PAC-Bayes bounds using Gaussian posterior distributions

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

Computer Vision and Machine Learning Group,
Institute for Science and Technology Austria,

Klosterneuburg, July 26, 2013

Supervised classification

We observe independent copies (X_i, Y_i), $1 \leq i \leq n$ of a couple of random variables (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \{-1, +1\}$. We would like to relate the *empirical error rate* of these labeled data

$$\int \mathbb{1}\Big(\langle x, \theta \rangle y \leq 0\Big) \mathrm{d}\overline{\mathbb{P}}(x, y)$$

where $\overline{\mathbb{P}} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \delta_{(X_i, Y_i)}$ is the empirical measure, with its expectation

$$\int \mathbb{1}\Big(\langle x, \theta \rangle y \leq 0\Big) \mathrm{d}\mathbb{P}(x, y),$$

where $\mathbb{P}$ is the joint distribution of (X, Y).

PAC-Bayes margin error bounds

after Langford, Shawe-Taylor, and McAllester

Let us introduce the *Gaussian perturbation of the parameter* θ

$$\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\theta'}(\theta') = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta\|\theta'-\theta\|^2}{2}\right), \qquad \theta \in \mathbb{R}^d.$$

It is such that

$$\mathscr{K}(\mu_\theta, \mu_0) = \int \log(\mathrm{d}\mu_\theta/\mathrm{d}\mu_0)\mathrm{d}\mu_\theta = \frac{\beta}{2}\|\theta\|^2.$$

The perturbation of the loss function can be computed explicitly :

$$\int \mathbb{1}\Big(\langle x, \theta' \rangle y \leq 0\Big) \mathrm{d}\mu_\theta(\theta') = \varphi\bigg(\frac{\sqrt{\beta}\langle x, \theta \rangle y}{\|x\|}\bigg),$$

where

$$\varphi(a) = \int_a^{+\infty} (2\pi)^{-1/2} \exp(-t^2/2) \mathrm{d}t.$$

More interestingly, the error function itself is upper-bounded by the perturbation of the *error function with margin*

$$\begin{aligned}
&\int \mathbb{1}\Big(\|x\|^{-1}\langle x,\theta'\rangle y \le 1\Big)\mathrm{d}\mu_\theta(\theta') \\
&\quad = \varphi\Big[\sqrt{\beta}\Big(\|x\|^{-1}\langle x,\theta\rangle y - 1\Big)\Big] \ge \varphi(-\sqrt{\beta})\mathbb{1}\Big(\langle x,\theta\rangle y \le 0\Big) \\
&\qquad\qquad = \Big[1-\varphi(\sqrt{\beta})\Big]\mathbb{1}\Big(\langle x,\theta\rangle y \le 0\Big),
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{P}\Big(\langle X,\theta\rangle\, Y \le 0\Big) &\le \Big[1-\varphi(\sqrt{\beta})\Big]^{-1} \\
&\quad \times \mathbb{E}\left(\int \underbrace{\mathbb{1}\Big(\|x\|^{-1}\langle X,\theta'\rangle\, Y \le 1\Big)}_{\text{error with margin}} \underbrace{\mathrm{d}\mu_\theta(\theta')}_{\text{perturbation}}\right).
\end{aligned}$$

This opens the possibility to use the following PAC-Bayes inequality :

$$\exp\left(\int h(\theta')\mu_\theta(\theta') - \mathscr{K}(\mu_\theta, \mu_0)\right) \leq \int \exp\big[h(\theta')\big]\mathrm{d}\mu_0(\theta'),$$

