PAC-BAYESIAN LEARNING BOUNDS

OLIVIER CATONI

September 15, 2011

SUPERVISED LEARNING THROUGH STATISTICAL INFERENCE

In this setting, we are given a large set of input-output data $w_1, \dots, w_N \in \mathcal{W}$. Our goal is to optimize some processing of these data. The changes we are ready to make to this processing is described by a set of tunable parameters $\theta \in \Theta$. The quality of the processing is described by some loss function $L(w, \theta) \in \mathbb{R}$. Our goal is to minimize

$$\frac{1}{N} \sum_{i=1}^{N} L(w_i, \theta)$$

with respect to $\theta \in \Theta$.

As we assume that N is potentially very large, we will draw at random some independent identically distributed sample $(W_1, \dots, W_n)$ according to the uniform distribution on $\{w_1, \dots, w_N\}$, where the size n of the statistical sample corresponds to the amount of computations we are ready to make. This sample will be used to choose the parameters.

Although in this scenario the marginal sample distribution is the atomic measure

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{w_i},$$

we will in the following deal with arbitrary i.i.d. sample distributions $\mathbb{P} \in \mathcal{M}_+^1(\mathcal{W})$, where $\mathcal{W}$ is assumed to be some arbitrary measurable space.

In this slightly more general setting, our goal is now to estimate

$$\theta(\mathbb{P}) \in \arg\min_{\theta \in \Theta} \int L(W, \theta) \, \mathrm{d}\mathbb{P},$$

where we assume that this is meaningful. For instance, we may consider some probability measure $\pi \in \mathcal{M}_+^1(\Theta)$, where Θ is a measurable space equipped with some σ-algebra and assume that $(w, \theta) \mapsto L(w, \theta) : \mathcal{W} \times \Theta \to \mathbb{R}$ belongs to $\mathbf{L}_1(\mathcal{W} \times \Theta, \mathbb{P} \otimes \pi)$ and is such that $\inf_{\theta \in \Theta} \int L(w, \theta) d\mathbb{P}(w) > -\infty$. The fact that

CNRS – UMR 8553, Département de Mathématiques et Applications, Ecole Normale Supérieure, 45, rue d'Ulm, F75230 Paris cedex 05, and INRIA Paris-Rocquencourt – CLASSIC team.

the minimum is reached and that $\theta(\mathbb{P})$ is uniquely defined is not crucial, since we will only be able to work out some approximation of it.

From a technical perspective, we will look for some estimator $\widehat{\theta} : \mathcal{W}^n \to \Theta$ such that $\widehat{\theta}(W_1, \dots, W_n)$ is an approximate minimizer in the sens that

$$\ell\big[\widehat{\theta}(W_1, \dots, W_n)\big] \stackrel{\text{def}}{=} \int L\big[w, \widehat{\theta}(W_1, \dots, W_n)\big] \, \mathrm{d}\mathbb{P}(w) - \inf_{\theta \in \Theta} \int L(w, \theta) \, \mathrm{d}\mathbb{P}(w)$$

is " as small as possible " (in some sense to be made more precise in the following).

EXAMPLES.

1. Supervised binary classification : $w = (x, y) \in \mathcal{W} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{-1, +1\}$, $\Theta \subset \mathcal{Y}^{\mathcal{X}}$ and $L(w, \theta) = \mathbb{1}\big[y \neq \theta(x)\big]$ is the classification error;
2. Least square linear regression : $w = (x, y) \in \mathcal{W} = \mathbb{R}^d \times \mathbb{R}$, $\Theta \subset \mathbb{R}^d$, and $L(w, \theta) = \big(y - \langle \theta, x \rangle\big)^2$ is the quadratic risk;
3. Density estimation : $\mathcal{W}$ is arbitrary, $\Theta \subset \big\{\theta \in \mathcal{M}_+^1\big(\mathcal{W}\big) \, ; \, \theta \ll \pi\big\}$, where $\pi \in \mathcal{M}_+^1(\mathcal{W})$ is some reference measure, and $L(w, \theta) = -\log\left[\dfrac{\mathrm{d}\theta}{\mathrm{d}\pi}(w)\right]$. In the case when $\mathbb{P} \ll \pi$,
$$\ell\big(\widehat{\theta}\big) = \mathcal{K}\big(\mathbb{P}, \widehat{\theta}\big),$$
where the Kullback Leibler divergence, also called relative entropy, is defined as
$$\mathcal{K}\big(\mathbb{P}, \mathbb{Q}\big) = \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P}, & \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{otherwise}, \end{cases}$$
for any probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_+^1(\mathcal{W})$.

A FEW NOTATIONS. Let $\overline{\mathbb{P}}$ be the empirical measure, defined as

$$\overline{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{W_i}.$$

We will use the following short notation of integrals :

$$f(\mathbb{P}, \rho, \pi) = \int f(w, \theta, \theta') \, \mathrm{d}\mathbb{P}(w) \, \mathrm{d}\rho(\theta) \, \mathrm{d}\pi(\theta'), \tag{1}$$

so that for instance

$$L(\mathbb{P}, \rho) = \int L(w, \theta) \, \mathrm{d}\mathbb{P}(w) \, \mathrm{d}\rho(\theta).$$

1. THE PAC-BAYES APPROACH PART I : CLASSIFICATION

In this section, we will derive margin bounds for linear classification in high dimension and its applications to *kernel methods*. PAC-Bayes theory was first developed in the framework of supervised classification (see [11, 12, 13, 14, 10]). We start with this simpler setting, and will show in the next section how to deal with more general loss functions.

1.1. A PAC-BAYES BOUND FOR 0-1 LOSS FUNCTIONS.

Let us assume that $L(w,\theta) \in \{0,1\}$. Given some parameter $\lambda \in \mathbb{R}$, let us consider the (normalized) log-Laplace transform of the Bernoulli distribution :

$$\Phi_\lambda(p) \stackrel{\text{def}}{=} -\frac{1}{\lambda}\log\bigl[1-p+p\exp(-\lambda)\bigr].$$

Let us also consider the Kullback-Leibler divergence of Bernoulli distributions

$$K(q,p) \stackrel{\text{def}}{=} q\log\left(\frac{q}{p}\right)+(1-q)\log\left(\frac{1-q}{1-p}\right).$$

Let us recall first Chernoff's bound.

PROPOSITION 1.1 *For any fixed value of the parameter $\theta \in \Theta$, the identity*

$$\int \exp\bigl[-\lambda L(\overline{\mathbb{P}},\theta)\bigr]\,\mathrm{d}\mathbb{P}^{\otimes n} = \exp\Bigl\{-\lambda\Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr]\Bigr\}$$

shows that with probability at least $1-\epsilon$,

$$\begin{aligned} &L(\mathbb{P},\theta) \le B_+\bigl[L(\overline{\mathbb{P}},\theta), \log(\epsilon^{-1})/n\bigr],\\ \textit{where}\quad &B_+(q,\delta) = \inf_{\lambda\in\mathbb{R}_+}\Phi_\lambda^{-1}\left(q+\frac{\delta}{\lambda}\right)\\ &\qquad = \sup\Bigl\{p\in[0,1]\,:\,K(q,p)\le\delta\Bigr\}, \qquad q\in[0,1],\ \delta\in\mathbb{R}_+. \end{aligned}$$

Moreover

$$-\delta q \le B_+(q,\delta) - q - \sqrt{2\delta q(1-q)} \le 2\delta(1-q).$$

In the same way, the identity

$$\int \exp\bigl[\lambda L(\overline{\mathbb{P}},\theta)\bigr]d\mathbb{P}^{\otimes n} = \exp\Bigl\{\lambda\Phi_{-\lambda}\bigl[L(\mathbb{P},\theta)\bigr]\Bigr\}$$

shows that with probability at least $1-\epsilon$

$$L(\overline{\mathbb{P}},\theta) \le B_-\bigl[L(\mathbb{P},\theta), \log(\epsilon^{-1})/n\bigr],$$

$$\textit{where} \quad B_-(q,\delta) = \inf_{\lambda\in\mathbb{R}_+} \Phi_{-\lambda}(q) + \frac{\delta}{\lambda}$$
$$= \sup\Big\{p\in[0,1] \,:\, K(p,q)\leq\delta\Big\}, \qquad q\in[0,1],\ \delta\in\mathbb{R}_+,$$

and

$$-\delta q \leq B_-(q,\delta) - q - \sqrt{2\delta q(1-q)} \leq 2\delta(1-q).$$

Let us mention here some important identity.

PROPOSITION 1.2 *For any probability measures π and ρ on some measurable space, such that $\mathcal{K}(\rho,\pi) < \infty$, and any bounded measurable function h, let us define the transformed probability measure $\pi_{\exp(h)} \ll \pi$ by its density*

$$\frac{\mathrm{d}\pi_{\exp(h)}}{\mathrm{d}\pi} = \frac{\exp(h)}{Z},$$

where $Z = \int \exp(h)\,\mathrm{d}\pi$. Let us moreover define

$$\mathbf{Var}\big(h\,\mathrm{d}\pi\big) = \int\big(h - \textstyle\int h\,\mathrm{d}\pi\big)^2\,\mathrm{d}\pi.$$

The expectations with respect to ρ and π of h and the log*-Laplace transform of h are linked by the identities*

$$\textstyle\int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi) + \mathcal{K}(\rho,\pi_{\exp(h)}) = \log\big[\int\exp(h)\,\mathrm{d}\pi\big] \tag{2}$$
$$= \textstyle\int h\,\mathrm{d}\pi + \int_0^1(1-\alpha)\,\mathbf{Var}\big[h\,\mathrm{d}\pi_{\exp(\alpha h)}\big]\,\mathrm{d}\alpha. \tag{3}$$

PROOF. The first identity is a straightforward consequence of the definitions of $\pi_{\exp(h)}$ and of the Kullback-Leibler divergence function. The second one is the Taylor expansion of order one with integral remainder of the function

$$f(\alpha) = \log\big[\textstyle\int\exp(\alpha h)\,\mathrm{d}\pi\big],$$

which says that $f(1) = f(0) + f'(0) + \int_0^1(1-\alpha)f''(\alpha)\,\mathrm{d}\alpha$. □

Exercise 1 *Prove that $f\in\mathcal{C}^\infty$. Hint : write*

$$\exp(\alpha h) = 1 + \int_0^{+\infty} \mathbb{1}(\gamma\leq\alpha)h\exp(\gamma h)\,\mathrm{d}\gamma$$

and use Fubini's theorem to show that $\alpha\mapsto\int\exp(\alpha h)\,\mathrm{d}\pi$ belongs to $\mathcal{C}^1$.

Let us come now to the proof of Proposition 1.1 (page 3). Chernoff's inequality reads

$$\Phi_\lambda\big[L(\mathbb{P},\theta)\big] - \frac{\log(\epsilon^{-1})}{n\lambda} \leq L(\overline{\mathbb{P}},\theta),$$

where the inequality holds with probability at least $1-\epsilon$. Since the left-hand side is non-random, it can be optimized in λ, giving $L(\mathbb{P},\theta) \leq B_+\big[L(\overline{\mathbb{P}},\theta), \log(\epsilon^{-1})/n\big]$.

Exercise 2 *Prove more precisely that*

$$\arg\max_{\lambda\in\mathbb{R}_+} \Phi_\lambda\big[L(\mathbb{P},\theta)\big] - \frac{\log(\epsilon^{-1})}{n\lambda}$$

is reached when $L(\mathbb{P},\theta) < 1$. *When* $L(\mathbb{P},\theta) = 1$, $L(\overline{\mathbb{P}},\theta) = 1$ *almost surely, and the result is trivial, since* $B_+(1,\delta) = 1$ *for any* $\delta \in \mathbb{R}_+$.

Since

$$\lim_{\lambda\to+\infty} \Phi_\lambda^{-1}\left(q + \frac{\delta}{\lambda}\right) = \lim_{\lambda\to+\infty} \frac{1-\exp(-\lambda q - \delta)}{1-\exp(-\lambda)} \leq 1,$$

$B_+(q,\delta) \leq 1$.

