Fondamentaux de l'Apprentissage Automatique

Lecturer: Liva Ralaivola
Scribe: Marius Roger

Lecture n°5 #
26/10/2020

Introduction

Reminders

Corollary of the Hoeffding Inequality :

For

— n IID variables $X_1, ..., X_n$ such that $\forall 1 \leq i \leq n, \mathbb{P}(0 \leq X_i \leq 1) = 1$
— $\mu = \mathbb{E}[X_1] \quad (= \mathbb{E}[X_2] = ... = \mathbb{E}[X_n])$
— $\epsilon > 0$

We have :

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right) \leq \exp(-2n\epsilon^2)$$

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq \epsilon\right) \leq \exp(-2n\epsilon^2)$$

Thus, by summing them, we obtain the "triangle inequality" :

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2)$$

This inequality is very present in the literature, and the important thing about it is not necessarily the constants, but rather the orders of magnitude (since the constants will disappear through logarithms and other operators). If we want to have the same bound, for the probability of the X_i and μ to be 10 times closer, we will need 100 times more data.

Goal of this lecture

For $S = \{(X_i, Y_i)\}_{i=1}^n$ an IID sample (very frequent assumption in ML), we want to obtain results like :
With probability at least $1 - \delta$ over the sampling of S :

$$\forall f \in \mathcal{F}, \mathcal{R}(f, \mathcal{D}) \leq \hat{\mathcal{R}}_n(f, S) + \epsilon(\delta, n, \mathcal{C}(\mathcal{F}))$$

Where $\mathcal{F}$ is a class of functions and $\epsilon(\delta, n, \mathcal{C}(\mathcal{F}))$ is some kind of measure on the capacity of the class of functions we are interested in.

This is a bound on the generalization error (= risk), and it is a **uniform** generalization error bound because it applies simultaneously to all f in the class $\mathcal{F}$, with probability at least $1 - \delta$.
Or equivalently, we have :

$$\mathbb{P}_{S\sim\mathcal{D}^n}(\ \exists f \in \mathcal{F}, \mathcal{R}(f, \mathcal{D}) \geq \hat{\mathcal{R}}_n(f, S) + \epsilon(\delta, n, \mathcal{C}(\mathcal{F}))\ ) \leq \delta$$

This expresses the same uniformity of the bound as :
The probability that a function f of $\mathcal{F}$ breaks this bound is lower than δ.

Today

We first study the case where $\mathcal{F}$ is countable and finite : $|\mathcal{F}| < \infty$. We want to understand the mechanics of how we can obtain a uniform generalization error bound when we only have "a few" functions to pick from.

Then, we study the case in which we don't have $|\mathcal{F}| < \infty$. This is somewhat similar to what we did with the Rademacher complexity, since it is relevant when dealing with an infinite number of functions. Since we cannot count the functions, we will try to restrict the "size", the capacity, of the class of functions. We will use the Vapnik-Chervonenkis dimension of classes of functions to this extent.

1 Case $|\mathcal{F}| < +\infty$

We assume that we consider binary classification.

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

$$\hat{\mathcal{R}}_n(f, S) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{f(X_i) \neq Y_i} \qquad \textit{(we can use other loss functions)}$$

$$\begin{aligned} \mathcal{R}(f, \mathcal{D}) &= \mathbb{E}[\mathbf{1}_{f(X_1) \neq Y_1}] && = \mathbb{P}(f(X_1) \neq Y_1) \\ &= \mathbb{E}_S[\hat{\mathcal{R}}_n(f, S)] && \textit{(by linearity of } \mathbb{E} \textit{ and IID properties of } S\textit{)} \end{aligned}$$

Where $\mathbf{1}$ is the indicator function.

