Machine Learning Theory (CS 6783)

Lecture 12: Statistical Learning, Lower Bounds and Uniform Convergence

1 Recap

1. For any statistical learning problem we have,

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

2. For any L-Lipchitz loss

$$\frac{1}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right] \le \frac{L}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t)\right]$$
$$\mathbb{E}_S\left[\hat{\mathcal{R}}_S(\ell\circ\mathcal{F})\right] \le L\ \mathbb{E}_S\left[\hat{\mathcal{R}}_S(\mathcal{F})\right]$$

3. Dudley Integral bound:

$$\hat{\mathcal{R}}_S(\mathcal{F}) \le \hat{D}_S(\mathcal{F}) := \inf_{\alpha>0}\left\{4\alpha + 12\int_\alpha^1\sqrt{\frac{\log\mathcal{N}_2(\mathcal{F},\beta;x_1,\ldots,x_n)}{n}}d\beta\right\}$$

4. We also have that

$$\frac{c}{12\log^2 n}\left(\mathcal{D}_S(\mathcal{F}) - \frac{4}{n}\right) \le \hat{\mathcal{R}}_S(\mathcal{F}) \le \mathcal{D}_S(\mathcal{F})$$

2 Lower Bounds on Supervised Learning for $\mathcal{Y}\subset\mathbb{R}$

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 $D = \text{Unif}\{(x_1,\epsilon_1),\ldots,(x_{kn},\epsilon_{kn})$. 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 1. *For any class $\mathcal{F}\subset[-1,1]^{\mathcal{X}}$ and for any $k\in\mathbb{N}$,*

$$\mathcal{V}_n^{\text{proper}}(\mathcal{F}) \ge \mathcal{R}_{kn} - \frac{1}{k}\mathcal{R}_n(\mathcal{F}) \qquad \textit{and} \qquad \mathcal{V}_n^{\text{improper}}(\mathcal{F}) \ge \mathcal{R}_{kn} - \frac{1}{k}$$

Proof.

$$\inf_{\hat{y}} \sup_{D} \mathbb{E}_S \left[L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \right]$$

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

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

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

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

$$= \sup_{x_1,\ldots,x_{kn}} \left\{ \inf_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t \hat{y}_S(x_t) \right] - \mathbb{E}_\epsilon \left[\inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t f(x_t) \right] \right\}$$

$$= \sup_{x_1,\ldots,x_{kn}} \left\{ \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t \hat{y}_S(x_t) \right] \right\}$$

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 :

$$\mathcal{V}_n^{\mathrm{stat}}(\mathcal{F}) \geq \sup_{x_1,\ldots,x_{kn}} \left\{ \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \frac{1}{kn} \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\sum_{t \in J} \epsilon_t \hat{y}_S(x_t) \right] \right\}$$

$$\geq \sup_{x_1,\ldots,x_{kn}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \frac{1}{kn} \sup_{x_1,\ldots,x_{kn}} \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\sum_{t \in J} \epsilon_t \hat{y}_S(x_t) \right]$$

$$= \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1,\ldots,x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right]$$

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

$$\mathcal{V}_n^{\mathrm{stat}}(\mathcal{F}) \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1,\ldots,x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right]$$

$$\geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1,\ldots,x_n} \mathbb{E}_\epsilon \left[\sup_{\hat{y} \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right]$$

$$= \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k} \mathcal{R}_n(\mathcal{F})$$

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

$$\mathcal{V}_n^{\mathrm{stat}}(\mathcal{F}) \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1,\ldots,x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right] \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k}$$

□

Using $k = 2$, in the above, we get that for proper learning algorithms, $\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \mathcal{R}_{2n}(\mathcal{F}) - \frac{1}{2}\mathcal{R}_n(\mathcal{F})$. If $\mathcal{R}_n(\mathcal{F}) = \Theta(n^{-p})$ for some $p \geq 2$ then, from this we conclude that if we consider minimax rate for proper learning,

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq 0.29\ \mathcal{R}_{2n}(\mathcal{F})$$

On the other hand if we consider improper learning as well, if $\mathcal{R}_n(\mathcal{F}) = \Omega(n^{-1/p})$ then picking $k = 2n^{1/(p-1)}$, in the lower bound above for improper learning we can conclude that,

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \Omega\left(n^{-\frac{1}{p-1}}\right)$$

3 Beyond Supervised Learning

If we consider problems beyond supervised learning with hinge loss, logistic loss etc. Is ERM still optimal, does it always work, do we need uniform convergence?

