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

Lecturer: A Rakhlin — Lectures 14-26
Scribe: A. Rakhlin — Spring 2020

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

First, we present a couple of motivating examples.

1. KOLMOGOROV'S GOODNESS-OF-FIT TEST

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

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

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

be the cdf of X, and let

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

be the empirical cdf obtained from n examples. The Glivenko-Cantelli Theorem (1933) states that

$$D_n = \sup_{\theta \in \mathbb{R}} |F_n(\theta) - F(\theta)| \to 0 \quad a.s.$$

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

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

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

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

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

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

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

2. STATISTICAL LEARNING

2.1 Empirical Risk Minimization

Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be n i.i.d. copies of a random variable (X, Y) with distribution $P = P_X \times P_{Y|X}$, where the X variable lives in some abstract space $\mathcal{X}$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$. Fix a loss function $\boldsymbol{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Fix a class of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$. Given the dataset S, the empirical risk minimization (ERM) method is defined as

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

Examples:

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

We now define expected loss (error) as

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

and empirical loss (error) as

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

For any $f^* \in \mathcal{F}$, The decomposition

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

holds true. By definition of ERM, the second term is nonpositive. If f^* is independent of the random sample, the third term is a difference between an average of random variables $\boldsymbol{\ell}(f^*(X_i), Y_i)$ and their expectation. Hence, this term is zero-mean, and its fluctuations can be controlled with the tail bounds we have seen in class. The first term, however, is not zero in expectation (why?).

Let us proceed by taking expectation (with respect to S) of both sides:

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \mathbf{L}(f^*) \le \mathbb{E}\left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] \le \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] \tag{2.1}$$

Here we "removed the hat" on $\widehat{f}$ by "supping out" this data-dependent choice. We are only using the knowledge that $f \in \mathcal{F}$, and nothing else about the method. We will see later that for "curved" loss functions, such as square loss, the supremum can be further localized within $\mathcal{F}$.

2.2 Classification

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

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

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

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]. \tag{2.2}$$

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

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

Substituting this choice, (2.2) becomes

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[P(X \le \theta) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \le \theta\}\right] = \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[F(\theta) - F_n(\theta)\right]. \tag{2.3}$$

which is precisely the quantity from the beginning of the lecture (albeit without absolute values and in expectation). Again, (2.3) is the expected largest pointwise (and one-sided) distance between the CDF and empirical CDF. Does it go to zero as $n \to \infty$? How fast?

Let's introduce the shorthand

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

$\{U_\theta\}_{\theta\in\mathbb{R}}$ is an uncountable collection of *correlated* random variables, so how does the maximum behave? We have already encountered the question (e.g. Lecture 5) in the context of linear forms $\langle X, \theta\rangle$, indexed by $\theta \in \mathsf{B}_2$ and we were able to use a covering argument to control the expected supremum. Recall the key step in that proof: we can introduce a cover $\theta_1, \dots, \theta_N$ such that control of $\sup U_\theta$ can be reduced to control of $\max_{j=1,\dots,N} U_{\theta_i}$. Does this idea work here? Problems with this approach start appearing immediately: how do we cover $\mathbb{R}$ by a finite collection?

We will now present two approaches for upper-bounding (2.3); both extend to the general case of (2.2).

2.2.1 The bracketing approach

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

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

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

and thus

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

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

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

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

2.2.2 The symmetrization approach

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

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

where X' is an independent copy of X.

Observe that

$$\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{X'_i \le \theta\right\}\right]$$

where $X_1', \dots, X_n'$ are n independent copies of X. We have the following upper bound on (2.3):

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.4}$$

$$\le \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.5}$$

by convexity of the sup. Now, since distribution of $\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}$ is the same as the distribution of $-\left(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}\right)$, we can insert arbitrary signs ϵ_i without changing the expected value:

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\})\right]. \tag{2.6}$$

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

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.7}$$

$$\le \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\})\right], \tag{2.8}$$

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

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i' \le \theta\right\}\right] + \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}-\epsilon_i\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.9}$$

$$= 2\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.10}$$

by symmetry of Rademacher random variables.

Now comes the key step. Let us condition on $X_1, \dots, X_n$ and think of the random variables

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

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

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

Since each V_θ is 1-subGaussian, and we get an overall upper bound

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

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

2.3 Discussion

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

In view of (2.1), the upper bounds we derived guarantee (modulo the fact that we omitted "$1 - 2Y$") that for empirical risk minimization,

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

It is worth stating the symmetrization lemma more formally:

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

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

Furthermore,

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

Proof. We only prove the second part since the first statement was proved earlier (for indicators). Write

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

Consider the first term on the RHS:

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

As for the second term,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathbb{E} f \right] \le \sup_{f \in \mathcal{F}} |\mathbb{E} f| \cdot \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right| \tag{2.11}$$

□

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

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

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

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

This justifies omitting $(1 - 2Y)$ for binary classification, at least with the symmetrization approach.

2.4 Empirical Process

Let us also define an empirical process:

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

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

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

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

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

Second, empirical process theory often employs the notation

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

where $\mathbb{P}$ is the distribution of X and $\mathbb{P}_n$ is the empirical measure. You may also see the notation

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

3. SUPREMA OF GAUSSIAN AND SUBGAUSSIAN PROCESSES

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

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

We will be interested in the behavior of the supremum of the stochastic process, and in particular

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

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

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

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

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

3.1 SubGaussian Processes

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

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

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

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

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

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

Rademacher process: Let $R_\theta = \langle \epsilon, \theta\rangle$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, ϵ i.i.d. Rademacher. Again, take $d(\theta, \theta') = \|\theta - \theta'\|$. Then

$$R_\theta - R'_\theta = \langle \epsilon, \theta - \theta'\rangle$$

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

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

Definition: We will call $\widehat{\mathcal{R}}(\Theta) = \mathbb{E}\sup_{\theta\in\Theta}\langle \epsilon, \theta\rangle$ the (empirical) Rademacher averages of Θ. The corresponding expected supremum of the Gaussian process will be called the Gaussian averages or the Gaussian width of Θ and denoted by $\widehat{\mathcal{G}}(\Theta)$.

3.1.1 A few examples

Let $U_\theta = \langle \epsilon, \theta\rangle$, $\Theta \subset \mathbb{R}^n$, and take Euclidean distance as the metric. We have

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

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

A few other sets:

$$\widehat{\mathcal{R}}(\mathsf{B}^n_2) = \mathbb{E} \sup_{\theta\in\mathsf{B}^n_2} \langle \epsilon, \theta\rangle = \mathbb{E}\,\|\epsilon\|_2 = \sqrt{n}$$

and

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

However, we observe that

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

and yet for the Gaussian process,

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

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

$$\widehat{\mathcal{R}}(\Theta) \lesssim \widehat{\mathcal{G}}(\Theta) \lesssim \sqrt{\log n} \cdot \widehat{\mathcal{R}}(\Theta). \tag{3.12}$$

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

For a singleton,

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

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

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

Some further properties of both Rademacher and Gaussian averages:

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

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

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

$$\mathbb{E} \max_{(\theta, \theta') \in A} U_\theta - U_{\theta'} \leq \max_{(\theta, \theta') \in A} d(\theta, \theta') \cdot \sqrt{2 \log \operatorname{card}(A)} \tag{3.13}$$

How do we go beyond finite cover?

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

As a simple consequence,

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

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

Proof. Observe that

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

Let $\widehat{\Theta}$ be a δ-cover of Θ. Then

$$\begin{aligned} U_\theta - U_{\theta'} &= U_\theta - U_{\hat\theta} + U_{\hat\theta} - U_{\hat\theta'} + U_{\hat\theta'} - U_{\theta'} && (3.14)\\ &\leq 2\sup_{d(\theta,\theta')\leq\delta}(U_\theta - U_{\theta'}) + \sup_{\hat\theta,\hat\theta'\in\widehat\Theta}(U_{\hat\theta} - U_{\hat\theta'}) && (3.15)\end{aligned}$$

The last term is

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

□

3.3 Example: Rademacher/Gaussian processes

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

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

Hence,

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta \leq 2\delta\sqrt{n} + 2\mathrm{diam}(\Theta)\sqrt{\log\mathcal{N}(\Theta, \|\cdot\|_2, \delta)} \tag{3.16}$$

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

Let's step back and ask what kind of generic statement we can say about a d-dimensional subset of a Euclidean ball. Suppose that $\Theta\subseteq \mathsf{B}_2^n$ and assume that Θ lives in a d-dimensional subspace. Then

$$\mathcal{N}(\Theta, \|\cdot\|_2, \delta) \leq \left(1 + \frac{2}{\delta}\right)^d$$

and by taking $\delta = \sqrt{d/n}$ the estimate in (3.16) becomes

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta \leq 2\sqrt{d} + 4\sqrt{d\log\left(1 + 2\sqrt{n/d}\right)} \lesssim \sqrt{d\log(n/d)}. \tag{3.17}$$

Here we tacitly assumed $d < n$. Recall that in Lecture 5 we obtained an upper bound of $O(\sqrt{d})$ in this setup by having a cover at scale $1/2$ and comparing the supremum to the maximum *multiplicatively*. Another way to see it is

$$\mathbb{E}\sup_{\theta\in\mathsf{B}_2^d}\langle\epsilon,\theta\rangle = \mathbb{E}\|\epsilon\| = \sqrt{d}$$

and similarly

$$\mathbb{E}\sup_{\theta\in\mathsf{B}_2^d}\langle g,\theta\rangle = \mathbb{E}\|g\| \le \sqrt{\sum_{i=1}^n \mathbb{E}g_i^2} \le \sqrt{d}$$

Hence, we lost a logarithmic factor by appealing to the general machinery of the previous section. We will also see that we can remove the extraneous logarithm by looking at a cover at multiple scales.

3.4 Function class

In particular, we will be interested in the following indexing set Θ. Let $x_1,\dots,x_n$ be fixed, and let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$. We call

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

a (scaled by $1/\sqrt{n}$) *projection* of $\mathcal{F}$ onto $x_1,\dots,x_n$. Take

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

where $\theta = (f(x_1),\dots,f(x_n))$ and $\theta' = (f'(x_1),\dots,f'(x_n))$, $f,f'\in\mathcal{F}$. With these definitions, we can define a Gaussian or Rademacher process with respect to Θ and d.

Important point: the symmetrization lemma allows us to relate supremum of the empirical process to supremum of a Rademacher process.

3.4.1 Example: Linear Function Class

We now focus on a specific example of linear functions

$$\mathcal{F} = \{x\mapsto\langle w,x\rangle : w\in\mathsf{B}_2^d\}.$$

Then for fixed $x_1,\dots,x_n\in\mathsf{B}_2^d$, a direct calculation yields

$$\mathbb{E}\sup_{w\in\mathsf{B}_2^d}\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\langle w,x_i\rangle = \frac{1}{\sqrt{n}}\mathbb{E}\left\|\sum_{i=1}^n\epsilon_i x_i\right\| \le \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}\left\|\sum_{i=1}^n\epsilon_i x_i\right\|^2} \le 1. \tag{3.18}$$

Let's see if we can recover this via our machinery. After all, the above object is precisely a supremum of a subgaussian process. Observe that

$$\Theta = \frac{1}{\sqrt{n}}\mathcal{F}|_{x_1,\dots,x_n} \subseteq \frac{1}{\sqrt{n}}\mathsf{B}_\infty^n \subset \mathsf{B}_2^n \tag{3.19}$$

and that

$$\mathcal{F}|_{x_1,\dots,x_n} = \left\{(\langle w,x_1\rangle,\dots,\langle w,x_n\rangle)\in\mathbb{R}^n : w\in\mathsf{B}_2^d\right\} = \{Xw : w\in\mathsf{B}_2^d\}$$

is a subset of a d-dimensional subspace. Hence, appealing to the previous example (3.17), we get an upper bound of $O(\sqrt{d\log(n/d)})$.

Looking back at (3.18), however, we see that we also gained an extra $\sqrt{d}$ factor, which can be a big loss in high-dimensional situations. Where did we gain it? We can see that the set $\frac{1}{\sqrt{n}}\mathcal{F}|_{x_1,\dots,x_n}$ in (3.19) is, in fact, much smaller than a d-dimensional Euclidean ball.

Theorem: Let $(U_\theta)_{\theta\in\Theta}$ be a (mean-zero) subGaussian stochastic process with respect to a metric d. Let $D = \operatorname{diam}(\Theta)$. Then for any $\delta \in [0, D]$,

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta \le 2\mathbb{E}\sup_{d(\theta,\theta')\le\delta}(U_\theta - U'_\theta) + 8\sqrt{2}\int_{\delta/4}^{D/2}\sqrt{\log\mathcal{N}(\Theta,d,\varepsilon)}d\varepsilon \tag{4.20}$$

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

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

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

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

For $\theta_N \in \Theta_N$,

$$U_{\theta_N} = \sum_{i=1}^{N} U_{\theta_i} - U_{\pi_{i-1}(\theta_i)} + U_{\theta_0} \tag{4.21}$$

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

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

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

and

$$\begin{aligned}
\mathbb{E}\max_{\theta,\theta'\in\Theta_N} U_\theta - U_{\theta'} &\le \sum_{i=1}^{N}\mathbb{E}\max_{\theta_i\in\Theta_i}(U_{\theta_i} - U_{\pi_{i-1}(\theta_i)}) + \sum_{i=1}^{N}\mathbb{E}\max_{\theta'_i\in\Theta_i}(U_{\pi_{i-1}(\theta'_i)} - U_{\theta'_i}) && (4.22)\\
&\le 2\sum_{i=1}^{N} D2^{-(i-1)}\sqrt{2\log\mathcal{N}(\Theta,d,2^{-i}D)} && (4.23)\\
&= 8\sum_{i=1}^{N} D2^{-(i+1)}\sqrt{2\log\mathcal{N}(\Theta,d,2^{-i}D)} && (4.24)\\
&\le 8\sum_{i=1}^{N}\int_{2^{-(i+1)}D}^{2^{-i}D}\sqrt{2\log\mathcal{N}(\Theta,d,\varepsilon)}d\varepsilon && (4.25)
\end{aligned}$$

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

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

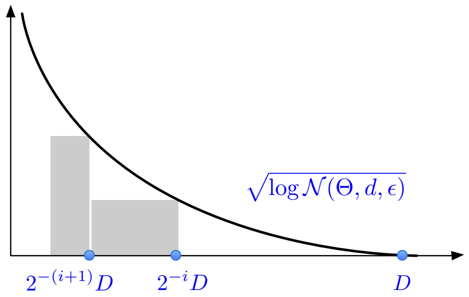


Figure 1: Illustration of the Dudley integral upper bound

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

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

for some constant C.

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

5. COVERING AND PACKING

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

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

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

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

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

Definition: An ε-net (or, ε-cover) of $\mathcal{F}$ with respect to $L^2(P)$ is a set of functions $f_1, \dots, f_N$ such that

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

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

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

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

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

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

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

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

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

Second, observe that

$$\mathcal{N}(\mathcal{F}, L^p(P_n), \varepsilon) \leq \mathcal{N}(\mathcal{F}, L^q(P_n), \varepsilon)$$

for $p \leq q$ since $\|f\|_{L^p(P_n)}$ increases with p. Note that this is different for unweighted metrics: e.g. $\|x\|_p$ is nonincreasing in p, and hence $\mathcal{N}(\Theta, \|\cdot\|_p, \varepsilon)$ is also nonincreasing in p.

