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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 24 & 25
May 5 & 7, 2020

1. TALAGRAND'S INEQUALITY AND APPLICATIONS

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

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

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

for some $\sigma > 0$. Let

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

Then for any $t \ge 0$,

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

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

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

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

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

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

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

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

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

where the expectation is over the data.

The following holds for Rademacher averages (proof via self-bounding, see [2]):

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

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

In particular, by using the inequality

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

we have

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

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

For this lecture, we will note that above theorems are at the heart of proving localization results for random design, both in the well-specified and misspecified settings. We will not flesh out all the details and instead refer to [1]. In particular, in the remainder of this lecture, we would like to develop tools for comparing random and population norms. This will allow us to go from fixed to random design. The tools are also useful more generally.

2. FROM FIXED TO RANDOM DESIGN

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

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

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

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

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

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

$$\|f - f^*\|^2_{L^2(P)} \leq 2\|f - f^*\|^2_{L^2(P_n)} + \psi(n, \mathcal{F}). \tag{2.1}$$

In that case, a guarantee for fixed-design regression *would* translate into a guarantee for random design regression as long as $\widehat{f} \in \mathcal{F}$ (for the Star Algorithm, just enlarge $\mathcal{F}$ appropriately). Furthermore, as long as $\psi(n, \mathcal{F})$ decays with n at least as fast as the rate of fixed

design regression, we would be able to conclude that random design is not harder than fixed design. Let's see if this can be shown.

Our plan of action for proving results of the form (2.1) is to view the inequality as an instance of a more general uniform comparison

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

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

Let $\hat{\delta}$ satisfy

$$\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}:\frac{1}{n}\sum_{i=1}^{n} g(X_i)\leq\delta^2} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i) \leq \delta^2/2 \tag{2.2}$$

conditionally on $X_1, \ldots, X_n$. Then the following result can be proved from the theorems in the previous section (see e.g. [3]):

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

$$\mathbb{E}g(X) \leq \frac{2}{n}\sum_{i=1}^{n} g(X_i) + c \cdot \hat{\delta}^2 + \frac{c' \cdot (t + \log\log n)}{n} \tag{2.3}$$

where $\hat{\delta} = \hat{\delta}(\mathcal{G})$ is any upper bound on the fixed point in (2.2).

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

$$\left\|f - f'\right\|^2_{L^2(P)} \leq 2\left\|f - f'\right\|^2_{L^2(P_n)} + c \cdot \hat{\delta}^2 + \frac{c' \cdot (t + \log\log n)}{n}. \tag{2.4}$$

A few remarks. First, $\mathcal{G} = (\mathcal{F} - \mathcal{F})^2$ can be replaced by $(\mathcal{F} - f^*)^2$, even if $f^* \notin \mathcal{F}$, as long as the resulting class is uniformly bounded. Second, we observe that (2.2) is defined with a localization restriction $\frac{1}{n}\sum_{i=1}^{n} g(X_i) \leq \delta^2$ rather than $\frac{1}{n}\sum_{i=1}^{n} g(X_i)^2 \leq \delta^2$ in the previous lecture. Since functions are bounded by 1, the set

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

and hence the set in (2.2) is smaller. Thus the fixed point (2.2) is potentially smaller than the one defined in the previous lecture.

Now, one can ask how to compute a suitable upper bound on the critical radius in (2.2) for particular classes of interest. As in the earlier lectures, the strategy is to upper bound the left-hand side of (2.2) in terms of some more tangible measures of complexity and δ, and then balance with $\delta^2/2$.

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

tempting to proceed with the help of contraction inequality and upper bound

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{F}^2 \cap \widehat{\mathcal{M}}} \frac{1}{n} \sum_{i=1}^n \epsilon_i g(X_i) \leq 2\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}: \|f\|_n^2 \leq \delta^2} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \tag{2.5}$$

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

Lemma: For any class $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$ of bounded functions, the critical radius in (2.2) for the class $\mathcal{G} = \mathcal{F}^2$ can be upper bounded by a solution to

$$\frac{12}{\sqrt{n}} \int_{\delta/16}^1 \sqrt{\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), u/2)} du \leq \delta/4. \tag{2.6}$$

Proof. We start upper bounding the left-hand side of (2.2), aiming to get an upper bound proportional to the scale δ. Observe that functions in $\mathcal{G}$ are nonnegative and bounded uniformly in $[0, 1]$. As discussed earlier, the restriction $\frac{1}{n} \sum_{i=1}^n g(X_i) \leq \delta^2$ implies $\|g\|_n \leq \delta$, and hence the left-hand-side of (2.2) is upper bounded by

$$\inf_\alpha \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\delta \sqrt{\log \mathcal{N}(\mathcal{G} \cap \widehat{\mathcal{M}}, L^2(P_n), \varepsilon)} d\varepsilon \right\}. \tag{2.7}$$

