Chapter 20
PAC-Bayes Bounds for Supervised Classification

Olivier Catoni

Abstract We present in this contribution a synthesis of Seeger's (PAC-Bayesian generalization error bounds for Gaussian process classification, 2002) and our own (Catoni, PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, 2007) approach of PAC-Bayes inequalities for 0–1 loss functions. We apply it to supervised classification, and more specifically to the proof of new margin bounds for support vector machines, in the spirit of the bounds established by Langford and Shawe-Taylor (Advances in Neural Information Processing Systems, 2002) and McAllester (Learning Theory and Kernel Machines, COLT 2003).

20.1 PAC-Bayes Bounds for 0–1 Loss Functions

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 $W_{1:n} \overset{\text{def}}{=} (W_i)_{i=1}^n$ such that the excess risk

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

O. Catoni (✉)
CNRS – UMR 8553, Département de Mathématiques et Applications,
École Normale Supérieure, 45, rue d'Ulm, 75230 Paris Cedex 05, France
e-mail: Olivier.Catoni@ens.fr

O. Catoni
INRIA Paris-Rocquencourt — CLASSIC Team, Le Chesnay Cedex, France

V. Vovk et al. (eds.), *Measures of Complexity*,
DOI 10.1007/978-3-319-21852-6_20

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

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

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

The point of view presented here is a synthesis of the approaches of [2, 8].

20.1.1 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:

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

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

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

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 first recall Chernoff's bound.

Proposition 20.1 *For any fixed value of the parameter $\theta \in \Theta$, the identity*

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

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

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

where

$$\begin{aligned} B_+(q, \delta) &= \inf_{\lambda \in \mathbb{R}_+} \Phi_\lambda^{-1}\biggl(q + \frac{\delta}{\lambda}\biggr) \\ &= \sup\Bigl\{p \in [0, 1] \,:\, K(q, p) \leq \delta\Bigr\}, \qquad q \in [0, 1],\ \delta \in \mathbb{R}_+, \end{aligned}$$

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

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

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

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

where

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

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 identities.

Proposition 20.2 *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 $Z = \int \exp(h)\,\mathrm{d}\pi$. Moreover, let us introduce the notation

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

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

$$\int h\,\mathrm{d}\rho - \mathcal{K}(\rho, \pi) + \mathcal{K}(\rho, \pi_{\exp(h)}) = \log\bigl[\int \exp(h)\,\mathrm{d}\pi\bigr] \tag{20.1}$$

$$= \int h\,\mathrm{d}\pi + \int_0^1 (1-\alpha)\,\mathbf{Var}\bigl[h\,\mathrm{d}\pi_{\exp(\alpha h)}\bigr]\,\mathrm{d}\alpha. \tag{20.2}$$

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

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

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

Exercise 20.1 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 20.1. Chernoff's inequality reads

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

where the inequality holds with probability at least $1 - \epsilon$. Since the left-hand side is non-random, it can be optimized in λ, giving

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

Exercise 20.2 Prove this statement in more detail. For any integer $k > 1$, consider the event

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

where $F(\lambda) = \Phi_\lambda\bigl[L(\mathbb{P}, \theta)\bigr] - \dfrac{\log(\epsilon^{-1})}{n\lambda}$. 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 $\mathbb{P}^{\otimes n}\bigl(\cup_k A_k\bigr) \leq \epsilon$. □

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 Eq. (20.1) 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

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

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

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

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 if $K(q,p) = \delta$ 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 20.3 The second part of Proposition 20.1 is proved in the same way and left as an exercise. □

20.1.2 PAC-Bayes Bounds

We are now going to make Proposition 20.1 uniform with respect to θ. The PAC-Bayes approach to this [3, 5–7] is to randomize θ, so we will now consider 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.[1] This posterior distribution ρ will be compared with a prior (meaning non-random) probability measure $\pi \in \mathcal{M}_+^1(\Theta)$.

[1] We will assume that ρ is a regular conditional probability kernel, meaning that for any measurable set A the map $(w_1, \dots, w_n) \mapsto \rho(w_1, \dots, w_n)(A)$ 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] (p. 50) for more details.

Proposition 20.3 *Let us introduce the notation*

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

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

$$\int \exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} n\lambda\Big\{\Phi_\lambda\big[L(\mathbb{P},\rho)\big] - L(\overline{\mathbb{P}},\rho)\Big\} - \mathcal{K}(\rho,\pi)\right] \mathrm{d}\mathbb{P}^{\otimes n} \leq 1, \qquad (20.3)$$

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

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

Proof The exponential moment inequality (20.3) is a consequence of Eq. (20.1), showing that

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

and of the fact that Φ_λ is convex, showing that

$$\Phi_\lambda\big[L(\mathbb{P},\rho)\big] \leq \int \Phi_\lambda\big[L(\mathbb{P},\theta)\big]\,\mathrm{d}\rho(\theta).$$

The deviation inequality follows as usual. □

We cannot take the infimum on $\lambda \in \mathbb{R}_+$ as in Proposition 20.1, because we can no longer 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 20.4 *Let us define the least increasing upper bound of the variance of a Bernoulli distribution of parameter* $p \in [0,1]$ *as*