1.1 Generalization error bound

According to the Hoeffding inequality,

$$\forall f \in \mathcal{F}, \quad \mathbb{P}_S(|\hat{\mathcal{R}}_n(f, S) - \mathcal{R}(f, \mathcal{D})| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

Since the $\mathbf{1}_{f(X_i) \neq Y_i}$ are IID, and $\mu = \mathbb{E}[\mathbf{1}_{f(X_i) \neq Y_i}]$

Or, using the one-sided inequality :

$$\forall f \in \mathcal{F}, \quad \mathbb{P}_S(\mathcal{R}(f, \mathcal{D}) - \hat{\mathcal{R}}_n(f, S) \geq \epsilon) \leq \exp(-2n\epsilon^2)$$

Given this, we have, $\forall f \in \mathcal{F}$, with probability at least $1 - \delta$:

$$\mathcal{R}(f, \mathcal{D}) \leq \hat{\mathcal{R}}_n(f, S) + \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\delta}\right)} \tag{1}$$

Proof of equation 1. By setting $\exp(-2n\epsilon^2) \leq \delta$, we obtain :

$$\begin{aligned} \exp(-2n\epsilon^2) = \delta &\iff -2n\epsilon^2 = \ln(\delta) \\ &\iff 2n\epsilon^2 = \ln\left(\frac{1}{\delta}\right) \\ &\iff \epsilon = \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\delta}\right)} \end{aligned}$$

□

Comments on equation 1 :
— The rate of the bound is $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$
— The generalization error bound is not uniform, since f is selected before "applying" the probability.

1.2 Uniform generalization error bound

To get a uniform generalization error bound, we want to achieve a result like

$$\mathbb{P}(\exists f \in \mathcal{F}, \mathcal{R}(f, \mathcal{D}) - \hat{\mathcal{R}}_n(f, S) \geq \epsilon) \leq \delta$$

We have :

$$\begin{aligned}
\mathbb{P}(\exists f \in \mathcal{F}, \mathcal{R}(f, \mathcal{D}) - \hat{\mathcal{R}}_n(f, S) \geq \epsilon) &= \mathbb{P}(\bigcup_{f \in \mathcal{F}} \{\mathcal{R}(f, \mathcal{D}) - \hat{\mathcal{R}}_n(f, S) \geq \epsilon\}) \\
&\underset{\textit{Union bound}}{\leq} \sum_{p=1}^{|\mathcal{F}|} \mathbb{P}(\mathcal{R}(f_p, \mathcal{D}) - \hat{\mathcal{R}}_n(f_p, S) \geq \epsilon) \\
&\underset{\textit{Hoeffding}}{\leq} \sum_{p=1}^{|\mathcal{F}|} \exp(-2n\epsilon^2) \\
&= |\mathcal{F}| \exp(-2n\epsilon^2)
\end{aligned}$$

Union bound : $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

Similarly to the first generalization bound, we solve :

$$\begin{aligned}
|\mathcal{F}| \exp(-2n\epsilon^2) = \delta &\iff -2n\epsilon^2 = \ln(\frac{\delta}{|\mathcal{F}|}) \\
&\iff 2n\epsilon^2 = \ln\left(\frac{|\mathcal{F}|}{\delta}\right) \\
&\iff \epsilon = \sqrt{\frac{1}{2n} \ln\left(\frac{|\mathcal{F}|}{\delta}\right)}
\end{aligned}$$

Given this ϵ, we obtain :

$$\mathbb{P}(\exists f \in \mathcal{F}, \mathcal{R}(f, \mathcal{D}) - \hat{\mathcal{R}}_n(f, S) \geq \sqrt{\frac{1}{2n} \ln\left(\frac{|\mathcal{F}|}{\delta}\right)}) \leq \delta$$

Theorem 1. *With a probability at least* $1 - \delta$ *:*

$$\forall f \in \mathcal{F}, \mathcal{R}(f, \mathcal{D}) \leq \hat{\mathcal{R}}_n(f, S) + \sqrt{\frac{1}{2n} \ln\left(\frac{|\mathcal{F}|}{\delta}\right)}$$

Comments :
— We used the "union bound"
— We used the fact that $|\mathcal{F}| < +\infty$
— $\mathcal{C}(\mathcal{F}) = |\mathcal{F}|$ here
— In practice, when doing machine learning, it is very rare to be in the case "$|\mathcal{F}| < +\infty$".
— VC-dimension will help us deal with the case "$|\mathcal{F}| < +\infty$ does not hold".

2 The Vapnik-Chervonenkis (VC) dimension

For a high-level idea of what the VC dimension measures, assume $\mathcal{F} \subseteq \{\mathcal{X} \to \{-1, 1\}\}$

Example : The set of linear binary classifiers in $\mathbb{R}^d$.

