Mach Learn (2007) 66:165–207
DOI 10.1007/s10994-006-7679-y

Model selection by bootstrap penalization for classification

Magalie Fromont

Received: 2 April 2005 / Revised: 16 December 2005 / Accepted: 22 December 2005 / Published online: 3 May 2006

Abstract We consider the binary classification problem. Given an i.i.d. sample drawn from the distribution of an $\mathcal{X} \times \{0, 1\}-$valued random pair, we propose to estimate the so-called *Bayes classifier* by minimizing the sum of the empirical classification error and a penalty term based on Efron's or i.i.d. weighted bootstrap samples of the data. We obtain exponential inequalities for such bootstrap type penalties, which allow us to derive non-asymptotic properties for the corresponding estimators. In particular, we prove that these estimators achieve the global minimax risk over sets of functions built from Vapnik-Chervonenkis classes. The obtained results generalize Koltchinskii (2001) and Bartlett et al.'s (2002) ones for Rademacher penalties that can thus be seen as special examples of bootstrap type penalties. To illustrate this, we carry out an experimental study in which we compare the different methods for an intervals model selection problem.

Keywords Model selection · Classification · Bootstrap penalty · Exponential inequality · Oracle inequality · Minimax risk

1 Introduction

Let (X, Y) be a random pair with values in a measurable space $\Xi = \mathcal{X} \times \{0, 1\}$, and with unknown distribution denoted by P. Given n independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), we aim at constructing a *classification rule* that is a function which would give the value of Y from the observation of X. More precisely, in statistical terms, we are interested in the estimation of the function s minimizing the classification error $\mathbb{P}[t(X) \neq Y]$ over all the measurable functions $t : \mathcal{X} \to \{0, 1\}$. The function s is called the *Bayes classifier* and it is also defined by $s(x) = \mathbb{I}_{\mathbb{P}[Y=1|X=x]>1/2}$.

Editor: Oliver Bousquet and Andre Elisseeff

M. Fromont (✉)
Laboratoire de Statistique, U.F.R. de Sciences Sociales–Département MASS, Université Rennes II, Place du Recteur H. Le Moal, 35 043 Rennes cedex, France
e-mail: magalie.fromont@uhb.fr

Given a class S of measurable functions from $\mathcal{X}$ to $\{0, 1\}$, an estimator $\hat{s}$ of s is determined by minimization of the empirical classification error $\gamma_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{I}_{t(X_i)\neq Y_i}$ over all the functions t in S. This method has been introduced in learning problems by Vapnik and Chervonenkis (1971). However, it poses the problem of the choice of the class S. To provide an estimator with classification error close to the optimal one, S has to be large enough so that the error of the best function in S is close to the optimal error, while it has to be small enough so that finding the best candidate in S from the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ is still possible. In other words, one has to choose a class S which achieves the best trade-off between the approximation error and the estimation error.

One approach proposed to solve this question is Grenander's (1981) *method of sieves* which allows to select one class among a nested sequence $S_1, S_2, \ldots$. The selection here is based only on the sample size n, in such a way that the complexity of the selected class grows with n, but that it does not grow too fast so that the estimation error may be controlled. However, the fact that the class is chosen independently of the data implies that the obtained estimator is not satisfactory for all distributions.

The method of *Structural Risk Minimization* (SRM) initiated by Vapnik (1982) and also known as *Complexity regularization* (see Barron 1985, 1991; Barron and Cover, 1991) fills this gap by using the data to choose the class over which the estimator is constructed. It consists in selecting among a given collection of functions sets the set S minimizing the sum of the empirical classification error of the estimator $\hat{s}$ and a penalty term taking the complexity of S into account. The quantities generally used to measure the complexity of some class S of functions from $\mathcal{X}$ to $\{0, 1\}$ are the *Shatter coefficients* of the associated class of sets $\mathcal{C} = \{\{x \in \mathcal{X},\ t(x) = 1\},\ t \in S\}$ given by:

$$\text{for } k \geq 1,\ \mathbb{S}(\mathcal{C}, k) = \max_{x_1,\ldots,x_k \in \mathcal{X}} |\{\{x_1, \ldots, x_k\} \cap C,\ C \in \mathcal{C}\}|,$$

and the *Vapnik-Chervonenkis dimension* of $\mathcal{C}$ defined as:

$$V(\mathcal{C}) = \infty \text{ if for all } k \geq 1,\ \mathbb{S}(\mathcal{C}, k) = 2^k,$$
$$V(\mathcal{C}) = \sup\{k \geq 1,\ \mathbb{S}(\mathcal{C}, k) = 2^k\} \text{ else.}$$

Considering a collection $\{S_m,\ m \in \mathbb{N}^*\}$ of classes of functions from $\mathcal{X}$ to $\{0, 1\}$ and setting $\mathcal{C}_m = \{\{x \in \mathcal{X},\ t(x) = 1\},\ t \in S_m\}$, Lugosi and Zeger (1996) study the standard penalties of the form

$$\text{pen}(m) = \kappa \sqrt{\frac{\log \mathbb{S}(\mathcal{C}_m, n^2) + m}{n}},$$

which are approximately $\kappa' \sqrt{(V(\mathcal{C}_m) \log n + m)/n}$. By using an inequality due to Devroye (1982), they prove that if all the classes $\mathcal{C}_m$ are Vapnik-Chervonenkis classes (that is if they have a finite VC dimension) such that the sequence $(V(\mathcal{C}_m))_{m\in\mathbb{N}^*}$ is strictly increasing, and if the Bayes classifier s belongs to the union of the S_m's, there exists an integer k such that the expected classification error of the rule obtained by SRM with such penalties differs from the optimal error $\mathbb{P}[s(X) \neq Y]$ by a term not larger than a constant times $\sqrt{V(\mathcal{C}_k) \log n/n}$. This upper bound is optimal in a global minimax sense up to a logarithmic factor. Given a class S of functions from $\mathcal{X}$ to $\{0, 1\}$ where $\mathcal{C} = \{\{x \in \mathcal{X},\ t(x) = 1\},\ t \in S\}$ is a VC class with VC dimension $V(\mathcal{C})$, Vapnik and Chervonenkis (1974) actually prove that there exist

some constants κ_1 and κ_2 such that for any classification rule $\hat{s}$ with classification error $L_{\hat{s}}$,

$$\sup_{P,\ s\in S} \mathbb{E}\left[L_{\hat{s}} - \mathbb{P}\left[s(X) \neq Y\right]\right] \geq \kappa_1 \sqrt{\frac{V(\mathcal{C})}{n}},\ \forall n \geq \kappa_2 V(\mathcal{C}).$$

We explain in the next section how the choice of the penalty terms is connected with the calibration of an upper bound for the quantity $\sup_{t\in S}|\gamma_n(t) - \mathbb{P}\left[t(X) \neq Y\right]|$. Unfortunately, in addition to the fact that their computation is generally complicated, the penalties based on the Shatter coefficients or the VC dimensions have the disadvantage to be deterministic and to overestimate this quantity for specific data distributions. This remark naturally leads to the idea of *data-driven* penalization. Buescher and Kumar (1996), Lugosi and Nobel (1999), Boucheron et al. (2000) introduce some penalties involving the related empirical coverings or empirical Shatter coefficients. Inspired by the method of Rademacher symmetrization commonly used in the empirical processes theory (see for instance Van der Vaart and Wellner, 1996), Koltchinskii (2001) and Bartlett et al. (2002) independently propose the so-called *Rademacher penalties* which are based on random variables of the form:

$$\mathbb{E}\left[\sup_{t\in S} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbb{I}_{t(X_i)\neq Y_i} \middle| \xi\right],$$

where ξ denotes the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, and $\varepsilon_1, \ldots, \varepsilon_n$ is a sequence of i.i.d. Rademacher random variables independent of ξ. They prove oracle type inequalities showing that such random penalties provide optimal classification rules in a global minimax sense over sets of functions built from Vapnik-Chervonenkis classes. Lozano (2000) gives the experimental evidence that, for the intervals model selection problem, Rademacher penalization outperforms SRM and cross validation over a wide range of sample sizes. Bartlett et al. (2002) also study Rademacher penalization from a practical point of view by comparing it with other kinds of data-driven methods.

Whereas the methods of Rademacher penalization are now frequently used in the statistical learning theory, they are not so popular yet in the applied statistics community. In fact, statisticians often prefer to stick with resampling tools such as bootstrap or jacknife in practice. We here aim at making the connection between the two approaches. From the asymptotic results due to Giné and Zinn (1990) about Efron's bootstrap or Praestgaard and Wellner (1993) about exchangeable weighted bootstrap, one can indeed expect to find sharp bootstrap upper bounds for $\sup_{t\in S}|\gamma_n(t) - \mathbb{P}\left[t(X) \neq Y\right]|$. We hence introduce and investigate a new family of penalties based on classical bootstrap processes such as Efron's or i.i.d. weighted bootstrap ones while attending to placing Rademacher penalties among this family.

The paper is organized as follows. In Section 2, we present the model selection by penalization approach and explain how to choose a penalty function. We study some penalties based on various symmetric variables, before dealing with the bootstrap type penalties. The results obtained for the corresponding classification rules generalize Koltchinskii (2001) and Bartlett et al.'s (2002) ones. As one can see in Section 5 which gives the details of the proofs, they essentially follow from some exponential inequalities established in Section 4. We furthermore devote Section 3 to an experimental comparison between our new penalization methods and the Rademacher one for an intervals model selection problem. We finally give in Section 6 a discussion about this work.

2 Model selection

We describe here the model selection by penalization approach to construct classification rules or estimators of the Bayes classifier s. In the following, we denote by $\mathcal{S}$ the set of all the measurable functions $t : \mathcal{X} \to \{0, 1\}$. Given a collection $\{S_m, m \in \mathcal{M}\}$ of countable classes of functions in $\mathcal{S}$ (the *models*) and $\rho_n \geq 0$, for any m in $\mathcal{M}$, we can construct some approximate minimum contrast estimator $\hat{s}_m$ in S_m satisfying:

$$\gamma_n(\hat{s}_m) \leq \inf_{t \in S_m} \gamma_n(t) + \rho_n/2.$$

We thus obtain a collection $\{\hat{s}_m, m \in \mathcal{M}\}$ of possible classification rules and at this stage, the issue is to choose among this collection the "best" rule in terms of risk minimization. Let l be the loss function defined by:

$$l(u, v) = \mathbb{E}\left[\mathbb{I}_{v(X)\neq Y} - \mathbb{I}_{u(X)\neq Y}\right], \text{ for all } u, v \text{ in } \mathcal{S}. \tag{1}$$

Notice that, by definition of s, $l(s, t)$ is nonnegative for every t in $\mathcal{S}$. The risk of any estimator $\hat{s}_m$ of s is given by $\mathbb{E}[l(s, \hat{s}_m)]$. Ideally, we would like to select some element $\bar{m}$ (the *oracle*) in $\mathcal{M}$ minimizing

$$\mathbb{E}[l(s, \hat{s}_m)] = l(s, s_m) + \mathbb{E}[l(s_m, \hat{s}_m)],$$

where for every m in $\mathcal{M}$, s_m denotes some function in S_m such that

$$l(s, s_m) = \inf_{t \in S_m} l(s, t).$$

However, such an oracle $\bar{m}$ necessarily depends on the unknown distribution of (X, Y). This leads us to use the method of model selection by penalization. The purpose of this method, that originates in Mallows' C_p and Akaike's heuristics, is actually to provide a criterion which allows to select, only from the data, an element $\hat{m}$ in $\mathcal{M}$ mimicking the oracle. Considering some *penalty function* pen : $\mathcal{M} \to \mathbb{R}_+$, we choose $\hat{m}$ such that:

$$\gamma_n(\hat{s}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \inf_{m \in \mathcal{M}} \{\gamma_n(\hat{s}_m) + \text{pen}(m)\} + \rho_n/2,$$

and we take as "best" rule the so-called *approximate minimum penalized contrast estimator*

$$\tilde{s} = \hat{s}_{\hat{m}}. \tag{2}$$

We then have to determine a penalty function such that the risk of the approximate minimum penalized contrast estimator $\tilde{s}$ is of the same order as

$$\inf_{m \in \mathcal{M}} \mathbb{E}[l(s, \hat{s}_m)] = \inf_{m \in \mathcal{M}} \{l(s, s_m) + \mathbb{E}[l(s_m, \hat{s}_m)]\}$$

or, failing that, at most of the same order as $\inf_{m\in\mathcal{M}}\{l(s, s_m) + \sqrt{V_m/n}\}$ when for each m in $\mathcal{M}$, $S_m = \{\mathbb{I}_C, C \in \mathcal{C}_m\}$, $\mathcal{C}_m$ being a VC class with VC dimension V_m. Indeed, as cited in the introduction, Vapnik and Chervonenkis (1974) proved that the global minimax risk over

such a class S_m defined by

$$\inf_{\hat{s}} \sup_{P, s \in S_m} \mathbb{E}\left[l(s, \hat{s})\right]$$

is of order $\sqrt{V_m/n}$ as soon as $n \geq \kappa V_m$, for some absolute constant κ.

The various strategies to determine adequate penalty functions rely on the same basic inequality that we present below. Let us fix m in $\mathcal{M}$ and introduce the centered empirical contrast defined for all t in $\mathcal{S}$ by

$$\overline{\gamma_n}(t) = \gamma_n(t) - \mathbb{E}\left[\mathbb{I}_{t(X) \neq Y}\right]. \tag{3}$$

By definition,

$$l(s_m, \hat{s}_{\hat{m}}) = \overline{\gamma_n}(s_m) - \gamma_n(s_m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}) + \gamma_n(\hat{s}_{\hat{m}}).$$

Noticing that

$$\begin{aligned} \gamma_n(\hat{s}_{\hat{m}}) + \text{pen}(\hat{m}) &\leq \gamma_n(\hat{s}_m) + \text{pen}(m) + \rho_n/2 \\ &\leq \gamma_n(s_m) + \text{pen}(m) + \rho_n, \end{aligned}$$

we derive

$$l(s, \tilde{s}) \leq l(s, s_m) + \overline{\gamma_n}(s_m) + \text{pen}(m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}) - \text{pen}(\hat{m}) + \rho_n, \tag{4}$$

which holds whatever the penalty function. Looking at the problem from a global minimax point of view, since $\mathbb{E}\left[\overline{\gamma_n}(s_m)\right] = 0$, it is a matter of choosing a penalty such that $\text{pen}(\hat{m})$ compensates for $-\overline{\gamma_n}(\hat{s}_{\hat{m}})$ and such that $\mathbb{E}\left[\text{pen}(m)\right]$ is of order at most $\sqrt{V_m/n}$ in the VC case. Hence, we need to control $-\overline{\gamma_n}(t)$ uniformly for t in S_m and m in $\mathcal{M}$ and the concentration inequalities appear as the appropriate tools.

Since we deal with a bounded contrast, we can use the following McDiarmid's (1989) inequality which derives from Azuma's (1967) result for martingale difference sequences.

Theorem 1 (McDiarmid). *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $\phi : A^n \to \mathbb{R}$ satisfies*:

$$\sup_{x_1, \ldots, x_n, x_i' \in A} |\phi(x_1, \ldots, x_n) - \phi(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i,$$

for all $i \in \{1, \ldots, n\}$. Then for all $x > 0$, the two following inequalities hold:

$$\mathbb{P}\left[\phi(X_1, \ldots, X_n) \geq \mathbb{E}\left[\phi(X_1, \ldots, X_n)\right] + x\right] \leq e^{-\frac{2x^2}{\sum_{i=1}^n c_i^2}},$$

$$\mathbb{P}\left[\phi(X_1, \ldots, X_n) \leq \mathbb{E}\left[\phi(X_1, \ldots, X_n)\right] - x\right] \leq e^{-\frac{2x^2}{\sum_{i=1}^n c_i^2}}.$$

From Theorem 1, we can see that for all m in $\mathcal{M}$, $\sup_{t\in S_m}(-\overline{\gamma_n}(t))$ concentrates around its expectation. A well-chosen estimator of an upper bound for $\mathbb{E}[\sup_{t\in S_m}(-\overline{\gamma_n}(t))]$, with expectation of order $\sqrt{V_m/n}$ in the VC case, may therefore be a good penalty.

In this paper, we focus on random penalty functions. This subject has been tackled by Buescher and Kumar (1996), Lugosi and Nobel (1999) and Boucheron et al. (2000) but the most interesting works for our approach are the ones due to Koltchinskii (2001) and Bartlett et al. (2002). Let ξ denote the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Starting from the symmetrization tools used in the empirical processes theory, Koltchinskii (2001) and Bartlett et al. (2002) propose a penalty based on the random variable

$$\hat{R}_m = \mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbb{I}_{t(X_i)\neq Y_i}\,\middle|\,\xi\right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is a sequence of independent identically distributed Rademacher variables such that $\mathbb{P}[\varepsilon_i = 1] = \mathbb{P}[\varepsilon_i = -1] = 1/2$ and the ε_i's are independent of ξ. More precisely, they take $\mathcal{M} = \mathbb{N}^*$ and they consider the approximate minimum penalized contrast estimator $\tilde{s}$ given by (2) with $\text{pen}(m) = 2\hat{R}_m + c_1\sqrt{\log m/n}$, for some absolute, positive constant c_1. Setting $L_t = \mathbb{P}[t(X) \neq Y]$, they prove that there exists some constant $c_2 > 0$ such that

$$\mathbb{E}[L_{\tilde{s}}] \leq \inf_{m\in\mathcal{M}}\left\{\inf_{t\in S_m} L_t + \mathbb{E}[\text{pen}(m)]\right\} + \frac{c_2}{\sqrt{n}} + \rho_n,$$

which can be translated in terms of risk bounds as follows:

$$\mathbb{E}[l(s,\tilde{s})] \leq \inf_{m\in\mathcal{M}}\left\{l(s,s_m) + \mathbb{E}[\text{pen}(m)]\right\} + \frac{c_2}{\sqrt{n}} + \rho_n.$$

Moreover, Koltchinskii notes that if the collection of models $\{S_m, m \in \mathcal{M}\}$ is taken such that

$$S_m = \{\mathbb{I}_C, C \in \mathcal{C}_m\},$$

where each $\mathcal{C}_m$ is a VC class of subsets of $\mathcal{X}$ with VC dimension V_m, then

$$\mathbb{E}[L_{\tilde{s}}] \leq \inf_{m\in\mathcal{M}}\left\{\inf_{t\in S_m} L_t + \kappa\left(\sqrt{\frac{V_m}{n}} + \sqrt{\frac{\log m}{n}} + \frac{1}{\sqrt{n}}\right)\right\} + \rho_n.$$

Our purpose is to extend this study by investigating penalty functions based on random variables of the form

$$\mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i}\,\middle|\,\xi\right],$$

with various random weights $Z_1, \ldots, Z_n$.

