E0 370 Statistical Learning Theory **Lecture 7 (Sep 6, 2011)**

Rademacher Averages

Lecturer: Shivani Agarwal *Scribe: Shivani Agarwal*

1 Introduction

So far we have seen how to obtain generalization error bounds for learning algorithms that pick a function from a function class of limited capacity or complexity, where the complexity of the class is measured using the growth function or VC-dimension in the binary case, and using covering numbers or the fat-shattering dimension in the real-valued case. These complexity measures however do not take into account the distribution of the data; regardless of whether the distribution generating the data is 'easy' or 'hard' to learn with respect to, one gets the same bound on the difference between the expected (generalization) error and empirical (sample) error. Although we saw that the results we obtained could be refined to give bounds in terms of the VC-entropy or the expected covering numbers, both of which are distribution-dependent quantities, estimating these quantities can be difficult in practice.[1]

In this lecture we will see an alternative notion of the capacity or complexity of a function class, namely the Rademacher averages, which are defined with respect to the distribution generating the data. We will see that the Rademacher averages of a function class can be estimated reliably from data, and that they yield generalization error bounds in terms of data-dependent quantities that can often be estimated or bounded in practice. In order to derive these results, we will need a more powerful concentration inequality than what we have used so far; in particular, we will use McDiarmid's inequality, which generalizes Hoeffding's inequality and bounds the probability of a large deviation for any function of independent random variables for which a change in a single random variable has bounded effect on the function value.

2 McDiarmid's Inequality

Theorem 2.1 (McDiarmid's inequality, also known as bounded differences inequality or Hoeffding-Azuma inequality; McDiarmid, 1989)**.** Let $X_1, \ldots, X_n$ be independent random variables, with X_i taking values in some set A_i. Let $\phi : A_1 \times \ldots \times A_n \rightarrow \mathbb{R}$ be such that $\exists\ c_1, \ldots, c_n > 0$ such that for all i,

$$\sup_{x_1,\ldots,x_n,x_i'} \Big| \phi(x_1, \ldots, x_i, \ldots, x_n) - \phi(x_1, \ldots, x_i', \ldots, x_n) \Big| \ \leq\ c_i \,.$$

Then for any $\epsilon > 0$,

$$\mathbf{P}\Big(\phi(X_1, \ldots, X_n) - \mathbf{E}\big[\phi(X_1, \ldots, X_n)\big] \geq \epsilon\Big) \leq e^{-2\epsilon^2 / \sum_{i=1}^n c_i^2} \,.$$

and

$$\mathbf{P}\Big(\phi(X_1, \ldots, X_n) - \mathbf{E}\big[\phi(X_1, \ldots, X_n)\big] \leq -\epsilon\Big) \leq e^{-2\epsilon^2 / \sum_{i=1}^n c_i^2} \,.$$

Note that Hoeffding's inequality follows as a special case of the above with $A_i = [a_i, b_i]$, $\phi(x_1, \ldots, x_m) = \sum_{i=1}^m x_i$, and $c_i = b_i - a_i$ (clearly, changing x_i can change the sum by at most $b_i - a_i$).

[1] We note that it is possible to show concentration results for certain empirical quantities related to the VC-entropy, which can then be used as empirical estimates for the VC-entropy; however in most cases, obtaining tight estimates or bounds for these empirical quantities is still difficult.

3 Rademacher Averages

Definition. Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. Let $x_1^m \in \mathcal{X}^m$. Then the *empirical Rademacher average* (or *empirical Rademacher complexity*) of $\mathcal{F}$ w.r.t. x_1^m is defined as[2]

$$R_{x_1^m}(\mathcal{F}) \;=\; \mathbf{E}_{\mathbf{r}\in\{\pm 1\}^m}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m r_i f(x_i)\right] .$$

