3. PAC-Bayes bounds for supervised classification

3.1. Introduction. PAC-Bayes theory was first developed in the framework of supervised classification (see [6, 7, 8, 9, 5]) and subsequently extended to other settings. We will not deal with these extensions in the present notes, but instead focus on supervised classification, a setting that plays a central role in statistical learning theory and requires specific techniques of proofs.

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 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, how small it is 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.

Let us remark that the point of view adopted in this section is different from the point of view of the previous section in some important ways. With the tools of the last section, we could have considered the loss function

$$L_c\big[(x, y), q_\theta\big] = -\log_2\big[q_\theta(y \,|\, x)\big],$$

where $q_\theta(y|x)$ is some family of conditional distributions indexed by θ.

We could then have obtained results concerned with

$$\sum_{i=1}^{n} L_c\left[(X_i, Y_i), \int q_\theta \, \mathrm{d}\rho_i(\theta)\right],$$

where $\rho_i \in \mathcal{M}_+^1(\Theta)$ is a random probability measure defined with the help of some prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ by its density

$$\frac{\mathrm{d}\rho_i}{\mathrm{d}\pi}(\theta) = \frac{\displaystyle\prod_{i=1}^{n} q_\theta(Y_i|X_i)}{\displaystyle\int \prod_{i=1}^{n} q_{\theta'}(Y_i|X_i) \, \mathrm{d}\pi(\theta')}.$$

Namely, the previous section provides the following almost sure upper bound:

$$\sum_{i=1}^{n} L_c\left[(X_i, Y_i), \int q_\theta \, \mathrm{d}\rho_i(\theta)\right] \leq \inf_{\theta\in\Theta} \sum_{i=1}^{n} L_c\big[(X_i, Y_i), q_\theta\big] - \log_2\big[\pi\big(\{\theta\}\big)\big]. \quad (7)$$

Importantly, the previous section is concerned with a *cumulated* notion of risk, whereas this one will deal with an *instantaneous* notion of risk. For the cumulated risk, we could obtain almost sure results, this will not be possible any more for the instantaneous risk we are considering here. We will have instead to make an hypothesis about the probabilistic nature of the observations, assuming that they are independent, and obtain results holding with a probability close to one but not equal to one.

Exercise 7 *To make the link with the previous section more specific, we can focus on*

$$q_\theta(y|x) = \frac{\exp\big\{-\beta \mathbb{1}\big[f_\theta(x) \neq y\big]\big\}}{\big[1 + \big(|\mathcal{Y}| - 1\big)\exp(-\beta)\big]}.$$

Show that

$$\big[1 - \exp(-\beta)\big] \sum_{i=1}^{n} \int \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big] \, \mathrm{d}\rho_i(\theta)$$
$$\leq \inf_{\theta\in\Theta} \beta \sum_{i=1}^{n} \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big] - \log_2\big[\pi\big(\{\theta\}\big)\big].$$

(Hint : consider a lower bound of the left-hand side of equation (7, page 25) *and an upper bound of its right-hand side, using the fact that* $\log(1+z) \leq z$*.) Conclude that, almost surely*

$$\sum_{i=1}^n \int L\big[(X_i,Y_i),\theta\big]\,\mathrm{d}\rho_i(\theta) \\ \leq \inf_{\theta\in\Theta} \frac{1}{1-\beta/2}\sum_{i=1}^n L\big[(X_i,Y_i),\theta\big] - \frac{\log_2\big[\pi\big(\{\theta\}\big)\big]}{\beta(1-\beta/2)}.$$

(Hint: use the inequality $\exp(-\beta)\leq 1-\beta+\beta^2/2$, $\beta\geq 0$*.) How would you choose the value of the parameter* β *?*

3.2. 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 :

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

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

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

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 $\mathbb{P}^{\otimes n}\in\mathcal{M}_+^1(\mathcal{W}^n)$. 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*

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

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

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

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

Moreover

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

In the same way, the identity

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

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

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

and

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

Let us mention here some important identity.

Proposition 3.2 *For any probability measures* π *and* ρ *on some measurable space, such that* $\mathcal{K}(\rho,\pi)<\infty$*, and any bounded measurable function* h*, let us define the transformed probability measure* $\pi_{\exp(h)} \ll \pi$ *by its density*

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

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

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

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

$$\int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi) + \mathcal{K}(\rho,\pi_{\exp(h)}) = \log\big[\textstyle\int \exp(h)\,\mathrm{d}\pi\big] \tag{8}$$
$$= \textstyle\int h\,\mathrm{d}\pi + \int_0^1 (1-\alpha)\,\mathbf{Var}\big[h\,\mathrm{d}\pi_{\exp(\alpha h)}\big]\,\mathrm{d}\alpha. \tag{9}$$

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

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

and 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 26). 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 9 *Prove this statement in more details. 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 equation (8, page 27) to Bernoulli distributions gives

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

where

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

This shows that

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

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

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

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

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 10 *The second part of Proposition 3.1 (page 26) is proved in the same way and left as an exercise.*

3.3. PAC-Bayes bounds. We are now going to make Proposition 3.1 uniform with respect to θ. The PAC-Bayes approach to this is to randomize θ, so we will consider now joint distributions on $(W_1,\dots,W_n,\theta)$, where the distribution of $(W_1,\dots,W_n)$ is still $\mathbb{P}^{\otimes n}$ and the conditional distribution of θ given the sample is given by some transition probability kernel $\rho : \mathcal{W}^n \to \mathcal{M}_+^1(\Theta)$, called in this context a posterior distribution*. This posterior distribution ρ will be compared with a prior (meaning non-random) probability measure $\pi \in \mathcal{M}_+^1(\Theta)$.

Proposition 3.3 *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 $\lambda \in \mathbb{R}_+$,

$$\int \exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} n\lambda\Big\{\Phi_\lambda\big[L(\mathbb{P},\rho)\big] - L(\overline{\mathbb{P}},\rho)\Big\} - \mathcal{K}(\rho,\pi)\right] \mathrm{d}\mathbb{P}^{\otimes n} \leq 1, \qquad (10)$$

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 (10) is a consequence of equation (8, page 27), showing that

*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][page 50] for more details

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

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 3.1 (page 26), 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

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

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

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

where

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

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

$$B_{\lfloor \log(n)^2 \rfloor - 1}(q, e, \epsilon) \leq B(q, e, \epsilon) \stackrel{\text{def}}{=} \sqrt{\frac{2\overline{v}(q)\big\{e + \log\big[\log(n)^2/\epsilon\big]\big\}}{n}} \cosh\big[\log(n)^{-1}\big] + \frac{2\big\{e + \log\big[\log(n)^2/\epsilon\big]\big\}}{n} \cosh\big[\log(n)^{-1}\big]^2, \quad (11)$$

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

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

Proof. Let us put

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

According to equation (9, page 27) 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 $\delta \geq \dfrac{\log\bigl[(m+1)/\epsilon\bigr]}{n}$,

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

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 $\log\bigl(\widehat{\lambda}\bigr)$ 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

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

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

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

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

$$p - q \leq \sqrt{2\delta\overline{v}(q)}\cosh(t/m) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2. \tag{12}$$

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

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

In conclusion, applying Proposition 3.3 (page 30) 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. □

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] F.M.J. Willems, Y.M. Shtarkov, and T.J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Trans. Inform. Theory*, 41(3):653–664, 1995.