

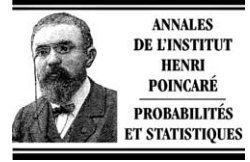


ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Ann. I. H. Poincaré – PR 40 (2004) 685–736



www.elsevier.com/locate/anihpb

Aggregated estimators and empirical complexity for least square regression

Jean-Yves Audibert^{a,b}

^a *Université Paris VI Pierre et Marie Curie, laboratoire de probabilités et modèles aléatoires, 175, rue du Chevaleret, 75013 Paris, France*

^b *CREST, laboratoire de finance et assurance, 15, bd Gabriel Péri, 92245 Malakoff Cedex, France*

Received 10 March 2003; accepted 10 November 2003

Available online 8 June 2004

Abstract

Numerous empirical results have shown that combining regression procedures can be a very efficient method. This work provides PAC bounds for the L^2 generalization error of such methods. The interest of these bounds are twofold.

First, it gives for any aggregating procedure a bound for the expected risk depending on the empirical risk and the empirical complexity measured by the Kullback–Leibler divergence between the aggregating distribution $\hat{\rho}$ and a prior distribution π and by the empirical mean of the variance of the regression functions under the probability $\hat{\rho}$.

Secondly, by structural risk minimization, we derive an aggregating procedure which takes advantage of the unknown properties of the best mixture \tilde{f} : when the best convex combination \tilde{f} of d regression functions belongs to the d initial functions (i.e. when combining does not make the bias decrease), the convergence rate is of order $(\log d)/N$. In the worst case, our combining procedure achieves a convergence rate of order $\sqrt{(\log d)/N}$ which is known to be optimal in a uniform sense when $d > \sqrt{N}$ (see [A. Nemirovski, in: Probability Summer School, Saint Flour, 1998; Y. Yang, Aggregating regression procedures for a better performance, 2001]).

As in AdaBoost, our aggregating distribution tends to favor functions which disagree with the mixture on mispredicted points. Our algorithm is tested on artificial classification data (which have been also used for testing other boosting methods, such as AdaBoost).

© 2004 Elsevier SAS. All rights reserved.

Résumé

De nombreuses études empiriques ont montré l'efficacité des méthodes consistant à combiner différentes procédures de régression. Ce travail fournit de nouvelles bornes "PAC" (probablement approximativement correct) pour l'erreur de généralisation L_2 de ces méthodes. Ces bornes présentent un double intérêt.

Tout d'abord, elles donnent une borne sur le risque de n'importe quelle procédure d'agrégation en termes du risque empirique et d'une mesure empirique de la complexité basée sur la divergence de Kullback–Leibler entre la probabilité d'agrégation $\hat{\rho}$ et la probabilité a priori et sur la moyenne empirique de la variance des fonctions de régression sous la probabilité $\hat{\rho}$.

Deuxièmement, par minimisation du "risque structurel", nous dérivons une procédure d'agrégation qui s'adapte aux propriétés pourtant inconnues du meilleur mélange \tilde{f} : quand la meilleure combinaison convexe \tilde{f} de d fonctions est une de ces d fonctions (c'est-à-dire quand combiner les fonctions ne permet pas de réduire le biais), le taux de convergence est

E-mail address: jyaudibe@ccr.jussieu.fr (J.-Y. Audibert).

d'ordre $(\log d)/N$. Dans le pire des cas, notre procédure d'agrégation a un taux de convergence d'ordre $\sqrt{(\log d)/N}$ qui est le taux minimax optimal quand $d > \sqrt{N}$ (cf. [A. Nemirovski, in : Summer School, Saint Flour, 1998 ; Y. Yang, Aggregating regression procedures for a better performance, 2001]).

Comme la méthode AdaBoost, le mélange obtenu par notre algorithme pondère les fonctions qui ne sont pas en accord avec le mélange sur les points mal classés de l'ensemble d'apprentissage. L'algorithme est testé sur des problèmes de classification artificiels (déjà utilisés pour tester des procédures de boosting telles qu'AdaBoost).

© 2004 Elsevier SAS. All rights reserved.

MSC: primary 62G08; secondary 62J02, 94A17, 62H30

Keywords: Nonparametric regression; Deviation inequalities; Adaptive estimator; Oracle inequalities; Boosting

1. Introduction

Boosting algorithms (AdaBoost introduced by Freund and Schapire in [4], Bagging and Arcing introduced by Breiman in [1,2]) have been successful in practical classification applications. With support vector machines, boosting is known to be one of the best off-the-shelf classification procedure. As a consequence, numerous researchers have studied the reasons of their efficiency and have looked for means to extend their application domain to regression problems.

Friedman, Hastie and Tibshirani have proved [5] that AdaBoost is a stage-wise estimation procedure for fitting an additive logistic regression model. From this idea, Friedman derive a “gradient boosting machine” to estimate a function for some specified loss criteria.

Rätsch et al. [10] have shown that boosting is similar to an iterative strategy which maximizes the minimum margin of the aggregated classifier using an exponential barrier. They also use their view to obtain a boosting technique for regression.

In [13], Yang has studied minimax properties of aggregating regression procedures. In particular, he has proved that when the number d of aggregated procedures is less than \sqrt{N} (where N is the size of the training set), the order of the convergence rate of the best mixture (in the minimax sense) is the same as the one of the best linear combination (i.e. d/N). When d is greater than \sqrt{N} , the convergence rate of the best convex combination attains $\sqrt{(\log d)/N}$ (see also [9]).

In this paper, we will obtain new bounds for any aggregating procedure (Section 4) and derive from these bounds a procedure which achieves the optimal minimax convergence rate. Before proving these bounds, we will review Catoni results [3] on randomization procedures (Section 3). The estimators obtained by minimization of the bound are tested on classification using common artificial data: Twonorm, Threenorm and Ringnorm (Section 5).

2. Framework

We assume that we observe an i.i.d. sample $Z_1^N \triangleq (X_i, Y_i)_{i=1}^N$ of random variables distributed according to a product probability measure $\mathbb{P}^{\otimes N}$, where \mathbb{P} is a probability distribution on $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}}) \triangleq (\mathcal{X} \otimes \mathcal{Y}, \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Y}})$, $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ is a measurable space, $\mathcal{Y} = \mathbb{R}$ and $\mathcal{B}_{\mathcal{Y}}$ is the Borel sigma algebra. Let $\mathbb{P}(dY|X)$ denote a regular version of the conditional probabilities (which we will use in the following without further mention).

We assume that we have no prior information about the distribution \mathbb{P} of (X, Y) , and that we have to guess it entirely from the training sample. We have to work with a prescribed set of regression functions since it is well known that there is generally no estimator $\hat{f} : \mathcal{Z}^N \rightarrow \mathcal{F}(\mathcal{X}, \mathcal{Y})$ such that

$$\lim_{N \rightarrow +\infty} \sup_{\mathbb{P} \in \mathcal{M}_+^1(\mathcal{Z})} \{ \mathbb{E}_{\mathbb{P}^{\otimes(N+1)}} L[Y_{N+1}, \hat{f}(Z_1^N)(X_{N+1})] - \inf_{f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} \mathbb{E}_{\mathbb{P}} L[Y, f(X)] \} = 0,$$

where $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ denotes the set of all the measurable functions from \mathcal{X} to \mathcal{Y} and L is a loss function. However, replacing $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ by the set of mixtures $\tilde{\mathcal{R}}$ of a set of functions \mathcal{R} in the previous equality makes the problem feasible (provided the model \mathcal{R} is not too big) with a speed of convergence depending on the capacity (or complexity) of \mathcal{R} . So we are interested in a particular non-parametric regression problem. For convenience of notation, we will index the functions in the model by the parameter θ :

$$\mathcal{R} \triangleq \{f_\theta \in \mathcal{F}(\mathcal{X}, \mathcal{Y}); \theta \in \Theta\}.$$

Note that the set \mathcal{R} (or equivalently the parameter set Θ) is not necessarily finite. Let $\pi(d\theta)$ denote a prior distribution on the measurable space (Θ, \mathcal{T}) , where \mathcal{T} is a σ -field on the parameter space Θ . In practice, the probability distribution π will be chosen according to our preferences (and to our prior knowledge had we some). For instance, if the model \mathcal{R} is the set of decision trees of depth lower than a certain limit and if we do not have any prior knowledge, we would like to favour small trees with respect to big ones since they are simpler and therefore more easily interpretable. To favour these trees, it suffices to give them a bigger π -probability. On the contrary, if a subset S of \mathcal{R} has a π -probability equal to one, then the functions in the π -negligible set $\mathcal{R} \setminus S$ are eliminated from the model.

We assume that the map $(\theta, x) \mapsto f_\theta(x)$ is $(\mathcal{B}_{\mathcal{X}} \otimes \mathcal{T})$ -measurable. The set of mixtures of the set \mathcal{R} is written as

$$\tilde{\mathcal{R}} \triangleq \{\mathbb{E}_{\rho(d\theta)} f_\theta; \rho \in \mathcal{M}_+^1(\Theta)\}.$$

The best possible guess is defined as the one minimizing the expected risk

$$R(\hat{f}) \triangleq \mathbb{E}_{\mathbb{P}} L(Y, \hat{f}(X)),$$

where L is the square loss: $L(Y, Y') = (Y - Y')^2$. The mean square loss has the distinguished property of being minimized by the conditional expectation of Y given X . More precisely, it decomposes into

$$R(\hat{f}) = \mathbb{E}_{\mathbb{P}} \{[Y - \mathbb{E}_{\mathbb{P}}(Y/X)]^2\} + \mathbb{E}_{\mathbb{P}} \{[\mathbb{E}_{\mathbb{P}}(Y/X) - \hat{f}(X)]^2\}.$$

Therefore, minimizing the mean square loss is equivalent to minimizing the quadratic distance to the conditional expectation.

Since the expected risk is not observable, we will have to use the empirical risk

$$r(\hat{f}) \triangleq \frac{1}{N} \sum_{i=1}^N L(Y_i, \hat{f}(X_i)) = \mathbb{E}_{\bar{\mathbb{P}}} L(Y, \hat{f}(X)),$$

where $\bar{\mathbb{P}}$ denotes the empirical distribution

$$\bar{\mathbb{P}} \triangleq \frac{1}{N} \sum_{i=1}^N \delta_{(X_i, Y_i)}.$$

Let $\Theta_1, \dots, \Theta_M$ be subsets of Θ such that their union is Θ . Consider a regression procedure which estimate the best θ among a subset of Θ . Using this procedure, we get $\hat{\theta}_1 \in \Theta_1, \dots, \hat{\theta}_M \in \Theta_M$.

- *Deterministic model selection* consists in choosing one of the $\hat{\theta}_i$ to estimate $\mathbb{E}_{\mathbb{P}}(Y/X)$.
- In *stochastic model selection* (or *randomized estimation*), the choice of $\hat{\theta}_i$ is randomized. This two-steps procedure (estimating the best θ in each sub-model Θ_i and choosing randomly the sub-model) can be seen as a one-step procedure if we allow \hat{f} to be drawn from \mathcal{R} according to some posterior distribution $\rho(d\theta)$ on the parameter set (Θ, \mathcal{T}) (see [3,8]).
- In *model averaging* (or *aggregated estimation*), the idea is to use a weighting average of the $f_{\hat{\theta}_i}$, in other words to combine the different estimators. This could also be done in a one-step procedure searching for the posterior distribution ρ on (Θ, \mathcal{T}) such that $\hat{f} = \mathbb{E}_{\rho(d\theta)} f_\theta$ is close to $\mathbb{E}_{\mathbb{P}}(Y/X)$.

In this paper, we give results concerning these last two estimation problems. Our assumptions are the two following ones. First the conditional expectation $\mathbb{E}_{\mathbb{P}}(Y/X)$ and the regression function in the models are relatively bounded in L^∞ -norm, i.e. for any f, g in $\mathcal{R} \cup \{E(Y/X = \cdot)\}$, for any $x \in \mathcal{X}$,

$$|f(x) - g(x)| \leq B. \tag{2.1}$$

Secondly, we assume that the noise has a uniform exponential moment conditionally to the explanatory variable, i.e. there exists $\alpha > 0, M > 0$ such that for any $x \in \mathcal{X}$,

$$\mathbb{E}_{\mathbb{P}(dY)} \exp(\alpha|Y - f^*(X)|/X = x) \leq M, \tag{2.2}$$

where $f^* \triangleq \mathbb{E}_{\mathbb{P}}(Y/X = \cdot)$ is the regression function associated with the distribution \mathbb{P} . Note that this second assumption is sufficiently weak to deal with the case in which the output is equal to a function of the input plus a gaussian noise.

Let \tilde{f} denote the best mixture (for the square loss) of the regression functions in the model \mathcal{R} :

$$\tilde{f} \triangleq \underset{f \in \mathcal{R}}{\operatorname{argmin}} R(f). \tag{2.3}$$

Finally, introduce a mixture distribution $\tilde{\rho} \in \mathcal{M}_+^1(\Theta)$ defined as $\mathbb{E}_{\tilde{\rho}(d\theta)} f_\theta = \tilde{f}$ (the probability distribution $\tilde{\rho}$ is not necessarily unique).

3. Randomization

3.1. PAC-Bayesian expected risk bound

The following theorems bound the expected risk of a randomized procedure in terms of the empirical risk and a term of empirical complexity relying on the Kullback–Leibler divergence between the randomizing distribution ρ and the prior distribution π . Introduce the functions

$$G(\lambda) \triangleq \frac{8M}{(\alpha B - 2\lambda)^2 e^2} + \frac{e^{2\lambda} - 1 - 2\lambda}{\lambda^2} \quad \text{and} \quad H(\lambda) \triangleq \frac{1}{1 - \lambda G(\lambda)}.$$

Theorem 3.1. *For any $\varepsilon > 0$ and $0 < \lambda < \frac{\alpha B}{2}$ such that $\lambda G(\lambda) < 1$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$, for any randomizing procedure $\hat{\rho}: \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$, we have*

$$\mathbb{E}_{\hat{\rho}(d\theta)} R(f_\theta) - R(\tilde{f}) \leq H(\lambda) \left(\mathbb{E}_{\hat{\rho}(d\theta)} r(f_\theta) - r(\tilde{f}) + \frac{B^2}{\lambda N} [K(\hat{\rho}, \pi) + \log(\varepsilon^{-1})] \right). \tag{3.1}$$

Proof. See Section 7.1. \square

To use this bound, one has to choose arbitrarily the parameter λ . To avoid this choice, one can use an union bound.

Theorem 3.2. *Introduce countable families $(\lambda_i)_{i \in I}$ and $(\eta_i)_{i \in I}$ such that $0 < \lambda_i < \alpha B/2$, $\lambda_i G(\lambda_i) < 1$, $\eta_i > 0$ and $\sum_{i \in I} \eta_i = 1$. For any $\varepsilon > 0$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$, for any randomizing procedure $\hat{\rho}: \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$, for any $i \in I$, we have*

$$\mathbb{E}_{\hat{\rho}(d\theta)} R(f_\theta) - R(\tilde{f}) \leq H(\lambda_i) \left(\mathbb{E}_{\hat{\rho}(d\theta)} r(f_\theta) - r(\tilde{f}) + \frac{B^2}{N \lambda_i} \{K(\hat{\rho}, \pi) + \log[(\eta_i \varepsilon)^{-1}]\} \right). \tag{3.2}$$

Proof. Introduce the event

$$A_i \triangleq \left\{ \frac{\mathbb{E}_{\hat{\rho}(d\theta)} R(f_\theta) - R(\tilde{f})}{H(\lambda_i)} > \mathbb{E}_{\hat{\rho}(d\theta)} r(f_\theta) - r(\tilde{f}) + \frac{B^2}{N\lambda_i} \{K(\hat{\rho}, \pi) + \log[(\eta_i \varepsilon)^{-1}]\} \right\}.$$

From Theorem 3.1, for any $i \in I$, we have $\mathbb{P}^{\otimes N}(A_i) < \eta_i \varepsilon$. Hence we have

$$\mathbb{P}^{\otimes N} \left(\bigcup_{i \in I} A_i \right) \leq \sum_{i \in I} \mathbb{P}^{\otimes N}(A_i) < \sum_{i \in I} \eta_i \varepsilon = \varepsilon. \quad \square$$

The problem is then to choose appropriately the parameter families $(\lambda_i)_{i \in I}$ and $(\eta_i)_{i \in I}$.

3.2. Optimal randomizing procedure

In this section we use Theorem 3.2 to define a randomizing procedure. The bounds in the previous theorems cannot be computed from the data only. However they can be upper bounded by replacing the empirical risk of the unknown best mixture $r(\tilde{f})$ by the infimum over the set $\tilde{\mathcal{R}}$ of the empirical risk $\inf_{\tilde{\mathcal{R}}} r$.

Introduce

$$\left\{ \begin{array}{l} \mathcal{Q}(\rho, \lambda, \eta) \triangleq \frac{\mathbb{E}_{\hat{\rho}(d\theta)} r(f_\theta) - \inf_{\tilde{\mathcal{R}}} r}{1 - \lambda G(\lambda)} + \frac{B^2}{N} \frac{K(\hat{\rho}, \pi) + \log[(\eta \varepsilon)^{-1}]}{\lambda[1 - \lambda G(\lambda)]}, \\ \mathcal{Q}(\rho, (\lambda_i)_{i \in I}, (\eta_i)_{i \in I}) \triangleq \inf_{i \in I} \mathcal{Q}(\rho, \lambda_i, \eta_i), \\ \mathcal{Q}(\rho) \triangleq \inf_{\substack{(\lambda_i)_{i \in I} \in \mathcal{P}_\lambda \\ (\eta_i)_{i \in I} \in \mathcal{P}_\eta}} \mathcal{Q}(\rho, (\lambda_i)_{i \in I}, (\eta_i)_{i \in I}), \end{array} \right.$$

where \mathcal{P}_λ and \mathcal{P}_η are respectively the set of parameter families $(\lambda_i)_{i \in I}$ and $(\eta_i)_{i \in I}$ such that $0 < \lambda_i < \frac{\alpha B}{2}$, $\lambda_i G(\lambda_i) < 1$, $\eta_i > 0$ and $\sum_{i \in I} \eta_i = 1$. Then the quantities $\mathcal{Q}(\rho, \lambda, 1)$ and $\mathcal{Q}(\rho, \lambda_i, \eta_i)$ are respectively slightly weakened version of the RHS of inequalities (3.1) and (3.2).

The quantity $\mathcal{Q}(\rho)$ can also be written as

$$\mathcal{Q}(\rho) = \inf_{0 < \lambda < \alpha B/2 \text{ such that } \lambda G(\lambda) < 1} \mathcal{Q}(\rho, \lambda, 1).$$

Let us define the optimal posterior distribution $\hat{\rho}_{\text{opt}}$ as

$$\hat{\rho}_{\text{opt}} = \operatorname{argmin}_{\rho \in \mathcal{M}_+^1(\Theta)} \mathcal{Q}(\rho).$$

For any $0 < \varepsilon < 1$, one may prove the existence of the “argmin” and that $\hat{\rho}_{\text{opt}}$ is a Gibbs distribution which can be written as

$$\hat{\rho}_{\text{opt}} = \frac{e^{-\frac{N\lambda_{\text{opt}}}{B^2} r(f)}}{\mathbb{E}_{\pi(d\theta)} e^{-\frac{N\lambda_{\text{opt}}}{B^2} r(f_\theta)}} \cdot \pi,$$

for an appropriate parameter $0 < \lambda_{\text{opt}} < \alpha B/2$ satisfying $\lambda_{\text{opt}} G(\lambda_{\text{opt}}) < 1$. Then the inverse temperature parameter of the Gibbs distribution is $\beta \triangleq N\lambda_{\text{opt}}/B^2$.

We would like to choose the parameter families such that the infimum $\inf_{\rho} \mathcal{Q}(\rho, (\lambda_i)_{i \in I}, (\eta_i)_{i \in I})$ is not “too far” from the optimal quantity $\mathcal{Q}(\hat{\rho}_{\text{opt}})$. The bound in Theorem 3.2 is appropriate when its order is $1/\sqrt{N}$. Therefore relevant values of λ are greater than $1/\sqrt{N}$. Let us define $0 < \Lambda < \alpha B/2$ such that $\Lambda G(\Lambda) = 1$. Consider the

family $(\lambda_i)_{i=1,\dots,p}$, where $\lambda_i \triangleq \Lambda/2^i$ and p is such that $\Lambda/2^{p+1} < 1/\sqrt{N} \leq \Lambda/2^p$. When the parameter λ_{opt} belongs to $[1/\sqrt{N}; \Lambda[$ (which is the case we are interested in), for any $\rho \in \mathcal{M}_+^1(\Theta)$, we have

$$\inf_{i=1,\dots,p} \mathcal{Q}(\rho, \lambda_i, 1) \leq 2\mathcal{Q}(\rho, \lambda_{\text{opt}}, 1).$$

So we just lose in the worst case a factor 2. It remains to choose the parameters η_i such that for any $\rho \in \mathcal{M}_+^1(\Theta)$, the quantity $\mathcal{Q}(\rho, \lambda_i, \eta_i)$ is not “too far” from the quantity $\mathcal{Q}(\rho, \lambda_i, 1)$. By taking $\eta_i = 1/p, i = 1, \dots, p$, we lose an additive $\log \log N$ factor in front of the Kullback–Leibler divergence $K(\rho, \pi)$ which, in general, would be for the optimal distribution at least of the same order as the Kullback–Leibler divergence (in practice, $\log \log N$ never exceeds 3).

Since the minimum over $\mathcal{M}_+^1(\Theta)$ of the quantity $\mathcal{Q}(\rho, \lambda, 1)$ (achieved for the probability distribution $\rho \propto e^{-\frac{N\lambda}{B^2}r(f)} \cdot \pi$) is

$$\frac{B^2}{N\lambda[1 - \lambda G(\lambda)]} \log\left[\left(\varepsilon \mathbb{E}_{\pi}(d\theta) e^{-\frac{N\lambda}{B^2}[r(f_\theta) - \inf_{\mathcal{R}} r]}\right)^{-1}\right],$$

let us introduce for any $i = 1, \dots, p$,

$$Q_i \triangleq \frac{1}{\lambda_i[1 - \lambda_i G(\lambda_i)]} \log\left(\frac{p}{\varepsilon \mathbb{E}_{\pi}(d\theta) e^{-\frac{N\lambda_i}{B^2}[r(f_\theta) - \inf_{\mathcal{R}} r]}}\right),$$

where $\lambda_i = \Lambda/2^i$. Finally, we obtain the following randomizing procedure

1. Compute

$$i_{\text{opt}} \triangleq \underset{i=1,\dots,p}{\operatorname{argmin}} Q_i.$$

2. Randomize using the probability distribution

$$\frac{e^{-\frac{N\Lambda}{B^2 2^{i_{\text{opt}}}}r(f)}}{\mathbb{E}_{\pi}(d\theta) e^{-\frac{N\Lambda}{B^2 2^{i_{\text{opt}}}}r(f_\theta)}} \cdot \pi.$$

Remark 3.1. Note that since our optimal randomizing procedure comes from a deviation inequality, the inverse temperature parameter β depends on the probability ε . Indeed, to get a higher confidence level, we need to have a bigger λ and therefore to take a bigger β (i.e. to be more selective). However in practice ε has little influence on the temperature.

Remark 3.2. Our optimal randomizing distribution is a Gibbs distribution. We find it in a minimax context. One may notice that the randomizing distribution minimizing the Bayesian risk in a gaussian noise context is also a Gibbs distribution. More precisely, consider that the output is given by

$$Y = f_\theta(X) + \eta,$$

where the random variable η is a centered gaussian with variance σ^2 independent of the input X . The Bayesian risk is

$$\begin{aligned} R_{\text{Bay}}(\hat{f}) &\triangleq \mathbb{E}_{\pi(d\theta/Z_1^N)} \mathbb{E}_{\mathbb{P}_\theta(dZ_{N+1})} [(Y_{N+1} - \hat{f}(X_{N+1}))^2] \\ &= \sigma^2 + \mathbb{E}_{\pi(d\theta/Z_1^N)} \mathbb{E}_{\mathbb{P}(dX_{N+1})} [(f_\theta(X_{N+1}) - \hat{f}(X_{N+1}))^2] \\ &= \sigma^2 + \mathbb{E}_{\mathbb{P}(dX_{N+1})} \mathbb{E}_{\pi(d\theta/Z_1^N)} [(f_\theta(X_{N+1}) - \hat{f}(X_{N+1}))^2]. \end{aligned}$$

Hence the optimal estimator is $\hat{f} = \mathbb{E}_{\pi(d\theta/Z_1^N)} f_\theta$. It is associated with the posterior distribution

$$\hat{\rho}(d\theta) = \pi(d\theta/Z_1^N) = \frac{e^{-\frac{N}{2\sigma^2}r(f_\theta)}}{\mathbb{E}_\pi e^{-\frac{N}{2\sigma^2}r(f)}} \cdot \pi(d\theta),$$

which is a Gibbs distribution with inverse temperature parameter $N/(2\sigma^2)$.

4. Aggregated estimators

4.1. PAC-Bayesian expected risk bound

In the least square regression framework, there exists a simple relation between the risk of an aggregated estimator and the one of the associated randomized estimator which is

$$R(\mathbb{E}_{\rho(d\theta)} f_\theta) = \mathbb{E}_{\rho(d\theta)} R(f_\theta) - \mathbb{E}_\mathbb{P} \text{Var}_{\rho(d\theta)} f_\theta(X). \tag{4.1}$$

This equality shows that aggregated regression procedures are more efficient than randomized ones and that the difference is measured by $\mathbb{E}_\mathbb{P} \text{Var}_{\rho(d\theta)} f_\theta(X)$. The first term of the RHS has already been bounded (see Theorem 3.1). So, to bound the expected risk of the aggregated estimator, it remains to bound the deviations of the variance term and this is done with similar techniques to those used for randomized estimators.

We obtain the following theorems which bound the expected risk of any aggregated estimator in terms of

- the empirical risk,
- the empirical complexity measured by the Kullback–Leibler divergence between the aggregating distribution $\hat{\rho}$ and the prior distribution π and by the empirical mean of the variance of the regression functions under the posterior distribution.

We still denote

$$G(\lambda) \triangleq \frac{8M}{(\alpha B - 2\lambda)^2 e^2} + \frac{e^{2\lambda} - 1 - 2\lambda}{\lambda^2} \quad \text{and} \quad H(\lambda) \triangleq \frac{1}{1 - \lambda G(\lambda)},$$

and we define

$$g(\beta) \triangleq \frac{e^\beta - 1 - \beta}{\beta^2} \quad \text{and} \quad h(\beta) \triangleq \frac{1}{1 + \beta g(\beta)}.$$

Theorem 4.1. For any $\varepsilon > 0$, $\beta > 0$ and $0 < \lambda < \alpha B/2$ such that $\lambda G(\lambda) < 1$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, for any aggregating procedure $\hat{\rho} : \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) &\leq H(\lambda) \left(\mathbb{E}_{\hat{\rho}(d\theta)} r(f_\theta) - r(\tilde{f}) + \frac{B^2}{N\lambda} [K(\hat{\rho}, \pi) + \log(\varepsilon^{-1})] \right) \\ &\quad + h(\beta) \left(-\bar{V} + \frac{B^2}{2N\beta} [2K(\hat{\rho}, \pi) + \log(\varepsilon^{-1})] \right) \\ &= H(\lambda) [r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - r(\tilde{f})] + [H(\lambda) - h(\beta)] \bar{V} \\ &\quad + \frac{B^2 H(\lambda)}{N\lambda} [K(\hat{\rho}, \pi) + \log(\varepsilon^{-1})] + \frac{B^2 h(\beta)}{2N\beta} [2K(\hat{\rho}, \pi) + \log(\varepsilon^{-1})], \end{aligned} \tag{4.2}$$

where $\bar{V} \triangleq \mathbb{E}_\mathbb{P} \text{Var}_{\hat{\rho}(d\theta)} f_\theta$.

