Machine Learning Theory (CS 6783)

Lecture 6: Properties of Rademacher Complexity

1 Rademacher Complexity Beyond Cover Style Result

We saw how Rademacher complexity and sequential Rademacher complexity came out naturally for bit prediction and betting games with binary outcomes. It turns out that the quantities are crucial complexity measures more generally for statistical learning and online learning problems. In fact, for the general online supervised learning problem with convex losses, in your assignment 1 you guys already make a connection to the sequential Rademacher complexity. Below to motivate why we need to understand properties of the two Rademacher complexities, I will mention results that make the connections of the two quantities to statistical and online learning. We will formally prove these results in latter lectures. But I will just mention them here.

First recall the statistical learning setting. We get instances $(x_1, y_1), \ldots, (x_n, y_n)$ iid from a fixed but unknown distribution D. In this case, a popular algorithm or method is to find a model amongst set of models $\mathcal{F}$ that minimizes training error. We will use $\hat{L}_S(f)$ to represent training loss w.r.t. to sample S for a model f. ERM returns,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{L}_S(f)$$

We will also use the notation $L_D(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(\hat{f}(x), y)\right]$. It turns out that the performance of ERM compared to the best in class over population loss (expected loss over future draws) is upper bounded by the Classical Rademacher complexity as follows.

$$\mathbb{E}_S\left[L_D(\hat{f}) - \min_{f \in \mathcal{F}} L_D(f)\right] = \mathbb{E}_S\left[L_D(\hat{f}) - \hat{L}_S(\hat{f}) + \hat{L}_S(\hat{f}) - \min_{f \in \mathcal{F}} L_D(f)\right]$$

Since for any model f, $\mathbb{E}_S\left[\hat{L}_S(f)\right] = L_D(f)$,

$$\begin{aligned}
&= \mathbb{E}_S\left[L_D(\hat{f}) - \hat{L}_S(\hat{f}) + \hat{L}_S(\hat{f})\right] - \min_{f \in \mathcal{F}} \mathbb{E}_S\left[\hat{L}_S(f)\right] \\
&\leq \mathbb{E}_S\left[L_D(\hat{f}) - \hat{L}_S(\hat{f}) + \hat{L}_S(\hat{f})\right] - \mathbb{E}_S\left[\min_{f \in \mathcal{F}} \hat{L}_S(f)\right] \\
&= \mathbb{E}_S\left[L_D(\hat{f}) - \hat{L}_S(\hat{f}) + \hat{L}_S(\hat{f})\right] - \mathbb{E}_S\left[\hat{L}_S(\hat{f})\right] \\
&= \mathbb{E}_S\left[L_D(\hat{f}) - \hat{L}_S(\hat{f})\right] \\
&\leq \mathbb{E}_S\left[\sup_{f \in \mathcal{F}} L_D(f) - \hat{L}_S(f)\right]
\end{aligned}$$

Now since $L_D(f) = \mathbb{E}_{S'}\left[\hat{L}_{S'}(f)\right]$ for any $S' = (x'_1, y'_1), \ldots, (x'_n, y'_n)$ drawn iid from distribution D, we have,

$$\begin{aligned}
\mathbb{E}_S\left[L_D(\hat{f}) - \min_{f\in\mathcal{F}} L_D(f)\right] &\leq \mathbb{E}_S\left[\sup_{f\in\mathcal{F}} \mathbb{E}_{S'}\left[\hat{L}_{S'}(f)\right] - \hat{L}_S(f)\right] \\
&\leq \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \hat{L}_{S'}(f) - \hat{L}_S(f)\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\left(\ell(f(x'_t), y'_t) - \ell(f(x_t), y_t)\right)\right]
\end{aligned}$$

Now note that S and S' are drawn iid from same distribution and so for each t, $\ell(f(x'_t), y'_t) - \ell(f(x_t), y_t)$ and $\ell(f(x_t), y_t) - \ell(f(x'_t), y'_t)$ have the same distribution and so for any sequence of signs $\epsilon_1, \ldots, \epsilon_n$,

$$\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\left(\ell(f(x'_t), y'_t) - \ell(f(x_t), y_t)\right)\right] = \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(\ell(f(x'_t), y'_t) - \ell(f(x_t), y_t)\right)\right]$$

