CUHK STAT6050: Statistical Learning Theory Spring 2022

Lecture 4: Rademacher complexity I

Symmetrization, Bousquet bound, and excess risk bounds

Lecturer: Ben Dai

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

1 Introduction

Recall the pre-mentioned aims:

- **A1.** The asymptotics of

$$A_1 = \mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \Big(l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - \mathbb{E} l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) \Big).$$

- **A2′.** Find a tight upper bound of

$$A_2 = \mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \Big(l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - \mathbb{E} l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) \Big)^2.$$

The solution to bound those two empirical processes is to introduce **Rademacher complexity** to measure the complexity of the functional space $\mathscr{F}$. The definition of Rademacher complexity is inspired by the one of the most important properties of the empirical processes, that is, **symmetrization**.

2 Symmetrization

We illustrate with the empirical process in **A1**. Define random variables $\tilde{\mathscr{D}}_n = (\tilde{\mathbf{X}}_i, \tilde{\mathbf{Y}}_i)_{i=1,\cdots,n}$ as the independent copy of $\mathscr{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1,\cdots,n}$, that is, $(\tilde{\mathbf{X}}_i, \tilde{\mathbf{Y}}_i) \stackrel{d}{=} (\mathbf{X}_i, \mathbf{Y}_i)$ and samples in $\{\mathscr{D}_n, \tilde{\mathscr{D}}_n\}$

are all independent.

$$
\begin{aligned}
A_1 &= \mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \Big(l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - \mathbb{E} l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) \Big) = \mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \Big(l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - \tilde{\mathbb{E}} l\big(\tilde{\mathbf{Y}}, f(\tilde{\mathbf{X}})\big) \Big) \\
&= \mathbb{E} \sup_{f \in \mathscr{F}} \Big(\frac{1}{n} \sum_{i=1}^n l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - \frac{1}{n} \sum_{i=1}^n \tilde{\mathbb{E}} l\big(\tilde{\mathbf{Y}}_i, f(\tilde{\mathbf{X}}_i)\big) \Big) \\
&= \mathbb{E} \sup_{f \in \mathscr{F}} \tilde{\mathbb{E}} \frac{1}{n} \sum_{i=1}^n \Big(l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - l\big(\tilde{\mathbf{Y}}_i, f(\tilde{\mathbf{X}}_i)\big) \Big) \le \mathbb{E}\tilde{\mathbb{E}} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \Big(l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - l\big(\tilde{\mathbf{Y}}_i, f(\tilde{\mathbf{X}}_i)\big) \Big) \qquad (1) \\
&= \mathbb{E}\tilde{\mathbb{E}} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \rho_i \Big(l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - l\big(\tilde{\mathbf{Y}}_i, f(\tilde{\mathbf{X}}_i)\big) \Big) \le 2\mathbb{E} \sup_{f \in \mathscr{F}} \Big| \frac{1}{n} \sum_{i=1}^n \rho_i l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) \Big|, \qquad (2)
\end{aligned}
$$

where $(\rho_i)_{i=1,\cdots,n}$ are i.i.d. Rademacher random variables independent with $\mathscr{D}_n$ and $\tilde{\mathscr{D}}_n$, with ρ_i taking the values $+1$ and -1 with probability $1/2$ each. The last equality follows from the fact that $(\tilde{\mathbf{X}}_i, \tilde{\mathbf{Y}}_i)$ is the independent copy of $(\mathbf{X}_i, \mathbf{Y}_i)$, thus the joint distribution of $(\mathscr{D}_n, \tilde{\mathscr{D}}_n)$ does not change by switching $(\tilde{\mathbf{X}}_i, \tilde{\mathbf{Y}}_i)$ and $(\mathbf{X}_i, \mathbf{Y}_i)$. Therefore, the equality holds for arbitrary choice of $\rho_i = +1$ or $\rho_i = -1$.

(1) and (2) are so-called *symmetrization inequalities*, and (2) indicates that the empirical risk excess process is upper bounded by the Rademacher process. Next, we summarize all the results for a **general empirical process**.