Proof. See Section 7.2. \square

Using an union bound, we get

Theorem 4.2. *Introduce countable families $(\lambda_i)_{i \in I}$, $(\eta_i)_{i \in I}$, $(\beta_j)_{j \in J}$ and $(\zeta_j)_{j \in J}$ such that $0 < \lambda_i < \alpha B/2$, $\lambda_i G(\lambda_i) < 1$, $\eta_i > 0$, $\sum_{i \in I} \eta_i = 1$, $\beta_j > 0$, $\zeta_j > 0$ and $\sum_{j \in J} \zeta_j = 1$. For any $\varepsilon > 0$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, for any aggregating procedure $\hat{\rho}: \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$, for any $i \in I$ and for any $j \in J$, we have*

$$\begin{aligned}
 R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) &\leq H(\lambda_i) [r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - r(\tilde{f})] + [H(\lambda_i) - h(\beta_j)] \bar{V} \\
 &\quad + \frac{B^2 H(\lambda_i)}{N \lambda_i} \{K(\hat{\rho}, \pi) + \log[(\eta_i \varepsilon)^{-1}]\} \\
 &\quad + \frac{B^2 h(\beta_j)}{2N \beta_j} \{2K(\hat{\rho}, \pi) + \log[(\zeta_j \varepsilon)^{-1}]\}.
 \end{aligned}
 \tag{4.3}$$

Proof. In the proof of Theorem 4.1 (see Section 7.2), we have obtained that with $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,

$$-\mathbb{E}_{\mathbb{P}} \text{Var}_{\rho(d\theta)} f_\theta \leq h(\beta) \left(-\mathbb{E}_{\mathbb{P}} \text{Var}_{\rho(d\theta)} f_\theta + \frac{B^2}{2N\beta} [2K(\rho, \pi) + \log(\varepsilon^{-1})] \right).$$

Instead of using an union bound directly on inequality (4.2), we use it on this inequation. We get that with $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$ and for any $j \in J$,

$$-\mathbb{E}_{\mathbb{P}} \text{Var}_{\rho(d\theta)} f_\theta \leq h(\beta_j) \left(-\mathbb{E}_{\mathbb{P}} \text{Var}_{\rho(d\theta)} f_\theta + \frac{B^2}{2N\beta_j} \{2K(\rho, \pi) + \log[(\zeta_j \varepsilon)^{-1}]\} \right),$$

where $(\beta_j)_{j \in J}$ and $(\zeta_j)_{j \in J}$ are parameter families such that $\beta_j > 0$, $\zeta_j > 0$ and $\sum_{j \in J} \zeta_j = 1$. It remains to add this inequation to inequality (3.2) to get the result. \square

Now let us introduce

$$\left\{ \begin{aligned}
 \mathbb{B}(\rho, \lambda, \eta, \beta, \zeta) &\triangleq H(\lambda) \left(\mathbb{E}_{\rho(d\theta)} r(f_\theta) - r(\tilde{f}) + \frac{B^2}{N\lambda} \{K(\rho, \pi) + \log[(\eta\varepsilon)^{-1}]\} \right) \\
 &\quad + h(\beta) \left(-\bar{V} + \frac{B^2}{2N\beta} \{2K(\rho, \pi) + \log[(\zeta\varepsilon)^{-1}]\} \right) \\
 \mathbb{B}(\rho, (\lambda_i)_{i \in I}, (\eta_i)_{i \in I}, (\beta_j)_{j \in J}, (\zeta_j)_{j \in J}) &\triangleq \kappa B^2 \wedge \inf_{\substack{i \in I \\ j \in J}} \mathbb{B}(\rho, \lambda_i, \eta_i, \beta_j, \zeta_j),
 \end{aligned} \right.
 \tag{4.4}$$

where $\kappa \triangleq 1 + \frac{4M}{e^2(\alpha B)^2}$.

By bounding the expected risk using assumptions (2.1) and (2.2), and from the previous theorem, we obtain

Corollary 4.3. *For any $\varepsilon > 0$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, for any aggregating procedure $\hat{\rho}: \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$, we have*

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) \leq \mathbb{B}(\rho, (\lambda_i)_{i \in I}, (\eta_i)_{i \in I}, (\beta_j)_{j \in J}, (\zeta_j)_{j \in J}).$$

Proof. From Theorem 4.1, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, for any aggregating procedure $\hat{\rho}: \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$, we have

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) \leq \inf_{\substack{i \in I \\ j \in J}} \mathbb{B}(\rho, \lambda_i, \eta_i, \beta_j, \zeta_j).
 \tag{4.5}$$

Since the noise has a conditional uniform exponential moment (assumption (2.2)), the expected risk is bounded. Specifically, we can write

$$\begin{aligned}
 R(\mathbb{E}_\rho f) &= \mathbb{E}_\mathbb{P}(Y - E(Y/X))^2 + \mathbb{E}_\mathbb{P}(E(Y/X) - \mathbb{E}_\rho f)^2 \\
 &\leq \mathbb{E}_\mathbb{P}(e^{\alpha|Y-E(Y/X)|} \sup_{u \in \mathbb{R}_+} \{u^2 e^{-\alpha u}\}) + B^2 \\
 &\leq \left(\frac{2}{\alpha e}\right)^2 M + B^2 \\
 &\leq \kappa B^2,
 \end{aligned}
 \tag{4.6}$$

where $\kappa \triangleq \frac{4M}{e^2(\alpha B)^2} + 1$. Since the quadratic risk $R(\tilde{f})$ is positive, for any probability distribution ρ , we have

$$\mathbb{E}_{\rho(d\theta)} R(\theta) - R(\tilde{f}) \leq \kappa B^2.
 \tag{4.7}$$

The result follows from equalities (4.5) and (4.7). \square

This corollary is the keystone of this work since

- by appropriately choosing the parameter families, one can deduce a parameter-free theorem which has the optimal minimax convergence rate except for a logarithmic factor (see Section 4.2.1),
- there exists an efficient procedure calculating one of the probability distributions minimizing the bound $\mathbb{B}(\rho, (\lambda_i)_{i \in I}, (\eta_i)_{i \in I}, (\beta_j)_{j \in J}, (\zeta_j)_{j \in J})$, when the sets I and J are finite (see Section 4.2.2).

4.2. Optimal aggregating procedure

4.2.1. Comparison with minimax bounds

In this section, we derive from Corollary 4.3 an aggregating procedure which is optimal in a minimax sense according to lower bounds established by Juditsky and Nemirovski [6] and by Yang [13]. We still denote $\tilde{\rho}$ a posterior distribution such that $R(\mathbb{E}_{\tilde{\rho}(d\theta)} f_\theta) = \min_{\tilde{\mathcal{R}}} R$.

Lemma 4.4. *For a well chosen finite parameter families, we have*

$$\mathbb{B}(\tilde{\rho}, (\lambda_i)_{i \in I}, (\eta_i)_{i \in I}, (\beta_j)_{j \in J}, (\zeta_j)_{j \in J}) \leq \gamma(\varepsilon),$$

where

$$\begin{cases}
 \gamma(\varepsilon) \triangleq 2\sqrt{\mathcal{C}_1 \bar{V}(\tilde{\rho})} + 6\sqrt{\mathcal{C}_2 \bar{V}(\tilde{\rho})} + 2\mathcal{C}_1 + 2\mathcal{C}_2, \\
 \bar{V}(\tilde{\rho}) \triangleq \mathbb{E}_\mathbb{P} \text{Var}_{\tilde{\rho}(d\theta)} f_\theta, \\
 \mathcal{C}_1 \triangleq \mathcal{C}_1(\varepsilon) \triangleq \frac{B^2}{N} \frac{K(\tilde{\rho}, \pi) + \log(L_1 \varepsilon^{-1})}{\kappa_1}, \\
 \mathcal{C}_2 \triangleq \mathcal{C}_2(\varepsilon) \triangleq \frac{B^2}{8N} \frac{2K(\tilde{\rho}, \pi) + \log(L_2 \varepsilon^{-1})}{\kappa_2},
 \end{cases}$$

and κ_1 and κ_2 , by definition, respectively satisfy $2\kappa_1 G(\kappa_1) = 1$ and $\kappa_2 g(\kappa_2) = 1$ and finally

$$\begin{cases}
 L_1 \triangleq \frac{\log(\frac{4\kappa_1 N}{\log(\varepsilon^{-1})})}{2 \log 2} \vee 2, \\
 L_2 \triangleq \frac{\log(\frac{8\kappa_2 N}{\log(\varepsilon^{-1})})}{2 \log 2} \vee 2.
 \end{cases}$$

The proof and the parameter families are given in Section 7.3. From this lemma and from Corollary 4.3, by using the same parameter families, we get

Theorem 4.5. *For any $\varepsilon > 0$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, any aggregating procedure $\hat{\rho}$ minimizing*

$$\mathbb{B}(\rho, (\lambda_i)_{i=0,\dots,p}, (\eta_i)_{i=0,\dots,p}, (\beta_j)_{j=0,\dots,q}, (\zeta_j)_{j=0,\dots,q})$$

satisfies

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) \leq \gamma(\varepsilon).$$

Proof. See Section 7.4. \square

In the worst case, the bound has the same order (when ε is fixed) as

$$\sqrt{\tilde{C}\tilde{V}} \vee \tilde{C},$$

where $\tilde{C} \triangleq \frac{K(\tilde{\rho}, \pi)}{N} B^2$ and $\tilde{V} \triangleq \sup_{x \in \mathcal{X}} \text{Var}_{\tilde{\rho}(d\theta)} f_\theta(x)$ (we neglect the $\log \log N$ term).

When the best mixture \tilde{f} belongs to the initial model \mathcal{R} , the variance term vanishes and the order of the bounds is given by \tilde{C} . A particular case of interest is when the parameter set Θ is finite: $\Theta = \{1, \dots, d\}$. Taking arbitrarily $\pi = \frac{1}{d} \sum_{i=1}^d \delta_i$ (uniform measure on Θ), one can check easily that for any $\rho \in \mathcal{M}_+^1(\Theta)$, we have

$$K(\rho, \pi) = \log d - H_s(\rho) \leq \log d,$$

where $H_s(\rho)$ denotes the Shannon entropy of ρ ($H_s(\rho) \triangleq - \sum_{i=1}^d \rho_i \log \rho_i$). In this case, when the best convex combination \tilde{f} belongs to the model \mathcal{R} ($\tilde{V} = 0$), the convergence rate of our estimator will be $\log d/N$, whereas when \tilde{f} is not too close to the regression functions in the model \mathcal{R} (i.e. when $\tilde{V} \geq K(\tilde{\rho}, \pi)/N$), the convergence rate will be $\sqrt{\frac{\log d}{N} \tilde{V}}$. In the worst case, the quantity \tilde{V} has the same order as B^2 , and we find a convergence rate $\sqrt{\frac{\log d}{N}}$ known to be optimal in the uniform sense as soon as $d > \sqrt{N}$ according to the following theorem

Theorem 4.6 (Yang, 2001). *Let $d = N^\tau$ for some $\tau > 0$. There exists a model*

$$\mathcal{R} = \{f_i \in \mathcal{F}(\mathcal{X}, \mathcal{Y}) : i = 1, \dots, d\}$$

such that for any aggregating procedure $\hat{\rho}$, one can find a function $\tilde{f} \in \tilde{\mathcal{R}} = \{\sum_{i=1}^d \tilde{\rho}_i f_i : \tilde{\rho} \in \mathcal{M}_+^1\{1, \dots, d\}\}$ satisfying

$$\mathbb{E}_{\mathbb{P}^{\otimes N}} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) \geq C \begin{cases} \frac{d}{N} & \text{when } \tau \leq \frac{1}{2}, \\ \sqrt{\frac{\log d}{N}} & \text{when } \tau > \frac{1}{2}, \end{cases}$$

where the constant C does not depend on N .

Remark 4.1. In [13], Yang also proposed an adaptive estimator. The advantage of the procedure designed here is to be feasible, to avoid splitting the data in many parts and to hold when the regression function wrt the unknown probability distribution is not in the model $\tilde{\mathcal{R}}$. Besides, our results also hold when the set of aggregated functions is infinite and under weaker assumptions (particularly on the noise).

Remark 4.2. Note that the unobservable term $r(\tilde{f})$ in the bound \mathbb{B} does not modify the probability distribution $\hat{\rho}_{\lambda,\eta,\beta,\xi}$ minimizing $\mathbb{B}(\rho, \lambda, \eta, \beta, \zeta)$. However the choice of λ among $(\lambda_i)_{i=0,\dots,p}$ depends on $r(\tilde{f})$. To circumvent this difficulty, one can, for instance, weaken the bound \mathbb{B} by replacing $\frac{r(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) - r(\tilde{f})}{1 - \lambda G(\lambda)}$ with

$$r(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) - r(\tilde{f}) + \frac{\lambda G(\lambda)}{1 - \lambda G(\lambda)} [r(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) - r(\hat{f}_{\text{ERM}})],$$

where the function \hat{f}_{ERM} minimizes the empirical risk among the functions in $\tilde{\mathcal{R}}$. For this algorithm, the first assertion of Theorem 4.5 becomes: for any $1/2 \geq \varepsilon > 0$,

$$\mathbb{P}^{\otimes N} (R(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) - R(\tilde{f}) \leq \gamma(\varepsilon) + r(\tilde{f}) - r(\hat{f}_{\text{ERM}})) \geq 1 - 2\varepsilon, \tag{4.8}$$

since

$$\sup_{\lambda \in (\lambda_i)_{i=0,\dots,p}} \left\{ \frac{\lambda G(\lambda)}{1 - \lambda G(\lambda)} \right\} = 1.$$

By using Theorem 4.1 (for a posterior distribution $\hat{\rho}_{\text{ERM}}$ satisfying $\mathbb{E}_{\hat{\rho}_{\text{ERM}}(d\theta)}f_\theta = \hat{f}_{\text{ERM}}$ and for λ and β of order $\sqrt{\frac{\log d}{N}}$), we get that the added term $r(\tilde{f}) - r(\hat{f}_{\text{ERM}})$ is at most of order $\sqrt{\frac{\log d}{N}}$ when the parameter set Θ is finite: $|\Theta| = d$.

Another solution to determine the right parameters is to cut the training sample into two parts, use the first part of the training sample to compute the distributions $\hat{\rho}_{\lambda,\eta,\beta,\xi}$ and use the second part of the training sample to select the best distribution among the $O[(\log N)^2]$ distributions (each distribution corresponds to a point in the (λ, β) -grid). This last step is almost free (since we neglect $\log \log N$ terms), so the convergence rate of the resulting procedure is effectively of order $\sqrt{\tilde{C} \tilde{V}} \vee \tilde{C}$.

Remark 4.3. Had we not been interested in having tight explicit constants, we could have written Theorem 4.1 in the following way (taking arbitrarily $\beta = \lambda$): there exists $C_1, C_2 > 0$ depending only on the constants B, α and M such that for any $\varepsilon > 0$ and $0 < \lambda' < C_1$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, for any aggregating procedure $\hat{\rho} : \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$,

$$R(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) - R(\tilde{f}) \leq (1 + \lambda') [r(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) - r(\tilde{f})] + 2\lambda' \tilde{V} + \frac{C_2}{N} \frac{K(\hat{\rho}, \pi) + \log(\varepsilon^{-1})}{\lambda'}$$

where we still have $\tilde{V} = \mathbb{E}_{\hat{\rho}} \text{Var}_{\hat{\rho}(d\theta)}f_\theta$. This inequation would have also led to the optimal convergence rate after optimization of the parameter λ' .

Theorem 4.6 also shows that a direct application of our aggregating procedure is not optimal when d is lower than \sqrt{N} , since then the convergence rate towards functions for which $\tilde{V} = \sup_{x \in \mathcal{X}} \text{Var}_{\hat{\rho}(d\theta)}f_\theta(x)$ has the same order as B^2 is

$$\sqrt{\frac{\log d}{N}} \gg \frac{d}{N}.$$

However, in this case ($d \leq \sqrt{N}$), one can consider a grid \mathcal{R}' on the simplex $\tilde{\mathcal{R}}$:

$$\mathcal{R}' \triangleq \left\{ \sum_{i=1}^d \frac{a_i}{\lfloor \sqrt{dN} \rfloor} f_i : a_i \in \mathbb{N} \text{ such that } \sum_{i=1}^d a_i = \lfloor \sqrt{dN} \rfloor \right\},$$

where $\lfloor x \rfloor$ denotes the integer satisfying $x - 1 < \lfloor x \rfloor \leq x$. We have $\tilde{\mathcal{R}}' = \tilde{\mathcal{R}}$. Then applying our aggregating procedure to the new initial model \mathcal{R}' for a uniform prior distribution π' on \mathcal{R}' , we obtain the desired convergence rate except for the logarithmic factor.

Proof. The best convex combination $\tilde{f} = \sum_{i=1}^d \tilde{\rho}_i f_i$ belongs to

$$\mathcal{S} \cap \left\{ \sum_{i=1}^d \frac{\lfloor \lfloor \sqrt{dN} \rfloor \tilde{\rho}_i \rfloor}{\lfloor \sqrt{dN} \rfloor} f_i + \frac{1}{\lfloor \sqrt{dN} \rfloor} C_d \right\},$$

where \mathcal{S} is the simplex $\{\sum_{i=1}^d \rho_i f_i : \rho_i \geq 0, \sum_{i=1}^d \rho_i = 1\}$ and C_d is the d -dimensional cube $\{\sum_{i=1}^d a_i f_i : 0 \leq a_i \leq 1\}$. This set is the convex combination of its vertices, so the function \tilde{f} can be written as a convex combination of the functions in

$$\mathcal{R}'' \triangleq \left\{ \sum_{i=1}^d \frac{\lfloor \lfloor \sqrt{dN} \rfloor \tilde{\rho}_i \rfloor + \varepsilon_i}{\lfloor \sqrt{dN} \rfloor} f_i : \varepsilon_i \in \{0, 1\} \right\} \cap \mathcal{R}'.$$

For any $f, g \in \mathcal{R}''$, we have

$$\|f - g\|_\infty \leq \frac{d}{2} \frac{B}{\lfloor \sqrt{dN} \rfloor},$$

hence¹

$$\tilde{V} \leq \frac{d^2}{16 \lfloor \sqrt{dN} \rfloor^2} B^2.$$

The number of functions in \mathcal{R}' is upper bounded by $(\lfloor \sqrt{dN} \rfloor + 1)^d$. Since we have $K(\tilde{\rho}, \pi') \leq \log \text{Card } \mathcal{R}'$ (because the distribution π' is uniform over the set \mathcal{R}'), we get $\tilde{C} \leq \frac{d \log(N^{3/4} + 1)}{N} B^2$. As a result, we have $\sqrt{\tilde{C}\tilde{V}} \vee \tilde{C} = O(\frac{d}{N} \log N)$, which is the desired convergence rate up to the logarithmic factor. \square

In fact, when $d \leq \sqrt{N}$, the optimal convergence rate can also be obtained by randomizing functions from the grid $\mathcal{R}' \subset \tilde{\mathcal{R}}$. To combine d regression functions is then equivalent (in terms of convergence rate) to randomizing with an appropriate Gibbs distribution on the grid \mathcal{R}' .

Remark 4.4. The previous idea of discretizing the model can be also used to find the best linear combination with bounded coefficients. Indeed, let A be the bound on the coefficients. Introduce the model

$$\mathcal{R}'_{\text{lin}} \triangleq \left\{ \sum_{i=1}^d \frac{q_i}{\lfloor \sqrt{dN} \rfloor} A f_i : q_i \in \mathbb{Z} \cap [-\lfloor \sqrt{dN} \rfloor; \lfloor \sqrt{dN} \rfloor] \right\}.$$

Then the best convex combination \tilde{f} of functions from $\mathcal{R}'_{\text{lin}}$ is also the best linear combination of functions from \mathcal{R} with coefficients bounded by A . Using our aggregating (or an appropriate randomizing) procedure on $\mathcal{R}'_{\text{lin}}$, we obtain that $\mathbb{E}_{\mathbb{P}^{\otimes N}} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f})$ is of order $\frac{d}{N} \log N$, which cannot be improved uniformly beyond the logarithmic factor.

Remark 4.5. Note that to obtain an algorithm with optimal convergence rate in the uniform sense, we need not have used sophisticated tools. We just need deviation inequalities, a simple union bound and to discretize the simplex $\tilde{\mathcal{R}}$. Indeed, any function f of $\tilde{\mathcal{R}}$ satisfies a deviation inequality similar to the one of Lemma 7.3: for any $0 \leq \lambda \leq \alpha B/2$ satisfying $8M\lambda \leq (\alpha B - 2\lambda)^2 e^2$, the deviations of

$$Z = -[Y - f(X)]^2 + [Y - \tilde{f}(X)]^2$$

¹ We use that for any random variable X such that $a \leq X \leq b$ a.s., the variance of X is bounded by $(b - a)^2/4$.

are given by

$$\log \mathbb{E}_{\mathbb{P}} e^{\lambda \frac{Z - \mathbb{E}_{\mathbb{P}} Z}{B^2}} \leq \lambda^2 \frac{\bar{R}(f)}{B^2} G(\lambda), \tag{4.9}$$

where $G(\lambda) \triangleq \frac{8M}{(\alpha B - 2\lambda)^2 e^2} + \frac{e^{2\lambda} - 1 - 2\lambda}{\lambda^2}$. The quantities $\bar{R}(f)$ and $\bar{r}(f)$ are still defined as

$$\begin{cases} \bar{R}(f) = R(f) - R(\tilde{f}) = \mathbb{E}_{\mathbb{P}}[(Y - f(X))^2] - \mathbb{E}_{\mathbb{P}}[(Y - \tilde{f}(X))^2], \\ \bar{r}(f) = r(f) - r(\tilde{f}) = \mathbb{E}_{\mathbb{P}}[(Y - f(X))^2] - \mathbb{E}_{\mathbb{P}}[(Y - \tilde{f}(X))^2]. \end{cases}$$

Hence, for any $0 \leq \lambda \leq \alpha B/2$ satisfying $\lambda G(\lambda) \leq 1$, we have successively

$$\mathbb{E}_{\mathbb{P}^{\otimes N}} e^{\frac{\lambda N}{B^2} \{\mathbb{E}_{\mathbb{P}} Z - \mathbb{E}_{\mathbb{P}} Z[1 - \lambda G(\lambda)]\}} \leq 1.$$

For any $\varepsilon > 0$,

$$\mathbb{P}^{\otimes N} \left\{ \frac{\lambda N}{B^2} \{\mathbb{E}_{\mathbb{P}} Z - \mathbb{E}_{\mathbb{P}} Z[1 - \lambda G(\lambda)]\} - \log(\varepsilon^{-1}) \geq 0 \right\} \leq \varepsilon.$$

With $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$,

$$\bar{R}(f) \leq \frac{\bar{r}(f)}{1 - \lambda G(\lambda)} + \frac{B^2}{N} \frac{\log(\varepsilon^{-1})}{\lambda[1 - \lambda G(\lambda)]}.$$

By using an union bound, for any discretized simplex $\mathcal{R}_{\text{disc}}$ with $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$, for any $f \in \mathcal{R}_{\text{disc}}$, we get

$$\bar{R}(f) \leq \frac{\bar{r}(f)}{1 - \lambda G(\lambda)} + \frac{B^2}{N} \frac{\log(\varepsilon^{-1} \text{Card } \mathcal{R}_{\text{disc}})}{\lambda[1 - \lambda G(\lambda)]}.$$

For some $m \in \mathbb{N}$ which will be chosen later, let us take

$$\mathcal{R}_{\text{disc}} = \left\{ \sum_{i=1}^d \frac{a_i}{m} f_i : a_i \in \mathbb{N} \text{ such that } \sum_{i=1}^d a_i = m \right\}.$$

Then we have

$$\text{Card } \mathcal{R}_{\text{disc}} = \binom{m+d}{d} \leq \begin{cases} 2 \times m^d & \text{when } d \leq m, \\ 2 \times d^m & \text{when } d \geq m, \end{cases}$$

and for any $g \in \tilde{\mathcal{R}}$ there exists $f \in \mathcal{R}_{\text{disc}}$ such that $\|f - g\|_{\infty} \leq \frac{B}{m}$. This last inequality implies that there exists $f \in \mathcal{R}_{\text{disc}}$ such that

$$\bar{r}(f) = \frac{1}{N} \sum_{i=1}^N [2Y_i - f(X_i) - \tilde{f}(X_i)][f(X_i) - \tilde{f}(X_i)] \leq \Sigma \frac{B}{m},$$

where

$$\Sigma \triangleq \frac{\sum_{i=1}^N |2Y_i - f(X_i) - \tilde{f}(X_i)|}{N} \leq 2 \frac{\sum_{i=1}^N |Y_i - f^*(X_i)|}{N} + 2B.$$

The algorithm which minimizes the empirical risk on the net $\mathcal{R}_{\text{disc}}$ satisfies with $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$, for any $f \in \mathcal{R}_{\text{disc}}$,

$$\bar{R}(\hat{f}) \leq \frac{\bar{r}(\hat{f}_{\text{disc}})}{1 - \lambda G(\lambda)} + \frac{B^2}{N} \frac{\log(\varepsilon^{-1} \text{Card } \mathcal{R}_{\text{disc}})}{\lambda[1 - \lambda G(\lambda)]},$$

where

$$\tilde{f}_{\text{disc}} \triangleq \underset{f \in \mathcal{R}_{\text{disc}}}{\operatorname{argmin}} R(f),$$

hence, by taking $\lambda = \kappa_1$ defined as $2\kappa_1 G(\kappa_1) = 1$,

$$R(\hat{f}) - R(\tilde{f}) \leq 2\Sigma \frac{B}{m} + \begin{cases} \frac{2B^2}{N\kappa_1} [d \log(m) + \log(2\varepsilon^{-1})] & \text{when } d \leq m, \\ \frac{2B^2}{N\kappa_1} [m \log(d) + \log(2\varepsilon^{-1})] & \text{when } d \geq m. \end{cases}$$

First, assume that the output data Y are bounded. Then we have $\Sigma \leq \kappa$ for some constant κ . By taking $m = \frac{N}{d}$ when $d \leq \sqrt{N}$ and $m = \sqrt{N/\log d}$ when $d > \sqrt{N}$, we obtain that with $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$,

$$R(\hat{f}) - R(\tilde{f}) \leq \begin{cases} \text{Cst } B^2 \left[\frac{d}{N} \log\left(\frac{N}{d}\right) + \frac{\log(2\varepsilon^{-1})}{N} \right] & \text{when } d \leq \sqrt{N}, \\ \text{Cst } B^2 \left[\sqrt{\frac{\log(d)}{N}} + \frac{\log(2\varepsilon^{-1})}{N} \right] & \text{when } d \geq \sqrt{N}. \end{cases} \tag{4.10}$$

In general, the output data Y are not bounded. However the quantity Σ behaves more or less like $2\mathbb{E}_{\mathbb{P}}|Y - f^*(X)| + 2B$. From assumption (2.2), this expectation is uniformly bounded wrt the distribution \mathbb{P} . Using once more deviation equalities, one can prove that with high probability Σ is bounded. So the bound (4.10) still holds. As a consequence, we have

$$\mathbb{P}^{\otimes N} R(\hat{f}) - R(\tilde{f}) \leq \begin{cases} \text{Cst } B^2 \frac{d}{N} \log\left(\frac{N}{d}\right) & \text{when } d \leq \sqrt{N}, \\ \text{Cst } B^2 \sqrt{\frac{\log d}{N}} & \text{whend } d \geq \sqrt{N}. \end{cases}$$

We have shown here that estimators having the optimal convergence rate (up to a logarithmic factor) can be constructed (but generally not easily implemented) using the ERM on an appropriate net of the model. It is interesting to notice that, in a different context [7,12], Mammen and Tsybakov similarly obtained optimal minimax convergence rate.

4.2.2. Aggregating procedure

We consider the aggregating procedure studied in Theorem 4.5: the algorithm minimizes the quantity $\mathbb{B}(\rho, (\lambda_i)_{i \in I}, (\eta_i)_{i \in I}, (\beta_j)_{j \in J}, (\zeta_j)_{j \in J})$ defined in (4.4) for well chosen parameter families.

This section explains how to minimize efficiently wrt the probability distribution ρ the quantity $\mathbb{B}(\rho, \lambda, \eta, \beta, \zeta)$ and shows that the resulting aggregated distribution has the same form as the optimal randomizing distribution (see Section 3.2), the difference being that the quantity that determines the weight given to each function is not just given by the empirical error but integrates a corrective factor that takes into account the errors made by the other weighted functions in a similar way as in Adaboost. Besides we will see that the corrective factor can be obtained by an algorithm in dual form which involves the choice of a N -dimensional real vector.

