PAC-Bayes learning bounds

Olivier Catoni

Lecture notes for the IFCAM Summer School
on Applied Mathematics
Indian Institute of Science,
Bangalore
July, 24-25, 2014

As a prerequisite, we will recall Chernoff's deviation bound for sums of independent random variables and its consequences. We will then proceed to PAC-Bayes uniform bounds on sums of independent random variables depending on a parameter. These bounds apply to parameter estimators based on the minimization of an empirical risk function. In the third section, we will focus on supervised classification, and in the last section more specifically on Support Vector Machines.

1. Deviations of sums of independent random variables

Let X_i, $1 \leq i \leq n$ be a sample of independent real valued random variables, and let

$$M \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} X_i$$

be their empirical mean. We will study the deviations of M with respect to its mean (under hypotheses ensuring that it exists). When this is meaningful, we will set

$$m \stackrel{\text{def}}{=} \mathbb{E}(M) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i).$$

Consider the moment generating functions

$$\psi_i(\lambda) = \log\Big\{ \mathbb{E}\big[\exp(\lambda X_i)\big] \Big\},$$

$$\psi(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \psi_i(\lambda).$$

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.

These are convex functions, taking their values in $\mathbb{R} \cup \{+\infty\}$, and sending zero to zero. Let us consider the dual function

$$\psi^*(x) = \sup_{\lambda \in \mathbb{R}_+} \lambda x - \psi(\lambda) \in \mathbb{R}_+ \cup \{+\infty\}.$$

PROPOSITION 1.1 *The deviations of the empirical mean M are such that*

$$\mathbb{P}\big(M \geq x\big) \leq \exp\big[-n\psi^*(x)\big].$$

PROOF. Let us remark that no hypothesis is needed, because m is not involved explicitly in this result and because $\psi^*(x) = 0$ is possible, in which case the result is trivial. The proof is based on the fact that for any $z \in \mathbb{R}_+$, $\mathbb{1}(z \geq 1) \leq z$. Indeed

$$\begin{aligned}\mathbb{P}\big(M \geq x\big) = \mathbb{E}\Big\{\mathbb{1}\big[\exp\big(n\lambda(M-x)\big) \geq 1\big]\Big\} \\ &\leq \mathbb{E}\big[\exp\big(n\lambda(M-x)\big)\big] \\ &= \exp\big\{n\big[\psi(\lambda) - \lambda x\big]\big\}, \qquad \lambda \in \mathbb{R}_+.\end{aligned}$$

Consequently,

$$\mathbb{P}\big(M \geq x\big) \leq \inf_{\lambda \in \mathbb{R}_+} \exp\big\{n\big[\psi(\lambda) - \lambda x\big]\big\} = \exp\big(-n\psi^*(x)\big).$$

□

PROPOSITION 1.2 *Let us put $\Lambda_i = \sup\big\{\lambda \in \mathbb{R}_+ : \psi_i(\lambda) < +\infty\big\}$ and $\Lambda = \min\big\{\Lambda_1, \ldots, \Lambda_n\big\}$. For any $\lambda \in [0, \Lambda_i[$, $\psi_i(\lambda) < +\infty$ and the function ψ_i is of class $\mathcal{C}^\infty$ on the interval $]0, \Lambda_i[$. If, moreover, $\mathbb{E}\big(|X_i|^k\big) < \infty$, the function ψ_i is of class $\mathcal{C}^k$ on $[0, \Lambda_i[$.*

PROOF. Let us put $\varphi_i(\lambda) = \mathbb{E}\big[\exp(\lambda X_i)\big]$. Let us consider some $\lambda \in [0, \Lambda_i[$. By definition of Λ_i, there is $\beta \in]\lambda, \Lambda_i[$ such that $\psi_i(\beta) < \infty$, and therefore, $\varphi_i(\beta) < \infty$. From Jensen's inequality,

$$+\infty > \mathbb{E}\big[\exp(\beta X_i)\big] = \mathbb{E}\Big\{\big[\exp(\lambda X_i)\big]^{\beta/\lambda}\Big\} \geq \Big\{\mathbb{E}\big[\exp(\lambda X_i)\big]\Big\}^{\beta/\lambda},$$

proving that $\varphi_i(\lambda) < \infty$, and therefore that $\psi_i(\lambda) < \infty$. Let us remark that

$$X_i^{j-1} \exp(\beta X_i) = X_i^{j-1} \exp(\alpha X_i) + \int_\alpha^\beta X_i^j \exp(\lambda X_i)\, \mathrm{d}\lambda,$$
$$0 < \alpha < \beta < \Lambda_i, \quad j \geq 1.$$

Moreover

$$\mathbb{E}\Big\{\sup_{\lambda\in[\alpha,\beta]}\big|X_i^j\exp\big(\lambda X_i\big)\big|\Big\}<\infty$$

Indeed, let us consider $\gamma\in]\beta,\Lambda_i[$ and

$$C_1=\sup_{x\in\mathbb{R}_+}x^j\exp\big[-(\gamma-\beta)x\big],$$
$$C_2=\sup_{x\in\mathbb{R}_+}x^j\exp(-\alpha x).$$

$$\big|X_i^j\exp(\lambda X_i)\big|\leq\begin{cases}C_1\exp(\gamma X_i), & X_i\geq 0,\\ C_2, & X_i\leq 0.\end{cases}$$

Consequently,

$$\mathbb{E}\Big\{\sup_{\lambda\in[\alpha,\beta]}\big|X_i^j\exp(\lambda X_i)\big|\Big\}\leq C_1\mathbb{E}\big[\exp(\gamma X_i)\big]+C_2<\infty.$$

We can therefore use Fubini's theorem and write

$$\begin{aligned}\mathbb{E}\big[X_i^{j-1}\exp(\beta X)\big]&=\mathbb{E}\big[X_i^{j-1}\exp(\alpha X_i)\big]+\mathbb{E}\bigg(\int_\alpha^\beta X_i^j\exp(\lambda X_i)\,\mathrm{d}\lambda\bigg)\\&=\mathbb{E}\big[X_i^{j-1}\exp(\alpha X_i)\big]+\int_\alpha^\beta\mathbb{E}\big[X_i^j\exp(\lambda X_i)\big]\,\mathrm{d}\lambda.\end{aligned}$$

From Lebesgue's dominated convergence theorem, $\lambda\mapsto\mathbb{E}\big(X_i^j\exp(\lambda X_i)\big]:[\alpha,\beta]\to\mathbb{R}$ is continuous, therefore $\beta\mapsto\mathbb{E}\big[X_i^{j-1}\exp(\beta X_i)\big]$ is of class $\mathcal{C}^1$, and its derivative is $\mathbb{E}\big[X_i^j\exp(\beta X_j)\big]$, therefore $\beta\mapsto\mathbb{E}\big[\exp(\beta X_i)\big]$ is of class $\mathcal{C}^\infty$ on $]0,\Lambda_i[$ and so is ψ_i.