Definition: An ε-packing of $\mathcal{F}$ with respect to $L^p(P_n)$ is a set $f_1, \ldots, f_N \in \mathcal{F}$ such that

$$\frac{1}{n}\sum_{i=1}^n |f_j(x_i) - f_k(x_i)|^p \geq \varepsilon^p$$

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

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

$$\mathcal{D}(\mathcal{F}, L^p(P), 2\varepsilon) \leq \mathcal{N}(\mathcal{F}, L^p(P), \varepsilon) \leq \mathcal{D}(\mathcal{F}, L^p(P), \varepsilon)$$

6. UPPER AND LOWER BOUNDS FOR RADEMACHER AVERAGES

As before, we let $U_\theta = \langle \epsilon, \theta \rangle$, $\Theta = \frac{1}{\sqrt{n}}\mathcal{F}|_{x_1,\ldots,x_n}$, and d Euclidean distance. Then from last lecture

$$\begin{aligned} \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(x_i) &= \mathbb{E}\sup_{\theta\in\Theta} U_\theta \\ &\leq 2\delta\sqrt{n} + 8\sqrt{2}\int_{\delta/4}^{D/2} \sqrt{\log \mathcal{N}(\Theta, d, \varepsilon)}d\varepsilon \end{aligned}$$

Trivially,

$$\mathcal{N}(\Theta, d, \varepsilon) = \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon).$$

Corollary: For any $X_1, \dots, X_n$,

$$\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \le \inf_{\delta \ge 0} \left\{ 8\delta + \frac{12}{\sqrt{n}} \int_{\delta}^{D/2} \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon)} d\varepsilon \right\}$$

with $D = \sup_{f,g \in \mathcal{F}} \|f - g\|_n \le 2 \sup_{f \in \mathcal{F}} \|f\|_n \le 2 \sup_{f \in \mathcal{F}} \|f\|_\infty$.

Putting together the symmetrization lemma and above Corollary, we have

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

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E} f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\} \le \mathbb{E} \inf_{\delta \ge 0} \left\{ 16\delta + \frac{24}{\sqrt{n}} \int_{\delta}^{D} \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon)} d\varepsilon \right\} \tag{6.26}$$

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

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

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

Corollary: For any $X_1, \dots, X_n$,

$$\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \ge \frac{c}{\sqrt{\log n}} \cdot \sup_{\alpha \ge 0} \alpha \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \alpha)}{n}}$$

for some absolute constant c.

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

7. PARAMETRIC AND NONPARAMETRIC CLASSES OF FUNCTIONS

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

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

Definition: We will say that a class $\mathcal{F}$ is *parametric* if for any empirical measure P_n,

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

We will say that $\mathcal{F}$ is *nonparametric* if for any empirical measure P_n,

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \asymp \left(\frac{1}{\epsilon}\right)^{p}. \tag{7.27}$$

The requirement that (7.27) holds for all measures P_n and values of n is quite strong. Yet, we will show that as an upper bound, it is true for a variety of function classes. However, one should keep in mind that there are also cases where dependence of the upper bound on n can lead to better overall estimates. The quantity

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

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

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

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

Here it's useful to note that

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

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

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

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

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

7.1 A phase transition

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

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

the integral becomes

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

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

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

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

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

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

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

yielding

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

We see that there is a transition at $p = 2$ in terms of the growth of Rademacher averages ("elbow" behavior). The phase transition will be important in the rest of the course when we study optimality of nonparametric least squares.

Remark that in the $p < 2$ regime, the rate $n^{-1/2}$ is the same rate CLT rate we would have if we simply considered $\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right|$ (or the average with random signs) with a single function. Hence, the payment for the supremum over class $\mathcal{F}$ is only in a constant that doesn't depend on n.

7.2 Single scale vs chaining

It is also worthwhile to compare the single-scale upper bound we obtained earlier to the tighter upper bound given by chaining. In other words, we are comparing

$$\delta + \sqrt{\frac{\log \mathcal{N}(\delta)}{n}}$$

versus

$$\delta + \int_\delta^{D/2} \sqrt{\frac{\log \mathcal{N}(\varepsilon)}{n}} d\varepsilon,$$

simplifying the notation for brevity.

In the parametric case, the single-scale bound becomes (with the choice of $\delta = 1/n$)

$$\sqrt{\frac{\dim\ \log n}{n}}$$

while chaining gives

$$\sqrt{\frac{\dim}{n}}.$$

In the nonparametric case, the difference is more stark:

$$\delta + \sqrt{\frac{\delta^{-p}}{n}} \asymp n^{-\frac{1}{2+p}}$$

vs

$$n^{-1/2}$$

for $p < 2$, and

$$\delta + \frac{\delta^{1-p/2}}{\sqrt{n}} \asymp n^{-1/p}$$

for $p > 2$.

7.3 Linear class: Parametric or Nonparametric?

Let's take a closer look at the function class

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathsf{B}_2^d\}$$

and take $\mathcal{X} = \mathsf{B}_2^d$. Recall that for a given $x_1, \ldots, x_n$,

$$\mathcal{F}|_{x_1,\ldots,x_n} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\} = \left\{Xw : w \in \mathsf{B}_2^d\right\}$$

where X is the $n \times d$ data matrix. As we have seen, the key quantity we need to compute is

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon).$$

What is a good upper bound for this quantity? What we had done in Lecture 16 was to discretize the set B_2^d to create a ε-net $w_1, \ldots, w_N$ of size $\mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon)$. Clearly, for any w and the corresponding ε-close element w_j of the cover,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n(\langle w, x_i\rangle - \langle w_j, x_i\rangle)^2 &\le \max_{i\in[n]}\langle w - w_j, x_i\rangle^2 \\ &\le \max_{i\in[n]} \|w - w_j\|^2 \cdot \|x_i\|^2 \\ &\le \varepsilon^2.\end{aligned}$$

Hence,

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \le \mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon). \tag{7.28}$$

In fact, a much stronger statement can be made: Since for any $x \in \mathcal{X}$

$$|\langle w, x\rangle - \langle w_j, x\rangle| \le \|w - w_j\| \, \|x\| \le \varepsilon,$$

the cover of the parameter space induces a cover of the function class *pointwise* (in the sup-norm $\|f - g\|_\infty = \sup_{x\in\mathcal{X}} |f(x) - g(x)|$) over the domain:

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \le \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon). \tag{7.29}$$

Recall that the covering number of B_2^d is

$$\left(1+\frac{2}{\varepsilon}\right)^d.$$

This gives a "parametric" growth of entropy

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \lesssim d \log(1+2/\varepsilon).$$

However, if d is large or infinite, this bound is loose. We will show that it also holds that

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \lesssim \varepsilon^{-2},$$

which is a nonparametric behavior. Hence, *the same class can be viewed as either parametric or nonparametric.* In fact, in the parametric behavior, it is not important that the domain of w is B_2^d since we would expect a similar estimate for other sets (including B_∞^d). In contrast, it will be crucial in nonparametric estimates that the norm of w is ℓ_2-bounded.

Jumping ahead, we will study neural networks and show a similar phenomenon: we can either count the number of neurons or connections (parameters) or we can calculate nonparametric "norm-based" estimates by looking at the norms of the layers in the network.

It's worth emphasizing again that (7.29) can lead to very loose bounds in high-dimensional situations. *A cover of function values on finite set of data can be significantly smaller than a cover with respect to sup norm.*

7.4 A more general result (Optional)

We have that for any fixed function

$$\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i)-\mathbb{E}f(X))\right| \le \mathrm{var}(f)^{1/2} = \|f-\mathbb{E}f\|_{L^2(P)}.$$

Obviously this implies

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

If we could ever prove

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

it would imply that we only paid $C(\mathcal{F})$ for having a statement uniform in $f \in \mathcal{F}$.

Next, rather than assuming that functions in $\mathcal{F}$ are uniformly bounded, it will be enough to assume that they have an $L_2(P)$-integrable envelope F:

$$F(x) = \sup_{f\in\mathcal{F}} |f(x)|.$$

Rather than assuming that $F(x) \le 1$, we shall assume that $\|F\|_{L^2(P)}^2 = \mathbb{E}F(X)^2 \le \infty$ and everything will be phrased in terms of $\|F\|_{L^2(P)}^2$.

Now, let $H : [0,\infty) \mapsto [0,\infty)$ is such that $H(z)$ is non-decreasing for $z > 0$ and $z\sqrt{H(1/z)}$ is non-decreasing for $z \in (0,1]$. Assume

$$\int_0^D \sqrt{H(1/x)}dx \le C_H D\sqrt{H(1/D)}$$

for all $D \in (0,1]$, and suppose that

$$\sup_Q \log 2\mathcal{N}(\mathcal{F}, L^2(Q), \tau \|F\|_{L^2(Q)}) \le H(1/\tau)$$

for all $\tau > 0$. With this control on Koltchinskii-Pollard entropy, it follows that

$$\mathbb{E}\sup \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - \mathbb{E}f(X)) \right| \lesssim \sigma \sqrt{H\left(\frac{2\|F\|_{L^2(P)}}{\sigma}\right)} \tag{7.30}$$

if n is large enough. We refer to [5] for more details, in particular Theorem 3.5.6 and the following corollaries.

Remarkably, under additional mild conditions on size of n, the inequality (7.30) can be reversed for a given P as soon as the entropy with respect to $L^2(P)$ indeed grows at least as $H\left(\frac{\|F\|_{L^2(P)}}{\sigma}\right)$.

Hence, the price we pay for uniformity in $f \in \mathcal{F}$ is truly

$$C(\mathcal{F}) \asymp \sqrt{H\left(\frac{\|F\|_{L^2(P)}}{\sigma}\right)}.$$

Of course, this expression is even simpler if $\sigma^2 = \sup_{f\in\mathcal{F}} \mathbb{E}(f(X) - \mathbb{E}f)^2$ is on the same order as $\|F\|^2_{L^2(P)} = \mathbb{E}\sup_f |f(X)|^2$.

8. COMBINATORIAL PARAMETERS

Let us gain some intuition for what can make $\widehat{\mathcal{R}}(\Theta)$ large. First, recall that

$$\widehat{\mathcal{R}}(\{\pm1\}^n) = \mathbb{E} \sup_{\theta\in\{\pm1\}^n} \langle \theta, \epsilon \rangle = n.$$

Next, suppose that for $\alpha > 0$ and $v \in \mathbb{R}^n$,

$$\alpha\{\pm1\}^n + v \subseteq \Theta.$$

Then

$$\widehat{\mathcal{R}}(\Theta) \ge \widehat{\mathcal{R}}(\alpha\{\pm1\}^n + v) = \widehat{\mathcal{R}}(\alpha\{\pm1\}^n) = \alpha\widehat{\mathcal{R}}(\{\pm1\}^n) \ge \alpha n$$

Hence, “large cubes” inside Θ make Rademacher averages large. It turns out, this is the only reason $\widehat{\mathcal{R}}(\mathcal{F}|_{x_1,\ldots,x_n})$ can be large!

The key question is whether $\mathcal{F}|_{x_1,\ldots,x_n}$ contains large cubes for a given class $\mathcal{F}$.

8.1 Binary-Valued Functions

Let's start with function classes of $\{0,1\}$-valued functions. In this case, $\mathcal{F}_{x_1,\ldots,x_n}$ is either a full $\{0,1\}^n$ cube or not. Consider the particular example of threshold functions on the real line. Take any point x_1. Clearly, $\mathcal{F}|_{x_1} = \{0,1\}$, which is a one-dimensional cube. Take two points x_1, x_2. We can only realize sign patters $(0,0),(0,1),(1,1)$, but not $(1,0)$. Hence, for no two points can we get a cube.

Definition: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$. We say that $\mathcal{F}$ shatters $x_1, \ldots, x_n \in \mathcal{X}$ if $\mathcal{F}|_{x_1,\ldots,x_n} = \{0,1\}^n$. The Vapnik-Chervonenkis dimension of $\mathcal{F}$ is

$$\mathrm{vc}(\mathcal{F}) = \max\{n : \mathcal{F} \text{ shatters some } \ x_1, \ldots, x_n\}$$

Lemma (Sauer-Shelah-Vapnik-Chervonenkis): If $\mathrm{vc}(\mathcal{F}) = d < \infty$,

$$\mathrm{card}\left(\mathcal{F}|_{x_1,\ldots,x_n}\right) \le \sum_{i=0}^{d} \binom{n}{i} \le \left(\frac{en}{d}\right)^d$$

This result is quite remarkable. It says that as soon as $n > \mathrm{vc}(\mathcal{F})$, the proportion of the cube that can be realized by $\mathcal{F}$ becomes very small (n^d vs 2^n). This combinatorial result is at the heart of empirical process theory and the early developments in pattern recognition.

In particular, the lemma can be interpreted as a covering number upper bound:

$$\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \le \left(\frac{en}{d}\right)^d$$

for any $\epsilon > 0$. Observe that these numbers are with respect to $L^\infty(P_n)$ rather than $L^2(P_n)$, and hence can be an overkill. Indeed, $L^\infty(P_n)$ covering numbers are necessarily n-dependent while we can hope to get dimension-independent $L^2(P_n)$ covering numbers. Indeed, this result (Dudley, Haussler) was already mentioned: for a binary-valued class with finite $\mathrm{vc}(\mathcal{F}) = d$,

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

Hence, a class with finite VC dimension is "parametric". On the other hand, if $\mathrm{vc}(\mathcal{F})$ is infinite, then $\mathcal{F}|_{x_1,\ldots,x_n}$ is a full cube for arbitrarily large n (for some appropriately chosen points). Hence, Rademacher averages of this set are too large and there is no uniform convergence for all P (to see this, consider P supported on the shattered set). Hence, finiteness of VC dimension is a characterization (of both distribution-free learnability and uniform convergence).

8.2 Real-Valued Functions

For binary-valued functions, the size of the cube contained in $\mathcal{F}|_{x_1,\ldots,x_n}$ was trivially 1, and we only varied n to see where the phase transition occurs. In contrast, for a general real-valued function class, it is feasible that $\mathcal{F}|_{x_1,\ldots,x_n}$ contains a cube of size α, but not larger than α; this extra parameter is in addition to the dimensionality of the cube. To deal with this extra degree of freedom, we fix the scale α and ask for the largest size n such that $\mathcal{F}|_{x_1,\ldots,x_n}$ contains a (translate of a) cube of size α. A true containment statement would read $s+(\alpha/2)\{-1,1\}^n \subseteq \mathcal{F}|_{x_1,\ldots,x_n}$. However, it is enough to ask that the equalities for the vertices are replaced with inequalities:

Definition: We say that $\mathcal{F}$ *shatters* a set of points $x_1,\ldots,x_n$ at scale α if there exists $s \in \mathbb{R}^n$ such that

$$\forall \epsilon \in \{\pm 1\}^n, \exists f \in \mathcal{F} \ \text{ s.t. } \begin{cases} f(x_t) \geq s_t + \alpha/2 & \text{if } \epsilon = +1 \\ f(x_t) \leq s_t - \alpha/2 & \text{if } \epsilon = -1 \end{cases}$$

The combinatorial dimension $\mathrm{vc}(\mathcal{F},\alpha)$ of $\mathcal{F}$ (on domain $\mathcal{X}$) at scale α is defined as the size n of the largest shattered set.

8.2.1 Example: non-decreasing functions

Consider the class of nondecreasing functions $f : \mathbb{R} \to [0,1]$. First, observe that a pointwise cover of this class does not exist ($\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) = \infty$ for any $\epsilon < 1/2$). However, $\mathcal{N}(\mathcal{F}, L^\infty(P_n), \epsilon)$ is necessarily finite. Let's calculate the scale-sensitive dimension of this class.

Claim: $\mathrm{vc}(\mathcal{F},\epsilon) \leq \epsilon^{-1}$. Indeed, fix any $x_1,\ldots,x_n$ and assume these are arranged in an increasing order. Suppose $\mathcal{F}$ shatters this set. Take the alternating sequence $\epsilon = (+1,-1,\ldots)$. We then must have a nondecreasing function that is at least $s_1+\alpha/2$ at x_1 but then no greater than $s_2-\alpha/2$ at x_2. The nondecreasing constraint implies that $s_2 \geq s_1+\alpha$. A similar argument then holds for the next point and so forth. Since functions are bounded, $n\alpha \leq 1$, which concludes the proof.

