5 - Combinatorial Measures of Complexity

Recap: - Rademacher complexity based generalization bound

w.p. at least $1-\delta$,

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

- If we bound RC of $\mathcal{G} = \{z \to \ell(z,f) \quad f \in \mathcal{F}\}$

we get a generalization bound.

Eg: Massart's Finite Lemma (MFL): If $\sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(z_i)^2 \leq k^2$

then, $\hat{\mathcal{R}}(\mathcal{F}) \leq k\sqrt{\frac{2\log|\mathcal{F}|}{n}} \longrightarrow |\mathcal{F}| < \infty.$

- $\hat{\mathcal{R}}(\mathcal{G}) \leq B\sqrt{\frac{2\log|\mathcal{G}|}{n}} \leq B\sqrt{\frac{2\log|\mathcal{F}|}{n}}$

since $|\mathcal{G}| \leq |\mathcal{F}|$.

$\Rightarrow$ Generalization:

RC + MFL : excess risk $\leq 4B\sqrt{\frac{2\log|\mathcal{F}|}{n}} + B\sqrt{\frac{2\log 2/\delta}{n}}$

- Shattering Coefficient

! When converting $\sup_{\mathcal{F}} \longrightarrow \sum_{\mathcal{F}}$ (last lecture)

$$\mathbb{E} \sup_{\mathcal{F}} \exp\left\{t \cdot \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\} \leq \mathbb{E}\sum_{\mathcal{F}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\} \Bigg|_{z_{1:n}}$$

- f enters this bound only through $[f(z_1) \cdots f(z_n)]$.
- $\mathcal{F}$ can be infinitely large, but $\{[f(z_1) \cdots f(z_n)]\}$ can be still small.

Ex: Data $z_i \in \mathcal{Z}$ and $\mathcal{F} = \{ z \to \sin(z \cdot \pi k) : k \in \mathbb{N} \}$

$\Rightarrow |\mathcal{F}| = \infty$, $f(z_i) = 0$

$\Rightarrow [f(z_1) \cdots f(z_n)] = [0 \cdots 0] \quad \forall f \in \mathcal{F}, \ \forall z_i \in \mathcal{Z}$

Ex: 0-1 loss : $\mathcal{G} = \{ g : z \longrightarrow \ell(z, f) : f \in \mathcal{F} \}$

$g(z_i) \in \{0, 1\}$

$$\left| \{ [g(z_1) \cdots g(z_n)] : g \in \mathcal{G} \} \right| \leq 2^n < \infty .$$

$[0 \ 0 \ \cdots \ 0]$

$[0 \ 1 \ 0 \cdots \ 1]$

$\vdots$

(we focus on 0-1 loss).

Continue from the step !

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

$$= \mathbb{E}\left[\sup_{[f_1 \cdots f_n] \in F} \exp\left\{ t \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_i \right\} \,\middle|\, z_{1:n} \right]$$

$[f_1 \cdots f_n] \in F = \{ [f(z_1) \cdots f(z_n)] : f \in \mathcal{F} \}$

- z_i's are fixed
- we vary $f \in \mathcal{F}$.

$$\leq \mathbb{E}\left[\sum_{[f_1 \cdots f_n] \in F} \exp\left\{ t \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_i \right\} \,\middle|\, z_{1:n} \right]$$

can take it out

$$\leq \sum_{F} \underbrace{\mathbb{E}\left[\exp\left\{ t \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_i \right\} \right]}_{\leq \exp\left\{ \frac{t^2 k^2}{2n} \right\}} \,\Bigg|\, z_{1:n}$$

$$\leq |F| \cdot \exp\left\{ \frac{t^2 k^2}{2n} \right\} \quad \Big| \, z_{1:n}$$

- We care about $R(\mathcal{F})$, not $\hat{R}(\mathcal{F})$. F depends on $\{z_1 \dots z_n\}$, so need to make the bound hold for all data.