For fixed λ and β , we need to minimize a bound of the following type

$$\bar{\psi}(\rho) \triangleq a(r(\mathbb{E}_{\rho(d\theta)} f_\theta) + b\mathbb{E}_{\mathbb{P}} \text{Var}_{\rho(d\theta)} f_\theta + cK(\rho, \pi)),$$

where $a > 0, 0 < b < 1$ and $c > 0$.²

² For our bound, we have $a = \frac{1}{1-\lambda G(\lambda)}$, $b = \frac{\beta g(\beta) + \lambda G(\lambda)}{1+\beta g(\beta)}$ and $c = \frac{B^2}{N\lambda} \left(1 + \frac{\lambda[1-\lambda G(\lambda)]}{\beta[1+\beta g(\beta)]}\right)$.

Writing the dual problem. For any measurable function such that e^h is π -integrable, introduce the probability distribution

$$\pi_h \triangleq \frac{e^h}{\mathbb{E}_{\pi(d\theta)} e^{h(\theta)}} \cdot \pi.$$

Since we have

$$\begin{cases} \mathbb{E}_{\rho} r(f_{\theta}) = r(\mathbb{E}_{\rho(d\theta)} f_{\theta}) + \mathbb{E}_{\bar{\pi}} \text{Var}_{\rho(d\theta)} f_{\theta}, \\ K(\rho, \pi_{-\frac{b}{c}r(f)}) = K(\rho, \pi) + \frac{b}{c} \mathbb{E}_{\rho} r(f_{\theta}) + \log \mathbb{E}_{\pi(d\theta)} e^{-\frac{b}{c}r(f_{\theta})}, \end{cases}$$

we can write

$$\begin{aligned} \bar{\psi}(\rho) &= a((1-b)r(\mathbb{E}_{\rho(d\theta)} f_{\theta}) + b\mathbb{E}_{\rho} r(f_{\theta}) + cK(\rho, \pi)) \\ &= a((1-b)r(\mathbb{E}_{\rho(d\theta)} f_{\theta}) + cK(\rho, \pi_{-\frac{b}{c}r(f)}) - c \log \mathbb{E}_{\pi(d\theta)} e^{-\frac{b}{c}r(f_{\theta})}) \\ &= ac \left(\frac{1-b}{Nc} \sum_{i=1}^N [Y_i - \mathbb{E}_{\rho(d\theta)} f_{\theta}(X_i)]^2 + K(\rho, \pi_{-\frac{b}{c}r(f)}) \right) - ac \log \mathbb{E}_{\pi(d\theta)} e^{-\frac{b}{c}r(f_{\theta})}. \end{aligned}$$

Hence minimizing $\bar{\psi}$ is equivalent to minimizing

$$\psi(\rho) \triangleq \frac{1}{2} \|\mathbb{E}_{\rho(d\theta)} h(\theta)\|^2 + K(\rho, \mu),$$

where $\mu \triangleq \pi_{-\frac{b}{c}r(f)}$, $\|\cdot\|$ the euclidean norm in \mathbb{R}^N and $h : \Theta \rightarrow \mathbb{R}^N$ is defined by

$$h_i(\theta) \triangleq \sqrt{\frac{2(1-b)}{Nc}} [Y_i - f_{\theta}(X_i)].$$

The minimization of the function ψ over the set of probability distributions has some distinctive features stressed in the following theorem.

Theorem 4.7. For any $\mu \in \mathcal{M}_+^1(\Theta)$ and any bounded function $h : \Theta \rightarrow \mathbb{R}^N$, the map ψ has a unique minimum $\bar{\rho}$ in $\mathcal{M}_+^1(\Theta)$. Besides, the probability distribution $\bar{\rho}$ is the only distribution satisfying

$$\bar{\rho}(d\theta) = \mu_{-\langle \mathbb{E}_{\bar{\rho}} h, h(\theta) \rangle}(d\theta) = \frac{e^{-\langle \mathbb{E}_{\bar{\rho}} h, h(\theta) \rangle}}{\mathbb{E}_{\mu(d\theta')} e^{-\langle \mathbb{E}_{\bar{\rho}} h, h(\theta') \rangle}} \cdot \mu(d\theta),$$

and we have

$$\psi(\rho) - \psi(\bar{\rho}) = K(\rho, \bar{\rho}) + \frac{1}{2} \|\mathbb{E}_{\rho} h - \mathbb{E}_{\bar{\rho}} h\|^2 \quad \text{for any } \rho \in \mathcal{M}_+^1(\Theta).$$

Proof. See Section 7.5 \square

Introduce $d_1 \triangleq \frac{b}{cN}$ and $d_2 \triangleq \frac{1-b}{cN}$. From assumption (2.1), the mappings h_i are bounded and we can apply the previous theorem. So the optimal distribution has the following form $\pi^w \triangleq \pi_{-d_1Nr(f) + \langle w, f(X) \rangle}$, where w is a N -dimensional vector to be determined. Note that in support vector machines, we have to solve a N -dimensional linearly constrained quadratic problem. Here we have a N -dimensional unconstrained minimization problem. Both methods come down to an N -dimensional optimization problem because they both write the dual of an initial learning problem.

For the optimal w , from the previous theorem, the posterior distribution is

$$\pi^w = \pi_{-d_1Nr(f) + 2d_2(Y - \mathbb{E}_{\pi^w(d\theta)} f_{\theta}(X), f(X) - Y)}.$$

So the optimal distribution π^w stresses on functions with low empirical risk and such that they make the opposite error as the combined estimator (since the bigger $\langle Y - \mathbb{E}_{\pi^w} f(X), f_\theta(X) - Y \rangle$ is, the more weight π^w gives to f_θ). This is precisely the idea that has lead to the first boosting methods, such as AdaBoost.

Solving the dual problem. Note that the unicity of the optimal probability distribution π^w according to Theorem 4.7 does not give the unicity of the vector w . We have $\pi_h = \pi_{h'}$ if and only if $h = h' + \text{Cst}$ π -a.s. Therefore we have $\pi^w = \pi^{w'}$ iff $\langle w - w', f(X) \rangle = \text{Cst}$ π -a.s.

Define

$$\bar{\varphi}(w) \triangleq \bar{\psi}(\pi^w) = ac \left[d_2 \left\| \mathbb{E}_{\pi^w} f(X) - Y \right\|^2 - \log \mathbb{E}_{\pi_{-\frac{1}{c}r(f)}} e^{\langle w, f(X) - \mathbb{E}_{\pi^w} f(X) \rangle} \right] - ac \log \mathbb{E}_{\pi} e^{-\frac{1}{c}r(f)}.$$

We have

$$\nabla \bar{\varphi}(w) = ac \mathbb{V}ar_{\pi^w} f(X) (2d_2 [\mathbb{E}_{\pi^w} f(X) - Y] + w),$$

where $\mathbb{V}ar_{\pi^w} f(X)$ is the covariance matrix of $f(X_i), i = 1, \dots, N$, wrt π^w . Denote r the rank of this matrix. Usually, we have $r = N$. Then there is no vector v such that $\langle v, f(X) \rangle = \text{Cst}$ π -a.s. Hence, in that case, there is a unique optimal w .

However, it may happen that $r < N$ (for instance when two input vectors are identical i.e. $X_i = X_j$ for some $i \neq j$). Even if it means numbering again, one may assume that $f(X_{r+1}), \dots, f(X_N)$ are π -linear combination of $f(X_1), \dots, f(X_r)$ to the extent that there exists $\alpha^i \in \mathbb{R}^r, \beta^i \in \mathbb{R}, i = r + 1, \dots, N$, such that for any $i \in \{r + 1, \dots, N\}$

$$f(X_i) = \langle \alpha^i, f(X) \rangle_r + \beta^i \quad \pi\text{-a.s.},$$

where $\langle \cdot, \cdot \rangle_r$ is the dot product in \mathbb{R}^r . From Theorem 4.7, we look for a N -dimensional vector w such that

$$\langle w, f(X) \rangle = 2d_2 \langle \mathbb{E}_{\pi^w} [Y - f(X)], f(X) \rangle + \text{Cst} \quad \pi\text{-a.s.} \tag{4.11}$$

Without constraints on w , there is an infinity of such vectors. Since we have

$$\begin{aligned} \langle \mathbb{E}_{\pi^w} [Y - f(X)], f(X) \rangle &= \sum_{j=1}^r \mathbb{E}_{\pi^w} [Y_j - f(X_j)] f(X_j) \\ &\quad + \sum_{i=r+1}^N \mathbb{E}_{\pi^w} [Y_i - \langle \alpha^i, f(X) \rangle_r - \beta^i] (\langle \alpha^i, f(X) \rangle_r + \beta^i) \\ &= \sum_{j=1}^r \left(\mathbb{E}_{\pi^w} [Y_j - f(X_j)] + \sum_{i=r+1}^N \alpha_j^i \mathbb{E}_{\pi^w} [Y_i - \langle \alpha^i, f(X) \rangle_r - \beta^i] \right) f(X_j) \\ &\quad + \sum_{i=r+1}^N \beta^i \mathbb{E}_{\pi^w} [Y_i - \langle \alpha^i, f(X) \rangle_r - \beta^i], \end{aligned}$$

one may set w_{r+1}, \dots, w_N to 0 and solve only a r -dimensional minimization problem for which the *unique* solution is

$$w = 2d_2 \left(Y - \mathbb{E}_{\pi^w} f(X) + \sum_{i=r+1}^N \alpha^i [Y_i - \langle \alpha^i, \mathbb{E}_{\pi^w} f(X) \rangle_r - \beta^i] \right). \tag{4.12}$$

Remark 4.6. In the case when none of the functions of the model discriminates X_i from X_j for some $i > j$ (i.e. $f_\theta(X_i) = f_\theta(X_j)$ for any $\theta \in \Theta$), we have $\alpha_j^i = 1$ and $\alpha_k^i = 0$ for $k \neq j$. Hence, in equality (4.12), there is no additional term in w_k for $k \neq j$ and the additional term in w_j is simply $Y_i - \mathbb{E}_{\pi^w} f(X_j)$.

Remark 4.7. From assumption (2.1), for any $x \in \mathcal{X}$, the mapping $[\theta \mapsto f_\theta(x)]$ is bounded. So we can write a bracketing of w . For instance, when $r = N$, we have

$$w_i \in [2d_2(Y_i - \sup_{\theta \in \Theta} f_\theta(X_i)); 2d_2(Y_i - \inf_{\theta \in \Theta} f_\theta(X_i))].$$

Remark 4.8. It follows from $w_{r+1} = \dots = w_N = 0$ that

$$\begin{aligned} \frac{1}{ac} \frac{\partial \bar{\varphi}}{\partial w_k}(w) &= \sum_{j=1}^r \text{Cov}_{\pi^w} [f(X_k), f(X_j)] (2d_2 \mathbb{E}_{\pi^w} [Y_j - f(X_j)] + w_j) \\ &\quad + \sum_{i=r+1}^N 2d_2 \text{Cov}_{\pi^w} [f(X_k), \langle \alpha^i, f(X) \rangle_r] \mathbb{E}_{\pi^w} [Y_i - \langle \alpha^i, f(X) \rangle_r - \beta^i] \\ &= \sum_{j=1}^r \text{Cov}_{\pi^w} [f(X_k), f(X_j)] \left(w_j + 2d_2 \mathbb{E}_{\pi^w} [Y_j - f(X_j)] \right. \\ &\quad \left. + 2d_2 \sum_{i=r+1}^N \alpha_j^i \mathbb{E}_{\pi^w} [Y_i - \langle \alpha^i, f(X) \rangle_r - \beta^i] \right), \end{aligned}$$

hence

$$\nabla_r \bar{\varphi}(w) = ac \mathbb{V}\text{ar}_{\pi^{w^l}} f(X)|_r \left[w - 2d_2 \left(Y - \mathbb{E}_{\pi^w} f(X) + \sum_{i=r+1}^N \alpha^i [Y_i - \langle \alpha^i, \mathbb{E}_{\pi^w} f(X) \rangle_r - \beta^i] \right) \right],$$

where $\nabla_r \bar{\varphi}$ is the vector $\frac{\partial \bar{\varphi}}{\partial w_k}$, $k = 1, \dots, r$, and $\mathbb{V}\text{ar}_{\pi^{w^l}} f(X)|_r$ is the covariance matrix of $f(X_1), \dots, f(X_r)$. This is another method of proving that an optimal w is given by (4.12). It is also the required formula to program a gradient descent algorithm in order to compute the optimal vector w . However, the variance matrix being computationally too expensive,³ we would prefer the following alternative minimization procedure.

Algorithm.

BEGIN

Start with $w^0 = 0$.

For $l = 0$ to maximum number of iterations do

- Set

$$w^{l+1} = 2d_2 \left(Y - \mathbb{E}_{\pi^{w^l}} f(X) + \sum_{i=r+1}^N \alpha^i [Y_i - \langle \alpha^i, \mathbb{E}_{\pi^{w^l}} f(X) \rangle_r - \beta^i] \right).$$

- Exit the loop if w^{l+1} is not “far” from w^l .
- While $\bar{\varphi}(w^{l+1}) > \bar{\varphi}(w^l)$ do

$$w^{l+1} = \frac{1}{2}(w^l + w^{l+1}).$$

END

³ In our numerical experiments described in Section 5, the order of the number of operations required to compute the N^2 covariances is $N^2 \times Nd$, where d is the dimensionality of the input vector (see Corollary 5.3 for details). In this framework, the gradient descent algorithm roughly loses a factor N in computational complexity wrt to the following procedure.

The stopping criteria in the loop comes from

Theorem 4.8. For any $w, w' \in \mathbb{R}^N$, we have

$$\begin{aligned} \bar{\varphi}(w) - \bar{\varphi}(w') &= ac(d_2 \|\mathbb{E}_{\pi^w} f(X) - \mathbb{E}_{\pi^{w'}} f(X)\|^2 + K(\pi^w, \pi^{w'})) \\ &\quad + \langle w' + 2d_2(\mathbb{E}_{\pi^{w'}} f(X) - Y), \mathbb{E}_{\pi^w} f(X) - \mathbb{E}_{\pi^{w'}} f(X) \rangle. \end{aligned}$$

In particular, we have

$$\bar{\psi}(\pi^{w'}) - \bar{\psi}(\bar{\rho}) \leq acB \left\| w' - 2d_2 \left(Y - \mathbb{E}_{\pi^{w'}} f(X) + \sum_{i=r+1}^N \alpha^i [Y_i - \langle \alpha^i, \mathbb{E}_{\pi^{w'}} f(X) \rangle_r - \beta^i] \right) \right\|.$$

Proof. See Section 7.6. \square

In Section 7.7, we prove that we exit the “While” loop in a finite number of iterations. Finally, we obtain an algorithm which derives directly from Corollary 4.3. However this procedure tends to regularize too much. The obtained bounds are upper bounds and even if a lot of care was taken to get sharp bounds, they still are quantitatively loose for small sample sizes. As a consequence, the regularization parameters coming from these bounds are too conservative. So in our numerical experiments, these parameters are tuned using validation sets. The previous minimization procedure will however be used to get the optimal aggregating distribution associated with a set of these parameters.

4.3. Expected risk bound for any aggregating procedure

From Corollary 4.3, we also derive an empirical bound on the expected risk of any aggregating procedure. One of the output of the algorithm described in the previous section is an upper bound of $R(\mathbb{E}_{\pi^{w_{\text{opt}}}} f) - R(\tilde{f})$. It can also be interesting to upper bound $R(\mathbb{E}_{\pi^{w_{\text{opt}}}} f)$ (since $R(\tilde{f})$ is unknown). The following corollary gives an observable upper bound of the expected risk of any aggregating procedure.

Corollary 4.9. For any $\varepsilon > e^{-\kappa_3 N}$, with $(\mathbb{P}^{\otimes N})_*$ -probability at least $1 - 3\varepsilon$, for any aggregating procedure $\hat{\rho} : \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) &\leq r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}' + \mathcal{L}^2 \frac{\log(\varepsilon^{-1})}{N} + \frac{4B^2 \log(\varepsilon^{-1})}{\kappa_1 N} \\ &\quad + 2\mathcal{L} \sqrt{\frac{\log(\varepsilon^{-1})}{N}} \sqrt{r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}' + \mathcal{L}^2 \frac{\log(\varepsilon^{-1})}{N}}, \end{aligned}$$

where

$$\left\{ \begin{array}{l} \mathbb{B}' \triangleq \inf_{\substack{i \in I \\ j \in J}} \mathbb{B}'(\hat{\rho}, \lambda_i, \eta_i, \beta_j, \zeta_j), \\ \mathbb{B}'(\rho, \lambda, \eta, \beta, \zeta) \triangleq H(\lambda) \left(\lambda G(\lambda) [\mathbb{E}_{\rho(d\theta)} r(f_\theta) - \inf_{\tilde{\mathcal{R}}} r] + B^2 \frac{K(\rho, \pi) + \log[(\eta\varepsilon)^{-1}]}{N\lambda} \right) \\ \quad + h(\beta) \left(\beta g(\beta) \bar{V}(\rho) + B^2 \frac{2K(\rho, \pi) + \log[(\zeta\varepsilon)^{-1}]}{2N\beta} \right) \\ = H(\lambda) \left(\lambda G(\lambda) [r(\mathbb{E}_{\rho(d\theta)} f_\theta) - \inf_{\tilde{\mathcal{R}}} r] + B^2 \frac{K(\rho, \pi) + \log[(\eta\varepsilon)^{-1}]}{N\lambda} \right) \\ \quad + [\lambda G(\lambda) H(\lambda) + \beta g(\beta) h(\beta)] \bar{V}(\rho) + B^2 h(\beta) \frac{2K(\rho, \pi) + \log[(\zeta\varepsilon)^{-1}]}{2N\beta}, \\ \mathcal{L} \triangleq \frac{1}{\sqrt{2\alpha}} \left[\log \left(\kappa_4 \frac{N}{\log(\varepsilon^{-1})} \right) \right]^2, \\ \bar{V}(\rho) \triangleq \mathbb{E}_{\mathbb{P}} \text{Var}_{\rho(d\theta)} f_\theta \end{array} \right.$$

and

$$\left\{ \begin{array}{l} \kappa_3 \triangleq \frac{M^2 e^{2(\alpha B - 1)}}{2[(\alpha B e)^2 + 4M]}, \\ \kappa_4 \triangleq \frac{M e^{\alpha B + 1}}{\alpha B} \sqrt{\frac{\kappa_1}{8}}, \quad \text{where by definition, } \kappa_1 \text{ satisfies } 2\kappa_1 G(\kappa_1) = 1. \end{array} \right.$$

Proof. See Section 7.8. \square

Remark 4.9. Once more, the threshold on ε is negligible, and κ_3 can be disregarded.

Remark 4.10. When $r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta)$ and $\bar{V}(\hat{\rho})$ are of order $1/N$, the bound on the expected risk $R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta)$ is of order $(\log N)^4/N$. For bounded noise (i.e. $Y - \mathbb{E}_{\mathbb{P}}(Y/X)$ uniformly bounded on \mathcal{X}), the argument in Section 7.8 can be easily adapted to get rid of the $(\log N)^4$ factor (since the deviations of the empirical risk of the best convex combination can be bounded using the first part of Lemma 7.2). This is the case in the classification context (see Corollary 4.11).

Remark 4.11. We will see in Section 7.8 that this corollary follows from Corollary 4.3 by controlling the deviations of the empirical risk $r(\tilde{f})$ of the best convex combination. A bound on the expected risk of any randomization procedure can be similarly deduced from this control.

Remark 4.12. The constants in Corollary 4.9 can be slightly improved by using Remark 7.4. Indeed, when $\tilde{f} = \mathbb{E}_{\mathbb{P}}(Y/X = \cdot)$, Lemma 7.6 holds for

$$\tilde{L} = \log \left(M e \sqrt{\frac{N}{2 \log(\varepsilon^{-1}) \alpha^2 R(\tilde{f})}} \right)$$

and $\kappa_3 = \frac{M^2 e^{-2}}{2(e^{2(\alpha B e)^2} + 4M)}$ (since inequality (7.14) can be improved by eliminating the $e^{\alpha B}$ factor). Therefore the corollary remains true for

$$\begin{cases} \kappa_3 = \frac{M^2 e^{-2}}{2[(\alpha B e)^2 + 4M]}, \\ \kappa_4 = \frac{M e}{\alpha B} \sqrt{\frac{\kappa_1}{8}}. \end{cases}$$

4.4. Application to binary classification

In binary classification, the output set is $\mathcal{Y} = \{0, 1\}$, and the model consists in a set of functions on the input space \mathcal{X} taking their values in $[0; 1]$. In this framework, the constants α and M in assumption (2.2) are not relevant since the output is bounded. Besides, we have $B = 1$. We still denote $g(\lambda) \triangleq \frac{e^\lambda - 1 - \lambda}{\lambda^2}$, $h(\beta) \triangleq \frac{1}{1 + \beta g(\beta)}$ and we define $\check{h}(\lambda) \triangleq \frac{1}{1 - 4\lambda g(\lambda)}$. Theorem 4.2 can be replaced by

Theorem 4.10. *Introduce countable families $(\lambda_i)_{i \in I}$, $(\eta_i)_{i \in I}$, $(\beta_j)_{j \in J}$ and $(\zeta_j)_{j \in J}$ such that $\lambda_i > 0$, $4\lambda_i g(\lambda_i) < 1$, $\eta_i > 0$, $\sum_{i \in I} \eta_i = 1$, $\beta_j > 0$, $\zeta_j > 0$ and $\sum_{j \in J} \zeta_j = 1$. For any $\varepsilon > 0$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, for any randomizing procedure $\hat{\rho} : \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$, for any $i \in I$ and for any $j \in J$, we have*

$$\begin{aligned} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) &\leq \check{h}(\lambda_i) [r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - r(\tilde{f})] + [\check{h}(\lambda_i) - h(\beta_j)] \bar{V} \\ &\quad + \frac{\check{h}(\lambda_i)}{N\lambda_i} \{K(\hat{\rho}, \pi) + \log[(\eta_i \varepsilon)^{-1}]\} \\ &\quad + \frac{h(\beta_j)}{2N\beta_j} \{2K(\hat{\rho}, \pi) + \log[(\zeta_j \varepsilon)^{-1}]\}, \end{aligned} \tag{4.13}$$

where $\bar{V}(\hat{\rho}) \triangleq \mathbb{E}_{\mathbb{P}} \text{Var}_{\hat{\rho}(d\theta)} f_\theta$.

Proof. The proof is similar to the ones which lead to Theorem 4.2. The only part to modify is in Section 7.2. Since we have trivially $B = 1$, the deviations of $Z_\theta = -(Y - f_\theta(X))^2 + (Y - \tilde{f}(X))^2 = [f_\theta(X) - \tilde{f}(X)][2Y - \tilde{f}(X) - f_\theta(X)]$ given by Lemma 7.3 can be obtained by using directly Lemma 7.2 to Z_θ ($b = 1$). We get

$$\log \mathbb{E}_{\mathbb{P}} e^{\lambda(Z_\theta - \mathbb{E}_{\mathbb{P}} Z_\theta)} \leq \lambda^2 \mathbb{E}_{\mathbb{P}} Z_\theta^2 g(\lambda) \leq 4\lambda^2 \bar{R}(\theta) g(\lambda),$$

Consequently, $G(\lambda)$ can be replaced by $4g(\lambda)$. \square

From Theorem 4.10, we may derive an empirical bound on the expected risk of any combining procedure.

Corollary 4.11. *For any countable families $(\lambda_i)_{i \in I}$, $(\eta_i)_{i \in I}$, $(\beta_j)_{j \in J}$ and $(\zeta_j)_{j \in J}$ such that $\lambda_i > 0$, $4\lambda_i g(\lambda_i) < 1$, $\eta_i > 0$, $\sum_{i \in I} \eta_i = 1$, $\beta_j > 0$, $\zeta_j > 0$ and $\sum_{j \in J} \zeta_j = 1$, for any $\varepsilon > 0$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, for any randomizing procedure $\hat{\rho} : \mathcal{Z}^N \rightarrow \mathcal{M}_+^1(\Theta)$, we have*

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) \leq r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'' + \sqrt{\frac{2\log(\varepsilon^{-1})}{N}} \left(\sqrt{r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'' + \frac{\log(\varepsilon^{-1})}{2N}} + \sqrt{\frac{\log(\varepsilon^{-1})}{2N}} \right),$$

where

$$\left\{ \begin{aligned} \mathbb{B}'' &\triangleq \inf_{\substack{i \in I \\ j \in J}} \mathbb{B}''(\hat{\rho}, \lambda_i, \eta_i, \beta_j, \zeta_j), \\ \mathbb{B}''(\rho, \lambda, \eta, \beta, \zeta) &\triangleq \check{h}(\lambda) \left(4\lambda g(\lambda) [\mathbb{E}_{\rho(d\theta)} r(f_\theta) - \inf_{\mathcal{R}} r] + \frac{K(\rho, \pi) + \log[(\eta\varepsilon)^{-1}]}{N\lambda} \right) \\ &\quad + h(\beta) \left(\beta g(\beta) \bar{V}(\rho) + \frac{2K(\rho, \pi) + \log[(\zeta\varepsilon)^{-1}]}{2N\beta} \right) \\ &= \check{h}(\lambda) \left(4\lambda g(\lambda) [r(\mathbb{E}_{\rho(d\theta)} f_\theta) - \inf_{\mathcal{R}} r] + \frac{K(\rho, \pi) + \log[(\eta\varepsilon)^{-1}]}{N\lambda} \right) \\ &\quad + [4\lambda g(\lambda) \check{h}(\lambda) + \beta g(\beta) h(\beta)] \bar{V}(\rho) + h(\beta) \frac{2K(\rho, \pi) + \log[(\zeta\varepsilon)^{-1}]}{2N\beta}. \end{aligned} \right.$$

Proof. The proof is similar to the one in Section 7.8. To control the deviations of the empirical risk $r(\tilde{f})$ of the best convex combination, we apply inequality (7.1) directly to $Z = (Y - \tilde{f}(X))^2 \in [0, 1]$. For any $\lambda > 0$ and any $\mu \in \mathbb{R}$, we have

$$\begin{aligned} \mathbb{P}^{\otimes N}(R(\tilde{f}) - r(\tilde{f}) > \mu) &\leq \mathbb{E}_{\mathbb{P}^{\otimes N}} e^{N\lambda(R(\tilde{f}) - r(\tilde{f}) - \mu)} \\ &\leq e^{-N\lambda\mu} (\mathbb{E}_{\mathbb{P}} e^{\lambda(\mathbb{E}_{\mathbb{P}} Z - Z)})^N \\ &\leq e^{N(-\lambda\mu + \frac{\lambda^2}{2} \mathbb{E}_{\mathbb{P}} Z)}. \end{aligned}$$

For $\mu = \frac{\log(\varepsilon^{-1})}{N\lambda} + \frac{\lambda}{2} R(\tilde{f})$, this last bound is equal to ε . The previous inequality holds for any $\lambda > 0$. To get a small μ , we take $\lambda = \sqrt{\frac{2\log(\varepsilon^{-1})}{NR(\tilde{f})}}$ (when $R(\tilde{f}) \neq 0$; otherwise the result is trivial). It follows that with $\mathbb{P}^{\otimes N}$ -probability at least $1 - \varepsilon$,

$$R(\tilde{f}) - r(\tilde{f}) \leq \sqrt{\frac{2\log(\varepsilon^{-1})R(\tilde{f})}{N}}.$$

Using Theorem 4.10, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 3\varepsilon$, we obtain

$$R(\tilde{f}) \leq R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) \leq \sqrt{\frac{2\log(\varepsilon^{-1})R(\tilde{f})}{N}} + r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'',$$

where \mathbb{B}'' is the quantity defined in Corollary 4.11. Hence, we have successively

$$\left(\sqrt{R(\tilde{f})} - \sqrt{\frac{\log(\varepsilon^{-1})}{2N}} \right)^2 \leq r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'' + \frac{\log(\varepsilon^{-1})}{2N},$$

$$\sqrt{R(\tilde{f})} \leq \sqrt{r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}''} + \frac{\log(\varepsilon^{-1})}{2N} + \sqrt{\frac{\log(\varepsilon^{-1})}{2N}},$$