We observe that for n points $S = \{x_1, ..., x_n\}$, with $\mathcal{F}_S = \{(f(x_1), ..., f(x_n)), f \in \mathcal{F}\}$,

$$|\mathcal{F}_S| \leq 2^n$$

Studying the VC dimension, we are interested in the situation :

$$\sup_{S, |S|=n} |\mathcal{F}_S| < 2^n$$

2.1 Definitions

Definition 1. *Restriction of $\mathcal{F}$ to a sample*
For
— $\mathcal{F} \subseteq \{\mathcal{X} \to \{-1, 1\}\}$
— $\forall 1 \leq i \leq n, x_i \in \mathcal{X}$
— $S = \{x_1, ..., x_n\}$

We define

$$\mathcal{F}_S = \{(f(x_1), ..., f(x_n)), f \in \mathcal{F}\}$$

Note : *Sometimes, in the literature, $\mathcal{F}_S$ is described using a "functional" expression :*

$$\mathcal{F}_S = \{(x_1, ..., x_n) \to (f(x_1), ..., f(x_n)), f \in \mathcal{F}\}$$

Definition 2. *Shattered set*
Let $S = \{x_1, ..., x_n\}$
*S is **shattered** by $\mathcal{F}$ if $|\mathcal{F}_S| = 2^n$*

Note : *This can be visualized as $\mathcal{F}$ being able to label S in all possible ways.*

Definition 3. *Vapnik-Chervonenkis (VC) dimension*
The VC dimension of $\mathcal{F}$ is the size of the largest set that is shattered by $\mathcal{F}$.

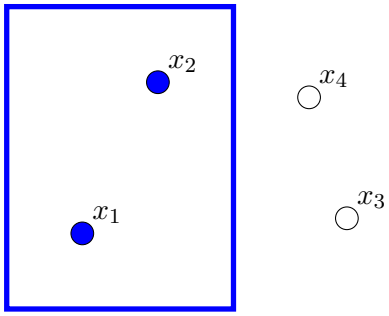
Note : *It is possible that VC-dim($\mathcal{F}$) $= +\infty$*

Comments :
— The VC dimension appeared in the 1970s.
— It is connected to "Computational ML".
— It is also connected to the PAC (Probably Approximately Correct) framework of learning, that took into consideration Complexity (from a computer science POV, studying complexity classes like NP, and decidability of problems).

2.2 Examples of VC dimension for some classes of functions

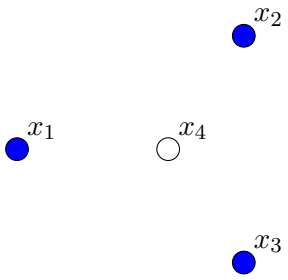
2.2.1 Axis-aligned rectangles

The VC dimension of axis-aligned rectangles is 4.



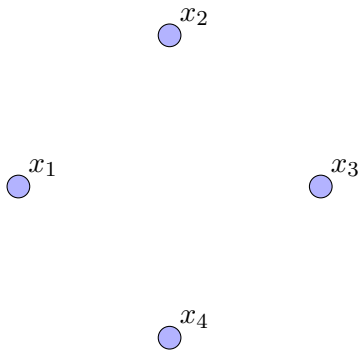
Samples that are classified as positive are exactly the samples inside the rectangle.

Proof of VC-dim$(AAR) = 4$. VC dimension analysis :

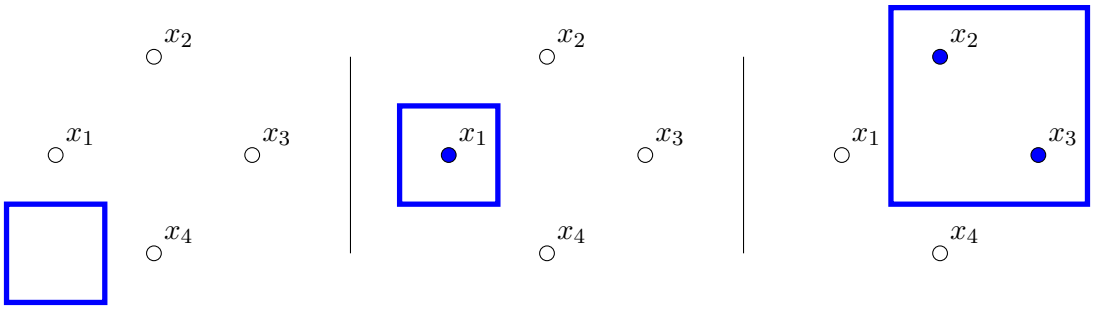


It is true that for this configuration of points you cannot realize *all labellings* (particularly the one depicted here).

