

## LECTURE 8: RADEMACHER COMPLEXITY

Recall that in the proof of the Glivenko-Cantelli theorem we used the Rademacher random variables  $\sigma_1, \dots, \sigma_n$  which are iid uniform  $\{\pm 1\}$  random variables.

**Definition.** Let  $\mu$  be a probability measure on  $\mathcal{X}$  and assume that  $X_1, \dots, X_n$  are independent random variables according to  $\mu$ . Let  $\mathcal{F}$  be a class of functions mapping from  $\mathcal{X}$  to  $\mathbb{R}$ . Define the random variable

$$\hat{\mathcal{R}}_n(\mathcal{F}) := \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \mid X_1, \dots, X_n \right],$$

where  $\sigma_1, \dots, \sigma_n$  are independent uniform  $\{\pm 1\}$ -valued random variables.  $\hat{\mathcal{R}}_n(\mathcal{F})$  is called the *empirical Rademacher averages* of  $\mathcal{F}$ . Note that it depends on the sample and can be actually computed. Essentially it measures the correlation between a random noise (labeling) and functions in the function class  $\mathcal{F}$ , in the supremum sense. The *Rademacher averages* of  $\mathcal{F}$  is

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}[\hat{\mathcal{R}}_n(\mathcal{F})].$$

For a function class  $\mathcal{F}$  and a sample  $S = \{x_1, \dots, x_n\}$ , we would like to bound the random quantity

$$\phi(S) := \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{P}_n f) = \sup_{f \in \mathcal{F}} \left( \mathbb{E}[f(X)] - \frac{1}{n} \sum_{i=1}^n f(x_i) \right).$$

First, we bound the difference between the random variable and its mean by using McDiarmid inequality. Consider another sample  $S'$  which only differs from  $S$  at one example. Then we have

$$|\phi(S) - \phi(S')| = \left| \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{P}_n f) - \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{P}'_n f) \right| \leq \frac{c}{n}$$

where we assume that  $f \in [a, b]$  with  $c = b - a$ . Then by McDiarmid inequality we have

$$P \left( \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{P}_n f) - \mathbb{E} \left[ \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{P}_n f) \right] \geq \epsilon \right) \leq \exp \left( -\frac{2n\epsilon^2}{c^2} \right).$$

Thus, by setting  $\delta = \exp(-2n\epsilon^2/c^2)$  we have  $\forall f \in \mathcal{F}$ ,

$$\phi(S) \leq \mathbb{E}[\phi(S)] + \sqrt{\frac{c^2 \log(1/\delta)}{2n}}.$$

Next, we related the  $\mathbb{E}[\phi(S)]$  with the Rademacher averages. From now on, we define  $S' = \{X'_1, \dots, X'_n\}$  to

be a ghost sample of  $S$  (not the same  $S'$  before). Note that

$$\begin{aligned}
\mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{P}_n f) \right] &= \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} \left( \mathbb{E}[f(X)] - \frac{1}{n} \sum_{i=1}^n f(x_i) \right) \right] \\
&= \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n f(X'_i) - \frac{1}{n} \sum_{i=1}^n f(X_i) \mid X_1, \dots, X_n \right] \right] \\
&\stackrel{\text{Jensen}}{\leq} \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(X'_i) - f(X_i)) \right] \\
&= \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i (f(X'_i) - f(X_i)) \right] \\
&\leq \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \left( \frac{1}{n} \sum_{i=1}^n \sigma_i f(X'_i) \right) + \sup_{f \in \mathcal{F}} \left( -\frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right) \right] \\
&= 2\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

So we have shown that  $\mathbb{E}_S[\phi(S)] \leq 2\mathcal{R}_n(\mathcal{F})$ . Combine it with the first step (with  $c = 1$ ), we have shown the first part of the following theorem:

**Theorem 8-1.** *Let  $\mathcal{F}$  be a set of binary-valued  $\{0, 1\}$  functions. For all  $\delta > 0$ , with probability at least  $1 - \delta$ ,*

$$\forall f \in \mathcal{F}, \quad \mathbb{P}f \leq \mathbb{P}_n f + 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

and also with probability at least  $1 - \delta$ ,

$$\forall f \in \mathcal{F}, \quad \mathbb{P}f \leq \mathbb{P}_n f + 2\hat{\mathcal{R}}_n(\mathcal{F}) + C\sqrt{\frac{\log(2/\delta)}{n}},$$

where  $C = \sqrt{2} + 1/\sqrt{2}$ .

PROOF.

First part has been proven above. For the second part, we apply the McDiarmid's inequality again to the empirical Rademacher averages

$$\hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \mid X_1, \dots, X_n \right].$$

Note that  $\hat{\mathcal{R}}_n(\mathcal{F})$  is a function of  $X_1, \dots, X_n$  and satisfies the condition of the McDiarmid's inequality with bounded difference at most  $1/n$ . So we have

$$P \left( 2\mathcal{R}_n(\mathcal{F}) - 2\hat{\mathcal{R}}_n(\mathcal{F}) > \epsilon \right) \leq \exp(-n\epsilon^2/2).$$

So with probability at least  $1 - \delta$ ,

$$2\mathcal{R}_n(\mathcal{F}) \leq 2\hat{\mathcal{R}}_n(\mathcal{F}) + \sqrt{\frac{2 \log(1/\delta)}{n}}.$$

So if we allow each step to be wrong with  $\delta/2$  probability, then we have with probability at least  $1 - \delta$ ,

$$\forall f \in \mathcal{F}, \quad \mathbb{P}f \leq \mathbb{P}_n f + 2\hat{\mathcal{R}}_n(\mathcal{F}) + C\sqrt{\frac{\log(2/\delta)}{n}}$$

where  $C = \sqrt{2} + 1/\sqrt{2}$ .

