CUHK STAT6050: Statistical Learning Theory Spring 2022

Lecture 5: Rademacher complexity II

Examples, covering number, and entropy bounds

Lecturer: Ben Dai

"*There is Nothing More Practical Than A Good Theory.*" — Kurt Lewin

1 Introduction

According to the Bousquet bound of Talagrand's inequality, it suffices to bound the Rademacher complexity of an empirical process. Let's recall the definition.

To bound the concentration for a general empirical process on i.i.d. samples $(\mathbf{Z}_i)_{i=1,\cdots,n}$ indexed by $h \in \mathscr{H}$:

$$\big\|\mathbb{P}_n - \mathbb{P}\big\|_{\mathscr{H}} = \sup_{h\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^n \Big(h(\mathbf{Z}_i) - \mathbb{E}h(\mathbf{Z}_i)\Big),$$

we consider its corresponding Rademacher process and Rademacher complexity:

$$\mathbf{Rad}_n(h) = \frac{1}{n}\sum_{i=1}^n \rho_i h(\mathbf{Z}_i), \quad h \in \mathscr{H}, \qquad \mathbb{E}\big\|\mathbf{Rad}_n(h)\big\|_{\mathscr{H}} = \mathbb{E}\sup_{h\in\mathscr{H}}\big|\mathbf{Rad}_n(h)\big|.$$

For example, suppose $\mathscr{H}$ is a finite class of functions, we can compute the Rademacher complexity.

Lemma 1.1 (Massart finite lemma)**.** *Suppose $\mathscr{H}$ is a finite class of functions uniformly bounded by U, then*

$$\mathbb{E}\big\|\mathbf{Rad}_n(h)\big\|_{\mathscr{H}} \leq U\sqrt{\frac{2\log\big(|\mathscr{H}|\big)}{n}},$$

where $|\mathscr{H}|$ is the total number of functions in $\mathscr{H}$.

In more general cases, we will try to bound Rademacher complexity of uncountable classes.

Recall the Remark 3.1 in Lecture 4, the Rademacher complexity is a criterion to measure the complexity of a functional space. Yet, directly compute the Rademacher complexity for a general class is not easy, and we tend to bound the Rademacher complexity by two steps. **Step 1:** we introduce **covering numbers** to qualify the complexity of the functional space; the reason is that **covering numbers** is usually more easy to understand and compute; **Step 2:** we introduce some entropy bounds to bridge the **covering numbers** and Rademacher complexity.

2 Covering numbers

To measure the complexity of the function class, we introduce covering numbers and packing numbers.

Definition 2.1 (Covering numbers)**.** Given a function class $\mathscr{H}$ with a pesudo metric μ, and $\varepsilon > 0$, $\mathscr{C} \subseteq \mathscr{H}$ is an ε-*cover* of $(\mathscr{H}, \mu)$, if for any $h \in \mathscr{H}$, there exists $g \in \mathscr{C}$ such that $\mu(h,g) \leq \varepsilon$. Moreover, the *covering numbers* of $(\mathscr{H}, \mu)$ are defined as:

$$N(\mathscr{H}, \mu, \varepsilon) = \inf\big\{|\mathscr{C}| : \mathscr{C} \text{ is an } \varepsilon\textit{-cover}\big\}.$$

Definition 2.2 (Packing numbers)**.** Given a function class $\mathscr{H}$ with a pesudo metric μ, and $\varepsilon > 0$, $\mathscr{P} \subseteq \mathscr{H}$ is an ε-*packing* of $(\mathscr{H}, \mu)$, if for any $g, g' \in \mathscr{P}$, such that $\mu(g,g') > \varepsilon$. Moreover, the *packing numbers* of $(\mathscr{H}, \mu)$ are defined as:

$$P(\mathscr{H}, \mu, \varepsilon) = \sup\big\{|\mathscr{P}| : \mathscr{P} \text{ is an } \varepsilon\textit{-packing}\big\}.$$

Note that covering numbers are the minimal number of balls of radius ε needed to cover $\mathscr{H}$, and the packing numbers are the maximal number of balls of radius ε packed inside $\mathscr{H}$.

