# 2 - Uniform Convergence

Today: A non-trivial setting where we use uniform conv.

- Recall the bound on excess risk:

$$\Rightarrow \underbrace{\mathbb{P}\left(R(\hat{f}) - R(f_*) \geq \epsilon\right)}_{\text{excess risk}} \leq \mathbb{P}\left(\left\{\sup_{\mathcal{F}} R(f) - \hat{R}(f) \geq \tfrac{\epsilon}{2}\right\} \cup \left\{\sup_{\mathcal{F}} \hat{R}(f) - R(f) \geq \tfrac{\epsilon}{2}\right\}\right)$$

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

- We'll bound $\mathbb{P}\left(\sup_{\mathcal{F}} \underbrace{\hat{R}(f) - R(f)}_{\text{empirical process}} \geq \tfrac{\epsilon}{2}\right)$ which will imply a bound on the first term by symmetry.

**Theorem** (Rademacher Complexity): Define $\mathcal{G} = \left\{(x,y) \to \ell((x,y), f) : f \in \mathcal{F}\right\}$. If loss satisfies $0 \leq \ell \leq 1$, then with probability at least $1-\delta$,

$$R(\hat{f}) - R(f_*) \leq 4 \underbrace{\mathcal{R}(\mathcal{G})}_{\text{Rademacher Complexity (RC)}} + \sqrt{\frac{2 \log 2/\delta}{n}}.$$

**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- ~~union bound~~ Symmetrization
3- Unif conv. $\Rightarrow$ generalization

Goal: Bound the empirical process.

## Step 1: Concentration

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

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

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

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

Application: (Hoeffding's Inequality)

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

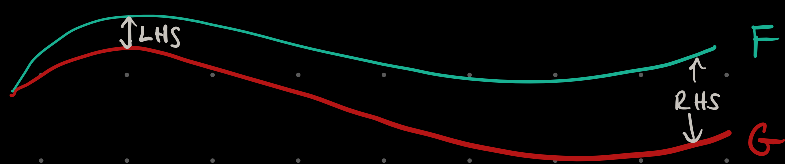
$$g(\underline{z_1},\dots,\underline{z_n}) = \sup_{\mathcal{F}} \hat{R}(f) - R(f)$$

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

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

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

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




$$* \leq \sup_{\mathcal{F}} \left| \frac{\ell(z_j,f) - \ell(z_j',f)}{n} \right| \leq \frac{1}{n} := c_j.$$

(Since $0 \leq \ell \leq 1$)

By McDiarmid's Inequality:

$$\mathbb{P}\Big( \underbrace{\sup_{\mathcal{F}} \hat{R}(f) - R(f)}_{} - \mathbb{E}\big[\sup_{\mathcal{F}} \hat{R}(f) - R(f)\big] \geq t \Big) \leq \exp\left\{-2nt^2\right\}$$

what we want

Need to show it is small.

Step 2: Symmetrization (To bound ↑)

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

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

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

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

- In our case, the data $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is r.v.
$= \{z_1, \ldots, z_n\}$

- Introduce iid copy of the dataset $D' = \{z_1', \ldots, z_n'\}$

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

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

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

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

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

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

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

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

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

$$\left\{\begin{array}{l}\text{Fact: } \sup_{x}\{F(x)+G(x)\} \\ \leq \sup_{x} F(x) + \sup_{x} G(x)\end{array}\right\}$$

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

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

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

**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

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

where $\sigma_i$'s are iid Rademacher r.v.'s.

* Empirical RC is defined as

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

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

no!

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

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

Step 3: Uniform convergence $\Rightarrow$ 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: $* \; \mathbb{P}\left( \sup_{\mathcal{F}} \hat{R}(f) - R(f) \geq \mathbb{E}\left[\sup_{\mathcal{F}} \hat{R}(f) - R(f)\right] + t \right) \leq e^{-2nt^2}$

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

- By Steps 1 and 2:

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

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

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

## — Generalization via RC

Goal: i) Relate $\mathcal{R}(\mathcal{G})$ to $\mathcal{R}(\mathcal{F})$.

ii) $\mathcal{R}(\mathcal{F})$ decays w/ n.

i) – **Theorem** (Talagrand's Contraction Principle): Let $g$ be a $L$-Lipschitz cont. func. and $g \circ \mathcal{F} = \{g \circ f : f \in \mathcal{F}\}$, then

$$\mathcal{R}(g \circ \mathcal{F}) \leq L \cdot \mathcal{R}(\mathcal{F}).$$

## – RC of Constrained Linear Models

ii) – Goal: $\mathcal{R}(\mathcal{F}) = O(1/\sqrt{n})$.

**Theorem** (RC of Linear Models): Let $\mathcal{F} = \{f(x) = \langle x, \theta\rangle : \|\theta\| \leq r\}$.

Then, i) $\hat{\mathcal{R}}(\mathcal{F}) \leq \frac{r}{n}\sqrt{\sum_{i=1}^{n} \|x_i\|^2}$

ii) If $\mathbb{E}[\|x_i\|^2] \leq k^2$, then $\mathcal{R}(\mathcal{F}) \leq \frac{r \cdot k}{\sqrt{n}}$

Remarks: 1- If we combine this bound w/ previous examples, we achieve generalization.

2 - Notice $K = O(\sqrt{d})$ so $\mathcal{R}(\mathcal{F}) \leq r\sqrt{\frac{d}{n}}$.

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

$$= \mathbb{E}\left[\sup_{\|\theta\|\leq r} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \langle x_i, \theta\rangle \,\middle|\, x_{1:n}\right]$$

$$= \mathbb{E}\left[\sup_{\|\theta\|\leq r} \left\langle \frac{1}{n}\sum_{i=1}^{n}\sigma_i x_i , \theta\right\rangle \,\middle|\, x_{1:n}\right]$$

$$\boxed{\sup_{\|\theta\|\leq r} \langle \theta, v\rangle = r\cdot\|v\|}$$

$$= r\,\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^{n}\sigma_i x_i\right\| \,\middle|\, x_{1:n}\right]$$

(by Jensen's Ineq.)

$$\leq \frac{r}{n}\,\mathbb{E}\left[\left\| \sum_{i=1}^{n}\sigma_i x_i\right\|^2 \,\middle|\, x_{1:n}\right]^{\frac{1}{2}}$$

$$= \frac{r}{n}\,\mathbb{E}\left[\sum_{i=1}^{n}\|x_i\|^2 + \sum_{i\neq j}\sigma_i\sigma_j\langle x_i, x_j\rangle \,\middle|\, x_{1:n}\right]^{\frac{1}{2}}$$

$$= \frac{r}{n}\sqrt{\sum_{i=1}^{n}\|x_i\|^2}$$

ii) $\mathcal{R}(\mathcal{F}) = \mathbb{E}\hat{\mathcal{R}}(\mathcal{F}) \leq \frac{r}{n}\,\mathbb{E}\sqrt{\sum_{i=1}^{n}\|x_i\|^2}$

(by Jensen's Ineq)

$$\leq \frac{r}{n}\sqrt{\mathbb{E}\sum_{i=1}^{n}\|x_i\|^2}$$

$$\leq \frac{r}{n}\sqrt{\sum_{i=1}^{n}\mathbb{E}[\|x_i\|^2]}$$

$$\leq \frac{r\cdot K}{\sqrt{n}}.$$

$\blacksquare$