THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS VI

Spécialité : Mathématiques

Thèse soumise au
Laboratoire de Probabilités et Modèles Aléatoires (U.M.R. CNRS 7599)
pour l'obtention du diplôme de
DOCTEUR DE L'UNIVERSITÉ PARIS VI
réalisée par

Jean-Yves AUDIBERT

Théorie Statistique de l'Apprentissage: une approche PAC-Bayésienne

PAC-Bayesian Statistical Learning Theory

Thèse soutenue le 29 juin 2004 devant le jury composé de

Lucien BIRGÉ	Examinateur
Olivier CATONI	Directeur de thèse
Gérard KERKYACHARIAN	Examinateur
Vladimir KOLTCHINSKII	Rapporteur
Dominique PICARD	Examinateur
Alain TROUVÉ	Rapporteur
Alexandre TSYBAKOV	Examinateur

Résumé. Cette thèse a pour objet l'étude et la conception d'algorithmes d'apprentissage, notamment pour la classification de données et la régression aux moindres carrés. Elle regroupe quatre articles. Le premier fournit une borne PAC-bayésienne sur l'erreur de régression aux moindres carrés qui est valable pour toute procédure d'agrégation. La minimisation de cette borne, qui est numériquement réalisable, conduit à un estimateur ayant la vitesse de convergence optimale au sens minimax.

Le deuxième article est le cœur de la thèse. Il présente de nouvelles bornes PAC-bayésiennes en classification et déduit de ces bornes des algorithmes originaux reposant, d'une part, sur les schémas de compression et, d'autre part, sur les lois de Gibbs.

La troisième partie illustre les bornes introduites dans la deuxième et montre l'influence de la distribution a priori sur la qualité des estimateurs de Gibbs. Ce travail discute également de manière approfondie les hypothèses de complexité et de marge proposées par Mammen et Tsybakov (E. Mammen and A.B. Tsybakov, Smooth discrimination analysis, Ann. Stat., 27, 1808–1829, 1999).

Enfin, le dernier article a pour but d'unifier les bornes existantes sur l'erreur de généralisation en classification. La borne proposée permet notamment d'établir un lien entre les complexités PAC-bayésiennes et les nombres de Rademacher.

Mots-clés. Théorie statistique de l'apprentissage, borne PAC-bayésienne, classification, régression aux moindres carrés, mesure empirique de complexité, mélange, combinaison convexe, estimateur randomisé, loi de Gibbs, schéma de compression, estimation adaptative, estimation non-paramétrique, inégalité de déviation, borne sur l'erreur de généralisation, théorie de Vapnik-Chervonenkis, hypothèse entropique, hypothèse de marge, borne sur le risque, chaînage, inégalité oracle, algorithme de "boosting".

Abstract. This PhD thesis is a mathematical study of the learning task – specifically classification and least square regression – in order to better understand why an algorithm works and to propose more efficient procedures. The thesis consists in four papers. The first one provides a PAC bound for the L^2 generalization error of methods based on combining regression procedures. This bound is tight to the extent that, for an appropriate aggregation procedure, we recover known optimal convergence rates. Besides, it is numerically tractable to derive an optimal aggregating procedure from the bound.

The second paper is the core of the thesis. It provides new PAC-Bayesian bounds in classification and put forward original algorithms based on compression schemes and Gibbs distributions.

The third paper illustrates the bounds developed in the second one and shows the influence of the prior distribution on the efficiency of Gibbs classifiers. It also discusses the complexity and margin assumptions proposed by Mammen and Tsybakov (E. Mammen and A.B. Tsybakov, Smooth discrimination analysis, Ann. Stat., 27, 1808–1829, 1999).

The fourth paper aims to unify the numerous generalization error bounds which have appeared these last decades. It makes in particular the link between Rademacher and PAC-Bayesian bounds.

Key words and phrases. Statistical learning theory, PAC-Bayesian bound, classification, least square regression, empirical complexity, mixture, convex aggregation, randomized estimator, Gibbs classifier, compression scheme, adaptive estimation, nonparametric estimation, deviation inequality, generalization error bound, VC theory, entropy assumption, margin assumption, risk bound, chaining, model selection, oracle inequality, boosting algorithm.

Acknowledgements

During my Masters, I had the opportunity to attend tremendous lectures from Jean Jacod, Annie Millet, Mark Yor, Dominique Picard, Fabienne Comte, Gérard Kerkyacharian, Gábor Lugosi and Olivier Catoni. I would like first to thank them all for their enthusiasm and the quality of their lectures.

I am very grateful to Olivier Catoni, my PhD advisor, for his constant and remarkable kindness, availability and support without which this work would not have been achieved.

I specially thank the following people for the time I spent with them and the seminars they have organized: Alexandre Tsybakov, Pascal Massart, Gilles Blanchard, Stéphane Boucheron and Nicolas Vayatis.

I am very grateful to the Direction du Personnel et des Services of the Ministère de l'Equipement, du Transport et du Logement for having accepted to support this work.

I am also very indebted to Professor Bernhard Schölkopf for inviting me at Max Planck Institute and to Olivier Bousquet for numerous discussions and hours of work together.

I would also like to thank my friends Ashkan Nikeghbali and Fulvio Pegoraro for both the time we have spent together and the thrilling discussions we have had. Finally, I thank my family, and in particular my parents, for their nonstop support.

Table of Contents

Introduction

Statistical Learning Theory is a research field devoted to the statistical analysis of algorithms for making predictions about the future based on past experiences. Since the pioneering work of Vapnik and Chervonenkis, the number of researchers working on this problem have known an exponential growth. A key reason for this development is that, with the computer revolution, we are able to collect huge complex data sets in many domains: bioinformatics, insurance, finance and so on, and many different tasks such as image processing, speech recognition, pattern recognition, requires efficient feasable algorithms.

Another reason for the development of the field is that the initial theoretical problem is ill-posed. Namely, we know that there is no uniformly consistent algorithm: without any knowledge of the probability generating the data (even when the data are assumed to be independent and identically distributed), there is no algorithm that guarantees to tend to the best possible prediction function with a given rate. This "no-free-lunch" theorem implies that we need to make assumptions (such as the unknown probability distribution is in some known large set of distributions) and/or to be less ambitious and to change the initial target into attempting to do as well as the best function in a given subset of prediction functions -called the model. So the underlying question is: what are these sets? Both sets depends intimately on the nature of the prediction task and on the way the data are represented. In other words, there is no general theory which will suit for any sets of data.

For a decade, several almost off-the-shelf efficient algorithms have arisen. The favourite ones are Support Vector Machines and Boosting algorithms. A less recent algorithm, the Neural Networks, is still much used by practitioners but requires much more knowledge to be properly implemented. For "reasonable" sets of data, these three classifiers predict rather accurately. However there is still often a gain to preprocess complex data having a peculiar form. This preprocessing step, as the selection of a limited number of features, is a common way to embody a prior knowledge on the underlying phenomenon.

This thesis provides a mathematical study of learning tasks – particularly, classification – in order to better understand algorithms and derive more efficient estimators. In supervised learning, we dispose of a set of training examples

$$Z_1^N \triangleq \big\{ Z_i \triangleq (X_i,Y_i) : X_i \in \mathcal{X}, Y_i \in \mathcal{Y}, i = 1,\dots,N \big\},$$

where $\mathcal{X}$ is some set of inputs (also called patterns, cases, instances or observations) and $\mathcal{Y}$ is some set of outputs (or targets). When $\mathcal{Y}$ is finite, the learning task is called classification (or pattern recognition). When $\mathcal{Y}$ is the real line, it is called regression.

In Statistical Learning Theory, we assume that the examples are generated independently from some unknown but fixed probability distribution $\mathbb{P}$. The goal is to construct a prediction function $f : \mathcal{X} \to \mathcal{Y}$ (also called decision function, hypothesis, estimator, procedure or algorithm[1]) based on the training set that minimizes

[1] The last three terms refer more to the way the prediction function is chosen.

the expected risk (or generalization error) defined as

$$R(f) \triangleq \mathbb{E}_{\mathbb{P}(dX,dY)} L[Y,f(X)],$$

where $L : \mathcal{Y} \times \mathcal{Y}$ is a dissimilarity measure on $\mathcal{Y}$. The classification task uses the Hamming distance $L(y,y') \triangleq \mathbb{1}_{y \neq y'}$, whereas L^2 regression looks at the square of the difference: $L(y,y') \triangleq (y-y')^2$. Since the probability distribution $\mathbb{P}$ is unknown, we need to estimate the expected risk R in order to assess the efficiency of a prediction function. This is done through the empirical risk (or empirical error)

$$r(f) \triangleq \mathbb{E}_{\bar{\mathbb{P}}(dX,dY)} L[Y,f(X)],$$

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

We can very easily choose the prediction function such that it fits perfectly all training points (provided that the inputs in the training set are pairwise distinct). However this is not sufficient to guarantee a small generalization error. This phenomenon is called overfitting. To avoid it, we need to restrict the class of functions on which the empirical error is minimized in order to have some guarantee on the efficiency of the algorithm.

This restriction to a prescribed set of functions – called the model – can lead to underfitting, i.e. to an estimator which has high empirical and expected risks. Therefore, to choose the adequate size (also called capacity or complexity) of the model is a key problem to build consistently efficient estimators. Vapnik-Chervonenkis theory considers that a huge model can be seen as the limit set of a nested sequence of subsets (or submodels) having increasing complexities. This sequence gives a structure on the model (and to some extent we expect that small submodels contain the best function in the model). Here again, it is a hidden way of incorporating prior knowledge.