If μ is a probability distribution on $\mathcal{X}$, then the (expected) *Rademacher averages* (or *Rademacher complexities*) of $\mathcal{F}$ w.r.t. μ are defined as[3]

$$R_{m,\mu}(\mathcal{F}) \;=\; \mathbf{E}_{x_1^m\sim\mu^m}\left[R_{x_1^m}(\mathcal{F})\right] .$$

The Rademacher averages can be viewed as measuring the extent to which, on average, functions in $\mathcal{F}$ (or vectors in $\mathcal{F}_{|x_1^m}$) can be correlated with a random sign vector. The Rademacher average should be compared with the VC-entropy or the expected covering numbers, which also provide a distribution-dependent measure of the complexity of a function class. The following result shows that the Rademacher averages can be estimated reliably from the empirical version on a single data sample:

Lemma 3.1. *Let $\mathcal{F} \subseteq [a,b]^{\mathcal{X}}$ for some $a \le b$. Let μ be any probability distribution on $\mathcal{X}$. For any $\epsilon > 0$,*

$$\mathbf{P}_{x_1^m\sim\mu^m}\left[\left|R_{x_1^m}(\mathcal{F}) - R_{m,\mu}(\mathcal{F})\right| \ge \epsilon\right] \;\le\; 2\, e^{-2m\epsilon^2/(b-a)^2} .$$

Proof. Define $\phi : \mathcal{X}^m \to \mathbb{R}$ as $\phi(x_1,\ldots,x_m) = R_{x_1^m}(\mathcal{F})$. Then clearly, $\mathbf{E}_{x_1^m\sim\mu^m}[\phi(x_1,\ldots,x_m)] = R_{m,\mu}(\mathcal{F})$. Moreover, for any $k\in[m]$ and any $x_1,\ldots,x_m,x_k' \in \mathcal{X}$, we have

$$\begin{aligned}
&\left|\phi(x_1,\ldots,x_k,\ldots,x_m) - \phi(x_1,\ldots,x_k',\ldots,x_m)\right| && (1)\\
&= \left|R_{x_1^m}(\mathcal{F}) - R_{(x_1,\ldots,x_k',\ldots,x_m)}(\mathcal{F})\right| && (2)\\
&= \left|\mathbf{E}_{\mathbf{r}\in\{\pm1\}^m}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m r_i f(x_i) - \sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i\neq k} r_i f(x_i) + \frac{1}{m} r_k f(x_k')\right)\right]\right| && (3)\\
&\le \frac{b-a}{m} \quad \text{(since for any } \mathbf{r}\text{, the change in the sum for any fixed } f\in\mathcal{F} \text{ is at most } \tfrac{b-a}{m}\text{)}. && (4)
\end{aligned}$$

The result follows by McDiarmid's inequality. □

In fact, we can show something stronger: the Rademacher averages can be estimated reliably from a single sample $x_1^m \sim \mu^m$ and a *single* instantiation of the random vector $\mathbf{r} \in \{\pm1\}^m$!

Lemma 3.2. *Let $\mathcal{F} \subseteq [a,b]^{\mathcal{X}}$ for some $a \le b$. Let μ be any probability distribution on $\mathcal{X}$. For any $\epsilon > 0$,*

$$\mathbf{P}_{x_1^m\sim\mu^m,\mathbf{r}\in\{\pm1\}^m}\left[\left|\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m r_i f(x_i) - R_{m,\mu}(\mathcal{F})\right| \ge \epsilon\right] \;\le\; 2\, e^{-2m\epsilon^2/((b-a)^2+4\max(a^2,b^2))} .$$

Proof. We will use McDiarmid's inequality again. Define now $\phi : \mathcal{X}^m \times \{\pm1\}^m \to \mathbb{R}$ as $\phi(x_1,\ldots,x_m,r_1,\ldots,r_m) = \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m r_i f(x_i)$. Clearly, $\mathbf{E}_{x_1^m\sim\mu^m,\mathbf{r}\in\{\pm1\}^m}[\phi(x_1,\ldots,x_m)] = R_{m,\mu}(\mathcal{F})$. Now fix any $x_1,\ldots,x_m \in \mathcal{X}$

[2] From here on, we will abbreviate the set $\{-1,1\}$ as $\{\pm1\}$.

