Uniform Convergence and Rademacher Complexity

Lecturer: Yann Chevaleyre
Scribe: BOUBAKER Arij

Lecture n°4 #
19/10/2023

1 Introduction

1.1 Reminder on Hoeffding's Inequality

Hoeffding's Inequality provides an upper bound on the probability that the sample mean deviates from the expected value:

$$\mathcal{P}\left(\left|\frac{1}{N}\sum_i \mathcal{Z}_i - \mathbb{E}(\mathcal{Z})\right| \geq \varepsilon\right) \leq 2\exp(-2N\varepsilon^2)$$

Alternatively, we can express it as:

$$\left|\frac{1}{N}\sum_i \mathcal{Z}_i - \mathbb{E}(\mathcal{Z})\right| < \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2N}}$$

This inequality holds with probability at least $1-\delta$.

1.2 Example with the Binomial Distribution

Let's begin by considering the scenario where 20 coins are drawn.
Each coin can take the value 0 or 1, representing tails and heads, respectively.
Our primary focus is on the number of 1s obtained, equivalent to counting the number of heads.
For instance, we might be interested in $t = 12$.
In this case, we inquire about the probability of obtaining at least 12 heads when drawing 20 coins. As we increase t, the probability decreases significantly.

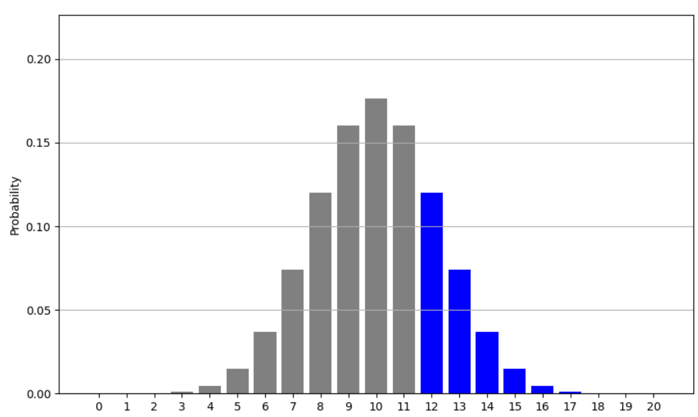


Figure 1: Binomial Distribution with $N = 20$ and $p = 0.5$.

The area to the right of the threshold t, denoted as $P(k > 12)$, is approximately 0.2517. According to Hoeffding's bound, $P(k > 12) < 0.6703$.

Exercise: From working with averages to dealing with the total sum

$$\mathcal{P}\left(\left|\frac{1}{N}\sum_i \mathcal{Z}_i - \mathbb{E}(\mathcal{Z})\right| \geq \varepsilon\right) \leq 2\exp(-2N\varepsilon^2)$$
$$\iff \mathcal{P}\left(\sum_i \mathcal{Z}_i \geq N \cdot \mathbb{E}(\mathcal{Z}) + N\varepsilon\right) \leq \exp(-2N\varepsilon^2)$$
$$\iff \mathcal{P}(k \geq t) \leq \exp\left(-2N\left(\frac{t}{N} - \mathbb{E}(\mathcal{Z})\right)^2\right)$$

Where:

$$t = N \cdot \mathbb{E}(\mathcal{Z}) + N\varepsilon$$
$$k = \sum_i Z_i$$
$$\varepsilon = \frac{t}{N} - \mathbb{E}(\mathcal{Z})$$

Definition 1. *A sequence of random variables $\mathcal{Z}_1, \ldots, \mathcal{Z}_N$ converges in probability to $\mathcal{Z}$ ($\mathcal{Z}_N \xrightarrow[N\to\infty]{\text{in proba}} \mathcal{Z}$) if and only if:*

For all $\varepsilon, \delta \in]0, 1[$, there exists an n such that if $N > n$, then

$$|\mathcal{Z}_N - \mathcal{Z}| < \varepsilon$$

with probability $1 - \delta$.

Equivalently, there exists a function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta \in]0, 1[$,

$$N > n(\varepsilon, \delta) \Rightarrow |\mathcal{Z}_N - \mathcal{Z}| < \varepsilon$$

with probability $1 - \delta$.

Exercise: Show in the Hoeffding setting that:

$$\frac{1}{N}\sum_i \mathcal{Z}_i \xrightarrow[N\to\infty]{\text{in proba}} \mathbb{E}(\mathcal{Z})$$

and provide the values for $n(\varepsilon, \delta)$.

