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

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

On veut minimiser

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

Exemples :

1. Classification linéaire à noyau (SVM).
   $\mathcal{W} = \mathcal{X} \times \{-1, +1\}$, où $k : \mathcal{X} \to \mathbb{R}$ est un noyau positif.

$$L\big[(x,y);(\alpha_k,x_k)_{k=1}^d\big] = \mathbb{1}\left[\left(\sum_{k=1}^d \alpha_k k(x_k,x) + b\right) y \le 0\right]$$
$$= \mathbb{1}\Big[\Big(\big\langle \Psi(x), \textstyle\sum \alpha_k \Psi(x_k)\big\rangle + b\Big) y \le 0\Big],$$

où $\Psi : \mathcal{X} \to \mathcal{H} = \left\{ \sum_{j=1}^{+\infty} \alpha_j k(x_i,\cdot)\,;\, \sum_{j=1}^{\infty} \alpha_j^2 k(x_j,x_j) < \infty \right\}$, espace de Hilbert autoreproduisant. Exemples de noyaux :

$$k(x_1,x_2) = \big(1 + \langle x_1, x_2\rangle\big)^s, \qquad \dim(\mathcal{H}) < \infty,$$
$$k(x_1,x_2) = \exp\big(-\|x_1 - x_2\|^2\big), \qquad \dim(\mathcal{H}) = \infty.$$

2. Classification linéaire simple en dimension finie.

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

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

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

3. Régression aux moindres carrés.

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

Couvre le cas fonctionnel où $x = \big[\varphi_j(z)\big]_{j=1}^{d}$, 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 $\pi = \mathcal{N}(0, \beta^{-1}\,\mathrm{Id})$.

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

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

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

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

Théorème

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

$$\lambda^{-1} \int \psi\left[\lambda f(w,\theta')\right] \mathrm{d}\pi_\theta(\theta')\, \mathrm{d}\overline{\mathbb{P}}_S(w)$$
$$\leq \int f(w,\theta')\, \mathrm{d}\pi_\theta(\theta')\, \mathrm{d}\mathbb{P}(w) + \frac{\lambda}{2} \int f(w,\theta')^2\, \mathrm{d}\pi_\theta(\theta')\, \mathrm{d}\mathbb{P}(w)$$
$$+ \frac{\mathcal{K}(\pi_\theta, \pi_{\theta_\star}) - \log(\epsilon)}{n\lambda},$$

où $\overline{\mathbb{P}}_{(w_1,\ldots,w_n)} = \dfrac{1}{n}\sum_{i=1}^{n} \delta_{w_i}$ *est la mesure empirique de l'échantillon et* $\mathcal{K}(\rho,\mu) = \displaystyle\int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\mu}\right) \mathrm{d}\rho$, *quand* $\rho \ll \mu$ *et* $+\infty$ *sinon.*

Corollaire

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

$$\lambda^{-1} \int \psi\left[\lambda f(w,\theta')\right] \mathrm{d}\pi_{\widehat{\theta}(s)}(\theta')\, \mathrm{d}\overline{\mathbb{P}}_s(w)\, \mathrm{d}\mathbb{P}^{\otimes n}(s)$$
$$\leq \int \left(\int f \mathrm{d}\pi_{\widehat{\theta}(s)} \mathrm{d}\mathbb{P} + \frac{\lambda}{2} \int f^2 \mathrm{d}\pi_{\widehat{\theta}(s)} \mathrm{d}\mathbb{P} + \frac{\mathcal{K}(\pi_{\widehat{\theta}(s)}, \pi_{\theta_\star}) + 1}{n\lambda} \right) \mathrm{d}\mathbb{P}^{\otimes n}(s),$$

Démonstration.

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

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

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

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

□

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

$$\int \mathcal{K}(\pi_{\widehat{\theta}(s)}, \pi_{\theta_\star})\, \mathrm{d}\mathbb{P}^{\otimes n}(s)$$

$$\geq \inf_{\mu \in \mathcal{M}_+^1(\Theta)} \int \mathcal{K}(\pi_{\widehat{\theta}(s)}, \mu)\, \mathrm{d}\mathbb{P}^{\otimes n}(s)$$

$$= \int \mathcal{K}\left(\pi_{\widehat{\theta}(s)}, \int \pi_{\widehat{\theta}(s')} \mathrm{d}\mathbb{P}^{\otimes n}(s')\right) \mathrm{d}\mathbb{P}^{\otimes n}(s)$$

$$= \text{ information mutuelle}$$

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

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

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

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 = Bernoulli de paramètre p, et

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

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

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

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

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

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

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

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

alors que $\inf_{\lambda \in \mathbb{R}_+} \Phi_\lambda^{-1} \left(q + \frac{\delta}{\lambda} \right) = \sup\Big\{ p \in [0,1] \, ; \, K(q,p) \leq \delta \Big\}$

$$\in q + \sqrt{2\delta q(1-q)} + [-\delta q, 2\delta(1-q)]].$$

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

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

Lemme

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

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

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

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

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

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

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

où

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

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

Théorème

Avec probabilité au moins $1 - \epsilon$,

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

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 :

