

Dimension dependent and dimension free PAC-Bayes bounds for the Gram matrix

Olivier Catoni

CNRS, INRIA (projet CLASSIC)

Département de Mathématiques et Applications,

ENS, 45 rue d'Ulm, 75 230 Paris Cedex 05,

`Olivier.Catoni@ens.fr`

International Workshop on Statistical Learning,

PREMO LAB,

INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS,

Moscow, June 27, 2013

Let $X_i \in \mathbb{R}^d$, $1 \leq i \leq n$ be vector valued i.i.d. random variables with common distribution $\mathbb{P} \in \mathcal{M}_+^1(\mathbb{R}^d)$.

Question : Estimate the quadratic form

$$N(\theta) = \int \langle x, \theta \rangle^2 d\mathbb{P}(x)$$

for all $\theta \in \mathbb{R}^d$, or equivalently all $\theta \in S_d$, the sphere of \mathbb{R}^d .

This is a way to estimate the expected Gram matrix

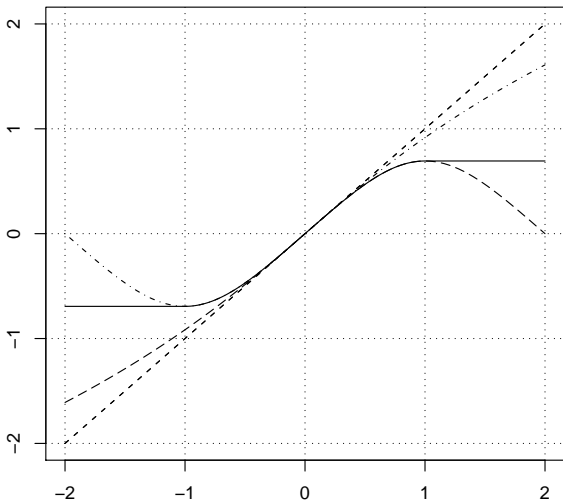
$$G \stackrel{\text{def}}{=} \int x x^\top d\mathbb{P}(x),$$

since $N(\theta) = \theta^\top G \theta$. We will assume that

$$\int \|x\|^2 d\mathbb{P}(x) = \mathbf{Tr}(G) < +\infty.$$

Let us introduce the **influence function**

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



$z \mapsto \psi(z)$, compared with $z \mapsto z$
 $z \mapsto \log(1 + z + z^2/2)$, and $z \mapsto -\log(1 - z + z^2/2)$

It is symmetric, non decreasing, bounded and satisfies for any $z \in \mathbb{R}$,

$$-\log(1 - z + z^2/2) \leq \psi(z) \leq \log(1 + z + z^2/2),$$

$$-\log(2) \leq \psi(z) \leq \log(2).$$

Dimension dependent bounds

Let $\bar{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ and

$$r_\lambda(\theta) = \lambda^{-1} \int \psi \left[\lambda (\langle \theta, x \rangle^2 - 1) \right] d\bar{\mathbb{P}}(x),$$

where $\lambda > 0$ will be chosen later. Let us consider the estimator

$$\hat{N}(\theta) = \hat{\alpha}(\theta)^{-2}, \quad \text{where} \quad \hat{\alpha}(\theta) = \sup \{ \alpha \in \mathbb{R}_+ : r_\lambda(\alpha\theta) \leq 0 \}.$$

This is justified by the fact that

$$\int \lim_{\lambda \rightarrow 0} r_\lambda(\theta) d\mathbb{P}^{\otimes n} = N(\theta) - 1.$$

Proposition

Let us assume that $\kappa = \sup_{\theta \neq 0} \frac{\int \langle \theta, x \rangle^4 d\mathbb{P}(x)}{\left(\int \langle \theta, x \rangle^2 d\mathbb{P}(x) \right)^2} < \infty$ and

$$\text{let us put } \lambda = \sqrt{\frac{2}{(\kappa - 1)n} [\log(\epsilon^{-1}) + 1.11 d]}.$$

For any $\epsilon > 0$, any n such that

$$n > \left(27\sqrt{\kappa d} + \frac{5\kappa - 4}{\sqrt{2(\kappa - 1)}} \sqrt{\log(\epsilon^{-1}) + 1.11 d} \right)^2,$$

with probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\left| \frac{N(\theta)}{\widehat{N}(\theta)} - 1 \right| \leq \frac{\mu}{1 - 2\mu},$$

where $\mu = \sqrt{\frac{2(\kappa - 1)}{n} [\log(\epsilon^{-1}) + 1.11 d]} + \sqrt{\frac{2\kappa \times 89 d}{n}}$.

