PAC-Bayes learning bounds

Olivier Catoni

CNRS, INRIA – CLASSIC
Département de Mathématiques et Applications,
ENS, 45 rue d'Ulm, 75 230 Paris Cedex 05,
Olivier.Catoni@ens.fr

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

Let X_i, $1 \leq i \leq n$ be n independent real valued random variables.

Let us introduce the empirical mean

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

and its expectation

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

Let us consider the moment generating functions

$$\psi_i(\lambda) = \log\{\mathbb{E}[\exp(\lambda X_i)]\},$$
$$\psi(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\psi_i(n).$$

They are convex, with values in $\mathbb{R} \cup \{+\infty\}$.

Consider the dual function

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

Proposition (Chernoff)

The deviations of the empirical mean M are such that

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

Proof.

We use the fact that $\mathbb{1}\big(z \geq 1\big) \leq z$, for any $z \in \mathbb{R}_+$.

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

Consequently,

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

$\square$

Let us remark that we have also proved that, for any $\lambda \in \mathbb{R}_+$, with probability at least $1 - \epsilon$,

$$M < \frac{\psi(\lambda)}{\lambda} + \frac{\log(\epsilon^{-1})}{n\lambda}.$$

Proposition

Let $\Lambda_i = \sup\{\lambda \in \mathbb{R}_+ : \psi_i(\lambda) < +\infty\}$,
and $\Lambda = \min\{\Lambda_1, \ldots, \Lambda_n\}$.

For any $\lambda \in [0, \Lambda_i[$, $\psi_i(\lambda) < +\infty$ *and the function* ψ_i *is of class* $\mathscr{C}^\infty$ *on the interval* $]0, \Lambda_i[$.

If, moreover, $\mathbb{E}(|X_i|^k) < \infty$, *the function* ψ_i *is of class* $\mathscr{C}^k$ *on* $[0, \Lambda_i[$.

Proof.

Based on the Fubini's theorem and Lebesgue's dominated convergence theorem, to prove that $\lambda \mapsto \mathbb{E}\big[\exp(\lambda X_i)\big]$ has the required regularity, starting from the identity

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

□

Proposition

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

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

moreover

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

Proof.

We know that ψ_i is $\mathscr{C}^2$, from the previous proposition. So we can compute ψ'' using the rules of composition of derivatives, and write a Taylor expansion of ψ_i to obtain the last statement. $\square$

Proposition

$$\textit{Let } \Lambda > 0 \textit{ and } \mathbb{E}(X_i^2) < \infty, 1 \leq i \leq n.$$

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

$$V(\lambda) \stackrel{\text{def}}{=} \sup_{\beta \in [0,\lambda]} \overline{V}(\beta) \in \mathbb{R}_+ \cup \{+\infty\},$$

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

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

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

Proof.

As $\psi^*(m+x) \geq \beta x - \frac{\beta^2}{2} V(\lambda)$,
$\mathbb{P}(M \geq m+x) \leq \exp\left[-n\left(\beta x - \frac{\beta^2}{2} V(\lambda)\right)\right]$. We can then choose
$\lambda = x/v$ and $\beta = x/V(\lambda) \leq \lambda$ to get the first inequality and
$\epsilon = \exp\left[-n\left(\beta x - \dfrac{\beta^2}{2} V(\lambda)\right)\right]$ to get
$\mathbb{P}\left(M \geq m + \dfrac{\beta}{2} V(\lambda) + \dfrac{\log(\epsilon^{-1})}{n\beta}\right) \leq \epsilon$, and then choose
$\lambda = \sqrt{\dfrac{2\log(\epsilon^{-1})}{nv}} \geq \beta = \sqrt{\dfrac{2\log(\epsilon^{-1})}{nV(\lambda)}}$ to get the second inequality.

$\square$

Proposition (Bennett's inequality)

Let us assume that $\mathbb{E}(X_i^2) < \infty$ *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,

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

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

Proof.

