

Petites perturbations des estimateurs et bornes PAC-Bayésiennes

Olivier Catoni

CNRS, INRIA (projet CLASSIC)
Département de Mathématiques et Applications,
École Normale Supérieure
45 rue d'Ulm, 75 230 Paris Cedex 05,
`Olivier.Catoni@ens.fr`

Séminaire de Probabilités et Statistiques,
Université de Nice - Sophia Antipolis,
jeudi 12 mai 2011

Introduction

Cadre : apprentissage supervisé par minimisation d'une fonction de coût.

On observe $S = (W_1, \dots, W_n) \sim \mathbb{P}^{\otimes n}$ où $\mathbb{P} \in \mathcal{M}_+^1(\mathcal{W})$.

On dispose de $L : \mathcal{W} \times \Theta \rightarrow \mathbb{R}$, où Θ est un espace de paramètres, (le plus souvent $\Theta \subset \mathbb{R}^d$).

On veut minimiser

$$\int L(w, \theta) d\mathbb{P}(w) \quad \text{en } \theta \in \Theta.$$

Exemples :

❶ Classification linéaire à noyau (SVM).

$\mathcal{W} = \mathcal{X} \times \{-1, +1\}$, où $k : \mathcal{X} \rightarrow \mathbb{R}$ est un noyau positif.

$$\begin{aligned} L[(x, y); (\alpha_k, x_k)_{k=1}^d] &= \mathbf{1} \left[\left(\sum_{k=1}^d \alpha_k k(x_k, x) + b \right) y \leq 0 \right] \\ &= \mathbf{1} \left[\left(\langle \Psi(x), \sum \alpha_k \Psi(x_k) \rangle + b \right) y \leq 0 \right], \end{aligned}$$

où $\Psi : \mathcal{X} \rightarrow \mathcal{H} = \left\{ \sum_{j=1}^{+\infty} \alpha_j k(x_j, \cdot); \sum_{j=1}^{\infty} \alpha_j^2 k(x_j, x_j) < \infty \right\}$, espace

de Hilbert autoreproduisant. Exemples de noyaux :

$$k(x_1, x_2) = (1 + \langle x_1, x_2 \rangle)^s, \quad \dim(\mathcal{H}) < \infty,$$

$$k(x_1, x_2) = \exp(-\|x_1 - x_2\|^2), \quad \dim(\mathcal{H}) = \infty.$$

② Classification linéaire simple en dimension finie.

On traitera le cas $\dim(\mathcal{H}) < \infty$ qui se ramène à $x \in \mathbb{R}^d$,
 $\Theta \subset \mathbb{R}^d$,

$$L[(x, y); \theta] = \mathbf{1}(\langle \theta, x \rangle y \leq 0).$$

Les bornes indépendantes de d resteront valables dans le cas où
 $\dim(\mathcal{H}) = \infty$.

3 Régression aux moindres carrés.

$$L[(x, y); \theta] = (\langle \theta, x \rangle - y)^2, \text{ où } x \in \mathbb{R}^d, y \in \mathbb{R} \text{ et } \Theta \subset \mathbb{R}^d.$$

Couvre le cas fonctionnel où $x = [\varphi_j(z)]_{j=1}^d$, le contraste

$$\left(y - \sum_{j=1}^d \theta_j \varphi_j(z) \right)^2$$

mesurant alors l'approximation de y par une fonction de z dans une base de fonctions.

Approche PAC-Bayésienne « perturbative »

On suppose ici que $\Theta \subset \mathbb{R}^d$.

On considère une perturbation de l'origine $\pi \in \mathcal{M}_+^1(\Theta)$ concentrée autour de 0, par exemple $\pi = \mathcal{N}(0, \beta^{-1} \text{Id})$.

On définit les translatées $\pi_\theta(A) = \pi(A - \theta)$, si bien que $\int h(\theta') d\pi_\theta(\theta') = \int h(\theta + \theta') d\pi(\theta')$

On introduit la fonction d'influence $\psi : \mathbb{R} \rightarrow \mathbb{R}$

$$\psi(z) = \begin{cases} \log(2), & z \geq 1, \\ -\log(1 - z + z^2/2), & 0 \leq z \leq 1, \\ -\psi(-z), & z \leq 0. \end{cases}$$

qui vérifie $-\log(1 - z + z^2/2) \leq \psi(z) \leq \log(1 + z + z^2/2)$.