This thesis collects four papers. A common point of these works is the way the model is structured. Unlike Vapnik-Chervonenkis work and its model selection approach through Structural Risk Minimization, the PAC-Bayesian approach proposes to structure the model by putting a prior distribution on it. We believe that viewing the model through the prior distribution is finer. This prior distribution has not the same meaning as in Bayesian learning since it does not represent the frequency according to which we expect to observe data produced by different probability distributions. It is a way of representing the model which is tightly related to the Minimum Description Length approach of Rissanen.

This thesis mainly concentrates on classification. However the first[2] of the four papers forming this thesis also concerns least square regression. It has been inspired by the success of boosting and by questions about convex aggregation of d regression functions. The main result of this work is to provide a tight PAC bound for the L^2 generalization error of methods based on combining regression procedures and to show how this bound can be used to build an adaptive estimator.

Specifically, let $\mathcal{R}$ be a class of regression functions indexed by a parameter $\theta \in \Theta$ $\big($i.e. $\mathcal{R} \triangleq \{f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta\}\big)$. Let $\mathcal{M}_+^1(\Theta)$ denote the set of probability distributions on the parameter set and $\pi \in \mathcal{M}_+^1(\Theta)$ be a prior distribution. In this

[2] It is a slightly revised version of the paper accepted by the Annales de l'Institut Henri Poincaré.

work, the model is the set of mixtures[3]

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

When the set Θ is finite, $\mathcal{C}(\mathcal{R})$ is just the set of all possible convex combinations of functions in $\mathcal{R}$.

In L^2 regression, the best prediction function is the conditional expectation $f^*_{\mathbb{P}} : x \mapsto \mathbb{E}_{\mathbb{P}}(Y/X = x)$. Let us assume that

- for any functions f, g in $\mathcal{R} \cup \{f^*_{\mathbb{P}}\}$, for any $x \in \mathcal{X}$,
$$|f(x) - g(x)| \le B.$$
- there exists $\alpha > 0$, $M > 0$ such that for any $x \in \mathcal{X}$,
$$\mathbb{E}_{\mathbb{P}(dY)} \exp\big(\alpha|Y - f^*_{\mathbb{P}}(X)|/X = x\big) \le M.$$

The constants α, B and M are assumed to be known by the statistician. Let $\tilde{f} \triangleq \operatorname{argmin}_{f\in\mathcal{C}(\mathcal{R})} R(f)$ denote the best mixture and let $\tilde{\rho} \in \mathcal{M}_+^1(\Theta)$ be such that $\tilde{f} = \mathbb{E}_{\tilde{\rho}(d\theta)} f_\theta$. The complexity of a mixture ρ is defined as the Kullbach-Leibler divergence $K(\rho,\pi) \triangleq \mathbb{E}_{\rho(d\theta)} \log \frac{\rho}{\pi}(\theta)$ when ρ is absolutely continuous wrt π and $K(\rho,\pi) \triangleq +\infty$ otherwise. The main result is:

Theorem 1. *There exist positive constants C_1 and C_2 (which can be explicited in terms of α, B and M) such that for any $\epsilon > 0$ and $0 < \lambda \le C_1$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any probability distribution $\rho \in \mathcal{M}_+^1(\Theta)$, we have*

$$\begin{aligned} R\big(\mathbb{E}_{\rho(d\theta)} f_\theta\big) - R(\tilde{f}) \quad &\le \quad (1+\lambda)\big[r\big(\mathbb{E}_{\rho(d\theta)} f_\theta\big) - r(\tilde{f})\big] \\ &\qquad + 2\lambda \mathbb{E}_{\bar{\mathbb{P}}(dX)} \mathbb{V}ar_{\rho(d\theta)} f_\theta(dX) + C_2 \tfrac{K(\rho,\pi)+\log(\epsilon^{-1})}{N\lambda}. \end{aligned}$$

Define $\hat{\rho}_\lambda$ as the minimizer of

$$\mathbb{B}_\lambda(\rho) \triangleq (1+\lambda) r\big(\mathbb{E}_{\rho(d\theta)} f_\theta\big) + 2\lambda \mathbb{E}_{\bar{\mathbb{P}}(dX)} \mathbb{V}\mathrm{ar}_{\rho(d\theta)} f_\theta(dX) + C_2 \frac{K(\tilde{\rho},\pi)}{N\lambda}.$$

Let Λ be a geometric grid of $\big[\frac{C_1}{\sqrt{N}}; C_1\big]$, $\mathcal{K} \triangleq \frac{K(\rho,\pi)+\log[\log(C_3 N)\epsilon^{-1}]}{N}$ for an appropriate constant C_3 (depending on C_1 and the radius of the grid) and let C denote a constant (possibly depending on α, B and M). From the previous result, we have

- For any $\epsilon > 0$, taking $\hat{\lambda}$ minimizing
$$\mathbb{B}_\lambda(\rho_\lambda) - (1+\lambda) \min_{\mathcal{C}(\mathcal{R})} r + C_2 \frac{\log[\log(C_3 N)\epsilon^{-1}]}{N\lambda}$$
over the grid Λ, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have
$$R(\mathbb{E}_{\hat{\rho}_{\hat{\lambda}}(d\theta)} f_\theta) - R(\tilde{f}) \le C\sqrt{\mathcal{K}},$$
- by cutting the training set into two pieces, building $\hat{\rho}_\lambda$ on the first sample only, and taking $\hat{\lambda}$ as the minimizer of the empirical error on the second sample of $\mathbb{E}_{\hat{\rho}_\lambda(d\theta)} f_\theta$ over the grid Λ, for any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,
$$R(\mathbb{E}_{\hat{\rho}_{\hat{\lambda}}(d\theta)} f_\theta) - R(\tilde{f}) \le C\Big(\sqrt{\mathcal{K}\mathbb{E}_{\mathbb{P}(dX)} \mathbb{V}\mathrm{ar}_{\tilde{\rho}(d\theta)} f_\theta(X)} \vee \mathcal{K}\Big).$$

[3] Here, the prior distribution is not put on the model but on the underlying set of functions $\mathcal{R}$. In the other papers of the thesis, the prior will effectively be on the model.

In particular, when Θ is finite with cardinal d, by taking a uniform prior distribution, we obtain:

Corollary 2. *The two-step procedure is adaptive to the extent that when the best convex combination belongs to the d initial functions, the convergence rate has the order of* $\frac{\log d}{N}$ *(which is the model selection rate). In the worst case, the procedure has the convergence rate* $\sqrt{\frac{\log d}{N}}$*, which is known to be optimal when* $d > \sqrt{N}$*.*

In the binary classification setting, we implemented the one-step (but non adaptive) procedure to classify input data in $\mathbb{R}^d$ by aggregating stumps and plugging in the regression function to obtain a decision rule. The accuracy of the prediction of the resulting algorithm competes with Adaboost in practice (in particular in noisy classification tasks). Since it is regularized (with the KL-divergence), the algorithm does not overfit unlike AdaBoost. However, for other basis functions than stumps, KL-Boost is not as computationally simple as Adaboost.

The second work presented in this thesis provides a better variance control in PAC-Bayesian bounds for classification and derive original algorithms from these bounds. The main idea of these classification procedures is to start with the function having the smallest complexity, and at each step take the function of smallest complexity having a smaller generalization error with high probability. To compare the efficiency of successive estimators leads to a better variance estimation.

We consider two types of complexity : a PAC-Bayesian one and a compression schemes one. The latter gives a simple way of adapting any overfitting estimator [4] into a well-regularized procedure, and also gives a simple criterion to pick the right algorithm into a family of algorithms.

Specifically, let $\mathcal{Z}$ be the product of the input space $\mathcal{X}$ and the label space $\mathcal{Y}$, and let $\hat{F} : \cup_{n=0}^{+\infty}\mathcal{Z}^n \times \Theta \times \mathcal{X} \to \mathcal{Y}$ denote the family of algorithms indexed by the parameter $\theta \in \Theta$. The associated model $\big\{\hat{F}_{z_1^n,\theta} : n \in \mathbb{N}, z_1^n \in \mathcal{Z}^n, \theta \in \Theta\big\}$ is huge. Compression schemes consider the small data-dependent subsets of the form: $\big\{\hat{F}_{z_1^n,\theta} : n \leq k, z_1^n \in Z_1^N, \theta \in \Theta\big\}$, k being small wrt the integer N.

For any subset $I \subset \{1,\dots,N\}$, define $I^c \triangleq \{1,\dots,N\} - I$ and $Z_I \triangleq (Z_i)_{i\in I}$. Let $\bar{\mathbb{P}}^I$ be the associated empirical distribution $\bar{\mathbb{P}}^I \triangleq \frac{1}{|I|}\sum_{i\in I}\delta_{Z_i}$. The law of the random variable Z_I will be denoted $\mathbb{P}^I$.