To avoid dealing with measurability issues, we assume that all the classes of functions considered in the paper are at most countable.

2.1 Penalties based on symmetric weights

Noticing that the symmetrization techniques used by Koltchinskii (2001) and Bartlett et al. (2002) can be applied to any symmetric variables (and not only Rademacher ones) with a finite first order moment, we focus on penalties based on some quantities

$$\mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^{n}Z_i\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right],$$

where $Z_1,\ldots,Z_n$ are i.i.d. symmetric random weights such that $\mathbb{E}[|Z_1|]<+\infty$.

We present here an upper bound for the risk of the approximate minimum penalized contrast estimators obtained via such penalties. In particular, provided that the weights $Z_1,\ldots,Z_n$ satisfy some moments conditions, if the collection of models is composed of sets of functions based on VC classes, we show that this risk is at most of the same order as the global minimax risk when the Bayes classifier is in some model of the collection.

The following result essentially derives from McDiarmid's inequality combined with a maximal inequality given in Section 5.1.

Theorem 2. *Let $\xi=(X_1,Y_1),\ldots,(X_n,Y_n)$ be a sample of n independent copies of a random pair (X,Y) with values in $\mathcal{X}\times\{0,1\}$, and let $Z_1,\ldots,Z_n$ be i.i.d. symmetric random variables independent of ξ such that $\mathbb{E}[|Z_1|]<+\infty$. Consider a countable collection $\{S_m, m\in\mathcal{M}\}$ of classes of functions in $\mathcal{S}$ and a family $(x_m)_{m\in\mathcal{M}}$ of nonnegative weights such that*

$$\sum_{m\in\mathcal{M}}e^{-x_m}\leq\Sigma,$$

for some absolute constant Σ. Assume that for the loss function defined by (1), for each m in $\mathcal{M}$, there exists a minimizer s_m of $l(s,.)$ over S_m. Choose a penalty function such that

$$\mathrm{pen}(m)=\frac{2}{n\mathbb{E}[|Z_1|]}\mathbb{E}\left[\sup_{t\in S_m}\sum_{i=1}^{n}Z_i\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right]+3\sqrt{\frac{x_m}{2n}}.$$

The approximate minimum penalized contrast estimator $\tilde{s}$ given by (2) satisfies:

$$\mathbb{E}[l(s,\tilde{s})]\leq\inf_{m\in\mathcal{M}}\{l(s,s_m)+\mathbb{E}\left[\mathrm{pen}(m)\right]\}+\frac{3\Sigma}{2}\sqrt{\frac{\pi}{2n}}+\rho_n.$$

Assume moreover that $Z_1,\ldots,Z_n$ satisfy the moments condition:

$$\forall k\geq 2,\ \mathbb{E}[|Z_1|^k]\leq\frac{k!}{2}vc^{k-2},\tag{5}$$

for some positive numbers v and c, and that $n\geq 4$. If for all m in $\mathcal{M}$, $S_m=\{\mathbb{I}_C,\ C\in\mathcal{C}_m\}$, where $\mathcal{C}_m$ is a VC class with VC dimension $V_m\geq 1$, there exist some positive, absolute

constants κ_1 and κ_2 such that

$$\mathbb{E}[l(s,\tilde{s})] \leq \inf_{m\in\mathcal{M}}\left\{l(s,s_m)+\frac{1}{\mathbb{E}[|Z_1|]}\left(\kappa_1\sqrt{v}\sqrt{\frac{V_m}{n}}+\kappa_2 c\frac{V_m}{n}\log^2 n\right)+3\sqrt{\frac{x_m}{2n}}\right\}+\frac{3\Sigma}{2}\sqrt{\frac{\pi}{2n}}+\rho_n,$$

and when $\mathbb{E}[e^{\lambda Z_1}] \leq e^{\lambda^2/2}$ for any $\lambda > 0$,

$$\mathbb{E}[l(s,\tilde{s})] \leq \inf_{m\in\mathcal{M}}\left\{l(s,s_m)+\frac{\kappa_1}{\mathbb{E}[|Z_1|]}\sqrt{\frac{V_m}{n}}+3\sqrt{\frac{x_m}{2n}}\right\}+\frac{3\Sigma}{2}\sqrt{\frac{\pi}{2n}}+\rho_n.$$

Comments:

(i) Since the Rademacher variables satisfy the subgaussian inequality $\mathbb{E}[e^{\lambda Z_1}] \leq e^{\lambda^2/2}$ for any $\lambda > 0$, the risk upper bound obtained here generalizes Koltchinskii (2001) and Bartlett et al.'s (2002) one for Rademacher penalization.

(ii) Consider a collection $\{S_m,\ m \in \mathcal{M}\}$ of at most n classes of functions in $\mathcal{S}$ such that for each m in $\mathcal{M}$, $S_m = \{\mathbb{I}_C,\ C \in \mathcal{C}_m\}$, $\mathcal{C}_m$ being a VC class with VC dimension $V_m \geq 1$. Assume that the Bayes classifier s associated with (X, Y) is in some S_{m_0} of the collection. If $\tilde{s}$ is the approximate minimum penalized contrast estimator obtained from the above penalty function based on symmetric weights $Z_1, \ldots, Z_n$ such that (5) holds, we deduce from Theorem 2 that

$$\mathbb{E}[l(s,\tilde{s})] \leq \nu(v,c,\mathbb{E}[|Z_1|])\left(\sqrt{\frac{V_{m_0}}{n}}+\sqrt{\frac{\log n}{n}}+\frac{V_{m_0}}{n}\log^2 n\right)+\rho_n.$$

When ρ_n is smaller than $n^{-1/2}$, this implies that if $\log n \leq V_{m_0} \leq n/\log^4 n$ then $\tilde{s}$ achieves, up to a constant, the global minimax risk over S_{m_0}.

(iii) We shall remark that the factor 2 in the expression of the penalty term, which comes from symmetrization inequalities, is surely pessimistic. All the experiments that we have carried out indeed lead us to think that the real constant is about 1 and to take in practice a penalty equal to $\mathbb{E}[\sup_{t\in S_m}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}|\xi]/(\mathbb{E}[|Z_1|]\,n)$.

2.2 Bootstrap penalization

Let P_n be the empirical process associated with the sample $\xi = (X_1, Y_1), \ldots, (X_n, Y_n)$ and defined by $P_n(f) = n^{-1}\sum_{i=1}^n f(X_i, Y_i)$, and set $P(f) = \mathbb{E}[f(X, Y)]$. For every m in $\mathcal{M}$, denote by $\mathcal{F}_m$ the class of functions $\{f : \Xi = \mathcal{X}\times\{0,1\} \to \{0,1\},\ f(x,y) = \mathbb{I}_{t(x)\neq y},\ t \in S_m\}$. As explained above with (4), we determine an adequate penalty function by controlling $\sup_{t\in S_m}(-\overline{\gamma_n}(t)) = \sup_{f\in\mathcal{F}_m}(P - P_n)(f)$ uniformly for m in $\mathcal{M}$. Since McDiarmid's inequality gives that each supremum concentrates around its expectation, we only need to estimate

$\mathbb{E}[\sup_{f\in\mathcal{F}_m}(P-P_n)(f)]$. Koltchinskii (2001) and Bartlett et al. (2002) consider the estimator

$$\hat{R}_m = \mathbb{E}\left[\sup_{f\in\mathcal{F}_m}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i,Y_i)\middle|\xi\right],$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. Rademacher variables independent of ξ. It is interesting here to notice that this estimator can be written as

$$\hat{R}_m = \mathbb{E}\left[\sup_{f\in\mathcal{F}_m}\left(P_n-P_n^b\right)(f)\middle|\xi\right],$$

where P_n^b denotes the weighted empirical process defined by $P_n^b(f)=n^{-1}\sum_{i=1}^{n}2B_i f(X_i,Y_i)$, the B_i's being i.i.d. random variables, independent of ξ, with a Bernoulli (with parameter $1/2$) distribution. This expression for $\hat{R}_m$ naturally leads to the idea of introducing estimators of the same form, but with an empirical process P_n^b based on weights more often used in practice, for instance multinomial weights with parameters $(n,n^{-1},\ldots,n^{-1})$. Such a multinomial weighted empirical process is actually well known by the applied statistics community since it corresponds to the Efron's bootstrap empirical process. More generally, if $W_n=(W_{n,1},\ldots,W_{n,n})$ denotes a vector of n exchangeable and nonnegative random variables independent of ξ and satisfying $\sum_{i=1}^{n}W_{n,i}=n$, we consider the exchangeably weighted bootstrap empirical process defined by

$$P_n^w(f)=\frac{1}{n}\sum_{i=1}^{n}W_{n,i}f(X_i,Y_i).$$

One of the most classical examples is the i.i.d. weighted bootstrap process with $W_{n,i}=V_i/\overline{V_n}$, where $V_1,\ldots,V_n$ are i.i.d. positive random variables independent of ξ and $\overline{V_n}=n^{-1}\sum_{j=1}^{n}V_j$.

Now, the question is whether $\mathbb{E}[\sup_{f\in\mathcal{F}_m}(P-P_n)(f)]$ is definitely well approximated by $\mathbb{E}[\sup_{f\in\mathcal{F}_m}(P_n-P_n^w)(f)|\xi]$.

We prove via a non-asymptotic approach that $\mathbb{E}[\sup_{f\in\mathcal{F}_m}\left(P_n-P_n^w\right)(f)|\xi]$ is in fact a sharp upper bound for $\mathbb{E}[\sup_{f\in\mathcal{F}_m}(P-P_n)(f)]$ (up to some constants), as Praestgaard and Wellner's (1993) asymptotic theory let think. This allows us to obtain some results in the same vein as Koltchinskii (2001) and Bartlett et al.'s (2002) ones.

The following theorem provides an upper bound for the risk of the approximate minimum penalized contrast estimator thus constructed. In the particular cases of Efron's and i.i.d. weighted bootstraps, one can moreover see that the estimators have some adaptive properties in a global minimax sense when the collection of models is based on VC classes.

Theorem 3. *Assume that $n\geq 4$ and let $\xi=(X_1,Y_1),\ldots,(X_n,Y_n)$ be a sample of n independent copies of a random pair (X,Y) with values in $\mathcal{X}\times\{0,1\}$. Introduce a vector $W_n=(W_{n,1},\ldots,W_{n,n})$ of n exchangeable and nonnegative random variables independent of ξ and satisfying $\sum_{i=1}^{n}W_{n,i}=n$. Let*

$$\gamma_n(t)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{t(X_i)\neq Y_i}\quad\text{and}\quad\gamma_n^w(t)=\frac{1}{n}\sum_{i=1}^{n}W_{n,i}\mathbb{I}_{t(X_i)\neq Y_i}.$$

Consider a countable collection $\{S_m, m \in \mathcal{M}\}$ *of classes of functions in* $\mathcal{S}$ *and a family* $(x_m)_{m\in\mathcal{M}}$ *of nonnegative weights such that for some absolute constant* Σ,

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma.$$

Assume that for the loss function defined by (1), for each m in $\mathcal{M}$, *there exists a minimizer* s_m *of* $l(s, .)$ *over* S_m. *Choose a penalty function such that*

$$pen(m) = \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{t\in S_m}\left(\gamma_n(t)-\gamma_n^w(t)\right)\Big|\xi\right] + \left(1+\frac{\mathbb{E}\left[|W_{n,1}-1|\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\right)\sqrt{\frac{x_m}{2n}}.$$

The approximate minimum penalized contrast estimator $\tilde{s}$ *given by (2) satisfies:*

$$\mathbb{E}\left[l(s,\tilde{s})\right] \leq \inf_{m\in\mathcal{M}}\left\{l(s,s_m)+\mathbb{E}\left[\text{pen}(m)\right]\right\} + \left(1+\frac{\mathbb{E}\left[|W_{n,1}-1|\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\right)\frac{\Sigma}{2}\sqrt{\frac{\pi}{2n}} + \rho_n.$$

Consider now the Efron's bootstrap case, where W_n *is a multinomial vector with parameters* $(n, n^{-1}, \ldots, n^{-1})$, *and the i.i.d. weighted bootstrap case where* $W_{n,i} = V_i/\overline{V_n}$, $V_1, \ldots, V_n$ *being i.i.d. positive random variables independent of* ξ, *with*

$$\forall k \geq 2,\ \mathbb{E}\left[V_1^k\right] \leq \frac{k!}{2}vc^{k-2}, \tag{6}$$

for some positive numbers v and c. If for all m in $\mathcal{M}$, $S_m = \{\mathbb{I}_C,\ C \in \mathcal{C}_m\}$, *where* $\mathcal{C}_m$ *is a VC class with VC dimension* $V_m \geq 1$, *there exists some positive constant v which may depend on v, c,* $\mathbb{E}\left[V_1\right]$, $\mathbb{E}\left[|V_1/\overline{V_n}-1|\right]$, *and* $\mathbb{E}\left[(V_1/\overline{V_n}-1)_+\right]$ *such that*

$$\mathbb{E}\left[l(s,\tilde{s})\right] \leq \inf_{m\in\mathcal{M}}\left\{l(s,s_m)+v\left(\sqrt{\frac{V_m}{n}}+\frac{V_m}{n}\log^2 n+\sqrt{\frac{x_m}{n}}\right)\right\}+\rho_n.$$

Comments:

(i) The structure of the risk upper bound obtained here in the Efron's or i.i.d. weighted bootstrap cases is essentially the same as the bound achieved by the approximate minimum penalized contrast estimator considered in Theorem 2. So one can see in the same way that it is optimal in a global minimax sense over classes of functions based on VC classes.

(ii) Furthermore, it is easy to see in the proof of Theorem 3 that the term $V_m \log^2 n/n$ in the risk upper bound for the i.i.d. weighted bootstrap can be removed when the $(V_i - \mathbb{E}[V_i])$'s satisfy a subgaussian inequality.

(iii) As in Theorem 2, we shall also remark that the constant $1/\mathbb{E}\left[(W_{n,1}-1)_+\right]$ which appears in the penalty term for technical reasons is probably not the optimal one. Our practical study actually tends to show that the real factor is closer to 1 in the Efron's bootstrap case, and $\sqrt{\mathbb{E}\left[V_1^2\right]/\text{Var}\left[V_1\right]}$ in the i.i.d. weighted bootstrap case, as expected from Giné and Zinn (1990) and Præstgaard and Wellner's (1993) asymptotic results.

3 Simulation study

In this section, we study the data-driven penalties proposed above from a practical point of view. As Lozano (2000) and Bartlett et al. (2002), we focus on the issue which is usually referred to as the intervals model selection problem and which can be described as follows.

Let us set for all u, v in $\mathbb{N}$, $u \leq v$, $[\![u, v]\!] = [u, v] \cap \mathbb{N}$. We consider $\mathcal{X} = \{1, \ldots, 2^N\}$ and some partition $\{[\![u_l, v_l]\!],\ l \in \mathcal{L}\}$ of $\mathcal{X}$. Let X be a random variable uniformly distributed on $\mathcal{X}$ and Y a $\{0, 1\}-$valued random variable such that

$$\mathbb{P}[Y = 1|X \in S_0] = \frac{1}{2} + h \quad \text{and} \quad \mathbb{P}[Y = 1|X \notin S_0] = \frac{1}{2} - h,$$

where h is some margin parameter in $]0, 1/2[$ and $S_0 = \cup_{l \in \mathcal{L}_0 \subset \mathcal{L}} [\![u_l, v_l]\!]$. Then the target is the piecewise constant function defined by $s(x) = \mathbb{I}_{S_0}(x)$ for x in $\mathcal{X}$.

We choose to distinguish two cases. The first one corresponds to a target based on a regular partition of $\mathcal{X}$ and the second one to a target based on some irregular partition.

The aim of this simulation study is to illustrate the theoretical results stated in the previous section. Since we consider here the problem from the global minimax point of view, we only evaluate the relevance of the studied data-driven penalties as estimators for global suprema of the form $\sup_{t \in S_m}(-\overline{\gamma_n}(t)) = \sup_{t \in S_m}(\mathbb{E}[\mathbb{I}_{t(X) \neq Y}] - n^{-1}\Sigma_{i=1}^n \mathbb{I}_{t(X_i) \neq Y_i})$. It is now well known that the estimation of these global suprema does not provide tight complexity measures, and that it does not lead to the appropriate penalties when some margin type conditions are considered (see Bartlett et al., 2005; Bartlett et al., 2004; or Koltchinskii, 2003, for instance for further details). Hence, what we call here the "ideal" penalty $\text{pen}_{id}(m) = \sup_{t \in S_m}(-\overline{\gamma_n}(t))$ is not in fact the real ideal penalty as soon as one assumes that some margin condition holds. However, this issue is out of the scope of the present paper, and it may be the subject of a future work.