[3] When the distribution μ is understood from context, the Rademacher averages $R_{m,\mu}(\mathcal{F})$ are often written simply as $R_m(\mathcal{F})$. Similarly, when x_1^m is understood to be random, the empirical Rademacher average $R_{x_1^m}(\mathcal{F})$, which then becomes a random variable, is often written as $\hat{R}_m(\mathcal{F})$. The notation we use here makes the dependences on μ and x_1^m explicit.

and $r_1, \ldots, r_k \in \{\pm 1\}$. For any $k \in [m]$ and $x'_k \in \mathcal{X}$, we have

$$\begin{aligned}
&\Big|\phi(x_1, \ldots, x_k, \ldots, x_m, r_1, \ldots, r_m) - \phi(x_1, \ldots, x'_k, \ldots, x_m, r_1, \ldots, r_m)\Big| && (5)\\
&= \left|\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m r_i f(x_i) - \sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i\neq k} r_i f(x_i) + \frac{1}{m} r_k f(x'_k)\right)\right| && (6)\\
&\le \frac{b-a}{m} \quad \text{as before.} && (7)
\end{aligned}$$

Moreover, for any $k \in [m]$ and $r'_k \in \{\pm 1\}$, we have

$$\begin{aligned}
&\Big|\phi(x_1, \ldots, x_m, r_1, \ldots, r_k, \ldots, r_m) - \phi(x_1, \ldots, x_m, r_1, \ldots, r'_k, \ldots, r_m)\Big| && (8)\\
&= \left|\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m r_i f(x_i) - \sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i\neq k} r_i f(x_i) + \frac{1}{m} r'_k f(x_k)\right)\right| && (9)\\
&\le \frac{2\max(|a|,|b|)}{m}. && (10)
\end{aligned}$$

The result follows by McDiarmid's inequality. □

Of course, the estimation of $R_{m,\mu}(\mathcal{F})$ using $\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m r_i f(x_i)$ still involves solving an optimization problem over $f \in \mathcal{F}$, which can be intractable. We will say more about this in the next section.

4 Generalization Error Bounds in Terms of Rademacher Averages

We have the following one-sided uniform bound on the generalization error of functions in $\mathcal{F}$ in terms of the Rademacher averages of the loss function class $\ell_{\mathcal{F}}$:

Theorem 4.1 (Bartlett and Mendelson, 2002; Koltchinkskii and Panchenko, 2002)**.** Let $\mathcal{Y}, \widehat{\mathcal{Y}} \subseteq \mathbb{R}$, and let $\mathcal{F} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$. Let $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to [0, B]$. Let D be any probability distribution on $\mathcal{X} \times \mathcal{Y}$. Let $0 < \delta \le 1$. Then with probability at least $1 - \delta$ (over $S \sim D^m$), all functions $f \in \mathcal{F}$ satisfy

$$\mathrm{er}^\ell_D[f] \;\le\; \mathrm{er}^\ell_S[f] + 2R_{m,D}(\ell_{\mathcal{F}}) + B\sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}\,.$$

Proof. Define $\phi : (\mathcal{X} \times \mathcal{Y})^m \to \mathbb{R}$ as $\phi(S) = \sup_{f\in\mathcal{F}}\Big(\mathrm{er}^\ell_D[f] - \mathrm{er}^\ell_S[f]\Big)$ for all $S \in (\mathcal{X} \times \mathcal{Y})^m$. Then clearly, a change in a single example (x_i, y_i) in S can change $\phi(S)$ by at most B/m. Therefore by McDiarmid's inequality, we have for any $\epsilon > 0$,

$$\mathbf{P}_{S\sim D^m}\left(\sup_{f\in\mathcal{F}}\Big(\mathrm{er}^\ell_D[f] - \mathrm{er}^\ell_S[f]\Big) \;\ge\; \mathbf{E}_{S\sim D^m}\left[\sup_{f\in\mathcal{F}}\Big(\mathrm{er}^\ell_D[f] - \mathrm{er}^\ell_S[f]\Big)\right] + \epsilon\right) \;\le\; e^{-2m\epsilon^2/B^2}\,.$$

