4 - Rademacher Complexity: Properties and Applications

Recall: z_i are iid data points, σ_i iid Rademacher r.v.'s

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

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

- Some properties of RC:
  - Monotonicity: $\mathcal{F}_1 \subseteq \mathcal{F}_2$, $\mathcal{R}(\mathcal{F}_1) \leq \mathcal{R}(\mathcal{F}_2)$
  - Linear combination: $\mathcal{F}_1 + \mathcal{F}_2 = \{ f_1 + f_2 : f_1 \in \mathcal{F}_1 \text{ and } f_2 \in \mathcal{F}_2 \}$
    $$\mathcal{R}(\mathcal{F}_1 + \mathcal{F}_2) = \mathcal{R}(\mathcal{F}_1) + \mathcal{R}(\mathcal{F}_2).$$
  - Scaling w/ scalar: $\mathcal{R}(\mathcal{F} \cdot c) = |c| \cdot \mathcal{R}(\mathcal{F})$.
  - Convex-hull: For $\mathcal{F}$ finite
    $$\mathcal{R}(\text{convex hull}(\mathcal{F})) = \mathcal{R}(\mathcal{F}).$$

Recall:

Theorem (Generalization due to RC): Loss is bounded by B, and $\mathcal{G} = \{ g(z) = \ell(z, f) \;\; f \in \mathcal{F} \}$, then w.p $1-\delta$,

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \leq 4 \cdot \mathcal{R}(\mathcal{G}) + B\sqrt{\frac{2 \log 2/\delta}{n}}$$

Also, $\sup_{\mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)| \leq 2\mathcal{R}(\mathcal{G}) + B\sqrt{\frac{2 \log 2/\delta}{n}}$.

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

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

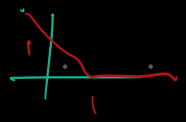
ii) $R(\mathcal{F})$ decays w/ n.

Ex (SVMs): - Data $z = (y, x)$ where $y \in \{-1, +1\}$, $f \in \mathcal{F}$

- Hinge loss: $\ell(z, f) = \max\{0, 1 - yf(x)\}$

$\phi : s \longrightarrow \max\{0, 1-s\}$ is 1-Lip.

* RC of less class



$$R(\mathcal{G}) \quad \text{where} \quad \mathcal{G} = \{ z = (y,x) \longrightarrow \ell(z,f) \quad f \in \mathcal{F} \}$$

$$= \{ z \longrightarrow \phi(y \cdot f(x)) \quad f \in \mathcal{F} \}$$

Define the intermediate fnc class

$$\mathcal{H} = \{ z = (y,x) \longrightarrow y \cdot f(x) \quad f \in \mathcal{F} \}$$

$$\Rightarrow \mathcal{G} = \phi \circ \mathcal{H} \quad \text{where } \phi \text{ is 1-Lip.}$$

(by Talagrand) $\Rightarrow R(\mathcal{G}) \le R(\mathcal{H}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\sigma_i y_i} f(x_i) \right]$

$\overset{d}{=} \sigma_i \perp\!\!\!\perp x_i$

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

$$= R(\mathcal{F})$$

Claim: $\sigma_i y_i \overset{d}{=} \sigma_i \perp\!\!\!\perp x_i$

proof: For any A,

$\mathbb{P}(x_i \in A, \sigma_i y_i = 1)$

$= \mathbb{P}(x_i \in A, \sigma_i = 1, y_i = 1) + \mathbb{P}(x_i \in A, \sigma_i = -1, y_i = -1)$

$= \mathbb{P}(x_i \in A, \sigma_i = 1, y_i = 1) + \mathbb{P}(x_i \in A, \sigma_i = 1, y_i = -1)$

$= \mathbb{P}(x_i \in A, \sigma_i = 1) = \mathbb{P}(x_i \in A)\,\mathbb{P}(\sigma_i = 1)$

$= \mathbb{P}(x_i \in A)\,\mathbb{P}(\sigma_i y_i = 1)$

* If we have RC of $\mathcal{F}$, we can characterize the generalization error of SVMs.

