1 - Uniform Convergence ⇒ Generalization

- Supervised learning $(y,x) \sim p(y,x)$ iid pairs.
  $y \in \mathcal{Y} = \mathbb{R}$, $x \in \mathcal{X} = \mathbb{R}^d$

- Observe data: $(y_i, x_i) \sim p$ for $i = 1, 2, \dots, n$,
- Goal: Find a function $f \in \mathcal{F}$ s.t. $f : \mathcal{X} \to \mathcal{Y}$.
  $y \approx f(x)$.

  * Need to choose $\mathcal{F} = \{f\}$, the function class
    and the loss fnc $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$

- Goal+: Find $f_* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} R(f) = \underset{\text{over } p}{\mathbb{E}}\left[\ell((y,x), f)\right]$

* Empirical Risk Minimization (ERM)
  - Observe data $D = \{(y_i, x_i) : i = 1 \dots n\}$, then

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell((y_i, x_i), f)$$

estimates

are these close?

$$f_* \xleftarrow{\text{argmin}} R(f) = \mathbb{E}\left[\ell((y,x), f)\right]$$

Ex (MLE): $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, $\ell = -\log p_\theta(y|x)$

- $\hat{\theta} = \underset{\Theta}{\operatorname{argmin}} \hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(y_i|x_i)$
- $\theta_* = \underset{\Theta}{\operatorname{argmin}} R(\theta) = \mathbb{E}\left[-\log p_\theta(y|x)\right]$

$\hat{R}(\hat{\theta})$: training error
$R(\hat{\theta})$: test error
Def (Excess risk) : $R(\hat{\theta}) - R(\theta_*)$

Generalization = small excess risk

— Uniform Convergence

Goal: Understand generalization (= small excess risk), in a non-asymptotic sense.

$$\underbrace{\mathbb{P}\Big(\underbrace{R(\hat{f}) - R(f_*)}_{\text{excess risk}} > \epsilon \Big)}_{\text{bad event}} \leq \underset{\text{small prob}}{\delta}.$$

Def (Uniform conv.) : $\sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \xrightarrow[n\to\infty]{P} 0$

$$\hat{f} := \operatorname*{argmin}_{\mathcal{F}} \hat{R}(f) = \frac{1}{n} \underbrace{\sum_{i=1}^{n} \ell((y_i, x_i), f)}_{\text{iid-avg for fixed } f}$$

random variable → $R(\hat{f})$

$\hat{R}(\hat{f})$: non-iid avg.
$f_* = \operatorname*{argmin}_{\mathcal{F}} R(f)$: not random.

$$\underline{\epsilon \leq} R(\hat{f}) - R(f_*) = \underbrace{\left\{ R(\hat{f}) - \hat{R}(\hat{f}) \right\}}_{\substack{\text{hard to handle} \\ \text{since } \hat{f} \text{ is random} \\ \Rightarrow \text{non-iid avg.}}} + \underbrace{\left\{ \hat{R}(\hat{f}) - \hat{R}(f_*) \right\}}_{\substack{\leq 0 \\ \text{since } \hat{f} = \operatorname*{argmin}_{\mathcal{F}} \hat{R}(f)}} + \underbrace{\left\{ \hat{R}(f_*) - R(f_*) \right\}}_{\substack{\text{iid avg} \\ \Rightarrow \text{ez to handle}}}$$

$$\leq \underbrace{\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f)}_{\geq \epsilon/2} + 0 + \underbrace{\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f)}_{\geq \epsilon/2}$$

$$\left[\leq 2 \cdot \sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \quad \text{if } \xrightarrow{P} 0 \text{ then generalization.} \right]$$

$$\Rightarrow \mathbb{P}\left(R(\hat{f}) - R(f_*) \geq \epsilon\right) \leq \mathbb{P}\left(\left\{ \sup_{\mathcal{F}} R(f) - \hat{R}(f) \geq \frac{\epsilon}{2} \right\} \cup \left\{ \sup_{\mathcal{F}} \hat{R}(f) - R(f) \geq \frac{\epsilon}{2} \right\} \right)$$

(by union bound) $\leq \mathbb{P}\left(\sup_{\mathcal{F}} R(f) - \hat{R}(f) \geq \frac{\epsilon}{2} \right) + \mathbb{P}\left(\sup_{\mathcal{F}} \hat{R}(f) - R(f) \geq \frac{\epsilon}{2} \right)$

$\Rightarrow$ Need to bound RHS.