For any $I,I' \subset \{1,\dots,N\}$, introduce

$$\begin{cases} \theta_I & \in & \text{argmin}_{\theta\in\Theta}\bar{\mathbb{P}}^I[Y \neq \hat{F}_{Z_I,\theta}(X)] \\ \hat{F}_I & \triangleq & \hat{F}_{Z_I,\theta_I} \\ R(I) & \triangleq & \mathbb{P}[Y \neq \hat{F}_I(X)] \\ r(I) & \triangleq & \bar{\mathbb{P}}^{I^c}[Y \neq \hat{F}_I(X)] \\ \mathbb{P}(I,I') & \triangleq & \mathbb{P}[\hat{F}_I(X) \neq \hat{F}_{I'}(X)] \\ \bar{\mathbb{P}}(I,I') & \triangleq & \bar{\mathbb{P}}^{(I\cup I')^c}[\hat{F}_I(X) \neq \hat{F}_{I'}(X)] \end{cases}$$

Let $\epsilon > 0$ be the desired confidence level (see the following theorem). Finally, for any $I,I' \subset \{1,\dots,N\}$, define $\mathcal{C}_{I,I'} \triangleq \frac{(|I|+|I'|)\log(2N)+\log[(2\epsilon)^{-1}]}{|(I\cup I')^c|}$ and

$$S(I,I') \triangleq \sqrt{2\mathcal{C}_{I,I'}\bar{\mathbb{P}}(I,I')} + \frac{7\mathcal{C}_{I,I'}}{3}.$$

[4] For instance, the 1-Nearest Neighbor algorithm, non pruned trees, kernel machines as SVM with heavily penalized errors.

The following algorithm appropriately chooses the primary algorithm $\theta \in \Theta$ and the compression set I.

Algorithm 1. *Let $I_0 \subset \{1,\dots,N\}$ of cardinal 2. For any $k \geq 1$, define I_k as the smallest subset of $\{1,\dots,N\}$ of cardinal greater or equal to $|I_{k-1}|$ such that*

$$r(I_k) - r(I_{k-1}) + S(I_k,I_{k-1}) \leq 0.$$

Classify using the function $\hat{F}_{I_K}$, where I_K is the compression set obtained at the last iteration.

The following theorem guarantees the efficiency of this procedure.

Theorem 3. *With $\mathbb{P}^{\otimes N}$-probability $1-\epsilon$, we have*

- *for any $k \in \{1,\dots,K\}$, $R(I_k) \leq R(I_{k-1})$,*
- *$R(I_K) \leq \inf_{I,\xi \geq 0} \sup_{I':|I'|\leq |I|} \Big\{(1+\xi)R(I) - \xi R(I') + 2(1+\xi)S(I,I')\Big\}$.*

This procedure can be useful to choose the similarity measure on the input data, and in particular to choose the kernel (its type and its parameter) of a SVM. It is an alternative to the commonly used cross-validation procedure which has the benefit to be theoretically justified.

PAC-Bayesian theorems and the study of randomized estimators lead to consider another measure of complexity based on the Kullbach-Leibler divergence. Let π denote the prior distribution on a given model indexed by the parameter set Θ. For any measurable real function h such that $\exp(h)$ is π-integrable, we define $\pi_h(d\theta) \triangleq \frac{\exp[h(\theta)]}{\mathbb{E}_{\pi(d\theta')}\exp[h(\theta')]} \cdot \pi(d\theta)$.

We propose an efficient way of choosing the temperature of the Gibbs estimator which classifies by drawing a function according to the posterior distribution $\pi_{-\lambda r}$.

Besides, we give the following bracketing of the efficiency of Gibbs classifiers.

Theorem 4. *For any $\lambda > 0$ and $0 < \chi \leq 1$, we have*

$$\mathbb{E}_{\pi_{-(1+\chi)\lambda R}} R - \frac{K(\pi_{-\lambda r},\pi_{-\lambda R})}{\chi\lambda} \leq \mathbb{E}_{\pi_{-\lambda r}} R \leq \mathbb{E}_{\pi_{-(1-\chi)\lambda R}} R + \frac{K(\pi_{-\lambda r},\pi_{-\lambda R})}{\chi\lambda},$$

and for any $\epsilon > 0$, $0 < \gamma < \frac{1}{2}$ and $0 < \lambda \leq 0.39\,\gamma N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$K(\pi_{-\lambda r},\pi_{-\lambda R}) \leq \tfrac{4}{1-\gamma} \log \mathbb{E}_{\pi_{-\lambda R}(d\theta)} \exp\Big(\tfrac{4.1\,\lambda^2}{\gamma N} \mathbb{E}_{\pi_{-\lambda R}(d\theta')} \mathbb{P}\big[f_\theta(X) \neq f_{\theta'}(X)\big]\Big) + \tfrac{5\gamma}{1-\gamma}\log(4\epsilon^{-1}).$$

The third paper in this thesis studies Gibbs classifiers, and other estimators linked to the empirical risk minimization, under variants of the complexity and margin assumptions introduced by Mammen and Tsybakov (E. Mammen and A.B. Tsybakov, Smooth discrimination analysis, Ann. Stat., 27, 1808–1829, 1999).

These assumptions assert that

- the entropy of the model wrt the pseudo-distance $(f_1,f_2) \mapsto \mathbb{P}[f_1(X) \neq f_2(X)]$ is bounded by a polynomial function of the inverse of the radius, i.e. for some $C' > 0$ and $q > 0$, for any $u > 0$, $H(u) \leq C'u^{-q}$
- the expected pseudo-distance between a function and the best function in the model is bounded by a polynomial function of the excess risk, i.e. for some $C'' > 0$ and $1 \leq \kappa \leq +\infty$, for any function f in the model,

$\mathbb{P}[f(X) \neq \tilde{f}(X)] \leq C''[R - R(\tilde{f})]^{\frac{1}{\kappa}}$, where $\tilde{f}$ minimizes the expected risk over the model.

Under these assumptions, we have

Theorem 5. *Let*

$$(v_N,a_N) \triangleq \begin{cases} \left(N^{-\frac{\kappa}{2\kappa-1+q}},\check{C}_1 N^{-\frac{\kappa-1+q}{q(2\kappa-1+q)}}\right) & \text{for } 0 < q < 1 \\ \left((\log N)N^{-\frac{1}{2}},\check{C}_1(\log N)^{-\frac{1}{2}}N^{-\frac{1}{2}}\right) & \text{for } q = 1 \\ \left(N^{-\frac{1}{1+q}},\check{C}_1 N^{-\frac{1}{1+q}}\right) & \text{for } q > 1 \end{cases}$$

For any classifier minimizing the empirical risk among a u_N-covering net $\mathcal{N}_{u_N}$ of the model such that $a_N \leq u_N \leq \check{C}_2 v_N$ and $\log\left|\mathcal{N}_{u_N}\right| \leq \check{C}_3 u_N^{-q}$ for some positive constants $\check{C}_i,i = 1,\dots,3$, we have

$$\mathbb{E}_{\mathbb{P}^{\otimes N}} R(\hat{f}) - R(\tilde{f}) \leq C v_N$$

for some constant $C > 0$ $\big($depending on C'', $\check{C}_i,i = 1,\dots,3\big)$.

Let $\lambda_N \geq \check{C}_4 \frac{h_q(u_N)}{v_N}$ for some constant $\check{C}_4$, and let π be the uniform distribution on the net $\mathcal{N}_{u_N}$. Then we have

$$\mathbb{E}_{\mathbb{P}^{\otimes N}} \mathbb{E}_{\pi_{-\lambda_N r}} R - R(\tilde{f}) \leq \breve{C} v_N$$

for some constant $\breve{C} > 0$ $\big($depending on C'', $\check{C}_i,i = 1,\dots,4\big)$.

The previous convergence rates are optimal to the extent that we prove associated lower bounds. The proof of this theorem requires the chaining trick introduced by Dudley (R.M. Dudley, Central limit theorems for empirical measures, Ann. Probab., 6, 899–929, 1978). This trick appears to be the only tool to properly take into account a polynomial entropy assumption, which holds for any radius. For complex classes (i.e. $q \geq 1$), we note that the optimal convergence rate is obtained since we upper bound the excess risk with an integral entropy which does *not* start from 0, but from the radius of the net we consider.

Consider the stronger margin assumption: for some $c'',C'' > 0$ and $1 \leq \kappa \leq +\infty$, for any functions f in the model,

$$c''[R - R(\tilde{f})]^{\frac{1}{\kappa}} \leq \mathbb{P}[f(X) \neq \tilde{f}(X)] \leq C''[R - R(\tilde{f})]^{\frac{1}{\kappa}}.$$

Under this assumption, two phenomenons occur

- we can prove that some particular estimators has the optimal convergence rate without having recourse to chaining.
- we no longer have discontinuities in results concerning $q < 1$ and $q \geq 1$. Specifically, let $\check{C}_1 N^{-\frac{\kappa-1+q}{q(2\kappa-1+q)}} \leq u_N \leq \check{C}_2 N^{-\frac{1}{2\kappa-1+q}}$. For any classifier minimizing the empirical risk among a u_N-covering net $\mathcal{N}_{u_N}$ with $\log\left|\mathcal{N}_{u_N}\right| \leq \check{C}_3 u_N^{-q}$, we have $\mathbb{E}_{\mathbb{P}^{\otimes N}} R(\hat{f}) - R(\tilde{f}) \leq C N^{-\frac{\kappa}{2\kappa-1+q}}$.

We also consider bracketing polynomial entropy assumptions. These are much more restrictive than covering ones. For instance, under these assumptions,

- with high probability, the empirical covering nets are “similar” to the expected ones,
- the ERM-classifier[5] is optimal whereas it was not even necessarily consistent under polynomial covering entropy assumptions.

[5] Empirical Risk Minimizer.

We deal also with the case of logarithmic entropy assumptions: without surprise[6], we find the same convergence rates as for VC classes. Once more, chaining was the key tool to get rid of the logarithmic factor appearing in classical Vapnik-Chervonenkis bounds.