Théorème

Soit $\theta_* \in \Theta$ une valeur déterministe du paramètre. Avec probabilité au moins $1 - \epsilon$, pour tout $\theta \in \Theta$,

$$\begin{aligned} & \lambda^{-1} \int \psi[\lambda f(w, \theta')] \, d\pi_\theta(\theta') \, d\bar{\mathbb{P}}_S(w) \\ & \leq \int f(w, \theta') \, d\pi_\theta(\theta') \, d\mathbb{P}(w) + \frac{\lambda}{2} \int f(w, \theta')^2 \, d\pi_\theta(\theta') \, d\mathbb{P}(w) \\ & \quad + \frac{\mathcal{K}(\pi_\theta, \pi_{\theta_*}) - \log(\epsilon)}{n\lambda}, \end{aligned}$$

où $\bar{\mathbb{P}}_{(w_1, \dots, w_n)} = \frac{1}{n} \sum_{i=1}^n \delta_{w_i}$ est la mesure empirique de l'échantillon et

$\mathcal{K}(\rho, \mu) = \int \log\left(\frac{d\rho}{d\mu}\right) d\rho$, quand $\rho \ll \mu$ et $+\infty$ sinon.

Corollaire

Soit $\hat{\theta}(S)$ un estimateur.

$$\begin{aligned} & \lambda^{-1} \int \psi[\lambda f(w, \theta')] \, d\pi_{\hat{\theta}(s)}(\theta') \, d\bar{\mathbb{P}}_s(w) \, d\mathbb{P}^{\otimes n}(s) \\ & \leq \int \left(\int f \, d\pi_{\hat{\theta}(s)} \, d\mathbb{P} + \frac{\lambda}{2} \int f^2 \, d\pi_{\hat{\theta}(s)} \, d\mathbb{P} \right. \\ & \quad \left. + \frac{\mathcal{K}(\pi_{\hat{\theta}(s)}, \pi_{\theta_*}) + 1}{n\lambda} \right) d\mathbb{P}^{\otimes n}(s), \end{aligned}$$

Démonstration.

C'est le résultat d'un simple contrôle de moment exponentiel :

$$\int \exp \left\{ \sup_{\theta \in \Theta} n \lambda \left[\lambda^{-1} \int \psi(\lambda f) d\pi_{\theta} d\bar{\mathbb{P}}_s - \int f d\pi_{\theta} d\mathbb{P} - \frac{\lambda}{2} \int f^2 d\pi_{\theta} d\mathbb{P} - \mathcal{K}(\pi_{\theta}, \pi_{\theta_*}) + \log(\epsilon) \right] \right\} d\mathbb{P}^{\otimes n}(s) \leq \epsilon,$$

par des arguments de convexité (Jensen) et d'indépendance. Le corollaire se déduit du théorème en utilisant l'inégalité

$$\mathbb{E}(W) \leq \mathbb{E}(W_+) = \int_0^{\infty} \mathbb{P}(W \geq \eta) d\eta.$$



Dans le cas de la classification, nous n'aurons pas besoin d'utiliser ψ car $L \in \{0, 1\}$ borné.

Interprétation en termes d'information mutuelle

$$\begin{aligned} & \int \mathcal{K}(\pi_{\hat{\theta}(s)}, \pi_{\theta_*}) \, d\mathbb{P}^{\otimes n}(s) \\ & \geq \inf_{\mu \in \mathcal{M}_+^1(\Theta)} \int \mathcal{K}(\pi_{\hat{\theta}(s)}, \mu) \, d\mathbb{P}^{\otimes n}(s) \\ & = \int \mathcal{K}\left(\pi_{\hat{\theta}(s)}, \int \pi_{\hat{\theta}(s')} \, d\mathbb{P}^{\otimes n}(s')\right) \, d\mathbb{P}^{\otimes n}(s) \\ & = \text{information mutuelle} \end{aligned}$$

entre s et θ sous la loi $d\mathbb{P}^{\otimes n}(s)d\pi_{\hat{\theta}(s)}(\theta)$, qui est la loi jointe du couple échantillon, estimateur perturbé.

