## II/ Application to classes $\mathcal{H}$ with finite VC-dimension

Recap': we proved the following result

**Theorem**: the ERM $\hat{h}_n \in \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \neq h(X_i)}$ satisfies

$$L(\hat{h}_n) - \inf_{h \in \mathcal{H}} L(h) \leq 4\sqrt{\frac{2\log(2S_n(\mathcal{H}))}{n}} + \sqrt{\frac{2\log 1/\delta}{n}} \quad \text{with prob} \geq 1-\delta \quad (1)$$

$$\mathbb{E}\left[L(\hat{h}_n)\right] - \inf_{h \in \mathcal{H}} L(h) \leq 4\sqrt{\frac{2\log(2S_n(\mathcal{H}))}{n}} \quad (2)$$

where $L(h) = \mathbb{P}(Y \neq h(X)) = \int \mathbb{1}_{y \neq h(x)} d\mathbb{P}^{(X,Y)}(x,y)$ ("risk of $h$")

and $S_n(\mathcal{H}) = \max_{(x_1,\dots,x_n) \in \mathcal{X}^n} \left| \{(h(x_1),\dots,h(x_n)) : h \in \mathcal{H}\} \right|$ ("shattering coefficient")

We proved directly (1) but (2) is even more direct.
Next we show a special case where the shattering coefficient $S_n(\mathcal{H})$ can be upper bounded in a simple fashion. We need a new notion:

**Definition**: the Vapnik-Chervonenkis dimension of $\mathcal{H}$ (or "VC-dimension") is defined by

$$V := \sup\{k \in \mathbb{N} : S_k(\mathcal{H}) = 2^k\} \quad (S_0(\mathcal{H}) := 1 \text{ by convention})$$

$$= \sup\left\{k \in \mathbb{N} : \exists (x_1,\dots,x_k) \in \mathcal{X}^k, \left|\{(h(x_1),\dots,h(x_k)) : h \in \mathcal{H}\}\right| = 2^k\right\}$$

$$\in \mathbb{N} \cup \{+\infty\}.$$

It is the maximal number of points that can be labelled in all possible ways by classifiers in $\mathcal{H}$.

Examples: ($\mathcal{X} = \mathbb{R}^d$ in both examples)

- affine classifiers: $\mathcal{H} = \{h(x) = \text{sign}(\langle a, x\rangle + b),\ (a,b) \in \mathbb{R}^d \times \mathbb{R}\}$
  has VC-dimension equal to $d+1$
- convex polygonal classifiers: $\mathcal{H} = \{h(x) = 2\mathbb{1}_A(x) - 1 : A \text{ convex polytope of } \mathbb{R}^d\}$
  has VC-dimension equal to $+\infty$.

A useful combinatorial result is:

**Sauer's lemma**

Let $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ be a set of classifiers with finite VC-dimension $V \in \mathbb{N}$. Then, the shattering coefficients $S_n(\mathcal{H})$ can be upper bounded as

$$S_n(\mathcal{H}) \leq \sum_{k=0}^{V} \binom{n}{k} \leq (n+1)^V$$

↑ a tighter upper bound of $\left(\frac{en}{V}\right)^V$ can also be proved.

Substituting the bound in Sauer's lemma into the excess risk bounds (1) and (2), we obtain:

**Corollary**: if $\mathcal{H}$ has VC-dimension $V \in \mathbb{N}^*$, then the ERM $\hat{h}_n$ satisfies

$$L(\hat{h}_n) - \inf_{h \in \mathcal{H}} L(h) \leq 8\sqrt{\frac{V \log(n+1)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \quad \text{with prob} \geq 1-\delta \qquad (1')$$

$$\mathbb{E}\left[L(\hat{h}_n)\right] - \inf_{h \in \mathcal{H}} L(h) \leq 8\sqrt{\frac{V \log(n+1)}{n}} \qquad (2')$$

In the next section we address the optimality of (2').

# III - Minimax lower bound for the misclassification excess risk with VC-classes

In this section, we prove that the risk bound of order $\sqrt{\frac{V \log n}{n}}$ is optimal in the worst case (up to log factors).

**Theorem 3**

We consider the classification problem with input space $\mathcal{X}$ and output space $\{-1,1\}$.

Let $\mathcal{G} \subseteq \{-1,1\}^{\mathcal{X}}$ be a class of classifiers with finite VC-dimension $V \in \mathbb{N}^*$.