Ex (Misclassification error):

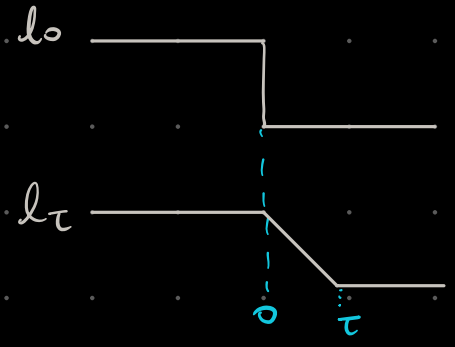
- Data $z=(y,x)$ where $y \in \{\pm 1\}$, $f \in \mathcal{F}$
- 0-1 loss: $\ell_0(z,f) = 1\{y \neq f(x)\} = 1\{y \cdot f(x) \leq 0\}$

if $f: \mathbb{R}^d \to \{\pm 1\}$

$$\ell_0(y \cdot f(x)) = 1\{y \cdot f(x) \leq 0\}$$

where $\ell_0(s) = \begin{cases} 1 & s \leq 0 \\ 0 & s > 0 \end{cases}$.

- ℓ_0 is not Lipschitz, we need surrogate losses (i.e. Hinge)



$$\ell_\tau(s) = \begin{cases} 1 & s < 0 \\ 1 - s/\tau & \text{otw} \\ 0 & s \geq \tau \end{cases}$$

ℓ_τ is $\frac{1}{\tau}$-Lip cont.

- What is the generalization error of $\hat{f}_\tau = \operatorname{argmin}_{\mathcal{F}} \hat{R}_\tau(f)$?

Assmp: $\mathbb{P}(0 \leq y \cdot f(x) \leq \tau) \leq C \cdot \tau$ for small τ.

Ex: $y = \operatorname{sign}(x)$ $x \sim N(0,1)$ $\Rightarrow \mathbb{P}(0 \leq y \cdot \operatorname{sign}(x) \leq \tau) = \sqrt{\frac{2}{\pi}} \cdot \tau + O(\tau^2)$

- Define $f_\tau^* = \operatorname{argmin}_{\mathcal{F}} R_\tau(f)$ (f_0^* minimizes $R_0(f)$).

- By Thm on RC, w.p at least $1-\delta$

$$(*) \quad R_\tau(\hat{f}_\tau) \leq R_\tau(f_\tau^*) + 4R(\mathcal{G}) + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}.$$

where $\mathcal{G} = \{z \to \ell_\tau(z,f) \ f \in \mathcal{F}\}$.

(ii) $\Rightarrow$ By Talagrand's Contraction: $\mathcal{R}(\mathcal{G}) \le \frac{1}{\tau} \cdot \mathcal{R}(\mathcal{F})$

(iii) $\ell_\tau \ge \ell_0 \Rightarrow R_\tau(f) \ge R_0(f) \Rightarrow R_\tau(\hat{f}_\tau) \ge R_0(\hat{f}_\tau)$

(iv) $|R_\tau(f) - R_0(f)| \le \mathbb{P}(0 \le Y \cdot f(x) \le \tau) \le C \cdot \tau.$

$\Rightarrow R_\tau(f_\tau^*) \le R_\tau(f_0^*) \le R_0(f_0^*) + C \cdot \tau.$

Then, by (i-iv), w.p $1-\delta$

$$\boxed{R_0(\hat{f}_\tau) - R_0(f_0^*) \le C \cdot \tau + \frac{4}{\tau}\mathcal{R}(\mathcal{F}) + \sqrt{\frac{2\log 2/\delta}{n}}}$$

- Typically, $\mathcal{R}(\mathcal{F}) = O\left(\frac{1}{\sqrt{n}}\right)$. (this is next)

$$R_0(\hat{f}_\tau) - R_0(f_0^*) \lesssim \tau + \frac{1}{\tau\sqrt{n}}$$

By optimizing over τ, $\lesssim \frac{1}{n^{1/4}}$ if $\tau = O\left(\frac{1}{n^{1/4}}\right)$

- We get generalization error $\le O(n^{-1/4})$ by solving the relaxation.

- RC of Constrained Linear Models
  - Goal: $\mathcal{R}(\mathcal{F}) = O\left(1/\sqrt{n}\right)$.