3.1 Regular case

We are interested here in the case where the joint law of (X, Y) is based on a regular partition of $\mathcal{X} = \{1, \ldots, 2^N\}$. More precisely, we take

$$S_0 = \bigcup_{k \in \{2p+1,\, p \in \mathbb{N},\, 2p+1 \leq 2^{J_0}-1\}} [\![(k-1)2^{N-J_0} + 1, k2^{N-J_0}]\!],$$

with $N = 8$, $J_0 = 2$, $h = 0.05$ first, $N = 8$, $J_0 = 6$, $h = 0.1$ then.

We choose a collection $\{S_m,\ m \in \mathcal{M}\}$ of *regular models* such that $\mathcal{M} = \{2, 2^2, \ldots, 2^N\}$ and for all $m = 2^J$ in $\mathcal{M}$,

$$S_{2^J} = \left\{ t : \mathcal{X} \to \{0, 1\},\ t = \sum_{k=1}^{2^J} c_k \mathbb{I}_{[\![(k-1)2^{N-J}+1, k2^{N-J}]\!]},\ c_1, \ldots, c_{2^J} \in \{0, 1\} \right\}.$$

Let $n \geq 2^N$ and introduce a sample $\xi = (X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from the distribution of (X, Y). The target function s is estimated by the *minimum penalized contrast estimator* $\tilde{s} = \hat{s}_{\hat{m}}$, where

– for each m in $\mathcal{M}$, $\hat{s}_m \in S_m$ and $n^{-1}\sum_{i=1}^n \mathbb{I}_{\hat{s}_m(X_i)\neq Y_i} \leq n^{-1}\sum_{i=1}^n \mathbb{I}_{t(X_i)\neq Y_i}$ for all t in S_m,
– $\hat{m} = \operatorname{argmin}_{m\in\mathcal{M}}\left\{n^{-1}\sum_{i=1}^n \mathbb{I}_{\hat{s}_m(X_i)\neq Y_i} + \operatorname{pen}(m)\right\}$.

The penalty terms that we propose in Section 2 are of the form

$$\mathbb{E}\left[\sup_{t\in S_m}\frac{c_1}{n}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right] + c_2\sqrt{\frac{x_m}{n}},$$

for various $(Z_1,\ldots,Z_n)$.

We consider the following Z_i's:

($\mathcal{R}$) $Z_1,\ldots,Z_n$ is a sequence of i.i.d. Rademacher variables.
($\mathcal{E}$) $Z_i = (1 - W_{n,i})$ where $(W_{n,1},\ldots,W_{n,n})$ is a multinomial vector with parameters $(n, n^{-1},\ldots,n^{-1})$.
(Γ) $Z_i = \sqrt{5}(1 - V_i/\overline{V_n})$ where $(V_1,\ldots,V_n)$ is a sample of i.i.d. random variables with a Gamma(4) distribution.

The case ($\mathcal{E}$) corresponds to the Efron's bootstrap type penalty term while (Γ) is a typical example of the i.i.d. weighted bootstrap type penalty terms (see Praestgaard and Wellner, 1993). Though the weights V_i are frequently taken as exponential random variables with parameter 1 (see Rubin, 1981; Lo, 1987), all the experiments that we have carried out have shown that such an option does not work in the present problem. The choice of the V_i's with a Gamma(4) distribution comes from the Bayesian bootstrap investigated by Weng (1989).

A usual choice for the sequence $(x_m)_{m\in\mathcal{M}}$ consists in $x_m = \log m$ for all m in $\mathcal{M}$. Since the VC dimension V_m of $\{\{x \in \mathcal{X}, t(x) = 1\},\ t \in S_m\}$ is known to be equal to m and since for the Z_i's that we consider, $\mathbb{E}[\sup_{t\in S_m} n^{-1}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}]$ is of the order of $\sqrt{V_m/n}$, we do not take the x_m's into account in the simulation study. Furthermore, the expression of the penalty is deduced from an evaluation of the quantity $\operatorname{pen}_{id}(m) = \sup_{t\in S_m}(-\overline{\gamma_n}(t))$. To optimize the method, the constant c_1 in the penalty term has to be chosen so that $\mathbb{E}[\sup_{t\in S_m} c_1 n^{-1}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}|\xi] \approx \operatorname{pen}_{id}(m)$. An experimental computation of the ratio $\mathbb{E}[\sup_{t\in S_m} c_1 n^{-1}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}]/\mathbb{E}[\operatorname{pen}_{id}]$ for the considered targets in the cases ($\mathcal{R}$), ($\mathcal{E}$) and (Γ) leads to a choice of $c_1 = 1$, as in Lozano (2000). Thus, the penalties are taken as

$$\operatorname{pen}(m) = \mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right].$$

In order to conduct the experiments, we need to compute for each m in $\mathcal{M}$,

$$\inf_{t\in S_m}\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{t(X_i)\neq Y_i} \quad \text{and} \quad \sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i},$$

for the considered $(Z_1, \ldots, Z_n)$. Our algorithm is constructed from the following observations. For $m = 2^J$ in $\mathcal{M}$,

$$\inf_{t\in S_m} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{t(X_i)\neq Y_i} = \frac{1}{n} \inf_{t\in S_m} \sum_{k=1}^{2^J} \sum_{X_i\in[\![(k-1)2^{N-J}+1,k2^{N-J}]\!]} \mathbb{I}_{t(X_i)\neq Y_i}$$

$$= \frac{1}{n} \min_{c_1,\ldots,c_{2^J}\in\{0,1\}} \sum_{k=1}^{2^J} \sum_{X_i\in[\![(k-1)2^{N-J}+1,k2^{N-J}]\!]} \mathbb{I}_{Y_i\neq c_k}.$$

Hence, setting for all p in $[\![1, 2^N]\!]$, q in $[\![p, 2^N]\!]$,

$$\begin{aligned} \Delta(p,q) &= \min_{c\in\{0,1\}} \sum_{i\in[\![1,n]\!], X_i\in[\![p,q]\!]} \mathbb{I}_{Y_i\neq c}, \\ \inf_{t\in S_{2^J}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{t(X_i)\neq Y_i} &= \frac{1}{n}\sum_{k=1}^{2^J} \Delta((k-1)2^{N-J}+1, k2^{N-J}). \end{aligned} \tag{7}$$

In the same way, if for p in $[\![1, 2^N]\!]$, q in $[\![p, 2^N]\!]$,

$$\begin{aligned} \Delta Z(p,q) &= \max_{c\in\{0,1\}} \sum_{i\in[\![1,n]\!], X_i\in[\![p,q]\!]} Z_i \mathbb{I}_{Y_i\neq c}, \\ \sup_{t\in S_{2^J}} \frac{1}{n}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i} &= \frac{1}{n}\sum_{k=1}^{2^J} \Delta Z((k-1)2^{N-J}+1, k2^{N-J}). \end{aligned} \tag{8}$$

The interest of such expressions lies in the fact that the $\Delta((k-1)2^{N-J}+1, k2^{N-J})$'s and $\Delta Z((k-1)2^{N-J}+1, k2^{N-J})$'s are very easy to calibrate.

The following figures present the experimental results. The penalty terms are estimated by 100 simulations. Figures 1, 2, 5 and 6 display the mean penalty terms pen(m) to be compared with the mean "ideal" penalty pen_{id}, and the mean criteria $\text{crit}(m) = n^{-1}\sum_{i=1}^{n} \mathbb{I}_{\hat{s}_m(X_i)\neq Y_i} + \text{pen}(m)$ as functions of the complexity m with a sample size equal to 500 for the two considered targets. Figures 3 and 7 display the risks (estimated by 200 experiments) of the chosen classification rules as functions of the sample size n while in Figs. 4 and 8, one can see the percentages of good model (or complexity) selection obtained over 200 experiments.

Comments: One can see in Figs. 1 and 5 that all the considered penalties track rather well the behavior of the bench mark quantity $\text{pen}_{id}(m) = \sup_{t\in S_m}(-\overline{\gamma_n}(t))$. While the Rademacher and Efron's bootstrap penalty terms underestimate pen_{id}, the i.i.d Gamma weigthed bootstrap one overevaluates it. This explains why Rademacher and Efron's bootstrap penalizations do not perform as well as i.i.d Gamma weigthed bootstrap penalization when the target is based on a partition with few pieces (see Figs. 3 and 4). When the problem of intervals model selection is more complex, that is when the target has 64 pieces, one can notice that Rademacher penalization has the best performance for sample sizes smaller than 1500, while i.i.d Gamma weighted bootstrap give the best results when $n \geq 1500$. In view of these results, we will choose i.i.d. Gamma weighted bootstrap penalization when the target is based on a regular

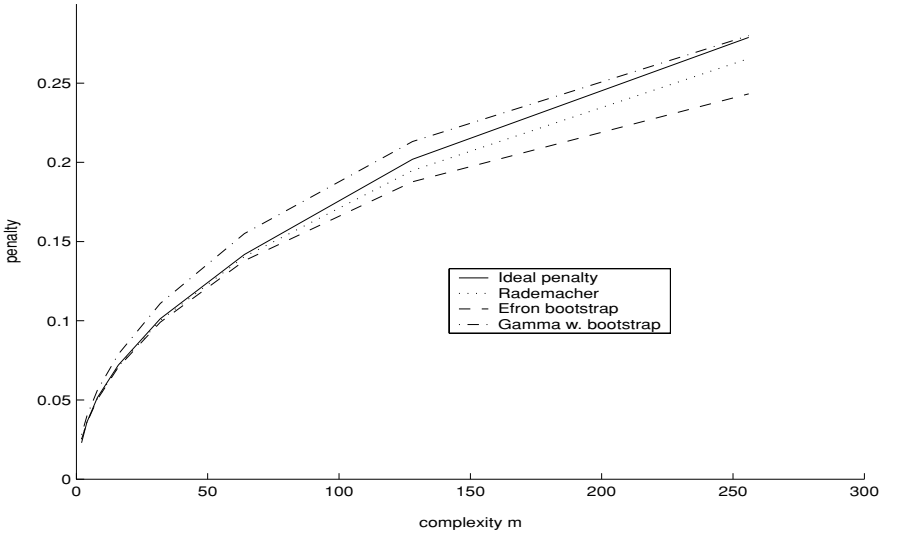

Fig. 1 Penalty terms for a target with 4 pieces over 256 points with $h = 0.05$, $n = 500$

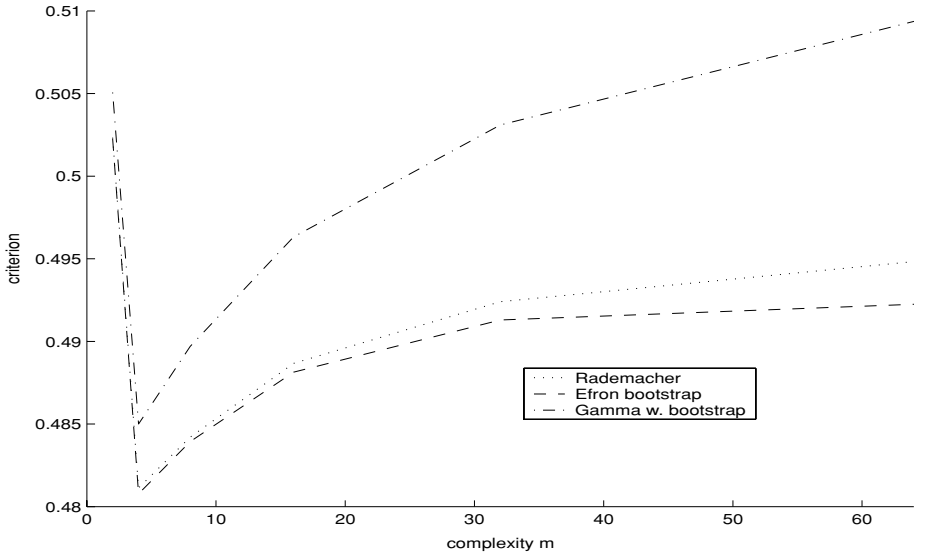

Fig. 2 Criteria for a target with 4 pieces over 256 points with $h = 0.05$, $n = 500$

partition with few pieces or when the sample size is more than 1500, Rademacher penalization otherwise.

3.2 Irregular case

When the target is not known to be based on some regular partition of $\mathcal{X} = \{1, \ldots, 2^N\}$, we can not only consider some collections of *regular models* any more. In fact, we take a

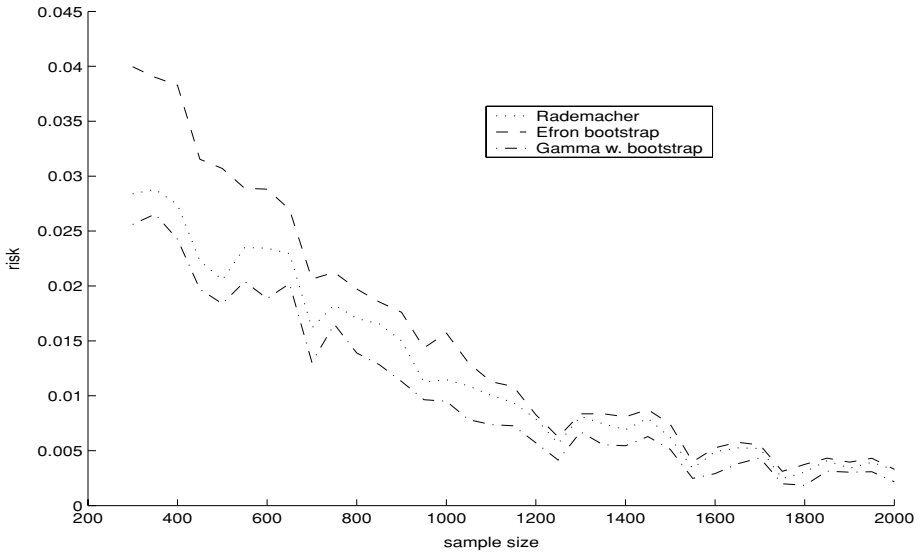

Fig. 3 Risks for a target with 4 pieces over 256 points with $h = 0.05$

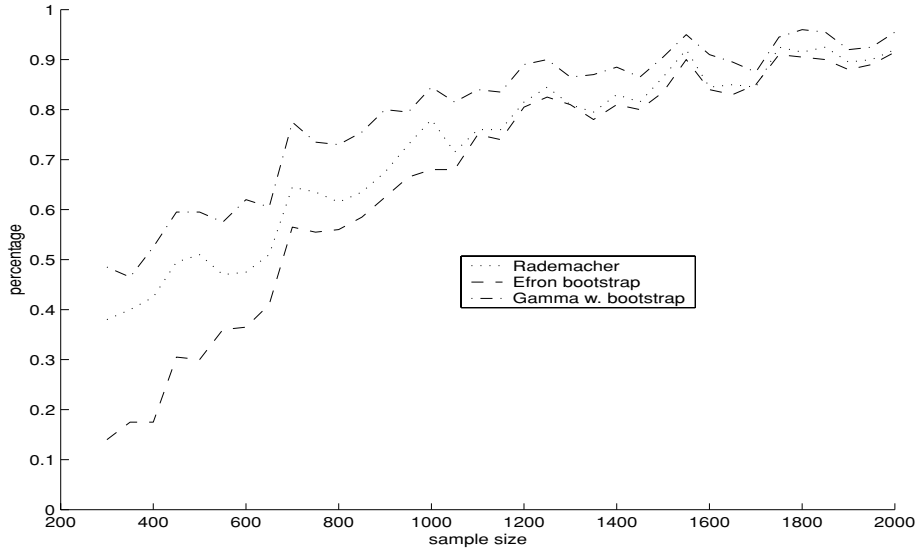

Fig. 4 Percentages of good complexity selection, target with 4 pieces, 256 points, $h = 0.05$

collection containing, for each complexity D in $\{2, 2^2, \ldots, 2^N\}$, all the models based on the partitions of $\mathcal{X}$ with D pieces.

