CMSC 35900 (Spring 2008) Learning Theory **Lecture: 9**

Rademacher Averages

Instructors: Sham Kakade and Ambuj Tewari

1 Bounded Differences Inequality

Suppose $Z_1, \ldots, Z_m$ are independent random variables taking values in some space $\mathcal{Z}$ and $f : \mathcal{Z}^m \to \mathbb{R}$ is a function that satisfies, for all i,

$$\sup_{z_1,\ldots,z_m,z_i'} |f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_m) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m)| \le c_i$$

for some constants $c_1, \ldots, c_m$. Then we have,

$$\mathbb{P}\left(f(Z_1^m) - \mathbb{E}\left[f(Z_1^m)\right] \ge t\right) \le \exp\left(\frac{-2t^2}{\sum_{i=1}^m c_i^2}\right) .$$

2 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 $\phi : \mathcal{D} \times \mathcal{Y} \to [0, 1]$, we consider the class

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

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. By the bounded differences inequality, 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 $\mathbb{P}(\epsilon_i = +1) = \mathbb{P}(\epsilon_i = -1) = 1/2$. 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 2.1. *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 2.2. *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 2.3. *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 ϕ-risk minimizer by $\hat{f}^*_\phi$. We refer to $L_\phi(\hat{f}^*_\phi) - \min_{f\in\mathcal{F}} L_\phi(f)$ as the estimation error. The next theorem bounds the estimation error using Rademacher averages.

Theorem 2.4. *Let $\phi_{\mathcal{F}}$ denote the loss class associated with $\mathcal{F}$. Then, we have, with probability at least $1 - 2\delta$,*

$$L_\phi(\hat{f}_\phi^*) - \min_{f \in \mathcal{F}} L_\phi(f) \le 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + 2\sqrt{\frac{\ln(1/\delta)}{2m}} \; .$$

Proof. Denote the function in $\mathcal{F}$ with minimum risk by $f_{\mathcal{F}}^*$. Since the loss function takes values in the interval $[0, 1]$, applying the previous corollary to the class $\phi_{\mathcal{F}}$, we get, with probability at least $1 - 2\delta$,

$$L_\phi(\hat{f}_\phi^*) - \hat{L}_\phi(\hat{f}_\phi^*) \le 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + \sqrt{\frac{\ln(1/\delta)}{2m}} \; .$$

Also, by the bounded differences inequality, we have with probability at least $1 - \delta$,

$$\hat{L}_\phi(f_{\mathcal{F}}^*) - L_\phi(f_{\mathcal{F}}^*) \le \sqrt{\frac{\ln(1/\delta)}{2m}} \; .$$

Thus we have, with probability at least $1 - 2\delta$,

$$\begin{aligned} L_\phi(\hat{f}_\phi^*) - L_\phi(f_{\mathcal{F}}^*) &\le \hat{L}_\phi(\hat{f}_\phi^*) - L_\phi(f_{\mathcal{F}}^*) + 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + \sqrt{\frac{\ln(1/\delta)}{2m}} \\ &\le \hat{L}_\phi(\hat{f}_\phi^*) - \hat{L}_\phi(f_{\mathcal{F}}^*) + 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + 2\sqrt{\frac{\ln(1/\delta)}{2m}} \\ &\le 0 + 2\mathfrak{R}_m(\phi_{\mathcal{F}}) + 2\sqrt{\frac{\ln(1/\delta)}{2m}} \end{aligned}$$

□