform empirical process i.e.,

$$X_n(t) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n I\{X_i \leq t\} - t \right) = \sqrt{n} \left(P_n I_{[0,t]} - P I_{[0,t]} \right)$$

where P_n is the empirical measure corresponding to the observations $X_1, \ldots, X_n$ which are i.i.d uniform on $[0,1]$ (also P is the distribution of $[0,1]$). We shall attempt to prove here that X_n satisfies stochastic equicontinuity. For this, first note that

$$\mathbb{E} \sup_{0 \le s,t \le 1 : |s-t| \le \eta} |X_n(s) - X_n(t)| = \mathbb{E} \sup_{0 \le s,t \le 1 : |s-t| \le \eta} \left| \sqrt{n}(P_n - P) I_{[0,t]} - \sqrt{n}(P_n - P) I_{[0,s]} \right| = \mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left| \sqrt{n}(P_n g - Pg) \right|$$

where we use the following notation:

$$\mathcal{F} := \left\{ I_{[0,t]} : 0 \le t \le 1 \right\} \quad \text{and} \quad \mathcal{G}_\eta := \left\{ I_{[0,t]} - I_{[0,s]} : 0 \le s, t \le 1, |s-t| \le \eta \right\} \text{ for } \eta \in [0,1].$$

We can use our earlier bounds on the expected suprema of empirical processes to control

$$\mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left| \sqrt{n}(P_n g - Pg) \right|.$$

One of our main bounds on the expected suprema of empirical processes (from Lecture 9) is

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left| \sqrt{n}(P_n h - Ph) \right| \le C \left\| H \right\|_{L^2(P)} J(H, \mathcal{H}) \tag{200}$$

where

$$J(H, \mathcal{H}) := \int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \left\| H \right\|_{L^2(Q)}, \mathcal{H}, L^2(Q))} d\epsilon.$$

where H is the envelope of $\mathcal{H}$ defined as $H(x) := \sup_{h \in \mathcal{H}} |h(x)|$. Applying this to $\mathcal{H} = \mathcal{G}_\eta$, we obtain

$$\mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left| \sqrt{n}(P_n g - Pg) \right| \le C \left\| G \right\|_{L^2(P)} \int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \left\| G \right\|_{L^2(Q)}, \mathcal{G}_\eta, L^2(Q))} d\epsilon$$

where G is the envelope of $\mathcal{G}_\eta$. It is now easy to see that G is the constant function that is equal to one. Note that $PG^2 = 1$ while $Pg^2 \le \eta$ for every $g \in \mathcal{G}_\eta$ (in other words, PG^2 is much larger than $\sup_{g \in \mathcal{G}_\eta} Pg^2$). Because $G \equiv 1$, the bound above becomes

$$\mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left| \sqrt{n}(P_n g - Pg) \right| \le C \int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon, \mathcal{G}_\eta, L^2(Q))} d\epsilon$$

We shall now show that the integral above is bounded from above by a constant. There are at least two ways of showing this. For the first way, argue that the class $\{I_{[0,s]} - I_{[0,t]} : 0 \le s, t \le 1\}$ has finite VC subgraph dimension (at most 3??) and use our earlier relations between packing numbers and VC subgraph dimensions. For the second way, the trivial inequality

$$\int \left(I_{[0,s]} - I_{[0,t]} - (f_1 - f_2) \right)^2 dQ \le 2 \int \left(I_{[0,s]} - f_1 \right)^2 dQ + 2 \int \left(I_{[0,t]} - f_2 \right)^2 dQ$$

which holds for every s, t and every pair of functions f_1 and f_2 implies that

$$N(2\delta, \mathcal{G}_\eta, L^2(Q)) \le \left(N(\delta, \mathcal{F}, L^2(Q)) \right)^2.$$

This is because we can cover functions $I_{[0,s]} - I_{[0,t]}$ to within 2δ in $L^2(Q)$ distance by covering the individual functions $I_{[0,t]}$ to within δ and then taking all pairs of functions in the cover. This gives

$$\int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon, \mathcal{G}_\eta, L^2(Q))} d\epsilon \le \int_0^1 \sqrt{1 + \log \sup_Q M(c\epsilon, \mathcal{F}, L^2(Q))} d\epsilon \le C \tag{201}$$

where c and C are positive constants (the latter inequality follows from the fact that $\mathcal{F}$ is a Boolean function class with VC dimension 1).

We have therefore proved that

$$\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\left|\sqrt{n}(P_ng-Pg)\right| \le C\int_0^1 \sqrt{1+\log\sup_Q M(\epsilon,\mathcal{G}_\eta,L^2(Q))}d\epsilon \le C.$$

Note that the second inequality above cannot be significantly improved because the integral is at least 1. Unfortunately, the bound above is not strong enough to yield

$$\lim_{\eta\downarrow 0}\limsup_{n\to\infty}\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\left|\sqrt{n}(P_ng-Pg)\right| = 0. \tag{202}$$

To improve our bounds in order to deduce the above, we need to use bounds that are better than (200). Recall (from Lecture 9) that although (200) was stated as a main bound, it actually follows from the following inequality:

$$\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\left|\sqrt{n}(P_ng-Pg)\right| \le C\mathbb{E}\int_0^{\sup_{g\in\mathcal{G}_\eta}\sqrt{P_ng^2}} \sqrt{1+\log M(\epsilon,\mathcal{G}_\eta,L^2(P_n))}d\epsilon. \tag{203}$$

From this bound, we can argue as follows. For every $\delta\in[0,1]$, we can write

$$\begin{aligned}\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\left|\sqrt{n}(P_ng-Pg)\right| &\le C\mathbb{E}\int_0^\delta \sqrt{1+\log M(\epsilon,\mathcal{G}_\eta,L^2(P_n))}d\epsilon + C\mathbb{E}\{\sup_{g\in\mathcal{G}_\eta}\sqrt{P_ng^2}>\delta\}\int_0^1\sqrt{1+\log M(\epsilon,\mathcal{G}_\eta,L^2(P_n))}d\epsilon\\ &\le C\int_0^\delta \sqrt{1+\log\sup_Q M(\epsilon,\mathcal{G}_\eta,L^2(Q))}d\epsilon + C\mathbb{P}\{\sup_{g\in\mathcal{G}_\eta}\sqrt{P_ng^2}>\delta\}\int_0^1\sqrt{1+\log\sup_Q M(\epsilon,\mathcal{G}_\eta,L^2(Q))}d\epsilon\end{aligned}$$

By (201), we can replace the second integral by a constant and the first integral by the integral of the covering numbers of $\mathcal{F}$ to get

$$\begin{aligned}\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\left|\sqrt{n}(P_ng-Pg)\right| &\le C\int_0^\delta \sqrt{1+\log\sup_Q M(c\epsilon,\mathcal{F},L^2(Q))}d\epsilon + C\mathbb{P}\{\sup_{g\in\mathcal{G}_\eta}\sqrt{P_ng^2}>\delta\}\\ &\le C\int_0^\delta \sqrt{1+\log\sup_Q M(c\epsilon,\mathcal{F},L^2(Q))}d\epsilon + \frac{C}{\delta}\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\sqrt{P_ng^2}.\end{aligned} \tag{204}$$

The last expected supremum can be controlled (as in the last lecture) via

$$\begin{aligned}\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\sqrt{P_ng^2} &= \mathbb{E}\sup_{g\in\mathcal{G}_\eta}\sqrt{Pg^2+(P_n-P)g^2}\\ &\le \left(\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\sqrt{Pg^2}+\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\sqrt{|(P_n-P)g^2|}\right)\end{aligned}$$

where we used the trivial inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$. As we mentioned earlier, $\sup_{g\in\mathcal{G}_\eta}Pg^2\le\eta$ so that

$$\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\sqrt{P_ng^2} \le \sqrt{\eta}+\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\sqrt{|(P_n-P)g^2|} \le \sqrt{\eta}+\sqrt{\mathbb{E}\sup_{g\in\mathcal{G}_\eta}|(P_n-P)g^2|}.$$

Note now that $\{g^2 : g\in\mathcal{G}_\eta\}$ is a Boolean class of VC dimension at most 2 so that

$$\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\sqrt{P_ng^2}\le\sqrt{\eta}+Cn^{-1/4}.$$

Combining this with (204), we obtain

$$\mathbb{E}\sup_{g\in\mathcal{G}_\eta}\left|\sqrt{n}(P_ng-Pg)\right| \le C\int_0^\delta\sqrt{1+\log\sup_Q M(c\epsilon,\mathcal{F},L^2(Q))}d\epsilon + \frac{C}{\delta}\left(\sqrt{\eta}+Cn^{-1/4}\right).$$

As a consequence,

$$\limsup_{n\to\infty} \mathbb{E} \sup_{g\in\mathcal{G}_\eta} \left|\sqrt{n}(P_n g - Pg)\right| \leq C \int_0^\delta \sqrt{1 + \log \sup_Q M(c\epsilon, \mathcal{F}, L^2(Q))} d\epsilon + \frac{C\sqrt{\eta}}{\delta}$$

and further

$$\lim_{\eta\downarrow 0} \limsup_{n\to\infty} \mathbb{E} \sup_{g\in\mathcal{G}_\eta} \left|\sqrt{n}(P_n g - Pg)\right| \leq C \int_0^\delta \sqrt{1 + \log \sup_Q M(c\epsilon, \mathcal{F}, L^2(Q))} d\epsilon.$$

Since this is true for every $\delta > 0$, the inequality will also hold if we take limit of the right hand side as $\delta \to 0$. It is now easy to show that this limit would be zero (this is because the integral from 0 to 1 is finite so the integral from 0 to δ should go to zero as $\delta \to 0$ by the dominated convergence theorem). We have thus proved (202) which is same as stochastic equicontinuity of $X_n(t), t \in [0,1]$. Combined with finite dimensional convergence, we have proved that the uniform empirical process converges in distribution to Brownian Bridge. This is Donsker's theorem for the uniform empirical process.

We had to do a bit of work above to go from (203) to (202). There exist other maximal inequalities which allow one to deduce of (202) more easily. The following theorem (taken from Kato [12, Theorem 8]) is one such result.

Theorem 21.3. *Let H be an envelope for the class $\mathcal{H}$ with $PH^2 < \infty$. Then*

$$\mathbb{E} \sup_{h\in\mathcal{H}} |\sqrt{n}(P_n h - Ph)| \leq C \left(\|H\|_{L^2(P)} J(\delta) + \frac{\Gamma}{\sqrt{n}} \frac{J^2(\delta)}{\delta^2} \right). \tag{205}$$

for every δ satisfying

$$\frac{\sup_{h\in\mathcal{H}} Ph^2}{PH^2} \leq \delta^2 \leq 1.$$

Here

$$J(\delta) := \int_0^\delta \sqrt{1 + \log \sup_Q M(\epsilon \|H\|_{L^2(Q)}, \mathcal{H}, L^2(Q))} d\epsilon \quad \text{and} \quad \Gamma := \sqrt{\mathbb{E} \max_{1\leq i\leq n} H^2(X_i)}.$$

Let us now demonstrate that Theorem 21.3 yields (202) quite easily. Indeed, applying inequality (205) to $\mathcal{H} = \mathcal{G}_\eta$ (with envelope $H \equiv 1$) and $\delta = \sqrt{\eta}$ (note that $\sup_{g\in\mathcal{G}_\eta} Pg^2 \leq \eta$ and $PG^2 = 1$), we get

$$\mathbb{E} \sup_{g\in\mathcal{G}_\eta} |\sqrt{n}(P_n g - Pg)| \leq C \left(J(\sqrt{\eta}) + \frac{1}{\sqrt{n}} \frac{J^2(\sqrt{\eta})}{\eta} \right)$$

which implies

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E} \sup_{g\in\mathcal{G}_\eta} |\sqrt{n}(P_n g - Pg)| &\leq C J(\sqrt{\eta}) \\ &= C \int_0^{\sqrt{\eta}} \sqrt{1 + \log \sup_Q M(\epsilon, \mathcal{G}_\eta, L^2(Q))} d\epsilon \leq C \int_0^{\sqrt{\eta}} \sqrt{1 + \log \sup_Q M(c\epsilon, \mathcal{F}, L^2(Q))} d\epsilon\end{aligned}$$

which, as before, goes to 0 as $\eta \downarrow 0$. This gives a shorter proof of (202) (albeit reliant on the nontrivial result from Theorem 21.3).

22 Lecture 22

In the last lecture, we proved that the uniform empirical process converges in distribution to Brownian Bridge in $\ell^\infty[0,1]$. The main ingredient for this is proving the stochastic equicontinuity condition for the uniform empirical process. This condition states that

$$\lim_{\eta\downarrow 0} \limsup_{n\to\infty} \mathbb{E} \sup_{s,t\in[0,1]:|s-t|\leq\eta} |X_n(t) - X_n(s)| = 0$$

where $X_n(t)$ is the uniform empirical process. We observed last time that this statement is equivalent to

$$\lim_{\eta \downarrow 0} \limsup_{n \to \infty} \mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left| \sqrt{n} \left(P_n g - P g \right) \right| = 0 \tag{206}$$

where

$$\mathcal{G}_\eta := \left\{ I_{[0,s]} - I_{[0,t]} : s, t \in [0,1], |s-t| \leq \eta \right\}.$$

In this lecture, we shall first generalize this argument to more general processes $\{\sqrt{n}(P_n f - Pf) : f \in \mathcal{F}\}$. Specifically, let $\mathcal{F}$ be a class of functions and let

$$\mathcal{G}_\eta := \left\{ f_1 - f_2 : f_1, f_2 \in \mathcal{F} \text{ and } P(f_1 - f_2)^2 \leq \eta \right\}.$$

We shall provide a sufficient condition on $\mathcal{F}$ for which (206) holds. To control the expected supremum in (206), we shall use the following maximal inequality (which was also stated in the previous lecture):

Theorem 22.1. *Let H be an envelope for the class $\mathcal{H}$ with $PH^2 < \infty$. Then*

$$\mathbb{E} \sup_{h \in \mathcal{H}} |\sqrt{n}(P_n h - Ph)| \leq C \left(\|H\|_{L^2(P)} J(\delta) + \frac{\Gamma}{\sqrt{n}} \frac{J^2(\delta)}{\delta^2} \right). \tag{207}$$

for every δ satisfying

$$\frac{\sup_{h \in \mathcal{H}} Ph^2}{PH^2} \leq \delta^2 \leq 1.$$

Here

$$J(\delta) := \int_0^\delta \sqrt{1 + \log \sup_Q M(\epsilon \|H\|_{L^2(Q)}, \mathcal{H}, L^2(Q))} d\epsilon \quad \text{and} \quad \Gamma := \sqrt{\mathbb{E} \max_{1 \leq i \leq n} H^2(X_i)}.$$

We shall apply Theorem 22.1 to $\mathcal{H} = \mathcal{G}_\eta$. Suppose F is an envelope for $\mathcal{F}$, then it is clear that $2F$ is an envelope for $\mathcal{G}_\eta$. We shall therefore take $H = 2F$ while applying Theorem 22.1. We get

$$\mathbb{E} \sup_{g \in \mathcal{G}_\eta} |\sqrt{n}(P_n g - Pg)| \leq C \left(2 \|F\|_{L^2(P)} J(\delta, 2F, \mathcal{G}_\eta) + \sqrt{\frac{\mathbb{E} \max_{1 \leq i \leq n} (4F^2(X_i))}{n}} \frac{J^2(\delta, 2F, \mathcal{G}_\eta)}{\delta^2} \right) \tag{208}$$

where

$$J(\delta, 2F, \mathcal{G}_\eta) := \int_0^\delta \sqrt{1 + \log \sup_Q M(\epsilon \|2F\|_{L^2(Q)}, \mathcal{G}_\eta, L^2(Q))} d\epsilon.$$

Here, as in Theorem 22.1, δ is any real number satisfying

$$\sup_{g \in \mathcal{G}_\eta} \frac{Pg^2}{P(2F)^2} \leq \delta^2 \leq 1.$$

Now because $\sup_{g \in \mathcal{G}_\eta} Pg^2 \leq \eta$, we can take

$$\delta^2 = \min \left(\frac{\eta}{4PF^2}, 1 \right)$$

so that $\delta \downarrow 0$ as $\eta \downarrow 0$. Because $\mathcal{G}_\eta$ is a subset of $\mathcal{F} - \mathcal{F}$ (which is the class of all functions $\{f_1 - f_2 : f_1, f_2 \in \mathcal{F}\}$), we can trivially bound the packing numbers of $\mathcal{G}_\eta$ by the square of the packing numbers of $\mathcal{F}$. More precisely,

$$M(\epsilon \|2F\|_{L^2(Q)}, \mathcal{G}_\eta, L^2(Q)) \leq \left(M(c\epsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q)) \right)^2$$

for a positive constant c. This gives

$$J(\delta, 2F, \mathcal{G}_\eta) := \int_0^\delta \sqrt{1 + \log \sup_Q M(\epsilon \|2F\|_{L^2(Q)}, \mathcal{G}_\eta, L^2(Q))} d\epsilon \leq C J(c\delta, F, \mathcal{F})$$

for two positive constants c and C. Plugging this in (208), we obtain

$$\mathbb{E}\sup_{g\in\mathcal{G}_\eta}|\sqrt{n}(P_ng-Pg)|\leq C\left(\|F\|_{L^2(P)}J(\delta,F,\mathcal{F})+\sqrt{\frac{\mathbb{E}\max_{1\leq i\leq n}F^2(X_i)}{n}}\frac{J^2(c\delta,F,\mathcal{F})}{\delta^2}\right).$$

As a result, we obtain

$$\limsup_{n\to\infty}\mathbb{E}\sup_{g\in\mathcal{G}_\eta}|\sqrt{n}(P_ng-Pg)|\leq C\left(\|F\|_{L^2(P)}J(\delta,F,\mathcal{F})+\frac{J^2(c\delta,F,\mathcal{F})}{\delta^2}\limsup_{n\to\infty}\sqrt{\frac{\mathbb{E}\max_{1\leq i\leq n}F^2(X_i)}{n}}\right).$$

Lemma 22.2 below then implies that the lim sup term on the right hand side above equals zero (as long as $J(c\delta,F,\mathcal{F})<\infty$) so that

$$\limsup_{n\to\infty}\mathbb{E}\sup_{g\in\mathcal{G}_\eta}|\sqrt{n}(P_ng-Pg)|\leq C\|F\|_{L^2(P)}J(\delta,F,\mathcal{F})$$

If we now assume that $\lim_{\delta\downarrow 0}J(\delta,F,\mathcal{F})=0$, then we establish (206). Note that $\lim_{\delta\downarrow 0}J(\delta,F,\mathcal{F})=0$ is a consequence of

$$J(1,F,\mathcal{F})=\int_0^1\sqrt{1+\log\sup_Q M(\epsilon\|F\|_{L^2(Q)},\mathcal{F},L^2(Q))}d\epsilon<\infty. \tag{209}$$

It remains to state and prove Lemma 22.2.

Lemma 22.2. *Suppose $Y_1,\ldots,Y_n$ are identically distributed random variables (no assumption of independence here) with $\mathbb{E}|Y_1|<\infty$. Then*

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\max_{1\leq i\leq n}|Y_i|=0. \tag{210}$$

Proof of Lemma 22.2. This is a consequence of the Dominated Convergence theorem. We write

$$\frac{1}{n}\mathbb{E}\max_{1\leq i\leq n}|Y_i|=\int_0^\infty\frac{1}{n}\mathbb{P}\left\{\max_{1\leq i\leq n}|Y_i|>x\right\}dx.$$

For each fixed x, it is clear that the integrand above converges to zero as $n\to\infty$. Further the integrand is bounded by (by the union bound and the identical distribution assumption) $\mathbb{P}\{|Y_1|>x\}$ which integrates to $\mathbb{E}|Y_1|<\infty$. The statement (210) therefore follows by the Dominated Convergence theorem. □

22.1 Donsker's Theorem under the Uniform Entropy Condition

We have proved therefore that under the condition (209) (known as the **uniform entropy condition**), the stochastic process $X_n(f):=\sqrt{n}(P_nf-Pf)$ satisfies the stochastic equicontinuity condition with respect to the metric $(f,g)\mapsto\|f-g\|_{L^2(P)}$ on $\mathcal{F}$. On the other hand, we also have finite dimensional convergence here via the usual multivariate central limit theorem i.e.,

$$(X_n(f_1),\ldots,X_n(f_k))\xrightarrow{L}N(0,\Sigma)$$

where $\Sigma(i,j)=\mathrm{Cov}(f_i(X_1),f_j(X_1))$ for every $k\geq 1$ and $f_1,\ldots,f_k\in\mathcal{F}$. We can thus apply our process convergence result from last class which gives that $X_n(f),f\in\mathcal{F}$ converges in distribution on $\ell^\infty(\mathcal{F})$. The theorem also guarantees that the limit process $X(f),f\in\mathcal{F}$ is a Gaussian process (i.e., $(X(f_1),\ldots,X(f_k))$ has a multivariate Gaussian distirbution for every $k\geq 1$ and $f_1,\ldots,f_k\in\mathcal{F}$) has continuous sample paths with respect to the metric $(f,g)\mapsto\|f-g\|_{L^2(P)}$. These conclusions are restated in the following theorem.

Theorem 22.3. *Assume the uniform entropy condition* (209). *Then there exists a Gaussian process $X(f),f\in\mathcal{F}$ with continuous sample paths (with respect to the metric $\|f-g\|_{L^2(P)}$) such that the sequence of stochastic processes $\{X_n\}$ defined by $X_n(f):=\sqrt{n}(P_nf-Pf),f\in\mathcal{F}$ converges in distribution to X in $\ell^\infty(\mathcal{F})$.*

Definition 22.4 (Donsker Class of Functions)**.** *Say that a class of functions $\mathcal{F}$ is Donsker with respect to a probability measure P (also written as P-Donsker) if the stochastic process $X_n(f) := \sqrt{n}(P_n f - Pf), f \in \mathcal{F}$ converges in distribution to a Gaussian process $X(f), f \in \mathcal{F}$ which has continuous sample paths with respect to the metric $(f, g) \mapsto \|f - g\|_{L^2(P)}$.*

Theorem (22.3) states therefore that if $\mathcal{F}$ satisfies the uniform entropy condition (209), then $\mathcal{F}$ is P-Donsker for every probability measure P.