Rôle de la perturbation : diminuer l'information mutuelle entre s et θ , l'information mutuelle étant en général infinie entre l'estimateur non perturbé $\hat{\theta}(S)$ et l'échantillon S .

Première partie : classification supervisée

On peut tirer partie du fait que $L(w, \theta) = \mathbb{1}(\langle \theta, x \rangle y \leq 0)$.

Introduisons

$$K(p, q) = p \log\left(\frac{p}{q}\right) + (1 - p) \log\left(\frac{1 - p}{1 - q}\right) = \mathcal{K}(B_p, B_q)$$

où $B_p = \text{Bernoulli de paramètre } p$, et

$$\begin{aligned}\Phi_\lambda(p) &= -\lambda^{-1} \log\left[1 - p + p \exp(-\lambda)\right] \\ &= -\lambda^{-1} \log\left[\int \exp(-\lambda\sigma) dB_p(\sigma)\right].\end{aligned}$$

Le résultat suivant est valable pour toute fonction de coût binaire $L(w, \theta) \in \{0, 1\}$.

Soit $\Lambda = \{\lambda_0 \exp(kt), k = 0, \dots, m\}$, où $t = m^{-1} \log(\sqrt{2}/\lambda_0)$. Avec probabilité au moins $1 - \epsilon$, pour tout $\rho \in \mathcal{M}_+^1(\Theta)$,

$$L(\mathbb{P}, \rho) \leq B_\Lambda \left(L(\bar{\mathbb{P}}_S, \rho), \frac{\mathcal{K}(\rho, \pi) + \log[(m+1)/\epsilon]}{n} \right),$$

où on a utilisé la notation abrégée $L(\mathbb{P}, \rho) = \int L(w, \theta) d\mathbb{P}(w) d\rho(\theta)$

et introduit la borne $B_\Lambda(q, \delta) = \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1} \left(q + \frac{\delta}{\lambda} \right)$.

On peut montrer que pour tout $\delta, q \in [0, 1/2]$ tels que

$$\frac{\lambda_0^2}{8} \leq \delta \leq q(1-q),$$

$$B_\Lambda(q, \delta) \leq q + \sqrt{2\delta q(1-q)} \cosh(t/2) + 2\delta(1-q) \cosh(t/2)^2,$$

$$\begin{aligned} \text{alors que } \inf_{\lambda \in \mathbb{R}_+} \Phi_\lambda^{-1} \left(q + \frac{\delta}{\lambda} \right) &= \sup \left\{ p \in [0, 1]; K(q, p) \leq \delta \right\} \\ &\in q + \sqrt{2\delta q(1-q)} + [-\delta q, 2\delta(1-q)]. \end{aligned}$$

Comme $\mathbb{1}(\langle \theta, \alpha x \rangle y \leq 0) = \mathbb{1}(\langle \theta, x \rangle y \leq 0)$, on supposera que $\|x\| = 1$ \mathbb{P} -presque sûrement, quitte à remplacer x par $\|x\|^{-1}x$.

Choisissons une perturbation gaussienne $\pi = \mathcal{N}(0, \beta^{-1} \text{Id})$ et $\theta_\star = 0$.

Lemme

$$L(w, \pi_\theta) = \varphi(\sqrt{\beta} \langle \theta, x \rangle y), \text{ où}$$

$$\begin{aligned} \varphi(z) &= \frac{1}{\sqrt{2\pi}} \int_z^{+\infty} \exp\left(-\frac{z^2}{2}\right) dz, & z \in \mathbb{R}, \\ &\leq \min\left\{\frac{1}{z\sqrt{2\pi}}, \frac{1}{2}\right\} \exp\left(-\frac{z^2}{2}\right), & z \in \mathbb{R}_+. \end{aligned}$$

Introduisons l'erreur avec marge $M(w, \theta) = \mathbb{1}(\langle \theta, x \rangle y \leq 1)$, qui indique soit que x est mal classé, soit qu'il se trouve à distance inférieure à $\|\theta\|^{-1}$ de la frontière.

$$M(w, \pi_\theta) = \varphi[\sqrt{\beta}(\langle \theta, x \rangle y - 1)] \geq \underbrace{\varphi(-\sqrt{\beta})}_{= 1 - \varphi(\sqrt{\beta})} L(w, \theta).$$

Théorème (J. Langford, J. Shawe-Taylor, D. McAllester, O.C.)

