Uniform convergence and Rademacher Complexity

I) Introduction

- Reminder on Hoeffding inequality

$$P\left(\left|\frac{1}{N}\sum z_i - \mathbb{E}z\right| \geq \varepsilon\right) \leq 2\exp(-2N\varepsilon^2)$$

equivalently,

$$\left|\frac{1}{N}\sum z_i - \mathbb{E}z\right| < \sqrt{\frac{\log 2/\delta}{2N}} \quad \text{w.p. at least } 1-\delta$$

- <u>def</u>: a sequence of r.v. $Z_1 \dots Z_N \dots$ converges in proba to Z $\left(Z_N \xrightarrow[N\to\infty]{\text{proba}} Z\right)$ iff:

$$\forall \varepsilon, \delta \in]0,1[, \ \exists n, \text{ if } N > n,$$
$$|Z_N - Z| < \varepsilon \text{ with proba } 1-\delta$$

. Equivalently, there exist a function $n(\varepsilon,\delta)$ such that $\forall \varepsilon, \delta \in]0,1[$,

$$N > n(\varepsilon,\delta) \Rightarrow |Z_N - Z| < \varepsilon \text{ with proba } 1-\delta$$

<u>exercice</u>: Show in the Hoeffding setting

$$\frac{1}{N}\sum_i Z_i \xrightarrow[N\to\infty]{\text{in proba}} \mathbb{E}Z, \quad \text{give } n(\varepsilon,\delta)$$

- In this lecture, $S = \{(x_1,y_1) \dots (x_N,y_N)\}$ is considered as a random variable

- The empirical risk

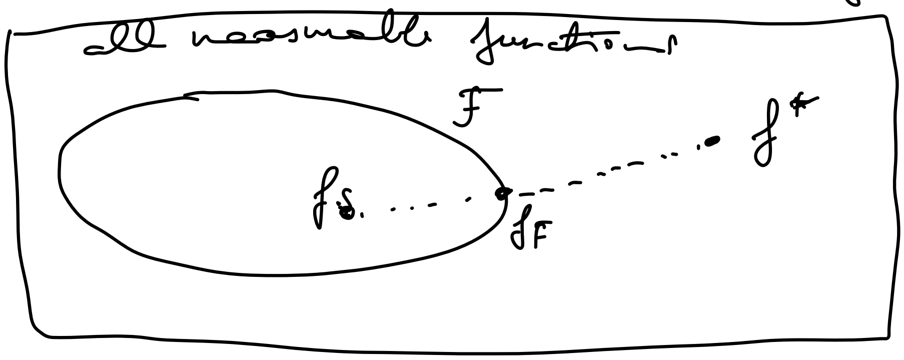
$$\hat{R}_S(f_S) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_S(x_i), y_i)$$

can be the 0/1 loss

II Notions of consistency

A learner f_S is ERM iff

$$f_S \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}_S(f)$$



$$f^* = \underset{f \in \text{measurable}}{\operatorname{argmin}} R(f)$$

$$f_\mathcal{F} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} R(f)$$

$$\underbrace{R(f_S) \geq R(f_\mathcal{F})}_{\text{estimation error}} \underbrace{\geq R(f^*)}_{\text{approximation error}}$$

Def: The learning algo f_S

- is universally Bayes consistent iff $\forall$ distribution $\mathcal{P}$, $R(f_S) \xrightarrow[N \to \infty]{\text{in proba}} R(f^*)$

  in other words, there is a function $m(\varepsilon, \delta, \mathcal{P})$ such that for any $\mathcal{P}$, ε, δ if $N > m(\varepsilon, \delta, \mathcal{P})$ then,
  for $S \sim \mathcal{P}^N$, $|R(f_S) - R(f^*)| < \varepsilon$ w.p. $1-\delta$

  ⚠ impossible for ERM

- is universally $\mathcal{F}$-consistent if $\forall \mathcal{P}$, $R(f_S) \xrightarrow[N \to \infty]{\text{in pr.}} R(f_\mathcal{F})$

- is a PAC-learner (Probably approximately correct) if there is a function $m(\varepsilon, \delta)$

such that $\forall$ distribution P,
$\forall \varepsilon, \delta \in]0,1[$, if $N > m(\varepsilon, \delta)$
then for $S \sim P^N$, $|R(f_S) - R(f_F)| < \varepsilon$ w.p. $1-\delta$

