Machine Learning Theory (CS 6783)

Lecture 5 : Minimax Rates, Statistical Learning and Uniform Convergence

1 Minimax Rate

How well does the best learning algorithm do in the worst case scenario?

$$\text{Minimax Rate} = \text{“Best Possible Guarantee”} = \min_{\text{Algo } \hat{\mathbf{y}}} \max_{\text{instance setting}} \text{Objective}$$

PAC framework:

$$\mathcal{V}_n^{PAC}(\mathcal{F}) := \inf_{\hat{\mathbf{y}}} \sup_{D_X, f^* \in \mathcal{F}} \mathbb{E}_{S:|S|=n} \left[\mathbb{P}_{x \sim D_x}\left(\hat{\mathbf{y}}(x) \neq f^*(x)\right)\right]$$

A problem is “PAC learnable” if $\mathcal{V}_n^{PAC} \to 0$. That is, there exists a learning algorithm that converges to 0 expected error as sample size increases.
Non-parametric Regression:

$$\mathcal{V}_n^{NR}(\mathcal{F}) := \inf_{\hat{\mathbf{y}}} \sup_{D_X, f^* \in \mathcal{F}} \mathbb{E}_{S:|S|=n} \left[\mathbb{E}_{x \sim D_X}\left[(\hat{\mathbf{y}}(x) - f^*(x))^2\right]\right]$$

A statistical estimation problem is consistent if $\mathcal{V}_n^{NR} \to 0$.
Statistical learning:

$$\mathcal{V}_n^{stat}(\mathcal{F}) := \inf_{\hat{\mathbf{y}}} \sup_{D} \mathbb{E}_{S:|S|=n} \left[L_D(\hat{\mathbf{y}}) - \inf_{f \in \mathcal{F}} L_D(f)\right]$$

A problem is “statistically learnable” if $\mathcal{V}_n^{stat} \to 0$.
Statistical learning:

$$\mathcal{V}_n^{stat}(\mathcal{F}) := \inf_{\hat{\mathbf{y}}} \sup_{D} \mathbb{E}_{S:|S|=n} \left[L_D(\hat{\mathbf{y}}) - \inf_{f \in \mathcal{F}} L_D(f)\right]$$

A problem is “statistically learnable” if $\mathcal{V}_n^{stat} \to 0$.
Online learning:

$$\mathcal{V}_n^{sq}(\mathcal{F}) := \sup_{x_1} \inf_{\hat{y}_1} \sup_{y_1} \sup_{x_2} \inf_{\hat{y}_2} \sup_{y_2} \ldots \sup_{x_n} \inf_{\hat{y}_n} \sup_{y_n} \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right\}$$

A problem is “online learnable” if $\mathcal{V}_n^{sq} \to 0$.

A statement in expectation implies statement in high probability by Markov inequality but more generally one can also easily convert to exponentially high probability.

1.1 Comparing the Minimax Rates

Proposition 1. *For any class* $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$,

$$4\mathcal{V}_n^{PAC}(\mathcal{F}) \leq \mathcal{V}_n^{NR}(\mathcal{F}) \leq \mathcal{V}_n^{stat}(\mathcal{F})$$

and for any $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$,

$$\mathcal{V}_n^{NR}(\mathcal{F}) \leq \mathcal{V}_n^{stat}(\mathcal{F})$$

That is, if a class is statistically learnable then it is learnable under either the PAC model or the statistical estimation setting

Proof. Let us start with the PAC learning objective. Note that,

$$\mathbb{1}_{\{\hat{\mathbf{y}}(x) \neq f^*(x)\}} = \frac{1}{4}(\hat{\mathbf{y}}(x) - f^*(x))^2$$

Now note that,

$$\begin{aligned}
\mathbb{P}_{x \sim D_x}\left(\hat{\mathbf{y}}(x) \neq f^*(x)\right) &= \mathbb{E}_{x \sim D_X}\left[\ \mathbb{1}_{\{\hat{\mathbf{y}}(x) \neq f^*(x)\}}\right] \\
&= \frac{1}{4}\mathbb{E}_{x \sim D_X}\left[(\hat{\mathbf{y}}(x) - f^*(x))^2\right]
\end{aligned}$$

Thus we conclude that

$$4\mathcal{V}_n^{PAC}(\mathcal{F}) \leq \mathcal{V}_n^{NR}(\mathcal{F})$$

Now to conclude the proposition we prove that the minimax rate for non-parametric regression is upper bounded by minimax rate for the statistical learning problem (under squared loss).

