Concentration et inégalités de marge

Olivier Catoni

4 avril 2013

1. Déviations des sommes de variables aléatoires indépendantes

Soit X_i, $1 \leq i \leq n$ un échantillon de variables indépendantes et

$$M \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} X_i$$

leur moyenne empirique. On se pose la question des déviations de M par rapport à sa moyenne

$$m \stackrel{\text{def}}{=} \mathbb{E}(M) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i).$$

Considérons les fonctions génératrices des moments

$$\psi_i(\lambda) = \log\Big\{ \mathbb{E}\big[\exp(\lambda X_i)\big] \Big\},$$
$$\psi(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \psi_i(\lambda).$$

Ce sont les fonctions convexes, nulles en zéro, à valeurs dans $\mathbb{R}_+ \cup \{+\infty\}$. Considérons la fonction duale

$$\psi^*(x) = \sup_{\lambda \in \mathbb{R}_+} \lambda x - \psi(\lambda) \in \mathbb{R}_+ \cup \{+\infty\}.$$

Proposition 1.1 *Les déviations de la moyennes empirique M vérifient*

$$\mathbb{P}\big(M \geq x\big) \leq \exp\big[-n\psi^*(x)\big].$$

Preuve.

CNRS – UMR 8553, Département de Mathématiques et Applications, Ecole Normale Supérieure, 45, rue d'Ulm, F75230 Paris cedex 05, and INRIA Paris-Rocquencourt – CLASSIC team.

$$\begin{aligned}\mathbb{P}\big(M \geq x\big) &= \mathbb{E}\Big\{\mathbb{1}\big[\exp\big(n\lambda(M-x)\big) \geq 1\big]\Big\} \\ &\leq \mathbb{E}\big[\exp\big(n\lambda(M-x)\big)\big] \\ &= \exp\big\{n\big[\psi(\lambda) - \lambda x\big]\big\}, \qquad \lambda \in \mathbb{R}_+.\end{aligned}$$

Par conséquent

$$\mathbb{P}\big(M \geq x\big) \leq \inf_{\lambda \in \mathbb{R}_+} \exp\big\{n\big[\psi(\lambda) - \lambda x\big]\big\} = \exp\big(-n\psi^*(x)\big).$$

□

Proposition 1.2 *Posons* $\Lambda_i = \sup\big\{\lambda \in \mathbb{R}_+ : \psi_i(\lambda) < +\infty\big\}$ *et* $\Lambda = \min\big\{\Lambda_1, \ldots, \Lambda_n\big\}$. *Pour tout* $\lambda \in [0, \Lambda_i[$, $\psi_i(\lambda) < +\infty$ *et la fonction* ψ_i *est de classe* $\mathcal{C}^\infty$ *sur* $]0, \Lambda_i[$ *lorsque* $\Lambda_i > 0$. *Si de plus* $\mathbb{E}\big(|X_i|^k\big) < \infty$, *la fonction* ψ_i *est de classe* $\mathcal{C}^k$ *sur* $[0, \Lambda_i[$.

Preuve. Posons $\varphi(\lambda) = \mathbb{E}\big[\exp(\lambda X_i)\big]$. Soit $\lambda \in [0, \Lambda_i[$. Par définition de Λ_i, il existe $\beta \in]\lambda, \Lambda_i[$ tel que $\psi_i(\beta) < \infty$, et donc $\varphi(\beta) < \infty$. D'après l'inégalité de Jensen

$$+\infty > \mathbb{E}\big[\exp(\beta X_i)\big] = \mathbb{E}\Big\{\big[\exp(\lambda X_i)\big]^{\beta/\lambda}\Big\} \geq \Big\{\mathbb{E}\big[\exp(\lambda X_i)\big]\Big\}^{\beta/\lambda},$$

ce qui prouve que $\varphi_i(\lambda) < \infty$, et donc que $\psi_i(\lambda) < \infty$.

Remarquons que

$$X_i^{j-1}\exp(\beta X_i) = X_i^{j-1}\exp(\alpha X_i) + \int_\alpha^\beta X_i^j \exp(\lambda X_i)\,\mathrm{d}\lambda, \qquad 0 < \alpha < \beta < \Lambda_i, \quad j \geq 1.$$

De plus

$$\mathbb{E}\Big\{\sup_{\lambda \in [\alpha,\beta]} \big|X_i^j \exp\big(\lambda X_i\big)\big|\Big\} < \infty$$

En effet, considérons $\gamma \in]\beta, \Lambda_i[$ et

$$\begin{aligned}C_1 &= \sup_{x \in \mathbb{R}_+} x^j \exp\big[-(\gamma - \beta)x\big], \\ C_2 &= \sup_{x \in \mathbb{R}_+} x^j \exp(-\alpha x).\end{aligned}$$

$$\big|X_i^j \exp(\lambda X_i)\big| \leq \begin{cases} C_1 \exp(\gamma X_i), & X_i \geq 0, \\ C_2, & X_i \leq 0. \end{cases}$$