$\Rightarrow$ take max over data: $|F| \leq \max\limits_{\{z_1 \dots z_n\} \subseteq \mathcal{Z}} \left| \{[f(z_1) \cdots f(z_n)] : f \in \mathcal{F}\} \right|$

$$\triangleq s(\mathcal{F}, n)$$

Def (Shattering Coefficient): Let $\mathcal{F} = \{ f : z \to \{0,1\} \}$

$$s(\mathcal{F}, n) = \underbrace{\max_{\{z_1 \dots z_n\} \subseteq \mathcal{Z}} \underbrace{\left| \{ \underbrace{[f(z_1) \cdots f(z_n)]}_{[0\ 1\ 0\ \cdots\ 0] \to \text{boolean vectors}} : f \in \mathcal{F} \} \right|}_{\text{cardinality for a fixed data.}}}_{\text{max of this card over all possible datasets.}}$$

$\Rightarrow$ In MFL, $|\mathcal{F}|$ is replaced by $s(\mathcal{F}, n)$.

New!!

Massart's "Infinite" Lemma: If $\sup\limits_{\mathcal{F}} \frac{1}{n} \sum\limits_{i=1}^{n} f(z_i)^2 \leq K^2$, then

$$\hat{R}(\mathcal{F}) \leq K \sqrt{\frac{2 \log s(\mathcal{F}, n)}{n}}.$$

Def: $\mathcal{F}$ is a class of boolean functions on $\mathcal{Z}$. We say that $\mathcal{F}$ shatters a subset $D \subseteq \mathcal{Z}$ if any fnc $g : D \to \{0,1\}$ can be obtained by restricting some $f \in \mathcal{F}$ to D.

Ex: $D = \{z_1 \dots z_n\}$, for $f \in \mathcal{F}$ look at the vectors $[f(z_1) \dots f(z_n)]$.

If we can get every 2^n combinations, $\mathcal{F}$ shatters D.

- For boolean fncs, if $s(\mathcal{F}, n) = 2^n$,

$\Leftrightarrow \exists D = \{z_1 \dots z_n\} \subseteq Z$ s.t $\mathcal{F}$ shatters D

$\Leftrightarrow \exists D = \{z_1 \dots z_n\} \subseteq Z$ s.t $\left| \{ [f(z_1) \dots f(z_n)] : f \in \mathcal{F} \} \right| = 2^n$.

- When this happens: $RC \leq k \sqrt{\frac{2 \log s(\mathcal{F}, n)}{n}} = O(1)$

$\Rightarrow$ no generalization!

$\Rightarrow$ For generalization, need $s(\mathcal{F}, n)$ to be subexp in n.

* First step: $\mathcal{G} \longrightarrow \mathcal{F}$ for <u>0-1 loss</u> $\quad f : \mathcal{X} \to \{-1, 1\}$, $\mathcal{Y} = \{-1, 1\}$

$$\ell((y,x), f) = \mathbb{1}_{\{y \neq f(x)\}} = -\frac{1}{2}(-1 + y f(x))$$

$$- \mathcal{F} = \{ f : \mathcal{X} \to \{\pm 1\} \} \text{ and } \mathcal{G} = \{ z = (y,x) \to \mathbb{1}_{\{y \neq f(x)\}} : f \in \mathcal{F} \}$$

Fix data:

$$f = [f(x_1) \dots f(x_n)]$$

$$\quad\; \downarrow \qquad\quad \downarrow$$

$$\quad\; y_1 \;\dots\; y_n$$

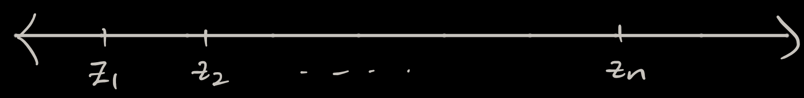
$$-\frac{1}{2}\left(-1 + [y_1 f(x_1) \dots y_n f(x_n)]\right) = [\ell((y_1, x_1), f) \dots \ell((y_n, x_n), f)] = g$$

$\Rightarrow$ bijection f and g

$\Rightarrow$ $\underline{s(\mathcal{F}, n)} = s(\mathcal{G}, n)$

$\quad \hookrightarrow$ we focus on this.

