PAC-Bayes bounds for the Gram matrix and least squares regression

Olivier Catoni

CREST – EXCESS,
CNRS UMR 9194
Olivier.Catoni@ensae.fr

SMILE seminar,

ParisTech,
ENS, 45 rue d'Ulm, 75230 Paris Cedex,

May 4, 2015

1. Estimate the Gram matrix

$$G \stackrel{\text{def}}{=} \int x\, x^\top \,\mathrm{dP}(x),$$

or equivalently the quadratic form

$$N(\theta) = \theta^\top G \theta = \int \langle x, \theta \rangle^2 \,\mathrm{dP}(x)$$

for all $\theta \in \mathbb{S}_d$, the sphere of $\mathbb{R}^d$, from an i.i.d. sample $(X_i)_{i=1}^n \sim \mathrm{P}^{\otimes n}$.

2. Estimate

$$\arg\min_{\theta \in \mathbb{R}^d} \int \big(y - \langle \theta, x \rangle\big)^2 \,\mathrm{dP}(x, y).$$

from an i.i.d. sample $(X_i, Y_i)_{i=1}^n \sim \mathrm{P}^{\otimes n}$.

3. We will assume that

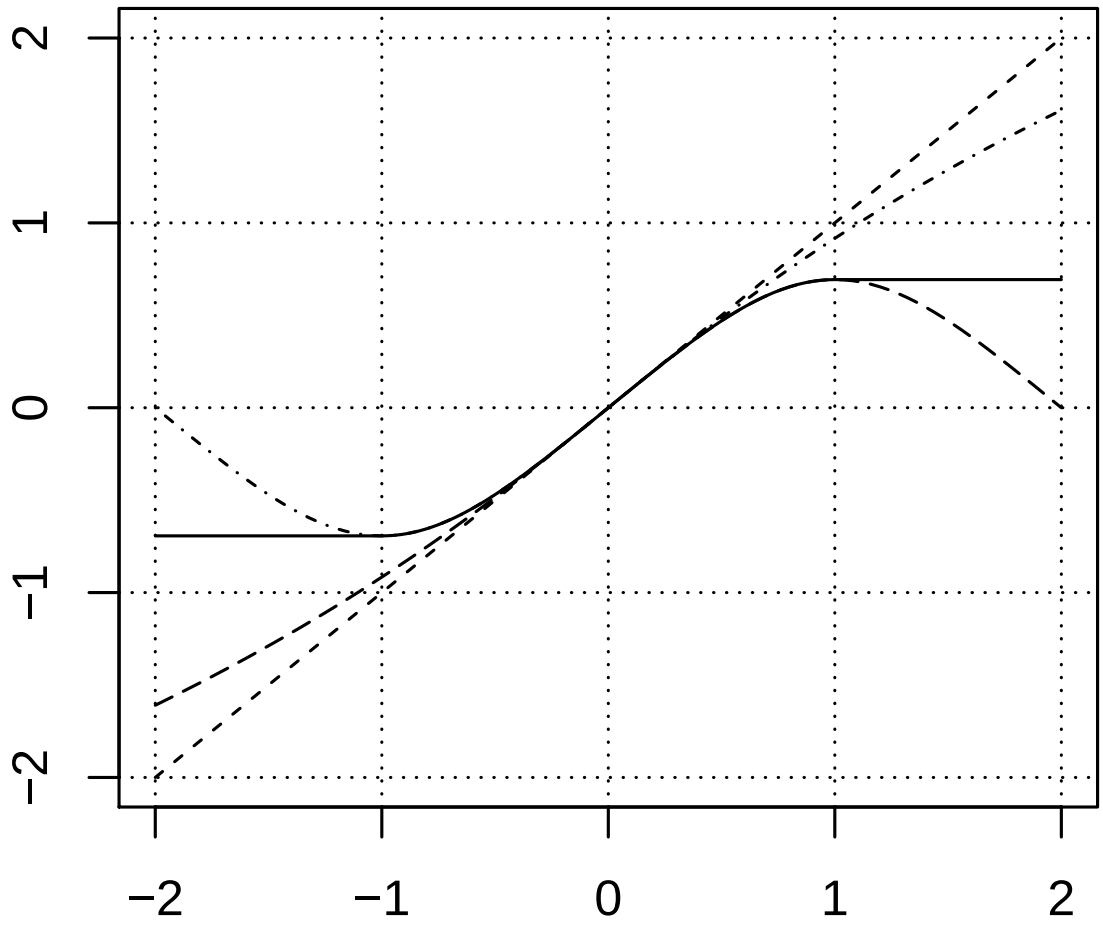
$$\int \|x\|^2 \,\mathrm{dP}(x) = \mathbf{Tr}(G) < +\infty.$$