To this end, in NR we assume that $y = f^*(x) + \varepsilon$ for zero-mean noise ε. Now note that, Now note that, for any $\hat{\mathbf{y}}$,

$$\begin{aligned}
(\hat{\mathbf{y}}(x) - f^*(x))^2 &= (\hat{\mathbf{y}}(x) - y - \varepsilon)^2 \\
&= (\hat{\mathbf{y}}(x) - y)^2 - 2\varepsilon(\hat{\mathbf{y}}(x) - y) + \varepsilon^2 \\
&= (\hat{\mathbf{y}}(x) - y)^2 - (f^*(x) - y)^2 + (f^*(x) - y)^2 - 2\varepsilon(\hat{\mathbf{y}}(x) - y) + \varepsilon^2 \\
&= (\hat{\mathbf{y}}(x) - y)^2 - (f^*(x) - y)^2 + 2\varepsilon^2 - 2\varepsilon(\hat{\mathbf{y}}(x) - y) \\
&= (\hat{\mathbf{y}}(x) - y)^2 - (f^*(x) - y)^2 + 2\varepsilon^2 - 2\varepsilon(\hat{\mathbf{y}}(x) - f^*(x) - \varepsilon) \\
&= (\hat{\mathbf{y}}(x) - y)^2 - (f^*(x) - y)^2 - 2\varepsilon(\hat{\mathbf{y}}(x) - f^*(x))
\end{aligned}$$

Taking expectation w.r.t. y (or ε) we conclude that,

$$\begin{aligned}
\mathbb{E}_{x \sim D_X}\left[(\hat{\mathbf{y}}(x) - f^*(x))^2\right] &= \mathbb{E}_{(x,y) \sim D}\left[(\hat{\mathbf{y}}(x) - y)^2\right] - \mathbb{E}_{(x,y) \sim D}\left[(f^*(x) - y)^2\right] - \mathbb{E}_{x \sim D_X}\left[\mathbb{E}_{\varepsilon}\left[2\varepsilon(\hat{\mathbf{y}}(x) - f^*(x))\right]\right] \\
&= \mathbb{E}_{(x,y) \sim D}\left[(\hat{\mathbf{y}}(x) - y)^2\right] - \mathbb{E}_{(x,y) \sim D}\left[(f^*(x) - y)^2\right] \\
&= L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f)
\end{aligned}$$

where in the above distribution D has marginal D_X over $\mathcal{X}$ and the conditional distribution $D_{Y|X=x} = N(f^*(x), \sigma)$. Hence we conclude that

$$\mathcal{V}_n^{NR}(\mathcal{F}) \leq \mathcal{V}_n^{stat}(\mathcal{F})$$

when we consider statistical learning under square loss. □

2 No Free Lunch Theorem

The more expressive the class $\mathcal{F}$ is, the larger is $\mathcal{V}_n^{PAC}(\mathcal{F}), \mathcal{V}_n^{NR}(\mathcal{F})$ and $\mathcal{V}_n^{stat}(\mathcal{F})$. The no free lunch theorem says that if $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$ the set of all function, then there is not convergence of minimax rates.

