SYS 6016/4582: Machine Learning **Spring 2017**

Lecture 5

Instructor: Quanquan Gu *Date: Feb 2^{nd}, 2017* *Scriber: Robert Carroll*

Last time we introduced **Bounded Difference Inequality** to bound the supremum of the difference between the empirical risk and expected risk, i.e.,

$$\sup_{h\in\mathcal{H}} \left|\hat{R}_n(h) - R(h)\right| \le \max\left\{\sup_{h\in\mathcal{H}} \hat{R}_n(h) - R(h), \sup_{h\in\mathcal{H}} R(h) - \hat{R}_n(h)\right\}.$$

With the fact that

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left|\hat{R}_n(h) - R(h)\right| \ge t\right) \le \mathbb{P}\Bigg(\underbrace{\sup_{h\in\mathcal{H}} \hat{R}_n(h) - R(h)}_{G_n(\mathcal{H})} \ge t\Bigg) + \mathbb{P}\Bigg(\underbrace{\sup_{h\in\mathcal{H}} R(h) - \hat{R}_n(h)}_{G_n'(\mathcal{H})} \ge t\Bigg),$$

we can bound the term $\sup_{h\in\mathcal{H}} \left|\hat{R}_n(h) - R(h)\right|$ by bounding $G_n(\mathcal{H})$ and $G_n'(\mathcal{H})$ respectively. Since $G_n(\mathcal{H})$ and $G_n'(\mathcal{H})$ can basically be bounded by the same term, we only show how to bound $G_n(\mathcal{H})$.

We restate the Bounded Difference Inequality here.

Theorem 1 (Bounded Difference Inequality) *Let f be a function satisfying*

$$|f(z_1, z_2, ..., z_i, ..., z_n) - f(z_1, z_2, ..., z_i', ..., z_n)| \le c_i$$

for all $i = 1, 2, ..., n$ and for all $z_1, z_2, ..., z_i, z_i', ..., z_n$. Let $Z_1, Z_2, ..., Z_n$ be independent random variables. We have that

$$\mathbb{P}\left(f(Z_1, Z_2, ..., Z_n) - \mathbb{E}[f(Z_1, Z_2, ..., Z_n)]\right) \ge t) \le 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Roadmap to bound $G_n(\mathcal{H})$

Step 1: Prove $G_n(\mathcal{H}) \le \mathbb{E}G_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}$ with probability at least $1-\delta$.
Step 2: Bound $\mathbb{E}G_n(\mathcal{H})$ by symmetrization.

Step 1: Bounding $G_n(\mathcal{H})$

Theorem 2 *Let $G_n(\mathcal{H}) = \sup_{h\in\mathcal{H}} \hat{R}_n(h) - R(h)$, we have the following holds with probability at least $1-\delta$*

$$G_n(\mathcal{H}) - \mathbb{E}\big[G_n(\mathcal{H})\big] \le \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof: Let $Z_i = (X_i, Y_i) \overset{i.i.d.}{\sim} P_{X,Y}$ and $f(Z_1, Z_2, ..., Z_n) = G_n(\mathcal{H})$. First we verify that the function $f(Z_1, \ldots, Z_n)$ satisfies the conditions in Theorem 1. For any i, the following equality holds:

$$\begin{aligned} &|f(Z_1, Z_2, ..., Z_i, ..., Z_n) - f(Z_1, Z_2, ..., Z_i', ..., Z_n)| \\ &= \left| \sup_{h\in\mathcal{H}} \left(\hat{R}_n(h) - R(h)\right) - \sup_{h\in\mathcal{H}} \left(\hat{R}_n'(h) - R(h)\right) \right|, \end{aligned} \tag{1}$$

where $\hat{R}_n'(h)$ is the empirical risk when the observation Z_i is replaced by Z_i', i.e.,

$$\hat{R}_n'(h) = \frac{1}{n} \left[\sum_{j\neq i}^{n} \mathbb{1}\{h(X_j) \neq Y_j\} + \mathbb{1}\{h(X_i') \neq Y_i'\} \right].$$

Thus we can obtain

$$\left| \sup_{h\in\mathcal{H}} \left(\hat{R}_n(h) - R(h)\right) - \sup_{h\in\mathcal{H}} \left(\hat{R}_n'(h) - R(h)\right) \right| \le \sup_{h\in\mathcal{H}} \left| \hat{R}_n(h) - \hat{R}_n'(h) \right|. \tag{2}$$