22.2 Bracketing Condition for Donsker Classes

Another sufficient condition for being P-Donsker is obtained by replacing the uniform entropies in (209) by bracketing entropy numbers. Specifically, assume that

$$J_{[]}(1, F, \mathcal{F}) = \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon \|F\|_{L^2(P)}, \mathcal{F}, L^2(P))} d\epsilon < \infty. \tag{211}$$

Note that this condition depends on the probability measure P (unlike (209)). The following theorem proves that, under the above condition, $\mathcal{F}$ is P-Donsker.

Theorem 22.5. *If $\mathcal{F}$ satisfies the bracketing condition* (211) *for the probability measure P, then $\mathcal{F}$ is P-Donsker.*

To prove this theorem, it is enough to show that the process $X_n(f) = \sqrt{n}(P_n f - Pf)$ satisfies the stochastic equicontinuity condition under (211). For this, we shall use the following maximal inequality from Van der Vaart [24, Lemma 19.34].

Theorem 22.6. *Suppose H is an envelope for a class of functions $\mathcal{H}$ and assume that $\|H\|_{L^2(P)} < \infty$. Then for every $\delta > 0$ satisfying*

$$\sup_{h \in \mathcal{H}} \frac{Ph^2}{PH^2} \leq \delta^2 \leq 1,$$

the following inequality holds:

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left(\sqrt{n}(P_n h - Ph) \right) \leq C \left(\|H\|_{L^2(P)} J_{[]}(\delta, H, \mathcal{H}) + \sqrt{n} \mathbb{E} \left[H(X_1) I\{H(X_1) > \sqrt{n} a(\delta)\} \right] \right)$$

where

$$J_{[]}(\delta, H, \mathcal{H}) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon \|H\|_{L^2(P)}, \mathcal{H}, L^2(P))} d\epsilon$$

and

$$a(\delta) = \frac{\delta \|H\|_{L^2(P)}}{\sqrt{\log N_{[]}(\delta \|H\|_{L^2(P)}, \mathcal{H}, L^2(P))}}.$$

Theorem 22.6 is an analogue of Theorem 21.3 for entropy with bracketing. Also, Theorem 22.6 can be seen an improvement of our earlier bracketing based maximal inequality:

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left(\sqrt{n}(P_n h - Ph) \right) \leq C \|H\|_{L^2(P)} J_{[]}(1, H, \mathcal{H}).$$

When δ is small, the bound given by Theorem 22.6 is much better than the above bound.

We shall now prove Theorem 22.5 using Theorem 22.6 (for proof of Theorem 22.6, refer to Van der Vaart [24, Lemma 19.34]).

Proof of Theorem 22.5. To prove $\mathcal{F}$ is P-Donsker, the key is to verify stochastic equicontinuity (finite dimensional convergence follows from the usual Central Limit Theorem). For stochastic equicontinuity, we need to prove that

$$\lim_{\eta \downarrow 0} \limsup_{n \to \infty} \mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left| \sqrt{n}(P_n g - Pg) \right| = 0$$

where

$$\mathcal{G}_\eta := \left\{ f_1 - f_2 : f_1, f_2 \in \mathcal{F}, P(f_1 - f_2)^2 \leq \eta \right\}.$$

For this, we use Theorem 22.6 with $\mathcal{H} := \mathcal{G}_\eta$ and $H = 2F$ (where F is the envelope of $\mathcal{F}$) to obtain

$$\mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left(\sqrt{n}(P_n g - Pg) \right) \leq C \left(\|2F\|_{L^2(P)} J_{[]}(\delta, 2F, \mathcal{G}_\eta) + \sqrt{n} \mathbb{E} \left[(2F(X_1)) I\{2F(X_1) > \sqrt{n} a(\delta)\} \right] \right)$$

with

$$J_{[]}(\delta, 2F, \mathcal{G}_\eta) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon \|2F\|_{L^2(P)}, \mathcal{G}_\eta, L^2(P))} d\epsilon.$$

Because the bracketing numbers of $\{f - g : f, g \in \mathcal{F}\}$ can be bounded by the squares of the bracketing numbers of $\mathcal{F}$, we get

$$J_{[]}(\delta, 2F, \mathcal{G}_\eta) \leq C \int_0^{c\delta} \sqrt{1 + \log N_{[]}(\epsilon \|F\|_{L^2(P)}, \mathcal{F}, L^2(P))} = C J_{[]}(c\delta, \mathcal{F}, L^2(P)) d\epsilon$$

for two positive constants c and C. We obtain thus

$$\mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left(\sqrt{n}(P_n g - Pg) \right) \leq C \left(\|F\|_{L^2(P)} J_{[]}(c\delta, F, \mathcal{F}) + \sqrt{n} \mathbb{E} \left[(2F(X_1)) I\{2F(X_1) > \sqrt{n} a(\delta)\} \right] \right)$$

Also

$$a(\delta) = \frac{\delta \|2F\|_{L^2(P)}}{\sqrt{\log N_{[]}(\delta \|2F\|_{L^2(P)}, \mathcal{G}_\eta, L^2(P))}} \geq \frac{\delta \|2F\|_{L^2(P)}}{\sqrt{2 \log N_{[]}(c\delta \|2F\|_{L^2(P)}, \mathcal{F}, L^2(P))}} =: a'(\delta).$$

We thus have

$$\mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left(\sqrt{n}(P_n g - Pg) \right) \leq C \left(\|F\|_{L^2(P)} J_{[]}(c\delta, F, \mathcal{F}) + \sqrt{n} \mathbb{E} \left[(2F(X_1)) I\{2F(X_1) > \sqrt{n} a'(\delta)\} \right] \right).$$

Note that $a'(\delta)$ does not depend on n. The second term above can be bounded as

$$\begin{aligned} \sqrt{n} \mathbb{E} \left[(2F(X_1)) I\{2F(X_1) > \sqrt{n} a'(\delta)\} \right] &\leq \sqrt{n} \mathbb{E} \left[(2F(X_1)) \frac{2F(X_1)}{\sqrt{n} a'(\delta)} I\{2F > \sqrt{n} a'(\delta)\} \right] \\ &= \frac{4}{a'(\delta)} \mathbb{E} F^2(X_1) I\{2F(X_1) > \sqrt{n} a'(\delta)\}. \end{aligned}$$

By the dominated convergence theorem, the above expectation converges to zero as $n \to \infty$ (note that we have assumed that $\mathbb{E}F^2(X_1) < \infty$). We have thus proved

$$\mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left(\sqrt{n}(P_n g - Pg) \right) \leq C \|F\|_{L^2(P)} J_{[]}(c\delta, \mathcal{F}, L^2(P))$$

for every $\delta > 0$. Under the assumption (211), the right hand side above converges to zero as $\delta \to 0$. This proves stochastic equicontinuity and consequently the fact that $\mathcal{F}$ is P-Donsker. □

22.3 Application to convergence rate of the sample median

We gave the example of finding the convergence rate of the sample median as one of the motivations for studying process convergence. Now that we understand what process convergence is, let us revisit this

example and rigorize the argument. The setting is as follows. We have i.i.d data $X_1, \ldots, X_n$ generated from, say, the $N(\theta_0, 1)$ distribution. Let

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} |X_i - \theta| \quad \text{and} \quad M(\theta) := \mathbb{E}|X_1 - \theta|$$

for $\theta \in \mathbb{R}$. The estimator $\hat{\theta}_n$ is defined as any minimizer of $M_n(\theta)$ over $\theta \in \mathbb{R}$. Also note that θ_0 uniquely maximizes $M(\theta), \theta \in \mathbb{R}$. We have proved earlier that $\hat{\theta}_n$ is consistent for θ_0 and that its rate of convergence is $n^{-1/2}$. To obtain the limiting distribution of $\hat{\theta}_n$, we considered the process:

$$\tilde{M}_n(h) := n\left(M_n(\theta_0 + n^{-1/2}h) - M_n(\theta_0)\right) = A_n(h) + B_n(h)$$

where

$$A_n(h) = n\left(M_n(\theta_0 + n^{-1/2}h) - M_n(\theta_0) - M(\theta_0 + n^{-1/2}h) + M(\theta_0)\right)$$

and

$$B_n(h) = n\left(M(\theta_0 + n^{-1/2}h) - M(\theta_0)\right).$$

We have earlier seen that $A_n(h) \xrightarrow{L} A(h) := hZ$ where $Z \sim N(0,1)$ and $B_n(h) \to B(h) := M''(\theta_0)h^2/2$ for every fixed $h \in \mathbb{R}$. The first convergence was an application of the Lindeberg-Feller CLT and the second convergence followed from Taylor expansion to second order. These convergence statements actually can be much strengthened. In fact, it holds that

$$A_n \xrightarrow{L} A \quad \text{in } \ell^\infty[-\Gamma, \Gamma] \text{ for every fixed } \Gamma > 0 \tag{212}$$

and

$$B_n \to B \quad \text{uniformly over } [-\Gamma, \Gamma] \text{ for every } \Gamma > 0. \tag{213}$$

The uniform convergence above means that $\sup_{|h| \le \Gamma} |B_n(h) - B(h)| \to 0$ as $n \to \infty$. The statement (213) is straightforward to prove via the usual Taylor expansion argument (left as exercise). We shall sketch the argument for (212) below. The process convergence statement (212) requires two ingredients: finite dimensional convergence and stochastic equicontinuity. For finite dimensional convergence, we need to prove that

$$(A_n(h_1), \ldots, A_n(h_k)) \xrightarrow{L} (A(h_1), \ldots, A(h_k))$$

for every $k \ge 1$ and $h_1, \ldots, h_k \in [-\Gamma, \Gamma]$. This can be proved via the multivariate Lindeberg-Feller Central Limit Theorme (left as exercise). For stochastic equicontinuity, we need to show that

$$\lim_{\eta \downarrow 0} \limsup_{n \to \infty} \mathbb{E} \sup_{h_1, h_2 \in [\Gamma, \Gamma] : |h_1 - h_2| \le \eta} |A_n(h_1) - A_n(h_2)| = 0. \tag{214}$$

For this, note that

$$\mathbb{E} \sup_{h_1, h_2 \in [\Gamma, \Gamma] : |h_1 - h_2| \le \eta} |A_n(h_1) - A_n(h_2)| = \mathbb{E} \sup_{g \in \mathcal{G}_\eta} \left(n\,|P_n g - Pg|\right)$$

where

$$\mathcal{G}_\eta := \left\{x \mapsto |x - \theta_0 - n^{-1/2}h_1| - |x - \theta_0 - n^{-1/2}h_2| : h_1, h_2 \in [-\Gamma, \Gamma], |h_1 - h_2| \le \eta\right\}.$$

The statement (214) can then be proved by using one of our bounds on the expected suprema of empirical process (say the bounds based on bracketing numbers). This is again left as exercise.

The two statements (212) and (213) can be added to yield (verify this):

$$\tilde{M}_n = A_n + B_n \xrightarrow{L} A + B =: \tilde{M} \quad \text{in } \ell^\infty[-\Gamma, \Gamma] \text{ for every fixed } \Gamma > 0.$$

The limit process of $\tilde{M}_n$ is therefore $\tilde{M}(h) := hZ + M''(\theta_0)h^2/2$ for $h \in \mathbb{R}$. The limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ now follows if we can prove that $\operatorname{argmin}_{h \in \mathbb{R}} \tilde{M}_n(h)$ converges in distribution to $\operatorname{argmin}_{h \in \mathbb{R}} \tilde{M}(h)$. This can be deduced from a general *argmax Continuous Mapping Theorem* which is proved next.

22.4 The Argmax Continuous Mapping Theorem

Theorem 22.7. *Let H be a metric space. Let $\{M_n(h), h \in H\}$ and $\{M(h), h \in H\}$ be stochastic processes indexed by H. Suppose the following conditions holds:*

1. *$M \xrightarrow{L} M$ in $\ell^\infty(K)$ for every compact subset K of H.*
2. *Every realization of M is continuous on H.*
3. *Let $\hat{h}_n$ maximize $M_n(h)$ over $h \in H$.*
4. *Let $\hat{h}$ be the **unique** maximizer of $M(h)$ over $h \in H$.*
5. ***Tightness**: For each $\epsilon > 0$, there exists a compact subset $K_\epsilon \subseteq H$ such that*

$$\mathbb{P}\left\{\hat{h} \notin K_\epsilon\right\} < \epsilon \quad \text{and} \quad \limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \notin K_\epsilon\right\} < \epsilon.$$

Then $\hat{h}_n \xrightarrow{L} \hat{h}$ in H i.e., for every bounded continuous function $f : H \to \mathbb{R}$, we have $\mathbb{E}f(\hat{h}_n) \to \mathbb{E}f(\hat{h})$ as $n \to \infty$.

Remark 22.1. *Usually Theorem 22.7 will applied to the process $\tilde{M}_n(h) := r_n^2\left(M_n(\theta_0 + hr_n^{-1}) - M_n(\theta_0)\right)$ and $\tilde{M}$ as the limit process of $\tilde{M}_n$. In this case, note that $\hat{h}_n = r_n\left(\hat{\theta}_n - \theta_0\right)$ and hence the tightness condition is equivalent to $\hat{\theta}_n - \theta_0 = O_P(r_n^{-1})$. Thus a preliminary rate result needs to be proved before applying Theorem 22.7 for obtaining the asymptotic distribution of $r_n(\hat{\theta}_n - \theta_0)$.*

Remark 22.2. *One can apply Theorem 22.7 to $M_n(\theta), \theta \in \Theta$ and $M(\theta), \theta \in \Theta$ as well (instead of $\tilde{M}_n$ and $\tilde{M}$). This will usually lead to a consistency result for $\hat{\theta}_n$.*

Proof of Theorem 22.7. It is enough to show that

$$\limsup_{n\to\infty} \mathbb{P}\{\hat{h}_n \in F\} \leq \mathbb{P}\{\hat{h} \in F\}$$

for every closed subset F of H. Fix a closed subset $F \subseteq H$ and also fix an arbitrary compact set K in H. Write

$$\mathbb{P}\left\{\hat{h}_n \in F\right\} \leq \mathbb{P}\left\{\hat{h}_n \in F \cap K\right\} + \mathbb{P}\left\{\hat{h}_n \notin K\right\} \leq \mathbb{P}\left\{\sup_{h\in F\cap K} M_n(h) - \sup_{h\in K} M_n(h) \geq 0\right\} + \mathbb{P}\left\{\hat{h}_n \notin K\right\}$$

which gives

$$\limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \in F\right\} \leq \limsup_{n\to\infty} \mathbb{P}\left\{\sup_{h\in F\cap K} M_n(h) - \sup_{h\in K} M_n(h) \geq 0\right\} + \limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \notin K\right\}$$

Note now that

$$\left\{m \in \ell^\infty(K) : \sup_{h\in F\cap K} m(h) - \sup_{h\in K} m(h) \geq 0\right\}$$

is a closed subset of $\ell^\infty(K)$. This follows because if $\sup_{h\in F\cap K} m_k(h) - \sup_{h\in K} m_k(h) \geq 0$ for each k and $m_k \to m$ uniformly in K, then $\sup_{h\in F\cap K} m(h) - \sup_{h\in K} m(h) \geq 0$. Thus from the convergence of M_n to M in $\ell^\infty(K)$, we get

$$\limsup_{n\to\infty} \mathbb{P}\left\{\sup_{h\in F\cap K} M_n(h) - \sup_{h\in K} M_n(h) \geq 0\right\} \leq \mathbb{P}\left\{\sup_{h\in F\cap K} M(h) - \sup_{h\in K} M(h) \geq 0\right\}.$$

We thus get

$$\limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \in F\right\} \leq \mathbb{P}\left\{\sup_{h\in F\cap K} M(h) - \sup_{h\in K} M(h) \geq 0\right\} + \limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \notin K\right\}$$
$$\leq \mathbb{P}\left\{\sup_{h\in F\cap K} M(h) - \sup_{h\in K} M(h) \geq 0, \hat{h} \in K\right\} + \mathbb{P}\{\hat{h} \notin K\} + \limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \notin K\right\}.$$

We now claim that

$$\left\{\sup_{h\in F\cap K} M(h) - \sup_{h\in K} M(h) \geq 0, \hat{h} \in K\right\} \subseteq \left\{\hat{h} \in F\right\}.$$

The reason for this is that when the right hand side holds, we have $\sup_{h\in F\cap K} M(h) \geq \sup_{h\in K} M(h) \geq \sup_{h\in H} M(h)$. The continuity of the sample paths of M and the closedness of F (which implies that $F\cap K$ is compact) implies that $\sup_{h\in F\cap K} M(h)$ is achieved at some point in $F\cap K$. The unique maximum assumption on M will imply that the point in $F\cap K$ achieving the maximum of M will have to equal $\hat{h}$ which implies that $\hat{h} \in F$. This therefore gives

$$\limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \in F\right\} \leq \mathbb{P}\left\{\hat{h} \in F\right\} + \mathbb{P}\{\hat{h} \notin K\} + \limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \notin K\right\}.$$

Note that this is true for every closed subset F of H and every compact subset K of H. Now fix $\epsilon > 0$ and choose K_ϵ as in the tightness condition. This will give

$$\limsup_{n\to\infty} \mathbb{P}\left\{\hat{h}_n \in F\right\} \leq \mathbb{P}\left\{\hat{h} \in F\right\} + 2\epsilon.$$

Let ϵ tend to zero to complete the proof. □

Use this result along with the process convergence results of the previous section to complete the proof of the result for the limiting distribution of the sample median.

23 Lecture 23

The main goal of this lecture is to prove the following theorem (from Van der Vaart [24, Theorem 5.23]) which proves asymptotic normality of M-estimators under some general conditions. To simplify the proof slightly, I have made some simplifications to the theorem (such as assuming that the criterion functions are indexed by $\mathbb{R}$; the full theorem in Van der Vaart [24, Theorem 5.23] applies to the case where the criterion functions are indexed by an open set in $\mathbb{R}^k$ for a fixed k).

23.1 An abstract M-estimation result

Theorem 23.1. *Suppose $\{m_\theta, \theta \in \mathbb{R}\}$ be a class of functions indexed by $\mathbb{R}$. Given i.i.d observations $X_1, \ldots, X_n$ having distribution P, we consider the estimator $\hat{\theta}_n$ defined as any maximizer of $P_n m_\theta$ over $\theta \in \mathbb{R}$. Let θ_0 be the population analogue of $\hat{\theta}_n$ defined as any maximizer of Pm_θ over $\theta \in \mathbb{R}$. Suppose that the following assumptions hold:*

1. *Assume that $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 with derivative $\dot{m}_{\theta_0}(x)$ for almost sure x (w.r.t P).*
2. *Assume that there exists a function $\Gamma(x)$ with $P\Gamma^2 < \infty$ (i.e., $\Gamma \in L^2(P)$) such that*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \Gamma(x)|\theta_1 - \theta_2| \tag{215}$$

for all θ_1, θ_2 and x.

3. *Suppose that $\theta \mapsto M(\theta) := Pm_\theta$ is twice continuously differentiable at θ_0 with $M''(\theta_0) < 0$.*

4. *$\hat{\theta}_n$ is consistent for θ_0 i.e., $\hat{\theta}_n \xrightarrow{P} \theta_0$ as $n \to \infty$.*

Then the following two conclusions holds:

1. *The rate of convergence of $\hat{\theta}_n$ to θ_0 is $n^{-1/2}$ i.e., $|\hat{\theta}_n - \theta_0| = O_P(n^{-1/2})$.*

2. *The following holds:*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N\left(0, \frac{\text{var}(\dot{m}_{\theta_0}(X_1))}{(M''(\theta_0))^2}\right) \qquad \text{as } n \to \infty. \tag{216}$$

Before proceeding to the proof of this Theorem, let us first look at the following remarks.

1. The conditions of the theorem hold when $m_\theta(x) = -|x - \theta|$ and thus this theorem can be viewed as a generalization of our limiting distribution result for the sample median from last class.

2. Note that the criterion function $\theta \mapsto m_\theta(x)$ is only assumed to be once differentiable with respect to θ at θ_0 (almost surely with respect to x). But the limit function $M(\theta) = Pm_\theta$ is assumed to be twice differentiable. If we insist on the criterion function to be twice differentiable, then the theorem will no longer be applicable to functions such as $m_\theta(x) = -|x - \theta|$. However, classical proofs for asymptotic normality of M-estimators will do Taylor expansions to second order and these arguments require existence of second derivatives (and some additional regularity).

3. θ_0 is not assumed to be a unique maximum of $M(\theta), \theta \in \mathbb{R}$. Instead of this, it is assumed that $\hat{\theta}_n$ is consistent for θ_0 i.e., it converges in probability to θ_0. This means that $\hat{\theta}_n$ will be close to θ_0 and to get detailed the asymptotic picture for $\hat{\theta}_n$, we can focus on local regions of θ_0. This theorem is therefore a local result where all attention is focussed on local regions of θ_0.

We shall now prove Theorem 23.1. It will use several ideas and results that we have seen so far in this course.