Solution:

Let us define Y_N as :

$$Y_N = \frac{1}{N}\sum_i \mathcal{Z}_i \Rightarrow \text{we need to show} \quad Y_N \xrightarrow[N\to\infty]{\text{in proba}} \mathbb{E}(\mathcal{Z})$$

First, let's calculate the expected value:

$$\mathbb{E}(Y_N) = \mathbb{E}\left(\frac{1}{N}\sum \mathcal{Z}_i\right) = \frac{1}{N}\sum \mathbb{E}(\mathcal{Z}_i) = \mathbb{E}(\mathcal{Z})$$

Now, we want to find $n(\varepsilon, \delta)$ such that:

$$\mathcal{P}\left(|Y_N - \mathbb{E}(\mathcal{Z})| > \varepsilon\right) < 2\exp(-2N\varepsilon^2) \leq \delta.$$

We can rewrite the condition as:

$$-2N\varepsilon^2 \leq \frac{\log\left(\frac{\delta}{2}\right)}{2N} \iff N \geq \frac{\log(\frac{2}{\delta})}{2\varepsilon^2}.$$

So, we find that $n(\varepsilon, \delta) = \frac{\log(\frac{2}{\delta})}{2\varepsilon^2}$, for $N > n(\varepsilon, \delta)$, then we have $|Y_n - \mathbb{E}(\mathcal{Z})| \leq \varepsilon$ with probability at least $1 - \delta$.

In the last form:

$$\mathcal{P}\left(|Y_n - \mathbb{E}(\mathcal{Z})| > \varepsilon\right) < 2\exp(-2N\varepsilon^2) \leq \delta$$

$$\Rightarrow \mathcal{P}\left(|Y_n - \mathbb{E}(\mathcal{Z})| > \varepsilon\right) < \delta$$

For $2\exp(-2N\varepsilon^2) \leq \delta$, it follows that $2N\varepsilon^2 \leq \frac{\log(\frac{2}{\delta})}{2}$, which implies $\varepsilon \leq \sqrt{\frac{\log(\frac{2}{\delta})}{2N}}$.

$$\Rightarrow \left|\frac{1}{N}\sum \mathcal{Z}_i - \mathbb{E}(\mathcal{Z})\right| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2N}} \quad \text{with probability at least } 1 - \delta$$

1.3 Example : True Risk and Empirical Risk on two Gaussians

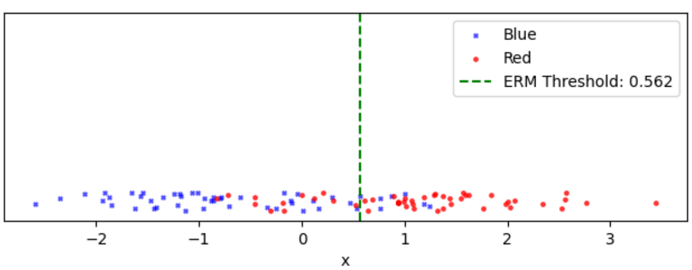
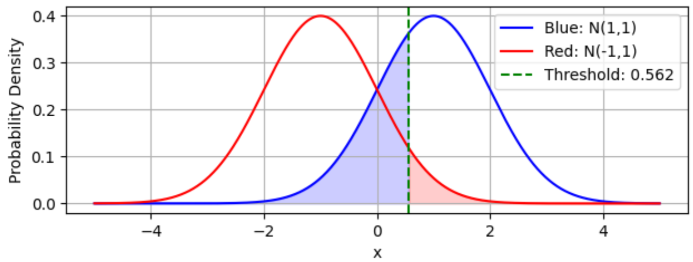


Figure 2: Generated Data Points.

The best classifier is one that cuts in the middle, setting a threshold at $x = 0$.

When the threshold is set at $x = 0$, we observe two normal distributions, one for class 1 and the other for class -1.