8.2.2 Control of covering numbers

The following generalization of the earlier result for binary-valued functions is due to Mendelson and Vershynin:

Theorem: Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to [-1,1]$. Then for any distribution P,

$$\mathcal{N}(\mathcal{F}, L_2(P), \varepsilon) \leq \left(\frac{c}{\varepsilon}\right)^{c\cdot \mathrm{vc}(\mathcal{F},\varepsilon/c)}$$

for all $\epsilon > 0$. Here c is an absolute constant.

In particular, plugging into the entropy integral yields

$$\int \sqrt{\mathrm{vc}(\mathcal{F},\varepsilon)\log(1/\varepsilon)}d\varepsilon$$

Rudelson-Vershynin: $\log(1/\epsilon)$ can be removed.

Back to the class of non-decreasing functions, we immediately get

$$\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \lesssim \varepsilon^{-1} \cdot \log\left(\frac{c}{\varepsilon}\right).$$

In particular, Rademacher averages of this class scale as $n^{-1/2}$ since this is a nonparametric class with entropy exponent $p < 2$.

8.3 Scale-sensitive dimension of linear class via Perceptron

In this section, we will prove that

Proposition: For

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathsf{B}_2^d\}$$

and $\mathcal{X} \subseteq \mathsf{B}_2^d$, it holds that

$$\mathrm{vc}(\mathcal{F}, \alpha) \lesssim 16\alpha^{-2}.$$

We turn to the Perceptron algorithm, defined as follows. We start with $\widehat{w}_0 = 0$. At time $t = 1, \ldots, T$, we observe $x_t \in \mathcal{X}$ and predict $\widehat{y}_t = \mathrm{sign}(\langle \widehat{w}_t, x_t\rangle)$, a *deterministic* guess of the label of x_t given the hypothesis $\widehat{w}_t$. We then observe the true label of the example $y_t \in \{\pm 1\}$. If $\widehat{y}_t \neq y_t$, we update

$$\widehat{w}_{t+1} = \widehat{w}_t + y_t x_t,$$

and otherwise $\widehat{w}_{t+1} = \widehat{w}_t$.

Lemma (Novikoff'62): For any sequence $(x_1, y_1), \ldots, (x_T, y_T) \in \mathsf{B}_2^d \times \{\pm 1\}$ the Perceptron algorithm makes at most γ^{-2} mistakes, where γ is the margin of the sequence, defined as

$$\gamma = \max_{w^* \in \mathsf{B}_2^d} \min_t y_t \langle w^*, x_t\rangle$$

Proof. If a mistake is made on round t,

$$\|\widehat{w}_{t+1}\|^2 = \|\widehat{w}_t + y_t x_t\|^2 \leq \|\widehat{w}_t\|^2 + 2y_t\langle \widehat{w}_t, x_t\rangle + 1 \leq \|\widehat{w}_t\|^2 + 1$$

Denote the number of mistakes at the end as m. Then $\|\widehat{w}_T\|^2 \leq m$. Next, for w^*,

$$\gamma \leq \langle w^*, y_t x_t\rangle = \langle w^*, \widehat{w}_{t+1} - \widehat{w}_t\rangle,$$

and so by summing and telescoping, $m\gamma \leq \langle w^*, \widehat{w}_T\rangle \leq \sqrt{m}$. This concludes the proof. □

Remarkably, the number of mistakes does not depend on the dimension d. We will now show that the mistake bound translates into a bound on the scale-sensitive dimension.

Proof of Proposition. Suppose there exist a shattered set $x_1, \ldots, x_m \in \mathsf{B}_2^d$: there exists $s_1, \ldots, s_m \in [-1, 1]$ such that for any sequence of signs $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ there exists a $w_\epsilon \in \mathsf{B}_2^d$ such that

$$\epsilon_i(\langle w_\epsilon, x_i\rangle - s_i) \geq \alpha/2.$$

Claim: we can reparametrize the problem so that $s_i = 0$. Indeed, take

$$\tilde{w}_\epsilon = [w_\epsilon, 1], \quad \tilde{x}_i = [x_i, -s_i].$$

Then we have

$$\epsilon_i \langle \tilde{w}_\epsilon, \tilde{x}_i \rangle \geq \alpha/2.$$

while the norms are at most $\sqrt{2}$:

$$\|\tilde{w}_\epsilon\|^2 = \|w_\epsilon\|^2 + 1 \leq 2, \qquad \|\tilde{x}_i\|^2 \leq 2$$

Now comes the key step. We run Perceptron on the sequence $\tilde{x}_1/\sqrt{2}, \ldots, \tilde{x}_m/\sqrt{2}$ and $y_i = -\widehat{y}_i$. That is, we force Perceptron to make mistakes on every round, no matter what the predictions are. It is important that Perceptron makes deterministic predictions for this argument to work. Note that the sequence of predictions of Perceptron defines the sequence $y = (y_1, \ldots, y_n)$ with

$$y_i \langle \tilde{w}_y/\sqrt{2}, \tilde{x}_i/\sqrt{2} \rangle \geq \alpha/4.$$

Hence, by Novikoff's result,

$$m \leq 16/\alpha^2.$$

□

Interestingly, both Perceptron and VC theory were developed in the 60's as distinct approaches (online vs batch), yet the connection between them runs deeper than was recognized, until recently. In particular, the above proof in fact shows that a stronger *sequential* version of $\text{vc}(\mathcal{F}, \alpha)$ is also bounded by $16\alpha^{-2}$, where (roughly speaking) sequential analogues allow the sequence to evolve as a predictable process with respect to a dyadic filtration. It turns out that there are sequential analogues of Rademacher averages, covering numbers, Dudley chaining, and combinatorial dimensions, and these govern *online* (rather than i.i.d.) learning. If there is time, we will mention these towards the end of the course.

9. REGRESSION. PREDICTION VS ESTIMATION

As before, let $\mathcal{S} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a set of i.i.d. pairs with distribution $P = P_X \times P_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$. Let $f^*(x) = \mathbb{E}[Y|X = x]$ be the *regression function.* One can show that

$$f^* \in \operatorname*{argmin}_f \mathbb{E}(f(X) - Y)^2$$

where minimization is over all measurable functions.

Given a class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$, we also define

$$f_{\mathcal{F}} \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2$$

to be the best predictor within the class $\mathcal{F}$.

Risk of a function f is defined as

$$\mathbb{E}(f(X) - f^*(X))^2 = \|f - f^*\|^2_{L^2(P)} = \|f - f^*\|^2$$

We will be interested in analyzing estimators $\widehat{f}$ constructed on the basis of n datapoints. The hat on $\widehat{f}$ reminds us about the dependence on $\mathcal{S}$.

Note that for any function f,

$$\begin{aligned}
\mathbb{E}(f(X) - Y)^2 - \min_h \mathbb{E}(h(X) - Y)^2 &= \mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\
&= \mathbb{E}(f(X) - f^*(X) + f^*(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\
&= \mathbb{E}(f(X) - f^*(X))^2
\end{aligned}$$

Question: given i.i.d. data $\mathcal{S}$, can we select estimator $\widehat{f}$ such that risk

$$\left\|\widehat{f}-f^*\right\|^2$$

is small in expectation or high-probability (with respect to the draw of $\mathcal{S}$)? Without further assumptions this is not possible.

Two standard scenarios:

- Well-specified case: given some class $\mathcal{F}$, assume $f^* \in \mathcal{F}$. More precisely, P is such that the regression function is in the class $\mathcal{F}$.
- Misspecified case (agnostic learning in CS community): Redefine goal as

$$\begin{aligned}&\left\|\widehat{f}-f^*\right\|^2-\min_{f\in\mathcal{F}}\|f-f^*\|^2 \\ &=\mathbb{E}(\widehat{f}(X)-Y)^2-\min_{f\in\mathcal{F}}\mathbb{E}(f(X)-Y)^2\end{aligned} \tag{9.31}$$

  but do not insist that $f^* \in \mathcal{F}$. Upper bounds on (9.31) are called Oracle Inequalities in statistics, while the prediction form has been also studied in statistical learning theory.

We see that the problem of prediction and the problem of estimation naturally coincide for square loss. Moreover, the misspecified problem arises naturally as a relaxation of an assumption on the form of the distribution.

Here, the road naturally forks into at least several paths: analyze the well-specified case, analyze the misspecified case, or change the loss function altogether. Let us briefly consider the last generalization.

10. PREDICTION WITH OTHER LOSS FUNCTIONS

This will be a brief but useful detour. Consider changing the loss function in the prediction problem (9.31) on the previous page:

$$\mathbb{E}\boldsymbol{\ell}(f(X),Y)-\min_{f\in\mathcal{F}}\mathbb{E}\boldsymbol{\ell}(f(X),Y) \tag{10.32}$$

for some $\boldsymbol{\ell}:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$. In Lecture 14 we already showed that ERM

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

enjoys

$$\mathbb{E}\boldsymbol{\ell}(\widehat{f}(X),Y)-\min_{f\in\mathcal{F}}\mathbb{E}\boldsymbol{\ell}(f(X),Y)\le\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{E}\boldsymbol{\ell}(f(X),Y)-\frac{1}{n}\sum_{i=1}^n\boldsymbol{\ell}(f(X_i),Y_i).$$

The latter is at most

$$2\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\epsilon_i\boldsymbol{\ell}(f(X_i),Y_i) \tag{10.33}$$

by symmetrization, which is Rademacher averages of the loss class

$$\boldsymbol{\ell}\circ\mathcal{F}|_{(X_1,Y_1),\ldots,(X_n,Y_n)}$$

We would like to further upper bound this with Rademacher averages of the function class itself. This can be done if $\boldsymbol{\ell}$ is Lipschitz in the first argument.

Lemma (Contraction): Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz, $i = 1, \ldots, n$. Let $\Theta \subset \mathbb{R}^n$ and $\phi \circ \theta = (\phi_1(\theta_1), \ldots, \phi_n(\theta_n))$ for $\theta \in \Theta$. Denote $\phi \circ \Theta = \{\phi \circ \theta : \theta \in \Theta\}$. Then

$$\widehat{\mathcal{R}}(\phi \circ \Theta) \leq \widehat{\mathcal{R}}(\Theta).$$

Proof. Conditionally on $\epsilon_1, \ldots, \epsilon_{n-1}$,

$$\begin{aligned}
\mathbb{E}_{\epsilon_n} \sup_{\theta\in\Theta} \langle \phi \circ \theta, \epsilon \rangle &= \frac{1}{2} \left(\sup_{\theta\in\Theta} \{\langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \phi_n(\theta_n)\} + \sup_{\theta'\in\Theta} \left\{\langle \phi \circ \theta'_{1:n-1}, \epsilon_{1:n-1} \rangle - \phi_n(\theta'_n)\right\} \right) \\
&\leq \frac{1}{2} \sup_{\theta,\theta'\in\Theta} \langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \langle \phi \circ \theta'_{1:n-1}, \epsilon_{1:n-1} \rangle + |\theta_n - \theta'_n| \\
&= \frac{1}{2} \sup_{\theta,\theta'\in\Theta} \langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \langle \phi \circ \theta'_{1:n-1}, \epsilon_{1:n-1} \rangle + \theta_n - \theta'_n \\
&= \frac{1}{2} \left(\sup_{\theta\in\Theta} \{\langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \theta_n\} + \sup_{\theta'\in\Theta} \left\{\langle \phi \circ \theta'_{1:n-1}, \epsilon_{1:n-1} \rangle - \theta'_n\right\} \right) \\
&= \mathbb{E}_{\epsilon_n} \sup_{\theta\in\Theta} \langle \phi \circ \theta_{1:n-1}, \epsilon_{1:n-1} \rangle + \epsilon_n \theta_n
\end{aligned}$$

The inequality follows from the Lipschitz condition and the following equality is justified because of the symmetry of the other two terms with respect to renaming θ and θ'. Proceeding to remove the other signs concludes the proof. □

We now apply this lemma to functions $\phi_i(\cdot) = \boldsymbol{\ell}(\cdot, Y_i)$. As long as these functions are L-Lipschitz, contraction lemma gives

$$\mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{\ell}(f(X_i), Y_i) \leq L \cdot \mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) = L \cdot \frac{1}{n} \mathbb{E} \widehat{\mathcal{R}}(\mathcal{F}|_{X_1,\ldots,X_n}), \tag{10.34}$$

the (expected) Rademacher averages of $\mathcal{F}$. The argument can be seen as a generalization of the argument we did in Lecture 14 for classification where we "erased" multipliers $(1-2Y_i)$.

The simple analysis we just performed applies to any Lipschitz loss function. For uniformly bounded $\mathcal{F}$ and $\mathcal{Y}$, square loss is Lipschitz, but that is no longer true for unbounded $\mathcal{Y}$ (e.g. for real-value prediction with Gaussian noise). Hence, such an analysis only goes so far.

Second, observe that one would only obtain rates $n^{-1/2}$ or worse with such an analysis, while we might hope to have faster decrease. For instance, in finite-dimensional regression, one can recall the classical $d \cdot n^{-1}$ rates for Least Squares.

A quick inspection tells us that the second step (see Lecture 14) in the sequence of inequalities

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \mathbf{L}(f^*) \leq \mathbb{E}\left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] \leq \mathbb{E} \sup_{f\in\mathcal{F}} \left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] \tag{10.35}$$

for ERM $\widehat{f}$ may be too loose. The second step only used the fact that $\widehat{f}$ belongs to $\mathcal{F}$. It turns out one can localize its place in $\mathcal{F}$ better than that.

Next few lectures will be on nonparametric regression. We will start with well-specified models.

11. NONPARAMETRIC REGRESSION: WELL-SPECIFIED CASE

We will start with “fixed design”: $x_1, \dots, x_n \in \mathcal{X}$ are fixed. Let

$$Y_i = f^*(x_i) + \eta_i$$

where η_i are zero-mean independent subGaussian. Suppose $f^* \in \mathcal{F}$. Goal: estimate f^* on the points $x_1, \dots, x_n$ (denoise the observed values). That is, the goal is to provide nonasymptotic bounds on

$$\mathbb{E}_\eta \left\| \widehat{f} - f^* \right\|^2_{L^2(P_n)},$$

where $\widehat{f}$ is the least squares (ERM) constrained to $\mathcal{F}$. In constrast, in random design the goal is w.r.t. $L^2(P)$ with P unknown, while here P_n is known. We write the $L^2(P_n)$ norm more succinctly as $\mathbb{E} \left\| \widehat{f} - f^* \right\|^2_n$.

Since

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - Y_i)^2 = \|f - Y\|_n^2$$

we have

$$\|f^* - Y\|_n^2 \geq \left\| \widehat{f} - Y \right\|_n^2 = \left\| \widehat{f} - f^* + f^* - Y \right\|_n^2 = \left\| \widehat{f} - f^* \right\|_n^2 + \|f^* - Y\|_n^2 + 2\langle \widehat{f} - f^*, f^* - Y \rangle_n$$

where $\langle a, b \rangle_n = \frac{1}{n} \langle a, b \rangle$. Thus,

$$\left\| \widehat{f} - f^* \right\|_n^2 \leq 2\langle \eta, \widehat{f} - f^* \rangle_n \tag{11.36}$$

which we will call *the basic inequality.*

11.1 Informal intuition for localization

Before developing the localization approach, we provide some intuition. The first intuition comes from viewing (11.36) as a fixed point.

Let’s assume for simplicity that η_i are 1-subGaussian. For $a \in \mathbb{R}^n$, we have that with high probability

$$\langle \eta, a \rangle \lesssim \|a\|$$

Hence, if it holds that

$$\|a\|^2 \leq \langle \eta, a \rangle,$$

then $\|a\| \lesssim 1$.