Robust Gram matrix estimate

Let us introduce the influence function

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

Robust Gram matrix estimate



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

Robust Gram matrix estimate

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

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

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

Let $\overline{\mathrm{P}} = \dfrac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and

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

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

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

This makes sense since

$$\lim_{\lambda \to 0} r_\lambda(\alpha\theta) = \alpha^2 \int \langle \theta, x\rangle^2 \mathrm{d}\overline{\mathrm{P}}(x) - 1$$

Proposition

Let us assume that $\kappa = \sup_{\theta \neq 0} \dfrac{\int \langle \theta, x \rangle^4 \, \mathrm{dP}(x),}{\left(\int \langle \theta, x \rangle^2 \, \mathrm{dP}(x) \right)^2} < \infty$ *and*

$$\textit{let us put } \lambda = \sqrt{\frac{2}{(\kappa - 1) n} \Big[\log(\epsilon^{-1}) + 0.73 \, d \Big]}.$$

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

$$n > \left(20 \sqrt{\kappa d} + \left(\frac{5}{2} + \frac{1}{2(\kappa - 1)} \right) \sqrt{2(\kappa - 1) \left[\log(\epsilon^{-1}) + 0.73 \, d \right]} \right)^2,$$

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 $\mu = \sqrt{\dfrac{2(\kappa - 1)}{n} \Big[\log(\epsilon^{-1}) + 0.73 \, d \Big]} + 6.81 \sqrt{\dfrac{2\kappa d}{n}}.$

Obtaining a true quadratic form

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

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

We can take for example

$$\widehat{N}(\theta)\left(1-\frac{\mu}{1-2\mu}\right) \le N(\theta) \le \widehat{N}(\theta)\left(1+\frac{\mu}{1-2\mu}\right)$$

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

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

Obtaining a true quadratic form

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

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

so that

$$\theta_2^\top \widehat{G}\theta_2 - \theta_1^\top \widehat{G}\theta_1 \le \|\theta_1+\theta_2\|\,\mathbf{Tr}(G)\|\theta_2-\theta_1\|, \qquad \theta_1,\theta_2 \in \mathbb{R}^d,$$

and

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

Proof

$\widehat{G}$ minimizes

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

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 $|\xi(\theta)| = \xi_+(\theta) + \xi_-(\theta)$.

Proposition

Assume that $B_{+/-}(\theta) = \|\theta\| B_{+/-}(\|\theta\|^{-1}\theta)$ *and are continuous functions. With probability at least* $1 - 2\epsilon$*, for any* $\eta > 0$*, there is* $\delta > 0$*, such that if* Θ *is a* δ*-net on the sphere* $\mathbb{S}_d$*, then*

$$B_-(\theta) - \eta\|\theta\|^2 \leq \theta^\top \widehat{G}\theta \leq B_+(\theta) + \eta\|\theta\|^2, \qquad \theta \in \mathbb{R}^d.$$

The empirical Gram matrix estimator

Here we consider

$$\overline{G} = \int x\, x^\top \, \mathrm{d}\overline{\mathrm{P}}(x) \text{ and } \overline{N}(\theta) = \theta^\top \overline{G}\theta.$$

Remark that, with probability $1-\epsilon$, for any $\theta \in \mathbb{S}_d$, $\widehat{N}(\theta) > 0$, and therefore

$$\int \psi\Big[\lambda(\langle \theta, x\rangle^2 \widehat{N}(\theta)^{-1} - 1)\Big]\, \mathrm{d}\overline{\mathrm{P}}(x) = 0,$$

so that

$$\frac{\overline{N}(\theta)}{\widehat{N}(\theta)} - 1 = \lambda^{-1} \int g\Big[\lambda(\langle \theta, x\rangle^2 \overline{N}(\theta)^{-1} - 1)\Big]\, \mathrm{d}\overline{\mathrm{P}}(x),$$

where $g(z) = z - \psi(z) \leq z_+^3/3$.

The empirical Gram matrix estimator

Proposition

With probability at least $1-\epsilon$,

$$-\frac{\lambda^2}{3}\int\Big(1-\langle\theta,x\rangle^2\widehat{N}(\theta)^{-1}\Big)_+^3\,\mathrm{d}\overline{\mathrm{P}}(x)\leq\frac{\overline{N}(\theta)}{\widehat{N}(\theta)}-1$$
$$\leq\frac{\lambda^2}{3}\int\Big(\langle\theta,x\rangle^2\widehat{N}(\theta)^{-1}-1\Big)_+^3\,\mathrm{d}\overline{\mathrm{P}}(x),$$

where

$$\lambda^2=\frac{2\big[\log(\epsilon^{-1})+0.73d\big]}{(\kappa-1)n}.$$

The empirical Gram matrix estimator

Considre $R = \max_{i=1,\ldots,n} \| G^{-1/2} X_i \|$. With probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$-\delta - \gamma_+ \leq \frac{N(\theta)}{\overline{N}(\theta)} - 1 \leq \frac{\delta + \gamma_-}{1 - \gamma_-},$$

where

$$\delta = \frac{\mu}{1 - 2\mu}, \quad \gamma_- = \frac{2\big[\log(\epsilon^{-1}) + 0.73\, d\big]}{3(\kappa - 1)n}, \quad \text{and } \gamma_+ = \gamma_- R^4 (1 + \delta^2).$$