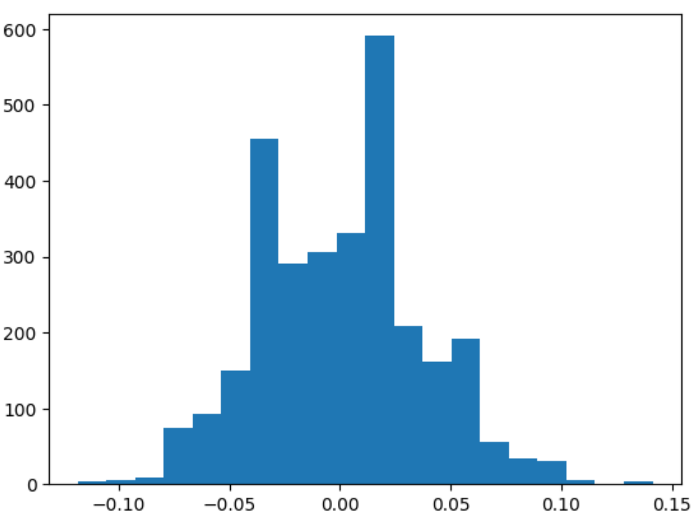
We have :
- $\hat{\mathcal{R}}(f_{\text{ERM}}) = 0.195$
- $\mathcal{R}(f_{\text{ERM}}) = 0.180$

When considering the 0/1 loss, the Bayes risk is simply the probability of making a mistake, which corresponds to 0.159

To find the optimal threshold, we draw points from these distributions and apply ERM to minimize the error, denoted as $\mathcal{W}$.

Instead of focusing solely on ERM, let's examine the classifier. Each time we run it, we obtain a different threshold. We can attempt to bound $\hat{\mathcal{R}}(f_{ERM}) - \mathcal{R}(f_{\text{ERM}})$.

To apply Hoeffding's inequality, we treat $\mathcal{R}(f_{\text{ERM}})$ as a random variable because the data is random. We assume that the dataset $\mathcal{S}$ is also a random variable. To be more concrete, we aim to estimate $\hat{\mathcal{R}(0)} - \mathcal{R}(0)$.



In each epoch, we draw a new dataset $\mathcal{S}$ and compute $\hat{\mathcal{R}}_0$. We calculate the true risk of the classifier.

Most of the time, the true risk and the empirical risk are close; however, they can occasionally deviate by as much as 10%. The empirical risk is bounded between 0 and 1, and with each new sample $\mathcal{S}$, this empirical risk is independent of the previous one. Thus, we can apply Hoeffding's bound.

In fact, $|\hat{\mathcal{R}}(f_0) - \mathcal{R}(f_0)| < \sqrt{\frac{\log(\frac{2}{\delta})}{2N}}$ with a probability of at least $1 - \delta$.

Here, $\hat{\mathcal{R}}(f_0)$ represents the empirical risk, defined as:

$\hat{\mathcal{R}}(f_0) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[f_0(x_i) \neq y_i\right]$

And $\mathcal{R}(f_0)$ represents the true risk, given by:

$\mathcal{R}(f_0) = \left|\frac{1}{N}\sum_{i=1}^{N} \mathcal{Z}_i - \mathbb{E}(\mathcal{Z})\right| < \sqrt{\frac{\log(\frac{2}{\delta})}{2N}}$

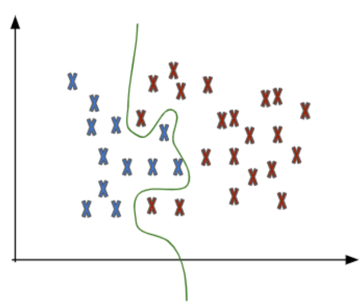
The probability that $|\hat{\mathcal{R}} - \mathcal{R}| < \varepsilon$ is greater than $1 - \delta$.

In this context, $\mathcal{Z}_i$ is defined as $\mathbf{1}\left[f_0(x_i) \neq y_i\right]$.

$\Rightarrow \hat{\mathcal{R}}_S(f_0) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{Z}_i$ and $\mathcal{R}(f_0) = \mathbb{E}(\mathcal{Z}_i)$.

1.4 Question: Can we bound $|\hat{\mathcal{R}}_S(f_{\mathbf{ERM}}) - \mathcal{R}(f_{\mathbf{ERM}})|$?

It turns out, no! This is because f_{ERM} is a best classifier computed on the data, implying that $\mathcal{Z}_i$ are not independent. The interdependence among the $\mathcal{Z}_i$ originates from the fact that f_{ERM} relies on all the $\mathcal{Z}_i$, and consequently, the error on one example becomes contingent on the probabilities of other examples.