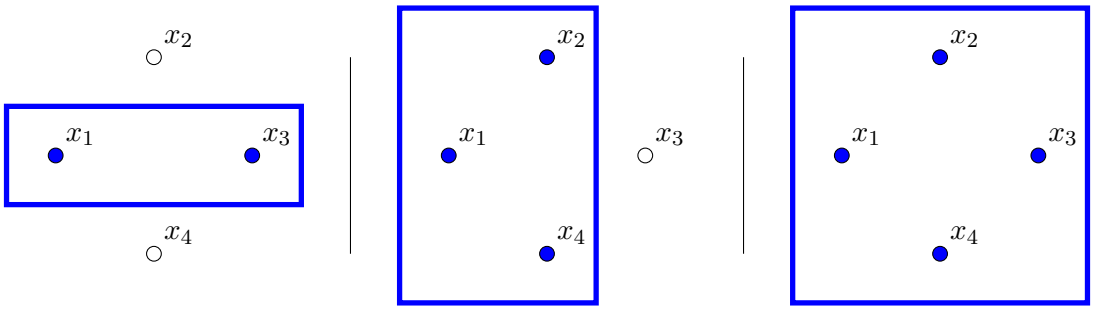
However, there exists a configuration of 4 points such that all labellings are possible :



This configuration is shattered by the class of axis-aligned rectangles.



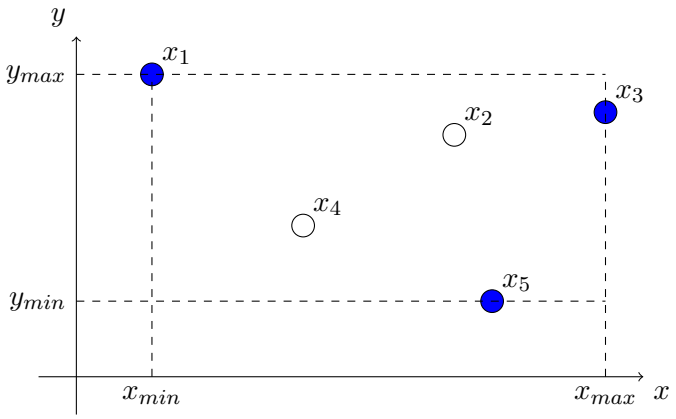
Examples of AAR classifiers for 0, 1, and 2 positives



Examples of AAR classifiers for 2 (alternative), 3, and 4 positives

However, there exists no configuration of 5 points that can be shattered by the class of axis-aligned rectangles.

For every 5-points configuration, there is no axis-aligned rectangle that can label as positive the points with maximum or minimum x/y value, and label as negative the points in the "inside". If there is no point in the inside, then all points are on the same maximal square and at least two points are on the same side of the square, thus one of them cannot be labeled as negative.



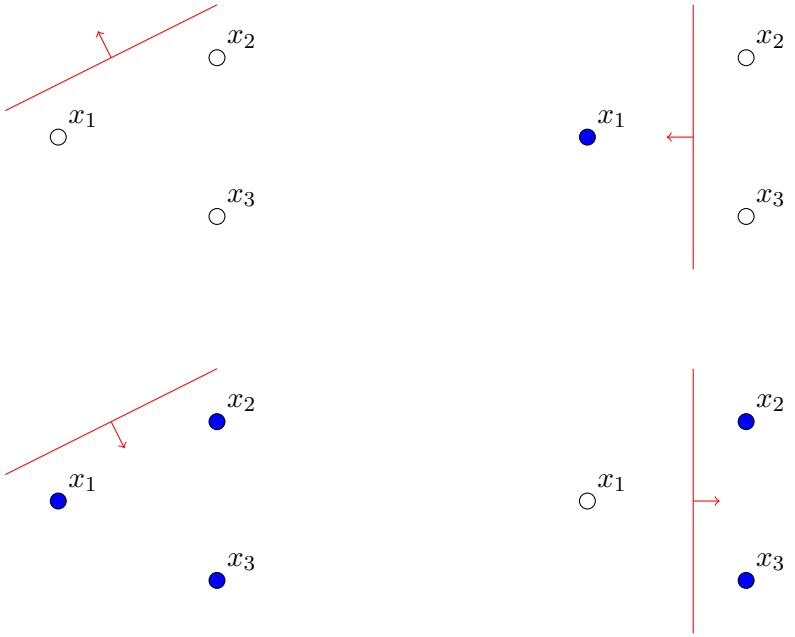
No AAR classifier can label this configuration.