⚠ PAC implies F consistency

III) PAC learning and Uniform convergence for ERM

- We want to bound $R(f_S) - R(f_F)$
- Hoeffding allows us to bound $R(f) - \hat{R}(f)$ for a fixed f !!!

$$R(f_S) - R(f_F) = R(f_S) - \hat{R}(f_S) + \underbrace{\hat{R}(f_S) - \hat{R}(f_F)}_{\leq 0} + \hat{R}(f_F) - R(f_F)$$

$$\leq 2 \sup_{f \in F} |R(f) - \hat{R}(f)|$$

def: The UnRepresentativeness of S w.r.t. F is
$$\mathrm{UnRep}(F, S) = \sup_{f \in F} |R(f) - \hat{R}(f)|$$

thm: If, for class F, there exist $m(\varepsilon, \delta)$ s.t.
for any distribution P, any $\varepsilon, \delta \in]0,1[$,
if $N > m(\varepsilon, \delta)$ then $\mathrm{UnRep}(F, S) < \varepsilon$ w.p. $1-\delta$
(which is called the uniform convergence property),
then, ERM is a PAC learner on F.

proof: if $N > m(\frac{\varepsilon}{2}, \delta)$ then $\mathrm{Unrep}(F, S) \leq \varepsilon/2$ w.p. $1-\delta$
and $R(f_S) - R(f_F) \leq 2\,\mathrm{UnRep}(F, S) \leq \varepsilon$ w.p. $1-\delta$, so f_S is a PAC learner

Application to finite class $\mathcal{F}$

- I want to show $\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| < \varepsilon$ w.p. $1-\delta$ for $N > m(\varepsilon, \delta)$.

- $P\left(\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \geq \varepsilon\right)$

$$= P(\exists f \in \mathcal{F}, |R(f) - \hat{R}(f)| \geq \varepsilon)$$

> **Union bound:**
>
> $P(A \cup B) \leq P(A) + P(B)$
>
> $P(\exists i, A_i) \leq \sum_i P(A_i)$

$$\leq \sum_{f \in \mathcal{F}} P(|R(f) - \hat{R}(f)| \geq \varepsilon)$$

this f does not depend on the data
$\Rightarrow$ Hoeffding!

note that: $\hat{R}(f) = \frac{1}{N} \sum_{i=1}^{N} \ell^{01}(f(x_i), y_i)$

$R(f) = \mathbb{E}_{S \sim P^N}[\hat{R}(f)]$

$$P\left(\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \geq \varepsilon\right) \leq \sum_{f \in \mathcal{F}} \overbrace{P(|R(f) - \hat{R}(f)| \geq \varepsilon)}^{\text{apply Hoeffding here}} \leq \delta \times |\mathcal{F}|$$

if $N > \frac{\log \frac{2}{\delta}}{2\varepsilon^2}$

$\Longrightarrow$ $\mathrm{Unif}(\mathcal{F}, S) \leq \varepsilon$ w.p. $1 - \delta \times |\mathcal{F}|$ when $N > \frac{\log 2/\delta}{2\varepsilon^2}$

let $\delta' = \delta \times |\mathcal{F}|$

$\Longrightarrow$ $\mathrm{Unif}(\mathcal{F}, S) \leq \varepsilon$ w.p. $1 - \delta'$ when $N > \frac{\log\left(\frac{2|\mathcal{F}|}{\delta'}\right)}{2\varepsilon^2}$

$\Rightarrow$ ERM on finite class is PAC-learnable.

Equivalently,

$$\text{Unrep}(\mathcal{F}, S) \leq \sqrt{\frac{\log \frac{2|\mathcal{F}|}{\delta'}}{2N}} \quad \text{w.p. at least } 1-\delta'$$

$$\Rightarrow \quad |R(f_S) - \hat{R}(f_S)| \leq \sqrt{\frac{\log \frac{2|\mathcal{F}|}{\delta'}}{N}} \quad \text{w.p. at least } 1-\delta'$$

(!) This only works for finite class because the union bound for ∞ nb of events looks like $P(\exists i, A_i) \leq \sum_{i=1}^{\infty} P(A_i) \approx \infty$

IV The case $|\mathcal{F}| = \infty$, Rademacher complexity

Goal: bound $\text{Unrep}(\mathcal{F}, S)$ for $|\mathcal{F}| = \infty$ without union bound

Many tools: Vapnik dim, covering numbers, gaussian complexity, Rademacher complexity, ...