Let us now assume moreover that $\mathbb{E}\big[|X_i|^k\big]<\infty$. In this case we can in the same way prove that for any $\beta\in]0,\Lambda_i[$,

$$\mathbb{E}\Big\{\sup_{\lambda\in[0,\beta]}\big|X_i^j\exp(\lambda X_i)\big|\Big\}\leq C_1\mathbb{E}\big[\exp(\gamma X_i)\big]+\mathbb{E}\big(|X_i|^j\big)<\infty,$$

Consequently, for any $\beta\in[0,\Lambda_i[$,

$$\mathbb{E}\big[X_i^{j-1}\exp(\beta X_i)\big]=\mathbb{E}\big(X_i^{j-1}\big)+\int_0^\beta\mathbb{E}\big[X_i^j\exp(\lambda X_i)\big]\,\mathrm{d}\lambda,\qquad 1\leq j\leq k,$$

so that φ_i and ψ_i are of class $\mathcal{C}^k$ on the interval $[0,\Lambda_i[$. □

PROPOSITION 1.3 *Let us assume that* $\mathbb{E}\big(X_i^2\big) < \infty$ *and that* $\Lambda_i > 0$. *The second derivative of* ψ_i *can be seen as a variance:*

$$\psi_i''(\lambda) = \frac{\mathbb{E}\big[X_i^2 \exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]} - \left(\frac{\mathbb{E}\big[X_i \exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]}\right)^2, \qquad 0 \leq \lambda < \Lambda_i,$$

moreover

$$\psi_i(\lambda) = \lambda \mathbb{E}(X_i) + \int_0^\lambda (\lambda - \alpha)\psi_i''(\alpha)\,\mathrm{d}\alpha, \qquad 0 \leq \lambda < \Lambda_i.$$

PROOF. According to the previous proposition, ψ_i is of class $\mathcal{C}^2$ on $[0, \Lambda_i[$ and

$$\psi_i'(\lambda) = \frac{\mathbb{E}\big[X_i \exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]}.$$

Taking one more derivative, we get the expression of ψ_i'' given in the proposition. Let us consider the random variable Y_i distributed according to the probability measure

$$\mathbb{P}\big(Y_i \in A\big) = \frac{\mathbb{E}\big[\mathbb{1}(X_i \in A)\exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]},$$

for any Borel set A. For any measurable function f satisfying $\mathbb{E}\big[|f(X_i)|\exp(\lambda X_i)\big] < \infty$, it is such that

$$\mathbb{E}\big[f(Y_i)\big] = \frac{\mathbb{E}\big[f(X_i)\exp(\lambda X_i)\big]}{\mathbb{E}\big[\exp(\lambda X_i)\big]}.$$

This shows that $\psi_i''(\lambda) = \mathbb{E}\big(Y_i^2\big) - \mathbb{E}\big(Y_i\big)^2$ can indeed be expressed as a variance. The second part of the proposition is obtained using a Taylor expansion with integral reminder, namely

$$\psi_i(\lambda) = \psi_i(0) + \lambda\psi_i'(0) + \int_0^\lambda (\lambda - \alpha)\psi_i''(\alpha)\,\mathrm{d}\alpha.$$

□

PROPOSITION 1.4 *Let us assume that* $\Lambda > 0$ *and that* $\mathbb{E}\big(X_i^2\big) < \infty$, $1 \leq i \leq n$. *Let us put*

$$\overline{V}(\lambda) \stackrel{\text{def}}{=} \frac{2}{\lambda^2}\big[\psi(\lambda) - \lambda m\big] = \frac{2}{\lambda^2}\int_0^\lambda (\lambda - \alpha)\psi''(\alpha)\,\mathrm{d}\alpha, \qquad 0 \leq \lambda < \Lambda$$

$$V(\lambda) \stackrel{\text{def}}{=} \sup_{\beta\in[0,\lambda]} \overline{V}(\beta),$$

$$v \stackrel{\text{def}}{=} V(0) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big\{ \big[X_i - \mathbb{E}(X_i)\big]^2 \Big\}$$

let us remark that V is a continuous non decreasing function with values in $\mathbb{R}+\cup\{+\infty\}$. Under these hypotheses

$$\mathbb{P}\big(M \geq m+x\big) \leq \exp\left(-\frac{nx^2}{2V(x/v)}\right),$$

$$\mathbb{P}\left(M \geq m + \sqrt{\frac{2\log(\epsilon^{-1})}{n} V\left(\sqrt{\frac{2\log(\epsilon^{-1})}{nv}}\right)}\ \right) \leq \epsilon.$$

Proof. For any $0 \leq \beta \leq \lambda$,

$$\psi^*(m+x) \geq \beta x - \frac{\beta^2}{2} V(\lambda),$$

so that

$$\mathbb{P}\big(M \geq m+x\big) \leq \exp\left[-n\left(\beta x - \frac{\beta^2}{2}V(\lambda)\right)\right].$$

We get the first part of the inequality, choosing $\lambda = x/v$ and $\beta = x/V(\lambda) \leq \lambda$. To get the second part, let us put $\epsilon = \exp\left[-n\left(\beta x - \frac{\beta^2}{2}V(\lambda)\right)\right]$. We get first that

$$\mathbb{P}\left(M \geq m + \frac{\beta}{2}V(\lambda) + \frac{\log(\epsilon^{-1})}{n\beta}\right) \leq \epsilon,$$

and from there, we conclude, choosing $\lambda = \sqrt{\dfrac{2\log(\epsilon^{-1})}{nv}} \geq \beta = \sqrt{\dfrac{2\log(\epsilon^{-1})}{nV(\lambda)}}$.

□

Proposition 1.5 (Bennett's inequality) *Let us assume that $\mathbb{E}\big(X_i^2\big) < \infty$ and that $X_i \leq \mathbb{E}(X_i) + b$, $1 \leq i \leq n$. Let us introduce the function*

$$h(u) = (1+u)\log(1+u) - u \geq \frac{u^2}{2(1+u/3)}, \qquad u \in \mathbb{R}_+.$$

Under these hypotheses,

$$\mathbb{P}\big(M \geq m+x\big) \leq \exp\left[-\frac{nv}{b^2}h\left(\frac{bx}{v}\right)\right] \leq \exp\left(-\frac{nx^2}{2v+\frac{2bx}{3}}\right),$$

$$\mathbb{P}\left(M \geq m + \sqrt{\frac{2v\log(\epsilon^{-1})}{n}}\left(1 - \frac{b}{3v}\sqrt{\frac{2v\log(\epsilon^{-1}}{n}}\right)^{-1/2}\right) \leq \epsilon.$$