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

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

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

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

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

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

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

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

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

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

(by Jensen's Ineq.) $$\le \frac{r}{n}\,\mathbb{E}\left[\left\|\sum_{i=1}^n \sigma_i x_i\right\|^2 \mid x_{1:n}\right]^{\frac{1}{2}}$$

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

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

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

(by Jensen's Ineq) $$\le \frac{r}{n}\sqrt{\mathbb{E}\sum_{i=1}^n \|x_i\|^2}$$

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

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

∎

– RC of Finite Fnc Classes

Thm (Massart's Finite Lemma): Let $z_1, \dots z_n$ are iid and $\mathcal{F}$ is finite. If $\sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(z_i)^2 \le K^2$ a.s., then

$$\hat{R}(\mathcal{F}) \le K\sqrt{\frac{2\log|\mathcal{F}|}{n}}.$$

proof: σ is a Rademacher r.v. w/ MGF $M_\sigma(t) = \mathbb{E}e^{\sigma t} = \frac{e^t + e^{-t}}{2}$
$= \cosh(t)$.

For $t \ge 0$, $\exp\{t\hat{R}(\mathcal{F})\} = \exp\{t\,\mathbb{E}[\sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i) \mid z_{1:n}]\}$

$z \to e^{tz}$ is cvx. By Jensen's $\le \mathbb{E}\left[\exp\{t \sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\} \mid z_{1:n}\right]$

$z \to e^{tz}$ is monotone $= \mathbb{E}\left[\sup_{\mathcal{F}} \exp\{t \cdot \frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\} \mid z_{1:n}\right]$

$\mathcal{F}$ is finite: $\sup_{\mathcal{F}} \le \sum_{\mathcal{F}}$ $\le \mathbb{E}\left[\sum_{f\in\mathcal{F}} \exp\{\frac{t}{n}\sum_{i=1}^n \sigma_i f(z_i)\} \mid z_{1:n}\right]$

$$= \sum_{f\in\mathcal{F}} \prod_{i=1}^n \underbrace{M_{\sigma_i}\left(\tfrac{t}{n} f(z_i)\right)}_{=\cosh} \mid z_{1:n}$$

Exercise: $\frac{x^2}{2} \ge \log\cosh(x)$

$\Rightarrow M_\sigma(t) \le \exp\{t^2/2\}$

$$\le \sum_{f\in\mathcal{F}} \prod_{i=1}^n \exp\left\{\frac{t^2}{2n^2} f(z_i)^2\right\}$$

$$= \sum_{f\in\mathcal{F}} \exp\left\{\frac{t^2}{2n} \cdot \frac{1}{n}\sum_{i=1}^n f(z_i)^2\right\}$$

$\le K^2$

$$\exp\{t\hat{R}(\mathcal{F})\} \le |\mathcal{F}| \exp\left\{\frac{t^2}{2n} K^2\right\}$$

Take log, divide by t $\Rightarrow$ $\hat{\mathcal{R}}(\mathcal{F}) \le \frac{\log|\mathcal{F}|}{t} + \frac{tK^2}{2n}$

Optimize over t $\left(t = K^{-1}\sqrt{2\log|\mathcal{F}| \cdot n}\right)$ $\Rightarrow$ $\hat{\mathcal{R}}(\mathcal{F}) \le K\sqrt{\frac{2\log|\mathcal{F}|}{n}}$ ☒

Remarks: Recall that w.p. $1-\delta$,

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

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

ℓ is bdd by B $\Rightarrow$ $\sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(z_i,f)^2 \le B^2 := K^2$

$\Rightarrow$ By MFL

$$\hat{\mathcal{R}}(\mathcal{G}) \le B\sqrt{\frac{2\log|\mathcal{F}|}{n}} \qquad (\text{since } |\mathcal{G}| \le |\mathcal{F}|)$$

$\Rightarrow$ Generalization bound: w.p $1-\delta$

- by MFL + RC: $4B\sqrt{\frac{2\log|\mathcal{F}|}{n}} + B\sqrt{\frac{2\log 2/\delta}{n}}$
- by union bound: $B\sqrt{\frac{2\log|\mathcal{F}|}{n} + \frac{2\log 1/\delta}{n}}$

The same order.