$|\underbrace{\hat{\mathcal{R}}_{\mathcal{S}}(f_{\text{ERM}})}_{=0} - \mathcal{R}(f_{\text{ERM}})|$ is significant.

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

The empirical risk $\hat{\mathcal{R}}_{\mathcal{S}}(f_s) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\ell\left(f_s(x_i), y_i\right)}_{\text{which can be the 0/1 loss}}$.

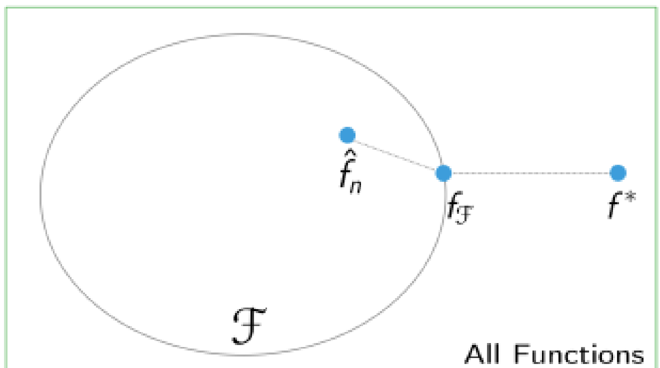
2 Notions of Consistency

A learner f_s is ERM if and only if:

$$f_s \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}_s(f)$$

Furthermore, we have:

$$f^* = \operatorname{argmin} \mathcal{R}(f), \quad f \in measurable$$
$$f_{\mathcal{F}} = \operatorname{argmin} \mathcal{R}(f), \quad f \in \mathcal{F}$$



Where : $\mathcal{R}(f_s) \geq \mathcal{R}(f_{\mathcal{F}}) \geq \mathcal{R}(f^*)$.

Definition 2. *The learning algorithm f_s is :*

- *universally Bayes consistent if and only if for all possible distributions $\mathcal{P}$,*

  $$\mathcal{R}(f_s) \xrightarrow[N \to \infty]{in\ proba} \mathcal{R}(f^*)$$

  In other words, there is a function $n(\varepsilon, \delta, \mathcal{P})$ such that for any ε, δ, and $\mathcal{P}$, if $N > n(\varepsilon, \delta, \mathcal{P})$, then for $\mathcal{S} \sim \mathcal{P}^N$:

  $$\left| \mathcal{R}(f_s) - \hat{\mathcal{R}}(f^*) \right| < \varepsilon$$

  with probability $1 - \delta$.

  ⚠ *Impossible for ERM*

- *Is universally $\mathcal{F}$-consistent if for any $\mathcal{P}$, $\mathcal{R}(f_s) \xrightarrow[N \to \infty]{in\ proba} \mathcal{R}(f_{\mathcal{F}})$*

- *Is a PAC-learner (Probably Approximately Correct) if there is a function $n(\varepsilon, \delta)$ such that for any distribution $\mathcal{P}$, for any $\varepsilon, \delta \in]0, 1[$, if $N > n(\varepsilon, \delta)$, then for $\mathcal{S} \sim \mathcal{P}^N$,*

  $$\left| \mathcal{R}(f_s) - \hat{\mathcal{R}}(f_{\mathcal{F})} \right| < \varepsilon \quad with\ probability\ 1 - \delta$$

  ⚠ *PAC implies $\mathcal{F}$-consistency*

3 PAC Learning and Uniform Convergence for ERM

We want to bound $\mathcal{R}(f_s) - \mathcal{R}(f_{\mathcal{F}})$.

However, Hoeffding allows us to bound $\mathcal{R}(f) - \hat{\mathcal{R}}(f_{\mathcal{F}})$ for a fixed f.

$$\begin{aligned}
\mathcal{R}(f_s) - \mathcal{R}(f_{\mathcal{F}}) &= \mathcal{R}(f_s) - \hat{\mathcal{R}}(f_s) + \hat{\mathcal{R}}(f_s) - \hat{\mathcal{R}}(f_{\mathcal{F}}) + \hat{\mathcal{R}}(f_{\mathcal{F}}) - \mathcal{R}(f_{\mathcal{F}}) \\
&\leq 2 \cdot \sup_{f \in \mathcal{F}} \left| \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right|
\end{aligned}$$

Definition 3. *The unrepresentativeness of $\mathcal{S}$ with respect to $\mathcal{F}$ is defined as*

$$\mathit{Unrep}(\mathcal{F}, \mathcal{S}) = \sup_{f \in \mathcal{F}} \left| \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right|$$