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) \leq r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'' + \sqrt{\frac{2\log(\varepsilon^{-1})}{N}} \left(\sqrt{r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}''} + \frac{\log(\varepsilon^{-1})}{2N} + \sqrt{\frac{\log(\varepsilon^{-1})}{2N}} \right). \quad \square$$

5. Numerical examples: binary classification

5.1. Setup and notations

The setting is quite simple: the input data are d -dimensional: $\mathcal{X} = \mathbb{R}^d$. In binary classification, the output set is $\mathcal{Y} = \{0, 1\}$. The model consists in all decision stumps. By definition, these stumps achieve a binary partition of \mathcal{X} along hyperplanes orthogonal to the axes in the canonical base of \mathcal{X} . In other words, they compare one component of the input data to a threshold. Hence the model is

$$\mathcal{R} = \left\{ \alpha_0 \mathbb{1}_{x_j < \tau} + \alpha_1 \mathbb{1}_{x_j \geq \tau} : j \in \{1, \dots, d\}, \tau \in \mathbb{R}, \alpha_0 \in [0, 1], \alpha_1 \in [0, 1] \right\}. \tag{5.1}$$

Recall that the set of all df (distribution functions) is the set of increasing càdlàg functions F such that

$$\begin{cases} \lim_{x \rightarrow -\infty} F(x) = 0, \\ \lim_{x \rightarrow +\infty} F(x) = 1. \end{cases}$$

Theorem 5.1. *The set $\tilde{\mathcal{R}}$ of mixtures of elements of \mathcal{R} is an additive model*

$$\tilde{\mathcal{R}} = \left\{ x \mapsto \sum_{j=1}^d \alpha_j h_j(x_j) : \text{for any } j \in \{1, \dots, d\}, h_j \in \mathcal{H}, \alpha_j \geq 0 \text{ and } \sum_{j=1}^d \alpha_j = 1 \right\}, \tag{5.2}$$

where

$$\mathcal{H} \triangleq \{ \alpha F + \beta(1 - G) + \gamma : \alpha \geq 0, \beta \geq 0, \gamma \geq 0, \alpha + \beta + \gamma \leq 1, F \text{ df}, G', \text{ df} \}.$$

$\tilde{\mathcal{R}}$ can also be written

$$\tilde{\mathcal{R}} = \left\{ x \mapsto \gamma + \sum_{j=1}^d (\alpha_j F_j(x_j) + \beta_j [1 - G_j(x_j)]) : \text{for any } j \in \{1, \dots, d\}, \right. \\ \left. F_j \text{ df}, G_j \text{ df}, \alpha_j \geq 0, \beta_j \geq 0 \text{ and } \gamma + \sum_{j=1}^d (\alpha_j + \beta_j) \leq 1 \right\}. \tag{5.3}$$

Proof. By definition, the set of mixtures of elements in \mathcal{R} is the set of functions which can be written as $\mathbb{E}_{\pi(dX)} X$, where π is a probability measure on \mathcal{R} . This definition requires to have put a sigma algebra on \mathcal{R} . In our context, we take the canonical one. Introduce the set

$$\mathcal{R}' \triangleq \{0_{\mathbb{R}}\} \cup \{1_{\mathbb{R}}\} \bigcup_{\substack{j \in \{1, \dots, d\} \\ \tau \in \mathbb{R}}} \{ \mathbb{1}_{x_j \geq \tau} \} \bigcup_{\substack{j' \in \{1, \dots, d\} \\ \tau' \in \mathbb{R}}} \{ \mathbb{1}_{x_{j'} < \tau'} \},$$

where $0_{\mathbb{R}} : x \mapsto 0$ and $1_{\mathbb{R}} : x \mapsto 1$. Let us put on \mathcal{R}' its canonical sigma algebra. Denote $\text{Mixt}(\mathcal{R}')$ the set of mixtures of elements in \mathcal{R}' . Since $\mathcal{R} \subset \text{Mixt}(\mathcal{R}')$ and $\mathcal{R}' \subset \mathcal{R}$, we have $\text{Mixt}(\mathcal{R}') = \text{Mixt}(\mathcal{R}) = \tilde{\mathcal{R}}$. Hence any element of $\tilde{\mathcal{R}}$ can be written $\mathbb{E}_{\rho(dX)} X$, where ρ is a probability distribution on \mathcal{R}' . Then define $\gamma = \rho(1_{\mathbb{R}})$, for any $j \in \{1, \dots, d\}$, $\alpha_j = \rho(j)$, for any $j' \in \{1, \dots, d\}$, $\beta_{j'} = \rho(j')$, $\mu_j(d\tau) = \rho(d\tau/j)$ the probability distribution on \mathbb{R} and $\nu_{j'}(d\tau') = \rho(d\tau'/j')$ the probability distribution on \mathbb{R} . Denote F_j the df of μ_j and $G_{j'}$ the df of $\nu_{j'}$. Then we have $\mathbb{E}_{\rho(dX)} X = \rho(0_{\mathbb{R}})0_{\mathbb{R}} + \rho(1_{\mathbb{R}})1_{\mathbb{R}} + \sum_{j=1}^d \rho(j) \mathbb{E}_{\rho(dX/j)} X + \sum_{j'=1}^d \rho(j') \mathbb{E}_{\rho(dX/j')} X$. Hence $\mathbb{E}_{\rho(dX)} X(x) = \gamma + \sum_{j=1}^d \alpha_j F_j(x_j) + \sum_{j'=1}^d \beta_{j'} [1 - G_{j'}(x_{j'})]$. From the definitions, it comes that for any $j \in \{1, \dots, d\}$, F_j and G_j are df, $\alpha_j \geq 0$, $\beta_j \geq 0$ and $\gamma + \sum_{j=1}^d (\alpha_j + \beta_j) \leq 1$. Therefore, we have

$$\tilde{\mathcal{R}} \subset \left\{ x \mapsto \gamma + \sum_{j=1}^d (\alpha_j F_j(x_j) + \beta_j [1 - G_j(x_j)]) : \text{for any } j \in \{1, \dots, d\}, \right. \\ \left. F_j \text{ df}, G_j \text{ df}, \alpha_j \geq 0, \beta_j \geq 0 \text{ and } \gamma + \sum_{j=1}^d (\alpha_j + \beta_j) \leq 1 \right\}.$$

Inversely, using the same ideas in the reverse order, one can prove the other inclusion. So equality (5.3) is true. Equality (5.2) directly comes from it. \square

Remark 5.1. The model $\tilde{\mathcal{R}}$ is additive. As any additive model, it cannot classify well data coming from certain simple generator. One of the simplest is the 4-checked draughtboard defined as

$$\begin{cases} \mathcal{L}(X) = \mathcal{U}[0; 1] \times \mathcal{U}[0; 1], \\ \mathcal{L}(Y/X = (x_1, x_2)) = \begin{cases} \delta_0 & \text{when } x_1 < 1/2 \text{ and } x_2 < 1/2, \\ \delta_1 & \text{when } x_1 < 1/2 \text{ and } x_2 \geq 1/2, \\ \delta_1 & \text{when } x_1 \geq 1/2 \text{ and } x_2 < 1/2, \\ \delta_0 & \text{when } x_1 \geq 1/2 \text{ and } x_2 \geq 1/2, \end{cases} \end{cases}$$

where δ_a denotes the Dirac distribution on point a . For this generator, the best additive model has a misclassification rate of $1/4$ whereas the Bayes classifier almost surely classifies well.

5.1.1. Data sets generators

The training sample will be drawn from the “twonorm”, “threenorm” and “ringnorm” generators. These generators introduced by Breiman in [2] have the following definitions

- *Twonorm*

Both classes have equal probabilities: $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = \frac{1}{2}$. The law of probability of $X \in \mathbb{R}^d$ conditional to $Y = 0$ is a multivariate normal distribution with unit covariance matrix and mean $m_- \triangleq (-\frac{2}{\sqrt{d}}, \dots, -\frac{2}{\sqrt{d}})$. The law of probability of X conditional to $Y = 1$ is a multivariate normal distribution with unit covariance matrix and mean $m_+ \triangleq (\frac{2}{\sqrt{d}}, \dots, \frac{2}{\sqrt{d}})$.

- *Threenorm*

Both classes have equal probabilities. The law of probability of $X \in \mathbb{R}^d$ conditional to $Y = 0$ is a multivariate normal distribution with unit covariance matrix and mean $m \triangleq (-\frac{2}{\sqrt{d}}, \frac{2}{\sqrt{d}}, -\frac{2}{\sqrt{d}}, \frac{2}{\sqrt{d}}, \dots)$. Conditional to $Y = 1$, X is drawn with equal probability from a multivariate normal distribution with unit covariance matrix and mean m_- and from a multivariate normal distribution with unit covariance matrix and mean m_+ .

- *Ringnorm*

Both classes have equal probabilities. The law of probability of $X \in \mathbb{R}^d$ conditional to $Y = 0$ is a multivariate normal distribution with unit covariance matrix and mean $\frac{m_+}{2}$. The law of probability of X conditional to $Y = 1$ is a multivariate centered normal distribution with covariance matrix four times the identity.

Denote G_μ the multivariate normal density wrt Lebesgue measure with mean μ and unit covariance matrix:

$$G_\mu(x) = \frac{e^{-\|x-\mu\|^2/2}}{(2\pi)^{d/2}}.$$

Introduce $n_1 \triangleq (0, 1, 0, 1, \dots)$, $n_2 \triangleq (1, 0, 1, 0, \dots)$ and $\text{Cst} \triangleq 8d \log 2$. The main characteristics of these generators are described in the following table.

5.1.2. Prior distribution

We are looking for the best classifying function among the functions of $\tilde{\mathcal{R}}$. In the proof of Theorem 5.1, we have noticed that $\tilde{\mathcal{R}}$ is the set of mixtures of elements in

$$\mathcal{R}' \triangleq \{0_{\mathbb{R}}\} \cup \{1_{\mathbb{R}}\} \cup \{f_{j,\tau}; j \in \{1, \dots, d\}, \tau \in \mathbb{R}\} \cup \{g_{j',\tau'}; j' \in \{1, \dots, d\}, \tau' \in \mathbb{R}\},$$

where $f_{j,\tau}(x) \triangleq \mathbb{1}_{x_j \geq \tau}$ and $g_{j',\tau'}(x) \triangleq \mathbb{1}_{x_{j'} < \tau'}$. Instead of putting the prior distribution π on \mathcal{R} , we will define it on \mathcal{R}' . For any $j \in \{1, \dots, d\}$, a probability distribution on $\{f_{j,\tau}; \tau \in \mathbb{R}\}$ or equivalently on $\{g_{j,\tau}; \tau \in \mathbb{R}\}$ can

Table 1

	Twonorm	Threenorm	Ringnorm
$\mathcal{L}(Y)$	$\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$	$\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$	$\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$
$\mathcal{L}(X/Y = 0)$	$N(m_-, \mathbf{I})$	$N(m, \mathbf{I})$	$N(\frac{m_+}{2}, \mathbf{I})$
$\mathcal{L}(X/Y = 1)$	$N(m_+, \mathbf{I})$	$\frac{G_{m_-}(x) + G_{m_+}(x)}{2} dx$	$N(0, 4\mathbf{I})$
$\mathcal{L}(X)$	$\frac{G_{m_-}(x) + G_{m_+}(x)}{2} dx$	$\frac{G_{m_-}(x) + G_{m_+}(x) + 2G_m(x)}{4} dx$	$\frac{G_{m_+/2}(x) + \frac{1}{2d}G_0(\frac{x}{2})}{2} dx$
$\mathbb{P}(Y = 1/X = x)$	$\frac{G_{m_+}}{G_{m_+} + G_{m_-}}(x) = \frac{1}{1 + e^{-2\langle x, m_+ \rangle}}$	$\frac{G_{m_-} + G_{m_+}}{G_{m_-} + G_{m_+} + 2G_m}(x)$	$\frac{G_0(\frac{x}{2})}{G_0(\frac{x}{2}) + 2^d G_{m_+/2}(x)}$
Frontier	$\langle x, m_+ \rangle = 0$	$e^{-\frac{4}{\sqrt{d}}\langle n_1, x \rangle} + e^{\frac{4}{\sqrt{d}}\langle n_2, x \rangle} = 2$	$\ 2x - m_+\ ^2 - \ x\ ^2 = \text{Cst}$

be seen as a probability distribution on the parameter $\tau \in \mathbb{R}$. We take arbitrarily the distribution π such that the law of the function $f \in \mathcal{R}'$ conditional to $f \in \{f_j, \tau; \tau \in \mathbb{R}\}$ and the law of the function $f \in \mathcal{R}'$ conditional to $f \in \{g_j, \tau; \tau \in \mathbb{R}\}$ are defined by the same law $G(d\tau)$ and such that

$$\left\{ \begin{array}{l} \pi(0_{\mathbb{R}}) = 1/4, \\ \pi(1_{\mathbb{R}}) = 1/4, \\ \pi\left(\bigcup_{\tau \in \mathbb{R}} \{f_j, \tau\}\right) = 1/4d \quad \text{for any } j \in \{1, \dots, d\}m \\ \pi\left(\bigcup_{\tau \in \mathbb{R}} \{g_j, \tau\}\right) = 1/4d \quad \text{for any } j \in \{1, \dots, d\}. \end{array} \right.$$

In our numerical examples, G will be a centered normal distribution with unit variance $N(0, 1)$:

$$G(d\tau) = \frac{e^{-\tau^2/2}}{\sqrt{2\pi}}.$$

5.2. Computation of the bound and of the classifier

Let $\mathbb{B}(\lambda_i, \beta_j, \rho)$ be equal to the RHS of inequality (4.13) in which we replace the unobservable quantity $r(\tilde{f})$ with $\inf_{\tilde{\mathcal{R}}} r$ and we take $\eta_i = \eta = 1/|I|$ and $\zeta_j = \zeta = 1/|J|$. Let d'_1 be some real and define $\hat{\rho}_{d'_1} \triangleq \pi_{-d'_1 N r(f) + (w, f(X))}$. Set

$$\left\{ \begin{array}{l} a \triangleq \frac{1}{1 - 4\lambda g(\lambda)}, \\ b \triangleq 1 - \frac{1 - 4\lambda g(\lambda)}{1 + \beta g(\beta)}, \\ c \triangleq \frac{1}{\lambda N} + \frac{1 - 4\lambda g(\lambda)}{\beta N [1 + \beta g(\beta)]}, \\ d_1 \triangleq \frac{b}{cN}, \\ d_2 \triangleq \frac{1 - b}{cN}, \\ d_3 \triangleq \frac{1}{N} \left(\frac{\log[(\eta\varepsilon)^{-1}]}{\lambda [1 - 4\lambda g(\lambda)]} + \frac{\log[(\zeta\varepsilon)^{-1}]}{2\beta [1 + \beta g(\beta)]} \right) - \frac{\inf\{r(f); f \in \tilde{\mathcal{R}}\}}{1 - 4\lambda g(\lambda)}. \end{array} \right.$$

We have $\mathbb{B}(\lambda, \beta, \hat{\rho}) = a[b\mathbb{E}_{\hat{\rho}(d\theta)} r(f_\theta) + (1 - b)r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + cK(\hat{\rho}, \pi)] + d_3$, hence

$$\begin{aligned} \mathbb{E}(\lambda, \beta, \hat{\rho}_{d'_1}) &= ac \left(d_2 \sum_{i=1}^N [Y_i - \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i)]^2 + d_1 \mathbb{E}_{\hat{\rho}_{d'_1}} \sum_{i=1}^N [Y_i - f(X_i)]^2 + K(\hat{\rho}_{d'_1}, \pi) \right) + d_3 \\ &= ac \left(d_2 \sum_{i=1}^N [Y_i - \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i)]^2 + (d_1 - d'_1) \sum_{i=1}^N (Y_i - 2Y_i \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i) + \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i)) \right. \\ &\quad \left. \sum_{i=1}^N w_i \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i) - \log \pi e^{-d'_1 N r(f) + (w, f(X))} \right) + d_3. \end{aligned} \tag{5.4}$$

We just need to compute $\mathbb{E}_{\pi} e^{-d'_1 N r(f) + (w, f)}$ and then use that for any $i \in \{1, \dots, N\}$, $\mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i) = \frac{\partial}{\partial w_i} \log \mathbb{E}_{\pi} e^{-d'_1 N r(f) + (w, f(X))}$ to calculate this bound.

For any input data $x \in \mathcal{X}$, the predicted output is

$$\mathbb{E}_{\hat{\rho}_{d'_1}} f(x) = \frac{\partial}{\partial u} \log \mathbb{E}_{\pi} e^{-d'_1 N r(f) + (w, f(X)) + u f(x)} \Big|_{u=0}.$$

The following theorem gives a simple expression of $\mathbb{E}_{\pi} e^{-d'_1 N r(f) + (w, f(X)) + u f(x)}$. We need first to introduce for any $j \in \{1, \dots, d\}$ the bijection σ_j onto $\{1, \dots, N\}$ such that

$$X_{\sigma_j(1),j} < \dots < X_{\sigma_j(N),j},$$

where $X_{i,j}$ denotes the j -th component of the i -th input vector of the training data. (We assume that the j -th component of the N input vectors are different.) By convention, put $X_{\sigma_j(0),j} \triangleq -\infty$ and $X_{\sigma_j(N+1),j} \triangleq +\infty$. Define

$$\phi(x_1, x_2) \triangleq \int_{x_1}^{x_2} G(\tau) d\tau$$

and for any $j \in \{1, \dots, d\}$ and $l \in \{0, \dots, N\}$,

$$\phi_{j,l} \triangleq \phi(X_{\sigma_j(l),j}, X_{\sigma_j(l+1),j}).$$

Introduce for any $j \in \{1, \dots, d\}$ and $x \in \mathcal{X}$, the integer $l_j(x) \in \{0, \dots, N\}$ satisfying

$$X_{\sigma_j[l_j(x)],j} \leq x < X_{\sigma_j[l_j(x)+1],j}.$$

Theorem 5.2. *We have*

$$\begin{aligned} &\mathbb{E}_{\pi} e^{-d'_1 N r(f) + (w, f(X)) + u f(x)} \\ &= \frac{1}{4} e^{-d'_1 \sum_{i=1}^N Y_i^2} + \frac{1}{4} e^{-d'_1 \sum_{i=1}^N (1-Y_i)^2 + \sum_{i=1}^N w_i + u} \\ &\quad + \frac{1}{4d} \sum_{j=1}^d \left\{ \sum_{l=0}^{l_j(x)-1} \phi_{j,l} \left[e^{-d'_1 \sum_{i=1}^l Y_{\sigma_j(i)}^2 - d'_1 \sum_{i=l+1}^N (1-Y_{\sigma_j(i)})^2 + \sum_{i=l+1}^N w_{\sigma_j(i)} + u} \right. \right. \\ &\quad \left. \left. + e^{-d'_1 \sum_{i=1}^l (1-Y_{\sigma_j(i)})^2 - d'_1 \sum_{i=l+1}^N Y_{\sigma_j(i)}^2 + \sum_{i=1}^l w_{\sigma_j(i)}} \right] \right. \\ &\quad \left. + \phi(X_{\sigma_j[l_j(x)],j}, x) \left[e^{-d'_1 \sum_{i=1}^{l_j(x)} Y_{\sigma_j(i)}^2 - d'_1 \sum_{i=l_j(x)+1}^N (1-Y_{\sigma_j(i)})^2 + \sum_{i=l_j(x)+1}^N w_{\sigma_j(i)} + u} \right. \right. \\ &\quad \left. \left. + e^{-d'_1 \sum_{i=1}^{l_j(x)} (1-Y_{\sigma_j(i)})^2 - d'_1 \sum_{i=l_j(x)+1}^N Y_{\sigma_j(i)}^2 + \sum_{i=1}^{l_j(x)} w_{\sigma_j(i)}} \right] \right. \\ &\quad \left. + \phi(x, X_{\sigma_j[l_j(x)+1],j}) \left[e^{-d'_1 \sum_{i=1}^{l_j(x)} Y_{\sigma_j(i)}^2 - d'_1 \sum_{i=l_j(x)+1}^N (1-Y_{\sigma_j(i)})^2 + \sum_{i=l_j(x)+1}^N w_{\sigma_j(i)}} \right] \right\} \end{aligned}$$

$$\begin{aligned}
 &+ e^{-d'_1 \sum_{i=1}^l (1-Y_{\sigma_j(i)})^2 - d'_1 \sum_{i=l+1}^N Y_{\sigma_j(i)}^2 + \sum_{i=1}^l w_{\sigma_j(i)} + u} \\
 &+ \sum_{l=l_j(x)+1}^N \phi_{j,l} \left[e^{-d'_1 \sum_{i=1}^l Y_{\sigma_j(i)}^2 - d'_1 \sum_{i=l+1}^N (1-Y_{\sigma_j(i)})^2 + \sum_{i=l+1}^N w_{\sigma_j(i)}} \right. \\
 &\left. + e^{-d'_1 \sum_{i=1}^l (1-Y_{\sigma_j(i)})^2 - d'_1 \sum_{i=l+1}^N Y_{\sigma_j(i)}^2 + \sum_{i=1}^l w_{\sigma_j(i)} + u} \right].
 \end{aligned}$$

As a consequence,

$$\begin{aligned}
 &\mathbb{E}_\pi e^{-d'_1 Nr(f) + \langle w, f(X) \rangle} \\
 &= \frac{1}{4} e^{-d'_1 \sum_{i=1}^N Y_i^2} + \frac{1}{4} e^{-d'_1 \sum_{i=1}^N (1-Y_i)^2 + \sum_{i=1}^N w_i} \\
 &+ \frac{1}{4d} \sum_{j=1}^d \sum_{l=0}^N \phi_{j,l} \left\{ e^{-d'_1 \sum_{i=1}^l Y_{\sigma_j(i)}^2 - d'_1 \sum_{i=l+1}^N (1-Y_{\sigma_j(i)})^2 + \sum_{i=l+1}^N w_{\sigma_j(i)}} \right. \\
 &\left. + e^{-d'_1 \sum_{i=1}^l (1-Y_{\sigma_j(i)})^2 - d'_1 \sum_{i=l+1}^N Y_{\sigma_j(i)}^2 + \sum_{i=1}^l w_{\sigma_j(i)}} \right\}.
 \end{aligned}$$

Proof. If l is the number of $X_{i,j}$, $i = 1, \dots, N$, lower than τ , we have

$$d'_1 Nr(f_{j,\tau}) + \langle w, f_{j,\tau} \rangle = d'_1 \sum_{k=1}^l Y_{\sigma_j(k)}^2 + \sum_{k=l+1}^N (1 - Y_{\sigma_j(k)})^2 + \sum_{k=l+1}^N w_{\sigma_j(k)}$$

and

$$d'_1 Nr(g_{j,\tau}) + \langle w, g_{j,\tau} \rangle = d'_1 \sum_{k=1}^l (1 - Y_{\sigma_j(k)})^2 + \sum_{k=l+1}^N Y_{\sigma_j(k)}^2 + \sum_{k=1}^l w_{\sigma_j(k)}.$$

The calculus is then straightforward. \square

Let N_0 (respectively N_1) be the number of class 0 data (respectively class 1 data) in the training sample. We have trivially $N_0 + N_1 = N$. Introduce $c_0^w \triangleq e^{-d'_1 N_1}$, $c_1^w \triangleq e^{-d'_1 N_0 + \sum_{i=1}^N w_i}$, for any $j \in \{1, \dots, d\}$ and $l \in \{0, \dots, N\}$,

$$\begin{cases}
 a_{j,l}^w \triangleq \phi_{j,l} e^{-d'_1 \sum_{i=1}^l Y_{\sigma_j(i)} - d'_1 \sum_{i=l+1}^N (1-Y_{\sigma_j(i)}) + \sum_{i=l+1}^N w_{\sigma_j(i)}} \\
 = \phi_{j,l} e^{-d'_1 (N_0 - l + 2 \sum_{i=1}^l Y_{\sigma_j(i)}) + \sum_{i=l+1}^N w_{\sigma_j(i)}}, \\
 b_{j,l}^w \triangleq \phi_{j,l} e^{-d'_1 \sum_{i=1}^l (1-Y_{\sigma_j(i)}) - d'_1 \sum_{i=l+1}^N Y_{\sigma_j(i)} + \sum_{i=1}^l w_{\sigma_j(i)}} \\
 = \phi_{j,l} e^{-d'_1 (N_1 + l - 2 \sum_{i=1}^l Y_{\sigma_j(i)}) + \sum_{i=1}^l w_{\sigma_j(i)}}
 \end{cases}$$

for any $x \in \mathcal{X}$,

$$c_{j,l}^w(x) \triangleq \begin{cases} a_{j,l}^w & \text{when } l < l_j(x), \\ \frac{\phi(X_{\sigma_j(l),j}, x_j) a_{j,l}^w + \phi(x_j, X_{\sigma_j(l+1),j}) b_{j,l}^w}{\phi_{j,l}} & \text{when } l = l_j(x), \\ b_{j,l}^w & \text{when } l > l_j(x), \end{cases}$$

and for any $x, y \in \mathcal{X}$,

$$c_{j,l}^w(x, y) \triangleq \begin{cases} a_{j,l}^w & \text{when } l < l_j(x) \wedge l_j(y), \\ \frac{\phi(X_{\sigma_j(l),j}, x_j \wedge y_j)}{\phi_{j,l}} a_{j,l}^w & \text{when } l = l_j(x) \wedge l_j(y), \\ \frac{\phi(x_j \vee y_j, X_{\sigma_j(l+1),j})}{\phi_{j,l}} b_{j,l}^w & \text{when } l = l_j(x) \vee l_j(y), \\ b_{j,l}^w & \text{when } l > l_j(x) \vee l_j(y), \end{cases}$$

with the following convention when $l_j(x) \vee l_j(y) = l_j(x) \wedge l_j(y)$:

$$c_{j,l_j(x) \vee l_j(y)}^w(x, y) \triangleq \frac{\phi(X_{\sigma_j(l),j}, x_j \wedge y_j)}{\phi_{j,l}} a_{j,l}^w + \frac{\phi(x_j \vee y_j, X_{\sigma_j(l+1),j})}{\phi_{j,l}} b_{j,l}^w.$$

Then

Corollary 5.3. For any constant d'_1 , we have

$$\mathbb{E}_\pi e^{-d'_1 N r(f) + \langle w, f(X) \rangle} = \frac{1}{4d} \left(dc_0^w + dc_1^w + \sum_{j=1}^d \sum_{l=0}^N (a_{j,l}^w + b_{j,l}^w) \right).$$

Let $\hat{\rho}_{d'_1} \triangleq \pi_{-d'_1 N r(f) + \langle w, f(X) \rangle}$. We have

$$\begin{cases} \mathbb{E}_{\hat{\rho}_{d'_1}} f(x) = \frac{dc_1^w + \sum_{j=1}^d \sum_{l=0}^N c_{j,l}^w(x)}{dc_0^w + dc_1^w + \sum_{j=1}^d \sum_{l=0}^N (a_{j,l}^w + b_{j,l}^w)}, \\ \mathbb{E}_{\hat{\rho}_{d'_1}} [f(x)f(y)] = \frac{dc_1^w + \sum_{j=1}^d \sum_{l=0}^N c_{j,l}^w(x, y)}{dc_0^w + dc_1^w + \sum_{j=1}^d \sum_{l=0}^N (a_{j,l}^w + b_{j,l}^w)}. \end{cases}$$

Proof. It comes from Theorem 5.2 and from

$$\begin{cases} \mathbb{E}_{\hat{\rho}_{d'_1}} f(x) = \frac{\partial}{\partial u} \log \mathbb{E}_\pi e^{-d'_1 N r(f) + \langle w, f(X) \rangle + u f(x)} \Big|_{u=0}, \\ \mathbb{Cov}_{\hat{\rho}_{d'_1}}(f(x), f(y)) = \frac{\partial^2}{\partial u \partial v} \log \mathbb{E}_\pi e^{-d'_1 N r(f) + \langle w, f(X) \rangle + u f(x) + v f(y)} \Big|_{u=0, v=0}. \end{cases} \quad \square$$

Remark 5.2. To compute $\mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i)$, we may note that $l_j(X_i) = \sigma_j^{-1}(i)$. Besides, there is a simple link between $a_{j,l}^w$ and $b_{j,l}^w$ since for any $j \in \{1, \dots, d\}$ and $l \in \{0, \dots, N\}$, we have

$$a_{j,l}^w b_{j,l}^w = \phi_{j,l}^2 c_0^w c_1^w.$$

Computation of the constant d_3 . We have

$$d_3 \triangleq \frac{1}{N} \left(\frac{\log[(\eta\varepsilon)^{-1}]}{\lambda[1 - 4\lambda g(\lambda)]} + \frac{\log[(\zeta\varepsilon)^{-1}]}{2\beta[1 + \beta g(\beta)]} \right) - \frac{\inf\{r(f); f \in \tilde{\mathcal{R}}\}}{1 - 4\lambda g(\lambda)}.$$

To compute the constant d_3 , we need to calculate $\inf\{r(f); f \in \tilde{\mathcal{R}}\}$. From Theorem 5.1, determining $\inf\{r(f); f \in \tilde{\mathcal{R}}\}$ is equivalent to solving the following convex quadratic (QP) problem

$$\min_{u_{i,j}, v_{i,j}} \sum_{i=1}^N \left(\sum_{j=1}^d (u_{i,j} + v_{i,j}) - Y_i \right)^2$$

under the linear constraints

$$\begin{cases} 0 \leq u_{\sigma_j(1),j} \leq \dots \leq u_{\sigma_j(N),j} & \text{for any } j \in \{1, \dots, d\}, \\ v_{\sigma_j(1),j} \geq \dots \geq v_{\sigma_j(N),j} \geq 0 & \text{for any } j \in \{1, \dots, d\}, \\ \sum_{j=1}^d (u_{\sigma_j(N),j} + v_{\sigma_j(1),j}) \leq 1. \end{cases}$$

The dimension of the QP-problem is dN and the number of linear constraints is $2dN + 1$. This is numerically untractable (since $dN \gg 1000$). Therefore, we can either weaken our bound by neglecting the term $-\frac{\inf\{r(f); f \in \tilde{\mathcal{R}}\}}{1-4\lambda g(\lambda)}$ or approximate this term by $-\frac{\inf\{r(\mathbb{E}_{\rho(d\theta)} f_\theta) + \delta K(\rho, \pi); \rho \in \mathcal{M}_+^1(\Theta)\}}{1-4\lambda g(\lambda)}$ for sufficiently small δ (since this last optimization problem has been proven to be tractable).

