Statistics 210B Lecture 9 Notes

Daniel Raban

February 15, 2022

1 Bounds on Rademacher Complexity of Function Classes

1.1 Bounding $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|]_{\mathcal{F}}$ in terms of Rademacher complexity

Last time, we were studying empirical processes defined by $X_i \overset{\text{iid}}{\sim} \mathbb{P} \in \mathcal{P}(\mathcal{X})$ and a function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R} : \mathbb{E}[|f(X)|] < \infty\}$. We want to bound the maximum of the empirical process,

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

We introduced the notion of Rademacher complexity for function classes: Given $\mathcal{F}$ and $\{x_i\}_{i \in [n]}$, we let

$$\mathcal{R}(\mathcal{F}(x_{1:n})/n) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$

Then, given $\mathcal{F}$ and $\mathbb{P}$,

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\varepsilon, X} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$

What is the relationship of Rademacher complexity and $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$? Define

$$\|\mathbb{S}_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right|.$$

Here is an upgraded version of what we showed last time.

Proposition 1.1. *For every convex, nondecreasing function $\Phi : \mathbb{R} \to \mathbb{R}$,*

$$\begin{aligned} \mathbb{E}_{X,\varepsilon}[\Phi(\tfrac{1}{2}\|\mathbb{S}_n\|_{\overline{\mathcal{F}}})] &\overset{\text{(a)}}{\leq} \mathbb{E}_X[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})] \\ &\overset{\text{(b)}}{\leq} \mathbb{E}_{X,\varepsilon}[\Phi(2\|\mathbb{S}_n\|_{\mathcal{F}})], \end{aligned}$$

where $\overline{\mathcal{F}} = \{f - \mathbb{E}[f] : f \in \mathcal{F}\}$.

Remark 1.1. Making $\Phi(t) = t$ retrieves the bound on $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ in terms of Rademacher complexity. We can also take the upper bound to also be $\overline{\mathcal{F}}$ because $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] = \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\overline{\mathcal{F}}}$.

Proof. For (b),

$$\mathbb{E}_X[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})] = \mathbb{E}_X\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}[f(Y_i)])\right|\right)\right]$$

Using Jensen's inequality,

$$\leq \mathbb{E}_{X,Y}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f(Y_i))\right|\right)\right]$$

Since $f(X_i) - f(Y_i)$ has a symmetric distribution,

$$\begin{aligned}
&= \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(Y_i))\right|\right)\right] \\
&\leq \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right|\right)\right]
\end{aligned}$$

Using Jensen's inequality again,

$$\begin{aligned}
&\leq \frac{1}{2}\,\mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\right] \\
&\qquad + \frac{1}{2}\,\mathbb{E}_{Y,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right|\right)\right] \\
&= \mathbb{E}_{X,\varepsilon}[\Phi(2\|\mathbb{S}_n\|_{\mathcal{F}})].
\end{aligned}$$

For (a),

$$\mathbb{E}_{X,\varepsilon}[\Phi(\tfrac{1}{2}\|S_n\|_{\mathcal{F}})] = \mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - \mathbb{E}[f(Y_i)])\right|\right)\right]$$

Using Jensen's inequality,

$$\leq \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(Y_i))\right|\right)\right]$$

Since $f(X_i) - f(Y_i)$ has a symmetric distribution,

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

$$\le \mathbb{E}_{X,Y}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}[f(X_i)])\right|\right.\right.$$
$$\left.\left.+\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Y_i)-\mathbb{E}[f(Y_i)])\right|\right)\right]$$

Using Jensen's inequality again,

$$=\frac{1}{2}\,\mathbb{E}_X\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}[f(X_i)])\right|\right)\right]$$
$$+\frac{1}{2}\,\mathbb{E}_Y\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Y_i)-\mathbb{E}[f(Y_i)])\right|\right)\right]$$
$$=\mathbb{E}_X[\Phi(\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}].$$

□

Suppose that for all $f\in\mathcal{F}$, $\|f\|_\infty\le b$. Then $\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}$ is $(2b/n,\dots,2b/n)$-bounded difference. The bounded difference inequality then gives that $\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}$ is $\mathrm{sG}(b/\sqrt{n})$. In other words,

$$|\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}-\mathbb{E}[\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}|\le b\sqrt{\frac{\log(2/\delta)}{n}}\qquad\text{with probability }1-\delta.$$