Theorem 1. *If, for class $\mathcal{F}$, there exist $n(\varepsilon, \delta)$ such that for any distribution $\mathcal{P}$, any $\varepsilon, \delta \in]0,1[$, if $N > n(\varepsilon, \delta)$ then Unrep($\mathcal{F}$, $\mathcal{S}$) $< \varepsilon$ with probability 1-δ (which is called the uniform convergence property), then , ERM is a PAC learner on $\mathcal{F}$.*

Proof .
If N$>$ $n(\frac{\varepsilon}{2}, \delta)$ then Unrep($\mathcal{F}$, $\mathcal{S}$) $\leq \dfrac{\varepsilon}{2}$ with probability 1-δ and $\mathcal{R}(f_s) - \mathcal{R}(f_{\mathcal{F}}) \leq 2 \cdot$ Unrep$(\mathcal{F}, S) \leq \varepsilon$ with probability $1 - \delta$, so f_s is a PAC learner

Application to finite class $\mathcal{F}$

I want to show :

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

with probability $1 - \delta$ for $N > n(\varepsilon, \delta)$.

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

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

⚠ f: doesn't depend on the data => Hoeffding!

Union bound: P$(A \cup B) \leq \mathcal{P}(A) + \mathcal{P}(B)$, P $(\exists i, A_i) \leq \sum_i \mathcal{P}(A_i)$

Note that:

$$\hat{\mathcal{R}}(f) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(f(x_i), y_i\right),$$
$$\mathcal{R}(f) = \mathbb{E}_{\mathcal{S}\sim\mathcal{P}^N}[\hat{\mathcal{R}}(f)].$$

We can bound the discrepancy between the true risk $R(f)$ and the empirical risk $\hat{R}(f)$ as follows:

$$\mathcal{P}\left(\sup_{f\in\mathcal{F}}\left|\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right| \geq \varepsilon\right) \leq \sum_{f\in\mathcal{F}} \mathcal{P}\left(\left|\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right| \geq \varepsilon\right) \leq \delta\cdot|\mathcal{F}| \quad \text{if } N > \frac{\log\left(\frac{2}{\delta}\right)}{2\varepsilon^2}.$$

Here, we notice that if we multiply δ by $|\mathcal{F}|$, we obtain δ'. Consequently, we can rewrite δ as $\delta = \frac{\delta'}{|\mathcal{F}|}$.

$$\text{Unrep}(\mathcal{F}, \mathcal{S}) \leq \varepsilon \text{ with probability at least } 1 - \delta\cdot|\mathcal{F}| \text{ when } N > \frac{\log\left(\frac{2}{\delta}\right)}{2\varepsilon^2}$$

Let $\delta' = \delta\cdot|\mathcal{F}|$.

$$\Rightarrow \text{Unrep}(\mathcal{F}, \mathcal{S}) \leq \varepsilon \text{ with probability 1-}\delta' \text{ when } N > \frac{\log\left(\frac{2\cdot|\mathcal{F}|}{\delta'}\right)}{2\varepsilon^2}$$

$$\Rightarrow \text{ERM on finite classes is PAC-learner}$$

Equivalently,

$$\text{Unrep}(\mathcal{F}, \mathcal{S}) \leq \frac{\log\left(\frac{2\cdot|\mathcal{F}|}{\delta'}\right)}{2N} \text{ with probability at least 1-}\delta'$$

$$\Rightarrow \left|R(f) - \hat{R}(f)\right| \leq \frac{\log\left(\frac{2\cdot|\mathcal{F}|}{\delta'}\right)}{N} \text{ with probability at least 1-}\delta'$$

This only works for finite classes because the union bound for an infinite number of events looks like:

$$\mathcal{P}\left(\exists i, A_i\right) \leq \sum_{i=1}^{\infty} \mathcal{P}(A_i) \approx \infty$$

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

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

There are many tools available for this purpose: Vapnik Dimension, Covering numbers, Gaussian Complexity, Rademacher Complexity, ...

Rademacher Complexity applies to arbitrary bounded losses, not limited to the 0/1 loss.

Notation:

$\mathcal{Z} = (X, Y)$ represents a labeled example.