The empirical Gram matrix estimator

Moreover, when, for some exponent $p \in]0,1]$,

$$\mathbb{E}\left\{\exp\left[\frac{\alpha}{2}\left(\|G^{-1/2}X\|^{2p} - d^p - \eta^p\right)\right]\right\} \leq 1,$$

with probability at least $1-\epsilon$,

$$R^4 \leq \left(\frac{2\log(n/\epsilon)}{\alpha} + d^p + \eta^p\right)^{2/p}.$$

When $X \in \mathbb{R}^d$ is a Gaussian vector, for any $\alpha \in]0,1[$, with probability at least $1-\epsilon$,

$$R^4 \leq \left(\frac{2\log(n/\epsilon) + d\log\left[(1-\alpha)^{-1}\right]}{\alpha}\right)^2.$$

The empirical Gram matrix estimator

Polynomial moment assumptions: Define

$$\breve{\gamma}_+ = \gamma_-(1+\delta)^3\left[\underbrace{\mathbb{E}(\|G^{-1/2}X\|^6)}_{\geq d^3} + \left(\frac{\mathbb{E}(\|G^{-1/2}X\|^{12})}{n\epsilon}\right)^{1/2}\right].$$

With probability at least $1-3\epsilon$, for any $\theta \in \mathbb{S}_d$,

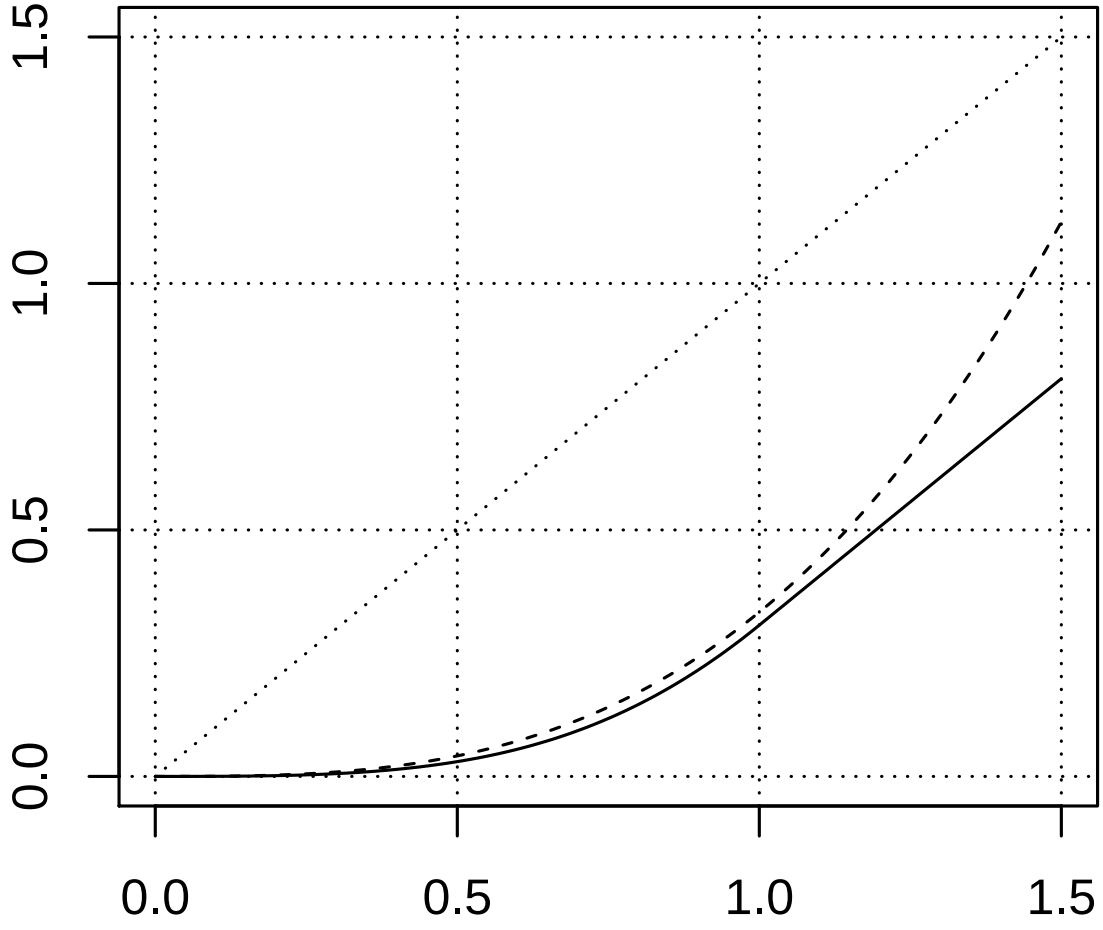
$$-\frac{\delta+\breve{\gamma}_+}{1+\breve{\gamma}_+} \leq \frac{N(\theta)}{\overline{N}(\theta)} - 1 \leq \frac{\delta+\gamma_-}{1-\gamma_-}$$

