1 Recap

PAC framework:

\[ V_{n}^{PAC}(F) := \inf \sup E_{S:|S|=n} [P_{x \sim D_{n}} (\hat{y}(x) \neq f^{*}(x))] \]

A problem is “PAC learnable” if \( V_{n}^{PAC} \to 0 \). That is, there exists a learning algorithm that converges to 0 expected error as sample size increases.

Non-parametric Regression:

\[ V_{n}^{NR}(F) := \inf \sup E_{S:|S|=n} [E_{x \sim D_{X}} [ (\hat{y}(x) - f^{*}(x))^2 ]] \]

A statistical estimation problem is consistent if \( V_{n}^{NR} \to 0 \).

Statistical learning:

\[ V_{n}^{stat}(F) := \inf \sup E_{S:|S|=n} [L_{D}(\hat{y}) - \inf_{f \in F} L_{D}(f)] \]

A problem is “statistically learnable” if \( V_{n}^{stat} \to 0 \).

Statistical learning:

\[ V_{n}^{stat}(F) := \inf \sup E_{S:|S|=n} [L_{D}(\hat{y}) - \inf_{f \in F} L_{D}(f)] \]

A problem is “statistically learnable” if \( V_{n}^{stat} \to 0 \).

Proposition 1. For any class \( F \subset \{\pm 1\}^{X} \),

\[ 4V_{n}^{PAC}(F) \leq V_{n}^{NR}(F) \leq V_{n}^{stat}(F) \]

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

\[ V_{n}^{NR}(F) \leq V_{n}^{stat}(F) \]

2 No Free Lunch Theorem

The more expressive the class \( F \) is, the larger is \( V_{n}^{PAC}(F), V_{n}^{NR}(F) \) and \( V_{n}^{stat}(F) \). 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 $|\mathcal{X}| \geq 2n$ 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
\[ \mathcal{V}_{n}^{PAC}(\mathcal{Y}^{X}) = \inf_{\hat{y}} \sup_{D_{X},f^{*} \in \mathcal{F}} \mathbb{E}_{S|S| = n} [\mathbb{P}_{x \sim D_{x}}(\hat{y}(x) \neq f^{*}(x))] \]
\[ \geq \inf_{\hat{y}} \mathbb{E}_{f^{*}} \left[ \mathbb{E}_{S|S| = n} [\mathbb{P}_{x \sim D_{x}}(\hat{y}(x) \neq f^{*}(x))] \right] = \inf_{\hat{y}} \mathbb{E}_{f^{*}} \left[ \mathbb{E}_{S|S| = n} \left[ \frac{1}{2n} \sum_{j=1}^{2n} 1(\hat{y}(x_{j}) \neq f^{*}(x_{j})) \right] \right] \]
\[ \geq \frac{1}{2n} \inf_{\hat{y}} \mathbb{E}_{f^{*}} \left[ \mathbb{E}_{i_{1},\ldots,i_{n} \sim \text{Unif}[2n]} \left[ \sum_{j \notin \{i_{1},\ldots,i_{n}\}} 1(\hat{y}(x_{j}) \neq f^{*}(x_{j})) \right] \right] \]
\[ = \frac{1}{2n} \inf_{\hat{y}} \mathbb{E}_{i_{1},\ldots,i_{n} \sim \text{Unif}[2n]} \left[ \mathbb{E}_{f^{*}} \left[ \sum_{j \notin \{i_{1},\ldots,i_{n}\}} 1(\hat{y}(x_{j}) \neq f^{*}(x_{j})) \right] \right] \]
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 $\mathcal{F}$ 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,
\[ \hat{y}_{\text{erm}} = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_{t}), y_{t}) \]

Claim 3. For any $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{F}$ and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ (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

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

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

This completes the proof. \( \square \)

- 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.

### 3.1 Example: Finite Class

For now and for most of this course we shall assume that the loss \( \ell \) is bounded by 1, that is \( \sup_{y,y' \in Y} |\ell(y', y)| \leq 1 \).

**Claim 4.** Consider the case when the hypothesis \( \mathcal{F} \) has finite cardinality, that is \( |\mathcal{F}| < \infty \). For any loss \( \ell \) bounded by 1, we have that

\[
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \sup_D \mathbb{E}_S \left[ L_D(\hat{y}_{\text{erm}}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \leq 8 \sqrt{\frac{\log n |\mathcal{F}|}{n}}
\]

**Proof.** By Claim 3 we have that

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

\[
\leq \mathbb{E}_S \left[ \max_{f \in \mathcal{F}} \left| \mathbb{E} [\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| \right]
\]
Now let us define for every $f \in \mathcal{F}$, the random variable $Z^f$ as $\ell(f(x), y)$ where $(x, y) \sim D$. Note that $\mathbb{E}[Z^f] = \mathbb{E}[\ell(f(x), y)]$ and $Z_1^f, \ldots, Z_n^f$ are $n$ iid copies of $Z^f$. Hence by Hoeffding’s inequality we have that:

$$P_S\left(\left|\mathbb{E}[Z^f] - \frac{1}{n} \sum_{t=1}^{n} Z_t^f\right| > \epsilon\right) \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Hence by union bound we have that

$$P_S\left(\max_{f \in \mathcal{F}} \left|\mathbb{E}[Z^f] - \frac{1}{n} \sum_{t=1}^{n} Z_t^f\right| > \epsilon\right) \leq 2|\mathcal{F}| \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Hence,

$$\mathbb{E}\left[\max_{f \in \mathcal{F}} \left|\mathbb{E}[Z^f] - \frac{1}{n} \sum_{t=1}^{n} Z_t^f\right|\right] \leq \epsilon P\left(\max_{f \in \mathcal{F}} \left|\mathbb{E}[Z^f] - \frac{1}{n} \sum_{t=1}^{n} Z_t^f\right| \leq \epsilon\right) + 2P\left(\max_{f \in \mathcal{F}} \left|\mathbb{E}[Z^f] - \frac{1}{n} \sum_{t=1}^{n} Z_t^f\right| > \epsilon\right)$$

$$\leq \epsilon + 4|\mathcal{F}| \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Choosing $\epsilon = \sqrt{\log(n|\mathcal{F}|^2)/n}$ we get,

$$\mathbb{E}\left[\max_{f \in \mathcal{F}} \left|\mathbb{E}[Z^f] - \frac{1}{n} \sum_{t=1}^{n} Z_t^f\right|\right] \leq 8\sqrt{\frac{\log n|\mathcal{F}|}{n}}$$

Hence for a finite class $\mathcal{F}$, using the ERM algorithm, one can achieve a rate of $O\left(\sqrt{\frac{\log n|\mathcal{F}|}{n}}\right)$ (in fact the $\sqrt{\log n}$ factor can be shaved off with more careful analysis).

What about infinite class $\mathcal{F}$?