For $u_1, \ldots, u_{D-1}$ in $\mathbb{N}$ with $1 \leq u_1 < u_2 < \ldots < u_{D-1} < 2^N$, we denote by $[u_1, \ldots, u_{D-1}]$ the partition of $\mathcal{X}$ such that:

$$[u_1, \ldots, u_{D-1}] = \{[\![1, u_1]\!], [\![u_1 + 1, u_2]\!], \ldots, [\![u_{D-1} + 1, 2^N]\!]\}.$$

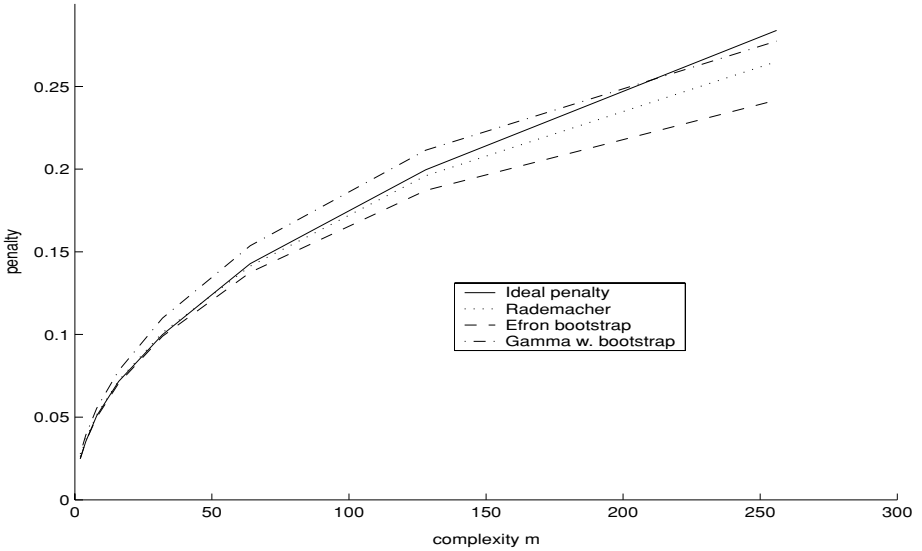

Fig. 5 Penalty terms for a target with 64 pieces over 256 points with $h = 0.1, n = 500$

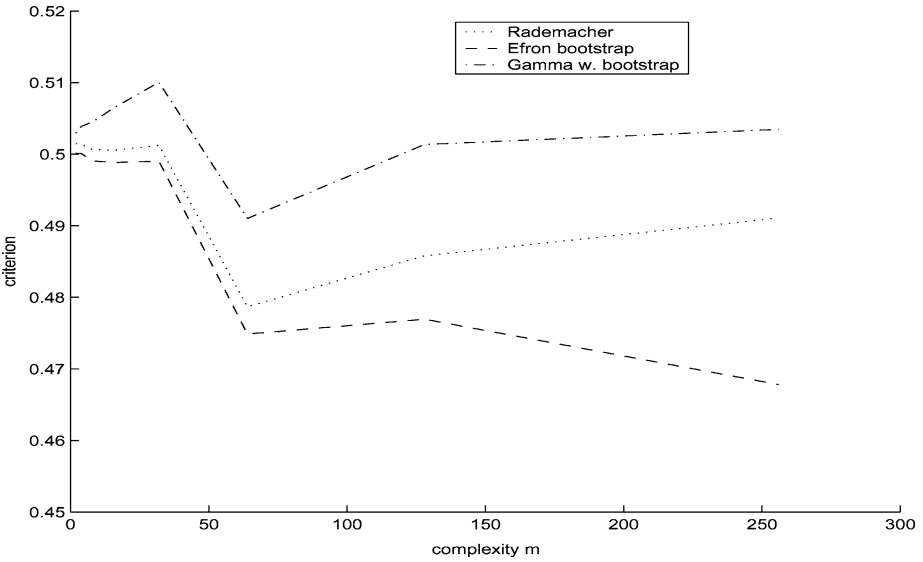

Fig. 6 Criteria for a target with 64 pieces over 256 points with $h = 0.1, n = 500$

Let $\mathcal{D} = \{2, 2^2, \ldots, 2^N\}$ and for all D in $\mathcal{D}$,

$$\mathcal{M}_D = \{[u_1, \ldots, u_{D-1}],\ 1 \leq u_1 < u_2 < \ldots < u_{D-1} < 2^N\}.$$

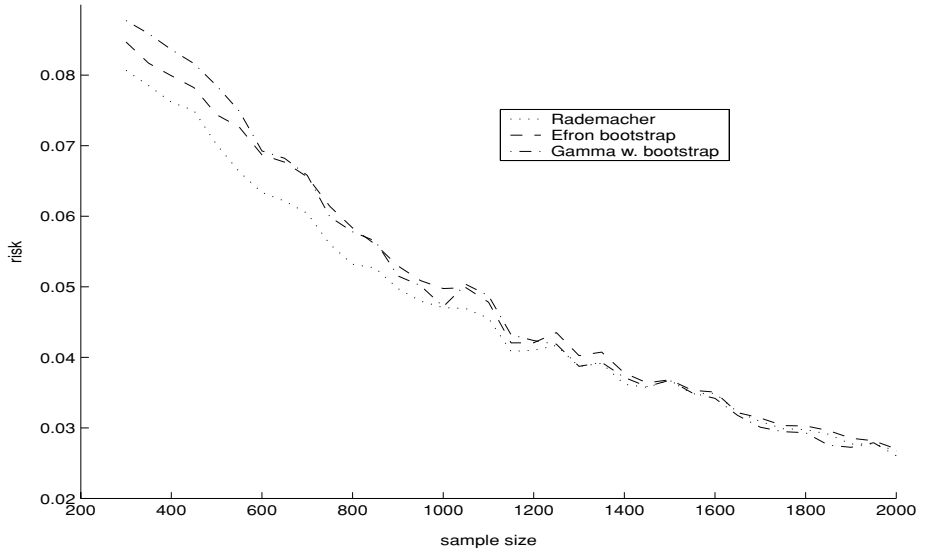

Fig. 7 Risks for a target with 64 pieces over 256 points with $h = 0.1$

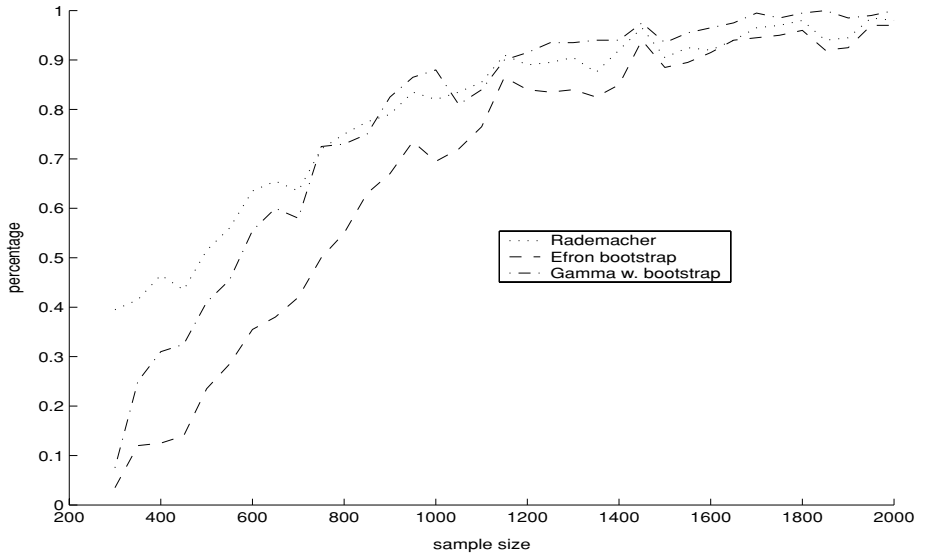

Fig. 8 Percentages of good complexity selection, target with 64 pieces, 256 points, $h = 0.1$

We choose the collection of models $\{S_m,\ m \in \mathcal{M}\}$ such that $\mathcal{M} = \cup_{D \in \mathcal{D}} \mathcal{M}_D$ and for all $m = [u_1, \ldots, u_{D-1}]$ in $\mathcal{M}$, setting $u_0 = 0$ and $u_D = 2^N$,

$$S_m = \left\{ t = \sum_{k=1}^{D} c_k \mathbb{I}_{[\![u_{k-1}+1, u_k]\!]},\ c_1, \ldots, c_D \in \{0, 1\} \right\}.$$

We are interested here in the case where the joint law of (X, Y) is based on

$$S_0 = \bigcup_{k\in\{2p+1,\, p\in\mathbb{N}, 2p+1\le 2^{J_0}-1\}} [\![U_{k-1}, U_k]\!],$$

with $U_0 = 1$ and $U_1, \ldots, U_{2^{J_0}-1}$ randomly chosen on $\{1, \ldots, 2^N - 1\}$ in such a way that $1 \le U_1 < \cdots < U_{2^{J_0}-1} < 2^N$. We take a target with $N = 6$, $J_0 = 2$, $h = 0.05$ first and $N = 6$, $J_0 = 4$, $h = 0.1$ then. Let $n \ge 2^N$ and introduce a sample $\xi = (X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from the distribution of (X, Y).

Since the penalties that we consider essentially depend on the complexities of the models in the collection, a natural idea is to group all the models with the same complexity together (see Lebarbier, 2002, for further details). It is then a matter of selecting a complexity in $\mathcal{D}$ instead of a model in $\{S_m,\ m \in \mathcal{M}\}$. Let us introduce for all D in $\mathcal{D}$,

$$\mathcal{S}_D = \bigcup_{1\le u_1<u_2<\cdots<u_{D-1}<2^N} S_{[u_1,\ldots,u_{D-1}]}.$$

The chosen classification rule is $\hat{s}_{\hat{D}}$ such that:

– for each D in $\mathcal{D}$, $\hat{s}_D$ belongs to $\mathcal{S}_D$ and satisfies for all t in $\mathcal{S}_D$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\hat{s}_D(X_i)\neq Y_i} \le \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{t(X_i)\neq Y_i},$$

– $\hat{D} = \operatorname{argmin}_{D\in\mathcal{D}}\{n^{-1}\sum_{i=1}^{n} \mathbb{I}_{\hat{s}_D(X_i)\neq Y_i} + \operatorname{pen}(D)\}$.

The penalty $\operatorname{pen}(D)$ is taken, as in the regular case, equal to $\mathbb{E}[\sup_{t\in\mathcal{S}_D} n^{-1}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i} | \xi]$, with the Z_i's of the cases $(\mathcal{R})$, $(\mathcal{E})$ and (Γ). Thus, we have to compute for each D in $\mathcal{D}$,

- $\inf_{t\in\mathcal{S}_D} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{t(X_i)\neq Y_i}$,
- $\hat{s}_D = \operatorname*{argmin}_{t\in\mathcal{S}_D} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{t(X_i)\neq Y_i}$,
- $\sup_{t\in\mathcal{S}_D} \frac{1}{n}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i}$ for various $(Z_1, \ldots, Z_n)$.

Let us see first how to compute $\inf_{t\in\mathcal{S}_D} n^{-1}\sum_{i=1}^{n} \mathbb{I}_{t(X_i)\neq Y_i}$ and $\sup_{t\in\mathcal{S}_D} n^{-1}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i}$. As in the regular case, for all D in $\mathcal{D}$, if $u_0 = 0$ and $u_D = 2^N$, with Δ and ΔZ defined by (7) and (8), we can prove that

$$\inf_{t\in\mathcal{S}_D} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{t(X_i)\neq Y_i} = \frac{1}{n} \min_{u_0<u_1<\cdots<u_{D-1}<u_D} \sum_{k=1}^{D} \Delta(u_{k-1}+1, u_k),$$

and

$$\sup_{t\in\mathcal{S}_D} \frac{1}{n}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i} = \frac{1}{n} \max_{u_0<u_1<\cdots<u_{D-1}<u_D} \sum_{k=1}^{D} \Delta Z(u_{k-1}+1, u_k).$$

Following Lebarbier (2002), we can then use the dynamical algorithm developed by Kay (1998) which can be described as follows. We set for $1 \leq p \leq q \leq 2^N$,

$$I(p,q) = \min_{u_0=0<u_1<\cdots<u_{p-1}<u_p=q} \sum_{k=1}^{p} \Delta(u_{k-1}+1, u_k),$$

and

$$IZ(p,q) = \max_{u_0=0<u_1<\cdots<u_{p-1}<u_p=q} \sum_{k=1}^{p} \Delta Z(u_{k-1}+1, u_k).$$

For all q in $[\![1, 2^N]\!]$,

$$I(1,q) = \Delta(1,q) \quad \text{and} \quad IZ(1,q) = \Delta Z(1,q),$$

and for all p in $[\![1, 2^N]\!]$, q in $[\![p, 2^N]\!]$,

$$I(p,q) = \min_{r \in [\![p-1, q-1]\!]} \{I(p-1,r) + \Delta(r+1,q)\}$$

and

$$IZ(p,q) = \max_{r \in [\![p-1, q-1]\!]} \{IZ(p-1,r) + \Delta Z(r+1,q)\}.$$

By computing all the $\Delta(p,q)$'s and $\Delta Z(p,q)$'s for $1 \leq p \leq q \leq 2^N$, we can obtain the values of:

1. $I(1,q)$ and $IZ(1,q)$ for all q in $[\![1, 2^N]\!]$
2. $I(2,q)$ and $IZ(2,q)$ for q in $[\![2, 2^N]\!]$ from the values of the $I(1,r)$'s and $IZ(1,r)$'s
3. and so on,

until we get

$$I(D, 2^N) = \inf_{t \in \mathcal{S}_D} \sum_{i=1}^{n} \mathbb{I}_{t(X_i) \neq Y_i} \quad \text{and} \quad IZ(D, 2^N) = \sup_{t \in \mathcal{S}_D} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i}.$$

To compute $\hat{s}_D$, we notice that if $\hat{u}_0 = 0$ and $\hat{u}_D = 2^N$, then $\hat{s}_D = \sum_{k=1}^{D} \hat{c}_k \mathbb{I}_{[\![\hat{u}_{k-1}+1, \hat{u}_k]\!]}$, with

- $[\hat{u}_1, \ldots, \hat{u}_{D-1}] = \underset{1 \leq u_1 < \ldots < u_{D-1} < 2^N}{\operatorname{argmin}} \left(\min_{t \in S_{[u_1, \cdots, u_{D-1}]}} \sum_{i=1}^{n} \mathbb{I}_{t(X_i) \neq Y_i} \right)$,
- $\hat{c}_k = \underset{c \in \{0,1\}}{\operatorname{argmin}} \left(\sum_{i \in [\![1,n]\!], X_i \in [\![\hat{u}_{k-1}+1, \hat{u}_k]\!]} \mathbb{I}_{Y_i \neq c} \right)$.

Hence, setting for $2 \leq p \leq q \leq 2^N$,

$$J(p,q) = \underset{r \in [\![p-1, q-1]\!]}{\operatorname{argmin}} \{I(p-1,r) + \Delta(r+1,q)\},$$

$$\hat{u}_{D-1} = J(D,N), \hat{u}_{D-2} = J(D-1, \hat{u}_{D-1}), \ldots, \hat{u}_1 = J(2, \hat{u}_2).$$

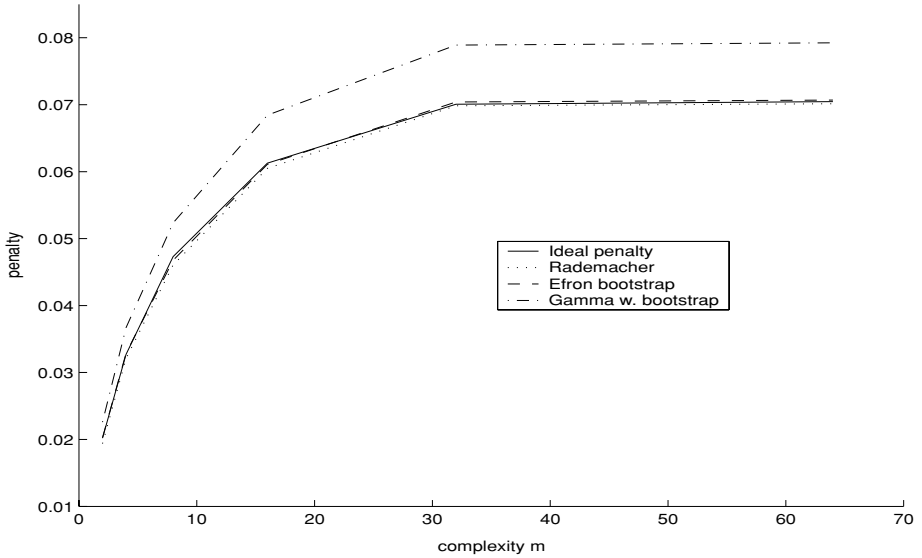

Fig. 9 Penalty terms for a target with 4 pieces over 64 points with $h = 0.05$, $n = 2000$

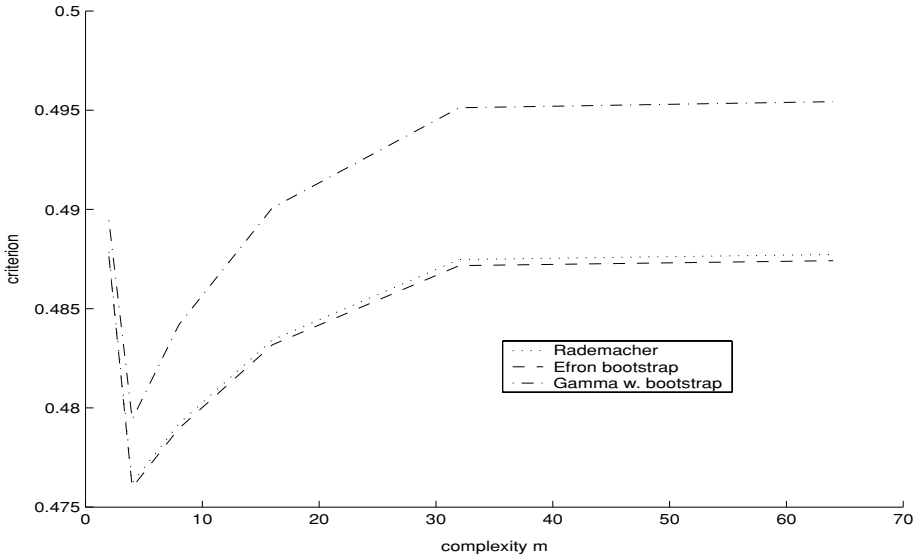

Fig. 10 Criteria for a target with 4 pieces over 64 points with $h = 0.05$, $n = 2000$

The values of the penalties are estimated by 100 simulations. Figures 9, 10, 13 and 14 display the mean penalty terms $\text{pen}(D)$ to be compared with the mean of $\text{pen}_{id}(D) = \sup_{t \in \mathcal{S}_D}(-\overline{\gamma_n}(t))$ and the mean criteria as functions of the complexity D for a sample size of 2000 for the two considered targets. We show in Figs. 11 and 15 the risks (estimated by 100 experiments) of the chosen classification rules as functions of the sample size. The percentages of good complexity selection are presented in Figs. 12 and 16.