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

$$\begin{aligned}\psi^*(m+x) &\geq \lambda(x+m) - \frac{1}{n}\sum_{i=1}^{n}\log\bigl[\mathbb{E}\bigl(\exp(\lambda X_i)\bigr)\bigr]\\ &= \lambda x - \frac{1}{n}\sum_{i=1}^{n}\log\Bigl\{\mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr)\bigr]\Bigr\},\end{aligned}$$

where $m_i \overset{\text{def}}{=} \mathbb{E}(X_i)$, and write

$$\begin{aligned}\mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr)\bigr] - 1 &= \mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr) - 1 - \lambda(X_i - m_i)\bigr]\\ &= \mathbb{E}\bigl[\lambda^2(X_i - m_i)^2 g\bigl(\lambda(X_i - m_i)\bigr)\bigr],\end{aligned}$$

where $g(y) = y^{-2}\bigl(\exp(y) - 1 - y\bigr)$.

Writing the Taylor expansion of $z \mapsto \exp(yz)$, we get

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

showing that the function g is non decreasing on $\mathbb{R}$.
Consequently, for any integer i such that $1 \le i \le n$,

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

Therefore,

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

Thus,

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

Let us choose $\lambda = b^{-1}\log\left(1+\frac{bx}{v}\right)$, to get $\psi^*(x) \geq \frac{v}{b^2}h\left(\frac{bx}{v}\right)$.
Chernoff's bound then gives the first inequality of the proposition.

Let us show now that $h(u) \geq \dfrac{u^2}{2(1+u/3)}$, $u > -1$, to get the second inequality. 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

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

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

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

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 third inequality of the proposition. □

Proposition (Hoeffding's inequality)

Let us assume that $a_i \leq X_i \leq b_i$, $1 \leq i \leq n$. *In this case,*

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

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,

$\psi(\lambda) \leq \lambda m + \dfrac{\lambda^2}{8}\displaystyle\sum_{i=1}^n (b_i - a_i)^2$, and therefore

$\psi^*(m + x) \geq \dfrac{2nx^2}{\sum_{i=1}^n (b_i - a_i)^2}$. □

Let $X_i \in \mathscr{X}$, $1 \leq i \leq n$ be independent, where $\mathscr{X}$ is a measurable space. Let Θ be a measurable parameter space and $f : \mathscr{X} \times \Theta \to \mathbb{R}$, a measurable function.
Assume that $\mathbb{E}\big[f(X_i, \theta)^2\big] < +\infty$, $\theta \in \Theta$, $1 \leq i \leq n$, and consider

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

Proposition

Let $\Lambda > 0$, *and* $\nu \in \mathscr{M}_+^1(\Theta)$. *For any* $\lambda \in [0, \Lambda[$,

$$\mathbb{E}\left[\exp\left(\sup\left\{\int_\Theta n\left[\lambda M(\theta) - \psi(\lambda,\theta)\right] \mathrm{d}\rho(\theta) - \mathscr{K}(\rho,\nu),\right.\right.\right.$$

$$\left.\left.\left.\rho \in \mathscr{M}_+^1(\Theta), \theta \mapsto \lambda M(\theta) - \psi(\lambda,\theta) \in \mathbb{L}^1(\rho), \mathscr{K}(\rho,\nu) < \infty\right\}\right)\right] \leq 1.$$

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

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

Proof. Let us recall that $\mathscr{K}(\rho,\nu) = \int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}\right)\mathrm{d}\rho$ whenever $\rho \ll \nu$, and is infinite otherwise. From Jensen's inequality, whenever ρ satisfies the hypotheses,

$$
\begin{aligned}
&\exp\left[\int_\Theta n\left[\lambda M(\theta) - \psi(\lambda,\theta)\right]\mathrm{d}\rho(\theta) - \mathscr{K}(\rho,\nu)\right] \\
&\leq \int_\Theta \exp\Big\{n\left[\lambda M(\theta) - \psi(\lambda,\theta)\right]\Big\}\mathbb{1}\left(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta) > 0\right)\left(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)\right)^{-1}\mathrm{d}\rho(\theta) \\
&\qquad = \int_\Theta \exp\Big\{n\left[\lambda M(\theta) - \psi(\lambda,\theta)\right]\Big\}\mathbb{1}\left(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta) > 0\right)\mathrm{d}\nu(\theta) \\
&\qquad\qquad \leq \int_\Theta \exp\Big\{n\left[\lambda M(\theta) - \psi(\lambda,\theta)\right]\Big\}\mathrm{d}\nu(\theta).
\end{aligned}
$$

