Stat 928: Statistical Learning Theory **Lecture: 11**

Symmetrization and Rademacher Averages

Instructor: Sham Kakade

1 Rademacher Averages

Recall that we are interested in bounding the difference between empirical and true expectations uniformly over some function class $\mathcal{G}$. In the context of classification or regression, we are typically interested in a class $\mathcal{G}$ that is the *loss class* associated with some function class $\mathcal{F}$. That is, given a *bounded* loss function $\ell : \mathcal{D} \times \mathcal{Y} \to [0, 1]$, we consider the class

$$\ell_{\mathcal{F}} := \{(x, y) \mapsto \ell(f(x), y) \mid f \in \mathcal{F}\} \ .$$

Rademacher averages give us a powerful tool to obtain uniform convergence results. We begin by examining the quantity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[g(Z)\right] - \frac{1}{m}\sum_{i=1}^{m} g(Z_i)\right)\right] \ ,$$

where $Z, \{Z_i\}_{i=1}^{m}$ are i.i.d. random variables taking values in some space $\mathcal{Z}$ and $\mathcal{G} \subseteq [a, b]^{\mathcal{Z}}$ is a set of bounded functions. We will later show that the random quantity we are interested in, namely

$$\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[g(Z)\right] - \frac{1}{m}\sum_{i=1}^{m} g(Z_i)\right) \ ,$$

will be close to the above expectation with high probability.

Let $\epsilon_1, \ldots, \epsilon_m$ be i.i.d. $\{\pm\}$-valued random variables with $\mathbb{P}\left(\epsilon_i = +1\right) = \mathbb{P}\left(\epsilon_i = -1\right) = 1/2$. These are also independent of the sample $Z_1, \ldots, Z_m$. Define the *empirical Rademacher average* of $\mathcal{G}$ as

$$\hat{\mathfrak{R}}_m(\mathcal{G}) := \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i g(Z_i) \,\middle|\, Z_1^m\right] \ .$$

The *Rademacher average* of $\mathcal{G}$ is defined as

$$\mathfrak{R}_m(\mathcal{G}) := \mathbb{E}\left[\hat{\mathfrak{R}}_m(\mathcal{G})\right] \ .$$

Theorem 1.1. *We have,*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[g(Z)\right] - \frac{1}{m}\sum_{i=1}^{m} g(Z_i)\right)\right] \le 2\mathfrak{R}_m(\mathcal{G}) \ .$$

Proof. Introduce the *ghost sample* $Z_1', \ldots, Z_m'$. By that we mean that Z_i''s are independent of each other and of Z_i's and have the same distribution as the latter. Then we have,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[g(Z)\right]-\frac{1}{m}\sum_{i=1}^{m}g(Z_i)\right)\right]\\
&=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}(\mathbb{E}\left[g(Z)\right]-g(Z_i))\right)\right]\\
&=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[g(Z_i')-g(Z_i)|Z_1^m\right]\right)\right]\\
&\le\mathbb{E}\left[\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}(g(Z_i')-g(Z_i))\right)\middle|Z_1^m\right]\right]\\
&=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}(g(Z_i')-g(Z_i))\right)\right]\\
&=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\epsilon_i(g(Z_i')-g(Z_i))\right)\right]\\
&\le\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\epsilon_i g(Z_i')\right]+\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\epsilon_i g(Z_i)\right]\\
&=2\mathfrak{R}_m(\mathcal{G})\ .
\end{aligned}$$

□

Since $\mathfrak{R}_m(-\mathcal{G})=\mathfrak{R}_m(\mathcal{G})$, we have the following corollary.

Corollary 1.2. *We have,*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}g(Z_i)-\mathbb{E}\left[g(Z)\right]\right)\right]\le 2\mathfrak{R}_m(\mathcal{G})\ .$$

Since $g(X_i)\in[a,b]$,

$$\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[g(Z)\right]-\frac{1}{m}\sum_{i=1}^{m}g(Z_i)\right)$$

does not change by more than $(b-a)/m$ if some Z_i is changed to Z_i'. Applying the bounded differences inequality, we get the following corollary.