Proof of Theorem 23.1. The first task is to prove that the rate of convergence is $n^{-1/2}$. For this, we can directly use the rate theorem. Letting $M_n(\theta) := P_n m_\theta$ and $d(\theta, \theta_0) = |\theta - \theta_0|$, it is easy to check that the conditions of the rate theorem hold (the key assumption is that $M(\theta_0) - M(\theta) \gtrsim d^2(\theta, \theta_0)$) which follows from the assumption that $M''(\theta_0) < 0$. To determine the rate, we have to bound

$$\mathbb{E} \sup_{\theta: |\theta - \theta_0| \le \delta} (M_n - M)(\theta - \theta_0)$$

and then equate the bound to δ^2. The above quantity equals

$$\mathbb{E} \sup_{\theta: |\theta - \theta_0| \le \delta} (P_n - P)(m_\theta - m_{\theta_0}) \le \mathbb{E} \sup_{\theta: |\theta - \theta_0| \le \delta} |(P_n - P)(m_\theta - m_{\theta_0})| .$$

To control the above expected supremum, we use the bracketing bound (from Lecture 12):

$$\mathbb{E} \sup_{h \in \mathcal{H}} \sqrt{n}|P_n h - Ph| \le C \|H\|_{L^2(P)} \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon \|H\|_{L^2(P)}, \mathcal{H}, L^2(P))} d\epsilon. \tag{217}$$

The relevant class $\mathcal{H}$ here is $\{m_\theta - m_{\theta_0} : |\theta - \theta_0| \le \delta\}$ and its envelope (by the Lipschitz condition (215)) can be taken to be $H(x) := \Gamma(x)\delta$. We thus have

$$\mathbb{E} \sup_{\theta: |\theta - \theta_0| \le \delta} |(P_n - P)(m_\theta - m_{\theta_0})| \le C \frac{\delta}{\sqrt{n}} \|\Gamma\|_{L^2(P)} \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon\delta \|\Gamma\|_{L^2(P)}, \{m_\theta - m_{\theta_0} : |\theta - \theta_0| \le \delta\}, L^2(P))}.$$

To control the bracketing numbers above, we use this result from Lecture 12: If $\Theta \subseteq \mathbb{R}^d$ is contained in a ball of radius R and if $\{g_\theta : \theta \in \Theta\}$ is a function class which satisfies $|g_{\theta_1}(x) - g_{\theta_2}(x)| \leq \Upsilon(x) \|\theta_1 - \theta_2\|$ for all x and $\theta_1, \theta_2 \in \Theta$. If $\Upsilon \in L^2(P)$, then

$$N_{[]}(\epsilon \|\Upsilon\|_{L^2(P)}, \{g_\theta, \theta \in \Theta\}, L^2(P)) \leq \left(1 + \frac{4R}{\epsilon}\right)^d \qquad \text{for every } \epsilon > 0. \tag{218}$$

Using this result with the class $\{g_\theta : |\theta - \theta_0| \leq \delta\}$ with $g_\theta := m_\theta - m_{\theta_0}$ and $\Upsilon = \Gamma$, we obtain

$$\log N_{[]}(\epsilon\delta \|\Gamma\|_{L^2(P)}, \{m_\theta - m_{\theta_0} : |\theta - \theta_0| \leq \delta\}, L^2(P)) \leq \log\left(1 + \frac{4\delta}{\epsilon\delta}\right) = \log\left(1 + \frac{4}{\epsilon}\right).$$

We thus obtain

$$\mathbb{E} \sup_{\theta:|\theta-\theta_0|\leq\delta} |(P_n - P)(m_\theta - m_{\theta_0})| \leq C\frac{\delta}{\sqrt{n}} \|\Gamma\|_{L^2(P)} \int_0^1 \sqrt{1 + \log\left(1 + \frac{4}{\epsilon}\right)} d\epsilon \leq C\frac{\delta}{\sqrt{n}} \|\Gamma\|_{L^2(P)} \leq \frac{C\delta}{\sqrt{n}}$$

because $\|\Gamma\|_{L^2(P)}$ is finite. Therefore to get a rate upper bound for $|\hat{\theta}_n - \theta_0|$, we can solve $\delta n^{-1/2} = \delta^2$ which gives $\delta = n^{-1/2}$. We have thus proved that $|\hat{\theta}_n - \theta_0| = O_P(n^{-1/2})$.

Now we shall attempt to prove (216). For this, we consider the process

$$\tilde{M}_n(h) := n\left(M_n(\theta_0 + hn^{-1/2}) - M_n(\theta_0)\right)$$

indexed by $h \in \mathbb{R}$ which can be decomposed as $\tilde{M}_n(h) = A_n(h) + B_n(h)$ where

$$A_n(h) := n\left(M_n(\theta_0 + hn^{-1/2}) - M_n(\theta_0) - M(\theta_0 + hn^{-1/2}) + M(\theta_0)\right)$$

and $B_n(h) = n\left(M(\theta_0 + hn^{-1/2}) - M(\theta_0)\right)$. By a second order Taylor expansion of M around θ_0 (note that we have assumed that $M(\theta)$ is twice continuously differentiable at its point of maximum θ_0 with $M''(\theta_0) < 0$; this also implies that $M'(\theta_0) = 0$), it can be proved that $B_n(h)$ converges to

$$B(h) := \frac{1}{2}M''(\theta_0)$$

for each fixed $h \in \mathbb{R}$. We will now show that A_n converges to a stochastic process $A(h)$ in $\ell^\infty[-K, K]$ for each fixed K. To prove this, the first step is to establish finite dimensional converges i.e., that $(A_n(h_1), \ldots, A_n(h_k))$ converges in distribution for a fixed k and $h_1, \ldots, h_k$. For this, we shall use the Lindeberg-Feller CLT. Observe that

$$(A_n(h_1), \ldots, A_n(h_k)) = \sum_{i=1}^n (Y_{ni} - \mathbb{E}Y_{ni})$$

where

$$Y_{ni} = \left(m_{\theta_0+h_1n^{-1/2}}(X_i) - m_{\theta_0}(X_i), m_{\theta_0+h_2n^{-1/2}}(X_i) - m_{\theta_0}(X_i), \ldots, m_{\theta_0+h_kn^{-1/2}}(X_i) - m_{\theta_0}(X_i).\right).$$

Because $X_1, \ldots, X_n$ are i.i.d, we have $\mathrm{Cov}(\sum_{i=1}^n Y_{ni}) = n\mathrm{Cov}(Y_{n1})$. Now for each fixed $h \in \mathbb{R}$,

$$n\mathrm{var}\left(m_{\theta_0+hn^{-1/2}}(X_1) - m_{\theta_0}(X_1)\right) = \mathbb{E}\left[\sqrt{n}\left(m_{\theta_0+hn^{-1/2}}(X_1) - m_{\theta_0}(X_1)\right)\right]^2 - \left[\mathbb{E}\sqrt{n}\left(m_{\theta_0+hn^{-1/2}}(X_1) - m_{\theta_0}(X_1)\right)\right]^2$$

Now by the almost sure first order derivative assumption on the criterion function $m_\theta(x)$ at θ_0, we have

$$\sqrt{n}\left(m_{\theta_0+hn^{-1/2}}(X_1) - m_{\theta_0}(X_1)\right) \to h\dot{m}_{\theta_0}(X_1) \qquad \text{almost surely.}$$

Also by the Lipschitz assumption (215), we have

$$\sqrt{n}\left(m_{\theta_0+hn^{-1/2}}(X_1) - m_{\theta_0}(X_1)\right) \leq |h|\Gamma(X_1).$$

Thus by the dominated convergence theorem, we obtain

$$n\text{var}\left(m_{\theta_0+hn^{-1/2}}(X_1) - m_{\theta_0}(X_1)\right) \to h^2\text{var}(\dot{m}_{\theta_0}(X_1)) \qquad \text{as } n \to \infty.$$

Similarly, for every fixed h_1 and h_2, we have

$$\begin{aligned} n\text{Cov}\left(m_{\theta_0+h_1n^{-1/2}}(X_1) - m_{\theta_0}(X_1), m_{\theta_0+h_2n^{-1/2}}(X_1) - m_{\theta_0}(X_1)\right) &\to \mathbb{E}h_1h_2\left(\dot{m}_{\theta_0}(X_1)\right)^2 - h_1h_2\left(\mathbb{E}\dot{m}_{\theta_0}(X_1)\right)^2 \\ &= h_1h_2\text{var}(\dot{m}_{\theta_0}(X_1)). \end{aligned}$$

Therefore

$$\sum_{i=1}^{n} \text{Cov}(Y_{ni}) \to \text{Cov}(A(h_1), \ldots, A(h_k)) \qquad \text{as } n \to \infty$$

where

$$A(h) := Zh\sqrt{\text{var}(\dot{m}_{\theta_0}(X_1))} \qquad \text{where } Z \sim N(0,1).$$

Further, note that

$$\|Y_{ni}\|^2 \le \sum_{j=1}^{k} \left|m_{\theta_0+h_jn^{-1/2}}(X_i) - m_{\theta_0}(X_i)\right|^2 \le \frac{\Gamma^2(X_i)}{n}\sum_{j=1}^{k} h_j^2.$$

Therefore

$$\sum_{i=1}^{n} \mathbb{E}\left(\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right) = n\mathbb{E}\left(\|Y_{n1}\|^2 I\{\|Y_{n1}\|^2 > \epsilon^2\}\right) \le \left(\sum_{j=1}^{k} h_j^2\right)\mathbb{E}\left(\Gamma^2(X_1) I\left\{\Gamma^2(X_1)\sum_{j=1}^{k} h_j^2 > n\epsilon^2\right\}\right)$$

which converges to zero as $n \to \infty$ by the Dominated Convergence theorem (note that we have assumed that $\mathbb{E}\Gamma^2(X_1) < \infty$). The assumptions of the Lindeberg-Feller CLT are all satisfied and we can thus conclude that

$$(A_n(h_1), \ldots, A_n(h_k)) \xrightarrow{L} (A(h_1), \ldots, A(h_k)) \qquad \text{as } n \to \infty$$

for every fixed $k \ge 1$ and $h_1, \ldots, h_k \in \mathbb{R}$.

To convert this finite-dimensional convergence to process level convergence in $\ell^\infty[-K, K]$ for each fixed K, we need to prove stochastic equicontinuity for which we need to bound

$$\mathbb{E}\sup_{h_1,h_2\in[-K,K]:|h_1-h_2|\le\eta} |A_n(h_1) - A_n(h_2)| = \mathbb{E}\sup_{g\in\mathcal{G}_\eta} |n(P_ng - Pg)|$$

where

$$\mathcal{G}_\eta := \left\{x \mapsto m_{\theta_0+h_1n^{-1/2}} - m_{\theta_0+h_2n^{-1/2}}(x) : h_1, h_2 \in [-K, K], |h_1 - h_2| \le \eta\right\}.$$

By the Lipschitz assumption (215), it is clear that the function $x \mapsto \eta n^{-1/2}\Gamma(x)$ is an envelope for $\mathcal{G}_\eta$. Thus the bound (217) gives

$$\begin{aligned} \mathbb{E}\sup_{g\in\mathcal{G}_\eta} |n(P_ng - Pg)| &\le \sqrt{n}\left(\|\Gamma\|_{L^2(P)}\frac{\eta}{\sqrt{n}}\right)\int_0^1 \sqrt{1 + \log N_{[]}\left(\frac{\epsilon\eta\|\Gamma\|_{L^2(P)}}{\sqrt{n}}, \mathcal{G}_\eta, L^2(P)\right)}d\epsilon \\ &= \eta\|\Gamma\|_{L^2(P)}\int_0^1 \sqrt{1 + \log N_{[]}\left(\frac{\epsilon\eta\|\Gamma\|_{L^2(P)}}{\sqrt{n}}, \mathcal{G}_\eta, L^2(P)\right)}d\epsilon. \end{aligned}$$

It is easy to see that for a small enough positive constant c,

$$N_{[]}\left(\frac{\epsilon\eta\|\Gamma\|_{L^2(P)}}{\sqrt{n}}, \mathcal{G}_\eta, L^2(P)\right) \le N_{[]}^2\left(\frac{c\epsilon\eta\|\Gamma\|_{L^2(P)}}{\sqrt{n}}, \{m_\theta - m_{\theta_0} : |\theta - \theta_0| \le Kn^{-1/2}\}, L^2(P)\right).$$

Thus by using (218) to control the bracketing numbers on the right hand side above, we obtain

$$\log N_{[]}\left(\frac{\epsilon\eta\left\|\Gamma\right\|_{L^2(P)}}{\sqrt{n}}, \mathcal{G}_\eta, L^2(P)\right) \leq 2\log\left(1+\frac{4K}{c\epsilon\eta}\right).$$

We thus obtain

$$\begin{aligned}\mathbb{E}\sup_{g\in\mathcal{G}_\eta}|n(P_n g - Pg)| &\leq \eta\left\|\Gamma\right\|_{L^2(P)}\int_0^1\sqrt{\log\left(1+\frac{4K}{c\epsilon\eta}\right)}d\epsilon \\ &\leq \eta\left\|\Gamma\right\|_{L^2(P)}\int_0^1\left(\sqrt{\log\left(1+\frac{4K}{\epsilon}\right)}+\sqrt{\log\frac{1}{\eta}}\right)d\epsilon \leq C_K\eta\left\|\Gamma\right\|_{L^2(P)}\sqrt{\log\frac{1}{\eta}}\end{aligned}$$

where C_K is a constant that only depends on K. The right hand side above clearly goes to zero as $\eta\to 0$. This proves stochastic equicontinuity of $\{A_n(h), -K\leq h\leq K\}$. Together with the finite dimensional convergence result established earlier, we can deduce that $A_n \xrightarrow{L} A$ in $\ell^\infty[-K,K]$ for every $K\geq 0$.

It can also be proved that the earlier convergence of $B_n(h)$ to $B(h)$ for each fixed $h\in\mathbb{R}$ can be improved to uniform convergence on $[-K,K]$. This is a consequence of twice continuous differentiability of M at θ_0.

Using $A_n\xrightarrow{L}A$ in $\ell^\infty[-K,K]$ and $B_n\to B$ uniformly on $[-K,K]$, we can deduce that $\tilde{M}_n = A_n + B_n \xrightarrow{L} \tilde{M} := A+B$ in $\ell^\infty[-K,K]$. We can therefore use the argmax continuous mapping theorem (all of whose conditions are met) to conclude that

$$\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{L}\operatorname*{argmax}_{h\in\mathbb{R}}\tilde{M}_n(h) = \frac{-\sqrt{\operatorname{var}(\dot{m}_{\theta_0}(X_1))}}{M''(\theta_0)}Z \sim N\left(0, \frac{\operatorname{var}(\dot{m}_{\theta_0}(X_1))}{(M''(\theta_0))^2}\right).$$

This completes the proof of Theorem 23.1. □

23.2 Application to MLE

Theorem 23.1 applies to maximum likelihood estimators. Suppose $\mathcal{P} = \{P_\theta, \theta\in\Theta\}$ denote a class of probability measures where Θ is an open subset of $\mathbb{R}$ and assume that $X_1,\ldots,X_n$ are i.i.d observations from P_{θ_0}. Assume that each P_θ has a density p_θ with respect to a common dominating measure μ. In this setting, Theorem 23.1 applies to $m_\theta(x) = \log p_\theta(x)$ and $P = P_{\theta_0}$. If the assumptions of Theorem 23.1 hold, then it follows that every MLE $\hat{\theta}_n$ has $\sqrt{n}$ rate of convergence and

$$\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{L}N\left(0,\frac{\operatorname{var}(\dot{m}_{\theta_0}(X_1))}{(M''(\theta_0))^2}\right).$$

The advantage of this result is that it only requires that $\log p_\theta$ is once differentiable at θ_0 for almost sure x (the function $\dot{m}_{\theta_0}(x)$ is called the score function). In comparison, traditional results on the asymptotic normality of the MLE require the existence of at least two derivatives of $\log p_\theta$ at θ_0. The asymptotic variance is given by

$$\frac{\operatorname{var}(\dot{m}_{\theta_0}(X_1))}{(M''(\theta_0))^2}.$$

The numerator here is the Fisher information $I(\theta_0)$. Under additional smoothness assumptions on $\dot{m}_{\theta_0}(x)$, it can be shown that $M''(\theta_0) = -I(\theta_0)$ so that the asymptotic variance is the familiar $1/I(\theta_0)$. The cleanest assumption involving the extra smoothness is Le Cam's differentiability in quadratic mean.

Definition 23.2 (Differentiability in Quadratic Mean (DQM))**.** *We say that $\{P_\theta, \theta\in\Theta\}$ is differentiable in quadratic mean at $\theta_0\in\Theta$ if there exists a function $\dot{\ell}_{\theta_0}\in L^2(P_{\theta_0})$ such that*

$$\left\|\sqrt{p_\theta}-\sqrt{p_{\theta_0}}-\frac{1}{2}(\theta-\theta_0)\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}\right\|_{L^2(\mu)} = o(|\theta-\theta_0|) \qquad \text{as } \theta\to\theta_0.$$

Under the DQM assumption, the function $\dot{\ell}_{\theta_0}$ plays the role of the score function and Fisher information will be defined by $I(\theta_0) = \text{var}_{P_{\theta_0}}(\dot{\ell}_{\theta_0}(X_1))$ (more details will be given in the next lecture). The next result asserts the $N(0, 1/I(\theta_0))$ asymptotic distribution of the MLE under DQM and an additional Lipschitz assumption on $\log p_\theta$. This is Van der Vaart [24, Theorem 5.39].

Theorem 23.3. *Suppose Θ is an open set with $\theta_0 \in \Theta$. Assume that $\{P_\theta, \theta \in \Theta\}$ satisfies DQM at θ_0. Assume also that*

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq \Gamma(x)|\theta_1 - \theta_2| \tag{219}$$

for all x and θ_1, θ_2 in a neighborhood of θ_0 with $P_{\theta_0}\Gamma^2 < \infty$. If $I(\theta_0) > 0$ and if $\hat{\theta}_n$ is consistent for θ_0, then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N\left(0, \frac{1}{I(\theta_0)}\right).$$

A crucial ingredient in the proof of Theorem 23.3 is the fact that the DQM property implies another propery known as Local Asymptotic Normality (LAN). We say that $\{P_\theta, \theta \in \Theta\}$ satisfies LAN at θ_0 if

$$\sum_{i=1}^{n} \log \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} = hS_n - \frac{1}{2}h^2 I + o_{P_{\theta_0}}(1) \tag{220}$$

where S_n converges in distribution to $N(0, I(\theta_0))$ under P_{θ_0}. It will be shown in the next lecture that DQM at θ_0 implies the LAN with $I = I(\theta_0)$. It should be clear that (220) along with the additional Lipschitz assumption (219) as well as the consistency of $\hat{\theta}_n$ implies (23.3). Indeed note first that the left hand side of (220) is $\tilde{M}_n(h)$. If we define $\tilde{M}(h) = hZ\sqrt{I} - h^2 I/2$ where $Z \sim N(0,1)$ and $I = I(\theta_0)$, then the (220) implies that the finite dimensional distributions of $\tilde{M}_n$ converge in distribution to those of $\tilde{M}$. Under the Lipschitz assumption (219), this finite dimensional convergence can be supplemented with process convergence to yield convergence in $\ell^\infty[-K, K]$ for every fixed $K \geq 0$. One can then use the argmax continuous mapping theorem to yield (23.3). This will complete the proof of Theorem 23.3. Therefore establishing (220) under the DQM assumption is key for the proof of Theorem 23.3. We shall prove this important fact (that DQM implies LAN) in the next lecture.

24 Lecture 24

This lecture will be about differentiability in quadratic mean (DQM) and local asymptotic normality (LAN). I am following the remarkably clean treatment in Pollard [19] and I recommend that you read this beautiful paper.

24.1 Differentiability in Quadratic Mean

The basic setting is the following. We have a class of probability measures $\mathcal{P} := \{P_\theta, \theta \in \Theta\}$ on some space that are indexed by a subset Θ of $\mathbb{R}$ (the extension to the case of $\Theta \subseteq \mathbb{R}^k$ for a fixed $k \geq 1$ is possible but we shall restrict to $k = 1$ for simplicity). Assume that there is a single sigma finite measure μ with respect to which each P_θ has a density which will be denoted by p_θ. The following is the definition of DQM.

Definition 24.1 (Differentiability in Quadratic Mean (DQM))**.** *We say that $\mathcal{P}$ is differentiable in quadratic mean at $\theta_0 \in \Theta$ if there exists a function $\dot{\ell}_{\theta_0} \in L^2(P_{\theta_0})$ such that*

$$\left\| \sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}} \right\|_{L^2(\mu)} = o(|\theta - \theta_0|) \qquad \text{as } \theta \to \theta_0.$$

In other words, if $\mathcal{P}$ satisfies DQM at θ_0, then we have the expansion:

$$\sqrt{p_\theta} = \sqrt{p_{\theta_0}} + \frac{1}{2}(\theta - \theta_0)\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}} + r_\theta$$

where r_θ satisfies

$$\lim_{\theta \to \theta_0} \frac{\|r_\theta\|_{L^2(\mu)}}{|\theta - \theta_0|} = 0.$$

Le Cam showed that, under DQM, classical asymptotic results in statistics (such as the asymptotic normality of maximum likelihood estimators) can be proved without requiring the densities $\theta \mapsto p_\theta(x)$ to be twice or thrice differentiable at θ_0.

The following lemma shows that if $\mathcal{P}$ satisfies DQM at θ_0 and if $\theta \mapsto p_\theta(x)$ is differentiable at θ_0 in the usual sense, then the function $\dot{\ell}_{\theta_0}$ given by the DQM coincides with the usual derivative of $\log p_\theta(x)$.

Lemma 24.2. *Suppose $\mathcal{P}$ satisfies DQM at $\theta_0 \in \Theta$ with the function $\dot{\ell}_{\theta_0}$. Assume also that $\theta \mapsto p_\theta(x)$ is differentiable at θ_0 with derivative $\dot{p}_{\theta_0}(x)$ for almost sure x with respect to the measureμ. Then*

$$\dot{\ell}_{\theta_0}(x) p_{\theta_0}(x) = \dot{p}_{\theta_0}(x) \qquad \text{for a.s } x \text{ (w.r.t } \mu).$$

Proof. Suppose $\{\theta_n\}$ is a sequence converging to θ_0. The DQM assumption allows us to write

$$\sqrt{p_{\theta_n}(x)} = \sqrt{p_{\theta_0}(x)} + \frac{1}{2}(\theta_n - \theta_0)\dot{\ell}_{\theta_0}(x)\sqrt{p_{\theta_0}(x)} + r_{\theta_n}(x) \tag{221}$$

for all x and n with

$$\lim_{n \to \infty} \frac{\|r_{\theta_n}\|_{L^2(\mu)}}{|\theta_n - \theta_0|} = 0. \tag{222}$$

By going to a subsequence if necessary, we shall assume that

$$\sum_{n \geq 0} \frac{\|r_{\theta_n}\|_{L^2(\mu)}}{|\theta_n - \theta_0|} < \infty. \tag{223}$$

This can be done, for example, by replacing $\{\theta_n\}$ by the subsequence $\{\theta_{n_k}\}$ where $\{n_k\}$ are chosen so that

$$\frac{\left\|r_{\theta_{n_k}}\right\|_{L^2(\mu)}}{|\theta_{n_k} - \theta_0|} \leq 2^{-k}.$$

We now use the fact that $\sum_{i=1}^\infty \|f_i\|_{L^1(\mu)} < \infty$ implies that $\sum_{i=1}^\infty |f_i| < \infty$ almost surely (by monotone convergence) which further implies that $f_i \to 0$ almost surely as $i \to \infty$. This gives that, under the assumption (223)

$$\frac{r_{\theta_n}(x)}{|\theta_n - \theta_0|} \to 0 \qquad \text{a.s (w.r.t } \mu) \text{ as } n \to \infty.$$

We can thus rewrite (221) as

$$\sqrt{p_{\theta_n}(x)} = \sqrt{p_{\theta_0}(x)} + \frac{1}{2}(\theta_n - \theta_0)\dot{\ell}_{\theta_0}(x)\sqrt{p_{\theta_0}(x)} + o(|\theta_n - \theta_0|) \qquad \text{a.s (w.r.t } \mu) \text{ as } n \to \infty. \tag{224}$$

Now let us use the fact that $\theta \mapsto p_\theta(x)$ is differentiable at θ_0 with derivative $\dot{p}_{\theta_0}(x)$ almost surely with respect to μ and write

$$p_{\theta_n}(x) = p_{\theta_0}(x) + (\theta_n - \theta_0)\dot{p}_{\theta_0}(x) + o(|\theta_n - \theta_0|) \qquad \text{a.s (w.r.t } \mu) \text{ as } n \to \infty. \tag{225}$$

Observe that (224) and (225) both hold for almost sure x (with respect to μ).