Obtaining a true quadratic form

Let us assume that with probability $1 - 2\epsilon$,

$$B_-(\theta) \leq N(\theta) \leq B_+(\theta), \quad \theta \in \mathbb{R}^d.$$

Let $\Theta \subset \mathbb{R}^d$ be any finite set (e.g. a δ -net of S_d). Let

$$\begin{aligned} \widehat{G} = \sum_{\theta \in \Theta} \xi(\theta) \theta \theta^\top, \text{ where } \xi \in \arg \min \frac{1}{2} \sum_{(\theta_1, \theta_2) \in \Theta^2} \xi(\theta_1) \xi(\theta_2) \langle \theta_1, \theta_2 \rangle^2 \\ - \sum_{\theta \in \Theta} \xi(\theta) \frac{B_-(\theta) + B_+(\theta)}{2} + |\xi(\theta)| \frac{B_+(\theta) - B_-(\theta)}{2}. \end{aligned}$$

Then with probability at least $1 - 2\epsilon$,

$$\|\widehat{G}\|_2^2 \stackrel{\text{def}}{=} \mathbf{Tr}(\widehat{G}\widehat{G}^\top) \leq \mathbf{Tr}(GG^\top) \leq \mathbf{Tr}(G)^2,$$

so that $\theta_2^\top \widehat{G} \theta_2 - \theta_1^\top \widehat{G} \theta_1 \leq 2 \mathbf{Tr}(G) \|\theta_2 - \theta_1\|$, $\theta_1, \theta_2 \in \mathbb{R}^d$

$$\text{and } B_-(\theta) \leq \theta^\top \widehat{G} \theta \leq B_+(\theta), \quad \theta \in \Theta.$$

\widehat{G} minimizes

$$\sup_{\xi_+ \geq 0, \xi_- \geq 0} \frac{1}{2} \|\widehat{G}\|_2^2 + \sum_{\theta \in \Theta} \xi_+(\theta) [B_-(\theta) - \theta^\top \widehat{G} \theta] + \xi_-(\theta) [\theta^\top \widehat{G} \theta - B_+(\theta)]$$

and there is no duality gap, so you can minimize in \widehat{G} first and then as $\widehat{\xi}_+(\theta)\widehat{\xi}_-(\theta) = 0$, you can assume that $\xi_-(\theta)\xi_+(\theta) = 0$ and introduce $\xi(\theta) = \xi_+(\theta) - \xi_-(\theta)$, so that $|\xi(\theta)| = \xi_+(\theta) + \xi_-(\theta)$.

Least square regression estimator

Let (X, Y) be a random vector, where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. Let (X_i, Y_i) , $1 \leq i \leq n$ be independent copies of (X, Y) and

$$R(\theta) = \mathbb{E}[(\langle \theta, X \rangle - Y)^2].$$

Introduce the quadratic form $N(\theta, \gamma) = \mathbb{E}[(\langle \theta, X \rangle - \gamma Y)^2]$.

Assume that $\widehat{N}(\theta, \gamma)$ is a quadratic estimator of $N(\theta, \gamma)$ and $\eta > 0$ a small real parameter such that with probability at least

$1 - 2\epsilon$, $\left| \frac{N(\theta, \gamma)}{\widehat{N}(\theta, \gamma)} - 1 \right| \leq \eta$. Let $\theta_* \in \arg \min R$ and

$\widehat{\theta} \in \arg \min_{\theta \in \mathbb{R}^d} \widehat{N}(\theta, 1)$. With probability at least $1 - 2\epsilon$,

$$\begin{aligned} (1 - \eta) \widehat{N}(\widehat{\theta} - \theta_*, 0) &\leq N(\widehat{\theta} - \theta_*, 0) = R(\widehat{\theta}) - R(\theta_*) \\ &\leq \frac{\eta^2}{1 - \eta} \widehat{N}(\widehat{\theta}, 1) \leq \frac{\eta^2}{1 - \eta} \widehat{N}(\theta_*, 1) \leq \frac{\eta^2}{(1 - \eta)^2} R(\theta_*). \end{aligned}$$

