

## 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

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

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

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

The solution to bound those two empirical processes is to introduce **Rademacher complexity** to measure the complexity of the functional space  $\mathcal{F}$ . 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  $\tilde{\mathcal{D}}_n = (\tilde{\mathbf{X}}_i, \tilde{\mathbf{Y}}_i)_{i=1, \dots, n}$  as the independent copy of  $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1, \dots, n}$ , that is,  $(\tilde{\mathbf{X}}_i, \tilde{\mathbf{Y}}_i) \stackrel{d}{=} (\mathbf{X}_i, \mathbf{Y}_i)$  and samples in  $\{\mathcal{D}_n, \tilde{\mathcal{D}}_n\}$

are all independent.

$$\begin{aligned}
A_1 &= \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left( l(\mathbf{Y}_i, f(\mathbf{X}_i)) - \mathbb{E} l(\mathbf{Y}_i, f(\mathbf{X}_i)) \right) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left( l(\mathbf{Y}_i, f(\mathbf{X}_i)) - \tilde{\mathbb{E}} l(\tilde{\mathbf{Y}}, f(\tilde{\mathbf{X}})) \right) \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \left( \frac{1}{n} \sum_{i=1}^n l(\mathbf{Y}_i, f(\mathbf{X}_i)) - \frac{1}{n} \sum_{i=1}^n \tilde{\mathbb{E}} l(\tilde{\mathbf{Y}}_i, f(\tilde{\mathbf{X}}_i)) \right) \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \tilde{\mathbb{E}} \frac{1}{n} \sum_{i=1}^n \left( l(\mathbf{Y}_i, f(\mathbf{X}_i)) - l(\tilde{\mathbf{Y}}_i, f(\tilde{\mathbf{X}}_i)) \right) \leq \mathbb{E} \tilde{\mathbb{E}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left( l(\mathbf{Y}_i, f(\mathbf{X}_i)) - l(\tilde{\mathbf{Y}}_i, f(\tilde{\mathbf{X}}_i)) \right)
\end{aligned} \tag{1}$$

$$= \mathbb{E} \tilde{\mathbb{E}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \rho_i \left( l(\mathbf{Y}_i, f(\mathbf{X}_i)) - l(\tilde{\mathbf{Y}}_i, f(\tilde{\mathbf{X}}_i)) \right) \leq 2 \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \rho_i l(\mathbf{Y}_i, f(\mathbf{X}_i)) \right|, \tag{2}$$

where  $(\rho_i)_{i=1, \dots, n}$  are i.i.d. Rademacher random variables independent with  $\mathcal{D}_n$  and  $\tilde{\mathcal{D}}_n$ , with  $\rho_i$  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  $(\mathcal{D}_n, \tilde{\mathcal{D}}_n)$  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  $(\mathbf{Z}_i)_{i=1, \dots, n}$  indexed by  $h \in \mathcal{H}$  as:

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

Its corresponding Rademacher process is defined as:

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

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

$$\frac{1}{2} \mathbb{E} \sup_{h \in \mathcal{H}} |\mathbf{Rad}_n(\tilde{h})| \leq \mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \left( h(\mathbf{Z}_i) - \mathbb{E} h(\mathbf{Z}_i) \right) \right| \leq 2 \mathbb{E} \sup_{h \in \mathcal{H}} |\mathbf{Rad}_n(h)|, \tag{3}$$

where  $\tilde{h}(\mathbf{Z}) = h(\mathbf{Z}) - \mathbb{E} h(\mathbf{Z})$ ,  $(\tilde{\mathbf{Z}}_1, \dots, \tilde{\mathbf{Z}}_n)$  is independent copy of  $(\mathbf{Z}_1, \dots, \mathbf{Z}_n)$ ,  $\mathbb{E} \sup_{h \in \mathcal{H}} \mathbf{Rad}_n(h)$  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  $\rho$  and  $h(\mathbf{Z})$ . Suppose  $h$  restrains only one constant, say  $h(\mathbf{z}) = 1$ ,

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

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

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  $l(\mathbf{Y}_i, f(\mathbf{X}_i)); f \in \mathcal{F}$ . Talagrand's contraction Lemma is proposed to address this target.

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

$$\mathbb{E} \sup_{h \in \mathcal{H}} |\mathbf{Rad}_n(\varphi \circ h)| \leq L \mathbb{E} \sup_{h \in \mathcal{H}} |\mathbf{Rad}_n(h)|. \quad (4)$$

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

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  $\mathcal{H}$  are uniformly bounded by a constant  $U$ , then*

$$\mathbb{E} \sup_{h \in \mathcal{H}} |\mathbf{Rad}_n(h^2)| \leq 2U \mathbb{E} \sup_{h \in \mathcal{H}} |\mathbf{Rad}_n(h)|.$$

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

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

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

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

## 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:

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

**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  $1 - \delta$*

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

where  $\sigma_{\mathcal{H}}^2$  is defined as

$$\sigma_{\mathcal{H}}^2 = \sup_{h \in \mathcal{H}} \mathbf{Var}(h(\mathbf{Z})).$$

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  $\varepsilon_n > 0$ ,*

$$\mathbb{P}\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{H}} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{H}} \geq \varepsilon_n\right) \leq \exp\left(-\frac{n\varepsilon_n^2}{2(\sigma_{\mathcal{H}}^2 + \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{H}} + U\varepsilon_n/3)}\right). \tag{6}$$

Furthermore, if

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

we have,

$$\mathbb{P}\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{H}} \geq \varepsilon_n\right) \leq \exp\left(-\frac{n\varepsilon_n^2}{8(\sigma_{\mathcal{H}}^2 + (1/2 + U/6)\varepsilon_n)}\right).$$

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

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

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

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

**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  $1 - \delta$*

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

where  $\sigma_{\mathcal{F}}^2$  is defined as

$$\sigma_{\mathcal{F}}^2 = \sup_{f \in \mathcal{F}} \mathbf{Var}(l(\mathbf{Y}, f(\mathbf{X}))).$$

Alternatively, for any  $\varepsilon_n > 0$ ,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\widehat{R}_n^c(f)| - \mathbb{E} \sup_{f \in \mathcal{F}} |\widehat{R}_n^c(f)| \geq \varepsilon_n\right) \leq \exp\left(-\frac{n\varepsilon_n^2}{2(\sigma_{\mathcal{F}}^2 + \mathbb{E} \sup_{f \in \mathcal{F}} |\widehat{R}_n^c(f)| + U\varepsilon_n/3)}\right).$$

Furthermore, if

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

we have,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| \geq \varepsilon_n\right) \leq \exp\left(-\frac{n\varepsilon_n^2}{8(\sigma_{\mathcal{F}}^2 + (1/2 + U/3)\varepsilon_n)}\right).$$

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

## 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.