We can try to repeat this argument with a being the values of $\widehat{f} - f^*$ on the data. However, since $\widehat{f}$ depends on η, we do not have the averaging that we need. Still, we can do the mental experiment of assuming that the dependence is “weak” (e.g. we fit linear regression in small d and large n). Then a bound on the size of $\left\| \widehat{f} - f^* \right\|_n$ would lead to an improved bound on the RHS of the basic inequality, which would in turn tighten the bound on the LHS of the basic inequality, suggesting some kind of a fixed point. It also seems intuitive that this fixed point likely depends on $\mathcal{F}$ and its richness.

11.2 1st approach to localization: ratio-type inequalities

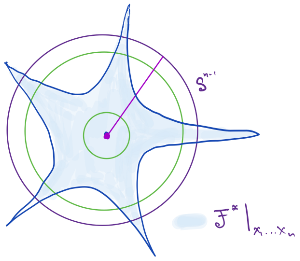
To simplify the proof somewhat, we will assume that $\eta_1, \ldots, \eta_n$ are independent standard normal $N(0, 1)$.

We proceed as in the linear case earlier in the course. First, we divide both sides of the Basic Inequality (11.36) by $\left\|\widehat{f} - f^*\right\|_n$ and further upper bound the right-hand side by a supremum over f, removing the dependence of the algorithm on the data:

$$\left\|\widehat{f} - f^*\right\|_n \leq 2 \sup_{f \in \mathcal{F}} \langle \eta, \frac{f - f^*}{\|f - f^*\|_n} \rangle_n \tag{11.37}$$

By squaring both sides, we would get an upper bound on the estimation error (in probability or in expectation).

Let us use the shorthand $\mathcal{F}^* = \mathcal{F} - f^*$. The rest of the discussion will be about complexity of the neighborhood around f^* in $\mathcal{F}$, or, equivalently, complexity of the neighborhood of 0 in $\mathcal{F}^*$. Observe that we only care about values of functions on the data $x_1, \ldots, x_n$, so the discussion is really about the set $\mathcal{F}^*|_{x_1,\ldots,x_n}$, drawn in blue below.



At this point, one can say that there is no difference from the linear case, and we should just go ahead and analyze

$$\sup_{g \in \mathcal{F}^*} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

After all, this is just the Gaussian width (normalized by $\sqrt{n}$) of the subset of the sphere obtained by rescaling all the functions:

$$K = \{v \in \mathsf{S}^{n-1} : \exists g \in \mathcal{F}^* \ \text{ s.t. } \ v = (g(x_1), \ldots, g(x_n))/(\sqrt{n}\,\|g\|_n)\}.$$

(here the normalization is because $\|g\|_n$ is scaled as $1/\sqrt{n}$ times the ℓ_2 norm.) How big is this subset of the sphere? Note: if the set is all of S^{n-1}, we are doomed since in that case

$$\sup_{g \in \mathcal{F}^*} \langle \eta, \frac{g}{\|g\|_n} \rangle_n = \sup_{v \in \mathsf{S}^{n-1}} \frac{1}{\sqrt{n}} \langle \eta, v \rangle = \frac{1}{\sqrt{n}} \|\eta\| \sim 1$$

and does not converge to zero. What we would need is that K is a *significantly smaller* subset of the sphere. In the linear case, this was easy: we simply used the fact that the subset is d-dimensional. However, for nonlinear functions, it is not easy to see what the set is.

There is a bigger problem, however. Upon rescaling every vector to the sphere, all the functions are treated equally even if their unscaled versions are very close to being zero (that is, close to f^* in the original class $\mathcal{F}$). In other words, the quantity

$$\sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

can be potentially much smaller than the unrestricted supremum. This is depicted in the above figure. If we look at functions within the smaller green sphere, its rescaled version is the whole sphere. However, at larger scales (e.g. the larger green sphere), the set can be much smaller. Understanding the map

$$u \mapsto \sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

will be key. In particular, we can break up the balance at scale u and instead have a better upper bound

$$\left\| \widehat{f} - f^* \right\|_n \leq u + 2 \sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n \tag{11.38}$$

Indeed, to show (11.38), write

$$\begin{aligned} \left\| \widehat{f} - f^* \right\|_n &= \left\| \widehat{f} - f^* \right\|_n \mathbf{1}\left\{ \left\| \widehat{f} - f^* \right\|_n < u \right\} + \left\| \widehat{f} - f^* \right\|_n \mathbf{1}\left\{ \left\| \widehat{f} - f^* \right\|_n \geq u \right\} \\ &\leq u + \left\| \widehat{f} - f^* \right\|_n \mathbf{1}\left\{ \left\| \widehat{f} - f^* \right\|_n \geq u \right\} \\ &\leq u + 2\langle \eta, \frac{\widehat{f} - f^*}{\left\| \widehat{f} - f^* \right\|_n} \rangle_n \times \mathbf{1}\left\{ \left\| \widehat{f} - f^* \right\|_n \geq u \right\} \\ &\leq u + 2 \sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n \end{aligned}$$

Consider the following assumption:

Definition: A class $\mathcal{H}$ is *star-shaped* (around 0) if $h \in \mathcal{H}$ implies $\lambda h \in \mathcal{H}$ for $h \in [0, 1]$. In particular, if $\mathcal{H}$ is convex and contains 0, it is star-shaped.

We will assume that $\mathcal{F}^*$ is star-shaped. In particular, if $\mathcal{F}$ is convex, then $\mathcal{F}^*$ is star-shaped. The key property of a star-shaped class is that by increasing the radius, the sets cannot become more complex, as for any function there is a scaled copy of it at a smaller magnitude.

In light of this last remark, we claim that the inequality $\|g\|_n \geq u$ in the supremum in (11.38) can be replaced with an *equality* if the class is star-shaped. Indeed, for any $g \in \mathcal{F}^*$ with $\|g\|_n \geq u$, there is a corresponding function $h = u\frac{g}{\|g\|_n}$ with norm $\|h\|_n = u$ and

$$\langle \eta, \frac{g}{\|g\|_n} \rangle_n = \langle \eta, \frac{h}{u} \rangle_n$$

Hence,

$$\langle \eta, \frac{g}{\|g\|_n} \rangle_n \leq \frac{1}{u} \sup_{h \in \mathcal{F}^* : \|h\|_n = u} \langle \eta, h \rangle_n$$

Taking a supremum on the LHS over g with $\|g\|_n \geq u$ gives an upper bound on (11.38) as

$$\begin{aligned} \left\| \widehat{f} - f^* \right\|_n &\leq u + \frac{2}{u} \sup_{g \in \mathcal{F}^* : \|g\|_n = u} \langle \eta, g \rangle_n \\ &\leq u + \frac{2}{u} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq u} \langle \eta, g \rangle_n \end{aligned} \tag{11.39}$$

where in the last step we included all the functions below level u. We will use concentration to replace the second term with its expectation. In particular, define

$$Z(u) = \sup_{g \in \mathcal{F}^* : \|g\|_n \leq u} \langle \eta, g \rangle_n$$

and

$$G(u) = \mathbb{E} Z(u).$$

If we were to replace $Z(u)$ on the RHS of (11.39) with $G(u)$, the natural balance between the two terms would be

$$u = \frac{2}{u} G(u)$$

Definition: The *critical radius* δ_n will be the minimum δ satisfying

$$G(\delta) \leq \delta^2/2$$

One can ask if this critical radius is actually well-defined. This follows from the following:

Lemma: If $\mathcal{F}^*$ is star-shaped, the function $u \mapsto G(u)/u$ is non-increasing.

Proof. Let $\delta' < \delta$. Take any $h \in \mathcal{F}^*$ with $\delta' < \|h\|_n \leq \delta$. By star-shapedness,

$$h' = \left(\frac{\delta'}{\delta} \right) h \in \mathcal{F}^*$$

and $\|h'\|_n = \frac{\delta'}{\delta} \|h\|_n \leq \delta'$. Hence,

$$\langle \eta, h \rangle_n = \frac{\delta}{\delta'} \langle \eta, h' \rangle_n \leq \frac{\delta}{\delta'} Z(\delta')$$

Taking supremum on the left-hand side over h with $\|h\|_n \leq \delta$, as well as expectation on both sides, finishes the proof. □

In particular, for any $u \geq \delta_n$,

$$G(u) \leq u^2/2$$

Indeed,

$$G(u) = u \frac{G(u)}{u} \leq u \frac{G(\delta_n)}{\delta_n} \leq u \delta_n / 2 \leq u^2/2. \tag{11.40}$$

To formally replace $Z(u)$ with $G(u)$ in the balancing equation, we need a concentration result.

Lemma (Gaussian Concentration): Let $\eta = (\eta_1, \ldots, \eta_n)$ be a vector of independent standard normals. Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz (w.r.t. Euclidean norm). Then for all $t > 0$

$$\mathbb{P}\left(\phi(\eta) - \mathbb{E}\phi \geq t\right) \leq \exp\left\{-\frac{t^2}{2L^2}\right\}$$

First, observe that $Z(u)$ is $(u/\sqrt{n})$-Lipschitz function of η. Omitting the argument u,

$$Z[\eta] - Z[\eta'] \leq \sup_{g \in \mathcal{F}^*, \|g\|_n \leq u} \langle \eta, g\rangle_n - \langle \eta', g\rangle_n \leq \left\|\eta - \eta'\right\|_n \sup_{g \in \mathcal{F}^*, \|g\|_n \leq u} \|g\|_n \leq \frac{u}{\sqrt{n}} \left\|\eta - \eta'\right\|$$

Hence, for any $u > 0$,

$$\mathbb{P}\left(Z(u) - \mathbb{E}Z(u) \geq t\right) \leq \exp\left\{-\frac{nt^2}{2u^2}\right\} \tag{11.41}$$

In particular, by setting $t = u^2$,

$$\mathbb{P}\left(Z(u) \geq G(u) + u^2\right) \leq \exp\left\{-\frac{nu^2}{2}\right\} \tag{11.42}$$

In light of (11.40), we have proved

Lemma: Assuming $\mathcal{F}^*$ is star-shaped, with probability at least $1 - \exp\left\{-\frac{nu^2}{2}\right\}$,

$$Z(u) \leq 1.5u^2 \tag{11.43}$$

for any $u \geq \delta_n$.

Thus, from (11.39), we have

$$\left\|\widehat{f} - f^*\right\|_n \leq 4u \tag{11.44}$$

with probability at least $1 - \exp\left\{-\frac{nu^2}{2}\right\}$, for any $u \geq \delta_n$. Squaring both sides, yields

Theorem: Assume $x_1, \ldots, x_n$ are fixed, $\eta_1, \ldots, \eta_n$ are i.i.d. standard normal, and $Y_i = f^*(x_i) + \eta_i$ with $f^* \in \mathcal{F}$. Assume $\mathcal{F} - f^*$ is star-shaped and δ_n the corresponding critical radius. Then constrained least squares $\widehat{f}$ satisfies

$$\mathbb{P}\left(\left\|\widehat{f} - f^*\right\|_n^2 \geq 16s\delta_n^2\right) \leq \exp\left\{-\frac{ns\delta_n^2}{2}\right\} \tag{11.45}$$

for any $s \geq 1$. In particular, this implies

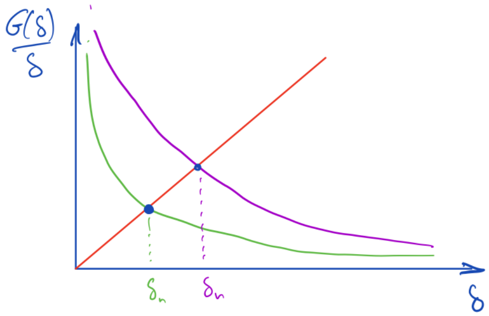
$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 \lesssim \delta_n^2 + \frac{1}{n}.$$

Note: in the literature, you will find a slightly different parametrization. Write $\psi(r) = \mathbb{E}Z(\sqrt{r})$. In other words, $\psi(u^2) = G(u)$. Then ψ has the *subroot* property:

$$\psi(ra) \leq \sqrt{a}\psi(r)$$

using the same type of proof as above. The fixed point then reads as the smallest r such that $\psi(r) \leq r$ (ignoring the constant).

Let's quickly discuss the behavior of $G(\delta)/\delta$.



The above sketch shows the function $\delta \mapsto G(\delta)/\delta$ for two classes of functions. The purple curve corresponds to a more complex class, since the Gaussian width (normalized by δ) grows faster as $\delta \to 0$. The corresponding fixed point is larger for a more rich class.

11.3 2nd approach to localization: offset

We start again with the basic inequality

$$\left\|\widehat{f} - f^*\right\|_n^2 \leq 2\langle \eta, \widehat{f} - f^* \rangle_n$$

and trivially write it as

$$\left\|\widehat{f} - f^*\right\|_n^2 \leq 4\langle \eta, \widehat{f} - f^* \rangle_n - \left\|\widehat{f} - f^*\right\|_n^2$$

Now take the supremum on both sides:

$$\begin{aligned}\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 &\leq \mathbb{E}\sup_{f\in\mathcal{F}} 4\langle \eta, f - f^* \rangle_n - \|f - f^*\|_n^2 \\ &= \mathbb{E}\sup_{g\in\mathcal{F}-f^*} \frac{1}{n}\sum_{i=1}^n 4\eta_i g(x_i) - g(x_i)^2\end{aligned}$$

which we shall call *the offset Rademacher (or Gaussian) averages.*

Contrast this approach with the first approach where we divided both sides by the norm $\left\|\widehat{f} - f^*\right\|_n$ and then upper bounded by supremum over an appropriately localized subset, then squared both sides.

Surprisingly, this somewhat simpler approach yields correct upper bounds. Note that the negative quadratic term annihilates the fluctuations of the term $\eta_i g(x_i)$ when the magnitude of g becomes large enough (beyond some critical radius). Hence, the supremum is achieved in a finite radius, no larger than the critical radius:

Lemma: Let δ_n be the critical radius. Then for any $c \geq 1$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{F}^*} 2c\langle \eta, g\rangle_n - \|g\|_n^2 > 2c^2\delta_n^2 + \frac{2c^2u}{n}\right) \leq \exp\{-u/2\} \tag{11.46}$$

In particular,

$$\mathbb{E}\sup_{g\in\mathcal{F}^*} 2\langle \eta, g\rangle_n - \|g\|_n^2 \lesssim \delta_n^2 + \frac{1}{n}.$$

Proof. By Gaussian concentration,

$$\mathbb{P}\left(Z(\delta_n) \geq \mathbb{E}Z(\delta_n) + t\delta_n\right) \leq \exp\left\{-\frac{nt^2}{2}\right\}. \tag{11.47}$$

We now condition on the complement of the above event. Take $g \in \mathcal{F}^*$. Consider two cases. First, if $\|g\|_n \leq \delta_n$ then

$$2c\langle \eta, g\rangle_n - \|g\|_n^2 \leq 2cZ(\delta_n) \leq 2c\left(\mathbb{E}Z(\delta_n) + t\delta_n\right) \leq 2c\left(\frac{\delta_n^2}{2} + t\delta_n\right) \leq c(t+\delta_n)^2 \tag{11.48}$$

Second, if $\|g\|_n \geq \delta_n$, we set $r = \delta_n / \|g\|_n \leq 1$. Then

$$2c\langle \eta, g\rangle_n - \|g\|_n^2 = \frac{2c}{r}\langle \eta, \frac{\delta_n}{\|g\|_n} g\rangle - \frac{\delta_n^2}{r^2} \leq \frac{2c}{r}Z(\delta_n) - \frac{\delta_n^2}{r^2} = \frac{2\delta_n}{r}\frac{cZ(\delta_n)}{\delta_n} - \frac{\delta_n^2}{r^2}. \tag{11.49}$$

Using $2ab - b^2 \leq a^2$, we get a further upper bound of

$$c^2\left(\frac{Z(\delta_n)}{\delta_n}\right)^2 \leq c^2\left(\frac{\delta_n^2/2 + t\delta_n}{\delta_n}\right)^2 = c^2(\delta_n/2 + t)^2 \tag{11.50}$$

□

11.3.1 Example: linear regression

To get a sense of the behavior of the offset process, consider the linear class $\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathbb{R}^d\}$. First, $\mathcal{F} - f^* = \mathcal{F}$. Second, note that functions are unbounded, and so Rademacher averages are unbounded too. However, offset averages are

$$\begin{aligned}\sup_{w\in\mathbb{R}^d} \sum_{i=1}^n \eta_i\langle w, x_i\rangle - c\langle w, x_i\rangle^2 &= \sup_{w\in\mathbb{R}^d} \langle w, \sum_{i=1}^n \eta_i x_i\rangle - c\|w\|_\Sigma^2 &(11.51)\\ &= \frac{1}{4c}\left\|\sum_{i=1}^n \eta_i x_i\right\|_{\Sigma^\dagger}^2 &(11.52)\end{aligned}$$

where $\Sigma = \sum_{i=1}^n x_i x_i^\mathsf{T}$ and $\Sigma^\dagger$ is the pseudoinverse. Assuming $\mathbb{E}\eta_i^2 \leq 1$,

$$\mathbb{E}\left\|\sum_{i=1}^n \eta_i x_i\right\|_{\Sigma^{-1}}^2 \leq \sum_{i=1}^n x_i^\mathsf{T}\Sigma^\dagger x_i = \mathrm{tr}(\Sigma\Sigma^\dagger) = \mathrm{rank}(\Sigma)$$

We see that, these offset Rademacher/Gaussian averages have the right behavior: we already saw in the first part of the course that the fast rate for linear regression is $O\left(\frac{\mathrm{rank}(\Sigma)}{n}\right)$ without further assumptions.

