Statistics 210B Lecture 12 Notes

Daniel Raban

February 24, 2022

1 The Metric Entropy Method for Function Spaces

1.1 Recap: controlling complexity via chaining

Last time, we were discussing the metric entropy method for obtaining bounds on empirical processes. We have a metric space (T, ρ), and we want to control

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \qquad \text{or} \qquad \mathbb{E}\left[\sup_{\theta \in T} |X_\theta|\right],$$

where X_θ is usually mean 0 and sub-Gaussian. We introduced the metric entropy is $\log N(\varepsilon; T, \rho)$, where $N(\varepsilon; T, \rho) = \inf\{N : |T_\varepsilon| = N, T_\varepsilon \text{ is an } \varepsilon\text{-cover}\}$ is the ε-covering number.

We had the one step discretization bound given by the maximal inequality

$$\mathbb{E}\left[\sup_{\theta \in T} |X_\theta|\right] \lesssim \inf_{\varepsilon} \sigma \sqrt{\log(N(\varepsilon; T, \rho)} + \mathbb{E}\left[\sup_{\rho(\theta, \widetilde{\theta}) \leq \varepsilon} |X_\theta - X_{\widetilde{\theta}}|\right]$$

We introduced the condition of a process to be sG(ρ):

$$\mathbb{E}[e^{\lambda(X_\theta - X_{\widetilde{\theta}})}] \leq \exp\left(\frac{\lambda^2}{2}\rho(\theta, \widetilde{\theta})^2 \sigma^2\right).$$

This condition allowed us to use the chaining bound

$$\mathbb{E}\left[\sup_{\theta \in T} |X_\theta|\right] \lesssim \inf_{\varepsilon} \sigma \int_{\varepsilon}^{D} \sqrt{\log N(u; T, \rho)}\, du + \mathbb{E}\left[\sup_{\rho(\theta,, \widetilde{\theta} \leq \varepsilon} |X_\theta - X_{\widetilde{\theta}}|\right].$$

Last time, we discussed examples where $T \subseteq \mathbb{R}^d$. We let $X_\theta = \langle \varepsilon, \theta \rangle$ or $X_\theta = \langle W, \theta \rangle$ to get bounds on the Rademacher/Gaussian complexity of Euclidean sets. Today, we will discuss examples where $T = \mathcal{F} \subseteq L^p$ for $1 \leq p \leq \infty$ is a function space. If we let

$$X_\theta = \frac{1}{n}\sum_{i=1}^{n} \varepsilon f(Z_i) \qquad \text{or }, X_\theta = \frac{1}{n}\sum_{i=1}^{n}(f(Z_i) - \mathbb{E}[f(Z_i)]),$$

then this gives us information about the Rademacher/Gaussian complexity of function spaces.

1.2 One step discretization and chaining bounds for Rademacher complexity of function classes

Recall that if $|mcF \subseteq L^1(\mathbb{P})$ and $\varepsilon_i \overset{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})$, then we defined the Rademacher complexity of function class as

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

where we can think of this as the expectation of the empirical Rademacher complexity,

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

where

$$\mathcal{F}(x_{1:n}) = (f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n.$$

Recall that VC theory tells us that when the value of f is binary, $\mathcal{F}(x_{1:n})$ is a finite set. Then we can use the maximal inequality.

This lecture, we will control this using the metric entropy method. Rewrite

$$\mathcal{R}(\mathcal{F}(x_{1:n})/n) = \frac{1}{\sqrt{n}}\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}|X_f|\right],$$

where

$$X_f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(x_i).$$

Hoeffding's inequality tells us that $X_f \sim \text{sG}(\sqrt{\frac{1}{n}\sum_{i=1}^n f(x_i)^2})$.

To apply Dudley's entropy intergral bound on $\mathbb{E}[\sup_{\theta\in T}|X_\theta|]$, we need

1. A metric ρ on $\mathcal{F}$,
2. X_f to be a sub-Gaussian process with respect to ρ,
3. An upper bound for $N(u; \mathcal{F}, \rho)$,
4. (Optional) An upper bound for the discretization error.

1.3 Useful metrics on $\mathcal{F} \subseteq L^1(\mathbb{P})$

Here are four useful metrics

(a) $L^2(\mathbb{P})$ metric:

$$\|f - g\|_{L^2(\mathbb{P})}^2 = \int_{\mathcal{X}} (f(x) - g(x))^2 \, d\mathbb{P}(x).$$

(b) L^∞ metric: If $\operatorname{supp} \mathbb{P} = \mathcal{X}$, then

$$\|f - g\|_{L^\infty} = \sup_{x \in \mathcal{X}} |f(x) - g(x)|.$$

(c) $L^2(\mathbb{P}_n)$ metric (given $x_{1:n}$):

$$\|f - g\|_{L^2(\mathbb{P}_n)}^2 = \int (f(x) - g(x))^2 \, d\mathbb{P}_n(x) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2.$$

We can make this a random metric by using $X_{1:n}$.