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

$$
\frac{1}{n} \sum_{i=1}^n \Big(h(\mathbf{Z}_i) - \mathbb{E} h(\mathbf{Z}_i) \Big), \quad h \in \mathscr{H}.
$$

Its corresponding Rademacher process is defined as:

$$
\mathbf{Rad}_n(h) = \frac{1}{n} \sum_{i=1}^n \rho_i h(\mathbf{Z}_i), \quad h \in \mathscr{H}.
$$

Theorem 2.1 (Symmetrization Inequalities)**.** *For any functional space h:*

$$
\frac{1}{2} \mathbb{E} \sup_{h \in \mathscr{H}} |\mathbf{Rad}_n(\tilde{h})| \le \mathbb{E} \sup_{h \in \mathscr{H}} \Big| \frac{1}{n} \sum_{i=1}^n \Big(h(\mathbf{Z}_i) - \mathbb{E} h(\mathbf{Z}_i) \Big) \Big| \le 2\mathbb{E} \sup_{h \in \mathscr{H}} \Big| \mathbf{Rad}_n(h) \Big|, \qquad (3)
$$

where $\tilde{h}(\mathbf{Z}) = h(\mathbf{Z}) - \mathbb{E} h(\mathbf{Z})$*,* $(\tilde{\mathbf{Z}}_1, \cdots, \tilde{\mathbf{Z}}_n)$ *is independent copy of* $(\mathbf{Z}_1, \cdots, \mathbf{Z}_n)$*,* $\mathbb{E} \sup_{h \in \mathscr{H}} \mathbf{Rad}_n(h)$ *is the Rademacher complexity of the function class h, the expectation* $\mathbb{E}$ *is taken with respect to all randomness.*

3 Rademacher complexity

Remark 3.1. Recall the definition of Rademacher process $\mathbf{Rad}_n(h)$, it can be considered as empirical *correlation* between ρ and $h(\mathbf{Z})$. Suppose h restrains only one constant, say $h(\mathbf{z}) = 1$,

$$\mathbf{Rad}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\rho_i = O_P(\frac{1}{\sqrt{n}});$$

if h is diverse enough, such that $h(\mathbf{z}_i) = \rho_i$:

$$\mathbf{Rad}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\rho_i^2 = O_P(1).$$

Therefore, the order of $\mathbb{E}\sup_{h\in\mathscr{H}}\mathbf{Rad}_n(h)$ is between $O(n^{-1/2})$ and $O(1)$, measuring the complexity of the function class $\mathscr{H}$.

In practice, we may want to bound the Rademacher complexity on $\varphi\circ f$. For example, in our case, we tend to investigate the complexity of $l(\mathbf{Y}_i, f(\mathbf{X}_i))$; $f\in\mathscr{F}$. Talagrand's contraction Lemma is proposed to address this target.

Lemma 3.2 (Talagrand's contraction Lemma [Ledoux and Talagrand, 1991]). *Let $\varphi:\mathbb{R}\to\mathbb{R}$ be a L-Lipschitz function, then*

$$\mathbb{E}\sup_{h\in\mathscr{H}}\left|\mathbf{Rad}_n(\varphi\circ h)\right| \le L\mathbb{E}\left|\sup_{h\in\mathscr{H}}\mathbf{Rad}_n(h)\right|. \tag{4}$$

Remark 3.3. Note that φ is a Lipschitz function, it is sensible to believe that the complexity of $\varphi\circ\mathscr{H}$ can be controlled by the complexity of $\mathscr{H}$.

One important application of Talagrand's contraction Lemma is to upper bound the "second moment" of empirical process (A_2 in our case).