We can view the negative term that extinguishes the fluctuations of the zero-mean process as coming from the curvature of the square loss. Without the curvature, the negative term is not there and we are left with the usual Rademacher/Gaussian averages.

12. LEAST SQUARES

12.0.1 Nonparametric

We would like to calculate the critical radius δ_n for some function clases of interest. Recall that δ_n is defined as the smallest number such that

$$\mathbb{E} \sup_{g\in\mathcal{F}^*:\|g\|_n\leq\delta} \langle \eta, g\rangle_n \leq \delta^2/2.$$

The strategy is to find upper bounds on the left-hand-side in terms of δ and then solve for the minimal δ. In particular, we know that for any $\alpha \geq 0$,

$$\mathbb{E} \sup_{g\in\mathcal{F}^*:\|g\|_n\leq\delta} \langle \eta, g\rangle_n \lesssim \alpha + \frac{1}{\sqrt{n}}\int_{\alpha/4}^{\delta} \sqrt{\log\mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon)}d\varepsilon$$

Suppose we have

$$\log\mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon) \lesssim \varepsilon^{-p}$$

for $p \in (0, 2)$. Then, taking $\alpha = 0$,

$$\mathbb{E} \sup_{g\in\mathcal{F}^*:\|g\|_n\leq\delta} \langle \eta, g\rangle_n \lesssim n^{-1/2}[\varepsilon^{1-p/2}]_0^\delta = n^{-1/2}\delta^{1-p/2}$$

Setting

$$n^{-1/2}\delta^{1-p/2} = \delta^2$$

yields

$$\delta_n = n^{-\frac{1}{2+p}}$$

and thus the rate of the least squares estimator is

$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 \lesssim n^{-\frac{2}{2+p}}$$

It can be shown that minimax optimal rates of estimation (for any estimator) for fixed design are given by the fixed point (see [16])

$$\frac{\log\mathcal{N}(\mathcal{F}, L^2(P_n), \delta_*)}{n} \asymp \delta_*^2 \tag{12.53}$$

If $\log\mathcal{N}(\mathcal{F}, L^2(P_n), \delta) \asymp \delta^{-p}$, the balance is

$$\delta_*^{-p}n^{-1} \asymp \delta_*^2$$

which gives the same rate of $\delta_*^2 = n^{-\frac{2}{2+p}}$. Hence, least squares are optimal in this minimax sense for $p \in (0, 2)$.

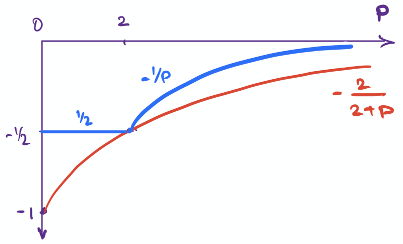


Figure 2: Optimal (in general) rates $n^{-\frac{2}{2+p}}$ (obtained with localization for $p \in (0, 2)$ by ERM) vs without localization (e.g. via global Rademacher averages)

Example: Convex L-Lipschitz functions on a compact domain in $\mathbb{R}^d$:

$$\log \mathcal{N}(\mathcal{F}_{\text{cvx,lip}}, L^2(P_n), \varepsilon) \le (L/\varepsilon)^{d/2}$$

Example: L-Lipschitz functions on a compact domain in $\mathbb{R}^d$:

$$\log \mathcal{N}(\mathcal{F}_{\text{lip}}, L^2(P_n), \varepsilon) \le (L/\varepsilon)^{d}$$

12.0.2 Parametric

Consider the parametric case,

$$\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2/\varepsilon)$$

Then

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \le \delta} \langle \eta, g \rangle_n \lesssim \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{d \log(1 + 2/\varepsilon)} d\varepsilon \tag{12.54}$$

Change of variables gives an upper bound

$$\sqrt{\frac{d}{n}} \delta \cdot \int_0^1 \sqrt{\log(1 + 2/(u\delta))} du \tag{12.55}$$

Unfortunately, this gives a pesky logarithmic factor that should not be there. However, for some parametric cases one can, in fact, prove that *local covering numbers* behave as

$$\log \mathcal{N}(\mathcal{F}^* \cap \{g : \|g\|_n \le \delta\}, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2\delta/\varepsilon) \tag{12.56}$$

In this case, the change-of-variables leads to

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \le \delta} \langle \eta, g \rangle_n \lesssim \sqrt{\frac{d}{n}} \delta \cdot \int_0^1 \sqrt{\log(1 + 2/\varepsilon)} d\varepsilon \lesssim \sqrt{\frac{d}{n}} \delta \tag{12.57}$$

Equating

$$\delta \sqrt{\frac{d}{n}} \asymp \delta^2$$

yields

$$\delta_n^2 \asymp \frac{d}{n}$$

Note that local covering numbers (12.56) are available in some parametric cases (e.g. when we discretize the parameter space of linear functions) but may not be available for some other classes (e.g. for VC classes, except under additional conditions).

12.1 Remarks

- to bound metric entropy of $\mathcal{F}^* = \mathcal{F} - f^*$, instead consider $\mathcal{F} - \mathcal{F}$. This often leads to only mild increase in a constant. For instance, if $\mathcal{F}$ is a class of L-Lipschitz functions, then $\mathcal{F} - \mathcal{F}$ is a subset of $2L$-Lipschitz functions.
- Note that the rate δ_n^2 depends on local covering numbers (or, local complexity) around f^*. This gives a path to proving adaptivity results (e.g. if f^* is convex but has only k linear pieces, the rate of estimation is parametric because its neighborhood is "simple").
- A simple counting argument (see Yang & Barron 1999, Section 7) shows that for rich enough classes (e.g. nonparametric) worst-case local entropy (worst-case location in the class) and global entropies behave similarly. This implies, in particular, that instead of constructing a local packing for a lower bound (via hypothesis testing), one can instead use global entropy with Fano inequality, justifying the LHS of (12.53) as the lower bound for estimation. See also Mendelson's "local vs global parameters" paper for an in-depth discussion.

13. ORACLE INEQUALITIES

What if we do not assume the regression function f^* is in $\mathcal{F}$? How can we prove an oracle inequality

$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 - \inf_{f\in\mathcal{F}} \|f - f^*\|_n^2 \leq \phi(\mathcal{F}, n)$$

Again, we will focus on fixed design.

13.1 Convex $\mathcal{F}$

Suppose $\mathcal{F}$ is convex (or, rather, $\mathcal{F}|_{x_1,\ldots,x_n}$ is convex). Let $\widehat{f}$ be the constrained least squares:

$$\widehat{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - Y_i)^2 = \operatorname*{argmin}_{f\in\mathcal{F}} \|f - Y\|_n^2$$

For the basic inequality we used

$$\left\|\widehat{f} - Y\right\|_n^2 \leq \|f^* - Y\|_n^2$$

but in the misspecified case this is no longer true. However, what is true is that

$$\left\|\widehat{f} - Y\right\|_n^2 \leq \|f_{\mathcal{F}} - Y\|_n^2$$

Unfortunately, this inequality is not strong enough to get us the desired result. Fortunately, we can do better. Since $\widehat{f}$ is a projection of Y onto $F = \mathcal{F}_{x_1,\ldots,x_n}$, it holds that

$$\left\|\widehat{f} - Y\right\|_n^2 \leq \|f - Y\|_n^2 - \left\|\widehat{f} - f\right\|_n^2 \tag{13.58}$$

for any $f \in \mathcal{F}$, and in particular for $f_{\mathcal{F}}$. This is a simple consequence of convexity and pythagorean theorem. The negative quadratic will give us the extra juice we need.

Adding and subtracting f^* on both sides and expanding,

$$\left\|\widehat{f}-f^*\right\|_n^2+\|f^*-Y\|_n^2+2\langle\widehat{f}-f^*,-\eta\rangle_n+\left\|f_{\mathcal{F}}-\widehat{f}\right\|_n^2 \le \|f_{\mathcal{F}}-f^*\|_n^2+\|f^*-Y\|_n^2+2\langle f_{\mathcal{F}}-f^*,-\eta\rangle_n$$

which leads to

$$\left\|\widehat{f}-f^*\right\|_n^2-\|f_{\mathcal{F}}-f^*\|_n^2 \le 2\langle\eta,\widehat{f}-f_{\mathcal{F}}\rangle_n-\left\|\widehat{f}-f_{\mathcal{F}}\right\|_n^2 \tag{13.59}$$

$$\le \sup_{h\in\mathcal{F}-f_{\mathcal{F}}} 2\langle\eta,h\rangle_n-\|h\|_n^2 \tag{13.60}$$

We conclude that for convex $\mathcal{F}$ and fixed design, the upper bounds we find for well-specified and misspecified cases match. Moreover, since the misspecified case is strictly more general and lower bounds for the well-specified case and polynomial entropy growth match the upper bounds, we conclude that constrained least squares are also minimax optimal for fixed design misspecified case.

Note: a crucial observation is that offset complexity would arise even if (13.58) had a different constant multiplier in front of $-\left\|f-\widehat{f}\right\|_n^2$. We will exploit this observation in a bit.

13.2 General $\mathcal{F}$

What if $\mathcal{F}$ is not convex? It turns out that least squares (ERM) can be suboptimal even if $\mathcal{F}$ is a finite class!

13.2.1 A lower bound for ERM (or any proper procedure)

The suboptimality can be illustrated on a very simple example. Suppose $\mathcal{X}=\{x\}$, Y is $\{0,1\}$-valued, and $\mathcal{F}=\{f_0,f_1\}$ such that $f_0(x)=0$ and $f_1(x)=1$. The marginal distribution is the trivial $P_X=\delta_x$ and suppose we have two conditional distributions $P_0(Y=1)=1/2-\alpha$ and $P_1(Y=1)=1/2+\alpha$. Clearly, the population minimizer for P_j is f_j. Also, under P_0 the regression function is $f_0^*=1/2-\alpha$ while under P_1 it is $f_1^*=1/2+\alpha$. Finally, ERM is a method that goes after the most frequent observation in the data $Y_1,\ldots,Y_n$.

However, if $\alpha\propto 1/\sqrt{n}$, there is a constant probability of error in determining whether P_0 or P_1 generated the data. Note that the oracle risk is $\min_{f\in\{f_0,f_1\}}\|f-f_i^*\|^2=(1/2-\alpha)^2$ while the risk of the estimator $p(1/2+\alpha)^2+(1-p)(1/2-\alpha)^2$ where p is the probability of making a mistake and not selecting f_i under the distribution P_i. Hence, the overall comparison to the oracle is at least $p((1/2+\alpha)^2-(1/2-\alpha)^2)=\Omega(\alpha)$ when p is constant.

Hence, ERM (or any "proper" method that selects from $\mathcal{F}$) cannot achieve excess loss smaller than $\Omega(n^{-1/2})$:

$$\max_{P_i\in\{P_0,P_1\}}\left\{\mathbb{E}\left\|\widehat{f}-f_i^*\right\|^2-\min_{f\in\{f_0,f_1\}}\|f-f_i^*\|^2\right\}=\Omega(n^{-1/2})$$

Yet, an improper method that selects $\widehat{f}$ outside $\mathcal{F}$ can achieve an $O(n^{-1})$ rate.

A similar simple lower bound can be constructed for ERM with random design.[1]

[1] For more detailed discussion, we refer to [8].

13.2.2 How about ERM over Convex Hull?

Given that the procedure has to be "improper" (select from outside of $\mathcal{F}$), one can hypothesize that doing ERM over $\mathrm{conv}(\mathcal{F})$ may work. Interestingly, this procedure is also rate-suboptimal for a finite $\mathcal{F}$ since $\mathrm{conv}(\mathcal{F})$ is too expressive.[2]

13.2.3 An improper procedure

Somewhat surprisingly, only a small modification of ERM is required to make it optimal for general classes. Consider the following two-step procedure[3] (*Star Estimator*):

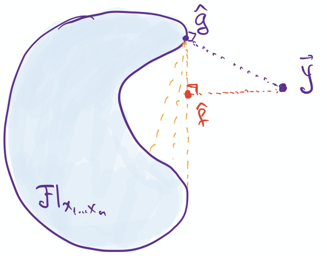
$$\widehat{g} = \operatorname*{argmin}_{f\in\mathcal{F}} \|f - Y\|_n^2 \tag{13.61}$$

$$\widehat{f} = \operatorname*{argmin}_{f\in\mathrm{star}(\mathcal{F},\widehat{g})} \|f - Y\|_n^2 \tag{13.62}$$

where

$$\mathrm{star}(\mathcal{F}, g) = \{\alpha f + (1-\alpha)g : f \in \mathcal{F}, \alpha \in [0, 1]\}.$$

Note that $\widehat{f}$ need not be in $\mathcal{F}$ but is an average of two elements of $\mathcal{F}$.



Note: the method is, in general, different from single ERM over a convex hull of $\mathcal{F}$, and so it is not clear that a version of (13.58) holds [9]:

Lemma: For any $f \in \mathcal{F}$,

$$\|f - Y\|_n^2 - \left\|\widehat{f} - Y\right\|_n^2 \geq \frac{1}{18}\left\|\widehat{f} - f\right\|_n^2. \tag{13.63}$$

The above inequality is an approximate version of (13.58), a generalization of the pythagorean relationship for convex sets.

As a consequence,

$$\left\|\widehat{f} - f^*\right\|_n^2 - \|f_{\mathcal{F}} - f^*\|_n^2 \leq 2\langle\eta, \widehat{f} - f_{\mathcal{F}}\rangle_n - \frac{1}{18}\left\|f_{\mathcal{F}} - \widehat{f}\right\|_n^2$$

and the same upper bounds hold as in the convex case, up to constants. The difference is that the supremum is now in $\mathrm{star}(\mathcal{F}, \widehat{g}) \subseteq \mathcal{F} - f^* + \mathrm{star}(\mathcal{F} - \mathcal{F})$ which is not significantly larger than $\mathcal{F}$ in terms of entropy (unless $\mathcal{F}$ is finite, which can be handled separately).

Remarks:

[2] Proof can be found in Lecué & Mendelson

[3] For a finite class, the above estimator was analyzed by J-Y. Audibert [1].

1. if the set is convex, $\widehat{f} = \widehat{g}$.
2. the Star Estimator can be viewed as one step of Frank-Wolfe. More steps can improve the constant.

