Improved Vapnik Cervonenkis bounds

Olivier Catoni

IMPROVED VAPNIK CERVONENKIS BOUNDS

OLIVIER CATONI

Abstract. We give a new proof of VC bounds where we avoid the use of symmetrization and use a shadow sample of arbitrary size. We also improve on the variance term. This results in better constants, as shown on numerical examples. Moreover our bounds still hold for non identically distributed independent random variables.

2000 Mathematics Subject Classification: 62H30, 68T05, 62B10.
Keywords: Statistical learning theory, PAC-Bayesian theorems, VC dimension.

1. Description of the problem

Let $(\mathcal{X}, \mathcal{B})$ be some measurable space and $\mathcal{Y}$ some finite set. Let $(\Theta, \mathcal{T})$ be a measurable parameter space and $\{f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}$ be a family of decision functions. Assume that

$$(\theta, x) \mapsto f_\theta(x) : (\Theta \times \mathcal{X}, \mathcal{T} \otimes \mathcal{B}) \to \mathcal{Y}$$

is measurable. Let

$$P_i \in \mathcal{M}_+^1\big(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \{0,1\}^{\mathcal{Y}}\big), \quad i = 1, \dots, N,$$

be some probability distributions on $\mathcal{X} \times \mathcal{Y}$ — where $\{0,1\}^{\mathcal{Y}}$ is the discrete sigma algebra of all the subsets of $\mathcal{Y}$. Let $(X_i, Y_i)_{i=1}^N$ be the canonical process on $(\mathcal{X} \times \mathcal{Y})^N$ — i.e. the coordinate process $(X_i, Y_i)(\omega) = \omega_i$, $\omega \in \big(\mathcal{X} \times \mathcal{Y}\big)^N$. Let

$$r(\theta) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big].$$

We are interested in bounding with $\bigotimes_{i=1}^N P_i$ probability at least $1 - \epsilon$ and for any $\theta \in \Theta$ the quantity $R(\theta) - r(\theta)$. This question has an interest both in statistical learning theory and in empirical process theory.

In the case when $|\mathcal{Y}| = 2$, introducing the notation

$$\mathcal{N}(X_1^{2N}) = \big|\big\{\big[f_\theta(X_i)\big]_{i=1}^{2N}; \theta \in \Theta\big\}\big|,$$

where $|A|$ is the number of elements of the set A, Vapnik proved in [10, page 138] that

Theorem 1.1. *For any probability distribution $P \in \mathcal{M}_+^1\big(\mathcal{X} \times \mathcal{Y}\big)$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq r(\theta) + \frac{2d'}{N}\left(1 + \sqrt{1 + \frac{Nr(\theta)}{d'}}\right),$$

where

$$d' = \log\Big\{P^{\otimes 2N}\big[\mathcal{N}(X_1^{2N})\big]\Big\} + \log\big(4\epsilon^{-1}\big).$$

It is also well known since the works of Vapnik and Cervonenkis that, in the case when $\mathcal{Y} = \{0,1\}$,

$$\log\big[\mathcal{N}(X_1^{2N})\big] \leq h \log\left(\frac{eN}{h}\right),$$

where

$$h = \max\big\{|A|; \mathcal{N}\big[(X_i)_{i\in A}\big] = 2^{|A|}\big\}.$$

Therefore when the VC dimension of $\{f_\theta; \theta \in \Theta\}$ is not greater than h, that is when by definition

$$\max\Big\{|A|; A \subset \mathcal{X}, \big|\big\{\big(f_\theta(x)\big)_{x\in A}; \theta \in \Theta\big\}\big| = 2^{|A|}\Big\} \leq h,$$

we have the following

Corollary 1.2. *When the VC dimension of $\big\{f_\theta; \theta \in \Theta\big\}$ is not greater than h, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq r(\theta) + \frac{2d'}{N}\left(1 + \sqrt{1 + \frac{Nr(\theta)}{d'}}\right),$$

where

$$d' = h\log\left(\frac{2eN}{h}\right) + \log\big(4\epsilon^{-1}\big).$$

The aim of this paper is to improve theorem 1.1 and its corollary, using PAC-Bayesian inequalities with data dependent priors.

We have already proved in [5] that with $P^{\otimes N}$ probability at least $1-\epsilon$,

$$R(\theta) \leq r(\theta) + \frac{\zeta d}{N}\left(1 + \sqrt{1 + \frac{4Nr(\theta)}{\zeta d}}\right), \tag{1.1}$$

where

$$d = P_{X_{N+1}^{2N}}^{\otimes N}\Big\{\log\big[\mathcal{N}(X_1^{2N})\big]\Big\} + \log\left(\frac{\log(2\zeta N)}{\epsilon\log(\zeta)}\right),$$

which brings an improvement when $r(\theta) \leq \frac{d}{N}$ and d is large.

Here we are going to generalize this theorem to arbitrary shadow sample sizes and non identically distributed independent random variables. We will also improve on the variance term in (1.1) and get rid of the (unwanted !) parameter ζ.

Moreover, we will derive VC bounds in the transductive setting in which the shadow sample error rate is bounded in terms of the empirical error rate (in this setting the shadow sample would more appropriately be described as a test set).

We will start with the transductive setting, since it has an interest of its own and will in the same time serve as a technical step towards more classical results.

2. The transductive setting

We will consider a shadow sample of size kN where k is some integer.
Let $(X_i, Y_i)_{i=1}^{(k+1)N}$ be the canonical process on $\big(\mathcal{X} \times \mathcal{Y}\big)^{(k+1)N}$.
We assume that we observe the first sample $(X_i, Y_i)_{i=1}^{N}$, that we may also observe the rest of the design $X_{N+1}^{(k+1)N}$, (this is a short notation for $(X_i)_{i=N+1}^{(k+1)N}$), but that we do not observe $Y_{N+1}^{(k+1)N}$.
Let $r_1(\theta)$ and $r_2(\theta)$ be the empirical error rates of the decision function f_θ on the training and test sets:

$$r_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\big[Y_i \neq f_\theta(X_i)\big],$$

$$r_2(\theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \mathbb{1}\big[Y_i \neq f_\theta(X_i)\big].$$

Let $\mathbb{P} \in \mathcal{M}_+^1\Big[\big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N}\Big]$ be some *partially exchangeable* probability distribution on $\big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N}$. What we mean by *partially exchangeable* will be precisely defined in the following. An important case is when $\mathbb{P} = \Big(\bigotimes_{i=1}^{N} P_i\Big)^{\otimes(k+1)}$, meaning that we have $(k+1)$ independent samples, each being distributed according to the same product of non identical probability distributions. Let as in the introduction

$$\mathcal{N}(X_1^{(k+1)N}) = \Big|\Big\{\big[f_\theta(X_i)\big]_{i=1}^{(k+1)N} : \theta \in \Theta\Big\}\Big|$$

be the number of distinct decision rules induced by the model on the design $(X_i)_{i=1}^{(k+1)N}$. We will prove

Theorem 2.1. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$r_2(\theta) \leq r_1(\theta) + \frac{d}{N} + \sqrt{\frac{2d(1+\frac{1}{k})r_1(\theta)}{N} + \frac{d^2}{N^2}},$$

where $d = \log\big[\mathcal{N}(X_1^{(k+1)N})\big] + \log(\epsilon^{-1})$.

Let us remind that when $|\mathcal{Y}| = 2$ and the VC dimension of $\{f_\theta; \theta \in \Theta\}$ is not greater than h,

$$d \leq h \log\left(\frac{e(k+1)N}{h}\right) + \log(\epsilon^{-1}).$$

Let us take some numerical example : when $N = 1000$, $h = 10$, $\epsilon = 0.01$ and $r_1(\theta) = 0.2$, we get $r_2(\theta) \leq 0.4872$ using $k = 4$ (whereas for $k = 1$ we get only $r_2(\theta) \leq 0.5098$, showing that increasing the shadow sample size is useful to get a bound less than 0.5)

Let us start the proof of theorem 2.1 with some notations and a few lemmas. Let $\chi_i = \mathbb{1}\big[Y_i \neq f_\theta(X_i)\big] \in \{0,1\}$. For any random variable $h : \Omega = \big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N} \to \mathbb{R}$ (we work on the canonical space), let the transformed random variable $\tau_i(h)$ be defined as

$$\tau_i(h) = \frac{1}{k+1} \sum_{j=0}^{k} h \circ \tau_i^j,$$

where $\tau_i^j : \Omega \to \Omega$ is defined by

$$\big[\tau_i^j(\omega)\big]_\ell = \begin{cases} \omega_{i+mn}, & \ell = i + [(m+j) \bmod (k+1)]N, \quad m = 0, \dots, k; \\ \omega_\ell, & \ell \notin \{i + mN : m = 0, \dots, k\}. \end{cases}$$

In other words, τ_i^j performs a circular permutation of the subset of indices $\{i + mN : m = 0, \dots, k\}$. Notice also that τ_i may be viewed as a regular conditional probability measure.

Definition 2.1. The joint distribution $\mathbb{P}$ is said to be partially exchangeable when for any $i = 1, \dots, N$, any $j = 0, \dots, k$, $\mathbb{P} \circ (\tau_i^j)^{-1} = \mathbb{P}$.

Equivalently, this means that for any bounded random variable h,

$$\mathbb{P}(h) = \mathbb{P}(h \circ \tau_i^1), \quad i = 1, \dots, N,$$

(since τ_i^j is the jth iterate of τ_i^1). As a result, any partially exchangeable distribution $\mathbb{P}$ is such that for any bounded random variable

$$\mathbb{P}(h) = \mathbb{P}\left\{\left[\bigcirc_{i=1}^N \tau_i\right](h)\right\},$$

where we have used the notation $\bigcirc_{i=1}^N \tau_i = \tau_1 \circ \tau_2 \circ \cdots \circ \tau_N$.