Using the definition of the empirical risk, the R.H.S. of (2) can be rewritten as:

$$\begin{aligned} &\sup_{h\in\mathcal{H}} \left| \hat{R}_n(h) - \hat{R}_n'(h) \right| \\ &= \sup_{h\in\mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\} - \frac{1}{n} \left[\sum_{j\neq i}^{n} \mathbb{1}\{h(X_j) \neq Y_j\} + \mathbb{1}\{h(X_i') \neq Y_i'\} \right] \right| \\ &= \sup_{h\in\mathcal{H}} \left| \frac{1}{n} \mathbb{1}\{h(X_i) \neq Y_i\} - \frac{1}{n} \mathbb{1}\{h(X_i') \neq Y_i'\} \right| = \frac{1}{n} \triangleq c_i, \end{aligned}$$

where in the third equality we use the fact that indicator functions are either 0 or 1, so the maximal difference between any two indicator functions is 1. Hence we proved that $G_n(\mathcal{H})$ satisfies the conditions in 1 with $c_i = 1/n$. Applying Theorem 1 gives

$$\begin{aligned} \mathbb{P}(G_n(\mathcal{H}) - \mathbb{E}G_n(\mathcal{H}) \ge t) &\le \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2} \right) \\ &= \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} (1/n)^2} \right) \\ &= \exp\left(-2nt^2\right). \end{aligned}$$

Let $\delta = \exp(-2nt^2)$. Solving for t gives

$$t = \sqrt{\frac{\log(1/\delta)}{2n}}.$$

We then have with probability at least $1 - \delta$

$$G_n(\mathcal{H}) - \mathbb{E}G_n(\mathcal{H}) \le \sqrt{\frac{\log(1/\delta)}{2n}},$$

equivalently

$$G_n(\mathcal{H}) \leq \mathbb{E}G_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

This completes the proof. ■

Step 2: Bounding $\mathbb{E}G_n(\mathcal{H})$

Before presenting the result we first lay out some definitions and lemmas that are essential in the proof of the main theorem.

Definition 1 (Symmetric Random Variable) *Z is a symmetric random variable if $\mathbb{P}(Z \geq t) = \mathbb{P}(-Z \geq t) = \mathbb{P}(-Z \leq -t)$. In other words, multiplying by -1 doesn't change the distribution.*

The following lemma shows how to construct a symmetric random variable from i.i.d. random variables.

Lemma 1 *If X, Y are two independent and identically distributed random variables, then $Z = X - Y$ is a symmetric random variable.*

The following lemma is essential to bound $\mathbb{E}G_n(\mathcal{H})$, which is called the symmetrization technique.

Lemma 2 *If $Z_1, Z_2, \ldots, Z_n$ are symmetric random variables, and $\sigma_1, \sigma_2, \ldots, \sigma_n$ are Rademacher random variables, i.e.,*

$$\sigma_i = \begin{cases} 1, & \textit{with probability } 1/2, \\ -1, & \textit{with probability } 1/2. \end{cases}$$

In addition, suppose $\sigma_1, \ldots, \sigma_n$ are independent of $Z_1 \ldots, Z_n$, then $\sum_{i=1}^n Z_i$ has the same distribution as $\sum_{i=1}^n \sigma_i Z_i$.

Based on the definitions and lemmas listed above, the bound of $\mathbb{E}G_n(\mathcal{H})$ can be represented as the following theorem.

Theorem 3 *We have*

$$\mathbb{E}G_n(\mathcal{H}) \leq 2\mathbb{E}_{X,Y}\mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbb{1}\left\{h(X_i) \neq Y_i\right\}\right].$$

Proof: We prove using symmetrization technique with a "ghost" sample $(\widetilde{X}_i, \widetilde{Y}_i)_{i=1}^n \overset{i.i.d}{\sim} P_{X,Y}$ that is independent of the training sample

First, we lay out several claims, which are useful to bound $\mathbb{E}G_n(\mathcal{H})$.

Claim 1 *Given the "ghost" sample* $(\widetilde{X}_i, \widetilde{Y}_i)_{i=1}^n$*, we have the following hold*

$$R(h) = \mathbb{E}_{\widetilde{X},\widetilde{Y}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{h(\widetilde{X}_i) \neq \widetilde{Y}_i\right\}\right].$$

By the definition of $G_n(\mathcal{H})$, we have

$$\begin{aligned}\mathbb{E}G_n(\mathcal{H}) &= \mathbb{E}_{X,Y}\left[\sup_{h\in\mathcal{H}}\left(\hat{R}_n(h) - R(h)\right)\right] \\ &= \mathbb{E}_{X,Y}\left[\sup_{h\in\mathcal{H}}\left(\hat{R}_n(h) - \mathbb{E}_{\widetilde{X},\widetilde{Y}}\left[\tilde{R}_n(h)\right]\right)\right],\end{aligned} \tag{3}$$

where

$$\tilde{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{h(\widetilde{X}_i) \neq \widetilde{Y}_i\right\} \quad \text{and} \quad (\widetilde{X}_i, \widetilde{Y}_i), \overset{i.i.d}{\sim} P_{X,Y}.$$

In addition we have $(\widetilde{X}_i, \widetilde{Y}_i)_{i=1}^n$ are independent from $(X_i, Y_i)_{i=1}^n$, where $(\widetilde{X}_i, \widetilde{Y}_i)_{i=1}^n$ is called the "ghost" sample.