Par conséquent

$$\mathbb{E}\Big\{\sup_{\lambda\in[\alpha,\beta]}\big|X_i^j\exp(\lambda X_i)\big|\Big\}\leq C_1\mathbb{E}\big[\exp(\gamma X_i)\big]+C_2<\infty.$$

On peut donc employer le théorème de Fubini et écrire

$$\begin{aligned}\mathbb{E}\big[X_i^{j-1}\exp(\beta X)\big]&=\mathbb{E}\big[X_i^{j-1}\exp(\alpha X_i)\big]+\mathbb{E}\bigg(\int_\alpha^\beta X_i^j\exp(\lambda X_i)\,\mathrm{d}\lambda\bigg)\\&=\mathbb{E}\big[X_i^{j-1}\exp(\alpha X_i)\big]+\int_\alpha^\beta\mathbb{E}\big[X_i^j\exp(\lambda X_i)\big]\,\mathrm{d}\lambda.\end{aligned}$$

D'après le théorème de convergence dominée, $\lambda\mapsto\mathbb{E}\big(X_i^j\exp(\lambda X_i)\big]:[\alpha,\beta]\to\mathbb{R}$ est continue, donc $\beta\mapsto\mathbb{E}\big[X_i^{j-1}\exp(\beta X_i)\big]$ est de classe $\mathcal{C}^1$, de dérivée $\mathbb{E}\big[X_i^j\exp(\beta X_j)\big]$, donc $\beta\mapsto\mathbb{E}\big[\exp(\beta X_i)\big]$ est de classe $\mathcal{C}^\infty$ sur $]0,\Lambda_i[$ et il en est de même de ψ_i.

Supposons maintenant de plus que $\mathbb{E}\big[|X_i|^k\big]<\infty$. On peut dans ce cas montrer de façon analogue que

$$\mathbb{E}\Big\{\sup_{\lambda\in[0,\beta]}\big|X_i^j\exp(\lambda X_i)\big|\Big\}<\infty,\qquad 0<\beta<\Lambda_i,$$

en déduire que

$$\mathbb{E}\big[X_i^{j-1}\exp(\beta X_i)\big]=\mathbb{E}\big(X_i^{j-1}\big)+\int_0^\beta\mathbb{E}\big[X_i^j\exp(\lambda X_i)\big]\,\mathrm{d}\lambda,\qquad 0\leq\beta<\Lambda_i,\quad 1\leq j\leq k,$$

et enfin que φ_i et ψ_i sont de classe $\mathcal{C}^k$ sur $[0,\Lambda_i[$. □

Proposition 1.3 *Supposons que $\mathbb{E}\big(X_i^2\big)<\infty$ et que $\Lambda_i>0$. La dérivée seconde de ψ_i prend la forme d'une variance :*

$$\psi_i''(\lambda)=\frac{\mathbb{E}\big[X_i^2\exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]}-\bigg(\frac{\mathbb{E}\big[X_i\exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]}\bigg)^2,\qquad 0\leq\lambda<\Lambda_i,$$

de plus

$$\psi_i(\lambda)=\lambda\mathbb{E}(X_i)+\int_0^\lambda(\lambda-\alpha)\psi_i''(\alpha)\,\mathrm{d}\alpha,\qquad 0\leq\lambda<\Lambda_i.$$

Preuve. D'après la proposition précédente, ψ_i est de classe $\mathcal{C}^2$ sur $[0,\Lambda_i[$ et

$$\psi_i'(\lambda)=\frac{\mathbb{E}\big[X_i\exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]},$$

en dérivant une fois de plus, on obtient l'expression de ψ_i'' indiquée dans la proposition. Considérons la variable aléatoire Y_i de loi

$$\mathbb{P}\big(Y_i \in A\big) = \frac{\mathbb{E}\big[\mathbb{1}(X_i \in A)\exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]},$$

pour tout borélien A. Elle vérifie pour toute fonction mesurable f telle que $\mathbb{E}\big[|f(X_i)|\exp(\lambda X_i)\big] < \infty$

$$\mathbb{E}\big[f(Y_i)\big] = \frac{\mathbb{E}\big[f(X_i)\exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]},$$

ce qui montre que $\psi_i''(\lambda) = \mathbb{E}\big(Y_i^2\big) - \mathbb{E}\big(Y_i\big)^2$ est bien une variance. La deuxième partie de la proposition s'obtient en écrivant la formule de Taylor $\psi_i(\lambda) = \psi_i(0) + \lambda\psi_i'(0) + \int_0^\lambda (\lambda - \alpha)\psi_i''(\alpha)\,\mathrm{d}\alpha$. □