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

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

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

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

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

where

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

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

$$\begin{aligned} B_{\lfloor \log(n)^2\rfloor - 1}(q,e,\epsilon) \leq B(q,e,\epsilon) \overset{\text{def}}{=} & \\ \sqrt{\frac{2\overline{v}(q)\bigl\{e+\log\bigl[\log(n)^2/\epsilon\bigr]\bigr\}}{n}} & \cosh\bigl[\log(n)^{-1}\bigr] \\ & + \frac{2\bigl\{e+\log\bigl[\log(n)^2/\epsilon\bigr]\bigr\}}{n}\cosh\bigl[\log(n)^{-1}\bigr]^2, \end{aligned} \qquad (20.4)$$

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

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

Proof Let us put

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

According to Eq. (20.2) applied to Bernoulli distributions, for any $\lambda \in \Lambda$,

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

Moreover, as $p_\alpha \le p$,

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

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 bounded by

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

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

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

so that in both cases

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

Let us now consider 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 20.3 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. □

20.2 Linear Classification and Support Vector Machines

In this section we are going 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 will follow the approach presented in [4, 5].

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 [9–11], 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

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

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

After [4, 5], let us consider 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

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

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

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

and

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

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

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

Unlike $L(w,\theta)$ which is independent of the norm of θ, the margin error $M(w,\theta)$ depends on $\|\theta\|$, counting a classification error each time x is at distance less than $\|x\|/\|\theta\|$ from the boundary $\{x' : \langle\theta, x'\rangle = 0\}$, so that the error with margin region is the complement of the open cone $\bigl\{x \in \mathbb{R}^d\,;\ y\langle\theta, x\rangle > \|x\|\bigr\}$.

Let us compute the randomized margin error

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

It satisfies the inequality

$$M(w,\pi_\theta) \geq \varphi(-\sqrt{\beta})L(w,\theta) = \bigl[1 - \varphi\bigl(\sqrt{\beta}\bigr)\bigr]L(w,\theta). \tag{20.5}$$

Applying previous results we obtain

Proposition 20.5 *With probability at least* $1-\epsilon$*, for any* $\theta \in \mathbb{R}^d$*,*

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

where

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

the bound B being defined by Eq. (20.4).

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

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

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

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

Indeed

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

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

Proposition 20.6 *With probability at least* $1-2\epsilon$,

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

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 $\|x\|\leq R$, $\|x\|_R=R$.

$$M(w,\pi_\theta)=\varphi\big[\sqrt{\beta}\big(y\|x\|^{-1}\langle\theta,x\rangle-1\big)\big]\leq\big(2-y\|x\|_R^{-1}\langle\theta,x\rangle\big)_+ +\varphi(\sqrt{\beta}). \quad (20.6)$$

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 (20.6), since the three intervals on which $g\geq f$ cover the whole real line.

Using the upper bounds (20.6) and (20.5), and Proposition 20.3, we obtain

Proposition 20.7 *With probability at least* $1-\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

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

where

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

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

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

where

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

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 also be used to optimize the values of λ and β, that is to optimize the strength of the regularizing factor $\dfrac{\beta R^2\|\theta\|^2}{n\lambda}$.

Here $\|\theta\|^{-1}$ 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 $\bigl(1 - y\langle\theta,x\rangle\bigr)_+$ 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 not necessarily centered ball, through a change of coordinates. More precisely, the previous proposition has the following corollary:

Corollary 20.1 *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* $\min_{i=1,\dots,n} \langle \theta, x_i \rangle \leq \gamma \leq \max_{i=1,\dots,n} \langle \theta, x_i \rangle$,

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

where

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

Proof Let us apply the previous result to $x' = (x - c, R)$, and $\theta' = \bigl[\theta, R^{-1}\bigl(\langle\theta, c\rangle - \gamma\bigr)\bigr]$. We get that $\|x'\|^2 \leq 2R^2$ and $\|\theta'\|^2 = 2\|\theta\|^2$, 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, $\bigl(\langle\theta, c\rangle - \gamma\bigr)^2 \leq R^2\|\theta\|^2$. 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. Catoni, O.: Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI—2001. Lecture Notes in Mathematics, vol. 1851. Springer, Berlin (2004)
2. Catoni, O.: PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. IMS Lecture Notes Monograph Series, vol. 56. Institute of Mathematical Statistics, Beachwood (2007)
3. Germain, P., Lacasse, A., Laviolette, F., Marchand, M.: PAC-Bayesian learning of linear classifiers. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML'09, pp. 353–360. ACM, New York (2009)
4. Langford, J., Shawe-Taylor, J.: PAC-Bayes & margins. Advances in Neural Information Processing Systems, pp. 423–430. MIT Press, Cambridge (2002)

5. McAllester, D.: Simplified PAC-Bayesian margin bounds. In: Schölkopf, B., Warmuth, M.K. (eds.) Learning Theory and Kernel Machines, COLT 2003. Lecture Notes in Artificial Intelligence, vol. 2777, pp. 203–215. Springer, Berlin (2003)
6. McAllester, D.A.: PAC-Bayesian model averaging. In: Proceedings of the 12th Annual Conference on Computational Learning Theory, pp. 164–170. ACM, New York (1999)
7. McAllester, D.A.: PAC-Bayesian stochastic model selection. Mach. Learn. **51**(1), 5–21 (2003)
8. Seeger, M.: PAC-Bayesian generalization error bounds for Gaussian process classification. Informatics report series EDI-INF-RR-0094, Division of Informatics, University of Edinburgh. http://www.inf.ed.ac.uk/publications/online/0094.pdf (2002)
9. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer, Berlin (1982)
10. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
11. Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. **16**(2), 264–280 (1971) (This volume, Chap. 3)