
Travaux dirigés n° 2 : Concentration, théorie de VC

Stephan Clémençon <stephan.clemencon@telecom-paris.fr>
Ekhine Irurozki <irurozki@telecom-paris.fr>

Exercice 1. On se place dans le cadre de la classification binaire : soient un descripteur aléatoire X à valeurs dans un espace mesurable $\mathcal{X} \subset \mathbb{R}^d$ ($d \in \mathbb{N}^*$) et un label aléatoire Y valant -1 ou 1. On considère une classe finie $\mathcal{G}$ de classifieurs $\mathcal{X} \longrightarrow \{-1, 1\}$ telle que les deux labels sont parfaitement séparables par un élément de $\mathcal{G}$, *i.e.* $\min_{g \in \mathcal{G}} \mathrm{L}(g) = 0$ pour le risque $\mathrm{L} : g \in \mathcal{G} \mapsto \mathbb{P}\left(g(\mathrm{X}) \neq \mathrm{Y}\right) \in [0, 1]$.

Soit $n \in \mathbb{N}^*$. On suppose que l'on dispose d'un échantillon i.i.d. $\{(\mathrm{X}_i, \mathrm{Y}_i)\}_{1 \leq i \leq n}$ suivant la même loi que (X, Y) et on note $\hat{g}_n$ un minimiseur de l'erreur empirique de classification :

$$\hat{g}_n \in \min_{g \in \mathcal{G}} \mathrm{L}_n(g) \quad \text{où} \quad \mathrm{L}_n : g \in \mathcal{G} \mapsto \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{g(\mathrm{X}_i) \neq \mathrm{Y}_i\}}.$$

1) Montrer que $\min_{g \in \mathcal{G}} \mathrm{L}_n(g) = 0$ presque-sûrement.
2) Montrer que $\mathbb{P}\left(\mathrm{L}(\hat{g}_n) > \epsilon\right) \leq |\mathcal{G}|(1-\epsilon)^n$ pour tout $\epsilon \in [0, 1]$.
   En déduire que $\mathbb{P}\left(\mathrm{L}(\hat{g}_n) > \epsilon\right) \leq |\mathcal{G}| e^{-n\epsilon}$ pour tout $\epsilon > 0$.

   Indication. Utiliser $\mathcal{G}_{\mathrm{B}} := \{g \in \mathcal{G} : \mathrm{L}(g) > \epsilon\}$ ainsi qu'une borne d'union.
3) Déduire de la question précédente que $\mathbb{E}\left(\mathrm{L}(\hat{g}_n)\right) \leq \dfrac{\log(e|\mathcal{G}|)}{n}$.

   Indication. Pour toute variable aléatoire Z positive, $\mathbb{E}\left(\mathrm{Z}\right) = \int_0^{+\infty} \mathbb{P}(\mathrm{Z} > t)\, dt$.

Solution:

1) Avant toute chose, rappelons que l'ensemble image (aléatoire) de L_n étant inclus par construction dans l'ensemble fini (déterministe) : $\{\frac{k}{n} \in \{0, n\}\}$, la variable aléatoire $\min_{g \in \mathcal{G}} \mathrm{L}_n(g)$ existe toujours. Il nous faut ici montrer que ce minimum est presque-sûrement nul.
   Pour cela, remarquons que par hypothèse, il existe $g^* \in \mathrm{G}$ tel que $\mathrm{L}(g^*) = \mathbb{P}(g^*(\mathrm{X}) \neq \mathrm{Y}) = 0$. Par conséquent,

$$\begin{aligned}\mathbb{P}(\mathrm{L}_n(g^*) = 0) = \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{\{g^*(\mathrm{X}_i) \neq \mathrm{Y}_i\}} = 0\right) &= \mathbb{P}(g^*(\mathrm{X}_1) = \mathrm{Y}_1, \ldots, g^*(\mathrm{X}_n) = \mathrm{Y}_n) \\ &= \mathbb{P}(g^*(\mathrm{X}) = \mathrm{Y})^n = (1 - \mathrm{L}(g^*))^n = 1.\end{aligned}$$

   Cela veut dire qu'avec probabilité 1, l'ensemble $\{\mathrm{L}_n(g) : g \in \mathcal{G}\}$ admet bien 0 pour minimum. En d'autres termes, $\min_{g \in \mathcal{G}} \mathrm{L}_n(g) = 0$ presque-sûrement.
2) (a) Remarquons tout d'abord que la variable aléatoire $\mathrm{L}(\hat{g}_n)$ est par construction à valeurs dans $[0, 1]$, donc si $\epsilon \geq 1$ on a directement $\mathbb{P}(\mathrm{L}(\hat{g}_n) > \epsilon) = 0$.

Supposons maintenant $\epsilon \in [0,1[$ et posons $\mathcal{G}_\epsilon := \{g \in \mathcal{G} : \mathrm{L}(g) > \epsilon\}$, de telle manière que $\mathbb{P}(\mathrm{L}(\hat{g}_n) > \epsilon) = \mathbb{P}(\hat{g}_n \in \mathcal{G}_\epsilon)$. D'après la question précédente, $\hat{g}_n$ étant un minimiseur de l'erreur empirique de classification, on a $\mathrm{L}_n(\hat{g}_n) = 0$ presque-sûrement. Ainsi,

$$\begin{aligned}
\mathbb{P}(\mathrm{L}(\hat{g}_n) > \epsilon) = \mathbb{P}(\hat{g}_n \in \mathcal{G}_\epsilon, \mathrm{L}_n(\hat{g}_n) = 0) \leq \mathrm{P}\left(\bigcup_{g \in \mathcal{G}_\epsilon} \{\mathrm{L}_n(g) = 0\}\right) &\leq \sum_{g \in \mathcal{G}_\epsilon} \mathbb{P}(\mathrm{L}_n(g) = 0) \\
&= \sum_{g \in \mathcal{G}_\epsilon} \mathbb{P}(g(\mathrm{X}) = \mathrm{Y})^n \quad \text{(comme en question 1)} \\
&= \sum_{g \in \mathcal{G}_\epsilon} (1 - \mathrm{L}(g))^n \leq \sum_{g \in \mathcal{G}_\epsilon} (1-\epsilon)^n \quad (\text{pour } g \in \mathcal{G}_\epsilon \text{ on a } \mathrm{L}(g) > \epsilon) \\
&\leq |\mathcal{G}_\epsilon|(1-\epsilon)^n \leq |\mathcal{G}|(1-\epsilon)^n \quad (\mathcal{G}_\epsilon \subset \mathcal{G} \text{ ensemble fini}).
\end{aligned}$$

Nous avons donc montré la première inégalité :

$$\mathbb{P}(\mathrm{L}(\hat{g}_n) > \epsilon) \leq |\mathcal{G}|(1-\epsilon)^n \mathbb{1}_{\{0 \leq \epsilon < 1\}}.$$

(b) Par convexité de la fonction exponentielle, on a $1 - \epsilon \leq e^{-\epsilon}$ avec $e^{-\epsilon} > 0$. Ainsi, on a toujours $(1-\epsilon)^n \mathbb{1}_{\{0 \leq \epsilon < 1\}} < e^{-n\epsilon}$. En partant de l'inégalité montrée à la question précédente, on obtient donc directement $\mathbb{P}(\mathrm{L}(\hat{g}_n) > \epsilon) \leq |\mathrm{G}|e^{-n\epsilon}$.