In the same way

Definition 2.2. A random variable $h : \big(\mathcal{X} \times \mathcal{Y}\big)^{(k+1)N} \to \mathbb{R}$ is said to be partially exchangeable when for any $i = 1, \dots, N$, $h \circ \tau_i^1 = h$.

Lemma 2.2. *For any $\theta \in \Theta$, any $\omega \in \big(\mathcal{X} \times \mathcal{Y}\big)^{(k+1)N}$, any positive partially exchangeable random variable λ, any partially exchangeable random variable η,*

$$\Big(\bigcirc_{i=1}^N \tau_i\Big)\Big\{\exp\Big[\lambda\big[r_2(\theta) - r_1(\theta)\big] - \eta\Big]\Big\}(\omega) \le \exp\Big[\tfrac{\lambda^2}{2N}\big[\tfrac{1}{k} r_1(\theta) + r_2(\theta)\big] - \eta\Big](\omega).$$

Proof.

$$\begin{aligned}
&\Big(\bigcirc_{i=1}^N \tau_i\Big)\Big\{\exp\Big[\lambda\big[r_2(\theta) - r_1(\theta)\big] - \eta\Big]\Big\} \\
&\qquad = \exp(-\eta) \prod_{i=1}^N \tau_i\left\{\exp\left(\frac{\lambda}{kN} \sum_{j=1}^k \chi_{i+jN} - \frac{\lambda}{N}\chi_i\right)\right\} \\
&\qquad = \exp(-\eta) \prod_{i=1}^N \exp\left(\frac{\lambda}{kN} \sum_{j=0}^k \chi_{i+jN}\right) \prod_{i=1}^N \tau_i\left\{\exp\left(-\frac{(k+1)\lambda}{kN}\chi_i\right)\right\}.
\end{aligned}$$

Let $p_i = \frac{1}{k+1} \sum_{j=0}^k \chi_{i+jN}$. Let χ be the identity (seen as the canonical process) on $\{0,1\}$ and B_p be the Bernoulli distribution on $\{0,1\}$ with parameter p, namely let $B_p(1) = 1 - B_p(0) = p$. It is easily seen that

$$\log\left\{\tau_i\left[\exp\left(-\frac{(k+1)\lambda}{kN}\chi_i\right)\right]\right\} = \log\left\{B_{p_i}\left[\exp\left(-\frac{(k+1)\lambda}{kN}\chi\right)\right]\right\}.$$

Moreover this last quantity can be bounded in the following way.

$$\log\Big\{B_p\big[\exp(-\alpha\chi)\big]\Big\} = -\alpha B_p(\chi) + \int_0^\alpha (1-\beta)\, \mathbb{V}\mathrm{ar}_{B_{f(\beta)}}(\chi) d\beta.$$

This is the Taylor expansion of order two of $\alpha \mapsto \log\Big\{B_p\big[\exp(-\alpha\chi)\big]\Big\}$, where

$$f(\beta) = \frac{B_p\big[\chi\exp(-\beta\chi)\big]}{B_p\big[\exp(-\beta\chi)\big]} = \frac{p\exp(-\beta)}{(1-p)+p\exp(-\beta)} \leq p.$$

Thus

$$\mathbb{V}\mathrm{ar}_{B_{f(\beta)}}(\chi) = f(\beta)\big[1-f(\beta)\big] \leq (p\wedge\tfrac{1}{2})\big[1-(p\wedge\tfrac{1}{2})\big] \leq (1-\tfrac{1}{k+1})p = \tfrac{k}{k+1}p,$$

for any $p \in \big(\frac{1}{k+1}\mathbb{N}\big)\cap[0,1]$. Hence

$$\log\Big\{B_p\big[\exp(-\alpha\chi)\big]\Big\} \leq -\alpha p + \frac{k\alpha^2}{2(k+1)}p,$$

and

$$\tau_i\bigg[\exp\bigg(-\frac{(k+1)\lambda}{kN}\chi_i\bigg)\bigg] \leq \exp\bigg(-\frac{(k+1)\lambda}{kN}p_i + \frac{(k+1)\lambda^2}{2kN^2}p_i\bigg).$$

Therefore

$$\begin{aligned}\Big(\bigcirc_{i=1}^N\tau_i\Big)\Big\{\exp\Big[\lambda\big[r_2(\theta)-r_1(\theta)\big]-\eta\Big]\Big\} &\leq \exp(-\eta)\exp\bigg(\frac{(k+1)\lambda^2}{2kN^2}\sum_{i=1}^N p_i\bigg)\\ &= \exp\bigg(\frac{\lambda^2}{2kN^2}\sum_{i=1}^{(k+1)N}\chi_i-\eta\bigg) = \exp\Big\{\frac{\lambda^2}{2N}\big[\tfrac{1}{k}r_1(\theta)+r_2(\theta)\big]-\eta\Big\}.\end{aligned}$$

□

Lemma 2.3. *For any $\theta\in\Theta$, for any positive partially exchangeable random variable λ, for any partially exchangeable random variable η,*

$$\mathbb{P}\Big\{\exp\Big[\lambda\big[r_2(\theta)-r_1(\theta)\big]-\eta\Big]\Big\} \leq \mathbb{P}\Big\{\exp\Big[\tfrac{\lambda^2}{2N}\big[\tfrac{1}{k}r_1(\theta)+r_2(\theta)\big]-\eta\Big]\Big\}.$$

Remark 2.1. Let us notice that we do not need integrability conditions, and that the previous inequality between expectations of positive random variables holds in $\mathbb{R}_+\cup\{+\infty\}$, meaning that both members may be equal to $+\infty$.

Remark 2.2. We can take $\eta = \log(\epsilon^{-1}) + \frac{\lambda^2}{2N}\big[\frac{1}{k}r_1(\theta)+r_2(\theta)\big]$ to get

$$\mathbb{P}\Big\{\exp\Big[\lambda\big[r_2(\theta)-r_1(\theta)\big]-\tfrac{\lambda^2}{2N}\big[\tfrac{1}{k}r_1(\theta)+r_2(\theta)\big]+\log(\epsilon)\Big]\Big\} \leq \epsilon.$$

Proof. According to the previous lemma,

$$\begin{aligned}&\mathbb{P}\Big\{\exp\Big[\lambda\big[r_2(\theta)-r_1(\theta)\big]-\eta\Big]\Big\}\\ &\qquad= \mathbb{P}\bigg\{\Big(\bigcirc_{i=1}^N\tau_i\Big)\Big\{\exp\Big[\lambda\big[r_2(\theta)-r_1(\theta)\big]-\eta\Big]\Big\}\bigg\}\\ &\qquad\leq \mathbb{P}\bigg\{\exp\bigg[\frac{\lambda^2}{2N}\big[\tfrac{1}{k}r_1(\theta)+r_2(\theta)\big]-\eta\bigg]\bigg\}.\end{aligned}$$

□

Let us now consider some *partially exchangeable prior distribution* $\pi\in\mathcal{M}_+^1(\Theta)$:

Definition 2.3. A regular conditional probability distribution $\pi : \big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N} \to \mathcal{M}_+^1(\Theta,\mathcal{T})$ is said to be partially exchangeable when for any $i = 1,\dots,N$, any $\omega \in \big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N}$, $\pi\big[\tau_i^1(\omega)\big] = \pi(\omega)$, this being an equality between probability measures in $\mathcal{M}_+^1(\Theta,\mathcal{T})$.

In the following, λ and η will be random variables depending on the parameter θ. We will say that a real random variable $h : \big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N}\times\Theta \to \mathbb{R}$ is partially exchangeable when $h(\omega,\theta) = h\big[\tau_i^1(\omega),\theta\big]$, $i = 1,\dots,N$, $\theta\in\Theta$, $\omega\in\big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N}$.

Lemma 2.4. *For any partially exchangeable prior distribution π, any positive partially exchangeable random variable $\lambda : \big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N}\times\Theta\to\mathbb{R}$, and any partially exchangeable random threshold function $\eta : \big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N}\times\Theta\to\mathbb{R}$,*

$$\mathbb{P}\biggl\{\pi\Bigl\{\exp\Bigl[\lambda(\theta)\bigl[r_2(\theta)-r_1(\theta)\bigr]-\eta(\theta)\Bigr]\Bigr\}\biggr\} \\ \leq \mathbb{P}\biggl\{\pi\Bigl\{\exp\Bigl[\frac{\lambda(\theta)^2}{2N}\bigl[\tfrac{1}{k}r_1(\theta)+r_2(\theta)\bigr]-\eta(\theta)\Bigr]\Bigr\}\biggr\}.$$

Proof. It is a consequence of lemma 2.2 and of the following identities:

$$\mathbb{P}\biggl\{\pi\Bigl\{\exp\Bigl[\lambda(\theta)\bigl[r_2(\theta)-r_1(\theta)\bigr]-\eta(\theta)\Bigr]\Bigr\}\biggr\} \\ = \mathbb{P}\biggl\{\Bigl(\bigcirc_{i=1}^N\tau_i\Bigr)\Bigl(\pi\Bigl\{\exp\Bigl[\lambda(\theta)\bigl[r_2(\theta)-r_1(\theta)\bigr]-\eta(\theta)\Bigr]\Bigr\}\Bigr)\biggr\} \\ = \mathbb{P}\biggl\{\pi\Bigl\{\bigl(\bigcirc_{i=1}^N\tau_i\bigr)\exp\Bigl[\lambda(\theta)\bigl[r_2(\theta)-r_1(\theta)\bigr]-\eta(\theta)\Bigr]\Bigr\}\biggr\}.$$