$\mathcal{S} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_N)$

Given $\mathcal{F}$, we define $\mathcal{G} = \ell \circ \mathcal{F} = \{(x, y) \to \ell(f(x), y) \mid f \in \mathcal{F}\}$.

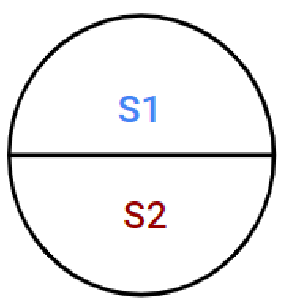
We can express the unrepresentativeness of $\mathcal{F}$ with respect to the sample $\mathcal{S}$ as follows:

$$\text{Unrep}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \left| \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right| = \sup_{g \in \mathcal{G}} \left| \frac{1}{N} \sum_{i=1}^{N} g(\mathcal{Z}_i) - \mathbb{E}_{\mathcal{Z} \sim \mathcal{P}} \left[g(\mathcal{Z}) \right] \right|.$$

Definition 4. *The empirical Rademacher complexity of $\mathcal{S}$ with respect to g is defined as*

$$\hat{Rad}_s(g) = \frac{1}{N} \mathbb{E}_{\sigma_1, \ldots, \sigma_N \sim Unif(\{-1,1\})} \sup \sum_{i=1}^{N} \sigma_i \cdot g(\mathcal{Z}_i)$$

Intuition 1: Suppose I have drawn two data sets $\mathcal{S}_1$ and $\mathcal{S}_2$.



We aim to calculate $\sup_{g \in \mathcal{G}} \left| \hat{\mathcal{R}}_{\mathcal{S}_1}(f) - \hat{\mathcal{R}}_{\mathcal{S}_2}(f) \right|$, which can be expressed as:

$$\sup_{g \in \mathcal{G}} \left[\frac{1}{N} \left(\sum_{(x,y) \in \mathcal{S}_1} g(\mathcal{Z}_i) - \sum_{(x,y) \in \mathcal{S}_2} g(\mathcal{Z}_i) \right) \right].$$

This can be further simplified as:

$$\sup_{g \in \mathcal{G}} \left[\frac{1}{N} \sum_{(x,y) \in \mathcal{S}_1 \cup \mathcal{S}_2} \sigma_i g(\mathcal{Z}_i) \right],$$

where σ_i is defined as:

$$\sigma_i = \begin{cases} 1 & \text{if } (x, y) \in \mathcal{S}_1, \\ -1 & \text{otherwise.} \end{cases}$$

Assuming $\mathcal{S}$ is given, we can average $\sup_{g \in \mathcal{G}} \left| \hat{\mathcal{R}}_{\mathcal{S}_1}(f) - \hat{\mathcal{R}}_{\mathcal{S}_2}(f) \right|$ over all partitions of $\mathcal{S}$ into $(\mathcal{S}_1, \mathcal{S}_2)$ to obtain the Rademacher complexity.

Intuition 2 : This intuition measures how effectively $\mathcal{F}$ can accommodate noisy labels.

Rademacher Lemma:

The concept of Rademacher complexity extends to cases with arbitrary bounded losses.

In this context, we define the *unrep* for the class $\mathcal{F}$ and the sample $\mathcal{S}$ as follows:

$$\mathbb{E}_{\mathcal{S} \sim \mathcal{P}^N}[\text{Unrep}(\mathcal{F}, \mathcal{S})] < 2\mathbb{E}_{\mathcal{S} \sim \mathcal{P}^N}[\hat{Rad}(g)].$$

Theorem 2. *Assume* $|\ell(f(x), y)| \leq c$ *for all* (x, y)*. For all* $f \in \mathcal{F}$*, if* $\mathcal{S} \sim \mathcal{P}^N$ *with probability* $1 - \delta$*,*

$$|\mathcal{R}(f) - \hat{\mathcal{R}}_s(f)| \leq 2 \cdot \hat{Rad}_s(\ell \circ \mathcal{F}) + 4 \cdot c \cdot \sqrt{\frac{2 \ln\left(\frac{4}{\delta}\right)}{N}}$$

We conclude the result:

$$\mathcal{R}(f) - \mathcal{R}(f_{\mathcal{F}}) \leq \text{?} \textit{ (exercise to be completedbe)}$$

5 Exercise 1

Let $g = \{\mathcal{Z} \to \alpha \mid \alpha \in [-1, 1]\}$. Determine $\hat{Rad}_s(g)$.

$$\hat{Rad}_s(g) = \frac{1}{N}\mathbb{E}_{\sigma_1,\dots,\sigma_N\sim\text{Unif}(\{-1,1\})} \sup \sum_{i=1}^{N} \sigma_i \cdot g(\mathcal{Z}_i)$$

Solution: 1

1.