Proof.

Remark that $\langle\theta, X_i\rangle^2 \leq N(\theta)\|G^{-1/2}X_i\|^2$, so that

$$\int\left(\langle\theta,x\rangle^2\widehat{N}(\theta)^{-1}-1\right)_+^3 \mathrm{d}\overline{\mathrm{P}}(x) \leq (1+\delta)^3\int\|G^{-1/2}X\|^6\,\mathrm{d}\overline{\mathrm{P}}(x),$$

and conclude using the Bienaymé Chebyshev inequality. □



$z \mapsto g(z)$, compared with $z \mapsto z^3/3$

Least squares regression with a random design

Solve

$$\inf_{\theta\in\mathbb{R}^d} \int \big(y - \langle\theta, x\rangle\big)^2 \,\mathrm{dP}(x,y), \text{ where } \mathrm{P} \in \mathscr{M}_+^1(\mathbb{R}^d \times \mathbb{R}),$$

knowing an i.i.d. sample $(X_i, Y_i)_{i=1,\dots,n}$, $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.

Introduce the risk function

$$R(\theta) = \int \big(y - \langle\theta, x\rangle\big)^2 \,\mathrm{dP}(x,y), \qquad \theta \in \mathbb{R}^d,$$

and the homogeneous quadratic form

$$\begin{aligned} N(\theta,\gamma) &= \int \big(\gamma y - \langle\theta, x\rangle\big)^2 \,\mathrm{dP}(x,y), \\ &= \begin{pmatrix}\theta\\ \gamma\end{pmatrix}^\top \underbrace{\left[\int \begin{pmatrix}x\\ -y\end{pmatrix}\begin{pmatrix}x\\ -y\end{pmatrix}^\top \mathrm{dP}(x,y)\right]}_{\stackrel{\text{def}}{=} G} \begin{pmatrix}\theta\\ \gamma\end{pmatrix}, \quad \theta \in \mathbb{R}^d, \gamma \in \mathbb{R}. \end{aligned}$$

Least squares regression with a random design

Let $\theta_* \in \arg\min_{\theta \in \mathbb{R}^d} R(\theta) = \arg\min_{\theta \in \mathbb{R}^d} N(\theta, 1)$.
Assume that we have some estimator $\widehat{N}(\theta, \gamma)$, such that with probability at least $1 - 2\epsilon$, for any $(\theta, \gamma) \in \mathbb{R}^{d+1}$,

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

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

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

Least squares regression with a random design

Let us remark that in any case,

$$R(\theta_*) \leq \int y^2 \, \mathrm{dP}(y),$$

and that in the case when $Y_i = f(X_i) + W_i$, where W_i is independent of X_i, centered and $\mathbb{E}(W_i) = \sigma^2$, then

$$R(\theta_*) \leq \int f(x)^2 \, \mathrm{dP}(x) + \sigma^2 \leq \|f\|_\infty^2 + \sigma^2,$$

whereas when $Y_i = \langle \theta_*, X_i \rangle + W_i$, then $R(\theta_*) = \sigma^2$. Remind that when $\widehat{N}(\theta, \gamma)$ is the robust estimator described above, η can be taken arbitrarily close to

$$\frac{2\mu}{1-2\mu}, \text{ where } \mu = \sqrt{\frac{2(\kappa-1)}{n}\left[\log(\epsilon^{-1}) + 0.73\, d\right]} + 6.81\sqrt{\frac{2\kappa d}{n}}.$$

Proof

Remark that $\begin{pmatrix}\theta_* \\ 1\end{pmatrix}^\top G \begin{pmatrix}\xi \\ 0\end{pmatrix} = 0$, for any $\xi \in \mathbb{R}^d$.