This gives with probability at least $1 - \delta$ over $S \sim D^m$,

$$\sup_{f\in\mathcal{F}}\Big(\mathrm{er}^\ell_D[f] - \mathrm{er}^\ell_S[f]\Big) \;\le\; \mathbf{E}_{S\sim D^m}\left[\sup_{f\in\mathcal{F}}\Big(\mathrm{er}^\ell_D[f] - \mathrm{er}^\ell_S[f]\Big)\right] + B\sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}\,.$$

We now bound the expectation using a symmetrization trick to deal with $\mathrm{er}^\ell_D[f]$, similar to what we have seen in previous uniform convergence proofs. We have,

$$\begin{aligned}
& \mathbf{E}_{S\sim D^m}\left[\sup_{f\in\mathcal{F}}\Big(\mathrm{er}^{\ell}_D[f]-\mathrm{er}^{\ell}_S[f]\Big)\right] && (11)\\
&= \mathbf{E}_{S\sim D^m}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^m\Big(\mathbf{E}_{(\tilde{x}_i,\tilde{y}_i)\sim D}\left[\ell_f(\tilde{x}_i,\tilde{y}_i)\right]-\ell_f(x_i,y_i)\Big)\right)\right] && (12)\\
&= \mathbf{E}_{S\sim D^m}\left[\sup_{f\in\mathcal{F}}\left(\mathbf{E}_{\tilde{S}\sim D^m}\left[\frac{1}{m}\sum_{i=1}^m\Big(\ell_f(\tilde{x}_i,\tilde{y}_i)-\ell_f(x_i,y_i)\Big)\right]\right)\right] && (13)\\
&\le \mathbf{E}_{(S,\tilde{S})\sim D^m\times D^m}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m\Big(\ell_f(\tilde{x}_i,\tilde{y}_i)-\ell_f(x_i,y_i)\Big)\right] \quad \text{(by Jensen's inequality)} && (14)\\
&= \mathbf{E}_{(S,\tilde{S})\sim D^m\times D^m,\mathbf{r}\in\{\pm1\}^m}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m r_i\Big(\ell_f(\tilde{x}_i,\tilde{y}_i)-\ell_f(x_i,y_i)\Big)\right] && (15)\\
&\le 2\,\mathbf{E}_{S\sim D^m,\mathbf{r}\in\{\pm1\}^m}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m r_i\ell_f(x_i,y_i)\right] && (16)\\
&= 2R_{m,D}(\ell_{\mathcal{F}})\,. && (17)
\end{aligned}$$

This completes the proof. □

4.1 Binary-Valued Function Classes

In the case of binary-valued functions and zero-one loss, the Rademacher averages of the loss function class are actually equal (upto a constant factor) to the Rademacher averages of the function class itself:

Lemma 4.2. Let $\mathcal{H}\subseteq\{\pm1\}^{\mathcal{X}}$. Let $\ell=\ell_{\text{0-1}}$. Let D be any distribution on $\mathcal{X}\times\{\pm1\}$ and let μ be the marginal of D on $\mathcal{X}$. Then