Ex: $\mathcal{F} = \{ z \to 1_{\{z \geq t\}} : t \in \mathbb{R} \}$ $\quad |\mathcal{F}| = \infty.$



$$\begin{array}{c|cccc}
t & 1_{\{z_1 \geq t\}} & 1_{\{z_2 \geq t\}} & \cdots & 1_{\{z_n \geq t\}} \\
\hline
t < z_1 & [\ 1 & 1 & \cdots & 1\ ] \\
z_1 \leq t < z_2 & [\ 0 & 1 & \cdots & 1\ ] \\
\vdots & \vdots & \vdots & & \vdots \\
z_n < t & [\ 0 & 0 & \cdots & 0\ ]
\end{array} \quad \Bigg\} \quad \begin{array}{c} s(\mathcal{F}, n) = n+1 \\ (\text{subexp}) \end{array}$$

$\Rightarrow \mathcal{F}$ cannot shatter $\{z_1 \ldots z_n\}$ for $n > 1$.

— VC Dimension (Vapnik – Chervonenkis)

Def (VC-dim): $VC(\mathcal{F})$ is the largest cardinality of a subset $D \subseteq \mathcal{Z}$ that can be shattered by $\mathcal{F}$.

$\Rightarrow$ Since $\mathcal{F} = \{ f \in \mathcal{Z} \to \{0,1\} \}$

$$VC(\mathcal{F}) = \sup \{ n : s(\mathcal{F}, n) = 2^n \}$$

Ex (above example): $s(\mathcal{F}, n) = n+1 = 2^n \Rightarrow VC(\mathcal{F}) = 1.$

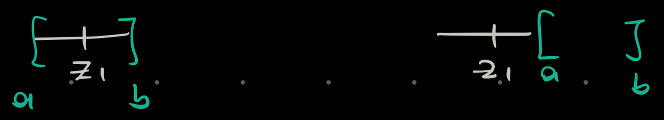
Remark: If $VC(\mathcal{F}) = d$,

i- $\exists D \subseteq \mathcal{Z}$ s.t. $\mathcal{F}$ shatters D and $|D| = d$.

ii- No subset $D \subseteq \mathcal{Z}$ of size $d+1$ can be shattered by $\mathcal{F}$.

Ex (Indicators of closed intervals): $\mathcal{F} = \left\{ z \to 1_{\{z \in [a,b]\}} : a, b \in \mathbb{R} \right\}$

$-\ n = 1$



$f(z_1)$ 1 0

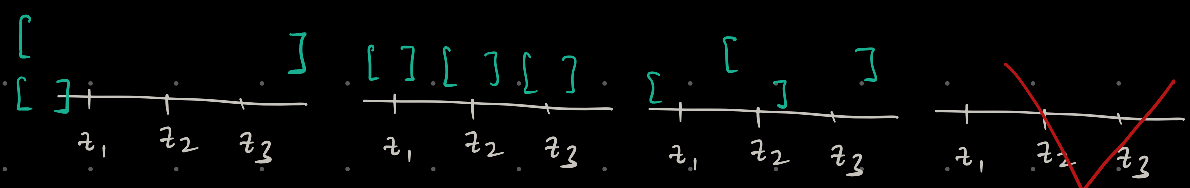
$$\Rightarrow \boxed{s(\mathcal{F},1) = 2}$$

$-\ n = 2$

$[f(z_1)\ f(z_2)]$ $[0\ 0]$ $[1\ 0]$ $[0\ 1]$ $[1\ 1]$

$$\Rightarrow \boxed{s(\mathcal{F},2) = 2^2}$$

$-\ n = 3$



$[f(z_1)\ f(z_2)\ f(z_3)]$ $[0\ 0\ 0]$, $[1\ 1\ 1]$ $[1\ 0\ 0]$, $[0\ 1\ 0]$, $[0\ 0\ 1]$ $[1\ 1\ 0]$, $[0\ 1\ 1]$ $\cancel{[1\ 0\ 1]}$

$$\Rightarrow \boxed{s(\mathcal{F},3) = 7}$$

$$\Rightarrow VC(\mathcal{F}) = 2.$$

Lemma (Sauer–Shelah): If $VC(\mathcal{F}) = d$, then

$$s(\mathcal{F}, n) \leq \begin{cases} 2^n & \text{if } n \leq d \\ \left(\frac{en}{d}\right)^d & \text{if } n > d \end{cases}.$$

