Machine Learning Theory (CS 6783)

Lecture 10 : Lower Bound and Optimality

1 Recap

1. For any statistical learning problem we have,

$$\mathbb{E}_S\left[L_D(\hat{f}_{\text{erm}}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \leq \frac{2}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right]$$

2. For any L-Lipchitz loss

$$\frac{1}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right] \leq \frac{L}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t)\right] = \mathbb{E}_S\left[\mathcal{R}_n(\mathcal{F}_{|x_1,\ldots,x_n})\right]$$

3. For online learning, there exists a learning algorithm such that:

$$\mathbb{E}\left[\text{Reg}_n\right] \leq 2\sup_{\mathbf{x},\mathbf{y}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(\mathbf{x}(\epsilon_{1:t-1})),\mathbf{y}(\epsilon_{1:t-1}))\right] = 2\sup_{\mathbf{x},\mathbf{y}}\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}_{|\mathbf{x},\mathbf{y}})$$

   where $\ell\circ\mathcal{F}_{|\mathbf{x},\mathbf{y}} = \{g_f : \forall t, \epsilon\in\{\pm1\}^n, g_f(\epsilon_{1:t-1}) = \ell(f(\mathbf{x}(\epsilon_{1:t-1})),\mathbf{y}(\epsilon_{1:t-1})), f\in\mathcal{F}\}$

4. Covering : V is an ℓ_p-cover of $\mathcal{F}$ on $x_1,\ldots,x_n$ at scale β if

$$\forall f\in\mathcal{F}, \epsilon\in\{\pm1\}^n, \exists\mathbf{v}\in V \text{ s.t. } \left(\frac{1}{n}\sum_{t=1}^{n}|f(\epsilon_{1:t-1}) - \mathbf{v}(\epsilon_{1:t-1})|^p\right)^{1/p} \leq \beta$$

   $\mathcal{N}_p(\mathcal{F},\beta) = \min\{|V| : V \text{ is an } \ell_p\text{-cover of } \mathcal{F} \text{ at scale } \beta\}$

5. Pollard bound:

$$\mathcal{R}_n^{sq}(\mathcal{F}) \leq \inf_{\beta>0}\left\{\beta + \sqrt{\frac{\log\mathcal{N}_1(\mathcal{F},\beta)}{n}}\right\}$$

6. Dudley Integral bound:

$$\mathcal{R}_n^{sq}(\mathcal{F}) \leq \mathcal{D}_n(\mathcal{F}) := \inf_{\alpha>0}\left\{4\alpha + 10\int_\alpha^1\sqrt{\frac{\log\mathcal{N}_2(\mathcal{F},\beta)}{n}}d\beta\right\}$$

2 Sudakov's Theorem and Partial Converse

In this section, we show that Dudley integral complexity and Rademacher complexity are within log factors of each other. We will show this for the classical or statistical learning versions. A similar statement is true for the sequential counter part but we will not prove this.

To prove the classical version we first start with a result called Sudakov Minoration which we will state below without proof.

Theorem 1. *Let $\mathcal{F} \subseteq \mathbb{R}^n$. There is a universal constant $c > 0$ such that*

$$\mathcal{R}_n(\mathcal{F}) \geq \frac{c}{\log n} \sup_{\alpha > 0} \alpha \sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \alpha)}{n}}$$

The above theorem (paraphrased) is due to Sudakov. We shall not go over its proof. But using the above, we shall prove that Dudley integral bound is a tight bound on Rademacher complexity.

Theorem 2.

$$\frac{c}{10 \log^2 n} \left(\mathcal{D}_n(\mathcal{F}) - \frac{4}{n} \right) \leq \mathcal{R}_n(\mathcal{F}) \leq \mathcal{D}_n(\mathcal{F})$$

Proof. We already showed that $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{D}_n(\mathcal{F})$. Now on the other hand, we have

$$\mathcal{D}_n(\mathcal{F}) = \inf_{\alpha > 0} \left\{ 4\alpha + \frac{10}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \left(\mathcal{N}_2(\mathcal{F}, \delta) \right)} d\delta \right\}$$

However by Sudakov's theorem we have that for any $\delta > 0$, we have

$$\sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \delta)}{n}} \leq \frac{\log(n)\ \mathcal{R}_n(\mathcal{F})}{c\ \delta}$$

Using this,

$$\begin{aligned} \mathcal{D}_n(\mathcal{F}) &\leq \inf_{\alpha > 0} \left\{ 4\alpha + \frac{10}{c} \log(n)\ \mathcal{R}_n(\mathcal{F}) \int_{\alpha}^{1} \frac{1}{\delta} d\delta \right\} \\ &= \inf_{\alpha > 0} \left\{ 4\alpha + \frac{10}{c} \log(n)\ \log(1/\alpha)\ \mathcal{R}_n(\mathcal{F}) \right\} \end{aligned}$$