3) Comme la variable aléatoire $\mathrm{L}(\hat{g}_n)$ est positive, à valeurs dans $[0,1]$, on a

$$\mathbb{E}(\mathrm{L}(\hat{g}_n)) = \int_0^{+\infty} \mathbb{P}(\mathrm{L}(\hat{g}_n) > \epsilon)\, d\epsilon.$$

En utilisant la seconde inégalité montrée à la question précédente et le fait qu'une probabilité est toujours plus petite que 1, quel que soit $\varepsilon \in \mathbb{R}^+$, on a

$$\mathrm{P}(\mathrm{L}(\hat{g}_n) > \varepsilon) \leq \min\left\{1, |\mathcal{G}|e^{-n\varepsilon}\right\}.$$

Or, pour tout $\varepsilon \in \mathbb{R}^+$, on a $|\mathcal{G}|e^{-n\varepsilon} \leq 1$ si et seulement si $\varepsilon \geq \frac{1}{n}\ln(|\mathcal{G}|)$. Ainsi,

$$\begin{aligned}
\mathbb{E}(\mathrm{L}(\hat{g}_n)) &\leq \int_0^{\frac{\ln(|\mathcal{G}|)}{n}} 1\, d\epsilon + \int_{\frac{\ln(|\mathcal{G}|)}{n}}^{+\infty} |\mathcal{G}|e^{-n\epsilon}\, d\epsilon = \frac{\ln(|\mathcal{G}|)}{n} + |\mathcal{G}| \left[\frac{-e^{-n\epsilon}}{n}\right]_{\frac{\ln(|\mathcal{G}|)}{n}}^{+\infty} \\
&= \frac{\ln(|\mathcal{G}|)}{n} + \frac{|\mathcal{G}|}{n}\frac{1}{|\mathcal{G}|} = \frac{1}{n}\left(\ln(|\mathcal{G}|) + 1\right) = \frac{1}{n}\ln(e|\mathcal{G}|).
\end{aligned}$$

On retrouve donc bien l'inégalité recherchée.

Remarque. Par construction, la variable aléatoire $\mathrm{L}(\hat{g}_n)$ est bornée, à valeurs dans $[0,1]$. Il en est donc de même pour son espérance. La borne obtenue n'a donc d'intérêt que si $\frac{1}{n}\ln(e|\mathcal{G}|) \leq 1$, c'est-à-dire si $|\mathcal{G}| \leq e^{n-1}$.

En utilisant le caractère borné de $\mathrm{L}(\hat{g}_n)$, on peut même raffiner cette borne. Si $|\mathcal{G}| \leq e^n$, alors

$$\begin{aligned}
\mathbb{E}(\mathrm{L}(\hat{g}_n)) &\leq \int_0^{\frac{\ln(|\mathcal{G}|)}{n}} 1\, d\epsilon + \int_{\frac{\ln(|\mathcal{G}|)}{n}}^{1} |\mathcal{G}|e^{-n\epsilon}\, d\epsilon \\
&= \frac{1}{n}\left(\ln(|\mathcal{G}|) + 1 - |\mathcal{G}|e^{-n}\right).
\end{aligned}$$

On retrouve la première borne en remarquant que

$$\begin{aligned}\frac{1}{n}\left(\ln(|\mathcal{G}|)+1-|\mathcal{G}|e^{-n}\right) &\leq \frac{1}{n}\left(\ln(|\mathcal{G}|)+1-\ln(|\mathcal{G}|)e^{-n}\right) \quad (|\mathcal{G}|>\ln(|\mathcal{G}|))\\ &= \frac{1}{n}\left(\ln(|\mathcal{G}|)(1-e^{-n})+1\right)\\ &\leq \frac{1}{n}\ln(e|\mathcal{G}|). \quad \text{for } 1-e^{-n}\leq 1 \text{ pour } n\geq 0.\end{aligned}$$

Interprétation. Tous ces résultats indiquent qu'à nombre d'observations n fixé, on peut d'autant mieux contrôler l'erreur empirique de classification que la classe $\mathcal{G}$ considérée est restreinte, i.e. que son cardinal est faible.

Alternativement, plus on enrichit $\mathcal{G}$, plus il faut de données (i.e. d'information) pour garantir une faible erreur empirique de classification. En particulier, on déduit de la question 2 que pour tous $\delta \in]0,1[$ et $\epsilon > 0$, dès que $n \geq \frac{1}{\epsilon}\ln\frac{|\mathcal{G}|}{\delta}$, on a $\mathbb{P}(\mathrm{L}(\hat{g}_n) \leq \epsilon) \geq 1-\delta$.

Exercice 2. On se place dans le cadre de la classification binaire. On utilisera les mêmes notations que dans l'exercice précédent. On pose $\mathrm{L}^* := \mathrm{L}(g^*)$ avec $g^* : x \in \mathcal{X} \mapsto 2\mathbb{1}_{\{\eta(x)\geq 1/2\}} - 1$ et on note $\eta : x \in \mathcal{X} \mapsto \mathbb{P}(\mathrm{Y}=1 \mid \mathrm{X}=x) \in [0,1]$ la fonction de régression. Soit $(\eta_n)_{n\in\mathbb{N}^*}$ une suite de fonctions définies sur $\mathcal{X}$ à valeurs dans $]0,1[$. Pour tout $n \in \mathbb{N}^*$ on considère le classifieur $g_n : x \in \mathcal{X} \mapsto 2\mathbb{1}_{\{\eta_n(x)\geq 1/2\}} - 1$.

1) On suppose qu'il existe $\delta > 0$ tel que $|\eta(x) - 1/2| \geq \delta$ pour tout $x \in \mathcal{X}$. Montrer que

$$\mathrm{L}(g_n) - \mathrm{L}^* \leq \frac{2\,\mathbb{E}\left((\eta_n(\mathrm{X})-\eta(\mathrm{X}))^2\right)}{\delta}.$$

2) Montrer que si $\mathrm{L}^* = 0$, alors quel que soit $q \in [1,+\infty[$

$$\mathrm{L}(g_n) \leq 2^q\,\mathbb{E}\left(|\eta_n(\mathrm{X})-\eta(\mathrm{X})|^q\right).$$

Soient maintenant $\eta' : \mathcal{X} \to]0,1[$ et $g : x \in \mathcal{X} \mapsto 2\mathbb{1}_{\{\eta'(x)\geq 1/2\}} - 1$.

