Mathematical Foundations of Machine Learning(CS 4783/5783)

Lecture 6: Properties of Rademacher Complexity, and Examples

1 Recap

1. For any $\delta > 0$, with probability at least $1-\delta$,

$$L_{\mathbf{D}}(\hat{f}_{\mathrm{ERM}}) - \min_{f\in\mathcal{F}} L_{\mathbf{D}}(f) \le 2\mathbb{E}_S\left[\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right|\right]\right] + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

2. The term $\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right|\right]$ is referred to as Rademacher complexity on a sample S. Further,

$$\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right|\right] = \mathbb{E}_\epsilon\left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}} \frac{1}{n}\left|\sum_{t=1}^n \epsilon_t \ell(\mathbf{f}[t], y_t)\right|\right] \le O\left(\sqrt{\frac{\log|\mathcal{F}_{|x_1,\ldots,x_n}|}{n}}\right)$$

2 Properties of Rademacher Complexity

Define empirical Rademacher complexity of a class $\mathcal{G}$, a set of functions on $\mathcal{Z}$, on a sample $S = \{z_1, \ldots, z_n\}$ as

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

In class we showed that $L_D(\hat{y}_{erm}) - \inf_{f\in\mathcal{F}} L_D(f) \le 2\ \mathbb{E}_S\left[\hat{\mathcal{R}}_S(\ell\circ\mathcal{F})\right] + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$, where $\ell\circ\mathcal{F} = \{(x,y)\mapsto \ell(f(x),y) : f\in\mathcal{F}\}$

We start with the following lemma called contraction lemma that is one of the most important property of the Rademacher complexity. It basically tells us that if we consider Rademacher complexity of a class functions got by taking a sequence of Lipschitz functions composed with any class of functions. This Rademacher complexity can be upper bounded by the Radmeacher complexity of the function class. That is the Lipschitz function can be removed. Before we begin, let us recall, a function $\phi : \mathbb{R} \mapsto \mathbb{R}$ is said to be an L-Lipschitz function if for any $a, b \in \mathbb{R}$,

$$|\phi(a) - \phi(b)| \le L|a-b|$$

L is called the Lipschitz constant. The property basically says that as points get close by, the function value at these points are also close.

Lemma 1. *For any $\phi_1, \ldots, \phi_n$ where each $\phi_i : \mathbb{R}\mapsto\mathbb{R}$ and is L-Lipschitz, and any $z_1, \ldots, z_n$, we have,*

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

Remark: Give $(x_1, y_1), \ldots, (x_n, y_n)$, let us define $\phi_t(a) = \ell(a, y_t)$. Now if the loss function is L Lipschitz in its first argument, then it is clear that ϕ_t's are Lipschitz and hence by the above contraction lemma, we can remove the loss and only have Rademacher complexity of $\mathcal{F}$. That is $\hat{\mathcal{R}}_S(\ell \circ \mathcal{F}) \leq L\ \hat{\mathcal{R}}_S(\mathcal{F})$

Proposition 2. *For any sample $S = \{z_1, \ldots, z_n\}$ and any classes $\mathcal{G}, \mathcal{H} \subset \mathbb{R}^{\mathcal{Z}}$:*

1. *If $\mathcal{H} \subset \mathcal{G}$, then $\hat{\mathcal{R}}_S(\mathcal{H}) \leq \hat{\mathcal{R}}_S(\mathcal{G})$*
2. *For any fixed function $h : \mathcal{Z} \mapsto \mathbb{R}$, $\hat{\mathcal{R}}_S(\mathcal{G} + h) = \hat{\mathcal{R}}_S(\mathcal{G})$*
3. *$\hat{\mathcal{R}}_S(\mathrm{cvx}(\mathcal{G})) = \hat{\mathcal{R}}_S(\mathcal{G})$*

Proof.

1. $\hat{\mathcal{R}}_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{g \in \mathcal{H}} \left|\sum_{t=1}^n \epsilon_t g(z_t)\right|\right] \leq \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{g \in \mathcal{G}} \left|\sum_{t=1}^n \epsilon_t g(z_t)\right|\right] \leq \hat{\mathcal{R}}_S(\mathcal{G})$.