We can then apply Fubini's theorem for non negative functions, to get

$$\begin{aligned}\mathbb{E}\Big\{\exp\Big[\sup_{\rho\in\mathscr{M}_+^1(\Theta)}\int_\Theta n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\,\mathrm{d}\rho(\theta)-\mathscr{K}(\rho,\nu)\Big]\Big\} \\ \leq \mathbb{E}\Big[\int_\Theta \exp\Big\{n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\Big\}\,\mathrm{d}\nu(\theta)\Big] \\ = \int_\Theta \mathbb{E}\Big[\exp\Big\{n\big[\lambda M(\theta)-\psi(\lambda,\theta)\big]\Big\}\Big]\,\mathrm{d}\nu(\theta) = 1.\end{aligned}$$

The second part of the proposition is a consequence of Markov's inequality. $\square$

$$
\begin{aligned}
\text{Let us put } m_i(\theta) &= \mathbb{E}\big[f(X_i,\theta)\big],\\
v(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big\{\big[f(X_i,\theta)-m_i(\theta)\big]^2\Big\},\\
\overline{V}(\lambda,\theta) &= \frac{2}{\lambda^2}\big[\psi(\lambda,\theta)-\lambda m(\theta)\big],\\
V(\lambda,\theta) &= \sup_{\beta\in[0,\lambda]}\overline{V}(\beta,\theta)
\end{aligned}
$$

and let us assume that $v \overset{\text{def}}{=} \sup_{\theta\in\Theta} v(\theta) < \infty$ and

$V(\lambda) \overset{\text{def}}{=} \sup_{\theta\in\Theta} V(\lambda,\theta) < \infty,\ 0 \leq \lambda < \Lambda'$.

Proposition

Under the previous hypotheses, for any positive constant c,

$$\mathbb{E}\Biggl(\sup\Bigl\{\int_{\Theta}\bigl[M(\theta)-m(\theta)\bigr]\,\mathrm{d}\rho(\theta);$$

$$\rho\in\mathscr{M}_{+}^{1}(\Theta),\theta\mapsto M(\theta)-m(\theta)\in\mathbb{L}^{1}(\rho),\mathscr{K}(\rho,\nu)\le c\Bigr\}\Biggr)$$

$$\le\inf_{\lambda\in[0,\Lambda'[}\frac{\lambda\overline{V}(\lambda)}{2}+\frac{c}{\lambda n}\le\sqrt{\frac{2c}{n}\,V\biggl(\sqrt{\frac{2c}{nv}}\biggr)}.$$

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

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

Proof.

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

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

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

Proposition

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

$$\sup\Big\{ \int_\Theta \big[M(\theta)-m(\theta)\big]\,\mathrm{d}\rho(\theta);$$

$$\rho \in \mathscr{M}_+^1(\Theta), \theta \mapsto M(\theta)-m(\theta) \in \mathbb{L}^1(\Theta), \mathscr{K}(\rho,\nu) \le c \Big\}$$

$$\le \inf_{\lambda \in [0,\Lambda'[} \frac{\lambda \overline{V}(\lambda)}{2} + \frac{c+\log(\epsilon^{-1})}{\lambda n}$$

$$\le \sqrt{\frac{2\big[c+\log(\epsilon^{-1})\big]}{n} V\Bigg(\sqrt{\frac{2\big[c+\log(\epsilon^{-1})\big]}{nv}}\Bigg)}.$$

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

$$\sup_{\theta\in\Theta}\left[M(\theta)-m(\theta)\right] \le \sqrt{\frac{2\log(|\Theta|/\epsilon)}{n} V\left(\sqrt{\frac{2\log(|\Theta|/\epsilon)}{nv}}\right)}.$$

Proof.

