The Learning Problem

Input: $x \in \mathcal{X} \longrightarrow$ features

Output: $y \in \mathcal{Y}$

Target function: $f : \mathcal{X} \longrightarrow \mathcal{Y}$

Data: $\underbrace{(x_1, y_1), \ldots\ldots, (x_n, y_n)}_{\text{Training Data}}, \;\Big|\; \underbrace{(x_{n+1}, y_{n+1}), \ldots\ldots, (x_m, y_m)}_{\text{Testing Data}}$

Loss fn: $L(x, y, f(x))$, $\quad L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \longrightarrow \mathbb{R}^{\geq 0}$

Risk: (True)

$$R_{L,P}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) \; \underbrace{dP(x,y)}_{\substack{\text{Generating pdf of input} \\ \text{\& output} \\ = p(y,x)\, dy\, dx}}$$

$$= \mathbb{E}\left[L(X, Y, f(X)) \right]$$

$$= \text{expected loss}$$

Here, $\mathbb{P}(X, Y) = \underbrace{\mathbb{P}(Y \mid X)}_{\substack{\text{probabilistic} \\ \text{model of i/p \& o/p}}} \; \underbrace{\mathbb{P}(X)}_{\substack{\text{probabilistic model} \\ \text{of data generation}}}$

Bayes risk:

$$R^*_{L,P}(f) = \inf_{f : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) \, dP(x, y)$$

$$= \inf_f \left[R_{L,P}(f) \right]$$

Goal Infer f_D using dataset D whose risk $R_{L,P}(f_D)$ is closest to $R^*_{L,P}(f)$

Consistent Learning

If f_D is the inferred model from data D, the learning is said to be universally consistent if

as $n \to \infty$ & $\forall$ $P(X, Y)$

$$R_{L,P}(f_D) \xrightarrow{n} R^*_{L,P}$$

$\longrightarrow$ Stone's theorem (1977)

No Free Lunch

$\forall$ consistent learning $\wedge$ $\forall$ convergence rate a_n,

$\exists\, P(X, Y)$ s.t. convergence rate of this learning method is slower than a_n.

$\longrightarrow$ Stochastic bound

Prove $\exists\, M$ $\forall\, \epsilon > 0$ s.t.

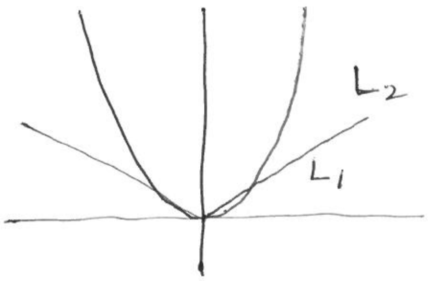
$$\mathbb{P}(|X_n| > M) < \epsilon \quad \forall n$$

Regression

$$x \in \mathcal{X} \subset \mathbb{R}^d$$

$$y \in \mathcal{Y} = [a, b] \subseteq \mathbb{R}$$

$$L(x, y, f(x)) = (y - f(x))^2$$



Huber's robust loss

$$L(x, y, f(x))$$

$$= \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{if } |y - f(x)| < 1 \\ |y - f(x)| - \frac{1}{2} & \text{otherwise} \end{cases}$$

$$R_{L,P}(f) = \mathbb{E}\left[(Y - f(X))^2\right]$$

Classification

$$x \in \mathcal{X} \subset \mathbb{R}^d$$

$$y \in \mathcal{Y} = \{0, 1\} \subset \mathbb{R}$$

$$L(x, y, f(x)) = \begin{cases} 1 & y \neq f(x) \\ 0 & y = f(x) \end{cases}$$

$$= 1 - \mathbb{1}(y = f(x))$$

$$R_{L,P}(f)$$

$$= 1 - \mathbb{E}\left[\mathbb{1}(y = f(x))\right]$$

$$= 1 - \mathbb{P}(y = f(x))$$

Empirical risk

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$$

$$= R_n^{train}(f)$$

$\boxed{\text{for simplicity, we write } L(x, y, f(x)) = L(y, f(x))}$

Goal: $$f^* = \underset{f}{\arg\min} \frac{1}{m-n} \sum_{i=n+1}^{m} L(\cancel{Y}_i, f(\cancel{X}_i))$$

$$= \underset{f}{\arg\min}\ R_{m-n}^{test}(f)$$

By LLN, $$R_n^{train} \xrightarrow{n \to \infty} R(f)$$

$$R_{m-n}^{test} \xrightarrow{m-n \to \infty} R(f)$$

Thus, we want to minimise R_n^{train} & R_{m-n}^{test}

But just minimising R_n^{train} to zero gives extreme overfitting.

eg 1.
$$f(x) = \begin{cases} Y_i & x = X_i \quad \forall\, i = 1, \dots, n \\ \text{any real value} & \text{otherwise} \end{cases}$$

$\Rightarrow R_n^{train} = 0$, $R(f) \neq 0$, R_{m-n}^{test} very high

$\hookrightarrow$ Overfitting

Soln: Minimise over a less complex hypotheses set & gradually increase the complexity of the set.