$$R_{m,D}(\ell_{\mathcal{H}}) \;=\; \frac{1}{2}R_{m,\mu}(\mathcal{H})\,.$$

Proof. We have

$$\begin{aligned}
R_{m,D}(\ell_{\mathcal{H}}) &= \mathbf{E}_{S\sim D^m,\mathbf{r}\in\{\pm\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m r_i\mathbf{1}(h(x_i)\neq y_i)\right] && (18)\\
&= \mathbf{E}_{S\sim D^m,\mathbf{r}\in\{\pm\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m r_i\frac{1-y_ih(x_i)}{2}\right] && (19)\\
&= \mathbf{E}_{S\sim D^m,\mathbf{r}\in\{\pm\}^m}\left[\frac{1}{m}\sum_{i=1}^m\frac{r_i}{2}+\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m\frac{(-r_iy_ih(x_i))}{2}\right] && (20)\\
&= \frac{1}{2}\mathbf{E}_{S\sim D^m,\mathbf{r}\in\{\pm\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m r_ih(x_i)\right] && (21)\\
&= \frac{1}{2}R_{m,\mu}(\mathcal{H})\,, && (22)
\end{aligned}$$

where the second equality makes use of the fact that $y_i,h(x_i)\in\{\pm1\}$, and the fourth equality makes use of the fact that the distribution of $-r_iy_i$ is the same as that of r_i. □

Corollary 4.3. Let $\mathcal{H}\subseteq\{\pm1\}^{\mathcal{X}}$. Let D be any probability distribution on $\mathcal{X}\times\{\pm1\}$, with marginal μ on $\mathcal{X}$. If h_S is selected from $\mathcal{H}$, then for any $0<\delta\le1$, with probability at least $1-\delta$ (over $S\sim D^m$),

$$\mathrm{er}^{\text{0-1}}_D[h_S] \;\le\; \mathrm{er}^{\text{0-1}}_S[h_S]+R_{m,\mu}(\mathcal{H})+\sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}\,. \tag{23}$$

Proof. Follows directly from Theorem 4.1 and Lemma 4.2, observing that $0 \le \ell_{\text{0-1}}(y,\hat{y}) \le 1$ and taking $B = 1$. □

We can also combine this with the concentration results for the empirical estimates of Rademacher averages to get the following results that involve data-dependent complexity measures:

Corollary 4.4. Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$. Let D be any probability distribution on $\mathcal{X} \times \{\pm 1\}$, with marginal μ on $\mathcal{X}$. If h_S is selected from $\mathcal{H}$, then for any $0 < \delta \le 1$, with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\text{er}_D^{\text{0-1}}[h_S] \;\le\; \text{er}_S^{\text{0-1}}[h_S] + R_{x_1^m}(\mathcal{H}) + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}\,. \tag{24}$$

Proof. Follows by combining the results of Corollary 4.3 and Lemma 3.1 (setting each to hold with probability $1-\delta/2$), observing we need only one side of Lemma 3.1 (allows us to remove a factor of 2 from that result). □

Corollary 4.5. Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$. Let D be any probability distribution on $\mathcal{X} \times \{\pm 1\}$, with marginal μ on $\mathcal{X}$. If h_S is selected from $\mathcal{H}$, then for any $0 < \delta \le 1$, with probability at least $1 - \delta$ (over $S \sim D^m, \mathbf{r} \in \{\pm 1\}^m$),

$$\text{er}_D^{\text{0-1}}[h_S] \;\le\; \text{er}_S^{\text{0-1}}[h_S] + \left(\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m r_i h(x_i)\right) + (1+2\sqrt{2})\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}\,. \tag{25}$$

Proof. Follows by combining the results of Corollary 4.3 and Lemma 3.2 (setting each to hold with probability $1-\delta/2$), observing we need only one side of Lemma 3.2 (allows to remove a factor of 2 from that result). □

For binary-valued function classes, we also have the following result which upper bounds the Rademacher averages in terms of the VC-entropy (and therefore the growth function), implying that bounds in terms of these averages can be tighter than those using the VC-entropy/growth function:[4]

Lemma 4.6. Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$. Let μ be any probability distribution on $\mathcal{X}$. Then

$$R_{m,\mu}(\mathcal{H}) \;\le\; \sqrt{\frac{(2\ln 2)\text{VC-entropy}_{\mathcal{H},\mu}(m)}{m}} \;\le\; \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(m)}{m}}\,.$$