$$R(\widehat{\theta}) - R(\theta_*) = \begin{pmatrix}\widehat{\theta}-\theta_* \\ 0\end{pmatrix}^\top G \begin{pmatrix}\widehat{\theta}-\theta_* \\ 0\end{pmatrix}$$

$$= \sup\left\{\rho^{-1}\begin{pmatrix}\widehat{\theta}-\theta_* \\ 0\end{pmatrix}^\top G \begin{pmatrix}\xi \\ 0\end{pmatrix}, \quad \xi \in \mathbb{R}^d, \begin{pmatrix}\xi \\ 0\end{pmatrix}^\top G \begin{pmatrix}\xi \\ 0\end{pmatrix} = \rho^2\right\}^2$$

$$= \sup\left\{\rho^{-1}\begin{pmatrix}\widehat{\theta} \\ 1\end{pmatrix}^\top G \begin{pmatrix}\xi \\ 0\end{pmatrix}, \quad \xi \in \mathbb{R}^d, N(\xi,0) = \rho^2\right\}^2$$

$$= \sup\left\{\frac{1}{4\rho}\Big[N(\widehat{\theta}+\xi,1) - N(\widehat{\theta}-\xi,1)\Big], \quad \xi \in \mathbb{R}^d, N(\xi,0) = \rho^2\right\}^2$$

Proof

$$\leq \sup\Bigg\{ \frac{1}{4\rho} \underbrace{\Big[\widehat{N}(\widehat{\theta}+\xi,1) - \widehat{N}(\widehat{\theta}-\xi,1)\Big]}_{=0}$$

$$+ \frac{\eta}{4\rho} \underbrace{\Big[\widehat{N}(\widehat{\theta}+\xi,1) + \widehat{N}(\widehat{\theta}-\xi,1)\Big]}_{=2\widehat{N}(\widehat{\theta},1)+2\widehat{N}(\xi,0)}, \quad \xi \in \mathbb{R}^d, N(\xi,0) = \rho^2 \Bigg\}^2$$

$$\leq \sup\Bigg\{ \frac{\eta}{2\rho} \Big[\widehat{N}(\widehat{\theta},1) + (1-\eta)^{-1} N(\xi,0)\Big], N(\xi,0) = \rho^2 \Bigg\}^2$$

$$= \left[\frac{\eta}{2}\left(\frac{\widehat{N}(\widehat{\theta},1)}{\rho} + \frac{\rho}{1-\eta}\right)\right]^2.$$

Taking $\rho = \widehat{N}(\widehat{\theta},1)^{1/2}(1-\eta)^{1/2}$, we get

$$R(\widehat{\theta}) - R(\theta_*) \leq \frac{\eta^2}{1-\eta} \widehat{N}(\widehat{\theta},1).$$

Estimation of parameters

The kurtosis coefficient κ can be deduced from a separate analysis of the distribution of X and of the distribution of the "noise" $Y - \langle \theta_*, X \rangle$. Indeed

$$\kappa = \sup_{\theta \in \mathbb{R}^d, \gamma \in \mathbb{R}} \frac{\mathbb{E}\big[\big(\gamma\, Y - \langle \theta, X \rangle\big)^4\big]}{\mathbb{E}\big[\big(\gamma\, Y - \langle \theta, X \rangle\big)^2\big]^2}$$

$$\leq 4 \max\left\{ \frac{\mathbb{E}\big[\big(\, Y - \langle \theta_*, X \rangle\big)^4\big]}{\mathbb{E}\big[\big(\, Y - \langle \theta_*, X \rangle\big)^2\big]^2}, \sup_{\theta \in \mathbb{R}^d} \frac{\mathbb{E}\big(\langle \theta, X \rangle^4\big)}{\mathbb{E}\big(\langle \theta, X \rangle^2\big)^2} \right\}.$$

Estimation of parameters

To analyse the empirical risk minimizer $\widehat{\theta}$, we need a bound on the empirical Gram estimator $\overline{N}(\theta,\gamma)$. Assume that, for some exponents $p, q \in]0,1]$ and some constants $\alpha > 0$ and $\beta > 0$,

$$\mathbb{E}\left\{\exp\left[\frac{\alpha}{2}\left(\|G^{-1/2}X\|^{2p} - d^p - \eta^p\right)\right]\right\} \leq 1,$$

$$\mathbb{E}\left\{\exp\left[\frac{\beta}{2}\left(\frac{(Y - \langle\theta_*, X\rangle)^{2q}}{\mathbb{E}\left[(Y - \langle\theta_*, X\rangle)^2\right]^q} - 1 - \gamma^q\right)\right]\right\} \leq 1.$$

Remark that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(\gamma Y_i - \langle\theta, X_i\rangle)^2}{\widehat{N}(\theta,\gamma)} - 1\right)_+^3$$

$$\leq \frac{\overline{N}(\theta,\gamma)}{\widehat{N}(\theta,\gamma)} \max_{i=1,\ldots,n} \frac{(\gamma Y_i - \langle\theta, X_i\rangle)^4}{N(\theta,\gamma)^2}(1+\delta)^2.$$