At last, we show that the efficiency of a Gibbs classifier essentially relies on the weight given by the prior distribution to the balls centered at the best function in the model and associated with the pseudo-distance $(f_1,f_2) \mapsto \mathbb{P}[f_1(X) \neq f_2(X)]$.

The last part of this thesis is a joint work with Olivier Bousquet[7] presented at the Neural Information Processing Systems conference in December 2003. The literature is abundant in generalization error bounds in classification, each one containing an improvement over the others for certain situations. The goal of this work is to combine these gains into a single bound.

The third work in this thesis had stressed on the usefulness of the chaining trick. In stochastic processes theory, it is well-known that the integral entropy, which is tight in many situations, does not capture exactly the expectation of the supremum of a sub-Gaussian process. A refinement of Dudley's chaining due to Fernique and Talagrand allows to be more precise and leads to the introduction of majorizing measures (M. Talagrand, Majorizing measures: the generic chaining, Ann. Probab., 24, (3), 1049–1103, 1996).

Our bound combines the generic chaining trick and the PAC-Bayesian bounds developed in the second paper in this thesis. We see that these two approachs are linked to the extent that majorizing measures can be seen as prior distributions on the model.

This paper gives a quick survey of generalization error bounds in classification and presents our bound from which we can deduce the previous ones up to some variations. Due to the complexity of the bound (which is inherent to the chaining technique), its practical use to design new algorithms is still a subject of future research.

[6] since VC classes have logarithmic empirical entropies.

[7] Max Planck Institute for Biological Cybernetics – Spemannstrasse 38 – D-72076 Tübingen – Germany.

AGGREGATED ESTIMATORS AND EMPIRICAL COMPLEXITY FOR LEAST SQUARE REGRESSION

JEAN-YVES AUDIBERT
jyaudibe@ccr.jussieu.fr

Université Paris VI Pierre et Marie Curie
Laboratoire de Probabilités et Modèles aléatoires
175 rue du Chevaleret
75013 Paris
FRANCE
and
Centre de Recherche en Économie et Statistique
Laboratoire de Finance et Assurance
15 Bd Gabriel Péri
92245 Malakoff Cedex
FRANCE

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 [10, 15]).

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).

Contents

Date: March 2003, revised in April 2004.

2000 *Mathematics Subject Classification.* Primary: 62G08; Secondary: 62J02,94A17,62H30.

Key words and phrases. Nonparametric regression, deviation inequalities, adaptive estimator, model selection, oracle inequalities, boosting.

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.

1. Introduction

Boosting algorithms (AdaBoost introduced by Freund and Schapire in [5], Bagging and Arcing introduced by Breiman in [2], [3]) 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 ([6]) that AdaBoost is a stagewise 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. ([11]) 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 [15], 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 [10]).

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 ([4]) 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 $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$ such that

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

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 $(\Theta, \mathcal{T})$, 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 $\mathcal{S}$ of $\mathcal{R}$ has a π-probability equal to one, then the functions in the π-negligible set $\mathcal{R} \setminus \mathcal{S}$ 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

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

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 $\Theta_1, ..., \Theta_M$ 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 $\hat{\theta}_1 \in \Theta_1, ..., \hat{\theta}_M \in \Theta_M$.

- *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 $(\Theta, \mathcal{T})$ (see [9, 4]).
- 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 $(\Theta, \mathcal{T})$ 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 $\mathcal{R}$:

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

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

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

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

Proof. See Section 7.1. □

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

Theorem 3.2. *Introduce countable 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 $\sum_{i\in I} \eta_i = 1$. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any randomizing procedure $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$, for any $i \in I$, we have*

(3.2)

$$\mathbb{E}_{\hat{\rho}(d\theta)} R(f_\theta) - R(\tilde{f}) \le H(\lambda_i)\bigg(\mathbb{E}_{\hat{\rho}(d\theta)} r(f_\theta) - r(\tilde{f}) + \frac{B^2}{N\lambda_i}\big\{K(\hat{\rho},\pi) + \log[(\eta_i\epsilon)^{-1}]\big\}\bigg).$$

Proof. Introduce the event

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

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

$$\mathbb{P}^{\otimes N}\Big(\underset{i\in I}{\cup} A_i\Big) \le \sum_{i\in I} \mathbb{P}^{\otimes N}(A_i) < \sum_{i\in I} \eta_i \epsilon = \epsilon.$$

□

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

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

where $\mathcal{P}_\lambda$ and $\mathcal{P}_\eta$ 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

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

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

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

For any $0<\epsilon<1$, 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 $0<\lambda_{\text{opt}}<\frac{\alpha B}{2}$ satisfying $\lambda_{\text{opt}} G(\lambda_{\text{opt}})<1$. Then the inverse temperature parameter of the Gibbs distribution is $\beta \triangleq \frac{N\lambda_{\text{opt}}}{B^2}$.

We would like to choose the parameter families such that the infimum $\inf_\rho \mathcal{Q}\big(\rho,(\lambda_i)_{i\in I},(\eta_i)_{i\in I}\big)$ 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 $\frac{1}{\sqrt{N}}$. Therefore relevant values of λ are greater than $\frac{1}{\sqrt{N}}$. Let us define $0<\Lambda<\frac{\alpha B}{2}$ such that $\Lambda G(\Lambda)=1$. Consider the family $(\lambda_i)_{i=1,\dots,p}$, where $\lambda_i \triangleq \frac{\Lambda}{2^i}$ and p is such that $\frac{\Lambda}{2^{p+1}} < \frac{1}{\sqrt{N}} \le \frac{\Lambda}{2^p}$. When the parameter λ_{opt} belongs to $[\frac{1}{\sqrt{N}};\Lambda[$ (which is the case we are interested in), for any $\rho\in\mathcal{M}^1_+(\Theta)$, we have

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

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 $\eta_i=\frac{1}{p}$, $i=1,\dots,p$, 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

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

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

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

where $\lambda_i = \frac{\Lambda}{2^i}$. Finally, we obtain the following randomizing procedure
1. Compute

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

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

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

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 $\frac{N}{2\sigma^2}$.

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

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

This equality shows that aggregated regression procedures are more efficient than randomized ones and that the difference is measured by $\mathbb{E}_{\mathbb{P}} \mathrm{Var}_{\rho(d\theta)} f_\theta(X)$. 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 $G(\lambda) \triangleq \frac{8M}{(\alpha B - 2\lambda)^2 e^2} + \frac{e^{2\lambda} - 1 - 2\lambda}{\lambda^2}$ and $H(\lambda) \triangleq \frac{1}{1 - \lambda G(\lambda)}$, and we define $g(\beta) \triangleq \frac{e^\beta - 1 - \beta}{\beta^2}$ and $h(\beta) \triangleq \frac{1}{1 + \beta g(\beta)}$.

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

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

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

Proof. See Section 7.2. □

Using a 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 $0 < \lambda_i < \frac{\alpha B}{2}$, $\lambda_i G(\lambda_i) < 1$, $\eta_i > 0$, $\sum_{i \in I} \eta_i = 1$, $\beta_j > 0$, $\zeta_j > 0$ and $\sum_{j \in J} \zeta_j = 1$. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, for any aggregating procedure $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$, for any $i \in I$ and for any $j \in J$, we have*

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

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

$$-\mathbb{E}_{\mathbb{P}} \mathrm{Var}_{\rho(d\theta)} f_\theta \leq h(\beta)\Big(- \mathbb{E}_{\bar{\mathbb{P}}} \mathrm{Var}_{\rho(d\theta)} f_\theta + \frac{B^2}{2N\beta}\big[2K(\rho,\pi) + \log(\epsilon^{-1})\big]\Big)$$

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

$$-\mathbb{E}_{\mathbb{P}} \mathrm{Var}_{\rho(d\theta)} f_\theta \leq h(\beta_j)\Big(- \mathbb{E}_{\bar{\mathbb{P}}} \mathrm{Var}_{\rho(d\theta)} f_\theta + \frac{B^2}{2N\beta_j}\big\{2K(\rho,\pi) + \log[(\zeta_j\epsilon)^{-1}]\big\}\Big)$$

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. □

Now let us introduce

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

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

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

Proof. From Theorem 4.1, with $\mathbb{P}^{\otimes N}$-probability at least $1-2\epsilon$, for any aggregating procedure $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$, 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 $\big($Assumption (2.2)$\big)$, the expected risk is bounded. Specifically, we can write

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

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). □

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 $\mathbb{B}\big(\rho, (\lambda_i)_{i\in I}, (\eta_i)_{i\in I}, (\beta_j)_{j\in J}, (\zeta_j)_{j\in J}\big)$, 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 ([7]) and by Yang ([15]). 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 independent from ϵ, for any $0 < \epsilon \le \frac{1}{2}$, we have*

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

where

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

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

$$\left\{\begin{array}{lcl} L_1 & \triangleq & \frac{\log\left(\frac{4\kappa_1 N}{\log 2}\right)}{2\log 2} \vee 2 \\ L_2 & \triangleq & \frac{\log\left(\frac{8\kappa_2 N}{\log 2}\right)}{2\log 2} \vee 2 \end{array}\right.$$

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. *Any aggregating procedure $\hat{\rho}$ minimizing*