Indeed for any positive random variable $h : \big(\mathcal{X}\times\mathcal{Y}\big)^{(k+1)N}\times\Theta\to\mathbb{R}$,

$$\pi(h)\circ\tau_i^j = (\pi\circ\tau_i^j)(h\circ\tau_i^j) = \pi(h\circ\tau_i^j).$$

Thus

$$\tau_i\bigl[\pi(h)\bigr] = \frac{1}{k+1}\sum_{j=0}^k\pi\bigl(h\circ\tau_i^j\bigr) = \pi\biggl(\frac{1}{k+1}\sum_{j=0}^k h\circ\tau_i^j\biggr) = \pi\bigl(\tau_i h\bigr).$$

□

As a consequence, we get the following learning theorem:

Theorem 2.5. *For any partially exchangeable prior distribution π, any positive partially exchangeable random variable λ, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\rho\in\mathcal{M}_+^1(\Theta)$,*

$$\rho\bigl[\lambda(\theta)r_2(\theta)\bigr]-\rho\bigl[\lambda(\theta)r_1(\theta)\bigr] \leq \rho\biggl\{\frac{\lambda(\theta)^2}{2N}\Bigl[\tfrac{1}{k}\bigl[r_1(\theta)+r_2(\theta)\bigr]\Bigr]\biggr\}+\mathcal{K}(\rho,\pi)+\log(\epsilon^{-1}).$$

Proof. Take $\eta(\theta) = \frac{\lambda(\theta)^2}{2N}\bigl[\frac{1}{k}r_1(\theta)+r_2(\theta)\bigr]+\log(\epsilon^{-1})$ and notice that it is indeed a partially exchangeable threshold function.

Thus

$$\mathbb{P}\Big\{ \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho\big[\lambda(\theta)r_2(\theta)\big] - \rho\big[\lambda(\theta)r_1(\theta)\big] - \rho\Big[\frac{\lambda(\theta)^2}{2N}\Big\{\tfrac{1}{k}\rho\big[r_1(\theta)\big] + \rho\big[r_2(\theta)\big]\Big\}\Big] - \mathcal{K}(\rho,\pi) + \log(\epsilon) \le 0\Big\}$$
$$= \mathbb{P}\Big\{\log\Big\{\pi\Big[\exp\big\{\lambda(\theta)\big[r_2(\theta)-r_1(\theta)\big] - \tfrac{\lambda(\theta)^2}{2N}\big[\tfrac{1}{k}r_1(\theta)+r_2(\theta)\big] + \log(\epsilon)\big\}\Big]\Big\} \le 0\Big\}$$
$$\le \mathbb{P}\Big\{\pi\Big[\exp\Big\{\lambda(\theta)\big[r_2(\theta)-r_1(\theta)\big] - \tfrac{\lambda(\theta)^2}{2N}\big[\tfrac{1}{k}r_1(\theta)+r_2(\theta)\big] + \log(\epsilon)\Big\}\Big]\Big\} \le \epsilon.$$

We have used the identity $\log\Big\{\pi\big[\exp(h)\big]\Big\} = \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho(h) - \mathcal{K}(\rho,\pi)$. See for instance [4, pages 159-160] or [5, lemma 4.2] for a proof. □

Let us consider the map $\Psi : \Theta \to \mathcal{Y}^{(k+1)N}$ which restricts each classification rule to the design: $\Psi(\theta) = \big[f_\theta(X_i)\big]_{i=1}^{(k+1)N}$. Let Θ/Ψ be the set of components of Θ for the equivalence relation $\{(\theta_1,\theta_2)\in\Theta^2; \Psi(\theta_1)=\Psi(\theta_2)\}$. Let $c : \{0,1\}^\Theta \to \Theta$ be such that $c(\theta')\in\theta'$ for each $\theta'\subset\Theta$ (the function c chooses some element from any subset of Θ). Let $\Theta' = c\big(\Theta/\Psi\big)$. Let us note that Ψ and therefore Θ/Ψ and Θ' are exchangeable random objects. Let

$$\pi = \frac{1}{|\Theta'|}\sum_{\theta\in\Theta'}\delta_\theta$$

be the uniform distribution on the finite subset Θ' of Θ.

Applying theorem 2.5 to π, and $\rho=\delta_\theta$, we get that for any positive partially exchangeable random variable λ, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta'\in\Theta'$,

$$\lambda(\theta')r_2(\theta') - \lambda(\theta')r_1(\theta') \le \frac{\lambda(\theta')^2}{2N}\Big[\tfrac{1}{k}r_1(\theta')+r_2(\theta')\Big] + \log\big|\Theta'\big| + \log(\epsilon^{-1}).$$

Let us choose

$$\lambda(\theta) = \left(\frac{2N\log\Big(\frac{|\Theta'|}{\epsilon}\Big)}{\frac{1}{k}r_1(\theta)+r_2(\theta)}\right)^{1/2},$$

with the convention that when $\frac{1}{k}r_1(\theta)+r_2(\theta)=0$, then $\lambda r_2(\theta)=\lambda r_1(\theta)=0$. This is legitimate, since $|\Theta'|$ and $\frac{1}{k}r_1(\theta')+r_2(\theta')$ are exchangeable random variables, and since when $\frac{1}{k}r_1(\theta)+r_2(\theta)=0$, then $r_1(\theta)=r_2(\theta)=0$.

Thus, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta'\in\Theta'$,

$$r_2(\theta') - r_1(\theta') \le \left(\frac{2\log\big(\frac{|\Theta'|}{\epsilon}\big)\big[\frac{1}{k}r_1(\theta')+r_2(\theta')\big]}{N}\right)^{1/2}.$$

Now we can remark that for each $\theta\in\Theta$, $\theta' = c\big[\Psi(\theta)\big]$ is such that $f_{\theta'}(X_i) = f_\theta(X_i)$, for $i=1,\dots,(k+1)N$. Therefore $r_1(\theta)=r_1(\theta')$ and $r_2(\theta)=r_2(\theta')$.

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

$$r_2(\theta) - r_1(\theta) \leq \left(\frac{2\log\left(\frac{|\Theta'|}{\epsilon}\right)\left[\frac{1}{k} r_1(\theta) + r_2(\theta)\right]}{N} \right)^{1/2}. \tag{2.1}$$

Putting for short $d = \log\left(\frac{|\Theta'|}{\epsilon}\right)$ and solving inequality (2.1) with respect to $r_2(\theta)$ proves theorem 2.1.

Note that we have in fact proved a more general version of theorem 2.1, where d can be taken to be $d = -\log\bigl[\bar\pi(\theta)\epsilon\bigr]$, where

$$\bar\pi(\theta) = \sup\{\pi(\theta') : \theta' \in \Theta, \Psi(\theta') = \Psi(\theta)\},$$

for any choice of partially exchangeable prior probability distribution π.

3. Improvement of the variance term

We will first improve the variance term in lemma 2.2 when $k=1$, and $\mathbb{P}$ is fully exchangeable. We will deal afterwards with the general case.

Theorem 3.1. *For any exchangeable probability distribution $\mathbb{P}$, with $\mathbb{P}$ probability $1-\epsilon$, for any $\theta \in \Theta$,*

$$r_2(\theta) \leq r_1(\theta) + \frac{d}{N}\bigl[1 - 2r_1(\theta)\bigr] + \sqrt{\frac{4d}{N}\bigl[1 - r_1(\theta)\bigr] r_1(\theta) + \frac{d^2}{N^2}\bigl[1 - 2r_1(\theta)\bigr]^2},$$

where $d = \inf\Bigl\{ -\log\bigl[\pi(\theta')\epsilon\bigr] : \theta' \in \Theta, \Psi(\theta') = \Psi(\theta) \Bigr\}$.

Let us pursue our numerical example : assuming that $|\mathcal{Y}| = 2$, $N = 1000$, $h = 10$, $\epsilon = 0.01$ and $r_1(\theta) = 0.2$, we get that $r_2(\theta) \leq 0.453$.

Proof. Proving theorem 3.1 will require some lemmas.

Let

$$\tau(h)(\omega) = \frac{1}{(2N)!} \sum_{\sigma \in \mathfrak{S}_{2N}} h(\omega \circ \sigma), \quad \omega \in \Omega,$$

where $\mathfrak{S}_{2N}$ is the set of permutations of $\{1, \dots, 2N\}$ and where $(\omega \circ \sigma)_i = \omega_{\sigma(i)}$. For any $\omega \in \Omega$, any $\sigma \in \mathfrak{S}_N$, let $\omega_{2,\sigma}$ be defined as

$$(\omega_{2,\sigma})_i = \begin{cases} \omega_i, & 1 \leq i \leq N, \\ \omega_{\sigma(i-N)}, & N < i \leq 2N. \end{cases}$$

Let

$$\tau'(h)(\omega) = \frac{1}{N!} \sum_{\sigma \in \mathfrak{S}_N} h(\omega_{2,\sigma}).$$

Let us remark that $\tau = \tau \circ \tau'$, and that $\tau'\bigl[r_k(\theta)\bigr] = r_k(\theta)$, $k = 1, 2$.

Moreover, we know from the previous section that $\tau\bigl[\exp(U)\bigr](\omega) \leq \epsilon$, where

$$U = \lambda\bigl[r_2(\theta) - r_1(\theta)\bigr] - \frac{\lambda^2}{2N^2} \sum_{i=1}^{N} \bigl(\chi_{i+N} - \chi_i\bigr)^2 + \log(\epsilon).$$

Thus $\tau\Bigl\{\exp\bigl[\tau'(U)\bigr]\Bigr\} \leq \tau \circ \tau'\bigl[\exp(U)\bigr] = \tau\bigl[\exp(U)\bigr] \leq \epsilon$, from the convexity of the exponential function and the fact that τ' is a (regular) conditional probability measure.

