Statistics 210B Lecture 13 Notes

Daniel Raban

March 1, 2022

1 Examples of Rademacher Complexity Bounds for Function Classes

1.1 Recap: chaining bounds for Rademacher complexity of function classes

Last time, we were using the metric entropy method to bound the Rademacher complexity of a function class. We considered 4 metrics on $\mathcal{F}$:

$$\|\cdot\|_{L^2(\mathbb{P})}, \qquad \|\cdot\|_{L^\infty}, \qquad \|\cdot\|_{L^2(\mathbb{P}_n)}, \qquad \rho \text{ on parameter space.}$$

Relationships of these metrics gave us relationships between the covering numbers:

$$\begin{aligned} N(\varepsilon; \mathcal{F}, \|\cdot\|_{L^2(\mathbb{P}_n)} &\le \sup_{\mathbb{P}} N(\varepsilon; \mathcal{F}, \|\cdot\|_{L^2(\mathbb{P})}) \\ &\le N(\varepsilon; \mathcal{F}, \|\cdot\|_{L^\infty}) \end{aligned}$$

And if the function class $\mathcal{F}$ is a Lipschitz parametrization,

$$\le N(\varepsilon; T, \rho)$$

If we let $X_f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(x_i)$, then we can show that

$$\mathbb{E}[e^{\lambda(X_f - X_g)}] \le e^{(\lambda^2/2)\|f-g\|_{\mathbb{P}_n}^2} \le e^{(\lambda^2/2)\|f-g\|_\infty^2},$$

which tells us that $\{X_f\}_{f\in\mathcal{F}}$ is a sub-Gaussian process with respect to the $L^2(\mathbb{P}_n)$ or L^∞ metric.

We showed two results:

Proposition 1.1. *Let $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\varepsilon_i, X_i}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]$. Then*

1. $$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{D_{\mathbb{P}}}{\sqrt{n}} \int_0^1 \sup_Q \sqrt{\log N(D_Q u; \mathcal{F}, L^2(Q))}\, du,$$

2.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{D_\infty}{\sqrt{n}} \inf_\varepsilon \varepsilon + \frac{1}{\sqrt{n}} \int_\varepsilon^1 \sup_Q \sqrt{\log N(D_\infty u; \mathcal{F}, L^\infty)}\, du,$$

where $D_{\mathbb{P}} = \sup_{f\in\mathcal{F}} \|f\|_{L^2(\mathbb{P})}$ *and* $D_\infty = \sup_{f\in\mathcal{F}} \|f\|_\infty$.

1.2 Examples of upper bounds for parametric and nonparametric function classes

Here are some examples for upper bounds of Rademacher complexity for function classes.

Example 1.1. Let $\mathcal{F} = \{f_\theta(x) = 1 - e^{-\theta x}, x \in [0,1] : \theta \in [0,1]\}$ be a parametric function class. Then taking the derivative gives us

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \le \sup_{\theta\in[theta_1,\theta_2]} \underbrace{|xe^{-\theta x}}_{\le x} |\theta_1 - \theta_2| \le |\theta_1 - \theta_2|$$

The covering number for the unit interval with $|\cdot|$ is bounded as

$$N(\varepsilon; [0,1], |\cdot|) \le \frac{1}{2\varepsilon} + 1,$$

so we get a covering number bound for the parametric function class

$$N(\varepsilon; \mathcal{F}, L^\infty) \le N(\varepsilon; [0,1], |\cdot|) \le \frac{1}{2\varepsilon} + 1.$$

Using the chaining bound with $D_\infty = \sup_{f\in\mathcal{F}} \|f\|_\infty = \sup_{f\in\mathcal{F}} \sup_{x\in[0,1]} |1 - e^{-\theta x}| \le 1$,

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &\le \frac{D_\infty}{\sqrt{n}} \int_0^1 \sqrt{\log N(uD_\infty; \mathcal{F}, L^\infty)}\, du \qquad = \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log N(u; \mathcal{F}, L^\infty)}\, du \\
&= \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log(\tfrac{1}{2u} + 1)}\, du \\
&\lesssim \frac{C}{\sqrt{n}}.
\end{aligned}$$

