Statistics 210B Lecture 11 Notes

Daniel Raban

February 22, 2022

1 Volume Bounds for Metric Entropy and the Chaining Method

1.1 Recap: one-step discretization bound

Last time, we began 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.

Here is the one-step discretization bound that the maximal inequality gives us:

Lemma 1.1. *If $X_\theta \sim \mathrm{sG}(\sigma)$ for all $\theta \in T$, then*

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta \in T} |X_\theta|\right] &\lesssim \inf_{\varepsilon} \inf_{\varepsilon\text{-cover } T_\varepsilon} \mathbb{E}\left[\sup_{\theta \in T_\varepsilon} |X_\theta|\right] + \mathbb{E}\left[\sup_{\rho(\theta, \widetilde{\theta}) \le \varepsilon} |X_\theta - X_{\widetilde{\theta}}|\right] \\ &\lesssim \inf_{\varepsilon} \sigma \sqrt{\log(N(\varepsilon; T, \rho)} + \mathbb{E}\left[\sup_{\rho(\theta, \widetilde{\theta}) \le \varepsilon} |X_\theta - X_{\widetilde{\theta}}|\right]\end{aligned}$$

Today, we will mostly discuss the case where $T \subseteq \mathbb{R}^d$ is Euclidean space, and X_θ is some canonical random variable, such as $X_\theta = \langle \varepsilon, \theta \rangle$ or $X_\theta = \langle W, \theta \rangle$, which give the Radamacher and Gaussian complexities. We will give a volume-based method for bounding the covering number, give some examples, and then introduce the chaining method, which will give us a sharper bound.

In the next few lecture, we will extend this discussion to $T = \mathcal{F} \subseteq L^p(\mathbb{P})$ for $1 \le p \le \infty$, with $X_\theta = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Z_i)$ or $X_\theta = \frac{1}{n} \sum_{i=1}^n (f(Z_i) - \mathbb{E}[f(Z_i)]$. We will also relate this to and extend our VC theory.

1.2 Volume bounds for metric entropy

We want to understand the ε-covering number for $T \subseteq \mathbb{R}^d$. The intuition is that

$$\log N(\varepsilon; T, \rho) \asymp \log \frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B_\rho(\varepsilon))},$$

so we can understand the covering number by understanding the volume. Here, we use the notation

$$B_\rho(\theta, \varepsilon) := \{\widetilde{\theta} \in \mathbb{R}^d : \rho(\theta, \widetilde{\theta}) \le \varepsilon\}, \qquad B_\rho(\varepsilon) := B_\rho(0, \varepsilon),$$

$$\mathrm{Vol}(T) = \int \mathbb{1}_{\{x \in T\}} \, dx,$$

where dx is Lebesgue measure.

Last time, we introduced the notion of the ε-packing number

$$M(\varepsilon; T, \rho) = \sup\{M : |\widetilde{T}_\varepsilon| = M, M \text{ is an } \varepsilon\text{-packing of } T\}.$$

This was related to the covering number by the following lemma.

Lemma 1.2. *For all $\varepsilon > 0$, we have*

$$M(2\varepsilon; T, \rho) \le N(\varepsilon; T, \rho) \le M(\varepsilon; T, \rho).$$

Lemma 1.3.

$$\frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B_\rho(\varepsilon))} \le N(\varepsilon; T, \rho) \le M(\varepsilon; T, \rho) \le \frac{\mathrm{Vol}(T + B_\rho(\varepsilon/2))}{\mathrm{Vol}(B_\rho(\varepsilon/2))},$$

where $T + B_\rho(\varepsilon/2) = \{a + b : a \in T, b \in B_\rho(\varepsilon/2)\}$.

