MATH 281C: Mathematical Statistics

Lecture 8

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

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

Then for every $t_0 \in T$,

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

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

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

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

where $\sigma_n = \sup_{t\in T}\|t\|_n$, $\|t\|_n = \sqrt{(1/n)\sum_{i=1}^n t_i^2}$ and $d_n(s,t) = \|s - t\|_n$. Note that if T is finite, then

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

and hence

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

This coincides with the bound on the expected maxima of sub-Gaussian random variables.

We next apply Dudley's entropy bound together with symmetrization to obtain our main bound for the expected suprema of an empirical process.

1 Main Bound on the Expected Suprema of Empirical Processes

Consider the usual empirical process setup. Our goal is to obtain upper bounds on Δ where

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

By symmetrization,

$$\begin{aligned}\Delta &\le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{f(X_i)-\mathbb{E}f(X_i)\}\right| \\ &= 2\mathbb{E}\left[\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\middle|X_1,\dots,X_n\right\}\right].\end{aligned}$$

The inner expectation above can be controlled via Dudley's entropy bound, giving

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

where $\mathcal{F}(X_1,\dots,X_n)=\{(f(X_1),\dots,f(X_n)) : f\in\mathcal{F}\}$ is a subset of $\mathbb{R}^n$, $\sigma_n=\sup_{f\in\mathcal{F}}\sqrt{P_nf^2}$ and d_n is the Euclidean metric on $\mathbb{R}^n$ scaled by $n^{-1/2}$.

We write

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

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

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

By the trivial inequality

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

We thus obtain

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

Taking expectations on both sides yields

$$\mathbb{E}\sup_{f\in\mathcal{F}}n^{1/2}|P_nf-Pf|\le C\mathbb{E}\left\{\int_0^{\sup_{f\in\mathcal{F}}\sqrt{P_nf^2}}\sqrt{1+\log M(\epsilon,\mathcal{F},L^2(P_n))}\,\mathrm{d}\epsilon\right\}.$$

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

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

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

$$
\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}} n^{1/2}|P_n f - Pf| &\le C\mathbb{E}\left\{\int_0^{\sup_{f\in\mathcal{F}}\sqrt{P_n f^2}} \sqrt{1+\log M(\epsilon,\mathcal{F},L^2(P_n))}\,\mathrm{d}\epsilon\right\}\\
&\le C\mathbb{E}\left\{\int_0^{\sqrt{P_nF^2}} \sqrt{1+\log M(\epsilon,\mathcal{F},L^2(P_n))}\,\mathrm{d}\epsilon\right\}\\
&\le C\mathbb{E}\left\{\sqrt{P_nF^2}\int_0^{1} \sqrt{1+\log M(\epsilon\sqrt{P_nF^2},\mathcal{F},L^2(P_n))}\,\mathrm{d}\epsilon\right\}\\
&\le C\mathbb{E}\left\{\sqrt{P_nF^2}\int_0^{1} \sqrt{1+\log \sup_Q M(\epsilon\sqrt{QF^2},\mathcal{F},L^2(Q))}\,\mathrm{d}\epsilon\right\}\\
&\le C\underbrace{\left\{\int_0^{1} \sqrt{1+\log \sup_Q M(\epsilon\sqrt{QF^2},\mathcal{F},L^2(Q))}\,\mathrm{d}\epsilon\right\}}_{\text{non-random}}\mathbb{E}\sqrt{P_nF^2}\\
&\le C\left\{\int_0^{1} \sqrt{1+\log \sup_Q M(\epsilon\sqrt{QF^2},\mathcal{F},L^2(Q))}\,\mathrm{d}\epsilon\right\}\sqrt{\mathbb{E}P_nF^2}\\
&= C\left\{\int_0^{1} \sqrt{1+\log \sup_Q M(\epsilon\sqrt{QF^2},\mathcal{F},L^2(Q))}\,\mathrm{d}\epsilon\right\}\sqrt{PF^2}.
\end{aligned}
$$

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

Theorem 1.1. Let F be an envelop for the class $\mathcal{F}$ such that $PF^2 < \infty$. Then

$$\mathbb{E}\sup_{f\in\mathcal{F}}(n^{1/2}|P_n f - Pf|) \le C\|F\|_{L^2(P)}J(F,\mathcal{F}),$$

where

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

1.1 Application to Boolean function classes with finite VC dimension

Let $\mathcal{F}$ be a Boolean function class with finite VC dimension, and let D denote its VC dimension. Recall that the VC dimension is defined as the maximum cardinality of a set in $\mathcal{X}$ that is shattered by the class $\mathcal{F}$. An important fact about VC dimension is the Sauer-Shelah-Vapnik-Chervonenkis lemma, which states that for every $n \ge 1$ and $x_1, \dots, x_n \in \mathcal{X}$,

$$|\mathcal{F}(x_1,\dots,x_n)| \le \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D}, \tag{8.1}$$

where $\mathcal{F}(x_1,\dots,x_n) = \{(f(x_1),\dots,f(x_n)) : f \in \mathcal{F}\}$. Note that $\binom{n}{k}$ in (8.1) is taken to be 0 if $n < k$. The RHS of (8.1) equals 2^D if $n < D$ and is bounded from above by $(en/D)^D$ if $n \ge D$.