Rademacher applies to arbitrary bounded losses.

Notation: $Z = (X, Y)$ a labelled example

$S = (Z_1 \ldots Z_N)$

Given $\mathcal{F}$

define $\mathcal{G} = \ell \circ \mathcal{F} = \{(x,y) \mapsto \ell(f(x), y), f \in \mathcal{F}\}$

$$\text{UnRep}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|$$

$$= \sup_{g \in \mathcal{G}} \left| \frac{1}{N} \sum_{i=1}^{N} g(z_i) - \mathbb{E}_{Z \sim P}\, g(Z) \right|$$

def: the empirical Rademacher complexity of S on $\mathcal{G}$ is

$$\boxed{\widehat{\mathrm{Rad}}_S(\mathcal{G}) = \frac{1}{N}\mathbb{E}_{\sigma_1\ldots\sigma_N\sim \mathrm{Unif}(\{-1,1\})} \sup_{g\in\mathcal{G}} \sum_{i=1}^{N} \sigma_i g(z_i)}$$

Intuition 1: suppose I have drawn 2 data sets S_1, S_2

$$\sup_g \left|\hat{R}_{S_1}(f) - \hat{R}_{S_2}(f)\right| = \sup_{g\in\mathcal{G}} \frac{1}{N}\left[\sum_{(x,y)\in S_1} g(z_i) - \sum_{(x,y)\in S_2} g(z_i)\right]$$

$$= \sup_g \frac{1}{N}\sum_{(x,y)\in S_1\cup S_2} \sigma_i g(z_i)$$

where $\sigma_i = \begin{cases} 1 & \text{if } (x,y)\in S_1, \\ -1 & \text{otherwise.}\end{cases}$

Assume S is given,
we average $\sup_g \left|\hat{R}_{S_1}(f) - \hat{R}_{S_2}(f)\right|$ over
<u>all</u> partitions of S in (S_1, S_2)
we get Rademacher complexity

<u>Intuition 2</u>: Measure how well $\mathcal{F}$ can fit noisy labels.

<u>Rademacher lemma</u>:

$$\mathbb{E}_{S\sim P^N}\left[\mathrm{UnRep}(\mathcal{F}, S)\right] \leq 2\,\mathbb{E}_{S\sim P^N}\left[\widehat{\mathrm{Rad}}(\mathcal{G})\right]$$

<u>Thm (PAC with Rademacher)</u>

Assume $|\ell(f(x), y)| \leq c$ for all (x,y)

For all $f \in \mathcal{F}$

if $S \sim P^N$, with proba $1-\delta$

$$R(f) - \hat{R}_S(f) \leq 2\, \hat{\mathrm{Rad}}_S(\ell \circ \mathcal{F}) + 4c\sqrt{\frac{2\ln 4/\delta}{N}}$$

we conclude the PAC result:

$$R(f) - R(f_{\mathcal{F}}) \leq \; ? \quad \text{(exercise)}$$

exercise: 1) let $\mathcal{G} = \{ z \mapsto \alpha \;:\; \alpha \in [-1,1] \}$

what is $\hat{\mathrm{Rad}}_S(\mathcal{G}) = ?$

$$\boxed{\hat{\mathrm{Rad}}_S(\mathcal{G}) = \frac{1}{N}\, \mathbb{E}_{\sigma_1 \dots \sigma_N \sim \mathrm{Unif}(\{-1,1\})} \sup_{g \in \mathcal{G}} \sum_{i=1}^N \sigma_i g(z_i)}$$

2) let $\mathcal{G} = \{$ decision trees which can output 1 or -1 at the leaves $\}$