Applying equation (2, page 4) to Bernoulli distributions gives

$$\lambda\Phi_\lambda(p) = \lambda q + K(q,p) - K(q,p_\lambda)$$

where

$$p_\lambda = \frac{p}{p + (1-p)\exp(\lambda)}.$$

This shows that

$$\begin{aligned}
B_+(q,\delta) &= \sup\Big\{p\in[0,1] \,:\, \Phi_\lambda(p) \leq q + \frac{\delta}{\lambda},\ \lambda\in\mathbb{R}_+\Big\} \\
&= \sup\Big\{p\in[q,1[\,:\, K(q,p) \leq \delta + K(q,p_\lambda), \lambda\in\mathbb{R}_+\Big\} \\
&= \sup\Big\{p\in[q,1[\,:\, K(q,p)\leq\delta\Big\} \\
&= \sup\Big\{p\in[0,1] \,:\, K(q,p)\leq\delta\Big\},
\end{aligned}$$

because when $q \leq p < 1$ then $\lambda = \log\left(\frac{q^{-1}-1}{p^{-1}-1}\right) \in \mathbb{R}_+$, $q = p_\lambda$ and therefore $K(q,p_\lambda) = 0$.

Let us remark now that $\dfrac{\partial^2}{\partial x^2} K(x,p) = x^{-1}(1-x)^{-1}$. Thus if $p \geq q \geq 1/2$, then

$$K(q,p) \geq \frac{(p-q)^2}{2q(1-q)},$$

so that if $K(q,p) \leq \delta$, then

$$p \leq q + \sqrt{2\delta q(1-q)}.$$

Now if $q \leq 1/2$ and $p \geq q$ then

$$K(q,p) \geq \left\{\begin{array}{ll} \dfrac{(p-q)^2}{2p(1-p)}, & p\leq 1/2 \\ 2(p-q)^2, & p\geq 1/2 \end{array}\right\} \geq \frac{(p-q)^2}{2p(1-q)},$$

so that if $K(q,p) \leq \delta$, then

$$(p-q)^2 \leq 2\delta p(1-q),$$

implying that

$$p - q \leq \delta(1-q) + \sqrt{2\delta q(1-q) + \delta^2(1-q)^2} \leq \sqrt{2\delta q(1-q)} + 2\delta(1-q).$$

On the other hand,

$$K(q,p) \leq \frac{(p-q)^2}{2\min\{q(1-q), p(1-p)\}} \leq \frac{(p-q)^2}{2q(1-p)},$$

thus when $K(q,p) = \delta$ with $p > q$, then

$$(p-q)^2 \geq 2\delta q(1-p),$$

implying that

$$p - q \geq -\delta q + \sqrt{2\delta q(1-q) + \delta^2 q^2} \geq \sqrt{2\delta q(1-q)} - \delta q.$$

Exercise 3 *The second part of Proposition 1.1 (page 3) is proved in the same way and left as an exercise.*

We are now going to make Proposition 1.1 uniform with respect to θ. The PAC-Bayes approach to this is to randomize θ, so we will consider now joint distributions on $(W_1, \ldots, W_n, \theta)$, where the distribution of $(W_1, \ldots, W_n)$ is still $\mathbb{P}^{\otimes n}$ and the conditional distribution of θ given the sample is given by some transition probability kernel $\rho : \mathcal{W}^n \to \mathcal{M}_+^1(\Theta)$, called in this context a posterior distribution*. This posterior distribution ρ will be compared with a prior (meaning non-random) probability measure $\pi \in \mathcal{M}_+^1(\Theta)$.

PROPOSITION 1.3 *Let us introduce the notation*

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

For any prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ and any $\lambda \in \mathbb{R}_+$,

$$\int \exp\left[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} n\lambda\Big\{\Phi_\lambda\big[L(\mathbb{P},\rho)\big] - L(\overline{\mathbb{P}},\rho)\Big\} - \mathcal{K}(\rho,\pi)\right] \mathrm{d}\mathbb{P}^{\otimes n} \leq 1, \quad (4)$$

*We will assume that ρ is a regular conditional probability kernel, meaning that for any measurable set A the map $(w_1, \ldots, w_n) \mapsto \rho(w_1, \ldots, w_n)(A)$ is assumed to be measurable. We will also assume that the σ-algebra we consider on Θ is generated by a countable family of subsets. See [6, page 50] for more details

and therefore for any finite set $\Lambda \subset \mathbb{R}_+$, with probability at least $1-\epsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,

$$L(\mathbb{P},\rho) \leq B_\Lambda\left(L(\overline{\mathbb{P}},\rho), \frac{\mathcal{K}(\rho,\pi)+\log\bigl(|\Lambda|/\epsilon\bigr)}{n}\right),$$

PROOF. The exponential moment inequality (4) is a consequence of equation (2, page 4), showing that

$$\exp\left\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)} n\lambda\int\Bigl\{\Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr]-L(\overline{\mathbb{P}},\theta)\Bigr\}\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\pi)\right\}$$
$$\leq \int \exp\left[n\lambda\Bigl\{\Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr]-L(\overline{\mathbb{P}},\theta)\Bigr\}\right]\mathrm{d}\pi(\theta),$$

and of the fact that Φ_λ is convex, showing that $\Phi_\lambda\bigl[L(\mathbb{P},\rho)\bigr] \leq \int \Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr]\,\mathrm{d}\rho(\theta)$. The deviation inequality follows as usual. □

We cannot take the infimum on $\lambda \in \mathbb{R}_+$ as in Proposition 1.1 (page 3), because we can no more cast our deviation inequality in such a way that λ appears on some non-random side of the inequality. Nevertheless, we can get a more explicit bound from some specific choice of the set Λ.

PROPOSITION 1.4 *Let us define the least increasing upper bound of the variance of a Bernoulli distribution of parameter $p \in [0,1]$ as*

$$\overline{v}(p) = \begin{cases} p(1-p), & p \leq 1/2, \\ 1/4, & \textit{otherwise.} \end{cases}$$

Let us choose some positive integer parameter m and let us put

$$t = \frac{1}{4}\log\left(\frac{n}{8\log\bigl[(m+1)/\epsilon\bigr]}\right).$$

With probability at least $1-\epsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,

$$L(\mathbb{P},\rho) \leq L(\overline{\mathbb{P}},\rho) + B_m\bigl[L(\overline{\mathbb{P}},\rho),\mathcal{K}(\rho,\pi),\epsilon\bigr],$$

where

$$B_m\bigl(q,e,\epsilon\bigr) = \max\left\{\sqrt{\frac{2\overline{v}(q)\bigl\{e+\log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n}}\cosh\bigl(t/m\bigr)\right.$$

$$
\begin{aligned}
&\qquad + \frac{2(1-q)\big\{e + \log\big[(m+1)/\epsilon\big]\big\}}{n} \cosh(t/m)^2, \\
&\qquad\qquad \frac{2\big\{e + \log\big[(m+1)/\epsilon\big]\big\}}{n} \Bigg\} \\
&\leq \sqrt{\frac{2\overline{v}(q)\big\{e + \log\big[(m+1)/\epsilon\big]\big\}}{n}} \cosh\big(t/m\big) \\
&\qquad\qquad + \frac{2\big\{e + \log\big[(m+1)/\epsilon\big]\big\}}{n} \cosh(t/m)^2.
\end{aligned}
$$

Moreover, as soon as $n \geq 5$*,*

$$
\begin{aligned}
B_{\lfloor \log(n)^2 \rfloor - 1}(q, e, \epsilon) \leq B(q, e, \epsilon) &\stackrel{\text{def}}{=} \sqrt{\frac{2\overline{v}(q)\big\{e + \log\big[\log(n)^2/\epsilon\big]\big\}}{n}} \cosh\big[\log(n)^{-1}\big] \\
&\quad + \frac{2\big\{e + \log\big[\log(n)^2/\epsilon\big]\big\}}{n} \cosh\big[\log(n)^{-1}\big]^2, \quad (5)
\end{aligned}
$$

so that with probability at least $1-\epsilon$*, for any* $\rho \in \mathcal{M}_+^1(\Theta)$*,*

$$
\begin{aligned}
&L(\mathbb{P}, \rho) \leq L(\overline{\mathbb{P}}, \rho) \\
&\quad + \sqrt{\frac{2\overline{v}\big[L(\overline{\mathbb{P}}, \rho)\big]\Big\{\mathcal{K}(\rho, \pi) + \log\big[\log(n)^2/\epsilon\big]\Big\}}{n}} \cosh\big[\log(n)^{-1}\big] \\
&\qquad + \frac{2\Big\{\mathcal{K}(\rho, \pi) + \log\big[\log(n)^2/\epsilon\big]\Big\}}{n} \cosh\big[\log(n)^{-1}\big]^2.
\end{aligned}
$$

PROOF. Let us put

$$
\begin{aligned}
q &= L(\overline{\mathbb{P}}, \rho), \\
\delta &= \frac{\mathcal{K}(\rho, \pi) + \log\big[(m+1)/\epsilon\big]}{n}, \\
\lambda_{\min} &= \sqrt{\frac{8\log\big[(m+1)/\epsilon\big]}{n}}, \\
\Lambda &= \Big\{\lambda_{\min}^{1-k/m}, k = 0, \ldots, m\Big\}, \\
p &= B_\Lambda(q, \delta) = \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}\bigg(q + \frac{\delta}{\lambda}\bigg), \\
\widehat{\lambda} &= \sqrt{\frac{2\delta}{\overline{v}(p)}}.
\end{aligned}
$$

According to equation (3, page 4) applied to Bernoulli distributions, for any $\lambda \in \Lambda$,

$$\Phi_\lambda(p) = p - \frac{1}{\lambda}\int_0^\lambda (\lambda - \alpha) p_\alpha (1 - p_\alpha)\,\mathrm{d}\alpha \leq q + \frac{\delta}{\lambda}.$$

As moreover $p_\alpha \leq p$,

$$p - q \leq \inf_{\lambda\in\Lambda} \frac{\lambda \overline{v}(p)}{2} + \frac{\delta}{\lambda} = \inf_{\lambda\in\Lambda} \sqrt{2\delta\overline{v}(p)}\cosh\left[\log\left(\frac{\widehat{\lambda}}{\lambda}\right)\right].$$

As $\overline{v}(p) \leq 1/4$ and $\delta \geq \dfrac{\log\big[(m+1)/\epsilon\big]}{n}$,

$$\sqrt{\frac{2\delta}{\overline{v}(p)}} = \widehat{\lambda} \geq \lambda_{\min} = \sqrt{\frac{8\log\big[(m+1)/\epsilon\big]}{n}}.$$

Therefore either $\lambda_{\min} \leq \widehat{\lambda} \leq 1$, or $\widehat{\lambda} > 1$. Let us consider these two cases separately.

If $\lambda_{\min} = \min\Lambda \leq \widehat{\lambda} \leq \max\Lambda = 1$, then $\log\big(\widehat{\lambda}\big)$ is at distance at most t/m from some $\log\big(\lambda\big)$ where $\lambda \in \Lambda$, because $\log(\Lambda)$ is a grid with constant steps of size $2t/m$. Thus

$$p - q \leq \sqrt{2\delta\overline{v}(p)}\cosh\big(t/m\big).$$

If moreover $q \leq 1/2$, then $\overline{v}(p) \leq p(1-q)$, so that we obtain a quadratic inequality in p, whose solution is less than

$$p \leq q + \sqrt{2\delta q(1-q)}\cosh\big(t/m\big) + 2\delta(1-q)\cosh\big(t/m\big)^2.$$

If on the contrary $q \geq 1/2$, then $\overline{v}(p) = \overline{v}(q) = 1/4$ and

$$p \leq q + \sqrt{2\delta\overline{v}(q)}\cosh\big(t/m\big),$$

so that in both cases

$$p - q \leq \sqrt{2\delta\overline{v}(q)}\cosh(t/m) + 2\delta(1-q)\cosh\big(t/m\big)^2. \tag{6}$$

Let us consider now the case when $\widehat{\lambda} > 1$. In this case

$$p - q \leq \sqrt{2\delta\overline{v}(p)}\;\widehat{\lambda} = 2\delta.$$

In conclusion, applying Proposition 1.3 (page 6) we see that with probability at least $1-\epsilon$, for any posterior distribution ρ,

$$L(\mathbb{P},\rho) \leq p \leq q + \max\Big\{2\delta,\, \sqrt{2\delta\overline{v}(q)}\cosh\big(t/m\big) + 2\delta(1-q)\cosh\big(t/m\big)^2\Big\},$$

which is precisely the statement to be proved.

