Motivation

$\downarrow$ CDF

$$\begin{cases} \text{ECDF} & \hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\} \\ \text{CDF} & F(x) = \mathbb{P}(X \le x) = \mathbb{E}[\mathbb{1}\{X \le x\}] \end{cases}$$

interest: $\gamma(F)$

$\downarrow$

plug-in estimator: $\gamma(F_n)$

$\|\hat{F}_n - F\|_\infty \xrightarrow{P} 0$

Hoeffding can only give point-wise bound $\mathbb{P}(|\hat{F}_n(x) - F(x)| \ge \varepsilon) \le C$

① LLN $\Rightarrow$ $\hat{F}_n(x) \xrightarrow{P/a.s.} F(x)$

② DKW $\Rightarrow$ $\mathbb{P}(\|F_n - F\|_\infty \ge \varepsilon) \le 2\exp(-2n\varepsilon^2)$

③ G-C thm $\Rightarrow$ $\boxed{\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{P} 0}$ $\to$ uniform converge

uniform

$\Longrightarrow$ **Empirical Process** $X_1, \dots, X_n \sim \mathbb{P}(\cdot)$

$$\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \int f \, d\mathbb{P}_n$$

$\downarrow$

$$\mathbb{P}(f) = \mathbb{E}_{\mathbb{P}}[f(X)] = \int f \, d\mathbb{P}$$

Answer this question $\longrightarrow$

Our interest: $\sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_F[f(X)] \right| \xrightarrow{P} 0$?

$$\Leftrightarrow \sup_{f \in \mathcal{F}} |\mathbb{P}_n(f) - \mathbb{P}(f)| \xrightarrow{P} 0$$

$$\Leftrightarrow \|\mathbb{P}_n - \mathbb{P}\|_{f \in \mathcal{F}} \xrightarrow{P} 0 \quad ?$$

Application: ERM $\quad (X_i, Y_i) \rightsquigarrow \theta^*$

$$\begin{cases} \hat{R}_n(\theta, \theta^*) = \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i, Y_i) \\ R(\theta, \theta^*) = \mathbb{E}_{(X,Y)\sim\theta^*}[\ell_\theta(X, Y)] \end{cases}$$

$$\begin{cases} \text{MLE} \Leftrightarrow \ell_\theta(x,y) = -\log \frac{P_{\theta^*}(x,y)}{P_\theta(x,y)} \\ \text{classify} \Rightarrow \ell_\theta(x,y) = \mathbb{1}\{f_\theta(x) \ne y\} \end{cases}$$

$$\begin{cases} \rightarrow \hat{\theta} = \underset{\theta \in \Theta}{\arg\min}\ \hat{R}_n(\theta, \theta^*) \iff \hat{f} = \underset{f \in \mathcal{H}}{\arg\min}\ \hat{R}_n(f, f^*) \\ \rightarrow \tilde{\theta} = \underset{\theta \in \Theta}{\arg\min}\ R(\theta, \theta^*) \iff \tilde{f} = \underset{f \in \mathcal{H}}{\arg\min}\ R(f, f^*) \end{cases}$$

Excess Risk

Consider $E(\hat{\theta}, \tilde{\theta}) = R(\hat{\theta}, \theta^*) - R(\tilde{\theta}, \theta^*)$

$$= \boxed{\left(R(\hat{\theta}, \theta^*) - R_n(\hat{\theta}, \theta^*)\right)} \rightarrow \text{r.v.}$$

$$+ \boxed{\left(R_n(\hat{\theta}, \theta^*) - R_n(\tilde{\theta}, \theta^*)\right)} \rightarrow \text{optimization} \leq 0$$

$$+ \boxed{\left(R_n(\tilde{\theta}, \theta^*) - R(\tilde{\theta}, \theta^*)\right)} \rightarrow \text{fixed Bound}$$

capacity of Hypothesis

need a uniform bound

$$\sup_{\theta \in \Theta} |\mathbb{P}_n(\ell_\theta) - \mathbb{P}(\ell_\theta)|$$

$= \boxed{\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}}$ where $\mathcal{F} = \{f : f = \ell(\theta) \quad \theta \in \Theta\}$

to solve this, VC-dim comes in

result like D-K-W

VC-dim attention

it focus on the question :

$$\begin{cases} \mathbb{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \in A\} \\ \mathbb{P}(A) = \mathbb{P}(X \in A) \end{cases} \rightarrow \boxed{\sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)|}$$

apply to Binary Classification ERM framework

'Toy' is because: linearity of (x, y) in estimator $\boxed{F(x) = F(X) \cdot Y}$

A Toy Example to explain: $\begin{cases} \text{Training error} \\ \text{Testing error} \end{cases}$

our problem consider the simple case that:
① fix x
② randomness comes from ε

3. Consider the fixed design nonparametric regression set up with observations $(x_1, Y_1), \ldots, (x_n, Y_n)$. Suppose further that the noise has constant variance, i.e. $\mathrm{Var}\{\epsilon_i\} = \sigma^2$. Both regressograms and Nadaraya-Watson kernel estimators are examples of *linear smoothers.* This means that the vector of predicted values $\hat{\mathbf{r}} = (\hat{r}_n(x_1), \ldots \hat{r}_n(x_n))$ is given by $\hat{\mathbf{r}} = \mathbf{L}\mathbf{y}$ where $\mathbf{y} = (Y_1, \ldots, Y_n)$. The training error can therefore be written as $\frac{1}{n}\|\mathbf{L}\mathbf{y} - \mathbf{y}\|_2^2$. What is the expected training error (your answer may contain $\mathrm{Trace}(\mathbf{L})$)? What is the average predictive risk, i.e. $\mathbb{E}\left\{\frac{1}{n}\|\mathbf{L}\mathbf{y} - \mathbf{y}^*\|_2^2\right\}$ where $\mathbf{y}^* = (Y_1^*, \ldots, Y_n^*)$ are new observations at each x_i (your answer may contain $\mathrm{Trace}(\mathbf{L})$)? The difference in the two values should be $2\sigma^2\mathrm{Trace}(\mathbf{L})/n$. In other words, this is the amount by which the training error is overly optimistic. *Hint: If M is any matrix, and ϵ a vector of independent random variables each with mean 0 and variance σ^2, then $\mathbb{E}\{\epsilon^T M\epsilon\} = Trace(M)\sigma^2$.*

Model: $\hat{y}_i = \hat{r}(x_i) + \varepsilon_i$

$$\begin{cases} \text{expected training error:} & \mathbb{E}_{\varepsilon}\left[\frac{1}{n}\left\|\left(L(\hat{r}+\varepsilon) - (\hat{r}+\varepsilon)\right\|_2^2\right] \\ \text{expected testing error:} & \mathbb{E}_{\varepsilon, \varepsilon^*}\left[\frac{1}{n}\left\| L\cdot(\hat{r}+\varepsilon) - (\hat{r}+\varepsilon^*)\right\|_2^2\right] \end{cases}$$