3) On suppose que $\mathbb{P}\{\eta'(\mathrm{X}) = 1/2\} = 0$ et que $\mathbb{E}\left(|\eta_n(\mathrm{X}) - \eta'(\mathrm{X})|\right) \to 0$ lorsque $n \to +\infty$. Montrer que $\mathrm{L}(g_n) \to \mathrm{L}(g)$ lorsque $n \to +\infty$.
4) On suppose que le label Y n'est plus observable, mais qu'une variable Z à valeurs dans $\{-1,+1\}$ l'est, telle que :

$$\begin{aligned}\mathbb{P}\left(\mathrm{Z}=1 \mid \mathrm{Y}=-1,\ \mathrm{X}\right) &= \mathbb{P}\left(\mathrm{Z}=1 \mid \mathrm{Y}=-1\right) = a < 1/2,\\ \mathbb{P}\left(\mathrm{Z}=-1 \mid \mathrm{Y}=1,\ \mathrm{X}\right) &= \mathbb{P}\left(\mathrm{Z}=-1 \mid \mathrm{Y}=1\right) = b < 1/2.\end{aligned}$$

On pose à présent $\eta' : x \in \mathcal{X} \mapsto \mathbb{P}\left(\mathrm{Z}=+1 \mid \mathrm{X}=x\right)$. Montrer que :

$$\mathrm{L}(g) \leq \mathrm{L}^*\left(1 + \frac{2\,|a-b|}{1-2\max(a,b)}\right).$$

Que peut-on en déduire lorsque $a = b$?

Solution:

1) Soit $n \in \mathbb{N}^*$. Tout d'abord, l'inégalité à démontrer est trivialement vraie si $L(g_n) = L^*$. Supposons donc que $L(g_n) \neq L^*$, i.e. que $L(g_n) - L^* > 0$ (par définition de L^*), alors

$$\begin{aligned}
L(g_n) - L^* &= \mathbb{P}(g_n(X) \neq Y) - \mathbb{P}(g^*(X) \neq Y) = \mathbb{E}\left(\mathbb{1}_{\{g_n(X)\neq Y\}} - \mathbb{1}_{\{g^*(X)\neq Y\}}\right) \\
&= \mathbb{E}\left(\mathbb{1}_{\{Y=-1\}}(\mathbb{1}_{\{g_n(X)=1\}} - \mathbb{1}_{\{g^*(X)=1\}}) + \mathbb{1}_{\{Y=1\}}(\mathbb{1}_{\{g_n(X)=-1\}} - \mathbb{1}_{\{g^*(X)=-1\}})\right) \\
&= \mathbb{E}\left((1-\eta(X))(\mathbb{1}_{\{g_n(X)=1\}} - \mathbb{1}_{\{g^*(X)=1\}}) + \eta(X)(\mathbb{1}_{\{g_n(X)=-1\}} - \mathbb{1}_{\{g^*(X)=-1\}})\right) \\
&= \mathbb{E}\left((2\eta(X)-1)(\mathbb{1}_{\{g_n(X)=-1\}} - \mathbb{1}_{\{g^*(X)=-1\}})\right)
\end{aligned}$$

Or pour tout $x \in \mathcal{X}$, si $g^*(x) = -1$ alors $2\eta(x) - 1 \leq 0$ et $\mathbb{1}_{\{g_n(x)=-1\}} - \mathbb{1}_{\{g^*(x)=-1\}} \leq 0$, puis si $g^*(x) = 1$ alors $2\eta(x) - 1 > 0$ et $\mathbb{1}_{\{g_n(x)=-1\}} - \mathbb{1}_{\{g^*(x)=-1\}} \geq 0$, d'où $(2\eta(x) - 1)(\mathbb{1}\{g_n(x) = -1\} - \mathbb{1}\{g^*(x) = -1\}) \geq 0$. Ainsi,

$$\begin{aligned}
L(g_n) - L^* &= \mathbb{E}\left[|2\eta(X)-1|\left|\mathbb{1}_{\{g_n(X)=-1\}} - \mathbb{1}_{\{g^*(X)=-1\}}\right|\right] \\
&= 2\mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right|\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}\right]. \qquad (1)
\end{aligned}$$

On remarque que pour tout $x \in X$, si $g_n(x) \neq g^*(x)$ alors $\eta_n(x)$ et $\eta(x)$ sont d'un côté et de l'autre de $\frac{1}{2}$, d'où $|\eta_n(x) - \eta(x)| \geq \left|\eta(x) - \frac{1}{2}\right|$. On en déduit que

$$\begin{aligned}
L(g_n) - L^* &\leq 2\mathbb{E}\left[|\eta_n(X) - \eta(X)|\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}\right] \qquad (2) \\
&\leq 2\mathbb{E}\left[(\eta_n(X) - \eta(X))^2\right]^{\frac{1}{2}}\mathbb{E}\left[\left(\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}\right)^2\right]^{\frac{1}{2}} \quad \text{(Cauchy-Schwartz).} \\
&= 2\mathbb{E}\left[(\eta_n(X) - \eta(X))^2\right]^{\frac{1}{2}}\mathbb{E}\left[\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}\right]^{\frac{1}{2}} \qquad (3)
\end{aligned}$$

En outre, par hypothèse, il existe $\delta > 0$ tel que $\left|\eta(x) - \frac{1}{2} \geq \delta\right|$ pour tout $x \in \mathcal{X}$. Alors en repartant de l'Eq. (1), on obtient

$$L(g_n) - L^* = 2\mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right|\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}\right] \geq 2\delta\mathbb{E}[\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}] \geq 0$$

d' ou

$$\mathbb{E}[\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}]^{\frac{1}{2}} \leq \left(\frac{1}{2\delta}(L(g_n) - L^*)\right)^{\frac{1}{2}}$$

En injectant ce dernier résultat dans l'Eq.(3), on obtient finalement

$$\begin{aligned}
& L(g_n) - L^* \leq 2\mathbb{E}\left[(\eta_n(X) - \eta(X))^2\right]^{\frac{1}{2}}\left(\frac{1}{2\delta}(L(g_n) - L^*)\right)^{1/2} \\
\Leftrightarrow \quad & (2\delta(L(g_n) - L^*))^{1/2} \leq 2\mathbb{E}\left[(\eta_n(X) - \eta(X))^2\right]^{\frac{1}{2}} \\
\Leftrightarrow \quad & L(g_n) - L^* \leq \frac{2}{\delta}\mathbb{E}\left[(\eta_n(X) - \eta(X))^2\right]. \quad \text{(tout est positif)}
\end{aligned}$$

2) Soit $n \in \mathbb{N}^*$ et supposons $L^* = 0$. Cela signifie que $g^*(X) = Y$ presque-sûrement. Comme à la question précédente, l'inégalité à démontrer est trivialement vraie si $L(g_n) = 0$. Supposons donc que $L(g_n) > 0$, alors en repartant de l'Eq. (2) et en appliquant l'inégalité de Hölder, pour tout $q \in [1, +\infty[$ on a

$$L(g_n) \leq 2\mathbb{E}\left[|\eta_n(X) - \eta(X)|^q\right]^{\frac{1}{q}}\mathbb{E}\left[\left(\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}\right)^{\frac{q}{q-1}}\right]^{\frac{q-1}{q}}$$

$$
\begin{aligned}
&= 2\mathbb{E}\left[|\eta_n(X) - \eta(X)|^q\right]^{\frac{1}{q}} \mathbb{E}\left[\mathbb{1}_{\{g_n(X)\neq g^*(X)\}}\right]^{\frac{q-1}{q}} \\
&= 2\mathbb{E}\left[|\eta_n(X) - \eta(X)|^q\right]^{\frac{1}{q}} \mathbb{E}\left[\mathbb{1}_{\{g_n(X)\neq Y\}}\right]^{\frac{q-1}{q}} \\
&= 2\mathbb{E}\left[|\eta_n(X) - \eta(X)|^q\right]^{\frac{1}{q}} L(g_n)^{\frac{q-1}{q}}
\end{aligned}
$$