In the special case when $m = \lfloor \log(n)^2 \rfloor - 1 \geq \log(n)^2 - 2$,

$$\frac{t}{m} \leq \frac{1}{4\big[\log(n)^2 - 2\big]} \log\left(\frac{n}{8 \log\big[\log(n)^2 - 1\big]} \right) \leq \log(n)^{-1}$$

as soon as the last inequality holds, that is as soon as $n \geq \exp(\sqrt{2}) \simeq 4.11$ to make $\log(n)^2 - 2$ positive and

$$3\log(n)^2 - 8 + \log(n) \log\Big\{ 8 \log\big[\log(n)^2 - 1\big] \Big\} \geq 0,$$

which holds true for any $n \geq 5$, as can be checked numerically. □

1.2. LINEAR CLASSIFICATION AND SUPPORT VECTOR MACHINES. We are going in this section to consider more specifically the case of linear binary classification. In this setting $\mathcal{W} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{-1, +1\}$, $w = (x, y)$, where $x \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, $\Theta = \mathbb{R}^d$, and

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

Although we will stick in this presentation to the case when $\mathcal{X}$ is a vector space of finite dimension, the results also apply to support vector machines, where the pattern space is some arbitrary space mapped to a Hilbert space $\mathcal{H}$ by some implicit mapping $\Psi : \mathcal{X} \to \mathcal{H}$, $\Theta = \mathcal{H}$ and $L(w, \theta) = \mathbb{1}\big(\langle \theta, \Psi(x) \rangle y \leq 0\big)$. It turns out that classification algorithms do not need to manipulate $\mathcal{H}$ itself, but only to compute scalar products of the form $k(x_1, x_2) = \langle \Psi(x_1), \Psi(x_2) \rangle$, defining a symmetric positive kernel k on the original pattern space $\mathcal{X}$. The converse is also true, any positive symmetric kernel k can be represented as a scalar product in some mapped Hilbert space (this is the Moore-Aronszajn theorem). Often used kernels on $\mathbb{R}^d$ are

$$\begin{aligned} k(x_1, x_2) &= \big(1 + \langle x_1, x_2 \rangle\big)^s, \text{ for which } \dim \mathcal{H} < \infty, \\ k(x_1, x_2) &= \exp\big(-\|x_1 - x_2\|^2\big), \text{ for which } \dim \mathcal{H} = +\infty. \end{aligned}$$

In the following, we will work in $\mathbb{R}^d$, which covers only the case when $\dim \mathcal{H} < \infty$, but extensions would be possible.

The loss function being homogeneous with respect to x and θ, we may replace x with $x/\|x\|$. Therefore, we will assume without loss of generality that $\|x\| = 1$.

Let us consider, after [10, 13] as prior probability measure π the centered Gaussian measure with covariance $\beta^{-1}\,\mathbf{Id}$, so that

$$\frac{\mathrm{d}\pi}{\mathrm{d}\theta}(\theta) = \left(\frac{\beta}{2\pi} \right)^{d/2} \exp\left(-\frac{\beta \|\theta\|^2}{2} \right).$$

Let us also consider the function

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp\bigl(-t^2/2\bigr)\,\mathrm{d}t, \qquad x \in \mathbb{R}$$
$$\leq \min\Bigl\{\frac{1}{x\sqrt{2\pi}}, \frac{1}{2}\Bigr\} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R}_+.$$

Let π_θ be the measure π shifted by θ, defined by the identity

$$\int h(\theta')\,\mathrm{d}\pi_\theta(\theta') = \int h(\theta+\theta')\,\mathrm{d}\pi(\theta').$$

In this case

$$\mathcal{K}(\pi_\theta, \pi) = \frac{\beta}{2}\|\theta\|^2,$$

and

$$L(w, \pi_\theta) = \varphi\bigl[\sqrt{\beta}\langle\theta, x\rangle y\bigr].$$

Thus the randomized loss function has an explicit expression : randomization replaces the indicator function of the negative real line by a smooth approximation. As we are eventually interested in $L(w,\theta)$, we will shift things a little bit, considering along with the classification error function L some *error with margin*

$$M(w, \theta) = \mathbb{1}\bigl[\langle\theta, x\rangle y \leq 1\bigr].$$

Unlike $L(w,\theta)$ which is independent of the norm of θ, the margin error $M(w,\theta)$ depends on $\|\theta\|$, counting a classification error each time x is at distance less than $\|\theta\|^{-1}$ from the boundary $\{x' : \langle\theta, x'\rangle = 0\}$.

Let us compute the randomized margin error

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

It satisfies the inequality

$$M(w, \pi_\theta) \geq \varphi(-\sqrt{\beta}\,)L(w,\theta) = \bigl[1 - \varphi\bigl(\sqrt{\beta}\,\bigr)\bigr]L(w,\theta).$$

Applying previous results and choosing some finite set of parameters $\Lambda \subset \mathbb{R}_+$, we obtain

PROPOSITION 1.5 *With probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,*

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

where

$$C_1(\theta) = \bigl[1 - \varphi\bigl(\sqrt{\beta}\,\bigr)\bigr]^{-1} B\left(M(\,\overline{\mathbb{P}}, \pi_\theta), \frac{\beta\|\theta\|^2}{2}, \epsilon\right),$$

the bound B being defined by equation (5, page 8).

We can now minimize this empirical upper-bound to define an estimator. Let us consider some estimator $\widehat{\theta}$ such that

$$C_1(\widehat{\theta}) \leq \inf_{\theta\in\mathbb{R}^d} C_1(\theta) + \zeta.$$

Then for any fixed parameter $\theta_\star$, $C_1(\theta) \leq C_1(\theta_\star) + \zeta$. On the other hand, with probability at least $1-\epsilon$

$$M(\overline{\mathbb{P}}, \pi_{\theta_\star}) \leq B_-\bigg(M(\mathbb{P}, \pi_{\theta_\star}), \frac{\log(\epsilon^{-1})}{n}\bigg).$$

Indeed

$$\int \exp\Big\{n\lambda\big[M(\overline{\mathbb{P}}, \pi_{\theta_\star}) - \Phi_{-\lambda}\big[M(\mathbb{P}, \pi_{\theta_\star})\big]\Big\}\,\mathrm{d}\mathbb{P}^{\otimes n}$$
$$\leq \int \exp\bigg\{n\lambda \int \Big\{M(\overline{\mathbb{P}}, \theta) - \Phi_{-\lambda}\big[M(\mathbb{P}, \theta)\Big\}\,\mathrm{d}\pi_{\theta_\star}(\theta)\bigg\}\,\mathrm{d}\mathbb{P}^{\otimes n} \leq 1,$$

because $p \mapsto -\Phi_{-\lambda}(p)$ is convex. As a consequence

PROPOSITION 1.6 *With probability at least* $1-2\epsilon$*,*

$$L(\mathbb{P}, \widehat{\theta}) \leq$$
$$\inf_{\theta_\star\in\Theta}\big[1-\varphi\big(\sqrt{\beta}\big)\big]^{-1} B\bigg(B_-\bigg(M(\mathbb{P}, \pi_{\theta_\star}), \frac{\log(\epsilon^{-1})}{n}\bigg), \frac{\beta\|\theta_\star\|^2}{2}, \epsilon\bigg) + \zeta.$$

It is also possible to state a result in terms of empirical margins. Indeed

$$M(w, \pi_\theta) \leq M(w, \theta/2) + \varphi(\sqrt{\beta}).$$

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

$$L(\mathbb{P}, \theta) \leq C_2(\theta),$$

where

$$C_2(\theta) = \big[1-\varphi\big(\sqrt{\beta}\big)\big]^{-1} B\bigg(M(\overline{\mathbb{P}}, \theta/2) + \varphi\big(\sqrt{\beta}\big), \frac{\beta\|\theta\|^2}{2}, \epsilon\bigg).$$

However, C_1 and C_2 are non-convex criterions, faster minimization algorithms are available for the usual SVN loss function, for which it is also possible to derive some generalization bound. Indeed

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

Thus we also have, using this time Proposition 1.3 (page 6)

PROPOSITION 1.7 *With probability at least* $1-\epsilon$*, for any* $\theta \in \mathbb{R}^d$*,*

$$L(\mathbb{P},\theta) \leq \big[1-\varphi\big(\sqrt{\beta}\big)\big]^{-1} B_\Lambda\bigg(\int \big(2-\langle\theta,x\rangle y\big)_+ d\overline{\mathbb{P}}(x,y) + \varphi(\sqrt{\beta}), \\ \frac{\beta\|\theta\|^2 + 2\log\big(|\Lambda|/\epsilon\big)}{2n}\bigg) \\ = \big[1-\varphi\big(\sqrt{\beta}\big)\big]^{-1} \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\bigg[C_3(\lambda,\theta) + \varphi\big(\sqrt{\beta}\,\big) + \frac{\log\big(|\Lambda|/\epsilon\big)}{n\lambda}\bigg],$$

where

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

The loss function $C_3(\lambda,\theta)$ is the most employed learning criterion for support vector machines, and is called the box constraint. It is convex in θ. There are fast algorithms to compute $\inf_\theta C_3(\lambda,\theta)$ for any fixed value of λ. Here we get an empirical criterion which could be used to optimize also the value of λ.

2. THE PAC-BAYES APPROACH PART II : ARBITRARY LOSS FUNCTIONS

2.1. ESTIMATE OF THE RISK AT SOME FIXED PARAMETER VALUE. Let us for short define the risk as

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

We want to minimize R. In the classification section, we started with estimates of $R(\theta)$ for a given value of θ. This is not however always the most effective way to handle the minimization of R. Estimating the variations of the criterion between to parameter values may give faster convergence rates. This is what we are going to explain here, before applying the results to the case of least square regression. Let

$$\begin{aligned} L'(w,\theta,\theta') &\stackrel{\text{def}}{=} L(w,\theta) - L(w,\theta'), && \theta,\theta' \in \Theta, \\ R'(\theta,\theta') &\stackrel{\text{def}}{=} R(\theta) - R(\theta') \\ &= \int L'(w,\theta,\theta')\,\mathrm{d}\mathbb{P}(w), && \theta,\theta' \in \Theta. \end{aligned}$$

Let us first discuss the estimation of $R'(\theta,\theta')$ for fixed values of θ and θ'. We will derive some bounds of the Chernoff kind.

Let us consider the following piecewise $\mathcal{C}^2$ influence function $\psi : \mathbb{R} \to \mathbb{R}$ satisfying the inequality

$$-\log\bigl(1 - x + x^2/2\bigr) \leq \psi(x) \leq \log\bigl(1 + x + x^2/2\bigr), \qquad x \in \mathbb{R}, \tag{7}$$

and defined as

$$\psi(x) = \begin{cases} -\log(2), & x \leq 1, \\ \log\bigl(1 + x + x^2/2\bigr), & -1 \leq x \leq 0, \\ -\log\bigl(1 - x + x^2/2\bigr), & 0 \leq x \leq 1, \\ \log(2), & x \geq 1, \qquad x \in \mathbb{R}, \end{cases}$$

Let us remark that

$$-\log\bigl(1 - x + x^2/2\bigr) = \log\left(\frac{1 + x + x^2/2}{1 + x^4/4}\right) \leq \log\bigl(1 + x + x^2/2\bigr),$$

showing that equation (7) holds. Let us consider for some parameter $\lambda \in \mathbb{R}_+^*$ the empirical counterpart of $R'(\theta, \theta')$ defined as

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

Let us remark that it is antisymmetric : $r_\lambda(\theta, \theta') = -r_\lambda(\theta', \theta)$ for any $\theta, \theta' \in \Theta$. The influence function ψ is chosen in order to satisfy the following lemma.