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

wrt the probability distribution ρ satisfies for any $\frac{1}{2} \ge \epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$,

$$R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) \le \gamma'(\epsilon),$$

where

$$\left\{\begin{array}{lcl} \gamma'(\epsilon) & = & 2\sqrt{\mathcal{C}_1[2V(\tilde{\rho}) + 4\mathcal{C}_2]} + 6\sqrt{\mathcal{C}_2[2V(\tilde{\rho}) + 4\mathcal{C}_2]} + 2\mathcal{C}_1 + 2\mathcal{C}_2 \\ V(\tilde{\rho}) & \triangleq & \mathbb{E}_{\mathbb{P}} \mathrm{Var}_{\tilde{\rho}(d\theta)} f_\theta . \end{array}\right.$$

Proof. see Section 7.4. □

For a given confidence level $\epsilon > 0$, this bound has the order of $\sqrt{\tilde{\mathcal{C}} V(\tilde{\rho})} \vee \tilde{\mathcal{C}}$, where $\tilde{\mathcal{C}} \triangleq \frac{K(\tilde{\rho},\pi)+\log\log N}{N}$. 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 $\tilde{\mathcal{C}}$. 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

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

where $H_s(\rho)$ denotes the Shannon entropy of ρ $\big(H_s(\rho) \triangleq -\sum_{i=1}^{d}\rho_i \log \rho_i\big)$. In this case, when the best convex combination $\tilde{f}$ belongs to the model $\mathcal{R}$ ($V(\tilde{\rho}) = 0$), the convergence rate of our estimator will be $\frac{\log d}{N}$ (we neglect $\log\log N$ terms), whereas when $\tilde{f}$ is not too close to the regression functions in the model $\mathcal{R}$ $\big($i.e. when $V(\tilde{\rho}) \ge \frac{K(\tilde{\rho},\pi)+\log\log N}{N}\big)$, the convergence rate will be $\sqrt{\frac{\log d}{N}V(\tilde{\rho})}$. In the worst case, the quantity $V(\tilde{\rho})$ 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*

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

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

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

where the constant C does not depend on N.

Remark 4.1. This theorem which strenghtens the one of Nemirovski ([10]) has been further improved by Tsybakov ([13]).

Remark 4.2. In [15], 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 4.3. Note that the unobservable term $r(\tilde{f})$ in the bound $\mathbb{B}$ does not modify the probability distribution $\hat{\rho}_{\lambda,\beta}$ minimizing $\mathbb{B}(\rho,\lambda,\eta,\beta,\zeta)$[1]. 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

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

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

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

[1] The distribution $\hat{\rho}_{\lambda,\beta}$ minimizes $H(\lambda)\mathbb{E}_{\rho(d\theta)} r(f_\theta) - h(\beta)\bar{V} + \frac{B^2}{N}\Big\{\frac{H(\lambda)}{\lambda} + \frac{h(\beta)}{\beta}\Big\}K(\rho,\pi)$ so that it does not depend on η, ζ and ϵ.

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 $\Big($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 $\sqrt{\frac{\log d}{N}}\Big)$, we get that the added term $r(\tilde{f}) - r(\hat{f}_{\text{ERM}})$ is at most of order $\sqrt{\frac{\log d}{N}}$ (we still neglect $\log\log N$ term).

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 $\hat{\rho}_{\lambda,\beta}$ and use the second part of the training sample to select the best distribution among the $\mathrm{O}\big[(\log N)^2\big]$ distributions (each distribution corresponds to a point in the (λ,β)-grid). From Catoni's theorem ([4]) concerning progressive mixtures (see also [1]) in least square regression, this last step is almost free (we just have to pay a negligible $\frac{\log\log N}{N}$ additive term), so the convergence rate of the resulting procedure is effectively of order $\sqrt{\tilde{\mathcal{C}} V(\tilde{\rho})} \vee \tilde{\mathcal{C}}$. From Theorem 3.1, this last step can also be done by simply taking the distribution $\hat{\rho}_{\lambda,\beta}$ having the smallest empirical risk on the second sample[2].

Remark 4.4. 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 $\epsilon > 0$ and $0 < \lambda' < C_1$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, for any aggregating procedure $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}^1_+(\Theta)$,

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

where we still have $\bar{V} = \mathbb{E}_{\overline{\mathbb{P}}}\mathrm{Var}_{\hat{\rho}(d\theta)} f_\theta$. 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 smaller than $\sqrt{N}$, since then the convergence rate towards functions for which $V(\tilde{\rho}) = \mathbb{E}_{\mathbb{P}}\mathrm{Var}_{\tilde{\rho}(d\theta)} f_\theta(x)$ has the same order as B^2 is

$$\sqrt{\frac{\log(dN)}{N}} \gg \frac{d}{N}.$$

However, in this case ($d \le \sqrt{N}$), 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 $x - 1 < \lfloor x \rfloor \le x$. 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

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

[2] See the appendix for details

where $\mathcal{S}$ is the simplex $\{\sum_{i=1}^d \rho_i f_i : \rho_i \ge 0, \sum_{i=1}^d \rho_i = 1\}$ and C_d is the d-dimensional cube $\{\sum_{i=1}^d a_i f_i : 0 \le a_i \le 1\}$. 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

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

For any $f, g \in \mathcal{R}''$, we have $\|f-g\|_\infty \le \frac{d}{2} \frac{B}{\lfloor \sqrt{dN} \rfloor}$, hence[3] $V(\tilde{\rho}) \le \frac{d^2}{16 \lfloor \sqrt{dN} \rfloor^2} B^2$. The number of functions in $\mathcal{R}'$ is upper bounded by $\big(\lfloor \sqrt{dN} \rfloor + 1\big)^d$. Since we have $K(\tilde{\rho}, \pi') \le \log \operatorname{Card} \mathcal{R}'$ (because the distribution π' is uniform over the set $\mathcal{R}'$), we get $\tilde{\mathcal{C}} \le \frac{d \log(N^{\frac{3}{4}}+1)}{N} B^2$. As a result, we have $\sqrt{\tilde{\mathcal{C}} V(\tilde{\rho})} \vee \tilde{\mathcal{C}} = O\big(\frac{d}{N} \log N\big)$, which is the desired convergence rate up to the logarithmic factor. □

In fact, when $d \le \sqrt{N}$, 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 4.5. 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 7.2: for any $0 \le \lambda \le \frac{\alpha B}{2}$ satisfying $8M\lambda \le (\alpha B - 2\lambda)^2 e^2$, the deviations of

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

are given by

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

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

$$\left\{ \begin{array}{ccccccc} \bar{R}(f) & = & R(f) & - & R(\tilde{f}) & = & \mathbb{E}_{\mathbb{P}}\big[\big(Y - f(X)\big)^2\big] - \mathbb{E}_{\mathbb{P}}\big[\big(Y - \tilde{f}(X)\big)^2\big] \\ \bar{r}(f) & = & r(f) & - & r(\tilde{f}) & = & \mathbb{E}_{\bar{\mathbb{P}}}\big[\big(Y - f(X)\big)^2\big] - \mathbb{E}_{\bar{\mathbb{P}}}\big[\big(Y - \tilde{f}(X)\big)^2\big] \end{array} \right.$$

Hence, for any $0 \le \lambda \le \frac{\alpha B}{2}$ satisfying $\lambda G(\lambda) \le 1$, we have successively

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

For any $\epsilon > 0$,

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

With $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, $\bar{R}(f) \le \frac{\bar{r}(f)}{1 - \lambda G(\lambda)} + \frac{B^2}{N} \frac{\log(\epsilon^{-1})}{\lambda[1 - \lambda G(\lambda)]}$. By using a union bound, for any discretized simplex $\mathcal{R}_{disc}$ with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$,

[3] we use that for any random variable X such that $a \le X \le b$ a.s., the variance of X is bounded by $(b-a)^2/4$.

for any $f \in \mathcal{R}_{disc}$, we get

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

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

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

Then we have

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

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

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

where $\Sigma \triangleq \frac{\sum_{i=1}^{N} |2Y_i - f(X_i) - \tilde{f}(X_i)|}{N} \le 2 \frac{\sum_{i=1}^{N} |Y_i - f^*(X_i)|}{N} + 2B$. The algorithm which minimizes the empirical risk on the net $\mathcal{R}_{disc}$ satisfies with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, for any $f \in \mathcal{R}_{disc}$,

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

where $\tilde{f}_{disc} \triangleq \underset{f \in \mathcal{R}_{disc}}{\text{argmin}} R(f)$, hence, by taking $\lambda = \kappa_1$ defined as $2\kappa_1 G(\kappa_1) = 1$,

$$R(\hat{f}) - R(\tilde{f}) \le 2\Sigma \frac{B}{m} + \begin{cases} \frac{2B^2}{N\kappa_1} \left[d \log(m) + \log(2\epsilon^{-1}) \right] & \text{when } d \le m \\ \frac{2B^2}{N\kappa_1} \left[m \log(d) + \log(2\epsilon^{-1}) \right] & \text{when } d \ge m \end{cases}$$

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

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

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

$$\mathbb{P}^{\otimes N} R(\hat{f}) - R(\tilde{f}) \le \begin{cases} \text{Cst } B^2 \frac{d}{N} \log(\frac{N}{d}) & \text{when } d \le \sqrt{N} \\ \text{Cst } B^2 \sqrt{\frac{\log d}{N}} & \text{when } d \ge \sqrt{N} \end{cases}$$

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 ([8, 14]), Mammen and Tsybakov similarly obtained optimal minimax convergence rate. Note that for linear and convex combination, simpler proofs exist under stronger assumptions (see [13]).