Exercise: for any $\varepsilon > 0$ and a set $F \subset \mathbb{R}^n$, the covering numbers satisfy

$$\log \mathcal{N}(F, \|\cdot\|, 2\varepsilon) \leq \log \mathcal{N}(\mathrm{star}(F), \|\cdot\|, 2\varepsilon) \leq \log(\mathrm{diam}(F)/\epsilon) + \log \mathcal{N}(F, \|\cdot\|, \varepsilon)$$

13.3 Offset Rademacher averages

For a set $V \subset \mathbb{R}^n$, the offset process indexed by V is defined as a stochastic process

$$v \mapsto \sum_{i=1}^{n} \epsilon_i v_i - c v_i^2 = \langle \epsilon, v \rangle - c \|v\|^2 .$$

Here ϵ_i are independent Rademacher, but the same results hold for any subGaussian random variables.

Lemma: Let $V \subset \mathbb{R}^n$ be a finite set of vectors, $\mathrm{card}(V) = N$. Then for any $c > 0$,

$$\mathbb{E}_\epsilon \max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 \leq \frac{\log N}{2c}.$$

Furthermore,

$$\mathbb{P}\left(\max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 \geq \frac{1}{2c}(\log N + \log(1/\delta)) \right) \leq \delta$$

Proof. Assuming the random variables are 1-subGaussian,

$$\begin{aligned} \mathbb{E} \max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 &= \frac{1}{\lambda} \mathbb{E} \log \exp \max \lambda \langle \epsilon, v \rangle - \lambda c \|v\|^2 \\ &\leq \frac{1}{\lambda} \log \sum_{v \in V} \mathbb{E} \exp\{\lambda \langle \epsilon, v \rangle - \lambda c \|v\|^2\} \\ &\leq \frac{1}{\lambda} \log \left(N \exp\{\lambda^2 \|v\|^2 /2 - \lambda c \|v\|^2\} \right) \\ &= \frac{1}{2c} \log N \end{aligned}$$

where we chose $\lambda = 2c$. □

Theorem: Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$. Then for any $x_1, \ldots, x_n \in \mathcal{X}$ and the corresponding empirical measure P_n,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i) - cf(x_i)^2 \tag{13.64}$$

$$\leq \inf_{\gamma\geq 0,\alpha\in[0,\gamma]}\left\{\frac{(2/c)\log\mathcal{N}(\mathcal{F},L^2(P_n),\gamma)}{n} + 4\alpha + \frac{12}{\sqrt{n}}\int_{\alpha}^{\gamma}\sqrt{\log\mathcal{N}(\mathcal{F},L^2(P_n),\delta)}d\delta\right\} \tag{13.65}$$

14. TALAGRAND'S INEQUALITY AND APPLICATIONS

The following version of Talagrand's inequality is due to Bousquet:

Theorem: Let $X_1,\ldots,X_n$ be i.i.d., and let $\mathcal{F}=\{f:\mathcal{X}\to[-1,1]\}$. Suppose $\mathbb{E}f(X)=0$ and let

$$\sup_{f\in\mathcal{F}}\mathbb{E}f^2(X)\leq\sigma^2$$

for some $\sigma>0$. Let

$$Z=\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}f(X_i), \qquad v=n\sigma^2+2\mathbb{E}Z$$

Then for any $t\geq 0$,

$$Z\leq\mathbb{E}Z+\sqrt{2tv}+\frac{t}{3}$$

with probability at least $1-e^{-t}$.

Consider a particular case of a singleton $\mathcal{F}=\{f\}$. Then $Z=\sum_{i=1}^{n}f(X_i)$, $\sigma^2=\mathbb{E}f^2$ and $v=n\mathbb{E}f^2$ because $\mathbb{E}Z=\mathbb{E}f=0$. Then the theorem says that

$$\mathbb{P}\left(\sum_{i=1}^{n}f(X_i)\geq\sigma\sqrt{2tn}+\frac{t}{3}\right)\leq e^{-t}$$

which is Bernstein's inequality. Moreover, the constants match those in Bernstein's inequality, which is remarkable.

Now, recall the definition of empirical Rademacher averages. In this lecture we will scale these averages by $1/n$:

$$\widehat{\mathcal{R}}(\mathcal{F})=\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i),$$

conditionally on $X_1,\ldots,X_n$ and its expectation

$$\mathcal{R}(\mathcal{F})=\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F})$$

where the expectation is over the data.

The following holds for Rademacher averages (proof via self-bounding, see [3]):

Theorem: Let $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$. Then

$$\mathbb{P}\left(\widehat{\mathcal{R}}(\mathcal{F}) \geq \mathcal{R}(\mathcal{F}) + \sqrt{\frac{2t\mathcal{R}(\mathcal{F})}{n}} + \frac{t}{3n}\right) \leq e^{-t}$$

In particular, by using the inequality

$$\forall x, y, \lambda > 0, \qquad \sqrt{xy} \leq \frac{\lambda}{2}x + \frac{1}{2\lambda}y,$$

we have

$$\mathbb{P}\left(\widehat{\mathcal{R}}(\mathcal{F}) \geq 2\mathcal{R}(\mathcal{F}) + \frac{5t}{6n}\right) \leq e^{-t}.$$

This and other deviation inequalities for empirical Rademacher averages around their expected value immediately result in data-dependent measures of complexity whenever one can derive a bound in terms of expected (over data) Rademacher averages. Specifically, Talagrand's inequality can be used to relate the random supremum of the empirical process to its expectation; then symmetrization can relate the expected supremum of the empirical process to the expected supremum of the Rademacher process; then above theorem can be employed to relate the latter to the random data-dependent Rademacher averages.

For this lecture, we will note that above theorems are at the heart of proving localization results for random design, both in the well-specified and misspecified settings. We will not flesh out all the details and instead refer to [2]. In particular, in the remainder of this lecture, we would like to develop tools for comparing random and population norms. This will allow us to go from fixed to random design. The tools are also useful more generally.

15. FROM FIXED TO RANDOM DESIGN

Recall that in fixed design regression we aim to prove that for a given set of points $x_1, \ldots, x_n$, an estimator (such as constrained least squares) attains

$$\left\|\widehat{f} - f^*\right\|^2_{L^2(P_n)} \leq \cdots$$

where on the right-hand side we have either a quantity that goes to zero with n or oracle risk as in the misspecified case. We would like to analyze random design regression where $X_1, \ldots, X_n$ are i.i.d from P. Importantly, we also measure the risk through the $L^2(P)$ norm. However,

$$\mathbb{E}\left\|\widehat{f} - f^*\right\|^2_{L^2(P_n)} \neq \mathbb{E}\left\|\widehat{f} - f^*\right\|^2_{L^2(P)}$$

since the algorithm $\widehat{f}$ depends on $X_1, \ldots, X_n$, and so lifting the results from the fixed design case is not straightforward.

Imagine, however, we could prove that with high probability, for all functions $f \in \mathcal{F}$,

$$\|f - f^*\|^2_{L^2(P)} \leq 2\,\|f - f^*\|^2_{L^2(P_n)} + \psi(n, \mathcal{F}). \tag{15.66}$$

In that case, a guarantee for fixed-design regression *would* translate into a guarantee for random design regression as long as $\widehat{f} \in \mathcal{F}$ (for the Star Algorithm, just enlarge $\mathcal{F}$ appropriately). Furthermore, as long as $\psi(n, \mathcal{F})$ decays with n at least as fast as the rate of fixed design regression, we would be able to conclude that random design is not harder than fixed design. Let's see if this can be shown.

Our plan of action for proving results of the form (15.66) is to view the inequality as an instance of a more general uniform comparison

$$\forall g \in \mathcal{G}, \quad \mathbb{E}g(X) \leq \frac{2}{n}\sum_{i=1}^{n} g(X_i) + \psi(n, \mathcal{G})$$

for a class $\mathcal{G}$ of uniformly bounded and *nonnegative* functions.

Let $\hat{\delta}$ satisfy

$$\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}:\frac{1}{n}\sum_{i=1}^{n} g(X_i)\leq\delta^2} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i) \leq \delta^2/2 \tag{15.67}$$

conditionally on $X_1, \ldots, X_n$. Then the following result can be proved from the theorems in the previous section (see e.g. [4]):

Lemma: Let $\mathcal{G}$ be a class of functions with values in $[0, 1]$. Then with probability at least $1 - e^{-t}$ for all $g \in \mathcal{G}$

$$\mathbb{E}g(X) \leq \frac{2}{n}\sum_{i=1}^{n} g(X_i) + c \cdot \hat{\delta}^2 + \frac{c' \cdot (t + \log\log n)}{n} \tag{15.68}$$

where $\hat{\delta} = \hat{\delta}(\mathcal{G})$ is any upper bound on the fixed point in (15.67).

Applying this inequality for the class $\mathcal{G} = \{(f - f')^2 : f, f' \in \mathcal{F}\}$, assuming $\mathcal{F}$ is a class of $[0, 1]$-valued functions, yields

$$\|f - f'\|^2_{L^2(P)} \leq 2\,\|f - f'\|^2_{L^2(P_n)} + c \cdot \hat{\delta}^2 + \frac{c' \cdot (t + \log\log n)}{n}. \tag{15.69}$$

A few remarks. First, $\mathcal{G} = (\mathcal{F} - \mathcal{F})^2$ can be replaced by $(\mathcal{F} - f^*)^2$, even if $f^* \notin \mathcal{F}$, as long as the resulting class is uniformly bounded. Second, we observe that (15.67) is defined with a localization restriction $\frac{1}{n}\sum_{i=1}^{n} g(X_i) \leq \delta^2$ rather than $\frac{1}{n}\sum_{i=1}^{n} g(X_i)^2 \leq \delta^2$ in the previous lecture. Since functions are bounded by 1, the set

$$\widehat{\mathcal{M}} := \left\{ g : \frac{1}{n}\sum_{i=1}^{n} g(X_i) \leq \delta^2 \right\} \subseteq \{\|g\|_n^2 \leq \delta^2\}$$

and hence the set in (15.67) is smaller. Thus the fixed point (15.67) is potentially smaller than the one defined in the previous lecture.

Now, one can ask how to compute a suitable upper bound on the critical radius in (15.67) for particular classes of interest. As in the earlier lectures, the strategy is to upper bound the left-hand side of (15.67) in terms of some more tangible measures of complexity and δ, and then balance with $\delta^2/2$.

In particular, we are interested in the case when $\mathcal{G} = \mathcal{F}^2$ (same analysis works for $(\mathcal{F} - \mathcal{F})^2$ or $(\mathcal{F} - f^*)^2$) for some class $\mathcal{F}$ of $[-1, 1]$-valued functions. In this case, it is tempting to proceed with the help of contraction inequality and upper bound

$$\mathbb{E}_\epsilon \sup_{g\in\mathcal{F}^2\cap\widehat{\mathcal{M}}} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i) \le 2\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}:\|f\|_n^2\le\delta^2} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \tag{15.70}$$

since square is 2-Lipschitz on $[-1,1]$. Balancing this with δ^2 gives, up to constants, the critical radius of $\mathcal{F}$ as defined in previous lectures. Interestingly, one can significantly improve upon this argument and show that the localization radius for $\mathcal{F}^2$ can be smaller than that of $\mathcal{F}$. In particular, a useful result is the following:

Lemma: For any class $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$ of bounded functions, the critical radius in (15.67) for the class $\mathcal{G} = \mathcal{F}^2$ can be upper bounded by a solution to

$$\frac{12}{\sqrt{n}}\int_{\delta/16}^1 \sqrt{\log\mathcal{N}(\mathcal{F}, L^\infty(P_n), u/2))}du \le \delta/4. \tag{15.71}$$

Proof. We start upper bounding the left-hand side of (15.67), aiming to get an upper bound proportional to the scale δ. Observe that functions in $\mathcal{G}$ are nonnegative and bounded uniformly in $[0,1]$. As discussed earlier, the restriction $\frac{1}{n}\sum_{i=1}^n g(X_i) \le \delta^2$ implies $\|g\|_n \le \delta$, and hence the left-hand-side of (15.67) is upper bounded by

$$\inf_\alpha \left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^\delta \sqrt{\log\mathcal{N}(\mathcal{G}\cap\widehat{\mathcal{M}}, L^2(P_n), \varepsilon)d\varepsilon}\right\}. \tag{15.72}$$

Let $V = \{\tilde{f}_1, \ldots, \tilde{f}_N\}$ be a proper $L^\infty(P_n)$-cover of $\mathcal{F}\cap\{\|f\|_n \le \delta\}$ at scale $\tau \le \delta$ (proper implies $\left\|\tilde{f}\right\|_n \le \delta$). Fix any $g = f^2 \in \mathcal{G}\cap\widehat{\mathcal{M}}$. Let $\tilde{f}$ be an element of V that is τ-close to f. Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (f(x_i)^2 - \tilde{f}(x_i)^2)^2 &= \frac{1}{n}\sum_{i=1}^n (f(x_i) - \tilde{f}(x_i))^2(f(x_i) + \tilde{f}(x_i))^2 \\ &\le \max_i (f(x_i) - \tilde{f}(x_i))^2 \cdot \frac{1}{n}\sum_{i=1}^n (f(x_i) + \tilde{f}(x_i))^2 \\ &\le \tau^2(2\|f\|_n^2 + 2\left\|\tilde{f}\right\|_n^2) \\ &\le 4\tau^2\delta^2 := \varepsilon^2\end{aligned}$$

We conclude that

$$\begin{aligned}\mathcal{N}(\mathcal{G}\cap\widehat{\mathcal{M}}, L^2(P_n), \varepsilon) &\le \mathcal{N}(\mathcal{F}\cap\{\|f\|_n \le \delta\}, L^\infty(P_n), \varepsilon/(2\delta)) \\ &\le \mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon/(2\delta))\end{aligned}$$

Substituting into (15.72), the upper bound on the right-hand side becomes

$$\begin{aligned}&\inf_{\alpha\ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^\delta \sqrt{\log\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon/(2\delta))}d\varepsilon\right\} \\ &\le \delta^2/4 + \delta \times \frac{12}{\sqrt{n}}\int_{\delta/16}^1 \sqrt{\log\mathcal{N}(\mathcal{F}, L^\infty(P_n), u/2))}du\end{aligned}$$

where we performed change-of-variables $u = \varepsilon/\delta$ and chose $\alpha = \delta^2/16$. Using this in (15.67) and balancing with $\delta^2/2$ yields (15.71). $\square$

A key outcome of the above lemma is that the critical radius of $\mathcal{F}^2$ (or $(\mathcal{F} - \mathcal{F})^2$) is much smaller than that of $\mathcal{F}$. The latter would have δ^2 rather than δ on the right-hand side of (15.71). In particular, if the left-hand side of (15.71) is of order $1/\sqrt{n}$, the solution is $\delta \propto 1/\sqrt{n}$ and hence the remainder in (15.69) is of the order $1/n$, a smaller order term as compared to the rate of estimation for fixed design. For instance, for a class that exhibits polynomial growth of entropy

$$\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \le \left(\frac{cn}{\varepsilon}\right)^d,$$

the localization radius of $\mathcal{G}$ can be upper bounded as

$$\hat{\delta}(\mathcal{G}) = C\sqrt{\frac{d}{n}\log\left(\frac{cn}{d}\right)}$$

and for a finite class we immediately have

$$\hat{\delta}(\mathcal{G}) \le C\sqrt{\frac{\log|\mathcal{F}|}{n}}.$$

We can also prove a general and useful result, albeit with extra log factors (due to its generality). Following [15], we have

Lemma: For any class $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$, the critical radius in (15.71) is at most

$$C\log^2 n \cdot \bar{\mathcal{R}}(\mathcal{F}),$$

where

$$\bar{\mathcal{R}}(\mathcal{F}) = \sup_{x_1,\ldots,x_n} \widehat{\mathcal{R}}(\mathcal{F}).$$

Proof. Substitute the following estimate for L^∞ covering numbers in terms of the scale-sensitive dimension (see e.g. [14]):

$$\log\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \le 2\mathrm{vc}(\mathcal{F}, c\varepsilon) \cdot \log n \cdot \left(\frac{cn}{\mathrm{vc}(\mathcal{F}, c\varepsilon) \cdot \varepsilon}\right) \tag{15.73}$$

and then use the following fact: for any $\varepsilon > \bar{\mathcal{R}}(\mathcal{F})$,

$$\mathrm{vc}(\mathcal{F}, \varepsilon) \le \frac{4n\bar{\mathcal{R}}(\mathcal{F})^2}{\varepsilon^2}. \tag{15.74}$$

This last inequality can be written in the more familiar form

$$\sup_{\varepsilon > \bar{\mathcal{R}}(\mathcal{F})} \varepsilon\sqrt{\frac{\mathrm{vc}(\mathcal{F}, \varepsilon)}{4n}} \le \bar{\mathcal{R}}(\mathcal{F}), \tag{15.75}$$

which bears similarity to Sudakov's minoration. This inequality is proved by taking the ε-shattered set, replicating it $\lceil n/\mathrm{vc}(\mathcal{F}, \varepsilon)\rceil$ times, and using our previous argument about

Rademacher averages being large when there is a cube inside the set. We leave it as an exercise.