2. For any fixed function h bounded by 1,

$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{G} + h) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{g \in \mathcal{G}}\left|\sum_{t=1}^n \epsilon_t(g(z_t) + h(z_t))\right|\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{g \in \mathcal{G}}\left|\sum_{t=1}^n \epsilon_t g(z_t)\right| + \left|\sum_{t=1}^n \epsilon_t h(z_t))\right|\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{g \in \mathcal{G}}\left|\sum_{t=1}^n \epsilon_t g(z_t)\right|\right] + \frac{1}{n}\mathbb{E}_\epsilon\left[\left|\sum_{t=1}^n \epsilon_t h(z_t))\right|\right] \\
&\leq \hat{\mathcal{R}}_S(\mathcal{G}) + O\left(\sqrt{\frac{1}{n}}\right)
\end{aligned}$$

3. $\mathrm{cvx}(\mathcal{G}) = \{\mathbf{z} \mapsto \mathbb{E}_{g \sim \pi}[g(z)] : \pi \in \Delta(\mathcal{G})\}$. That is, instead of only considering functions in $\mathcal{G}$ we are allowed to also pick any distribution over $\mathcal{G}$ and consider the expected function under the distribution.

$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathrm{cvx}(\mathcal{G})) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{\pi \in \Delta(\mathcal{G})}\left|\sum_{t=1}^n \epsilon_t \mathbb{E}_{g \in \pi}[g(z_t)]\right|\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{\pi \in \Delta(\mathcal{G})}\left|\mathbb{E}_{g \in \pi}\left[\sum_{t=1}^n \epsilon_t g(z_t)\right]\right|\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{\pi \in \Delta(\mathcal{G})} \mathbb{E}_{g \in \pi}\left[\left|\sum_{t=1}^n \epsilon_t g(z_t)\right|\right]\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{g \in \mathcal{G}} \sum_{t=1}^n \epsilon_t g(z_t)\right] = \hat{\mathcal{R}}_S(\mathcal{G})
\end{aligned}$$

However, we also have that $\mathcal{G} \subseteq \mathrm{cvx}(\mathcal{G})$ and so from earlier shown property, $\hat{\mathcal{R}}_S(\mathcal{G}) \leq \hat{\mathcal{R}}_S(\mathrm{cvx}(\mathcal{G}))$ and so overall we have shown that

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \hat{\mathcal{R}}_S(\mathrm{cvx}(\mathcal{G}))$$

$\square$

3 Example : Rademacher complexity of linear function classes

1. L1 regularizer : Let $\mathcal{F}_R = \{x \mapsto f^\top x : f \in \mathbb{R}^d, \|f\|_1 \leq R\}$, where $\|f\|_1 = \sum_{i=1}^d |f[i]|$. In this case we have

$$\begin{aligned}\hat{\mathcal{R}}_S(\mathcal{F}_R) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{f:\|f\|_1\leq R}\left|\sum_{t=1}^n \epsilon_t f^\top x_t\right|\right] \\ &= \frac{R}{n}\mathbb{E}_\epsilon\left[\max_{f:\|f\|_1\leq 1}\left|\sum_{t=1}^n \epsilon_t f^\top x_t\right|\right] \\ &= R\,\hat{\mathcal{R}}_S(\mathcal{F}_1)\end{aligned}$$

Consider the class $\mathcal{G} = \{x \mapsto g^\top x : g \in \{e_1, -e_1, e_2, -e_2, \ldots, d_d, -e_d\}$ whose cardinality is clearly $2d$. That is, the 2d functions where each one returns one chosen coordinate of input vector x along with a chosen sign. Now we first claim that $\mathcal{F}_1 = \mathrm{cvx}(\mathcal{G})$. Why is this?

Hence by Proposition 2 property (3) we have that

$$\begin{aligned}\hat{\mathcal{R}}_S(\mathcal{F}_R) &= R\,\hat{\mathcal{R}}_S(\mathcal{G}) \\ &\leq O\left(R\max_{x\in\mathcal{X}}\|x\|_\infty\sqrt{\frac{\log(2d)}{n}}\right)\end{aligned}$$

2. ℓ_2 regularizer : Let $\mathcal{F} = \{x \mapsto \langle f, x\rangle : \|f\|_2 \leq R\}$. For this case we have that,

$$\begin{aligned}\hat{\mathcal{R}}_S(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{f:\|f\|_2\leq R}\left|\sum_{t=1}^n \epsilon_t f^\top x_t\right|\right] \\ &= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{f:\|f\|_2\leq R}\left|f^\top\left(\sum_{t=1}^n \epsilon_t x_t\right)\right|\right] \\ &= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{f:\|f\|_2\leq R}\|f\|_2\left|\frac{f}{\|f\|_2}^\top\left(\sum_{t=1}^n \epsilon_t x_t\right)\right|\right] \\ &= \frac{R}{n}\mathbb{E}_\epsilon\left[\left\|\sum_{t=1}^n \epsilon_t x_t\right\|_2\right] \\ &= \frac{R}{n}\mathbb{E}_\epsilon\left[\sqrt{\left\|\sum_{t=1}^n \epsilon_t x_t\right\|_2^2}\right]\end{aligned}$$