Avec probabilité $1 - \epsilon$, pour tout $\theta \in \mathbb{R}^d$,

$$L(\mathbb{P}, \theta) \leq [1 - \varphi(\sqrt{\beta})]^{-1} M(\mathbb{P}, \pi_\theta) \leq [1 - \varphi(\sqrt{\beta})]^{-1} C_1(\theta),$$

où

$$C_1(\theta) = B_\Lambda \left(\underbrace{M(\bar{\mathbb{P}}_S, \pi_\theta)}_{\substack{\leq M(\bar{\mathbb{P}}_S, \theta/2) \\ + \varphi(\sqrt{\beta})}}, \frac{\beta \|\theta\|^2 + 2 \log[(m+1)/\epsilon]}{2n} \right).$$

Soit $\hat{\theta}$ un estimateur tel que $C_1(\hat{\theta}) \leq \inf_{\theta \in \mathbb{R}^d} C_1(\theta) + \zeta$.

Théorème

Avec probabilité au moins $1 - \epsilon$,

$$L(\mathbb{P}, \hat{\theta}) \leq \inf_{\theta \in \mathbb{R}^d} [1 - \varphi(\sqrt{\beta})]^{-1} B_{\Lambda} \left[B_{-} \left(M(\mathbb{P}, \pi_{\theta}), \frac{\log(\epsilon^{-1})}{n} \right), \frac{\beta \|\theta\|^2 + 2 \log[(m+1)/\epsilon]}{2n} \right],$$

où

$$\begin{aligned} B_{-}(q, \delta) &= \sup \left\{ p \in [0, 1]; K(p, q) \leq \delta \right\}, \\ &\leq q + \sqrt{2\delta q(1-q)} + 2\delta(1-q). \end{aligned}$$

Comparaison avec un critère convexe :

$$M(w, \pi_\theta) = \varphi[\sqrt{\beta}(\langle \theta, x \rangle y - 1)] \leq (2 - \langle \theta, x \rangle y)_+ + \varphi(\sqrt{\beta}).$$

Corollaire

Avec probabilité au moins $1 - \epsilon$, pour tout $\theta \in \mathbb{R}^d$,

$$L(\mathbb{P}, \theta) \leq [1 - \varphi(\sqrt{\beta})]^{-1} \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1} \left[C_3(\lambda, \theta) + \varphi(\sqrt{\beta}) + \frac{\log[(m+1)/\epsilon]}{n\lambda} \right],$$

où

$$C_3(\lambda, \theta) = \int (2 - \langle \theta, x \rangle y)_+ d\bar{\mathbb{P}}_S(x, y) + \frac{\beta \|\theta\|^2}{2n\lambda}.$$

Ce résultat donne un critère pour choisir λ , et donc l'intensité de la régularisation.

Deuxième partie : Régression aux moindres carrés

Ici $L(w, \theta) = (\langle \theta, x \rangle - y)^2$, $x, \theta \in \mathbb{R}^d, y \in \mathbb{R}$. Posons

$$R(\theta) = \int L(w, \theta) d\mathbb{P}(w),$$

$$R'(\theta, \theta') = R(\theta) - R(\theta'),$$

$$\begin{aligned} L'(w, \theta, \theta') &= L(w, \theta) - L(w, \theta') \\ &= \langle \theta - \theta', x \rangle^2 + 2\langle \theta - \theta', x \rangle (\langle \theta', x \rangle - y). \end{aligned}$$

Soit Θ un convexe fermé de \mathbb{R}^d et $\theta_\star = \arg \min_{\Theta} R$.