This upper bound is typically smaller than $\mathcal{F}_n(\mathcal{F})$. This tells us that

$$\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}\begin{cases}\le 2\mathcal{R}_n(\mathcal{F})+b\sqrt{\frac{\log(2/\delta}{n}}\\ \ge \frac{1}{2}\mathcal{R}_n(\overline{\mathcal{F}})-b\sqrt{\frac{\log(2/\delta}{n}}.\end{cases}$$

Note that

$$\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}=\|\mathbb{P}_n-\mathbb{P}\|_{\overline{\mathcal{F}}}\lesssim 2\mathcal{R}_n(\overline{\mathcal{F}}).$$

1.2 Aside: the maximal inequality

How do we upper bound the Rademacher complexity? Let's take a higher level picture and try to bound $\mathbb{E}[\sup_{\theta\in\Theta}X_\theta]$. In many cases, X_θ is sub-Gaussian for each fixed θ.

The simplest case is when Θ is finite. In this case, we have a **maximal inequality**: If for all $\theta\in\Theta$, $X_\theta\in\mathrm{sG}(\sigma)$, then

$$\mathbb{E}\left[\max_{\theta\in\Theta}X_\theta\right]\le\sigma\sqrt{2\log|\Theta|}.$$

However, typically, this set Θ is infinite, so the maximal inequality cannot handle this case.

In the next lecture, we will discuss the metric entropy method, in which we approximate Θ by Θ_ε, where $|\Theta_\varepsilon|<\infty$ and

$$\sup_{\theta\in\Theta_\varepsilon}X_\theta\xrightarrow{\varepsilon\to0}\sup_{\theta\in\Theta}X_\theta.$$

We will make this statement quantitative and precise. We will also introduce a different reduction, based on the concept of VC dimension.

1.3 Bounding Rademacher complexity using the maximal inequality

Use the special structure

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{2}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \\ &= \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{2}\sum_{i=1}^n \varepsilon_i f(X_i)\right| \mid X_{1:n}\right]\right] \\ &= \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\sup_{\nu\in\mathcal{F}(X_{1:n})}\left|\frac{1}{n}\langle\varepsilon,\nu\rangle\right| \mid X_{1:n}\right]\right]\end{aligned}$$

Bound the expectation by the supremum.

$$\leq \sup_{X_{1:n}} \mathbb{E}_\varepsilon\left[\sup_{\nu\in\mathcal{F}(X_{1:n})}\left|\frac{1}{n}\langle\varepsilon,\nu\rangle\right| \mid X_{1:n}\right]$$

If, for example, $\mathcal{F}\subseteq\{f:\mathcal{X}\to\{\pm1\}\}$, then

$$\mathcal{F}(X_{1:n}) = \{(f(X_1),\dots,f(X_n)) : f\in\mathcal{F}\}\subseteq\{\pm1\}^n.$$

Sometimes $|\mathcal{F}|=\infty$, but $|\mathcal{F}(X_{1:n})|<\infty$.

Example 1.1. Suppose $\mathcal{F}=\{\mathbb{1}_{\{X\leq t\}} : t\in\mathbb{R}\}$, so

$$\mathcal{F}(X_{1:n}) = \{(\mathbb{1}_{\{X_1\leq t\}},\mathbb{1}_{\{X_2\leq t\}},\dots,\mathbb{1}_{\{X_n\leq t\}}) : t\in\mathbb{R}\}.$$

Then if $X_1<X_2<\cdots<X_n$,

$$\mathcal{F}(X_{1:n}) = \{(0,0,\dots,0),(1,0,\dots,0),(1,1,0,\dots,0),\dots,(1,1,\dots,1)\},$$

so

$$\sup_{X_{1:n}}|\mathcal{F}(X_{1:n})| = n+1.$$

Let's return to bounding

$$\mathbb{E}_\varepsilon\left[\sup_{\nu\in\mathcal{F}(X_{1:n})}\left|\frac{1}{n}\langle\varepsilon,\nu\rangle\right| \mid X_{1:n}\right].$$

We have that $\frac{1}{n}\langle\varepsilon,\nu\rangle = \frac{1}{n}\sum_{i=1}^n\varepsilon_i\nu_i$ is $\mathrm{sG}(\sigma_n)$, where

$$\sigma_n = \sup_{\nu\in\mathcal{F}(X_{1:n})}\frac{1}{n}\|\nu\|_2 = \sup_{f\in\mathcal{F}}\frac{1}{n}\sqrt{\sum_{i=1}^n f(X_i)^2}.$$