according to it

$$\mathbb{E}\Bigg(\exp\Bigg\{\sup_{\theta\in\mathbb{R}^d} n\lambda \int \Bigg[\Phi_\lambda\Big(\mathbb{P}\Big(\|X\|^{-1}\langle X,\theta'\rangle\, Y \leq 1\Big)\Big) - \int \mathbb{1}\Big(\|x\|^{-1}\langle x,\theta'\rangle y \leq 1\Big)\mathrm{d}\overline{\mathbb{P}}(x,y)\Bigg]\mathrm{d}\mu_\theta(\theta') - \mathscr{K}(\mu_\theta,\mu_0)\Bigg\}\Bigg) \leq \mathbb{E}\Bigg(\int \exp\Bigg[\Phi_\lambda\Big(\mathbb{P}\Big(\|X\|^{-1}\langle X,\theta'\rangle\, Y \leq 1\Big)\Big) - \int \mathbb{1}\Big(\|x\|^{-1}\langle x,\theta'\rangle y \leq 1\Big)\mathrm{d}\overline{\mathbb{P}}(x,y)\Bigg]\mathrm{d}\mu_0(\theta')\Bigg) \leq 1,$$

where $\Phi_\lambda(p) = -\lambda^{-1}\log\big[1-p+p\exp(-\lambda)\big]$.

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

$$\mathbb{P}\Big[\mathbb{1}\big(\langle X,\theta\rangle\, Y \le 0\big)\Big] \le \big[1-\varphi(\sqrt{\beta})\big]^{-1}$$

$$\times \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\Bigg\{\int \varphi\big[\sqrt{\beta}\big(\|x\|^{-1}\langle x,\theta\rangle y - 1\big)\big]\,\mathrm{d}\overline{\mathbb{P}}(x,y)$$

$$+ \frac{\beta\|\theta\|^2/2 + \log\big(|\Lambda|/\epsilon\big)}{n\lambda}\Bigg\}$$

$$\le \big[1-\varphi(\sqrt{\beta})\big]^{-1} B\left(\int \varphi\big[\sqrt{\beta}\big(\|x\|^{-1}\langle x,\theta\rangle y - 1\big)\big]\,\mathrm{d}\overline{\mathbb{P}}(x,y), \frac{\beta\|\theta\|^2}{2}, \epsilon\right),$$

where $B(q,e,\epsilon) = \sqrt{\dfrac{2q(1-q)\big\{e+\log\big[\log(n)^2/\epsilon\big]\big\}}{n}}\cosh\big[\log(n)^{-1}\big]$

$$+ \frac{2\big\{e+\log\big[\log(n)^2/\epsilon\big]\big\}}{n}\cosh\big[\log(n)^{-1}\big]^2, \text{ for a suitable choice of } \Lambda.$$

As a consequence, for any radius $R > 0$, with probability at least $1 - \epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\mathbb{P}\big[\mathbb{1}\big(\langle X, \theta\rangle\, Y \leq 0\big)\big] \leq \big[1 - \varphi(\sqrt{\beta})\big]^{-1}$$
$$\times \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}\left\{ \int \Big(2 - \|x\|_R^{-1} \langle \theta, x\rangle y\Big)_+ \mathrm{d}\overline{\mathbb{P}}(x,y) + \frac{\beta \|\theta\|^2}{2n\lambda} \right.$$
$$\left. + \frac{\log(|\Lambda|/\epsilon)}{n\lambda} + \varphi(\sqrt{\beta}) \right\}, \text{ where } \|x\|_R = \max\{R, \|x\|\}.$$

Assume that $\mathbb{P}\big(\|X - c\| \leq R\big) = 1$, for some $c \in \mathbb{R}^d$ and $R \in \mathbb{R}_+$. With probability at least $1 - \epsilon$, for any $\theta \in \mathbb{R}^d$, any $\gamma \in \mathbb{R}$ such that $\min_{i=1,\ldots,n} \langle \theta, X_i \rangle \leq \gamma \leq \max_{i=1,\ldots,n} \langle \theta, X_i \rangle$,

$$\mathbb{P}\big[\big(\langle \theta, X \rangle - \gamma\big)\, Y \leq 0\big] \leq \inf_{\beta \in \Xi} \big[1 - \varphi(\sqrt{\beta})\big]^{-1}$$

$$\inf_{\lambda \in \Lambda} \Phi_\lambda^{-1} \Bigg[\int \big[1 - y\big(\langle \theta, x \rangle - \gamma\big)\big]_+ \, \mathrm{d}\overline{\mathbb{P}}(x, y) + \frac{4\beta R^2 \|\theta\|^2}{n\lambda} + \frac{\log\big(|\Xi||\Lambda|/\epsilon\big)}{n\lambda} + \varphi(\sqrt{\beta}) \Bigg].$$