4.2.2. *Aggregating procedure.* We consider the aggregating procedure studied in Theorem 4.5: the algorithm minimizes the quantity $\mathbb{B}\big(\rho, (\lambda_i)_{i\in I}, (\eta_i)_{i\in I}, (\beta_j)_{j\in J}, (\zeta_j)_{j\in J}\big)$ 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

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

where $a > 0$, $0 < b < 1$ and $c > 0$[4].

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

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

we can write

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

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 euclidian norm in $\mathbb{R}^N$ and $h : \Theta \to \mathbb{R}^N$ 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.

[4] 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 $c = \frac{B^2}{N\lambda}\Big(1 + \frac{\lambda[1-\lambda G(\lambda)]}{\beta[1+\beta g(\beta)]}\Big)$.

Theorem 4.7. *For any $\mu \in \mathcal{M}_+^1(\Theta)$ and any bounded function $h : \Theta \to \mathbb{R}^N$, 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*

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

and we have

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

Proof. See Section 7.5 □

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 $\pi^w \triangleq \pi_{-d_1 N r(f) + \langle w, f(X) \rangle}$, 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

$$\pi^w = \pi_{-d_1 N r(f) + 2 d_2 \langle Y - \mathbb{E}_{\pi^w(d\theta)} f_\theta(X), f(X) - Y \rangle}.$$

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

$$\begin{aligned} \bar{\varphi}(w) \triangleq \bar{\psi}(\pi^w) \quad = \quad & ac \Big[d_2 \|\mathbb{E}_{\pi^w} f(X) - Y\|^2 - \log \mathbb{E}_{\pi_{-\frac{b}{c} r(f)}} e^{\langle w, f(X) - \mathbb{E}_{\pi^w} f(X) \rangle} \Big] \\ & - ac \log \mathbb{E}_\pi e^{-\frac{b}{c} r(f)}. \end{aligned}$$

We have

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

where $\mathbb{V}\mathrm{ar}_{\pi^w} f(X)$ is the covariance matrix of $f(X_i), i = 1, \ldots, N$ 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 $f(X_{r+1}), \ldots, f(X_N)$ are π-linear combination of $f(X_1), \ldots, f(X_r)$ to the extent that there exists $\alpha^i \in \mathbb{R}^r, \beta^i \in \mathbb{R}$, $i = r+1, \ldots, N$ such that for any $i \in \{r+1, \ldots, N\}$

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
\tfrac{1}{ac}\tfrac{\partial \bar\varphi}{\partial w_k}(w) &= \textstyle\sum_{j=1}^{r} \mathbb{C}\mathrm{ov}_{\pi^w}[f(X_k), f(X_j)]\big(2d_2\mathbb{E}_{\pi^w}[Y_j - f(X_j)] + w_j\big) \\
&\quad + \textstyle\sum_{i=r+1}^{N} 2d_2 \mathbb{C}\mathrm{ov}_{\pi^w}[f(X_k), \langle \alpha^i, f(X)\rangle_r]\mathbb{E}_{\pi^w}[Y_i - \langle \alpha^i, f(X)\rangle_r - \beta^i] \\
&= \textstyle\sum_{j=1}^{r} \mathbb{C}\mathrm{ov}_{\pi^w}[f(X_k), f(X_j)]\bigg(w_j + 2d_2\mathbb{E}_{\pi^w}[Y_j - f(X_j)] \\
&\qquad\qquad + 2d_2 \textstyle\sum_{i=r+1}^{N} \alpha_j^i\ \mathbb{E}_{\pi^w}[Y_i - \langle \alpha^i, f(X)\rangle_r - \beta^i]\bigg),
\end{aligned}$$

hence

$$\begin{aligned}
\triangledown_r \bar\varphi(w) = ac\mathbb{V}\mathrm{ar}_{\pi^{w^l}} f(X)\,\Big|_r \bigg[w - 2d_2\bigg(&Y - \mathbb{E}_{\pi^w} f(X) \\
&+ \textstyle\sum_{i=r+1}^{N} \alpha^i [Y_i - \langle \alpha^i, \mathbb{E}_{\pi^w} f(X)\rangle_r - \beta^i]\bigg)\bigg],
\end{aligned}$$