* It usually suffices to bound only one.
(by symmetry)

– Generalization for Finite Function Classes. ($|\mathcal{F}| < \infty$)
(warm-up)

Theorem: If $|\mathcal{F}| < \infty$ and $\ell \in [0,1]$, then

$$\mathbb{P}\left(R(\hat{f}) - R(f_*) \geq \sqrt{\underbrace{\frac{2}{n}}_{\text{sample size}} \Big(\underbrace{\log 2|\mathcal{F}|}_{\text{complexity of } \mathcal{F}} + \underbrace{\log \frac{1}{\delta}}_{\text{confidence level}} \Big)} \right) \leq \delta.$$

Remark: – Generalization error rate: $O\left(\sqrt{\frac{\log|\mathcal{F}|}{n}}\right)$

Proof: Strategy: 1- Concentration

2- Union bound

3- Generalization

* Lemma (Hoeffding's Ineq.): Let $z_1, z_2, \dots z_n$ be indep. r.v.'s such that $a_i \le z_i \le b_i$ almost surely.

Then, $\forall \epsilon > 0$ for $S_n = \frac{1}{n}\sum_{i=1}^{n} z_i$

1. $$\mathbb{P}(S_n - \mathbb{E}S_n \ge \epsilon) \le \exp\left\{-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}$$

2. $$\mathbb{P}(|S_n - \mathbb{E}S_n| \ge \epsilon) \le 2\exp\left\{-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}.$$

Note that $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell((y_i, x_i), f)$ is like

$$S_n = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad \underset{a_i}{0} \le \ell \le \underset{b_i}{1}$$

1- Concentration: Fix $f \in \mathcal{F}$, then

(by Hoeffding) $$\mathbb{P}\left(\hat{R}(f) - R(f) \ge \frac{\epsilon}{2}\right) \le \exp\left\{-\frac{2n^{\cancel{2}}\epsilon^2}{\underbrace{\sum_{i=1}^{n} 1}_{=n}}\right\}$$

$$= \exp\left\{-\frac{2n\epsilon^2}{4}\right\}$$

$$= \exp\left\{-\frac{n\epsilon^2}{2}\right\}$$

2- Uniform Convergence (via union bound):

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f) \geq \frac{\varepsilon}{2} \right) = \mathbb{P}\left(\bigcup_{f \in \mathcal{F}} \left\{ \hat{R}(f) - R(f) \geq \frac{\varepsilon}{2} \right\} \right)$$

(by union bound) $\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(\hat{R}(f) - R(f) \geq \varepsilon/2 \right)$

(by Hoeffding) $\leq |\mathcal{F}| \, e^{-n\varepsilon^2/2}$

3 — Generalization

$$\mathbb{P}\left(R(\hat{f}) - R(f_*) \geq \varepsilon \right) \leq \mathbb{P}\left(\sup_{\mathcal{F}} R(f) - \hat{R}(f) \geq \frac{\varepsilon}{2} \right) + \mathbb{P}\left(\sup_{\mathcal{F}} \hat{R}(f) - R(f) \geq \frac{\varepsilon}{2} \right)$$

$$\leq 2 \cdot |\mathcal{F}| \exp\left\{ -\frac{n\varepsilon^2}{2} \right\} := \delta$$

$$\Rightarrow \log \frac{\delta}{2|\mathcal{F}|} = -\frac{n\varepsilon^2}{2} \Rightarrow \varepsilon = \sqrt{\frac{2}{n} \log \frac{2|\mathcal{F}|}{\delta}}$$

Remarks:

1. Means: with prob $1-\delta$, $R(\hat{f}) - R(f_*) \leq O\left(\sqrt{\frac{\log|\mathcal{F}| + \log \delta^{-1}}{n}} \right)$
2. Choose $\delta = O(|\mathcal{F}|^{-1})$
   Conv. rate becomes $O\left(\sqrt{\frac{\log|\mathcal{F}|}{n}} \right)$
3. Covers bounded loss, cannot cover square loss.
4. Bound fails when $|\mathcal{F}| = \infty$!