LEMMA 2.1 *Let us consider some fixed pair of parameters $(\theta, \theta') \in \Theta^2$. Let us assume that $w \mapsto L'(w, \theta, \theta') \in \mathbf{L}_2(\mathbb{P})$. For any $\lambda \in \mathbb{R}_+$,*

$$\int \exp\bigl[n\lambda r'_\lambda(\theta, \theta')\bigr] \, \mathrm{d}\mathbb{P}^{\otimes n} \leq \exp\left\{ n \log\left[1 + \lambda R'(\theta, \theta') + \frac{\lambda^2}{2} \int L'(w, \theta, \theta')^2 \, \mathrm{d}\mathbb{P}(w)\right]\right\}.$$

COROLLARY 2.2 *Under the same hypotheses, with probability at least $1 - \epsilon$,*

$$\begin{aligned} r'_\lambda(\theta, \theta') &\leq \frac{1}{\lambda} \log\left[1 + \lambda R'(\theta, \theta') + \frac{\lambda^2}{2} \int L'(w, \theta, \theta')^2 \, \mathrm{d}\mathbb{P}(w)\right] + \frac{\log(\epsilon^{-1})}{n\lambda} \\ &\leq R'(\theta, \theta') + \frac{\lambda}{2} \int L'(w, \theta, \theta')^2 \, \mathrm{d}\mathbb{P} + \frac{\log(\epsilon^{-1})}{n\lambda}. \end{aligned}$$

Let us assume for simplicity that $\inf_{\theta\in\Theta} R(\theta)$ is reached at $\theta_\star$. Let us choose some slope parameter $a \in \mathbb{R}_+$ and let us put

$$b = \sup\biggl\{ \int L'\bigl(w,\theta,\theta_\star\bigr)^2 \,\mathrm{d}\mathbb{P}(w) - aR'(\theta,\theta_\star)\,;\, \theta\in\Theta \biggr\},$$

so that

$$\int L'(w,\theta,\theta_\star)^2 \,\mathrm{d}\mathbb{P}(w) \le aR'(\theta,\theta_\star) + b.$$

COROLLARY 2.3 *For any $\lambda \in \mathbb{R}_+$, any $\theta\in\Theta$, with probability at least $1-\epsilon$,*

$$R(\theta) - R(\theta_\star) \le \frac{r'_\lambda(\theta,\theta_\star) + \dfrac{b\lambda}{2} + \dfrac{\log(\epsilon^{-1})}{n\lambda}}{1-\dfrac{a\lambda}{2}} \le \frac{\sup\limits_{\theta'\in\Theta} r'_\lambda(\theta,\theta') + \dfrac{b\lambda}{2} + \dfrac{\log(\epsilon^{-1})}{n\lambda}}{1-\dfrac{a\lambda}{2}}.$$

We would like to proceed by minimizing the right-hand side of this last inequality, something we are not allowed to do, because this "probably approximately correct" inequality is not uniform with respect to $\theta\in\Theta$.

2.2. PAC-BAYES BOUNDS. As in the classification case, the PAC-Bayes approach consists in considering randomized values of θ, possibly depending on the sample $(W_1,\dots,W_n)$, whose conditional distribution with respect to $(W_1,\dots,W_n)$ is described by a posterior distribution $\rho : \mathcal{W}^n \to \mathcal{M}_+^1(\Theta)$. As we may for commodity wish to randomize in a larger parameter set, let us introduce the target set $\Theta_\star \subset \Theta$ and let us assume that the infimum of the criterion is reached on this set, so that we may consider $\theta_\star \in \arg\min_{\Theta_\star} R$.

LEMMA 2.4 *Let $\pi \in \mathcal{M}_+^1(\Theta)$ be some prior probability measure.*

$$\int \exp\biggl\{ \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \int \biggl[-n\log\biggl(1-\lambda R'(\theta,\theta') + \frac{\lambda^2}{2}\int L'(w,\theta,\theta')^2\,\mathrm{d}\mathbb{P}(w)\biggr) - n\lambda r'_\lambda(\theta,\theta')\biggr]\mathrm{d}\rho(\theta)\,\mathrm{d}\pi(\theta') - \mathcal{K}(\rho,\pi)\biggr\}\mathrm{d}\mathbb{P}^{\otimes n} \le 1. \quad (8)$$

This lemma is a consequence of the following corollary of equation (2, page 4).

LEMMA 2.5 *For any upper-bounded real valued measurable function h and any probability measure π,*

$$\log\left(\int \exp(h)\,\mathrm{d}\pi\right) = \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi). \tag{9}$$

PROOF. When h is upper-bounded, $\exp(h)$ is bounded and we can define $\pi_{\exp(h)}$ as in Proposition 1.2 (page 4). If $\mathcal{K}(\rho,\pi) = +\infty$, then $\int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi) = -\infty$. If on the other hand $\mathcal{K}(\rho,\pi) < +\infty$, $\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \in \mathbf{L}_1(\rho)$ and

$$\mathcal{K}\left(\rho,\pi_{\exp(h)}\right) = \mathcal{K}(\rho,\pi) - \int h\,\mathrm{d}\rho - \log\left(\int \exp(h)\,\mathrm{d}\pi\right). \tag{10}$$

If $\int h\,\mathrm{d}\rho = -\infty$, then again $\int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi) = -\infty$, otherwise $h \in \mathbf{L}_1(\rho)$, and the above equation shows that in this case

$$\log\left(\int \exp(h)\,\mathrm{d}\pi\right) \geq \int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi). \tag{11}$$

In the case when $\rho = \pi_{\exp(h)}$, the left-hand side of equation (10) is null, so that $h \in \mathbf{L}_1(\rho)$, and equation (10) shows that equality is reached in equation (11) when $\rho = \pi_{\exp(h)}$. □

Let us now come to the proof of Lemma 2.4 (page 15). The left-hand side of the inequality to be proved is well defined, if suitably interpreted. Indeed, even if $L'(\cdot,\theta,\theta') \notin \mathbf{L}_1(\mathbb{P})$, we may still define by convention

$$-\lambda R'(\theta,\theta') + \frac{\lambda^2}{2}\int L'(w,\theta,\theta')^2\,\mathrm{d}\mathbb{P}(w)$$

as

$$\int\left[-\lambda L'(w,\theta,\theta') + \frac{\lambda^2}{2}L'(w,\theta,\theta')^2\right]\mathrm{d}\mathbb{P}(w).$$

Indeed, we are now taking the expectation of a measurable function lower-bounded by $-1/2$, which always makes sense in $\mathbb{R}\cup\{+\infty\}$. Making this interpretation, we get that

$$-n\log\left(1 - \lambda R'(\theta,\theta') + \frac{\lambda^2}{2}\int L'(w,\theta,\theta')\,\mathrm{d}\mathbb{P}(w)\right) - n\lambda r'_\lambda(\theta,\theta')$$

takes its values in $[-\infty, 2n\log(2)]$, due to the fact that ψ is upper bounded by $\log(2)$. Thus we can define its generalized expectation, with values in $\mathbb{R}\cup\{-\infty\}$,

with respect to any probability measure bearing on (θ, θ'). Thus the left-hand side of equation (8) is well defined if we adopt these conventions, and according to Lemma 2.5 (page 16), it is not greater than

$$\int \left(1 - \lambda R'(\theta,\theta') + \frac{\lambda^2}{2}(L')^2(\mathbb{P},\theta,\theta')\right)^{-n} \exp\bigl[-n\lambda r'_\lambda(\theta,\theta')\bigr] d\pi(\theta) d\pi(\theta') d\mathbb{P}^{\otimes n}.$$

According to Fubini's theorem for positive functions, this is equal to

$$\int \left(1 - \lambda R'(\theta,\theta') + \frac{\lambda^2}{2}(L')^2(\mathbb{P},\theta,\theta')\right)^{-n} \exp\bigl[-n\lambda r'_\lambda(\theta,\theta')\bigr] d\mathbb{P}^{\otimes n} d\pi(\theta) d\pi(\theta'),$$

which is in turn not greater than 1, since

$$-\psi(x) \leq \min\Bigl\{\log(2), -\log\bigl(1 - x + x^2/2\bigr)\Bigr\},$$

and therefore

$$\exp\bigl[-n\lambda r'_\lambda(\theta,\theta')\bigr] \leq \prod_{i=1}^n \min\left\{2, \left(1 - \lambda L'(W_i,\theta,\theta') + \frac{\lambda^2}{2} L'(W_i,\theta,\theta')^2\right)\right\}.$$

This ends the proof of Lemma 2.4 (page 15).

In order to go further while staying reasonably explicit, let us assume from now on that $\Theta = \mathbb{R}^d$. Let us consider some probability measure $\pi \in \mathcal{M}_+^1\bigl(\mathbb{R}^d\bigr)$, which we will choose below to be concentrated near the origin. Let π_θ be the shifted measure defined by the formula

$$\int h(\theta')\, \mathrm{d}\pi_\theta(\theta') = \int h(\theta + \theta')\, \mathrm{d}\pi(\theta'),$$

for any bounded measurable function h. Let us consider as already explained, some closed subparameter set $\Theta_\star \subset \Theta$, and $\theta_\star \in \arg\min_{\Theta_\star} R$, assuming for simplicity that it exists.

Simplifying somehow the previous lemma gives

PROPOSITION 2.6 *Let us assume that for any $\theta \in \Theta$, $L(\cdot,\theta) \in \mathbf{L}_1(\mathbb{P})$, so that $R(\theta) = L(\mathbb{P},\theta)$ is well defined, and that $\theta_\star \in \arg\min_{\Theta_\star} R$.*

With probablity at least $1-\epsilon$, for any $\theta \in \Theta$, as soon as

$$(L')^2(\mathbb{P}, \pi_\theta, \pi_{\theta_\star}) < +\infty,$$

$(\theta_1,\theta_2) \mapsto R'(\theta_1,\theta_2) \in \mathbf{L}_1\bigl(\pi_\theta \otimes \pi_{\theta_\star}\bigr)$, and

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

PROOF. Indeed, when $(L')^2\big(\mathbb{P}, \pi_\theta, \pi_{\theta_\star}\big) < +\infty$,

$$\int R'(\theta_1,\theta_2)^2 \,\mathrm{d}\pi_\theta(\theta_1)\mathrm{d}\pi_{\theta_\star}(\theta_2) = \int \left(\int L'(w,\theta_1,\theta_2)\,\mathrm{d}\mathbb{P}(w) \right)^2 \mathrm{d}\pi_\theta(\theta_1)\,\mathrm{d}\pi_{\theta_\star}(\theta_2) \leq (L')^2(\mathbb{P}, \pi_\theta, \pi_{\theta_\star}) < +\infty,$$

so that $R' \in \mathbf{L}_2(\pi_\theta \otimes \pi_{\theta_\star}) \subset \mathbf{L}_1(\pi_\theta \otimes \pi_{\theta_\star})$, and the proposition follows from Lemma 2.4 (page 15) and the fact that $\log(1+x) \leq x$. □

In order to use this proposition, let us modify our generalized margin assumption, assuming instead that for any $D \in \mathbb{R}_+$,

$$(L')^2(\mathbb{P}, \pi_\theta, \pi_{\theta_\star}) \leq a_D R'(\theta,\theta_\star) + b_D, \qquad \theta \in \Theta_\star, R'(\theta,\theta_\star) \leq D^2.$$

Let us also make an assumption linking the excess risk $R'(\theta,\theta_\star)$ with the entropy $\mathcal{K}(\pi_\theta, \pi_{\theta_\star})$. Namely, let us assume that for any $D \in \mathbb{R}_+$ and some $p_D \in \mathbb{R}_+^*$ and $q_D \in \mathbb{R}$,

$$\mathcal{K}(\pi_\theta, \pi_{\theta_\star}) \leq p_D R'(\theta,\theta_\star) + q_D, \qquad \theta \in \Theta_\star, R'(\theta,\theta_\star) \leq D^2.$$

Let us assume in the same way that for some $\xi \in \mathbb{R}$,

$$R'(\theta,\theta_\star) \leq R(\pi_\theta, \pi_{\theta_\star}) + \xi, \qquad \theta \in \Theta_\star.$$

(In the application to least square regression, we can take $\xi = 0$.)

COROLLARY 2.7 *With probability at least $1-\epsilon$, for any $\theta \in \Theta_\star$ and any $D \in \mathbb{R}_+$ such that $R'(\theta,\theta_\star) \leq D^2$,*

$$R'(\theta,\theta_\star) \leq \left(1 - \frac{a_D\lambda}{2} - \frac{p_D}{n\lambda}\right)^{-1} \left(r'_\lambda(\pi_\theta, \pi_{\theta_\star}) + \frac{\lambda b_D}{2} + \frac{q_D + \log(\epsilon^{-1})}{n\lambda} + \xi \right).$$

Let us consider some pseudo-estimator $\widetilde{\theta}(W_1, \dots, W_n)$ such that

$$r'_\lambda(\pi_{\widetilde{\theta}}, \pi_{\theta_\star}) = \inf_{\theta \in \Theta_\star} r'_\lambda(\pi_\theta, \pi_{\theta_\star}).$$

We assume for simplicity that it exists. (If not, we can find some estimator which reaches the infimum up to some arbitrary small margin.) This is not a legitimate estimator, since it cannot be computed from the observations, but it will serve nevertheless to state the following proposition.