Ainsi,

$$
\begin{aligned}
&L(g_n) \leq 2\mathbb{E}\left[|\eta_n(X) - \eta(X)|^q\right]^{\frac{1}{q}} L(g_n)^{\frac{q-1}{q}} \\
&\qquad \iff L(g_n)^{1-\frac{q-1}{q}} = L(g_n)^{\frac{1}{q}} \leq 2\mathbb{E}\left(|\eta_n(X) - \eta(X)|^q\right)^{1/q} \quad (L(g_n) > 0 \text{ par hypothèse}) \\
&\qquad \iff L(g_n) \leq 2^q\mathbb{E}\left(|\eta_n(X) - \eta(X)|^q\right). \quad (\text{Tout est positif})
\end{aligned}
$$

3) Soit $n \in \mathbb{N}^*$. En reprenant les mêmes calculs qu'à la question 1, on obtient

$$
\begin{aligned}
|L(g_n) - L(g)| &= 2\left|\mathbb{E}\left((\eta(X)-1)\left(\mathbb{1}_{\{g_n(X)=-1\}} - \mathbb{1}_{\{g(X)=-1\}}\right)\right)\right| \\
&\leq 2\mathbb{E}\left(|\eta(X)-1|\left|\mathbb{1}_{\{g_n(X)=-1\}} - \mathbb{1}_{\{g(X)=-1\}}\right|\right) \quad (\text{inégalité triangulaire}) \\
&\leq \mathbb{E}\left(\left|\mathbb{1}_{\{g_n(X)=-1\}} - \mathbb{1}_{\{g(X)=-1\}}\right|\right) \quad (|\eta(X)-1| \leq 2 \text{ par définition}) \\
&= \mathbb{E}\left(\mathbb{1}_{\{g_n(X)\neq g(X)\}}\right) = P(g_n(X) \neq g(X)) \\
&= P\left(\eta_n(X) > \frac{1}{2}, \eta(X) \leq \frac{1}{2}\right) + P\left(\eta_n(X) \leq \frac{1}{2}, \eta(X) > \frac{1}{2}\right)
\end{aligned}
$$

Or, quel que soit $\lambda \in \left[0, \frac{1}{2}\right]$, on a les relations d'événements suivantes :

$$
\begin{aligned}
\left\{\eta'(X) > \frac{1}{2}\right\} &= \left\{\eta'(X) \geq \frac{1}{2} + \lambda\right\} \cup \left\{\frac{1}{2} < \eta'(X) < \frac{1}{2} + \lambda\right\}, \text{ puis} \\
\left\{\eta_n(X) \leq \frac{1}{2}\right\} &\cap \left\{\eta'(X) \geq \frac{1}{2} + \lambda\right\} \subset \left\{|\eta_n(X) - \eta'(X)| \geq \lambda\right\}, \\
\left\{\eta_n(X) \leq \frac{1}{2}\right\} &\cap \left\{1 < \eta'(X) < \frac{1}{2} + \lambda\right\} \subset \left\{\left|\eta'(X) - \frac{1}{2}\right| < \lambda\right\}.
\end{aligned}
$$

d'où

$$
\mathbb{P}\left(\eta_n(X) \leq \frac{1}{2}, \eta'(X) > \frac{1}{2}\right) \leq \mathbb{P}\left(\left|\eta'(X) - \frac{1}{2}\right| < \lambda\right) + \mathbb{P}\left(|\eta_n(X) - \eta'(X)| \geq \lambda\right),
$$

$$
\begin{aligned}
\left\{\eta'(X) \leq \frac{1}{2}\right\} &= \left\{\eta'(X) \leq \frac{1}{2} + \lambda\right\} \cup \left\{\frac{1}{2} < \eta'(X) \leq \frac{1}{2} + \lambda\right\}, \text{ puis} \\
\left\{\eta_n(X) > \frac{1}{2}\right\} &\cap \left\{\eta'(X) \leq \frac{1}{2} + \lambda\right\} \subset \left\{|\eta_n(X) - \eta'(X)| \geq \lambda\right\}, \\
\left\{\eta_n(X) > \frac{1}{2}\right\} &\cap \left\{1 < \eta'(X) \leq \frac{1}{2} + \lambda\right\} \subset \left\{\left|\eta'(X) - \frac{1}{2}\right| < \lambda\right\}.
\end{aligned}
$$

d'où

$$
\mathbb{P}\left(\eta_n(X) > \frac{1}{2}, \eta'(X) \leq \frac{1}{2}\right) \leq \mathbb{P}\left(\left|\eta'(X) - \frac{1}{2}\right| < \lambda\right) + \mathbb{P}\left(|\eta_n(X) - \eta'(X)| \geq \lambda\right),
$$

Soit $\epsilon > 0$, il s'agit maintenant de trouver $N \in \mathbb{N}^*$ tel que pour tout $n \geq N$, on a $|L(g_n) - L(g)| < \epsilon$. Tout d'abord, puisque $\mathbb{P}(\eta'(X) = \frac{1}{2}) = \mathbb{P}(\left|\eta'(X) - \frac{1}{2}\right| = 0) = 0$ par hypothèse, on peut choisir $\lambda \in \left(0, \frac{1}{2}\right)$ tel que $\mathbb{P}(\left|\eta'(X) - \frac{1}{2}\right| < \lambda) \leq \frac{\epsilon}{2}$ (si on avait une masse $m \in]0, 1]$ en 0, on aurait toujours $\mathbb{P}(\left|\eta'(X) - \frac{1}{2}\right| < \lambda) \geq m$ et on ne pourrait pas descendre en dessous).

Ensuite, d'après l'inégalité de Markov, $\mathbb{P}(|\eta_n(X) - \eta'(X)| \geq \lambda) \leq \frac{1}{\lambda}\mathbb{E}(|\eta_n(X) - \eta'(X)|)$ et par hypothèse, il existe $N = N(\epsilon, \lambda) \in \mathbb{N}^*$ tel que pour tout $n \geq N$, $\mathbb{E}(|\eta_n(X) - \eta'(X)|) \leq \frac{\lambda\epsilon}{2}$, et finalement $|L(g_n) - L(g)| < \epsilon$.

Nous avons donc bien montré que $L(g_n) \to L(g)$ quand $n \to +\infty$.

4) Commençons par remarquer que pour tout $x \in \mathcal{X}$, on a

$$\begin{aligned}\eta'(x) &= \mathbb{P}(Z = 1 \mid X = x)\\ &= \mathbb{P}(Z = 1 \mid Y = 1, X = x)\mathbb{P}(Y = 1 \mid X = x) + \mathbb{P}(Z = 1 \mid Y = -1, X = x)\mathbb{P}(Y = -1 \mid X = x)\\ &= (1-b)\eta(x) + a(1-\eta(x)) = a + (1-b-a)\eta(x).\end{aligned}$$

Maintenant, d'après l'Eq.(2), on a

$$\begin{aligned}L(g) - L^* &\leq 2\mathbb{E}\left(|\eta'(X) - \eta(X)|\mathbb{1}_{\{g(X)\neq g^*(X)\}}\right)\\ &= 2\mathbb{E}\left(|\eta'(X) - \eta(X)| \mid g(X) \neq g^*(X)\right)\mathbb{P}(g(X) \neq g^*(X)).\end{aligned}$$

Il s'agit de majorer les deux termes à droite de cette dernière inégalité (l'espérance conditionnelle et la probabilité).