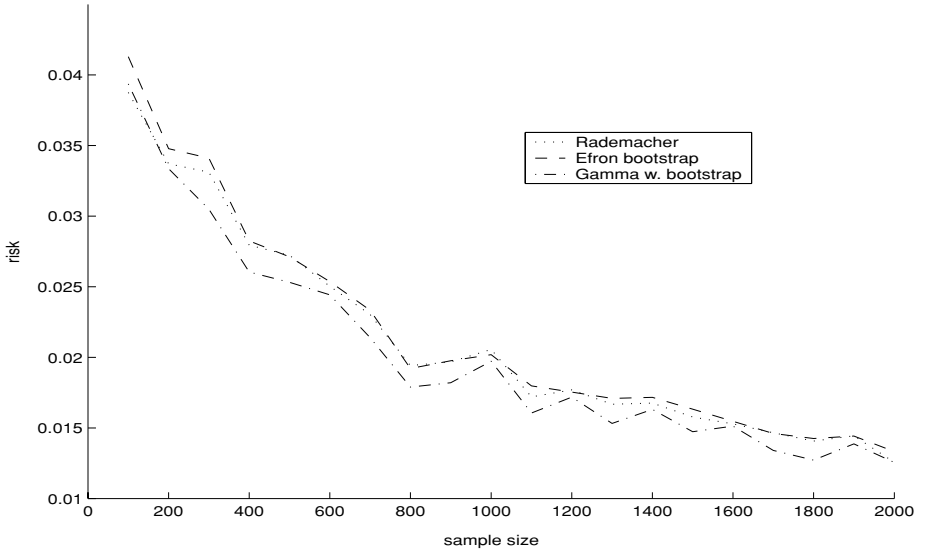

Fig. 11 Risks for a target with 4 pieces over 64 points with $h = 0.05$

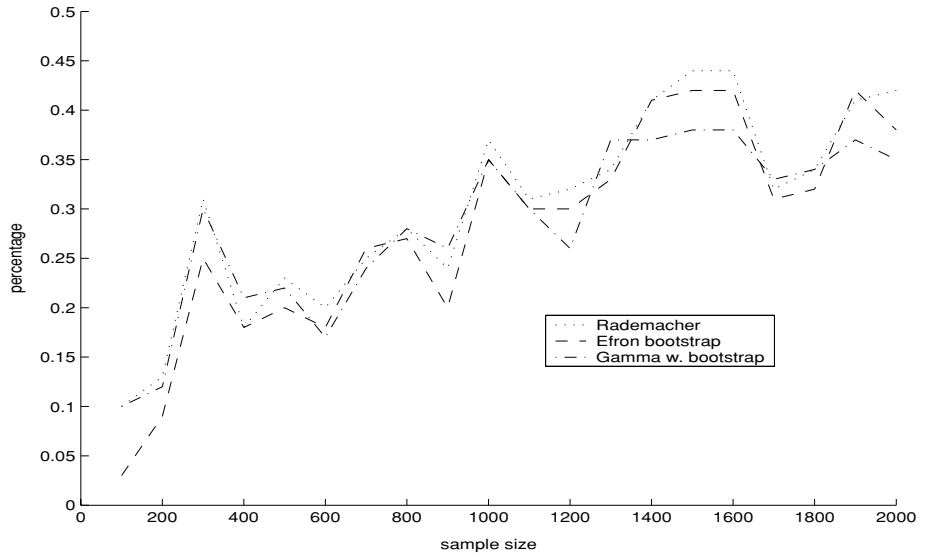

Fig. 12 Percentages of good complexity selection, target with 4 pieces, 64 points, $h = 0.05$

Comments: Overall, one can observe the same behaviors as in the regular case, though the complexity of the true model is, as expected, more difficult to select, whatever the penalization method. The three considered penalty terms resemble closely pen_{id}. Figure 11 show that Rademacher and Efron's bootstrap penalizations, which have a very similar behavior here, are less good in terms of risks than i.i.d. Gamma weighted bootstrap penalization for a target

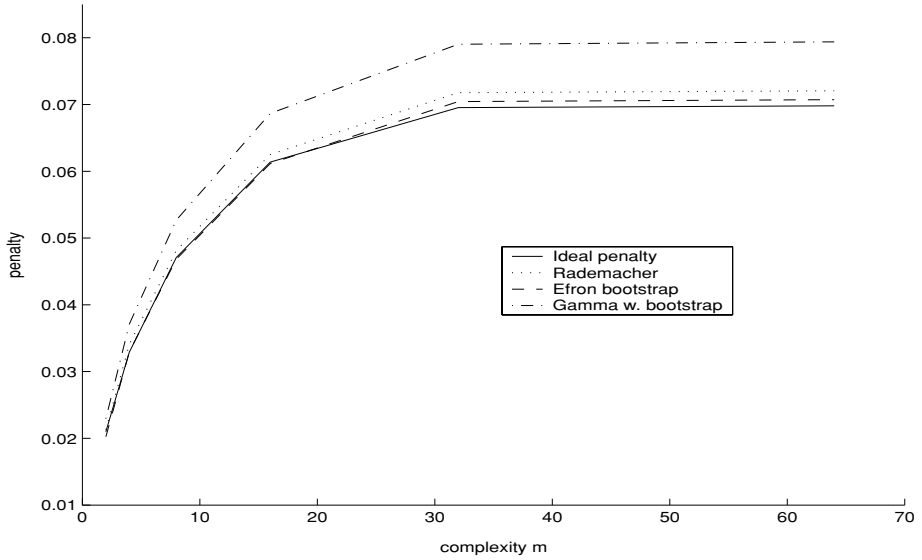

Fig. 13 Penalty terms for a target with 16 pieces over 64 points with $h = 0.1$, $n = 2000$

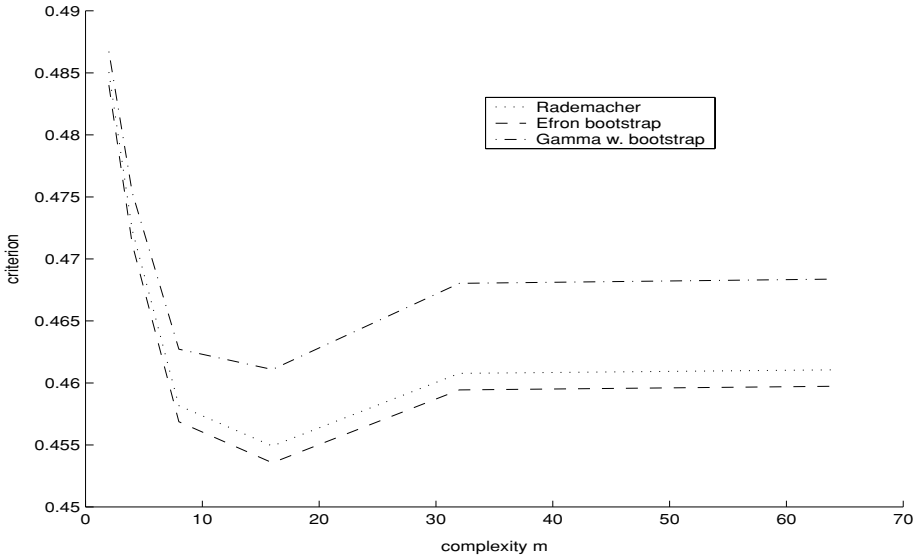

Fig. 14 Criteria for a target with 16 pieces over 64 points with $h = 0.1$, $n = 2000$

with 4 pieces. However, this trend is reversed when the target is more complex (see Fig. 15). Consequently, we would rather choose Rademacher or Efron's bootstrap penalizations if we had to face a complex problem of classification, i.i.d. Gamma weighted bootstrap penalization in simpler cases.

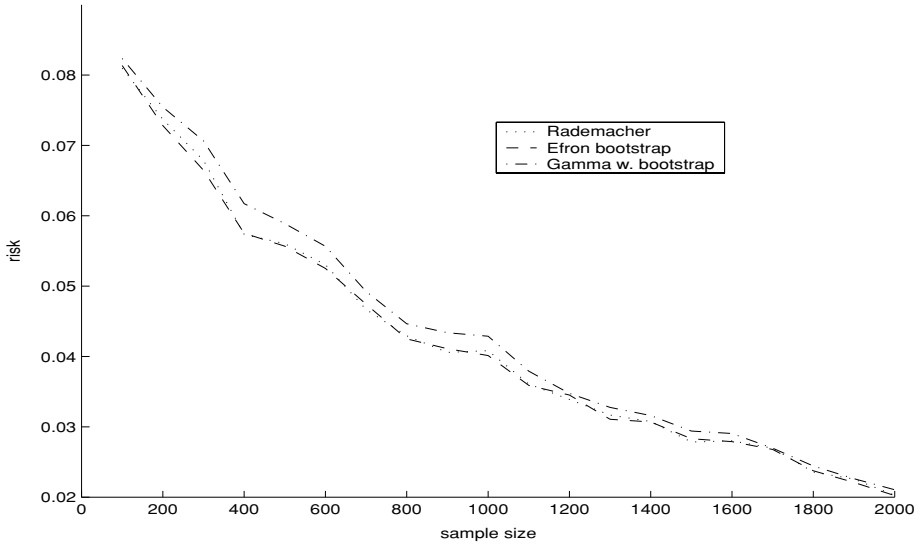

Fig. 15 Risks for a target with 16 pieces over 64 points with $h = 0.1$

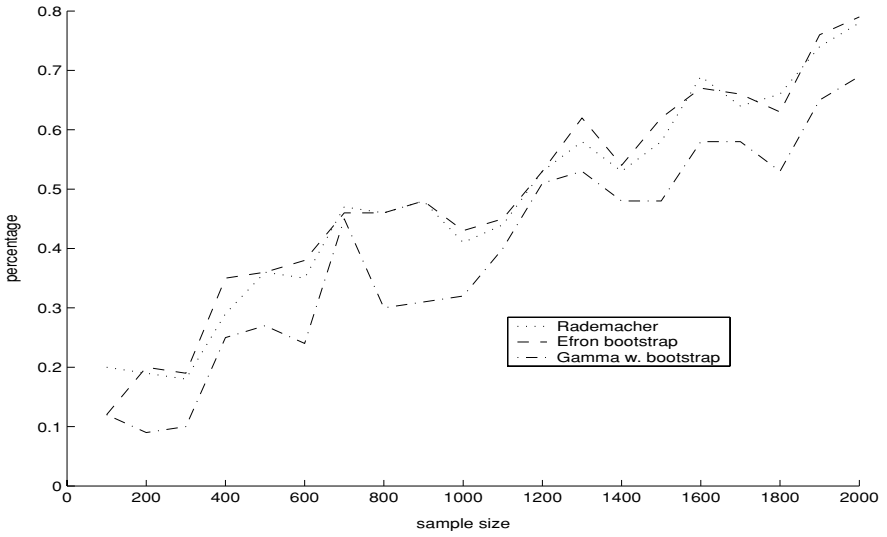

Fig. 16 Percentages of good complexity selection, target with 16 pieces, 64 points, $h = 0.1$

4 Exponential inequalities

As we have seen in the beginning of the paper, to obtain the expected upper bounds for the risk of our estimators, we need to control $\sup_{t\in S_m}(-\overline{\gamma_n}(t))$ uniformly for m in $\mathcal{M}$, where $\overline{\gamma_n}$ is defined by (3). We then use some exponential inequalities that we present in this section.

4.1 An exponential inequality based on symmetrization

We propose here a generalization of the exponential inequality for $\sup_{t\in S_m}(-\overline{\gamma_n}(t))$ used by Koltchinskii (2001) and Bartlett et al. (2002). It is based on symmetrization arguments combined with McDiarmid's inequality.

Let $\xi = (\xi_1, \ldots, \xi_n)$ be a sample of n independent identically distributed random variables with values in some probability space Ξ and with common distribution P. Let P_n be the corresponding empirical process defined by

$$P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(\xi_i),$$

and let

$$P(f) = \mathbb{E}\left[f(\xi_1)\right].$$

• *Symmetrization inequality*

The following lemma directly derives from a symmetrization tool which has been introduced in the empirical processes theory by Koltchinskii (1981), Pollard (1982) and especially Giné and Zinn (1984) (see Appendix for the proof).

Lemma 1. *Let $\mathcal{F}$ be a countable set of measurable functions from Ξ to $[0, 1]$. For every sequence of i.i.d. symmetric variables $Z_1, \ldots, Z_n$ independent of ξ such that $\mathbb{E}[|Z_1|] < +\infty$,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P - P_n)(f)\right] \leq \frac{2}{n\mathbb{E}[|Z_1|]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} Z_i f(\xi_i)\right].$$

• *Exponential bound*

Proposition 1. *Let $\xi = (\xi_1, \ldots, \xi_n)$ be a sample of n i.i.d. random variables with values in some probability space Ξ and with common distribution P. Denote by P_n the corresponding empirical process. Consider some countable set $\mathcal{F}$ of measurable functions from Ξ to $[0, 1]$. Let $Z_1, \ldots, Z_n$ be a sequence of i.i.d. symmetric variables independent of ξ and such that $\mathbb{E}[|Z_1|] < +\infty$. For any $x > 0$, the following inequality holds*:

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P - P_n)(f) - \frac{2}{n\mathbb{E}[|Z_1|]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} Z_i f(\xi_i)\middle|\xi\right] \geq 3\sqrt{\frac{x}{2n}}\right] \leq e^{-x}.$$

Proof: Let

$$T_n = \frac{2}{n\mathbb{E}[|Z_1|]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} Z_i f(\xi_i)\middle|\xi\right].$$

Lemma 1 leads to the inequality:

$$\sup_{f\in\mathcal{F}}(P-P_n)(f)-T_n \le \sup_{f\in\mathcal{F}}(P-P_n)(f)-T_n-\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P-P_n)(f)-T_n\right]. \quad (9)$$

We now use McDiarmid's concentration inequality stated in Theorem 1. Since every f in $\mathcal{F}$ has its values in [0, 1], we have that for all i in $\{1,\dots,n\}$, for all $x_1,\dots,x_n,x_i'$ in Ξ,

$$\left|\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{j=1}^{n}(P(f)-f(x_j))-\sup_{f\in\mathcal{F}}\frac{1}{n}\left(\sum_{1\le j\le n,\, j\ne i}(P(f)-f(x_j))+(P(f)-f(x_i'))\right)\right|\le\frac{1}{n}.$$

Let us introduce some copy $(\xi_1',\dots,\xi_n')$ of $\xi=(\xi_1,\dots,\xi_n)$, independent of ξ and $Z_1,\dots,Z_n$. Setting $x=(x_1,\dots,x_n)$, $x^i=(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)$ and $\xi^i=(\xi_1,\dots,\xi_{i-1},\xi_i',\xi_{i+1},\dots,\xi_n)$, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}Z_jf(\xi_j)\middle|\xi=x\right]-\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}Z_jf(\xi_j)\middle|\xi=x^i\right]\right|\\
&\quad\le\left|\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}Z_jf(\xi_j)\middle|\xi=x,\xi_i'=x_i'\right]-\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}Z_jf\left(\xi_j^i\right)\middle|\xi=x,\xi_i'=x_i'\right]\right|\\
&\quad\le\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|Z_i(f(\xi_i)-f(\xi_i'))\right|\middle|\xi=x,\xi_i'=x_i'\right]\\
&\quad\le\mathbb{E}\left[|Z_1|\right].
\end{aligned}$$

Hence,

$$\left|\frac{2}{n\mathbb{E}\left[|Z_1|\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}Z_jf(\xi_j)\middle|\xi=x\right]-\frac{2}{n\mathbb{E}\left[|Z_1|\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}Z_jf(\xi_j)\middle|\xi=x^i\right]\right|\le\frac{2}{n},$$

and the variable $\phi(\xi_1,\dots,\xi_n)=(\sup_{f\in\mathcal{F}}(P-P_n)(f)-T_n)$ satisfies the assumptions of McDiarmid's inequality with $c_i=3/n$ for all i in $\{1,\dots,n\}$. With (9), this directly leads to the result. □

4.2 A bootstrap exponential inequality

We still consider a sample $\xi=(\xi_1,\dots,\xi_n)$ of n independent identically distributed random variables with values in some probability space Ξ and with common distribution P. Recall that P_n denotes the empirical process defined by

$$P_n(f)=\frac{1}{n}\sum_{i=1}^{n}f(\xi_i),$$

and that

$$P(f)=\mathbb{E}\left[f(\xi_1)\right].$$

If $W_n = (W_{n,1}, \ldots, W_{n,n})$ denotes a vector of n exchangeable and nonnegative random variables independent of ξ and satisfying $\sum_{i=1}^n W_{n,i} = n$, the corresponding exchangeably weighted bootstrap empirical process is defined by

$$P_n^w(f) = \frac{1}{n}\sum_{i=1}^n W_{n,i} f(\xi_i). \tag{10}$$

Proposition 2. *Let $\xi = (\xi_1, \ldots, \xi_n)$ be a sample of n independent identically distributed random variables with values in some probability space Ξ and with common distribution P. Introduce a vector $W_n = (W_{n,1}, \ldots, W_{n,n})$ of n exchangeable and nonnegative random variables independent of ξ and satisfying $\sum_{i=1}^n W_{n,i} = n$. Denote by P_n the empirical process associated with ξ and consider the exchangeably weighted bootstrap empirical process defined by (10). Let $\mathcal{F}$ be some countable set of measurable functions from Ξ to $[0,1]$. For any $x > 0$, the following inequality holds:*

$$\mathbb{P}\Bigg[\sup_{f\in\mathcal{F}}(P-P_n)(f) - \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\Bigg[\sup_{f\in\mathcal{F}}\left(P_n - P_n^w\right)(f)\Bigg|\xi\Bigg] \geq \left(1 + \frac{\mathbb{E}\left[|W_{n,1}-1|\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\right)\sqrt{\frac{x}{2n}}\ \Bigg] \leq e^{-x}.$$

Proof: Since we do not deal with symmetrized variables any more, we here need to replace the symmetrization inequality of Lemma 1 by another argument.

