1 Minimax Rate

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

Minimax Rate = “Best Possible Guarantee”

PAC framework:

\[ \mathcal{V}_n^{PAC}(\mathcal{F}) := \inf \sup_{\hat{y}} \mathbb{E}_{S:|S|=n} [\mathbb{P}_{x \sim D_x} (\hat{y}(x) \neq f^*(x))] \]

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:

\[ \mathcal{V}_n^{NR}(\mathcal{F}) := \inf \sup_{\hat{y}} \mathbb{E}_{S:|S|=n} \mathbb{E}_{x \sim D_x} [(\hat{y}(x) - f^*(x))^2] \]

A statistical estimation problem is consistent if \( \mathcal{V}_n^{NR} \to 0 \).

Statistical learning:

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

A problem is “statistically learnable” if \( \mathcal{V}_n^{stat} \to 0 \).

Statistical learning:

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

A problem is “statistically learnable” if \( \mathcal{V}_n^{stat} \to 0 \).

Online learning:

\[ \mathcal{V}_n^{sq}(\mathcal{F}) := \sup \inf \sup \sup \sup \ldots \sup \sup \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right\} \]

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 \( \mathcal{F} \subset \{\pm 1\}^X \),

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

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

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

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,

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

Now note that,

\[
P_{x \sim D_x}(\hat{y}(x) \neq f^*(x)) = \mathbb{E}_{x \sim D_x}\left[ \mathbf{1}_{\{\hat{y}(x) \neq f^*(x)\}} \right]
\]

\[
= \frac{1}{4}\mathbb{E}_{x \sim D_x}\left[ (\hat{y}(x) - f^*(x))^2 \right]
\]

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 \( \varepsilon \). Now note that, Now note that, for any \( \hat{y} \),

\[
(\hat{y}(x) - f^*(x))^2 = (\hat{y}(x) - y - \varepsilon)^2
\]

\[
= (\hat{y}(x) - y)^2 - 2\varepsilon(\hat{y}(x) - y) + \varepsilon^2
\]

\[
= (\hat{y}(x) - y)^2 - (f^*(x) - y)^2 + (f^*(x) - y)^2 - 2\varepsilon(\hat{y}(x) - y) + \varepsilon^2
\]

\[
= (\hat{y}(x) - y)^2 - (f^*(x) - y)^2 + 2\varepsilon^2 - 2\varepsilon(\hat{y}(x) - y)
\]

\[
= (\hat{y}(x) - y)^2 - (f^*(x) - y)^2 + 2\varepsilon^2 - 2\varepsilon(\hat{y}(x) - f^*(x) - \varepsilon)
\]

\[
= (\hat{y}(x) - y)^2 - (f^*(x) - y)^2 - 2\varepsilon(\hat{y}(x) - f^*(x))
\]

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

\[
\mathbb{E}_{x \sim D_x}\left[ (\hat{y}(x) - f^*(x))^2 \right] = \mathbb{E}_{(x,y) \sim D}\left[ (\hat{y}(x) - y)^2 \right] - \mathbb{E}_{(x,y) \sim D}\left[ (f^*(x) - y)^2 \right] - \mathbb{E}_{x \sim D_x}\left[ 2\varepsilon(\hat{y}(x) - f^*(x)) \right]
\]

\[
= \mathbb{E}_{(x,y) \sim D}\left[ (\hat{y}(x) - y)^2 \right] - \mathbb{E}_{(x,y) \sim D}\left[ (f^*(x) - y)^2 \right] - L_D(\hat{y}) + \inf_{f \in \mathcal{F}} L_D(f)
\]

where in the above distribution \( D \) has marginal \( D_X \) over \( X \) 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. \( \Box \)
2 No Free Lunch Theorem

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

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

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

**Proof.** Consider $D_X$ to be the uniform distribution over $2n$ points. Also let $f^* \in \mathcal{Y}^X$ 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

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

$$\geq \inf_{\hat{y}} \mathbb{E}_{f^*} \left[ \mathbb{E}_{S: |S| = n} \left[ \frac{1}{2n} \sum_{j=1}^{2n} 1\{\hat{y}(x_j) \neq f^*(x_j)\} \right] \right]$$

$$= \inf_{\hat{y}} \mathbb{E}_{f^*} \left[ \mathbb{E}_{i_1, \ldots, i_n \sim \text{Unif}[2n]} \left[ \sum_{j \notin \{i_1, \ldots, i_n\}} 1\{\hat{y}(x_j) \neq f^*(x_j)\} \right] \right]$$

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

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

This shows that we need some restriction on $F$ 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 $F$ 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 $F$ that best fits the sample, or in other words:

$$\hat{y}_{\text{erm}} = \arg\min_{f \in F} \sum_{t=1}^{n} \ell(f(x_t), y_t)$$
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

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \sup_D \mathbb{E}_S \left[ L_D(\hat{y}_{\text{erm}}) - \inf_{f \in \mathcal{F}} L_D(f) \right]$$

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

Proof. Note that

$$\mathbb{E}_S [L_D(\hat{y}_{\text{erm}})] - \inf_{f \in \mathcal{F}} L_D(f)$$

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

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

$$\leq \mathbb{E}_S \left[ L_D(\hat{y}_{\text{erm}}) - \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}_{\text{erm}}(x_t), y_t) \right]$$

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

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

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

This completes the proof. 

- The question of whether minimax value converges to 0, or equivalently whether the problem is learnable can now be understood by studying if, uniformly over class $\mathcal{F}$ does average converge to expected loss ?

- For bounded losses, for any fixed $f \in \mathcal{F}$, the difference of average loss and expected loss for a given $f \in \mathcal{F}$ goes to 0 by Hoeffding bound.

- The difference of average loss and expected loss is an empirical process indexed by class $\mathcal{F}$. We study supremum (over $\mathcal{F}$) of these empirical processes. This is the main question of interest in empirical process theory.