PROOF. Let us remark first that for any $\lambda \in \mathbb{R}+$,

$$\psi^*(m+x) \geq \lambda(x+m) - \frac{1}{n}\sum_{i=1}^{n} \log\big[\mathbb{E}\big(\exp(\lambda X_i)\big)\big] \\ = \lambda x - \frac{1}{n}\sum_{i=1}^{n} \log\Big\{\mathbb{E}\big[\exp\big(\lambda(X_i - m_i)\big)\big]\Big\},$$

where $m_i \stackrel{\text{def}}{=} \mathbb{E}(X_i)$. We can then write

$$\mathbb{E}\big[\exp\big(\lambda(X_i - m_i)\big)\big] - 1 = \mathbb{E}\big[\exp\big(\lambda(X_i - m_i)\big) - 1 - \lambda(X_i - m_i)\big] \\ = \mathbb{E}\big[\lambda^2 (X_i - m_i)^2 g\big(\lambda(X_i - m_i)\big)\big]$$

where $g(y) = y^{-2}\big(\exp(y) - 1 - y\big)$. The function g is non decreasing on $\mathbb{R}$. Using a Taylor expansion of order two of the function $z \mapsto \exp(yz)$ between 0 and 1, we see that it can indeed be written as

$$g(y) = \int_0^1 (1-z)\exp(yz)\,\mathrm{d}z, \qquad y \in \mathbb{R}.$$

Consequently

$$\mathbb{E}\big[\lambda^2 (X_i - m_i)^2 g\big(\lambda(X_i - m_i)\big)\big] \leq \mathbb{E}\big[\lambda^2 (X_i - m_i)^2 g(\lambda b)\big], \qquad 1 \leq i \leq n.$$

Therefore,

$$\log\Big\{\mathbb{E}\big[\exp\big(\lambda(X_i - m_i)\big)\big]\Big\} \leq \lambda^2 g(\lambda b)\mathbb{E}\big[(X_i - m_i)^2\big].$$

Thus,

$$\psi^*(m+x) \geq \lambda x - \lambda^2 v g(\lambda b) = \lambda x - \frac{v}{b^2}\big(\exp(\lambda b) - 1 - \lambda b\big).$$

Let us choose $\lambda = b^{-1}\log\left(1 + \dfrac{bx}{v}\right)$, to get

$$\psi^*(x) \geq \frac{v}{b^2} h\left(\frac{bx}{v}\right).$$

Let us show now that $h(u) \geq \dfrac{u^2}{2(1+u/3)}$, $u > -1$. Let us compute the derivatives of h, $h'(u) = \log(1+u)$, $h''(u) = 1/(1+u)$, and then the derivatives

of $f(u) = (1+u/3)h(u) - u^2/2$. We get $f'(u) = h'(u)(1+u/3) + h(u)/3 - u$. Thus $f'(0) = 0$ and

$$f''(u) = h''(u)(1+u/3) + 2h'(u)/3 - 1 = \frac{1+u/3}{1+u} + \frac{2}{3}\log(1+u) - 1$$
$$= \frac{2}{3}\log(1+u) - \frac{2u}{3(1+u)} = \frac{2h(u)}{3(1+u)} \geq 0, \qquad u > -1.$$

The convex function f, sending zero to zero, with a null first derivative at zero, is therefore everywhere non negative.

Let us put $\epsilon = \exp\left(-\frac{nx^2}{2v + \frac{2bx}{3}}\right)$. We get

$$x^2 = \frac{2v\log(\epsilon^{-1})}{n}\left(1 + \frac{bx^2}{3vx}\right)$$
$$\leq \frac{2v\log(\epsilon^{-1}}{n}\left(1 + \frac{bx^2}{3v}\left(\frac{2v\log(\epsilon^{-1})}{n}\right)^{-1/2}\right).$$

We deduce that

$$x^2 \leq \frac{2v\log(\epsilon^{-1})}{n}\left(1 - \frac{b}{3v}\sqrt{\frac{2v\log(\epsilon^{-1})}{n}}\right)^{-1},$$

proving the second inequality of the proposition. □

Proposition 1.6 (Hoeffding's inequality) *Let us assume that $a_i \leq X_i \leq b_i$, $1 \leq i \leq n$. In this case,*

$$\mathbb{P}\big(M \geq m + x\big) \leq \exp\left(-\frac{2n^2x^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$
$$\mathbb{P}\left(M \geq m + \sqrt{\frac{\sum_{i=1}^n (b_i - a_i)^2 \log(\epsilon^{-1})}{2n^2}}\right) \leq \epsilon.$$

Proof. The second derivative of ψ_i is the variance of a random variable taking its values in the interval $[a_i, b_i]$. It cannot therefore be larger than $(b_i - a_i)^2/4$. Consequently, $\psi(\lambda) \leq \lambda m + \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2$, and therefore $\psi^*(m+x) \geq \frac{2nx^2}{\sum_{i=1}^n (b_i - a_i)^2}$. □

2. PAC-Bayes bounds on the uniform deviations of empirical means with respect to their expectations

Let us consider n independent random variables X_i, $1 \leq i \leq n$ taking their values in a measurable space $\mathcal{X}$. Let us also consider a measurable parameter space Θ and a measurable function $f : \mathcal{X} \times \Theta \to \mathbb{R}$ (which can be seen as a family of functions from $\mathcal{X}$ to $\mathbb{R}$ depending on the parameter θ). Let us assume that

$$\mathbb{E}\big[f(X_i,\theta)^2\big] < +\infty, \qquad \theta \in \Theta, \quad 1 \leq i \leq n,$$

and let us put

$$\begin{aligned}
M(\theta) &= \frac{1}{n}\sum_{i=1}^n f(X_i,\theta),\\
m(\theta) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(X_i,\theta)\big],\\
\psi_i(\lambda,\theta) &= \log\Big\{\mathbb{E}\exp\big[\lambda f(X_i,\theta)\big]\Big\},\\
\psi(\lambda,\theta) &= \frac{1}{n}\sum_{i=1}^n \psi_i(\lambda,\theta),\\
\Lambda &= \sup\big\{\lambda : \psi(\lambda,\theta) < \infty, \theta \in \Theta\big\}
\end{aligned}$$

Proposition 2.1 *Let us assume that $\Lambda > 0$. Let $\nu \in \mathcal{M}_+^1(\Theta)$ be a reference measure on the parameter space Θ. For any $\lambda \in [0,\Lambda[$,*

$$\mathbb{E}\Bigg[\exp\Bigg(\sup\Bigg\{\int_\Theta n\big[\lambda M(\theta) - \psi(\lambda,\theta)\big]\,\mathrm{d}\rho(\theta) - \mathcal{K}(\rho,\nu),$$
$$\rho \in \mathcal{M}_+^1(\Theta), \theta \mapsto \lambda M(\theta) - \psi(\lambda,\theta) \in \mathbb{L}^1(\rho), \mathcal{K}(\rho,\nu) < \infty\Bigg\}\Bigg)\Bigg] \leq 1.$$

Consequently, with probability at least $1-\epsilon$, for any probability measure $\rho \in \mathcal{M}_+^1(\Theta)$, such that $\theta \mapsto \lambda M(\theta) - \psi(\lambda,\theta) \in \mathbb{L}^1(\rho)$ and $\mathcal{K}(\rho,\nu) < \infty$,

$$\int M(\theta)\,\mathrm{d}\rho(\theta) \leq \frac{1}{\lambda}\int \psi(\lambda,\theta)\,\mathrm{d}\rho(\theta) + \frac{\mathcal{K}(\rho,\nu) + \log(\epsilon^{-1})}{n\lambda}.$$