By Jensen's inequality, we get:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P-P_n)(f)\right] \\
&\quad\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}\left[(W_{n,i}-1)\mathbb{I}_{W_{n,i}\geq 1}\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}(P(f)-f(\xi_i))\right] \\
&\quad\leq \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (W_{n,i}-1)\mathbb{I}_{W_{n,i}\geq 1}(P(f)-f(\xi_i))\Bigg|\xi\right]\right] \\
&\quad\leq \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n (W_{n,i}-1)\mathbb{I}_{W_{n,i}\geq 1}(P(f)-f(\xi_i))\right] \\
&\quad\leq \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n (W_{n,i}-1)\mathbb{I}_{W_{n,i}\geq 1}(P(f)-f(\xi_i))\Bigg|W_n\right]\right].
\end{aligned}$$

It is well known that if U and V are independent random variables such that for all g in a class of functions $\mathcal{G}$, $\mathbb{E}\left[g(V)\right] = 0$, then

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} g(U)\right] \leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}(g(U)+g(V))\right]. \tag{11}$$

To see this, we only notice that

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}} g(U)\Big] = \mathbb{E}\Big[\sup_{g\in\mathcal{G}} \mathbb{E}[g(U)+g(V)|U]\Big],$$

and use Jensen's inequality. Since W_n is independent of ξ, conditionnally given W_n, the variables $(\xi_1\mathbb{I}_{W_{n,1}\geq 1},\ldots,\xi_n\mathbb{I}_{W_{n,n}\geq 1})$ and $(\xi_1\mathbb{I}_{W_{n,1}<1},\ldots,\xi_n\mathbb{I}_{W_{n,n}<1})$ are independent and for all f in $\mathcal{F}$, the variable $\sum_{i=1}^n (W_{n,i}-1)\mathbb{I}_{W_{n,i}<1}(P(f)-f(\xi_i))$ is centered. So, applying (11) conditionnally given W_n, one gets:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P-P_n)(f)\right] \leq \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n (W_{n,i}-1)(P(f)-f(\xi_i))\right],$$

that is

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P-P_n)(f)\right] \leq \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(P_n-P_n^w\right)(f)\right]. \tag{12}$$

We now need to prove that

$$\begin{aligned}&\sup_{f\in\mathcal{F}}(P-P_n)(f) - \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(P_n-P_n^w\right)(f)\middle|\xi\right]\\ &\quad= \sup_{f\in\mathcal{F}}(P-P_n)(f) - \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n (1-W_{n,i})f(\xi_i)\middle|\xi\right]\end{aligned}$$

concentrates around its expectation.

As in the proof of Proposition 1, one can easily see that

$$\phi(\xi_1,\ldots,\xi_n) = \sup_{f\in\mathcal{F}}(P-P_n)(f) - \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n (1-W_{n,i})f(\xi_i)\middle|\xi\right]$$

satisfies the assumptions of McDiarmid's inequality with $c_i = (1+\frac{\mathbb{E}\left[|W_{n,1}-1|\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]})/n$. This completes the proof of Proposition 2.

□

5 Proofs of Theorems 2 and 3

Theorems 2 and 3 both involve some penalties of the form:

$$\mathrm{pen}(m) = c_1\mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right] + c_2\sqrt{\frac{x_m}{n}},$$

with specific random weights $Z_1,\ldots,Z_n$. To derive adaptive properties for the proposed estimators, we need to compute $\mathbb{E}\left[\mathrm{pen}(m)\right]$, and a fortiori $\mathbb{E}[\sup_{t\in S_m} n^{-1}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}]$.

5.1 A maximal inequality

This section is devoted to the calibration of an upper bound for $\mathbb{E}\big[\sup_{t\in S_m}\sum_{i=1}^n Z_i\mathbb{I}_{t(X_i)\neq Y_i}/n\big]$ in a VC framework, provided that the Z_i's satisfy independence and moments conditions precised below.

Using a result stated in Massart (2003), we can prove the following lemma:

Lemma 2. *Let $\mathcal{A}$ be some finite subset of $[0,1]^n$ and $Z_1,\ldots,Z_n$ be i.i.d. centered real random variables satisfying the moments condition:*

$$\forall k\geq 2,\ \mathbb{E}\left[|Z_1|^k\right]\leq \frac{k!}{2}vc^{k-2}.$$

If $\sup_{(a_1,\ldots,a_n)\in\mathcal{A}}\sum_{i=1}^n a_i^2\leq\delta^2$ and $sup_{(a_1,\ldots,a_n)\in\mathcal{A}}sup_{1\leq i\leq n}|a_i|\leq\beta$ for some positive numbers δ and β, denoting by $|\mathcal{A}|$ the cardinality of $\mathcal{A}$, one has:

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^n a_iZ_i\right]\leq\delta\sqrt{2v\log|\mathcal{A}|}+c\beta\log|\mathcal{A}|.$$

If the Z_i's satisfy the subgaussian inequality $\mathbb{E}[e^{\lambda Z_1}]\leq e^{\lambda^2/2}$ for all $\lambda\geq 0$, one has in fact

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^n a_iZ_i\right]\leq\delta\sqrt{2\log|\mathcal{A}|}.$$

Proof: It is easy to see that under the assumptions of Lemma 2, for all $k\geq 2$,

$$\sum_{i=1}^n\mathbb{E}\left[|a_iZ_i|^k\right]\leq\frac{k!}{2}v\delta^2(c\beta)^{k-2}.$$

From the special version of Bernstein's inequality due to Birgé and Massart (1998), we can deduce that for all $\lambda\in[0,1/(c\beta)[$,

$$\log\mathbb{E}\left[e^{\lambda\sum_{i=1}^n a_iZ_i}\right]\leq\frac{\lambda^2v\delta^2}{2(1-c\beta\lambda)}.$$

Furthermore, in the subgaussian case, for all $\lambda\geq 0$,

$$\log\mathbb{E}\left[e^{\lambda\sum_{i=1}^n a_iZ_i}\right]\leq\frac{\lambda^2\delta^2}{2}.$$

Hence, using Lemma 2.3 in Massart (2003) which follows from an argument due to Pisier (see also Massart and Rio (1998)) we get the expected upper bound.

Let us consider $S_m=\{\mathbb{I}_C,\ C\in\mathcal{C}_m\}$ where $\mathcal{C}_m$ is some countable class of subsets of $\mathcal{X}$.

For each m in $\mathcal{M}$, denote by H_m the Vapnik-Chervonenkis entropy of $\mathcal{C}_m$ that is

$$H_m = \log |\{C \cap \{X_1, \ldots, X_n\}, C \in \mathcal{C}_m\}| .$$

Let $Z_1, \ldots, Z_n$ be i.i.d. centered real random variables satisfying the moments condition:

$$\forall k \geq 2,\ \mathbb{E}[|Z_1|^k] \leq \frac{k!}{2} v c^{k-2}.$$

Applying Lemma 2 with

$$\mathcal{A} = \left\{ \left(\mathbb{I}_{t(X_1) \neq Y_1}, \ldots, \mathbb{I}_{t(X_n) \neq Y_n} \right),\ t \in \{ \mathbb{I}_C,\ C \in \mathcal{C}_m \} \right\}$$

gives

$$\mathbb{E}\left[\sup_{t \in S_m} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i} \right] \leq (\sqrt{2vn\mathbb{E}[H_m]} + c\mathbb{E}[H_m]).$$

If in addition, each $\mathcal{C}_m$ is assumed to be a VC class with VC dimension V_m, Sauer's bound (see Sauer, 1972) provides the following upper bound:

$$\mathbb{E}\left[\sup_{t \in S_m} \frac{1}{n} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i} \right] \leq \sqrt{2v} \sqrt{\frac{V_m}{n} \left(1 + \log \left(\frac{n}{V_m} \right) \right)} + c \frac{V_m}{n} \left(1 + \log \left(\frac{n}{V_m} \right) \right).$$

In fact, by refining this result via some quite classical chaining arguments, we can see that the factor $\log(n/V_m)$ in the dominating term is avoidable. This is the object of the following theorem. □

Theorem 4. *Let $\xi = (X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of n independent copies of a random pair (X, Y) with values in $\mathcal{X} \times \{0, 1\}$. Introduce n i.i.d. real random variables $Z_1, \ldots, Z_n$ centered, independent of ξ and satisfying the moments condition*:

$$\forall k \geq 2,\ \mathbb{E}[|Z_1|^k] \leq \frac{k!}{2} v c^{k-2}, \tag{13}$$

for some positive constants v and c. Let $S_m = \{\mathbb{I}_C, C \in \mathcal{C}_m\}$ where $\mathcal{C}_m$ is a VC class with VC dimension V_m and assume that $n \geq 4$. There exist some absolute constants κ_1 and κ_2 such that:

$$\mathbb{E}\left[\frac{1}{n} \sup_{t \in S_m} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i} \right] \leq \kappa_1 \sqrt{v} \sqrt{\frac{V_m}{n}} + \kappa_2 c \frac{V_m}{n} \log^2 n,$$

and if for all $\lambda \geq 0$, $\mathbb{E}[e^{\lambda Z_1}] \leq e^{\lambda^2/2}$, then

$$\mathbb{E}\left[\frac{1}{n} \sup_{t \in S_m} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i} \right] \leq \kappa_1 \sqrt{\frac{V_m}{n}}.$$

Proof: The proof of Theorem 4 is based on the following lemma. Let for all $a = (a_1, \ldots, a_n)$ in $\mathbb{R}^n$, $\|a\|_2^2 = \sum_{i=1}^n a_i^2$. For all $\varepsilon > 0$ and all subset $\mathcal{A}$ of $\mathbb{R}^n$, let $H_2(\varepsilon, \mathcal{A})$ denote the logarithm of the maximal number N of elements $\{a^{(1)}, \ldots, a^{(N)}\}$ in $\mathcal{A}$ such that for every $l, l' \in \{1, \ldots, N\}$, $l \neq l'$, $\|a^{(l)} - a^{(l')}\|_2^2 > \varepsilon^2$. □

Lemma 3. *Let $\mathcal{A}$ be some subset of $[0,1]^n$ and $Z_1, \ldots, Z_n$ i.i.d. centered real random variables. Let $\delta > 0$ such that $\sup_{a\in\mathcal{A}}\|a\|_2 \leq \delta$ and assume that there exist some positive constants v and c such that the Z_i's satisfy the moments condition (13). Then, one has*

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^n a_i Z_i\right] \leq 3\sum_{j=0}^{+\infty}\left(\delta\sqrt{v}2^{-j}\sqrt{H_2\left(2^{-(j+1)}\delta, \mathcal{A}\right)} + c(2^{-j}\delta \wedge 1)H_2\left(2^{-(j+1)}\delta, \mathcal{A}\right)\right),$$

and if for all $\lambda \geq 0$, $\mathbb{E}\left[e^{\lambda Z_1}\right] \leq e^{\lambda^2/2}$,

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^n a_i Z_i\right] \leq 3\sum_{j=0}^{+\infty} 2^{-j}\delta\sqrt{H_2\left(2^{-(j+1)}\delta, \mathcal{A}\right)}.$$

The proof of this lemma is directly inspired by Lemma 6.1 in Massart (2003) (see Appendix for further details).

Considering

$$\mathcal{B}_m = \{\{(x, y) \in \mathcal{X} \times \{0, 1\},\ \mathbb{I}_C(x) \neq y\},\ C \in \mathcal{C}_m\}$$

and the set

$$\mathcal{A}_m = \{(\mathbb{I}_B(X_1, Y_1), \ldots, \mathbb{I}_B(X_n, Y_n)),\ B \in \mathcal{B}_m\},$$

one has

$$\mathbb{E}\left[\sup_{t\in S_m}\sum_{i=1}^n Z_i \mathbb{I}_{t(X_i)\neq Y_i}\right] = \mathbb{E}\left[\sup_{a\in\mathcal{A}_m}\sum_{i=1}^n a_i Z_i\right].$$

Moreover,

$$\sup_{a\in\mathcal{A}_m}\|a\|_2 \leq \sqrt{n},$$

and by definition of $\mathcal{A}_m$, for all $\varepsilon > 0$,

$$H_2(\sqrt{n}\varepsilon, \mathcal{A}_m) = H(\varepsilon, \mathcal{B}_m, P_n),$$

where $H(\varepsilon, \mathcal{B}_m, P_n)$ is the $\varepsilon-$metric entropy of $\mathcal{B}_m$ with respect to the empirical measure $n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$. For any probability measure Q, the $\varepsilon-$metric entropy $H(\varepsilon, \mathcal{B}_m, Q)$ of $\mathcal{B}_m$ with respect to Q is the logarithm of the maximal number N of elements $\{b^{(1)}, \ldots, b^{(N)}\}$

in $\{\mathbb{I}_B,\ B \in \mathcal{B}_m\}$ such that for all $l, l' \in \{1, \ldots, N\},\ l \neq l'$, $\mathbb{E}_Q(b^{(l)} - b^{(l')})^2 > \varepsilon^2$. Let us denote by $H(\varepsilon, \mathcal{B}_m)$ the universal $\varepsilon-$metric entropy of $\mathcal{B}_m$ that is

$$H(\varepsilon, \mathcal{B}_m) = \sup_Q H(\varepsilon, \mathcal{B}_m, Q),$$

where the supremum is taken over all the probabilty measures on $\mathcal{X} \times \{0, 1\}$. Then, for all $\varepsilon > 0$,

$$H_2(\sqrt{n}\varepsilon, \mathcal{A}_m) \leq H(\varepsilon, \mathcal{B}_m),$$

and from Lemma 3, we get

$$\mathbb{E}\left[\sup_{t \in S_m} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i}\right] \leq 3 \sum_{j=0}^{+\infty} \left(\sqrt{v} 2^{-j} \sqrt{n H\left(2^{-(j+1)}, \mathcal{B}_m\right)} + c(2^{-j}\sqrt{n} \wedge 1) H\left(2^{-(j+1)}, \mathcal{B}_m\right)\right).$$

Since $\mathcal{B}_m$ is a VC class with VC dimension not larger than V_m, Haussler's (1995) bound leads to

$$H\left(2^{-(j+1)}, \mathcal{B}_m\right) \leq \kappa V_m (1 + (j+1)\log 2)\ \forall j \in \mathbb{N},$$

for some positive constant κ. Hence,

$$\mathbb{E}\left[\sup_{t \in S_m} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i}\right] \leq 3 \sum_{j=0}^{+\infty} (\sqrt{v\kappa} 2^{-j} \sqrt{n V_m (1 + (j+1)\log 2)} + c\kappa (2^{-j}\sqrt{n} \wedge 1) V_m (1 + (j+1)\log 2)).$$

On the one hand, since the function $x \mapsto 2^{-x}\sqrt{1 + (x+1)\log 2}$ is decreasing on $]0, +\infty[$, one has

$$\sum_{j=0}^{+\infty} 2^{-j}\sqrt{1 + (j+1)\log 2} \leq \sqrt{1 + \log 2} + \int_0^{+\infty} 2^{-x}\sqrt{1 + (x+1)\log 2}\, dx \leq \sqrt{1 + \log 2} + \frac{\sqrt{1 + \log 2}}{\log 2} + \frac{\sqrt{\pi}}{2\log 2}.$$

On the other hand, one has

$$\sum_{j=0}^{+\infty} (2^{-j}\sqrt{n} \wedge 1)(1 + (j+1)\log 2) = \sum_{j \leq \frac{\log n}{2\log 2}} (1 + (j+1)\log 2) + \sqrt{n} \sum_{j > \frac{\log n}{2\log 2}} 2^{-j}(1 + (j+1)\log 2).$$

It is easy to see that if $n \geq 4$,

$$\sum_{j \leq \frac{\log n}{2 \log 2}} (1 + (j+1)\log 2) \leq \frac{1}{2 \log 2}\left(\frac{3}{2} + \frac{1}{\log 2}\right) \log^2 n,$$

and since $x \mapsto 2^{-x}(1 + (x+1)\log 2)$ is decreasing,

$$\begin{aligned}\sum_{j > \frac{\log n}{2 \log 2}} 2^{-j}(1 + (j+1)\log 2) &\leq \int_{\frac{\log n}{2 \log 2} - 1}^{+\infty} 2^{-x}(1 + (x+1)\log 2) dx \\ &\leq \frac{1}{\log 2}\left(1 + \frac{2}{\log 2}\right)\frac{\log n}{\sqrt{n}},\end{aligned}$$

which gives

$$\mathbb{E}\left[\frac{1}{n} \sup_{t \in S_m} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i}\right] \leq \kappa_1 \sqrt{v}\sqrt{\frac{V_m}{n}} + \kappa_2 c \frac{V_m}{n} \log^2 n.$$

The upper bound in the subgaussian case is obtained in the same way.

We can now prove the results stated in Section 2.