PROPOSITION 2.8 *Let us consider any estimator $\widehat{\theta} \in \Theta_\star$ such that for some $\zeta > 0$,*

$$\sup_{\theta' \in \Theta_\star} r'_\lambda(\pi_{\widehat{\theta}}, \pi_{\theta'}) \leq \inf_{\theta \in \Theta_\star} \sup_{\theta' \in \Theta_\star} r'_\lambda(\pi_\theta, \pi'_\theta) + \zeta.$$

With probability at least $1-\epsilon$*, for any* $D \in \mathbb{R}_+$*, as soon as*

$$\max\bigl\{R'(\widehat{\theta},\theta_\star), R'(\widetilde{\theta},\theta_\star)\bigr\} \leq D^2, \tag{12}$$

$$R'(\widehat{\theta},\theta_\star) \leq \left(1-\frac{\lambda a_D}{2}-\frac{p_D}{n\lambda}\right)^{-1}\biggl[\sup_{\theta'\in\Theta_\star} r'_\lambda(\pi_{\widehat{\theta}},\pi_{\theta'}) + \frac{\lambda b_D}{2}+\frac{q_D+\log(\epsilon^{-1})}{n\lambda}+\xi\biggr]$$

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

PROOF. Let us mention that $(\theta,\theta') \mapsto r'_\lambda(\theta,\theta')$ is a bounded measurable function, since $-\lambda^{-1}\log(2) \leq r'_\lambda(\theta,\theta') \leq \lambda^{-1}\log(2)$, thus the suprema and infima appearing in the proposition are all finite real numbers. Let us remark now that, according to the assumption made on the estimator $\widehat{\theta}$,

$$\begin{aligned} r'_\lambda(\pi_{\widehat{\theta}},\pi_{\theta_\star}) &\leq \sup_{\theta'\in\Theta_\star} r'_\lambda(\pi_{\widehat{\theta}},\pi_{\theta'}) \leq \sup_{\theta'\in\Theta_\star} r'_\lambda(\pi_{\theta_\star},\pi_{\theta'})+\zeta \\ &= -\inf_{\theta'\in\Theta_\star} r'_\lambda(\pi_{\theta'},\pi_{\theta_\star})+\zeta = -r'_\lambda(\pi_{\widetilde{\theta}},\pi_{\theta_\star})+\zeta. \end{aligned}$$

Applying Corollary 2.7 (page 18) we deduce that with probability at least $1-\epsilon$, as soon as condition (12) is fullfilled,

$$\begin{aligned} R'(\widehat{\theta},\theta_\star) &\leq \left(1-\frac{\lambda a_D}{2}-\frac{p_D}{n\lambda}\right)^{-1}\biggl[\sup_{\theta'\in\Theta_\star} r'_\lambda(\pi_{\widetilde{\theta}},\pi_{\theta'}) + \frac{\lambda b_D}{2}+\frac{q_D+\log(\epsilon^{-1})}{n\lambda}+\xi\biggr], \\ &\leq \left(1-\frac{\lambda a_D}{2}-\frac{p_D}{n\lambda}\right)^{-1}\biggl[-r'_\lambda(\pi_{\widetilde{\theta}},\pi_{\theta_\star})+\zeta + \frac{\lambda b_D}{2}+\frac{q_D+\log(\epsilon^{-1})}{n\lambda}+\xi\biggr], \\ R'(\widetilde{\theta},\theta_\star) &\leq \left(1-\frac{\lambda a_D}{2}-\frac{p_D}{n\lambda}\right)^{-1}\biggl[r'_\lambda(\pi_{\widetilde{\theta}},\pi_{\theta_\star}) + \frac{\lambda b_D}{2}+\frac{q_D+\log(\epsilon^{-1})}{n\lambda}+\xi\biggr], \end{aligned}$$

so that the last result of the proposition is obtained by summing up these two inequalities. □

2.3. SIMPLIFIED CRITERION. As we will see below, in some situations, $r'_\lambda(\pi_\theta, \pi_{\theta'})$ can be compared with the simpler criterion

$$\widetilde{r}_\lambda(\theta, \theta') \stackrel{\text{def}}{=} \lambda^{-1} \int \psi\big[\lambda L'(w, \pi_\theta, \pi_{\theta'})\big] d\overline{\mathbb{P}}(w),$$

where the integration with respect to π_θ and $\pi_{\theta'}$ is performed inside the influence function ψ.

Indeed let us introduce the numerical constant $c = 3/\log(4) \leq 2.17$ and the (hopefully) small quantity

$$\eta_D = c\lambda \int \sup\Big\{ \mathbf{Var}\big[L(w,\theta_1)\,\mathrm{d}\pi_\theta(\theta_1)\big], \theta \in \Theta_\star, R'(\theta,\theta_\star) \leq D^2 \Big\}\, \mathrm{d}\mathbb{P}(w) \\ + c\lambda \int \mathbf{Var}\big[L(w,\theta_2)\,\mathrm{d}\pi_{\theta_\star}(\theta_2)\big]\, \mathrm{d}\mathbb{P}(w) + \frac{\log(\epsilon^{-1})}{n\lambda}.$$

We assume that it makes sense, namely that $\int L(w,\theta')^2\,\mathrm{d}\pi_\theta(\theta')\,\mathrm{d}\mathbb{P}(w) < \infty$ for all $\theta \in \Theta_\star$ and that

$$w \mapsto \sup\Big\{ \mathbf{Var}\big[L(w,\theta_1)\,\mathrm{d}\pi_\theta(\theta_1)\big], \theta \in \Theta_\star, R'(\theta,\theta_\star) \leq D^2 \Big\} \in \mathbf{L}_1(\mathbb{P}).$$

LEMMA 2.9 *For any $D \in \mathbb{R}_+$, with probability at least $1-\epsilon$, for any $\theta \in \Theta_\star$ such that $R'(\theta,\theta_\star) \leq D^2$,*

$$r'_\lambda(\pi_\theta, \pi_{\theta_\star}) \leq \widetilde{r}_\lambda(\theta, \theta_\star) + \eta_D.$$

PROOF. Let us remark first that for any $x \in [0,1]$,

$$\psi'(x) = \frac{1-x}{1-x+x^2/2},$$
$$\psi''(x) = -2\frac{1-(1-x)^2}{\big[1+(1-x)^2\big]^2} \geq -2.$$

Using the symmetries of ψ, we deduce from this inequality that

$$x \mapsto \overline{\psi}_m(x) = \psi(x) + (x-m)^2$$

is convex on the whole real line for any value of $m \in \mathbb{R}$. Jensen's inequality tells us that for any probability measure $\rho \in \mathcal{M}_+^1(\Theta)$ and any function $h \in \mathbf{L}_1(\rho)$,

$$\overline{\psi}_m\left[\int h(\theta)\,\mathrm{d}\rho(\theta)\right] \leq \int \overline{\psi}_m\big[h(\theta)\big]\,\mathrm{d}\rho(\theta).$$

Choosing $m = \int h(\theta)\,\mathrm{d}\rho(\theta)$ gives

$$\psi\Big[\int h(\theta)\,\mathrm{d}\rho(\theta)\Big] \leq \int \psi\big[h(\theta)\big]\,\mathrm{d}\rho(\theta) + \mathbf{Var}\big[h(\theta)\,\mathrm{d}\rho(\theta)\big].$$

Using the symmetry $\psi(-h) = -\psi(h)$ and working now with $-h$ instead of h proves the reversed inequality, so that

LEMMA 2.10 *For any probability measure $\rho \in \mathcal{M}_+^1(\Theta)$ and any function $h \in \mathbf{L}_1(\rho)$,*

$$\begin{aligned}\left|\int \psi\big[h(\theta)\big]\,\mathrm{d}\rho(\theta) - \psi\Big[\int h(\theta)\,\mathrm{d}\rho(\theta)\Big]\right| &\leq \min\Big\{\log(4), \mathbf{Var}\big[h(\theta)\,\mathrm{d}\rho(\theta)\big]\Big\} \\ &\leq \log\Big\{1 + c\,\mathbf{Var}\big[h(\theta)\,\mathrm{d}\rho(\theta)\big]\Big\}.\end{aligned}$$

PROOF. The last inequality is a consequence of the inequality $\min\big\{\log(4), x\big\} \leq \log(1+cx)$, which, according to the fact that $x \mapsto \log(1+cx)$ is concave has to be checked only when $x = \log(4)$, where we get an equality for the chosen value of c. □

Applying this lemma, we see that

$$\begin{aligned}&r'_\lambda(\pi_\theta, \pi_{\theta_\star}) - \widetilde{r}_\lambda(\theta, \theta_\star) \\ &\qquad \leq \lambda^{-1} \int \log\Big\{1 + c\lambda^2\,\mathbf{Var}\big[L'(w, \theta_1, \theta_2)\,\mathrm{d}\pi_\theta(\theta_1)\,\mathrm{d}\pi_{\theta_\star}(\theta_2)\big]\Big\}\,\mathrm{d}\overline{\mathbb{P}}(w).\end{aligned}$$

We can now remark that

$$\begin{aligned}&\mathbf{Var}\big[L'(w, \theta_1, \theta_2)\,\mathrm{d}\pi_\theta(\theta_1)\,\mathrm{d}\pi_{\theta_\star}(\theta_2)\big] \\ &\qquad = \mathbf{Var}\big[L(w, \theta_1)\,\mathrm{d}\pi_\theta(\theta_1)\big] + \mathbf{Var}\big[L(w, \theta_2)\,\mathrm{d}\pi_{\theta_\star}(\theta_2)\big].\end{aligned}$$

To end the proof of Lemma 2.9 (page 20), it is enough now to use the fact that for any $h \in \mathbf{L}_1(\mathbb{P})$, with probability at least $1 - \epsilon$

$$\begin{aligned}\int \log\big[1 + h(w)\big]\,\mathrm{d}\overline{\mathbb{P}}(w) &\leq \log\Big[1 + \int h(w)\,\mathrm{d}\mathbb{P}(w)\Big] + \frac{\log(\epsilon^{-1})}{n} \\ &\leq \int h(w)\,\mathrm{d}\mathbb{P}(w) + \frac{\log(\epsilon^{-1})}{n}.\end{aligned}$$

□

PROPOSITION 2.11 *Let us consider some pseudo-estimator $\widetilde{\theta} \in \Theta_\star$ such that*

$$\widetilde{r}_\lambda(\widetilde{\theta}, \theta_\star) = \inf_{\theta \in \Theta_\star} \widetilde{r}_\lambda(\theta, \theta_\star)$$

and any estimator $\widehat{\theta}$ such that

$$\sup_{\theta' \in \Theta_\star} \widetilde{r}_\lambda(\widehat{\theta}, \theta') \leq \inf_{\theta \in \Theta_\star} \sup_{\theta' \in \Theta_\star} \widetilde{r}_\lambda(\theta, \theta') + \zeta.$$

For any $D \in \mathbb{R}_+$, with probability at least $1 - 2\epsilon$, as soon as

$$\max\bigl\{R'(\widehat{\theta}, \theta_\star), R'(\widetilde{\theta}, \theta_\star)\bigr\} \leq D^2, \tag{13}$$

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

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

PROOF. Combining Lemma 2.9 (page 20) with Corollary 2.7 (page 18), we obtain with probability at least $1 - 2\epsilon$ that for any $\theta \in \Theta_\star$ such that

$$R'(\theta, \theta_\star) \leq D^2,$$

$$R'(\theta, \theta_\star) \leq \left(1 - \frac{\lambda a_D}{2} - \frac{p_D}{n\lambda}\right)^{-1} \left(\widetilde{r}_\lambda(\theta, \theta_\star) + \frac{\lambda b_D}{2} + \frac{q_D + \log(\epsilon^{-1})}{n\lambda} + \eta_D + \xi \right).$$

The end of the proof is the same as in Proposition 2.8 (page 19). □

2.4. THE EXAMPLE OF LINEAR LEAST SQUARE REGRESSION. Let us apply the previous propositions to the case where $w = (x, y) \in \mathbb{R}^d \times \mathbb{R}$, and $L(w, \theta) = \bigl(\langle \theta, x \rangle - y\bigr)^2$. It is representative of the local behaviour of any smooth loss function and explicit computations can be performed. Let us work with Gaussian perturbations, choosing as reference measure the Gaussian centered measure

$$\frac{\mathrm{d}\pi}{\mathrm{d}\theta}(\theta) = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\bigl(-\beta \|\theta\|^2 / 2\bigr).$$

Let us assume that $\displaystyle\int y^4 \, \mathrm{d}\mathbb{P}(y) < +\infty$ and $\displaystyle\int \|x\|^4 \, \mathrm{d}\mathbb{P}(x) < +\infty$. Let us also assume that $\Theta_\star$ is a closed convex set and that $R(\theta_\star) = \inf\{R(\theta), \theta \in \Theta_\star\}$.