But $$\tau'(U) = \lambda\big[r_2(\theta) - r_1(\theta)\big] - \frac{\lambda^2}{2N}\tau'(V) + \log(\epsilon^{-1}),$$
where $V = \frac{1}{N}\sum_{i=1}^{N}(\chi_{i+N} - \chi_i)^2$. Noticing that

$$\tau'(V) = \frac{1}{N}\sum_{i=1}^{N}(\chi_i + \chi_{i+N}) - 2\left(\frac{1}{N}\sum_{i=1}^{N}\chi_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\chi_{i+N}\right) = r_1(\theta) + r_2(\theta) - 2r_1(\theta)r_2(\theta),$$

we get

Lemma 3.2. *For any exchangeable random variable η,*

$$\tau\left\{\exp\left[\lambda\big[r_2(\theta) - r_1(\theta)\big] - \frac{\lambda^2}{2N}\Big[r_1(\theta) + r_2(\theta) - 2r_1(\theta)r_2(\theta)\Big] - \eta\right]\right\}(\omega) \leq \exp(-\eta)(\omega), \qquad \omega \in \Omega.$$

As a consequence,

Lemma 3.3. *For any exchangeable probability distribution $\mathbb{P}$, any exchangeable prior distribution π, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$r_2(\theta) \leq r_1(\theta) + \frac{\lambda}{2N}\Big[r_1(\theta) + r_2(\theta) - 2r_1(\theta)r_2(\theta)\Big] + \frac{d}{\lambda},$$

where $d = \inf\Big\{-\log\big[\pi(\theta')\epsilon\big] : \theta' \in \Theta, \Psi(\theta') = \Psi(\theta)\Big\}$.

Remark 3.1. As a special case, we can take $d = \log\big[\mathcal{N}(X_1^{2N})\big] - \log(\epsilon)$. This corresponds to the case when π is chosen to be the uniform distribution on Θ', using the remark that each f_θ, $\theta \in \Theta$ coincides with some $f_{\theta'}$, $\theta' \in \Theta'$ on the design $\{X_i : i = 1, \ldots, 2N\}$.

We would like to prove a little more, showing that it is legitimate to take in the previous equation

$$\lambda = \left(\frac{2Nd}{r_1(\theta) + r_2(\theta) - 2r_1(\theta)r_2(\theta)}\right)^{1/2} = \sqrt{\frac{2Nd}{\tau'(V)}}.$$

This is not so clear, since this quantity is not (even partially) exchangeable. Anyhow we can write the following:

$$\sqrt{\frac{2Nd}{\tau'(V)}}\big|r_2(\theta) - r_1(\theta)\big| \leq \tau'(V^{-1/2})\sqrt{2Nd}\big|r_2(\theta) - r_1(\theta)\big| = \tau'\left(\sqrt{\frac{2Nd}{V}}\big|r_2(\theta) - r_1(\theta)\big|\right),$$

because $r \mapsto r^{-1/2}$ is convex. Moreover, using successively the fact that $\tau'(V)$ is a symmetric function of $r_1(\theta)$ and $r_2(\theta)$, the fact that cosh is an even function, the previous inequality, the convexity of cosh, the invariance $\tau = \tau \circ \tau'$, the invariance of V under $\omega \mapsto \bigcirc_{i=1}^{N}\tau_i^1(\omega)$, and the fact that V is almost surely constant under each τ_i, we get the following chain of inequalities:

$$
\begin{aligned}
&\tau\biggl\{\exp\biggl[\sqrt{\tfrac{2Nd}{\tau'(V)}}\bigl[r_2(\theta)-r_1(\theta)\bigr]-d+\log(\epsilon)\biggr]\biggr\}\\
&\qquad=\tau\biggl\{\cosh\biggl[\sqrt{\tfrac{2Nd}{\tau'(V)}}\bigl[r_2(\theta)-r_1(\theta)\bigr]\biggr]\biggr\}\exp\bigl[-d+\log(\epsilon)\bigr]\\
&\qquad=\tau\biggl\{\cosh\biggl[\sqrt{\tfrac{2Nd}{\tau'(V)}}\bigl|r_2(\theta)-r_1(\theta)\bigr|\biggr]\biggr\}\exp\bigl[-d+\log(\epsilon)\bigr]\\
&\qquad\le\tau\biggl\{\cosh\biggl[\tau'\biggl[\sqrt{\tfrac{2Nd}{V}}\bigl|r_2(\theta)-r_1(\theta)\bigr|\biggr]\biggr]\biggr\}\exp\bigl[-d+\log(\epsilon)\bigr]\\
&\qquad\le\tau\biggl\{\tau'\biggl[\cosh\biggl[\sqrt{\tfrac{2Nd}{V}}\bigl|r_2(\theta)-r_1(\theta)\bigr|\biggr]\biggr]\biggr\}\exp\bigl[-d+\log(\epsilon)\bigr]\\
&\qquad=\tau\biggl\{\cosh\biggl[\sqrt{\tfrac{2Nd}{V}}\bigl[r_2(\theta)-r_1(\theta)\bigr]\biggr]\biggr\}\exp\bigl[-d+\log(\epsilon)\bigr]\\
&\qquad=\tau\biggl\{\exp\biggl[\sqrt{\tfrac{2Nd}{V}}\bigl[r_2(\theta)-r_1(\theta)\bigr]-d+\log(\epsilon)\biggr]\biggr\}\\
&\qquad=\tau\circ\bigcirc_{i=1}^N\tau_i\biggl\{\exp\biggl[\sqrt{\tfrac{2Nd}{V}}\bigl[r_2(\theta)-r_1(\theta)\bigr]-d+\log(\epsilon)\biggr]\biggr\}\le\epsilon.
\end{aligned}
$$

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

$$
r_2(\theta)\le r_1(\theta)+\sqrt{\frac{2d\tau'(V)}{N}}=r_1(\theta)+\sqrt{\frac{2d\bigl[r_1(\theta)+r_2(\theta)-2r_1(\theta)r_2(\theta)\bigr]}{N}}.
$$

Solving this inequality in $r_2(\theta)$ ends the proof of theorem 3.1. □

In the general case when $\mathbb{P}$ is only partially exchangeable and k is arbitrary, we will obtain the following

Theorem 3.4. *Let* $d=\inf\Bigl\{-\log\bigl[\pi(\theta)\epsilon\bigr]:\theta'\in\Theta,\Psi(\theta')=\Psi(\theta)\Bigr\}$ *and*

$$
\begin{aligned}
B(\theta)=\left(1+\frac{2d}{N}\right)^{-1}\biggl\{&r_1(\theta)+\frac{d}{N}\Bigl\{1+k^{-1}\bigl[1-2r_1(\theta)\bigr]\Bigr\}\\
&+(1+k^{-1})\sqrt{\frac{2d}{N}r_1(\theta)\bigl[1-r_1(\theta)\bigr]+\frac{d^2}{N^2}}\biggr\}.
\end{aligned}
$$

For any partially exchangeable probability distribution $\mathbb{P}$*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any* $\theta\in\Theta$ *such that* $r_1(\theta)<1/2$ *and* $B(\theta)\le 1/2$*,* $r_2(\theta)\le B(\theta)$*.*

As a special case, the theorem holds with $d=\log\bigl[\mathcal{N}\bigl(X_1^{(k+1)N}\bigr)\bigr]+\log(\epsilon^{-1})$. When using a set of binary classification rules $\{f_\theta:\theta\in\Theta\}$ whose VC dimension is not greater than h, we can use the bound $d\le h\log\left(\dfrac{e(k+1)N}{h}\right)-\log(\epsilon)$. The result is satisfactory when k is large, because in this case $(1+k^{-1})$ is close to one. This will be useful in the inductive case.

Let us carry on our numerical example in the binary classification case: taking $N=1000$, $h=10$, $\epsilon=0.01$ and $r_1(\theta)=0.2$, we get a bound $B(\theta)\le 0.4203$ for values of k ranging from 15 to 18, showing that increasing the size of the shadow sample has an increased impact when the improved variance term is used.

Proof. Let $\Phi(p) = (p \wedge \frac{1}{2})\big[1 - (p \wedge \frac{1}{2})\big]$. This is obviously a concave function. We have proved that

$$\left(\bigcirc_{i=1}^N \tau_i\right)\left\{\exp\Big[\lambda\big[r_2(\theta) - r_1(\theta)\big] - \eta\Big]\right\} \leq \exp\left[\frac{(1+k^{-1})^2\lambda^2}{2N^2}\sum_{i=1}^N \Phi(p_i) - \eta\right].$$

As

$$\frac{1}{N}\sum_{i=1}^N \Phi(p_i) \leq \Phi\left(\frac{1}{N}\sum_{i=1}^N p_i\right) = \Phi\left(\frac{r_1(\theta) + kr_2(\theta)}{k+1}\right),$$

this shows that

$$\left(\bigcirc_{i=1}^N \tau_i\right)\left\{\exp\Big[\lambda\big[r_2(\theta) - r_1(\theta)\big] - \eta\Big]\right\} \leq \exp\left[\frac{(1+k^{-1})^2\lambda^2}{2N}\Phi\left(\frac{r_1(\theta) + kr_2(\theta)}{k+1}\right) - \eta\right].$$

Taking $\eta = \dfrac{(1+k^{-1})^2\lambda^2}{2N}\Phi\left(\dfrac{r_1(\theta) + kr_2(\theta)}{k+1}\right) - \log(\epsilon)$, and

$$\lambda = \left(\frac{2Nd}{(1+k^{-1})^2\Phi\left(\frac{r_1(\theta)+kr_2(\theta)}{k+1}\right)}\right)^{1/2},$$

where $d = \inf\Big\{-\log\big[\pi(\theta)\epsilon\big] : \theta' \in \Theta, \Psi(\theta') = \Psi(\theta)\Big\}$, we get that with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,

$$r_2(\theta) - r_1(\theta) \leq \left(\frac{2(1+k^{-1})^2\Phi\left(\frac{r_1(\theta)+kr_2(\theta)}{k+1}\right)d}{N}\right)^{1/2}.$$

Solving this inequality in $r_2(\theta)$ ends the proof of theorem 3.4. □

4. The inductive setting

We will integrate with respect to $\mathbb{P}(\cdot|Z_1^N)$ theorem 2.1 and its variants. Let us start with theorem 3.4. Let us consider the non identically distributed independent case, assuming thus that $\mathbb{P} = \Big[\bigotimes_{i=1}^N P_i\Big]^{\otimes(k+1)}$.

