CUHK STAT6050: Statistical Learning Theory Spring 2026

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 of a general empirical process on i.i.d. samples $(\mathbf{Z}_i)_{i=1,\dots,n}$ indexed by $h \in \mathscr{H}$:

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

we consider its corresponding Rademacher process and Rademacher complexity:

$$\mathbf{Rad}_n(h) = \frac{1}{n}\sum_{i=1}^n \rho_i h(\mathbf{Z}_i), \quad h \in \mathscr{H}, \qquad \mathbb{E}\big\|\mathbf{Rad}_n(h)\big\|_{\mathscr{H}} = \mathbb{E}\sup_{h\in\mathscr{H}}\big|\mathbf{Rad}_n(h)\big|. \tag{2}$$

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*

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

where $|\mathscr{H}|$ is the cardinality of $\mathscr{H}$.

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

Recall Remark 3.1 in Lecture 4, the Rademacher complexity is a criterion to measure the complexity of a function space. Yet, directly computing the Rademacher complexity for a general class is not easy, and we tend to bound it in two steps. **Step 1:** we introduce **covering numbers** to quantify the complexity of the function space; the reason is that **covering numbers** are usually easier 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 pseudo 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 number* of $(\mathscr{H}, \mu)$ is defined as:

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

Definition 2.2 (Packing numbers)**.** Given a function class $\mathscr{H}$ with a pseudo 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 number* of $(\mathscr{H}, \mu)$ is defined as:

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

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

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

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

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 a finite dimensional parameter 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 $\varepsilon \leq 1$,*

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

Lemma 2.7 (Lipschitz 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 $\|\cdot\|_\Theta$ *as the norm for* $\boldsymbol{\theta} \in \Theta$*, and* $\|\cdot\|_{\mathscr{H}}$ *as 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,

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

This result is useful for the function class with Lipschitz parametrization, where the Lipschitz constant is c*.*

A Sub-gaussian random variables

Definition A.1 (Sub-gaussian random variable)**.** A random variable Y is said to be sub-gaussian with parameters (μ, σ^2), denoted $Y \in \mathrm{SG}_\mu(\sigma^2)$, if its moment generating function satisfies for all $t \in \mathbb{R}$:

$$\mathbb{E}[\exp(t(Y-\mu))] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right).$$

When $\mu = 0$, we simply denote $Y \in \mathrm{SG}(\sigma^2)$.

Lemma A.2. *The following random variables are sub-gaussian:*

- *Gaussian random variables with mean 0 and variance* σ^2 *are in* $SG(\sigma^2)$
- *Rademacher random variables (taking values* ± 1 *with probability 1/2) are in* $SG(1)$

Lemma A.3. *Suppose* $Y_j \in SG(\sigma_j^2)$ *for* $j = 1, \ldots, n \geq 2$ *are independent random variables, then we have the following properties of sub-gaussian random variables:*

- $\sum_{j=1}^n Y_j \in SG(\sum_{j=1}^n \sigma_j^2)$.
- $\mathbb{E} \max_{1 \leq j \leq n} |Y_j| \leq 2 \max_{1 \leq j \leq n} \sigma_j \sqrt{1 + \log(2n)}/3$.