$\hat{\mathrm{Rad}}_S(\mathcal{G}) = ?$

1) $$\sup_{\alpha \in [-1,1]} \sum_{i=1}^N \sigma_i \alpha = \sup_{\alpha \in \{-1,1\}} \sum_{i=1}^N \sigma_i \alpha = \left| \sum_{i=1}^N \sigma_i \right|$$

$$\hat{\mathrm{Rad}}_S(\mathcal{G}) = \frac{1}{N}\, \mathbb{E}_{\sigma_1 \dots \sigma_N} \left| \sum_{i=1}^N \sigma_i \right| = \frac{1}{N}\, \mathbb{E}_{\sigma_1 \dots \sigma_N} \sqrt{(\Sigma \sigma_i)^2}$$

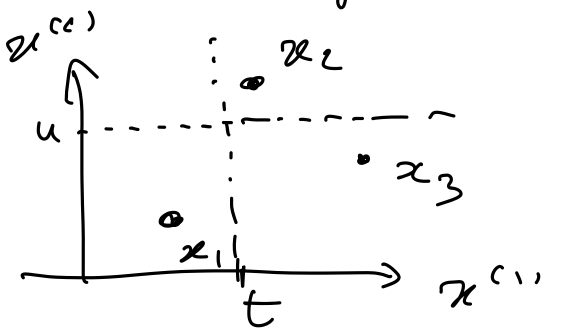
$$\leq \frac{1}{N} \sqrt{\mathbb{E}_{\sigma_1 \dots \sigma_N} \Big(\sum_{i=1}^N \sigma_i\Big)^2} \quad \text{(jensen inequality)}$$

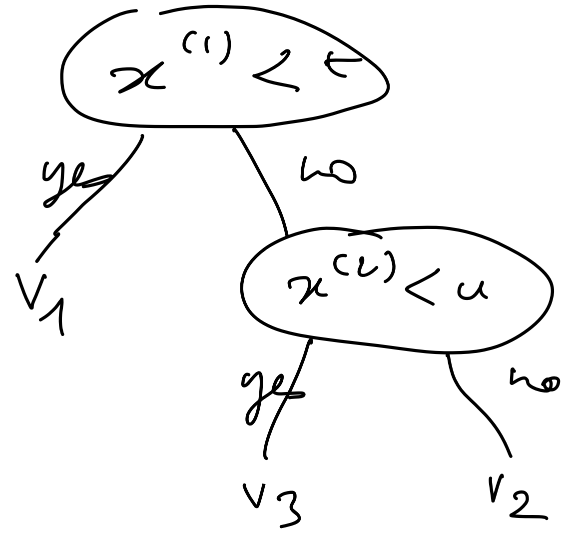
$$= \frac{1}{N} \sqrt{\mathrm{Var}(\Sigma \sigma_i)}$$

$$= \frac{1}{N} \sqrt{N\, \mathrm{Var}(\sigma_1)}$$

$$= \frac{1}{\sqrt{N}}$$

2) let us choose a dataset of 3 points.





σ_1	σ_2	σ_3	g	$\sum \sigma_i g(x_i)$
-1	-1	-1	$T(-1,-1,-1)$	3
-1	-1	1	$T(-1,-1,1)$	3
-1	1	-1	$T(-1,1,-1)$	3
-1	1	1	$T(-1,1,1)$	$\vdots$
1	-1	-1	$T(1,-1,-1)$	
1	-1	1	$\vdots$	
1	1	-1		
1	1	1		

Call $T(j,k,l)$ the above tree where $v_1 = j$, $v_2 = k$, $v_3 = l$

Here, $\forall \sigma_1 .. \sigma_N$, $\sup_g \sum_{i=1}^{N} \sigma_i g(x_i) = N$

$\widehat{Rad}_S(\mathcal{G}) = 1$.

$\Longrightarrow$ in the above thm this bound is completely useless!

Thm: for any $\mathcal{P}$, for $\mathcal{F} = \{x \mapsto x^T\theta \,;\, \|\theta\|_2 \leq W_2\}$

for any 1-lipschitz loss (hinge, logistic loss...)

$$\mathrm{UnRep}(\mathcal{F}, S) \leq \underbrace{\frac{W_2 X_2}{\sqrt{N}}}_{\text{Rademacher complexity}} + 4 X_2 \sqrt{\frac{2}{N} \ln \frac{2}{\delta}}$$

$$X_2 = \sup_{x \in \mathcal{X}} \|x\|_2$$