4.2 Real-Valued Function Classes

For general real-valued function classes and 'normalized' Lipschitz losses, we have the following relation between the Rademacher averages of the loss function class and those of the function class itself:[5]

Lemma 4.7. Let $\mathcal{Y}, \widehat{\mathcal{Y}} \subseteq \mathbb{R}$ and $\mathcal{F} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$. Let $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to [0,\infty)$ be such that $\ell(y,0) = 0$ for all $y \in \mathcal{Y}$ and $\ell(y,\hat{y})$ is L-Lipschitz in its second argument for some $L > 0$. Let D be any probability distribution on $\mathcal{X} \times \mathcal{Y}$, and let μ be the marginal of D on $\mathcal{X}$. Then

$$R_{m,D}(\ell_{\mathcal{F}}) \;\le\; L\,R_{m,\mu}(\mathcal{F})\,.$$

The loss functions we have studied do not usually satisfy the constraint $\ell(y,0) = 0\ \forall\ y$. However we can obtain a result similar to that of Theorem 4.1 in terms of the loss function class corresponding to a *shifted* loss $\tilde{\ell}(y,\hat{y}) = \ell(y,\hat{y}) - \ell(y,0)$ which does satisfy the constraint:

Theorem 4.8. (Bartlett and Mendelson, 2002) Let $\mathcal{Y}, \widehat{\mathcal{Y}} \subseteq \mathbb{R}$, and let $\mathcal{F} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$. Let $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to [0,B]$. Let D be any probability distribution on $\mathcal{X} \times \mathcal{Y}$. Let $0 < \delta \le 1$. Then with probability at least $1 - \delta$ (over $S \sim D^m$), all functions $f \in \mathcal{F}$ satisfy

$$\text{er}_D^{\ell}[f] \;\le\; \text{er}_S^{\ell}[f] + 2R_{m,D}(\tilde{\ell}_{\mathcal{F}}) + 2B\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}\,,$$

where $\tilde{\ell}(y,\hat{y}) = \ell(y,\hat{y}) - \ell(y,0)$.

[4] Proof via Massart's finite class lemma.

[5] Proof via Ledoux-Talagrand contraction inequality.

Proof. Note that we can write

$$\sup_{f\in\mathcal{F}}\left(\mathrm{er}_D^{\ell}[f]-\mathrm{er}_S^{\ell}[f]\right) \;=\; \sup_{f\in\mathcal{F}}\left(\mathrm{er}_D^{\tilde{\ell}}[f]-\mathrm{er}_S^{\tilde{\ell}}[f]\right)+\left(\mathbf{E}_{(x,y)\sim D}[\ell(y,0)]-\frac{1}{m}\sum_{i=1}^{m}\ell(y_i,0)\right). \tag{26}$$

The first term can be bounded using the same argument as in Theorem 4.1 to get with probability at least $1-\delta/2$ (over $S\sim D^m$):

$$\sup_{f\in\mathcal{F}}\left(\mathrm{er}_D^{\tilde{\ell}}[f]-\mathrm{er}_S^{\tilde{\ell}}[f]\right) \;\le\; 2R_{m,D}(\tilde{\ell}_{\mathcal{F}})+B\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}\,. \tag{27}$$

The second term can be bounded using Hoeffding's (or McDiarmid's) inequality to get with probability at least $1-\delta/2$ (over $S\sim D^m$):

$$\mathbf{E}_{(x,y)\sim D}[\ell(y,0)]-\frac{1}{m}\sum_{i=1}^{m}\ell(y_i,0) \;\le\; B\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}\,. \tag{28}$$