Example 1.2 (Lipschitz parameterization)**.** Consider a function class $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathbb{R} : \theta \in B_2^d(1)$ with $\|f_0(x)\|_\infty = c_0 = 0$. If $|f_{\theta_1}(x) - f_{\theta_2}(x)| \le L\|\theta_1 - \theta_2\|_2$, then we can use the bound

$$\log N(\varepsilon; \mathcal{F}, L^\infty) \le \log N(\varepsilon; B_2^d(1), \|\cdot\|_2 \lesssim d \log \lesssim d \log\left(\frac{1}{\varepsilon} + 1\right)$$

to get

$$\mathcal{R}_n(\mathcal{F}) \lesssim L\frac{D_\infty}{\sqrt{n}} \int_0^1 \sqrt{\log N(\varepsilon; \mathcal{F}, L^\infty)}\, d\varepsilon,$$

where $D_\infty = \sup_\theta \|f_\theta\|_\infty \le c_0 + L = L$

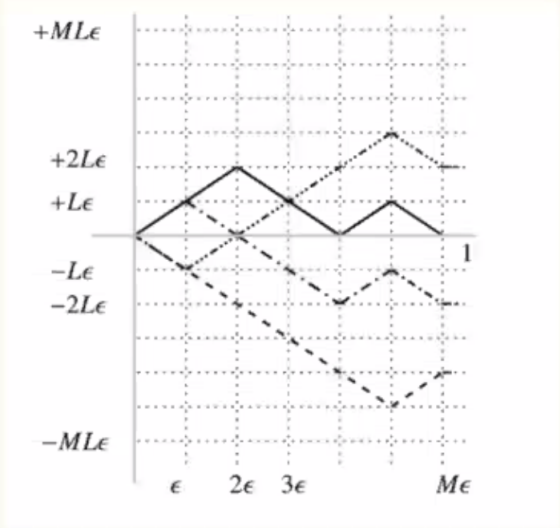
$$\begin{aligned} &\lesssim \frac{L}{\sqrt{n}} \int_0^1 \frac{1}{\sqrt{n}} \int_0^1 \sqrt{d \log(1/n)}\, du \\ &\lesssim L\sqrt{\frac{d}{n}}. \end{aligned}$$

If we have a nonparametric function class, it may have infinite Rademacher complexity. So in general, we will want some sort of smoothness condition to make the complexity finite.

Example 1.3 (Nonparametric class with smoothness/convexity)**.** Consider the non-parametric function class $\mathcal{F}_L = \{g : [0,1] \to \mathbb{R} \mid g(0) = 0,\ g \text{ is } L\text{-Lip}\}$. Then

$$\log N(\varepsilon; \mathcal{F}_L, L^\infty) \asymp \frac{L}{\varepsilon},$$

which we can see from the following figure in Wainwright's book that shows how to bound the packing number:



In particular, if $f \neq g$, then

$$M(L\varepsilon; \mathcal{F}, L^\infty) \ge 2^{1/\varepsilon}.$$

Taking log and rescaling ε, we get

$$\log M(L\varepsilon; \mathcal{F}, L^\infty) \ge 2^{1/\varepsilon} \ge \frac{2L}{\varepsilon} \log 2.$$

On the other hand, we can get an upper bound by seeing that these functions cover the function class.

Here, we have $\|\mathcal{F}\|_\infty = \sup_{f\in\mathcal{F}} |f| = L$, so the one-step discretization bound gives

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}_L) &\lesssim \inf_\varepsilon \varepsilon + \frac{1}{\sqrt{n}}\sqrt{\log N(\varepsilon;\mathcal{F},\|\cdot\|_\infty)} \\ &= \inf_\varepsilon \varepsilon + \frac{1}{\sqrt{n\varepsilon}} \\ &\asymp \frac{1}{n^{1/3}}\end{aligned}$$

The chaining bound gives

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}_L) &\lesssim \inf_\varepsilon \varepsilon + \frac{1}{\sqrt{n}}\int_\varepsilon^1 \sqrt{\frac{1}{u}}\,du \\ &\asymp \frac{1}{\sqrt{n}}.\end{aligned}$$

So in this case, the one-step discretization bound gives a sharper bound than the chaining method.

Example 1.4 (Nonparametric class, general d)**.** Consider a nonparametric function class with general d:

$$\mathcal{F}_L^d = \{g : [0,1]^d \to \mathbb{R} : g(0) = 0, g \text{ is } L\text{-Lip in } \|\cdot\|_\infty\}.$$

We can show that

$$\log N(\varepsilon;\mathcal{F}_L^d.L^\infty) \asymp \left(\frac{L}{\varepsilon}\right)^d.$$

The calculation of the resulting bounds on the Rademacher complexity is left for homework.