This is equivalent to $\|\cdot\|_2$ on $\mathcal{F}(x_{1:n})/\sqrt{n} \subseteq \mathbb{R}^n$. Recall that

$$\mathcal{F}(x_{1:n}/\sqrt{n}) = \{\tfrac{1}{\sqrt{n}}(f(x_1), \dots, f(x_n)) \in \mathbb{R}^n : f \in \mathcal{F}\}.$$

Then if $f(x_{1:n})/\sqrt{n}, g(x_{1:n})/\sqrt{n} \in \mathcal{F}(x_{1:n})/\sqrt{n}$,

$$\|f(x_{1:n})/\sqrt{n} - g(x_{1:n})/\sqrt{n}\|_2^2 = \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2.$$

(d) Parametric metric: If $\mathcal{F} = \{f_\theta : \theta \in T \subseteq \mathbb{R}^d\}$, a metric ρ on T induces a metric ρ on $\mathcal{F}$ by

$$\rho(f_\theta, f_{\widetilde{\theta}}) := \rho(\theta, \widetilde{\theta}).$$

Here are the relationships between these metrics:

- For any measure $\mathbb{P}$, $\|f - g\|_{\mathbb{P}} \le \|f - g\|_\infty$. In particular, this says that $\|f - g\|_{\mathbb{P}_n} \le \|f - g\|_\infty$ for all $x_{1:n}$.
- When $\mathcal{F} = f_\theta : \theta \in T \subseteq \mathbb{R}^d\}$, suppose that $|f_{\theta_1} - f_{\theta_2}(x)| \le \Gamma(x)\rho(\theta_1, \theta_2)$. Then

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^2(\mathbb{P})} \le \|\Gamma\|_{L^2(\mathbb{P})} \rho(\theta_1, \theta_2),$$

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^\infty} \le \|\Gamma\|_{L^\infty} \rho(\theta_1, \theta_2).$$

Example 1.1. Let $\mathcal{F} = \{f_\theta(x) = 1 - e^{-\theta x}, x \in [0,1] : \theta \in [0,1]\}$. Then, using Taylor expansion and the intermediate value theorem,

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| = \left| xe^{-\xi x}|\theta_1 - \theta_2| \right| \leq |x| \cdot |\theta_1 - \theta_2|.$$

This tells us that

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^2(\mathbb{P})} \leq \|x\|_{L^2(\mathbb{P})}|\theta_1 - \theta_2|.$$
$$\|f_{\theta_1} - f_{\theta_2}\|_{L^\infty} \leq |\theta_1 - \theta_2|.$$

When x is not restricted to a bounded domain, we will not get a bound for the L^∞ norm

We care about inequalities between metrics because they introduce inequalities between covering numbers.

Lemma 1.1. *If ρ_1, ρ_2 are two metrics on T and $\rho_1(\theta_1, \theta_2) \leq \rho_2(\theta_1, \theta_2)$ for all $\theta_1, \theta_2 \in T$, then*

$$N(\varepsilon; T, \rho_1) \leq N(\varepsilon; T, \rho_2).$$

As a consequence,

$$N(\varepsilon; \mathcal{F}, L^2(\mathbb{P}_n)) \leq N(\varepsilon; \mathcal{F}, L^\infty), \qquad N(\varepsilon; \mathcal{F}, L^2(\mathbb{P})) \leq N(\varepsilon; \mathcal{F}, L^\infty).$$

If $|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \Gamma(x)\rho(\theta_1, \theta_2)$, then

$$N(\varepsilon; \mathcal{F}, L^\infty) \leq N(\varepsilon; T, \|\Gamma\|_\infty \rho), \qquad N(\varepsilon; \mathcal{F}, L^\infty) \leq N(\varepsilon; T, \|\Gamma\|_{L^2(\mathbb{P})}\rho).$$

Note that we can express this rescaling either in the metric or as a scaling factor in front of ε.

1.4 The uniform entropy bound for empirical processes

In what metrics might $X_f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i)$ be a sub-Gaussian process?

$$\begin{aligned}
\mathbb{E}[e^{\lambda(X_f - X_g)} \mid X_{1:n}] &= \mathbb{E}[e^{(\lambda/\sqrt{n})|X_{1:n} \sum_{i=1}^n \varepsilon_i (f(X_i) - g(X_i))} \mid X_{1:n}] \\
&= \prod_{i=1}^n \mathbb{E}[e^{(\lambda/\sqrt{n})|X_{1:n} \varepsilon_i (f(X_i) - g(X_i))} \mid X_i] \\
&\leq \prod_{i=1}^n e^{(\lambda^2/n)(f(X_i) - g(X_i)^2/2} \\
&= e^{(\lambda^2/2)\frac{1}{n}\sum_{i=1}^n (f(X_i) - g(X_i)^2/2}
\end{aligned}$$

Since $\frac{1}{n}\sum_{i=1}^n (f(X_i) - g(X_i)^2/2 - \|f - g\|_{\mathbb{P}_n} \leq \|f - g\|_\infty$,

$$\leq e^{(\lambda^2/2)\|f-g\|_\infty}.$$

This tells us that $(X_f)_{f\in\mathcal{F}}$ is a sub-Gaussian process with respect to the metric $\|\cdot\|_{L^2(\mathbb{P}_n)}$. The inequalities between metrics tell us that this is also then sub-Gaussian with respect to $\|\cdot\|_{L^\infty}$.