Corollary 1.3. *With probability at least* $1-\delta$,

$$\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[g(Z)\right]-\frac{1}{m}\sum_{i=1}^{m}g(Z_i)\right)\le 2\mathfrak{R}_m(\mathcal{G})+(b-a)\sqrt{\frac{\ln(1/\delta)}{2m}}$$

Recall that we denote the empirical ℓ-loss minimizer by $\hat{f}_\ell^*$. We refer to $L_\ell(\hat{f}_\ell^*)-\min_{f\in\mathcal{F}}L_\ell(f)$ as the estimation error. The next theorem bounds the estimation error using Rademacher averages.

2 Expected Regret

Now let us examine the expected regret of the empirical risk minimizer (e.g. analogous to the statistical risk). Let

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} g(Z_i)$$

where τ is the training set and

$$g^* = \arg\min_{g \in \mathcal{G}} \mathbb{E}\left[g(Z)\right]$$

which is true minimizer.

Lemma 2.1. *The expected regret is:*

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\hat{g}(Z)\right] - \mathbb{E}\left[g^*(Z)\right]\right] &\leq 2\mathfrak{R}_m(\mathcal{G}) + \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} g^*(Z_i) - \mathbb{E}\left[g^*(Z)\right]\right] \\
&\leq 4\mathfrak{R}_m(\mathcal{G})
\end{aligned}$$

where the expectation is with respect $\hat{g}$ (due to randomness in the training set).

Proof. Let

$$\hat{g}$$

The expected regret is:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\hat{g}(Z)\right] - \mathbb{E}\left[g^*(Z)\right]\right] &\leq \mathbb{E}\left[\mathbb{E}\left[\hat{g}(Z)\right] - \frac{1}{m}\sum_{i=1}^{m}\hat{g}(Z_i) + \frac{1}{m}\sum_{i=1}^{m}\hat{g}(Z_i) - \mathbb{E}\left[g^*(Z)\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\hat{g}(Z)\right] - \frac{1}{m}\sum_{i=1}^{m}\hat{g}(Z_i) + \frac{1}{m}\sum_{i=1}^{m}g^*(Z_i) - \mathbb{E}\left[g^*(Z)\right]\right] \\
&\leq \mathbb{E}\left[\sup g \in \mathcal{G}\left(\mathbb{E}\left[\hat{g}(Z)\right] - \frac{1}{m}\sum_{i=1}^{m}\hat{g}(Z_i)\right)\right] + \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}g^*(Z_i) - \mathbb{E}\left[g^*(Z)\right]\right] \\
&\leq 2\mathfrak{R}_m(\mathcal{G}) + \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}g^*(Z_i) - \mathbb{E}\left[g^*(Z)\right]\right]
\end{aligned}$$

The final claim is straightforward. □

3 Growth function

Consider the case $\mathcal{Y} = \{\pm 1\}$ (classification). Let ℓ be the 0-1 loss function and $\mathcal{F}$ be a class of ± 1-valued functions. We can relate the Rademacher average of $\ell_{\mathcal{F}}$ to that of $\mathcal{F}$ as follows.

Lemma 3.1. *Suppose $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ and let $\ell(y', y) = \mathbf{1}\left[y' \neq y\right]$ be the 0-1 loss function. Then we have,*

$$\mathfrak{R}_m(\ell_{\mathcal{F}}) = \frac{1}{2}\mathfrak{R}_m(\mathcal{F}) \ .$$