Proof : Let us put $\widehat{R}(\theta) = \widehat{N}(\theta, 1)$.

$$\begin{aligned} 0 &= \sup \left\{ \frac{1}{4} \left(\widehat{R}(\widehat{\theta} + \theta') - \widehat{R}(\widehat{\theta} - \theta') \right), \quad \theta' : \widehat{N}(\theta', 0) \leq \widehat{R}(\widehat{\theta}) \right\} \\ &\geq \sup \left\{ \frac{1}{4} \left(R(\widehat{\theta} + \theta') - R(\widehat{\theta} - \theta') \right) - \frac{\eta}{4} \left(\widehat{R}(\widehat{\theta} + \theta') + \widehat{R}(\widehat{\theta} - \theta') \right), \right. \\ &\quad \left. \theta' : \widehat{N}(\theta', 0) \leq \widehat{R}(\widehat{\theta}) \right\} \\ &= \sup \left\{ \mathbb{E} \left((\langle \widehat{\theta}, X \rangle - Y) \langle \theta', X \rangle \right) - \frac{\eta}{2} \left(\widehat{R}(\widehat{\theta}) + \widehat{N}(\theta', 0) \right), \right. \\ &\quad \left. \theta' : \widehat{N}(\theta', 0) \leq \widehat{R}(\widehat{\theta}) \right\} \\ &\geq \sup \left\{ \mathbb{E} \left(\langle \widehat{\theta} - \theta_*, X \rangle \langle \theta', X \rangle \right) - \eta \widehat{R}(\widehat{\theta}), \right. \\ &\quad \left. \theta' : N(\theta', 0) \leq (1 - \eta) \widehat{R}(\widehat{\theta}) \right\}, \end{aligned}$$

Taking $\theta' = c(\widehat{\theta} - \theta_*)$, and using $N(\widehat{\theta} - \theta_*, 0) = R(\widehat{\theta}) - R(\theta_*)$, we get

$$0 \geq \sqrt{(1 - \eta) \widehat{R}(\widehat{\theta}) [R(\widehat{\theta}) - R(\theta_*)]} - \eta \widehat{R}(\widehat{\theta}).$$

Dimension free bounds (*Joint work with Ilaria Giulini*)

With probability $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$, and some estimator \widehat{N} to be described in the proof,

$$\mathbb{1}(4\mu < 1) \left| \frac{\widehat{N}(\theta)}{N(\theta)} - 1 \right| \leq \frac{\mu}{1 - 4\mu},$$

where, for $n \leq 10^{20}$,

$$\mu = \sqrt{\frac{2.07(\kappa - 1)}{n} \left[\log(\epsilon^{-1}) + 4.3 + \frac{1.6 \times \|\theta\|^2 \mathbf{Tr}(G)}{N(\theta)} \right]} + \sqrt{\frac{2\kappa}{n} \times \frac{92 \mathbf{Tr}(G)}{N(\theta)}}.$$

Let us recall that $\mathbf{Tr}(G) = \int \|x\|^2 d\mathbb{P}(x) = \sum_{i=1}^d N(\theta_i)$, for any orthogonal basis $(\theta_i, 1 \leq i \leq n)$.

Proof

Let $r_\lambda(\theta) = \int \psi(\langle \theta, x \rangle^2 - \lambda) d\bar{\mathbb{P}}(x)$. Consider the Gaussian parameter perturbations $\pi_\theta = \mathcal{N}(\theta, \beta^{-1}\mathbb{I}_d)$, where \mathbb{I}_d is the identity matrix of size $d \times d$. Let

$$\hat{\alpha}(\theta) = \sup\{\alpha \in \mathbb{R}_+ : r_\lambda(\alpha\theta) \leq 0\}.$$

Proposition

Let $c = \frac{15}{\log(4)} \leq 11$.