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

Corollaire

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

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

où

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

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

Deuxième partie : Régression aux moindres carrés

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

$$
\begin{aligned}
R(\theta) &= \int L(w,\theta)\, \mathrm{d}\mathbb{P}(w), \\
R'(\theta,\theta') &= R(\theta) - R(\theta'), \\
L'(w,\theta,\theta') &= L(w,\theta) - L(w,\theta') \\
&= \langle \theta - \theta', x\rangle^2 + 2\langle \theta - \theta', x\rangle \big(\langle \theta', x\rangle - y\big).
\end{aligned}
$$

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

Proposition

Avec probabilité au moins $1-\epsilon$,

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

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

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

$$
\begin{aligned}
\left(L'\right)^2\left(\mathbb{P}, \pi_\theta, \pi_{\theta_\star}\right) &\leq a_D R'(\theta, \theta_\star) + b_D, \\
\mathcal{K}(\pi_\theta, \pi_{\theta_\star}) &\leq p_D R'(\theta, \theta_\star) + q_D, \\
R'(\theta, \theta_\star) &\leq R(\pi_\theta, \pi_{\theta_\star}) + \xi, \\
r'_\lambda(\pi_\theta, \pi_{\theta_\star}) &\leq r'(\theta, \theta_\star) + \eta_D, \qquad \text{avec probabilité } 1 - \epsilon \text{ pour } D \text{ fixé.}
\end{aligned}
$$

Théorème (J.-Y. Audibert, O.C.)

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

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

$$\text{soit } \max\left\{R'(\widetilde{\theta}, \theta_\star), R'(\widehat{\theta}, \theta_\star)\right\} > D^2, \text{ soit}$$

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

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

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 $1-\epsilon$,

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

où

$$B_1 = \left(1+\frac{2^7 c s_4^4}{n}\right)\left[2^8 + 2^6\times 3\,c + 2^{11}(7+c)\sqrt{\kappa}\left(4+\sqrt{\kappa}\frac{\Delta^2}{\sigma_4^2}\right)\frac{s_4^2}{n}\right]$$
$$+ \frac{2^{14}c^2 s_4^6\Delta^2}{n\sigma_4^2}, \qquad B_2 = 2^8\left(1+\frac{2^7 c\,s_4^4}{n}\right),$$
$$B_3 = 2^6\left(1+\frac{2^7 c\,s_4^4}{n}\right), \qquad B_4 = 2\left(1+\frac{2^7 c\,s_4^4}{n}\right),$$

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

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

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

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

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

Théorème (J.-Y. Audibert, O.C.)

Soit $\widehat{\theta} = \arg\min_{\theta\in\Theta} L(\overline{\mathbb{P}}_S, \theta)$, *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*

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

avec probabilité au moins $1-\epsilon$,

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

Autres hypothèses sur le bruit : supposons

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

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

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

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

avec probabilité au moins $1-\epsilon$,

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

Estimation de la matrice de Gram

On veut estimer uniformément

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

Introduisons

$$r_\lambda(\theta) = \lambda^{-1} \int \psi\Big\{\lambda\big[\langle \theta, x \rangle^2 - 1\big]\Big\},$$
$$\widehat{\alpha}(\theta) = \sup\Big\{\alpha \in \mathbb{R}_+, r_\lambda(\alpha\theta) \leq 0\Big\} \in \mathbb{R}_+ \cup +\infty,$$
$$\widehat{N}(\theta) = \widehat{\alpha}(\theta)^{-2}.$$

Lemme

$$\psi\Big\{\lambda\big[\langle\theta,x\rangle^2-1\big]\Big\}\leq\int\log\Bigg\{1+\lambda\bigg(\langle\theta',x\rangle^2-1-\frac{\|x\|^2}{\beta}\bigg)+\frac{\lambda^2}{2}\bigg(\langle\theta',x\rangle-1-\frac{\|x\|^2}{\beta}\bigg)^2+\frac{2c\lambda^2\|x\|^2}{\beta}\bigg(4\langle\theta',x\rangle^2+\frac{5\|x\|^2}{\beta}\bigg)\Bigg\}\,\mathrm{d}\pi_\theta(\theta'),$$

où $c=\dfrac{15}{\log(4)}\leq 10,83.$

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 $\pi_{\theta_\star}$. Le lemme précédent permet un contrôle PAC-Bayésien uniforme en θ de $\psi\Big\{\lambda\big[\langle\theta, x\rangle^2 - 1\big]\Big\}$ et conduit au résultat suivant.

Théorème

Pour tous $n \in \mathbb{N}$ *et* $\epsilon \in]0,1]$ *tels que* $2(\eta+\gamma) + \dfrac{\kappa\eta}{2(\kappa-1)} < 1$, *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* $\theta \in \mathbb{R}^d$,

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

Dans le cas où $\mathrm{d}\mathbb{P}(x)$ 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 $\widehat{\theta}$.
- Calculer un préestimateur en interprétant $R(\theta)$ comme le carré d'une norme en dimension $d+1$ suivant l'identité

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

  Utiliser ensuite la méthode précédente d'estimation de $\arg\min_{\Theta_\star} R$, dans un voisinage (aléatoire) $\Theta_\star$ de $\theta_\star$ 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 $\sqrt{R(\widehat{\theta}) - R(\theta_\star)}$.


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.