We shall now work with two separate cases. The first case is when $p_{\theta_0}(x) > 0$. In this case, we can rewrite (225) as (note that $p_{\theta_0}(x)$ does not depend on n so that $o(|\theta - \theta_0|/p_{\theta_0(x)}) = o(|\theta - \theta_0|)$):

$$p_{\theta_n}(x) = p_{\theta_0}(x)\left(1 + (\theta_n - \theta_0)\frac{\dot{p}_{\theta_0}(x)}{p_{\theta_0}(x)} + o(|\theta_n - \theta_0|)\right)$$

Taking square roots on both sides, we obtain

$$\sqrt{p_{\theta_n}(x)} = \sqrt{p_{\theta_0}(x)} \left(1 + (\theta_n - \theta_0)\frac{\dot{p}_{\theta_0}(x)}{p_{\theta_0}(x)} + o(|\theta_n - \theta_0|)\right)^{1/2}.$$

By a Taylor expansion of $x \mapsto \sqrt{x}$ at $x = 0$ up to first order, we deduce from above that

$$\sqrt{p_{\theta_n}(x)} = \sqrt{p_{\theta_0}(x)} \left(1 + \frac{1}{2}(\theta_n - \theta_0)\frac{\dot{p}_{\theta_0}(x)}{p_{\theta_0}(x)} + o(|\theta_n - \theta_0|)\right) = \sqrt{p_{\theta_0}(x)}1 + \frac{1}{2}(\theta_n - \theta_0)\frac{\dot{p}_{\theta_0}(x)}{\sqrt{p_{\theta_0}(x)}} + o(|\theta_n - \theta_0|).$$

Comparing the above with (224), we deduce that

$$\dot{\ell}_{\theta_0}(x) = \frac{\dot{p}_{\theta_0}(x)}{p_{\theta_0}(x)}.$$

Let us now consider the case when $p_{\theta_0}(x) = 0$. In this case, (224) and (225) become respectively

$$\sqrt{p_{\theta_n}(x)} = o(|\theta_n - \theta_0|) \implies p_{\theta_n}(x) = o(|\theta_n - \theta_0|^2)$$

and

$$p_{\theta_n}(x) = (\theta_n - \theta_0)\dot{p}_{\theta_0}(x) + o(|\theta_n - \theta_0|).$$

Equating the above two equations, we obtain

$$o(|\theta_n - \theta_0|^2) = (\theta_n - \theta_0)\dot{p}_{\theta_0}(x) + o(|\theta_n - \theta_0|).$$

Dividing through by $|\theta_n - \theta_0|$ and letting $n \to \infty$, we obtain $\dot{p}_{\theta_0}(x) = 0$ i.e., the equation $\dot{\ell}_{\theta_0}(x)p_{\theta_0}(x) = \dot{p}_{\theta_0}(x)$ is satisfied in this case as well. This completes the proof. □

The above lemma implies that

$$\dot{\ell}_{\theta_0}(x) = \frac{\dot{p}_{\theta_0}(x)}{p_{\theta_0}(x)} \qquad \text{whenever } p_{\theta_0}(x) > 0.$$

The right hand side above is the classical *score function*. Thus when the DQM holds, we shall refer to the function $\dot{\ell}_{\theta_0}$ as the score function.

A standard fact about the classical score function is that its expectation with respect to the probability measure P_{θ_0} equals zero. The classical proof for this involves interchanging the order of differentiation w.r.t θ and the integral:

$$\int \dot{\ell}_{\theta_0}(x)p_{\theta_0}(x)d\mu(x) = \int \dot{p}_{\theta_0}(x)d\mu(x) = \left(\int \dot{p_\theta}(x)d\mu(x)\right) = 0.$$

The following lemma shows that the DQM assumption implies this fact directly.

Lemma 24.3. *Suppose $\mathcal{P}$ satisfies DQM at θ_0 with score function $\dot{\ell}_{\theta_0}$. Then*

$$\int \dot{\ell}_{\theta_0}(x)p_{\theta_0}(x)d\mu(x) = 0. \tag{226}$$

Proof. Let θ_n be a sequence converging to θ_0. By the DQM representation, we can write (221) with the remainder term r_{θ_n} satisfying (222). Note then that

$$1 = \int p_{\theta_n} d\mu = \int \left(\sqrt{p}_{\theta_0}(x) + \frac{1}{2}(\theta_n - \theta_0)\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}} + r_{\theta_n}\right)^2 d\mu.$$

We now expand the square in the right hand side above which will lead to six terms. One of the terms equals $\int p_{\theta_0} = 1$ which cancels with the left hand side. We thus obtain

$$0 = (\theta_n - \theta_0) \int \dot{\ell}_{\theta_0} p_{\theta_0} d\mu + 2 \int \sqrt{p_{\theta_0}} r_{\theta_n} d\mu + \frac{1}{4}(\theta_n - \theta_0)^2 \int (\dot{\ell}_{\theta_0})^2 p_{\theta_0} d\mu + (\theta_n - \theta_0) \int \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} r_{\theta_n} d\mu + \int r_{\theta_n}^2 d\mu \tag{227}$$

The first term in the right hand side above is clearly $O(|\theta_n - \theta_0|)$ in absolute value. The third term is $O((\theta_n - \theta_0)^2)$. The final term (by (222)) equals $o((\theta_n - \theta_0)^2)$. The remaining two terms (second and fourth) can be controlled via the Cauchy-Schwarz inequality as

$$2 \int \sqrt{p_{\theta_0}} |r_n| d\mu \leq 2 \sqrt{\int p_{\theta_0} d\mu} \sqrt{\int r_{\theta_n}^2 d\mu} = 2 \sqrt{\int r_{\theta_n}^2 d\mu} = 2o(|\theta_n - \theta_0|)$$

by (222) and

$$\left| (\theta_n - \theta_0) \int \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} r_{\theta_n} d\mu \right| \leq |\theta_n - \theta_0| \sqrt{\int (\dot{\ell}_{\theta_0})^2 p_{\theta_0} d\mu} \sqrt{\int r_{\theta_n}^2 d\mu} = o(|\theta_n - \theta_0|^2)$$

again by (222). It is clear therefore that the leading term on the right hand side in (227) is the first term. By dividing the equation (227) through by $|\theta_n - \theta_0|$ and letting $n \to \infty$, we deduce (226). □

We shall now define *Fisher Information*. Assume that $\mathcal{P}$ satisfies DQM with score function $\dot{\ell}_{\theta_0}$. Then the Fisher Information at θ_0 is given by

$$I(\theta_0) = \text{var}_{P_{\theta_0}}(\dot{\ell}_{\theta_0}(X)) = \mathbb{E}_{P_{\theta_0}} \left(\dot{\ell}_{\theta_0}(X) \right)^2 = \int \left(\dot{\ell}_{\theta_0}(x) \right)^2 p_{\theta_0}(x) d\mu(x).$$

The argument used in the proof of Lemma 24.3 above leads to an interesting and important fact involving $\dot{\ell}_{\theta_0}$ and the Fisher Information. Because $\int \dot{\ell}_{\theta_0} p_{\theta_0} d\mu = 0$, we can plug this into (227) to obtain (also using the fact that the last two terms in (227) are $o(|\theta_n - \theta_0|^2)$):

$$2 \int \sqrt{p_{\theta_0}} r_{\theta_n} d\mu + \frac{1}{4}(\theta_n - \theta_0)^2 \int (\dot{\ell}_{\theta_0})^2 p_{\theta_0} d\mu = o(|\theta_n - \theta_0|^2).$$

Plugging in the fact that $\int (\dot{\ell}_{\theta_0})^2 p_{\theta_0} d\mu = I(\theta_0)$, we obtain

$$2 \int \sqrt{p_{\theta_0}} r_{\theta_n} d\mu = -\frac{1}{4}(\theta_n - \theta_0)^2 I(\theta_0) + o(|\theta_n - \theta_0|^2). \tag{228}$$

This fact is crucial for establishing that DQM implies LAN. The interesting aspect about (228) is the following. The statement (222) implies that $\|r_{\theta_n}\|_{L^2(\mu)} = o(|\theta_n - \theta_0|)$. Therefore, if we use the Cauchy-Schwarz inequality on the left hand side above, we obtain that the left hand side is $o(|\theta_n - \theta_0|)$. But the equality above implies that the right hand side is $O((\theta_n - \theta_0)^2)$ which is a much stronger conclusion that what can be derived from Cauchy-Schwarz. Therefore $\int r_{\theta_n} \sqrt{p_{\theta_0}} d\mu$ is much smaller in comparison to the $L^2(\mu)$ norm of r_{θ_n}. Pollard [19] attributes this phenomenon to the fact that the functions $\sqrt{p_{\theta_n}}$ in $L^2(\mu)$ all have norm one (this is clear from the above proof of (228)) and argues that this is the main reason behind the magic of the DQM.

24.2 Local Asymptotic Normality

The DQM is a statement about the first differentiability of the densities p_θ at θ_0. There is of course no mention of second order differentiability in DQM. Yet, remarkably, the DQM assumption implies that the log-likelihood function has a second-order Taylor expansion around θ_0 at a scale of $n^{-1/2}$. Such a local Taylor expansion is known as Local Asymptotic Normality (LAN) and is proved in the following result.

Theorem 24.4. *Suppose $\mathcal{P}$ satisfies DQM at θ_0 with score function $\dot{\ell}_{\theta_0}$ and Fisher information $I(\theta_0)$. Then for every fixed $h \in \mathbb{R}$, we have*

$$\left| \sum_{i=1}^{n} \log \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} - \frac{h}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + \frac{1}{2} h^2 I(\theta_0) \right| \xrightarrow{P_{\theta_0}} 0 \qquad \text{as } n \to \infty.$$

Equivalently, the conclusion of the above theorem can be written

$$\sum_{i=1}^{n} \log \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} = \frac{h}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2} h^2 I(\theta_0) + o_{P_{\theta_0}}(1) \qquad \text{as } n \to \infty. \tag{229}$$

We say that $\mathcal{P}$ satisfies the LAN property at θ_0 if the above holds for every $h \in \mathbb{R}$. Why is this called local asymptotic normality? To see this, note first that, by the CLT, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) \xrightarrow{L} N(0, I(\theta_0)).$$

Therefore, as a consequence of (229), we obtain that for every $h \in \mathbb{R}$,

$$\sum_{i=1}^{n} \log \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} \xrightarrow{L} hN(0, I(\theta_0)) - \frac{1}{2} h^2 I(\theta_0) \qquad \text{under } X_1, \ldots, X_n \sim^{i.i.d} P_{\theta_0}.$$

Now consider a second estimation problem where we have one observation Y whose density belongs to the family $\{\mathcal{Q}_h, h \in \mathbb{R}\}$ were $\mathcal{Q}_h$ has the density q_{sh} which is the density of the normal distribution with mean h and variance $1/I(\theta_0)$. It is easy to see then that

$$\log \frac{q_h(Y)}{q_0(Y)} \sim hN(0, I(\theta_0)) - \frac{1}{2} h^2 I(\theta_0) \qquad \text{under } Y \sim N(0, 1/I(\theta_0)).$$

Therefore (229) effectively says that the likelihood ratios of $\{P_\theta, \theta \in \Theta\}$ (which can be arbitrary as long as $\mathcal{P}$ satisfies DQM) behave like the likelihood ratios of a Normal Experiment $\{Q_h, h \in \mathbb{R}\}$ where $Q_h = N(h, 1)$. Hence asymptotically around θ_0 at the scale $n^{-1/2}$, the original statistical problem $\mathcal{P}$ becomes a Normal mean estimation problem. This is why (229) is referred to as Local Asymptotic Normality.

We shall now prove Theorem 24.4.

Proof of Theorem 24.4. All expectations and probabilities in this proof are with respect to the probability measure P_{θ_0}. Write

$$L_n := \sum_{i=1}^{n} \log \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} = 2 \sum_{i=1}^{n} \log \sqrt{\frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)}} = 2 \sum_{i=1}^{n} \log (1 + W_{ni})$$

where

$$W_{ni} := \sqrt{\frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)}} - 1.$$

We will use the fact that

$$\log(1 + y) = y - \frac{y^2}{2} + \frac{1}{2} \beta(y) \qquad \text{where } \lim_{y \to 0} \frac{\beta(y)}{y^2} = 0$$

or equivalently, $\beta(y) = o(y^2)$ as $y \to 0$. This gives

$$L_n = 2 \sum_{i-1}^{n} W_{ni} - \sum_{i=1}^{n} W_{ni}^2 + \sum_{i=1}^{n} \beta(W_{ni}).$$

Using the DQM representation, we can write

$$W_{ni} = \frac{\sqrt{p_{\theta_0+hn^{-1/2}}(X_i)} - \sqrt{p_{\theta_0}(X_i)}}{\sqrt{p_{\theta_0}(X_i)}} = \frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i) + \frac{r_{\theta_0+hn^{-1/2}}(X_i)}{\sqrt{p_{\theta_0}(X_i)}} = \frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i) + R_{ni}$$

where

$$R_{ni} = \frac{r_{\theta_0+hn^{-1/2}}(X_i)}{\sqrt{p_{\theta_0}(X_i)}}.$$

We thus get

$$\begin{aligned} L_n &= \frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + 2\sum_{i=1}^n R_{ni} - \sum_{i=1}^n \left(\frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i) + R_{ni}\right)^2 + \sum_{i=1}^n \beta\left(\frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i) + R_{ni}\right) \\ &= \frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + 2\sum_{i=1}^n R_{ni} - \frac{h^2}{4n}\sum_{i=1}^n (\dot{\ell}_{\theta_0}(X_i))^2 - \frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)R_{ni} - \sum_{i=1}^n R_{ni}^2 + \sum_{i=1}^n \beta\left(\frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i) + R_{ni}\right). \end{aligned}$$

Observe now that by the DQM, we know the following about the random variables R_{ni}:

$$\mathbb{E}_{\theta_0} R_{ni}^2 = \mathbb{E}_{\theta_0}\frac{r^2_{\theta_0+hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} = \left\|r_{\theta_0+hn^{-1/2}}(x)\right\|_{L^2(\mu)} = o\left(\frac{h^2}{n}\right) = o(n^{-1}).$$

This gives that $\sum_{i=1}^n \mathbb{E}_{\theta_0} R_{ni}^2 = o(1)$ and hence $\sum_{i=1}^n R_{ni}^2 \to 0$ in $L^1(P_{\theta_0})$ which further implies that $\sum_{i=1}^n R_{ni}^2 \xrightarrow{P} 0$. Also, by the Cauchy-Schwarz inequality, we have

$$\left|\frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)R_{ni}\right| \le h\sqrt{\frac{1}{n}\sum_{i=1}^n (\dot{\ell}_{\theta_0}(X_i))^2}\sqrt{\sum_{i=1}^n R_{ni}^2} \xrightarrow{P} h\sqrt{I(\theta_0)}\sqrt{0} = 0.$$

We thus have

$$L_n = \frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + 2\sum_{i=1}^n R_{ni} - \frac{h^2}{4n}\sum_{i=1}^n (\dot{\ell}_{\theta_0}(X_i))^2 + \sum_{i=1}^n \beta\left(\frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i) + R_{ni}\right) + o_{P_{\theta_0}}(1).$$

We shall prove later that

$$\sum_{i=1}^n \beta\left(\frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i) + R_{ni}\right) = o_{P_{\theta_0}}(1) \tag{230}$$

so that we have

$$L_n = \frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + 2\sum_{i=1}^n R_{ni} - \frac{h^2}{4n}\sum_{i=1}^n (\dot{\ell}_{\theta_0}(X_i))^2 + o_{P_{\theta_0}}(1).$$

The third term in the right hand side above clearly converges to $-h^2 I(\theta_0)/4$ in probability (by the Strong Law of Large Numbers) so to complete the proof of Theorem 24.4, we only need to show that

$$2\sum_{i=1}^n R_{ni} \xrightarrow{P} -\frac{h^2}{4}I(\theta_0). \tag{231}$$

For this, write

$$2\sum_{i=1}^n R_{ni} = 2\sum_{i=1}^n \mathbb{E}_{\theta_0} R_{ni} + 2\sum_{i=1}^n (R_{ni} - \mathbb{E}R_{ni}).$$

Because

$$\mathbb{E}\left(2\sum_{i=1}^{n}(R_{ni}-\mathbb{E}R_{ni})\right)^2 = 4\sum_{i=1}^{n}\text{var}(R_{ni}) \leq 4\sum_{i-1}^{n}\mathbb{E}R_{ni}^2 = o(1),$$

we get

$$2\sum_{i=1}^{n}R_{ni} = 2\sum_{i=1}^{n}\mathbb{E}R_{ni} + o_{P_{\theta_0}}(1) \qquad \text{as } n\to\infty. \tag{232}$$

Note that

$$2\sum_{i=1}^{n}\mathbb{E}R_{ni} = 2n\mathbb{E}R_{n1} = 2n\mathbb{E}\frac{r_{\theta_0+hn^{-1/2}}(X_1)}{\sqrt{p_{\theta_0}(X_1)}} = 2n\int r_{\theta_0+hn^{-1/2}}\sqrt{p_{\theta_0}}d\mu.$$

We shall now use the fact (228) which gives

$$2\int r_{\theta_0+hn^{-1/2}}\sqrt{p_{\theta_0}}d\mu = -\frac{h^2}{4n}I(\theta_0)+o(n^{-1})$$

so that

$$2\sum_{i=1}^{n}\mathbb{E}R_{ni} = -\frac{h^2}{4}I(\theta_0)+o(1).$$

Combining with (232), we obtain (231). To finish the proof of Theorem 24.4, we only need to verify (230). This is mainly a consequence of $\beta(y)=o(y^2)$ as $y\to 0$. It turns out that in order to prove (230), it is enough to show that

$$\max_{1\leq i\leq n}\left|\frac{\dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}}\right| = o_{P_{\theta_0}}(1) \quad \text{and} \quad \max_{1\leq i\leq n}|R_{ni}| = o_{P_{\theta_0}}(1). \tag{233}$$

Indeed, if these statements hold, then (as $\beta(y)=o(y^2)$), we can write (rigorize this):

$$\begin{aligned}\left|\sum_{i=1}^{n}\beta\left(\frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i)+R_{ni}\right)\right| &\leq o_{P_{\theta_0}}(1)\sum_{i=1}^{n}\left(\frac{h}{2\sqrt{n}}\dot{\ell}_{\theta_0}(X_i)+R_{ni}\right)^2\\ &\leq 2o_{P_{\theta_0}}(1)\left(\frac{h^2}{n}\sum_{i=1}^{n}\dot{\ell}_{\theta_0}(X_i)+\sum_{i=1}^{n}R_{ni}^2\right) = o_{P_{\theta_0}}(1).\end{aligned}$$

We shall complete the proof now by proving the assertions in (233). For the first assertion in (233), write (for a fixed $\epsilon>0$),

$$\mathbb{P}_{\theta_0}\left\{\max_{1\leq i\leq n}\left|\frac{\dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}}\right|>\epsilon\right\} \leq \sum_{i=1}^{n}\mathbb{P}\left\{\left|\frac{\dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}}\right|>\epsilon\right\} = n\mathbb{P}\left\{\left|\frac{\dot{\ell}_{\theta_0}(X_1)}{\sqrt{n}}\right|>\epsilon\right\} \leq \frac{1}{\epsilon^2}\mathbb{E}(\dot{\ell}_{\theta_0}(X_1))^2 I\left\{\left|\frac{\dot{\ell}_{\theta_0}(X_1)}{\sqrt{n}}\right|>\epsilon\right\}$$

which converges to zero as $n\to\infty$ by the Dominated Convergence Theorem.

For the second assertion in (233), write

$$\mathbb{P}\left\{\max_{1\leq i\leq n}|R_{ni}|>\epsilon\right\} \leq n\mathbb{P}\left\{|R_{n1}|>\epsilon\right\} \leq \frac{n}{\epsilon^2}\mathbb{E}R_{n1}^2 \to 0.$$

This completes the proof of Theorem 24.4. □

25 Lecture 25

The next (and last) topic in this class is about minimax lower bounds. In this lecture, we shall mainly motivate the study of minimax lower bounds. We start with the basic decision-theoretic setting under which these are studied.

25.1 Decision Theoretic Framework

Minimaxity can be studied in the general and abstract decision theoretic framework. This framework is described here. A classical reference for this is the book Ferguson [8, Chapter 1 and 2].

We have an unknown parameter θ. θ can be a real-number or a vector or a function or a matrix etc. We assume that θ takes values in a known set Θ which we refer to as the Parameter Space.

The data will be generically be denoted by X. Again X can be a real-number, vector or function or matrix etc. We assume that X takes values in a set $\mathcal{X}$ which is referred to as the sample space.

The connection between X and θ is that the distribution of X depends on θ through a known probability measure P_θ. The class of all probability measures $\{P_\theta, \theta \in \Theta\}$ will be denoted by $\mathcal{P}$. We assume that each P_θ has a density p_θ with respect to a single sigma finite measure μ.

Next, we have an action space $\mathcal{A}$ which corresponds to the actions that the statistician needs to take in the problem (for example, in estimation problems, $\mathcal{A}$ will be equal to or larger than Θ, in testing problems with a null and an alternative hypothesis, $\mathcal{A}$ will correspond to the two hypotheses etc. specific examples are given below).