Proposition 1.4 *Supposons que* $\Lambda > 0$ *et que* $\mathbb{E}\big(X_i^2\big) < \infty$, $1 \le i \le n$. *Posons*

$$\overline{V}(\lambda) \stackrel{\text{def}}{=} \frac{2}{\lambda^2}\big[\psi(\lambda) - \lambda m\big] = \frac{2}{\lambda^2}\int_0^\lambda (\lambda - \alpha)\psi''(\alpha)\,\mathrm{d}\alpha, \qquad 0 \le \lambda < \Lambda$$

$$V(\lambda) \stackrel{\text{def}}{=} \sup_{\beta \in [0,\lambda]} \overline{V}(\beta),$$

$$v \stackrel{\text{def}}{=} V(0) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big\{\big[X_i - \mathbb{E}(X_i)\big]^2\Big\}$$

et remarquons que V *est une fonction continue et croissante à valeurs dans* $\mathbb{R} + \cup\{+\infty\}$. *Sous ces hypothèses*

$$\mathbb{P}\big(M \ge m + x\big) \le \exp\bigg(-\frac{nx^2}{2V(x/v)}\bigg),$$

$$\mathbb{P}\left(M \ge m + \sqrt{\frac{2\log(\epsilon^{-1})}{n}V\left(\sqrt{\frac{2\log(\epsilon^{-1})}{nv}}\right)}\ \right) \le \epsilon.$$

Preuve. Pour tout $0 \le \beta \le \lambda$,

$$\psi^*(m + x) \ge \beta x - \frac{\beta^2}{2}V(\lambda),$$

si bien que

$$\mathbb{P}\big(M \ge m + x\big) \le \exp\bigg[-n\bigg(\beta x - \frac{\beta^2}{2}V(\lambda)\bigg)\bigg].$$

On obtient la première inégalité en choisissant $\lambda = x/v$ et $\beta = x/V(\lambda) \leq \lambda$. Pour obtenir la seconde posons $\epsilon = \exp\biggl[-n\Bigl(\beta x - \frac{\beta^2}{2}V(\lambda)\Bigr)\biggr]$, pour obtenir dans un premier temps

$$\mathbb{P}\biggl(M \geq m + \frac{\beta}{2}V(\lambda) + \frac{\log(\epsilon^{-1})}{n\beta}\biggr) \leq \epsilon.$$

Choisissons alors $\lambda = \sqrt{\dfrac{2\log(\epsilon^{-1})}{nv}} \geq \beta = \sqrt{\dfrac{2\log(\epsilon^{-1})}{nV(\lambda)}}$ pour conclure. □

Proposition 1.5 (Inégalité de Bennett) *Supposons que* $\mathbb{E}\bigl(X_i^2\bigr) < \infty$ *et que* $X_i \leq \mathbb{E}(X_i) + b$, $1 \leq i \leq n$. *Introduisons la fonction*

$$h(u) = (1+u)\log(1+u) - u \geq \frac{u^2}{2(1+u/3)}, \qquad u \in \mathbb{R}_+.$$

Sous ces hypothèses

$$\mathbb{P}\bigl(M \geq m + x\bigr) \leq \exp\biggl[-\frac{nv}{b^2}h\Bigl(\frac{bx}{v}\Bigr)\biggr] \leq \exp\biggl(-\frac{nx^2}{2v + \frac{2bx}{3}}\biggr),$$
$$\mathbb{P}\left(M \geq m + \sqrt{\frac{2v\log(\epsilon^{-1})}{n}}\biggl(1 - \frac{b}{3v}\sqrt{\frac{2v\log(\epsilon^{-1}}{n}}\biggr)^{-1/2}\right) \leq \epsilon.$$

Preuve. Remarquons tout d'abord que pour tout $\lambda \in \mathbb{R}+$,

$$\begin{aligned}\psi^*(m+x) &\geq \lambda(x+m) - \frac{1}{n}\sum_{i=1}^n \log\bigl[\mathbb{E}\bigl(\exp(\lambda X_i)\bigr)\bigr] \\ &= \lambda x - \frac{1}{n}\sum_{i=1}^n \log\Bigl\{\mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr)\bigr]\Bigr\},\end{aligned}$$

où $m_i \stackrel{\text{def}}{=} \mathbb{E}(X_i)$. On peut alors écrire

$$\begin{aligned}\mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr)\bigr] - 1 &= \mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr) - 1 - \lambda(X_i - m_i)\bigr] \\ &= \mathbb{E}\bigl[\lambda^2(X_i - m_i)^2 g\bigl(\lambda(X_i - m_i)\bigr)\bigr]\end{aligned}$$

où $g(y) = y^{-2}\bigl(\exp(y) - 1 - y\bigr)$. La fonction g est croissante sur $\mathbb{R}$. En effectuant un développement de Taylor à l'ordre deux de la fonction $z \mapsto \exp(yz)$ entre 0 et 1, on voit en effet qu'elle peut s'écrire

$$g(y) = \int_0^1 (1-z)\exp(yz)\,\mathrm{d}z, \qquad y \in \mathbb{R}.$$

On en déduit que

$$\mathbb{E}\big[\lambda^2(X_i-m_i)^2 g\big(\lambda(X_i-m_i)\big)\big] \le \mathbb{E}\big[\lambda^2(X_i-m_i)^2 g(\lambda b)\big], \qquad 1\le i\le n,$$

et donc que

$$\log\Big\{\mathbb{E}\big[\exp\big(\lambda(X_i-m_i)\big)\big]\Big\} \le \lambda^2 g(\lambda b)\mathbb{E}\big[(X_i-m_i)^2\big].$$