This is a direct consequence of Equation (1) and of the inequality $\psi(\lambda,\theta) \le \frac{\lambda^2 V(\lambda)}{2} + \lambda m(\theta)$. $\square$

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

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

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

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

Proposition

With probability at least $1-\epsilon$,

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

Thus, the quality of the estimation depends on the ratio d/n.

Proof. Let us put $f(x,\theta) = f(x, \theta/\|\theta\|)$, $\theta \in \mathbb{R}^d \setminus \mathbb{B}_d$.
Let $\delta > 0$ and ν the uniform measure on the ball $(1+\delta)\mathbb{B}_d$ of radius $1+\delta$.
For any $\theta \in \mathbb{B}_d$, let ρ_θ be the uniform probability measure 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,

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

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

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

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

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

Let $\theta_* \in \arg\min_{\theta\in\mathbb{B}_d} m(\theta)$ (reached because $\mathbb{B}_d$ is compact).

With probability at least $1-\epsilon$, $M(\theta_*) \leq m(\theta_*) + B\sqrt{\frac{\log\left(\epsilon^{-1}\right)}{2n}}$.

By construction of $\widehat{\theta}$, $M(\widehat{\theta}) \leq M(\theta_*)$. Consequently, with probability at least $1-2\epsilon$,

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

To conclude, choose $\delta = \frac{B}{4g}\sqrt{\frac{d}{2n}}$ and replace ϵ with $\epsilon/2$.

Let $\Theta = \mathbb{R}^d$. Assume that for some measurable function $(x,\theta) \mapsto \nabla f(x,\theta) \in \mathbb{R}^d$, and some positive constants g and H, for any $x \in \mathscr{X}$ and any $\theta, \theta' \in \mathbb{R}^d$,

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

$$|f(x,\theta') - f(x,\theta) - \langle \nabla f(x,\theta), \theta' - \theta\rangle| \leq \frac{H}{2}\|\theta' - \theta\|^2.$$

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

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

Proposition

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* $1 - \epsilon$

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

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

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.

Proof. Let $\rho_\theta = \mathscr{N}(\theta, \beta^{-1} I)$ and $\nu = \rho_{\theta_*}$.
Let us remark that $\mathscr{K}(\rho_\theta, \nu) = \dfrac{\beta}{2} \|\theta - \theta_*\|^2$.
Let us apply Equation (1) to the function $(x,\theta) \mapsto f(x,\theta_*) - f(x,\theta)$. From Hoeffding's inequality,

$$\log \mathbb{E} \exp\Big\{ \lambda \big[f(X,\theta_*) - f(X,\theta) \big] \Big\} - \lambda \big[m(\theta_*) - m(\theta) \big] \\ \leq \frac{\lambda^2 g^2 \|\theta - \theta_*\|^2}{2}.$$

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,

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

In the same way, $\displaystyle\int M(\theta') \,\mathrm{d}\rho_\theta(\theta') \leq M(\theta) + \frac{Hd}{2\beta}.$

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

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

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 $1-\epsilon$

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

Thus

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

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

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

This gives the first upper bound of the proposition.

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

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

We conclude in the same way, replacing λ and β by their values.
$\square$

Let $W_{1:n} \in \mathscr{W}^n$ be an i.i.d. sample, on a measurable space $\mathscr{W}$.
Let $\mathbb{P}^{\otimes n} \in \mathscr{M}_+^1(\mathscr{W}^n)$ be the distribution of $W_{1:n}$.
Let Θ be a measurable parameter space, and $L : \mathscr{W} \times \Theta \to \{0, 1\}$ a binary measurable loss function.

Our aim will be to minimize the expected loss $\displaystyle\int L(w, \theta) \, \mathrm{d}\mathbb{P}(w)$.

In the setting of supervised classification, $\mathscr{W} = \mathscr{X} \times \mathscr{Y}$, where $\mathscr{X}$ is a pattern space and $\mathscr{Y}$ a finite set of classes. Accordingly, $W_i = (X_i, Y_i)$ are input-output pairs.

