MATH 281C: Mathematical Statistics

Lecture 5

1 Bounds for the Expected Suprema

The next major topic of the course involves bounding the quantity

$$\mathbb{E} \sup_{f\in\mathcal{F}} |P_n f - Pf|. \tag{5.1}$$

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

Symmetrization bounds (5.1) from above using the *Rademacher complexity* of the class $\mathcal{F}$. Let us first denote the Rademacher complexity. A Rademacher random variable is a random variable that takes the two values $+1$ and -1 with probability $1/2$ each. For a subset $A \subseteq \mathbb{R}^n$, its Rademacher average is defined by

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

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

Example 1.1. For $A = \{(1, \dots, 1)\} \subseteq \mathbb{R}^n$, we have $R_n(A) = \mathbb{E}|(1/n)\sum_{i=1}^n \epsilon_i| \approx \Theta(n^{-1/2})$.

Example 1.2. Let $A = \{-1, 1\}^n$ with cardinality $|A| = 2^n$. For each realization $(\epsilon_1, \dots, \epsilon_n) \in \{-1, 1\}^n$, the maximum $|(1/n)\sum_{i=1}^n \epsilon_i a_i|$ is achieved at $a_i = \epsilon_i$ for all i. This implies $R_n(A) = \mathbb{E}\sup_{a\in A} |(1/n)\sum_{i=1}^n \epsilon_i a_i| = 1$.

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

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

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

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

It is easy to see that

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

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

The next result shows that the expectation in (5.1) is bounded from above by twice the Rademacher complexity $R_n(\mathcal{F})$.

Theorem 1.1 (Symmetrization)**.** We have

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

where the expectation on the left-hand side is taken with respect to $X_1, \dots, X_n$ that are iid with distribution P, while the expectation on the right-hand side is taken with respect to both X_i's and independent Rademachers ϵ_i's.

Proof. Let $X_1', \dots, X_n'$ be random variables that are independent copies of $X_1, \dots, X_n$. In other words, $X_1, \dots, X_n, X_1', \dots, X_n'$ are iid with distribution P. We can then write

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

As a result, we have

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

The method used above is called symmetrization. We now introduce iid Rademacher variables $\epsilon_1, \dots, \epsilon_n$. Because X_i is an independent copy of X_i', it is clear that the distribution of $f(X_i) - f(X_i')$ is the same as that of $\epsilon_i \{ f(X_i) - f(X_i') \}$. As a result, we have

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

□

Theorem 1.1 implies that we can control (5.1) by bounding $R_n(\mathcal{F})$ from above. The usual strategy used for bounding $R_n(\mathcal{F})$ is the following. One first fixes points $x_1, \dots, x_n \in \mathcal{X}$, and bounds the Rademacher average of the set

$$\mathcal{F}(x_1, \dots, x_n) := \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}. \tag{5.2}$$

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

The main technique for bounding $R_n(\mathcal{F}(x_1, \dots, x_n))$ will be *chaining*. Before we get to chaining, let us first look at a more elementary bound that works well in certain situations for Boolean classes $\mathcal{F}$ (i.e., $f(x) \in \{0, 1\}$). As we will see later, this bound will not be as accurate/sharp as the bounds given by chaining.

2 Simple Bounds on the Rademacher Average $R_n(\mathcal{F}(x_1, \dots, x_n))$

These bounds are based on the following simple result.

Proposition 2.1. Suppose A is a finite subset of $\mathbb{R}^n$ with cardinality $|A|$. Then

$$R_n(A) = \mathbb{E} \max_{a \in A} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i a_i \right| \le \sqrt{6} \sqrt{\frac{\log(2|A|)}{n}} \max_{a \in A} \sqrt{\frac{1}{n} \sum_{i=1}^n a_i^2}. \tag{5.3}$$

Proof. For every nonnegative random variable X, one has

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)\, \mathrm{d}t,$$

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

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

$$\begin{aligned}
\mathbb{E} \exp\left\{ \frac{(\sum_{i=1}^n a_i \epsilon_i)^2}{6 \sum_{i=1}^n a_i^2} \right\} &= \int_0^\infty \mathbb{P}\left[\exp\left\{ \frac{(\sum_{i=1}^n a_i \epsilon_i)^2}{6 \sum_{i=1}^n a_i^2} \right\} > t \right] \mathrm{d}t \\
&\le 1 + \int_1^\infty \mathbb{P}\left(\left| \sum_{i=1}^n a_i \epsilon_i \right| > \sqrt{6 \sum_{i=1}^n a_i^2} \sqrt{\log t} \right) \mathrm{d}t \\
&\le 1 + 2 \int_1^\infty \exp\left\{ -\frac{6 \log(t) \sum_{i=1}^n a_i^2}{2 \sum_{i=1}^n a_i^2} \right\} \mathrm{d}t \quad \text{(Hoeffding's inequality)} \\
&= 1 + 2 \int_1^\infty t^{-3}\, \mathrm{d}t = 2.
\end{aligned}$$

From the above, we have

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

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

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

Note that the function $x\mapsto e^{x^2}$ is convex, applying Jensen's inequality yields ($e^{(\mathbb{E}Y)^2}\le \mathbb{E}e^{Y^2}$)

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

so that

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

From here, the inequality given in (5.3) follows by the trivial inequality:

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

□

Alternatively, we can also prove Proposition 2.1 by directly using moment generating function.