According to (3), we have

$$\begin{aligned}&\mathbb{E}_{X,Y}\left[\sup_{h\in\mathcal{H}}\left(\hat{R}_n(h) - \mathbb{E}_{\widetilde{X},\widetilde{Y}}\left[\tilde{R}_n(h)\right]\right)\right] \\ &= \mathbb{E}_{X,Y}\left[\sup_{h\in\mathcal{H}}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\left(\hat{R}_n(h) - \tilde{R}_n(h)\right)\right] \\ &\leq \mathbb{E}_{X,Y}\left[\mathbb{E}_{\widetilde{X},\widetilde{Y}}\sup_{h\in\mathcal{H}}\left(\hat{R}_n(h) - \tilde{R}_n(h)\right)\right] \\ &= \mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{h(X_i) \neq Y_i\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{h(\widetilde{X}_i) \neq \widetilde{Y}_i\right\}\right)\right],\end{aligned}$$

where the the second inequality comes from Jensen's Inequality and the fact that the supremum function is convex. This implies that

$$\begin{aligned}\mathbb{E}G_n(\mathcal{H}) &\leq \mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{h(X_i) \neq Y_i\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{h(\widetilde{X}_i) \neq \widetilde{Y}_i\right\}\right)\right] \\ &= \mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{1}\left\{h(X_i) \neq Y_i\right\} - \mathbb{1}\left\{h(\widetilde{X}_i) \neq \widetilde{Y}_i\right\}\right]\right)\right].\end{aligned} \tag{4}$$

Since $\mathbb{1}\left\{h(X_i) \neq Y_i\right\}$, and $\mathbb{1}\left\{h(\widetilde{X}_i) \neq \widetilde{Y}_i\right\}$ are two independent and identically distributed random variables, according to Lemma 1, we have $\mathbb{1}\left\{h(X_i) \neq Y_i\right\} - \mathbb{1}\left\{h(\widetilde{X}_i) \neq \widetilde{Y}_i\right\}$

is symmetric. According to Lemma 2, (4) can be further upper bounded by

$$
\begin{aligned}
\mathbb{E}G_n(\mathcal{H}) &\le \mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{1}\left\{h(X_i)\neq Y_i\right\}-\mathbb{1}\left\{h(\widetilde{X}_i)\neq \widetilde{Y}_i\right\}\right]\right)\right]. \\
&= \mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[\mathbb{1}\left\{h(X_i)\neq Y_i\right\}-\mathbb{1}\left\{h(\widetilde{X}_i)\neq \widetilde{Y}_i\right\}\right]\right)\right] \\
&\le \mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{1}\left\{h(X_i)\neq Y_i\right\}+\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)\mathbb{1}\left\{h(\widetilde{X}_i)\neq \widetilde{Y}_i\right\}\right],
\end{aligned}
$$

where the last line comes from the inequality that $\sup_x\big(f(x)+g(x)\big)\le\sup_x f(x)+\sup_x g(x)$. By the linearity of expectaion we have

$$
\begin{aligned}
\mathbb{E}G_n(\mathcal{H}) \le\ & \mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{1}\left\{h(X_i)\neq Y_i\right\}\right] \\
&+\mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)\mathbb{1}\left\{h(\widetilde{X}_i)\neq \widetilde{Y}_i\right\}\right].
\end{aligned}
$$

Since σ_i's are symmetric random variables, $-\sigma_i$ has the same distribution with σ_i. Then changing $-\sigma_i$ into σ_i in the second term does not change the value of the whole expression. Hence we have Furthermore, we can get

$$
\begin{aligned}
\mathbb{E}G_n(\mathcal{H}) &\le \mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{1}\left\{h(X_i)\neq Y_i\right\}\right] \\
&\quad+\mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{X},\widetilde{Y}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)\mathbb{1}\left\{h(\widetilde{X}_i)\neq \widetilde{Y}_i\right\}\right] \\
&= \mathbb{E}_{\sigma}\left[\mathbb{E}_{X,Y}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{1}\left\{h(X_i)\neq Y_i\right\}+\mathbb{E}_{\widetilde{X},\widetilde{Y}}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{1}\left\{h(\widetilde{X}_i)\neq \widetilde{Y}_i\right\}\right] \\
&= 2\mathbb{E}_{X,Y}\mathbb{E}_{\sigma}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{1}\left\{h(X_i)\neq Y_i\right\}. \qquad (5)
\end{aligned}
$$

This completes the proof. ■

Next, we introduce the concept of Rademacher Complexity, which measures the complexity of a class of functions and has a wide range of applications in machine learning theory. Note that the following definition is tailored to 0-1 loss function.

Definition 2 (Rademacher Complexity) *For a given function class $\mathcal{F}$, its Rademacher complexity is defined as*

$$
\mathcal{R}_n(\mathcal{F})=\mathbb{E}_Z\mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right],
$$

where σ_i's are i.i.d. Rademacher random variables, and Z_i's are i.i.d. random variables.

Based on the definition of Rademacher Complexity, we can also define the Empirical Rademacher Complexity.

Definition 3 (Empirical Rademacher Complexity) *For a given hypothesis class $\mathcal{F}$ and a set of observations Z_i's, the empirical Rademacher complexity is*

$$\hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \quad \middle| Z_{1:n} \right].$$

where σ_i's are i.i.d. Rademacher random variables, and $Z_{1:n}$ is a shorthand notation for $Z_1, Z_2, ..., Z_n$.