5.3. Experiments

5.3.1. Our algorithm: KL-Boost

In KL-Boost algorithm, we cross-validate on the Kullback–Leibler regularization parameter and neglect the variance term. For any couple (λ, β) , the vector w_{opt} in the procedure derived from Corollary 4.3 is solution of the minimization problem

$$\min_{w \in \mathbb{R}^N} \frac{1}{2} r(\mathbb{E}_{\pi^w(d\theta)} f_\theta) + \alpha' \mathbb{E}_{\mathbb{P}} \text{Var}_{\pi^w(d\theta)} f_\theta + \alpha K(\pi^w, \pi),$$

for $\alpha = 2c$ and $\alpha' = 2b$. The variance term in this minimization problem is useful only when the best regression function \tilde{f} in the model $\tilde{\mathcal{R}}$ is in (or very close to) the initial model \mathcal{R} . Generally, this is not the case in applications. So let us forget the variance term ($\alpha' = 0$). Finally, we look for the adequate parameter α by using cross-validation. After having chosen the parameter, the algorithm is calibrated on all the training set for this regularization parameter.

According to Theorem 4.10, the quantity $\mathbb{B}(\lambda, \beta, \hat{\rho}_0)$ (see (5.4)) gives a risk guarantee. From Section 4.2.2, the final aggregating distribution is $\hat{\rho} = \pi_{(w,f)}$, where the vector w satisfies $w_i = \frac{1}{\alpha N} [Y_i - \mathbb{E}_{\pi_{(w,f)}} f(X_i)]$ for any $i \in \{1, \dots, N\}$.

In our experiments, we have taken

- maximum number of iterations used to optimize the bound $m = 300$,
- absolute error accepted when minimizing the bound $err = 0.0001$,
- number of blocks used in the cross-validation = 2,
- set of values of the regularization parameter α :

$$\{0.0002, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.2\}.$$

Note that this set is inspired from the bound and takes into account the fact that the bound is conservative (i.e. tends to regularize too much). Strictly speaking, it should depend on N .

In our simulations, the value 0.0002 of the parameter α leads to a procedure close to the empirical risk minimizer on the set of mixtures $\tilde{\mathcal{R}}$ and thus is used to approximate d_3 .

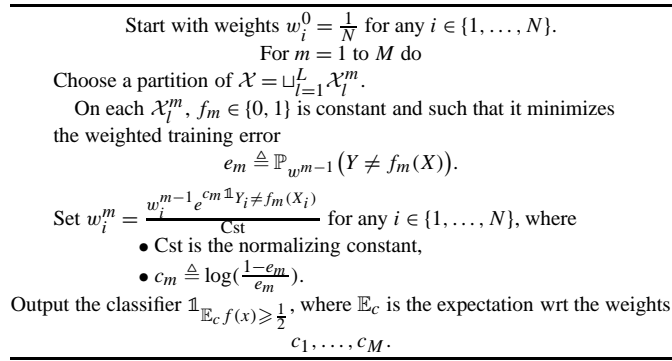


Fig. 1. “Discrete” AdaBoost using domain-partitioning functions (Freund and Schapire [4]).

5.3.2. AdaBoost using domain-partitioning functions [4,5,11]

The first boosting methods train functions on weighted versions of the training sample, giving higher weights to cases that are currently misclassified. In AdaBoost (Freund and Schapire [4]), the functions trained are classifiers, that is to say functions taking their values in $\{0, 1\}$ in the two-class classification setting. We describe the original algorithm in Fig. 1 where \mathbb{E}_{w^m} denotes the empirical expectation wrt the weights w_1^m, \dots, w_N^m .

The weights c_m are positive since by construction of the classifier f_m , we have $e_m \leq \frac{1}{2}$. The choice of the partition can be done in several different ways. In standard boosting methods, one can choose the split which causes the greatest drop in the value of a criterion to be specified. This greedy procedure is sometimes replaced by randomizing methods. For instance, one can draw a set of splits and choose the split among this set which minimizes the criterion. Another way of randomizing is to draw a subset of the training sample and then take the split which minimizes the criterion on this subset.

Introduce $F_m \triangleq \sum_{j=1}^m c_j f_j$. Define $\bar{Y} \triangleq -1 + 2Y \in \{-1, 1\}$, $\bar{f} \triangleq -1 + 2f$ and $\bar{F}_m \triangleq -1 + 2F_m$. Then we have: $\bar{F}_m = \sum_{j=1}^m c_j \bar{f}_j$. Introduce $f_{m,l} \in \{0, 1\}$ such that

$$f_m(x) = \sum_{l=1}^L f_{m,l} \mathbb{1}_{x \in \mathcal{X}_l^m},$$

where $\{\mathcal{X}_l^m\}_{1, \dots, L}$ is the chosen partition during the m -th step of the procedure (described in Fig. 1).

Lemma 5.4. *Once the partition has been chosen, the positive real c_m and the family $f_{m,l} \in \{0, 1\}$, $l = 1, \dots, L$, are chosen in order to minimize $\mathbb{E}_{\mathbb{P}}(e^{-\frac{1}{2} \bar{Y} \bar{F}_m(X)})$.*

The link between AdaBoost and this criterion has been introduced by Friedman, Hastie and Tibshirani [5].

Proof. By induction on m , one may easily prove that for any $m \in \{0, \dots, M\}$,

$$\mathbb{P}_{w^m} = \frac{e^{-\frac{1}{2} \bar{Y} \bar{F}_m(X)}}{\mathbb{E}_{\mathbb{P}}(e^{-\frac{1}{2} \bar{Y} \bar{F}_m(X)})} \cdot \bar{\mathbb{P}}.$$

Then we have

$$\frac{\mathbb{E}_{\bar{\mathbb{P}}}(e^{-\frac{1}{2} \bar{Y} \bar{F}_m(X)})}{\mathbb{E}_{\bar{\mathbb{P}}}(e^{-\frac{1}{2} \bar{Y} \bar{F}_{m-1}(X)})} = \mathbb{E}_{w^{m-1}}(e^{-\frac{1}{2} \bar{Y} c_m \bar{f}_m(X)})$$

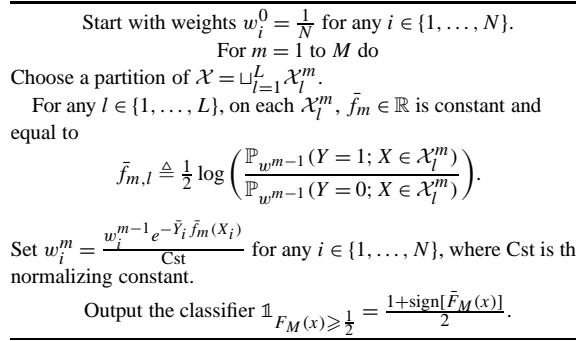


Fig. 2. “Real” AdaBoost using domain-partitioning functions (Schapire and Singer [11]).

$$\begin{aligned} &= \sum_{l=1}^L \mathbb{P}(X \in \mathcal{X}_l^m) \mathbb{E}_{w^{m-1}}(e^{-\frac{1}{2} \tilde{Y} c_m \tilde{f}_{m,l}} / X \in \mathcal{X}_l^m) \\ &= \sum_{l=1}^L (\mathbb{P}_{w^{m-1}}(Y = 1; X \in \mathcal{X}_l^m) e^{-\frac{1}{2} c_m \tilde{f}_{m,l}} + \mathbb{P}_{w^{m-1}}(Y = 0; X \in \mathcal{X}_l^m) e^{\frac{1}{2} c_m \tilde{f}_{m,l}}). \end{aligned}$$

For any $l \in \{1, \dots, L\}$ and for fixed $c_m \geq 0$, the l -th term of this last sum is minimized for $\tilde{f}_{m,l}$ equal to the most w^{m-1} -popular class on \mathcal{X}_l^m , hence

$$f_{m,l} = \operatorname{argmax}_{u \in \{0,1\}} \mathbb{P}_{w^{m-1}}(Y = u / X \in \mathcal{X}_l^m) = \operatorname{argmin}_{u \in \{0,1\}} \mathbb{E}_{w^{m-1}} \mathbb{1}_{\{Y \neq u; X \in \mathcal{X}_l^m\}}.$$

Since we have

$$\mathbb{E}_{w^{m-1}}(e^{-\frac{1}{2} \tilde{Y} c_m \tilde{f}_m(X)}) = e^{\frac{1}{2} c_m} \mathbb{P}_{w^{m-1}}[Y \neq f_m(X)] + e^{-\frac{1}{2} c_m} \mathbb{P}_{w^{m-1}}[Y = f_m(X)],$$

the optimal c_m is

$$c_m = \log \left(\frac{1 - e_m}{e_m} \right),$$

where $e_m = \mathbb{P}_{w^{m-1}}(Y \neq f_m(X))$. \square

As Friedman, Hastie and Tibshirani pointed out, this algorithm produces adaptive Newton updates for minimizing $[\tilde{F} \mapsto \mathbb{E}_{\tilde{\mathbb{P}}} e^{-\tilde{Y} \tilde{F}(X)}]$, which are stage-wise contributions to an additive logistic model.

In [11], Schapire and Singer suggests to use real-valued functions rather than classifiers (which, by definition, take their values in $\{-1, 1\}$). This leads to the algorithm described in Fig. 2 which outperforms the “discrete” AdaBoost when L is small (especially when we use stumps: $L = 2$).

In this procedure, at the m -th step, the family $\tilde{f}_{m,l}, l = 1, \dots, L$, is chosen such that it minimizes

$$\mathbb{E}_{\tilde{\mathbb{P}}} e^{-\tilde{Y} \tilde{F}_m(X)} = \mathbb{E}_{\tilde{\mathbb{P}}} e^{-\tilde{Y} \tilde{F}_{m-1}(X)} \mathbb{E}_{w^{m-1}} e^{-\tilde{Y} \tilde{f}_m(X)}.$$

Besides, we have

$$\begin{aligned} \mathbb{E}_{w^{m-1}} e^{-\tilde{Y} \tilde{f}_m(X)} &= \sum_{l=1}^L \mathbb{P}_{w^{m-1}}(Y = 0; X \in \mathcal{X}_l^m) e^{\tilde{f}_{m,l}} + \mathbb{P}_{w^{m-1}}(Y = 1; X \in \mathcal{X}_l^m) e^{-\tilde{f}_{m,l}} \\ &= 2 \sum_{l=1}^L \sqrt{\mathbb{P}_{w^{m-1}}(Y = 0; X \in \mathcal{X}_l^m) \mathbb{P}_{w^{m-1}}(Y = 1; X \in \mathcal{X}_l^m)}. \end{aligned}$$

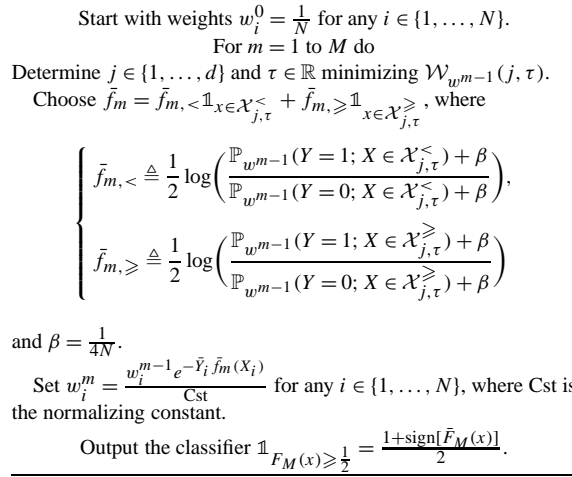


Fig. 3. “Real” AdaBoost using stumps (Schapire and Singer [11]).

Therefore, as Schapire and Singer stresses, a natural criterion to partition the input space \mathcal{X} is to minimize this last sum. This is more coherent to use it instead of the Gini index or an entropy function since it aims, as the rest of the procedure, to minimize the functional $[\bar{F} \mapsto \mathbb{E}_{\mathbb{P}} e^{-\bar{Y} \bar{F}(X)}]$.

It may happen that one of the predictions $\tilde{f}_{m,l}$ is very large or even infinite, which leads to numerical problems. To limit the magnitude of the predictions, Schapire and Singer define

$$\tilde{f}_{m,l} \triangleq \frac{1}{2} \log \left(\frac{\mathbb{P}_{w^{m-1}}(Y = 1; X \in \mathcal{X}_l^m) + \beta}{\mathbb{P}_{w^{m-1}}(Y = 0; X \in \mathcal{X}_l^m) + \beta} \right),$$

where β is a small positive real arbitrarily defined as $\beta = \frac{1}{4N}$.

In our numerical examples, we are interested in decision stumps $x \mapsto \alpha_0 \mathbb{1}_{x_j < \tau} + \alpha_1 \mathbb{1}_{x_j \geq \tau}$ which partition \mathcal{X} into $\mathcal{X}_{j,\tau}^{<} \triangleq \{x_j < \tau\}$ and $\mathcal{X}_{j,\tau}^{\geq} \triangleq \{x_j \geq \tau\}$. For any $j \in \{1, \dots, d\}$ and $\tau \in \mathbb{R}$, introduce

$$\begin{aligned} \mathcal{W}_w(j, \tau) \triangleq & \sqrt{\mathbb{P}_w(Y = 0; x \in \mathcal{X}_{j,\tau}^{<}) \mathbb{P}_w(Y = 1; x \in \mathcal{X}_{j,\tau}^{<})} \\ & + \sqrt{\mathbb{P}_w(Y = 0; x \in \mathcal{X}_{j,\tau}^{\geq}) \mathbb{P}_w(Y = 1; x \in \mathcal{X}_{j,\tau}^{\geq})}. \end{aligned}$$

The AdaBoost used in our numerical examples is described in Fig. 3. After having tested different values for the number of stumps aggregated, we have taken $M = 100$.

Remark 5.3. The set of (j, τ) minimizing $\mathcal{W}_{w^{m-1}}(j, \tau)$ has the following form

$$\bigcup_{j=1}^d \left(\{j\} \times \bigcup_{k=1}^{k_j} [a_j; b_j] \right),$$

where a_j and b_j belong to $\{-\infty, X_{1,j}, \dots, X_{N,j}, +\infty\}$ and k_1, \dots, k_d are positive integers. We take arbitrarily the smallest j to make the split (i.e. the smallest integer j such that $k_j > 0$). Then τ is chosen in $]X_{\sigma_j(l),j}; X_{\sigma_j(l+1),j}]$, where l is the smallest integer such that $(j, X_{\sigma_j(l+1),j})$ minimizes $\mathcal{W}_{w^{m-1}}(j, \tau)$. We take arbitrarily

$$\tau = \frac{X_{\sigma_j(l),j} + X_{\sigma_j(l+1),j}}{2} \in \bar{\mathbb{R}}.$$

We use the convention $\mathcal{X}_{j,-\infty}^{<} \triangleq \emptyset$, $\mathcal{X}_{j,-\infty}^{\geq} \triangleq \mathbb{R}$, $\mathcal{X}_{j,+\infty}^{<} \triangleq \mathbb{R}$ and $\mathcal{X}_{j,+\infty}^{\geq} \triangleq \emptyset$. Hence $\tau = +\infty$ and $\tau = -\infty$ give the same partition and consequently, the same function f_m .

Remark 5.4. Since $\mathbb{E}e^{-\tilde{Y}\tilde{F}(X)}$ is minimized for $\tilde{F}(x) = \frac{1}{2} \log\left(\frac{\mathbb{P}(Y=1/X=x)}{\mathbb{P}(Y=0/X=x)}\right)$ and since the AdaBoost procedure aims to minimize the functional $[\tilde{F} \mapsto \mathbb{E}_{\mathbb{P}}e^{-\tilde{Y}\tilde{F}(X)}]$, the quantity $\frac{1}{1+e^{-2\tilde{F}_M(x)}}$ is an estimate of the regression function $\mathbb{E}(Y/X = x) = \mathbb{P}(Y = 1/X = x)$.

Remark 5.5. The “real” AdaBoost algorithm using stumps as a weak learner leads to a classifier which belongs to

$$\text{sign}(\tilde{\mathcal{R}}) \triangleq \{g : \mathcal{X} \rightarrow \{-1; 1\} : \text{there exists } f \in \tilde{\mathcal{R}} \text{ such that } g = \text{sign } f\}.$$

So it is not associated with a larger model than the one used in KL-Boost. “Discrete” AdaBoost using stumps has trivially this property (final classifier belongs to $\text{sign}(\tilde{\mathcal{R}})$) since the estimates f_m aggregated belongs to \mathcal{R}' . To prove the property for the “real” Adaboost algorithm, we just need to notice that

$$\mathbb{1}_{F_M(x) \geq \frac{1}{2}} = \mathbb{1}_{\mathbb{E}_{\mu} f'_m(x) \geq \frac{1}{2}},$$

where

$$f'_m \triangleq \frac{1 + \tilde{f}'_m}{2}, \quad \tilde{f}'_m \triangleq \frac{\tilde{f}_m}{\max\{|f_{m,k}|; k \in \{<, \geq\}, m \in \{1, \dots, M\}\}}$$

and μ is the uniform distribution on $\{1, \dots, M\}$, and to check that f'_m belongs to \mathcal{R} (see equality (5.1) for the definition of \mathcal{R}).

However, in KL-Boost, the additive model is put on the conditional expectation rather than the logit transformation

$$\frac{1}{2} \log\left(\frac{\mathbb{P}(Y = 1/X)}{\mathbb{P}(Y = 0/X)}\right) = \frac{1}{2} \log\left(\frac{\mathbb{E}(Y/X)}{1 - \mathbb{E}(Y/X)}\right).$$

Therefore, as algorithms estimating the conditional expectation $\mathbb{E}(Y/X)$, AdaBoost and KL-Boost are associated with very different models.

5.4. Numerical results and comments

In our experiments, we compare KL-Boost with Adaboost. It appears that KL-Boost is more efficient than AdaBoost on noisy data, and the results are more balanced in low noise frameworks. For the lines of the tables in which the training sample is of size 100 or 500 and in which the dimension is 3, we generated 100 training sets. For the other lines, 25 training sets have been simulated. The errors which appear in Tables 2 to 10 are averaged errors over the 100 or 25 simulations. Below, in brackets, we put twice the associated standard deviations over the square root of the number of simulations to give the usual approximations of the confidence intervals. In the numerical simulations, the input dimension was either 3 or 6 or 20. In the tables, the parameter 3, 6 (respectively 10, 20) in the “dimension” column means that the input is 6-dimensional (respectively 10-dimensional) but the output only depends on 3 (respectively 10) components of the input (the other 3 (respectively 10) components of the input being generated by a centered normal distribution with unit variance independently of the output).

For ringnorm generators without noise, AdaBoost is definitely more efficient than KL-Boost. We have to bear in mind that even if the underlying classification model is the same for all the algorithms (that is to say the set $\text{sign}(-1 + 2\tilde{\mathcal{R}})$ where $\tilde{\mathcal{R}}$ is described in Theorem 5.1 and when the classes are $\{-1; +1\}$), the regression models are different in Adaboost and KL-Boost procedures. Let us denote $\tilde{\mathcal{R}}_{\text{ada}}$ the regression function model associated with Adaboost. On the one hand, Adaboost will tend to classify as $C_{\text{ada}} \triangleq \text{sign}(-1 + 2\tilde{f}_{\text{ada}})$, where

$$\tilde{f}_{\text{ada}} \triangleq \underset{f \in \tilde{\mathcal{R}}_{\text{ada}}}{\text{argmin}} R(f)$$

Table 2

Comparison between Adaboost and KL-Boost: classification and quadratic errors for different twonorm generators

N	Dimension	Classif. gen. errors		Classif. emp. errors		L ² gen. errors		L ² emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3	5, 1%	3,8%	0, 0%	2, 0%	0,050	0, 085	0, 000	0, 077
		(±0, 3%)	(±0, 3%)	(±0, 0%)	(±0, 3%)	(±0, 003)	(±0, 008)	(±0, 000)	(±0, 010)
500	3	3, 2%	2,9%	0, 0%	2, 6%	0,029	0, 100	0, 000	0, 099
		(±0, 1%)	(±0, 1%)	(±0, 0%)	(±0, 1%)	(±0, 001)	(±0, 010)	(±0, 000)	(±0, 010)
2000	3	2, 8%	2,7%	1, 3%	2, 7%	0,023	0, 131	0, 009	0, 131
		(±0, 2%)	(±0, 1%)	(±0, 1%)	(±0, 1%)	(±0, 001)	(±0, 018)	(±0, 001)	(±0, 018)
100	6	5, 4%	4,2%	0, 0%	2, 6%	0,052	0, 106	0, 000	0, 095
		(±0, 3%)	(±0, 5%)	(±0, 0%)	(±0, 6%)	(±0, 004)	(±0, 014)	(±0, 000)	(±0, 016)
500	6	3, 6%	3,0%	0, 0%	2, 6%	0,032	0, 129	0, 000	0, 127
		(±0, 2%)	(±0, 1%)	(±0, 0%)	(±0, 3%)	(±0, 001)	(±0, 016)	(±0, 000)	(±0, 016)
2000	6	2, 9%	2,8%	0, 7%	2, 8%	0,024	0, 156	0, 005	0, 156
		(±0, 1%)	(±0, 1%)	(±0, 1%)	(±0, 1%)	(±0, 001)	(±0, 015)	(±0, 001)	(±0, 015)
100	20	7, 8%	7,3%	0, 0%	2, 4%	0,073	0, 152	0, 000	0, 129
		(±0, 6%)	(±1, 1%)	(±0, 0%)	(±0, 6%)	(±0, 005)	(±0, 008)	(±0, 000)	(±0, 011)
500	20	4, 5%	3,7%	0, 0%	3, 0%	0,041	0, 160	0, 000	0, 156
		(±0, 2%)	(±0, 2%)	(±0, 0%)	(±0, 3%)	(±0, 001)	(±0, 008)	(±0, 000)	(±0, 008)
2000	20	3, 6%	3,1%	0, 1%	3, 0%	0,030	0, 167	0, 002	0, 167
		(±0, 1%)	(±0, 1%)	(±0, 1%)	(±0, 2%)	(±0, 001)	(±0, 010)	(±0, 000)	(±0, 010)

Table 3

Comparison between Adaboost and KL-Boost: classification and quadratic errors for twonorm generators with superfluous features

N	Dimension	Classif. gen. errors		Classif. emp. errors		L ² gen. errors		L ² emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3,6	12, 4%	10,8%	0, 0%	7, 2%	0,119	0, 123	0, 000	0, 108
		(±0, 7%)	(±0, 8%)	(±0, 0%)	(±1, 2%)	(±0, 006)	(±0, 012)	(±0, 000)	(±0, 015)
500	3,6	10, 4%	9,5%	1, 0%	8, 4%	0,086	0, 130	0, 009	0, 127
		(±0, 3%)	(±0, 2%)	(±0, 2%)	(±0, 4%)	(±0, 002)	(±0, 018)	(±0, 002)	(±0, 018)
2000	3,6	9,0%	9, 1%	6, 3%	8, 7%	0,069	0, 168	0, 044	0, 168
		(±0, 2%)	(±0, 2%)	(±0, 2%)	(±0, 2%)	(±0, 001)	(±0, 020)	(±0, 001)	(±0, 020)
100	10,20	15, 2%	14,7%	0, 0%	6, 7%	0,144	0, 170	0, 000	0, 143
		(±0, 8%)	(±1, 3%)	(±0, 0%)	(±1, 1%)	(±0, 008)	(±0, 011)	(±0, 000)	(±0, 017)
500	10,20	11, 5%	10,5%	0, 0%	8, 5%	0,099	0, 169	0, 000	0, 165
		(±0, 3%)	(±0, 2%)	(±0, 0%)	(±0, 5%)	(±0, 002)	(±0, 009)	(±0, 000)	(±0, 010)
2000	10,20	10, 1%	9,3%	4, 9%	8, 9%	0,079	0, 183	0, 034	0, 180
		(±0, 3%)	(±0, 2%)	(±0, 3%)	(±0, 2%)	(±0, 001)	(±0, 011)	(±0, 002)	(±0, 010)

and $R(f)$ still denotes the quadratic risk. On the other hand, KL-Boost algorithm will tend to classify as $C_{KL} \triangleq \text{sign}(-1 + 2\tilde{f})$, where

$$\tilde{f} \triangleq \underset{f \in \tilde{\mathcal{R}}}{\text{argmin}} R(f).$$

Usually, the function \tilde{f} is different from \tilde{f}_{ada} . Therefore the classifiers C_{ada} and C_{KL} are in general different and the type of the classification task (which is determined by the unknown probability distribution \mathbb{P}) will decide which of these two classifiers outperforms the other. The performance of the algorithms will utterly come from the performance of these classifiers.