Proposition

Avec probabilité au moins $1 - \epsilon$,

$$R'(\pi_\theta, \pi_{\theta_\star}) \leq r'_\lambda(\pi_\theta, \pi_{\theta_\star}) + \frac{\lambda}{2} (L')^2(\mathbb{P}, \pi_\theta, \pi_{\theta_\star}) + \frac{\mathcal{K}(\pi_\theta, \pi_{\theta_\star}) + \log(\epsilon^{-1})}{n\lambda},$$

où $r'_\lambda(\theta, \theta') = \lambda^{-1} \int \psi[\lambda L'(w, \theta, \theta')] d\bar{\mathbb{P}}_S(w)$.

Hypothèses : pour tout θ tel que $R'(\theta, \theta_*) \leq D^2$,

$$(L')^2 (\mathbb{P}, \pi_\theta, \pi_{\theta_*}) \leq a_D R'(\theta, \theta_*) + b_D,$$

$$\mathcal{K}(\pi_\theta, \pi_{\theta_*}) \leq p_D R'(\theta, \theta_*) + q_D,$$

$$R'(\theta, \theta_*) \leq R(\pi_\theta, \pi_{\theta_*}) + \xi,$$

$$r'_\lambda(\pi_\theta, \pi_{\theta_*}) \leq r'(\theta, \theta_*) + \eta_D, \quad \text{avec probabilité } 1 - \epsilon \text{ pour } D \text{ fixé.}$$

Théorème (J.-Y. Audibert, O.C.)

Soit $\tilde{\theta}(S) \in \Theta$ tel que $r'_\lambda(\tilde{\theta}, \theta_\star) = \inf_{\theta \in \Theta} r'_\lambda(\theta, \theta_\star)$, et $\hat{\theta} \in \Theta$ (notre estimateur) tel que $\sup_{\theta' \in \Theta} r'_\lambda(\hat{\theta}, \theta') = \inf_{\theta \in \Theta} \sup_{\theta' \in \Theta} r'_\lambda(\theta, \theta') + \zeta$.

Pour tout $D > 0$, avec probabilité au moins $1 - \epsilon$,

soit $\max\{R'(\tilde{\theta}, \theta_\star), R'(\hat{\theta}, \theta_\star)\} > D^2$, soit

$$R'(\hat{\theta}, \theta_\star) \leq \left(1 - \frac{a_D \lambda}{2} - \frac{p_D}{n\lambda}\right)^{-1} \left(\sup_{\theta' \in \Theta} r'_\lambda(\hat{\theta}, \theta') + \frac{\lambda b_D}{2} + \frac{q + \log(\epsilon^{-1})}{n\lambda} + \eta + \xi\right),$$

$$\text{et } R'(\hat{\theta}, \theta_\star) + R'(\tilde{\theta}, \theta_\star) \leq \left(1 - \frac{a_D \lambda}{2} - \frac{p_D}{n\lambda}\right)^{-1} \left(b_D \lambda + \frac{2[q_D + \log(\epsilon^{-1})]}{n\lambda} + 2\eta + 2\xi + \zeta\right).$$

Théorème (J.-Y. Audibert, O.C.)

Pour tout n tel que

$$n \geq 2^7 \times 3 \sqrt{\kappa} s_4^2,$$

avec probabilité au moins $1 - \epsilon$,

$$R'(\hat{\theta}, \theta_*) \leq \frac{[B_1 s_4^2 + B_2 \sqrt{\kappa} \log(4/\epsilon)] \sigma_4^2}{n} + \frac{B_3 \kappa \log(4/\epsilon) \Delta^2}{n} + B_4 \zeta,$$

où

$$B_1 = \left(1 + \frac{2^7 c s_4^4}{n}\right) \left[2^8 + 2^6 \times 3 c + 2^{11} (7 + c) \sqrt{\kappa} \left(4 + \sqrt{\kappa} \frac{\Delta^2}{\sigma_4^2}\right) \frac{s_4^2}{n}\right]$$

$$+ \frac{2^{14} c^2 s_4^6 \Delta^2}{n \sigma_4^2}, \quad B_2 = 2^8 \left(1 + \frac{2^7 c s_4^4}{n}\right),$$