The loss function L is a nonnegative function defined on $\Theta \times \mathcal{A}$ i.e., for every parameter $\theta \in \Theta$ and action $a \in \mathcal{A}$, there is associated a nonnegative loss $L(\theta, a)$.

A nonrandomized decision rule d is a function from $\mathcal{X}$ to $\mathcal{A}$. In other words, d associates an action to every $x \in \mathcal{X}$. The risk of a decision rule d at a particular parameter value θ is defined by

$$R(\theta, d) := \mathbb{E}_\theta L(\theta, d(X))$$

where the expectation above is taken with respect to $X \sim P_\theta$.

The goal of a statistician in a decision problem is to choose a decision rule d whose risk $R(\theta, d)$ is small. This statement of "small risk" needs to be qualified further however because the risk $R(\theta, d)$ depends on the unknown θ. In other words, the risk of a decision rule d depends on what the unknown parameter value (or state of nature) is. So when we say small risk, we need to specify if we mean uniformly small risk over θ or small average risk or small worst case risk. We shall come back to this issue shortly after seeing some examples of decision-theoretic problems.

Example 25.1. *Consider the problem of estimating a vector $\theta \in \mathbb{R}^n$ from an observation $Y \sim N_n(\theta, I_n)$ under squared error loss. Suppose it is known that θ is k-sparse i.e., the number of non-zero entries in θ is at most k. This can be be put in the decision theoretic framework outlined above by taking Θ to be the set of all k-sparse vectors in $\mathbb{R}^n$, $\mathcal{X} = \mathbb{R}^n$, $\mathcal{A} = \Theta$ or $\mathcal{A} = \mathbb{R}^n$ (depending on whether we want our estimators to be k-sparse or not) and $L(\theta, a) = \|\theta - a\|^2$. Also P_θ is the $N_n(\theta, I_n)$ distribution. Decision rules are simply estimators for θ and the risk of an estimator $\hat{\theta}$ is given by*

$$R(\theta, \hat{\theta}) := \mathbb{E}_\theta \left\|\theta - \hat{\theta}\right\|^2.$$

Example 25.2. *Consider the same setting as the last problem but suppose now that we want to estimate the L^1-norm of θ (and not the entire vector θ): $\|\theta\|_1 := |\theta_1| + \cdots + |\theta_n|$. Then $\Theta, \mathcal{X}$ and P_θ remain the same as in the previous example but $\mathcal{A} = \mathbb{R}$ and the loss function is $L(\theta, a) = (\|\theta\|_1 - a)^2$.*

Example 25.3. *Consider the problem of estimating a Lipschitz function $f : [0,1] \to \mathbb{R}$ from independent observations $Y_1, \ldots, Y_n$ with $Y_i \sim N(f(i/n), 1)$ for $i = 1, \ldots, n$. In this problem, we can take Θ to be the class of all Lipschitz functions from $[0,1]$ to $\mathbb{R}$. For $f \in \Theta$, the probability measure P_f is the multivariate normal distribution with mean $(f(1/n), \ldots, f(n/n))$ and covariance matrix I_n. The action space can be taken to be the space of all real-valued functions on $[0,1]$ and the loss function can be either:*

$$L(f,g) := \int_0^1 (f(x) - g(x))^2 \, dx \quad \text{or} \quad \frac{1}{n} \sum_{i=1}^n (f(i/n) - g(i/n))^2 .$$

Example 25.4. *Consider the problem of testing $H_0 : \theta = 0$ against $H_1 : \theta = 1$ from n independent observations $X_1, \ldots, X_n$ drawn from $N(\theta, 1)$. Because we are testing $\theta = 0$ against $\theta = 1$, we believe that only these two values are possible so we take $\Theta = \{0, 1\}$. The action space is then also $\mathcal{A} = \{0, 1\}$. A natural loss function is $L(\theta, a) = I\{\theta \neq a\}$. Given a decision rule d (test), its risk is given by*

$$R(\theta, d) = P_\theta\{\theta \neq d(X)\}.$$

Note that if $\theta = 0$, then the risk is given by $R(0, d) = P_0\{d(X) = 1\}$ which is the usual Type I error. When $\theta = 1$, the risk is given by $R(1, d) = P_1\{d(X) = 0\}$ which is the Type II error.

Example 25.5. *Consider the problem of testing the hypothesis $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ on the basis of n i.i.d observations $X_1, \ldots, X_n$ from the $N(\theta, 1)$ distribution. In this case, $\Theta = \Theta_0 \cup \Theta_1$, $\mathcal{A} = \{0, 1\}$ and*

$$L(\theta, a) = I\{\theta \in \Theta_0, a = 1\} + I\{\theta \in \Theta_1, a = 0\}.$$

The risk of a decision rule (test) d is given by $R(\theta, d) = P_\theta\{d(X) = 1\}$ if $\theta \in \Theta_0$ and $R(\theta, d) = P_\theta\{d(X) = 0\}$ if $\theta \in \Theta_1$. These can again be treated as type I and type II errors respectively. Note that these depend on θ (i.e., there is a family of type I errors for each $\theta \in \Theta_0$ and a family of type II errors for each $\theta \in \Theta_1$).

25.2 How to evaluate decision rules

As mentioned earlier, the risk $R(\theta, d)$ of a decision rule d depends on θ. It turns out that usually it is impossible to find a single decision rule d^* such that

$$R(\theta, d^*) \leq R(\theta, d) \qquad \text{for every } \theta \in \Theta \text{ and decision rule } d. \tag{234}$$

For example, in an estimation problem with $\Theta = \mathcal{A} \subseteq \mathbb{R}^k$ and $L(\theta, a) = \|\theta - a\|^2$. Consider the estimator $d_0(X) = \theta_0$ for a fixed $\theta_0 \in \Theta$. This estimator clearly has risk equal to 0 at $\theta = \theta_0$ (i.e., $R(\theta_0, d_0) = 0$). Thus if there existed a decision rule d^* which satisfies (234), then $R(\theta_0, d^*) \leq R(\theta_0, d_0) = 0$. Since $\theta_0 \in \Theta$ is arbitrary here, this must mean that

$$R(\theta, d^*) = \mathbb{E}_\theta \|\theta - d^*(X)\|^2 = 0$$

for every $\theta \in \Theta$. This implies that $d^*(X) = \theta$ almost surely under P_θ for every $\theta \in \Theta$. This obviously cannot happen for general classes $\{P_\theta, \theta \in \theta\}$.

Therefore we cannot hope for an optimal decision rule d^* in the strong sense (234). There are three common ways of getting a relaxed notion of optimality:

1. The first way involves restricting to a subclass of all decision rules. For example, in parametric estimation problems, it is common to restrict attention to unbiased or equivariant estimators. In parametric testing problems, it is natural to restrict attention to level α tests or unbiased level α tests. This approach was taken in STAT 210A and we will not pursue it here.
2. Bayes approach
3. Minimax approach

We will study the Bayes and Minimax approaches in detail here.

25.3 Bayes Approach

Here we fix a probability measure w on Θ and evaluate decision rules by their average risk (where the averaging is done with respect to w). In other words, we evaluate decision rules d by their average risk:

$$\int R(\theta, d) dw(\theta) \tag{235}$$

with respect to the probability measure w. The probability measure w is also referred to as a proper prior or simply prior. The smallest achievable average risk is called the Bayes risk with respect to w and is denoted by

$$R_{\text{Bayes}}(w) := \inf_d \int_\Theta R(\theta, d) w(d\theta).$$

and the estimator d which minimizes (235) is known as the Bayes estimator with respect to w.

The obvious problem with this approach of evaluating decision rules is its dependence on the prior w and, in many situations, it is not clear what a reasonable choice of the prior is. For example, in the Lipschitz regression problem of Example 25.3, one would need to choose a prior on the class of all Lipschitz functions on $[0, 1]$ and it is not clear how one can do this.

The above issue notwithstanding, the Bayes approach has the important advantage in that finding the Bayes rule (the rule which minimizes (235)) is, in principle, tractable. Indeed, we can write (235) as

$$\int R(\theta, d) dw(\theta) = \int \mathbb{E}_\theta L(\theta, d(X)) dw(\theta) = \int_\Theta \int_{\mathcal{X}} L(\theta, d(x)) dP_\theta(x) dw(\theta) = \int_\Theta \int_{\mathcal{X}} L(\theta, d(x)) p_\theta(x) d\mu(x) dw(\theta).$$

We now interchange the order of integration above (this is allowed because the loss function is nonnegative) to get

$$\int R(\theta, d) dw(\theta) = \int_{\mathcal{X}} \left\{ \int_\Theta L(\theta, d(x)) p_\theta(x) dw(\theta) \right\} d\mu(x).$$

From the simple inequality

$$\int_\Theta L(\theta, d(x)) p_\theta(x) dw(\theta) \geq \inf_{a \in \mathcal{A}} \int_\Theta L(\theta, a) p_\theta(x) dw(\theta)$$

which holds for every $x \in \mathcal{X}$, it should be clear that the rule which minimizes (235) is given by

$$d^*(x) := \operatorname*{argmin}_{a \in \mathcal{A}} \int_\Theta L(\theta, a) p_\theta(x) dw(\theta) = \operatorname*{argmin}_{a \in \mathcal{A}} \int_\Theta L(\theta, a) \frac{p_\theta(x) dw(\theta)}{\int_\Theta p_\theta(x) dw(\theta)}.$$

The density

$$x \mapsto \frac{p_\theta(x) dw(\theta)}{\int_\Theta p_\theta(x) dw(\theta)}$$

is simply the posterior density of θ given $X = x$ in the model $X|\theta \sim p_\theta$ and $\theta \sim w$. We thus obtain the well-known fact that Bayes rule minimizes the posterior expectation of the Loss function. For example, in the case of the squared error loss $L(\theta, a) = \|\theta - a\|^2$, the Bayes rule is simply the expectation of the posterior distribution.

The above calculation also gives an exact expression for the Bayes risk with respect to w:

$$R_{\text{Bayes}}(w) = \int_{\mathcal{X}} \inf_{a \in \mathcal{A}} \left\{ \int_\Theta L(\theta, a) p_\theta(x) dw(\theta) \right\} d\mu(x). \tag{236}$$

25.4 Minimax Approach

In the minimax approach, we evaluate decision rules by their worst case (supremum) risk over $\theta \in \Theta$. In other words, we aim to select a decision rule d for which $\sup_{\theta\in\Theta} R(\theta, d)$ is small. The advantage with this approach is that one does not need to select a specific prior distribution. The disadvantage is that it focuses on worst case behavior and might be regarded as too pessimistic. Nevertheless, this is most widely used optimality criterion currently.

The minimax risk is defined as

$$R_{\text{Minimax}} := \inf_{d} \sup_{\theta\in\Theta} \mathbb{E}_\theta L(\theta, d(X))$$

where the infimum is taken over all decision rules d. A decision rule d^* is said to be minimax if

$$\frac{\sup_{\theta\in\Theta} R(\theta, d^*)}{R_{\text{Minimax}}} = 1.$$

Finding minimax estimators is quite difficult in many problems so one is often with approximate minimaxity. There are two commonly used notions of approximate minimaxity. These are defined in terms of a "sample size" or "dimension" parameter n that is present in most decision problems. Specifically, we assume that possibly all the ingredients of the decision problem (i.e., Θ, $\mathcal{A}$, $L(\theta, a)$ and P_θ) depend on a sample size or dimension parameter n and we are interested in the problem only for large values of n. In this context, we have the following two definitions:

1. **Sharp Asymptotically Minimaxity**: We say that a decision rule d^* is sharp asymptotically minimax if
$$\frac{\sup_{\theta\in\Theta} R(\theta, d^*)}{R_{\text{Minimax}}} \to 1$$
as $n \to \infty$. This is equivalent to
$$\sup_{\theta\in\Theta} R(\theta, d^*) = R_{\text{Minimax}}\,(1 + o(1)) \qquad \text{as } n \to \infty.$$

2. **Rate Minimaxity**: We say that a decision rule d^* is rate minimax if
$$\frac{\sup_{\theta\in\Theta} R(\theta, d^*)}{R_{\text{Minimax}}} \leq C$$
for a constant C that does not depend on n. This is equivalent to saying that
$$\sup_{\theta\in\Theta} R(\theta, d^*) = O(R_{\text{Minimax}}) \qquad \text{as } n \to \infty.$$

Consider now the following situation. Suppose we have a decision rule d^* which we have constructed (say by some M-estimation method) and we have a good understanding of its performance in the sense that we have an upper bound u_n on its supremum risk over Θ i.e., we know that

$$\sup_{\theta\in\Theta} R(\theta, d^*) \leq u_n.$$

How then would we show that d^* is minimax (in one of the above senses: minimax, sharp asympotically minimax or rate minimax)? It is obvious that in order to do this we need to bound R_{Minimax} from below. Indeed, if we prove the minimax lower bound:

$$R_{\text{Minimax}} \geq \ell_n \tag{237}$$

then we can assert

1. minimaxity if $\ell_n = u_n$ for every n.
2. sharp asymptotic minimaxity if $u_n/\ell_n \to 1$ as $n \to \infty$.
3. rate minimaxity if $u_n/\ell_n = O(1)$ as $n \to \infty$.

Of course the key to doing this is to be able to prove minimax lower bounds (i.e., bounds of the form (237)) which we shall study now.

25.5 Minimax Lower Bounds

There is basically only one technique for proving lower bounds on the minimax risk. This involves bounding the minimax risk from below by a Bayes risk. Indeed

$$R_{\text{Minimax}} \geq R_{\text{Bayes}}(w) \qquad \text{for every probability measure } w \text{ on } \Theta. \tag{238}$$

As we indicated earlier, the Bayes risk $R_{\text{Bayes}}(w)$ is a much more tractable object (compared to R_{Minimax}) and it has the exact expression (236).

Inequality (238) can be rewritten as

$$R_{\text{Minimax}} \geq \sup_w R_{\text{Bayes}}(w) = \sup_w \inf_d \int \mathbb{E}_\theta L(\theta, d(X)) dw(\theta)$$

where the supremum is taken over all probability measures w on Θ. On the other hand, it is easy to see that the minimax risk satisfies:

$$R_{\text{Minimax}} = \inf_d \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, d(X)) = \inf_d \sup_w \int \mathbb{E}_\theta L(\theta, d(X)) dw(\theta).$$

We thus have

$$\inf_d \sup_w \int \mathbb{E}_\theta L(\theta, d(X)) dw(\theta) = R_{\text{Minimax}} \geq \sup_w R_{\text{Bayes}}(w) = \sup_w \inf_d \int \mathbb{E}_\theta L(\theta, d(X)) dw(\theta).$$

Suppose now that

$$\inf_d \sup_w \int \mathbb{E}_\theta L(\theta, d(X)) dw(\theta) = \sup_w \inf_d \int \mathbb{E}_\theta L(\theta, d(X)) dw(\theta),$$

then we would have $R_{\text{Minimax}} = \sup_w R_{\text{Bayes}}(w)$ which would imply that the only way to bound the minimax risk from below is via a Bayes risk for an appropriate prior w. An infimum and a supremum obviously cannot always be interchanged; for example,

$$R_{\text{Minimax}} = \inf_d \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, d(X)) \neq \sup_{\theta \in \Theta} \inf_d \mathbb{E}_\theta L(\theta, d(X)) = 0.$$

However, under some conditions, they can be interchanged. Theorems which guarantee the interchange are known as minimax theorems. There exist a variety of such minimax theorems in the literature one example of which is the following (known sometime as Kneser's minimax theorem):

Theorem 25.6. *Let K be a convex subset of a vector space $\mathcal{X}$ and let L be a compact convex subset of a Hausdorff Topological Vector Space $\mathcal{Y}$. Suppose $f : K \times L \to \mathbb{R}$ is a function such that*

1. *$x \mapsto f(x, y)$ is convex for each fixed $y \in L$.*
2. *$y \mapsto f(x, y)$ is concave and continuous for each fixed $x \in K$.*

Then

$$\inf_{x \in K} \sup_{y \in L} f(x, y) = \sup_{y \in L} \inf_{x \in K} f(x, y).$$

Potentially this theorem can be applied to verify (25.5). For this, we can take K to be the class of all decision rules d and L to be the set of all probability measures w on Θ. We would then need

$$(d, w) \mapsto \int \mathbb{E}_\theta L(\theta, d(X)) dw(\theta)$$

to be convex in d for each fixed w and concave in w for each fixed d. The concavity in w is alright but in order to ensure convexity in d, we need to switch to randomized decision rules and extend the notion of risk to randomized decision rules. In order to verify the compactness assumptions, one needs to put a topology on the space of all probability measures on Θ. These can be done in quite some generality but the details are quite involved. You can see Le Cam and Yang [15] or Le Cam [14] for full details.

To summarize this section, the inequality (238) is always true. Also usually, $R_{\text{Minimax}} = \sup_w R_{\text{Bayes}}(w)$ so (238) is really the only way of obtaining minimax lower bounds. Because of the exact expression (236) for the Bayes risk, we have

$$R_{\text{Minimax}} \geq R_{\text{Bayes}}(w) = \int_{\mathcal{X}} \inf_{a \in \mathcal{A}} \left\{ \int_\Theta L(\theta, a) p_\theta(x) dw(\theta) \right\} d\mu(x). \tag{239}$$

We shall see many examples of (239) in the sequel. A simple example is the following where we can use (239) to prove exact minimaxity.

Example 25.7 (Multivariate Normal Model)**.** *Consider the problem of estimating $\theta \in \mathbb{R}^n$ from $X \sim N_n(\theta, I_n)$ under loss*

$$L(\theta, a) = \frac{1}{n} \|\theta - a\|^2.$$

In this case, the parameter space is $\Theta = \mathbb{R}^n$. The estimator $d(X) = X$ has risk equal to 1. It turns out that this is the minimax risk over Θ. To see this, let w to be the normal distribution on $\mathbb{R}^n$ with mean vector μ and covariance matrix $\tau^2 I_n$. The Bayes risk $R_{\text{Bayes}}(w)$ can then be explicitly calculated. To see this, note that the posterior distribution is given by

$$\theta | X \sim N_n \left(\frac{\mu/\tau^2 + X}{1/\tau^2 + 1}, \frac{I_n}{1/\tau^2 + 1} \right)$$

so that

$$R_{\text{Bayes}}(w) = \frac{1}{n} \mathbb{E} \left\| \theta - \frac{\mu/\tau^2 + X}{1/\tau^2 + 1} \right\|^2 = \frac{\tau^2}{\tau^2 + 1}.$$

This gives

$$R_{\text{Minimax}} \geq \frac{\tau^2}{\tau^2 + 1} \qquad \text{for every } \tau > 0.$$

Letting $\tau \to \infty$, we obtain $R_{\text{Minimax}} \geq 1$. Because the supremum risk of X over $\mathbb{R}^n$ is at most 1, this proves that X is minimax.

26 Lecture 26

In the last lecture, we saw the important minimax lower bound:

$$R_{\text{Minimax}} \geq R_{\text{Bayes}}(w) \qquad \text{for every probability measure } w \text{ on } \Theta.$$

In this lecture, we shall look at two non-trivial examples where the above bound can be used to establish sharp asymptotic minimaxity of natural estimators. The first example involves sparse normal mean estimation. The second example involves estimation of a normal mean under a power (L^2 norm) constraint (this result is known as a finite-dimensional Pinsker's theorem).

26.1 Sparse Normal Mean Estimation

Consider the problem of estimating a k-sparse vector $\theta \in \mathbb{R}^n$ from the observation $Y \sim N_n(\theta, I_n)$ under squared error loss. This estimation problem can be put in the decision-theoretic framework with Θ being the set of all k-sparse vectors in $\mathbb{R}^n$, $\mathcal{A} = \mathbb{R}^n$ and $L(\theta, a) = \|\theta - a\|^2$. Also P_θ is the probability measure $N_n(\theta, I_n)$. We shall assume throughout this section that the sparsity level k satisfie $k = o(n)$ (i.e., $k/n \to 0$ as $n \to \infty$).

From our earlier results, we have seen that the LASSO (which is same as soft-thresholding) estimator with tuning parameter $\lambda = \sqrt{2\log(n/k)}$ satisfies

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}_\lambda) \leq (2k \log(n/k)) (1 + o(1)) \qquad \text{as } n \to \infty.$$

We shall now show that this estimator is sharp asymptotically minimax by proving that

$$R_{\text{Minimax}} \geq (2k \log(n/k)) (1 + o(1)) \qquad \text{as } n \to \infty. \tag{240}$$

The argument below is taken from Johnstone [11, Section 8.6].

We first work with $k = 1$ (and then later argue for general $k = o(n)$). For $k = 1$, the parameter Θ is particularly simple and consists of 1-sparse vectors in $\mathbb{R}^n$. Here a natural prior is the uniform prior on the finite parameter set $\{\tau e_1, \ldots, \tau e_n\}$ where e_i is the usual standard unit vector and $\tau \geq 0$ will be chosen (depending on n) appropriately. Let us denote this prior by w.