Let $$R(\theta) = \frac{1}{N}\sum_{i=1}^N P_i\big[Y_i \neq f_\theta(X_i)\big]$$

and $$\overline{r}(\theta) = \frac{r_1(\theta) + kr_2(\theta)}{k+1}.$$

Let $$\mathbb{P}'(h) = \mathbb{P}(h|Z_1^N).$$

Lemma 4.1. *For any partially exchangeable prior distribution π, any partially exchangeable positive function $\zeta : \Theta \to \mathbb{R}_+^*$,*

$$\mathbb{P}\biggl\{\sup_{\theta\in\Theta}\int_{\lambda=0}^{+\infty}\zeta\exp\biggl[\lambda\bigl[R(\theta)-r_1(\theta)-\zeta\bigr] - \frac{(1+k^{-1})^2\lambda^2}{2N}\Phi\biggl(\frac{r_1(\theta)+kR(\theta)}{k+1}\biggr)+\mathbb{P}'\Bigl[\log\bigl[\bar\pi(\theta)\epsilon\bigr]\Bigr]\biggr]d\lambda\biggr\}\leq\epsilon,$$

where $\bar\pi(\theta)=\sup\Bigl\{\pi(\theta'):\theta'\in\Theta,\Psi(\theta')=\Psi(\theta)\Bigr\}$.

Proof. Let

$$U'=\lambda\bigl[R(\theta)-r_1(\theta)-\zeta\bigr]-\frac{(1+k^{-1})^2\lambda^2}{2N}\Phi\biggl(\frac{r_1(\theta)+kR(\theta)}{k+1}\biggr)+\mathbb{P}'\Bigl[\log\bigl[\bar\pi(\theta)\epsilon\bigr]\Bigr].$$

Let

$$U=\lambda\bigl[r_2(\theta)-r_1(\theta)-\zeta\bigr]-\frac{(1+k^{-1})^2\lambda^2}{2N}\Phi\bigl[\overline{r}(\theta)\bigr]+\log\bigl[\bar\pi(\theta)\epsilon\bigr].$$

The function Φ being concave,

$$\zeta\exp(U')\leq\zeta\exp\bigl[\mathbb{P}'(U)\bigr]\leq\mathbb{P}'\bigl[\zeta\exp(U)\bigr].$$

Thus

$$\begin{aligned}\sup_\theta\int_{\lambda=0}^{+\infty}\zeta\exp(U')d\lambda&\leq\sup_{\theta\in\Theta}\int_{\lambda=0}^{+\infty}\mathbb{P}'\bigl[\zeta\exp(U)\bigr]d\lambda\\&\leq\sup_{\theta\in\Theta}\mathbb{P}'\biggl(\int_{\lambda=0}^{+\infty}\zeta\exp(U)d\lambda\biggr)\leq\mathbb{P}'\biggl(\int_{\lambda=0}^{+\infty}\sup_{\theta\in\Theta}\bigl[\zeta\exp(U)\bigr]d\lambda\biggr)\end{aligned}$$

Moreover

$$\sup_{\theta\in\Theta}\bigl[\zeta\exp(U)\bigr]\leq\pi\bigl[\zeta\exp(S)\bigr],$$

where

$$S=U-\log\bigl[\pi(\theta)\bigr]=\lambda\bigl[r_2(\theta)-r_1(\theta)-\zeta\bigr]-\frac{(1+k^{-1})^2\lambda^2}{2N}\Phi\bigl[\overline{r}(\theta)\bigr]+\log(\epsilon).$$

Thus

$$\begin{aligned}\mathbb{P}\biggl(\sup_{\theta\in\Theta}\int_{\lambda=0}^{+\infty}\zeta\exp(U')d\lambda\biggr)&\leq\mathbb{P}\biggl[\mathbb{P}'\biggl(\int_{\lambda=0}^{+\infty}\pi\bigl[\zeta\exp(S)\bigr]d\lambda\biggr)\biggr]\\&=\mathbb{P}\biggl(\int_{\lambda=0}^{+\infty}\pi\bigl[\zeta\exp(S)\bigr]d\lambda\biggr)\\&=\mathbb{P}\biggl[\Bigl(\bigcirc_{i=1}^N\tau_i\Bigr)\biggl(\int_{\lambda=0}^{+\infty}\pi\bigl[\zeta\exp(S)\bigr]d\lambda\biggr)\biggr]\\&=\mathbb{P}\biggl\{\pi\biggl[\int_{\lambda=0}^{+\infty}\zeta\Bigl(\bigcirc_{i=1}^N\tau_i\Bigr)\bigl[\exp(S)\bigr]d\lambda\biggr]\biggr\}.\end{aligned}$$

But we have established on the occasion of the proof of theorem 3.4 that

$$\Bigl(\bigcirc_{i=1}^N\tau_i\Bigr)\bigl[\exp(S)\bigr]\leq\epsilon\exp(-\zeta\lambda).$$

This proves that

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\int_{\lambda=0}^{+\infty}\zeta\exp(U')d\lambda\right)\leq\epsilon,$$

as stated in the lemma. □

Theorem 4.2. *Let*

$$B(\theta)=\left(1+\frac{2d'}{N}\right)^{-1}\left\{r_1(\theta)+\frac{d'}{N}+\sqrt{\frac{2d'r_1(\theta)\big[1-r_1(\theta)\big]}{N}+\frac{d'^2}{N^2}}\right\},$$

where

$$d'=d(1+k^{-1})^2\left(1-\frac{\log(\alpha)}{2d}+\frac{\alpha}{\sqrt{\pi d}}\right)^2,$$

and

$$d=-\mathbb{P}\Big\{\log\big[\epsilon\bar{\pi}(\theta)\big]|Z_1^N\Big\}.$$

Let us notice that it covers the case when

$$d=\mathbb{P}\Big\{\log\big[\mathcal{N}(X_1^{(k+1)N})\big]|Z_1^N\Big\}+\log(\epsilon^{-1}).$$

In this case, when $|\mathcal{Y}|=2$ and the set of classification rules has a VC dimension not greater than h,

$$d\leq h\log\left(\frac{e(k+1)N}{h}\right)+\log(\epsilon^{-1}).$$

With $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$, $R(\theta)\leq B(\theta)$ when $r_1(\theta)<1/2$ and $B(\theta)\leq 1/2$.

In the case when the model has a VC dimension not greater than h, we can bound as mentioned in the theorem the random variable d with the constant

$$d^\star=h\log\left(\frac{e(k+1)N}{h}\right)+\log(\epsilon^{-1}).$$

We can then optimize the choice of α by taking $\alpha=\frac{1}{2}\sqrt{\frac{\pi}{d^\star}}$. This leads to

$$d'\leq d^\star(1+k^{-1})^2\left[1+\frac{1}{2d^\star}\log\left(2e\sqrt{\frac{d^\star}{\pi}}\right)\right]^2.$$

We can also approximately optimize

$$(1+k^{-1})^2\log\left(\frac{eN(k+1)}{h}\right)$$

by taking $k=2\log\left(\frac{eN}{h}\right)$.

Let us resume our numerical example to illustrate theorem 4.2. Assume that $N=1000$, $h=10$ and $\epsilon=10^{-2}$. For $r_1(\theta)=0.2$, we get $B(\theta)\leq 0.4257$ for $k=19$. More generally, we get

$$B(\theta)\leq 0.828\left\{r_1(\theta)+0.105+\sqrt{0.209\big[1-r_1(\theta)\big]r_1(\theta)+0.011}\right\}.$$

For comparison, Vapnik's corollary 1.2 in the same situation gives a bound greater than 0.610, and therefore not significant (since a random classification has a better expected error rate of 0.5).

Proof. Let

$$V = (1+k^{-1})^2 \Phi\left(\frac{r_1(\theta)+kR(\theta)}{k+1}\right),$$
$$d = -\mathbb{P}'\Big\{\log\big[\bar\pi(\theta)\epsilon\big]\Big\},$$
$$\Delta = R(\theta) - r_1(\theta) - \zeta.$$

Let us remark that

$$U' = -\frac{V}{2N}\left(\lambda - \frac{N\Delta}{V}\right)^2 + \frac{N\Delta^2}{2V} - d.$$

Thus

$$\int_{\lambda=0}^{+\infty} \zeta \exp(U') d\lambda \geq \mathbb{1}(\Delta \geq 0) W,$$

where

$$W = \sqrt{\frac{\pi N}{2V}} \zeta \exp\left(\frac{N\Delta^2}{2V} - d\right).$$

Thus, according to the previous lemma,

$$\mathbb{P}\Big\{\sup_{\theta\in\Theta}\big[\mathbb{1}(\Delta\geq 0)W\big]\Big\} \leq \epsilon.$$

This proves that with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\sup_{\theta\in\Theta}\big[\mathbb{1}(\Delta\geq 0)W\big] \leq 1.$$

Translated into a logical statement this says that with $\mathbb{P}$ probability at least $1-\epsilon$, either $\Delta < 0$, or $\log(W) \leq 0$.