We are given a family of measurable classification rules $\{f_\theta : \mathscr{X} \to \mathscr{Y}, \theta \in \Theta\}$, and L is defined as $L\big[(x, y), \theta\big] = \mathbb{1}\big(f_\theta(x) \neq y\big)$, so that the loss $\int L(w, \theta) \, \mathrm{d}\mathbb{P}(w) = \mathbb{P}_{X,Y}\big(f_\theta(X) \neq Y\big)$ is equal to the expected classification error.

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

For any $\lambda \in \mathbb{R}$, let $\Phi_\lambda(p) \overset{\text{def}}{=} -\frac{1}{\lambda} \log\bigl[1 - p + p \exp(-\lambda)\bigr]$, and
$K(q,p) \overset{\text{def}}{=} q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$.
Let $\overline{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n \delta_{W_i}$.
For any $\rho, \pi \in \mathscr{M}_+^1(\Theta)$ and any integrable function $f \in \mathbb{L}_1\bigl(\mathscr{W} \times \Theta^2, \mathbb{P} \otimes \pi \otimes \rho\bigr)$, let

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

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

For any probability measures π and ρ defined on the same measurable space, such that $\mathscr{K}(\rho, \pi) < \infty$, 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

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

Proposition

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

$$\int h \,\mathrm{d}\rho - \mathscr{K}(\rho,\pi) + \mathscr{K}(\rho,\pi_{\exp(h)}) = \log\left[\int \exp(h)\,\mathrm{d}\pi\right] \quad (2)$$

$$= \int h\,\mathrm{d}\pi + \int_0^1 (1-\alpha)\,\mathbf{Var}\left[h\,\mathrm{d}\pi_{\exp(\alpha h)}\right]\mathrm{d}\alpha. \quad (3)$$

Proof.

Equation (2) is a straightforward consequence of the definitions. Equation (3) is the Taylor expansion of of the function $\alpha \mapsto \log\left[\int \exp(\alpha h)\,\mathrm{d}\pi\right]$. □

$$\begin{aligned}
\text{Let} \qquad B_+(q,\delta) &= \inf_{\lambda \in \mathbb{R}_+} \Phi_\lambda^{-1}\left(q + \frac{\delta}{\lambda}\right) \\
&= \sup\Big\{ p \in [0,1] \,:\, K(q,p) \leq \delta \Big\}, \qquad q \in [0,1],\ \delta \in \mathbb{R}_+, \\
\text{and} \qquad B_-(q,\delta) &= \inf_{\lambda \in \mathbb{R}_+} \Phi_{-\lambda}(q) + \frac{\delta}{\lambda} \\
&= \sup\Big\{ p \in [0,1] \,:\, K(p,q) \leq \delta \Big\}, \qquad q \in [0,1],\ \delta \in \mathbb{R}_+,
\end{aligned}$$

Proposition

For any non random $\theta \in \Theta$, *with probability at least* $1-\epsilon$,

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

Moreover

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

In the same way, with probability at least $1-\epsilon$

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

and

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

Proof. From Chernoff's bound, with probability at least $1-\epsilon$,

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

Since the left-hand side is non-random, it can be optimized in λ, giving

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

Since $\displaystyle\lim_{\lambda\to+\infty} \Phi_\lambda^{-1}\left(q+\frac{\delta}{\lambda}\right) = \lim_{\lambda\to+\infty} \frac{1-\exp(-\lambda q-\delta)}{1-\exp(-\lambda)} \leq 1$, $B_+(q,\delta) \leq 1$. Applying equation (2) 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

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

because when $q \le p < 1$ we can choose $\lambda = \log\left(\dfrac{q^{-1}-1}{p^{-1}-1}\right) \in \mathbb{R}_+$, for which $q = p_\lambda$ and therefore $K(q,p_\lambda) = 0$.

Let us remark now that $\frac{\partial^2}{\partial x^2} K(x,p) = x^{-1}(1-x)^{-1}$. 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

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

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

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

Reverse inequalities are proved in the same way.

Proposition

Given any set $\Lambda \subset \mathbb{R}_+$, *let* $B_\Lambda(q, \delta) = \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}\left(q + \frac{\delta}{\lambda}\right)$.
For any prior probability measure $\pi \in \mathscr{M}_+^1(\Theta)$ *and any* $\lambda \in \mathbb{R}_+$,

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

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

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

Proof.