Let the posterior distribution be denoted by $(p_{1n}(Y), p_{2n}(Y), \ldots, p_{nn}(Y))$ (i.e., $p_{in}(Y)$ is the posterior probability associated with τe_i). It is easy to see that

$$p_{in}(Y) \propto \exp\left(\frac{-\|Y - \tau e_i\|^2}{2}\right) \propto \exp\left(\langle Y, \tau e_i \rangle\right) = \exp(\tau Y_i)$$

and thus

$$p_{in}(Y) = \frac{\exp(\tau Y_i)}{\sum_j \exp(\tau Y_j)} \qquad \text{for } i = 1, \ldots, n.$$

The posterior mean (which is the Bayes estimator) is therefore given by $(\tau p_{1n}(Y), \ldots, \tau p_{nn}(Y))$. The Bayes risk with respect to this prior is thus

$$R_{\text{Bayes}}(w) := \frac{1}{n} \sum_{I=1}^n \sum_{i=1}^n \mathbb{E}_{\tau e_I} \left(\tau p_{in}(Y) - \tau e_{In}(i)\right)^2.$$

By replacing the inner sum above by only the term corresponding to $i = I$, we obtain the lower bound

$$R_{\text{Bayes}}(w) \geq \frac{\tau^2}{n} \sum_{I=1}^n \mathbb{E}_{\tau e_I} \left(p_{In}(Y) - 1\right)^2$$

By symmetry each term above will take the same value and we therefore get

$$R_{\text{Bayes}}(w) \geq \tau^2 \mathbb{E}_{\tau e_1} \left(p_{1n}(Y) - 1\right)^2 = \tau^2 \mathbb{E}_{\tau e_1} \left(\frac{e^{\tau Y_1}}{\sum_{i=1}^n e^{\tau Y_i}} - 1\right)^2.$$

To compute the above expectation (note that τ is not a constant but it changes with n), we switch to standard gaussian random variables $z_1, \ldots, z_n$ by taking $Y_1 = \tau + z_1$ and $Y_i = z_i$ for $i \geq 2$. Then we need to compute:

$$S := \mathbb{E}\left(\frac{e^{\tau^2} e^{\tau z_1}}{\sum_{i=2}^n e^{\tau z_i} + e^{\tau^2} e^{\tau z_1}} - 1\right)^2.$$

We shall show that for

$$\tau_n = \sqrt{2\log n} - \log\left(\sqrt{2\log n}\right),$$

the quantity S converges to 1. This would then imply that

$$R_{\text{Minimax}} \geq R_{\text{Bayes}}(w) \geq \left(\sqrt{2\log n} - \log(\sqrt{2\log n})\right)^2 (1+o(1)) = (2\log n)(1+o(1))$$

which will prove the required lower bound (240) for $k = 1$.

For ease of notation, let us denote $\lambda_n = \sqrt{2\log n}$ so that $\exp(-\lambda_n^2/2) = n^{-1}$. To prove that $S = 1 + o(1)$ for $\tau_n := \lambda_n - \log\lambda_n$, we only need to show that the sequence of random variables

$$A_n := \frac{e^{\tau_n^2} e^{\tau_n z_1}}{e^{\tau_n^2} e^{\tau_n z_1} + \sum_{i=2}^n e^{\tau_n z_i}}$$

converges to 0 in probability. Note that A_n is precisely the posterior probability of τe_1 (and we are working in the case when the truth is τe_1). Thus A_n converging to zero in probability means that the posterior probability of τe_1 (when the truth is τe_1) goes to zero which intuitively means that the spike will be missed.

To prove that $A_n \xrightarrow{P} 0$, rewrite

$$A_n = \frac{1}{1 + e^{-\tau_n^2} e^{-\tau_n z_1} \sum_{i=2}^n e^{\tau_n z_i}} = \frac{1}{1 + V_n W_{n-1}}$$

where

$$V_n = (n-1)e^{-\tau^2/2} e^{-\tau z_1} \quad \text{and} \quad W_{n-1} := \frac{1}{n-1} e^{-\tau^2/2} \sum_{i=2}^n e^{\tau z_i}.$$

It is easy to see that V_n converges to $+\infty$ in probability. To see this, write

$$V_n = \frac{n-1}{n} \exp\left(\log n - \frac{\tau^2}{2} - \tau z_1\right)$$

Because $\lambda = \sqrt{2\log n}$,

$$V_n == \frac{n-1}{n} \exp\left(\frac{\lambda^2}{2} - \frac{\tau^2}{2} - \tau z_1\right)$$

Now

$$\lambda^2 - \tau^2 - 2\tau z_1 \geq \lambda^2 - \tau^2 - (\lambda+\tau) z_1^+ \geq (\lambda - \tau - z_1^+)(\lambda + \tau) \to \infty$$

as $n \to \infty$ (almost surely) because $\lambda_n - \tau_n \to \infty$.

Therefore in order to prove that A_n goes to 0 in probability, we only need to show that W_{n-1} goes to 1 in probability. Reindex and call $n-1$ to be n for simplicity.

$$W_n := \frac{1}{n} e^{-\tau_n^2/2} \sum_{k=1}^n e^{\tau_n z_k}.$$

This is just an average of i.i.d mean one random variables. But τ_n depends on n so we cannot apply the usual weak law of large numbers. We can use however the following version of the weak law.

Theorem 26.1. *For each n, let $X_{nk}, 1 \leq k \leq n$ be independent random variables. Let $b_n > 0$ with $b_n \to \infty$ and let $\tilde{X}_{nk} := X_{nk}\{|X_{nk}| \leq b_n\}$. Suppose that as $n \to \infty$*

1. $\sum_{k=1}^n \mathbb{P}\{|X_{nk}| > b_n\} \to 0$, *and*
2. $b_n^{-2} \sum_{k=1}^n \mathbb{E}\tilde{X}_{nk}^2 \to 0$.

Let $S_n := X_{n1} + \cdots + X_{nn}$ *and put* $a_n := \sum_{k=1}^n \mathbb{E}\tilde{X}_{nk}$. *Then*

$$\frac{S_n - a_n}{b_n} \to 0 \qquad \text{in probability as } n \to \infty.$$

We shall apply the above theorem with $X_{nk} := e^{\tau_n z_k}$ and $b_n = e^{\tau_n \lambda_n}$ (recall $\lambda_n := \sqrt{2\log n}$). The first condition in Theorem 26.1 can be checked as follows. Note that $\{|X_{nk}| \le b_n\} = \{z_k \le \lambda_n\}$ so that

$$\sum_{k=1}^n \mathbb{P}\left\{|X_{nk}| > b_n\right\} = \sum_{k=1}^n \mathbb{P}\left\{z_k > \lambda_n\right\} = n\left(1 - \Phi(\lambda_n)\right) \le \frac{n\phi(\lambda_n)}{\lambda_n} = \frac{1}{\sqrt{2\pi}\lambda_n} \to 0.$$

To verify the second condition in Theorem 26.1, we need to compute $\mathbb{E}\tilde{X}_{nk}^r$ for $r = 2$:

$$\mathbb{E}\tilde{X}_{nk}^r = \mathbb{E}e^{r\tau_n z_k}\left\{z_k \le \lambda_n\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\lambda_n} e^{r\tau_n x}\exp(-x^2/2)dx = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{r^2\tau_n^2}{2}\right)\int_{-\infty}^{\lambda_n}\exp\left(-\frac{1}{2}\left(x - r\tau_n\right)^2\right)dx.$$

Therefore

$$\mathbb{E}\tilde{X}_{nk}^r = e^{r^2\tau_n^2/2}\Phi(\lambda_n - r\tau_n). \tag{241}$$

As a result

$$b_n^{-2}\sum_{k=1}^n \mathbb{E}\tilde{X}_{nk}^2 = \frac{ne^{2\tau_n^2}\Phi(\lambda_n - 2\tau_n)}{e^{2\tau_n\lambda_n}}.$$

Observe that λ_n will be smaller than $2\tau_n$ eventually so that $\lambda_n - 2\tau_n$ will be negative. Therefore

$$\Phi(\lambda_n - 2\tau_n) = 1 - \Phi(2\tau_n - \lambda_n) \le \frac{\phi(2\tau_n - \lambda_n)}{2\tau_n - \lambda_n}.$$

Thus

$$b_n^{-2}\sum_{k=1}^n \mathbb{E}\tilde{X}_{nk}^2 = \frac{ne^{2\tau_n^2}}{e^{2\tau_n\lambda_n}}\frac{\phi(2\tau_n - \lambda_n)}{2\tau_n - \lambda_n} \le \frac{1}{\sqrt{2\pi}(2\tau_n - \lambda_n)} \to 0$$

as $n \to \infty$.

The two conditions of Theorem 26.1 have been verified so we can apply it now. We need to calculate a_n for which we can simply use (241) with $r = 1$. This will give

$$a_n = ne^{\tau_n^2/2}\Phi(\lambda_n - \tau_n) \tag{242}$$

Theorem 26.1 therefore gives

$$\frac{\sum_{k=1}^n e^{\tau_n z_k} - a_n}{e^{\lambda_n\tau_n}} = o_P(1)$$

which is the same as

$$\sum_{k=1}^n e^{\tau_n z_k} = a_n + e^{\lambda_n\tau_n}o_P(1).$$

Therefore

$$W_n = \frac{1}{n}e^{-\tau_n^2/2}\sum_{k=1}^n e^{\tau_n z_k} = \frac{1}{n}e^{-\tau_n^2/2}a_n + \frac{1}{n}e^{-\tau_n^2/2}e^{\lambda_n\tau_n}o_P(1).$$

From (242), we have

$$W_n = \Phi(\lambda_n - \tau_n) + \exp\left(-\frac{1}{2}(\lambda_n + \tau_n)^2\right)o_P(1).$$

Because $\lambda_n - \tau_n \to \infty$ and $\lambda_n + \tau_n \to \infty$, we have

$$W_n = 1 + o_P(1)$$

which is what we wanted to prove. This proves (240) for $k = 1$.

To prove (240) for general $k = o(n)$, the idea is to use an *independent blocks* prior. Divide the indices $\{1, \ldots, n\}$ into k_n blocks each of size $m = m_n = \lfloor n/k_n \rfloor$. On each block, use a single spike prior as in the case of $k = 1$. The overall prior would then make these k_n blocks independent. Because of independence, the Bayes risk adds up and we obtain the overall lower bound of $2k \log(n/k)(1 + o(1))$. We would need the assumption that $k/n \to 0$ because in each block we need the number of observations m_n to go to infinity. This completes the proof of (240).

26.2 Normal Mean Estimation under Power Constraint (Finite-dimensional Pinsker's Theorem)

Consider the problem of estimating a vector $\theta \in \mathbb{R}^n$ from $Y \sim N_n(\theta, I_n)$ in the loss function

$$L(\theta, a) = \frac{1}{n} \|\theta - a\|^2$$

under the following constraint:

$$\frac{1}{n} \sum_{i=1}^{n} \theta_i^2 \leq c^2 \qquad \text{for a fixed } c > 0. \tag{243}$$

Let Θ denote the class of all $\theta \in \mathbb{R}^n$ satisfying the above constraint (the constraint is often referred to as the Power Constraint in signal processing and information theory). Let $\mathcal{A} = \mathbb{R}^n$ and as usual P_θ is the $N_n(\theta, I_n)$ distribution. The risk of an estimator $\hat{\theta}$ is given by

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta L(\theta, \hat{\theta}) = \mathbb{E}_\theta \frac{1}{n} \left\| \hat{\theta} - \theta \right\|^2.$$

What are good candidate estimators for θ under the power constraint (243)? The most natural estimator is the projection of Y onto Θ. This is an M-estimator which can be analyzed by our earlier techniques. In fact, in this case, the projection has the explicit form

$$\hat{\theta} = \begin{cases} \frac{Yc\sqrt{n}}{\|Y\|} & \text{if } \|Y\| > c\sqrt{n} \\ Y & \text{if } \|Y\| \leq c\sqrt{n}. \end{cases}$$

Note that this is a non-linear estimator. It turns out that in this problem, simple linear estimators of the form αY for appropriate $\alpha > 0$ perform very well. This will be demonstrated below. First of all, note that the risk of αY is given by

$$R(\theta, \alpha Y) = (1 - \alpha)^2 \frac{\|\theta\|^2}{n} + \alpha^2.$$

Under the power constraint (243), we have

$$\sup_{\theta \in \Theta} R(\theta, \alpha Y) = (1 - \alpha)^2 c^2 + \alpha^2.$$

The value of α that minimizes the right hand side is

$$\alpha^* = \frac{c^2}{1 + c^2}$$

and its risk is given by

$$\sup_{\theta \in \Theta} R(\theta, \alpha^* Y) = c^2 \left(\frac{1}{1 + c^2} \right)^2 + \left(\frac{c^2}{1 + c^2} \right)^2 = \frac{c^2}{1 + c^2}.$$

It turns out that the linear estimator $\alpha^* Y$ is sharp asymptotically minimax in this problem. To prove this, we shall show below that

$$\liminf_{n\to\infty} R_{\text{Minimax}} \geq \frac{c^2}{1+c^2}. \tag{244}$$

The first step in proving (244) is to choose an appropriate prior w on Θ. The most natural choice for w might seem to be the uniform prior on Θ. However, it is slightly simpler to work with the following prior. Let π denote the $N_n(0, \delta^2 c^2 I_n)$ distribution on $\mathbb{R}^n$ where $\delta \in (0,1)$. We shall take w to be the conditional probability measure under π conditioned to be in Θ i.e.,

$$w(A) := \frac{\pi(A \cap \Theta)}{\pi(\Theta)}$$

for Borel subsets A of $\mathbb{R}^n$. Let $\hat{\theta}_B(w)$ denote the Bayes estimator with respect to the prior w. Note that because w is supported on the convex set Θ, the estimator $\hat{\theta}_B(w)$ will belong to Θ with probability one. We then have

$$\begin{aligned}
R_{\text{Bayes}}(w) &= \int R(\theta, \hat{\theta}_B(w)) dw(\theta) \\
&= \frac{1}{\pi(\Theta)} \int_\Theta R(\theta, \hat{\theta}_B(w)) d\pi(\theta) \\
&= \frac{1}{\pi(\Theta)} \left(\int R(\theta, \hat{\theta}_B(w)) d\pi(\theta) - \int_{\Theta^c} R(\theta, \hat{\theta}_B(w)) d\pi(\theta) \right) \geq \frac{1}{\pi(\Theta)} \left(R_{\text{Bayes}}(\pi) - \int_{\Theta^c} R(\theta, \hat{\theta}_B(w)) d\pi(\theta) \right).
\end{aligned}$$

Because π is the $N_n(0, \delta^2 c^2 I_n)$ prior, its Bayes risk can be easily computed in closed form as

$$R_{\text{Bayes}}(\pi) = \frac{\delta^2 c^2}{1+\delta^2 c^2}$$

so that we obtain

$$R_{\text{Bayes}}(w) \geq \frac{1}{\pi(\Theta)} \left(\frac{\delta^2 c^2}{1+\delta^2 c^2} - \int_{\Theta^c} R(\theta, \hat{\theta}_B(w)) d\pi(\theta) \right).$$

Note now that by the elementary inequality $\|a-b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we have

$$R(\theta, \hat{\theta}_B(w)) = \frac{1}{n} \mathbb{E}_\theta \left\| \hat{\theta}_B(w) - \theta \right\|^2 \leq \frac{2}{n} \left(\mathbb{E}_\theta \left\| \hat{\theta}_B(w) \right\|^2 + \|\theta\|^2 \right) \leq \frac{2}{n} \left(nc^2 + \|\theta\|^2 \right) = 2c^2 + \frac{2\|\theta\|^2}{n}.$$

This gives

$$\int_{\Theta^c} R(\theta, \hat{\theta}_B(w)) d\pi(\theta) \leq 2c^2 \pi(\Theta^c) + \frac{2}{n} \int_{\Theta^c} \|\theta\|^2 d\pi(\theta) \leq 2c^2 \pi(\Theta^c) + \frac{2\sqrt{\pi(\Theta^c)}}{n} \left(\int \|\theta\|^4 d\pi(\theta) \right)^{1/2}$$

by the Cauchy-Schwarz inequality. Now under π,

$$\frac{\|\theta\|^2}{\delta^2 c^2} \sim \chi^2_n$$

so that

$$\int \|\theta\|^4 d\pi(\theta) = \delta^4 c^4 \left(\left(\mathbb{E}\chi^2_n \right)^2 + \text{var}(\chi^2_n) \right) = \delta^4 c^4 \left(n^2 + 2n \right).$$

Putting things together, we obtain

$$\begin{aligned}
R_{\text{Bayes}}(w) &\geq \frac{1}{\pi(\Theta)} \left(\frac{\delta^2 c^2}{1+\delta^2 c^2} - 2c^2 \pi(\Theta^c) - \frac{2\sqrt{\pi(\Theta^c)}}{n} \delta^2 c^2 \sqrt{n^2+2n} \right) \\
&\geq \frac{\delta^2 c^2}{1+\delta^2 c^2} - 2c^2 \pi(\Theta^c) - 2\sqrt{\pi(\Theta^c)} \delta^2 c^2 \sqrt{1+\frac{2}{n}}
\end{aligned}$$

because $\pi(\Theta) \leq 1$. We complete the argument, we just need lower bounds on $\pi(\Theta^c)$. For this, note that

$$\pi(\Theta^c) = \pi\{\|\theta\|^2 > nc^2\} = \mathbb{P}\left\{\frac{\chi^2_n}{n} > \frac{1}{\delta^2}\right\} \leq \mathbb{P}\left\{\left|\frac{\chi^2_n}{n} - 1\right| > \frac{1}{\delta^2} - 1\right\}.$$

Using the chi-squared concentration inequality

$$\mathbb{P}\left\{\left|\frac{\chi^2_n}{n} - 1\right| \geq t\right\} \leq 2\exp\left(\frac{-nt^2}{8}\right) \qquad \text{for } 0 \leq t \leq 1,$$

we obtain

$$\pi(\Theta^c) \leq 2\exp\left(\frac{-n}{8}\left(\frac{1}{\delta^2} - 1\right)^2\right) \qquad \text{when } \frac{1}{2} \leq \delta^2 \leq 1.$$

Thus for $0.5 \leq \delta^2 \leq 1$, we obtain

$$R_{\text{Bayes}}(w) \geq \frac{\delta^2 c^2}{1 + \delta^2 c^2} - 4c^2 \exp\left(\frac{-n}{8}\left(\frac{1}{\delta^2} - 1\right)^2\right) - 2\sqrt{2}\delta^2 c^2 \left(\sqrt{1 + \frac{2}{n}}\right) \exp\left(\frac{-n}{16}\left(\frac{1}{\delta^2} - 1\right)^2\right).$$

As a result, we have

$$\liminf_{n\to\infty} R_{\text{Bayes}}(w) \geq \frac{\delta^2 c^2}{1 + \delta^2 c^2}$$

which implies that

$$\liminf_{n\to\infty} R_{\text{Minimax}} \geq \frac{\delta^2 c^2}{1 + \delta^2 c^2}.$$

Since $\delta^2 \in [0.5, 1]$ here is arbitrary, we can let $\delta^2 \to 1$ to obtain (244). This completes the proof of the sharp asymptotic minimaxity of the linear estimator $\alpha^* Y$.

From the above two examples of sharp asymptotic minimaxity, it should be clear that the key to these arguments is the choice of an appropriate prior w. Also once the prior w is chosen, the argument is usually intricate because we do not even want to lose constant factors in n.

We shall next study arguments for rate minimaxity where it will be okay to lose constant factors while bounding the minimax risk from below. These arguments are much simpler and one uses discrete priors (most often uniform priors on a finite subset of the parameter space). We shall study these (which are related to bounds in multi-hypothesis testing problems) next week.

27 Lecture 27

We shall spend today's lecture on uniform Bayes risk lower bounds in multi-hypotheses testing problems. It turns out (as will be seen in the next lecture) that the minimax risk in general decision theoretic problems can always be bounded from below by this testing risk and this is quite useful for establishing rate minimaxity.

27.1 The Multi-Hypothesis Testing Problem

Suppose we observe data X taking values in a space $\mathcal{X}$ (as usual, X can be a vector, matrix, function etc.). We have the following N hypotheses for the distribution of X:

$$H_1 : X \sim P_1, \quad H_2 : X \sim P_2, \quad \ldots \quad H_N : X \sim P_N.$$

Here $P_1, \ldots, P_N$ are probability measures on $\mathcal{X}$. We need to choose one of these hypotheses based on the observation X. A test T is any function from $\mathcal{X}$ to $\{1, \ldots, N\}$. Given a test T, its type i error is defined

by $P_i\{T \neq i\}$ for $i = 1, \ldots, n$. We shall evaluate tests by the average of their type i errors for $i = 1, \ldots, N$. Specifically, let

$$R(T) := \frac{1}{N} \sum_{i=1}^{N} P_i\{T \neq i\}.$$

One can treat this problem in the general decision-theoretic framework by taking $\Theta = \mathcal{A} = \{1, \ldots, N\}$ and $L(\theta, a) = I\{\theta \neq a\}$. In this case, $R(T)$ will simply be the average risk of the test T averaged with respect to the discrete uniform prior on Θ.

It is easy to see (shown below) that the test T^* which minimizes $R(T)$ is given by the maximum likelihood test. To see this, let p_i denote the density of P_i with respect to a common dominating measure μ. We can then write

$$R(T) = \frac{1}{N} \sum_{i=1}^{N} P_i\{T \neq i\} = 1 - \frac{1}{N} \sum_{i=1}^{N} P_i\{T = i\} = 1 - \frac{1}{N} \int \sum_{i=1}^{N} I\{T(x) = i\} p_i(x) d\mu(x).$$

It is easy to see then that for every test T and $x \in \mathcal{X}$, we have

$$\sum_{i=1}^{n} I\{T(x) = i\} p_i(x) \leq \max_{1 \leq i \leq N} p_i(x)$$

with equality being achieved for the maximum likelihood test defined by $T^*(x) := \text{argmax}_{1 \leq i \leq N} p_i(x)$. This proves that T^* minimizes $R(T)$ and also that

$$B(P_1, \ldots, P_N) := \inf_T R(T) = 1 - \frac{1}{N} \int \max_{1 \leq i \leq N} p_i(x) d\mu(x). \tag{245}$$

It is usually difficult to compute $B(P_1, \ldots, P_N)$ exactly. We shall focus on obtaining lower bounds for $B(P_1, \ldots, P_N)$. As will be seen later, these lower bounds will yield lower bounds on the minimax risk in general decision theoretic problems.

In order to motivate lower bounds for $B := B(P_1, \ldots, P_N)$, let us first provide an intuitive meaning for B. Because B is the smallest possible average error (Bayes risk) in the testing problem, it should be clear that it measures, in some sense, the degree of separation between the probability measures $P_1, \ldots, P_N$. Indeed, if $P_1, \ldots, P_N$ are far from each other, the testing problem should be easier and B will be small. On the other hand, if $P_1, \ldots, P_N$ are close to each other, the testing problem will be harder and B will be large. Note also that we always have

$$1 \leq \int \max_i p_i(x) d\mu(x) \leq N$$

so that

$$0 \leq B(P_1, \ldots, P_N) \leq 1 - \frac{1}{N}.$$

Also, it is easy to see that the $B(P_1, \ldots, P_N)$ takes the maximum possible value $1-(1/N)$ when $P_1 = \cdots = P_N$. This makes sense because when $P_1 = \cdots = P_N$, identifying i based on $X \sim P_i$ is impossible and hence the testing Bayes risk $B(P_1, \ldots, P_N)$ takes its maximum possible value.