Back to the estimate, we have

$$\frac{1}{\sqrt{n}}\int_{\delta/64}^{1/4}\sqrt{\log\mathcal{N}(\mathcal{F},L^\infty(P_n),\varepsilon))}d\varepsilon \lesssim \frac{\sqrt{\log n}}{\sqrt{n}}\int_{\delta/64}^{1/4}\sqrt{\mathrm{vc}(\mathcal{F},c\varepsilon)\log\left(\frac{cn}{\varepsilon}\right)}d\varepsilon \tag{15.76}$$

$$\lesssim \sqrt{\log n}\bar{\mathcal{R}}(\mathcal{F})\int_{\delta/64}^{1/4}\frac{1}{\varepsilon}\sqrt{\log\left(\frac{cn}{\varepsilon}\right)}d\varepsilon \tag{15.77}$$

To finish the proof, choose $\delta = 64\bar{\mathcal{R}}(\mathcal{F})$ and observe that

$$\int_{\bar{\mathcal{R}}(\mathcal{F})}^{1}\frac{1}{\varepsilon}\sqrt{\log\left(\frac{cn}{\varepsilon}\right)}d\varepsilon \lesssim \log^2(cn/\bar{\mathcal{R}}(\mathcal{F})).$$

□

Hence, ignoring logarithmic factors, $\hat{\delta}(\mathcal{G}) \leq \tilde{O}(n^{-1})$ when $\bar{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/2}$ and $\hat{\delta}(\mathcal{G}) \leq \tilde{O}(n^{-2/p})$ when $\bar{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/p}$, which is *smaller* than the rate of estimation for least squares, ignoring logarithmic factors.

We conclude that rates of estimation for fixed design translate into rates for estimation with random design, at least for bounded functions. It is worth emphasizing that the extra factors one gains from comparing $\|f - f^*\|^2_{L^2(P)}$ to $2\|f - f^*\|^2_{L^2(P_n)}$ is typically of smaller order than what one gets from denoising for fixed design. The next section explains why this happens.

16. BEYOND BOUNDEDNESS: THE SMALL-BALL METHOD

This approach was pioneered by [6] and then developed by Mendelson in a series of papers starting with [10].

Roughly speaking, the realization is that whenever the population norm $\|f\|_{L^2(P)}$ is large enough, it is highly unlikely that the random empirical norm $\|f\|_{L^2(P_n)}$ can be smaller than a fraction of the population norm. Moreover, conditions for such a statement to be true are rather weak and definitely do not require boundedness.

We first recall the Paley-Zygmund inequality (1932) stating that for a nonnegative random variable Z with finite variance,

$$\mathbb{P}(Z \geq t\mathbb{E}Z) \geq (1-t)^2\frac{(\mathbb{E}Z)^2}{\mathbb{E}Z^2}$$

for any $0 \leq t \leq 1$.

Let us use the following shorthand. We will write $\|f\|_2 = \|f\|_{L^2(P)} = (\mathbb{E}f(X)^2)^{1/2}$ and $\|f\|_4 = \|f\|_{L^4(P)} = (\mathbb{E}f(X)^4)^{1/4}$. Then

$$\mathbb{P}(|f(X)| \geq t\|f\|_2) = \mathbb{P}\left(f(X)^2 \geq t^2\|f\|_2^2\right) \geq (1-t^2)^2\frac{\|f\|_2^4}{\|f\|_4^4}$$

Now, we make an assumption that for every $f \in \mathcal{F}$,

$$\mathbb{E}f(X)^4 \leq c(\mathbb{E}f(X)^2)^2$$

for some c.

Under this $L^4 - L^2$ norm comparison, it holds that

$$\mathbb{P}\left(|f(X)| \geq t \left\|f\right\|_2\right) \geq (1-t^2)^2 c$$

More generally, the condition

$$\mathbb{P}\left(|f(X)| \geq c \left\|f\right\|_2\right) \geq c' \tag{16.78}$$

for some c, c' is called the small-ball property.

Let's see how we can compare the empirical and population norms, uniformly over $\mathcal{F}$, given such a condition. First, let's consider any function with norm $\left\|f\right\|_2 = 1$. Observe that if we could show with high probability

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{|f(X_i)| \geq c_1\right\} \geq c_2 \tag{16.79}$$

for some constants c_1, c_2, we would be done since such a lower bound implies a constant lower bound on $\frac{1}{n}\sum_{i=1}^n f(X_i)^2 \geq c_3 \left\|f\right\|_2 = c_3$). By rescaling and assuming star-shapedness, we would extend the result to all functions in $\mathcal{F}$ (above some critical level for which we can prove (16.79)).

For a given $c > 0$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{|f(X_i)| \geq c\right\} &= \mathbb{E}\mathbf{1}\left\{|f(X)| \geq 2c\right\} - \left(\mathbb{E}\mathbf{1}\left\{|f(X)| \geq 2c\right\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{|f(X_i)| \geq c\right\}\right) \\ &\geq \mathbb{E}\mathbf{1}\left\{|f(X)| \geq 2c\right\} - \left(\mathbb{E}\phi(|f(X)|) - \frac{1}{n}\sum_{i=1}^n \phi(|f(X_i)|)\right)\end{aligned}$$

for $\phi(u) = 0$ on $(-\infty, c]$, $\phi(u) = u/c - 1$ on $[c, 2c]$, and $\phi(u) = 1$ on $[2c, \infty)$.

$$\geq \inf_{f\in\mathcal{F}} \mathbb{P}\left(|f(X)| \geq 2c \left\|f\right\|_2\right) - \sup_{f\in\mathcal{F}, \left\|f\right\|_2=1}\left(\mathbb{E}\phi(|f|) - \frac{1}{n}\sum_{i=1}^n \phi(|f(X_i)|)\right)$$

Now, using concentration (since $\phi(|f|)$ are in $[0,1]$), the random supremum

$$\sup_{f\in\mathcal{F}, \left\|f\right\|_2=1}\left(\mathbb{E}\phi(|f|) - \frac{1}{n}\sum_{i=1}^n \phi(|f(X_i)|)\right)$$

can be upper bounded with probability at least $1 - e^{-2u^2}$ by its expectation

$$\mathbb{E}\sup_{f\in\mathcal{F}, \left\|f\right\|_2=1}\left(\mathbb{E}\phi(|f|) - \frac{1}{n}\sum_{i=1}^n \phi(|f(X_i)|)\right) + \frac{u}{\sqrt{n}}$$

which, in turn, can be upper bounded via symmetrization and contraction inequality (since ϕ is $1/c$-Lipschitz) by

$$\frac{4}{c}\mathbb{E}\sup_{f\in\mathcal{F}, \left\|f\right\|_2=1}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) + \frac{u}{\sqrt{n}}$$

By choosing $u = \sqrt{n} \cdot c''$, we can make the additive term an arbitrarily small constant c''. Now, we see that (16.79) will hold with a non-zero constant c_2 as long as

$$\mathbb{E} \sup_{f \in \mathcal{F}, \|f\|_2 = 1} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \leq c''$$

for an appropriately small constant c''. We now need to extend this control to all $\|f\|_2$ above some critical radius. The key observation is that the critical radius β^* can be defined as the smallest β such that

$$\mathbb{E} \sup_{f \in \mathcal{F}, \|f\|_2 \leq \beta} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \leq c'' \beta \tag{16.80}$$

Assuming that $\mathcal{F}$ is star-shaped around 0, the control extends for all $\beta \geq \beta^*$.

To summarize, with probability at least e^{-cn},

$$\inf_{f \in \mathcal{F} : \|f\|_2 \geq \beta^*} \frac{\|f\|_n}{\|f\|_2} \geq c'$$

for some constants c, c'. Alternatively, we have with probability at least e^{-cn}, for all $f \in \mathcal{F}$,

$$\|f\|_2^2 \leq C \|f\|_n^2 + (\beta^*)^2.$$

Observe that β^* can be significantly smaller than if (16.80) were defined with β^2 on the right-hand side, as before.

17. EXAMPLE: INTERPOLATION

Suppose we observe *noiseless* values $y_i = f^*(X_i)$ at i.i.d. locations $X_1, \ldots, X_n$. Let $\widehat{f}$ be an ERM with respect to square loss over $\mathcal{F}$ and assume $f^* \in \mathcal{F}$. Clearly, $\widehat{f}$ achieves zero error, and the question is what the expected deviation from f^* is. This is a question of a "version space size" – what is the $L^2(P)$ diameter of the random subset of $\mathcal{F}$ that matches f^* on a set of data points. More precisely, define the interpolation set

$$\mathcal{I}_{X_1,\ldots,X_n} = \{f \in \mathcal{F} : f(X_i) = f^*(X_i)\},$$

a random subset of the class $\mathcal{F}$, and its diameter as

$$\mathrm{diam}_2(\mathcal{I}_{X_1,\ldots,X_n}) = \sup_{f, f' \in \mathcal{I}_{X_1,\ldots,X_n}} \left\|f - f'\right\|_{L^2(P)}.$$

Of course, from the earlier calculations, we have that with high probability

$$\left\|f - f'\right\|_{L^2(P)} \lesssim \hat{\delta}^2$$

where $\hat{\delta}$ is the localization radius for $(\mathcal{F} - \mathcal{F})^2$ and can be upper bounded by $\sup_{x_{1:n}} \widehat{\mathcal{R}}(\mathcal{F})^2$. Alternatively, we can use the fixed point $(\beta^*)^2$ under the small ball property.

18. EXAMPLE: RANDOM PROJECTIONS AND JOHNSON-LINDENSTRAUSS LEMMA

The development here can be seen as a nonlinear generalization of the random projection method and the Johnson–Lindenstrauss lemma. Let $\Gamma \in \mathbb{R}^{n\times d}$ be an appropriately scaled random matrix. We then prove that for any fixed $v \in \mathbb{R}^d$, with high probability

$$(1-\varepsilon)^2 \|v\|_2^2 \le \|\Gamma v\|_2^2 \le (1+\varepsilon)^2 \|v\|_2^2 .$$

Of particular interest in applications is the lower side of this inequality:

$$\frac{\|\Gamma v\|_2^2}{\|v\|_2^2} \ge 1-\alpha$$

where $\alpha \in (0,1)$. A corresponding *uniform* statement over a set $V \subset \mathbb{R}^d$ asks that with high probability,

$$\inf_{v\in V} \frac{\|\Gamma v\|_2^2}{\|v\|_2^2} \ge 1-\alpha.$$

Statements of this form are very useful in statistics, signal processing, etc. The lower isometry says that the energy of the signal is preserved under random measurement. Or, the null space of the random matrix Γ is likely to miss (in a quantitative way) the set V. Of course, if V is too large, it's not possible to miss it, and so complexity of V (as quantified by the measures we have studied) enters the picture.

The connection to today's lecture can be seen by taking

$$\Gamma = \frac{1}{\sqrt{n}} \begin{pmatrix} -X_1- \\ \dots \\ -X_n- \end{pmatrix}$$

with $X_1, \dots, X_n$ i.i.d. from an isotropic distribution. Then

$$\|\Gamma v\|_2^2 = \frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle^2$$

while $\|v\| = \mathbb{E}_x \langle v, X\rangle^2$. Each $v \in V$ then corresponds to $f \in \mathcal{F}$ in our earlier notation.

19. LARGE MARGIN THEORY

We end this lecture with a result from large margin classification, because its proof utilizes the same technique (not surprisingly, the authors of [7] and [6] have a nonzero intersection).

Let $\mathcal{F}$ be a class of $\mathbb{R}$-valued functions. Consider a classification problem with binary $Y \in \{\pm 1\}$. Fix $\gamma > 0$ as a margin parameter.

Let $\phi : \mathbb{R} \to \mathbb{R}$ be defined by $\phi(a) = 0$ on $(-\infty, 0]$, $\phi(a) = a/\gamma$ on $[0,\gamma]$, and $\phi(a) = 1$ on $[\gamma, \infty)$. Then with probability at least $1 - e^{-2u^2}$, for any $f \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}\mathbf{1}\{Yf(X) \ge 0\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Y_i f(X_i) \ge \gamma\} &\le \sup_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)) - \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i)) \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)) - \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i)) + \frac{u}{\sqrt{n}}\end{aligned}$$

since ϕ is in $[0,1]$. By symmetrization, the above expectation is at most

$$2\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi(Y_if(X_i)) \le \frac{2}{\gamma}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_iY_if(X_i) = \frac{2}{\gamma}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_if(X_i) \le \frac{2}{\gamma}\mathcal{R}(\mathcal{F})$$

Hence, with probability at least $1-e^{-2u^2}$, for any $f\in\mathcal{F}$,

$$\mathbb{E}\mathbf{1}\left\{Yf(X)\ge 0\right\} \le \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{Y_if(X_i)\ge\gamma\right\} + \frac{2}{\gamma}\mathcal{R}(\mathcal{F}) + \frac{u}{\sqrt{n}}$$

As an example, consider the class of linear functions

$$\mathcal{F} = \{x\mapsto\langle x,w\rangle : w\in\mathsf{B}_2^d\}$$

and $\mathcal{X}\in\mathsf{B}_2^d$. We saw earlier that

$$\mathcal{R}(\mathcal{F}) \le \frac{1}{\sqrt{n}}$$

(recall that here we normalized Rademacher averages by $1/n$). Thus, one can derive an upper bound on classification out-of-sample performance that does not depend on the dimensionality of the space despite the fact that the VC dimension of the set of hyperplanes in $\mathbb{R}^d$ is d and covering numbers of $\text{sign}(\mathcal{F})$ necessarily grow with d. Similarly, one can prove margin bounds for neural networks in terms of norms of the weight matrices and without any dependence on the number of neurons.

20. TIME SERIES

Suppose we observe a sequence

$$\boldsymbol{x}_{t+1} = f^*(\boldsymbol{x}_t) + \eta_t, \quad t = 1,\dots,n$$

where $\boldsymbol{x}_t\in\mathbb{R}^d$ and η_t are independent zero mean vectors. The function f^* is unknown, but we assume it is a member of a known class $\mathcal{F}$. Let us treat this problem as a fixed-design regression problem, except that the outcomes are now vectors rather than reals, and the sequence $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ is a sequence of *dependent* random variables.