5.2 Proof of Theorem 2

• *A risk upper bound*

Recall that for m in $\mathcal{M}$, $\tilde{s}$ satisfies the inequality (4):

$$l(s, \tilde{s}) \leq l(s, s_m) + \overline{\gamma_n}(s_m) + \text{pen}(m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}) - \text{pen}(\hat{m}) + \rho_n,$$

with

$$\overline{\gamma_n}(t) = \gamma_n(t) - \mathbb{E}\left[\mathbb{I}_{t(X) \neq Y}\right].$$

Applying Proposition 1 with $\xi = (X_1, Y_1), \ldots, (X_n, Y_n)$ and $\mathcal{F} = \{(x, y) \mapsto \mathbb{I}_{t(x) \neq y}, t \in S_{m'}\}$ gives that for all m' in $\mathcal{M}$, for all $x > 0$:

$$\mathbb{P}\left[\sup_{t \in S_{m'}} (-\overline{\gamma_n}(t)) - \frac{2}{n\mathbb{E}\left[|Z_1|\right]}\mathbb{E}\left[\sup_{t \in S_{m'}} \sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i) \neq Y_i} \Big| \xi\right] \geq 3\sqrt{\frac{x}{2n}}\right] \leq e^{-x}.$$

Introduce a family $(x_m)_{m \in \mathcal{M}}$ of nonnegative weights such that for some absolute constant Σ,

$$\sum_{m \in \mathcal{M}} e^{-x_m} \leq \Sigma.$$

For $\zeta > 0$, we can deduce from the above inequality that except on a set of probability not larger than $\Sigma e^{-\zeta}$, one has for every m' in $\mathcal{M}$,

$$\sup_{t\in S_{m'}}(-\overline{\gamma_n}(t)) \le \frac{2}{n\mathbb{E}[|Z_1|]}\mathbb{E}\left[\sup_{t\in S_{m'}}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right] + 3\sqrt{\frac{x_{m'}+\zeta}{2n}}.$$

This implies that if

$$\hat{\Sigma}_m = \frac{1}{n\mathbb{E}[|Z_1|]}\mathbb{E}\left[\sup_{t\in S_m}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right],$$

except on a set of probability not larger than $\Sigma e^{-\zeta}$,

$$l(s,\tilde{s}) \le l(s,s_m) + \overline{\gamma_n}(s_m) + \mathrm{pen}(m) + 2\hat{\Sigma}_{\hat{m}} + 3\sqrt{\frac{x_{\hat{m}}}{2n}} - \mathrm{pen}(\hat{m}) + \rho_n + 3\sqrt{\frac{\zeta}{2n}}$$

holds. Therefore, if for all m' in $\mathcal{M}$,

$$\mathrm{pen}(m') = \frac{2}{n\mathbb{E}[|Z_1|]}\mathbb{E}\left[\sup_{t\in S_{m'}}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right] + 3\sqrt{\frac{x_{m'}}{2n}} = 2\hat{\Sigma}_{m'} + 3\sqrt{\frac{x_{m'}}{2n}},$$

then

$$\mathbb{P}\left[l(s,\tilde{s}) \ge l(s,s_m) + \overline{\gamma_n}(s_m) + \mathrm{pen}(m) + \rho_n + 3\sqrt{\frac{\zeta}{2n}}\right] \le \Sigma e^{-\zeta}.$$

By integration with respect to ζ, this leads to:

$$\mathbb{E}[(l(s,\tilde{s}) - l(s,s_m) - \overline{\gamma_n}(s_m) - \mathrm{pen}(m) - \rho_n)_+] \le \frac{3\Sigma}{2}\sqrt{\frac{\pi}{2n}}.$$

Since $\mathbb{E}[\overline{\gamma_n}(s_m)] = 0$, we obtain that

$$\mathbb{E}[l(s,\tilde{s})] \le l(s,s_m) + \mathbb{E}[\mathrm{pen}(m)] + \frac{3\Sigma}{2}\sqrt{\frac{\pi}{2n}} + \rho_n,$$

which gives, since m can be taken arbitrarily in $\mathcal{M}$, the final risk bound:

$$\mathbb{E}[l(s,\tilde{s})] \le \inf_{m\in\mathcal{M}}\{l(s,s_m) + \mathbb{E}[\mathrm{pen}(m)]\} + \frac{3\Sigma}{2}\sqrt{\frac{\pi}{2n}} + \rho_n.$$

• *Adaptive properties in a minimax sense*

Let for all m in $\mathcal{M}$, $S_m = \{\mathbb{I}_C,\ C \in \mathcal{C}_m\}$, where $\mathcal{C}_m$ is a VC class with VC dimension $V_m \ge 1$ and assume that $n \ge 4$.

Since the Z_i's satisfy the appropriate independence and moments conditions, a direct application of the maximal inequality given in Theorem 4 gives that for all $x > 0$:

$$\mathbb{E}\left[\frac{1}{n}\sup_{t\in S_m}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i}\right] \leq \kappa_1\sqrt{v}\sqrt{\frac{V_m}{n}} + \kappa_2 c\frac{V_m}{n}\log^2 n,$$

and if for all $\lambda \geq 0$, $\mathbb{E}[e^{\lambda Z_1}] \leq e^{\lambda^2/2}$, then

$$\mathbb{E}\left[\frac{1}{n}\sup_{t\in S_m}\sum_{i=1}^{n} Z_i \mathbb{I}_{t(X_i)\neq Y_i}\right] \leq \kappa_1\sqrt{\frac{V_m}{n}}.$$

This concludes the proof of Theorem 2.

5.3 Proof of Theorem 3

• *A risk upper bound*

From Proposition 2 with $\xi = (X_1, Y_1), \ldots, (X_n, Y_n)$ and $\mathcal{F} = \{(x, y) \mapsto \mathbb{I}_{t(x)\neq y},\ t \in S_{m'}\}$, we derive that for all m' in $\mathcal{M}$, for all $x > 0$:

$$\mathbb{P}\Bigg[\sup_{t\in S_{m'}}(-\overline{\gamma_n}(t)) - \frac{1}{\mathbb{E}[(W_{n,1}-1)_+]}\mathbb{E}\left[\sup_{t\in S_{m'}}(\gamma_n(t)-\gamma_n^w(t))\Big|\xi\right] \geq \left(1+\frac{\mathbb{E}[|W_{n,1}-1|]}{\mathbb{E}[(W_{n,1}-1)_+]}\right)\sqrt{\frac{x}{2n}}\Bigg] \leq e^{-x}.$$

Introduce a family $(x_m)_{m\in\mathcal{M}}$ of nonnegative weights such that for some absolute constant Σ,

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma.$$

If $\zeta > 0$, it follows from the above inequality that except on a set of probability not larger than $\Sigma e^{-\zeta}$, one has for every m' in $\mathcal{M}$,

$$\sup_{t\in S_{m'}}(-\overline{\gamma_n}(t)) \leq \frac{1}{\mathbb{E}[(W_{n,1}-1)_+]}\mathbb{E}\left[\sup_{t\in S_{m'}}\left(\gamma_n(t)-\gamma_n^w(t)\right)\Big|\xi\right] + \left(1+\frac{\mathbb{E}[|W_{n,1}-1|]}{\mathbb{E}[(W_{n,1}-1)_+]}\right)\sqrt{\frac{x_{m'}+\zeta}{2n}}.$$

This implies as the proof of Theorem 2 that if for all m' in $\mathcal{M}$,

$$\mathrm{pen}(m') = \frac{1}{\mathbb{E}[(W_{n,1}-1)_+]}\mathbb{E}\left[\sup_{t\in S_{m'}}\left(\gamma_n(t)-\gamma_n^w(t)\right)\Big|\xi\right] + \left(1+\frac{\mathbb{E}[|W_{n,1}-1|]}{\mathbb{E}[(W_{n,1}-1)_+]}\right)\sqrt{\frac{x_{m'}}{2n}},$$

then for any m in $\mathcal{M}$,

$$\mathbb{E}[l(s,\tilde{s})] \leq l(s,s_m) + \mathbb{E}[\mathrm{pen}(m)] + \left(1+\frac{\mathbb{E}[|W_{n,1}-1|]}{\mathbb{E}[(W_{n,1}-1)_+]}\right)\frac{\Sigma}{2}\sqrt{\frac{\pi}{2n}} + \rho_n,$$

which gives, by an appropriate choice of m in $\mathcal{M}$, the expected risk bound.

• *Adaptive properties in a minimax sense*

Let for all m in $\mathcal{M}$, $S_m = \{\mathbb{I}_C,\ C \in \mathcal{C}_m\}$, where $\mathcal{C}_m$ is a VC class with VC dimension $V_m \geq 1$ and assume that $n \geq 4$.
As in the proof of the minimax properties in Theorem 2, we here aim at using the maximal inequality given in Theorem 4. The main difficulty that we have to deal with lies in the fact that the weights $(1 - W_{n,i})$ involved in the penalty are not independent any more.

Consider first the Efron's bootstrap case, where W_n is a multinomial vector with parameters $(n, n^{-1}, \ldots, n^{-1})$. To remove the dependence of the $(1 - W_{n,i})$'s, we use the classical tool of Poissonization. Let $(U_1, \ldots, U_n)$ be a sample of n i.i.d. random variables uniformly distributed on $]0, 1[$ independent of ξ such that $W_{n,i} = \sum_{j=1}^{n} \mathbb{I}_{U_j \in](i-1)/n, i/n]}$. Introduce a Poisson random variable N with parameter n independent of ξ and $(U_1, \ldots, U_n)$ and for all i in $\{1, \ldots, n\}$, $N_i = \sum_{j=1}^{N} \mathbb{I}_{U_j \in](i-1)/n, i/n]}$. It is easy to check that the N_i's are independent identically distributed Poisson random variables with parameter 1. Furthermore, one has:

$$\begin{aligned}
&\left| \mathbb{E}\left[\sup_{t \in S_m} \frac{1}{n} \sum_{i=1}^{n} (1 - W_{n,i}) \mathbb{I}_{t(X_i) \neq Y_i} | \xi \right] - \mathbb{E}\left[\sup_{t \in S_m} \frac{1}{n} \sum_{i=1}^{n} (1 - N_i) \mathbb{I}_{t(X_i) \neq Y_i} | \xi \right] \right| \\
&\quad \leq \mathbb{E}\left[\sup_{t \in S_m} \left| \frac{1}{n} \sum_{i=1}^{n} (N_i - W_{n,i}) \mathbb{I}_{t(X_i) \neq Y_i} \right| \middle| \xi \right] \\
&\quad \leq \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} |N_i - W_{n,i}| \right].
\end{aligned}$$

Since according to the definition of the N_i's,

$$\sum_{i=1}^{n} |N_i - W_{n,i}| = \left| \sum_{i=1}^{n} (N_i - W_{n,i}) \right| = |N - n|,$$

we get:

$$\left| \mathbb{E}\left[\sup_{t \in S_m} \frac{1}{n} \sum_{i=1}^{n} (1 - W_{n,i}) \mathbb{I}_{t(X_i) \neq Y_i} \middle| \xi \right] - \mathbb{E}\left[\sup_{t \in S_m} \frac{1}{n} \sum_{i=1}^{n} (1 - N_i) \mathbb{I}_{t(X_i) \neq Y_i} \middle| \xi \right] \right| \leq \frac{1}{\sqrt{n}}.$$

If

$$\hat{W}_m = \mathbb{E}\left[\sup_{t \in S_m} \left(\gamma_n(t) - \gamma_n^w(t) \right) \middle| \xi \right],$$

then

$$\mathbb{E}\left[\hat{W}_m \right] \leq \mathbb{E}\left[\sup_{t \in S_m} \frac{1}{n} \sum_{i=1}^{n} (1 - N_i) \mathbb{I}_{t(X_i) \neq Y_i} \right] + \frac{1}{\sqrt{n}}.$$

The fact that N_1 is a Poisson variable with parameter 1 implies that for all $k \geq 3$,

$$\frac{\mathbb{E}\left[|1-N_1|^k\right]}{k!} \leq \mathbb{E}\left[e^{|1-N_1|}\right] - 1 - \mathbb{E}\left[|1-N_1|\right] - \frac{\mathbb{E}\left[(1-N_1)^2\right]}{2} \leq \frac{e^e-1}{e^2} - \frac{2}{e} - \frac{1}{2}.$$

Hence the $(1-N_i)$'s are i.i.d centered real random variables satisfying the moments condition (13) with $v=1$ and $c=1$ and Theorem 4 leads to:

$$\mathbb{E}\left[\hat{W}_m\right] \leq \left(\kappa_1\sqrt{\frac{V_m}{n}} + \kappa_2\frac{V_m}{n}\log^2 n + \frac{1}{\sqrt{n}}\right).$$

The equalities $\mathbb{E}\left[(W_{n,1}-1)_+\right] = (1-1/n)^n$ and $\mathbb{E}\left[|W_{n,1}-1|\right] = 2(1-1/n)^n$ allow to complete the proof.

Consider then the i.i.d. weighted bootstrap case, where $W_{n,i} = V_i/\overline{V_n}$, $V_1, \ldots, V_n$ being i.i.d. positive random variables independent of ξ. We here prove that

$$\hat{W}_m = \mathbb{E}\left[\sup_{t\in S_m}\left(\gamma_n(t) - \gamma_n^w(t)\right)\middle|\xi\right]$$

is bounded from above by

$$\frac{v'(v,c,\mathbb{E}\left[V_1\right])}{\sqrt{n}} + \frac{1}{\mathbb{E}\left[V_1\right]}\mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n(\mathbb{E}\left[V_1\right]-V_i)\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right].$$

Since we have that

$$\begin{aligned}
\hat{W}_m &= \mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n\left(1-\frac{V_i}{\overline{V_n}}\right)\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right] \\
&\leq \mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n\left(\frac{V_i}{\mathbb{E}\left[V_1\right]}-\frac{V_i}{\overline{V_n}}\right)\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right] + \mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n\left(1-\frac{V_i}{\mathbb{E}\left[V_1\right]}\right)\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right] \\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left|\frac{V_i}{\mathbb{E}\left[V_1\right]}-\frac{V_i}{\overline{V_n}}\right|\right] + \frac{1}{\mathbb{E}\left[V_1\right]}\mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n(\mathbb{E}\left[V_1\right]-V_i)\,\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right] \\
&\leq \frac{1}{\mathbb{E}\left[V_1\right]}\mathbb{E}\left[\frac{V_1}{\overline{V_n}}|\overline{V_n}-\mathbb{E}\left[V_1\right]|\right] + \frac{1}{\mathbb{E}\left[V_1\right]}\mathbb{E}\left[\sup_{t\in S_m}\frac{1}{n}\sum_{i=1}^n(\mathbb{E}\left[V_1\right]-V_i)\,\mathbb{I}_{t(X_i)\neq Y_i}\middle|\xi\right],
\end{aligned}$$

it is a matter in fact of controlling $\mathbb{E}\left[(V_1/\overline{V_n})|\overline{V_n}-\mathbb{E}\left[V_1\right]|\right]$.
Assuming that the V_i's satisfy the moments condition (6), the special version of Bernstein's inequality proposed by Birgé and Massart (1998) shows that for all $x>0$,

$$\mathbb{P}\left[|\overline{V_n}-\mathbb{E}\left[V_1\right]| \geq \sqrt{2v}\sqrt{\frac{x}{n}} + c\frac{x}{n}\right] \leq 2e^{-x}, \tag{14}$$

and

$$\mathbb{P}\left[\overline{V_n} \leq \mathbb{E}[V_1] - \sqrt{2v}\sqrt{\frac{x}{n}} - c\frac{x}{n}\right] \leq e^{-x}. \tag{15}$$

Let $a = 3\sqrt{2v}/\mathbb{E}[V_1]$ and $\sqrt{b} = 2(\sqrt{2v} + \sqrt{c\mathbb{E}[V_1]})/\mathbb{E}[V_1]$. From (14), we deduce that for all $x > 0$,

$$\begin{aligned}&\mathbb{P}\left[\frac{V_1}{\overline{V_n}}|\overline{V_n} - \mathbb{E}[V_1]| \geq \left(\sqrt{2v}\sqrt{\frac{x}{n}} + c\frac{x}{n}\right)(1 + a\sqrt{x} + bx)\right]\\ &\quad\leq \mathbb{P}\left[\frac{V_1}{\overline{V_n}} \geq 1 + a\sqrt{x} + bx\right] + 2e^{-x}.\end{aligned}$$

Moreover, since $V_1/\overline{V_n} \leq n$,

$$\mathbb{P}\left[\frac{V_1}{\overline{V_n}} \geq 1 + a\sqrt{x} + bx\right] = \mathbb{P}\left[\frac{V_1}{\overline{V_n}} \geq 1 + a\sqrt{x} + bx,\ a\sqrt{x} \leq n,\ bx \leq n\right],$$

and the exponential inequality (15) leads to:

$$\begin{aligned}&\mathbb{P}\left[\frac{V_1}{\overline{V_n}} \geq 1 + a\sqrt{x} + bx\right]\\ &\quad\leq \mathbb{P}\left[V_1 \geq (1 + a\sqrt{x} + bx)\left(\mathbb{E}[V_1] - \sqrt{2v}\sqrt{\frac{x}{n}} - c\frac{x}{n}\right), a\sqrt{x} \leq n,\ bx \leq n\right] + e^{-x}.\end{aligned}$$

For $n \geq 4$, we have that

$$\begin{aligned}&(1 + a\sqrt{x} + bx)\left(\mathbb{E}[V_1] - \sqrt{2v}\sqrt{\frac{x}{n}} - c\frac{x}{n}\right)\\ &\quad\geq \mathbb{E}[V_1] + \left(a\mathbb{E}[V_1] - \frac{1}{2}\sqrt{2v} - a\sqrt{\frac{x}{n}}\sqrt{2v}\right)\sqrt{x}\\ &\quad+ \left(b\mathbb{E}[V_1] - b\sqrt{\frac{x}{n}}\sqrt{2v} - \frac{c}{4} - ac\frac{\sqrt{x}}{n} - bc\frac{x}{n}\right)x.\end{aligned}$$

If in addition, $a\sqrt{x} \leq n$ and $bx \leq n$, we get:

$$\begin{aligned}&(1 + a\sqrt{x} + bx)\left(\mathbb{E}[V_1] - \sqrt{2v}\sqrt{\frac{x}{n}} - c\frac{x}{n}\right)\\ &\quad\geq \mathbb{E}[V_1] + \left(a\mathbb{E}[V_1] - \frac{1}{2}\sqrt{2v} - \frac{a}{\sqrt{b}}\sqrt{2v}\right)\sqrt{x} + \left(b\mathbb{E}[V_1] - \sqrt{2vb} - \frac{9c}{4}\right)x,\end{aligned}$$

and by definition of a and b, since $\sqrt{b} \geq 2\sqrt{2v}/\mathbb{E}[V_1]$,

$$(1 + a\sqrt{x} + bx)\left(\mathbb{E}[V_1] - \sqrt{2v}\sqrt{\frac{x}{n}} - c\frac{x}{n}\right)$$
$$\geq \mathbb{E}[V_1] + \sqrt{2v}\sqrt{x} + \left(\sqrt{b}(\sqrt{b}\mathbb{E}[V_1] - \sqrt{2v}) - \frac{9c}{4}\right)x.$$