Combining the two bounds gives the desired result. □

Corollary 4.9. Let $\mathcal{Y},\widehat{\mathcal{Y}}\subseteq\mathbb{R}$, and let $\mathcal{F}\subseteq\widehat{\mathcal{Y}}^{\mathcal{X}}$. Let $\ell:\mathcal{Y}\times\widehat{\mathcal{Y}}\to[0,B]$ be such that $\ell(y,\hat{y})$ is L-Lipschitz in its second argument for some $L>0$. Let D be any probability distribution on $\mathcal{X}\times\mathcal{Y}$, with marginal μ on $\mathcal{X}$. If f_S is selected from $\mathcal{F}$, then for any $0<\delta\le 1$, with probability at least $1-\delta$ (over $S\sim D^m$),

$$\mathrm{er}_D^{\ell}[f_S] \;\le\; \mathrm{er}_S^{\ell}[f_S]+2L\,R_{m,\mu}(\mathcal{F})+2B\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}\,.$$

Proof. Follows directly from Theorem 4.8 and Lemma 4.7, observing that if ℓ is L-Lipschitz, then so is $\tilde{\ell}$. □

Again, we can combine this with the concentration results for the empirical estimates of Rademacher averages to get results involving data-dependent complexity measures, for example:

Corollary 4.10. Let $\mathcal{Y}\subseteq\mathbb{R}$, and let $\mathcal{F}\subseteq[a,b]^{\mathcal{X}}$ for some $a\le b$. Let $\ell:\mathcal{Y}\times[a,b]\to[0,B]$ be such that $\ell(y,\hat{y})$ is L-Lipschitz in its second argument for some $L>0$. Let D be any probability distribution on $\mathcal{X}\times\mathcal{Y}$, with marginal μ on $\mathcal{X}$. If f_S is selected from $\mathcal{F}$, then for any $0<\delta\le 1$, with probability at least $1-\delta$ (over $S\sim D^m$),

$$\mathrm{er}_D^{\ell}[f_S] \;\le\; \mathrm{er}_S^{\ell}[f_S]+2L\,R_{x_1^m}(\mathcal{F})+2B\sqrt{\frac{\ln(\frac{4}{\delta})}{2m}}+(b-a)\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}\,. \tag{29}$$

Proof. Follows by combining the results of Corollary 4.9 and Lemma 3.1 (setting each to hold with probability $1-\delta/2$), observing we need only one side of Lemma 3.1 as before. □

4.3 Computation of (Estimates of) Rademacher Averages

As noted above, the computation of the empirical estimates of Rademacher averages still involves solving an optimization problem over the function class. In the case of binary classification, for $\mathcal{H}\subseteq\{\pm1\}^{\mathcal{X}}$, the computation of the estimate $\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}r_ih(x_i)$ (used in Lemma 3.2 and Corollary 4.5) involves minimizing the empirical zero-one loss on a sample consisting of the points x_i labeled with $r_i\in\{\pm1\}$:

$$\begin{aligned}
\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}r_ih(x_i) \;&=\; -\inf_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}-r_ih(x_i)\right) && (30)\\
&=\; -\inf_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1-r_ih(x_i)}{2}\cdot 2-1\right) && (31)\\
&=\; 1-2\inf_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(h(x_i)\neq r_i)\right). && (32)
\end{aligned}$$

For simple function classes $\mathcal{H}$, this may be tractable; however, as we have noted before, for many function classes of interest, this optimization problem is NP-hard. One possibility then is to obtain an approximate solution, e.g. using the same learning algorithm being used for the classification task; this will not normally lead to an exact bound, but could be used to obtain an approximate bound.