Proposition 2. *If $|\mathcal{X}| \geq 2n$ then,*

$$\mathcal{V}_n^{PAC}(\mathcal{Y}^{\mathcal{X}}) \geq \frac{1}{4}$$

Proof. Consider D_X to be the uniform distribution over $2n$ points. Also let $f^* \in \mathcal{Y}^{\mathcal{X}}$ be a random choice of the possible 2^{2n} function on these points. Now if we obtain sample S of size at most n, then

$$\begin{aligned}
\mathcal{V}_n^{PAC}(\mathcal{Y}^{\mathcal{X}}) &= \inf_{\hat{\mathbf{y}}} \sup_{D_X, f^* \in \mathcal{F}} \mathbb{E}_{S:|S|=n}\left[\mathbb{P}_{x\sim D_x}\left(\hat{\mathbf{y}}(x) \neq f^*(x)\right)\right] \\
&\geq \inf_{\hat{\mathbf{y}}} \mathbb{E}_{f^*}\left[\mathbb{E}_{S:|S|=n}\left[\mathbb{P}_{x\sim D_x}\left(\hat{\mathbf{y}}(x) \neq f^*(x)\right)\right]\right] \\
&= \inf_{\hat{\mathbf{y}}} \mathbb{E}_{f^*}\left[\mathbb{E}_{S:|S|=n}\left[\frac{1}{2n}\sum_{j=1}^{2n} \mathbb{1}_{\{\hat{\mathbf{y}}(x_j)\neq f^*(x_j)\}}\right]\right] \\
&\geq \frac{1}{2n} \inf_{\hat{\mathbf{y}}} \mathbb{E}_{f^*}\left[\mathbb{E}_{i_1,\ldots,i_n\sim \mathrm{Unif}[2n]}\left[\sum_{j\notin\{i_1,\ldots,i_n\}} \mathbb{1}_{\{\hat{\mathbf{y}}(x_j)\neq f^*(x_j)\}}\right]\right] \\
&= \frac{1}{2n} \inf_{\hat{\mathbf{y}}} \mathbb{E}_{i_1,\ldots,i_n\sim \mathrm{Unif}[2n]}\left[\mathbb{E}_{f^*}\left[\sum_{j\notin\{i_1,\ldots,i_n\}} \mathbb{1}_{\{\hat{\mathbf{y}}(x_j)\neq f^*(x_j)\}}\right]\right]
\end{aligned}$$

But outside of sample S, on each x, $f^*(x)$ can be ± 1 with equal probability. Hence,

$$\mathcal{V}_n^{PAC}(\mathcal{Y}^{\mathcal{X}}) \geq \frac{1}{2n} \inf_{\hat{\mathbf{y}}} \mathbb{E}_{i_1,\ldots,i_n\sim \mathrm{Unif}[2n]}\left[\mathbb{E}_{f^*}\left[\sum_{j\notin\{i_1,\ldots,i_n\}} \mathbb{1}_{\{\hat{\mathbf{y}}(x_j)\neq f^*(x_j)\}}\right]\right] \geq \frac{1}{2n}\frac{n}{2} = \frac{1}{4}$$

□

This shows that we need some restriction on $\mathcal{F}$ even for the realizable PAC setting. We cannot learn arbitrary set of hypothesis, there is no free lunch.

This tells us that we need to restrict the set of models $\mathcal{F}$ we consider,

3 Empirical Risk Minimization and The Empirical Process

One algorithm/principle/ learning rule that is natural for statistical learning problems is the Empirical Risk Minimizer (ERM) algorithm. That is pick the hypothesis from model class $\mathcal{F}$ that best fits the sample, or in other words,:

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

Claim 3. *For any $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{F}$ and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ (subject to mild regularity conditions required for measurability), we have that*

$$\begin{aligned}\mathcal{V}_n^{\text{stat}}(\mathcal{F}) &\leq \sup_D \mathbb{E}_S \left[L_D(\hat{y}_{\text{erm}}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \\ &\leq \sup_D \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| \right]\end{aligned}$$