Let $V' = (1+k^{-1})\Phi\big[R(\theta)\big]$. Consider setting $\zeta = \alpha\sqrt{\frac{2V'}{\pi N}}$, where α is some positive real number.

We have proved that with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\frac{N\Delta^2}{2V} \leq d - \log(\alpha) + \frac{1}{2}\log\left(\frac{V}{V'}\right),$$

when $\Delta \geq 0$. But Φ is increasing and when $\Delta \geq 0$, $R(\theta) \geq r_1(\theta)$, thus in this case $V' \geq V$, and we can weaken and simplify our statement to

$$\frac{N\Delta^2}{2V'} \leq d - \log(\alpha).$$

Equivalently, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,

$$R(\theta) - r_1(\theta) \leq \sqrt{\frac{2V'd}{N}}\left(\sqrt{1 - \frac{\log(\alpha)}{d}} + \frac{\alpha}{\sqrt{\pi d}}\right)$$

Using the fact that $\sqrt{1+x} \leq 1 + \frac{x}{2}$, we get that

$$\big[R(\theta) - r_1(\theta)\big]^2 \leq \frac{2d'\Phi\big[R(\theta)\big]}{N},$$

where $d' = d(1+k^{-1})^2\left(1 - \frac{\log(\alpha)}{2d} + \frac{\alpha}{\sqrt{\pi d}}\right)^2$. Since $\Phi(R) = R(1-R)$ when $R \leq 1/2$, this can be solved in $R(\theta)$ in this case to end the proof of theorem 4.2. □

With a little more work we could have kept

$$\frac{N\Delta^2}{2V} \leq d - \log(\alpha),$$

leading to

$$R(\theta) - r_1(\theta) \leq \sqrt{\frac{2V\big[d-\log(\alpha)\big]}{N}} + \alpha\sqrt{\frac{2V'}{\pi N}},$$

$$\text{and } \big[R(\theta) - r_1(\theta)\big]^2 \leq \frac{2V\big[d-\log(\alpha)\big]}{N} + 4\alpha\frac{V'}{N}\sqrt{\frac{d-\log\alpha}{\pi}} + \alpha^2\frac{2V'}{\pi N}.$$

This leads to the following

Theorem 4.3. *Let us put*

$$\begin{aligned}
c &= 2\alpha\sqrt{\frac{d-\log(\alpha)}{\pi}} + \frac{\alpha^2}{\pi},\\
d_1 &= d - \log(\alpha) + (1+k^{-1})^2 c,\\
d_2 &= (1+k^{-1})\Big\{\big[d-\log(\alpha)\big]\big[1-\tfrac{2r_1(\theta)}{1+k}\big] + (1+k^{-1})c\Big\},\\
d_3 &= (1+k^{-1})^2\Big\{d - \log(\alpha) + c + 2\tfrac{c}{Nk^2}\big[d-\log(\alpha)\big]\Big\},\\
d_4 &= (1+k^{-1})\big[d - \log(\alpha) + (1+k^{-1})c\big].
\end{aligned}$$

Theorem 4.2 still holds when the bound $B(\theta)$ is strengthened to

$$B(\theta) = \left(1 + \frac{2d_1}{N}\right)^{-1}\left\{r_1(\theta) + \frac{d_2}{N} + \sqrt{\frac{2d_3 r_1(\theta)\big[1-r_1(\theta)\big]}{N} + \frac{{d_4}^2}{N^2}}\right\}.$$

On the previous numerical example ($N = 1000$, $h = 10$, $\epsilon = 10^{-2}$, $k = 19$, $\alpha = \frac{1}{2}\sqrt{\frac{\pi}{d^\star}}$, $r_1(\theta) = 0.2$), we get a bound $B(\theta) \leq 0.4248$, instead of $B(\theta) \leq 0.4257$, showing that the improvement brought to theorem 4.2 is not so strong, and therefore that theorem 4.2 is a satisfactory approximation of theorem 4.3.

Starting from lemma 2.2, we can make the same kind of computations taking $V = (1+k^{-1})\frac{r_1(\theta)+kR(\theta)}{k+1}$, to obtain that with $\mathbb{P}$ probability at least $1-\epsilon$,

$$R(\theta) - r_1(\theta) \leq \sqrt{\frac{2V'd}{N}}\left(1 - \frac{\log(\alpha)}{2d} + \frac{\alpha}{\sqrt{\pi d}}\right),$$

where $V' = (1+k^{-1})R(\theta)$. This proves the following

Theorem 4.4. *For any positive constant α, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq r_1(\theta) + \frac{d'}{N} + \sqrt{\frac{2d' r_1(\theta)}{N} + \frac{{d'}^2}{N^2}},$$

where $d' = (1+k^{-1})\left(1 - \frac{\log(\alpha)}{2d} + \frac{\alpha}{\sqrt{\pi d}}\right)^2 d.$

Our previous numerical application gives in this case a non significant bound $R(\theta) \leq 0.516$, (for the best value of $k = 9$), showing that the improvement of the variance term has a decisive impact when $r_1(\theta)$ is not small.

In the fully exchangeable case, when $k = 1$, a slightly better result can be obtained, using lemma 3.2, and thus putting

$$V = r_1(\theta) + R(\theta) - 2r_1(\theta)R(\theta),$$
$$V' = 2R(\theta)\big[1 - R(\theta)\big].$$

It leads to the following theorem

Theorem 4.5. *Let* $d' = d - \log(\alpha)$ *and*

$$c = 2\alpha\sqrt{\frac{d - \log(\alpha)}{\pi}} + \frac{\alpha^2}{\pi}.$$

Theorem 4.2 still holds when the bound is tightened to

$$B(\theta) = \left(1 + \frac{4c}{N}\right)^{-1} \left\{ r_1(\theta) + \big[1 - 2r_1(\theta)\big]\frac{d'}{N} + \frac{2c}{N} \right.$$
$$\left. + \sqrt{\frac{4(d'+c)r_1(\theta)\big[1 - r_1(\theta)\big]}{N} + \frac{{d'}^2}{N^2}\big[1 - 2r_1(\theta)\big]^2 + \frac{4c(d'+c)}{N^2}} \right\}.$$

Remark 4.1. Our previous numerical example gives in this case a bound $B(\theta) \leq 0.460$, (for $\alpha = \frac{1}{2}\sqrt{\frac{\pi}{d}}$). This shows that the improvement brought by a better variance term is significant, but that the optimization of the size of the shadow sample is also interesting.

Remark 4.2. Note that we can take $\alpha = 1$. In this case, $d' = d$ and $c = 2\sqrt{\frac{d}{\pi}} + \frac{1}{\pi}$. Note that we can also take $\alpha = d^{-1/2}$, leading to $d' = d + \frac{1}{2}\log(d)$ and

$$c = \frac{2}{\sqrt{\pi}}\sqrt{1 + \frac{\log(d)}{2d}} + \frac{1}{\pi d} \leq \frac{2}{\sqrt{\pi}}\left(1 + \frac{\log(d)}{4d}\right) + \frac{1}{\pi d} \leq \frac{3}{2}.$$

Remark 4.3. Note also that the bound can be weakened and simplified to

$$B(\theta) \leq r_1(\theta) + \big[1 - 2r_1(\theta)\big]\frac{d''}{N} + \sqrt{\frac{4d''r_1(\theta)\big[1 - r_1(\theta)\big]}{N} + \frac{{d''}^2}{N^2}\big[1 - 2r_1(\theta)\big]^2},$$

where $d'' = d' + 2c$. Taking $\alpha = d^{-1/2}$ gives $d'' \leq d + \frac{1}{2}\log(d) + 3$.

Another technical possibility to get inductive bounds is to choose some near optimal value for λ, instead of averaging over some exponential prior distribution on λ.

This leads to the following theorem

Theorem 4.6. *Let*

$$\bar{d} = \mathbb{P}\Big\{\log\big[\bar{\pi}(\theta)^{-1}\epsilon^{-1}\big]\Big\},$$
$$d = \mathbb{P}\Big\{\log\big[\bar{\pi}(\theta)^{-1}\epsilon^{-1}\big] | Z_1^N\Big\},$$
$$d' = \frac{1}{4}(1 + k^{-1})^2(\bar{d} + d)\Big(1 + \frac{d}{\bar{d}}\Big).$$

Theorem 4.2 still holds when the bound is tightened to

$$B(\theta) = \left(1 + \frac{2d'}{N}\right)^{-1} \left\{ r_1(\theta) + \frac{d'}{N} + \sqrt{\frac{2d'r_1(\theta)\big[1 - r_1(\theta)\big]}{N} + \frac{{d'}^2}{N^2}} \right\}.$$

Moreover, putting $d^\star = \operatorname{ess\,sup}_{\mathbb{P}} \log\big[\bar{\pi}(\theta)^{-1}\epsilon^{-1}\big]$, d' *can be replaced with* $d^\star(1+k^{-1})^2$ *in the previous bound. In the case of a VC class of dimension* h, $d^\star$ *can be bounded by* $h\log(\frac{eN}{h})$.