Proof. For the first inequality, let T_ε be an ε-covering, so $T \subseteq \bigcup_{\theta \in T_\varepsilon} B_\rho(\theta, \varepsilon)$. This tells us that

$$\begin{aligned}
\mathrm{Vol}(T) &\le \mathrm{Vol}\left(\bigcup_{\theta \in T_\varepsilon} B_\rho(\theta, \varepsilon)\right) \\
&\le \sum_{\theta \in T_\varepsilon} \mathrm{Vol}(B_\rho(\theta, \varepsilon)) \\
&\le |T_\varepsilon| \, \mathrm{Vol}(B_\rho(\varepsilon)).
\end{aligned}$$

For the second inequality, let $\widetilde{T}_\varepsilon$ be a ε-packing, so the union of all the balls in the packing is contained in the set augmented by $\varepsilon/2$. That is, $\bigcup_{\theta \in \widetilde{T}_\varepsilon} B(\theta, \varepsilon/2) \subseteq T + B_\rho(\varepsilon/2)$. This tells us that

$$\mathrm{Vol}(T + B_\rho(\varepsilon/2)) \ge \mathrm{Vol}\left(\bigcup_{\theta \in \widetilde{T}_\varepsilon} B(\theta, \varepsilon/2)\right)$$

$$= |\widetilde{T}_\varepsilon| \operatorname{Vol}(B_\rho(\varepsilon)).$$

Now take the sup over all packings. □

Example 1.1. Let $\rho = \|\cdot\|_p$ and $T = B_p(1) = \{x \in \mathbb{R}^d : \|x\|_p \le 1\}$. Then

$$N(\varepsilon; T, \rho) \le \frac{\operatorname{Vol}(T + B(\varepsilon/2))}{\operatorname{Vol}(B(\varepsilon/2))} = \frac{\operatorname{Vol}(B_p(1 + \varepsilon/2))}{\operatorname{Vol}(B_p(\varepsilon/2))}$$

Note that $\operatorname{Vol}(B_\rho(r)) = c_{d,p} r^d$ for some constant $c_{d,p}$. We do not need to know the value of $c_{d,p}$ because we are looking at ratios of volumes. This gives

$$N(\varepsilon; T, \rho) \le \frac{(1 + \varepsilon/2)^d}{(\varepsilon/2)^d} = \left(\frac{2}{\varepsilon} + 1\right)^d.$$

We also get the lower bound

$$N(\varepsilon; T, \rho) \ge \frac{\operatorname{Vol}(B_p(1))}{\operatorname{Vol}(B_p(\varepsilon))} = \frac{1^d}{\varepsilon^d} = \left(\frac{1}{\varepsilon}\right)^d.$$

So we get bounds on the metric entropy

$$d \log\left(\frac{1}{\varepsilon}\right) \le \log N(\varepsilon; T, \rho) \le d \log\left(\frac{2}{\varepsilon} + 1\right).$$

These bounds are of the same order. Note that the bounds do not depend on p because we are looking at the p-ball in the p-norm.

Example 1.2. Consider $W_i \overset{\text{iid}}{\sim} N(0, 1)$, so $\langle W, \theta\rangle \sim \text{sG}(\|\theta\|_2)$. Then we know that

$$\mathcal{G}(B_2(1) = \mathbb{E}\left[\sup_{\theta \in B_2(1)} \langle W, \theta\rangle\right] = \mathbb{E}[\|W\|_2\| \simeq \sqrt{d}.$$

Here is another way to get this computation:

$$\begin{aligned}
\mathcal{G}(B_2(1)) &\le C\left[\sup_{\theta \in B_2(1)} \underbrace{\|\theta\|_2}_{=1} \underbrace{\sqrt{\log N(\varepsilon; B_2(1), \|\cdot\|_2)}}_{\le \sqrt{d \log(1 + 2/\varepsilon)}} + \mathbb{E}_W\left[\sup_{\|\theta - \theta'\|_2 \le \varepsilon} |W_\theta - W_{\theta'}|\right]\right] \\
&\le C\left[\sqrt{d \log(1 + 2/\varepsilon)} + \mathbb{E}_W\left[\sup_{\|\theta - \widetilde{\theta}\|_2 \le \varepsilon} \langle W, \theta - \theta'\rangle\right]\right] \\
&= C\left[\sqrt{d \log(1 + 2/\varepsilon)} + \mathbb{E}_W\left[\sup_{\|r\|_2 \le \varepsilon} \langle W, r\rangle\right]\right]
\end{aligned}$$

$$= C \left[\sqrt{d \log(1 + 2/\varepsilon)} + \varepsilon \underbrace{\mathbb{E}_W \left[\sup_{\|\widetilde{r}\|_2 \le 1} \langle W, \widetilde{r} \rangle \right]}_{\mathcal{G}(B_2(1))} \right].$$