Exercise 4 *Show that*

$$R'(\theta,\theta_\star) \geq \int \langle \theta-\theta_\star, x\rangle^2 \,\mathrm{d}\mathbb{P}(x), \qquad \theta\in\Theta_\star,$$
$$R(\pi_\theta) = R(\theta) + \beta^{-1}\int \|x\|^2 \,\mathrm{d}\mathbb{P}(x), \qquad \theta\in\mathbb{R}^d,$$

so that $R'(\pi_\theta,\pi_{\theta_\star}) = R'(\theta,\theta_\star)$ *and we can take* $\xi = 0$.

Let us also recall that

$$\mathcal{K}(\pi_\theta,\pi_{\theta_\star}) = \frac{\beta}{2}\|\theta-\theta_\star\|^2.$$

Let us now compute legitimate values for a_D and b_D.

$$\begin{aligned}(L')^2(\mathbb{P},\pi_\theta,\pi_{\theta_\star}) &= \int \Big[\langle\theta_1-\theta_0,x\rangle^2 \\ &\qquad + 2\langle\theta_1-\theta_0,x\rangle\big(\langle\theta_0,x\rangle - y\big)\Big]^2 \,\mathrm{d}\pi_\theta(\theta_1)\,\mathrm{d}\pi_{\theta_\star}(\theta_0)\,\mathrm{d}\mathbb{P}(x,y)\\ &\leq \int \Big[2\langle\theta_1-\theta_0,x\rangle^4 \\ &\qquad + 8\langle\theta_1-\theta_0,x\rangle^2\big(\langle\theta_0,x\rangle - y\big)^2 \,\mathrm{d}\pi_\theta(\theta_1)\,\mathrm{d}\pi_{\theta_\star}(\theta_0)\,\mathrm{d}\mathbb{P}(x,y).\end{aligned}$$

Let us put

$$\begin{aligned} g_1 &= \langle\theta_1-\theta,x\rangle, & g_0 &= \langle\theta_0-\theta_\star,x\rangle,\\ m_1 &= \langle\theta-\theta_\star,x\rangle, & m_0 &= \langle\theta_\star,x\rangle - y.\end{aligned}$$

With these notations

$$\begin{aligned}&\int \langle\theta_1-\theta_0,x\rangle^4 \,\mathrm{d}\pi_\theta(\theta_1)\,\mathrm{d}\pi_{\theta_\star}(\theta_0)\\ &\quad = \mathbb{E}\big[(g_1-g_0+m_1)^4\big] = \mathbb{E}\big[(g_1-g_0)^4\big] + 6\mathbb{E}\big[(g_1-g_0)^2\big]m_1^2 + m_1^4\\ &\qquad = \frac{12\|x\|^4}{\beta^2} + \frac{12\|x\|^2}{\beta}\langle\theta-\theta_\star,x\rangle^2 + \langle\theta-\theta_\star,x\rangle^4,\end{aligned}$$

and

$$\begin{aligned}&\int \langle\theta_1-\theta_0,x\rangle^2\big(\langle\theta_0,x\rangle-y\big)^2 \,\mathrm{d}\pi_\theta(\theta_1)\,\mathrm{d}\pi_{\theta_\star}(\theta_0)\\ &\qquad = \mathbb{E}\big[(g_1-g_0+m_1)^2(g_0+m_0)^2\big]\\ &\quad = \mathbb{E}\Big\{\big[(g_1-g_0)^2 + 2(g_1-g_0)m_1 + m_1^2\big]\big[g_0^2+2g_0m_0+m_0^2\big]\Big\}\\ &\quad = \mathbb{E}\big[g_0^4 + g_0^2g_1^2 + (g_1-g_0)^2m_0^2 - 4g_0^2m_0m_1 + m_1^2g_0^2 + m_1^2m_0^2\big]\end{aligned}$$

$$\leq \frac{4\|x\|^4}{\beta^2} + \frac{4\|x\|^2}{\beta}\big(\langle\theta_\star, x\rangle - y\big)^2 + \frac{3\|x\|^2}{\beta}\langle\theta-\theta_\star, x\rangle^2 + \langle\theta-\theta_\star, x\rangle^2\big(\langle\theta_\star, x\rangle - y\big)^2.$$

Thus

$$(L')^2(\mathbb{P}, \pi_\theta, \pi_{\theta_\star}) \leq \int \bigg(2\langle\theta-\theta_\star, x\rangle^4 + 8\langle\theta-\theta_\star, x\rangle^2\big(\langle\theta_\star, x\rangle - y\big)^2 + \frac{48\|x\|^2}{\beta}\langle\theta-\theta_\star, x\rangle^2 + \frac{32\|x\|^2}{\beta}\big(\langle\theta_\star, x\rangle - y\big)^2 + \frac{56\|x\|^4}{\beta^2}\bigg)\,\mathrm{d}\mathbb{P}(x,y).$$

Let us define

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

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

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

assuming all these quantities are finite. We obtain that

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

where

$$a_D = 8\sqrt{\kappa}\sigma_4^2 + 2\kappa D^2 + \frac{48\sqrt{\kappa}s_4^2}{\beta},$$

$$b_D = \frac{32 s_4^2}{\beta}\sigma_4^2 + \frac{56 s_4^4}{\beta^2}.$$

Let us now discuss the simplified criterion. Let us notice first that

$$L(w, \pi_\theta) = \int\big(\langle\theta', x\rangle - y\big)^2\,\mathrm{d}\pi_\theta(\theta') = \big(\langle\theta, x\rangle - y\big)^2 + \frac{\|x\|^2}{\beta},$$

so that

$$L'(w, \pi_\theta, \pi_{\theta'}) = L'(w, \theta, \theta')$$

and

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

As the criterion $\widetilde{r}_\lambda$ which serves to compute our estimator $\widehat{\theta}$ requires only to compute scalar products of the form $\langle \theta, x \rangle$, we are entitled to make in the previous computations any change of representation which preserves this scalar product. We can thus consider the Gram matrix

$$G = \int x x^t \, \mathrm{d}\mathbb{P}(x),$$

and make the change of representation $(x, \theta) \mapsto (G^{-1/2}x, G^{1/2}\theta)$. If G is not invertible, we can restrict $\mathcal{X}$ and Θ to the linear subspace generated by the eigenvectors of G with non zero eigenvalues, since $X_1, \dots, X_n$ almost surely belong to this subspace. Therefore we can assume without loss of generality that G is the identity, because this becomes true after we restrict the space and make the proposed change of representation. Now that we made this change of representation and assume that $G = \mathbf{Id}$, we have

$$\begin{aligned}
\int \|x\|^2 \, \mathrm{d}\mathbb{P}(x) &= d, \\
\mathcal{K}(\pi_\theta, \pi_{\theta_\star}) &= \frac{\beta}{2}\|\theta - \theta_\star\|^2 = \frac{\beta}{2}\int \langle \theta - \theta_\star, x \rangle^2 \, \mathrm{d}\mathbb{P}(x) \\
&\leq \frac{\beta}{2} R'(\theta, \theta_\star),
\end{aligned}$$

so that we can take $p_D = \dfrac{\beta}{2}$, and $q_D = 0$.

Let us now compute η_D. The variance of the sum of two uncorrelated random variable being the sum of their variances,

$$\begin{aligned}
\mathbf{Var}\big[L(w, \theta') \, \mathrm{d}\pi_\theta(\theta')\big] &= \mathbf{Var}\Big\{ \big[\langle \theta' - \theta, x \rangle^2 + 2\langle \theta - \theta', x \rangle \big(\langle \theta, x \rangle - y \big)\big] \, \mathrm{d}\pi_\theta(\theta') \Big\} \\
&= \mathbf{Var}\big[\langle \theta' - \theta, x \rangle^2 \, \mathrm{d}\pi_\theta(\theta')\big] \\
&\qquad + 4\big(\langle \theta, x \rangle - y\big)^2 \, \mathbf{Var}\big[\langle \theta' - \theta, x \rangle \, \mathrm{d}\pi_\theta(\theta')\big] \\
&= \frac{2\|x\|^4}{\beta^2} + 4\big(\langle \theta, x \rangle - y\big)^2 \frac{\|x\|^2}{\beta} \\
&\leq \frac{2\|x\|^4}{\beta^2} + 8\big(\langle \theta_\star, x \rangle - y\big)^2 \frac{\|x\|^2}{\beta} + 8\langle \theta - \theta_\star, x \rangle^2 \frac{\|x\|^2}{\beta} \\
&\leq 8\big(\langle \theta_\star, x \rangle - y\big)^2 \frac{\|x\|^2}{\beta} + 8D^2 \frac{\|x\|^4}{\beta} + \frac{2\|x\|^4}{\beta^2}.
\end{aligned}$$

Thus we can take

$$\eta_D = c\lambda \left[\frac{12 s_4^2 \sigma_4^2}{\beta} + \left(\frac{8D^2}{\beta} + \frac{4}{\beta^2} \right) s_4^4 \right] + \frac{\log(\epsilon^{-1})}{n\lambda}.$$

PROPOSITION 2.12 *With probability at least* $1-2\epsilon$*, as soon as*

$$\max\big\{R'(\widehat{\theta},\theta_\star), R'(\widetilde{\theta},\theta_\star)\big\} \le D^2,$$

$$R'(\widehat{\theta},\theta_\star) + R'(\widetilde{\theta},\theta_\star) \le \left(1 - \frac{a_D\lambda}{2} - \frac{\beta}{2n\lambda}\right)^{-1}\left(b\lambda + \frac{2\log(\epsilon^{-1})}{n\lambda} + 2\eta_D + \zeta\right),$$

where

$$\begin{aligned} a_D &= 8\sqrt{\kappa}\sigma_4^2 + 2\kappa D^2 + \frac{48\sqrt{\kappa}s_4^2}{\beta},\\ b &= \frac{32 s_4^2\sigma_4^2}{\beta} + \frac{56 s_4^4}{\beta^2},\\ \eta_D &= c\lambda\left[\frac{12 s_4^2\sigma_4^2}{\beta} + \left(\frac{8D^2}{\beta} + \frac{4}{\beta^2}\right)s_4^4\right] + \frac{\log(\epsilon^{-1})}{n\lambda}. \end{aligned}$$

Let us put

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

and let us take

$$\begin{aligned} \lambda^{-1} &= 8\big(4\sqrt{\kappa}\sigma_4^2 + \kappa\Delta^2\big),\\ \beta^{-1} &= \frac{4}{n\lambda}. \end{aligned}$$

COROLLARY 2.13 *For any* n *such that*

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

for any $D\in\mathbb{R}_+$*, with probability at least* $1-2\epsilon$*, as soon as*

$$\max\big\{R'(\widehat{\theta},\theta_\star), R'(\widetilde{\theta},\theta_\star)\big\} \le D^2,$$

$$\begin{aligned} R'(\widehat{\theta},\theta_\star) + R'(\widetilde{\theta},\theta_\star) \le{}& 2^6(4+3c)\frac{s_4^2\sigma_4^2}{n} + 2^6\sqrt{\kappa}\Big(4\sigma_4^2 + \sqrt{\kappa}\Delta^2\Big)\frac{\log(\epsilon^{-1})}{n}\\ &+ 2^{11}(7+c)\sqrt{\kappa}\big(4\sigma_4^2 + \sqrt{\kappa}\Delta^2\big)\frac{s_4^4}{n^2} + \frac{2^7 c\, s_4^4 D^2}{n} + 2\zeta. \end{aligned}$$

Starting with $D=\Delta$ and applying the above corollary twice we obtain

PROPOSITION 2.14 *For any* n *such that*

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

with probability at least $1-\epsilon$*,*

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

where

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

Exercise 5 *Deduce from the previous proposition that*

$$\begin{aligned}
\int R'(\widehat{\theta},\theta_\star)\,\mathrm{d}\mathbb{P}^{\otimes n} \leq{}& \frac{\big[B_1\, s_4^2 + B_2\,\sqrt{\kappa}\big(1+\log(4)\big)\big]\sigma_4^2}{n} \\
&+ \frac{B_3\,\kappa\big(1+\log(4)\big)\Delta^2}{n} + B_4\,\zeta.
\end{aligned}$$

Let us remark that the order of magnitude of the bound depends on the values of σ_4^2, which measures the size of the noise, Δ, which measures the size of the target parameter set $\Theta_\star$, and of two remaining quantities, s_4^2 and κ, which are linked with the dimension d.