The exponential moment inequality (4) is a consequence of Equation (2), showing that

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

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. □

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

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

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

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

Let us define

$$B_m(q,e,\epsilon) = \max\Bigg\{ \sqrt{\frac{2\overline{v}(q)\{e+\log[(m+1)/\epsilon]\}}{n}}\cosh(t/m) + \frac{2(1-q)\{e+\log[(m+1)/\epsilon]\}}{n}\cosh(t/m)^2, \frac{2\{e+\log[(m+1)/\epsilon]\}}{n} \Bigg\}$$

$$\leq \sqrt{\frac{2\overline{v}(q)\{e+\log[(m+1)/\epsilon]\}}{n}}\cosh(t/m) + \frac{2\{e+\log[(m+1)/\epsilon]\}}{n}\cosh(t/m)^2.$$

Let us also consider

$$B(q, e, \epsilon) \stackrel{\text{def}}{=} \sqrt{\frac{2\overline{v}(q)\{e + \log[\log(n)^2/\epsilon]\}}{n}} \cosh[\log(n)^{-1}] + \frac{2\{e + \log[\log(n)^2/\epsilon]\}}{n} \cosh[\log(n)^{-1}]^2, \quad (5)$$

Proposition

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

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

Moreover, as soon as $n \geq 5$, $B_{\lfloor \log(n)^2 \rfloor - 1}(q,e,\epsilon) \leq B(q,e,\epsilon)$, *so that with probability at least* $1-\epsilon$, *for any* $\rho \in \mathscr{M}_+^1(\Theta)$,

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

Let us put

$$
\begin{aligned}
q &= L(\overline{\mathbb{P}}, \rho), \\
\delta &= \frac{\mathscr{K}(\rho, \pi) + \log\bigl[(m+1)/\epsilon\bigr]}{n}, \\
\lambda_{\min} &= \sqrt{\frac{8\log\bigl[(m+1)/\epsilon\bigr]}{n}}, \\
\Lambda &= \Bigl\{\lambda_{\min}^{1-k/m}, k = 0, \dots, m\Bigr\}, \\
p &= B_\Lambda(q, \delta) = \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}\left(q + \frac{\delta}{\lambda}\right), \\
\widehat{\lambda} &= \sqrt{\frac{2\delta}{\overline{v}(p)}}.
\end{aligned}
$$

According to equation (3) 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 $\delta \geq \dfrac{\log[(m+1)/\epsilon]}{n}$,

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

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 $\log(\widehat{\lambda})$ is at distance at most t/m from some $\log(\lambda)$ where $\lambda \in \Lambda$, because $\log(\Lambda)$ is a grid with constant steps of size $2t/m$. Thus

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

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

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

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

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

so that in both cases

$$p - q \le \sqrt{2\delta\overline{v}(q)}\cosh(t/m) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2. \qquad (6)$$

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

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

In conclusion, applying Proposition 14 we see that with probability at least $1-\epsilon$, for any posterior distribution ρ,

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

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.

Let $\mathscr{W} = \mathscr{X} \times \mathscr{Y} = \mathbb{R}^d \times \{-1, +1\}$, and
$L(w, \theta) = L((x, y), \theta) = \mathbb{1}\big[\langle \theta, x \rangle\, y \leq 0\big]$.

We will follow the approach presented in [5] and [8].

The bounds that does not depend on d can be generalized to the case where the pattern space $\mathscr{X}$ is a Hilbert space of infinite dimension. They apply to Support Vector Machines, where we have an implicit mapping $\Psi : \mathscr{X} \to \mathscr{H}$, into a Hilbert space $\mathscr{H}$, where $\Theta = \mathscr{H}$ and where $L(w, \theta) = \mathbb{1}\big(\langle \theta, \Psi(x) \rangle\, y \leq 0\big)$.