Ainsi

$$\psi^*(m+x) \ge \lambda x - \lambda^2 v g(\lambda b) = \lambda x - \frac{v}{b^2}\big(\exp(\lambda b) - 1 - \lambda b\big).$$

Choisissons $\lambda = b^{-1}\log\left(1+\dfrac{bx}{v}\right)$ pour obtenir

$$\psi^*(x) \ge \frac{v}{b^2} h\left(\frac{bx}{v}\right).$$

Montrons maintenant que $h(u) \ge \dfrac{u^2}{2(1+u/3)}$, $u>-1$. Calculons les dérivées de h, $h'(u)=\log(1+u)$, $h''(u)=1/(1+u)$, puis les dérivées de $f(u) = (1+u/3)h(u) - u^2/2$. On obtient $f'(u) = h'(u)(1+u/3) + h(u)/3 - u$ qui vérifie $f'(0)=0$ et

$$\begin{aligned} f''(u) = h''(u)(1+u/3) + 2h'(u)/3 - 1 &= \frac{1+u/3}{1+u} + \frac{2}{3}\log(1+u) - 1 \\ &= \frac{2}{3}\log(1+u) - \frac{2u}{3(1+u)} = \frac{2h(u)}{3(1+u)} \ge 0, \qquad u>-1. \end{aligned}$$

La fonction f, convexe, nulle en zéro et de dérivée première nulle en zéro est donc positive.

Posons $\epsilon = \exp\left(-\dfrac{nx^2}{2v+\frac{2bx}{3}}\right)$. On obtient

$$\begin{aligned} x^2 = \frac{2v\log(\epsilon^{-1})}{n}\left(1+\frac{bx^2}{3vx}\right) \\ \le \frac{2v\log(\epsilon^{-1}}{n}\left(1+\frac{bx^2}{3v}\left(\frac{2v\log(\epsilon^{-1})}{n}\right)^{-1/2}\right). \end{aligned}$$

On en déduit que

$$x^2 \le \frac{2v\log(\epsilon^{-1})}{n}\left(1-\frac{b}{3v}\sqrt{\frac{2v\log(\epsilon^{-1})}{n}}\right)^{-1},$$

ce qui prouve la deuxième inégalité de la proposition. □

PROPOSITION 1.6 (INÉGALITÉ DE HOEFFDING) *Supposons que* $a_i \leq X_i \leq b_i$, $1 \leq i \leq n$. *Dans ce cas*

$$\mathbb{P}\big(M \geq m + x\big) \leq \exp\bigg(-\frac{2n^2x^2}{\sum_{i=1}^n (b_i - a_i)^2}\bigg),$$
$$\mathbb{P}\left(M \geq m + \sqrt{\frac{\sum_{i=1}^n (b_i - a_i)^2 \log(\epsilon^{-1}}{2n^2}}\right) \leq \epsilon.$$

PREUVE. La dérivée seconde de ψ_i est la variance d'une variable aléatoire à valeurs dans l'intervalle $[a_i, b_i]$, et ne peut donc excéder $(b_i - a_i)^2/4$. Par conséquent $\psi(\lambda) \leq \lambda m + \dfrac{\lambda^2}{8}\displaystyle\sum_{i=1}^n (b_i - a_i)^2$ d'où $\psi^*(m + x) \geq \dfrac{2nx^2}{\sum_{i=1}^n (b_i - a_i)^2}$. □

2. BORNES PAC-BAYÉSIENNES POUR LES DÉVIATIONS UNIFORMES DES MOYENNES EMPIRIQUES PAR RAPPORT À LEURS ESPÉRANCES

Considérons n variables aléatoires indépendantes X_i, $1 \leq i \leq n$ à valeurs dans un espace mesurable $\mathcal{X}$, un espace mesurable de paramètres Θ et une fonction mesurable $f : \mathcal{X} \times \Theta \to \mathbb{R}$ (qui peut être vue comme une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ indexée par θ). Supposons que

$$\mathbb{E}\big[f(X, \theta)^2\big] < +\infty, \qquad \theta \in \Theta,$$

et posons

$$M(\theta) = \frac{1}{n}\sum_{i=1}^n f(X_i, \theta),$$
$$m(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(X_i, \theta)\big],$$
$$\psi_i(\lambda, \theta) = \log\Big\{\mathbb{E}\exp\big[\lambda f(X_i, \theta)\big]\Big\},$$
$$\psi(\lambda, \theta) = \frac{1}{n}\sum_{i=1}^n \psi_i(\lambda, \theta),$$
$$\Lambda = \sup\big\{\lambda : \psi(\lambda, \theta) < \infty, \theta \in \Theta\big\}$$

PROPOSITION 2.1 *Supposons que* $\Lambda > 0$. *Soit* $\nu \in \mathcal{M}_+^1(\Theta)$ *une mesure de référence sur l'espace des paramètres* Θ. *Pour tout* $\lambda \in [0, \Lambda[$,

$$\mathbb{E}\Bigg[\exp\Bigg(\sup\Big\{\int_\Theta n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\nu),$$
$$\rho\in\mathcal{M}_+^1(\Theta),\theta\mapsto\lambda M(\theta)-\psi(\lambda,\theta)\in\mathbb{L}^1(\rho),\mathcal{K}(\rho,\nu)<\infty\Big\}\Bigg)\Bigg]\leq 1.$$