In general, if the function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ of interest can be viewed as being appropriately composed of some simpler function classes, one can make use of some structural results/composition theorems that allow one to bound the Rademacher averages of the final class in terms of those for the simpler classes; see for example [1] for such results. As an example of a function class for which the empirical Rademacher averages can be bounded directly, let $\mathcal{X} \subseteq \mathbb{R}^n$, and consider the class of linear functions with bounded-norm weight vectors: $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathbb{R} \mid f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} \text{ for some } \mathbf{w} \in \mathbb{R}^n, \|\mathbf{w}\|_2 \leq W\}$. For this class, we have for any $\mathbf{x}_1^m = (\mathbf{x}_1 \ldots, \mathbf{x}_m) \in \mathcal{X}^m$ and $\mathbf{r} \in \{\pm 1\}^m$,

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} r_i f(\mathbf{x}_i) &= \sup_{\mathbf{w}:\|\mathbf{w}\|_2 \leq W} \left(\frac{1}{m} \sum_{i=1}^{m} r_i \mathbf{x}_i \right) \cdot \mathbf{w} && (33) \\
&\leq W \left\| \frac{1}{m} \sum_{i=1}^{m} r_i \mathbf{x}_i \right\|_2 \quad \text{(by Cauchy-Schwarz inequality)} && (34) \\
&= \frac{W}{m} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} r_i r_j (\mathbf{x}_i \cdot \mathbf{x}_j)}\,. && (35)
\end{aligned}$$

This further gives

$$\begin{aligned}
R_{\mathbf{x}_1^m}(\mathcal{F}) &\leq \frac{W}{m} \mathbf{E}_{\mathbf{r} \in \{\pm 1\}^m} \left[\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} r_i r_j (\mathbf{x}_i \cdot \mathbf{x}_j)} \right] && (36) \\
&\leq \frac{W}{m} \sqrt{\mathbf{E}_{\mathbf{r} \in \{\pm 1\}^m} \left[\sum_{i=1}^{m} \sum_{j=1}^{m} r_i r_j (\mathbf{x}_i \cdot \mathbf{x}_j) \right]} \quad \text{(by Jensen's inequality)} && (37) \\
&= \frac{W}{m} \sqrt{\sum_{i=1}^{m} \|\mathbf{x}_i\|_2^2}\,. && (38)
\end{aligned}$$

The same technique can also be used to bound the empirical Rademacher averages of classes of bounded-norm functions in a reproducing kernel Hilbert space (RKHS), as used by SVMs when learning a kernel-based classifier. We will discuss RKHSs in more detail in the next lecture.

5 Maximum Discrepancy and Gaussian Complexities

One can also define other measures of the complexity or capacity of a function class that can be estimated reliably from data. Two such measures are the *Gaussian averages* (or *Gaussian complexities*), which are defined similarly to the Rademacher averages except that the Rademacher random variables $r_i \in \{\pm 1\}$ are replaced with Gaussian random variables $g_i \sim \mathcal{N}(0, 1)$, and the *maximum discrepancy*, which measures the maximum difference between the average values of a function in the class on two halves of a sample:

$$\begin{aligned}
D_{x_1^m}(\mathcal{F}) &= \sup_{f \in \mathcal{F}} \left(\frac{2}{m} \sum_{i=1}^{m/2} f(x_i) - \frac{2}{m} \sum_{i=m/2+1}^{m} f(x_i) \right) && (39) \\
D_{m,\mu}(\mathcal{F}) &= \mathbf{E}_{x_1^m \sim \mu^m} \left[D_{x_1^m}(\mathcal{F}) \right]\,. && (40)
\end{aligned}$$

Both these measures are closely related to the Rademacher averages and can be used to obtain similar generalization error bounds; see for example [1] for more details.[6]

[6] Note that the Rademacher and Gaussian averages are defined slightly differently in [1] (difference of a factor of 2 and an absolute value in the definitions); the differences are not important in deriving the results we have discussed (the definitions we have used are the ones more commonly used in the literature now).

6 Next Lecture

In the next lecture we will see a different technique for bounding the generalization error of a learned function, which will involve stability properties of the learning algorithm rather than complexity measures of the function class searched by the algorithm; in particular, we will depart from the uniform bounds we have seen so far, which bound the probability of a large deviation uniformly over all functions in the class, and instead will obtain bounds directly for the function learned by the algorithm. We will see that McDiarmid's inequality will be useful in deriving these results as well.

References

[1] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.