This tells us that the maximum of $|\mathcal{F}(X_{1:n})|$ is the number of mean 0 sG(σ_n) random variables. So the maximum inequality tells us that

$$\begin{aligned}\mathbb{E}_\varepsilon\left[\sup_{\nu\in\mathcal{F}(X_{1:n})}\left|\frac{1}{n}\langle\varepsilon,\nu\rangle\right| \mid X_{1:n}\right] &\leq \sigma_n\sqrt{2\log(2|\mathcal{F}(X_{1:n})|)}\\ &\approx \underbrace{\sup_{f\in\mathcal{F}}\sqrt{\frac{\sum_{i=1}^n f(X_i)^2}{n}}}_{D_{\mathcal{F}}(X_{1:n})}\sqrt{\frac{2\log(2|\mathcal{F}(X_{1:n})|)}{n}}\end{aligned}$$

Example 1.2. Let $\mathcal{F} = \{\mathbb{1}_{\{X\leq t\}} : t \in \mathbb{R}\}$ be the function class in the Glivenko-Cantelli theorem. Then

$$\sup_{X_{1:n}} |\mathcal{F}(X_{1:n})| = n+1,$$

$$\sup_{X_{1:n}} D_{\mathcal{F}}(X_{1:n}) = \sup_{f\in\mathcal{F}}\sqrt{\frac{\sum_{i=1}^n 1^2}{n}} = 1.$$

So we get

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log(2(n+1))}{n}},$$

which bounds

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim 2\sqrt{\frac{2\log(2(n+1))}{n}} + \sqrt{\frac{\log(2/\delta)}{n}} \qquad \text{with probability } 1-\delta.$$

Remark 1.2. The above example gives a proof of the Glivenko-Cantelli theorem.

Remark 1.3. This $\log n$ factor is not sharp. Using other arguments, we will be able to show that the bound is actually of order $\sqrt{1/n}$. The issue here is that the maximal inequality is only sharp when the terms are independent. If X_i are sG(1), then

$$\sup_{i\in[n]} X_i = \begin{cases} O(\sqrt{\log n}) & \text{if the } X_i \text{ are independent}\\ X_1 = O(1) & \text{if } X_1 = X_2 = \cdots = X_n.\end{cases}$$

Look at the bound

$$\Delta = \underbrace{D_{\mathcal{F}}(X_{1:n})}_{\text{typically } O(1)}\underbrace{\sqrt{\frac{2\log(2|\mathcal{F}(X_{1:n})|)}{n}}}_{\text{want to vanish as } n\to\infty}.$$

Let's restricut our attention to $\mathcal{F} \subseteq \{f : \mathcal{X} \to \{\pm 1\}\}$. Here are two frequent behaviors of $|\mathcal{F}(X_{1:n})|$:

(a) If $|\mathcal{F}(X_{1:n})| \lesssim O(n^\nu)$, then $\Delta = O(\sqrt{\frac{\nu \log n}{n}})$. This will go to 0 as $n \to \infty$, so this situation is good.

(b) If $|\mathcal{F}(X_{1:n})| \lesssim O(\nu^n)$, then $\Delta = O(\sqrt{\frac{n \log \nu}{n}}) = O(\sqrt{\log \nu})$. This will not go to 0 as $n \to \infty$, so this situation is not good.

We want to be able to discriminate between these two cases. Since $\mathcal{F}(X_{1:n}) \subseteq \{\pm 1\}^n$, $|\mathcal{F}(X_{1:n})| \leq 2^n$. But when can we give a sharper upper bound?

Definition 1.1. $\mathcal{F}$ has **polynomial discrimination** of order $\nu \geq 1$ if for all n and $X_{1:n}$,

$$|\mathcal{F}(X_{1:n})| \lesssim (n+1)^\nu.$$

Lemma 1.1. *Suppose $\mathcal{F}$ has* PD(ν)*. Then*

$$\mathcal{R}_n(\mathcal{F}) \leq 4 \left(\sup_{X_{1:n}} D_{\mathcal{F}}(X_{1:n}) \right) \sqrt{\frac{\nu \log(n+1)}{n}}.$$

Example 1.3. The function class $\{\mathbb{1}_{\{X \leq t\}} : t \in \mathbb{R}\}$ has PD(1), which implies the Glivenko-Cantelli theorem.

What kind of function classes have polynomial discrimination? Let $\psi : \mathcal{X} \to \mathbb{R}^d$.

Example 1.4. If $\mathcal{F} = \{\langle \psi(x), \theta \rangle + b : \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$, then $|\mathcal{F}(X_{1:n})| = \infty$. So this does not have polynomial discrimination.

Example 1.5. If $\mathcal{F} = \{\mathbb{1}_{\{\langle \psi(x), \theta \rangle \geq b\}} : \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$, then $\mathcal{F}$ has PD$(d+1)$.