Par conséquent, avec probabilité au moins $1-\epsilon$*, pour toute probabilité* $\rho\in\mathcal{M}_+^1(\Theta)$*, telle que* $\theta\mapsto\lambda M(\theta)-\psi(\lambda,\theta)\in\mathbb{L}^1(\rho)$ *et* $\mathcal{K}(\rho,\nu)<\infty$*,*

$$\int M(\theta)\,\mathrm{d}\rho(\theta)\leq\frac{1}{\lambda}\int\psi(\lambda,\theta)\,\mathrm{d}\rho(\theta)+\frac{\mathcal{K}(\rho,\nu)+\log(\epsilon^{-1})}{n\lambda}.$$

Preuve. D'après l'inégalité de Jensen, lorsque ρ vérifie les hypothèses,

$$\begin{aligned}
&\exp\Big[\int_\Theta n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\nu)\Big]\\
&\quad\leq\int_\Theta\exp\Big\{n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\Big\}\mathbb{1}\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)>0\Big)\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)\Big)^{-1}\mathrm{d}\rho(\theta)\\
&\quad=\int_\Theta\exp\Big\{n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\Big\}\mathbb{1}\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)>0\Big)\,\mathrm{d}\nu(\theta)\\
&\quad\leq\int_\Theta\exp\Big\{n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\Big\}\,\mathrm{d}\nu(\theta).
\end{aligned}$$

On peut ensuite appliquer le théorème de Fubini pour les fonctions positives.

$$\begin{aligned}
&\mathbb{E}\Big\{\exp\Big[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\int_\Theta n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\nu)\Big]\Big\}\\
&\qquad\leq\mathbb{E}\Big[\int_\Theta\exp\Big\{n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\Big\}\,\mathrm{d}\nu(\theta)\Big]\\
&\qquad=\int_\Theta\mathbb{E}\Big[\exp\Big\{n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\Big\}\Big]\,\mathrm{d}\nu(\theta)=1.
\end{aligned}$$

L'espérance indiquée dans la proposition ne porte pas forcément sur une fonction mesurable, mais cette fonction est majorée par une fonction mesurable d'espérance inférieure ou égale à 1 : c'est ce que montre la preuve et c'est le sens technique à donner à la proposition. La deuxième partie de la proposition s'obtient en appliquant l'inégalité de Markov. Là encore, l'événement en question n'est pas forcément mesurable : il faut comprendre qu'il contient un événement mesurable de probabilité au moins égale à $1-\epsilon$. □

Posons $m_i(\theta)=\mathbb{E}\big[f(X_i,\theta)\big]$,

$$v(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big\{ \big[f(X_i,\theta) - m_i(\theta)\big]^2 \Big\},$$
$$\overline{V}(\lambda,\theta) = \frac{2}{\lambda^2}\big[\psi(\lambda,\theta) - \lambda m(\theta)\big],$$
$$V(\lambda,\theta) = \sup_{\beta\in[0,\lambda]} \overline{V}(\beta,\theta)$$

et supposons que $v \stackrel{\text{def}}{=} \sup_{\theta\in\Theta} v(\theta) < \infty$ et $V(\lambda) \stackrel{\text{def}}{=} \sup_{\theta\in\Theta} V(\lambda,\theta) < \infty$, $0 \le \lambda < \Lambda'$.

Proposition 2.2 *Sous les hypothèses précédentes, pour toute constante positive c,*

$$\mathbb{E}\Bigg(\sup\Big\{ \int_\Theta \big[M(\theta) - m(\theta)\big]\, \mathrm{d}\rho(\theta);$$
$$\rho \in \mathcal{M}_+^1(\Theta), \theta \mapsto M(\theta) - m(\theta) \in \mathbb{L}^1(\rho), \mathcal{K}(\rho,\nu) \le c \Big\} \Bigg)$$
$$\le \inf_{\lambda\in[0,\Lambda'[} \frac{\lambda\overline{V}(\lambda)}{2} + \frac{c}{\lambda n} \le \sqrt{\frac{2c}{n} V\Bigg(\sqrt{\frac{2c}{nv}}\Bigg)}.$$

En particulier quand Θ est fini, en prenant $c = \log(|\Theta|)$, $\rho = \delta_\theta$ et $\nu(\theta) = |\Theta|^{-1}$, $\theta \in \Theta$, on obtient

$$\mathbb{E}\Big\{ \sup_{\theta\in\Theta} \big[M(\theta) - m(\theta)\big] \Big\} \le \sqrt{\frac{2\log\big(|\Theta|\big)}{n} V\Bigg(\sqrt{\frac{2\log\big(|\Theta|\big)}{nv}}\Bigg)}.$$