Picking $\alpha = \frac{1}{n}$ we conclude that $\mathcal{D}_n(\mathcal{F}) \leq \frac{4}{n} + \frac{10}{c} \log^2(n)\ \mathcal{R}_n(\mathcal{F})$. Rewriting we get that

$$\frac{c}{10 \log^2 n} \left(\mathcal{D}_n(\mathcal{F}) - \frac{4}{n} \right) \leq \mathcal{R}_n(\mathcal{F})$$

□

3 Lower Bound for Online Supervised Learning With Absolute Loss

In this section we show that if one considers online supervised learning where the loss is $\ell(y', y) = |y - y'|$, then one can obtain a lower bound for the optimal rate in terms of sequential Rademacher Complexity of model class $\mathcal{F}$.

Lemma 3. *For any class $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, there is an adversary strategy that ensures that irrespective of what (possibly randomized) algorithm the learner uses, the expected regret of the learner is lower bounded as:*

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

Proof. Pick any mapping $\mathbf{x}$. Now let the adversary play at round t, the input instance $\mathbf{x}(y_{1:t-1})$ and choose label $y_t = \epsilon_t$ drawn as Rademacher random variable. In this case, note that expected regret against this adversary for any learning algorithm is given by

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\mathbb{E}\left[\mathrm{Reg}_n\right]\right] &= \mathbb{E}_\epsilon\left[\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}|\epsilon_t - \hat{y}_t| - \min_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}|f(\mathbf{x}(\epsilon_{1:t-1})) - \epsilon_t|\right]\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_\epsilon\left[|\epsilon_t - \hat{y}_t|\right] - \mathbb{E}_\epsilon\left[\min_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}|f(\mathbf{x}(\epsilon_{1:t-1})) - \epsilon_t|\right]\right]
\end{aligned}$$

However since $\hat{y}_t$ is trying to predict a random coin flip, $\mathbb{E}_\epsilon\left[|\epsilon_t - \hat{y}_t|\right] = 1$ and so,

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\mathbb{E}\left[\mathrm{Reg}_n\right]\right] &= 1 - \mathbb{E}_\epsilon\left[\min_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}|f(\mathbf{x}(\epsilon_{1:t-1})) - \epsilon_t|\right] \\
&= \mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\left\{1 - \frac{1}{n}\sum_{t=1}^{n}|f(\mathbf{x}(\epsilon_{1:t-1})) - \epsilon_t|\right\}\right] \\
&= \mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\left(1 - |f(\mathbf{x}(\epsilon_{1:t-1})) - \epsilon_t|\right)\right]
\end{aligned}$$

However for any $y \in \{\pm 1\}$ and any $a \in [-1,1]$, $|a - y| = 1 - a\cdot y$ and so,

$$\begin{aligned}
&= \mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\left(1 - (1 - \epsilon_t f(\mathbf{x}(\epsilon_{1:t-1})))\right)\right] \\
&= \mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}(\epsilon_{1:t-1}))\right]
\end{aligned}$$

Since the choice of $\mathbf{x}$ was arbitrary, we can take a supremum over all such mappings. □

4 Lower Bounds For Supervised Learning in Statistical Setting

Basic idea : To show lower bound, we pick $k \cdot n$ points $x_1, \ldots, x_{kn}$ and signs $\epsilon_1, \ldots, \epsilon_{kn}$. The signs are not revealed to the learner. We use the uniform distribution over the kn pairs of instances as the distribution D. That is $D = \mathrm{Unif}\{(x_1, \epsilon_1), \ldots, (x_{kn}, \epsilon_{kn})$. Learner can even know this fact, only learner does not get to see the ϵ_t's before hand. Now we sample n points from this distribution and provide this to the learner. Clearly the learner sees at most n labels and so on the the remaining $kn - n$ points learner has no way to predict anything meaningful. The rest is simply massaging the math.

We shall consider the absolute loss $\ell(y', y) = |y - y'|$. However similar analysis can be extended to other commonly used supervised learning losses (called margin losses) like all ℓ_p losses, logistic loss, hinge loss etc.