Lemma 2.3 (Covering-packing duality)**.** *Given a function class $\mathscr{H}$ with a pesudo metric μ, and $\varepsilon > 0$*

$$N(\mathscr{H}, \mu, \varepsilon) \leq P(\mathscr{H}, \mu, \varepsilon) \leq N(\mathscr{H}, \mu, \varepsilon/2).$$

In practice, the pesudo metric $\mu(h,h')$ is often replaced by a norm $\|h - h'\|$. On this ground, $N(\mathscr{H}, \|\cdot\|, \varepsilon)$ denotes the covering numbers on a normed space $(\mathscr{H}, \|\cdot\|)$.

Lemma 2.4. *Given a function class $\mathscr{H}$ with pesudo metrics μ and μ', such that*

$$\mu(h,h') \leq c\mu'(h,h'), \quad \textit{for any } h, h' \in \mathscr{H}.$$

Then

$$N(\mathscr{H}, \mu, \varepsilon) \leq N(\mathscr{H}, \mu', \varepsilon/c).$$

Based on the definition of a norm, we have the following properties of covering numbers.

Lemma 2.5. *Given a normed space $(\mathscr{H}, \|\cdot\|)$, for any $h_0 \in \mathscr{H}$ and $c > 0$, then*

$$N(c\mathscr{H} + h_0, \mu, \varepsilon) = N(c\mathscr{H}, \mu, \varepsilon) = N(\mathscr{H}, \mu, \varepsilon/c).$$

One typical example is finite dimensional parametric space.

Lemma 2.6 (Euclidean balls)**.** *Consider $\mathscr{H} = \mathbb{R}^d$ with a norm $\|\cdot\|$, denote $\mathscr{B}$ as a unit Euclidean ball in d dimension, then for $\varepsilon \leq 1$,*

$$(\frac{1}{\varepsilon})^d \leq N(\mathscr{B}, \|\cdot\|, \varepsilon) \leq P(\mathscr{B}, \|\cdot\|, \varepsilon) \leq \big(\frac{3}{\varepsilon}\big)^d.$$

Lemma 2.7 (Lipschtiz parametrization)**.** *Consider the following function class parametrized by $\boldsymbol{\theta} \in \Theta$:*

$$\mathscr{H} := \left\{ h_{\boldsymbol{\theta}}(\cdot) : \ \boldsymbol{\theta} \in \Theta \right\}.$$

Denote $\|\cdot\|_\Theta$ is the norm for $\boldsymbol{\theta} \in \Theta$, and $\|\cdot\|_{\mathscr{H}}$ is the norm for $h \in \mathscr{H}$, if

$$\left\| h_{\boldsymbol{\theta}} - h_{\boldsymbol{\theta}'} \right\|_{\mathscr{H}} \leq c \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\Theta.$$

Then,

$$N(\mathscr{H}, \|\cdot\|_{\mathscr{H}}, \varepsilon) \leq N(\Theta, \|\cdot\|_\Theta, \varepsilon/c).$$

3 Entropy bounds

Now, we give the connection between covering number and Rademacher complexity.

Theorem 3.1 (Pollard's bounds)**.**

$$\mathbb{E}_\rho \|\mathbf{Rad}_n(h)\|_{\mathscr{H}} \leq \inf_{\varepsilon > 0} \left(U \sqrt{\frac{2 \log \left(N(\mathscr{H}, L_2(\mathbb{P}_n), \varepsilon) \right)}{n}} + \varepsilon \right),$$

where $U = \sup_{h \in \mathscr{H}} \|h\|_{L^2}$ and $\|h\|_{L_2(\mathbb{P}_n)} = \left(n^{-1} \sum_{i=1}^n h^2(\mathbf{Z}_i) \right)^{1/2}$.