1.3 Boolean function classes

Consider a Boolean function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to \{0,1\}\}$, VC theory tells us that $\mathcal{F}$ has PD(ν), where $\nu = \mathrm{VC}(\mathcal{F})$. Usng the maximal inequality, we have the bound

$$\mathcal{R}_n(\mathcal{F}) \lesssim \sqrt{\frac{\nu\log(n+1)}{n}}.$$

We have mentioned that the log factor in this bound makes the bound not tight.

Proposition 1.2. *For a boolean function class with $\nu = \mathrm{VC}(\mathcal{F})$,*

$$\sup_{\mathbb{P}} \log(N(\varepsilon;\mathcal{F},L^2(\mathbb{P}))) \lesssim \nu\log\left(\frac{e}{\varepsilon}\right)$$

for $\varepsilon < 1$.

For a sharp but difficult proof of this bound, see theorem 2.6.4 from [Van der Vaart and Wellner, 1996]. A weaker but easier version of this bound can be found in the notes [Sen, Theorem 7.9].

If we use the chaining argument, we get the bound

$$\mathcal{F}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\nu \log(e/\varepsilon)}\, d\varepsilon \propto \sqrt{\frac{\nu}{n}}.$$

Example 1.5. Specialize to the function class $\mathcal{F} = \{\mathbb{1}_{x \le t} : t \in \mathbb{R}$, which we first examined when looking at empirical processes. This has VC-dimension 1, so

$$\mathcal{R}_n(\mathcal{F}) \lesssim \sqrt{\frac{1}{n}}.$$

This tells us that

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \ge \frac{c}{\sqrt{n}} + \frac{\varepsilon}{\sqrt{n}}\right) \le 2e^{-\varepsilon^2/2}.$$

Remark 1.1. This is not the tightest version of this bound. The tighest bound, given by Dvoretzky, Kiefer, Wolfowitz, and Massart, is

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \ge \frac{\varepsilon}{\sqrt{n}}\right) \le 2e^{-\varepsilon^2/2}.$$

1.4 Contraction inequalities

Consider d functions $\phi_j : \mathbb{R} \to \mathbb{R}$ which are L-Lipschitz with $\phi_j(0) = 0$. We can think off $\phi_j(\theta)$ as a loss function $L(y; \theta)$.

Proposition 1.3 (Talagrand-Ledoux concentration)**.** *Let $T \subseteq \mathbb{R}^d$, and let $\{\phi_j\}$ be centered Lipschitz. Then*

$$\mathbb{E}\left[\sup_{\theta \in T} \sum_{j=1}^d \varepsilon_j \phi_j(\theta_j)\right] \le L\, \mathbb{E}\left[\sup_{\theta \in T} \sum_{j=1}^d \varepsilon_j \theta_j\right],$$

$$\mathbb{E}\left[\sup_{\theta \in T} \left|\sum_{j=1}^d \varepsilon_j \phi_j(\theta_j)\right|\right] \le 2L\, \mathbb{E}\left[\sup_{\theta \in T} \left|\sum_{j=1}^d \varepsilon_j \theta_j\right|\right].$$

The interpretation is that the right hand side is $\mathcal{R}(T)$. The left hand side is $\mathcal{R}(\phi(T))$. This says that if we apply a contraction map to a space, the Rachemacher complexity will not increase.

The textbook has a proof for when ε_i are iid Gaussian random variables. This is given by the Gaussian comparison inequality.

Example 1.6. Let $Z_i = (X_i, Y_i) \overset{\text{iid}}{\sim} \mathbb{P} \in \mathcal{P}(B_2(M) \times \{\pm 1\})$ for $i \in [n]$. For logistic regression, we want a logistic loss function:

$$M_\theta(Z) := \log(1 + \exp(-y\theta^\top x)).$$

Taking the expectation gives

$$M(\theta) = \mathbb{E}_Z[m_\theta(Z)].$$

We also let $\Theta = B_2(r)$. Compare the empirical and population risk:

$$\begin{aligned} E &:= \mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n m_\theta(Z_i) - M(\theta)\right|\right] \\ &\le 2\,\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i m_\theta(Z_i)\right|\right] \end{aligned}$$