Lemma 4. *For any class $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$ and for any $k \in \mathbb{N}$, when we consider statistical learning with absolute loss, if one considers all proper learning algorithms, (where $\hat{y}$ returned by the algorithm is within model class $\mathcal{F}$), then we have the following lower bound.*

$$\inf_{\hat{y} \in \text{Proper}} \sup_{D} \mathbb{E}_S \left[L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k} \mathcal{R}_n(\mathcal{F})$$

Similarly, if one considers all algorithms including improper ones, we get the bound:

$$\inf_{\hat{y} \in \text{All}} \sup_{D} \mathbb{E}_S \left[L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k}$$

Proof.

$$\begin{aligned}
&\inf_{\hat{y}} \sup_{D} \mathbb{E}_S \left[L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \\
&\quad \geq \inf_{\hat{y}} \sup_{x_1, \ldots, x_{kn}} \mathop{\mathbb{E}}_{\epsilon_1, \ldots, \epsilon_{kn}} \mathbb{E}_{S \sim \text{Unif}\{(x_1, \epsilon_1), \ldots, (x_{kn}, \epsilon_{kn})\}} \left[\frac{1}{kn} \sum_{t=1}^{kn} |\hat{y}_S(x_t) - \epsilon_t| - \inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} |f(x_t) - \epsilon_t| \right] \\
&\quad \geq \sup_{x_1, \ldots, x_{kn}} \inf_{\hat{y}} \mathop{\mathbb{E}}_{\epsilon_1, \ldots, \epsilon_{kn}} \mathbb{E}_{S \sim \text{Unif}\{(x_1, \epsilon_1), \ldots, (x_{kn}, \epsilon_{kn})\}} \left[\frac{1}{kn} \sum_{t=1}^{kn} |\hat{y}_S(x_t) - \epsilon_t| - \inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} |f(x_t) - \epsilon_t| \right]
\end{aligned}$$

For any $y' \in [-1,1]$, $|y' - \epsilon_t| = 1 - y'\epsilon_t$ and so,

$$\begin{aligned}
&= \sup_{x_1, \ldots, x_{kn}} \inf_{\hat{y}} \mathop{\mathbb{E}}_{\epsilon_1, \ldots, \epsilon_{kn}} \mathbb{E}_{S \sim \text{Unif}\{(x_1, \epsilon_1), \ldots, (x_{kn}, \epsilon_{kn})\}} \left[\frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t \hat{y}_S(x_t) - \inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t f(x_t) \right] \\
&= \sup_{x_1, \ldots, x_{kn}} \left\{ \inf_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t \hat{y}_S(x_t) \right] - \mathbb{E}_\epsilon \left[\inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t f(x_t) \right] \right\} \\
&= \sup_{x_1, \ldots, x_{kn}} \left\{ \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t \hat{y}_S(x_t) \right] \right\}
\end{aligned}$$

Now define $J \subset [2n]$ as, $J_S = \{i : (x_i, \epsilon_i) \in S\}$. Notice that for any $i \in J_S^c$, ,because $\hat{y}_S$ is only a function of sample S, we have $\mathbb{E}_S \left[\mathbb{E}_{\epsilon_i} \left[\epsilon_i \hat{y}_S(x_i) \right] \right] = \mathbb{E}_S \left[\mathbb{E}_{\epsilon_i} \left[\epsilon_i \right] \hat{y}_S(x_i) \right] = 0$. Hence :

$$\begin{aligned}
&\inf_{\hat{y}} \sup_{D} \mathbb{E}_S \left[L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \\
&\quad \geq \sup_{x_1, \ldots, x_{kn}} \left\{ \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \frac{1}{kn} \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\sum_{t \in J} \epsilon_t \hat{y}_S(x_t) \right] \right\} \\
&\quad \geq \sup_{x_1, \ldots, x_{kn}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \frac{1}{kn} \sup_{x_1, \ldots, x_{kn}} \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\sum_{t \in J} \epsilon_t \hat{y}_S(x_t) \right] \\
&\quad = \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1, \ldots, x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right]
\end{aligned}$$

Now if we consider minimax rates with respect to only *proper learning algorithms*, that is $\hat{y}_S \in \mathcal{F}$, then

$$\begin{aligned}
&\inf_{\hat{y}\in\text{Proper}} \sup_D \mathbb{E}_S \left[L_D(\hat{y}) - \inf_{f\in\mathcal{F}} L_D(f) \right] \\
&\qquad \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1,\ldots,x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right] \\
&\qquad \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1,\ldots,x_n} \mathbb{E}_\epsilon \left[\sup_{\hat{y}\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right] \\
&\qquad = \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k} \mathcal{R}_n(\mathcal{F})
\end{aligned}$$

On the other hand if we consider *improper learning algorithms* as well, then

$$\begin{aligned}
&\inf_{\hat{y}\in\text{All}} \sup_D \mathbb{E}_S \left[L_D(\hat{y}) - \inf_{f\in\mathcal{F}} L_D(f) \right] \\
&\qquad \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1,\ldots,x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right] \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k}
\end{aligned}$$

□