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

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

We conclude that

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

Substituting into (2.7), the upper bound on the right-hand side becomes

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

where we performed change-of-variables $u = \varepsilon/\delta$ and chose $\alpha = \delta^2/16$. Using this in (2.2) and balancing with $\delta^2/2$ yields (2.6). □

A key outcome of the above lemma is that the critical radius of $\mathcal{F}^2$ (or $(\mathcal{F}-\mathcal{F})^2$) is much smaller than that of $\mathcal{F}$. The latter would have δ^2 rather than δ on the right-hand side of (2.6). In particular, if the left-hand side of (2.6) is of order $1/\sqrt{n}$, the solution is $\delta \propto 1/\sqrt{n}$ and hence the remainder in (2.4) is of the order $1/n$, a smaller order term as compared to the rate of estimation for fixed design. For instance, for a class that exhibits polynomial growth of entropy

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

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

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

and for a finite class we immediately have

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

We can also prove a general and useful result, albeit with extra log factors (due to its generality). Following [8], we have

Lemma: For any class $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$, the critical radius in (2.6) is at most

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

where

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

Proof. Substitute the following estimate for L^∞ covering numbers in terms of the scale-sensitive dimension (see e.g. [7]):

$$\log\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \leq 2\mathrm{vc}(\mathcal{F}, c\varepsilon)\cdot\log n\cdot\left(\frac{cn}{\mathrm{vc}(\mathcal{F}, c\varepsilon)\cdot\varepsilon}\right) \tag{2.8}$$

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

$$\mathrm{vc}(\mathcal{F}, \varepsilon) \leq \frac{4n\bar{\mathcal{R}}(\mathcal{F})^2}{\varepsilon^2}. \tag{2.9}$$

This last inequality can be written in the more familiar form

$$\sup_{\varepsilon > \bar{\mathcal{R}}(\mathcal{F})} \varepsilon\sqrt{\frac{\mathrm{vc}(\mathcal{F},\varepsilon)}{4n}} \leq \bar{\mathcal{R}}(\mathcal{F}), \tag{2.10}$$

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

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

Back to the estimate, we have

$$\frac{1}{\sqrt{n}}\int_{\delta/64}^{1/4}\sqrt{\log \mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon))}d\varepsilon \lesssim \frac{\sqrt{\log n}}{\sqrt{n}}\int_{\delta/64}^{1/4}\sqrt{\mathrm{vc}(\mathcal{F}, c\varepsilon)\log\Big(\frac{cn}{\varepsilon}\Big)}d\varepsilon \tag{2.11}$$

$$\lesssim \sqrt{\log n}\bar{\mathcal{R}}(\mathcal{F})\int_{\delta/64}^{1/4}\frac{1}{\varepsilon}\sqrt{\log\Big(\frac{cn}{\varepsilon}\Big)}d\varepsilon \tag{2.12}$$

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

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

□

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

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

3. BEYOND BOUNDEDNESS: THE SMALL-BALL METHOD

This approach was pioneered by [4] and then developed by Mendelson in a series of papers starting with [6].

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

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

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

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

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

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

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

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

for some c.

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

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

More generally, the condition

$$\mathbb{P}\left(|f(X)| \geq c \|f\|_2\right) \geq c' \tag{3.13}$$

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

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

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{|f(X_i)| \geq c_1\right\} \geq c_2 \tag{3.14}$$

for some constants c_1, c_2, we would be done since such a lower bound implies a constant lower bound on $\frac{1}{n}\sum_{i=1}^n f(X_i)^2 \geq c_3 \|f\|_2 = c_3$). By rescaling and assuming star-shapedness, we would extend the result to all functions in $\mathcal{F}$ (above some critical level for which we can prove (3.14)).

For a given $c > 0$, we have

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

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

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

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

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

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

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

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

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

By choosing $u = \sqrt{n} \cdot c''$, we can make the additive term an arbitrarily small constant c''. Now, we see that (3.14) will hold with a non-zero constant c_2 as long as

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

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

$$\mathbb{E} \sup_{f \in \mathcal{F}, \|f\|_2 \leq \beta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \leq c''\beta \tag{3.15}$$

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

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

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

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

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

Observe that β^* can be significantly smaller than if (3.15) were defined with β^2 on the right-hand side, as before.

4. EXAMPLE: INTERPOLATION

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

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

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

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

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

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

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

5. EXAMPLE: RANDOM PROJECTIONS AND JOHNSON-LINDENSTRAUSS LEMMA

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

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

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

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

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

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

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

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

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

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

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

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

6. LARGE MARGIN THEORY

We end this lecture with a result from large margin classification, because its proof utilizes the same technique (not surprisingly, the authors of [5] and [4] have a nonzero intersection).

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

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

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

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

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

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

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

As an example, consider the class of linear functions

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

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

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

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

References

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

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

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

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

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

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

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

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