Following our numerical example ($N = 1000$, $h = 10$, $r_1(\theta) = 0.2$), we get an optimal value of $B(\theta) \leq 0.4213$ for k ranging from 17 to 19. This shows that in this case going from the transductive setting to the inductive one was done with an insignificant loss of 0.001. Although making use of a rather cumbersome flavor of entropy term in the general case, theorem 4.6 provides the tightest bound in the case of a VC class.

Proof. Starting from

$$\mathbb{P}\left\{\sup_{\theta\in\Theta}\exp\left[\lambda(\theta)\big[R(\theta)-r_1(\theta)\big] - \frac{\lambda(\theta)^2}{2N}(1+k^{-1})^2\Phi\left(\frac{r_1(\theta)+kR(\theta)}{1+k}\right) + \mathbb{P}'\Big\{\log\big[\bar{\pi}(\theta)\epsilon\big]\Big\}\right]\right\} \leq \epsilon,$$

we can choose

$$\lambda = \left[\frac{-2N\mathbb{P}\Big\{\log\big[\bar{\pi}(\theta)\epsilon\big]\Big\}}{(1+k^{-1})^2\Phi\big[R(\theta)\big]}\right]^{1/2}.$$

We get with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\lambda\big[R(\theta)-r_1(\theta)\big] \leq \bar{d}\frac{\Phi\left(\frac{r_1(\theta)+kR(\theta)}{1+k}\right)}{\Phi\big[R(\theta)\big]} + d.$$

We can then remark that whenever $R(\theta) \geq r_1(\theta)$, then $\frac{\Phi\left(\frac{r_1(\theta)+kR(\theta)}{1+k}\right)}{\Phi\big[R(\theta)\big]} \leq 1$, to get

$$R(\theta) - r_1(\theta) \leq (1+k^{-1})\frac{\bar{d}+d}{\sqrt{\bar{d}}}\sqrt{\frac{\Phi\big[R(\theta)\big]}{2N}}.$$

Solving this inequality in $R(\theta)$ ends the proof of theorem 4.6. □

In the same way, in the fully exchangeable case, starting from

$$\mathbb{P}\left\{\sup_{\theta\in\Theta}\exp\left[\lambda(\theta)\big[R(\theta)-r_1(\theta)\big] - \frac{\lambda(\theta)^2}{2N}\big[R(\theta)+r_1(\theta)-2r_1(\theta)R(\theta)\big] + d\right]\right\} \leq \epsilon,$$

we can take

$$\lambda(\theta) = \sqrt{\frac{N\bar{d}}{R(\theta)\big[1-R(\theta)\big]}},$$

to get

Theorem 4.7. *Let* $d' = \frac{1}{2}\bar{d}\left(1+\frac{d}{\bar{d}}\right)^2$, *and assume that* $\mathbb{P}$ *is fully exchangeable. Theorem 4.2 still holds when the bound is tightened to*

$$B(\theta) = \left(1+\frac{2d'}{N}\right)^{-1}\left\{r_1(\theta) + \frac{d'}{N} + \sqrt{\frac{2d' r_1(\theta)\big[1-r_1(\theta)\big]}{N} + \frac{d'^2}{N^2}}\right\}.$$

Moreover, putting $d^\star = \operatorname{ess\,sup}_{\mathbb{P}} \log\big[\bar{\pi}(\theta)^{-1}\epsilon^{-1}\big]$, d' *can be replaced with* $2d^\star$ *in the previous bound. In the case of a VC class of dimension* h, $d^\star$ *can be bounded by* $h\log\big(\frac{eN}{h}\big)$.

Our numerical example ($N = 1000$, $h = 10$, $\epsilon = 0.0.1$ and $r_1(\theta) = 0.2$), gives a bound $B(\theta) \leq 0.445$.

5. Using relative bounds

Relative bounds were introduced in the PhD thesis of our student Jean-Yves Audibert [2]. Here we will use them to sharpen Vapnik's bounds when $r_1(\theta)$ and N are large (a flavor of how large they should be is given in the numerical application at the end of this section). Audibert showed that chaining relative bounds can be used to remove $\log(N)$ terms in Vapnik bounds. Here, we will generalize relative bounds to increased shadow samples and will use only one step of the chaining method (lest we would spoil the constants too much, the price to pay being a trailing $\log\big[\log(N)\big]$ term which anyhow behaves like a constant in practice).

Let us assume that $\mathbb{P}$ is partially exchangeable. Let θ, $\theta' \in \Theta$, and let

$$\begin{aligned}
\chi_i &= \mathbb{1}\big[Y_i \neq f_\theta(X_i)\big] - \mathbb{1}\big[Y_i \neq f_{\theta'}(X_i)\big],\\
r_1'(\theta,\theta') &= r_1(\theta) - r_1(\theta') = \frac{1}{N}\sum_{i=1}^{N}\chi_i,\\
r_2'(\theta,\theta') &= r_2(\theta) - r_2(\theta') = \frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\chi_i.
\end{aligned}$$

For any real number x, let $g(x) = x^{-2}\big[\exp(x) - 1 - x\big]$. As it is well known, $x \mapsto g(x) : \mathbb{R} \to \mathbb{R}$ is an increasing function. This is the key argument in the proof of Bernstein's deviation inequality.

Let

$$\ell(\theta,\theta') = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\mathbb{1}\big[f_\theta(X_i) \neq f_{\theta'}(X_i)\big].$$

Lemma 5.1. *For any partially exchangeable random variable* $\lambda : \Omega \to \mathbb{R}$,

$$\Big(\bigcirc_{i=1}^N \tau_i\Big)\exp\Big\{\lambda\big[r_2'(\theta,\theta') - r_1'(\theta,\theta')\big] - g\big[(1+k^{-1})\tfrac{2\lambda}{N}\big](1+k^{-1})^2\frac{\lambda^2}{N}\ell(\theta,\theta') + \log(\epsilon)\Big\} \leq \epsilon.$$

Proof. For any partially exchangeable random variable η,

$$\log\Big\{ \big(\bigcirc_{i=1}^N \tau_i\big) \exp\Big[\lambda\big[r_2'(\theta,\theta') - r_1'(\theta,\theta')\big] - \eta\Big]\Big\}$$
$$= -\eta + \sum_{i=1}^N \log\Big\{ \tau_i \exp\Big[\frac{\lambda}{N}\Big(\frac{1}{k}\sum_{j=1}^k \chi_{i+jN} - \chi_i\Big)\Big]\Big\}$$
$$= -\eta + \sum_{i=1}^N \log\Big\{ \exp\Big(\frac{\lambda}{kN}\sum_{j=0}^N \chi_{i+jN}\Big)\tau_i \exp\Big(-(1+k^{-1})\frac{\lambda}{N}\chi_i\Big)\Big\}$$
$$= -\eta + \frac{\lambda}{kN}\sum_{i=1}^{(k+1)N} \chi_i + \sum_{i=1}^N \log\Big\{ \tau_i\Big[\exp\Big(-(1+k^{-1})\frac{\lambda}{N}\chi_i\Big)\Big]\Big\}.$$

Now we can apply Bernstein's inequality to

$$\log\Big\{ \tau_i \exp\Big[-(1+k^{-1})\frac{\lambda}{N}\chi_i\Big]\Big\},$$

to show that

$$\log\Big\{ \tau_i \exp\Big[-(1+k^{-1})\frac{\lambda}{N}\chi_i\Big]\Big\} \le \frac{\lambda}{kN}\sum_{j=0}^k \chi_{i+jN}$$
$$+ (1+k^{-1})^2\frac{\lambda^2}{N^2}\tau_i\Big[(\chi_i - p_i)^2\Big] g\left[\frac{2\lambda}{N}(1+k^{-1})\right],$$

where we have put $p_i = \tau_i(\chi_i) = \dfrac{1}{k+1}\displaystyle\sum_{j=0}^k \chi_{i+jN}$. Anyhow, let us reproduce the proof of this statement here, for the sake of completeness. Let us put $\alpha = (1+k^{-1})\frac{\lambda}{N}$.

$$\log\Big\{ \tau_i\Big[\exp(-\alpha\chi_i)\Big]\Big\} = -\alpha p_i$$
$$+ \log\Big\{ 1 + \tau_i\Big[\exp\Big[-\alpha(\chi_i - p_i)\Big] - 1 - \alpha(\chi_i - p_i)\Big\}$$
$$\le -\alpha p_i + \tau_i\Big[\alpha^2(\chi_i - p_i)^2 g\big[-\alpha(\chi_i - p_i)\big]\Big]$$
$$\le -\alpha p_i + g(2\alpha)\tau_i\Big[\alpha^2(\chi_i - p_i)^2\Big].$$

We can now use the bound $\tau_i\Big[(\chi_i - p_i)^2\Big] \le \tau(\chi_i^2)$ and remark that $\chi_i^2 \le \mathbb{1}\big[f_\theta(X_i) \ne f_{\theta'}(X_i)\big]$, to get

$$\log\Big\{ \big(\bigcirc_{i=1}^N \tau_i\big) \exp\Big[\lambda\big[r_2'(\theta,\theta') - r_1'(\theta,\theta')\big] - \eta\Big]\Big\}$$
$$\le g(2\alpha)\alpha^2 \sum_{i=1}^N \tau_i\Big\{\mathbb{1}\big[f_\theta(X_i) \ne f_{\theta'}(X_i)\big]\Big\} - \eta$$
$$= \frac{\lambda^2}{N}(1+k^{-1})^2 g\left(\frac{2\lambda(1+k^{-1})}{N}\right)\ell(\theta,\theta') - \eta.$$

We end the proof by choosing

$$\eta = g\left(\frac{2(1+k^{-1})\lambda}{N}\right)(1+k^{-1})^2\frac{\lambda^2}{N}\ell(\theta,\theta') - \log(\epsilon).$$

□

We deduce easily from the previous lemma the following

Proposition 5.2. *For any partially exchangeable prior distributions π, $\pi' : \Omega \to \mathcal{M}^1_+(\Theta)$, for any partially exchangeable probability measure $\mathbb{P} \in \mathcal{M}^1_+(\Omega)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any θ, $\theta' \in \Theta$,*

$$r'_2(\theta,\theta') - r'_1(\theta,\theta') \leq g\left(\frac{2(1+k^{-1})\lambda}{N}\right)(1+k^{-1})^2\frac{\lambda}{N}\ell(\theta,\theta') - \frac{1}{\lambda}\log\big[\bar\pi(\theta)\bar\pi'(\theta')\epsilon\big],$$

where $\bar\pi(\theta) = \sup\{\pi(\theta'') : \theta'' \in \Theta, \Psi(\theta'') = \Psi(\theta)\}$, and an analogous definition is used for $\bar\pi'$.