□

2.2.2 Hyperplanes

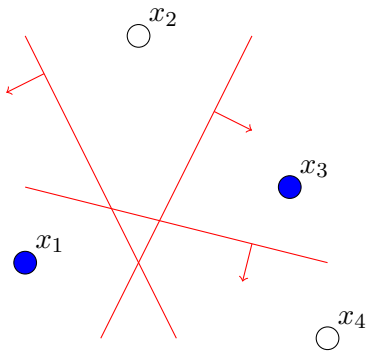
The VC dimension of hyperplanes in dimension d is $d + 1$. The proof can be done by induction on the dimension.

Example for $d = 2$



This configuration is shattered by the class of hyperplanes in dimension 2.

However, a configuration that would include 4 points cannot be shattered by the class of hyperplanes of dimension 2 :



No hyperplane can handle the XOR situation.

Definition 4. *Growth function*
The growth function $\Pi_{\mathcal{F}} : \mathbb{N} \to \mathbb{N}$ *is defined as*

$$\Pi_{\mathcal{F}}(n) = \max_{S \subseteq \mathcal{X}, |S|=n} |\mathcal{F}_S|$$

Note : *If VC-dim*$(\mathcal{F}) = d$*, then* $\forall n \leq d, \Pi_{\mathcal{F}}(n) = 2^n$

3 VC dimension and generalization error bound

Let $\mathcal{F} \in \{-1,1\}^{\mathcal{X}}$ such that $d = \text{VC-dim}(\mathcal{F}) < +\infty$

Theorem 2. *With a probability at least* $1-\delta$ *:*

$$\forall f \in \mathcal{F}, \mathcal{R}(F,\mathcal{D}) \leq \hat{\mathcal{R}}_n(f,S) + \sqrt{\frac{2d\ln\left(\frac{\mathbf{e}n}{d}\right)}{n}} + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{8}\right)}\right)$$

Where $\ln(\mathbf{e}) = 1$

Reminder : With the Rademacher complexity,

$$\forall f \in \mathcal{F}, \mathcal{R}(F,\mathcal{D}) \leq \hat{\mathcal{R}}_n(f,S) + \hat{Rad}(\mathcal{F},S) + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{8}\right)}\right)$$

The proof of the Theorem is composed of :
— Massart's Lemma
— Bounding the growth function using the Rademacher complexity (somewhat anti-chronological, since the Rademacher complexity appeared later)
— Sauer's Lemma, which will not be proven here

3.1 Massart's Lemma

Lemma 1. *Massart's Lemma*

Let $A \subseteq \mathbb{R}^n$, $\sigma_1, ..., \sigma_n$ *independant Rademacher variables (*$\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$*).*
With $r = \sup_{a\in A} ||a||_2$*, we have :*

$$\mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{a\in A}\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right] \leq r\frac{\sqrt{2\ln(|A|)}}{n}$$

Proof of Massart's Lemma.