Proof. From Jensen's inequality, whenever ρ satisfies the hypotheses,

$$\exp\biggl[\int_\Theta n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\nu)\biggr]$$
$$\le\int_\Theta \exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\mathbb{1}\biggl(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)>0\biggr)\biggl(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)\biggr)^{-1}\mathrm{d}\rho(\theta)$$
$$=\int_\Theta \exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\mathbb{1}\biggl(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)>0\biggr)\,\mathrm{d}\nu(\theta)$$
$$\le\int_\Theta \exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\,\mathrm{d}\nu(\theta).$$

We can then apply Fubini's theorem for non negative functions.

$$\mathbb{E}\biggl\{\exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\int_\Theta n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\nu)\biggr]\biggr\}$$
$$\le\mathbb{E}\biggl[\int_\Theta \exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\,\mathrm{d}\nu(\theta)\biggr]$$
$$=\int_\Theta \mathbb{E}\biggl[\exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\biggr]\,\mathrm{d}\nu(\theta)=1.$$

The expectation used in the proposition is taken on a function that may not be measurable, but that is upper bounded by a measurable function with an expectation not greater than one, this is what the proof shows and this is also how the proposition should be understood. The second part of the proposition is a consequence of Markov's inequality. Here again, the involved event may not be a measurable set, it should be understood that it contains a measurable set of probability at least equal to $1-\epsilon$. □

$$\text{Let us put } m_i(\theta)=\mathbb{E}\bigl[f(X_i,\theta)\bigr],$$
$$v(\theta)=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\Bigl\{\bigl[f(X_i,\theta)-m_i(\theta)\bigr]^2\Bigr\},$$
$$\overline{V}(\lambda,\theta)=\frac{2}{\lambda^2}\bigl[\psi(\lambda,\theta)-\lambda m(\theta)\bigr],$$
$$V(\lambda,\theta)=\sup_{\beta\in[0,\lambda]}\overline{V}(\beta,\theta)$$

and let us assume that $v\overset{\text{def}}{=}\sup_{\theta\in\Theta}v(\theta)<\infty$ and $V(\lambda)\overset{\text{def}}{=}\sup_{\theta\in\Theta}V(\lambda,\theta)<\infty$, $0\le\lambda<\Lambda'$.

Proposition 2.2 *Under the previous hypotheses, for any positive constant c,*

$$\mathbb{E}\Bigg(\sup\Big\{\int_\Theta \big[M(\theta)-m(\theta)\big]\,\mathrm{d}\rho(\theta);$$
$$\rho\in\mathcal{M}_+^1(\Theta), \theta\mapsto M(\theta)-m(\theta)\in\mathbb{L}^1(\rho), \mathcal{K}(\rho,\nu)\le c\Big\}\Bigg)$$
$$\le \inf_{\lambda\in[0,\Lambda'[}\frac{\lambda\overline{V}(\lambda)}{2}+\frac{c}{\lambda n}\le\sqrt{\frac{2c}{n}V\bigg(\sqrt{\frac{2c}{nv}}\bigg)}.$$

In particular, when Θ is a finite set, taking $c=\log(|\Theta|)$, $\rho=\delta_\theta$ and $\nu(\theta)=|\Theta|^{-1}$, $\theta\in\Theta$, we get

$$\mathbb{E}\Big\{\sup_{\theta\in\Theta}\big[M(\theta)-m(\theta)\big]\Big\}\le\sqrt{\frac{2\log\big(|\Theta|\big)}{n}V\bigg(\sqrt{\frac{2\log\big(|\Theta|\big)}{nv}}\bigg)}.$$

Proof. From the proof of the previous proposition, the argument of the expectation to be bounded is not greater than

$$\frac{1}{n\lambda}\log\bigg\{\int\exp\Big[n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\Big]\,\mathrm{d}\nu(\theta)\bigg\}+\frac{\lambda\overline{V}(\lambda)}{2}+\frac{c}{\lambda n},$$

and we conclude with the help of Jensen's inequality. We get in this way the first upper bound $\inf_{\lambda\in[0,\Lambda'[}\frac{\lambda\overline{V}(\lambda)}{2}+\frac{c}{\lambda n}$ that we can weaken to get $\inf_{0\le\lambda\le\beta}\frac{\lambda V(\beta)}{2}+\frac{c}{\lambda n}$. To get the second upper bound, we should choose $\beta=\sqrt{\frac{2c}{nv}}$ and $\lambda=\sqrt{\frac{2c}{nV(\beta)}}\le\beta$. □

Proposition 2.3 *Under the previous hypotheses, for any positive constant c, with probability at least $1-\epsilon$,*

$$\sup\Big\{\int_\Theta \big[M(\theta)-m(\theta)\big]\,\mathrm{d}\rho(\theta);$$
$$\rho\in\mathcal{M}_+^1(\Theta), \theta\mapsto M(\theta)-m(\theta)\in\mathbb{L}^1(\Theta), \mathcal{K}(\rho,\nu)\le c\Big\}$$
$$\le\inf_{\lambda\in[0,\Lambda'[}\frac{\lambda\overline{V}(\lambda)}{2}+\frac{c+\log(\epsilon^{-1})}{\lambda n}\le\sqrt{\frac{2\big[c+\log(\epsilon^{-1})\big]}{n}V\bigg(\sqrt{\frac{2\big[c+\log(\epsilon^{-1})\big]}{nv}}\bigg)}.$$

In particular, when Θ is a finite set, with probability at least $1-\epsilon$

$$\sup_{\theta\in\Theta}\bigl[M(\theta)-m(\theta)\bigr]\leq\sqrt{\frac{2\log\bigl(|\Theta|/\epsilon\bigr)}{n}V\biggl(\sqrt{\frac{2\log\bigl(|\Theta|/\epsilon\bigr)}{nv}}\biggr)}.$$

PROOF. This is a direct consequence of the second part of Proposition 2.1 (page 8) and of the inequality $\psi(\lambda,\theta)\leq\dfrac{\lambda^2V(\lambda)}{2}+\lambda m(\theta)$. □

PROPOSITION 2.4 *Let us assume that* $\Theta=\mathbb{B}_d=\bigl\{\theta\in\mathbb{R}^d;\|\theta\|\leq 1\bigr\}$ *and that there exist two positive constants* B *and* g *such that*

$$\begin{aligned}\sup_{x\in\mathcal{X}}f(x,\theta)-\inf_{x\in\mathcal{X}}f(x,\theta)&\leq B, &&\theta\in\mathbb{B}_d,\\ |f(x,\theta)-f(x,\theta')|&\leq g\|\theta-\theta'\|, &&x\in\mathcal{X},\quad\theta,\theta'\in\mathbb{B}_d.\end{aligned}$$

Let us consider the value of the parameter where the empirical risk takes its minimum value

$$\widehat{\theta}\in\arg\min_{\theta\in\mathbb{B}_d}M(\theta).$$

With probability at least $1-\epsilon$,

$$m(\widehat{\theta})\leq\inf_{\theta\in\mathbb{B}_d}m(\theta)+B\left\{\sqrt{\frac{d}{2n}\log\biggl(1+\frac{4g}{B}\sqrt{\frac{2n}{d}}\biggr)+\frac{\log\bigl(2/\epsilon\bigr)}{2n}}\right.$$
$$\left.+\sqrt{\frac{d}{8n}}+\sqrt{\frac{\log\bigl(2/\epsilon\bigr)}{2n}}\right\}.$$

Under those simple hypotheses, we see that the quality of the estimation of $\inf_{\theta\in\Theta}m(\theta)$ by $\widehat{\theta}$ depends on the dimension d of the parameter space, and more precisely of the ratio d/n between this dimension and the sample size.