Support Vector Machine algorithms are defined in terms of the scalar product $k(x_1, x_2) = \langle \Psi(x_1), \Psi(x_2) \rangle$, defining a positive symmetric kernel k on the original pattern space $\mathscr{X}$. According to the Moore-Aronszajn theorem, k may be any positive symmetric kernel. Popular kernels on $\mathscr{X} = \mathbb{R}^d$ are

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

Let us consider, after [5, 8] as prior probability measure π the centered Gaussian measure with covariance $\beta^{-1}\,\mathrm{Id}$, 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

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

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

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

and

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

To get an insight on $L(w, \theta)$ itself, let us introduce the *error with margin*

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

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

$$M(w, \pi_\theta) \geq \varphi(-\sqrt{\beta}) L(w, \theta) = \big[1 - \varphi(\sqrt{\beta})\big] L(w, \theta). \tag{7}$$

Proposition

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

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

where

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

the bound B *being defined by equation* (5).

Let $\widehat{\theta}$ be any estimator satisfying

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

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

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

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

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

the function $p \mapsto -\Phi_{-\lambda}(p)$ being convex.

As a consequence

Proposition

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

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

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

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

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

The criterions C_1 and C_2 are non-convex, faster minimization algorithms are available for the usual SVM loss function, that we are going to study now.

Let us choose some positive radius R and let us put $\|x\|_R = \max\{R, \|x\|\}$, so that in the case when $\|x\| \leq R$, $\|x\|_R = R$.

$$M(w, \pi_\theta) = \varphi\big[\sqrt{\beta}\big(y\|x\|^{-1}\langle\theta, x\rangle - 1\big)\big] \\ \leq \big(2 - y\|x\|_R^{-1}\langle\theta, x\rangle\big)_+ + \varphi(\sqrt{\beta}). \quad (8)$$

Using the upper bounds (8) and (7), and Proposition 14, we obtain

Proposition

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

$$L(\mathbb{P},\theta) \leq \left[1-\varphi(\sqrt{\beta})\right]^{-1} B_\Lambda\Bigg(\int \left(2-y\|x\|_R^{-1}\langle\theta,x\rangle\right)_+ d\overline{\mathbb{P}}(x,y) + \varphi(\sqrt{\beta}), \frac{\beta\|\theta\|^2 + 2\log(|\Lambda|/\epsilon)}{2n}\Bigg)$$

$$= \left[1-\varphi(\sqrt{\beta})\right]^{-1} \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\left[C_3(\lambda,\theta) + \varphi(\sqrt{\beta}) + \frac{\log(|\Lambda|/\epsilon)}{n\lambda}\right],$$

where

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

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 also that $L(\mathbb{P}, \theta) = L(\mathbb{P}, 2R\theta)$,

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

Proposition

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

$$L(\mathbb{P},\theta) \le \inf_{\beta\in\Xi} \left[1-\varphi(\sqrt{\beta})\right]^{-1} \inf_{\lambda\in\Lambda} \Phi_{\lambda}^{-1}\biggl[2\,C_4(\beta,\lambda,\theta) + \varphi(\sqrt{\beta}) + \frac{\log\bigl(|\Xi|\,|\Lambda|/\epsilon\bigr)}{n\lambda}\biggr],$$

where

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

The loss function $C_4(\lambda,\theta)$ 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 reguralizing factor, $\|\theta\|^{-1}$ plays the role of a 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 .

The speed of convergence depends on $R^2\|\theta\|^2/n$, where $R^2\|\theta\|^2$, plays the role of the dimension and is independent of d.

Corollary

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* $\min_{i=1,\ldots,n} \langle \theta, x_i \rangle \leq \gamma \leq \max_{i=1,\ldots,n} \langle \theta, x_i \rangle$*,*

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

where

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

Proof.

Let us apply the previous result to $x' = (x - c, R)$, and $\theta' = \left[\theta, R^{-1}\left(\langle \theta, c \rangle - \gamma\right)\right]$.
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,
$\left(\langle \theta, c \rangle - \gamma\right)^2 \leq R^2 \|\theta\|^2.$
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. □

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.

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.

T. Cover and J. Thomas.
Elements of Information Theory.

Wiley and Sons, New York, second edition, 2006.

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.

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

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

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

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

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.