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

2.

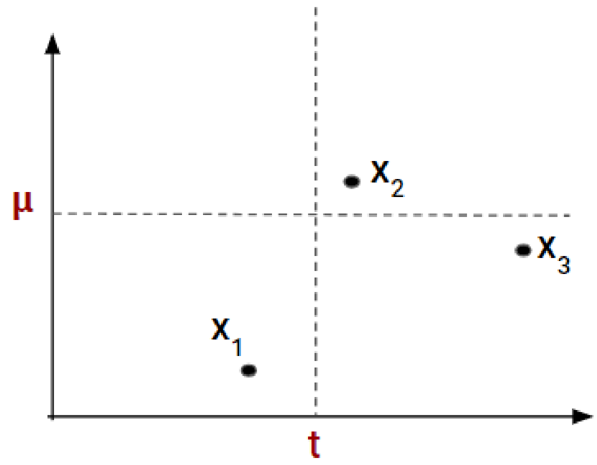
$$\begin{aligned}
\hat{Rad}_s(\mathcal{G}) &= \frac{1}{N}\mathbb{E}_{\sigma_1,\dots,\sigma_N}\left|\sum_{i=1}^{N} \sigma_i\right| \\
&= \frac{1}{N}\mathbb{E}_{\sigma_1,\dots,\sigma_N}\sqrt{(\sum_{i=1}^{N} \sigma_i)^2} \\
&\leq \frac{1}{N}\sqrt{\mathbb{E}_{\sigma_1,\dots,\sigma_N}\left(\sum_{i=1}^{N} \sigma_i\right)^2} \quad \text{(by Jensen's Inequality)} \\
&= \frac{1}{N}\sqrt{\text{var}\left(\sum_{i=1}^{N} \sigma_i\right)} \\
&= \frac{1}{N}\sqrt{N\cdot\text{var}(\sigma)} \\
&= \frac{1}{\sqrt{N}}
\end{aligned}$$

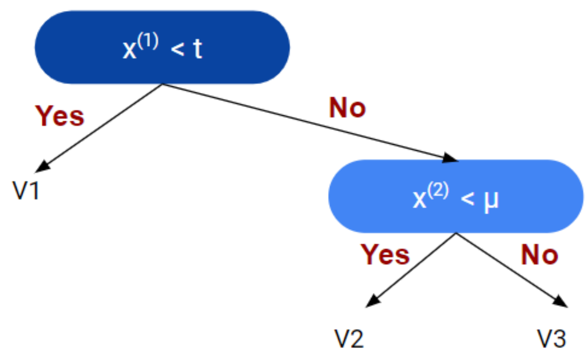
6 Exercise 2

Consider the set $\mathcal{G} = \{$decision trees that can output 1 or -1 at the leaves$\}$. Find $\hat{Rad}_s(g)$.

Solution 2:

Let's work with a dataset of 3 points. We define T(jk, l) as the tree where $v_1 = j$, $v_2 = k$, and $v_3 = l$.





Here, for all $\sigma_1, \dots, \sigma_N$, we have $\sup_g \sum_{i=1}^{N} \sigma_i \cdot g(x_i)$.

In this context, the bound provided by the previous theorem appears to be of limited utility!

σ_1	σ_2	σ_3	g	$\sum_{i=1}^{N} \sigma_i \cdot g(x_i)$
-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)	3
1	-1	-1	T(1,-1,-1)	3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Theorem 3. ***PAC with Rademacher***
For any $\mathcal{P}$ and a class $\mathcal{F}$ defined as $\mathcal{F} = \{x \to x^T \cdot \theta \mid \|\theta\|_2 \leq W_2\}$,
where $\mathcal{F}$ is associated with a 1-Lipschitz loss (such as hinge or logistic loss), we have:

$$Unrep(\mathcal{F}, \mathcal{S}) \leq \frac{W_2 \cdot X_2}{\sqrt{n}} + 4 \cdot X_2 \cdot \sqrt{\frac{2}{n} \cdot \ln\left(\frac{2}{\delta}\right)}$$

Where

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