We are looking at the function class $\mathcal{F} = \{m_\theta(z_i) : \theta \in \Theta\}$. If we wanto replace $m_\theta(z_i)$ by $\theta^\top x_i$, then we can use the contraction inequality. This is because $\log(1 + e^x)$ is 1-Lipschitz (by $\frac{d}{dx}\log(1 + e^x) = \frac{e^x}{1+e^x} \le 1$. So we can write $\phi_i(\widetilde{\theta}_i) = \log(1 + \exp(-y_i\widetilde{\theta}_i)) - \log 2$. This depends on Y_i, and $\widetilde{\theta}_i = \theta^\top X_i$ depends on X_i, to use the contraction inequality, we first condition on Y and X:

$$\begin{aligned} &= 2\,\mathbb{E}_{Y,X}\left[\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i m_\theta(Z_i)\right| \mid Y, X\right]\right] \\ &= 2\,\mathbb{E}_{Y,X}\left[\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\phi(\widetilde{\theta}_i) + \log 2)\right| \mid Y, X\right]\right] \end{aligned}$$

First, use the triangle inequality to get rid of the $\log 2$:

$$\le 2\,\mathbb{E}_{Y,X}\left[\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\phi(\widetilde{\theta}_i) + \log 2)\right| \mid Y, X\right]\right] + (\cdots)$$

Now apply the contraction inequality with $\widetilde{\Theta} = \{(\langle\theta, x_i\rangle, \ldots, \langle\theta, x_n\rangle) : \theta \in \Theta\} \subseteq \mathbb{R}^n$.

$$\begin{aligned} &\le 4\,\mathbb{E}_{Y,X}\left[\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\widetilde{\theta}_i\right| \mid Y, X\right]\right] + (\cdots) \\ &= 4\,\mathbb{E}_{Y,X}\left[\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle X_i, \theta\rangle\right| \mid Y, X\right]\right] + (\cdots) \\ &= 4\,\mathbb{E}_{\varepsilon,X}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle X_i, \theta\rangle\right|\right] + (\cdots) \\ &= 4\,\mathbb{E}_{\varepsilon,X}\left[\sup_{\theta\in\Theta}\left|\left\langle\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i, \theta\right\rangle\right|\right] + (\cdots) \end{aligned}$$

$$
\begin{aligned}
&= 4r\,\mathbb{E}_{\varepsilon,X}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\right\|_2\right] + (\cdots) \\
&\le 4r\,\mathbb{E}_{\varepsilon,X}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\right\|_2^2\right]^{1/2} + (\cdots) \\
&= 4r\left(\frac{\mathbb{E}[\|X\|_2^2]}{n}\right)^{1/2} + (\cdots) \\
&\le 4\frac{rM}{\sqrt{n}} + (\cdots).
\end{aligned}
$$

1.5 Further topics: Orlicz processes and bracketing numbers

There is a generalization of sub-Gaussian using the Orlicz norm.

Definition 1.1. Let $\psi_q(t) := \exp(t^q) - 1$ for $q \in [1,2]$. The **q-Orlicz norm** is

$$\|X\|_{\psi_q} := \inf\{\lambda > 0 : \mathbb{E}[\psi_q(|X|/\lambda)] \le 1\}.$$

We can prove concentration inequalities, the maximal inequality, the one-step discretization bound, and the chaining bounding in terms of Orlicz norms.

In empirical process theory, there is another notion of covering called the bracketing number. This is discussed in the notes by Sen and in Chapter 2 of Van der Waart and Wellner.

Definition 1.2. Given two functions $\ell(\cdot)$ and $u(\cdot)$, the **bracket**

$$[L,u] = \{f \in \mathcal{F} : \ell(x) \le f(x) \le u(x)\ \forall x \in \mathcal{X}\}.$$

An **ε-bracket** is a bracket $[L,u]$ with $\|\ell - u\| \le \varepsilon$.

Definition 1.3. The **bracketing number** $N_{[]}(\varepsilon; \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$, i.e.

$$N_{[]}(\varepsilon; \mathcal{F}, \|\cdot\|) = \min\{N : \{[\ell_i, u_i]_{i\in[N]} \text{ covers } \mathcal{F} \text{ and } \|\ell_i - u_i\| \le \varepsilon\}.$$

Proposition 1.4. *Let* $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\varepsilon_i, X_i}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]$. *Then*

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{D_{\mathbb{P}}}{\sqrt{n}}\int_0^1 \sqrt{\log N_{[]}(D_{\mathbb{P}}u; \mathcal{F}, L^2(\mathbb{P}))}.$$

Notice that here, unlike the our bound in terms of covering numbers, does not require us to take the sup over distributions Q. Regardless, usually, if you can prove a bound using the bracketing number, you can prove it using the covering number.