Hence taking expectation over ϵ's we conclude that:

$$\begin{aligned}
\mathbb{E}_S\left[L_D(\hat{f}) - \min_{f\in\mathcal{F}} L_D(f)\right] &\leq \mathbb{E}_{S,S'}\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(\ell(f(x'_t), y'_t) - \ell(f(x_t), y_t)\right)\right]\right] \\
&= \mathbb{E}_{S,S'}\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x'_t), y'_t) + \frac{1}{n}\sum_{t=1}^{n} -\epsilon_t\ell(f(x_t), y_t)\right)\right]\right] \\
&\leq \mathbb{E}_{S,S'}\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x'_t), y'_t)\right) + \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{t=1}^{n} -\epsilon_t\ell(f(x_t), y_t)\right)\right]\right] \\
&= \mathbb{E}_{S'}\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x'_t), y'_t)\right]\right] + \mathbb{E}_S\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} -\epsilon_t\ell(f(x_t), y_t)\right]\right] \\
&= 2\mathbb{E}_S\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t), y_t)\right]\right]
\end{aligned}$$

Now to proceed note that,

$$\begin{aligned}
\mathbb{E}_S\left[L_D(\hat{f}) - \min_{f\in\mathcal{F}} L_D(f)\right] &\leq 2\mathbb{E}_S\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t), y_t)\right]\right] \\
&= \mathbb{E}_S\left[\mathbb{E}_\epsilon\left[\sup_{g\in\ell\circ\mathcal{F}_{|(x_1,y_1),\ldots,(x_n,y_n)}} \frac{1}{n}\sum_{t=1}^{n}\epsilon_t g_t\right]\right]
\end{aligned}$$

where the set $\ell\circ\mathcal{F}_{|(x_1,y_1),\ldots,(x_n,y_n)} \subset \mathbb{R}^n$ is simply given by

$$\ell\circ\mathcal{F}_{|(x_1,y_1),\ldots,(x_n,y_n)} = \{(\ell(f(x_1), y_1), \ldots, \ell(f(x_n), y_n)) : f \in \mathcal{F}\}$$

Notice that the right hand side is the style of Rademacher complexity you guys have already encountered where each element of the set is an n dimensional vector.

Similarly, in the online setting where x_t, y_t are adversarially produced, it turns out that one can show bound on regret against any model class $\mathcal{F}$ in terms of the sequential Rademacher Complexity as follows:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\hat{y}_t, y_t) - \min_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t), y_t)\right] &\leq 2\sup_{\mathbf{x},\mathbf{y}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(\mathbf{x}(\epsilon_{1:t-1})), \mathbf{y}(\epsilon_{1:t-1}))\right] \\
&= 2\sup_{\mathbf{x},\mathbf{y}}\mathbb{E}_\epsilon\left[\sup_{g\in\ell\circ\mathcal{F}_{|\mathbf{x},\mathbf{y}}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t g(\epsilon_{1:t-1})\right]
\end{aligned}$$

where given $\mathbf{x}, \mathbf{y}$ that are mapping from binary strings of length 0 up to $n-1$ to $\mathcal{X}$ and $\mathcal{Y}$ respectively, we define $\ell \circ \mathcal{F}_{|\mathbf{x},\mathbf{y}}$ as set of mapping from $\bigcup_{t=1}^{n}\{\pm 1\}^{t-1} \mapsto \mathbb{R}$ given by:

$$\ell \circ \mathcal{F}_{|\mathbf{x},\mathbf{y}} = \{\forall t \in [n], \epsilon_{1:t-1} \mapsto \ell(f(\mathbf{x}(\epsilon_{1:t-1})), \mathbf{y}(\epsilon_{1:t-1})) : f \in \mathcal{F}\}$$

This is true for very general loss functions and $\mathcal{X}$ and $\mathcal{Y}$. Notice that the RHS above is the style of sequential Rademacher complexity you encountered in the linear betting game.

2 Properties of Rademacher Complexity

Recall that the Sequential Rademacher complexity is defined as:

$$\mathcal{R}_n(\mathcal{F}) := \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(\epsilon_{1:t-1})\right]$$

Proposition 1. *For any classes $\mathcal{G}, \mathcal{H}$:*

1. *If $\mathcal{H} \subset \mathcal{G}$, then $\mathcal{R}_n(\mathcal{H}) \leq \mathcal{R}_n(\mathcal{G})$*
2. *For any fixed function h, $\mathcal{R}_n(\mathcal{G}+h) = \mathcal{R}_n(\mathcal{G})$*
3. *$\mathcal{R}_n(\text{cvx}(\mathcal{G})) = \mathcal{R}_n(\mathcal{G})$*

Proof.