PROOF. Let us start by extending the domain of f to $\mathbb{R}^d$, putting

$$f(x,\theta)=f(x,\theta/\|\theta\|),\qquad\theta\in\mathbb{R}^d\setminus\mathbb{B}_d.$$

Let $\delta>0$ be a positive real parameter to be set later and ν the uniform measure on the ball $(1+\delta)\mathbb{B}_d$ of radius $1+\delta$. Let us consider, for all $\theta\in\mathbb{B}_d$, the uniform probability measure ρ_θ on the ball $\theta+\delta\mathbb{B}_d$ centered at θ and of radius δ. As the volume of a ball in $\mathbb{R}^d$ is proportional to its radius raised to the power d, we see that

$$\mathcal{K}(\rho_\theta,\nu)=d\log\biggl(\frac{1+\delta}{\delta}\biggr),\qquad\theta\in\mathbb{B}_d.$$

From the previous proposition and Hoeffding's inequality, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{B}_d$,

$$\int m(\theta')\,\mathrm{d}\rho_\theta(\theta') \leq \int M(\theta')\,\mathrm{d}\rho_\theta(\theta') + B\sqrt{\frac{d\log\bigl(1+\delta^{-1}\bigr)+\log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

We deduce, still with probability at least $1-\epsilon$, that

$$m(\widehat{\theta}) \leq M(\widehat{\theta}) + 2g\delta + B\sqrt{\frac{d\log\bigl(1+\delta^{-1}\bigr)+\log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

Let $\theta_* \in \mathbb{B}_d$, such that $m(\theta_*) = \inf_{\theta\in\mathbb{B}_d} m(\theta)$ (it exists, since $\theta \mapsto m(\theta)$ is continuous on the compact set $\mathbb{B}_d$). With probability at least $1-\epsilon$

$$M(\theta_*) \leq m(\theta_*) + B\sqrt{\frac{\log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

According to the definition of the estimator $\widehat{\theta}$, $M(\widehat{\theta}) \leq M(\theta_*)$. Consequently, with probability at least $1-2\epsilon$,

$$m(\widehat{\theta}) \leq m(\theta_*) + B\left\{\sqrt{\frac{d\log(1+\delta^{-1})+\log\bigl(\epsilon^{-1}\bigr)}{2n}} + \sqrt{\frac{\log(\epsilon^{-1})}{2n}}\right\} + 2g\delta.$$

To conclude, it is enough to choose $\delta = \dfrac{B}{4g}\sqrt{\dfrac{d}{2n}}$ and to replace ϵ with $\epsilon/2$. □

PROPOSITION 2.5 *Let us assume that $\Theta = \mathbb{R}^d$, and that for some measurable function $(x,\theta) \mapsto \nabla f(x,\theta) \in \mathbb{R}^d$ and some positive constants g and H,*

$$|f(x,\theta) - f(x,\theta')| \leq g\|\theta-\theta'\|,$$

$$|f(x,\theta') - f(x,\theta) - \langle \nabla f(x,\theta), \theta'-\theta\rangle| \leq \frac{H}{2}\|\theta'-\theta\|^2, \quad x\in\mathcal{X}, \quad \theta,\theta'\in\mathbb{R}^d.$$

Let $\theta_* \in \arg\min_{\theta\in\mathbb{B}_d} m(\theta)$. Let us consider, for any $h>0$, the function

$$\chi(h) = \sup_{\theta\in\mathbb{B}_d} \frac{h}{2}\|\theta-\theta_*\|^2 - m(\theta) + m(\theta_*),$$

Under these hypotheses, the empirical minimizer, $\widehat{\theta} \in \arg\min_{\theta\in\mathbb{B}_d} M(\theta)$ of m on the unit ball is such that with probability at least $1-\epsilon$

$$\|\widehat{\theta}-\theta_*\|^2 \leq \frac{8g^2}{nh^2}\left[\left(\frac{8H}{h}+1\right)d + 2\log\bigl(\epsilon^{-1}\bigr)\right] + \frac{4\chi(h)}{h}$$

$$\text{and } m(\widehat{\theta}) - m(\theta_*) \leq \frac{4g^2}{nh}\left[\left(\frac{8H}{h}+1\right)d + 2\log\bigl(\epsilon^{-1}\bigr)\right] + \chi(h).$$

In the case when there is $h > 0$ such that $\chi(h) = 0$, we thus get a convergence speed of order d/n instead of $\sqrt{d/n}$, under stronger hypotheses than in the previous proposition.

Exercice 1 *In the case when* $m(\theta) - m(\theta_*) \geq c\,\|\theta - \theta_*\|^\alpha$, $\theta \in \mathbb{B}_d$, *where* $c > 0$ *and* $\alpha > 2$ *what speed do we get?*

Proof. Let us choose $\rho_\theta = \mathcal{N}(\theta, \beta^{-1}I)$ and $\nu = \rho_{\theta_*}$. Let us remark that $\mathcal{K}(\rho_\theta, \nu) = \dfrac{\beta}{2}\|\theta - \theta_*\|^2$. We are going to apply Proposition 2.1 (page 8) to the function $(x, \theta) \mapsto f(x, \theta_*) - f(x, \theta)$. From Hoeffding's inequality,

$$\log \mathbb{E}\exp\Bigl\{\lambda\bigl[f(X,\theta_*) - f(X,\theta)\bigr]\Bigr\} - \lambda\bigl[m(\theta_*) - m(\theta)\bigr] \leq \frac{\lambda^2 g^2 \|\theta - \theta_*\|^2}{2}$$

Consequently, with probability at least $1 - \epsilon$, for any $\theta \in \mathbb{B}_d$,

$$\int m(\theta')\,\mathrm{d}\rho_\theta(\theta') - m(\theta_*) \leq \int M(\theta')\,\mathrm{d}\rho_\theta(\theta') - M(\theta_*) + \frac{\lambda g^2}{2}\int \|\theta' - \theta_*\|^2\,\mathrm{d}\rho_\theta(\theta') + \frac{\beta\|\theta - \theta_*\|^2}{2n\lambda} + \frac{\log(\epsilon^{-1})}{n\lambda}.$$