Preuve. D'après la preuve de la proposition précédente, l'argument de l'espérance à majorer est inférieur ou égal à

$$\frac{1}{n\lambda}\log\Bigg\{ \int \exp\Big[n\big[\lambda M(\theta) - \psi(\lambda,\theta)\big]\Big]\, \mathrm{d}\nu(\theta) \Bigg\} + \frac{\lambda\overline{V}(\lambda)}{2} + \frac{c}{\lambda n},$$

et on conclut à l'aide de l'inégalité de Jensen. On otient ainsi le premier majorant $\inf_{\lambda\in[0,\Lambda'[} \frac{\lambda\overline{V}(\lambda)}{2} + \frac{c}{\lambda n}$ que l'on peut affaiblir en $\inf_{0\le\lambda\le\beta} \frac{\lambda V(\beta)}{2} + \frac{c}{\lambda n}$. On obtient le second majorant en choisissant $\beta = \sqrt{\frac{2c}{nv}}$ et $\lambda = \sqrt{\frac{2c}{nV(\beta)}} \le \beta$.

□

PROPOSITION 2.3 *Sous les hypothèses précédentes, pour toute constante positive c, avec probabilité au moins $1-\epsilon$,*

$$\sup\Bigg\{\int_\Theta \big[M(\theta)-m(\theta)\big]\,\mathrm{d}\rho(\theta);\\ \rho\in\mathcal{M}_+^1(\Theta), \theta\mapsto M(\theta)-m(\theta)\in\mathbb{L}^1(\Theta), \mathcal{K}(\rho,\nu)\le c\Bigg\}$$

$$\le \inf_{\lambda\in[0,\Lambda'[}\frac{\lambda\overline{V}(\lambda)}{2}+\frac{c+\log(\epsilon^{-1})}{\lambda n}\le\sqrt{\frac{2\big[c+\log(\epsilon^{-1})\big]}{n}V\Bigg(\sqrt{\frac{2\big[c+\log(\epsilon^{-1})\big]}{nv}}\Bigg)}.$$

En particulier quand Θ est fini, avec probabilité au moins $1-\epsilon$

$$\sup_{\theta\in\Theta}\big[M(\theta)-m(\theta)\big]\le\sqrt{\frac{2\log\big(|\Theta|/\epsilon\big)}{n}V\Bigg(\sqrt{\frac{2\log\big(|\Theta|/\epsilon\big)}{nv}}\Bigg)}.$$

PREUVE. C'est une conséquence directe de la seconde partie de la proposition 2.1 (page 7) et de la majoration $\psi(\lambda,\theta)\le\dfrac{\lambda^2V(\lambda)}{2}+\lambda m(\theta)$. □

PROPOSITION 2.4 *Supposons que $\Theta=\mathbb{B}_d=\big\{\theta\in\mathbb{R}^d;\|\theta\|\le 1\big\}$ et qu'il existe deux constantes positives B et g telles que*

$$\sup_{x\in\mathcal{X}}f(x,\theta)-\inf_{x\in\mathcal{X}}f(x,\theta)\le B,\qquad \theta\in\mathbb{B}_d,$$
$$|f(x,\theta)-f(x,\theta')|\le g\|\theta-\theta'\|,\qquad x\in\mathcal{X},\quad \theta,\theta'\in\mathbb{B}_d.$$

Considérons le minimiseur du risque empirique

$$\widehat{\theta}\in\arg\min_{\theta\in\mathbb{B}_d}M(\theta).$$

Avec probabilité au moins $1-\epsilon$,

$$m(\widehat{\theta})\le\inf_{\theta\in\mathbb{B}_d}m(\theta)+B\Bigg\{\sqrt{\frac{d}{2n}\log\Bigg(1+\frac{4g}{B}\sqrt{\frac{2n}{d}}\Bigg)+\frac{\log\big(2/\epsilon\big)}{2n}}\\ +\sqrt{\frac{d}{8n}}+\sqrt{\frac{\log\big(2/\epsilon\big)}{2n}}\Bigg\}.$$

Sous ces hypothèses très simples, on voit que la qualité de l'estimation de $\inf_{\theta\in\Theta} m(\theta)$ par $\widehat{\theta}$ dépend de la dimension d de l'espace des paramètres, et plus précisément du rapport d/n entre cette dimension et la taille de l'échantillon.

PREUVE. Commençons par étendre le domaine de définition de f à $\mathbb{R}^d$ en posant

$$f(x,\theta) = f(x,\theta/\|\theta\|), \qquad \theta \in \mathbb{R}^d \setminus \mathbb{B}_d.$$

Soit $\delta > 0$ un paramètre réel positif dont nous choisirons par la suite la valeur et ν la mesure uniforme sur la boule $(1+\delta)\mathbb{B}_d$ de rayon $1+\delta$. Considérons pour tout $\theta \in \mathbb{B}_d$ la probabilité uniforme ρ_θ sur la boule $\theta + \delta\mathbb{B}_d$ centrée en θ de rayon θ. Le volume d'une boule de $\mathbb{R}^d$ étant proportionnel à son rayon élevé à la puissance d, on voit immédiatement que

$$\mathcal{K}(\rho_\theta, \nu) = d\log\left(\frac{1+\delta}{\delta}\right), \qquad \theta \in \mathbb{B}_d.$$