Estimation of parameters

Remark also that

$$\frac{\big(\gamma\, Y_i + \langle\theta, X_i\rangle\big)^4}{N(\theta,\gamma)^2} \le 4\max\Big\{ \frac{\big(Y_i - \langle\theta_*, X_i\rangle\big)^4}{\big(\int (y - \langle\theta_*, x\rangle)^2 \,\mathrm{dP}(x,y)\big)^2}, \|G^{-1/2} X_i\|^4 \Big\}$$

$$\le 4\max\Big\{ \Big(d^p + \eta^p + 2\alpha^{-1}\log(n/\epsilon)\Big)^{2/p}, \Big(1 + \gamma^q + 2\beta^{-1}\log(n/\epsilon)\Big)^{2/q} \Big\}$$

Consider

$$\widetilde{\gamma}_+ = 4\gamma_-(1+\delta)^2$$

$$\times \max\Big\{ \Big(d^p + \eta^p + 2\alpha^{-1}\log(n/\epsilon)\Big)^{2/p}, \Big(1 + \gamma^q + 2\beta^{-1}\log(n/\epsilon)\Big)^{2/q} \Big\},$$

With probability at least $1 - 4\epsilon$, for any $(\theta,\gamma) \in \mathbb{R}^{d+1}$,

$$-\delta - \widetilde{\gamma}_+ \le \frac{N(\theta,\gamma)}{\overline{N}(\theta,\gamma)} - 1 \le \frac{\delta + \gamma_-}{1 - \gamma_-}.$$

Estimation of parameters

Polynomial moment assumptions. Define $G_d = \mathbb{E}(XX^\top)$, so that $\theta^\top G_d \theta = \begin{pmatrix}\theta\\0\end{pmatrix}^\top G \begin{pmatrix}\theta\\0\end{pmatrix}$ and consider

$$\breve{\gamma}_+ = 8\gamma_-(1+\delta)^3 \max\Bigg\{ \mathbb{E}(\|G_d^{-1/2}X\|^6) + \left(\frac{\mathbb{E}(\|G_d^{-1/2}X\|^{12})}{n\epsilon}\right)^{1/2}, \\ \frac{\mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^6\big]}{\mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^2\big]^3} + \left(\frac{\mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^{12}\big]}{n\epsilon\mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^2\big]^6}\right)^{1/2}\Bigg\}$$

With probability at least $1-4\epsilon$, for any $(\theta,\gamma) \in \mathbb{R}^d \times \mathbb{R}$,

$$-\frac{\delta+\breve{\gamma}_+}{1+\breve{\gamma}_+} \le \frac{N(\theta,\gamma)}{\overline{N}(\theta,\gamma)} - 1 \le \frac{\delta+\gamma_-}{1-\gamma_-}.$$

Dimension free bounds (*Ilaria Giulini*)

With probability $1-2\epsilon$, for any $\theta \in \mathbb{S}_d$, and some robust estimator $\widehat{N}$,

$$\left| \frac{\widehat{N}(\theta) \vee \sigma}{N(\theta) \vee \sigma} - 1 \right| \leq \frac{\mu\big(N(\theta) \vee \sigma\big)}{1 - 4\mu\big(N(\theta) \vee \sigma\big)},$$

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

$$\mu(t) = \sqrt{\frac{2.13(\kappa-1)}{n}\left[\log(\epsilon^{-1}) + 4.64 + \frac{0.73 \times \mathbf{Tr}(G)}{t}\right]} + \sqrt{\frac{118\kappa\,\mathbf{Tr}(G)}{nt}},$$

and σ is such that $8\mu(\sigma) \leq 1$. Let us recall that $\mathbf{Tr}(G) = \int \|x\|^2 \,\mathrm{dP}(x) = \sum_{i=1}^{d} N(\theta_i)$, for any orthonormal basis $(\theta_i, 1 \leq i \leq n)$.

Proof of main result

Replacing $\mathbb{R}^d$ by $\mathbf{Im}(G)$, we can assume that $\mathbf{Ker}(G) = \{0\}$, without loss of generality. For any $\lambda > 0$, any $x, \theta \in \mathbb{R}^d$,

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

where $\pi_\theta = \mathcal{N}(\theta, \beta^{-1} G^{-1})$ is a Gaussian perturbation of the parameter θ and

$$c = \frac{15}{8\log(2)(\sqrt{2}-1)} \exp\bigg(\frac{1+2\sqrt{2}}{2}\bigg) \leq 44.3.$$

Proof of main result

Let $\nu \in \mathscr{M}_+^1(\Theta)$ be a prior probability measure on the parameter space Θ. Under suitable assumptions, with probability at least $1-\epsilon$ for any posterior probability measure $\rho \in \mathscr{M}_+^1(\Theta)$ such that $\mathscr{K}(\rho,\nu) < +\infty$,

