Concentration et inégalités de marge

Olivier Catoni

28 mars 2013

1. Déviations des sommes de variables aléatoires indépendantes

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

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

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

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

Considérons les fonctions génératrices des moments

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

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*

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

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.

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

Par conséquent

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

□

PROPOSITION 1.2 *Posons* $\Lambda_i = \sup\bigl\{\lambda \in \mathbb{R}_+ : \psi_i(\lambda) < +\infty\bigr\}$ *et* $\Lambda = \min\bigl\{\Lambda_1, \dots, \Lambda_n\bigr\}$*. 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* $\mathbb{E}\bigl(|X_i|^k\bigr) < \infty$*, la fonction* ψ_i *est de classe* $\mathcal{C}^k$ *sur* $[0, \Lambda_i[$*.*

PREUVE. Posons $\varphi(\lambda) = \mathbb{E}\bigl[\exp(\lambda X_i)\bigr]$. 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

$$+\infty > \mathbb{E}\bigl[\exp(\beta X_i)\bigr] = \mathbb{E}\Bigl\{\bigl[\exp(\lambda X_i)\bigr]^{\beta/\lambda}\Bigr\} \geq \Bigl\{\mathbb{E}\bigl[\exp(\lambda X_i)\bigr]\Bigr\}^{\beta/\lambda},$$

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

$$\mathbb{E}\Bigl\{\sup_{\lambda \in [\alpha, \beta]} \bigl|X_i^j \exp\bigl(\lambda X_i\bigr)\bigr|\Bigr\} < \infty$$

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

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

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

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 $[0, \Lambda_i[$. $\square$

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

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

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

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

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

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

$$\begin{aligned}
\overline{V}(\lambda) &\overset{\text{def}}{=} \frac{2}{\lambda^2}\bigl[\psi(\lambda) - \lambda m\bigr] = \frac{2}{\lambda^2}\int_0^\lambda (\lambda - \alpha)\psi''(\alpha)\,\mathrm{d}\alpha, \qquad 0 \leq \lambda < \Lambda \\
V(\lambda) &\overset{\text{def}}{=} \sup_{\beta \in [0,\lambda]} \overline{V}(\beta), \\
v &\overset{\text{def}}{=} V(0) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Bigl\{\bigl[X_i - \mathbb{E}(X_i)\bigr]^2\Bigr\}
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\bigl(M \geq m + x\bigr) \leq \exp\biggl(-\frac{nx^2}{2V(x/v)}\biggr), \\
&\mathbb{P}\Biggl(M \geq m + \sqrt{\frac{2\log(\epsilon^{-1})}{n} V\biggl(\sqrt{\frac{2\log(\epsilon^{-1})}{nv}}\biggr)}\Biggr) \leq \epsilon.
\end{aligned}$$

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

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

si bien que

$$\mathbb{P}\bigl(M \geq m + x\bigr) \leq \exp\biggl[-n\biggl(\beta x - \frac{\beta^2}{2}V(\lambda)\biggr)\biggr].$$

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 $\lambda = \sqrt{\frac{2\log(\epsilon^{-1})}{nv}} \geq \beta = \sqrt{\frac{2\log(\epsilon^{-1})}{nV(\lambda)}}$ 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 $m_i \overset{\text{def}}{=} \mathbb{E}(X_i)$. 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

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

et donc que

$$\log\Bigl\{\mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr)\bigr]\Bigr\} \leq \lambda^2 g(\lambda b)\mathbb{E}\bigl[(X_i - m_i)^2\bigr].$$

Ainsi

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

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

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

Montrons maintenant que $h(u) \geq \dfrac{u^2}{2(1 + u/3)}$, $u > -1$. 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

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

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

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

On en déduit que

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

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*

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

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 $\psi^*(m + x) \geq \dfrac{2nx^2}{\sum_{i=1}^n (b_i - a_i)^2}$.
$\square$