Gram matrix estimate

Let $X_i \in \mathbb{R}^d$, $1 \leq i \leq n$ be vector valued i.i.d. random variables with common distribution $\mathbb{P} \in \mathscr{M}_+^1(\mathbb{R}^d)$.
Question : Estimate the quadratic form

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

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

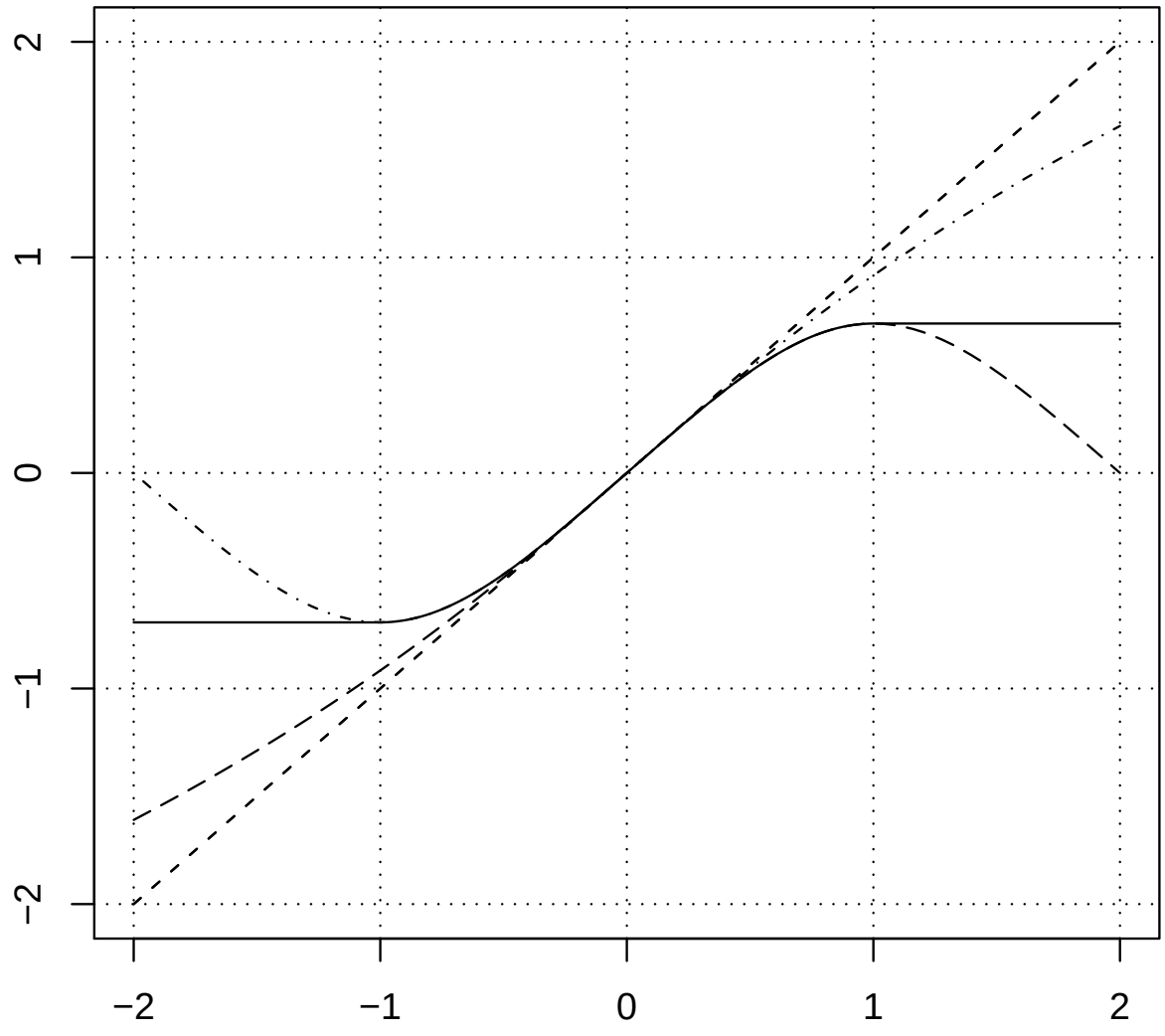
$$G \stackrel{\text{def}}{=} \int x \, x^\top \, \mathrm{d}\mathbb{P}(x),$$