Then, for all $n \geq c_1 V$,

$$\inf_{\hat{g}_n} \sup_{P \in \mathcal{P}_1^+(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[R(\hat{g}_n)\right] - \inf_{g \in \mathcal{G}} R(g) \right\} \geq c_2 \sqrt{\frac{V}{n}}$$

where $c_1, c_2 > 0$ are absolute constants.

Our proof is inspired from Massart and Nédélec (A.o.S '06) who actually proved a finer lower bound in terms of the margin $h = \operatorname{ess\,inf} |2\mathbb{P}(Y=1|X) - 1|$. Here, we don't care about the margin, but the interested reader can note that we consider the particular case $h \approx \sqrt{\frac{V}{n}}$.

**Proof:** As in Section II, we restrict the sup over an appropriately chosen finite set $P$, which we define next:

- Since $\mathcal{G}$ has VC-dimension equal to $V$, there exists $\{x_1, \dots, x_V\} \subseteq \mathcal{X}$ such that $\mathcal{G}$ shatters $\{x_1, \dots, x_V\}$, i.e., $\left|\{(g(x_1), \dots, g(x_V)) : g \in \mathcal{G}\}\right| = 2^V$. In particular, $x_i \neq x_j$ for all $1 \leq i \neq j \leq V$.
- Denote by $\mu$ the uniform probability measure over $\{x_1, \dots, x_V\}$.

- Define, for all $b \in \{-1,1\}^V$, the function $p_b$ on $\{x_1, \dots, x_V\}$ by

$$p_b(x_i) = \frac{1}{2}(1 + b_i h) \quad , 1 \le i \le V.$$

where $h \in (0,1)$ will be chosen later.

Now, for all $b \in \{-1,1\}^V$, define $P_b$ as the joint probability distribution on $\mathcal{X} \times \{-1,1\}$ such that, under $P_b$, $X$ has distribution $\mu$ and $Y | X = x \sim p_b(x)\,\delta_1 + (1 - p_b(x))\,\delta_{-1}$, for all $x \in \{x_1, \dots, x_V\}$.

(In other words: for all nonnegative measurable mapping $\varphi : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}_P\left[\varphi(X)\, \mathbb{1}_{\{Y=1\}}\right] = \frac{1}{V} \sum_{i=1}^{V} \varphi(x_i)\, p_b(x_i)$$

$$\mathbb{E}_P\left[\varphi(X)\, \mathbb{1}_{\{Y=-1\}}\right] = \frac{1}{V} \sum_{i=1}^{V} \varphi(x_i)\, (1 - p_b(x_i))$$

Note that the Bayes classifier $t_b^* : \mathcal{X} \to \{-1,1\}$ under $P_b$ is given by

$$\begin{cases} +1 & \text{if } p_b(x_i) \ge \frac{1}{2} \\ -1 & \text{if } p_b(x_i) < \frac{1}{2} \end{cases} \quad , \quad 1 \le i \le V$$

i.e. $t_b^*(x_i) = b_i$ for all $1 \le i \le V$.

Note also that, since $\mathcal{G}$ shatters $\{x_1, \dots, x_V\}$:

$$\forall b \in \{-1,1\}^V, \ \exists g_b \in \mathcal{G} : (g_b(x_1), \dots, g_b(x_V)) = b.$$

Therefore, $g_b$ is the Bayes classifier under $P_b$, so that

$$R(g_b) = \inf_{\substack{\text{measurable} \\ g : \mathcal{X} \to \{-1,1\}}} R(g) = \inf_{g \in \mathcal{G}} R(g) \qquad (5)$$

We are now ready to derive the lower bound.

Let $\mathcal{H} = 2\Gamma - 1 = \{2x - 1 : x \in \Gamma\} \subseteq \{-1,1\}^V$ be the set provided by Vershamov-Gilbert's lemma rescaled to $\{-1,1\}^V$.