Proof. Note that

$$\begin{aligned}&\mathbb{E}_S\left[L_D(\hat{y}_{\text{erm}})\right] - \inf_{f \in \mathcal{F}} L_D(f) \\ &\qquad = \mathbb{E}_S\left[L_D(\hat{y}_{\text{erm}})\right] - \inf_{f \in \mathcal{F}} \mathbb{E}_S \left[\frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\ &\qquad \leq \mathbb{E}_S \left[L_D(\hat{y}_{\text{erm}}) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\ &\qquad \leq \mathbb{E}_S \left[L_D(\hat{y}_{\text{erm}}) - \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}_{\text{erm}}(x_t), y_t) \right]\end{aligned}$$

since $\hat{y}_{\text{erm}} \in \mathcal{F}$, we can pass to upper bound by replacing with supremum over all $f \in \mathcal{F}$ as

$$\begin{aligned}&\leq \mathbb{E}_S \sup_{f \in \mathcal{F}} \left[\mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\ &\leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| \right]\end{aligned}$$

This completes the proof. □

Fact: Consider r.v. $Z_1, \ldots, Z_n$ drawn iid from some fixed distribution, assume Z_t's are bounded by 1. Let $\mu = \mathbb{E}[Z]$ be their expectation. We have the following bound on the average of these random variables.

$$P\left(\left| \mu - \frac{1}{n} \sum_{t=1}^{n} Z_t \right| > \theta \right) \leq 2 \exp\left(-\frac{n\theta^2}{2} \right)$$

Now for any $f \in \mathcal{F}$, let $Z_t^f = \ell(f(x_t), y_t)$ where (x_t, y_t) is drawn from D. Note that $\mathbf{E}[Z^f] = \mathbb{E}_{(x,y)\sim D} \ell(f(x), y)$. Hence note that for any single $f \in \mathcal{F}$,

$$P_S\left(\left| \mathbb{E}_{(x,y)\sim D} \ell(f(x), y) - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| > \theta \right) \leq 2 \exp\left(-\frac{n\theta^2}{2} \right)$$

Taking a union bound we conclude that:

$$P_S\left(\max_{f \in \mathcal{F}} \left| \mathbb{E}_{(x,y)\sim D} \ell(f(x), y) - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| > \theta \right) \leq 2|\mathcal{F}| \exp\left(-\frac{n\theta^2}{2} \right)$$

Now using the fact that for a non-negative random variable X, $\mathbb{E}[X] = \int_0^\infty P(X > x)dx$ we have that for any choice of $\epsilon > 0$:

$$
\begin{aligned}
&\mathbb{E}_S\left[\max_{f\in\mathcal{F}}\left|\mathbb{E}_{(x,y)\sim D}\ell(f(x),y) - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|\right] \\
&= \int_0^\infty P_S\left(\max_{f\in\mathcal{F}}\left|\mathbb{E}_{(x,y)\sim D}\ell(f(x),y) - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right| > \theta\right)d\theta \\
&\le \int_0^\epsilon d\theta + \int_\epsilon^\infty P_S\left(\max_{f\in\mathcal{F}}\left|\mathbb{E}_{(x,y)\sim D}\ell(f(x),y) - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right| > \theta\right)d\theta \\
&\le \epsilon + 2|\mathcal{F}|\int_\epsilon^\infty \exp\left(-\frac{n\theta^2}{2}\right)d\theta \\
&= \epsilon + \frac{2|\mathcal{F}|}{\sqrt{n}}\int_{\sqrt{n}\epsilon}^\infty \exp\left(-\frac{x^2}{2}\right)dx \\
&\le \epsilon + \frac{2|\mathcal{F}|}{\sqrt{n}}e^{-n\epsilon^2}
\end{aligned}
$$

Using $\epsilon = \sqrt{\log(2|\mathcal{F}|)/n}$ we have:

$$
\begin{aligned}
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) &\le \sup_D \mathbb{E}_S\left[L_D(\hat{y}_{\text{erm}}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \\
&\le \sup_D \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|\right] \\
&\le O\left(\sqrt{\frac{\log|\mathcal{F}|}{n}}\right)
\end{aligned}
$$