Soit $x \in \mathcal{X}$, alors $g(x) \neq g^*(x)$ signifie (i) $\eta(x) \geq \frac{1}{2} > \eta'(x)$ ou (ii) $\eta(x) < \frac{1}{2} \leq \eta'(x)$. Puisque $\eta'(x) = a + (1-a-b)\eta(x)$, cela donne

(a) $\dfrac{1-2a}{2(1-a-b)} > \eta(x) \geq \frac{1}{2}$, ce qui n'est possible que si $a < b$, et dans ce cas

$$\begin{aligned}|\eta'(x) - \eta(x)| &= \eta(x) - \eta'(x) = \eta(x)(a+b) - a\\ &\quad\frac{1-2a}{2(1-a-b)}(a+b) - a = \frac{b-a}{2(1-a-b)} = \frac{|a-b|}{2(1-a-b)}.\end{aligned}$$

(b) $\dfrac{1-2a}{2(1-a-b)} \leq \eta(x) < \frac{1}{2}$, ce qui n'est possible que si $a > b$, et dans ce cas

$$\begin{aligned}|\eta'(x) - \eta(x)| &= \eta'(x) - \eta(x) = a - \eta(x)(a+b)\\ &\quad a - \frac{1-2a}{2(1-a-b)}(a+b) = \frac{a-b}{2(1-a-b)} = \frac{|a-b|}{2(1-a-b)}.\end{aligned}$$

On peut donc d'ores et déjà majorer l'espérance conditionnelle :

$$\mathbb{E}(|\eta'(X) - \eta(X)| \,|\, g(X) \neq g^*(X)) \leq \frac{|a-b|}{2(1-a-b)}.$$

Pour majorer la probabilité, rappelons-nous que d'après le TD1 on a

$$\begin{aligned}L^* &= \mathbb{E}(\eta(X) \wedge (1-\eta(X)))\\ &\geq \mathbb{E}(\eta(X) \wedge (1-\eta(X))\mathbb{1}_{g(X)\neq g^*(X)})\\ &= \mathbb{E}(\eta(X) \wedge (1-\eta(X))|g(X) \neq g^*(X))\mathbb{P}(g(X) \neq g^*(X))\\ &\geq \begin{cases}\frac{1-2b}{2(1-a-b)}\mathbb{P}(g(X) \neq g^*(X)) & \text{si } a < b\\ = \frac{1-2a}{2(1-b-a)}\mathbb{P}(g(X) \neq g^*(X)) & \text{si } a > b\end{cases}\\ &= \frac{1-2(a \vee b)}{2(1-a-b)}\mathbb{P}(g(X) \neq g^*(X)).\end{aligned}$$

d'où

$$\mathbb{P}(g(\mathrm{X}) \neq g^*(\mathrm{X})) \leq 2\mathrm{L}^* \frac{1-a-b}{1-2(a \vee b)}.$$

Finalement, on obtient

$$\mathrm{L}(g) - \mathrm{L}^* \leq \frac{|a-b|}{1-a-b} 2\mathrm{L}^* \frac{1-a-b}{1-2(a \vee b)} = \mathrm{L}^* \frac{2|a-b|}{1-2(a \vee b)}.$$

d'où le résultat recherché

$$\mathrm{L}(g) \leq \mathrm{L}^* \left(1 + \frac{2|a-b|}{1-2(a \vee b)}\right)$$

Lorsque $a = b$, on a $\eta'(x) = a + (1-2a)\eta(x)$. Ainsi, $\eta'(x) > \frac{1}{2} \iff \eta(x) > \frac{1}{2} \quad \frac{1-2a}{1-2a} = 1$ et donc $g = g^*$; on retombe sur le classifieur de Bayes.

Exercice 3. Calculer la VC dimension des classes $\mathcal{A}$ d'ensembles suivantes :

1) $\mathcal{A} = \left\{ \,]-\infty, x_1] \times \ldots \times]-\infty, x_d] \ : \ (x_1, \ldots, x_d) \in \mathbb{R}^d \right\}$,
2) $\mathcal{A}$ est constituée des rectangles de $\mathbb{R}^d$.

Solution:

Commençons par quelques rappels de cours. Soient $\mathcal{A}$ une classe d'ensembles de $\mathbb{R}^d$ et $n \in \mathbb{N}^*$. Quels que soient $x_1, \ldots, x_n \in \mathbb{R}^d$, on note

$$\mathrm{N}_{\mathcal{A}}(x_1, \ldots, x_n) := \mathrm{Card}\left\{\{x_1, \ldots, x_n\} \cap \mathrm{A} : \mathrm{A} \in \mathcal{A}\right\}$$

le nombre d'ensembles différents que l'on peut former en intersectant $\{x_1, \ldots, x_n\}$ avec les éléments de $\mathcal{A}$. Par construction, il y en a au plus 2^n, et si $\mathrm{N}_{\mathcal{A}}(x_1, \ldots, x_n) = 2^n$, on dit que $\mathcal{A}$ pulvérise ou éclate $\{x_1, \ldots, x_n\}$.

Le n-ème coefficient de pulvérisation ou d'éclatement de $\mathcal{A}$ est alors défini comme la quantité

$$\mathrm{S}_{\mathcal{A}}(n) := \max_{x_1, \ldots, x_n \in \mathbb{R}^d} \mathrm{N}_{\mathcal{A}}(x_1, \ldots, x_n),$$

qui donne le plus grand nombre d'ensembles différents que l'on peut obtenir en intersectant les éléments de $\mathcal{A}$ avec n'importe quel ensemble de $n \in \mathbb{N}^*$ éléments de $\mathbb{R}^d$. Par construction, il jouit des propriétés suivantes :

(i) $\mathrm{S}_{\mathcal{A}}(n) \in [\![1, 2^n]\!]$ (on a toujours soit l'ensemble vide soit au moins un des n points considérés),
(ii) $\mathrm{S}_{\mathcal{A}}(n) = 2^n$ ssi il existe un sous-ensemble de n éléments de $\mathbb{R}^d$ éclaté par $\mathcal{A}$,
(iii) si $\mathrm{S}_{\mathcal{A}}(n) < 2^n$, alors pour tout $k \geq n$ on a aussi $\mathrm{S}_{\mathcal{A}}(k) < 2^k$ (si $\mathcal{A}$ ne peut pas éclater un ensemble à n éléments, alors elle ne peut pas éclater d'ensemble plus grand encore).

Les coefficients d'éclatement permettent ainsi de mesurer la richesse de la classe $\mathcal{A}$.