Alternative Proof of Proposition 2.1. In this proof, we will use a basic inequality: for any $x\in\mathbb{R}$,

$$\frac{1}{2}(e^x+e^{-x})\le e^{x^2/2}.$$

This inequality is easily proved by comparing the coefficients of Taylor series of both sides. For $a=(a_1,\dots,a_n)$, let $Z_a=(1/n)\sum_{i=1}^n \epsilon_i a_i$, and consider its MGF: for $\lambda\in\mathbb{R}$,

$$\mathbb{E}e^{\lambda R_n(A)} = e^{\lambda\mathbb{E}\max_{a\in A}|Z_a|} \overset{\text{(i)}}{\le} \mathbb{E}e^{\lambda\max_{a\in A}|Z_a|} = \mathbb{E}\max_{a\in A}e^{\lambda|Z_a|} \le \mathbb{E}\sum_{a\in A}(e^{\lambda Z_a}+e^{-\lambda Z_a}) \overset{\text{(ii)}}{=} 2\sum_{a\in A}\mathbb{E}e^{\lambda Z_a},$$

where inequality (i) is based on Jensen's inequality, and equality (ii) uses the symmetry property that $Z_a\overset{d}{=}-Z_a$. Next, using the independence of ϵ_i and the basic inequality we obtain

$$\mathbb{E}e^{\lambda Z_a}=\prod_{i=1}^n \mathbb{E}e^{\lambda\epsilon_i a_i/n}=\prod_{i=1}^n\frac{1}{2}(e^{\lambda a_i/n}+e^{-\lambda a_i/n})\le\prod_{i=1}^n e^{\lambda^2 a_i^2/(2n^2)}=e^{\lambda^2\sum_{i=1}^n a_i^2/(2n^2)}.$$

Combine both equations, we immediately have $e^{\lambda R_n(A)}\le 2|A|e^{\lambda^2\max_{a\in A}\sum_{i=1}^n a_i^2/(2n^2)}$. Taking logarithm on both sides yields

$$R_n(A)\le\frac{\log(2|A|)}{\lambda}+\frac{\lambda\max_{a\in A}\sum_{i=1}^n a_i^2}{2n^2},\quad\text{valid for any }\lambda>0.$$

Picking the optimal

$$\lambda^*=\sqrt{\frac{2n^2\log(2|A|)}{\max_{a\in A}\sum_{i=1}^n a_i^2}}$$

to minimize the RHS finishes the proof. □

Let us now apply Proposition 2.1 to control the Rademacher complexity of Boolean Function Classes. We say that $\mathcal{F}$ is a Boolean class if $f(x)$ takes only the two values 0 and 1 for every function f and every $x \in \mathcal{X}$. Boolean classes $\mathcal{F}$ arise in the problem of classification (where $\mathcal{F}$ can be taken to consist of all functions f of the form $I\{g(X) \neq Y\}$). They are also important for historical reasons: empirical process theory has its origins in the study of $\sup_x\{F_n(x) - F(x)\}$, which corresponds to taking $\mathcal{F} = \{I(-\infty, t] : t \in \mathbb{R}\}$.

Let us now fix a Boolean class $\mathcal{F}$ and points $x_1, \ldots, x_n$. The set $\mathcal{F}(x_1, \ldots, x_n)$ (defined in (5.2)) is obviously finite, so that we can apply Proposition 2.1 to control $R_n(\mathcal{F}(x_1, \ldots, x_n))$. This gives

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

Because $\mathcal{F}$ is Boolean, we can simply bound each $f^2(x_i)$ by 1 to obtain

$$R_n(\mathcal{F}(x_1, \ldots, x_n)) \leq \sqrt{\frac{6\log(2|\mathcal{F}(x_1, \ldots, x_n)|)}{n}}. \tag{5.4}$$

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

Definition 2.1. The class of Boolean functions $\mathcal{F}$ is said to have **polynomial discrimination** if there exists a polynomial $\rho(\cdot)$ such that for every $n \geq 1$ and every set of n points $x_1, \ldots, x_n$ in $\mathcal{X}$, the cardinality of $\mathcal{F}(x_1, \ldots, x_n)$ is at most $\rho(n)$.

