3 - Symmetrization

Last time:

- Generalization by
  1- concentration
  2- discretization + union bound
  3- uniform convergence $\Rightarrow$ generalization

Today: A new technique to replace step 2, namely symmetrization.

- Slight modification uniform conv $\Rightarrow$ generalization

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

which was due to

$$\epsilon \leq R(\hat{f}) - R(f_*) \leq R(f) - \hat{R}(f) + 0 + \hat{R}(f_*) - R(f_*)$$
$$\leq \underbrace{\sup_{\mathcal{F}} R(\hat{f}) - \hat{R}(\hat{f})}_{\leq \frac{\epsilon}{2}} + \underbrace{\sup_{\mathcal{F}} \hat{R}(f) - R(f)}_{\leq \frac{\epsilon}{2}}$$

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

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

(by symmetry) $$= 2\, \mathbb{P}\left(\sup_{\mathcal{F}} \underbrace{\hat{R}(f) - R(f)}_{\text{empirical process}} \geq \frac{\epsilon}{2} \right)$$

$\Rightarrow$ Need to bound RHS.

Theorem (Generalization by RC): For a fnc class $\mathcal{F}$, define $\mathcal{G} = \{ (x,y) \to \ell((x,y), f) : f \in \mathcal{F} \}$. If the loss fnc satisfies $0 \le \ell \le B$, then with prob at least $1-\delta$, we have

$$R(\hat{f}) - R(f_*) \le 4 \mathcal{R}(\mathcal{G}) + B\sqrt{\frac{2 \log 2/\delta}{n}}.$$

Remarks:
- RC is a complexity measure of a fnc class.
- Rate depends on $\mathcal{R}(\mathcal{G})$.
  - $g_f \in \mathcal{G}$ depends on $f \in \mathcal{F}$. We expect $\mathcal{R}(\mathcal{G}) \approx \mathcal{R}(\mathcal{F})$?
  - We hope, as $n \uparrow$ $\mathcal{R}(\mathcal{G}) \downarrow$.

proof:

Strategy:
1. Concentration (~~Hoeffding~~, McDiarmid's)
2. ~~Discretization + union bound~~ Symmetrization
3. Unif conv. $\Rightarrow$ generalization

Goal: Bound the empirical process.

Step 1: Concentration

Theorem (McDiarmid's Inequality): Let g be a function satisfying the "bounded difference" property,

$$\forall j \in [n] \quad |g(x_1, \dots, x_j, \dots, x_n) - g(x_1, \dots, x_j', \dots, x_n)| \le c_j.$$

Then, for $z_1, z_2, \dots z_n$ independent r.v.'s

$$\mathbb{P}\left(g(z_1, \dots, z_n) - \mathbb{E} g(z_1, \dots, z_n) \ge \epsilon \right) \le \exp\left\{ - \frac{2\epsilon^2}{\sum_{j=1}^{n} c_j^2} \right\}.$$

Example (Hoeffding's Inequality): $z_1 \dots z_n$ independent and $\forall i \quad a_i \le z_i \le b_i$.

$$g(z_1, \dots z_n) = \frac{1}{n}\sum_{i=1}^{n} z_i$$

$$|g(z_1, \dots z_j, \dots z_n) - g(z_1 \dots z_j', \dots z_n)| = \left| \cancel{\frac{1}{n}\sum_{i=1}^{n} z_i} - \cancel{\frac{1}{n}\sum_{i=1}^{n} z_i} - \frac{1}{n}(z_j' - z_j)\right|$$

$$= \frac{1}{n}|z_j - z_j'| \le \frac{b_j - a_j}{n} := c_j$$

By McDiarmid's inequality

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

Goal: Bound $\sup_{\mathcal{F}} \hat{R}(f) - R(f)$

$$g(z_1, \dots, z_n) = \sup_{\mathcal{F}} \hat{R}(f) - R(f)$$

$$= \sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(\underbrace{(x_i, y_i)}_{z_i}, f) - \mathbb{E}\,\ell(\underbrace{(x,y)}_{z}, f)$$

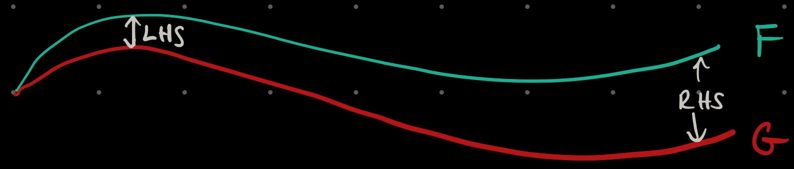
$$= \sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, f) - \mathbb{E}\,\ell(z, f).$$

$$|g(\dots z_j \dots) - g(\dots z_j' \dots)|$$

$$* = \Bigg| \sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, f) - \mathbb{E}\,\ell(z,f)$$

$$- \left\{\sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, f) - \mathbb{E}\,\ell(z,f) - \frac{\ell(z_j, f) - \ell(z_j', f)}{n}\right\}\Bigg|$$

Fact: $\left| \sup_x F(x) - \sup_x G(x) \right| \le \sup_x \left| F(x) - G(x) \right|$



$$* \le \sup_{\mathcal{F}} \left| \frac{\ell(z_j, f) - \ell(z_j', f)}{n} \right| \le \frac{B}{n} := c_j .$$

(Since $0 \le \ell \le B$)

By McDiarmid's Inequality:

$$\mathbb{P}\Big(\underbrace{\sup_{\mathcal{F}} \hat{R}(f) - R(f)}_{\text{what we want}} - \underbrace{\mathbb{E}\Big[\sup_{\mathcal{F}} \hat{R}(f) - R(f)\Big]}_{\text{Need to show it is small.}} \ge t \Big) \le \exp\left\{ - \frac{2nt^2}{B^2} \right\}$$

Step 1: Symmetrization (To bound $\uparrow$)

- Basic idea: X is a r.v. and X' is its iid copy.