where $\triangledown_r \bar\varphi$ is the vector $\frac{\partial \bar\varphi}{\partial w_k}, k = 1, \dots, r$ and $\mathbb{V}\mathrm{ar}_{\pi^{w^l}} f(X)\,\big|_r$ 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 computationnally too expensive[5], 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
        $w^{l+1} = 2d_2\Big(Y - \mathbb{E}_{\pi^{w^l}} f(X) + \sum_{i=r+1}^{N} \alpha^i [Y_i - \langle \alpha^i, \mathbb{E}_{\pi^{w^l}} f(X)\rangle_r - \beta^i]\Big).$
    • 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
```

The stopping criteria in the loop comes from

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

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

In particular, we have

$$\begin{aligned}&\bar\psi(\pi^{w^l}) - \bar\psi(\bar\rho) \\ &\quad \le acB\bigg\| w^l - 2d_2\Big(Y - \mathbb{E}_{\pi^{w^l}} f(X) + \textstyle\sum_{i=r+1}^{N} \alpha^i [Y_i - \langle \alpha^i, \mathbb{E}_{\pi^{w^l}} f(X)\rangle_r - \beta^i]\Big)\bigg\|.\end{aligned}$$

Proof. See Section 7.6. □

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 $R(\mathbb{E}_{\pi^{w_{\rm opt}}} f) - R(\tilde f)$. It can also be interesting to upper bound $R(\mathbb{E}_{\pi^{w_{\rm opt}}} f)$ (since $R(\tilde f)$ is unknown). The following corollary gives an observable upper bound of the expected risk of any aggregating procedure.

[5] 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.

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

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

where

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

and

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

Proof. See Section 7.8. □

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 $\frac{1}{N}$, the bound on the expected risk $R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta)$ is of order $\frac{(\log N)^4}{N}$. 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 7.1). 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 7.5 holds for

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

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

$$\left\{\begin{array}{rcl} \kappa_3 & = & \frac{M^2 e^{-2}}{2[(\alpha B e)^2+4M]} \\ \kappa_4 & = & \frac{Me}{\alpha B}\sqrt{\frac{\kappa_1}{8}} \end{array}\right. .$$

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 $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-2\epsilon$, for any randomizing procedure $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}^1_+(\Theta)$, for any $i \in I$ and for any $j \in J$, we have*

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

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

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 $Z_\theta = -\big(Y - f_\theta(X)\big)^2 + \big(Y - \tilde{f}(X)\big)^2 = [f_\theta(X) - \tilde{f}(X)][2Y - \tilde{f}(X) - f_\theta(X)]$ given by Lemma 7.2 can be obtained by using directly Lemma 7.1 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 $4g(\lambda)$. □

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 $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-2\epsilon$, for any randomizing procedure $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}^1_+(\Theta)$, we have*

$$\begin{array}{rcl} R(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) & \leq & r(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) + \mathbb{B}'' \\ & & +\sqrt{\frac{2\log(\epsilon^{-1})}{N}}\left(\sqrt{r(\mathbb{E}_{\hat{\rho}(d\theta)}f_\theta) + \mathbb{B}'' + \frac{\log(\epsilon^{-1})}{2N}} + \sqrt{\frac{\log(\epsilon^{-1})}{2N}}\right) \end{array}$$

where

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

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 $Z = \big(Y - \tilde{f}(X)\big)^2 \in [0;1]$. For any $\lambda > 0$ and any $\mu \in \mathbb{R}$, we have

$$\begin{array}{rcl} \mathbb{P}^{\otimes N}(R(\tilde{f}) - r(\tilde{f}) > \mu) & \leq & \mathbb{E}_{\mathbb{P}^{\otimes N}} e^{N\lambda(R(\tilde{f}) - r(\tilde{f}) - \mu)} \\ & \leq & e^{-N\lambda\mu}\big(\mathbb{E}_{\mathbb{P}}\, e^{\lambda(\mathbb{E}_{\mathbb{P}} Z - Z)}\big)^N \\ & \leq & e^{N\left(-\lambda\mu + \frac{\lambda^2}{2}\mathbb{E}_{\mathbb{P}} Z\right)}, \end{array}$$

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

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

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

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

where $\mathbb{B}''$ are the quantities defined in Corollary 4.11. Hence, we have successively

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

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

$$\begin{array}{rcl} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) & \leq & r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'' \\ & & + \sqrt{\frac{2\log(\epsilon^{-1})}{N}}\left(\sqrt{r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + \mathbb{B}'' + \frac{\log(\epsilon^{-1})}{2N}} + \sqrt{\frac{\log(\epsilon^{-1})}{2N}}\right). \end{array}$$

□

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

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

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

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

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

$$\begin{aligned} \tilde{\mathcal{R}} = \Big\{ x \mapsto \textstyle\sum_{j=1}^d \alpha_j h_j(x_j) \ :\ \text{for any } j \in \{1, \ldots, d\}, h_j \in \mathcal{H}, \alpha_j \geq 0 \\ \text{and } \textstyle\sum_{j=1}^d \alpha_j = 1 \Big\}, \end{aligned} \tag{5.2}$$

where

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

$\tilde{\mathcal{R}}$ *can also be written*
(5.3)

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

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

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

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. □

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

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

where δ_a denotes the Dirac distribution on point a. For this generator, the best additive model has a misclassification rate of $\frac{1}{4}$ 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 [3] 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 :

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

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

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

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

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 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

$$\begin{cases} \pi(0_{\mathbb{R}}) = \frac{1}{4} \\ \pi(1_{\mathbb{R}}) = \frac{1}{4} \\ \pi\big(\underset{\tau\in\mathbb{R}}{\cup}\{f_{j,\tau}\}\big) = \frac{1}{4d} \text{ for any } j \in \{1,\dots,d\} \\ \pi\big(\underset{\tau\in\mathbb{R}}{\cup}\{g_{j,\tau}\}\big) = \frac{1}{4d} \text{ for any } j \in \{1,\dots,d\} \end{cases}$$

	Twonorm
$\mathcal{L}(Y)$	$\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$
$\mathcal{L}(X/Y=0)$	$N(m_-, \mathbf{I})$
$\mathcal{L}(X/Y=1)$	$N(m_+, \mathbf{I})$
$\mathcal{L}(X)$	$\frac{G_{m_-}(x)+G_{m_+}(x)}{2}dx$
$\mathbb{P}(Y=1/X=x)$	$\frac{G_{m_+}}{G_{m_+}+G_{m_-}}(x) = \frac{1}{1+e^{-2\langle x, m_+\rangle}}$
frontier	$\langle x, m_+\rangle = 0$

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

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

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

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 $\eta_i = \eta = \frac{1}{|I|}$ and $\zeta_j = \zeta = \frac{1}{|J|}$. Let d'_1 be some real and define $\hat{\rho}_{d'_1} \triangleq \pi_{-d'_1 N r(f) + \langle w, f(X)\rangle}$. Set

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

We have $\mathbb{B}(\lambda, \beta, \hat{\rho}) = a\big[b\mathbb{E}_{\hat{\rho}(d\theta)} r(f_\theta) + (1-b) r(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) + cK(\hat{\rho}, \pi)\big] + d_3$, hence
(5.4)

$$\begin{array}{rcl} \mathbb{B}(\lambda, \beta, \hat{\rho}_{d'_1}) & = & ac\Big(d_2 \sum_{i=1}^N [Y_i - \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i)]^2 + d_1 \mathbb{E}_{\hat{\rho}_{d'_1}} \sum_{i=1}^N [Y_i - f(X_i)]^2 \\ & & \qquad\qquad\qquad\qquad\qquad\qquad + K(\hat{\rho}_{d'_1}, \pi)\Big) + d_3 \\ & = & ac\Big(d_2 \sum_{i=1}^N [Y_i - \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i)]^2 \\ & & \quad + (d_1 - d'_1) \sum_{i=1}^N \big(Y_i - 2Y_i \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i) + \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i)\big) \\ & & \quad \sum_{i=1}^N w_i \mathbb{E}_{\hat{\rho}_{d'_1}} f(X_i) - \log \pi e^{-d'_1 N r(f) + \langle w, f(X)\rangle}\Big) + d_3 \end{array}$$

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

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

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

The following theorem gives a simple expression of $\mathbb{E}_\pi e^{-d'_1 N r(f) + \langle w, f(X)\rangle + u f(x)}$. 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 $l \in \{0, \dots, N\}$,

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

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

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

Theorem 5.2. *We have*

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

As a consequence,

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

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. □

Let N_0 (resp. N_1) be the number of class 0 data (resp. 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*

$$\mathbb{E}_\pi e^{-d_1' Nr(f) + \langle w, f(X)\rangle} \quad = \quad \tfrac{1}{4d}\left(dc_0^w + dc_1^w + \textstyle\sum_{j=1}^d \sum_{l=0}^N \left(a_{j,l}^w + b_{j,l}^w\right)\right).$$

Let $\hat\rho_{d_1'} \triangleq \pi_{-d_1' Nr(f) + \langle w, f(X)\rangle}$. We have

$$\begin{cases} \mathbb{E}_{\hat\rho_{d_1'}} f(x) & = & \frac{dc_1^w + \sum_{j=1}^d \sum_{l=0}^N c_{j,l}^w(x)}{dc_0^w + dc_1^w + \sum_{j=1}^d \sum_{l=0}^N \left(a_{j,l}^w + b_{j,l}^w\right)} \\ \mathbb{E}_{\hat\rho_{d_1'}} [f(x) f(y)] & = & \frac{dc_1^w + \sum_{j=1}^d \sum_{l=0}^N c_{j,l}^w(x,y)}{dc_0^w + dc_1^w + \sum_{j=1}^d \sum_{l=0}^N \left(a_{j,l}^w + b_{j,l}^w\right)} \end{cases}$$

Proof. It comes from Theorem 5.2 and from

$$\begin{cases} \mathbb{E}_{\hat{\rho}_{d_1'}} f(x) & = \frac{\partial}{\partial u} \log \mathbb{E}_\pi e^{-d_1' N r(f) + \langle w, f(X)\rangle + u f(x)} \Big|_{u=0} \\ \mathbb{C}\mathrm{ov}_{\hat{\rho}_{d_1'}} \big(f(x), f(y)\big) & = \frac{\partial^2}{\partial u \partial v} \log \mathbb{E}_\pi e^{-d_1' N r(f) + \langle w, f(X)\rangle + u f(x) + v f(y)} \Big|_{u=0, v=0} \end{cases}$$

□

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 $a^w_{j,l}$ and $b^w_{j,l}$ since for any $j \in \{1, \dots, d\}$ and $l \in \{0, \dots, N\}$, we have

$$a^w_{j,l} b^w_{j,l} = \phi^2_{j,l} c^w_0 c^w_1.$$

Computation of the constant d_3

We have

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

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

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

under the linear constraints

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

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

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

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 $\hat{\rho} = \pi_{\langle w,f \rangle}$, where the vector w satisfies $w_i = \frac{1}{\alpha N}[Y_i - \mathbb{E}_{\pi_{\langle w,f\rangle}} f(X_i)]$ 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.

5.3.2. *AdaBoost using domain-partitioning functions (*[5, 12, 6]*).* 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 [5]), 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 figure 2 where $\mathbb{E}_{w^m}$ denotes the empirical expectation wrt the weights $w_1^m, \dots, w_N^m$.

FIGURE 1. "Discrete" AdaBoost using domain-partitioning functions (Freund and Schapire [5])

Start with weights $w_i^0 = \frac{1}{N}$ for any $i \in \{1, \dots, N\}$.

For $m = 1$ to M do

Choose a partition of $\mathcal{X} = \sqcup_{l=1}^{L} \mathcal{X}_l^m$.

On each $\mathcal{X}_l^m$, $f_m \in \{0, 1\}$ is constant and such that it minimizes the weighted training error

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

Set $w_i^m = \frac{w_i^{m-1} e^{c_m \mathbb{1}_{Y_i \neq f_m(X_i)}}}{\text{Cst}}$ for any $i \in \{1, \dots, N\}$, where

- Cst is the normalizing constant,