Table 4

Comparison between Adaboost and KL-Boost: classification and quadratic errors for different threenorm generators

N	Dimension	Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3	16, 5%	16, 5%	0, 0%	14, 4%	0,159	0, 165	0, 001	0, 146
		($\pm 0, 7\%$)	($\pm 0, 8\%$)	($\pm 0, 0\%$)	($\pm 0, 8\%$)	($\pm 0, 004$)	($\pm 0, 003$)	($\pm 0, 000$)	($\pm 0, 005$)
500	3	15, 2%	13, 2%	8, 6%	14, 2%	0,113	0, 156	0, 058	0, 152
		($\pm 0, 3\%$)	($\pm 0, 3\%$)	($\pm 0, 3\%$)	($\pm 0, 4\%$)	($\pm 0, 001$)	($\pm 0, 002$)	($\pm 0, 002$)	($\pm 0, 002$)
2000	3	14, 9%	12, 6%	13, 1%	14, 4%	0,099	0, 153	0, 091	0, 152
		($\pm 0, 4\%$)	($\pm 0, 1\%$)	($\pm 0, 4\%$)	($\pm 0, 4\%$)	($\pm 0, 001$)	($\pm 0, 002$)	($\pm 0, 002$)	($\pm 0, 002$)
100	6	20, 6%	27, 5%	0, 0%	16, 1%	0, 233	0,187	0, 000	0, 160
		($\pm 1, 6\%$)	($\pm 1, 2\%$)	($\pm 0, 0\%$)	($\pm 1, 8\%$)	($\pm 0, 009$)	($\pm 0, 006$)	($\pm 0, 000$)	($\pm 0, 013$)
500	6	18, 2%	23, 9%	8, 3%	19, 0%	0,178	0, 180	0, 056	0, 177
		($\pm 0, 6\%$)	($\pm 0, 6\%$)	($\pm 0, 6\%$)	($\pm 0, 8\%$)	($\pm 0, 003$)	($\pm 0, 004$)	($\pm 0, 004$)	($\pm 0, 005$)
2000	6	18, 0%	23, 6%	14, 3%	19, 2%	0,156	0, 173	0, 099	0, 172
		($\pm 0, 4\%$)	($\pm 0, 4\%$)	($\pm 0, 4\%$)	($\pm 0, 4\%$)	($\pm 0, 002$)	($\pm 0, 002$)	($\pm 0, 002$)	($\pm 0, 003$)
100	20	28, 1%	31, 4%	0, 0%	13, 5%	0, 273	0,209	0, 009	0, 153
		($\pm 1, 2\%$)	($\pm 1, 0\%$)	($\pm 0, 0\%$)	($\pm 1, 6\%$)	($\pm 0, 008$)	($\pm 0, 003$)	($\pm 0, 013$)	($\pm 0, 010$)
500	20	24, 9%	26, 5%	4, 4%	21, 3%	0, 209	0,208	0, 034	0, 200
		($\pm 0, 6\%$)	($\pm 0, 8\%$)	($\pm 0, 6\%$)	($\pm 0, 8\%$)	($\pm 0, 003$)	($\pm 0, 004$)	($\pm 0, 003$)	($\pm 0, 006$)
2000	20	23, 1%	24, 3%	15, 7%	22, 0%	0,170	0, 202	0, 107	0, 200
		($\pm 0, 3\%$)	($\pm 0, 4\%$)	($\pm 0, 3\%$)	($\pm 0, 4\%$)	($\pm 0, 002$)	($\pm 0, 002$)	($\pm 0, 002$)	($\pm 0, 003$)

Table 5

Comparison between Adaboost and KL-Boost: classification and quadratic errors for threenorm generators with superfluous features

N	Dimension	Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3,6	30, 1%	27, 4%	0, 0%	21, 1%	0, 268	0,205	0, 001	0, 171
		($\pm 1, 2\%$)	($\pm 1, 3\%$)	($\pm 0, 1\%$)	($\pm 2, 1\%$)	($\pm 0, 009$)	($\pm 0, 005$)	($\pm 0, 001$)	($\pm 0, 011$)
500	3,6	27, 4%	23, 1%	14, 1%	24, 1%	0, 192	0,191	0, 095	0, 183
		($\pm 0, 6\%$)	($\pm 0, 6\%$)	($\pm 1, 0\%$)	($\pm 1, 2\%$)	($\pm 0, 003$)	($\pm 0, 004$)	($\pm 0, 005$)	($\pm 0, 005$)
2000	3,6	25, 0%	21, 0%	20, 8%	22, 9%	0,161	0, 185	0, 142	0, 183
		($\pm 0, 4\%$)	($\pm 0, 3\%$)	($\pm 0, 3\%$)	($\pm 0, 4\%$)	($\pm 0, 001$)	($\pm 0, 002$)	($\pm 0, 002$)	($\pm 0, 001$)
100	10,20	36, 1%	35, 6%	0, 0%	20, 4%	0, 333	0,228	0, 000	0, 180
		($\pm 1, 4\%$)	($\pm 2, 1\%$)	($\pm 0, 0\%$)	($\pm 2, 9\%$)	($\pm 0, 010$)	($\pm 0, 004$)	($\pm 0, 000$)	($\pm 0, 013$)
500	10,20	32, 5%	29, 1%	8, 2%	25, 7%	0, 241	0,215	0, 061	0, 203
		($\pm 0, 7\%$)	($\pm 0, 6\%$)	($\pm 0, 6\%$)	($\pm 0, 8\%$)	($\pm 0, 003$)	($\pm 0, 004$)	($\pm 0, 004$)	($\pm 0, 006$)
2000	10,20	30, 1%	27, 2%	21, 3%	27, 2%	0,196	0, 214	0, 142	0, 210
		($\pm 0, 3\%$)	($\pm 0, 3\%$)	($\pm 0, 3\%$)	($\pm 0, 4\%$)	($\pm 0, 002$)	($\pm 0, 005$)	($\pm 0, 002$)	($\pm 0, 006$)

Using big training sets, one gets an idea of the efficiency of these classifiers. Numerical results (for training sets of size $N = 2000$) tend to say that the classifier C_{ada} is “closer” to the Bayes rule than C_{KL} for non-noisy ringnorm generators. The opposite occurs for non-noisy twonorm generators. In the other cases, the situation is balanced but globally in favor of C_{KL} .

To cross-validate a parameter of the algorithm using the classification error plays a key role for the twonorm generators since in this context, KL-Boost works better than AdaBoost whereas its least square generalization errors is worse than AdaBoost ones and increases when the training set size N increases.

In KL-Boost, the theoretical bound given by Theorem 4.10 is still far away from the real value. When the number of training points is lower than 500, it often gets irrelevant values, i.e. values bigger than $1/4$. This is not surprising

Table 6
Comparison between Adaboost and KL-Boost: classification and quadratic errors for different ringnorm generators

N	Dimension	Classif. gen. errors		Classif. emp. errors		L ² gen. errors		L ² emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3	26,7% (±0,5%)	30,4% (±0,5%)	0,3% (±0,1%)	23,9% (±0,9%)	0,232 (±0,003)	0,209 (±0,002)	0,007 (±0,001)	0,188 (±0,005)
500	3	22,5% (±0,2%)	27,0% (±0,3%)	13,1% (±0,3%)	25,0% (±0,3%)	0,166 (±0,001)	0,199 (±0,001)	0,090 (±0,002)	0,193 (±0,002)
2000	3	21,0% (±0,5%)	25,1% (±0,5%)	17,6% (±0,2%)	24,4% (±0,5%)	0,148 (±0,001)	0,194 (±0,001)	0,122 (±0,001)	0,192 (±0,002)
100	6	20,1% (±0,8%)	30,4% (±1,4%)	0,0% (±0,0%)	20,6% (±1,2%)	0,186 (±0,007)	0,211 (±0,003)	0,000 (±0,000)	0,182 (±0,008)
500	6	14,7% (±0,4%)	24,7% (±0,5%)	4,6% (±0,5%)	23,2% (±0,5%)	0,120 (±0,002)	0,200 (±0,002)	0,032 (±0,003)	0,196 (±0,002)
2000	6	13,2% (±0,3%)	23,7% (±0,4%)	9,5% (±0,3%)	23,0% (±0,3%)	0,099 (±0,001)	0,198 (±0,001)	0,067 (±0,001)	0,195 (±0,001)
100	20	12,4% (±1,1%)	28,9% (±2,6%)	0,0% (±0,0%)	13,9% (±1,7%)	0,116 (±0,011)	0,217 (±0,003)	0,000 (±0,000)	0,183 (±0,008)
500	20	4,9% (±0,2%)	21,2% (±2,0%)	0,0% (±0,0%)	16,5% (±1,6%)	0,041 (±0,002)	0,210 (±0,003)	0,000 (±0,000)	0,201 (±0,005)
2000	20	3,3% (±0,2%)	17,7% (±1,0%)	0,1% (±0,0%)	16,5% (±0,8%)	0,026 (±0,001)	0,205 (±0,002)	0,001 (±0,000)	0,205 (±0,003)

Table 7
Comparison between Adaboost and KL-Boost: classification and quadratic errors for ringnorm generators with superfluous features

N	Dimension	Classif. gen. errors		Classif. emp. errors		L ² gen. errors		L ² emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3,6	25,9% (±0,7%)	29,0% (±0,9%)	0,0% (±0,0%)	20,8% (±1,5%)	0,236 (±0,005)	0,206 (±0,006)	0,000 (±0,000)	0,177 (±0,012)
500	3,6	21,6% (±0,5%)	25,1% (±0,7%)	10,0% (±0,6%)	23,3% (±0,6%)	0,167 (±0,002)	0,188 (±0,003)	0,068 (±0,003)	0,183 (±0,005)
2000	3,6	19,6% (±0,3%)	22,9% (±0,5%)	15,9% (±0,2%)	22,2% (±0,5%)	0,142 (±0,001)	0,183 (±0,002)	0,110 (±0,001)	0,182 (±0,002)
100	10,20	16,7% (±0,9%)	28,7% (±1,7%)	0,0% (±0,0%)	15,9% (±1,2%)	0,157 (±0,008)	0,214 (±0,004)	0,000 (±0,000)	0,178 (±0,012)
500	10,20	9,7% (±0,2%)	20,9% (±0,7%)	0,0% (±0,0%)	17,9% (±0,6%)	0,085 (±0,002)	0,201 (±0,004)	0,000 (±0,000)	0,194 (±0,005)
2000	10,20	8,1% (±0,2%)	19,2% (±0,5%)	3,4% (±0,2%)	18,4% (±0,4%)	0,065 (±0,001)	0,202 (±0,004)	0,024 (±0,001)	0,200 (±0,005)

since we use the minimax approach, which considers the worst possible probability distribution and consequently leads to very conservative bounds.

To add noise, we just flip the output with probability 20%. Then the frontier between the classes is not altered but the regression function f is transformed into $0.2 + 0.6f$ which implies that it is always between 0.2 and 0.8. In this case, results are much more in favor of KL-Boost. Here the loss of performance of AdaBoost does not seem to come from overfitting since the empirical risks are no longer close to 0. It is due to the model itself, which is not enough complex to take into account a regression function which is bounded away from 0 and 1.

For the 6-dimensional twonorm generator with 3 superfluous components in the input, KL-Boost gives better results than AdaBoost for small training sets, whereas for large training sets, both methods lead to similar results.

Table 8

Comparison between Adaboost and KL-Boost: classification and quadratic errors for noisy twonorms generators

N	Dimension	Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3	31, 8%	23,4%	1, 0%	20, 6%	0, 269	0,191	0, 015	0, 172
		(±0, 7%)	(±0, 4%)	(±0, 3%)	(±0, 9%)	(±0, 004)	(±0, 003)	(±0, 002)	(±0, 006)
500	3	26, 0%	21,9%	18, 0%	21, 6%	0, 198	0,190	0, 124	0, 187
		(±0, 3%)	(±0, 1%)	(±0, 3%)	(±0, 3%)	(±0, 001)	(±0, 003)	(±0, 002)	(±0, 004)
2000	3	23, 2%	21,6%	21, 1%	21, 5%	0,181	0, 185	0, 157	0, 184
		(±0, 3%)	(±0, 1%)	(±0, 4%)	(±0, 4%)	(±0, 001)	(±0, 007)	(±0, 002)	(±0, 007)
100	6	32, 4%	24,1%	0, 0%	19, 7%	0, 287	0,198	0, 001	0, 172
		(±1, 0%)	(±0, 9%)	(±0, 0%)	(±1, 8%)	(±0, 008)	(±0, 005)	(±0, 001)	(±0, 013)
500	6	28, 4%	22,1%	15, 6%	21, 6%	0, 213	0,197	0, 104	0, 194
		(±0, 6%)	(±0, 1%)	(±0, 6%)	(±0, 6%)	(±0, 003)	(±0, 006)	(±0, 003)	(±0, 007)
2000	6	24, 2%	21,8%	21, 2%	21, 7%	0,187	0, 194	0, 154	0, 195
		(±0, 4%)	(±0, 1%)	(±0, 4%)	(±0, 4%)	(±0, 001)	(±0, 007)	(±0, 002)	(±0, 007)
100	20	34, 7%	28,2%	0, 0%	17, 6%	0, 322	0,210	0, 000	0, 166
		(±1, 0%)	(±1, 8%)	(±0, 0%)	(±2, 0%)	(±0, 008)	(±0, 005)	(±0, 000)	(±0, 014)
500	20	31, 5%	23,0%	8, 8%	21, 8%	0, 245	0,213	0, 061	0, 209
		(±0, 7%)	(±0, 3%)	(±0, 5%)	(±0, 8%)	(±0, 003)	(±0, 006)	(±0, 003)	(±0, 007)
2000	20	27, 2%	22,0%	20, 4%	21, 9%	0,202	0, 216	0, 141	0, 215
		(±0, 4%)	(±0, 1%)	(±0, 4%)	(±0, 4%)	(±0, 001)	(±0, 005)	(±0, 002)	(±0, 005)

Table 9

Comparison between Adaboost and KL-Boost: classification and quadratic errors for noisy threenorm generators

N	Dimension	Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3	38, 0%	32,8%	2, 1%	25, 2%	0, 307	0,222	0, 026	0, 188
		(±0, 7%)	(±0, 9%)	(±0, 3%)	(±1, 0%)	(±0, 005)	(±0, 002)	(±0, 002)	(±0, 004)
500	3	32, 3%	28,1%	21, 5%	27, 7%	0, 225	0,211	0, 145	0, 204
		(±0, 3%)	(±0, 2%)	(±0, 3%)	(±0, 4%)	(±0, 001)	(±0, 001)	(±0, 002)	(±0, 002)
2000	3	29, 5%	27,5%	26, 5%	27, 9%	0,205	0, 207	0, 180	0, 205
		(±0, 4%)	(±0, 2%)	(±0, 4%)	(±0, 4%)	(±0, 001)	(±0, 001)	(±0, 002)	(±0, 002)
100	6	39, 0%	38,2%	0, 0%	26, 0%	0, 350	0,231	0, 004	0, 194
		(±1, 2%)	(±1, 1%)	(±0, 1%)	(±1, 6%)	(±0, 009)	(±0, 003)	(±0, 002)	(±0, 009)
500	6	35, 2%	34,2%	18, 5%	29, 9%	0, 257	0,219	0, 127	0, 212
		(±0, 6%)	(±0, 4%)	(±0, 5%)	(±1, 0%)	(±0, 002)	(±0, 003)	(±0, 003)	(±0, 004)
2000	6	32,6%	33, 5%	27, 0%	30, 8%	0, 227	0,214	0, 181	0, 212
		(±0, 4%)	(±0, 2%)	(±0, 5%)	(±0, 4%)	(±0, 001)	(±0, 001)	(±0, 002)	(±0, 001)
100	20	42, 6%	41,9%	0, 0%	24, 6%	0, 388	0,241	0, 000	0, 188
		(±1, 0%)	(±1, 9%)	(±0, 0%)	(±4, 2%)	(±0, 007)	(±0, 003)	(±0, 000)	(±0, 014)
500	20	39, 8%	36,8%	12, 3%	30, 2%	0, 290	0,230	0, 091	0, 215
		(±0, 5%)	(±0, 7%)	(±0, 7%)	(±1, 1%)	(±0, 003)	(±0, 002)	(±0, 004)	(±0, 006)
2000	20	36, 6%	34,9%	26, 0%	32, 7%	0, 240	0,229	0, 172	0, 227
		(±0, 4%)	(±0, 3%)	(±0, 3%)	(±0, 4%)	(±0, 001)	(±0, 002)	(±0, 002)	(±0, 007)

This is also true for the 6-dimensional noisy threenorm and ringnorm generators. The reverse has not occurred in our simulations. So KL-Boost seems to be well-adapted to small training set situations.

It seems that KL-Boost is in general more trustworthy than Adaboost since

Table 10
Comparison between Adaboost and KL-Boost: classification and quadratic errors for noisy ringnorm generators

N	Dimension	Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
		AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost	AdaBoost	KL-Boost
100	3	39, 3%	36, 5%	2, 1%	28, 5%	0, 318	0, 231	0, 026	0, 203
		(±0, 5%)	(±0, 7%)	(±0, 4%)	(±1, 1%)	(±0, 004)	(±0, 002)	(±0, 002)	(±0, 005)
500	3	33, 9%	32, 3%	22, 0%	30, 6%	0, 233	0, 219	0, 149	0, 211
		(±0, 3%)	(±0, 2%)	(±0, 3%)	(±0, 4%)	(±0, 001)	(±0, 002)	(±0, 002)	(±0, 004)
2000	3	31, 7%	30, 8%	27, 5%	30, 2%	0, 214	0, 213	0, 187	0, 210
		(±0, 5%)	(±0, 2%)	(±0, 4%)	(±0, 4%)	(±0, 001)	(±0, 002)	(±0, 002)	(±0, 002)
100	6	37, 3%	36, 6%	0, 0%	25, 0%	0, 327	0, 232	0, 003	0, 196
		(±1, 0%)	(±1, 9%)	(±0, 0%)	(±2, 2%)	(±0, 009)	(±0, 004)	(±0, 001)	(±0, 009)
500	6	32, 6%	31, 5%	17, 2%	29, 4%	0, 233	0, 219	0, 117	0, 213
		(±0, 5%)	(±0, 3%)	(±0, 5%)	(±0, 8%)	(±0, 003)	(±0, 003)	(±0, 003)	(±0, 004)
2000	6	29, 3%	30, 5%	24, 8%	30, 0%	0, 206	0, 213	0, 171	0, 211
		(±0, 5%)	(±0, 2%)	(±0, 4%)	(±0, 4%)	(±0, 001)	(±0, 001)	(±0, 002)	(±0, 001)
100	20	34, 7%	39, 5%	0, 0%	24, 1%	0, 324	0, 237	0, 066	0, 203
		(±1, 0%)	(±2, 2%)	(±0, 0%)	(±3, 9%)	(±0, 008)	(±0, 004)	(±0, 066)	(±0, 013)
500	20	30, 5%	30, 7%	8, 5%	27, 0%	0, 240	0, 225	0, 062	0, 216
		(±0, 7%)	(±1, 0%)	(±0, 4%)	(±0, 8%)	(±0, 004)	(±0, 002)	(±0, 002)	(±0, 005)
2000	20	26, 7%	28, 2%	19, 7%	27, 1%	0, 199	0, 222	0, 139	0, 218
		(±0, 5%)	(±0, 5%)	(±0, 3%)	(±0, 4%)	(±0, 001)	(±0, 002)	(±0, 002)	(±0, 002)

- Adaboost clearly overfits (note that it does not prevent the algorithm from classifying well; it will not overfit when the model is too simple to explain the learning sample; in other cases, it is bound to overfit since it is based on the empirical risk minimization principle).
- KL-Boost behaves well on small training sets and on noisy data.
- Adaboost minimizes a criterion (the exponential risk) using a model which is not at all suited to do it.⁴

6. Conclusion

To get an upper bound on the misclassification rate of any aggregating procedure, we introduce the Kullback–Leibler distance between the aggregating distribution and an arbitrary chosen prior distribution. Then we obtain bounds of optimal order in the minimax sense. We use these bounds to derive the KL-Boost procedure that competes with Adaboost in practice (in particular in noisy classification tasks) and which does not suffer from wild overfitting as Adaboost. KL-Boost is an aggregating procedure regularized by the Kullback–Leibler distance between the aggregating distribution and a prior distribution. A full description of the algorithm has been given when stumps are aggregated.

Future work may concentrate on:

- Describing the general algorithm when the functions aggregated are not stumps: due to the simplicity of stumps, it has been possible to compute explicitly terms which are not computable in general.
- Tightening the bounds: even if these theoretical bounds are much tighter than most of the existing bounds, there is still a gap between theoretical bounds of the misclassification error and the actual misclassification error. Part of this gap clearly comes from the minimax approach. The target would be to reduce the other part.
- Reducing the computational cost of the algorithm.

⁴ Numerical results show that this criterion is minimized much more efficiently by KL-Boost!

7. Proofs

7.1. Proof of Theorem 3.1

The proof relies on deviation inequalities and on Legendre formula.

7.1.1. First step: deviation inequalities

Let $\bar{R}(\theta)$ denote the expected risk of f_θ relatively to the reference one: $\bar{R}(\theta) \triangleq R(f_\theta) - R(\tilde{f})$. Similarly, we define $\bar{r}(\theta) \triangleq r(f_\theta) - r(\tilde{f})$. Putting $Z_\theta(X, Y) \triangleq -(Y - f_\theta(X))^2 + (Y - \tilde{f}(X))^2$, we have $\bar{R}(\theta) = -\mathbb{E}_{\mathbb{P}} Z_\theta$. We will need a deviation lemma for Z_θ . Let us start with general deviation lemmas for random variables:

Lemma 7.1. *Let W be a random variable bounded by $b \in \mathbb{R}$. Then for any $\eta > 0$, we have*

$$\log \mathbb{E} e^{\eta(W - \mathbb{E}W)} \leq \eta^2 \mathbb{E}W^2 g(\eta b),$$

where $g(u) \triangleq \frac{e^u - 1 - u}{u^2}$.

Proof. We have

$$e^{\eta W} = 1 + \eta W + \eta^2 W^2 g(\eta W).$$

Using that $\log(1 + x) \leq x$ and that $g(\eta W) \leq g(\eta b)$, we obtain

$$\log \mathbb{E} e^{\eta W} \leq \eta \mathbb{E}W + \eta^2 g(\eta b) \mathbb{E}W^2,$$

which is the desired result. \square

Lemma 7.2. *Let Z be a random variable.*

- If $Z \leq b$ a.s., then for any $\eta \geq 0$,

$$\log \mathbb{E} e^{\eta(Z - \mathbb{E}Z)} \leq \eta^2 \mathbb{E}Z^2 g(\eta b), \quad (7.1)$$

where $g : u \mapsto \frac{e^u - 1 - u}{u^2}$ is a positive convex increasing function such that $g(0) = \frac{1}{2}$ by continuity.

- If $\mathbb{E} e^{\alpha|Z - \mathbb{E}Z|} \leq M$ for some $\alpha > 0$ and $M > 0$, then for any $0 \leq \eta < \alpha$,

$$\log \mathbb{E} e^{\eta(Z - \mathbb{E}Z)} \leq \eta^2 g_1(\eta), \quad (7.2)$$

where $g_1(\eta) \triangleq \frac{2M}{(\alpha - \eta)^2 e^2}$.

Proof.

- We have

$$e^{\eta Z} = 1 + \eta Z + \eta^2 Z^2 g(\eta Z).$$

Using that $\log(1 + x) \leq x$ and that $g(\eta Z) \leq g(\eta b)$, we obtain

$$\log \mathbb{E} e^{\eta Z} \leq \eta \mathbb{E}Z + \eta^2 g(\eta b) \mathbb{E}Z^2,$$

which leads to inequality (7.1).

- From the bound on the exponential moment of \bar{Z} , we can easily deduce bounds for the moments of \bar{Z} . By straightforward computation, one can show that the maximum of $[u \mapsto ue^{-\beta u}]$ on \mathbb{R}_+ is $\frac{1}{\beta e}$, hence, for any $q > 0$:

$$\mathbb{E}|\bar{Z}|^q \leq \left(\sup_{u \in \mathbb{R}_+} ue^{-\frac{\alpha}{q}u} \right)^q \mathbb{E} e^{\alpha|\bar{Z}|} \leq \left(\frac{q}{\alpha e} \right)^q \mathbb{E} e^{\alpha|\bar{Z}|} \leq \left(\frac{q}{\alpha e} \right)^q M.$$

According to the Taylor series expansion, for any $\eta \geq 0$, for any $x \in \mathbb{R}$, there exists $\gamma \in]0; \eta[$ such that $e^{\eta x} - 1 - \eta x = \frac{\eta^2 x^2}{2} e^{\gamma x}$, hence for any $x \in \mathbb{R}$,

$$e^{\eta x} - 1 - \eta x \leq \frac{\eta^2 x^2}{2} e^{\eta|x|}.$$

Then for any $\eta \in [0; \alpha[$, we have

$$\begin{aligned} \log \mathbb{E} e^{\eta \bar{Z}} &\leq \mathbb{E}(e^{\eta \bar{Z}} - 1 - \eta \bar{Z}) \leq \mathbb{E}\left(\frac{\eta^2 \bar{Z}^2}{2} e^{\eta|\bar{Z}|}\right) \leq \frac{\eta^2}{2} \mathbb{E}(\bar{Z}^2 e^{\eta|\bar{Z}|}) \\ &\leq \frac{\eta^2}{2} (\mathbb{E}|\bar{Z}|^{\frac{2\alpha}{\alpha-\eta}})^{\frac{\alpha-\eta}{\alpha}} (\mathbb{E}e^{\alpha|\bar{Z}|})^{\frac{\eta}{\alpha}} \quad (\text{by Hölder's inequality}) \\ &\leq \frac{\eta^2}{2} \left(\frac{2}{\alpha - \eta} e\right)^2 M \leq \eta^2 g_1(\eta). \quad \square \end{aligned}$$

The deviations of $Z_\theta = -(Y - f_\theta(X))^2 + (Y - \tilde{f}(X))^2$ are given by:

Lemma 7.3. For any $0 < \lambda < \frac{\alpha B}{2}$ satisfying

$$8M\lambda \leq (\alpha B - 2\lambda)^2 e^2, \tag{7.3}$$

we have

$$\log \mathbb{E}_{\mathbb{P}} e^{\lambda \frac{Z_\theta - \mathbb{E}_{\mathbb{P}} Z_\theta}{B^2}} \leq \lambda^2 \frac{\mathbb{E}_{\mathbb{P}}(\tilde{f} - f_\theta)^2}{B^2} G(\lambda), \tag{7.4}$$

where

$$G(\lambda) \triangleq \frac{8M}{(\alpha B - 2\lambda)^2 e^2} + \frac{e^{2\lambda} - 1 - 2\lambda}{\lambda^2}.$$

Remark 7.1. The condition $\lambda < \frac{\alpha B}{2}$ is unavoidable since we have not put strong assumptions on the noise (i.e. $Y - E(Y/X)$) distribution. The result will be applied for small values of λ . So the conditions on λ are not harmful and can be disregarded, and we will have

$$G(\lambda) \approx G(0) = \frac{8M}{(\alpha B e)^2} + 2.$$

Note that G is adimensional since it is expressed in terms of M and αB .

Remark 7.2. The first term in the deviation function G comes from the noise whereas the second one takes into account the deviations of f_θ with respect to the reference regression function \tilde{f} . When the noise is gaussian, specifically when $Y - f^*(X)$ is a centered gaussian random variable with variance σ^2 , the deviation function is

$$G(\lambda) = \frac{\sigma^2}{2B^2} + \frac{e^{2\lambda} - 1 - 2\lambda}{\lambda^2}.$$

Remark 7.3. The inequality is tight to the extent that for f_θ sufficiently close to \tilde{f} , the bound is close to 0.