Hence,

$$(1 + a\sqrt{x} + bx)\left(\mathbb{E}[V_1] - \sqrt{2v}\sqrt{\frac{x}{n}} - c\frac{x}{n}\right)$$
$$\geq \mathbb{E}[V_1] + \sqrt{2v}\sqrt{x} + \left(2\sqrt{\frac{c}{\mathbb{E}[V_1]}}\left(\sqrt{2v} + 2\sqrt{c\mathbb{E}[V_1]}\right) - \frac{9c}{4}\right)x$$
$$\geq \mathbb{E}[V_1] + \sqrt{2v}\sqrt{x} + cx.$$

Therefore,

$$\mathbb{P}\left[V_1 \geq (1 + a\sqrt{x} + bx)\left(\mathbb{E}[V_1] - \sqrt{2v}\sqrt{\frac{x}{n}} - c\frac{x}{n}\right),\ a\sqrt{x} \leq n,\ bx \leq n\right] \leq e^{-x}.$$

Finally, we obtain that

$$\mathbb{P}\left[\frac{V_1}{\overline{V_n}}|\overline{V_n} - \mathbb{E}[V_1]| \geq \left(\sqrt{2v}\sqrt{\frac{x}{n}} + c\frac{x}{n}\right)(1 + a\sqrt{x} + bx)\right] \leq 4e^{-x},$$

which implies:

$$\mathbb{P}\left[\frac{V_1}{\overline{V_n}}\left|\overline{V_n} - \mathbb{E}[V_1]\right| \geq (1 + a + b)\left(\frac{\sqrt{2v}}{\sqrt{n}} + \frac{c}{n}\right)\sqrt{x} \vee x^2\right] \leq 4e^{-x}.$$

By integration with respect to $x > 0$, this leads to:

$$\mathbb{E}\left[\frac{V_1}{\overline{V_n}}|\overline{V_n} - \mathbb{E}[V_1]|\right] \leq 4(1 + a + b)\left(\frac{\sqrt{2v}}{\sqrt{n}} + \frac{c}{n}\right)\left(\int_0^1 \frac{1}{2\sqrt{x}}e^{-x}dx + \int_1^{+\infty} 2xe^{-x}dx\right)$$
$$\leq 9(1 + a + b)\left(\frac{\sqrt{2v}}{\sqrt{n}} + \frac{c}{n}\right).$$

Furthermore, since the $(\mathbb{E}[V_1] - V_i)$'s satisfy the moments condition of Theorem 4 with $2v$ instead of v, we have that

$$\frac{1}{\mathbb{E}[V_1]}\mathbb{E}\left[\sup_{t \in S_m}\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[V_1] - V_i)\mathbb{I}_{t(X_i) \neq Y_i}\right] \leq \frac{1}{\mathbb{E}[V_1]}\left(\kappa_1\sqrt{2v}\sqrt{\frac{V_m}{n}} + \kappa_2 c\frac{V_m}{n}\log^2 n\right),$$

which gives the expected bound.

6 Conclusion

In this conclusion, we wish to point out that the theoretical results presented here do not allow to come out in favour of one of the investigated penalization schemes. In particular, as we consider the problem from the global minimax point of view, we can not decide between Rademacher and bootstrap type penalties.

However, it is now admitted that the global minimax risk is not an ideal bench mark to evaluate the relevance of classification rules, since it may overestimate the risk in some situations. Vapnik and Chervonenkis' (1974) results in the so called *zero-error case* first raised this question. Lugosi (2002) and Devroye and Lugosi (1995) then confirmed these reservations by studying the interpolation case where the best classification error $\inf_{t\in S}\mathbb{P}[t(X)\neq Y]$ of a given class S is nonzero but small. By further analyzing the problem, Mammen and Tsybakov (1999), Tsybakov (2004) and Massart and Nédélec (2005) show that the behavior of the regression function $\eta : x \mapsto \mathbb{P}[Y = 1|X = x]$ around $1/2$ is crucial. They indeed introduce some margin conditions that can be written in the following general way:

$$\exists h > 0,\ l(s,t) \geq h\mathbb{E}_X[|t(X) - s(X)|]^\theta,\ \forall t : \mathcal{X} \to \{0,1\}. \tag{16}$$

Consider moreover the following complexity condition: let s belong to $S = \{\mathbb{I}_C, C \in \mathcal{C}\}$, $\mathcal{C}$ being a VC class with VC dimension $V(\mathcal{C})$. Massart and Nédélec (2005) prove in particular that under the margin condition (16) with $\theta \geq 1$ and $h \geq (V(\mathcal{C})/n)^{1/(2\theta)}$, the Empirical Risk Minimizer $\hat{s}$ over S satisfies

$$\mathbb{E}\left[l(s,\hat{s})\right] \leq \kappa_1\left(\frac{V(\mathcal{C})(1+\log(nh^{2\theta}/V(\mathcal{C})))}{nh}\right)^{\frac{\theta}{2\theta-1}}.$$

Then they discuss the optimality of this upper bound in a minimax sense for the special margin condition $|2\eta(x) - 1| \geq h$ for all x in $\mathcal{X}$ (that leads to (16) with $\theta = 1$). They obtain that if $\mathcal{P}(h)$ denotes the set of the distributions P satisfying the above complexity and margin conditions with $2 \leq V(\mathcal{C}) \leq n$, for any classification rule $\hat{s}$,

$$\sup_{P\in\mathcal{P}(h)} \mathbb{E}\left[l(s,\hat{s})\right] \geq \kappa_2\left(\frac{V(\mathcal{C})}{nh} \wedge \sqrt{\frac{V(\mathcal{C})}{n}}\right).$$

In view of these works, we now aim at developing some model selection procedures which lead to classification rules adapting better to the margin. Following the ideas initially introduced by Koltchinskii and Panchenko (1999), Bartlett et al. (2005) and Bartlett et al. (2004) propose some localized versions of Rademacher averages as tight data-dependent measures of complexity. Recently, it has been proved that these localized Rademacher averages can be used to construct margin-adaptive model selection procedures (see Boucheron et al., 2005, for a brief survey, or Koltchinskii, 2003, for a more complete study). In the same spirit, we could introduce localized versions of our bootstrap penalties. This would entail improving the inequality given in Proposition 2 under propitious conditions, namely margin type conditions. Boucheron et al.'s (2000) concentration inequality seems to be the adequate tool, though it can not be directly applied because of the dependence between the weights involved in the bootstrap processes. Some refined Poissonization techniques may allow us to overcome this difficulty, and this may be the subject of a future work. The results that we may obtain would

however not sort out the main criticism made of all these penalization techniques, that is that they essentially have theoretical interest.
Nevertheless, we are hopeful that the connection made here between Rademacher penalization and the bootstrap approach, which takes advantage of its intuitive qualities, provides new lines of research towards more operational methods for the construction of margin-adaptive classification rules.

Appendix: Proofs of Lemmas 1 and 3

Proof of Lemma 1

Introducing some independent copy $\xi' = (\xi'_1, \ldots, \xi'_n)$ of ξ and denoting by P'_n the corresponding empirical process, we have by Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P-P_n)(f)\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{E}\left[\left(P'_n-P_n\right)(f)\,|\xi\right]\right]\\ &\le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(P'_n-P_n\right)(f)\right].\end{aligned}$$

Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ be a sequence of n i.i.d. Rademacher variables independent of ξ, ξ' and $Z_1, \ldots, Z_n$. Since for any symmetric random variable Z independent of ε_1, the variables $\varepsilon_1 Z$, $\varepsilon_1|Z|$ and Z are identically distributed, we get:

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P-P_n)(f)\right] &\le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(\xi'_i)-f(\xi_i))\right]\\ &\le \frac{2}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i f(\xi_i)\right]\\ &\le \frac{2}{n\mathbb{E}\left[|Z_1|\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{E}\left[\sum_{i=1}^n \varepsilon_i|Z_i|f(\xi_i)\,\middle|\,\varepsilon,\xi\right]\right]\\ &\le \frac{2}{n\mathbb{E}\left[|Z_1|\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i|Z_i|f(\xi_i)\right].\end{aligned}$$

By using the same symmetrization argument as above, we finally obtain:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P-P_n)(f)\right] \le \frac{2}{n\mathbb{E}\left[|Z_1|\right]}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n Z_i f(\xi_i)\right].$$

Proof of Lemma 3

As in Massart (2003), since $\mathcal{A} \mapsto H_2(., \mathcal{A})$ is nondecreasing with respect to the inclusion ordering, by continuity of $a \mapsto \sum_{i=1}^n a_i Z_i$ and separability of $\mathcal{A}$, we only need to consider the case where $\mathcal{A}$ is a finite subset of $[0, 1]^n$.

Let us consider for all j in $\mathbb{N}^*$ a mapping Π_j from $\mathcal{A}$ to $\mathcal{A}$ such that

$$\log |\Pi_j(\mathcal{A})| \leq H_2(2^{-j}\delta, \mathcal{A})$$

and

$$\|a - \Pi_j(a)\|_2 \leq 2^{-j}\delta \quad \text{for all } a \text{ in } \mathcal{A}.$$

Since $\mathcal{A}$ is finite, there exists some integer J such that for all a in $\mathcal{A}$, $\Pi_J(a) = a$. Thus, if Π_0 denotes the mapping which is identically equal to 0, for all $a = (a_1, \ldots, a_n)$ in $\mathcal{A}$,

$$\sum_{i=1}^n a_i Z_i = \sum_{j=0}^{J-1} \left(\sum_{i=1}^n [\Pi_{j+1}(a)]_i Z_i - \sum_{i=1}^n [\Pi_j(a)]_i Z_i \right)$$

and

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}} \sum_{i=1}^n a_i Z_i \right] \leq \sum_{j=0}^{J-1} \mathbb{E}\left[\sup_{a\in\mathcal{A}} \sum_{i=1}^n ([\Pi_{j+1}(a)]_i - [\Pi_j(a)]_i) Z_i \right].$$

Moreover, for all j in $\mathbb{N}$,

$$\sup_{a\in\mathcal{A}} \|\Pi_{j+1}(a) - \Pi_j(a)\|_2^2 \leq \left(\frac{3}{2} 2^{-j}\delta \right)^2,$$

$$\sup_{a\in\mathcal{A}} \sup_{1\leq i\leq n} |[\Pi_{j+1}(a)]_i - [\Pi_j(a)]_i| \leq \|\Pi_{j+1}(a) - \Pi_j(a)\|_2 \wedge 1 \leq 1 \wedge 3(2^{-j-1}\delta),$$

and

$$|\{\Pi_{j+1}(a) - \Pi_j(a),\ a \in \mathcal{A}\}| \leq e^{2H_2(2^{-(j+1)}\delta, \mathcal{A})},$$

by the definition of the Π_j's and the fact that $x \mapsto H_2(x, \mathcal{A})$ is nonincreasing. We can then apply Lemma 2 to the set $\{\Pi_{j+1}(a) - \Pi_j(a),\ a \in \mathcal{A}\}$ and we obtain in the general case:

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}} \sum_{i=1}^n ([\Pi_{j+1}(a)]_i - [\Pi_j(a)]_i) Z_i \right] \leq 3\sqrt{v H_2(2^{-(j+1)}\delta, \mathcal{A})} 2^{-j}\delta + 2c\left(\frac{3}{2} 2^{-j}\delta \wedge 1 \right) H_2\big(2^{-(j+1)}\delta, \mathcal{A}\big),$$

and in the subgaussian case:

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^{n}([\Pi_{j+1}(a)]_i - [\Pi_j(a)]_i)Z_i\right] \leq 3\sqrt{H_2(2^{-(j+1)}\delta, \mathcal{A})}2^{-j}\delta.$$

This concludes the proof of Lemma 3.

Acknowledgments The author wishes to thank Stéphane Boucheron and Pascal Massart for many interesting and helpful discussions.

References

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Math. J.*, *19*, 357–367.

Barron, A. (1985). Logically smooth density estimation. Technical Report 56, Dept. of Statistics, Stanford University.

Barron, A. (1991). Complexity regularization with application to artificial neural networks. *Nonparametric functional estimation and related topics (NATO ASI Ser.)*, *C, 335*, 561–576.

Barron, A., & Cover, T. (1991). Minimum complexity density estimation. *IEEE Trans. Inf. Theory*, *37*, 1034–1054.

Bartlett, P., Boucheron, S., & Lugosi, G. (2002). Model selection and error estimation. *Mach. Learn.*, *48*(1–3), 85–113.

Bartlett, P., Bousquet, O., & Mendelson, S. (2005). Localized Rademacher complexities. *Ann. Stat.*, *33*(4), 1497–1537.

Bartlett, P., Mendelson, S., & Philips, P. (2004). Local complexities for empirical risk minimization. In *Learning theory: 17th annual conference on learning theory, COLT 2004*, volume 3120 of *lecture notes in computer science* (pp. 270–284). Springer-Verlag.

Birgé, L., & Massart, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, *4*(3), 329–375.

Boucheron, S., Bousquet, O., & Lugosi, G. (2005). Theory of classification: A survey of recent advances. To appear in *ESAIM: probability and statistics*.

Boucheron, S., Lugosi, G., & Massart, P. (2000). A sharp concentration inequality with applications. *Random Struct. Algorithms*, *16*(3), 277–292.

Buescher, K., & Kumar, P. (1996). Learning by canonical smooth estimation. I: Simultaneous estimation, II: Learning and choice of model complexity. *IEEE Trans. Autom. Control*, *41*(4), 545–556, 557–569.

Devroye, L. (1982). Bounds for the uniform deviation of empirical measures. *J. Multivariate Anal.*, *12*, 72–79.

Devroye, L., & Lugosi, G. (1995). Lower bounds in pattern recognition and learning. *Pattern Recognition 28*, 1011–1018.

Giné, E., & Zinn, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.*, *12*(4), 929–998.

Giné, E., & Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.*, *18*(2), 851–869.

Grenander, U. (1981). *Abstract inference.* New York: Wiley.

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Comb. Theory*, *A 69*(2), 217–232.

Kay, S. (1998). *Fundamentals of statistical signal processing—Detection theory*. Prentice Hall Signal Processing Series.

Koltchinskii, V. (1981). On the central limit theorem for empirical measures. *Prob. Theory Math. Statist.*, *24*, 71–82.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory*, *47*(5), 1902–1914.

Koltchinskii, V. (2003). Local Rademacher complexities and oracle inequalities in risk minimization. Technical report, University of New Mexico. To appear in *Ann. Stat.*

Koltchinskii, V., & Panchenko, D. (1999). Rademacher processes and bounding the risk of function learning. In Giné, E. (Ed.), *High dimensional probability II. 2nd international conference,* (pp. 443–459). Univ. of Washington, DC, USA, Volume Prog. Probab. 47, Boston, Birkhäuser.

Lebarbier, E. (2002). *Quelques approches pour la détection de ruptures à horizon fini*. Ph. D. thesis, Université Paris XI.

Lo, A. (1987). A large sample study of the Bayesian bootstrap. *Ann. Stat.*, *15*(1), 360–375.

Lozano, F. (2000). Model selection using Rademacher penalization. In *Proceedings of the 2nd ICSC Symp. on Neural Computation (NC2000). Berlin, Germany.* ICSC Academic Press.

Lugosi, G. (2002). Pattern classification and learning theory. In L. Györfi (Ed.), *Principles of nonparametric learning* (pp. 1–56). New York: Springer, Wien.

Lugosi, G., & Nobel, A. (1999). Adaptive model selection using empirical complexities. *Ann. Stat.*, *27*(6), 1830–1864.

Lugosi, G., & Zeger, K. (1996). Concept learning using complexity regularization. *IEEE Trans. Inf. Theory*, *42*(1), 48–54.

Mammen, E., & Tsybakov, A. (1999). Smooth discrimination analysis. *Ann. Stat.*, *27*(6), 1808–1829.

Massart, P. (2003). Concentration inequalities and model selection. Lectures given at the Saint-Flour summer school of probability theory, To appear in *Lect. Notes Math.*

Massart, P., & Nédélec, E. (2005). Risk bounds for statistical learning. To appear in *Ann. Stat.*

Massart, P., & Rio, E. (1998). A uniform Marcinkiewicz-Zygmund strong law of large numbers for empirical processes. In *Asymptotic methods in probability and statistics* (pp. 199–211) (Ottawa, ON, 1997).

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in combinatorics (Lond. Math. Soc. Lect. Notes)* (vol. 141, pp. 148–188).

Pollard, D. (1982). A central limit theorem for empirical processes. *J. Austral. Math. Soc.*, *A 33*(2), 235–248.

Præstgaard, J., & Wellner, J. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, *21*(4), 2053–2086.

Rubin, D. (1981). The Bayesian bootstrap. *Ann. Stat.*, *9*, 130–134.

Sauer, N. (1972). On the density of families of sets. *J. Combinatorial Theory*, *A 13*, 145–147.

Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Stat.*, *32*(1), 135–166.

Van der Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes*. New York: Springer.

Vapnik, V. (1982). *Estimation of dependences based on empirical data*. Springer-Verlag: New York.

Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theor. Probab. Appl.*, *16*, 264–280.

Vapnik, V., & Chervonenkis, A. (1974). *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya (Theory of pattern recognition. Statistical problems of learning)*. Moscow: Nauka.

Weng, C.-S. (1989). On a second-order asymptotic property of the Bayesian bootstrap mean. *Ann. Stat.*, *17*(2), 705–710.