We have seen previously that

$$\mathbb{E}\sup_{f\in\mathcal{F}}(n^{1/2}|P_n f - Pf|) \le C\sqrt{D\log(en/D)} \quad \text{for } n \ge D, \tag{8.2}$$

and this bound was proved by symmetrization and the elementary bound on Rademacher averages. This elementary bound involved the cardinality of $\mathcal{F}(X_1, \dots, X_n)$ which we bound via (8.1).

It turns out that the logarithmic factor is redundant in (8.2), and one actually has the bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} (n^{1/2} |P_n f - Pf|) \le C D^{1/2}. \tag{8.3}$$

This can be deduced as a consequence of Theorem 1.1 as we will demonstrate in this section. Since Theorem 1.1 gives bounds in terms of packing numbers, it becomes necessary to relate the packing numbers of $\mathcal{F}$ to its VC dimension. This is done in the following important result due to Dudley.

Theorem 1.2. Suppose $\mathcal{F}$ is a Boolean function class with VC dimension D. Then

$$\sup_Q M(\epsilon, \mathcal{F}, L^2(Q)) \le \left(\frac{c_1}{\epsilon}\right)^{c_2 D} \quad \text{for all } 0 < \epsilon \le 1. \tag{8.4}$$

Here $c_1, c_2 > 0$ are universal constants, and the supremum is taken over all probability measures Q on $\mathcal{X}$.

Note that Theorem 1.2 gives upper bounds for the ϵ-packing numbers when $\epsilon \le 1$. Since the functions in $\mathcal{F}$ take only values 0 and 1, it is clear that $M(\epsilon, \mathcal{F}, L^2(Q)) = 1$ for all $\epsilon \ge 1$.

Proof. Fix $0 < \epsilon \le 1$ and a probability measure Q on $\mathcal{X}$. Write $N = M(\epsilon, \mathcal{F}, L^2(Q))$ and let $\{f_1, \dots, f_N\}$ be a maximal ϵ-separated subset of $\mathcal{F}$ in the $L^2(Q)$ metric. This means that for every $1 \le i \ne j \le N$,

$$\delta := \epsilon^2 < \int (f_i - f_j)^2 \mathrm{d}Q = \int I(f_i \ne f_j) \mathrm{d}Q = QI(f_i \ne f_j).$$

Let $Z_1, Z_2, \dots$ be iid random variables from Q. Then

$$\mathbb{P}\{f_i(Z_1) = f_j(Z_1)\} = 1 - QI(f_i \ne f_j) < 1 - \delta.$$

By independence, it holds for every $k \ge 1$ that

$$\mathbb{P}\{f_i(Z_1) = f_j(Z_1), f_i(Z_2) = f_j(Z_2), \dots, f_i(Z_k) = f_j(Z_k)\} < (1 - \delta)^k.$$

In other words, the probability that f_i and f_j agree on every $Z_1, \dots, Z_k$ is at most $(1-\delta)^k \le e^{-k\epsilon^2}$. By the union bound,

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

It follows immediately that

$$\mathbb{P}\{|\mathcal{F}(Z_1, \dots, Z_k)| \ge N\} \ge 1 - \frac{N^2}{2} e^{-k\delta}.$$

If we take

$$k = \left\lceil \frac{2 \log N}{\delta} \right\rceil \ge \frac{2 \log N}{\delta}, \tag{8.5}$$

then $\mathbb{P}\{|\mathcal{F}(Z_1,\ldots,Z_k)| \geq N\} \geq 1/2 > 0$. For this particular choice of k, there exists a subset $\{z_1,\ldots,z_k\}$ such that

$$N \leq |\mathcal{F}(z_1,\ldots,z_k)| \leq \binom{k}{0} + \binom{k}{1} + \cdots + \binom{k}{D}, \tag{8.6}$$

where the second inequality is due to the Sauer-Shelah-VC lemma.
Case 1. If $k \leq D$, then (8.6) gives

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

which proves (8.4).
Case 2. Assume $k \geq D$. Together, (8.6) and (8.5) imply

$$N \leq \left(\frac{ek}{D}\right)^D \leq \left(\frac{3e \log N}{\delta D}\right)^D.$$

It follows that

$$N^{1/D} \leq \frac{3e \log N}{\delta D} = \frac{6e}{\delta} \log N^{1/(2D)} \leq \frac{6e}{\delta} N^{1/(2D)},$$

where we used $\log x \leq x$ $(\forall x \geq 1)$ in the last step. Consequently, $N \leq (6e/\delta)^{2D}$ and hence

$$N \leq \left(\frac{6e}{\delta} \log(6e/\delta)\right)^D \leq \left(\frac{6e}{\delta} \frac{6e}{4\delta}\right)^D = \left(\frac{3e}{\delta}\right)^{2D},$$

where the last inequality is based on the bound $\log x \leq x/4$ for $x \geq 9$.