Remarks: – For $n \leq d$, $R(\mathcal{F}) \leq \sqrt{\frac{n \log 2}{n}} = O(1)$.

– For $n > d$, $s(\mathcal{F}, n) \leq \mathrm{poly}(n) \Rightarrow R(\mathcal{F}) \lesssim \sqrt{\frac{\log n}{n}}$.

In fact, $R(\mathcal{F}) \leq \sqrt{\frac{2 \log s(\mathcal{F},n)}{n}} \leq \sqrt{\frac{2d \log \frac{en}{d}}{n}}$

$$\leq \sqrt{\frac{2d\,(1 + \log n - \log d)}{n}} \leq \sqrt{\frac{3\, VC(\mathcal{F}) \cdot \log n}{n}}.$$

- For 0-1 loss and binary classifier, when $n \geq VC(\mathcal{F})$,

$$R(\mathcal{G}) \leq R(\mathcal{F}) \leq \sqrt{\frac{3\,VC(\mathcal{F})\log n}{n}}.$$

$$\Rightarrow R(\hat{f}) - R(f_*) \leq 4 \cdot \sqrt{\frac{3\,VC(\mathcal{F})\log n}{n}} + \sqrt{\frac{2\log 1/\delta}{n}}.$$

proof: i) $n \leq d$ is trivial.

ii) $n > d$:

- Let $z_1^*, \dots z_n^*$ be s.t. $s(\mathcal{F}, n) = |\{[f(z_1^*) \dots f(z_n^*)] : f \in \mathcal{F}\}|$.
- Define $\mathcal{Z}^* = \{z_1^*, \dots z_n^*\}$ and restrict $\mathcal{F}$ onto $\mathcal{Z}^*$ and refer to it as $\mathcal{F}^*$.
- $\mathcal{F}^*$ is finite and $|\mathcal{F}^*| = s(\mathcal{F}, n)$.

Pajor's lemma: $\mathcal{F}^*$ is a class of fncs on a finite domain $\mathcal{Z}^*$. Then, $|\mathcal{F}^*| \leq |\{\Lambda \subseteq \mathcal{Z}^* : \Lambda \text{ is shattered by } \mathcal{F}^*\}|$

— proved in HW3 —

- By Pajor's lemma, $s(\mathcal{F}, n) = |\mathcal{F}^*| \leq \sum_{i=0}^{d^*} \binom{n}{i}$ (*)

where $d^* = VC(\mathcal{F}^*)$ (if $d^* \leq n$).

- But if $\Lambda \subseteq \mathcal{Z}^* \subseteq \mathcal{Z}$, and $\mathcal{Z}^*$ is shattered by $\mathcal{F}^*$, it is also shattered by $\mathcal{F}$.

$$\Rightarrow VC(\mathcal{F}^*) = d^* \leq VC(\mathcal{F}) = d < n.$$

(*) $$s(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

$$\leq \left(\frac{en}{d}\right)^d \quad \text{by the below lemma.}$$

Lemma: $\displaystyle\sum_{i=0}^{d} \binom{n}{i} \le \left(\frac{en}{d}\right)^d$ for $n \ge d$.

proof:
$$\sum_{i=0}^{d} \binom{n}{i} = \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^i \left(\frac{d}{n}\right)^i$$
$$\le \left(\frac{n}{d}\right)^d \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i$$
$$= \left(\frac{n}{d}\right)^d \left(1+\frac{d}{n}\right)^n$$
$$\le \left(\frac{n}{d}\right)^d e^d .$$