$\Rightarrow \quad X \overset{d}{=} X'$ let g be any fnc.

$$\Rightarrow \quad g(X) - g(X') \overset{d}{=} g(X') - g(X)$$
$$\overset{d}{=} -1\,(g(X) - g(X'))$$
$$\overset{d}{=} \sigma\,(g(X) - g(X'))$$

where σ is a Rademacher r.v.: $\mathbb{P}(\sigma = +1) = \frac{1}{2}$.
$\mathbb{P}(\sigma = -1) = \frac{1}{2}$.

- In our case, data $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is r.v.

$$= \{z_1, \ldots\ldots, z_n\}$$

- Introduce iid copy of the dataset (ghost data)

$$D' = \{ z_1', \dots, z_n' \}$$

z_i's and z_i''s are iid.

- Now, we have 2 empirical risks, 1 population risk.

i) $\hat{R}(f;D) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f)$

ii) $\hat{R}(f;D') = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i', f)$

$\mathbb{E}\hat{R}(f;D) = \mathbb{E}\hat{R}(f;D') = R(f)$.

- Notice that $R(f) = \mathbb{E}\left[\hat{R}(f;D)\right] = \mathbb{E}\left[\hat{R}(f;D') \mid D\right]$.

Goal: Bound $\mathbb{E}\left[\sup_{\mathcal{F}} \hat{R}(f) - R(f)\right]$.

$$\mathbb{E}\left[\sup_{\mathcal{F}} \hat{R}(f;D) - R(f)\right] = \mathbb{E}\left[\sup_{\mathcal{F}} \left\{\hat{R}(f;D) - \mathbb{E}\left[\hat{R}(f;D')\right]\right\}\right]$$

$$= \mathbb{E}\left[\sup_{\mathcal{F}} \left\{\hat{R}(f;D) - \mathbb{E}\left[\hat{R}(f;D') \mid D\right]\right\}\right]$$

$$= \mathbb{E}\left[\sup_{\mathcal{F}} \mathbb{E}\left[\hat{R}(f;D) - \hat{R}(f;D') \mid D\right]\right]$$

$\{$by $\sup \mathbb{E} \le \mathbb{E} \sup\}$ $\le \mathbb{E}\left[\mathbb{E}\left[\sup_{\mathcal{F}} \hat{R}(f;D) - \hat{R}(f;D') \mid D\right]\right]$

$\left\{\begin{array}{l}\text{by tower property} \\ \mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]\end{array}\right\}$ $= \mathbb{E}\left[\sup_{\mathcal{F}} \hat{R}(f;D) - \hat{R}(f;D')\right]$

$$= \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(z_i, f) - \ell(z_i', f)}_{\stackrel{d}{=}\ \sigma_i \{\ell(z_i, f) - \ell(z_i', f)\}}\right]$$

$$= \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left\{ \ell(z_i, f) - \ell(z_i', f) \right\} \right]$$

Fact: $\sup_{x} \{ F(x) + G(x) \}$
$\leq \sup_{x} F(x) + \sup_{x} G(x)$

$$\leq \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(z_i, f) \right] + \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} -\sigma_i \ell(z_i', f) \right]$$

$\sigma_i \overset{d}{=} -\sigma_i$

$$= 2\, \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(z_i, f) \right] = **$$

Definition (Rademacher Complexity): For a fnc class $\mathcal{F} = \{ f : \mathcal{Z} \longrightarrow \mathbb{R} \}$ and a dataset $D = \{ z_1 \dots z_n \}$,

* RC is defined as

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right]$$

where σ_i's are iid Rademacher r.v.'s.

* Empirical RC is defined as

$$\hat{\mathcal{R}}(\mathcal{F}) = \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \,\middle|\, z_{1:n} \right]$$

$$** = 2\, \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(z_i, f) \right] \leq 2 \mathcal{R}(\cancel{\mathcal{F}})$$

no!

$$\leq 2 \mathcal{R}(\mathcal{G})$$

where $\mathcal{G} = \{ z \rightarrow \ell(z, f) : f \in \mathcal{F} \}$.

Step 3 : Uniform convergence ⇒ generalization

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

- By Step 1: $* \ \mathbb{P}\left(\sup_{\mathcal{F}} \hat{R}(f) - R(f) \geq \mathbb{E}\left[\sup_{\mathcal{F}} \hat{R}(f) - R(f)\right] + t \right) \leq e^{-\frac{2nt^2}{B^2}}$

- By Step 2: $* \ \mathbb{E}\left[\sup_{\mathcal{F}} \hat{R}(f) - R(f) \right] \leq 2 \cdot \mathcal{R}(\mathcal{G})$.

- By Steps 1 and 2:

$$2\,\mathbb{P}\left(\sup_{\mathcal{F}} \hat{R}(f) - R(f) \geq \underbrace{t + 2\mathcal{R}(\mathcal{G})}_{\epsilon/2} \right) \leq \underbrace{2 e^{-\frac{2nt^2}{B^2}}}_{:= \delta}$$

$$\Rightarrow \delta = 2 e^{-\frac{2nt^2}{B^2}} \implies t = B\sqrt{\frac{\log 2/\delta}{2n}}$$

$$\Rightarrow \frac{\epsilon}{2} := t + 2\mathcal{R}(\mathcal{G}) \Rightarrow \epsilon = 4\mathcal{R}(\mathcal{G}) + B\sqrt{\frac{2\log 2/\delta}{n}}$$

∎