On the other hand, $B(P_1, \ldots, P_N)$ takes its minimum value of 0 when $P_1, \ldots, P_N$ are mutually singular (so that $\max_i p_i = p_1 + \ldots p_N$ almost surely w.r.t μ). In this case, one can perfectly identify i based on $X \sim P_i$ so that the testing Bayes risk is at its lowest possible value.

The intuition that $B(P_1, \ldots, P_N)$ measures the degree of separation between $P_1, \ldots, P_N$ suggests that we can bound it via other natural quantities for measuring the degree of separation or spread of $P_1, \ldots, P_N$. For real numbers $a_1 \ldots, a_N$, the most natural way of measuring their spread is their variance:

$$\frac{1}{N} \sum_{i=1}^{N} (a_i - \bar{a})^2.$$

We can try to extend this idea to probability measures by defining

$$I(P_1, \ldots, P_N) := \frac{1}{N} \sum_{i=1}^{N} D(P_i \| \bar{P}) \qquad \text{with } \bar{P} := \frac{1}{N} \sum_{i=1}^{N} P_i. \tag{246}$$

D here refers to a notion of discrepancy/divergence between probability measures (analogous to the squared Euclidean distance between real numbers). Various choices for D are possible but the most common one is the Kullback-Leibler divergence. Given two probability measures P and Q having densities p and q respectively with respect to a common dominating measure μ, the Kullback-Leibler divergence between them is defined as

$$D(P\|Q) := \int p \log \left(\frac{p}{q} \right) d\mu.$$

Based on the above discussion, it should be clear that there should be some connection between $B(P_1, \ldots, P_N)$ and $I(P_1, \ldots, P_N)$ (because both are measuring the spread or degree of separation between $P_1, \ldots, P_N$). The following lemma describes a simple relation between the two. This is often used in the statistics literature to prove Minimax Lower Bounds where it is referred to as Fano's Inequality or Fano's lemma. In fact, this is a weaker form of Fano's inequality; there is a stronger version which we shall describe later today.

Lemma 27.1. *The following inequality holds for every $N \geq 1$ and probability measures $P_1, \ldots, P_N$:*

$$B(P_1, \ldots, P_N) \geq 1 - \frac{\log 2 + I(P_1, \ldots, P_N)}{\log N}. \tag{247}$$

Proof of Lemma 27.1. This elegant proof is due to Kemperman [13, Page 135].

Using the formula (245) for $B(P_1, \ldots, P_N)$ and the definition of $I(P_1, \ldots, P_N)$, it is clear that (247) is equivalent to

$$\frac{1}{N} \int \max_i p_i d\mu \leq \frac{\log 2}{\log N} + \frac{1}{\log N} \frac{1}{N} \sum_{i=1}^{N} \int p_i \log \left(\frac{p_i}{\bar{p}} \right) d\mu$$

It is easy to see that this is further equivalent to (multiplying both sides above by $N \log N$ and using $\int (\sum_i p_i) d\mu = N$),

$$\int (\log N) \max_i p_i d\mu \leq \int (\log 2) \left(\sum_{i=1}^{N} p_i \right) d\mu + \int \sum_{i=1}^{N} p_i \log \left(\frac{p_i}{\bar{p}} \right) d\mu$$

which is identical to

$$\int (\log N) \max_i p_i d\mu \leq \int \sum_{i=1}^{N} p_i \log \left(\frac{2p_i}{\bar{p}} \right) d\mu.$$

To prove this, it is obviously enough to prove the following fact involving nonnegative real numbers. For every set of nonnegative real numbers $a_1, \ldots, a_N$, the following inequality holds:

$$(\log N) \max_{1 \leq i \leq N} a_i \leq \sum_{i=1}^{N} a_i \log \left(\frac{2a_i}{\bar{a}} \right). \tag{248}$$

To prove this inequality, note first that we can assume, without loss of generality, $\sum_{i=1}^{N} a_i = 1$ (so that $a_1, \ldots, a_N$ becomes a probability vector) and that $a_1 = \max_i a_i$. Inequality (248) is then equivalent to

$$(\log N) a_1 \leq \sum_{i=1}^{N} a_i \log \left(\frac{2a_i}{(1/N)} \right) = a_1 \log(2Na_1) + \sum_{i=2}^{N} a_i \log(2Na_i)$$

and this be rearranged as

$$a_1 \log(2a_1) + \sum_{i=2}^{N} a_i \log(2Na_i) \geq 0.$$

The above inequality is true because (note that $2N \geq 2(N-1)$)

$$a_1 \log(2a_1) + \sum_{i=2}^{N} a_i \log\left(\frac{a_i}{1/(2N)}\right) \geq a_1 \log(2a_1) + \sum_{i=2}^{N} a_i \log\left(\frac{a_i}{1/(2(N-1))}\right)$$
$$= D\left((a_1, \ldots, a_N) \| (1/2, 1/(2(N-1)), \ldots, 1/(2(N-1)))\right) \geq 0.$$

This completes the proof of (248) which gives (247). □

27.2 Mutual Information

The quantity $I(P_1, \ldots, P_N)$ (defined in (246)) is known as *Mutual Information*. Mutual Information is a term coming from information theory. It is usually defined for a pair of random variables Y and Z. Formally, the mutual information $I(Y, Z)$ between Y and Z is defined as the Kullback-Leibler divergence between the joint distribution of Y and Z and the product of the marginal distributions of Y and Z. Specifically,

$$I(Y, Z) := D(P_{(Y,Z)} \| P_Y \times P_Z).$$

Consider now two random variables Θ and X such that Θ is uniformly distributed on $\{1, \ldots, N\}$ and the conditional distribution of X given $\Theta = i$ is P_i. Then it is easy to see that

$$D(P_{(\Theta,X)} \| P_\Theta \times P_X) = \frac{1}{N} \sum_{i=1}^{N} D(P_i \| \bar{P}). \tag{249}$$

Thus $I(P_1, \ldots, P_N)$ defined as in (246) is just the mutual information between Θ and X. For this reason, $I(P_1, \ldots, P_N)$ is referred to as the mutual information term. Fano's inequality therefore gives a bound for the Bayes risk in terms of Mutual Information.

The following fact about $I(P_1, \ldots, P_N)$ will be useful in the sequel.

Lemma 27.2. *For every $N \geq 1$ and probability measures $P_1, \ldots, P_N$, we have*

$$I(P_1, \ldots, P_N) = \frac{1}{N} \sum_{i=1}^{N} D(P_i \| \bar{P}) = \inf_Q \sum_{i=1}^{N} D(P_i \| Q) \tag{250}$$

where the infimum is taken over all probability measures Q.

Proof. The proof is simple and based on the following identity:

$$\frac{1}{N} \sum_{i=1}^{N} D(P_i \| Q) = \frac{1}{N} \sum_{i=1}^{N} D(P_i \| \bar{P}) + D(\bar{P} \| Q)$$

which is a consequence of:

$$\log\left(\frac{p_i}{q}\right) = \log\left(\frac{p_i}{\bar{p}}\right) + \log\left(\frac{\bar{p}}{q}\right).$$

Here $\bar{p} = (p_1 + \cdots + p_N)/N$ is the density of $\bar{P}$ with respect to μ and q is the density of Q with respect to μ. □

27.3 Application to Sparse Normal Mean Estimation

Fix $n \geq 1$ and let P_i be the $N_n(\tau e_i, I_n)$ distribution for $i = 1, \ldots, n$. Here τ is a positive real number that depends on n and e_i is the vector with 1 in the i^{th} position and 0 elsewhere. We are interested in $B := B(P_1, \ldots, P_n)$ and how it depends on n and τ. Fano's inequality (specifically inequality (247)) gives

$$B \geq 1 - \frac{\log 2 + I}{\log n} \qquad \text{where } I := I(P_1, \ldots, P_n).$$

To get an explicit bound from here, we need upper bounds for I. Here Lemma 27.2 is very useful because it states that

$$I \leq \frac{1}{N}\sum_{i=1}^{N} D(P_i\|Q)$$

for every probability measure Q. A natural choice for Q which enables explicit computation is $Q = N_n(0, I_n)$. We then obtain (using the fact that $D(N_n(\mu_1, \Sigma)\|N_n(\mu_2, \Sigma)) = (\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)/2$)

$$I \leq \frac{1}{N}\sum_{i=1}^{N} D(N_n(\tau e_i, I_n)\|N_n(0, I_n)) = \frac{\tau^2}{2}.$$

This allows us to deduce that

$$B \geq 1 - \frac{\log 2 + \tau^2/2}{\log n} = 1 - \frac{\log 2}{\log n} - \frac{\tau^2}{2\log n}. \tag{251}$$

This gives interesting corollaries such as:

$$\liminf_{n\to\infty} B \geq \frac{1}{2} \qquad \text{for } \tau = \sqrt{\log n}.$$

However, inequality (251) is not strong enough to yield anything nontrivial when τ is close to $\lambda_n := \sqrt{2\log n}$. For example, when $\tau_n := \lambda_n - \log(\lambda_n)$, then (251) does not give anything useful. However, by a direct calculation (as shown below), it can be shown that

$$\lim_{n\to\infty} B = 1 \qquad \text{when } \tau_n = \lambda_n - \log(\lambda_n). \tag{252}$$

This shows an important weakness of using Fano's inequality to obtain lower bounds for B. To prove (252), first note that

$$B = 1 - \frac{1}{n}\int \max_{1\leq i\leq N} p_i(x)dx \qquad \text{where } p_i(x) = (2\pi)^{-n/2}\exp\left(\frac{-1}{2}\left\|x - \tau e_i\right\|^2\right).$$

From this, we can obtain

$$B = 1 - \frac{1}{n}\int e^{-\tau^2/2}\left(\max_{1\leq i\leq n} e^{\tau x_i}\right)\phi_d(x)dx \qquad \text{where } \phi_d(x) := (2\pi)^{-d/2}\exp\left(-\left\|x\right\|^2/2\right).$$

Thus if $z_1, \ldots, z_n$ are independent standard normal random variables, then

$$B = 1 - \frac{e^{-\tau^2/2}}{n}\mathbb{E}\max_{1\leq i\leq n} e^{\tau z_i}.$$

To further bound this from below, we need to bound $\mathbb{E}\max_i e^{\tau z_i}$ from above which we do in the following way (recall $\lambda_n = \sqrt{2\log n}$):

$$\begin{aligned}
\mathbb{E}\max_{1\leq i\leq n} e^{\tau z_i} &= \mathbb{E}e^{\tau \max_{1\leq i\leq n} z_i} \\
&\leq e^{\tau\lambda} + \mathbb{E}e^{\tau\max_i z_i}\left\{\max_i z_i > \lambda\right\} \\
&\leq e^{\tau\lambda} + \mathbb{E}\sum_{i=1}^{n} e^{\tau z_i} I\{z_i > \lambda\} \\
&= e^{\tau\lambda} + n\mathbb{E}e^{\tau z_1} I\{z_1 > \lambda\} \\
&= e^{\tau\lambda} + n\int_{\lambda}^{\infty} e^{\tau x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\
&= e^{\tau\lambda} + ne^{\tau^2/2}\int_{\lambda}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x-\tau)^2/2}dx = e^{\tau\lambda} + ne^{\tau^2/2}\left(1 - \Phi(\lambda - \tau)\right) \leq e^{\tau\lambda} + ne^{\tau^2/2}\frac{\phi(\lambda-\tau)}{\lambda - \tau}.
\end{aligned}$$

Note that the last inequality above (the Mill's ratio bound) requires that $\tau < \lambda$. We thus get

$$B \geq 1 - \frac{e^{-\tau^2/2}}{n}e^{\tau\lambda} + \frac{(2\pi)^{-1/2}}{\lambda - \tau}\exp\left(\frac{-1}{2}(\lambda - \tau)^2\right) = 1 - \exp\left(\frac{-1}{2}(\lambda - \tau)^2\right) + \frac{(2\pi)^{-1/2}}{\lambda - \tau}\exp\left(\frac{-1}{2}(\lambda - \tau)^2\right).$$

For $\tau = \lambda - \log(\lambda)$, we have $\lambda - \tau \to \infty$ as $n \to \infty$ and then, from the above, we immediately obtain (252).

27.4 Fano's Lemma via the Data Processing Inequality

The Data Processing Inequality is a standard fact about the Kullback-Leibler divergence. It states the following. Suppose P and Q are two probability measures on a space $\mathcal{X}$. Let $\Gamma : \mathcal{X} \to \mathcal{Y}$ be any function. Let $P\Gamma^{-1}$ denote the image of the probability measure P under the map Γ i.e.,

$$P\Gamma^{-1}(A) := P\left\{\Gamma \in A\right\}.$$

Similarly define $Q\Gamma^{-1}$. The Data Processing Inequality then states that

$$D(P\|Q) \geq D\left(P\Gamma^{-1}\|Q\Gamma^{-1}\right).$$

This is true for every pair of probability measures P and Q and every function Γ. We will not give a proof of this fact here (this is standard and can be found in many places). We shall outline a simple proof of Fano's inequality in Lemma 27.1 (actually we shall derive a stronger version of Fano's inequality than in (247)) via the Data Processing Inequality.

Consider the setting of Fano's inequality where we have N probability measures $P_1, \ldots, P_N$ on a space $\mathcal{X}$ having densities $p_1, \ldots, p_N$ respectively with respect to μ. Consider two random variables Θ and X such that Θ is uniformly distributed on $\{1, \ldots, N\}$ and the conditional distribution of X given $\Theta = i$ is P_i. Let $\mathbb{P}$ be the joint distribution of Θ and X. Also let $\mathbb{Q}$ be the joint distribution that is the product of the marginal distributions of Θ and X. We have seen (in (249)) that

$$I(P_1, \ldots, P_N) = D(\mathbb{P}\|\mathbb{Q}).$$

Now fix a test T i.e., T is a function from $\mathcal{X}$ to $\{1, \ldots, N\}$. We will then apply the Data Processing Inequality to the map $\Gamma : \{1, \ldots, N\} \times \mathcal{X} \to \{0, 1\}$ defined by

$$\Gamma(j, x) := I\{T(x) \neq j\} \qquad \text{for } j \in \{1, \ldots, N\} \text{ and } x \in \mathcal{X}.$$

The Data Processing Inequality will then give

$$I(P_1, \ldots, P_N) = D(\mathbb{P}\|\mathbb{Q}) \geq D(\mathbb{P}\Gamma^{-1}\|\mathbb{Q}\Gamma^{-1}) = \mathbb{P}\Gamma^{-1}\{1\} \log \frac{\mathbb{P}\Gamma^{-1}\{1\}}{\mathbb{Q}\Gamma^{-1}\{1\}} + \left(1 - \mathbb{P}\Gamma^{-1}\{1\}\right) \log \frac{1 - \mathbb{P}\Gamma^{-1}\{1\}}{1 - \mathbb{Q}\Gamma^{-1}\{1\}}.$$

It is now easy to see that

$$\mathbb{P}\Gamma^{-1}\{1\} = \frac{1}{N}\sum_{j=1}^{N} P_j\{T(x) \neq j\} = R(T)$$

and

$$\mathbb{Q}\Gamma^{-1}\{1\} = \frac{1}{N}\sum_{j=1}^{N} \bar{P}\{T(x) \neq j\} = \bar{P}\left(\frac{1}{N}\sum_{j=1}^{N}\{T(x) \neq j\}\right) = \bar{P}\left(\frac{N-1}{N}\right) = 1 - \frac{1}{N}.$$

We have therefore proved that for every test T,

$$I(P_1, \ldots, P_N) \geq R(T) \log\left(\frac{R(T)}{1 - (1/N)}\right) + (1 - R(T)) \log\left(\frac{1 - R(T)}{(1/N)}\right).$$

Because this is true for every test T, we can take $T = T^*$ (the maximum likelihood test which minimizes $R(T)$ over all T) so that $R(T^*) = B = B(P_1, \ldots, P_N)$. This will then give

$$I(P_1, \ldots, P_N) \geq B \log\left(\frac{NB}{N-1}\right) + (1 - B) \log\left(N(1 - B)\right). \tag{253}$$

This inequality can be treated as a stronger version of Fano's inequality. It is easy to prove that (253) implies (247). To see this, just note that the right hand side of (253) equals:

$$B \log B + (1-B)\log(1-B) + B \log\left(\frac{N}{N-1}\right) + (1-B)(\log N) \geq -\log 2 + (1-B)(\log N)$$

because $\inf_{x \in (0,1)} (x \log x + (1-x)\log(1-x)) = -\log 2$ and $\log(N/(N-1)) \geq 0$.

An important advantage of this proof of Fano's inequality (via the Data Processing Inequality) is that it generalizes to f-divergences. f-divergences are a general class of divergences between probability measures that include the Kullback-Leibler divergence as a special case. They are defined in the following way. Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. It is then easy to show that the following limits exist (even though they may be $+\infty$. Suppose P and Q are two probability measures on a space $\mathcal{X}$ having densities p and q with respect to a common dominating measure μ. The f-divergence between P and Q is denoted by $D_f(P\|Q)$ and is defined in the following way:

$$D_f(P\|Q) := \int f\left(\frac{p}{q}\right) q d\mu + f'(\infty) P\{q = 0\}.$$

Different choices of f lead to different specific divergences. For example, KL divergence corresponds to $f(x) = x \log x$, total variation distance corresponds to $f(x) = |x-1|/2$, squared Hellinger distance corresponds to $f(x) = 1 - \sqrt{x}$ or $f(x) = (\sqrt{x} - 1)^2/2$, chi-squared divergence corresponds to $f(x) = x^2 - 1$ and so on.

It turns out that the data processing inequality is satisfied for every f-divergence. Using this, it is possible to prove the following generalization of Fano's inequality for every f-divergence:

$$\inf_Q \frac{1}{N} \sum_{i=1}^N D_f(P_i\|Q) \geq D_f\left((B, 1-B)\|(1-(1/N), 1/N)\right).$$

With specific choices for f, this leads to more explicit lower bounds for B. For example, for $f(x) = x^2 - 1$, one obtains

$$B(P_1, \ldots, P_n) \geq 1 - \frac{1}{N} - \sqrt{\frac{1}{N^2} \inf_Q \sum_{i=1}^N \chi^2(P_i\|Q)}$$

where $\chi^2(P\|Q) = D_f(P\|Q)$ for $f(x) = x^2 - 1$. See Gushchin [10] or Guntuboyina [9] and Chen et al. [4] for more details.

28 Lecture 28

We shall study the method of proving rate minimaxity results via Fano's inequality. The first step is to bound from below the minimax risk in a general decision-theoretic problem via the Bayes risk in a testing problem.

28.1 Minimax Lower Bound via Testing

Consider the general decision-theoretic setting with a parameter space Θ, action space $\mathcal{A}$ and nonnegative loss function $L(\theta, a)$. We observe data X whose distribution belongs to the family $\{P_\theta, \theta \in \Theta\}$.

Let F be a finite subset of Θ. We say that F is η-separated for a positive real number η if

$$\inf_{a \in \mathcal{A}} (L(\theta_1, a) + L(\theta_2, a)) \geq \eta \qquad \text{for every } \theta_1, \theta_2 \in F \text{ with } \theta_1 \neq \theta_2.$$

Let w denote the uniform prior on F. The lemma below shows that, when F is η-separated, the Bayes risk $R_{\text{Bayes}}(w)$ is bounded from below by $(\eta/2)$ times the Bayes risk in the testing problem corresponding to the probability measures $P_\theta, \theta \in F$.

Lemma 28.1. *Suppose F is η-separated. Then*

$$R_{\text{Bayes}}(w) \geq \frac{\eta}{2} B\left(\{P_\theta, \theta \in F\}\right). \tag{254}$$

Note here that

$$R_{\text{Bayes}}(w) = \inf_d \frac{1}{|F|} \sum_{\theta \in F} \mathbb{E}_\theta L(\theta, d(X)) \quad \text{and} \quad B\left(\{P_\theta, \theta \in F\}\right) = \inf_T \frac{1}{|F|} \sum_{\theta \in F} P_\theta\{T \neq \theta\}$$

where the infimum is over all decision rules d in $R_{\text{Bayes}}(w)$ and over all tests T (i.e., functions from $\mathcal{X}$ to F) in $B(\{P_\theta, \theta \in F\})$.

Proof of Lemma 28.1. Using $L(\theta, a) \geq (\eta/2) I\{L(\theta, a) \geq \eta/2\}$, we obtain

$$R_{\text{Bayes}}(w) \geq \frac{\eta}{2} \inf_d \frac{1}{|F|} \sum_{\theta \in F} P_\theta \left\{ L(\theta, d(X)) \geq \frac{\eta}{2} \right\}.$$

For each decision rule d, we now associate a test T in the following way. Define $T(X)$ as equal to θ provided there exists a $\theta \in F$ such that $L(\theta, d(X)) < \eta/2$ (note that because F is η-separated, there exists at most one $\theta \in F$ such that $L(\theta, d(X)) < \eta/2$). If there is no such $\theta \in F$, then we take $T(X)$ to be an arbitrary point in F. With this construction, it is easy to see that

$$I\left\{ L(\theta, d(X)) \geq \frac{\eta}{2} \right\} \geq I\left\{\theta \neq T(X)\right\} \qquad \text{for every } \theta \in F.$$

From here, inequality (254) immediately follows. □

In the last class, we proved the following inequality (known as Fano's inequality)

$$B(\{P_\theta, \theta \in F\}) \geq 1 - \frac{\log 2 + I(\{P_\theta, \theta \in F\})}{\log |F|}.$$

Combining this with Lemma 28.1 and the fact that $R_{\text{Minimax}} \geq R_{\text{Bayes}}(w)$ for every prior w, we obtain the following minimax lower bound:

$$R_{\text{Minimax}} \geq \frac{\eta}{2} \left(1 - \frac{\log 2 + I(\{P_\theta, \theta \in F\})}{\log |F|} \right) \qquad \text{for every finite } F \subseteq \Theta. \tag{255}$$

We shall see two examples of this bound below: to sparse normal mean estimation and Lipschitz regression. The main challenge in using (255) is to make an appropriate choice of F.