En cherchant l'entier n tel que à partir duquel $\mathrm{S}_{\mathcal{A}}(n+k) < 2^{n+k}$ pour tout $k \in \mathbb{N}^*$, on définit la dimension de Vapnik-Chervonenkis (VC dimension) de la classe $\mathcal{A}$:

$$\mathrm{V}_{\mathcal{A}} := \max\{n \in \mathbb{N}^* : \mathrm{S}_{\mathcal{A}}(n) = 2^n\} \in \mathbb{N}^* \cup \{+\infty\}.$$

Elle donne le plus grand nombre de points que l'on peut éclater avec $\mathcal{A}$.

1) Quel que soit $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, notons $\mathrm{A}_\mathbf{x} :=]-\infty, x_1] \times \cdots \times]-\infty, x_d]$, de telle manière que $\mathcal{A} = \{\mathrm{A}_\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\}$. Nous allons montrer que $\mathrm{V}_\mathcal{A} = d$. Pour cela, il faut d'abord trouver un ensemble de d points de $\mathbb{R}^d$ éclatés par $\mathcal{A}$. Considérons les vecteurs $e_1, \ldots, e_d$ de la base canonique de $\mathbb{R}^d$, tels que pour tout $i \in \{1, \ldots, d\}$, toutes les composantes du vecteur e_i valent 0, sauf la i-ème qui vaut 1.

On remarque que quels que soient $i \in [\![1, d]\!]$ et $x \in \mathbb{R}^d$,

$$e_i \in \mathrm{A}_\mathbf{x} \quad \text{ssi} \quad \begin{cases} x_1, \ldots, x_n \geq 0 and \\ x_i \geq 1. \end{cases}$$

Par conséquent, pour tout $x \in \mathbb{R}^d$, on a d'abord $\{e_1, \ldots, e_d\} \cap \mathrm{A}_\mathbf{x} = \emptyset$ ssi $x_1, \ldots, x_n < 1$ puis pour tous $m \in [\![1, d]\!]$ et $i_1, \ldots, i_m \in [\![1, d]\!]$,

$$\{e_1, \ldots, e_d\} \cap \mathrm{A}_x = \{e_{i_1}, \ldots, e_{i_m}\} \quad \text{ssi} \quad \begin{cases} \forall k \in [\![1, d]\!] \setminus \{i_1, \ldots, i_m\} : 0 \leq x_k < 1, \\ x_{i_1}, \ldots, x_{i_m} \geq 1. \end{cases}$$

En d'autres termes, pour tout sous-ensemble $\{e_{i_1}, \ldots, e_{i_m}\}$, on est capable de trouver au moins un $\mathrm{A}_\mathbf{x} \in \mathcal{A}$ qui permet de l'isoler. Cela signifie exactement que $\mathcal{A}$ éclate $\{e_1, \ldots, e_d\}$, d'où $\mathrm{S}_\mathcal{A}(d) = 2^d$. Il nous faut maintenant vérifier que $\mathcal{A}$ ne peut pas éclater $d+1$ points. Pour cela, prenons $\mathbf{z}_1, \ldots, \mathbf{z}_{d+1} \in \mathbb{R}^d$ quelconques. Quel que soit $i \in [\![1, d]\!]$, notons $\mathbf{z}_k = (z_{k,1}, \ldots, z_{k,d})$ et définissons $\bar{z}_i$ comme l'élément de $\{\mathbf{z}_1, \ldots, \mathbf{z}_{d+1}\}$ ayant la plus grande i-ème coordonnée, c'est-à-dire $\bar{z}_{i,i} = \max\{z_{1,i}, \ldots, z_{d+1,i}\}$. Alors, quel que soit $\mathbf{x} \in \mathbb{R}^d$, on a

$$\begin{aligned} \{\bar{z}_1, \ldots, \bar{z}_d\} \subset \mathrm{A}_\mathbf{x} &\iff \forall i \in [\![1, d]\!],\ \bar{z}_{1,i}, \ldots, \bar{z}_{d,i} \leq x_i \\ &\iff \forall i \in [\![1, d]\!],\ \bar{z}_{i,i} = \max\{z_{1,i}, \ldots, z_{d+1,i}\} \leq x_i \\ &\iff \forall i \in [\![1, d]\!],\ z_{1,i}, \ldots, z_{d+1,i} \leq x_i \\ &\iff \{z_1, \ldots, z_{d+1}\} \subset \mathrm{A}_\mathbf{x}. \end{aligned}$$

En d'autres termes, on ne peut pas trouver $\mathrm{A}_\mathbf{x} \in \mathcal{A}$ qui isole le sous-ensemble de points $\{\bar{z}_1, \ldots, \bar{z}_d\}$, d'où $\mathrm{S}_\mathcal{A}(d+1) < 2^{d+1}$.

Par définition de la dimension VC, on en conclut que $\mathrm{V}_\mathcal{A} = d$. La Figure 2 donne une illustration des résultats précédents quand $d = 2$.

Remarque. Pour avoir l'intuition de la démonstration, on peut regarder d'abord les cas $d = 1$ et $d = 2$, qui amènent à penser que $\mathrm{V}_\mathcal{A} = d$ et donnent l'idée de travailler avec les maxima des coordonnées.

2) Quels que soient $\mathbf{x} = (x_1, \ldots, x_d), \mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d$, on note

$$\mathbf{x} \wedge \mathbf{y} := (x_1 \wedge y_1, \ldots, x_d \wedge y_d) \quad \text{et} \quad \mathbf{x} \vee \mathbf{y} := (x_1 \vee y_1, \ldots, x_d \vee y_d)$$

les minimum et maximum composante par composante. Lorsque $\mathbf{x} = \mathbf{x} \wedge \mathbf{y}$ et $\mathbf{y} = \mathbf{x} \vee \mathbf{y}$, on note

$$[\mathbf{x}, \mathbf{y}] := [x_1, y_1] \times \cdots \times [x_d, y_d].$$

Nous allons montrer que $\mathrm{V}_\mathcal{A} = 2d$.

Pour cela, il faut d'abord trouver un ensemble de $2d$ points de $\mathbb{R}^d$ éclatés par $\mathcal{A}$. Considérons les vecteurs $e_1, \ldots, e_d$ de la base canonique de $\mathbb{R}^d$, comme à la question précédente, ainsi que leurs symétriques $-e_1, \ldots, -e_d$. On remarque que pour tous $i \in [\![1, d]\!]$ et $\varepsilon \in]0, 1[$,

$$\pm e_i \in [\pm e_i - \varepsilon, \pm e_i + \varepsilon].$$

Par conséquent, pour tout sous-ensemble de $\{-e_1, \ldots, -e_d, e_1, \ldots, e_d\}$, on peut trouver un hyper-rectangle qui l'isole des autres points : quel que soit $\varepsilon \in]0, 1[$,

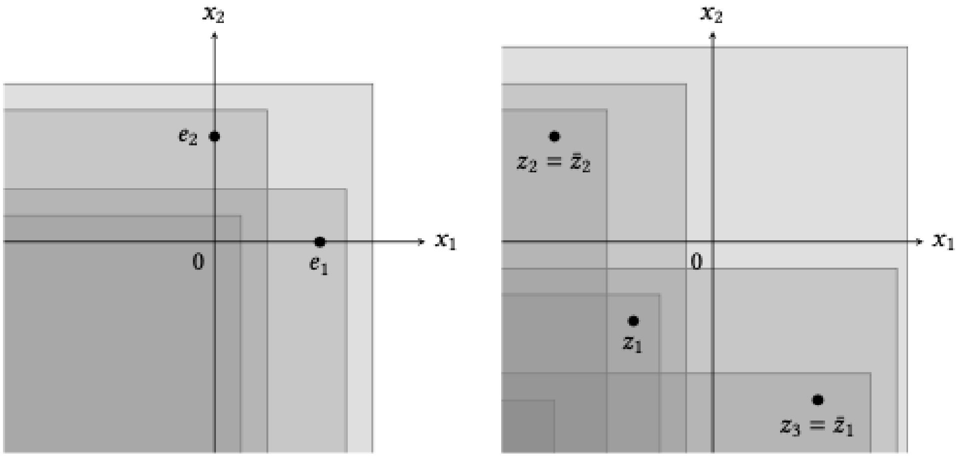

FIGURE 1 – Illustration of VC dimension for $d = 2$.