Let us now assume that we use a set of binary classification rules $\{f_\theta : \theta \in \Theta\}$ with VC dimension not greater than h.

Let us consider in the following the values

$$\xi_j = \frac{\lfloor (k+1)N\exp(-j)\rfloor}{(k+1)N} \simeq \exp(-j),$$

where $\lfloor x \rfloor$ is the lower integer part of the real number x. Let us define

$$d_j = h\log\left[\frac{2e^2(k+1)N}{h\xi_j}\right] + \log\big[e\log(N)(h+1)\epsilon^{-1}\big].$$

Proposition 5.3. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, any $j \in \{1,\dots,\lfloor\log(N)\rfloor\}$, there is $\theta'_j \in \Theta'_j$ such that*

$$r_2(\theta) - r_1(\theta) \leq r_2(\theta'_j) - r_1(\theta'_j) + \left[g\left(\sqrt{\frac{8d_j}{\xi_j N}}\right) + \frac{1}{2}\right]\sqrt{\frac{2(1+k^{-1})^2\xi_j d_j}{N}}.$$

Proof. Let us recall a lemma due to David Haussler [6] : when the VC dimension of $\{f_\theta : \theta \in \Theta\}$ is not greater than h, then, for any $\xi = \frac{m}{(k+1)N}$, we can find some ξ-covering net $\Theta'_\xi \subset \Theta$ for the distance ℓ (which is a random exchangeable object), such that

$$|\Theta'_\xi| \leq e(h+1)\left(\frac{2e}{\xi}\right)^h.$$

Let us put on

$$\bigsqcup_{j,1\leq j\leq\log(N)} \Theta'_{\xi_j}$$

the prior probability distribution defined by

$$\pi'(\theta'_j) = \big(\lfloor\log(N)\rfloor|\Theta'_{\xi_j}|\big)^{-1} \geq \left[\log(N)e(h+1)\left(\frac{2e}{\xi_j}\right)^h\right]^{-1}, \qquad \theta'_j \in \Theta'_j.$$

We see that with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, any j, $1 \leq j \leq \log(N)$ there is $\theta'_j \in \Theta'_{\xi_j}$ such that $\ell(\theta,\theta'_j) \leq \xi_j$, and therefore such that

$$r_2(\theta,\theta'_j) - r_1(\theta,\theta'_j) \le g\left[(1+k^{-1})\frac{2\lambda}{N}\right](1+k^{-1})^2\frac{\lambda}{N}\xi_j + \frac{1}{\lambda}\log\bigl[\bar\pi(\theta)^{-1}\bar\pi'(\theta'_j)^{-1}\epsilon^{-1}\bigr],$$

where we can take $\bar\pi(\theta) \ge \left(\frac{e(k+1)N}{h}\right)^{-h}$ and where $\pi'(\theta'_j)$ has been defined earlier. We can then choose

$$\lambda(\theta,\theta'_j) = \left[\frac{2N\log\Bigl\{\bigl[\bar\pi(\theta)\bar\pi'(\theta'_j)\epsilon\bigr]^{-1}\Bigr\}}{(1+k^{-1})^2\xi_j}\right],$$

to prove proposition 5.3. □

On the other hand, from theorem 3.4 applied to $\bigsqcup\Theta'_{\xi_j}$ and π', we see that with $\mathbb{P}$ probability at least $1-\epsilon$, for any j, $1 \le j \le \log(N)$ and any $\theta'_j \in \Theta'_j$, putting

$$d'_j = h\log\left(\frac{2e}{\xi_j}\right) + \log\bigl[e\log(N)(h+1)\bigr] - \log(\epsilon),$$

we have

$$r_2(\theta'_j) - r_1(\theta'_j) \le \sqrt{\frac{2}{N}(1+k^{-1})^2 d'_j \Phi\left(\frac{r_1(\theta'_j)+kr_2(\theta'_j)}{1+k}\right)}.$$

We can then remark that when $\ell(\theta,\theta'_j) \le \xi_j$,

$$\frac{1}{k+1}\bigl[r_1(\theta'_j)+kr_2(\theta'_j)\bigr] \le \frac{1}{k+1}\bigl[r_1(\theta)+kr_2(\theta)\bigr]+\ell(\theta,\theta'_j) \le \frac{1}{k+1}\bigl[r_1(\theta)+kr_2(\theta)\bigr]+\xi_j.$$

We have proved the following

Theorem 5.4. *With $\mathbb{P}$ probability at least $1-2\epsilon$,*

$$r_2(\theta) - r_1(\theta) \le \inf_{j\in\mathbb{N}^*, 1\le j\le\log(N)} \left[g\left(\sqrt{\frac{8d_j}{\xi_j N}}\right)+\frac{1}{2}\right]\sqrt{\frac{2(1+k^{-1})^2\xi_j d_j}{N}} + \sqrt{\frac{2}{N}(1+k^{-1})^2 d'_j \Phi\left(\frac{r_1(\theta)+kr_2(\theta)}{1+k}+\xi_j\right)}.$$

Remark 5.1. To use this theorem, we have to solve equations of the type

$$r_2 - r_1 \le a + b\left[\Phi\left(\frac{r_1+kr_2}{1+k}+\xi\right)\right]^{1/2}.$$

Whenever r_1 and the bound are less than $1/2$, this is equivalent to

$$r_2 \le \frac{B+\sqrt{B^2-AC}}{A},$$

where

$$A = 1+\left(\frac{kb}{1+k}\right)^2,$$

$$B = r_1 + a + \frac{kb^2}{2(1+k)^2}\bigl[(1+k)(1-2\xi)-2r_1\bigr],$$

$$C = (r_1+a)^2 - \frac{b^2}{(1+k)^2}\bigl[(1+k)\xi+r_1\bigr]\bigl[(1+k)(1-\xi)-r_1\bigr].$$

Let us make some numerical application. We should take N pretty large, because the expected benefit of this last theorem is to improve on the $\log(N)$ term (the optimization in ξ_j allows to kill the $\log(N)$ term in d_j and be left only with $\log\big[\log(N)\big]$ terms). So let us take $N = 10^6$, $h = 10$, $r_1(\theta) = 0.2$ and $\epsilon = 0.005$. For these values, theorem 3.4 gives a bound greater than 0.2075 and less than 0.2076 when k ranges from 24 to 46. Here we obtain a bound less than 0.2070 for k ranging from 24 to 46, the optimal values for (k, j) being $(257, 7)$, giving a bound less than 0.20672. The bound is less than 0.2068 for k ranging from 42 to 19470, showing that we can really use big shadow samples with theorem 5.4 !

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale sensitive dimensions, uniform convergence and learnability, J. of ACM **44**(4):615-631, 1997.

[2] J.-Y. Audibert, PAC-Bayesian statistical learning theory, Thèse de doctorat de l'Université Paris 6, 29 juin 2004.

[3] A. Blum and J. Langford, PAC-MDL Bounds, *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, Lecture Notes in Computer Science 2777, Springer 2003, pp. 344-357.

[4] O. Catoni, Statistical learning theory and stochastic optimization, *Ecole d'été de Probabilités de Saint-Flour XXXI - 2001, J. Picard Ed., Lecture notes in mathematics* , **1851**, pp. 1-272, Springer, 2004.

[5] O. Catoni, A PAC-Bayesian approach to adaptive classification, *preprint*, (2003).

[6] D. Haussler, Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. J. Combin. Theory Ser. A 69 (1995), no. 2, 217–232.

[7] J. Langford and D. McAllester, Computable Shell Decomposition Bounds, *Journal of Machine Learning Research*, **5**, 2004, pp. 529-547 (communicated at COLT 2000).

[8] J. Langford and M. Seeger, Bounds for Averaging Classifiers, *technical report CMU-CS-01-102*, Carnegie Mellon University, jan. 2001, `www.cs.cmu.edu/~jcl`.

[9] J. Langford, M. Seeger and N. Megiddo, An Improved Predictive Accuracy Bound for Averaging Classifiers, *International Conference on Machine Learning* **18** (2001), 290-297.

[10] V. N. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.

[11] D. A. McAllester, Some PAC-Bayesian Theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, 230–234 (electronic), ACM, New York, 1998;

[12] D. A. McAllester, PAC-Bayesian Model Averaging, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, 164–170 (electronic), ACM, New York, 1999;

[13] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, *Journal of Machine Learning Research* **3** (2002), 233–269.

[14] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, *Informatics Research Report, Division of Informatics, University of Edinburgh* EDI-INF-RR-0094 (2002), 1–42. `http://www.informatics.ed.ac.uk`

[15] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, *IEEE Trans. Inform. Theory* **44** (1998), no. 5, 1926–1940.

CNRS – Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6 (site Chevaleret), 4 place Jussieu – Case 188, 75 252 Paris Cedex 05.