It turns out that for general statistical learning problems, it is not the case. Consider the following general learning problem.

Say $\mathcal{X} = \{0,1\}^d$ and say $\mathcal{Y} = \{y \in \mathbb{R}^d : \|y\|_2 \leq 1\}$ and let $\mathcal{F} = \{x \mapsto x \odot \mathbf{w} : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq 1\}$ and set

$$\ell(f_{\mathbf{w}}(x), y) = \frac{1}{2}\|f_{\mathbf{w}}(x) - y\|_2^2 = \frac{1}{2}\|\mathbf{w} \odot x - y\|_2^2$$

We make two claims:

1. This general learning problem is learnable and that too with a dimension (d) independent rate. We will show this using Stochastic Gradient Descent and with a rate that looks like $O(1/\sqrt{n})$.
2. This problem is not learnable using ERM, and uniform convergence fails for this example when d is very large (We will show for $d > 2^n$).

Together the two claims will show that there are problems that are (distribution free) learnable with nice rates but ERM is not the right algorithm and uniform convergence, Rademacher Complexity are not the right tools for the general learning setting.

Lemma 2. *There is a distribution over $\mathcal{X} \times \mathcal{Y}$ under which ERM fails and hence uniform convergence fails when $d > 2^n$. Specifically, there exists ERM such that*

$$\mathbb{E}_S\left[L_D(f_{\hat{\mathbf{w}}_{\text{ERM}}})\right] - \min_{f \in \mathcal{F}} L_D(f) \geq 1/4$$

Proof. Take distribution D to be uniform distribution over $\mathcal{X}$ and $\mathcal{Y}$ independently. Now note that

$$\begin{aligned}
L_D(f_{\mathbf{w}}) &= \frac{1}{2}\mathbb{E}_{x,y\sim D}\left[\|\mathbf{w}\odot x - y\|_2^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{d}\mathbb{E}_{x,y\sim D}\left[(\mathbf{w}[i]\cdot x[i] - y[i])^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{d}\mathbb{E}_{x,y\sim D}\left[(\mathbf{w}[i])^2\cdot x[i] + y[i]^2 - 2\mathbf{w}[i]\cdot x[i]\cdot y[i]\right] \\
&= \frac{1}{2}\sum_{i=1}^{d}\frac{1}{2}(\mathbf{w}[i])^2 + \mathbb{E}_{y\sim D}\left[y[i]^2\right] \\
&= \frac{1}{4}\|\mathbf{w}\|_2^2 + \|\mathbb{E}_{y\sim D}\left[y\right]\|^2
\end{aligned}$$

Clearly $\mathbf{w} = 0$ is the solution of minimizer of L_D. Now on the other hand, say we have a sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn iid from this distribution. In this case, notice that when $d > 2^n$ then, there is at least a constant probability that there exists one coordinate $\hat{i} \in [d]$ such that $\forall t \in [n], x_t[\hat{i}] = 0$. Now note that if we pick $\hat{\mathbf{w}}_{\text{ERM}} = e_{\hat{i}}$, such a solution is clearly an ERM (0 and $e_{\hat{i}}$ are both ERMs with same empirical loss). On the other hand,

$$L_D(e_{\hat{i}}) = \frac{1}{4} + \|\mathbb{E}_{y\sim D}\left[y\right]\|^2$$

Hence

$$L_D(e_{\hat{i}}) - L_D(0) \geq 1/4$$

and so clearly $e_{\hat{i}}$ has a sub-optimality of $1/4$ and hence is a bad solution. Hence, clearly there is an ERM what has bad suboptimality (as long as dimensionality is very large).

□

On the other hand, we claim that this problem is learnable using stochastic gradient descent because it is a convex problem.