Indeed, in any case

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

Let us give more precisions in some special cases.

1. The distribution of x under $\mathbb{P}$ is a multivariate Gaussian measure. In this case, it is still so after the linear change of representation we used to turn the Gram matrix into the identity matrix. Thus in this case

$$s_4^4 = \int \|x\|^4\,\mathrm{d}\mathbb{P}(x) = \int\left(\sum_{i=1}^d x_i^2\right)^2\mathrm{d}\mathbb{P}(x)$$

$$= \sum_{i=1}^{d} \int x_i^4 \, \mathrm{d}\mathbb{P}(x) + 2 \sum_{1 \leq i < j \leq d} \int x_i^2 x_j^2 \, \mathrm{d}\mathbb{P}(x)$$
$$= 3d + d(d-1) = d^2 + 2d,$$

so that $s_4^2 = d\sqrt{1 + 2/d}$. Moreover, $\langle \theta, x \rangle$ is Gaussian under $\mathbb{P}$ for any θ, so that $\kappa = 3$ in this case.

2. The distribution of x under $\mathbb{P}$ is such that $x_1 = 1$ almost surely and $x_2, \ldots, x_d$ are independent. In this case, the linear change of representation made to normalize the Gram matrix will renormalize $x_2, \ldots, x_d$ independently by substracting their means and making a change of scale to change their variance to 1, while x_1 will remain unchanged. Let us introduce

$$\chi = \max_{i=1,\ldots,d} \frac{\displaystyle\int x_i^4 \, \mathrm{d}\mathbb{P}(x)}{\displaystyle\left(\int x_i^2 \, \mathrm{d}\mathbb{P}(x) \right)^2}.$$

We get $s_4^4 \leq \chi d + d(d-1)$, so that $s_4^2 \leq d\sqrt{1 + (\chi - 1)/d}$.

Let θ be such that $\|\theta\| = 1$. As

$$\int \langle \theta, x \rangle^4 \, \mathrm{d}\mathbb{P}(x) = \int \left(\sum_{i=1}^{d} \theta_i^4 x_i^4 + 6 \sum_{1 \leq i < j \leq d} \theta_i^2 \theta_j^2 x_i^2 x_j^2 + 4 \, \theta_1 \sum_{i=2}^{d} \theta_j^3 x_1 x_i^3 \right) \mathrm{d}\mathbb{P}(x)$$
$$\leq \chi \sum_{i=1}^{d} \theta_i^4 + 6 \sum_{1 \leq i < j \leq d} \theta_i^2 \theta_j^2 + 4\sqrt{\chi} \, |\theta_1| \sum_{i=2}^{d} |\theta_i|^3$$
$$= (\chi - 3) \sum_{i=1}^{d} \theta_i^4 + 3 \left(\sum_{i=1}^{d} \theta_i^2 \right)^2 + 4 \sqrt{\chi} \, |\theta_1| \sum_{i=2}^{d} |\theta_i|^3$$
$$\leq \max\{\chi, 3\} + 4\sqrt{\chi} \sup_{p \in [0,1]} p(1 - p^2)^{3/2},$$

$$\kappa \leq \max\{\chi, 3\} + \frac{3^{3/2}}{4} \sqrt{\chi}.$$

3. The distribution of x under $\mathbb{P}$ is almost surely bounded and nearly orthogonal. More precisely let us assume that for some positive constants A and

B,

$$\mathbb{P}\big(\|x\| \leq B\big) = 1, \qquad \text{and} \qquad \sup_{\theta \neq 0} \frac{\|\theta\|^2}{\int \langle \theta, x \rangle^2 \, \mathrm{d}\mathbb{P}(x)} \leq A^2.$$

In this situation, $\kappa \leq A^2 B^2$, since

$$\int \langle \theta, x \rangle^4 \, \mathrm{d}\mathbb{P}(x) \leq B^2 \|\theta\|^2 \int \langle \theta, x \rangle^2 \, \mathrm{d}\mathbb{P}(x) \leq A^2 B^2 \bigg(\int \langle \theta, x \rangle^2 \, \mathrm{d}\mathbb{P}(x) \bigg)^2 .$$

Moreover $s_4^2 \leq AB\sqrt{d}$. Indeed,

$$\begin{aligned} s_4^4 &= \int \sup_{\theta \neq 0} \frac{\langle \theta, x \rangle^4}{\Big(\int \langle \theta, u \rangle^2 \, \mathrm{d}\mathbb{P}(u) \Big)^2} \, \mathrm{d}\mathbb{P}(x) \\ &\qquad \leq A^2 B^2 \int \sup_{\theta \neq 0} \frac{\langle \theta, x \rangle^2}{\int \langle \theta, u \rangle^2 \, \mathrm{d}\mathbb{P}(u)} \, \mathrm{d}\mathbb{P}(x) = A^2 B^2 d. \end{aligned}$$

The condition involving A means that the smallest eigen-value of the Gram matrix is not smaller than A^{-2}. It is for instance the case if

$$\begin{aligned} \int x_i^2 \, \mathrm{d}\mathbb{P}(x) &\geq 1, & & i = 1, \ldots, d, \\ \left| \int x_i x_j \, \mathrm{d}\mathbb{P}(x) \right| &\leq \frac{1 - A^{-2}}{d-1}, & & 1 \leq i < j \leq d. \end{aligned}$$

Let us remark that in this setting, necessarily $A^2 B^2 \geq s_4^4 / d \geq d$, so that this cannot yield a bound for κ lower than d. This situation is met when estimating functions in orthogonal or nearly orthogonal bases. For instance let us consider the Fourier basis on the unit interval, defined for any $u \in [0,1]$ as

$$\begin{aligned} \varphi_0(u) &= 1, \\ \varphi_{2i-1}(u) &= \sqrt{2} \cos\big(2k\pi u\big), & & 1 \leq i \leq r, \\ \varphi_{2i}(u) &= \sqrt{2} \sin\big(2k\pi u\big), & & 1 \leq i \leq r. \end{aligned}$$

Let $\mathrm{d}\mathbb{P}(x)$ be the image of the uniform probability measure $\mathbb{U}$ on the unit interval $[0,1]$, by the map φ, so that for any measurable set $E \subset \mathbb{R}^{2r+1}$, $\mathbb{P}(x \in E) = \mathbb{U}\big[\varphi^{-1}(E)\big]$. In this case $A = 1$, because we have an orthogonal basis, and $B = \sqrt{1+2r} = \sqrt{d}$. So in this case we get a d/n convergence rate.

The same rate can be achieved with localized bases, the most simple being the even histogram basis, defined as

$$\varphi_i(u) = \sqrt{d}\,\mathbb{1}\big(u \in \big[(i-1)/d, i/d\big[\big), \qquad u \in [0,1], \quad 1 \leq i \leq d.$$

Setting here again $\mathbb{P}(x \in E) = \mathbb{U}\big[\Phi^{-1}(E)\big]$ for any measurable set E, we obtain as in the previous case that $A = 1$ and $B = \sqrt{d}$.

3. ORDINARY LEAST SQUARE ESTIMATOR

We are going to study in this section the ordinary least square estimator $\widehat{\theta} = \arg\min_{\theta \in \Theta_\star} L(\overline{\mathbb{P}}, \theta)$, the setting being the same as in the end of the previous section.

A slight variant of Proposition 2.6 (page 18) is obtained by replacing $R'(\pi_\theta, \pi_{\theta_\star})$ with $R'(\pi_\theta, \theta_\star)$ and accordingly $r'_\lambda(\pi_\theta, \pi_{\theta_\star})$ with $r'_\lambda(\pi_\theta, \theta_\star)$. We still obtain with probability at least $1-\epsilon$ that for any $\theta \in \Theta$ such that

$$\big(L'\big)^2(\mathbb{P}, \pi_\theta, \theta_\star) < +\infty,$$

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

Moreover it is straightforward to deduce from the properties of ψ that

$$r'_\lambda(\pi_\theta, \theta_\star) \leq L'(\overline{\mathbb{P}}, \pi_\theta, \theta_\star) + \frac{\lambda}{2}\big(L'\big)^2(\overline{\mathbb{P}}, \pi_\theta, \theta_\star).$$

Let us introduce the empirical counterparts of σ_4, s_4 and κ, defined as

$$\begin{aligned}
\overline{\sigma}_4^4 &= \int \big(\langle \theta_\star, x\rangle - y\big)^4 \,\mathrm{d}\overline{\mathbb{P}}(x,y),\\
\overline{s}_4^4 &= \int \|x\|^4 \,\mathrm{d}\overline{\mathbb{P}}(x),\\
\overline{\kappa} &= \sup\left\{\int \langle \theta, x\rangle^4 \,\mathrm{d}\overline{\mathbb{P}}(x), \theta \in \mathbb{R}^d, \|\theta\| = 1\right\}.
\end{aligned}$$

Here we assume as in the end of the previous section that we have made the necessary linear change of variables on θ and x, replacing if necessary (θ, x) with $(G^{1/2}\theta, G^{-1/2}x)$, to turn the Gram matrix into the identity matrix, while letting the scalar product $\langle \theta, x\rangle$ unchanged.

$$\big(L'\big)^2(w, \pi_\theta, \theta_\star) = \int \big[\langle \theta' - \theta_\star, x\rangle^2 + 2\langle \theta' - \theta_\star, x\rangle\big(\langle \theta_\star, x\rangle - y\big)\big]^2 \,\mathrm{d}\pi_\theta(\theta')$$

$$\leq \int \big[2\langle\theta'-\theta+\theta-\theta_\star,x\rangle^4 + 8\langle\theta'-\theta+\theta-\theta_\star,x\rangle^2\big(\langle\theta_\star,x\rangle-y\big)^2\big]\,\mathrm{d}\pi_\theta(\theta')$$
$$\leq \frac{6\|x\|^4}{\beta^2} + \frac{12\|x\|^2}{\beta}\langle\theta-\theta_\star,x\rangle^2 + 2\langle\theta-\theta_\star,x\rangle^4$$
$$+ \frac{8\|x\|^2}{\beta}\big(\langle\theta_\star,x\rangle-y\big)^2 + 8\langle\theta-\theta_\star,x\rangle^2\big(\langle\theta_\star,x\rangle-y\big)^2.$$

Thus

$$\big(L'\big)^2(\overline{\mathbb{P}},\pi_\theta,\theta_\star) \leq 8\|\theta-\theta_\star\|^2\sqrt{\overline{\kappa}}\,\overline{\sigma}_4^2 + 2\|\theta-\theta_\star\|^4\overline{\kappa}$$
$$+ \frac{12\overline{s}_4^2\sqrt{\kappa}}{\beta}\|\theta-\theta_\star\|^2 + \frac{8\overline{s}_4^2\overline{\sigma}_4^2}{\beta} + \frac{6\overline{s}_4^4}{\beta^2}.$$

In the same way

$$\big(L'\big)^2(\mathbb{P},\pi_\theta,\theta_\star) \leq 8\|\theta-\theta_\star\|^2\sqrt{\kappa}\,\sigma_4^2 + 2\|\theta-\theta_\star\|^4\kappa$$
$$+ \frac{12s_4^2\sqrt{\kappa}}{\beta}\|\theta-\theta_\star\|^2 + \frac{8s_4^2\sigma_4^2}{\beta} + \frac{6s_4^4}{\beta^2}.$$