$$
\begin{aligned}
&\le \frac{R}{n}\sqrt{\mathbb{E}_\epsilon\left[\left\|\sum_{t=1}^n \epsilon_t x_t\right\|_2^2\right]} \\
&= \frac{R}{n}\sqrt{\mathbb{E}_\epsilon\left[\sum_{t=1}^n \|x_t\|_2^2 + 2\sum_{t=1}^n\sum_{s>t}\epsilon_t\epsilon_s x_t^\top x_s\right]} \\
&= \frac{R}{n}\sqrt{\sum_{t=1}^n \|x_t\|_2^2} \le \frac{R\max_{x\in\mathcal{X}}\|x\|_2}{\sqrt{n}}
\end{aligned}
$$

4 Applications

Example applications : Lasso, SVM, ridge regression, Logistic Regression (including kernel methods), ℓ_1 neural networks, matrix completion (max norm, trace norm), graph prediction

Observation : Hinge loss given by $\ell(y', y) = \max\{1 - y'y, 0\}$ is 1-Lipschitz. Logistic loss given by $\ell(y', y) = \log(1 + e^{-y'y})$ is 1-Lipchitz. Squared loss $\ell(y', y) = (y' - y)^2$ is $4B$ Lipschtiz when $|y|, |y'| \le B$. Absolute loss $\ell(y', y) = |y - y'|$ is 1-Lipchitz. In all these cases, using contraction lemma we can remove the loss function and using the bound for ERM conclude that with probability $1 - \delta$,

$$
L_D(\hat{f}_{\mathrm{ERM}}) - \inf_{f\in\mathcal{F}} L_D(f) \le 2L\ \mathbb{E}_S\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t f(x_t)\right] + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)
$$

where L is the corresponding Lipchitz constant of the loss.

1. SVM :

$$
\begin{aligned}
&\text{minimize} \sum_{t=1}^n \max\{1 - \langle f, x_t\rangle \cdot y_t, 0\} \\
&\text{subject to } \|f\|_2 \le R
\end{aligned}
$$

   This corresponds to class F given by linear predictors with Hilbert norm constrained by R

2. Lasso :

$$
\begin{aligned}
&\text{minimize} \sum_{t=1}^n (y - \langle f, x_t\rangle)^2 \\
&\text{subject to } \|f\|_1 \le R
\end{aligned}
$$

   Corresponds to linear predictor with ℓ_1 norm constrained by 1

3. ℓ_1 neural network with K layers. Loss could be squared loss or logistic loss. Let $\mathcal{F}_1$ be some arbitrary base class of predictors. Recursively define the subsequent i layer neural network predictor as follows

$$\mathcal{F}_i = \{x \mapsto \sum_j w_j^i \sigma(f_j(x)) : \forall j, \quad f_j \in \mathcal{F}_{i-1}, \quad \left\|w^i\right\|_1 \leq B_i\}$$

where σ is a 1-Lipchitz loss function. Then

$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{F}_i) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{\substack{\|w^i\|_1 \leq B_i \\ \forall j,\ f_j \in \mathcal{F}_{i-1}}} \sum_{t=1}^n \sum_j \epsilon_t w_j^i \sigma(f_j(x_t))\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{\substack{\|w^i\|_1 \leq B_i \\ \forall j,\ f_j \in \mathcal{F}_{i-1}}} \left\|w^i\right\|_1 \max_j \left|\sum_{t=1}^n \epsilon_t \sigma(f_j(x_t))\right|\right] \\
&\leq \frac{B_i}{n}\mathbb{E}_\epsilon\left[\max_{\forall j,\ f_j \in \mathcal{F}_{i-1}} \max_j \left|\sum_{t=1}^n \epsilon_t \sigma(f_j(x_t))\right|\right] \\
&= \frac{B_i}{n}\mathbb{E}_\epsilon\left[\max_{f \in \mathcal{F}_{i-1}} \left|\sum_{t=1}^n \epsilon_t \sigma(f_j(x_t))\right|\right] \\
&\leq \frac{2B_i}{n}\mathbb{E}_\epsilon\left[\max_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n \epsilon_t \sigma(f_j(x_t))\right] \\
&= 2B_i \hat{\mathcal{R}}_S(\sigma \circ \mathcal{F}_{i-1}) \\
&\leq 2B_i \hat{\mathcal{R}}_S(\mathcal{F}_{i-1})
\end{aligned}$$

Hence we can conclude that

$$\hat{\mathcal{R}}_S(\mathcal{F}_i) \leq \left(\prod_{i=1}^k 2B_i\right) \hat{\mathcal{R}}_S(\mathcal{F}_1)$$