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

Indeed,

$$\psi\left(\int h \, d\rho\right) \leq \int \psi(h) \, d\rho + \min\{\log(4), \mathbf{Var}(h \, d\rho)\},$$

because $y \mapsto \psi(y) + (y - \int h \, d\rho)^2$ is convex, since $\psi''(y) \geq -2$.
As moreover

$$\langle \theta, x \rangle^2 - \lambda = \int \langle \theta', x \rangle^2 \, d\pi_\theta(\theta') - \lambda - \frac{\|x\|^2}{\beta},$$

we get

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

Lemma

If $W \sim \mathcal{N}(0, \sigma^2)$,

$$\min\{a, bm^2 + c\} \leq \mathbb{E}\left(\min\{2a, 2b(m + W)^2 + 2b\sigma^2 + c\}\right),$$

$a, b, c \in \mathbb{R}_+, m \in \mathbb{R}.$

The proof of this lemma is based on the inequalities $m^2 \leq 2(m + W)^2 + 2W^2$ and

$$\begin{aligned} \min\{a, y + z\} &\leq \min\{a, y\} + \min\{a, z\} \\ &\leq \min\{2a, y + z\}, \quad a, y, z \in \mathbb{R}_+. \end{aligned}$$

Accordingly

$$\begin{aligned} \psi(\langle \theta, x \rangle^2 - \lambda) &\leq \int \psi\left(\langle \theta', x \rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\right) d\pi_{\theta}(\theta') \\ &+ \int \min\left\{4\log(2), \frac{8\|x\|^2\langle \theta', x \rangle^2}{\beta} + \frac{10\|x\|^4}{\beta^2}\right\} d\pi_{\theta}(\theta'). \end{aligned}$$

We will now use

Lemma

For any $a, b, y, \in \mathbb{R}_+$ and $c = \frac{a}{b} [\exp(b) - 1]$,

$$\log(a) + \min\{b, y\} \leq \log(a + cy).$$

Applying this lemma to $a \leq 2$, $b = 4 \log(2)$, and the corresponding $c = \frac{15}{\log(4)}$ ends the proof of the previous proposition.

PAC-Bayes bound

Proposition

For any measure $\nu \in \mathcal{M}_+^1(\Theta)$, real number $a > -1$ and any measurable function $f : \mathcal{X} \times \Theta \rightarrow [a, +\infty[$, with probability at least $1 - \epsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$ such that $\mathcal{K}(\rho, \nu) < \infty$,

$$\int \log[1 + f(x, \theta)] d\rho(\theta) d\bar{\mathbb{P}}(x) \leq \int f(x, \theta) d\rho(\theta) d\mathbb{P}(x) + \frac{\mathcal{K}(\rho, \nu) - \log(\epsilon)}{n}.$$

The proof is based on the fact that

$$\int h d\rho - \mathcal{K}(\rho, \nu) \leq \log\left(\int \exp(h) d\nu\right),$$

for any upper-bounded measurable function $h : \Theta \rightarrow \mathbb{R}$.

We get with probability $1 - \epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned} \int \psi(\langle \theta, x \rangle^2 - \lambda) d\bar{\mathbb{P}}(x) &\leq \int \left[\langle \theta, x \rangle^2 - \lambda \right. \\ &\quad \left. + \frac{1}{2} \left((\langle \theta, x \rangle^2 - \lambda)^2 + \frac{4}{\beta} \langle \theta, x \rangle^2 \|x\|^2 + \frac{2}{\beta^2} \|x\|^4 \right) \right. \\ &\quad \left. + \frac{2c\|x\|^2}{\beta} \left(\frac{9\|x\|^2}{\beta} + 4\langle \theta, x \rangle^2 \right) \right] d\mathbb{P}(x) \\ &\quad + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n}. \end{aligned}$$

Using the Cauchy-Schwartz inequality will make the following quantities appear

$$s_4 = \left(\int \|x\|^4 d\mathbb{P}(x) \right)^{1/4},$$

$$\kappa = \sup \left\{ \int \langle \theta, x \rangle^4 d\mathbb{P}(x), \theta \in \mathbb{R}^d \text{ s. t. } \int \langle \theta, x \rangle^2 d\mathbb{P}(x) = 1 \right\},$$