$$B_3 = 2^6 \left(1 + \frac{2^7 c s_4^4}{n}\right), \quad B_4 = 2 \left(1 + \frac{2^7 c s_4^4}{n}\right),$$

$$\Delta = \sup\{\|\theta - \theta'\| : (\theta, \theta') \in \Theta_\star^2\},$$

$$\kappa = \sup_{\theta \in \Theta_\star} \frac{\int \langle \theta - \theta_\star, x \rangle^4 \mathrm{d}\mathbb{P}(x)}{[\int \langle \theta - \theta_\star, x \rangle^2 \mathrm{d}\mathbb{P}(x)]^2},$$

$$\sigma_4^2 = \sqrt{\int (\langle \theta_\star, x \rangle - y)^4 \mathrm{d}\mathbb{P}(x, y)},$$

$$s_4^2 = \sqrt{\int \|x\|^4 \mathrm{d}\mathbb{P}(x)},$$

$$c = \frac{3}{\log(4)} \leq 2, 17.$$

Théorème (J.-Y. Audibert, O.C.)

Soit $\hat{\theta} = \arg \min_{\theta \in \Theta} L(\bar{\mathbb{P}}_S, \theta)$, l'estimateur des moindres carrés ordinaires. Il existe un entier N (dépendant de \mathbb{P}) tel que pour tout $n \geq N$ et tout $\epsilon \in]0, 1]$ tels que

$$n > \frac{16\sqrt{\kappa}}{15} [25s_4^2 + 33\sqrt{\kappa} \log(\epsilon^{-1})],$$

avec probabilité au moins $1 - \epsilon$,

$$R'(\hat{\theta}, \theta_*) \leq \frac{4[25s_4^2 + 33\sqrt{\kappa} \log(\epsilon^{-1})] \sigma_4^2}{n}.$$

Autres hypothèses sur le bruit : supposons

$$\sigma_2^2 = \operatorname{ess\,sup}_{d\mathbb{P}(x)} \int (\langle \theta_*, x \rangle - y)^2 d\mathbb{P}(y|x) < \infty,$$

$$\text{et } \int \|x\|^4 \mathbb{P}(x) < \infty.$$

Théorème (J.-Y. Audibert, O.C.)

Il existe un entier N tel que pour tout $n > N$ et tout $\epsilon \in]0, 1]$ tels que

$$n > \frac{16\kappa}{15} [25d + 33 \log(\epsilon^{-1})],$$

avec probabilité au moins $1 - \epsilon$,

$$R'(\hat{\theta}, \theta_*) \leq \frac{4[25d + 33 \log(\epsilon^{-1})] \sigma_2^2}{n}.$$

Estimation de la matrice de Gram

On veut estimer uniformément

$$\begin{aligned} N(\theta) &= \int \langle \theta, x \rangle^2 d\mathbb{P}(x) \\ &= \theta^t G \theta, \text{ où } G = \int x x^t d\mathbb{P}(x). \end{aligned}$$

Introduisons

$$\begin{aligned} r_\lambda(\theta) &= \lambda^{-1} \int \psi \left\{ \lambda [\langle \theta, x \rangle^2 - 1] \right\}, \\ \hat{\alpha}(\theta) &= \sup \left\{ \alpha \in \mathbb{R}_+, r_\lambda(\alpha \theta) \leq 0 \right\} \in \mathbb{R}_+ \cup +\infty, \\ \hat{N}(\theta) &= \hat{\alpha}(\theta)^{-2}. \end{aligned}$$