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

Definition 2.2 (VC dimension)**.** The VC dimension of a class of Boolean functions $\mathcal{F}$ on $\mathcal{X}$ is defined as the maximum integer D for which there exists a finite subset $\{x_1, \ldots, x_D\}$ of $\mathcal{X}$ satisfying

$$\mathcal{F}(x_1, \ldots, x_D) = \{0, 1\}^D, \quad \text{or equivalently,} \quad |\mathcal{F}(x_1, \ldots, x_D)| = 2^D.$$

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

Example 2.1. For Boolean function class $\mathcal{F} = \{I(-\infty, t] : t \in \mathbb{R}\}$, its VD dimension is 1. Since we can easily verify that for $x_1 = 0$, $\mathcal{F}(x_1) = \{0, 1\}$; for any x_1, x_2 (w.l.o.g. assume $x_1 \leq x_2$), then $(0, 1) \notin \mathcal{F}(x_1, x_2)$.

Definition 2.3 (Shattering)**.** A finite subset $\{x_1, \ldots, x_m\}$ of $\mathcal{X}$ is said to be **shattered** by the Boolean class $\mathcal{F}$ if $\mathcal{F}(x_1, \ldots, x_m) = \{0, 1\}^m$. By convention, we extend the definition of shattering to empty subsets as well by saying that the empty set is shattered by every nonempty class $\mathcal{F}$.

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

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

Lemma 2.1 (Sauer-Shelah-Vapnik-Chervonenkis)**.** Suppose that the the VC dimension of a Boolean class $\mathcal{F}$ of functions on $\mathcal{X}$ is D. Then for every $n \geq 1$ and $x_1, \ldots, x_n \in \mathcal{X}$, we have

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

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

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

Combining (5.4) with Lemma 2.1, we obtain the following bound on the Rademacher complexity and expected suprema for Boolean classes with finite VC dimension.

Proposition 2.2. Suppose $\mathcal{F}$ is a Boolean function class with VC dimension D. Then

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

Here C is a universal positive constant.

Remark 2.1. It turns out that the logarithmic term is not needed in the bounds given by the above proposition. We will see later that the bounds given by *chaining* do not have the superfluous logarithmic factor.

We leave the proof of Lemma 2.1 to the next lecture. Here we give two examples of Boolean classes with finite VC dimension.

Example 2.2. Let $\mathcal{V}$ be a D-dimensional vector space of real functions on $\mathcal{X}$. Let $\mathcal{F} := \{I(f \geq 0) : f \in \mathcal{V}\}$. Then VC dimension of $\mathcal{F}$ is at most D.

Proof. For any $D+1$ points $\{x_1, \ldots, x_{D+1}\}$, consider the set $T = \{(f(x_1), \ldots, f(x_{D+1}) : f \in \mathcal{V}\}$. Since $\mathcal{V}$ is a D-dimensional vector space, T is a linear subspace of $\mathbb{R}^{D+1}$ with dimension at most D. Therefore, there exists $y \in \mathbb{R}^{D+1}$ and $y \neq 0$ such that y is orthogonal to the subspace T, i.e.,

$$\sum_i y_i f(x_i) = 0 \quad \text{for all} \quad f \in \mathcal{V}. \tag{5.5}$$

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

$$f(x_i) < 0 \;\; (\Leftrightarrow I\{f(x_i) \geq 0\} = 0) \quad \text{for all } i \text{ such that } \; y_i > 0;$$
$$f(x_i) \geq 0 \;\; (\Leftrightarrow I\{f(x_i) \geq 0\} = 1) \quad \text{for all } i \text{ such that } \; y_i \leq 0.$$

Then we have $\sum_i y_i f(x_i) < 0$, which is a contradiction to (5.5). Thus $\mathcal{F}$ cannot shatter $\{x_1, \ldots, x_{D+1}\}$, and so the VC dimension is at most D. □

Example 2.3. Let $\mathcal{H}_k$ denote the indicators of all closed half-spaces in $\mathbb{R}^k$, i.e. $\mathcal{H}_k = \{x \mapsto I(\langle a, x\rangle + b \leq 0) : a \in \mathbb{R}^k, b \in \mathbb{R}\}$. The VC dimension of $\mathcal{H}_k$ is exactly equal to $k+1$.

Example 2.4 (Spheres in $\mathbb{R}^k$)**.** Consider the sphere $S_{a,b} = \{x \in \mathbb{R}^k : \|x - a\|_2 \leq b\}$, where $(a, b) \in \mathbb{R}^k \times \mathbb{R}_+$ specify its center and radius, respectively. Define the function $f_{a,b}(x) = \|x\|_2^2 - 2\sum_{j=1}^k a_j x_j + \|a\|_2^2 - b^2$, so that $S_{a,b} = \{x \in \mathbb{R}^k : f_{a,b}(x) \leq 0\}$. Let $\mathcal{S}_k = \{x \mapsto I\{f_{a,b}(x) \leq 0\} : a \in \mathbb{R}^k, b \geq 0\}$. The VC dimension of $\mathcal{S}_k$ is at most $k+2$.

We leave the verification of the two examples above as homework.