We assume that $V \ge 6$ so that $|\mathcal{H}| \ge e^{6/8} \ge 2$.

$$\inf_{\hat{g}_n} \sup_{P \in \mathcal{P}_\gamma^+(\mathcal{X}\times\{-1,1\})} \left\{ \mathbb{E}[R(\hat{g}_n)] - \inf_{g \in \mathcal{G}} R(g) \right\}$$

$$\geq \inf_{\hat{g}_n} \max_{b \in \mathcal{H}} \left\{ \mathbb{E}\left[R(\hat{g}_n)\right] - \underbrace{\inf_{g \in \mathcal{G}} R(g)}_{= \inf_{\substack{\text{measurable} \\ g : \mathcal{X} \to \{-1,1\}}} R(g) \text{ by (5)}} \right\}$$

$$= \mathbb{E}_b\left[ |2\eta_{P_b}(X) - 1| \cdot \mathbb{1}_{\{\hat{g}_n(X) \neq g_b(X)\}} \right]$$

($\mathbb{E}_b$ denotes expectation when $P_b$ is the underlying prob. distribution)

$$\geq h \inf_{\hat{g}_n} \max_{b \in \mathcal{H}} \mathbb{E}_b\left[ \frac{|\hat{g}_n(X) - g_b(X)|}{2} \right] \quad \text{since } |2\eta_{P_b} - 1| \geq h \text{ by construction}$$

and $\mathbb{1}_{u \neq v} = \frac{|u-v|}{2}$ for all $u, v \in \{-1,1\}$

$$= \frac{h}{2} \inf_{\hat{g}_n} \max_{b \in \mathcal{H}} \mathbb{E}_b\left[ \|\hat{g}_n - g_b\|_{L^1(\mu)} \right] \quad (6)$$

Let $\hat{b}_n \in \operatorname{argmin}_{b \in \mathcal{H}} \|\hat{g}_n - g_b\|_{L^1(\mu)}$ ($\hat{b}_n$ is a random variable)

As previously, we get from the triangle inequality that

$$\left\{ \hat{b}_n \neq b \right\} \subseteq \left\{ \|\hat{g}_n - g_b\|_{L^1(\mu)} \geq \frac{\varepsilon}{2} \right\}$$

where $\varepsilon > 0$ is such that $\|g_{b_1} - g_{b_2}\|_{L^1(\mu)} > \varepsilon$ for all $b_1 \neq b_2 \in \mathcal{H}$.

(Otherwise, we would have $\|g_{\hat{b}_n} - g_b\|_{L^1(\mu)} \leq \|g_{\hat{b}_n} - \hat{g}_n\|_{L^1(\mu)} + \|\hat{g}_n - g_b\|_{L^1(\mu)}$

$\leq 2\|\hat{g}_n - g_b\|_{L^1(\mu)}$ by def of $\hat{b}_n$

$< \varepsilon$, which is impossible.)

Here, we on the $\varepsilon = \frac{1}{2}$ because

[see below]

$\forall b \neq b' \in \mathcal{H}$, $b = 2x-1$, $b' = 2x'-1$, $x \neq x' \in \Gamma$,

$$\|g_b - g_{b'}\|_{L^1(\mu)} = \frac{1}{V} \sum_{i=1}^{V} |g_b(x_i) - g_{b'}(x_i)|$$

$$= \frac{1}{V} \sum_{i=1}^{V} |b_i - b'_i| \quad \text{by def of } g_b \text{ and } g_{b'}$$

$$= \frac{2}{V} \sum_{i=1}^{V} \mathbb{1}_{\{b_i \neq b'_i\}}$$

$$> \frac{2}{V} \cdot \frac{V}{4} = \frac{1}{2},$$

Getting back to (6), we obtain

$$\inf_{\hat{g}_n} \sup_{P} \left\{ \mathbb{E}[R(\hat{g}_n)] - \inf_{g \in \mathcal{G}} R(g) \right\} \geq \frac{h}{2} \inf_{\hat{g}_n} \max_{b \in \mathcal{H}} \frac{\varepsilon}{2} \mathbb{P}_b(\hat{b}_n \neq b)$$

$$= \frac{h}{8} \left( 1 - \sup_{\hat{g}_n} \min_{b \in \mathcal{H}} \mathbb{P}_b(\hat{b}_n = b) \right)$$

Fix $\hat{g}_n$. We use Fano's inequality to upper bound $\min_{b \in \mathcal{H}} \mathbb{P}_b(\hat{b}_n = b)$:

- We have $|\mathcal{H}| \geq 2$ since we assumed that $V \geq 6$.
- Let $\tilde{b} \in \mathcal{H}$. We have: $\forall b \in \mathcal{H} \setminus \{\tilde{b}\}$,

$$KL\left(P_b^{\otimes n}, P_{\tilde{b}}^{\otimes n}\right) = n \cdot KL\left(P_b, P_{\tilde{b}}\right)$$

$$= n \left( KL(\mu, \mu) + \int_{\mathcal{X}} d\mu(x) \, KL\left(\mathcal{B}(p_b(x)), \mathcal{B}(p_{\tilde{b}}(x))\right) \right)$$

by the chain rule for the Kullback-Leibler entropy, and since under $P_b$, $Y | X = x \sim p_b(x) \, \delta_1 + (1 - p_b(x)) \, \delta_{-1}$

$$= \frac{n}{V} \sum_{i=1}^{V} \underbrace{KL\left(\mathcal{B}(p_b(x_i)), \mathcal{B}(p_{\tilde{b}}(x_i))\right)}_{\text{if } b_i \neq \tilde{b}_i, \text{ this quantity is equal to either } KL\left(\mathcal{B}\left(\frac{1+h}{2}\right), \mathcal{B}\left(1-\frac{h}{2}\right)\right) \text{ or } KL\left(\mathcal{B}\left(1-\frac{h}{2}\right), \mathcal{B}\left(\frac{1+h}{2}\right)\right)} \mathbb{1}_{\{b_i \neq \tilde{b}_i\}}$$

$p_b(x_i) \neq p_{\tilde{b}}(x_i) \iff b_i \neq \tilde{b}_i$

i.e. either $KL(\mathcal{B}(q), \mathcal{B}(1-q))$ or $KL(\mathcal{B}(1-q), \mathcal{B}(q))$, with $q := \frac{1+h}{2}$

But these two quantities are equal to $(2q-1)\log\frac{q}{1-q} = h\log\left(\frac{1+h}{1-h}\right)$

Hence:

$$KL\left(P_b^{\otimes n}, P_{\tilde{b}}^{\otimes n}\right) = \frac{n}{V}\sum_{i=1}^{V} h\log\left(\frac{1+h}{1-h}\right)\mathbb{1}_{\{b_i \neq \tilde{b}_i\}}$$

$$\leqslant \frac{2nh^2}{1-h}\cdot\frac{1}{V}\sum_{i=1}^{V}\mathbb{1}_{\{\tilde{b}_i \neq b_i\}}$$

$$\leqslant 4nh^2 \quad \text{if } h \in (0, \tfrac{1}{2}]. \text{ (condition to be checked later)}$$

Applying Fano's inequality, we get: [see footnote]

$$\min_{b\in\mathcal{H}} P_b\left(\hat{b}_n = b\right) \leqslant \max\left\{\frac{2e}{2e+1}, \frac{4nh^2}{\ln|\mathcal{H}|}\right\}$$

$$\leqslant \max\left\{\frac{2e}{2e+1}, \frac{32nh^2}{V}\right\} \quad \text{since } \ln|\Gamma| \geqslant \frac{V}{8}$$

Choosing $h = \frac{1}{4}\sqrt{\frac{e}{2e+1}\cdot\frac{V}{n}}$, we get $h \in [0, \frac{1}{2}]$ when $n \geqslant \underbrace{\frac{e}{4(2e+1)}}_{c_1}\cdot V$, and:

$$\inf_{\hat{g}_n}\sup_{P\in\mathcal{C}_1^+(\mathcal{X}\times\{-1,1\})}\left\{\mathbb{E}[R(\hat{g}_n)] - \inf_{g\in\mathcal{G}} R(g)\right\} \geqslant \frac{h}{8}\left(1 - \frac{2e}{2e+1}\right)$$

$$\geqslant \frac{1}{32}\sqrt{\frac{2e}{(2e+1)^3}\cdot\frac{V}{n}}.$$

The case $V \in \{1, 2, \dots, 5\}$ can be handled similarly (easier actually!) via Pinsker's inequality. The constants $c_1$ and $c_2$ may change a little. ∎

Remark: I used a slightly different version of Fano's inequality than the one we studied in Lecture 2, namely:

Fano's inequality (Birgé's version)

Assume • the events $A_1, \dots, A_N$ form a partition of $\Omega$ (with $N \geqslant 2$)
• $P_1, \dots, P_N$ are probability distributions on $(\Omega, \mathcal{F})$.

Then,

$$\min_{1\leqslant j\leqslant N} P_j(A_j) \leqslant \max\left\{\frac{2e}{2e+1}, \frac{\overline{KL}}{\log N}\right\} \quad \text{where } \overline{KL} := \frac{1}{N-1}\sum_{\substack{j=2\\ \equiv}}^{N} KL(P_j, P_1).$$

Exercise: prove a similar lower bound with the version of Fano involving $\log(2)$.