□

Assume that  $\mathcal{H}$  is the hypothesis space and  $\mathcal{F} = \ell \circ \mathcal{H} = \{f : f(x, y) = \ell(y, h(x)), \forall h \in \mathcal{H}\}$  is the class induced from  $\mathcal{H}$ . Then the Rademacher averages of  $\mathcal{H}$  and  $\mathcal{F}$  are quite related. In fact, if we assume  $\mathcal{Y} = \{\pm 1\}$ , then we have

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}) &= \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i I(Y_i \neq h(X_i)) \right] \\ &= \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \frac{1}{2} (1 - Y_i h(X_i)) \right] \\ &= \frac{1}{2} \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i Y_i h(X_i) \right] \\ &= \frac{1}{2} \mathcal{R}_n(\mathcal{H}). \end{aligned}$$

The Rademacher average  $\mathcal{R}_n(\mathcal{H})$  can actually be computed. Notice that

$$\begin{aligned} \frac{1}{2} \mathcal{R}_n(\mathcal{H}) &= \frac{1}{2} \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i Y_i h(X_i) \right] \\ &= \frac{1}{2} + \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n -\frac{1 - \sigma_i h(X_i)}{2} \right] \\ &= \frac{1}{2} - \mathbb{E} \left[ \inf_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \frac{1 - \sigma_i h(X_i)}{2} \right] \\ &= \frac{1}{2} - \mathbb{E} \left[ \inf_{h \in \mathcal{H}} \hat{R}_n(h, \sigma) \right] \end{aligned}$$

where  $\hat{R}_n(h, \sigma)$  is the empirical risk of classifier  $h$  with respect to random label  $\sigma = [\sigma_1, \dots, \sigma_n]$ . When  $\mathcal{H}$  is so large that it can fit every random labeling perfectly, we have  $\mathcal{R}_n(\mathcal{H}) = 1/2$  and the bound becomes meaningless.

The Rademacher average is related to the growth function and VC dimension. One can bound the Rademacher average by the growth function or VC dimension. We could estimate Rademacher averages for function classes which are built from simpler classes. The following is a list of properties about Rademacher averages.

1. If  $\mathcal{F} \subset \mathcal{G}$  then  $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{G})$ . It follows from the definition.
2.  $\mathcal{R}_n(c \cdot \mathcal{F}) = |c| \mathcal{R}_n(\mathcal{F})$ , where  $c \cdot \mathcal{F} = \{x \mapsto cf(x) : f \in \mathcal{F}\}$ . Since we have

$$\mathcal{R}_n(c\mathcal{F}) = \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n c\sigma_i f(X_i) \right]$$

where  $\sigma_i$ 's are iid Rademacher random variables. Since  $|c|\sigma_i$  has the same distribution as  $c\sigma_i$ , we have

$$\mathcal{R}_n(c\mathcal{F}) = |c| \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right] = |c| \mathcal{R}_n(\mathcal{F}).$$

3.  $\mathcal{R}_n(\mathcal{F} + g) = \mathcal{R}_n(\mathcal{F})$ , where  $\mathcal{F} + g$  is defined as  $\{f + g : \forall f \in \mathcal{F}\}$  and  $g$  is a fixed function. To show

this, we have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F} + g) &= \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i [f(X_i) + g(X_i)] \right] \\
&= \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right] + \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right] \\
&= \mathcal{R}_n(\mathcal{F})
\end{aligned}$$

since the second term is zero.

4. Let the convex hull of a set of functions  $\mathcal{F}$  be defined as

$$\text{conv}(\mathcal{F}) = \left\{ \sum_{i=1}^k \alpha_i f_i : k \geq 1, \alpha_i \geq 0, \sum_{i=1}^k \alpha_i = 1, f_1, \dots, f_k \in \mathcal{F} \right\}.$$

Then we have  $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(\text{conv}(\mathcal{F}))$  since

$$\begin{aligned}
\mathcal{R}_n(\text{conv}(\mathcal{F})) &= \mathbb{E} \left[ \sup_{f_j \in \mathcal{F}, \|\alpha\|_1=1} \frac{1}{n} \sum_{i=1}^n \sigma_i \sum_{j=1}^k \alpha_j f_j(X_i) \right] \\
&= \mathbb{E} \left[ \sup_{f_j \in \mathcal{F}} \sup_{\|\alpha\|_1=1} \sum_{j=1}^k \alpha_j \left( \frac{1}{n} \sum_{i=1}^n \sigma_i f_j(X_i) \right) \right] \\
&= \mathbb{E} \left[ \sup_{f_j \in \mathcal{F}} \max_j \frac{1}{n} \sum_{i=1}^n \sigma_i f_j(X_i) \right] \\
&= \mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

5. Ledoux-Talagrand contraction inequality: If  $\phi_i$  is Lipschitz, i.e. it satisfies  $|\phi_i(a) - \phi_i(b)| \leq L|a - b|$ , then

$$\mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i \phi_i(f(X_i)) \right] \leq L \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right] = L \mathcal{R}_n(\mathcal{F}).$$