This tells us that

$$\mathcal{G}(B_2(1)) \le C\sqrt{d \log(1 + 2/\varepsilon)} + C\varepsilon \mathcal{G}(B_2(1)).$$

If we take $\varepsilon \le \frac{1}{2C}$, then we get

$$\mathcal{G}(B_2(1)) \le 2C\sqrt{d \log(1 + 4C)} \asymp \sqrt{d},$$

which is the same order as before.

1.3 The chaining method

We have been using the bound

$$\mathbb{E}\left[\sup_{\theta \in T} |X_\theta|\right] \lesssim \inf_{\varepsilon} \underbrace{\inf_{\varepsilon\text{-cover } T_\varepsilon} \mathbb{E}\left[\sup_{\theta \in T_\varepsilon} |X_\theta|\right]}_{\text{bdd by covering number}} + \underbrace{\mathbb{E}\left[\sup_{\rho(\theta,\widetilde{\theta}) \le \varepsilon} |X_\theta - X_{\widetilde{\theta}}|\right]}_{\text{how to give tight control?}}$$

Controlling the right term can require ad-hoc arguments. The chaining method gives a way to bound this effectively.

Definition 1.1. $\{X_\theta\}_{\theta \in T}$ is a **sub-Gaussian process** with respect to ρ on T if

$$\mathbb{E}[e^{\lambda(X_\theta - X_{\theta'})}] \le e^{\lambda^2 \rho(\theta,\theta')^2/2},$$

or, equivalently, $X_\theta - X_{\theta'}$ is $\mathrm{sG}(\rho(\theta, \theta'))$.

Example 1.3. Let $T \subseteq \mathbb{R}^d$ with $\rho = \|\cdot\|_2$. Look at $X_\theta = \langle W, \theta \rangle$, where $W \sim N(0, I_d)$. To bound, the Gaussian complexity, we want to bound $\mathbb{E}[\sup_{\theta \in T} X_\theta]$. Then $X_\theta - X'_\theta = \langle W, \theta - \theta' \rangle \sim N(0, \|\theta - \theta'\|_2^2) \sim \mathrm{sG}(\|\theta - \theta'\|_2)$.

Proposition 1.1. *Let $\{X_\theta, \theta \in T\}$ be a mean 0 sub-Gaussian process with metric ρ. Then if $D = \sup_{\theta, \theta \in T}$,*

$$\mathbb{E}\left[\sup_{\theta, \widetilde{\theta}} (X_\theta - X_{\widetilde{\theta}})\right] \le \inf_{\varepsilon \le D} 2\left[\sup_{\rho(r,r') \le \varepsilon} (X_r - X_{r'})\right] + 32 \underbrace{\int_\varepsilon^D \sqrt{\log N(u; T, \rho)}\, du}_{=: J(\varepsilon; D; T, \rho)}.$$

Here, $J(\varepsilon; D; T, \rho)$ is known as **Dudley's entropy integral**.

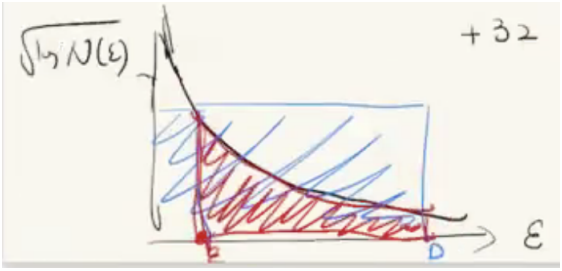
Remark 1.1. This gives an upper bound for $\mathbb{E}[\sup_{\theta\in T} X_\theta]$ because by the 0 mean condition and Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right] &= \mathbb{E}\left[\sup_{\theta,\theta'\in T} (X_\theta - \mathbb{E}_{\theta'}[X_{\theta'}])\right] \\ &\le \mathbb{E}\left[\sup_{\theta,\widetilde{\theta}} (X_\theta - X_{\widetilde{\theta}})\right].\end{aligned}$$

