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.
  * ——— the loss $\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$

- **Goal+:** Find $f_* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} R(f) = \mathbb{E}\left[\ell\left((y,x), f\right)\right]$ ($\ell(y, f(x))$)
  ↑ over $(y,x) \sim p$
  cannot!

  * Empirical Risk Minimization (ERM)
    - Observe data $\mathcal{D} = \{(y_i, x_i) : i = 1 \dots n\}$, then
      $\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell((y_i, x_i), f)$
      hope ☺ ↓ estimator
      $f_* \leftarrow \operatorname{argmin}$ — $R(f) = \mathbb{E}[\ell((y,x), f)]$

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

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

$\hat{R}(\hat{\theta})$: 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.

$$\mathbb{P}\Big(\underbrace{R(\hat{f}) - R(f_*)}_{\text{excess risk}} > \epsilon \Big) \le \delta \quad (\delta \downarrow \text{small prob})$$

bad event.

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

$\hat{f} = \arg\min_{\mathcal{F}} \hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, z_i), f)$

iid-avg for fixed f → random variable → $R(\hat{f})$

$\hat{R}(\hat{f})$: non-iid avg.

$f_* = \arg\min_{\mathcal{F}} R(f)$: not random.

$$R(\hat{f}) - R(f_*) = \{R(\hat{f}) - \hat{R}(\hat{f})\} + \{\hat{R}(\hat{f}) - \hat{R}(f_*)\} + \{\hat{R}(f_*) - R(f_*)\}$$

$\{\hat{R}(\hat{f}) - \hat{R}(f_*)\} \le 0$ since $\hat{f} = \arg\min_{\mathcal{F}} \hat{R}(f)$

hard to handle ($\hat{f}$ is random – non-iid avg –)

ez to handle as f_* is fixed – iid avg –

$$\le \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| + 0 + \sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)|$$

$$\le 2 \cdot \sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)|$$

If $\xrightarrow[P]{} 0$ (uniform convergence)

we have small excess risk (generalization)

$$\underbrace{\left\{ R(\hat{f}) - R(f_*) \geq \epsilon \right\}}_{A} \subseteq \underbrace{\left\{ \sup_{f \in \mathcal{F}} \left| \hat{R}(f) - R(f) \right| \geq \frac{\epsilon}{2} \right\}}_{B}$$

(a) (b)

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

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

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

$$\mathbb{P}\left(R(\hat{f}) - R(f_*) \geq B\sqrt{\frac{2}{n}\left(\log 2|\mathcal{F}| + \log \frac{1}{\delta} \right)} \right) \leq \delta.$$

sample size — complexity of $\mathcal{F}$ — confidence level

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

Proof:

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

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

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

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

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

$S_n = \frac{1}{n}\sum_{i=1}^{n} z_i$ $\qquad$ $0 \leq \ell \leq B$

a_i $\qquad$ b_i

Strategy: 1- Concentration
2- Union bound
3- Generalization

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

$$\mathbb{P}\left(|\hat{R}(f) - R(f)| \geq \frac{\epsilon}{2}\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{\sum_{i=1}^{n} B^2 4}\right\}$$

(by Hoeffding)

$$= nB^2$$

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

Remark: looks exponential in n, infact it implies $\frac{1}{\sqrt{n}}$ rate.

2- Uniform Convergence (via union bound)

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

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

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

by Hoeffding $\leq 2|\mathcal{F}|\exp\left\{-\frac{n\epsilon^2}{2B^2}\right\}$

3 – Generalization

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

$$\leq 2 |\mathcal{F}| \exp\left\{ - \frac{n\epsilon^2}{2B^2} \right\} := \delta.$$

$$\Rightarrow \log \frac{\delta}{2|\mathcal{F}|} = - \frac{n\epsilon^2}{2B^2} \quad \Rightarrow \epsilon = B\sqrt{\frac{2}{n} \log \frac{2|\mathcal{F}|}{\delta}}. \quad \blacksquare$$

Remarks: 1. Means: with prob $1-\delta$, $R(\hat{f}) - R(f_*) \leq B\sqrt{\frac{2}{n} \log \frac{2|\mathcal{F}|}{\delta}}$

$$= O\left(\sqrt{\frac{\log|\mathcal{F}| + \log \frac{1}{\delta}}{n}} \right)$$

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 $|\mathcal{F}| = \infty$!