Combining the two cases completes the proof. □

The bound (8.3) immediately follows from Theorems 1.1 and 1.2 as shown below.

Theorem 1.3. Suppose $\mathcal{F}$ is a Boolean class of functions with VC dimension D. Then

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - P f| \leq C \sqrt{\frac{D}{n}}. \tag{8.7}$$

Proof. Since $\mathcal{F}$ is a Boolean class, we can apply Theorem 1.1 with $F(x) \equiv 1$. This gives

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

The packing numbers above can be bounded by Theorem 1.2, implying

$$J(1, \mathcal{F}) \leq \int_0^1 \sqrt{1 + 2D \log \frac{3e}{\epsilon^2}}\, d\epsilon.$$

Given $A \geq e$ and $v > 0$, we wish to bound

$$\int_0^1 \sqrt{1 + v \log(A/\epsilon)}\, d\epsilon \leq A \sqrt{v} \int_A^\infty \frac{\sqrt{1 + \log \epsilon}}{\epsilon^2} d\epsilon.$$

An integration by parts gives

$$\int_A^\infty \frac{\sqrt{1+\log\epsilon}}{\epsilon^2}d\epsilon = -\frac{\sqrt{1+\log\epsilon}}{\epsilon}\Big|_A^\infty + \frac{1}{2}\int_A^\infty \frac{1}{\epsilon^2\sqrt{1+\log\epsilon}}d\epsilon$$
$$\leq \frac{\sqrt{\log(eA)}}{A} + \frac{1}{2}\int_A^\infty \frac{\sqrt{1+\log\epsilon}}{\epsilon^2}d\epsilon \quad (\text{if } A \geq e),$$

from which it follows

$$\int_A^\infty \frac{\sqrt{1+\log\epsilon}}{\epsilon^2}d\epsilon \leq \frac{2\sqrt{\log(eA)}}{A}.$$

Consequently,

$$J(1,\mathcal{F}) \leq CD^{1/2}$$

for some absolute constant $C > 0$. Putting together the pieces completes the proof of Theorem 1.3.

□

The following examples are immediate applications of Theorem 1.3.

Example 1.1. Suppose $X_1, \ldots, X_n$ are iid real-valued random variables having a common CDF F. Let F_n denote the empirical CDF. Then Theorem 1.3 immediately gives

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

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

One can also obtain a high probability upper bound on $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ using the bounded differences inequality that we discussed previously. Combined with above bound, it gives

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

Example 1.2 (Classification with VC classes)**.** Consider the classification problem where we observe iid data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $X_i \in \mathcal{X}$ and $Y_i \in \{0, 1\}$. Let C be class of functions from $\mathcal{X}$ to $\{0, 1\}$ (these are classifiers). For a classifier g, define its test error and training error by

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

respectively. The ERM (empirical risk minimization) classifier is given by

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

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

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

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

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

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

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

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

Using the bounded differences inequality and the bound given by Theorem 1.3, we obtain that for every $\alpha \in (0, 1)$,

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

with probability at least $1 - \alpha$.

It can further be shown that $\mathrm{VC}(\mathcal{F}) \leq \mathrm{VC}(\mathcal{C})$. To see this, it suffices to argue that if $\mathcal{F}$ can shatter $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, then $\mathcal{C}$ can shatter $x_1, \dots, x_n$. For this, let $\eta_1, \dots, \eta_n$ be arbitrary in $\{0, 1\}$. We need to obtain a function $g \in \mathcal{C}$ for which $g(x_i) = \eta_i$. Define $\delta_1, \dots, \delta_n$ by

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

Because $\mathcal{F}$ can shatter $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, there exists a function $f \in \mathcal{F}$ with $f(x_i, y_i) = \delta_i$ for $i = 1, \dots, n$. If $f(x, y) = I\{g(x) \neq y\}$ for some $g \in \mathcal{C}$, then

$$\eta_i I(y_i = 0) + (1 - \eta_i) I(y_i = 1) = I\{g(x_i) \neq y_i\},$$

indicating $g(x_i) = \eta_i$. This proves that $\mathcal{C}$ shatters $x_1, \dots, x_n$, and hence proves the claim.

Finally, we conclude that for every $\alpha \in (0, 1)$,

$$L(\widehat{g}_n) - L(g^*) \leq C \sqrt{\frac{\mathrm{VC}(\mathcal{C})}{n}} + \sqrt{\frac{8 \log(1/\alpha)}{n}}$$

with probability at least $1 - \alpha$. Thus, as long as $\mathrm{VC}(\mathcal{C}) = o(n)$, the test error of $\widehat{g}_n$ relative to the best test error in $\mathcal{C}$ converges to zero as $n \to \infty$.