$$\begin{aligned}
\exp\left(\lambda\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right]\right) &\leq \mathbb{E}_\sigma\left[\exp\left(\lambda\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right)\right] && \text{(Jensen inequality)}\\
&= \mathbb{E}_\sigma\left[\sup_{a\in A}\exp\left(\lambda\sum_{i=1}^{n}\sigma_i a_i\right)\right] && \text{(exp is increasing)}\\
&\leq \mathbb{E}_\sigma\left[\sum_{a\in A}\exp\left(\lambda\sum_{i=1}^{n}\sigma_i a_i\right)\right] && \text{(exp is positive)}\\
&= \sum_{a\in A}\mathbb{E}_\sigma\left[\exp\left(\lambda\sum_{i=1}^{n}\sigma_i a_i\right)\right]\\
&= \sum_{a\in A}\mathbb{E}_\sigma\left[\prod_{i=1}^{n}\exp(\lambda\sigma_i a_i)\right]\\
&= \sum_{a\in A}\prod_{i=1}^{n}\mathbb{E}_{\sigma_i}[\exp(\lambda\sigma_i a_i)] && \text{(independence of the } \sigma_i)\\
&= \sum_{a\in A}\prod_{i=1}^{n}\left(\frac{1}{2}\exp(-\lambda a_i) + \frac{1}{2}\exp(\lambda a_i)\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{a\in A} \prod_{i=1}^{n} \exp\left(\frac{\lambda^2 a_i^2}{2}\right) && \cosh(z) \leq \exp\left(\frac{z^2}{2}\right)\\
&= \sum_{a\in A} \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^{n} a_i^2\right)\\
&= \sum_{a\in A} \exp\left(\frac{\lambda^2}{2}||a||_2^2\right)\\
&\leq \sum_{a\in A} \exp\left(\frac{\lambda^2}{2}r^2\right)\\
&= |A|\exp\left(\frac{\lambda^2}{2}r^2\right)
\end{aligned}
$$

Thus we obtain :

$$
\begin{aligned}
\exp\left(\lambda\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right]\right) &\leq |A|\exp\left(\frac{\lambda^2}{2}r^2\right)\\
\lambda\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right] &\leq \ln(|A|) + \frac{\lambda^2}{2}r^2\\
\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right] &\leq \frac{\ln(|A|)}{\lambda} + \frac{\lambda}{2}r^2
\end{aligned}
$$

We must find a λ that minimizes

$$\frac{\ln(|A|)}{\lambda} + \frac{\lambda}{2}r^2$$

Derivative :

$$-\frac{\ln(|A|)}{\lambda^2} + \frac{r^2}{2}$$

$$
\begin{aligned}
-\frac{\ln(|A|)}{\lambda^2} + \frac{r^2}{2} = 0 &\iff \frac{\lambda^2 r^2}{2} = \ln(|A|)\\
&\iff \lambda^2 = \frac{2\ln(|A|)}{r^2}\\
&\iff \lambda = \pm\frac{\sqrt{2\ln(|A|)}}{r}
\end{aligned}
$$

We inject it into the bound (only $\lambda > 0$ is a minimizer) :

$$
\begin{aligned}
\frac{\ln(|A|)r}{\sqrt{2\ln(|A|)}} + \frac{\sqrt{2\ln(|A|)}}{2r}r^2 &= \sqrt{\frac{\ln(|A|)}{2}}r + \sqrt{\frac{\ln(|A|)}{2}}r\\
&= r\sqrt{2\ln(|A|)}
\end{aligned}
$$

Dividing both sides by n, we get :

$$\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{a\in A}\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right] \leq r\frac{\sqrt{2\ln(|A|)}}{n}$$

□

3.2 Bounding the growth function using the Rademacher complexity

Lemma 2.

$$\hat{Rad}(\mathcal{F}, S) \leq \sqrt{\frac{2\ln(\Pi_{\mathcal{F}}(n))}{n}}$$

With

$$\hat{Rad}(\mathcal{F}, S) = \mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right]$$

Where $\sigma_1, ..., \sigma_n$ are independent Rademacher variables, and $S = \{x_1, ..., x_n\}$

Clarification of $\mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right] = \mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{a\in\mathcal{F}_S} \frac{1}{n}\sum_{i=1}^{n} \sigma_i a_i\right]$

— $\mathcal{F} \subseteq \{-1, 1\}^{\mathcal{X}}$ is set beforehand : we want to measure the capacity of this given set of functions.

— $\mathcal{F}_S$ is defined (1) as $\{(f(x_1), ..., f(x_n)), f \in \mathcal{F}\} \subseteq \{-1, 1\}^n$
We have $|\mathcal{F}_S| \leq 2^n$ since $|\{-1, 1\}^n| = 2^n$
Example for $n = 5$:

$$\begin{aligned}\mathcal{F}_S = \{&(-1,-1,+1,-1,+1),\\ &(-1,+1,-1,+1,+1),\\ &(+1,+1,+1,+1,+1),\\ &(-1,+1,+1,+1,+1),\\ &(+1,+1,-1,-1,-1),\\ &(+1,+1,-1,+1,+1)\}\end{aligned}$$

We have $|\mathcal{F}_S| = 6$ in this example.