Now, if $D = \sup_{f,g\in\mathcal{F}} \|f-g\|_{L^2(\mathbb{P}_n)} =: \|\mathcal{F}\|_{\mathbb{P}_n}$ is the diameter,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |X_f|\right] \leq \int_0^D \sqrt{\log N(u;\mathcal{F}, L^2(\mathbb{P}_n)}\, du.$$

Then the empirical Rademacher complexity is bounded above by

$$\mathcal{R}(\mathcal{F}(X_{1:n})/n) = \frac{1}{\sqrt{n}}\,\mathbb{E}\left[\sup_{f\in\mathcal{F}} |X_f|\right]$$

Using the change of variables $u = \|\mathcal{F}\|_{\mathbb{P}_n}\widetilde{u}$,

$$\begin{aligned}
&\lesssim \frac{1}{\sqrt{n}} \int_0^{\|\mathcal{F}\|_{\mathbb{P}_n}} \sqrt{\log N(\|\mathcal{F}\|_{\mathbb{P}_n}\widetilde{u};\mathcal{F}, L^2(\mathbb{P}_n))}\, d\|\mathcal{F}\|_{\mathbb{P}_n}\widetilde{u} \\
&= \frac{\|\mathcal{F}\|_{\mathbb{P}_n}}{\sqrt{n}} \int_0^1 \sqrt{\log N(\|\mathcal{F}\|_{\mathbb{P}_n}u;\mathcal{F}, L^2(\mathbb{P}_n))}\, du \\
&\leq \frac{\|\mathcal{F}\|_{\mathbb{P}_n}}{\sqrt{n}} \int_0^1 \sup_Q \sqrt{\log N(\|\mathcal{F}\|_Q u;\mathcal{F}, L^2(Q))}\, du.
\end{aligned}$$

When we take the expectation of the empirical Rademacher complexity and use Cauchy-Schwarz, we get

$$\mathbb{E}[\mathcal{F}(\mathcal{F}(X_{1:n})/\sqrt{n})] \leq \frac{\|\mathcal{F}\|_{\mathbb{P}}}{\sqrt{n}} \int_0^1 \sup_X \sqrt{\log N(\|\mathcal{F}\|_Q u;\mathcal{F}, L^2(Q))}\, du.$$

We can summarize this in the following proposition:

Proposition 1.1 (Uniform entropy bound)**.**

$$\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \lesssim \mathcal{R}_n(\mathcal{F}) \lesssim \frac{\|\mathcal{F}\|_{\mathcal{P}}}{\sqrt{n}} \int_0^1 \sup_Q \sqrt{\log N(\|\mathcal{F}\|_Q u;\mathcal{F}, L^2(Q))}\, du.$$

This is not in Wainwright's textbook, but you can find it as Theorem 4.7 in *A Gentle Introduction to Empirical Process Theory and Applications* by Bodhisattva Sen.

1.5 Examples of bounding Rademacher complexity for different covering numbers

Example 1.2. Suppose we have $\log N(u) \asymp d\log(1+1/u)$. Then

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{1}{\sqrt{n}} \int_0^1 \sqrt{d\log(1+1/u)}\, du \lesssim \sqrt{\frac{d}{n}}.$$

Example 1.3. If $\log N(u) \asymp 1/u$, then

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\frac{1}{u}}\, du \lesssim \frac{1}{\sqrt{n}}.$$

Example 1.4. If $\log N(u) \asymp \frac{1}{u^d}$, where $d \geq 2$, then

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\frac{1}{u^d}}\, du = \infty.$$

However, we can get a better bound in the last example by using the following proposition.

Proposition 1.2.

$$\sup_{\mathbb{P}} \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim \mathcal{R}_n(\mathcal{F}) \lesssim \|\mathcal{F}\|_\infty \inf_{\varepsilon} \varepsilon + \frac{1}{\sqrt{n}} \int_\varepsilon^1 \sqrt{\log N(\|\mathcal{F}\|_\infty u; \mathcal{F}, L^\infty)}\, du.$$

How can we upper bound $\mathbb{E}_{\varepsilon_i}[\sup_{\|f-g\|_{\mathbb{P}_n} \leq \varepsilon} |\sum_{i=1}^n \varepsilon_i(f(X_i) - g(X_i))|]$? We know that we can bound

$$\mathbb{E}_{\varepsilon_i}\left[\sup_{\|f-g\|_{L^\infty} \leq \varepsilon} \sum_{i=1}^n \varepsilon_i(f(X_i) - g(X_i))|\right] \leq \sqrt{n}\varepsilon.$$

If we use this bound, then when $\log N(u) \lesssim \frac{1}{u^d}$ with $d \geq 2$, we get

$$\mathcal{R}_n(\mathcal{F}) \lesssim \inf_{\varepsilon} \varepsilon + \frac{1}{\sqrt{n}} \int_\varepsilon^1 \frac{1}{u^{d/2}}\, du.$$