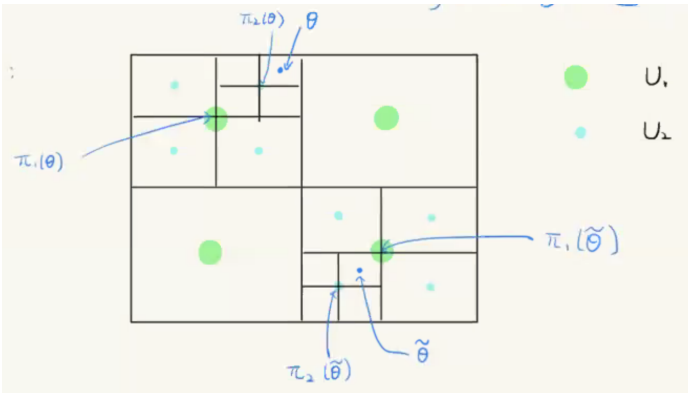
Remark 1.2. Compare this to the bound

$$\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}}(X_\theta - X_{\widetilde{\theta}})\right] \le \inf_{\varepsilon\le D} 2\left[\sup_{\rho(r,r')\le\varepsilon}(X_r - X_{r'})\right] + 32D\sqrt{\log N(\varepsilon;T,\rho)}.$$

The integration gives a better bound because $\sqrt{\log N(\varepsilon)}$ is decreasing in ε.



Proof. Take a sequence of ε-coverings corresponding to $\varepsilon_m = D/2^m$ for $m = 0, 1, 2, 3, \ldots, L$. Let U_m be the minimal ε_m-covering of T, so $|U_m| \le N(\varepsilon_m; T_\rho)$. Then define the projection operation $\pi_m(\theta) = \arg\min_{\beta\in U_m} \rho(\theta,\beta)$.



This allows us to bound

$$|X_\theta - X_{\widetilde{\theta}}| \le |X_\theta - X_{\pi_2(\theta)}| + |X_{\pi_2(\theta)} - X_{\pi_1(\theta)}| + |X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}|$$

$$+ |X_{\pi_1(\widetilde{\theta})} - X_{\pi_2(\widetilde{\theta})}| + |X_{\pi_2(\widetilde{\theta})} - X_{\widetilde{\theta}}|.$$

Then we can take the expectation of $\sup_{\theta,\widetilde{\theta}}$ on both sides. What is the purpose of having all these interpolation points? The first and the last terms have infinitely many choices, so these are the discretization terms, while the middle terms have only finitely many choices, so we can apply the maximal inequality.

$$\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in T} |X_\theta - X_{\widetilde{\theta}}|\right] \le \mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in T} |X_{\pi_1(\theta)} - X_{\pi_1(\widetilde{\theta})}|\right] + 2\,\mathbb{E}\left[\sup_{\theta\in T} |X_{\pi_2(\theta)} - X_{\pi_1(\theta)}|\right]$$
$$+ \cdots + 2\,\mathbb{E}\left[\sup_{\theta\in T} |X_{\pi_L(\theta)} - X_{\pi_{L-1}(\theta)}|\right] + 2\,\mathbb{E}\left[\sup_{\theta\in T} |X_\theta X_{\pi_L(\theta)}|\right]$$

These terms on the right correspond to $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{L-1}, \varepsilon_*$, respectively. This process will define a Riemann sum. For the remaining details, see the textbook. □

Example 1.4. We want to bound the Gaussian complexity $\mathcal{G}(B_2(1)) = \mathbb{E}[\sup_{\theta\in B_2(1)}\langle W, \theta\rangle]$ using chaining. We get the bound

$$\begin{aligned}
\mathcal{G}(B_2(1)) &\le C\int_0^2 \sqrt{\underbrace{\log N(u; B_2(1), \|\cdot\|_2)}_{\le d\log(2/u+1)}}\,du \\
&\le C\int_0^2 \sqrt{d\log(2/u+1)}\,du \\
&= C\sqrt{d}\underbrace{\int_0^2 \sqrt{\log(2/u+1)}\,du}_{C'} \\
&\asymp \sqrt{d}.
\end{aligned}$$