Lemme

$$\begin{aligned} \psi \left\{ \lambda [\langle \theta, x \rangle^2 - 1] \right\} &\leq \int \log \left\{ 1 + \lambda \left(\langle \theta', x \rangle^2 - 1 - \frac{\|x\|^2}{\beta} \right) \right. \\ &\quad + \frac{\lambda^2}{2} \left(\langle \theta', x \rangle - 1 - \frac{\|x\|^2}{\beta} \right)^2 \\ &\quad \left. + \frac{2c\lambda^2\|x\|^2}{\beta} \left(4\langle \theta', x \rangle^2 + \frac{5\|x\|^2}{\beta} \right) \right\} d\pi_{\theta}(\theta'), \end{aligned}$$

$$\text{où } c = \frac{15}{\log(4)} \leq 10,83.$$

Posons

$$\lambda = \sqrt{\frac{2}{(\kappa - 1)n} \left[\log(\epsilon^{-1}) + \frac{(1 + 18c)s_4^2}{4\sqrt{\kappa}(1 + 4c)} \right]},$$

$$\beta = 2\lambda s_4 \kappa^{1/4} \sqrt{(1 + 4c)n},$$

$$\eta = (\kappa - 1)\lambda = \sqrt{\frac{2(\kappa - 1)}{n} \left[\log(\epsilon^{-1}) + \frac{(1 + 18c)s_4^2}{4\sqrt{\kappa}(1 + 4c)} \right]},$$

$$\gamma = 2\sqrt{\frac{(1 + 4c)s_4^2\sqrt{\kappa}}{n}},$$

et choisissons $\theta_\star = 0$ pour définir la loi a priori π_{θ_\star} . Le lemme précédent permet un contrôle PAC-Bayésien uniforme en θ de $\psi \left\{ \lambda [\langle \theta, x \rangle^2 - 1] \right\}$ et conduit au résultat suivant.

Théorème

Pour tous $n \in \mathbb{N}$ et $\epsilon \in]0, 1]$ tels que $2(\eta + \gamma) + \frac{\kappa\eta}{2(\kappa - 1)} < 1$, à savoir tels que

$$n > \left[4\kappa^{1/4}s_4\sqrt{1 + 4c} + \left(2 + \frac{\kappa}{2(\kappa - 1)}\right) \sqrt{2(\kappa - 1) \left[\log(\epsilon^{-1}) + \frac{(1 + 18c)s_4^2}{4\sqrt{\kappa}(1 + 4c)} \right]} \right]^2,$$

avec probabilité au moins $1 - 2\epsilon$, pour tout $\theta \in \mathbb{R}^d$,

$$|N(\theta) - \hat{N}(\theta)| \leq 2(\eta + \gamma)\hat{N}(\theta) \leq \frac{2(\eta + \gamma)}{1 - 2(\eta + \gamma)}N(\theta).$$

Dans le cas où $d\mathbb{P}(x)$ est une gaussienne, $\kappa = 3$ et $s_4^2 = d\sqrt{1 + 2/d}$, on obtient donc une vitesse uniforme en $\sqrt{d/n}$, dans le cas des bases de fonctions, on s'attend typiquement à ce que κ et s_4^2 soient tous les deux d'ordre d , ce qui donne une vitesse en $\sqrt{d^{3/2}/n}$.

Utilisations possibles :

- Donner une région de confiance autour de $\hat{\theta}$.
- Calculer un préestimateur en interprétant $R(\theta)$ comme le carré d'une norme en dimension $d + 1$ suivant l'identité

$$R(\theta) = \int \langle (\theta, -1), (x, y) \rangle^2 d\mathbb{P}(x, y).$$

Utiliser ensuite la méthode précédente d'estimation de $\arg \min_{\Theta_\star} R$, dans un voisinage (aléatoire) Θ_\star de θ_\star de diamètre Δ faible (en $\sqrt{d^{3/2}/n}$), la méthode d'estimation fine servant ensuite à garantir le passage à une vitesse en $\sqrt{d/n}$ pour $\sqrt{R(\hat{\theta}) - R(\theta_\star)}$.



J.-Y. Audibert and O. Catoni.

Robust linear least squares regression, 2010.

arXiv.



O. Catoni.

Challenging the empirical mean and empirical variance : a deviation study, 2010.

arXiv :1009.2048v1 [math.ST].



O. Catoni.

Pac-bayesian learning bounds, 2011.

Lecture notes, <http://www.math.ens.fr>.



J. Langford and J. Shawe-Taylor.

PAC-bayes & margins.

In *Advances in Neural Information Processing Systems*, pages 423–430, 2002.



David Mcallester.

Simplified pac-bayesian margin bounds.

In *In COLT*, pages 203–215, 2003.