In many applications, $\Theta \subseteq \mathcal{A}$ and $L(\theta, a) = d^2(\theta, a)$ for some pseudometric d on $\mathcal{A}$. In this case, note that

$$\min_{\theta_1, \theta_2 \in F : \theta_1 \neq \theta_2} d(\theta_1, \theta_2) \geq \tau \implies F \text{ is } (\tau^2/2)\text{-separated}. \tag{256}$$

This is because for every $\theta_1, \theta_2 \in F$ with $\theta_1 \neq \theta_2$ and $a \in \mathcal{A}$, we have

$$L(\theta_1, a) + L(\theta_2, a) = d^2(\theta_1, a) + d^2(\theta_2, a) \geq \frac{1}{2} \left(d(\theta_1, a) + d(\theta_2, a) \right)^2 \geq \frac{1}{2} d^2(\theta_1, \theta_2) \geq \frac{\tau^2}{2}.$$

28.2 Sparse Normal Mean Estimation

Consider the problem of estimating a 1-sparse vector $\theta \in \mathbb{R}^n$ in squared Euclidean loss from $Y \sim N_n(\theta, I_n)$. Here Θ is the class of all 1-sparse vectors in $\mathbb{R}^n$, $\mathcal{A} = \mathbb{R}^n$ and $L(\theta, a)$ is the squared Euclidean distance between θ and a. Also P_θ is the $N_n(\theta, I_n)$ distribution.

It is natural here to apply (255) with $F = \{\tau e_1, \ldots, \tau e_n\}$ for some $\tau > 0$ (chosen later). Because $\|\tau e_i - \tau e_j\| = \tau\sqrt{2}$ for every $i \neq j$, it follows that F is η separated with $\eta = \tau^2$ (using (256)). Inequality (255) then gives

$$R_{\text{Minimax}} \geq \frac{\tau^2}{2}\left(1 - \frac{\log 2 + I(\{P_{\theta,\theta\in F}\})}{\log n}\right) \qquad \text{where } I = I(P_\theta, \theta \in F).$$

In the last class, we saw that

$$I \leq \frac{1}{n}\sum_{\theta\in F} D(P_\theta \| P_0) \leq \tau^2$$

and this gives

$$R_{\text{Minimax}} \geq \frac{\tau^2}{2}\left(1 - \frac{\log 2 + (\tau^2/2)}{\log n}\right).$$

Taking $\tau^2 = \log n$ will give that $R_{\text{Minimax}} \geq c \log n$ for a positive constant c (for n large). This result is good enough to yield rate minimaxity of soft thresholding with $\lambda = \sqrt{2\log n}$. However it is not strong enough to yield sharp asymptotic minimaxity.

28.3 Lipschitz Regression

Let $\mathcal{F}$ denote the class of all functions $f : [0, .1] \to \mathbb{R}$ that are bounded in absolute value by 1 and 1-Lipschitz. Consider the problem of estimating $f \in \mathcal{F}$ from i.i.d observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ where

$$X_i \sim \text{Unif}[0, 1] \quad \text{and} \quad Y_i | X_i \sim N(f(X_i), 1).$$

We take $\Theta = \mathcal{F}$, $\mathcal{A}$ to be the class of all real-valued functions on $[0, 1]$ and use the loss function

$$L(f, g) := \int_0^1 (f(x) - g(x))^2 \, dx.$$

We shall denote by P_f the joint distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Note that P_f has the following density on $[0, 1]^n \times \mathbb{R}^n$:

$$p_f((x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y_i - f(x_i))^2}{2}\right).$$

We are interested in the minimax risk:

$$R_{\text{Minimax}} := \inf_{\hat{f}} \sup_{f\in\mathcal{F}} \mathbb{E}_f \int_0^1 \left(f - \hat{f}\right)^2.$$

It can be shown that $R_{\text{Minimax}} \leq Cn^{-2/3}$ for a universal positive constant C. This can be done by studying the least sqaures estimator over $\mathcal{F}$ (using methods that we looked at previously in the class). One can also consider simpler kernel regression estimators (see, for example, Tsybakov [23, Chapter 1]). Here we shall prove (using (255)) that $R_{\text{Minimax}} \geq cn^{-2/3}$ for a positive constant c. This will prove, in particular, that the least squares estimator is minimax rate optimal for estimating functions in $\mathcal{F}$.

The main challenge is to construct a suitable finite subset F of $\mathcal{F}$. The standard construction is as follows. Fix a small $\delta > 0$. For a closed subinterval I of $[0, 1]$ of length δ, let $T_I : I \to [0, \delta]$ denote the piecewise linear *tent* function which equals its maximum value δ at the midpoint of the interval I (specifically T_I linearly

increases from 0 at the left end point of I to δ at the midpoint of I and then linearly decreases to 0 at the right end point of I). Now consider the m intervals:

$$I_j := [(j-1)\delta, j\delta] \qquad \text{for } j = 1, \ldots, m \text{ with } m := \lfloor 1/\delta \rfloor \gtrsim 1/\delta.$$

We now construct 2^m functions in $\mathcal{F}$. These functions will be indexed by $\tau \in \{0,1\}^m$ and will be denoted by $\{f_\tau, \tau \in \{0,1\}^m\}$. Specifically, for each $\tau \in \{0,1\}^m$, we define f_τ to equal the tent function T_{I_j} on the interval I_j if $\tau_j = 1$ and to equal zero on I_j if $\tau_j = 0$. Also each f_τ will equal zero outside $\cup_j I_j$.

We shall apply (255) with this collection $\{P_{f_\tau} : \tau \in \{0,1\}^m\}$. The first step is find a suitable value η for which $F := \{f_\tau, \tau \in \{0,1\}^m\}$ is η-separated. For this, fix $\tau, \tau' \in \{0,1\}^m$ with $\tau \neq \tau'$. Then $\tau_j \neq \tau'_j$ for some $j \in \{1, \ldots, m\}$ and then it is easy to see that

$$\int (f_\tau - f_{\tau'})^2 \geq \int_{I_j} T^2_{I_j}(x)dx \gtrsim \delta^2 \times \delta = \delta^3.$$

It follows (from (256)) that F is η-separated with $\eta \gtrsim \delta^3$. Inequality (255) then says that

$$R_{\text{Minimax}} \geq \eta \left(1 - \frac{\log 2 + I}{\log(2^m)}\right) \gtrsim \delta^3 \left(1 - \frac{\log 2 + I}{m \log 2}\right) \qquad \text{with } I := I(P_{f_\tau}, \tau \in \{0,1\}^m).$$

We next need to bound I from above. For this, we shall use

$$I \leq \frac{1}{2^m} \sum_{\tau \in \{0,1\}^m} D(P_{f_\tau} \| Q)$$

with $Q = P_0$ (i.e., P_f corresponding to $f \equiv 0$). Note that for two functions f, g,

$$D(P_f \| P_g) = nD\left(P_{(X_1,Y_1)} \| P_{(\tilde{X}_1, \tilde{Y}_1)}\right)$$

where

$$X_1 \sim \text{Unif}[0,1];\ Y_1|X_1 \sim N(f(X_1), 1) \quad \text{and} \quad \tilde{X}_1 \sim \text{Unif}[0,1];\ \tilde{Y}_1|\tilde{X}_1 \sim N(g(\tilde{X}_1), 1).$$

One can then also compute that

$$D\left(P_{(X_1,Y_1)} \| P_{(\tilde{X}_1, \tilde{Y}_1)}\right) = \int_0^1 (f-g)^2.$$

We thus have

$$D(P_{f_\tau}, P_0) = n \int_0^1 f^2_\tau(x)dx \leq n\delta^2$$

and consequently $I \leq n\delta^2$. We thus have

$$R_{\text{Minimax}} \gtrsim \delta^3 \left(1 - \frac{\log 2 + n\delta^2}{m \log 2}\right).$$

Because $m \geq c_1/\delta$ for a positive constant c_1, we have

$$R_{\text{Minimax}} \gtrsim \delta^3 \left(1 - \frac{\log 2 + n\delta^2}{(c_1 \log 2)/\delta}\right) = \delta^3 \left(1 - \frac{\delta}{c_1} - \frac{n\delta^3}{c_1 \log 2}\right).$$

Taking $\delta^3 = (c_1 \log 2)/(2n)$, we obtain $R_{\text{Minimax}} \gtrsim n^{-1}$ for all large n. Note however that we set out to prove $R_{\text{Minimax}} \gtrsim n^{-2/3}$ so the above argument yields a suboptimal lower bound for R_{Minimax}. The reason where this argument becomes weak is in the separation calculation. Indeed, we used the fact that

$$\int_0^1 (f_\tau - f_{\tau'})^2 \gtrsim \delta^3 \qquad \text{for every } \tau, \tau' \in \{0,1\}^m \text{ with } \tau \neq \tau'.$$

It is also easy to see that the above bound is tight up to a constant multiplicative factor when τ and τ' differ in only one coordinate (i.e., $H(\tau,\tau') := \sum_j I\{\tau_j \neq \tau'_j\}$ equals exactly 1). However, in general, the L^2 distance between f_τ and $f_{\tau'}$ depends on $H(\tau,\tau')$. Precisely, it is easily seen that

$$\int_0^1 (f_\tau - f_{\tau'})^2 \gtrsim \delta^3 H(\tau,\tau') \qquad \text{for all } \tau,\tau' \in \{0,1\}^m. \tag{257}$$

The Gilbert-Varshamov Lemma (stated and proved next) proves the existence of a subset W of $\{0,1\}^m$ with cardinality $|W| \geq \exp(m/8)$ and such that $H(\tau,\tau') > m/4$ for every $\tau,\tau' \in W$ with $\tau \neq \tau'$. The idea then is to apply (255) to $F = \{f_\tau : \tau \in W\}$ as opposed to $F = \{f_\tau : \tau \in \{0,1\}^m\}$. The inequality (257) above along with $H(\tau,\tau') \gtrsim m \gtrsim (1/\delta)$ for $\tau,\tau' \in W$ with $\tau \neq \tau'$ implies then that $F = \{f_\tau : \tau \in W\}$ is η-separated with $\eta \gtrsim \delta^2$. The mutual information bound remains the same as before. We would then obtain

$$R_{\text{Minimax}} \gtrsim \delta^2 \left(1 - \frac{\delta}{c_1} - \frac{n\delta^3}{c_1 \log 2}\right).$$

The choice $\delta^3 = (c_1 \log 2)/(2n)$ would then give $R_{\text{Minimax}} \gtrsim n^{-2/3}$ for all large n.

28.4 Gilbert-Varshamov Lemma

Lemma 28.2. *For every $m \geq 1$, there exists a subset W of $\{0,1\}^m$ with cardinality $|W| \geq \exp(m/8)$ such that $H(\tau,\tau') > m/4$ for every $\tau,\tau' \in W$ with $\tau \neq \tau'$.*

Proof. The following elementary probability bound will be used here:

$$\mathbb{P}\{Bin(m,1/2) \leq m/4\} = \mathbb{P}\{Bin(m,1/2) \geq 3m/4\} \leq \exp\left(\frac{-m}{8}\right). \tag{258}$$

To prove (258), note that (the first equality follows by symmetry)

$$\mathbb{P}\{Bin(m,1/2) \geq 3m/4\} \leq \inf_{\lambda>0} e^{-3m\lambda/4}\mathbb{E}e^{\lambda Bin(m,1/2)} = \inf_{\lambda>0} e^{-3m\lambda/4}\left(\frac{1}{2} + \frac{1}{2}e^{\lambda}\right)^m.$$

Taking $\lambda = \log 3$, we get

$$\mathbb{P}\{Bin(m,1/2) \geq 3m/4\} \leq 3^{-3m/4}2^m = \exp\left(m\log 2 - \frac{3\log 3}{4}m\right) = \exp(-0.130812m) \leq \exp(-m/8).$$

Now let W be a maximal subset of $\{0,1\}^m$ for which $H(\tau,\tau') > m/4$ for every $\tau,\tau' \in W$ with $\tau \neq \tau'$. Maximal here means that the separation condition will be violated if any other element of $\{0,1\}^m$ is added to W. This implies then that

$$\bigcup_{\tau \in W} B_H(\tau, m/4) = \{0,1\}^m \qquad \text{where } B_H(\tau,m/4) := \{\omega \in \{0,1\}^m : H(\tau,\omega) \leq m/4\}$$

so that

$$2^m = \sum_{\tau \in W} |B_H(\tau,m/4)| \leq |W| \max_{\tau \in W} |B_H(\tau,m/4)|. \tag{259}$$

Now for every $A \subseteq \{0,1\}^m$, we have

$$|A| = 2^m \mathbb{P}\{(T_1,\ldots,T_m) \in A\} \qquad \text{where } T_1,\ldots,T_m \text{ are i.i.d } Ber(1/2).$$

Thus

$$\begin{aligned} 2^{-m}|B_H(\tau,m/4)| &= \mathbb{P}\{(T_1,\ldots,T_m) \in B_H(\tau,m/4)\} \\ &= \mathbb{P}\left\{\sum_{i=1}^m I\{T_i \neq \tau_i\} \leq m/4\right\} = \mathbb{P}\{Bin(m,1/2) \leq m/4\} \leq \exp(-m/8). \end{aligned}$$

Inequality (259) then immediately gives $|W| \geq \exp(m/8)$ which completes the proof of Lemma 28.2. □

28.5 Yang-Barron Method for Avoiding Explicit Construction of F

As we mentioned earlier, the main difficulty in applying (255) is the construction of a finite subset F of Θ. Yang and Barron [26] had a nice idea of avoiding the explicit construction of F provided results on packing and covering numbers of Θ are available. Here are the details behind this idea.

For $\eta > 0$, suppose $N(\eta, \Theta)$ is any positive real number such that there exists an η-separated finite subset F of Θ with cardinality $|F| \geq N(\eta, \Theta)$. Applying inequality (255) to such an F, we get

$$R_{\text{Minimax}} \geq \frac{\eta}{2}\left(1 - \frac{\log 2 + I(\{P_\theta, \theta \in F\})}{\log N(\eta, \Theta)}\right). \tag{260}$$

We now bound $I = I(\{P_\theta, \theta \in F\})$ from above in the following way. We know that

$$I \leq \frac{1}{|F|}\sum_{\theta \in F} D(P_\theta \| Q) \tag{261}$$

for every probability measure Q on $\mathcal{X}$. Suppose now that $Q_1, \ldots, Q_M$ are arbitrary probability measures on $\mathcal{X}$ and apply (261) with $Q = \bar{Q} = (Q_1 + \cdots + Q_M)/M$. This gives the bound

$$I \leq \frac{1}{|F|}\sum_{\theta \in F} D(P_\theta \| \bar{Q}).$$

Now for each $\theta \in F$, if $q_1, \ldots, q_M$ denote the densities of $Q_1, \ldots, Q_M$ w.r.t μ respectively (and p_θ denote the density of P_θ w.r.t μ), then

$$D(P_\theta \| \bar{Q}) = \int p_\theta \log \frac{p_\theta}{(q_1 + \cdots + q_M)/M} d\mu = \int p_\theta \log \frac{p_\theta}{q_1 + \ldots q_M} d\mu + \log M.$$

Now for every $1 \leq j \leq M$, we have $q_1 + \cdots + q_M \geq q_j$ so that

$$D(P_\theta \| \bar{Q}) \leq \int p_\theta \log \frac{p_\theta}{q_j} d\mu + \log M = D(P_\theta \| Q_j) + \log M.$$

Since this is true for every $1 \leq j \leq M$, we deduce

$$D(P_\theta \| \bar{Q}) \leq \min_{1 \leq j \leq M} D(P_\theta \| Q_j) + \log M.$$

Since this is true for every $\theta \in F$, we obtain

$$I \leq \frac{1}{|F|}\sum_{\theta \in F} D(P_\theta \| \bar{Q}) \leq \frac{1}{|F|}\sum_{\theta \in F} \min_{1 \leq j \leq M} D(P_\theta \| Q_j) + \log M \leq \sup_{\theta \in F} \min_{1 \leq j \leq M} D(P_\theta \| Q_j) + \log M.$$

Now for $\epsilon > 0$ and a subset S of Θ, let $M(\epsilon, S)$ denote the minimal number M of probability measures $Q_1, \ldots, Q_M$ on $\mathcal{X}$ such that

$$\sup_{\theta \in S} \min_{1 \leq j \leq M} D(P_\theta \| Q_j) \leq \epsilon^2.$$

The above argument then gives

$$I = I(\{P_\theta, \theta \in F\}) \leq \epsilon^2 + \log M(\epsilon, F) \qquad \text{for every } \epsilon > 0.$$

Using this bound in (260), we obtain that for every $\eta > 0$ and $\epsilon > 0$,

$$R_{\text{Minimax}} \geq \frac{\eta}{2}\left(1 - \frac{\log 2 + \log M(\epsilon, F) + \epsilon^2}{\log N(\eta, \Theta)}\right).$$

The inequality $M(\epsilon, F) \leq M(\epsilon, \Theta)$ (this is only useful if $M(\epsilon, \Theta) < \infty$) then gives

$$R_{\text{Minimax}} \geq \frac{\eta}{2}\left(1 - \frac{\log 2 + \log M(\epsilon, \Theta) + \epsilon^2}{\log N(\eta, \Theta)}\right). \tag{262}$$

The advantage with this bound is that it only depends on properties of Θ. For example, in the Lipschitz regression example, it is known (Lecture 6) that the packing numbers of $\mathcal{F}$ (here $\mathcal{F}$ is the class of all real-valued functions on $[0,1]$ that are 1-Lipschitz and bounded by 1) under the L^2 metric satisfy:

$$\exp\left(C_2\epsilon^{-1}\right) \leq M(\epsilon, \mathcal{F}, L_2[0,1]) \leq \exp\left(C_1\epsilon^{-1}\right).$$

Using this, it is easy to show that we can take

$$\log N(\eta, \Theta) = \frac{c_2}{\sqrt{\eta}} \quad \text{and} \quad \log M(\epsilon, \Theta) \leq c_1 \frac{\sqrt{n}}{\epsilon}$$

in (262). This gives

$$R_{\text{Minimax}} \geq \frac{\eta}{2}\left(1 - \frac{\log 2 + c_1\sqrt{n}/\epsilon + \epsilon^2}{c_2/\sqrt{\eta}}\right).$$

From here taking $\epsilon \sim n^{1/6}$ and $\eta \sim n^{-2/3}$ (and adjusting the underlying constants appropriately), we can immediately derive $R_{\text{Minimax}} \gtrsim n^{-2/3}$. Note that no explicit construction of a finite subset F has been used in this argument (that work is implicitly done in the proof of the packing number bounds).

More generally, recall the smoothness class $\mathcal{S}_{d,\alpha}$ from Lecture 6. Consider the estimation of a function $f \in \mathcal{S}_{d,\alpha}$ from n i.i.d observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ with

$$X_i \sim \text{Unif}[0,1] \quad \text{and} \quad Y_i | X_i \sim N(f(X_i), 1).$$

Consider the integral L^2 loss function on $[0,1]^d$:

$$L(f,g) = \int_{[0,1]^d} (f(x) - g(x))^2 \, dx.$$

In this case, using the fact that the packing numbers of $\mathcal{S}_{d,\alpha}$ satisfy (stated in Lecture 6):

$$\exp\left(C_2\epsilon^{-d/\alpha}\right) \leq M(\epsilon, \mathcal{S}_{d,\alpha}, L_2[0,1]^d) \leq \exp\left(C_1\epsilon^{-d/\alpha}\right),$$

we can take (the constants here all depend on d)

$$\log N(\eta, \Theta) = c_2\left(\frac{1}{\sqrt{\eta}}\right)^{d/\alpha} \quad \text{and} \quad \log M(\epsilon, \Theta) \leq c_1\left(\frac{\sqrt{n}}{\epsilon}\right)^{d/\alpha}$$

in (262). This gives

$$R_{\text{Minimax}} \geq \frac{\eta}{2}\left(1 - \frac{\log 2 + c_1(\sqrt{n}/\epsilon)^{d/\alpha} + \epsilon^2}{c_2(1/\sqrt{\eta})^{d/\alpha}}\right).$$

Taking $\epsilon \sim n^{d/(2(2\alpha+d))}$ and $\eta = n^{-2\alpha/(2\alpha+d)}$, we obtain that

$$R_{\text{Minimax}} \geq n^{-2\alpha/(2\alpha+d)}.$$

References

[1] Baraniuk, R., M. Davenport, R. DeVore, and M. Wakin (2008). A simple proof of the restricted isometry property for random matrices. *Constr. Approx. 28*(3), 253–263.

[2] Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

[3] Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

[4] Chen, X., A. Guntuboyina, and Y. Zhang (2016). On Bayes risk lower bounds. *Journal of Machine Learning Research 17*(219), 1–58.

[5] Dudley, R. M. (1989). *Real Analysis and Probability*. Belmont, Calif: Wadsworth.

[6] Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

[7] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume 1. New York: Wiley.

[8] Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Boston: Academic Press.

[9] Guntuboyina, A. (2011). Lower bounds for the minimax risk using f divergences, and applications. *IEEE Transactions on Information Theory 57*, 2386–2399.

[10] Gushchin, A. A. (2003). On Fano's lemma and similar inequalities for the minimax risk. *Theor. Probability and Math. Statist. 67*, 29–41.

[11] Johnstone, I. M. (2017). Gaussian estimation: Sequence and wavelet models. *Manuscript, August 2017*. available at `http://statweb.stanford.edu/~imj/GE_08_09_17.pdf`.

[12] Kato, K. (2017). Empirical process theory. Lecture notes available at `https://sites.google.com/site/kkatostat/home/research`.

[13] Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. In *Probability and Information Theory*. Springer-Verlag. Lecture Notes in Mathematics, 89, pages 126–169.

[14] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

[15] Le Cam, L. and G. L. Yang (2000). *Asymptotics in Statistics: Some Basic Concepts* (2nd ed.). Springer-Verlag.

[16] Mendelson, S. and R. Vershynin (2003). Entropy and the combinatorial dimension. *Inventiones mathematicae 152*, 37–55.

[17] Oymak, S. and B. Hassibi (2013). Sharp MSE bounds for proximal denoising. *Foundations of Computational Mathematics*, 1–65.

[18] Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

[19] Pollard, D. (1997). Another look at differentiability in quadratic mean. In D. Pollard, E. Torgersen, and G. L. Yang (Eds.), *A Festschrift for Lucien Le Cam*, pp. 305–314. New York: Springer-Verlag.

[20] Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

[21] Rudelson, M. and R. Vershynin (2006). Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, 603–648.

[22] Talagrand, M. (1996). A new look at independence. *Annals of Probability 24*, 1–34.

[23] Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer-Verlag.

[24] Van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

[25] Van der Vaart, A. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

[26] Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.