Corollary 3.4. *Suppose that functions in $\mathscr{H}$ are uniformly bounded by a constant U, then*

$$\mathbb{E}\sup_{h\in\mathscr{H}}\left|\mathbf{Rad}_n(h^2)\right| \le 2U\mathbb{E}\sup_{h\in\mathscr{H}}\left|\mathbf{Rad}_n(h)\right|.$$

Now, we apply the results to A_1 and A_2. Denote $h(\mathbf{Z}_i) = l(\mathbf{Y}_i, f(\mathbf{X}_i))$, and suppose the loss function l is uniformly bounded by U, then

$$A_1 = \mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\big(h(\mathbf{Z}_i) - \mathbb{E}(h(\mathbf{Z}_i))\big) \le 2\mathbb{E}\sup_{f\in\mathscr{F}}\left|\mathbf{Rad}_n(h)\right|.$$

Denote $\tilde{h}(\mathbf{Z}_i) = l(\mathbf{Y}_i, f(\mathbf{X}_i)) - \mathbb{E}l(\mathbf{Y}_i, f(\mathbf{X}_i))$, then

$$
\begin{aligned}
A_2 &= \mathbb{E}\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{h}^2(\mathbf{Z}_i) - \mathbb{E}(\tilde{h}^2(\mathbf{Z}_i)) + \mathbb{E}(\tilde{h}^2(\mathbf{Z}_i))\right) \\
&\leq \mathbb{E}\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{h}^2(\mathbf{Z}_i) - \mathbb{E}(\tilde{h}^2(\mathbf{Z}_i))\right) + \sup_{f\in\mathscr{F}} \mathbb{E}\tilde{h}^2(\mathbf{Z}) \\
&\leq 2\mathbb{E}\sup_{f\in\mathscr{F}} \left|\mathbf{Rad}_n(\tilde{h}^2)\right| + \sup_{f\in\mathscr{F}} \mathbb{E}\tilde{h}^2(\mathbf{Z}) \leq 4U\mathbb{E}\sup_{f\in\mathscr{F}} \left|\mathbf{Rad}_n(\tilde{h})\right| + \sup_{f\in\mathscr{F}} \mathbf{Var}\big(h(\mathbf{Z})\big) \\
&\leq 8UA_1 + \sup_{f\in\mathscr{F}} \mathbf{Var}\big(h(\mathbf{Z})\big) \leq 16U\mathbb{E}\sup_{f\in\mathscr{F}} \left|\mathbf{Rad}_n(h)\right| + \sup_{f\in\mathscr{F}} \mathbf{Var}\big(h(\mathbf{Z})\big). \qquad (5)
\end{aligned}
$$

4 Bousquet bound

Now, we combine all results to have a new updated form of Talagrand's inequality, namely Bousquet bound. For simplicity, we denote:

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

Theorem 4.1 (Bousquet bound of Talagrand's inequality [Bousquet, 2002])**.** *Suppose $h(\mathbf{Z})$ is uniformly bounded by a constant U almost surely, then for $t > 0$, with probability at least $1-\delta$*

$$
\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{H}} \leq \mathbb{E}\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{H}} + \sqrt{\frac{2\log(1/\delta)}{n}\left(\sigma^2_{\mathscr{H}} + \mathbb{E}\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{H}}\right)} + \frac{U\log(1/\delta)}{3n},
$$

where $\sigma^2_{\mathscr{H}}$ is defined as

$$
\sigma^2_{\mathscr{H}} = \sup_{h\in\mathscr{H}} \mathbf{Var}\big(h(\mathbf{Z})\big).
$$

Theorem 4.1 implies the following corollary.

Corollary 4.2. *Suppose $h(\mathbf{Z})$ is uniformly bounded by a constant U almost surely, then for any $\varepsilon_n > 0$,*

$$
\mathbb{P}\Big(\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{H}} - \mathbb{E}\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{H}} \geq \varepsilon_n\Big) \leq \exp\left(-\frac{n\varepsilon_n^2}{2\big(\sigma^2_{\mathscr{F}} + \mathbb{E}\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{H}} + U\varepsilon_n/3\big)}\right). \qquad (6)
$$