Lemma 3. *Let* $\mathbf{w}_1 = 0$ *and for* $t > 1$*, define the SGD update:*

$$\mathbf{w}_{t+1} = \Pi(\mathbf{w}_t - \eta\nabla\ell(\mathbf{w}_t\odot\mathbf{x}_t, y_t))$$

where Π *is projection on to unit ball and* $\eta = 1/\sqrt{n}$*. Then we have that for the algorithm that returns* $\bar{\mathbf{w}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}_t$*, we have that for any arbitrary distribution on instance space,*

$$\mathbb{E}_S\left[L_D(\bar{\mathbf{w}}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \leq \sqrt{\frac{1}{n}}$$

Proof. Say $\mathbf{w}^* = \text{argmin}_{\mathbf{w}:\|\mathbf{w}\|_2 \leq 1} L_D(\mathbf{w})$, note that

$$\begin{aligned}
\mathbb{E}_S\left[\sum_{t=1}^{n}\left(L_D(\mathbf{w}_t) - L_D(\mathbf{w}^*)\right)\right] &\leq \mathbb{E}_S\left[\sum_{t=1}^{n}\langle \nabla L_D(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^*\rangle\right] \\
&= \mathbb{E}_S\left[\sum_{t=1}^{n}\langle \mathbb{E}_{x_t,y_t}\left[\nabla \ell(\mathbf{w}_t \odot x_t, y_t)\right], \mathbf{w}_t - \mathbf{w}^*\rangle\right] \\
&= \mathbb{E}_S\left[\sum_{t=1}^{n}\langle \nabla \ell(\mathbf{w}_t \odot x_t, y_t), \mathbf{w}_t - \mathbf{w}^*\rangle\right]
\end{aligned}$$

However note that

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 &= \|\Pi(\mathbf{w}_t - \eta \nabla \ell(\mathbf{w}_t \odot \mathbf{x}_t, y_t)) - \mathbf{w}^*\|_2^2 \\
&\leq \|\mathbf{w}_t - \eta \nabla \ell(\mathbf{w}_t \odot \mathbf{x}_t, y_t) - \mathbf{w}^*\|_2^2 \\
&\leq \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta^2 \|\nabla \ell(\mathbf{w}_t \odot \mathbf{x}_t, y_t)\|_2^2 - 2\eta \langle \nabla \ell(\mathbf{w}_t \odot \mathbf{x}_t, y_t), \mathbf{w}_t - \mathbf{w}^*\rangle
\end{aligned}$$

Hence using this we have:

$$\begin{aligned}
\mathbb{E}_S\left[\sum_{t=1}^{n}\left(L_D(\mathbf{w}_t) - L_D(\mathbf{w}^*)\right)\right] &\leq \mathbb{E}_S\left[\sum_{t=1}^{n}\frac{\eta}{2}\|\nabla \ell(\mathbf{w}_t \odot x_t, y_t)\|_2^2 + \frac{1}{2\eta}\left(\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2\right)\right] \\
&\leq \frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 + \frac{\eta}{2}\mathbb{E}_S\left[\sum_{t=1}^{n}\|\nabla \ell(\mathbf{w}_t \odot \mathbf{x}_t, y_t)\|_2^2\right]
\end{aligned}$$

However note that

$$\|\nabla \ell(\mathbf{w}_t \odot \mathbf{x}_t, y_t)\|_2^2 = \|x_t \odot (\mathbf{w}_t - y_t)\|^2 \leq 1$$

Hence we have:

$$\mathbb{E}_S\left[\sum_{t=1}^{n}\left(L_D(\mathbf{w}_t) - L_D(\mathbf{w}^*)\right)\right] \leq \frac{1}{2\eta}\|\mathbf{w}^*\|_2^2 + \eta n \leq \frac{1}{2\eta} + \frac{n}{2}\eta$$

Hence we conclude that using $\eta = 1/\sqrt{n}$ we have

$$\mathbb{E}_S\left[\sum_{t=1}^{n}\left(L_D(\mathbf{w}_t) - L_D(\mathbf{w}^*)\right)\right] \leq \sqrt{n}$$

By Convexity of loss and Jensen's inequality,

$$\mathbb{E}_S\left[\left(L_D\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{w}_t\right) - L_D(\mathbf{w}^*)\right)\right] \leq \mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n}\left(L_D(\mathbf{w}_t) - L_D(\mathbf{w}^*)\right)\right] \leq \sqrt{\frac{1}{n}}$$

Hence returning $\bar{\mathbf{w}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}_t$ gives our algorithm. □

Thus we see that stepping beyond supervised learning problem, ERM or uniform-convergence is not the right tool. In that case what is?