Moreover,

$$\int m(\theta')\,\mathrm{d}\rho_\theta(\theta') = m(\theta) + \mathbb{E}\left[\int \Bigl[f(X,\theta') - f(X,\theta) - \langle \nabla f(X,\theta), \theta' - \theta\rangle\Bigr]\,\mathrm{d}\rho_\theta(\theta')\right] \geq m(\theta) - \frac{H}{2}\int \|\theta' - \theta\|^2\,\mathrm{d}\rho_\theta(\theta') = m(\theta) - \frac{Hd}{2\beta}.$$

In the same way, $\displaystyle\int M(\theta')\,\mathrm{d}\rho_\theta(\theta') \leq M(\theta) + \frac{Hd}{2\beta}$. We deduce that with probability at least $1 - \epsilon$, for any $\theta \in \mathbb{B}_d$,

$$m(\theta) - m(\theta_*) \leq M(\theta) - M(\theta_*) + \frac{Hd}{\beta} + \frac{\lambda g^2 d}{2\beta} + \frac{\lambda g^2}{2}\|\theta - \theta_*\|^2 + \frac{\beta\|\theta - \theta_*\|^2}{2n\lambda} + \frac{\log\bigl(\epsilon^{-1}\bigr)}{n\lambda}.$$

We can then use the fact that $m(\theta) - m(\theta_*) \geq \frac{h}{2}\|\theta - \theta_*\|^2 - \chi(h)$ and that by construction $M(\widehat{\theta}) \leq M(\theta_*)$. We conclude that with probability at least $1-\epsilon$

$$\frac{h}{2}\|\widehat{\theta} - \theta_*\|^2 \leq \chi(h) + \frac{d}{\beta}\left(H + \frac{\lambda g^2}{2}\right) + \left(\frac{\lambda g^2}{2} + \frac{\beta}{2n\lambda}\right)\|\widehat{\theta} - \theta_*\|^2 + \frac{\log\bigl(\epsilon^{-1}\bigr)}{n\lambda}.$$

Thus

$$\|\widehat{\theta} - \theta_*\|^2\left(1 - \frac{\lambda g^2}{h} - \frac{\beta}{n\lambda h}\right) \leq \frac{2\chi(h)}{h} + \frac{2d}{\beta h}\left(H + \frac{\lambda g^2}{2}\right) + \frac{2\log\bigl(\epsilon^{-1}\bigr)}{hn\lambda}.$$

Let us then choose $\lambda = \dfrac{h}{4g^2}$ and $\beta = \dfrac{n\lambda h}{4} = \dfrac{nh^2}{16g^2}$. We get

$$\frac{1}{2}\|\widehat{\theta} - \theta_*\|^2 \leq \frac{2\chi(h)}{h} + \frac{32g^2 d}{nh^3}\left(H + \frac{h}{8}\right) + \frac{8g^2\log\bigl(\epsilon^{-1}\bigr)}{nh^2}.$$

This gives the first upper bound of the proposition.

To prove the second upper bound, let us use the fact that $\|\widehat{\theta} - \theta_*\|^2 \leq \dfrac{2}{h}\bigl[m(\widehat{\theta}) - m(\theta_*) + \chi(h)\bigr]$, to obtain

$$m(\widehat{\theta}) - m(\theta_*) \leq \frac{d}{\beta}\left(H + \frac{\lambda g^2}{2}\right) + \left(\frac{\lambda g^2}{2} + \frac{\beta}{2n\lambda}\right)\frac{2}{h}\bigl[m(\widehat{\theta}) - m(\theta_*) + \chi(h)\bigr] + \frac{\log\bigl(\epsilon^{-1}\bigr)}{n\lambda}.$$

We conclude in the same way, replacing λ and β by their values. □

3. PAC-Bayes bounds for supervised classification

In this section, we are given some i.i.d. sample $(W_i)_{i=1}^n \in \mathcal{W}^n$, where $\mathcal{W}$ is a measurable space, and some binary measurable loss function $L : \mathcal{W} \times \Theta \to \{0,1\}$, where Θ is a measurable parameter space. Our aim is to minimize with respect to $\theta \in \Theta$ the expected loss

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

where $\mathbb{P}$ is the marginal distribution of the observed sample $(W_i)_{i=1}^n$. More precisely, assuming that $\mathbb{P}$ is unknown, we would like to find an estimator $\widehat{\theta}(W_{1:n})$ depending on the observed sample $W_{1:n} \stackrel{\text{def}}{=} (W_i)_{i=1}^n$ such that the excess risk

$$\int L(w, \widehat{\theta}) \, \mathrm{d}\mathbb{P}(w) - \inf_{\theta \in \Theta} \int L(w, \theta) \, \mathrm{d}\mathbb{P}(w)$$

is small. The previous quantity is random, since $\widehat{\theta}$ depends on the random sample $W_{1:n}$. Therefore its size can be understood in different ways. Here we will focus on the *deviations* of the excess risk. Accordingly, we will look for estimators providing a small risk with a probability close to one.

A typical example of such a problem is provided by supervised classification. In this setting $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is a finite set, $W_i = (X_i, Y_i)$, where (X_i, Y_i) are input-output pairs, a family of measurable classification rules $\big\{ f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta \big\}$ is considered and the loss function $L(w, \theta)$ is defined as the classification error

$$L\big[(x, y), \theta\big] = \mathbb{1}\big[f_\theta(x) \neq y\big].$$

Accordingly the aim is to minimize the expected classification error

$$\mathbb{P}_{X,Y}\big[f_\theta(X) \neq Y\big]$$

in view of a sample $(X_i, Y_i)_{i=1}^n$ of observations.

The point of view exposed here is a synthesis of the approaches of [9] and [2].

3.1. Deviation bounds for sums of Bernoulli random variables. Given some parameter $\lambda \in \mathbb{R}$, let us consider the (normalized) log-Laplace transform of the Bernoulli distribution :

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

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

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

In the sequel $\overline{\mathbb{P}}$ will be the empirical measure

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

of an i.i.d. sample $(W_i)_{i=1}^n$ drawn from $\mathbb{P}^{\otimes n} \in \mathcal{M}_+^1(\mathcal{W}^n)$ (the set of probability measures on $\mathcal{W}^n$). We will use a short notation for integrals, putting for any $\rho, \pi \in \mathcal{M}_+^1(\Theta)$ and any integrable function $f \in \mathbb{L}_1\big(\mathcal{W} \times \Theta^2, \mathbb{P} \otimes \pi \otimes \rho\big)$

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

so that for instance $L(\mathbb{P}, \rho) = \displaystyle\int L(w, \theta) \, \mathrm{d}\mathbb{P}(w) \mathrm{d}\rho(\theta)$.

Let us recall first Chernoff's bound.

Proposition 3.1 *For any fixed value of the parameter $\theta \in \Theta$, the identity*

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

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

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

Moreover

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

In the same way, the identity

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

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

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

and

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

Before proving this proposition, let us mention some important identity.