$$\xi = \frac{\kappa\lambda}{2},$$

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

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

$$\delta = \frac{\beta}{2n\lambda}.$$

Proposition

With probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\frac{r_\lambda(\alpha\theta)}{\lambda} \leq \xi \left(\frac{N(\theta)}{\lambda} \alpha^2 - 1 \right)^2 + (1 + \gamma) \left(\frac{N(\theta)}{\lambda} \alpha^2 - 1 \right) + \eta + \delta \|\theta\|^2 \alpha^2,$$

$$\frac{r_\lambda(\alpha\theta)}{\lambda} \geq -\xi \left(\frac{\alpha^2 N(\theta)}{\lambda} - 1 \right)^2 + \left(1 - \gamma - \frac{\lambda \|\theta\|^2 \delta}{N(\theta)} \right) \left(\frac{N(\theta)}{\lambda} \alpha^2 - 1 \right) - \eta - \frac{\lambda \|\theta\|^2 \delta}{N(\theta)}.$$

Let

$$\Phi_-(z) = z \left(1 - \frac{\eta + \frac{\delta\lambda}{z}}{1 + \gamma - \eta - \frac{\delta\lambda}{z}} \right) \mathbb{1} \left(\xi - \gamma + \eta + \frac{\delta\lambda}{z} < 1 \right),$$
$$\Phi_+(z) = z \left(1 + \frac{\eta + \frac{\delta\lambda}{z}}{1 - \gamma - \eta - \frac{2\delta\lambda}{z}} \right)^{-1} \mathbb{1} \left(\xi + \gamma + \eta + \frac{2\delta\lambda}{z} < 1 \right).$$

Replacing $N(\theta)$ with $\Phi_-(\lambda/\widehat{\alpha}^2)$ in the first inequality of the previous proposition and $\lambda/\widehat{\alpha}^2$ with $\Phi_+[N(\theta)]$ in the second inequality, we can prove that

$$\Phi_-\left(\frac{\lambda}{\widehat{\alpha}^2}\right) \leq N(\theta),$$
$$\Phi_+[N(\theta)] \leq \frac{\lambda}{\widehat{\alpha}^2}.$$

Proposition

With probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$B_- = \sup_{\lambda, \beta} \Phi_- \left(\frac{\lambda}{\hat{\alpha}^2} \right) \leq N(\theta) \leq \inf_{\lambda, \beta} \Phi_+^{-1} \left(\frac{\lambda}{\hat{\alpha}^2} \right) = B_+.$$

Let us put $\hat{N}(\theta) = \frac{B_- + B_+}{2}$, we get with probability at least $1 - 2\epsilon$, for any $\lambda, \beta \in \mathbb{R}_+$,

$$N(\theta) - \hat{N}(\theta) \leq \frac{N - B_-}{2} = \frac{N - \Phi_-(\lambda/\hat{\alpha}^2)}{2} \leq \frac{N(\theta) - \Phi_- \circ \Phi_+[N(\theta)]}{2},$$
$$\hat{N}(\theta) - N(\theta) \leq \frac{\Phi_+^{-1}(\lambda/\hat{\alpha}^2) - N(\theta)}{2} \leq \frac{\Phi_+^{-1} \circ \Phi_-^{-1}[N(\theta)] - N(\theta)}{2}.$$

Putting $r(z) = \frac{\eta + \frac{\delta\lambda}{z}}{1 - \gamma - \eta - \frac{2\delta\lambda}{z}}$, and $c(z) = \eta + \frac{\delta\lambda}{z}$, we can prove that

$$\Phi_-(z) \geq z[1 - r(z)] \mathbf{1}[4c(z) < 1],$$

$$\Phi_+(z) \geq z[1 + r(z)]^{-1} \mathbf{1}[4c(z) < 1]$$

$$\Phi_-^{-1}(z) \mathbf{1}[4c(z) < 1] \leq z[1 - r(z)]^{-1},$$

$$\Phi_+^{-1}(z) \mathbf{1}[4c(z) < 1] \leq z[1 + r(z)].$$

From this we can deduce

Proposition

With probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\mathbb{1}[4c(N(\theta)) < 1] \left| N(\theta) - \widehat{N}(\theta) \right| \leq N(\theta) \frac{c(N(\theta))}{1 - 4c(N(\theta))}.$$

The announced result is then obtained by optimizing the value of $c(N(\theta))$ by appropriate choices of λ and β .