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

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

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

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

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

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

This is justified by the fact that

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

Proposition

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

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

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

Obtaining a true quadratic form

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

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

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

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

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

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

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

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

Proof

$\widehat{G}$ minimizes

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

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

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

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

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

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 $1 - 2\epsilon$, $\left|\dfrac{N(\theta,\gamma)}{\widehat{N}(\theta,\gamma)} - 1\right| \leq \eta$. 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 $\widehat{R}(\theta) = \widehat{N}(\theta, 1)$.

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

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

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

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

Proof

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

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

Proposition

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

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

Indeed,

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

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

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

we get

$$\psi(\langle\theta, x\rangle^2 - \lambda \leq \int \psi\Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big)\,\mathrm{d}\pi_\theta(\theta') + \min\Big\{\log(4), \frac{4\|x\|^2\langle\theta, x\rangle^2}{\beta} + \frac{2\|x\|^4}{\beta^2}\Big\}$$

Lemma

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

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

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

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

$$\min\{a, y + z\} \leq \min\{a, y\} + \min\{a, z\}$$

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

Accordingly

$$\psi(\langle\theta,x\rangle^2-\lambda)\le\int\psi\bigg(\langle\theta',x\rangle^2-\frac{\|x\|^2}{\beta}-\lambda\bigg)\,\mathrm{d}\pi_\theta(\theta')$$
$$+\int\min\bigg\{4\log(2),\frac{8\|x\|^2\langle\theta',x\rangle^2}{\beta}+\frac{10\|x\|^4}{\beta^2}\bigg\}\,\mathrm{d}\pi_\theta(\theta').$$

We will now use

Lemma

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

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

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

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

The proof is based on the fact that

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

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

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

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

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

$$
\begin{aligned}
s_4 &= \left(\int \|x\|^4 \, \mathrm{d}\mathbb{P}(x) \right)^{1/4}, \\
\kappa &= \sup \left\{ \int \langle \theta, x \rangle^4 \, \mathrm{d}\mathbb{P}(x), \theta \in \mathbb{R}^d \text{ s. t. } \int \langle \theta, x \rangle^2 \, \mathrm{d}\mathbb{P}(x) = 1 \right\}, \\
\xi &= \frac{\kappa \lambda}{2}, \\
\gamma &= \lambda(\kappa - 1) + \frac{2}{\beta}(1 + 4c) s_4^2 \sqrt{\kappa}, \\
\eta &= \frac{\lambda}{2}(\kappa - 1) + \frac{2}{\beta}(1 + 4c) s_4^2 \sqrt{\kappa} + \frac{(1 + 18c) s_4^4}{\beta^2 \lambda} - \frac{\log\left[\nu(\lambda, \beta)\epsilon\right]}{n\lambda}, \\
\delta &= \frac{\beta}{2n\lambda}.
\end{aligned}
$$

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

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

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

Proposition

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

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

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

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

$$\widehat{N}(\theta) - N(\theta) \leq \frac{\Phi_+^{-1}\left(\lambda/\widehat{\alpha}^2\right) - N(\theta)}{2} \leq \frac{\Phi_+^{-1} \circ \Phi_-^{-1}\left[N(\theta)\right] - N(\theta)}{2}.$$

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

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

From this we can deduce

Proposition

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

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

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