Proposition 3.2 *For any probability measures π and ρ defined on the same 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 introduce the notation

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

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

$$\int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi) + \mathcal{K}(\rho,\pi_{\exp(h)}) = \log\big[\textstyle\int \exp(h)\,\mathrm{d}\pi\big] \tag{1}$$

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

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*

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

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

Let us come now to the proof of Proposition 3.1 (page 16). Chernoff's inequality reads

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

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

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

Exercise 2 *Prove this statement in more details. For any integer $k > 1$, consider the event*

$$A_k = \Big\{ \sup_{\lambda \in \mathbb{R}_+} F(\lambda) - k^{-1} > L(\overline{\mathbb{P}}, \theta) \Big\},$$

where $F(\lambda) = \Phi_\lambda\big[L(\mathbb{P}, \theta)\big] - \dfrac{\log(\epsilon^{-1})}{n\lambda}$. *Show that* $\mathbb{P}^{\otimes n}(A_k) \leq \epsilon$ *by choosing some suitable value of* λ. *Remark that* $A_k \subset A_{k+1}$ *and conclude that* $\mathbb{P}^{\otimes n}\big(\cup_k A_k\big) \leq \epsilon$.

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 (1, page 17) 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 $\lambda = \log\left(\dfrac{q^{-1} - 1}{p^{-1} - 1}\right) \in \mathbb{R}_+$, $q = p_\lambda$ 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 3.1 (page 16) is proved in the same way and left as an exercise.*

3.2. PAC-Bayes bounds. We are now going to make Proposition 3.1 uniform with respect to θ. The PAC-Bayes approach to this [6, 7, 8, 4] is to randomize θ, so we will consider now joint distributions on $(W_{1:n}, \theta)$, where the distribution of $W_{1:n}$ 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 $\pi \in \mathcal{M}_+^1(\Theta)$.

*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 [1][page 50] for more details

Proposition 3.3 *Let us introduce the notation*

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

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

$$\int \exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} n\lambda\Bigl\{\Phi_\lambda\bigl[L(\mathbb{P},\rho)\bigr] - L(\overline{\mathbb{P}},\rho)\Bigr\} - \mathcal{K}(\rho,\pi)\biggr]\,\mathrm{d}\mathbb{P}^{\otimes n} \leq 1, \qquad (3)$$

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

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

Proof. The exponential moment inequality (3) is a consequence of equation (1, page 17), showing that

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

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 3.1 (page 16), 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 3.4 *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

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

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

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

where

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

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

$$\begin{aligned} B_{\lfloor \log(n)^2 \rfloor - 1}(q,e,\epsilon) &\leq B(q,e,\epsilon) \stackrel{\text{def}}{=} \\ &\sqrt{\frac{2\overline{v}(q)\big\{e+\log\big[\log(n)^2/\epsilon\big]\big\}}{n}}\cosh\big[\log(n)^{-1}\big] \\ &\qquad + \frac{2\big\{e+\log\big[\log(n)^2/\epsilon\big]\big\}}{n}\cosh\big[\log(n)^{-1}\big]^2, \end{aligned} \tag{4}$$

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

$$q = L(\overline{\mathbb{P}},\rho),$$

$$\delta = \frac{\mathcal{K}(\rho,\pi) + \log\bigl[(m+1)/\epsilon\bigr]}{n},$$
$$\lambda_{\min} = \sqrt{\frac{8\log\bigl[(m+1)/\epsilon\bigr]}{n}},$$
$$\Lambda = \Bigl\{\lambda_{\min}^{1-k/m}, k = 0, \dots, m\Bigr\},$$
$$p = B_\Lambda(q,\delta) = \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\biggl(q + \frac{\delta}{\lambda}\biggr),$$
$$\widehat{\lambda} = \sqrt{\frac{2\delta}{\overline{v}(p)}}.$$

According to equation (2, page 17) applied to Bernoulli distributions, for any $\lambda \in \Lambda$,

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

As moreover $p_\alpha \le p$,

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

As $\overline{v}(p) \le 1/4$ and $\delta \ge \dfrac{\log\bigl[(m+1)/\epsilon\bigr]}{n}$,

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

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

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

$$p - q \le \sqrt{2\delta\overline{v}(p)}\cosh\bigl(t/m\bigr).$$

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

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

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

$$p \le q + \sqrt{2\delta\overline{v}(q)}\cosh\bigl(t/m\bigr),$$

so that in both cases

$$p - q \leq \sqrt{2\delta\overline{v}(q)}\cosh(t/m) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2. \tag{5}$$

Let us consider now the case when $\widehat{\lambda} > 1$. In this case $\overline{v}(p) < 2\delta$, so that

$$p - q \leq \frac{\overline{v}(p)}{2} + \delta \leq 2\delta.$$

In conclusion, applying Proposition 3.3 (page 20) we see that with probability at least $1-\epsilon$, for any posterior distribution ρ,

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

which is precisely the statement to be proved.

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

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

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

$$3\log(n)^2 - 8 + \log(n)\log\Bigl\{8\log\bigl[\log(n)^2-1\bigr]\Bigr\} \geq 0,$$

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

4. 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

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

We are going to follow in this section the approach presented in [5] and [8].

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 [11, 10, 12], 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 $L(w,\theta) = \mathbb{1}\bigl(\langle\theta, \Psi(x)\rangle y \leq 0\bigr)$. 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

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

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.

Let us consider, after [5, 8] 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

$$\begin{aligned}\varphi(x) &= \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp\big(-t^2/2\big)\,\mathrm{d}t, \qquad x \in \mathbb{R} \\ &\leq \min\Big\{\frac{1}{x\sqrt{2\pi}}, \frac{1}{2}\Big\} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R}_+.\end{aligned}$$

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

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

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*

$$M(w, \theta) = \mathbb{1}\big[y\|x\|^{-1}\langle \theta, x \rangle \leq 1\big].$$

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 $\|x\|/\|\theta\|$ from the boundary $\{x' : \langle\theta, x'\rangle = 0\}$, so that the error with margin region is the complement of the open cone $\big\{x \in \mathbb{R}^d\,;\, y\langle\theta, x\rangle > \|x\|\big\}$.

Let us compute the randomized margin error

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

It satisfies the inequality

$$M(w, \pi_\theta) \geq \varphi(-\sqrt{\beta}\,)L(w, \theta) = \big[1 - \varphi\big(\sqrt{\beta}\,\big)\big]L(w, \theta). \tag{6}$$

Applying previous results we obtain

Proposition 4.1 *With probability at least $1 - \epsilon$, for any $\theta \in \mathbb{R}^d$,*

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

where

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

the bound B being defined by equation (4, page 21).