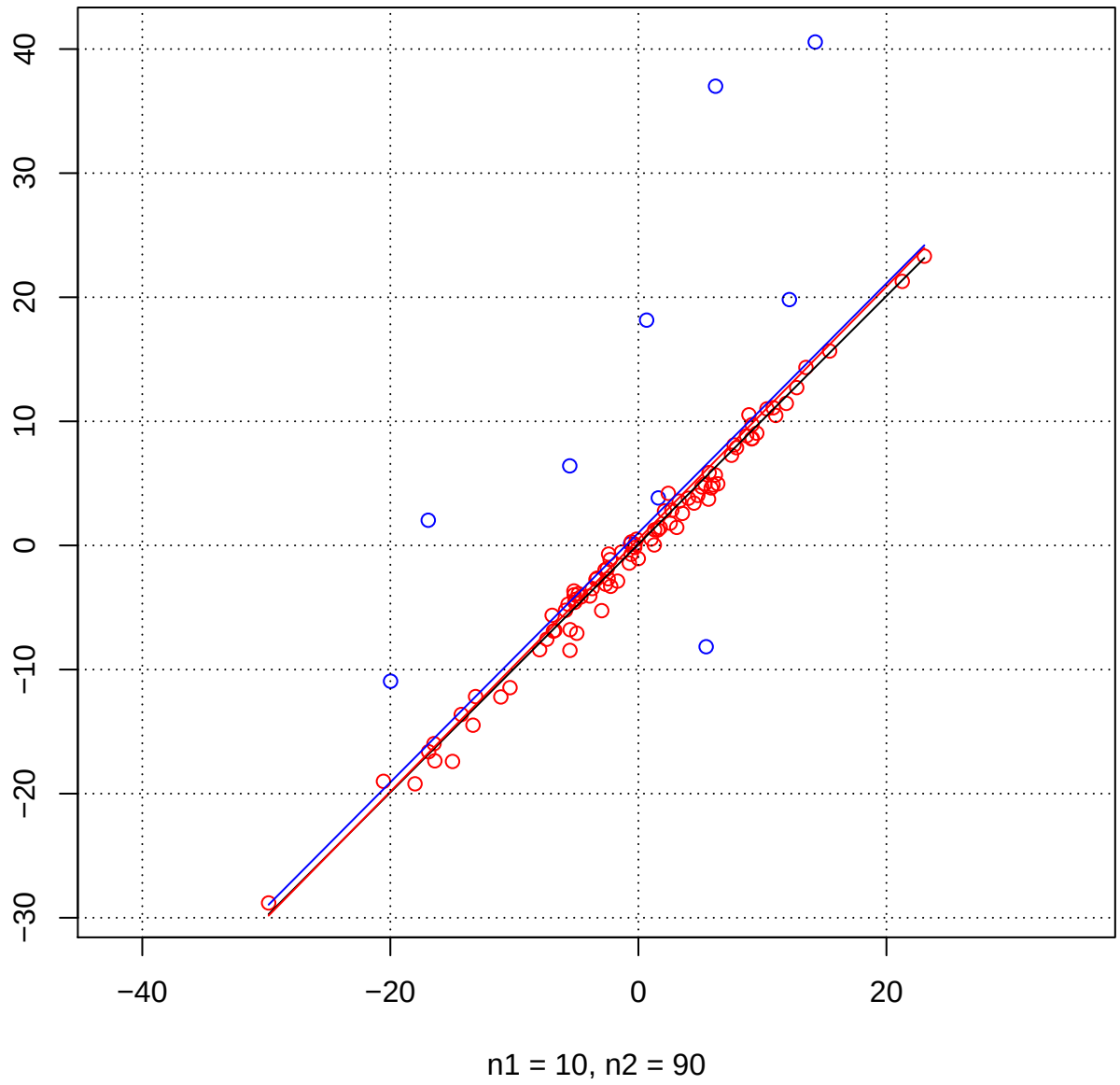
$$\iint \log\bigl[1+f(x,\theta)\bigr]\,\mathrm{d}\rho(\theta)\,\mathrm{d}\overline{\mathrm{P}}(x) \leq \iint f(x,\theta)\,\mathrm{dP}(x)\,\mathrm{d}\rho(\theta) + \frac{\mathscr{K}(\rho,\nu)-\log(\epsilon)}{n}.$$