Theorem 3.2 (Dudley's Theorem)**.** *Let $\sigma_n^2 = \sup_{h \in \mathscr{H}} \|h\|^2_{L_2(\mathbb{P}_n)}$, then*

$$\mathbb{E}_\rho \|\mathbf{Rad}_n(h)\|_{\mathscr{H}} \leq 12 \int_0^{\sigma_n} \sqrt{\frac{\log N(\mathscr{H}, L_2(\mathbb{P}_n), \varepsilon)}{n}} d\varepsilon$$

3.1 Some examples for uniform bounded classes

In this section, we give some examples of covering number and Rademacher average on some uniform bounded classes, and the difference between Pollard's and Dudley's bounds as well.

Note that

$$\|h - h'\|_{L_2(\mathbb{P}_n)} \leq \sup_{\mathbf{x} \in \mathscr{X}} |h(\mathbf{x}) - h'(\mathbf{x})| =: \|h - h'\|_{L_\infty}.$$

Then, according to Lemma 2.4, we have

$$N(\mathscr{H}, L_2(\mathbb{P}_n), \varepsilon) \leq N(\mathscr{H}, L_\infty, \varepsilon).$$

Lemma 3.3 (VC-type classes)**.** *Suppose $\mathscr{H}$ is uniformly bounded by U, and*

$$N(\mathscr{H}, L_2(\mathbb{P}_n), \varepsilon) \leq \left(\frac{cU}{\varepsilon} \right)^d, \quad \varepsilon > 0.$$

Then,

$$\mathbb{E} \|\mathbf{Rad}_n(h)\|_{\mathscr{H}} \leq c \sqrt{\frac{d}{n}}.$$

Example 3.4 (linear function class)**.** *Let* $\mathscr{X} = [-1,1]^d$ *and* $\mathscr{H} = \{h(\mathbf{x}) = \boldsymbol{\theta}^\intercal \mathbf{x} : \|\boldsymbol{\theta}\|_1 \leq U\}$*. Then,*

$$\mathbb{E}\|\mathbf{Rad}_n(h)\|_{\mathscr{H}} \leq c\sqrt{\frac{d}{n}}.$$

Example 3.5 (Sparse function class)**.** *Let* $\mathscr{X} = [-1,1]^d$ *and* $\mathscr{H} = \{h(\mathbf{x}) = \boldsymbol{\theta}^\intercal \mathbf{x} : \|\boldsymbol{\theta}\| \leq U, \|\boldsymbol{\theta}\|_0 \leq K\}$*. Then,*

$$N(\mathscr{H}, L_2(\mathbb{P}_n), \varepsilon) \leq \binom{d}{K}\left(\frac{cU}{\varepsilon}\right)^K \leq \left(\frac{ed}{K}\right)^K\left(\frac{cU}{\varepsilon}\right)^K,$$

and

$$\mathbb{E}\|\mathbf{Rad}_n(h)\|_{\mathscr{H}} \leq c\sqrt{\frac{K\log(d)}{n}}.$$

Example 3.6 (Lipschitz functions)**.** *Suppose* $\mathscr{H}$ *is L-Lipschitz function class from* $\mathscr{X} = [0,1]^d$ *to* $[0,1]$*, then*

$$N(\mathscr{H}, L_2(\mathbb{P}_n), \varepsilon) \leq c(1/\varepsilon)3^{L/\varepsilon},$$

and

$$\mathbb{E}\|\mathbf{Rad}_n(h)\|_{\mathscr{H}} \leq c(L/n)^{1/2}.$$

Example 3.7 (Non-decreasing function class)**.** *Suppose* $\mathscr{H}$ *is a non-decreasing function class from* $\mathbb{R}$ *to* $[0,1]$*. Then,*

$$N(\mathscr{H}, L_2(\mathbb{P}_n), \varepsilon) \leq n^{1/\varepsilon},$$

and

$$\mathbb{E}\|\mathbf{Rad}_n(h)\|_{\mathscr{H}} \leq c\left(\frac{\log(n)}{n}\right)^{1/2}.$$

Please check [Bartlett and Mendelson, 2002] for more examples, including decision trees, neural networks, and kernel methods.

References

[Bartlett and Mendelson, 2002] Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.