D'après la proposition précédente et l'inégalité de Hoeffding, avec probabilité au moins $1-\epsilon$, pour tout $\theta \in \mathbb{B}_d$,

$$\int m(\theta')\,\mathrm{d}\rho_\theta(\theta') \le \int M(\theta')\,\mathrm{d}\rho_\theta(\theta') + B\sqrt{\frac{d\log\bigl(1+\delta^{-1}\bigr) + \log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

On en déduit, toujours avec probabilité $1-\epsilon$,

$$m(\widehat{\theta}) \le M(\widehat{\theta}) + 2g\delta + B\sqrt{\frac{d\log\bigl(1+\delta^{-1}\bigr) + \log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

Soit $\theta_* \in \mathbb{B}_d$, tel que $m(\theta_*) = \inf_{\theta\in\mathbb{B}_d} m(\theta)$ (qui existe car $\theta \mapsto m(\theta)$ est continue sur le compact $\mathbb{B}_d$). Avec probabilité $1-\epsilon$

$$M(\theta_*) \le m(\theta_*) + B\sqrt{\frac{\log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

Par construction de l'estimateur $\widehat{\theta}$, $M(\widehat{\theta}) \le M(\theta_*)$. On en déduit donc qu'avec probabilité au moins $1-2\epsilon$,

$$m(\widehat{\theta}) \le m(\theta_*) + B\left\{\sqrt{\frac{d\log(1+\delta^{-1}) + \log\bigl(\epsilon^{-1}\bigr)}{2n}} + \sqrt{\frac{\log(\epsilon^{-1})}{2n}}\right\} + 2g\delta.$$

On conclut en choisissant $\delta = \dfrac{B}{4g}\sqrt{\dfrac{d}{2n}}$ et en remplaçant ϵ par $\epsilon/2$. □

PROPOSITION 2.5 *Supposons que $\Theta = \mathbb{R}^d$, qu'il existe une fonction mesurable $(x,\theta) \mapsto \nabla f(x,\theta) \in \mathbb{R}^d$ et des constantes positives g et H telles que*

$$|f(x,\theta) - f(x,\theta')| \leq g\|\theta - \theta'\|,$$
$$|f(x,\theta') - f(x,\theta) - \langle \nabla f(x,\theta), \theta' - \theta\rangle| \leq \frac{H}{2}\|\theta' - \theta\|^2, \quad x \in \mathcal{X}, \quad \theta, \theta' \in \mathbb{R}^d.$$

Soit $\theta_* \in \arg\min_{\theta\in\mathbb{B}_d} m(\theta)$. Introduisons la fonction

$$\chi(h) = \sup_{\theta\in\mathbb{B}_d} \frac{h}{2}\|\theta - \theta_*\|^2 - m(\theta) + m(\theta_*),$$

Dans ces conditions, le minimiseur empirique $\widehat{\theta} \in \arg\min_{\theta\in\mathbb{B}_d} M(\theta)$ de m sur la boule unité vérifie avec probabilité au moins $1-\epsilon$

$$\|\widehat{\theta} - \theta_*\|^2 \leq \frac{8g^2}{nh^2}\left[\left(\frac{8H}{h} + 1\right)d + 2\log\bigl(\epsilon^{-1}\bigr)\right] + \frac{4\chi(h)}{h}$$
$$\text{et } m(\widehat{\theta}) - m(\theta_*) \leq \frac{4g^2}{nh}\left[\left(\frac{8H}{h} + 1\right)d + 2\log\bigl(\epsilon^{-1}\bigr)\right] + \chi(h).$$

Dans le cas où il existe $h > 0$ tel que $\chi(h) = 0$, on obtient donc une vitesse de convergence en d/n au lieu d'une vitesse en $\sqrt{d/n}$ sous les hypothèses plus faibles de la proposition précédente.

Exercice 1 *Dans le cas où $m(\theta) - m(\theta_*) \geq c\,\|\theta - \theta_*\|^\alpha$, $\theta \in \mathbb{B}_d$, avec $c > 0$ et $\alpha > 2$ quelle vitesse obtient-on ?*

PREUVE. Choisissons $\rho_\theta = \mathcal{N}(\theta, \beta^{-1}I)$ et $\nu = \rho_{\theta_*}$. Remarquons que $\mathcal{K}(\rho_\theta, \nu) = \dfrac{\beta}{2}\|\theta - \theta_*\|^2$. Nous allons appliquer la proposition 2.1 (page 7) à la fonction $(x,\theta) \mapsto f(x,\theta_*) - f(x,\theta)$. D'après l'inégalité de Hoeffding

$$\log \mathbb{E}\exp\Bigl\{\lambda\bigl[f(X,\theta_*) - f(X,\theta)\bigr]\Bigr\} - \lambda\bigl[m(\theta) - m(\theta_*)\bigr] \leq \frac{\lambda^2 g^2\|\theta - \theta_*\|^2}{2}$$