Proof. We can write

$$Z_\theta = -(\tilde{f} - f_\theta)^2 - 2(Y - f^*)(\tilde{f} - f_\theta) - 2(f^* - \tilde{f})(\tilde{f} - f_\theta),$$

where f refers to $f(X)$ in order to simplify notations and $f^* \triangleq \mathbb{E}_{\mathbb{P}}(Y/X = \cdot)$ is the regression function associated with the distribution \mathbb{P} . Hence, using the deviation inequality (7.2) and introducing

$$\kappa(\lambda) \triangleq \frac{4\lambda}{B^2} g_1\left(\frac{2\lambda}{B}\right) = \frac{8M\lambda}{(\alpha B - 2\lambda)^2 e^2} \leq 1$$

for any λ satisfying (7.3),

$$\begin{aligned} \mathbb{E}_{\mathbb{P}(dY/X)} e^{\lambda \frac{Z_\theta - \mathbb{E}_{\mathbb{P}} Z_\theta}{B^2}} &= e^{\frac{\lambda}{B^2} (\bar{R}(\theta) - (\tilde{f} - f_\theta)^2 - 2(f^* - \tilde{f})(\tilde{f} - f_\theta))} \mathbb{E}_{\mathbb{P}(dY/X)} e^{-\frac{2\lambda}{B^2} (\tilde{f} - f_\theta)(Y - f^*)} \\ &\leq e^{\frac{\lambda}{B^2} (\bar{R}(\theta) - (\tilde{f} - f_\theta)^2 - 2(f^* - \tilde{f})(\tilde{f} - f_\theta))} e^{\left[\frac{2\lambda}{B^2} (\tilde{f} - f_\theta)\right]^2 g_1\left(\frac{2\lambda}{B}\right)} \\ &= e^{\frac{\lambda}{B^2} (\mathbb{E}_{\mathbb{P}}(\tilde{f} - f_\theta)^2 + 2\mathbb{E}_{\mathbb{P}}\{(f^* - \tilde{f})(\tilde{f} - f_\theta)\} - [1 - \frac{4\lambda}{B^2} g_1\left(\frac{2\lambda}{B}\right)](\tilde{f} - f_\theta)^2 - 2(f^* - \tilde{f})(\tilde{f} - f_\theta))} \\ &= e^{\frac{\lambda}{B^2} [\mathbb{E}_{\mathbb{P}}(\tilde{f} - f_\theta)^2 + 2\mathbb{E}_{\mathbb{P}}\{(f^* - \tilde{f})(\tilde{f} - f_\theta)\} - (\tilde{f} - f_\theta)([1 - \kappa(\lambda)](\tilde{f} - f_\theta) + 2(f^* - \tilde{f}))]} \\ &= e^{\frac{\lambda}{B^2} \kappa(\lambda) \mathbb{E}_{\mathbb{P}}(\tilde{f} - f_\theta)^2 + \frac{\lambda}{B^2} (\bar{Z}_\theta - \mathbb{E}_{\mathbb{P}} \bar{Z}_\theta)}, \end{aligned}$$

where $\bar{Z}_\theta \triangleq -(\tilde{f} - f_\theta)\{2f^* - [1 + \kappa(\lambda)]\tilde{f} - [1 - \kappa(\lambda)]f_\theta\} \leq 2B^2$. From the deviation inequality (7.1), we get

$$\begin{aligned} \log \mathbb{E}_{\mathbb{P}} e^{\frac{\lambda}{B^2} (Z_\theta - \mathbb{E}_{\mathbb{P}} Z_\theta)} &\leq \frac{\lambda \kappa(\lambda)}{B^2} \mathbb{E}_{\mathbb{P}}(\tilde{f} - f_\theta)^2 + \left(\frac{\lambda}{B^2}\right)^2 \mathbb{E}_{\mathbb{P}} \bar{Z}_\theta^2 g(2\lambda) \\ &\leq \frac{\lambda \kappa(\lambda)}{B^2} \mathbb{E}_{\mathbb{P}}(\tilde{f} - f_\theta)^2 + \frac{\lambda^2}{B^4} \mathbb{E}_{\mathbb{P}}(\tilde{f} - f_\theta)^2 4B^2 g(2\lambda) \\ &\leq \lambda^2 \frac{\mathbb{E}_{\mathbb{P}}(\tilde{f} - f_\theta)^2}{B^2} \left[\frac{\kappa(\lambda)}{\lambda} + 4g(2\lambda) \right]. \end{aligned}$$

7.1.2. Second step: Legendre formula

Let us remind the definition of the Kullback–Leibler divergence between two probability distributions on a measurable set (A, \mathcal{A}) :

$$K(v, \mu) \triangleq \begin{cases} \mathbb{E}_v \log(v/\mu) & \text{if } v \ll \mu, \\ +\infty & \text{otherwise.} \end{cases}$$

The Legendre transform of the convex function $v \mapsto K(v, \mu)$ is given by the following formula: for any measurable function $h : A \mapsto \mathbb{R}$,

$$\sup_{v \in \mathcal{M}_+^1(A)} \left\{ \mathbb{E}_{v(da)} h(a) - K(v, \mu) \right\} = \log \mathbb{E}_{\mu(da)} e^{h(a)}, \tag{7.5}$$

where, by convention:

$$\begin{cases} \mathbb{E}_{v(da)} h(a) \triangleq \sup_{H \in \mathbb{R}} \mathbb{E}_{v(da)} [H \wedge h(a)], \\ \mathbb{E}_{v(da)} h(a) - K(v, \mu) = -\infty & \text{if } K(v, \mu) = +\infty. \end{cases}$$

Moreover, when e^h is μ -integrable, the probability distribution

$$v(da) \triangleq \frac{e^{h(a)}}{\mathbb{E}_{\mu(da')} e^{h(a')}} \cdot \mu(da)$$

achieves the supremum.

For any $\varepsilon > 0$ and $\lambda > 0$ such that $\lambda G(\lambda) < 1$, the event

$$\left\{ \begin{array}{l} \text{there exists } \rho \in \mathcal{M}_+^1(\Theta) \text{ such that} \\ \mathbb{E}_{\rho(d\theta)} R(f_\theta) - R(\tilde{f}) > \frac{\mathbb{E}_{\rho(d\theta)} r(f_\theta) - r(\tilde{f})}{1 - \lambda G(\lambda)} + \frac{B^2}{N} \frac{K(\rho, \pi) + \log(\varepsilon^{-1})}{\lambda[1 - \lambda G(\lambda)]} \end{array} \right\}$$

is successively equal to

$$\left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \mathbb{E}_\rho \bar{R} - \frac{\mathbb{E}_\rho \bar{f}}{1 - \lambda G(\lambda)} - \frac{B^2}{N} \frac{K(\rho, \pi) + \log(\varepsilon^{-1})}{\lambda[1 - \lambda G(\lambda)]} \right\} > 0 \right\},$$

$$\left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \mathbb{E}_\rho \left([1 - \lambda G(\lambda)] \bar{R} - \bar{f} \right) - \frac{B^2}{N\lambda} [K(\rho, \pi) + \log(\varepsilon^{-1})] \right\} > 0 \right\},$$

$$\left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \mathbb{E}_\rho \left[\frac{N\lambda}{B^2} ([1 - \lambda G(\lambda)] \bar{R} - \bar{f}) - \log(\varepsilon^{-1}) \right] - K(\rho, \pi) \right\} > 0 \right\},$$

$$\{ \log \mathbb{E}_\pi e^{\frac{N\lambda}{B^2} ([1 - \lambda G(\lambda)] \bar{R} - \bar{f}) - \log(\varepsilon^{-1})} > 0 \},$$

$$\{ \mathbb{E}_\pi e^{\frac{N\lambda}{B^2} ([1 - \lambda G(\lambda)] \bar{R} - \bar{f}) - \log(\varepsilon^{-1})} > 1 \}.$$

Therefore its $\mathbb{P}^{\otimes N}$ -probability is strictly lower than

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}^{\otimes N}} \mathbb{E}_\pi e^{\frac{N\lambda}{B^2} ([1 - \lambda G(\lambda)] \bar{R} - \bar{f}) - \log(\varepsilon^{-1})} \\ &= \mathbb{E}_\pi \mathbb{E}_{\mathbb{P}^{\otimes N}} e^{\frac{N\lambda}{B^2} ([1 - \lambda G(\lambda)] \bar{R} - \bar{f}) - \log(\varepsilon^{-1})} \quad (\text{by Fubini's theorem}) \\ &= \varepsilon \mathbb{E}_\pi \mathbb{E}_{\mathbb{P}^{\otimes N}} e^{\frac{N\lambda}{B^2} [\mathbb{E}_{\mathbb{P}} Z_\theta - \mathbb{E}_{\mathbb{P}} Z_\theta - \lambda G(\lambda) \bar{R}]} \quad (\text{since } Z_\theta \triangleq (Y - \tilde{f})^2 - (Y - f_\theta)^2) \\ &\leq \varepsilon \mathbb{E}_\pi \left[e^{-\frac{N\lambda^2 G(\lambda) \bar{R}}{B^2}} (\mathbb{E}_{\mathbb{P}} e^{\frac{\lambda}{B^2} (Z_\theta - \mathbb{E}_{\mathbb{P}} Z_\theta)})^N \right] \quad (\text{since the training sample is i.i.d}) \\ &\leq \varepsilon \mathbb{E}_\pi \left[e^{\frac{N\lambda^2 G(\lambda) [\mathbb{E}_{\mathbb{P}} (\tilde{f} - f_\theta)^2 - \bar{R}]}{B^2}} \right] \quad (\text{from Lemma 7.3}) \\ &\leq \varepsilon, \end{aligned}$$

where at the last step we use that we have $\mathbb{E}_{\mathbb{P}} (\tilde{f} - f_\theta)^2 \leq \bar{R}(\theta)$ since the function \tilde{f} is the best convex combination.

Remark 7.4. Theorems 3.1 and 3.2 remain true for any reference estimator \tilde{f} satisfying $\mathbb{E}_{\mathbb{P}} \{ [f^*(X) - \tilde{f}(X)] [\tilde{f}(X) - f_\theta(X)] \} \geq 0$. Naturally, this property holds for the best mixture. When the reference estimator is the regression function associated with the distribution \mathbb{P} : $\tilde{f} = f^*$, we have $\bar{Z}_\theta = -[1 - \kappa(\lambda)] [f^* - f_\theta]^2 \in [-B^2; 0]$. Consequently, in this case, Theorems 3.1 and 3.2 hold with a smaller deviation function: $G(\lambda) = \frac{8M}{(\alpha B - 2\lambda)^2 e^2} + \frac{1}{2}$.

7.2. Proof of Theorem 4.1

The decomposition

$$R(\mathbb{E}_{\rho(d\theta)} f_\theta) = \mathbb{E}_{\rho(d\theta)} R(f_\theta) - \mathbb{E}_{\mathbb{P}} \text{Var}_{\rho(d\theta)} f_\theta(X) \tag{7.6}$$

shows that aggregating regression procedures is more efficient than randomizing and that the difference is measured by $\mathbb{E}_{\mathbb{P}} \text{Var}_{\rho(d\theta)} f_\theta(X)$. We will use this decomposition to bound the expected risk of the aggregated regression procedure by successively bounded the two terms on the right-hand side. The first term has already been bounded (see Theorem 3.1). It remains to bound the variance term. Once more, we use deviation inequalities and Legendre formula.

7.2.1. First step: deviation inequalities

Let us introduce $Z_{\theta, \theta'} \triangleq (f_\theta - f_{\theta'})^2 \in [0; B^2]$. We have

$$\text{Var}_{\rho(d\theta)} f_\theta(X) = \frac{1}{2} \mathbb{E}_{\rho \otimes \rho(d\theta, d\theta')} Z_{\theta, \theta'}.$$

The deviations of $Z_{\theta, \theta'}$ are given by

Lemma 7.4. For any $\lambda \geq 0$,

$$\log \mathbb{E}_{\mathbb{P}} e^{\lambda \frac{Z_{\theta, \theta'} - \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}}{B^2}} \leq \lambda^2 \frac{\mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}}{B^2} g(\lambda),$$

where $g(\lambda) \triangleq \frac{e^{\lambda} - 1 - \lambda}{\lambda^2}$.

Remark 7.5. Recall that g is a positive convex increasing function such that $g(0) = \frac{1}{2}$ by continuity.

Proof. For any $\lambda \geq 0$,

$$\begin{aligned} \log \mathbb{E}_{\mathbb{P}} e^{\lambda \frac{Z_{\theta, \theta'} - \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}}{B^2}} &\leq \mathbb{E}_{\mathbb{P}} \left[e^{\lambda \frac{Z_{\theta, \theta'} - \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}}{B^2}} - 1 - \lambda \frac{Z_{\theta, \theta'} - \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}}{B^2} \right] \\ &= \mathbb{E}_{\mathbb{P}} \left[\left(\lambda \frac{Z_{\theta, \theta'} - \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}}{B^2} \right)^2 g \left(\lambda \frac{Z_{\theta, \theta'} - \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}}{B^2} \right) \right] \\ &\leq \frac{\lambda^2}{B^4} \mathbb{E}_{\mathbb{P}} [Z_{\theta, \theta'}^2 g(\lambda)] \leq \frac{\lambda^2}{B^2} g(\lambda) \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}, \end{aligned}$$

since $Z_{\theta, \theta'}^2 \leq B^2 Z_{\theta, \theta'}$. \square

7.2.2. Second step: Legendre formula

Introduce $V = \mathbb{E}_{\mathbb{P}} \text{Var}_{\hat{\rho}(d\theta)} f_{\theta}$ and $\bar{V} = \mathbb{E}_{\bar{\mathbb{P}}} \text{Var}_{\hat{\rho}(d\theta)} f_{\theta}$. For any $\varepsilon > 0$ and $\beta > 0$, the event

$$\left\{ \begin{array}{l} \text{there exists } \rho \in \mathcal{M}_+^1(\Theta) \text{ such that} \\ -V > -\frac{\bar{V}}{1 + \beta g(\beta)} + \frac{B^2}{2N} \frac{2K(\rho, \pi) + \log(\varepsilon^{-1})}{\beta[1 + \beta g(\beta)]} \end{array} \right\}$$

is equal to

$$\left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ -\mathbb{E}_{\rho \otimes \rho}(d\theta, d\theta') \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'} + \frac{\mathbb{E}_{\rho \otimes \rho}(d\theta, d\theta') \mathbb{E}_{\bar{\mathbb{P}}} Z_{\theta, \theta'}}{1 + \beta g(\beta)} - \frac{B^2}{N} \frac{2K(\rho, \pi) + \log(\varepsilon^{-1})}{\beta[1 + \beta g(\beta)]} \right\} > 0 \right\},$$

which is included in the event

$$\left\{ \sup_{\mu \in \mathcal{M}_+^1(\Theta \times \Theta)} \left\{ \mathbb{E}_{\mu}(d\theta, d\theta') [\mathbb{E}_{\bar{\mathbb{P}}} Z_{\theta, \theta'} - [1 + \beta g(\beta)] \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}] - \frac{B^2}{N} \frac{K(\mu, \pi \otimes \pi) + \log(\varepsilon^{-1})}{\beta} \right\} > 0 \right\}.$$

This last event can be written successively as

$$\begin{aligned} &\left\{ \sup_{\mu \in \mathcal{M}_+^1(\Theta \times \Theta)} \left\{ \mathbb{E}_{\mu}(d\theta, d\theta') \left[\frac{N\beta}{B^2} (\mathbb{E}_{\bar{\mathbb{P}}} Z_{\theta, \theta'} - [1 + \beta g(\beta)] \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}) - \log(\varepsilon^{-1}) \right] - K(\mu, \pi \otimes \pi) \right\} > 0 \right\}, \\ &\left\{ \log \mathbb{E}_{\pi \otimes \pi}(d\theta, d\theta') e^{\frac{N\beta}{B^2} (\mathbb{E}_{\bar{\mathbb{P}}} Z_{\theta, \theta'} - [1 + \beta g(\beta)] \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}) - \log(\varepsilon^{-1})} > 0 \right\}, \\ &\left\{ \mathbb{E}_{\pi \otimes \pi}(d\theta, d\theta') e^{\frac{N\beta}{B^2} (\mathbb{E}_{\bar{\mathbb{P}}} Z_{\theta, \theta'} - [1 + \beta g(\beta)] \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}) - \log(\varepsilon^{-1})} > 1 \right\}. \end{aligned}$$

Therefore its $\mathbb{P}^{\otimes N}$ -probability is strictly lower than

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}^{\otimes N}} \mathbb{E}_{\pi \otimes \pi} (d\theta, d\theta') e^{\frac{N\beta}{B^2} (\mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'} - [1 + \beta g(\beta)] \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}) - \log(\varepsilon^{-1})} \\ &= \varepsilon \mathbb{E}_{\pi \otimes \pi} (d\theta, d\theta') \mathbb{E}_{\mathbb{P}^{\otimes N}} e^{\frac{N\beta}{B^2} (\mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'} - \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'} - \beta g(\beta) \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'})} \quad (\text{by Fubini's theorem}) \\ &\leq \varepsilon \mathbb{E}_{\pi} \left[e^{-\frac{N\beta^2 g(\beta) \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'}}{B^2}} (\mathbb{E}_{\mathbb{P}} e^{\frac{\beta}{B^2} (Z_{\theta, \theta'} - \mathbb{E}_{\mathbb{P}} Z_{\theta, \theta'})})^N \right] \quad (\text{i.i.d. training sample}) \\ &\leq \varepsilon \quad (\text{from Lemma 7.4}). \end{aligned}$$

7.3. Proof of Lemma 4.4

We will take the following parameter families

- $(\lambda_i)_{i=0, \dots, p}$, where $\lambda_i \triangleq \lambda_{\max}/2^i$, p is such that $\lambda_{\max}/2^p < \lambda_{\min} \leq \lambda_{\max}/2^{p-1}$ and λ_{\min} and λ_{\max} will be determined later,
- $(\eta_i)_{i=0, \dots, p}$, where $\eta_i \triangleq \eta \triangleq 1/(p+1)$,
- $(\beta_j)_{j=0, \dots, q}$, where $\beta_j \triangleq \beta_{\max}/2^j$, q is such that $\beta_{\max}/2^q < \beta_{\min} \leq \beta_{\max}/2^{q-1}$ and β_{\min} and β_{\max} will be determined later,
- $(\zeta_j)_{j=0, \dots, q}$, where $\zeta_j \triangleq \zeta \triangleq 1/(q+1)$.

The exponential form of the parameters λ_i and β_j allows us to have a grid on which for any probability distribution ρ , the minimum of $\mathbb{B}(\rho, \lambda, \eta, \beta, \zeta)$ has the same order as

$$\inf_{\substack{\lambda \in [\lambda_{\min}; \lambda_{\max}] \\ \beta \in [\beta_{\min}; \beta_{\max}]}} \mathbb{B}(\rho, \lambda, \eta, \beta, \zeta).$$

We will choose the parameters λ_{\min} and λ_{\max} (respectively β_{\min} and β_{\max}) such that the constant η (respectively ζ) is large (in order that the bound is not significantly affected by the union bound term $\log[(\eta\varepsilon)^{-1}]$ (respectively $\log[(\zeta\varepsilon)^{-1}]$). We will see a posteriori that $\mathbb{B}(\tilde{\rho}, \lambda, \eta, \beta, \zeta)$ will just differ from $\mathbb{B}(\tilde{\rho}, \lambda, 1, \beta, 1)$ by a $\log \log N$ factor.

We have

$$\begin{aligned} \mathbb{B}(\tilde{\rho}, \lambda, \eta, \beta, \zeta) &= \left(\frac{1}{1 - \lambda G(\lambda)} - \frac{1}{1 + \beta g(\beta)} \right) \bar{V}(\tilde{\rho}) + \frac{B^2}{N} \frac{K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{\lambda[1 - \lambda G(\lambda)]} \\ &\quad + \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta\varepsilon)^{-1}]}{\beta[1 + \beta g(\beta)]}. \end{aligned} \tag{7.7}$$

In general, the quantity $\bar{V}(\tilde{\rho}) = \mathbb{E}_{\mathbb{P}} \text{Var}_{\tilde{\rho}(d\theta)} f_{\theta}$ is of order 1 (i.e. B^2). Consequently, to make the second term small, we need to take both parameters λ and β small. However, these parameters must not be too small since the two last terms are respectively proportional to $\frac{1}{\lambda}$ and $\frac{1}{\beta}$. In the particular case when $\bar{V}(\tilde{\rho})$ is close to 0, we need not taking λ and β small. So we take arbitrarily

$$\begin{cases} \lambda_{\max} = \kappa_1, \\ \beta_{\max} = \kappa_2, \end{cases}$$

where κ_1 and κ_2 are respectively defined as $2\kappa_1 G(\kappa_1) = 1$ and $\kappa_2 g(\kappa_2) = 1$.

We will consider separately the terms of (7.7) depending on λ and on β . We start with the β terms. Since g is an increasing function such that $g(0) = \frac{1}{2}$ and since for any $0 < x \leq 1$, $1 - x < \frac{1}{1+x} \leq 1 - \frac{x}{2}$, we have for any $0 < \beta \leq \beta_{\max}$,

$$-\frac{\bar{V}(\tilde{\rho})}{1 + \beta g(\beta)} + \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta\varepsilon)^{-1}]}{\beta[1 + \beta g(\beta)]}$$

$$\begin{aligned}
&\leq -[1 - \beta g(\beta_{\max})] \bar{V}(\tilde{\rho}) + \left(1 - \frac{\beta}{4}\right) \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta} \\
&= -\bar{V}(\tilde{\rho}) - \frac{B^2}{8N} (2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]) + \frac{\beta}{\beta_{\max}} \bar{V}(\tilde{\rho}) + \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta}.
\end{aligned} \tag{7.8}$$

The last RHS is minimum when

$$\beta = \beta_{\text{opt}} \triangleq \sqrt{\frac{B^2 \beta_{\max}}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\bar{V}(\tilde{\rho})}} \geq \sqrt{\frac{2\beta_{\max} \log(\varepsilon^{-1})}{N}},$$

since $\bar{V}(\tilde{\rho}) \leq B^2/4$ according to assumption (2.1). Therefore, let us take

$$\beta_{\min} \triangleq \sqrt{\frac{2\beta_{\max} \log(\varepsilon^{-1})}{N}} \wedge \beta_{\max}.$$

Let us define the event

$$E_1 \triangleq \left\{ \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\bar{V}(\tilde{\rho})} \leq \beta_{\max} \right\}.$$

General case: E_1 occurs. Then we have $\beta_{\text{opt}} \leq \beta_{\max}$. So there exists an integer $0 \leq j \leq q$ such that $\beta_j \leq \beta_{\text{opt}} < 2\beta_j$. For this integer j , using inequality (7.8), we get

$$\begin{aligned}
&-\frac{\bar{V}(\tilde{\rho})}{1 + \beta_j g(\beta_j)} + \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_j [1 + \beta_j g(\beta_j)]} \\
&\leq -\bar{V}(\tilde{\rho}) - \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{4} + \frac{\beta_{\text{opt}}}{\beta_{\max}} \bar{V}(\tilde{\rho}) + \frac{B^2}{N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_{\text{opt}}} \\
&= -\bar{V}(\tilde{\rho}) - \frac{B^2}{8N} (2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]) + 3\sqrt{\frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_{\max}}} \bar{V}(\tilde{\rho}).
\end{aligned}$$

Particular case: $(E_1)^c$ occurs. Then, for $j = 0$, we have

$$-\frac{\bar{V}(\tilde{\rho})}{1 + \beta_j g(\beta_j)} + \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_j [1 + \beta_j g(\beta_j)]} = -\frac{\bar{V}(\tilde{\rho})}{2} + \frac{B^2}{4N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_{\max}}.$$

Besides, we have

$$\sqrt{\frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_{\max}}} \bar{V}(\tilde{\rho}) \geq \bar{V}(\tilde{\rho}).$$

So, in both cases, there exists an integer $0 \leq j \leq q$ such that

$$\begin{aligned}
&-\frac{\bar{V}(\tilde{\rho})}{1 + \beta_j g(\beta_j)} + \frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_j [1 + \beta_j g(\beta_j)]} \\
&\leq -\bar{V}(\tilde{\rho}) + \frac{B^2}{4N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_{\max}} + 3\sqrt{\frac{B^2}{2N} \frac{2K(\tilde{\rho}, \pi) + \log[(\zeta \varepsilon)^{-1}]}{\beta_{\max}}} \bar{V}(\tilde{\rho}).
\end{aligned} \tag{7.9}$$

Now let us deal with the λ terms of (7.7). Since G is an increasing function and the inequality $\frac{1}{1-x} \leq 1 + 2x$ holds for any $0 < x \leq \frac{1}{2}$, we have for any $0 < \lambda \leq \lambda_{\max}$

$$\begin{aligned} & \frac{\bar{V}(\tilde{\rho})}{1 - \lambda G(\lambda)} + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda [1 - \lambda G(\lambda)]} \\ & \leq [1 + 2\lambda G(\lambda_{\max})] \left(\bar{V}(\tilde{\rho}) + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda} \right) \\ & = \bar{V}(\tilde{\rho}) + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}} + \lambda \frac{\bar{V}(\tilde{\rho})}{\lambda_{\max}} + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda}. \end{aligned}$$

The last RHS is minimum when

$$\lambda = \lambda_{\text{opt}} \triangleq \sqrt{\frac{B^2 \lambda_{\max} K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \bar{V}(\tilde{\rho})}} > 2 \sqrt{\frac{\lambda_{\max} \log(\varepsilon^{-1})}{N}}.$$

Therefore, let us take

$$\lambda_{\text{min}} \triangleq 2 \sqrt{\frac{\lambda_{\max} \log(\varepsilon^{-1})}{N}} \wedge \lambda_{\max}.$$

Introduce the event

$$E_2 = \left\{ \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \bar{V}(\tilde{\rho})} \leq \lambda_{\max} \right\}.$$

By convention, the event E_2^c contains the case when $\bar{V}(\tilde{\rho}) = 0$ ($\lambda_{\text{opt}} = +\infty$).

General case: E_2 occurs. Then we have $\lambda_{\text{opt}} \leq \lambda_{\max}$. So there exists an integer $0 \leq i \leq p$ such that $\lambda_i \leq \lambda_{\text{opt}} < 2\lambda_i$. For this integer i , we have

$$\begin{aligned} & \frac{\bar{V}(\tilde{\rho})}{1 - \lambda_i G(\lambda_i)} + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_i [1 - \lambda_i G(\lambda_i)]} \\ & \leq \bar{V}(\tilde{\rho}) + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}} + \lambda_{\text{opt}} \frac{\bar{V}(\tilde{\rho})}{\lambda_{\max}} + \frac{2B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\text{opt}}} \\ & = \bar{V}(\tilde{\rho}) + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}} + 3 \sqrt{\frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}}} \bar{V}(\tilde{\rho}) \\ & \leq \bar{V}(\tilde{\rho}) + \frac{2B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}} + 2 \sqrt{\frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}}} \bar{V}(\tilde{\rho}). \end{aligned}$$

Particular case: $(E_2)^c$ occurs. For $i = 0$, we have

$$\frac{\bar{V}(\tilde{\rho})}{1 - \lambda_i G(\lambda_i)} + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_i [1 - \lambda_i G(\lambda_i)]} = 2\bar{V}(\tilde{\rho}) + \frac{2B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}}$$

and

$$\sqrt{\frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}}} \bar{V}(\tilde{\rho}) \geq \bar{V}(\tilde{\rho}).$$

Therefore, in both subcases, there exists an integer $0 \leq i \leq p$ such that

$$\begin{aligned} & \frac{\bar{V}(\tilde{\rho})}{1 - \lambda_i G(\lambda_i)} + \frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_i [1 - \lambda_i G(\lambda_i)]} \\ & \leq \bar{V}(\tilde{\rho}) + \frac{2B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}} + 2 \sqrt{\frac{B^2 K(\tilde{\rho}, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda_{\max}}} \bar{V}(\tilde{\rho}). \end{aligned} \tag{7.10}$$

To prove the first inequation of Corollary 4.3, it remains to lower bound $\eta = \frac{1}{p+1}$ and $\zeta = \frac{1}{q+1}$.

$$\begin{cases} p = \left\lfloor \frac{\log \frac{\lambda_{\max}}{\lambda_{\min}}}{\log 2} + 1 \right\rfloor, \\ q = \left\lfloor \frac{\log \frac{\beta_{\max}}{\beta_{\min}}}{\log 2} + 1 \right\rfloor, \end{cases}$$

hence

$$\begin{cases} (\eta)^{-1} = \left\lfloor \frac{\log \frac{4\lambda_{\max}}{\lambda_{\min}}}{\log 2} \right\rfloor \leq L_1, \\ (\zeta)^{-1} = \left\lfloor \frac{\log \frac{4\beta_{\max}}{\beta_{\min}}}{\log 2} \right\rfloor \leq L_2, \end{cases}$$

where $\lfloor x \rfloor$ denotes the integer part of x .