Proof. Note that we can write $\ell(y', y)$ as $(1 - yy')/2$. Then we have,

$$\begin{aligned}
\mathfrak{R}_m(\ell_{\mathcal{F}}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i \frac{1 - Y_i f(X_i)}{2} \middle| X_1^m, Y_1^m\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i \frac{Y_i f(X_i)}{2} \middle| X_1^m, Y_1^m\right] \qquad (1) \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} (-\epsilon_i Y_i) f(X_i) \middle| X_1^m, Y_1^m\right] \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i f(X_i) \middle| X_1^m, Y_1^m\right] \qquad (2) \\
&= \frac{1}{2}\mathfrak{R}_m(\mathcal{F}) \ .
\end{aligned}$$

Equation (1) follows because $\mathbb{E}\left[\epsilon_i | X_1^m, Y_1^m\right] = 0$. Equation (2) follows because $-\epsilon_i Y_i$'s jointly have the same distribution as ϵ_i's. □

Note that the Rademacher average of the class $\mathcal{F}$ can also be written as

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}\left[\sup_{a\in\mathcal{F}_{|X_1^m}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i a_i\right] \ ,$$

where $\mathcal{F}_{|X_1^m}$ is the function class $\mathcal{F}$ restricted to the set $X_1, \ldots, X_m$. That is,

$$\mathcal{F}_{|X_1^m} := \{((f(X_1), \ldots, f(X_m)) \mid f \in \mathcal{F}\} \ .$$

Note that $\mathcal{F}_{|X_1^m}$ is finite and

$$|\mathcal{F}_{|X_1^m}| \leq \min\{|\mathcal{F}|, 2^m\} \ .$$

Thus we can define the *growth function* as

$$\Pi_{\mathcal{F}}(m) := \max_{x_1^m \in \mathcal{X}^m} |\mathcal{F}_{|x_1^m}| \ .$$

The following lemma due to Massart allows us to bound the Rademacher average in terms of the growth function.

Lemma 3.2. *(Finite Class Lemma) Let $\mathcal{A}$ be some finite subset of $\mathbb{R}^m$ and $\epsilon_1, \ldots, \epsilon_m$ be independent Rademacher random variables. Let $r = \sup_{a\in\mathcal{A}} \|a\|$. Then, we have,*

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i a_i\right] \leq \frac{r\sqrt{2\ln|\mathcal{A}|}}{m} \ .$$

Proof. Let

$$\mu = \mathbb{E}\left[\sup_{a\in\mathcal{A}} \sum_{i=1}^{m} \epsilon_i a_i\right] \ .$$

We have, for any $\lambda > 0$,

$$
\begin{aligned}
e^{\lambda\mu} &\leq \mathbb{E}\left[\exp\left(\lambda \sup_{a\in\mathcal{A}} \sum_{i=1}^{m} \epsilon_i a_i\right)\right] && \text{Jensen's inequality} \\
&= \mathbb{E}\left[\sup_{a\in\mathcal{A}} \exp\left(\lambda \sum_{i=1}^{m} \epsilon_i a_i\right)\right] \\
&\leq \mathbb{E}\left[\sum_{a\in\mathcal{A}} \exp\left(\lambda \sum_{i=1}^{m} \epsilon_i a_i\right)\right] \\
&= \sum_{a\in\mathcal{A}} \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{m} \epsilon_i a_i\right)\right] \\
&= \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} \mathbb{E}\left[\exp\left(\lambda \epsilon_i a_i\right)\right] \\
&\leq \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} e^{\lambda^2 a_i^2/2} && \because \text{Hoeffding's lemma} \\
&= \sum_{a\in\mathcal{A}} e^{\lambda^2 \|a\|^2/2} \\
&\leq |\mathcal{A}| e^{\lambda^2 r^2/2}
\end{aligned}
$$

Taking logs and dividing by λ, we get that, for any $\lambda > 0$,

$$\mu \leq \frac{\ln|\mathcal{A}|}{\lambda} + \frac{\lambda r^2}{2} \ .$$

Setting $\lambda = \sqrt{2\ln|\mathcal{A}|/r^2}$ gives,

$$\mu \leq r\sqrt{2\ln|\mathcal{A}|} \ ,$$

which proves the lemma. □