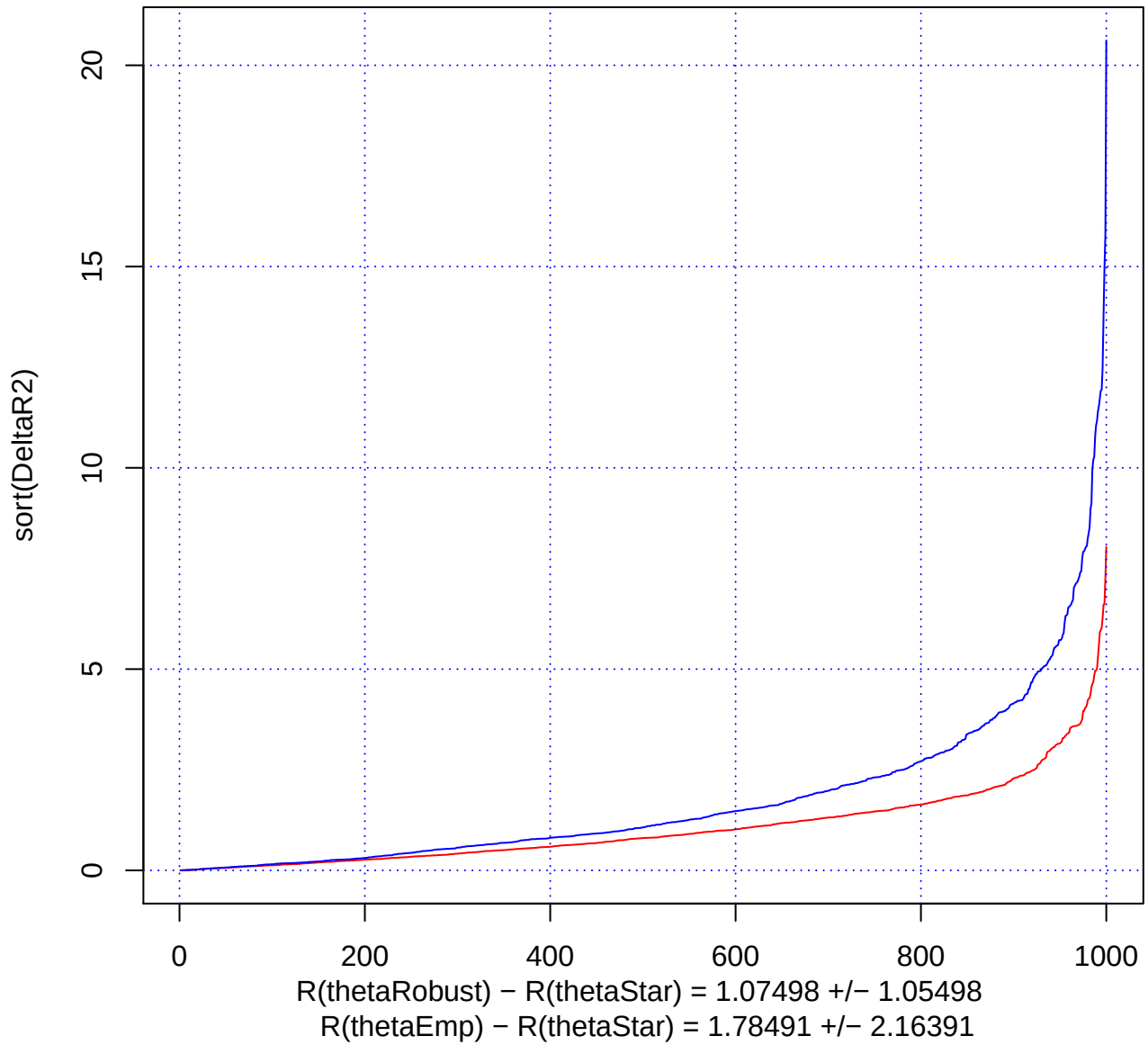
Proof of main result

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

$$r_\lambda(\theta) \leq \xi\big[N(\theta)-1\big]^2 + (1+\mu)\big[N(\theta)-1\big] + \mu,$$
$$r_\lambda(\theta) \geq -\xi\big[N(\theta)-1\big]^2 + (1-\mu)\big[N(\theta)-1\big] - \mu,$$

where $\xi = \dfrac{\kappa\lambda}{2}$. We end the proof by showing that when $N(\theta) > 0$, with probability $1-2\epsilon$, $r_\lambda(\widehat{\alpha}\theta) = 0$ and solving the resulting inequalities in $N(\widehat{\alpha}\theta) = \dfrac{N(\theta)}{\widehat{N}(\theta)}$.


40
30
20
10
0
−10
−20
−30
−40
−20
0
20
n1 = 10, n2 = 90


sort(DeltaR2)
0
5
10
15
20
0
200
400
600
800
1000
R(thetaRobust) − R(thetaStar) = 1.07498 +/− 1.05498
R(thetaEmp) − R(thetaStar) = 1.78491 +/− 2.16391