(a) $\{-e_1, \ldots, -e_d, e_1, \ldots, e_d\} \cap [-\varepsilon, \varepsilon]^d = \emptyset$,

(b) pour tous $m \in [\![1, d]\!]$ et $i_1, \ldots, i_m \in [\![1, d]\!]$,

$$\{-e_1, \ldots, -e_d, e_1, \ldots, e_d\} \cap \left[\bigwedge_{k=1}^{m} \pm e_{i_k} - \varepsilon, \bigvee_{k=1}^{m} \pm e_{i_k} + \varepsilon\right] = \{\pm e_{i_1}, \ldots, \pm e_{i_m}\}.$$

On a donc bien $\mathrm{S}_{\mathcal{A}}(2d) = 2^{2d}$.
Il nous faut maintenant vérifier que $\mathcal{A}$ ne peut pas éclater $2d + 1$ points. Pour cela, prenons $\mathbf{z}_1, \ldots, \mathbf{z}_{2d+1} \in \mathbb{R}^d$ quelconques. Quel que soit $i \in \{1, \ldots, d\}$, définissons $\bar{z}_i$ et $\underline{z}_i$ comme les éléments de $\{\mathbf{z}_1, \ldots, \mathbf{z}_{2d+1}\}$ ayant respectivement la plus grande et la plus petite i-ème coordonnée, i.e.

$$\bar{z}_{i,i} = \max\{z_{1,i}, \ldots, z_{2d+1,i}\} \quad \text{et} \quad \underline{z}_{i,i} = \min\{z_{1,i}, \ldots, z_{2d+1,i}\}.$$

Alors, quels que soient $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ tels que $\mathbf{x} = \mathbf{x} \wedge \mathbf{y}$ et $\mathbf{y} = \mathbf{x} \vee \mathbf{y}$, on a

$$\begin{aligned}
\{\bar{z}_1, \ldots, \bar{z}_d, \underline{z}_1, \ldots, \underline{z}_d\} \subset [\mathbf{x}, \mathbf{y}] &\iff \forall i, k \in [\![1, d]\!],\ \bar{z}_{i,k}, \underline{z}_{i,k} \in [x_k, y_k] \\
&\iff \forall i \in [\![1, d]\!],\ \bar{z}_{i,i}, \underline{z}_{i,i} \in [x_i, y_i] \\
&\iff \forall i \in [\![1, d]\!],\ z_{1,i}, \ldots, z_{2d+1,i} \in [x_i, y_i] \\
&\iff \{\mathbf{z}_1, \ldots, \mathbf{z}_{2d+1}\} \subset [\mathbf{x}, \mathbf{y}].
\end{aligned}$$

En d'autres termes, on ne peut pas trouver d'hyper-rectangle qui isole le sous-ensemble de points $\{\bar{z}_1, \ldots, \bar{z}_d, \underline{z}_1, \ldots, \underline{z}_d\}$, d'où $\mathrm{S}_{\mathcal{A}}(2d + 1) < 2^{2d+1}$.
Par définition de la dimension VC, on en conclut que $\mathrm{V}_{\mathcal{A}} = 2d$. La Figure 2 donne une illustration des résultats précédents quand $d = 2$.

EXERCICE 4. Donner une borne supérieure de la VC dimension de la classe des boules fermées dans $\mathbb{R}^d$:

$$\mathcal{A} = \left\{\left\{x = (x_1, \ldots, x_d) \in \mathbb{R}^d \ : \ \sum_{i=1}^{d} |x_i - a_i|^2 \leq b\right\} : a_1, \ldots, a_d, b \in \mathbb{R}\right\}.$$

Solution:

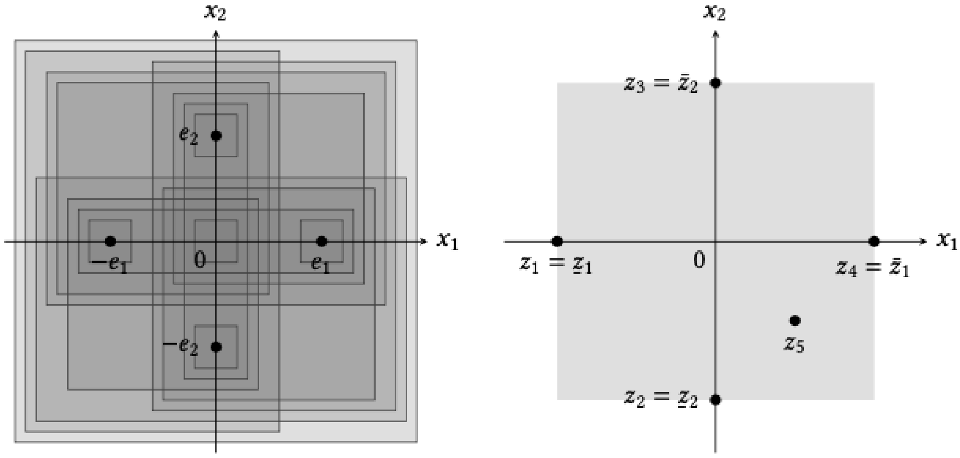

FIGURE 2 – Illustration of VC dimension for $d = 2$.

Pour répondre à la question posée, nous allons d'abord démontrer le théorème suivant.
Soit G un espace vectoriel de fonctions $\mathbb{R}^d \to \mathbb{R}$ ($d \in \mathbb{N}^*$), de dimension finie $m \in \mathbb{N}^*$. Alors la classe d'ensembles

$$\mathcal{A} := \{\{x \in \mathbb{R}^d : g(x) \geq 0\} \ : \ g \in \mathcal{G}\}$$

a pour VC dimension $\mathrm{V}_{\mathcal{A}} \leq m$.