Consider the least squares solution:

$$\widehat{f}\in\operatorname*{argmin}_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\left\|\boldsymbol{x}_{t+1}-f(\boldsymbol{x}_t)\right\|_2^2,$$

where the norm is the euclidean norm. This is a natural generalization of least squares to vector-valued regression. As before, we denote

$$\|f-g\|_n^2 = \frac{1}{n}\sum_{t=1}^{n}\|f(\boldsymbol{x}_t)-g(\boldsymbol{x}_t)\|_2^2$$

The basic inequality can now be written as (exercise):

$$\left\|\widehat{f}-f^*\right\|_n^2 \le 2\frac{1}{n}\sum_{t=1}^{n}\langle\eta_t,\widehat{f}(\boldsymbol{x}_t)-f^*(\boldsymbol{x}_t)\rangle.$$

Choosing the offset-style approach covered in previous lectures, we have

$$\left\|\widehat{f} - f^*\right\|_n^2 \le \sup_{g \in \mathcal{F}-f^*} \frac{1}{n}\sum_{t=1}^n 4\langle \eta_t, g(\boldsymbol{x}_t)\rangle - \|g(\boldsymbol{x}_t)\|^2 .$$

Up until now, the statement is conditional on $\{\eta_1, \ldots, \eta_n\}$. What happens if we take expectations on both sides? On the left-hand side we have a denoising guarantee on the sequence. On the right-hand side, we have a "dependent version" of offset Gaussian/Rademacher complexity where $\boldsymbol{x}_t$ is measurable with respect to $\sigma(\eta_1, \ldots, \eta_{t-1})$. To analyze this object, we first need to understand the simpler $\mathbb{R}$-valued version without the offset: what is the behavior of

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \epsilon_t f(\boldsymbol{x}_t)$$

where $\boldsymbol{x}_t$ is $\sigma(\epsilon_1, \ldots, \epsilon_{t-1})$-measurable, $\mathcal{F}$ is a class of real-valued functions $\mathcal{X} \to \mathbb{R}$, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables.

21. SEQUENTIAL COMPLEXITIES

We choose to study the random process generated by Rademacher random variables for several reasons. First, just as in the classical case, conditioning on the data will lead to a simpler object (binary tree) and, second, other noise processes can be reduced to the Rademacher case, under moment assumptions on the noise. The development here is based on [13], and we refer also to [12] for an introduction.

Let us elaborate on the first point. Note that $\boldsymbol{x}_t$ being measurable with respect to $\sigma(\epsilon_1, \ldots, \epsilon_{t-1})$ simply means $\boldsymbol{x}_t$ is a function of $\epsilon_1, \ldots, \epsilon_{t-1}$ (in other words, it's a predictable process). Note that the collection $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ can be "summarized" as a depth-n binary tree decorated with elements of $\mathcal{X}$ at the nodes. Indeed, $\boldsymbol{x}_1 \in \mathcal{X}$ is a constant (root), $\boldsymbol{x}_2 = \boldsymbol{x}_2(\epsilon_1)$ takes on two possible values depending on the sign of ϵ_1 (left or right), and so forth. It is useful to think of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ as a tree, even though it doesn't bring any more information into the picture. We shall denote the collection of n functions $\boldsymbol{x}_i : \{\pm 1\}^{i-1} \to \mathcal{X}$ as $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and call it simply as an $\mathcal{X}$-valued *tree.* We shall refer to $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ as a *path* in the tree. We will also talk about $\mathbb{R}$-valued trees, such as $f \circ \boldsymbol{x}$ for $f : \mathcal{X} \to \mathbb{R}$.

Given a tree $\boldsymbol{x}$, we shall call

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}, \boldsymbol{x}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \epsilon_t f(\boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1}))$$

the *sequential Rademacher complexity* of $\mathcal{F}$ on the tree $\boldsymbol{x}$.

Comparing to the classical version,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \epsilon_t f(x_t)$$

where $x_1, \ldots, x_n$ are constant values, we see that it is a special case of a tree with constant levels $\boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1}) = x_t$. Hence, sequential Rademacher complexity is a generalization of the classical notion.

To ease the notation, we will write $\boldsymbol{x}_t$ without explicit dependence on ϵ, or for brevity write $\boldsymbol{x}_t(\epsilon)$ even though $\boldsymbol{x}_t$ only depends on the prefix $\epsilon_{1:t-1}$.

Observe that for any $f \in \mathcal{F}$, the variable

$$\nu_f = \frac{1}{n}\sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t)$$

is zero mean. Moreover, it is an average of martingale differences $\epsilon_t f(\boldsymbol{x}_t)$, and so we expect $1/\sqrt{n}$ behavior from Azuma-Hoeffding's inequality. It should be clear that, say, for $\mathcal{F}$ consisting of a finite collection of $[-1,1]$-valued functions on $\mathcal{X}$, we have

$$\mathbb{E}\max_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t) \le \sqrt{\frac{2\log \operatorname{card}(\mathcal{F})}{n}}$$

Given that there is no difference with the classical case, one may wonder if we can just reduce everything to the classical Rademacher averages. The answer is no, and the differences already start to appear when we attempt to define covering numbers.

More precisely, since any tree $\boldsymbol{x}$ is defined by $2^n - 1$ values, one might wonder if we could define a notion of pseudo-distance between f and f' as an ℓ_2 distance on these $2^n - 1$ values. It is easy to see that this is a huge overkill. Perhaps one of the key points to understand here is: what is the equivalent of the projection $\mathcal{F}|_{x_1,\ldots,x_n}$ for the tree case? Spoiler: it's not $\mathcal{F}|_{\boldsymbol{x}}$. The following turns out to be the right definition:

Definition: A set V of $\mathbb{R}$-valued trees is an 0-cover of $\mathcal{F}$ on a tree $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ if

$$\forall f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad \text{s.t.} \quad f(\boldsymbol{x}_t(\epsilon_{1:t-1})) = \boldsymbol{v}_t(\epsilon_{1:t-1}) \quad \forall t \in [n]$$

The size of the smallest 0-cover of $\mathcal{F}$ on a tree $\boldsymbol{x}$ will be denoted by $\mathcal{N}(\mathcal{F}, \boldsymbol{x}, 0)$.

The key aspect of this definition is that $\boldsymbol{v} \in V$ can be chosen based on the sequence $\epsilon \in \{\pm 1\}^n$. In other words, in contrast with the classical definition, for the same function f different elements $\boldsymbol{v} \in V$ can provide a cover on different paths. This results in the needed reduction in the size of V.

As an example, take a set of 2^{n-1} functions that take a value of 1 on one of the 2^{n-1} leaves of $\boldsymbol{x}$ and zero everywhere else. Then the projection $\mathcal{F}|_{\boldsymbol{x}}$ is of size 2^{n-1} but the size of the 0-cover is only 2, corresponding to our intuition that the class is simple (as it only varies on the last example). Indeed, the size of the 0-cover is the analogue of the size of $\mathcal{F}|_{x_1,\ldots,x_n}$ in the binary-valued case.

For real-valued functions, consider the following definition.

Definition: A set V of $\mathbb{R}$-valued trees is an α-cover of $\mathcal{F}$ on a tree $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ with respect to ℓ_2 if

$$\forall f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad \text{s.t.} \quad \frac{1}{n}\sum_{i=1}^{n} (f(\boldsymbol{x}_t(\epsilon_{1:t-1})) - \boldsymbol{v}_t(\epsilon_{1:t-1}))^2 \le \alpha^2$$

The size of the smallest α-cover of $\mathcal{F}$ on a tree $\boldsymbol{x}$ with respect to ℓ_2 will be denoted by $\mathcal{N}_2(\mathcal{F}, \boldsymbol{x}, \alpha)$.

A similar definition can be stated for cover with respect to ℓ_p.

The following is an analogue of the chaining bound:

Theorem: For any class of $[-1,1]$-valued functions $\mathcal{F}$,

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F},\boldsymbol{x}) \le \inf_{\alpha\ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{\log \mathcal{N}_2(\mathcal{F},\boldsymbol{x},\varepsilon)}d\varepsilon\right\}$$

Recall the definition of VC dimension and a shattered set. Here is the right sequential analogue:

Definition: Function class $\mathcal{F}$ of $\{\pm 1\}$-valued functions shatters a tree $\boldsymbol{x}$ of depth d if

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f\in\mathcal{F}, \quad \text{s.t.} \quad \forall t\in[d], \qquad f(\boldsymbol{x}_t(\epsilon)) = \epsilon_t$$

The largest depth d for which there exists a shattered $\mathcal{X}$-valued tree is called the *Littlestone dimension* and denoted by $\text{ldim}(\mathcal{F})$.

To contrast with the classical definition, the path on which the signs should be realized is given by the path itself. But it's clear that the definition serves the same purpose: if $\boldsymbol{x}$ is shattered by $\mathcal{F}$ then $\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F},\boldsymbol{x}) = 1$. It is also easy to see that $\text{vc}(\mathcal{F}) \le \text{ldim}(\mathcal{F})$, and the gap can be infinite.

The following is an analogue of the Sauer-Shelah-Vapnik-Chervonenkis lemma.

Theorem: For a class of binary-valued functions $\mathcal{F}$ with Littlestone dimension $\text{ldim}(\mathcal{F})$,

$$\mathcal{N}(\mathcal{F},\boldsymbol{x},0) \le \sum_{i=0}^{d}\binom{n}{i} \le \left(\frac{en}{d}\right)^d$$

Scale-sensitive sequential versions are defined as follows:

Definition: Function class $\mathcal{F}$ of $\mathbb{R}$-valued functions shatters a tree $\boldsymbol{x}$ of depth d at scale α if there exists a witness $\mathbb{R}$-valued tree $\boldsymbol{s}$ such that

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f\in\mathcal{F}, \quad \text{s.t.} \quad \forall t\in[d], \qquad \epsilon_t(f(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{s}_t(\epsilon)) \ge \alpha/2$$

The largest depth d for which there exists an α-shattered $\mathcal{X}$-valued tree is called sequential scale-sensitive dimension and denoted $\text{ldim}(\mathcal{F},\alpha)$.

We note that the above definitions reduce to the classical ones if we consider only trees $\boldsymbol{x}$ with constant levels.

Theorem: For any class of $[-1,1]$-valued functions $\mathcal{F}$ and $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n

$$\mathcal{N}_\infty(\mathcal{F},\boldsymbol{x},\alpha) \le \left(\frac{2en}{\alpha}\right)^{\mathrm{ldim}(\mathcal{F},\alpha)}$$

Finally, it is possible to show an analogue of symmetrization lemma: for any joint distribution of $(X_1,\dots,X_n)$,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[f(X_t)|X_{1:t-1}] - f(X_t) \le 2\sup_{\boldsymbol{x}}\widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F},\boldsymbol{x})$$

If the sequence $(X_1,\dots,X_n)$ is i.i.d., the left-hand side is the expected supremum of the empirical process. The present version provides a martingale generalization. Furthermore, if we take supremum over all joint distributions on the left-hand-side, then the lower bound is also matching the upper bound, up to a constant.

The offset Rademacher complexity has been analyzed in [11].

22. ONLINE LEARNING

Consider the following online classification problem. On each of n rounds $t=1,\dots,n$, the learner observes $x_t\in\mathcal{X}$, makes a prediction $\widehat{y}_t\in\{\pm1\}$, and observes the outcome $y_t\in\{\pm1\}$. The learner models the problem by fixing a class $\mathcal{F}$ of possible models $f:\mathcal{X}\to\{\pm1\}$, and aims to predict nearly as well as the best model in $\mathcal{F}$ in the sense of keeping *regret*

$$\mathrm{Reg}(\mathcal{F}) = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}\left\{\widehat{y}_t\neq y_t\right\}\right] - \inf_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}\left\{f(x_t)\neq y_t\right\}\right] \tag{22.81}$$

small for any sequence $(x_1,y_1),\dots,(x_n,y_n)$. At least visually, this looks like oracle inequalities for misspecified models. The distinguishing feature of this online framework is that (a) data arrives sequentially, and (b) we aim to have low regret for any sequence without assuming any generative process.

It is also worth noting that in the above protocol there is no separation of training and test data: the online nature of the problem allows us to first test our current hypothesis by making a prediction, then observe the outcome and incorporate the datum in to our dataset.

The expectation on the first term in (22.81) is with respect to learner's internal randomization. More specifically, let Q_t be the distribution on $\{\pm1\}$ that the learner uses to predict $\widehat{y}_t\sim Q_t$. Let $q_t=\mathbb{E}\widehat{y}_t$ be the (conditional) mean of this distribution. In other words, $q_t=0$ would correspond to the learner tossing a fair coin.

A note about the protocol. The results below hold even if the sequence is chosen based on learner's past predictions. However, in this case, y_t may only depend on q_t but not on the realization $\widehat{y}_t$. To simplify the presentation, let us just assume that the sequence $(x_1,y_1),\dots,(x_n,y_n)$ is fixed in advanced (this turns out not to matter).

We will answer the following question: what is the best achievable $\mathrm{Reg}(\mathcal{F})$ for a given $\mathcal{F}$ by any prediction strategy?

Let us first rewrite $\mathbf{1}\{\widehat{y}_t \neq y_t\} = (1 - \widehat{y}_t y_t)/2$ and do the same for the oracle term. Cancelling 1/2, we have

$$2\mathrm{Reg}(\mathcal{F}) = \frac{1}{n}\sum_{t=1}^{n} -q_t y_t - \inf_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} -y_t f(x_t)\right] \tag{22.82}$$

$$= \sup_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} y_t f(x_t)\right] - \frac{1}{n}\sum_{t=1}^{n} q_t y_t \tag{22.83}$$

Now, consider a particular stochastic process for generating the data sequence: fix any $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n, and on round t let $x_t = \boldsymbol{x}_t(y_1, \ldots, y_{t-1})$ and $y_t = \epsilon_t$ be an independent Rademacher random variable. This defines a stochastic process with 2^n possible sequences $(x_1, y_1), \ldots, (x_n, y_n)$. Now, clearly

$$\sup_{(x_1,y_1),\ldots,(x_n,y_n)} 2\mathrm{Reg}(\mathcal{F}) \geq 2\mathbb{E}_\epsilon \mathrm{Reg}(\mathcal{F}).$$

Observe that $q_t = q_t(\epsilon_1, \ldots, \epsilon_{t-1})$ and thus

$$\mathbb{E}_\epsilon \left[\frac{1}{n}\sum_{t=1}^{n} q_t \epsilon_t\right] = 0.$$

Hence,

$$\mathbb{E}_\epsilon \mathrm{Reg}(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t)\right]. \tag{22.84}$$

Since the argument holds for any $\boldsymbol{x}$, we have proved that the optimal value of $\mathrm{Reg}(\mathcal{F})$ is lower bounded by half of

$$\bar{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}) = \sup_{\boldsymbol{x}} \widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}, \boldsymbol{x}).$$

It turns out that this lower bound is within a factor of 2 from optimal. Define the minimax value

$$\mathcal{V} = \min_{\text{Algo}} \max_{\{(x_t,y_t)\}_{t=1}^n} \mathrm{Reg}(\mathcal{F})$$

Theorem: For a binary-valued class $\mathcal{F}$,

$$\frac{1}{2}\bar{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}) \leq \mathcal{V} \leq \bar{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F})$$

Similar results also holds for absolute value and other Lipschitz loss functions. For square loss, the sequential Rademacher averages are replaced by offset sequential Rademacher averages (again, as both upper and lower bounds).

In short, sequential complexities in online learning play a role similar to the role played by i.i.d. complexities as studied in this course. However, quite a large number of questions still remains open. But that's a topic for a different course.

References

[1] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2008.

[2] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[4] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Biologische Kybernetik, 2002.

[5] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

[6] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23): 12991–13008, 2015.

[7] V. Koltchinskii, D. Panchenko, et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[8] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

[9] T. Liang, A. Rakhlin, and K. Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

[10] S. Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[11] A. Rakhlin and K. Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

[12] A. Rakhlin and K. Sridharan. On martingale extensions of Vapnik–Chervonenkis theory with applications to online learning. In *Measures of Complexity*, pages 197–215. Springer, 2015.

[13] A. Rakhlin, K. Sridharan, and A. Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

[14] M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

[15] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

[16] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.