Avec probabilité au moins $1-\epsilon$, pour tout $\theta \in \mathbb{B}_d$,

$$\int m(\theta')\,\mathrm{d}\rho_\theta(\theta') - m(\theta_*) \leq \int M(\theta')\,\mathrm{d}\rho_\theta(\theta') - M(\theta_*) + \frac{\lambda g^2}{2}\int \|\theta' - \theta_*\|^2\,\mathrm{d}\rho_\theta(\theta') + \frac{\beta\|\theta - \theta_*\|^2}{2n\lambda} + \frac{\log(\epsilon^{-1})}{n\lambda}.$$

De plus

$$\int m(\theta')\,\mathrm{d}\rho_\theta(\theta') = m(\theta) + \mathbb{E}\left[\int \Big[f(X,\theta') - f(X,\theta) - \langle \nabla f(X,\theta), \theta'-\theta\rangle\Big]\,\mathrm{d}\rho_\theta(\theta')\right] \geq m(\theta) - \frac{H}{2}\int \|\theta'-\theta\|^2\,\mathrm{d}\rho_\theta(\theta') = m(\theta) - \frac{Hd}{2\beta}.$$

De même $\int M(\theta')\,\mathrm{d}\rho_\theta(\theta') \leq M(\theta) + \frac{Hd}{2\beta}$. On en déduit qu'avec probabilité au moins $1-\epsilon$, pour tout $\theta \in \mathbb{B}_d$,

$$m(\theta) - m(\theta_*) \leq M(\theta) - M(\theta_*) + \frac{Hd}{\beta} + \frac{\lambda g^2 d}{2\beta} + \frac{\lambda g^2}{2}\|\theta-\theta_*\|^2 + \frac{\beta\|\theta-\theta_*\|^2}{2n\lambda} + \frac{\log\big(\epsilon^{-1}\big)}{n\lambda}.$$

On peut alors utiliser le fait que $m(\theta) - m(\theta_*) \geq \frac{h}{2}\|\theta-\theta_*\|^2 - \chi(h)$ et que par construction $M(\widehat{\theta}) \leq M(\theta_*)$. On en conclut avec probabilité au moins $1-\epsilon$

$$\frac{h}{2}\|\widehat{\theta}-\theta_*\|^2 \leq \chi(h) + \frac{d}{\beta}\left(H + \frac{\lambda g^2}{2}\right) + \left(\frac{\lambda g^2}{2} + \frac{\beta}{2n\lambda}\right)\|\widehat{\theta}-\theta_*\|^2 + \frac{\log\big(\epsilon^{-1}\big)}{n\lambda}.$$

Ainsi

$$\|\widehat{\theta}-\theta_*\|^2\left(1 - \frac{\lambda g^2}{h} - \frac{\beta}{n\lambda h}\right) \leq \frac{2\chi(h)}{h} + \frac{2d}{\beta h}\left(H + \frac{\lambda g^2}{2}\right) + \frac{2\log\big(\epsilon^{-1}\big)}{hn\lambda}.$$

Choisissons alors $\lambda = \frac{h}{4g^2}$ et $\beta = \frac{n\lambda h}{4} = \frac{nh^2}{16g^2}$. On obtient

$$\frac{1}{2}\|\widehat{\theta}-\theta_*\|^2 \leq \frac{2\chi(h)}{h} + \frac{32g^2 d}{nh^3}\left(H + \frac{h}{8}\right) + \frac{8g^2\log\big(\epsilon^{-1}\big)}{nh^2},$$

qui donne la première majoration de la proposition.

Pour prouver la seconde, on utilise $\|\widehat{\theta}-\theta_*\|^2 \leq \frac{2}{h}\big[m(\widehat{\theta}) - m(\theta_*) + \chi(h)\big]$, pour obtenir

$$m(\widehat{\theta}) - m(\theta_*) \leq \frac{d}{\beta}\left(H + \frac{\lambda g^2}{2}\right) + \left(\frac{\lambda g^2}{2} + \frac{\beta}{2n\lambda}\right)\frac{2}{h}\left[m(\widehat{\theta}) - m(\theta_*) + \chi(h)\right] + \frac{\log\left(\epsilon^{-1}\right)}{n\lambda}.$$

On conclut de même en remplaçant λ et β par leurs valeurs. □

RÉFÉRENCES

[1] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. Pages 1–269.

[2] O. Catoni. *PAC-Bayesian Supervised Classification : The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007. Pages i-xii, 1-163.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley and Sons, New York, second edition, 2006.

[4] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, New York, NY, USA, 2009. ACM.

[5] J. Langford and J. Shawe-Taylor. PAC-bayes & margins. In *Advances in Neural Information Processing Systems*, pages 423–430, 2002.

[6] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*. Morgan Kaufmann, 1999.

[7] D. A. McAllester. PAC-Bayesian stochastic model selection. *Mach. Learn.*, 51(1) :5–21, April 2003.

[8] David Mcallester. Simplified pac-bayesian margin bounds. In *In COLT*, pages 203–215, 2003.

[9] M. Seeger. PAC-Bayesian generalization error bounds for gaussian process classification. Informatics report series EDI-INF-RR-0094, Division of Informatics, University of Edinburgh, 2002.