Furthermore, if

$$
\varepsilon_n \geq 4\mathbb{E}\sup_{h\in\mathscr{H}} |\mathbf{Rad}_n(h)|, \quad (\textit{thus } \varepsilon_n \geq 2\mathbb{E}\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{H}})
$$

we have,

$$
\mathbb{P}\Big(\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{H}} \geq \varepsilon_n\Big) \leq \exp\left(-\frac{n\varepsilon_n^2}{8\big(\sigma^2_{\mathscr{H}} + (1/2 + U/3)\varepsilon_n\big)}\right).
$$

Remark 4.3. When $\mathscr{H} = \{h\}$ (only one function), let $W_i = h(\mathbf{Z}_i)$, then $\sigma^2_{\mathscr{H}} = \mathbf{Var}(W) =: \sigma^2$, (6) yields that

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} W_i - \mathbb{E}(W) \geq \varepsilon_n\Big) \leq \exp\Big(-\frac{n\varepsilon_n^2}{2\big(\sigma^2 + U\varepsilon_n/3\big)}\Big),$$

which is Bernstein inequality. This fact partially indicates that Bousquet bound of Talagrand's inequality is tight.

5 Excess risk bounds

Next, we apply the uniform concentration inequalities to our excess risks. For simplicity, we denote

$$\widehat{R}_n^c(f) = \widehat{R}_n(f) - R(f) = \frac{1}{n}\sum_{i=1}^{n}\Big(l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big) - \mathbb{E}l\big(\mathbf{Y}_i, f(\mathbf{X}_i)\big)\Big)$$

Corollary 5.1. *Suppose the loss function $l(\cdot,\cdot)$ is uniformly bounded by a constant U, then for $t > 0$, with probability at least $1-\delta$*

$$\sup_{f\in\mathscr{F}} |\widehat{R}_n^c(f)| \leq \mathbb{E}\sup_{f\in\mathscr{F}} |\widehat{R}_n^c(f)| + \sqrt{\frac{2\log(1/\delta)}{n}\big(\sigma^2_{\mathscr{F}} + \mathbb{E}\sup_{f\in\mathscr{F}} |\widehat{R}_n^c(f)|\big)} + \frac{U\log(1/\delta)}{3n},$$

where $\sigma^2_{\mathscr{F}}$ is defined as

$$\sigma^2_{\mathscr{F}} = \sup_{f\in\mathscr{F}} \mathbf{Var}\big(l(\mathbf{Y}, f(\mathbf{X}))\big).$$

Alternatively, for any $\varepsilon_n > 0$,

$$\mathbb{P}\Big(\sup_{f\in\mathscr{F}} |\widehat{R}_n^c(f)| - \mathbb{E}\sup_{f\in\mathscr{F}} |\widehat{R}_n^c(f)| \geq \varepsilon_n\Big) \leq \exp\Big(-\frac{n\varepsilon_n^2}{2\big(\sigma^2_{\mathscr{F}} + \mathbb{E}\sup_{f\in\mathscr{F}} |\widehat{R}_n^c(f)| + U\varepsilon_n/3\big)}\Big).$$

Furthermore, if

$$\varepsilon_n \geq 4\mathbb{E}\sup_{f\in\mathscr{F}} |\mathbf{Rad}_n(l\bullet f)|, \quad (l\bullet f)(\mathbf{Z}) = l\big(\mathbf{Y}, f(\mathbf{X})\big)$$

we have,

$$\mathbb{P}\Big(\sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)| \geq \varepsilon_n\Big) \leq \exp\Big(-\frac{n\varepsilon_n^2}{8\big(\sigma^2_{\mathscr{F}} + (1/2 + U/3)\varepsilon_n\big)}\Big).$$

From Corollary 5.1, to derive a probabilistic bound for an excess risk, it suffices to compute and upper bound the Rademacher complexity of $(l\bullet f)(\mathbf{Z}) = l\big(\mathbf{Y}, f(\mathbf{X})\big); f\in\mathscr{F}$.

References

[Bousquet, 2002] Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500.

[Ledoux and Talagrand, 1991] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media.