Il suffit pour cela de prouver qu'aucun ensemble de $m+1$ points de $\mathbb{R}^d$ ne peut être éclaté par $\mathcal{A}$. Prenons donc $x_1, \ldots, x_{m+1} \in \mathbb{R}^d$ quelconques, puis considérons l'application linéaire

$$h : g \in \mathcal{G} \mapsto (g(x_1), \ldots, g(x_{m+1})) \in \mathbb{R}^{m+1}.$$

Puisque $\mathcal{G}$ est un espace vectoriel, $h(\mathrm{G})$ est un sous-espace vectoriel de $\mathbb{R}^{m+1}$, de dimension inférieure ou égale à celle de $\mathcal{G}$, i.e. m. Il existe donc un vecteur non nul $\gamma = (\gamma_1, \ldots, \gamma_{m+1}) \in \mathbb{R}^{m+1}$ orthogonal à $h(\mathcal{G})$, c'est-à-dire tel que pour tout $g \in \mathcal{G}$, on ait

$$\langle \gamma, h(g) \rangle = \gamma_1 g(x_1) + \cdots + \gamma_{m+1} g(x_{m+1}) = 0. \tag{4}$$

Soient $\mathrm{P} := \{i \in [\![1, m+1]\!] : \gamma_i \geq 0\}$ et $\mathrm{N} := \{i \in [\![1, m+1]\!] : \gamma_i < 0\}$, alors pour tout $g \in \mathcal{G}$, on a

$$\sum_{i \in \mathrm{P}} \gamma_i g(x_i) = -\sum_{i \in \mathrm{N}} \gamma_i g(x_i). \tag{4}$$

Quitte à remplacer, si besoin, γ par $-\gamma$ (qui satisfait aussi l'Eq. 4), on peut supposer que $\mathrm{N} \neq \emptyset$.

Supposons maintenant par l'absurde qu'il existe $g \in \mathcal{G}$ telle que $\{x \in \mathbb{R}^d : g(x) \geq 0\}$ contient $\{x_i : i \in \mathrm{P}\}$ mais pas $\{x_i : i \in \mathrm{N}\}$. Alors la partie de gauche dans l'Eq. 4 est positive ou nulle, tandis que celle de droite est forcément strictement négative, ce qui amène à une contradiction. Il n'existe donc pas d'élément de $\mathcal{G}$ qui permette de séparer $\{x_i : i \in \mathrm{P}\}$ et $\{x_i : i \in \mathrm{N}\}$. Par conséquent, $\mathcal{S}_{\mathrm{A}}(m+1) < 2^{m+1}$ puis $\mathrm{V}_{\mathcal{A}} < m+1$, i.e. $\mathrm{V}_{\mathcal{A}} \leq m$.

Revenons maintenant à la question posée. Quels que soient $a_1, \ldots, a_d, b \in \mathbb{R}$ et $(x_1, \ldots, x_d) \in \mathbb{R}^d$, on remarque que

$$\sum_{i=1}^{d} |x_i - a_i|^2 = \sum_{i=1}^{d} x_i^2 - 2\sum_{i=1}^{d} a_i x_i + \left(\sum_{i=1}^{d} a_i^2 - b\right).$$

Ainsi, si l'on considère $\mathcal{G}$, l'espace vectoriel engendré par les fonctions $\varphi_1, \dots, \varphi_{d+2} : \mathbb{R}^d \to \mathbb{R}$ qui à tout $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ associent

$$\varphi_1(x) = \sum_{i=1}^{d} x_i^2, \quad \varphi_2(x) = x_1, \dots, \varphi_{d+1}(x) = x_d, \quad \varphi_{d+2}(x) = 1,$$

on a $\mathcal{A} \subset \{\{x \in \mathbb{R}^d : g(x) \leq 0\} : g \in \mathcal{G}\} =: \mathcal{B}$.
Or, comme $\mathcal{G}$ est un espace vectoriel, $g \in \mathcal{G}$ ssi $-g \in \mathcal{G}$, et

$$\mathcal{B} = \{\{x \in \mathbb{R}^d : -g(x) \geq 0\} : g \in \mathcal{G}\} = \{\{x \in \mathbb{R}^d : g(x) \geq 0\} : g \in \mathcal{G}\}.$$

En appliquant le théorème que l'on vient de démontrer, on a donc $\mathrm{V}_{\mathcal{A}} \leq \mathrm{V}_{\mathcal{B}} \leq d + 2$.

Exercice 5. Soit $\mathcal{A}$ une classe d'ensembles de $\mathbb{R}^d$ de VC dimension $\mathrm{V} < +\infty$ et de coefficients d'éclatement $s(\mathcal{A}, n)$, $\forall n \in \mathbb{N}^*$.

1) Montrer que : $\forall n \geq 1$, $s(\mathcal{A}, n) \leq (n+1)^{\mathrm{V}}$.
2) Montrer que : $\forall n \geq \mathrm{V}$, $s(\mathcal{A}, n) \leq (ne/\mathrm{V})^{\mathrm{V}}$.

Indication. On utilisera le lemme de Sauer : $\forall n \geq 1$, $s(\mathcal{A}, n) \leq \sum_{k=0}^{\mathrm{V}} \binom{n}{k}$.

Solution:

1) Soit $n \in \mathbb{N}^*$, alors

$$\begin{aligned}(n+1)^{\mathrm{V}} &= \sum_{k=0}^{\mathrm{V}} \binom{\mathrm{V}}{k} n^k = \sum_{k=0}^{\mathrm{V}} \frac{\mathrm{V}!}{(\mathrm{V}-k)!k!} n^k = 1 + \sum_{k=1}^{\mathrm{V}} \frac{(\mathrm{V}-k+1)\cdots\mathrm{V}}{k!} n^k \\ &\geq 1 + \sum_{k=1}^{\mathrm{V}} \frac{n^k}{k!} = 1 + \sum_{k=1}^{\mathrm{V}} \binom{n}{k} n^k \frac{(n-k)!}{n!} = 1 + \sum_{k=1}^{\mathrm{V}} \binom{n}{k} \frac{n^k}{(n-k+1)\cdots n} \\ &\geq 1 + \sum_{k=1}^{\mathrm{V}} \binom{n}{k} = \sum_{k=0}^{\mathrm{V}} \binom{n}{k} \geq \mathrm{S}_{\mathcal{A}}(n)\end{aligned}$$

par le lemme de Sauer.

2) Soit un entier $n \geq \mathrm{V}$, alors $1 \geq \frac{\mathrm{V}}{n}$. On cherche à montrer que

$$\sum_{k=0}^{\mathrm{V}} \binom{n}{k} \leq \left(\frac{ne}{\mathrm{V}}\right)^{\mathrm{V}} = \left(\frac{n}{\mathrm{V}}\right)^{\mathrm{V}} e^{\mathrm{V}}$$

ou de manière équivalente que

$$\left(\frac{\mathrm{V}}{n}\right)^{\mathrm{V}} \sum_{k=0}^{\mathrm{V}} \binom{n}{k} \leq e^{\mathrm{V}},$$

pour pouvoir ensuite appliquer le lemme de Sauer et conclure. Or

$$\left(\frac{n}{\mathrm{V}}\right)^{\mathrm{V}} \sum_{k=0}^{n} \binom{n}{k} \leq \sum_{k=0}^{\mathrm{V}} \binom{n}{k} \left(\frac{\mathrm{V}}{n}\right)^{\mathrm{V}} \leq \sum_{k=0}^{n} \binom{n}{k} \left(\frac{\mathrm{V}}{n}\right)^{k} = \left(\frac{\mathrm{V}}{n} + 1\right)^{n} \leq \left(e^{\frac{\mathrm{V}}{n}}\right)^{n} = e^{\mathrm{V}}.$$