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 $\theta_\star$, $C_1(\widehat{\theta}) \leq C_1(\theta_\star) + \zeta$. On the other hand, with probability at least $1 - \epsilon$

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

Indeed

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

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

PROPOSITION 4.2 *With probability at least* $1-2\epsilon$,

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

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

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

However, C_1 and C_2 are non-convex criterions, and faster minimization algorithms are available for the usual SVM loss function, for which we are going to derive some generalization bounds now. Indeed, let us choose some positive radius R and let us put $\|x\|_R=\max\big\{R,\|x\|\big\}$, so that in the case when $\|x\|\leq R$, $\|x\|_R=R$.

$$M(w,\pi_\theta)=\varphi\big[\sqrt{\beta}\big(y\|x\|^{-1}\langle\theta,x\rangle-1\big)\big]\leq\big(2-y\|x\|_R^{-1}\langle\theta,x\rangle\big)_+ +\varphi(\sqrt{\beta}). \quad (7)$$

To check that this is true, consider the functions

$$f(z)=\varphi\big[\sqrt{\beta}\big(\|x\|^{-1}z-1\big)\big],$$
$$g(z)=\big(2-\|x\|_R^{-1}z\big)_+ +\varphi(\sqrt{\beta}),\qquad z\in\mathbb{R}.$$

Let us remark that they are both non increasing, that f is convex on the interval $z\in\big(\|x\|,\infty\big($ (because φ is convex on $\mathbb{R}_+$), and that $\sup f=\sup\varphi=1$. Since $\|x\|_R\geq\|x\|$, for any $z\in]-\infty,\|x\|]$, $g(z)\geq 1\geq f(z)$. Moreover, $g(2\|x\|_R)=\varphi(\sqrt{\beta})\geq\varphi\big[\sqrt{\beta}\big(2\|x\|^{-1}\|x\|_R-1\big)\big]=f(z)$. Since on the interval $\big(\|x\|,2\|x\|_R\big)$, the function g is linear, the function f is convex and g is not smaller than f at the two ends, this proves that g is not smaller than f on the whole interval. Finally, on the interval $z\in\big(2\|x\|_R,+\infty\big($, the function g is constant and the function f is decreasing, so that on this interval also g is not smaller than f, and this ends the proof of (7), since the three intervals on which $g\geq f$ cover the whole real line.

Using the upper bounds (7) and (6, page 25), and Proposition 3.3 (page 20), we obtain

PROPOSITION 4.3 *With probability at least* $1-\epsilon$*, for any* $\theta \in \mathbb{R}^d$*,*

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

where

$$C_3(\lambda,\theta) = \int \big(2-y\|x\|_R^{-1}\langle\theta,x\rangle\big)_+ \,\mathrm{d}\overline{\mathbb{P}}(x,y) + \frac{\beta\|\theta\|^2}{2n\lambda}.$$

Let us assume now that the patterns x are in a ball, so that $\|x\| \leq R$ almost surely. In this case $\|x\|_R = R$ almost surely. Let us remark that $L(\mathbb{P},\theta) = L(\mathbb{P}, 2R\,\theta)$, and let us make the previous result uniform in $\beta \in \Xi$. This leads to

PROPOSITION 4.4 *Let us assume that* $\|x\| \leq R$ *almost surely. With probability at least* $1-\epsilon$*, for all* $\theta \in \mathbb{R}^d$*,*

$$L(\mathbb{P},\theta) \leq \inf_{\beta\in\Xi} \big[1-\varphi(\sqrt{\beta})\big]^{-1} \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\bigg[2C_4\big(\beta,\lambda,\theta\big) \\ + \varphi(\sqrt{\beta}) + \frac{\log\big(|\Xi|\,|\Lambda|/\epsilon\big)}{n\lambda}\bigg],$$

where

$$C_4(\beta,\lambda,\theta) = \frac{1}{2}\,C_3(\lambda, 2R\,\theta) = \int \big(1-y\langle\theta,x\rangle\big)_+ \,\mathrm{d}\overline{\mathbb{P}}(x,y) + \frac{\beta R^2\|\theta\|^2}{n\lambda},$$

and

$$\Phi_\lambda^{-1}(q) = \frac{1-\exp(-\lambda q)}{1-\exp(-\lambda)} \leq \frac{q}{1-\dfrac{\lambda}{2}}.$$

The loss function $C_4(\lambda,\theta)$ 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 $\inf_\theta C_4(\lambda,\theta)$ for any fixed values of λ and β. Here we get an empirical criterion which could be used to optimize also the values of λ and β, that is to optimize the strength of the regularizing factor $\dfrac{\beta R^2\|\theta\|^2}{n\lambda}$.

In this criterion, $\|\theta\|^{-1}$ can be interpreted as the margin width, that is the minimal distance of x from the separating hyperplane $\{x' : \langle\theta, x'\rangle = 0\}$ beyond which the error term $\big(1-y\langle\theta, x\rangle\big)_+$ vanishes (for data x that are on the right side of the separating hyperplane). The speed of convergence depends on $R^2\|\theta\|^2/n$. For this reason, $R^2\|\theta\|^2$, the square of the ratio between the radius of the ball containing the data and the margin, plays the role of the dimension. The bound does not depend on d, showing that with separating hyperplanes and more generally Support Vector Machines, we can get low error rates while choosing to represent the data in a Reproducing Kernel Hilbert Space with a large, or even infinite, dimension.

We considered so far only linear hyperplanes and data centered around 0. Anyhow, this also covers affine hyperplanes and data contained in a non necessarily centered ball, through a change of coordinates. More precisely, the previous proposition has the following corollary:

Corollary 4.5 *Assume that almost surely $\|x - c\| \leq R$, 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 $\min_{i=1,\dots,n} \langle\theta, x_i\rangle \leq \gamma \leq \max_{i=1,\dots,n} \langle\theta, x_i\rangle$,*

$$\int \mathbb{1}\big[y\big(\langle\theta, x\rangle - \gamma\big) \leq 0\big]\, \mathrm{d}\mathbb{P}(x, y) \leq \inf_{\beta\in\Xi}\big[1 - \varphi(\sqrt{\beta})\big]^{-1} \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\left[2C_5(\beta, \lambda, \theta, \gamma) + \varphi(\sqrt{\beta}) + \frac{\log\big(|\Xi|\,|\Lambda|/\epsilon\big)}{n\lambda}\right],$$

where

$$C_5(\beta, \lambda, \theta, \gamma) = \int \big[1 - y\big(\langle\theta, x\rangle - \gamma\big)\big]_+ \,\mathrm{d}\overline{\mathbb{P}}(x, y) + \frac{4\beta R^2\|\theta\|^2}{n\lambda}.$$

Proof. Let us apply the previous result to $x' = (x - c, R)$, and $\theta' = \big[\theta, R^{-1}\big(\langle\theta, c\rangle - \gamma\big)\big]$. We get that $\|x'\|^2 \leq 2R^2$ and $\|\theta'\|^2 \leq 2\|\theta\|^2$, because almost surely $-\|\theta\|R \leq \operatorname{ess\,inf}\langle\theta, x-c\rangle \leq \gamma - \langle\theta, c\rangle \leq \operatorname{ess\,sup}\langle\theta, x-c\rangle \leq \|\theta\|R$, so that almost surely, for the allowed values of γ, $\big(\langle\theta, c\rangle - \gamma\big)^2 \leq R^2\|\theta\|^2$. This proves that $C_4(\beta, \lambda, \theta') \leq C_5(\beta, \lambda, \theta, \gamma)$, as required to deduce the corollary from the previous proposition. □

References

[1] 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.

[2] 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.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley and Sons, New York, second edition, 2006.

[4] 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.

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

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

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

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

[9] 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.

[10] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, 1982.

[11] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[12] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.