1. $\mathcal{R}_n(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{H}}\sum_{t=1}^{n}\epsilon_t g(\epsilon_{1:t-1})\right] \leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t g(\epsilon_{1:t-1})\right] = \mathcal{R}_n(\mathcal{G})$

2. $$\begin{aligned}\mathcal{R}_n(\mathcal{G}+h) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t(g(\epsilon_{1:t-1})+h(\epsilon_{1:t-1}))\right]\\ &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n}\epsilon_t g(\epsilon_{1:t-1})\right\}+\sum_{t=1}^{n}\epsilon_t h(\epsilon_{1:t-1}))\right]\\ &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t g(\epsilon_{1:t-1})\right]+0 = \mathcal{R}_n(\mathcal{G})\end{aligned}$$

3. $\text{cvx}(\mathcal{G}) = \{\mathbb{E}_{g\sim\pi}[g(\cdot)] : \pi \in \Delta(\mathcal{G})\}$

$$\begin{aligned}\mathcal{R}_n(\text{cvx}(\mathcal{G})) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\pi\in\Delta(\mathcal{G})}\sum_{t=1}^{n}\epsilon_t\mathbb{E}_{g\in\pi}[g](\epsilon_{1:t-1})\right]\\ &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\pi\in\Delta(\mathcal{G})}\sum_{t=1}^{n}\epsilon_t\mathbb{E}_{g\in\pi}[g(\epsilon_{1:t-1})]\right]\\ &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\pi\in\Delta(\mathcal{G})}\mathbb{E}_{g\in\pi}\left[\sum_{t=1}^{n}\epsilon_t g(\epsilon_{1:t-1})\right]\right]\end{aligned}$$

Max is attained at vertex of simplex (max element amongst support of distribution)

$$= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t g(z_t)\right] = \mathcal{R}_n(\mathcal{G})$$

□

Next we will prove a very important lemma called the contraction lemma. We will prove it specifically for classical Rademacher complexity. While the statement is also true for the sequential version, we will only prove it for the classical version.

Lemma 2. *For any $\phi_1, \ldots, \phi_n$ where each $\phi_i : \mathbb{R} \mapsto \mathbb{R}$ is L-Lipschitz, and any $\mathcal{F} \subset \mathbb{R}^n$, we have,*

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t\phi_t\left(g_t\right)\right] \leq \frac{L}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t g_t\right]$$

Proof.

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\epsilon_{1:n}}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t\phi_t\left(g_t\right)\right] \\
&= \mathbb{E}_{\epsilon_{1:n-1}}\frac{\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g_t\right)+\phi_n(g_n)\right\}+\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g_t\right)-\phi_n(g_n)\right\}}{2} \\
&= \mathbb{E}_{\epsilon_{1:n-1}}\left[\frac{\sup_{g,g'\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\left(\phi_t\left(g_t\right)+\phi_t\left(g'_t\right)\right)+\phi_n(g_n)-\phi_n(g'_n)\right\}}{2}\right] \\
&\leq \mathbb{E}_{\epsilon_{1:n-1}}\left[\frac{\sup_{g,g'\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\left(\phi_t\left(g_t\right)+\phi_t\left(g'_t\right)\right)+L|g_n-g'_n|\right\}}{2}\right] \\
&= \mathbb{E}_{\epsilon_{1:n-1}}\left[\frac{\sup_{g,g'\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\left(\phi_t\left(g_t\right)+\phi_t\left(g'_t\right)\right)+L(g_n-g'_n)\right\}}{2}\right] \\
&= \mathbb{E}_{\epsilon_{1:n-1}}\frac{\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g_t\right)+Lg_n\right\}+\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g_t\right)-Lg_n\right\}}{2} \\
&= \frac{1}{n}\mathbb{E}_{\epsilon_{1:n}}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g_t\right)+L\epsilon_n g_n\right]
\end{aligned}$$

Repeating the above argument we remove $\phi_1, \ldots, \phi_{n-1}$ and so, we conclude that

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t\phi_t\left(g_t\right)\right] \leq \frac{L}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t g_t\right]$$

□

To see an example of the above contraction lemma application, think of the statistical learning setting where given $(x_1, y_1), \ldots, (x_n, y_n)$ we want to bound the Rademacher complexity of set $\ell \circ \mathcal{F}_{|(x_1,y_1),\ldots,(x_n,y_n)}$. To this end, we cant think of $\phi_t(a) = \ell(a, y_t)$ and we can think of each $f_t = f(x_t)$. Thus the above contraction lemma tells us that:

$$\mathcal{R}_n(\ell \circ \mathcal{F}_{|(x_1,y_1),\ldots,(x_n,y_n)}) \leq L\mathcal{R}_n(\mathcal{F}_{|x_1,\ldots,x_n})$$

where $\mathcal{F}_{|x_1,\ldots,x_n} = \{(f(x_1), \ldots, f(x_n))) : f \in \mathcal{F}\}$. In other words, if we have a Lipschitz loss, Rademacher complexity with loss composed, can be upper bounded by Rademacher complexity of just the class of models.