— $\hat{Rad}(\mathcal{F}, S) = \mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right]$
By definition of $\mathcal{F}_S$:

$$\forall f \in \mathcal{F}, (f(x_1), ..., f(x_n)) \in \mathcal{F}_S$$

Thus :

$$\begin{aligned}\hat{Rad}(\mathcal{F}, S) &= \mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right]\\ &= \mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{a\in\{(f(x_1),...,f(x_n)),f\in\mathcal{F}\}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i a_i\right]\\ &= \mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{a\in\mathcal{F}_S} \frac{1}{n}\sum_{i=1}^{n} \sigma_i a_i\right]\end{aligned}$$

Proof of Lemma 2.

$$\hat{Rad}(\mathcal{F}, S) = \mathbb{E}_{\sigma_1,...,\sigma_n}\left[\sup_{a\in\mathcal{F}_S} \frac{1}{n}\sum_{i=1}^{n} \sigma_i a_i\right]$$

Since $\mathcal{F} \subseteq \{-1, 1\}^{\mathcal{X}}$, we have $\forall a \in \mathcal{F}_S, \forall 1 \leq i \leq n, a_i^2 = 1$
Thus $\forall a \in \mathcal{F}_S, ||a||_2 = \sqrt{\sum_{i=1}^{n} a_i^2} = \sqrt{n}$

$$
\begin{aligned}
\hat{Rad}(\mathcal{F}, S) &= \mathbb{E}_{\sigma_1,\dots,\sigma_n}\left[\sup_{a\in\mathcal{F}_S}\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right] && \text{(Massart's Lemma (1))}\\
&\leq \sqrt{n}\frac{\sqrt{2\ln(|\mathcal{F}_S|)}}{n}\\
&\leq \sqrt{\frac{2\ln(|\mathcal{F}_S|)}{n}}\\
&\leq \sqrt{\frac{2\ln(|\Pi_F(n)|)}{n}}
\end{aligned}
$$

□

3.3 Sauer's Lemma

Lemma 3. *Sauer's Lemma*
Let $\mathcal{F} \subseteq \{-1,1\}^{\mathcal{X}}$ such that VC-dim$(\mathcal{F}) \leq d < +\infty$. Then, $\forall n \geq d$:

$$\Pi_{\mathcal{F}}(n) \leq \sum_{i=1}^{d}\binom{n}{i} \leq \left(\frac{\mathbf{e}n}{d}\right)^d$$

Where $\ln(\mathbf{e}) = 1$, *and*

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}$$

3.4 Gluing the proof together

Proof of Theorem 2.

$$
\begin{aligned}
\Pi_{\mathcal{F}}(n) &\leq \left(\frac{\mathbf{e}n}{d}\right)^d && \text{(Sauer's Lemma)}\\
\ln(\Pi_{\mathcal{F}}(n)) &\leq \ln\left(\left(\frac{\mathbf{e}n}{d}\right)^d\right) && (x \mapsto \ln(x) \text{ increasing})\\
\frac{2\ln(\Pi_{\mathcal{F}}(n))}{n} &\leq \frac{2d\ln\left(\frac{\mathbf{e}n}{d}\right)}{n} && (\frac{2}{n} > 0)\\
\sqrt{\frac{2\ln(\Pi_{\mathcal{F}}(n))}{n}} &\leq \sqrt{\frac{2d\ln\left(\frac{\mathbf{e}n}{d}\right)}{n}} && (x \mapsto \sqrt{x} \text{ increasing})
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{R}(F,\mathcal{D}) &\leq \hat{\mathcal{R}}_n(f,S) + \hat{Rad}(\mathcal{F},S) + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{8}\right)}\right) && \text{(Rademacher bound)}\\
&\leq \hat{\mathcal{R}}_n(f,S) + \sqrt{\frac{2\ln(\Pi_{\mathcal{F}}(n))}{n}} + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{8}\right)}\right) && \text{(Lemma 2)}\\
&\leq \hat{\mathcal{R}}_n(f,S) + \sqrt{\frac{2d\ln\left(\frac{\mathbf{e}n}{d}\right)}{n}} + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{8}\right)}\right)
\end{aligned}
$$

□