If $\mathcal{H}_t$ is the hypothesis set at an iteration t,

$$\underbrace{R(f)}_{\text{risk of classifier}} - \underbrace{R^*(f^*)}_{\text{Bayesian risk}} = \underbrace{R(f) - \inf_{f \in \mathcal{H}_t} R(f)}_{\text{Term 1}} + \underbrace{\inf_{f \in \mathcal{H}_t} R(f) - R(f^*)}_{\text{Term 2}}$$

Term 1 = Estimation error

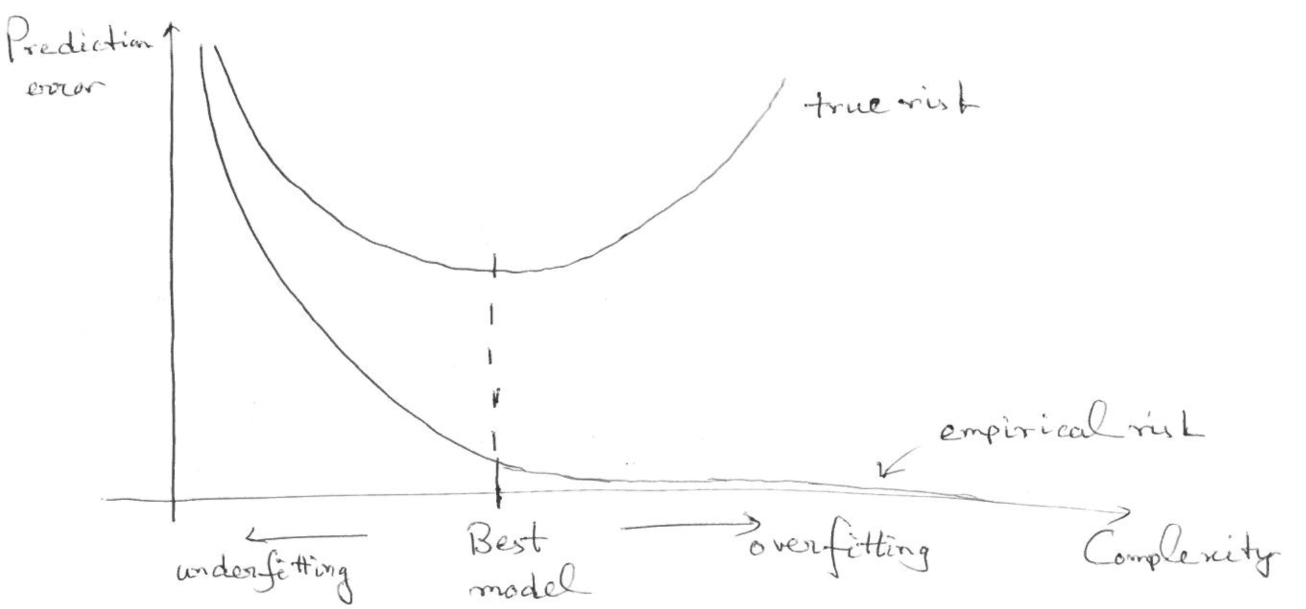
$$= R(f) - \inf_{f \in \mathcal{H}_t} R(f) = R(f) - R^*(f \mid f \in \mathcal{H}_t)$$

Term 2 = Approximation error

$$= \inf_{f \in \mathcal{H}_t} R(f) - R(f^*) = R^*(f \mid f \in \mathcal{H}_t) - R(f^*)$$

Term 2 should be non-negative.

Thus, there is this trade-off between complexity of $\mathcal{H}_t$ & minimising the estimation error.



$\longrightarrow$ Empirical risk alone is not a good performance measure & all loss functions are not eligible for learning.

→ Optimization Problem

$$R^*_{n,\mathcal{H}} = \inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$$