Moreover

$$R'(\pi_\theta,\theta_\star) = L'(\mathbb{P},\pi_\theta,\theta_\star) = \frac{d}{\beta} + R'(\theta,\theta_\star),$$
$$\text{and}\quad L'(\overline{\mathbb{P}},\pi_\theta,\theta_\star) = \frac{\overline{d}}{\beta} + L'(\overline{\mathbb{P}},\theta,\theta_\star),\ \text{where}$$
$$\overline{d} = \int \|x\|^2\,\mathrm{d}\overline{\mathbb{P}}(x).$$

PROPOSITION 3.1 *With probability at least* $1-\epsilon$*, for any* $\theta\in\Theta_\star$*,*

$$\lambda(\kappa+\overline{\kappa})R'(\theta,\theta_\star)^2$$
$$- R'(\theta,\theta_\star)\left\{1-\lambda\left[4\big(\sigma_4^2\sqrt{\kappa}+\overline{\sigma}_4^2\sqrt{\overline{\kappa}}\big)+\frac{6}{\beta}\big(s_4^2\sqrt{\kappa}+\overline{s}_4^2\sqrt{\overline{\kappa}}\big)\right]-\frac{\beta}{2n\lambda}\right\}$$
$$+ \frac{\log(\epsilon^{-1})}{n\lambda} + \frac{4\lambda}{\beta}\big(s_4^2\sigma_4^2+\overline{s}_4^2\overline{\sigma}_4^2\big) + \frac{3\lambda}{\beta^2}\big(s_4^4+\overline{s}_4^4\big) + \frac{\overline{d}-d}{\beta} + L'(\overline{\mathbb{P}},\theta,\theta_\star) \geq 0$$

The random coefficients $\overline{\kappa}$, $\overline{\sigma}_4$, $\overline{s}_4$ and $\overline{d}$, according to the weak law of large numbers, converge to their deterministic counterparts κ, σ_4, s_4 and d. That it is the case for $\overline{\kappa}$ is not completely straightforward, since its defintion involves a supremum, but can be seen from the inequality

$$\begin{aligned}
\big|\overline{\kappa}-\kappa\big| &\leq \sup\Big\{\Big|\Big(\sum_{i=1}^d \theta_i x_i\Big)^4 \,\mathrm{d}(\overline{\mathbb{P}}-\mathbb{P})(x)\Big|, \theta\in\mathbb{R}^d, \|\theta\|=1\Big\} \\
&\leq \sup\Big\{\sum_{\substack{\alpha\in\mathbb{N}^{[\![1,d]\!]},\\ |\alpha|=4}} \frac{|\alpha|!}{\prod_{i=1}^d \alpha_i!}\prod_{i=1}^d|\theta_i|^{\alpha_i}\Big|\int\prod_{i=1}^d x_i^{\alpha_i}\,\mathrm{d}\big(\overline{\mathbb{P}}-\mathbb{P}\big)(x)\Big|, \theta\in\mathbb{R}^d, \|\theta\|=1\Big\} \\
&\leq \sup\Big\{\Big(\sum_{i=1}^d|\theta_i|\Big)^4, \theta\in\mathbb{R}^d, \|\theta\|=1\Big\} \\
&\quad\times\max\Big\{\Big|\int\prod_{i=1}^d x_i^{\alpha_i}\,\mathrm{d}\big(\overline{\mathbb{P}}-\mathbb{P}\big)(x)\Big|, \alpha\in\mathbb{N}^{[\![1,d]\!]}, |\alpha|=4\Big\} \\
&\leq d^2\max\Big\{\Big|\int\prod_{i=1}^d x_i^{\alpha_i}\,\mathrm{d}\big(\overline{\mathbb{P}}-\mathbb{P}\big)(x)\Big|, \alpha\in\mathbb{N}^{[\![1,d]\!]}, |\alpha|=4\Big\}.
\end{aligned}$$

Here $[\![1,d]\!] \stackrel{\text{def}}{=} [1,d]\cap\mathbb{N}$ is an integer interval and $|\alpha|=\sum_{i=1}^d\alpha_i$. Thus the weak law of large numbers applied to each term of the last maximum shows that

$$\lim_{n\to+\infty}\mathbb{P}^{\otimes n}\Big\{|\overline{\kappa}-\kappa|\geq\eta\Big\}=0, \qquad \eta\in\mathbb{R}_+^*.$$

Thus, putting

$$\begin{aligned}
\overline{m}=\frac{1}{2}\max\Big\{&\overline{\sigma}_4^2\sqrt{\overline{\kappa}}-\sigma_4^2\sqrt{\kappa}, \overline{s}_4^2\sqrt{\overline{\kappa}}-s_4^2\sqrt{\kappa}, \\
&\overline{\kappa}-\kappa, \overline{s}_4^2\overline{\sigma}_4^2-s_4^2\sigma_4^2, \overline{s}_4^4-s_4^4, \overline{d}-d\Big\},
\end{aligned}$$

we see that for any $\eta>0$,

$$\lim_{n\to+\infty}\mathbb{P}^{\otimes n}\big(\,\overline{m}\geq\eta\,\big)=0.$$

Choosing, for some small enough value of $\eta>0$,

$$\begin{aligned}
\lambda &= \frac{1}{32\big(\sigma_4^2\sqrt{\kappa}+\eta\big)}, \\
\beta &= \frac{n\lambda}{3},
\end{aligned}$$

we see that

PROPOSITION 3.2 *There is some integer N such that for any $n\geq N$ and any $\epsilon>0$, with probability at least $1-\epsilon$, for any $\theta\in\Theta_\star$,*

$$aR'(\theta,\theta_\star)^2-\frac{1}{2}R'(\theta,\theta_\star)+c+L'(\overline{\mathbb{P}},\theta,\theta_\star)\geq 0,$$

where

$$a = \frac{\sqrt{\kappa}}{15\sigma_4^2},$$
$$c = \frac{\big[25s_4^2 + 33\sqrt{\kappa}\log(\epsilon^{-1})\big]\sigma_4^2}{n}.$$

Let $\widehat{\theta}$ be the ordinary least square estimator on $\Theta_\star$, defined by the equation

$$L(\overline{\mathbb{P}}, \widehat{\theta}) = \inf\Big\{L(\overline{\mathbb{P}}, \theta); \theta \in \Theta_\star\Big\},$$

so that $L'(\overline{\mathbb{P}}, \widehat{\theta}, \theta_\star) \leq 0$. Let us consider the discriminant $\Delta = 1/4 - 4ac$ of the quadratic equation $ax^2 - x/2 + c = 0$ and let us assume that $\Delta > 0$, or equivalently that

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

Let us consider $\widetilde{\theta} \in \Theta_\star$ satisfying

$$R'(\widetilde{\theta}, \theta_\star) \leq \frac{1}{4a} \quad \text{and}$$
$$L(\overline{\mathbb{P}}, \widetilde{\theta}) = \inf\left\{L(\overline{\mathbb{P}}, \theta); \theta \in \Theta_\star, R'(\theta, \theta_\star) \leq \frac{1}{4a}\right\}.$$

We see from the previous proposition that $R'(\widetilde{\theta}, \theta_\star) \leq \dfrac{1}{4a} - \dfrac{\Delta}{2a}$. Let us now consider for any $\alpha \in [0,1]$,

$$\widetilde{\theta}_\alpha = (1-\alpha)\widetilde{\theta} + \alpha\widehat{\theta}.$$

From the convexity of $\theta \mapsto L(\overline{\mathbb{P}}, \theta)$, we deduce that

$$L(\overline{\mathbb{P}}, \widetilde{\theta}_\alpha) \leq (1-\alpha)L(\overline{\mathbb{P}}, \widetilde{\theta}) + \alpha L(\overline{\mathbb{P}}, \widehat{\theta}) \leq L(\overline{\mathbb{P}}, \theta_\star).$$

Thus either $R'(\widetilde{\theta}_\alpha, \theta_\star) \leq \dfrac{1}{4a} - \dfrac{\sqrt{\Delta}}{2a}$ or $R'(\widetilde{\theta}_\alpha, \theta_\star) \geq \dfrac{1}{4a} + \dfrac{\sqrt{\Delta}}{2a}$. Since $R'(\widetilde{\theta}_0, \theta_\star) \leq \dfrac{1}{4a} - \dfrac{\sqrt{\Delta}}{2a}$ and $\alpha \mapsto R'(\widetilde{\theta}_\alpha, \theta_\star)$ is continuous, this proves that $R'(\widetilde{\theta}_1, \theta_\star) \leq \dfrac{1}{4a} - \dfrac{\Delta}{2a}$. As $\widetilde{\theta}_1 = \widehat{\theta}$, we have proved the following proposition.

PROPOSITION 3.3 *There is some N (depending on $\mathbb{P}$) such that for any $n \geq N$ and $\epsilon \in]0,1]$ such that*

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

with probability according to the sample distribution $\mathbb{P}^{\otimes n}$ *at least* $1-\epsilon$*,*

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

PROOF. Since $\sqrt{\Delta} = \dfrac{\sqrt{1-16ac}}{2} \geq \dfrac{1-16ac}{2}$, $\dfrac{1}{4a} - \dfrac{\sqrt{\Delta}}{2a} \leq 4c$. □

Exercise 6 (Another set of hypotheses) *Let us put*

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

Show that there is N *such that for any* $n > N$ *and* $\epsilon \in]0,1]$ *such that*

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

with probability at least $1-\epsilon$*,*

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

Thus, asymptotically, we obtain a $\dfrac{d}{n}$ *rate of convergence in this situation, under the only assumptions that*

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

$$\textit{and} \qquad \int \|x\|^4 \,\mathrm{d}\mathbb{P}(x) < +\infty.$$

Exercise 7 *Let us make the same hypotheses as in the previous exercise, and consider now the estimator* $\widehat{\theta}$ *of Proposition 2.11 (page 22). Let us take*

$$\lambda^{-1} = 8\big(4\sigma_2^2 + \kappa\Delta^2\big) \quad \textit{and} \quad \frac{\lambda}{\beta} = \frac{4}{n},$$

and let us assume that

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

Show that with probability at least $1-\epsilon$*,*

$$R'(\widehat{\theta},\theta_\star) \leq \frac{\big[B_1 d + B_2\log(4/\epsilon)\big]\sigma_2^2}{n} + \frac{B_3\kappa\log(4/\epsilon)\Delta^2}{n} + B_4\zeta,$$

where

$$B_1 = \left(1 + \frac{2^7 c s_4^4}{n}\right)\left[2^8 + 2^6 \times 3\, c + 2^{11}(7+c)\left(4 + \frac{\kappa\Delta^2}{\sigma_2^2}\right)\frac{s_4^4}{dn}\right] + \frac{2^{14} c^2 s_4^6 \Delta^2}{n\sigma_2^2},$$

and the other constants are as in Proposition 2.14 (page 27).

REFERENCES

[1] P. Alquier. Iterative feature selection in least square regression estimation. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 2008.

[2] P. Alquier. PAC-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

[3] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 40(6):685–736, 2004.

[4] J.-Y. Audibert and O. Catoni. Robust linear least squares regression, 2010. arXiv.

[5] J.-Y. Audibert and O. Catoni. Robust linear regression through PAC-Bayesian truncation, 2010. arXiv.

[6] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. Pages 1–269.

[7] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007. Pages i-xii, 1-163.

[8] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study, 2010. arXiv:1009.2048v1 [math.ST].

[9] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, New York, NY, USA, 2009. ACM.

[10] J. Langford and J. Shawe-Taylor. PAC-bayes & margins. In *Advances in Neural Information Processing Systems*, pages 423–430, 2002.

[11] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*. Morgan Kaufmann, 1999.

[12] D. A. McAllester. PAC-Bayesian stochastic model selection. *Mach. Learn.*, 51(1):5–21, April 2003.

[13] David Mcallester. Simplified pac-bayesian margin bounds. In *In COLT*, pages 203–215, 2003.

[14] M. Seeger. PAC-Bayesian generalization error bounds for gaussian process classification. Informatics report series EDI-INF-RR-0094, Division of Informatics, University of Edinburgh, 2002.