7.4. Proof of Theorem 4.5

The result directly comes from Lemma 4.4 and Corollary 4.3 since an aggregating procedure minimizing

$$\mathbb{B}(\rho, (\lambda_i)_{i=0,\dots,p}, (\eta_i)_{i=0,\dots,p}, (\beta_j)_{j=0,\dots,q}, (\zeta_j)_{j=0,\dots,q})$$

wrt the probability distribution ρ is such that

$$\mathbb{B}(\hat{\rho}, (\lambda_i), (\eta_i), (\beta_j), (\zeta_j)) \leq \mathbb{B}(\tilde{\rho}, (\lambda_i), (\eta_i), (\beta_j), (\zeta_j)). \quad (7.11)$$

So, for any $0 < \varepsilon \leq 1/2$, with $\mathbb{P}^{\otimes N}$ -probability at least $1 - 2\varepsilon$, we have

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f\theta) - R(\tilde{f}) \leq \gamma(\varepsilon).$$

7.5. Proof of Theorem 4.7

We will first notice that the infimum of $\psi(\rho) \triangleq \frac{1}{2} \|\mathbb{E}_{\rho(d\theta)} h(\theta)\|^2 + K(\rho, \mu)$ can be searched in the set of probabilities which are equivalent to μ . It is clear that we do not change the infimum by considering only distributions absolutely continuous wrt μ . Inversely, consider ρ such that $\text{supp}(\rho)$ is strictly included $\text{supp}(\mu)$. Let $A \triangleq \text{supp}(\mu) - \text{supp}(\rho)$. We have $\rho(A) = 0$ and $\mu(A) > 0$. Our aim is then to build $\rho' \in \mathcal{M}_+^1(\Theta)$ such that $\psi(\rho') \leq \psi(\rho)$ and $\text{supp}(\rho') = \text{supp}(\mu)$. Define $\rho_A(d\theta) \triangleq \mu(\cdot/A) = \frac{\mathbb{1}_{\theta \in A}}{\mu(A)} \cdot \mu(d\theta)$ and $\rho' \triangleq \lambda \rho_A + (1 - \lambda)\rho$ for some $\lambda \in]0; 1[$ to be determined. We have

$$\begin{aligned} \psi(\rho') - \psi(\rho) &= \frac{1}{2} \|\lambda \mathbb{E}_{\rho_A} h + (1 - \lambda) \mathbb{E}_{\rho} h\|^2 + \lambda \mathbb{E}_{\rho_A} \log \frac{\lambda}{\mu(A)} + (1 - \lambda) \mathbb{E}_{\rho} \log \frac{(1 - \lambda)\rho}{\mu} \\ &\quad - \frac{1}{2} \|\mathbb{E}_{\rho} h\|^2 - \mathbb{E}_{\rho} \log \frac{\rho}{\mu} \\ &= \frac{1}{2} \|\mathbb{E}_{\rho} h\|^2 (\lambda^2 - 2\lambda) + \frac{\lambda^2}{2} \|\mathbb{E}_{\rho_A} h\|^2 + \lambda(1 - \lambda) \langle \mathbb{E}_{\rho_A} h, \mathbb{E}_{\rho} h \rangle \\ &\quad + \lambda \log[\mu(A)^{-1}] + \lambda \log \lambda + (1 - \lambda) \log(1 - \lambda) \underset{\lambda \rightarrow 0}{\sim} \lambda \log \lambda. \end{aligned}$$

Therefore, for sufficiently small λ , we have $\psi(\rho') < \psi(\rho)$.

We will now prove that for any $\rho \in \mathcal{M}_+^1(\Theta)$ equivalent to μ , there exists $z \in \mathbb{R}^N$ such that $\mathbb{E}_{\mu_{(z,h)}} h = \mathbb{E}_\rho h$. With this end in view, we introduce

$$\chi_\rho(v) = \log \mathbb{E}_\mu e^{(v,h - \mathbb{E}_\rho h)},$$

for any $v \in \mathbb{R}^N$. Let us show that χ_ρ admits a minimum. Without loss of generality, one may assume that the $h_i, i = 1, \dots, N$, are linearly independent wrt to μ , or equivalently wrt to ρ (since μ and ρ are equivalent).⁵ So, for any $z \in \mathbb{R}^N, \rho(\langle z, h \rangle - \mathbb{E}_\rho \langle z, h \rangle) > 0$, hence $\mu(\langle z, h \rangle - \mathbb{E}_\rho \langle z, h \rangle) > 0$. Introduce, for $\beta > 0$, the mappings η_β from $\mathcal{S}(0, 1) \triangleq \{u \in \mathbb{R}^N : \|u\| = 1\}$ to \mathbb{R} defined as

$$\eta_\beta(u) = \mu(\langle u, h - \mathbb{E}_\rho h \rangle > \beta).$$

We first claim that there exists β such that the mapping is lower bounded by β . Otherwise one can build a sequence $u_n \in \mathcal{S}(0, 1)$ such that $\eta_{1/n}(u_n) \geq 1/n$. Since the sphere $\mathcal{S}(0, 1)$ is compact, there exists a converging subsequence $u_{\alpha(n)}$. Denote u its limit. By Fatou's theorem, we have

$$\begin{aligned} \mu(\langle u, h - \mathbb{E}_\rho h \rangle > 0) &\leq \mathbb{E}_\mu(\liminf_{n \rightarrow +\infty} \mathbb{1}_{\langle u_n, h - \mathbb{E}_\rho h \rangle > 1/n}) \\ &\leq \liminf_{n \rightarrow +\infty} \mu(\langle u_n, h - \mathbb{E}_\rho h \rangle > 1/n) \\ &= 0, \end{aligned}$$

which is absurd. For this real β , we have

$$\chi_\rho(z) = \log \mathbb{E}_\mu e^{\|z\| \langle \frac{z}{\|z\|}, h - \mathbb{E}_\rho h \rangle} \geq \beta \|z\| + \log \beta \xrightarrow{\|z\| \rightarrow +\infty} +\infty.$$

Now, by Lebesgue's theorem, the mapping χ_ρ is continuous. Consequently, it admits a minimum which we will denote z . By differentiation under the expectation, we have $\mathbb{E}_{\mu_{(z,h)}} h - \mathbb{E}_\rho h = \nabla \chi_\rho(z) = 0$. Hence,

$$\begin{aligned} \psi(\rho) - \psi(\mu_{(z,h)}) &= K(\rho, \mu) - K(\mu_{(z,h)}, \mu) \\ &= K(\rho, \mu) - \langle z, \mathbb{E}_{\mu_{(z,h)}} h \rangle + \log \mathbb{E}_\mu e^{\langle z, h \rangle} \\ &= K(\rho, \mu_{(z,h)}) \geq 0. \end{aligned}$$

So the infimum of ψ could be searched among $\{\mu_{(z,h)} : z \in \mathbb{R}^N\}$.

Now, let $(z'_n)_{n \in \mathbb{N}}$ be a sequence of \mathbb{R}^N such that

$$\psi(\mu_{(z'_n,h)}) \xrightarrow{n \rightarrow +\infty} \inf_{\mathcal{M}_+^1(\Theta)} \psi. \tag{7.12}$$

Let $p_{\{x_1, \dots, x_m\}^\perp}$ denote the orthogonal projection into the orthogonal of the system $\{x_1, \dots, x_m\}$ (by convention, $p_{\emptyset^\perp} \triangleq \text{Id}_{\mathbb{R}^N}$). By compacity of the sphere $\mathcal{S}(0, 1)$, there exists a subsequence $(z_n)_{n \in \mathbb{N}}$ such that there exists $L \in \{1, \dots, N\}$ and an orthonormal system $\mathcal{V}_L \triangleq \{v_1, \dots, v_L\}$ satisfying

$$\frac{p_{\{v_1, \dots, v_{l-1}\}^\perp}(z_n)}{\|p_{\{v_1, \dots, v_{l-1}\}^\perp}(z_n)\|} \xrightarrow{n \rightarrow +\infty} v_l$$

for any $l \in \{1, \dots, L\}$ and $z_n \in \text{Span}(v_1, \dots, v_L)$. Let $(\lambda_{n,l})_{l=1, \dots, L}$ denote the components of z_n in the system \mathcal{V}_L : $z_n = \sum_{l=1}^L \lambda_{n,l} v_l$. By definition of the system \mathcal{V}_L , we have $\lambda_{n,1} \gg \lambda_{n,2} \gg \dots \gg \lambda_{n,L}$, where $a_n \gg b_n$ means that $b_n = o(a_n)$. Even if it means to consider a subsequence of $(z_n)_{n \in \mathbb{N}}$, one can assume that for any $l \in \{1, \dots, L\}$,

⁵ For $h = \text{Cst}$ μ -a.s., the result is trivial.

$\lambda_{n,l} \rightarrow_{n \rightarrow +\infty} \lambda_l \in \mathbb{R}_+ \cup \{+\infty\}$. Let $\lambda_0 \triangleq +\infty$ and $L' \triangleq \max\{l \in \{0, \dots, L\}: \lambda_l = +\infty\}$. Introduce the following family of subsets of Θ :

$$\begin{cases} \tilde{A}_0 \triangleq \Theta, \\ \tilde{A}_l \triangleq \{\theta \in \tilde{A}_{l-1}: \langle v_l, h(\theta) \rangle = \operatorname{ess\,sup}_{\mu(\cdot/\tilde{A}_{l-1})} \langle v_l, h \rangle\}, \end{cases}$$

where

$$\mu(\cdot/\tilde{A}_{l-1}) \triangleq \frac{\mathbb{1}_{\tilde{A}_{l-1}}}{\mu(\tilde{A}_{l-1})} \cdot \mu$$

makes sense since one can prove (by induction and using that $\limsup_{n \rightarrow +\infty} K(\mu_{\langle z_n, h \rangle}, \mu) < +\infty$) that $\mu(\tilde{A}_{l-1}) > 0$. Then, one can prove that $\mu_{\langle \lambda_{L'+1} v_{L'+1}, h \rangle}(\cdot/\tilde{A}_{L'})$ minimizes ψ (where $\lambda_{L'+1} v_{L'+1} \triangleq 0$ when $L' = L$). Now, we have necessarily $L' = 0$. Indeed, if $L' > 0$, from the linear independency of the functions $h_i, i = 1, \dots, N$, we have $\mu(\tilde{A}_{L'}) < 1$, hence, the optimal distribution is not equivalent to μ . This is in contradiction with what we proved at the beginning of this section.

So the function $\varphi: z \mapsto \psi(\mu_{\langle z, h \rangle})$ admits a minimum denoted $\bar{z} = \lambda_1 v_1$. Let $\bar{\rho} \triangleq \mu_{\langle \bar{z}, h \rangle}$. By differentiation under the expectation, $\nabla \varphi(z) = \nabla \operatorname{Var}_{\mu_{\langle z, h \rangle}} h(\mathbb{E}_{\mu_{\langle z, h \rangle}} h + z)$, where $\nabla \operatorname{Var}_{\mu_{\langle z, h \rangle}} h$ denotes the covariance matrix of the $h_i, i = 1, \dots, N$, wrt $\mu_{\langle z, h \rangle}$. Since the functions $h_i, i = 1, \dots, N$, are linearly independent wrt to $\mu_{\langle z, h \rangle}$, the matrix $\nabla \operatorname{Var}_{\mu_{\langle z, h \rangle}} h$ is invertible. Therefore, we have $\bar{z} = -\mathbb{E}_{\bar{\rho}} h$. It remains to prove the uniqueness. It follows from the following equality which holds for any $\rho \in \mathcal{M}_+^1(\Theta)$ and comes from $\bar{\rho} = \mu_{-\langle \mathbb{E}_{\bar{\rho}} h, h \rangle}$:

$$\begin{aligned} \psi(\rho) - \psi(\bar{\rho}) &= \frac{1}{2} \|\mathbb{E}_{\rho} h\|^2 + K(\rho, \mu) - \frac{1}{2} \|\mathbb{E}_{\bar{\rho}} h\|^2 - K(\bar{\rho}, \mu) \\ &= \frac{1}{2} \|\mathbb{E}_{\rho} h\|^2 + K(\rho, \bar{\rho}) - \langle \mathbb{E}_{\bar{\rho}} h, \mathbb{E}_{\rho} h \rangle - \log \mathbb{E}_{\mu} e^{-\langle \mathbb{E}_{\bar{\rho}} h, h \rangle} - \frac{1}{2} \|\mathbb{E}_{\bar{\rho}} h\|^2 - \log \mathbb{E}_{\mu} e^{\langle \mathbb{E}_{\bar{\rho}} h, h - \mathbb{E}_{\bar{\rho}} h \rangle} \\ &= K(\rho, \bar{\rho}) + \frac{1}{2} \|\mathbb{E}_{\rho} h - \mathbb{E}_{\bar{\rho}} h\|^2. \end{aligned}$$

7.6. Proof of Theorem 4.8

For any $w, w' \in \mathbb{R}^N$, we have

$$\begin{aligned} \frac{\bar{\varphi}(w) - \bar{\varphi}(w')}{ac} &= d_2(\|\mathbb{E}_{\pi^w} f(X) - Y\|^2 - \|\mathbb{E}_{\pi^{w'}} f(X) - Y\|^2) \\ &\quad + \log \mathbb{E}_{\pi_{-\frac{1}{2}r(f)}} e^{\langle w', f(X) - \mathbb{E}_{\pi^{w'}} f(X) \rangle} - \log \mathbb{E}_{\pi_{-\frac{1}{2}r(f)}} e^{\langle w, f(X) - \mathbb{E}_{\pi^w} f(X) \rangle} \\ &= d_2(\|\mathbb{E}_{\pi^w} f(X) - Y\|^2 - \|\mathbb{E}_{\pi^{w'}} f(X) - Y\|^2) - \langle w', \mathbb{E}_{\pi^{w'}} f(X) - Y \rangle \\ &\quad + \langle w, \mathbb{E}_{\pi^w} f(X) - Y \rangle + \log \mathbb{E}_{\pi_{-\frac{1}{2}r(f)}} e^{\langle w', f(X) - Y \rangle} - \log \mathbb{E}_{\pi_{-\frac{1}{2}r(f)}} e^{\langle w, f(X) - Y \rangle} \\ &= d_2(\|\mathbb{E}_{\pi^w} f(X) - Y\|^2 - \|\mathbb{E}_{\pi^{w'}} f(X) - Y\|^2) - \langle w', \mathbb{E}_{\pi^{w'}} f(X) - Y \rangle \\ &\quad + \langle w, \mathbb{E}_{\pi^w} f(X) - Y \rangle + K(\pi^w, \pi^{w'}) \\ &= d_2(\|\mathbb{E}_{\pi^w} f(X) - Y\|^2 - \|\mathbb{E}_{\pi^{w'}} f(X) - Y\|^2 - 2\langle \mathbb{E}_{\pi^{w'}} f(X) - Y, \mathbb{E}_{\pi^w} f(X) - \mathbb{E}_{\pi^{w'}} f(X) \rangle) \\ &\quad + \langle w' + 2d_2(\mathbb{E}_{\pi^{w'}} f(X) - Y), \mathbb{E}_{\pi^w} f(X) - \mathbb{E}_{\pi^{w'}} f(X) \rangle + K(\pi^w, \pi^{w'}) \\ &= d_2\|\mathbb{E}_{\pi^w} f(X) - \mathbb{E}_{\pi^{w'}} f(X)\|^2 + K(\pi^w, \pi^{w'}) \\ &\quad + \langle w' + 2d_2(\mathbb{E}_{\pi^{w'}} f(X) - Y), \mathbb{E}_{\pi^w} f(X) - \mathbb{E}_{\pi^{w'}} f(X) \rangle. \end{aligned}$$

The second inequality of the theorem is obtained by choosing $w = \bar{w} \triangleq -\mathbb{E}_{\bar{\rho}}h$ and $w' = w^l$ and by using assumption (2.1).

7.7. Proof of the exit of the “While” loop

The w^{l+1} tested by the loop are

$$w^{l+1} = w^l - \alpha z^l,$$

where

$$z^l \triangleq w^l - 2d_2 \left(Y - \mathbb{E}_{\pi^{w^l}} f(X) + \sum_{i=r+1}^N \alpha^i [Y_i - \langle \alpha^i, \mathbb{E}_{\pi^{w^l}} f(X) \rangle_r - \beta^i] \right)$$

and $\alpha \in \{\frac{1}{2^n} : n \in \mathbb{N}\}$. We have

$$\nabla_r \bar{\varphi}(w^l) = ac \nabla \text{Var}_{\pi^{w^l}} f(X) |_r z^l$$

hence

$$\begin{aligned} \bar{\varphi}(w^{l+1}) - \bar{\varphi}(w^l) &= \langle w^{l+1} - w^l, \nabla \bar{\varphi}(w^l) \rangle + o(\|w^{l+1} - w^l\|) \\ &= -ac\alpha(z^l)' \nabla \text{Var}_{\pi^{w^l}} f(X) |_r z^l + o(\alpha). \end{aligned}$$

The covariance matrix $\text{Var}_{\pi^{w^l}} f(X) |_r$ is definite positive by definition of r . So there exists $\alpha \in \{\frac{1}{2^n} : n \in \mathbb{N}\}$ such that $\bar{\varphi}(w^l - \alpha z^l) - \bar{\varphi}(w^l) < 0$.

7.8. Proof of Corollary 4.9

To deduce Corollary 4.9 from Corollary 4.3, we need to control the deviations of the empirical risk $r(\tilde{f})$ of the best convex combination. We begin with the following deviation inequality.

Lemma 7.5. *Let Z be a positive random variable. If $\mathbb{E}e^{\alpha\sqrt{Z}} \leq M'$ for some $\alpha > 0$ and $M' > 0$, then, for any $\eta \geq 0$ and $A \geq (2/\alpha)^2$,*

$$\log \mathbb{E}e^{\eta(\mathbb{E}Z - Z)} \leq \eta M' A e^{-\alpha\sqrt{A}} + \frac{\eta^2}{2} A \mathbb{E}Z.$$

Proof. For any $A \geq (2/\alpha)^2$,

$$\begin{aligned} \mathbb{E}Z - Z &\leq \mathbb{E}(Z\mathbb{1}_{Z \geq A}) + \mathbb{E}(Z\mathbb{1}_{Z < A}) - Z\mathbb{1}_{Z < A} \\ &\leq \mathbb{E}\left(e^{\alpha\sqrt{Z}} \sup_{u \geq A} ue^{-\alpha\sqrt{u}}\right) + \mathbb{E}(Z\mathbb{1}_{Z < A}) - Z\mathbb{1}_{Z < A} \\ &\leq M' A e^{-\alpha\sqrt{A}} + \mathbb{E}(Z\mathbb{1}_{Z < A}) - Z\mathbb{1}_{Z < A} \end{aligned}$$

since the mapping $[u \mapsto ue^{-\alpha\sqrt{u}}]$ is decreasing on $[(2/\alpha)^2; +\infty[$. Applying the previous deviation inequality to $Z\mathbb{1}_{Z < A} \in [0; A]$, we obtain

$$\log \mathbb{E}e^{\eta(\mathbb{E}Z - Z)} \leq \eta M' A e^{-\alpha\sqrt{A}} + \frac{\eta^2}{2} A \mathbb{E}Z. \quad \square$$

The deviations of the empirical risk of the best mixture \tilde{f} are given by

Lemma 7.6. For any $\varepsilon \geq e^{-\kappa_3 N}$, we have

$$\mathbb{P}^{\otimes N} \left[R(\tilde{f}) - r(\tilde{f}) > \tilde{L}^2 \sqrt{\frac{2\log(\varepsilon^{-1})R(\tilde{f})}{\alpha^2 N}} \right] \leq \varepsilon, \tag{7.13}$$

where

$$\tilde{L} \triangleq \log \left(M e^{\alpha B + 1} \sqrt{\frac{N}{2\log(\varepsilon^{-1})\alpha^2 R(\tilde{f})}} \right)$$

and

$$\kappa_3 \triangleq \frac{M^2 e^{2(\alpha B - 1)}}{2[(\alpha B e)^2 + 4M]}.$$

Proof. For any $\lambda > 0$ and any $\mu \in \mathbb{R}$,

$$\mathbb{P}^{\otimes N} (R(\tilde{f}) - r(\tilde{f}) > \mu) \leq \mathbb{E}_{\mathbb{P}^{\otimes N}} e^{N\lambda(R(\tilde{f}) - r(\tilde{f}) - \mu)} \leq e^{-N\lambda\mu} (\mathbb{E}_{\mathbb{P}} e^{\lambda(\mathbb{E}Z - Z)})^N,$$

where $Z \triangleq (Y - \tilde{f}(X))^2 \geq 0$. We have

$$\mathbb{E}_{\mathbb{P}} e^{\alpha\sqrt{Z}} = \mathbb{E}_{\mathbb{P}} e^{\alpha|Y - \tilde{f}(X)|} \leq \mathbb{E}_{\mathbb{P}} e^{\alpha(|Y - \mathbb{E}_{\mathbb{P}}(Y/X)| + |\mathbb{E}_{\mathbb{P}}(Y/X) - \tilde{f}(X)|)} \leq M e^{\alpha B} \triangleq M'. \tag{7.14}$$

From the previous lemma, we get for any $A \geq (2/\alpha)^2$,

$$\mathbb{P}^{\otimes N} (R(\tilde{f}) - r(\tilde{f}) > \mu) \leq \exp \left\{ -N\lambda\mu + N\lambda M' A e^{-\alpha\sqrt{A}} + N \frac{\lambda^2}{2} A R(\tilde{f}) \right\} \leq \varepsilon,$$

when $\mu = \frac{\log(\varepsilon^{-1})}{N\lambda} + M' A e^{-\alpha\sqrt{A}} + \frac{\lambda}{2} A R(\tilde{f})$. The previous inequality holds for any $\lambda > 0$ and $A \geq (2/\alpha)^2$. To get a small μ , we take $\lambda = \sqrt{\frac{2\log(\varepsilon^{-1})}{ANR(\tilde{f})}}$ (when $R(\tilde{f}) \neq 0$; otherwise the result is trivial) and $A = ((\tilde{L} - 1)/\alpha)^2$. To fulfill the condition $A \geq (2/\alpha)^2$, we need that ε should be not too small. More precisely, the condition $(\tilde{L} - 1)^2 \geq 4$ is satisfied when

$$\log \left(M e^{\alpha B + 1} \sqrt{\frac{N}{2\log(\varepsilon^{-1})\alpha^2 R(\tilde{f})}} \right) \geq 3,$$

equivalently,

$$2M^2 e^{2\alpha B} \frac{N}{2\log(\varepsilon^{-1})\alpha^2 R(\tilde{f})} \geq e^4,$$

$$\frac{M^2 e^{2\alpha B - 4}}{2\alpha^2 R(\tilde{f})} N \geq \log(\varepsilon^{-1}).$$

Now, from inequality (4.6), the expected risk of any function in the model $\tilde{\mathcal{R}}$ is bounded by κB^2 where $\kappa \triangleq \frac{4M}{e^{2(\alpha B)^2}} + 1$. Therefore, for any $\varepsilon \geq e^{-\kappa_3 N}$, we have $(\tilde{L} - 1)^2 \geq 4$ as required. \square

From Corollary 4.3, using that $r(\tilde{f}) \geq \inf_{\tilde{\mathcal{R}}} r$, we have

$$R(\tilde{f}) \leq R(\mathbb{E}_{\hat{\rho}(d\theta)} f_{\theta}) \leq R(\tilde{f}) - r(\tilde{f}) + r(\mathbb{E}_{\hat{\rho}(d\theta)} f_{\theta}) + \mathbb{B}',$$

where

$$\left\{ \begin{aligned} \mathbb{B}' &\triangleq \inf_{\substack{i \in I \\ j \in J}} \mathbb{B}'(\rho, \lambda_i, \eta_i, \beta_j, \zeta_j), \\ \mathbb{B}'(\rho, \lambda, \eta, \beta, \zeta) &\triangleq \frac{\lambda G(\lambda)}{1 - \lambda G(\lambda)} [\mathbb{E}_{\rho(d\theta)} r(f_\theta) - \inf_{\tilde{\mathcal{R}}} r] + \frac{B^2 K(\rho, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda [1 - \lambda G(\lambda)]} \\ &\quad + \frac{\beta g(\beta)}{1 + \beta g(\beta)} \bar{V} + \frac{B^2 2K(\rho, \pi) + \log[(\zeta\varepsilon)^{-1}]}{2N \beta [1 + \beta g(\beta)]} \\ &= \frac{\lambda G(\lambda)}{1 - \lambda G(\lambda)} [r(\mathbb{E}_{\rho(d\theta)} f_\theta) - \inf_{\tilde{\mathcal{R}}} r] + \frac{B^2 K(\rho, \pi) + \log[(\eta\varepsilon)^{-1}]}{N \lambda [1 - \lambda G(\lambda)]} \\ &\quad + \left(\frac{\lambda G(\lambda)}{1 - \lambda G(\lambda)} + \frac{\beta g(\beta)}{1 + \beta g(\beta)} \right) \bar{V} + \frac{B^2 2K(\rho, \pi) + \log[(\zeta\varepsilon)^{-1}]}{2N \beta [1 + \beta g(\beta)]}. \end{aligned} \right.$$

Then, using Lemma 7.6, we obtain that with probability at least $1 - 3\varepsilon$,

$$R(\tilde{f}) \leq R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) \leq \tilde{L}^2 \sqrt{\frac{2 \log(\varepsilon^{-1}) R(\tilde{f})}{\alpha^2 N}} + r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'.$$

Now, using simple computations, one can show that a positive number x such that $x \leq 2c\sqrt{x} + a$ for some $a, c > 0$ satisfies $\sqrt{x} \leq c + \sqrt{a + c^2}$. Applying this result for $x = R(\tilde{f})$, $a = r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'$ and $c = \tilde{L}^2 \sqrt{\frac{\log(\varepsilon^{-1})}{2\alpha^2 N}}$, we get

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) \leq \tilde{L}^2 \sqrt{\frac{2 \log(\varepsilon^{-1})}{\alpha^2 N}} (c + \sqrt{a + c^2}) + a.$$

The remaining unobservable term in this bound is \tilde{L} which depends on $R(\tilde{f})$. We will consider two cases:

General case: $R(\tilde{f}) \geq \frac{4}{\kappa_1} \frac{\log(\varepsilon^{-1})}{N} B^2$ occurs. The constant $\frac{4}{\kappa_1}$ in this threshold is arbitrary (it has been chosen since it looks like the second term in \mathbb{B}'). Then we have

$$\tilde{L} \leq \log \left(\frac{M e^{\alpha B + 1}}{\alpha B} \sqrt{\frac{\kappa_1}{8}} \frac{N}{\log(\varepsilon^{-1})} \right),$$

hence

$$\tilde{L}^2 \sqrt{\frac{2 \log(\varepsilon^{-1})}{\alpha^2 N}} \leq 2\mathcal{L} \sqrt{\frac{\log(\varepsilon^{-1})}{N}},$$

where

$$\mathcal{L} \triangleq \frac{1}{\sqrt{2\alpha}} \left[\log \left(\kappa_4 \frac{N}{\log(\varepsilon^{-1})} \right) \right]^2 \quad \text{and} \quad \kappa_4 \triangleq \frac{M e^{\alpha B + 1}}{\alpha B} \sqrt{\frac{\kappa_1}{8}}.$$

This leads to the desired result.

Particular case: $R(\tilde{f}) < \frac{4}{\kappa_1} \frac{\log(\varepsilon^{-1})}{N} B^2$ occurs. From Corollary 4.3, with probability at least $1 - 2\varepsilon$, we have

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) \leq r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}' + \frac{4}{\kappa_1} \frac{\log(\varepsilon^{-1})}{N} B^2.$$

The announced inequality is also true in this case.

Acknowledgements

I would like to thank Professor Olivier Catoni – my PhD advisor – for his constant and remarkable kindness, availability and support without which this work would not have been achieved.

References

- [1] L. Breiman, Bagging predictors, *Machine Learning* 24 (2) (1996) 123–140.
- [2] L. Breiman, Arcing classifiers, *Ann. Statist.* 26 (3) (1998) 801–849.
- [3] O. Catoni, *Statistical Learning Theory and Stochastic Optimization*, in: *Probability Summer School, Saint Flour, 2001*, Springer-Verlag, submitted for publication.
- [4] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: *Machine Learning: Proceedings of the Thirteenth International Conference, 1996*, pp. 148–156.
- [5] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Technical Report, Dept. of Statistics, Stanford University, 1998.
- [6] A. Juditsky, A. Nemirovski, Functional aggregation for nonparametric estimation, *Ann. Statist.* 28 (2000) 681–712.
- [7] E. Mammen, A.B. Tsybakov, Smooth discrimination analysis, *Ann. Statist.* 27 (1999) 1808–1829.
- [8] D.A. McAllester, PAC-bayesian stochastic model selection, *Machine Learning J.* (2001), submitted for publication.
- [9] A. Nemirovski, *Lectures on Probability Theory and Statistics. Part II: Topics in Non-Parametric Statistics*, in: *Probability Summer School, Saint Flour, Springer-Verlag, Berlin, 1998*.
- [10] G. Rätsch, M. Warmuth, S. Mika, T. Onoda, S. Lemm, K.-R. Müller, Barrier boosting, in: *Proc. COLT'00, Morgan Kaufmann, Palo Alto, 2000*, pp. 170–179.
- [11] R.E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, 1998, pp. 80–91.
- [12] A.B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, 2001.
- [13] Y. Yang, Aggregating regression procedures for a better performance, 2001.