- $c_m \triangleq \log\left(\frac{1-e_m}{e_m}\right)$.

Output the classifier $\mathbb{1}_{\mathbb{E}_c f(x) \geq \frac{1}{2}}$, where $\mathbb{E}_c$ is the expectation wrt the weights $c_1, \dots, c_M$.

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 figure 2).

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

The link between AdaBoost and this criterion has been introduced by Friedman, Hastie and Tibshirani [6].

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

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

Then we have

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

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 $\bar{f}_{m,l}$ equal to the most w^{m-1}-popular class on $\mathcal{X}_l^m$, hence

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

Since we have

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

the optimal c_m is

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

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

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

In [12], 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 figure 3 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 $\bar{f}_{m,l}$, $l = 1,\dots,L$ is chosen such that it minimizes

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

FIGURE 2. "Real" AdaBoost using domain-partitioning functions (Schapire and Singer[12])

Start with weights $w_i^0 = \frac{1}{N}$ for any $i \in \{1, \dots, N\}$.

For $m = 1$ to M do

Choose a partition of $\mathcal{X} = \sqcup_{l=1}^{L} \mathcal{X}_l^m$.

For any $l \in \{1, \dots, L\}$, on each $\mathcal{X}_l^m$, $\bar{f}_m \in \mathbb{R}$ is constant and equal to

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

Set $w_i^m = \frac{w_i^{m-1} e^{-\bar{Y}_i \bar{f}_m(X_i)}}{\text{Cst}}$ for any $i \in \{1, \dots, N\}$, where Cst is the normalizing constant.

Output the classifier $\mathbb{1}_{F_M(x) \geq \frac{1}{2}} = \frac{1+\text{sign}[\bar{F}_M(x)]}{2}$.

Besides, we have

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

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 $[\bar{F} \mapsto \mathbb{E}_{\bar{\mathbb{P}}} e^{-\bar{Y}\bar{F}(X)}]$.

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

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

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

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

The AdaBoost used in our numerical examples is described in figure 4. 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

$$\cup_{j=1}^{d}\Big(\{j\} \times \cup_{k=1}^{k_j}]a_j; b_j]\Big),$$

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

FIGURE 3. "Real" AdaBoost using stumps (Schapire and Singer[12])

Start with weights $w_i^0 = \frac{1}{N}$ for any $i \in \{1, \dots, N\}$.
For $m = 1$ to M do

Determine $j \in \{1, \dots, d\}$ and $\tau \in \mathbb{R}$ minimizing $\mathcal{W}_{w^{m-1}}(j, \tau)$.

Choose $\bar{f}_m = \bar{f}_{m,<} \mathbb{1}_{x \in \mathcal{X}^{<}_{j,\tau}} + \bar{f}_{m,\geq} \mathbb{1}_{x \in \mathcal{X}^{\geq}_{j,\tau}}$, where

$$\begin{cases} \bar{f}_{m,<} & \triangleq & \frac{1}{2} \log \left(\frac{\mathbb{P}_{w^{m-1}}(Y=1; X \in \mathcal{X}^{<}_{j,\tau}) + \beta}{\mathbb{P}_{w^{m-1}}(Y=0; X \in \mathcal{X}^{<}_{j,\tau}) + \beta} \right) \\ \bar{f}_{m,\geq} & \triangleq & \frac{1}{2} \log \left(\frac{\mathbb{P}_{w^{m-1}}(Y=1; X \in \mathcal{X}^{\geq}_{j,\tau}) + \beta}{\mathbb{P}_{w^{m-1}}(Y=0; X \in \mathcal{X}^{\geq}_{j,\tau}) + \beta} \right) \end{cases}$$

and $\beta = \frac{1}{4N}$.

Set $w_i^m = \frac{w_i^{m-1} e^{-\bar{Y}_i \bar{f}_m(X_i)}}{\mathrm{Cst}}$ for any $i \in \{1, \dots, N\}$, where Cst is the normalizing constant.

Output the classifier $\mathbb{1}_{F_M(x) \geq \frac{1}{2}} = \frac{1 + \mathrm{sign}[\bar{F}_M(x)]}{2}$.

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

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

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

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 $\mathrm{sign}(\tilde{\mathcal{R}})$) 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 $f'_m \triangleq \frac{1+\bar{f}'_m}{2}$, $\bar{f}'_m \triangleq \frac{\bar{f}_m}{\max\left\{ |f_{m,k}|; k \in \{<, \geq\}, m \in \{1, \dots, M\} \right\}}$ 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 the tables 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 $3,6$ (resp. $10,20$) in the "dimension" column means that the input is 6-dimensional (resp. 10-dimensional) but the output only depends on 3 (resp. 10) components of the input (the other 3 (resp. 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 $\tilde{\mathcal{R}}_{ada}$ the regression function model associated with Adaboost. On the one hand, Adaboost will tend to classify as $C_{ada} \triangleq \text{sign}(-1+2\tilde{f}_{ada})$, where

$$\tilde{f}_{ada} \triangleq \operatorname*{argmin}_{f\in\tilde{\mathcal{R}}_{ada}} R(f)$$

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

$$\tilde{f} \triangleq \operatorname*{argmin}_{f\in\tilde{\mathcal{R}}} R(f).$$

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.

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-Booost, 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 since we use

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

		Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
N	dimension	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)

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

		Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
N	dimension	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)

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

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

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

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

Comparaison between Adaboost and KL-Boost:
classification and quadratic errors for different ringnorm generators

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

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

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

Comparaison between Adaboost and KL-Boost:
classification and quadratic errors for noisy twonorm generators

		Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
N	dimension	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)

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

		Classif. gen. errors		Classif. emp. errors		L^2 gen. errors		L^2 emp. errors	
N	dimension	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)

Comparaison between Adaboost and KL-Boost:
classification and quadratic errors for noisy ringnorm generators

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

the minimax approach, which condiders 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. This is also true for the 6-dimensional noisy threenorm and ringnorm generators. The reverse has not occured 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

- 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[6].

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.

[6] 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 $Z_\theta(X,Y) \triangleq -\big(Y - f_\theta(X)\big)^2 + \big(Y - \tilde{f}(X)\big)^2$, 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 Z be a random variable.*

- *If $Z \le b$ a.s., then for any $\eta \ge 0$,*

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

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

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

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

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

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 $q > 0$:

$$\begin{aligned} \mathbb{E}|\bar{Z}|^q &\le \Big(\sup_{u\in\mathbb{R}_+} ue^{-\frac{\alpha}{q}u}\Big)^q \mathbb{E}e^{\alpha|\bar{Z}|} \\ &\le \Big(\frac{q}{\alpha e}\Big)^q \mathbb{E}e^{\alpha|\bar{Z}|} \\ &\le \Big(\frac{q}{\alpha e}\Big)^q M. \end{aligned}$$

According to the Taylor series expansion, for any $\eta \ge 0$, 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 $x \in \mathbb{R}$,

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

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

$$\begin{aligned}\log \mathbb{E}e^{\eta\bar{Z}} &\le \mathbb{E}(e^{\eta\bar{Z}} - 1 - \eta\bar{Z}) \\ &\le \mathbb{E}\Big(\tfrac{\eta^2\bar{Z}^2}{2}e^{\eta|\bar{Z}|}\Big) \\ &\le \tfrac{\eta^2}{2}\mathbb{E}\Big(\bar{Z}^2e^{\eta|\bar{Z}|}\Big) \\ &\le \tfrac{\eta^2}{2}\Big(\mathbb{E}|\bar{Z}|^{\frac{2\alpha}{\alpha-\eta}}\Big)^{\frac{\alpha-\eta}{\alpha}}\Big(\mathbb{E}e^{\alpha|\bar{Z}|}\Big)^{\frac{\eta}{\alpha}} && \text{(by Hőlder's inequality)} \\ &\le \tfrac{\eta^2}{2}\Big(\tfrac{2}{(\alpha-\eta)e}\Big)^2 M \\ &\le \eta^2 g_1(\eta).\end{aligned}$$

□

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

Lemma 7.2. *For any $0 < \lambda < \frac{\alpha B}{2}$ satisfying*

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

we have

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

where

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

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

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

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

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

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

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

for any λ satisfying (7.3),

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

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

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

□

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 $(A, \mathcal{A})$:

$$K(\nu,\mu) \triangleq \begin{cases} \mathbb{E}_\nu \log\left(\frac{\nu}{\mu}\right) & \text{if } \nu \ll \mu, \\ +\infty & \text{otherwise.} \end{cases}$$

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

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

where, by convention:

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

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

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

achieves the supremum.

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

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

is successively equal to

$$\begin{aligned}
&\left\{\sup_{\rho\in\mathcal{M}^1_+(\Theta)}\left\{\mathbb{E}_\rho \bar{R} - \tfrac{\mathbb{E}_\rho \bar{r}}{1-\lambda G(\lambda)} - \tfrac{B^2}{N}\tfrac{K(\rho,\pi)+\log(\epsilon^{-1})}{\lambda[1-\lambda G(\lambda)]}\right\} > 0\right\},\\
&\left\{\sup_{\rho\in\mathcal{M}^1_+(\Theta)}\left\{\mathbb{E}_\rho\Big([1-\lambda G(\lambda)]\bar{R}-\bar{r}\Big) - \tfrac{B^2}{N\lambda}\big[K(\rho,\pi)+\log(\epsilon^{-1})\big]\right\} > 0\right\},\\
&\left\{\sup_{\rho\in\mathcal{M}^1_+(\Theta)}\left\{\mathbb{E}_\rho\left[\tfrac{N\lambda}{B^2}\Big([1-\lambda G(\lambda)]\bar{R}-\bar{r}\Big) - \log(\epsilon^{-1})\right] - K(\rho,\pi)]\right\} > 0\right\},\\
&\left\{\log \mathbb{E}_\pi e^{\frac{N\lambda}{B^2}\left([1-\lambda G(\lambda)]\bar{R}-\bar{r}\right)-\log(\epsilon^{-1})} > 0\right\},\\
&\left\{\mathbb{E}_\pi e^{\frac{N\lambda}{B^2}\left([1-\lambda G(\lambda)]\bar{R}-\bar{r}\right)-\log(\epsilon^{-1})} > 1\right\}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^{\otimes N}}\mathbb{E}_\pi e^{\frac{N\lambda}{B^2}\left([1-\lambda G(\lambda)]\bar{R}-\bar{r}\right)-\log(\epsilon^{-1})} \\
&= \mathbb{E}_\pi\mathbb{E}_{\mathbb{P}^{\otimes N}} e^{\frac{N\lambda}{B^2}\left([1-\lambda G(\lambda)]\bar{R}-\bar{r}\right)-\log(\epsilon^{-1})} && \text{(by Fubini's theorem)}\\
&= \epsilon\mathbb{E}_\pi\mathbb{E}_{\mathbb{P}^{\otimes N}} e^{\frac{N\lambda}{B^2}[\mathbb{E}_{\bar{\mathbb{P}}}Z_\theta-\mathbb{E}_{\mathbb{P}}Z_\theta-\lambda G(\lambda)\bar{R}]} && \Big(\text{since } Z_\theta \triangleq (Y-\tilde{f})^2-(Y-f_\theta)^2\Big)\\
&\le \epsilon\mathbb{E}_\pi\left[e^{-\frac{N\lambda^2 G(\lambda)\bar{R}}{B^2}}\left(\mathbb{E}_{\mathbb{P}}\, e^{\frac{\lambda}{B^2}(Z_\theta-\mathbb{E}_{\mathbb{P}}Z_\theta)}\right)^N\right] && \text{(since the training sample is i.i.d)}\\
&\le \epsilon\mathbb{E}_\pi\left[e^{\frac{N\lambda^2 G(\lambda)[\mathbb{E}_{\mathbb{P}}(\tilde{f}-f_\theta)^2-\bar{R}]}{B^2}}\right] && \text{(from Lemma 7.2)}\\
&\le \epsilon,
\end{aligned}$$

where at the last step we use that we have $\mathbb{E}_{\mathbb{P}}(\tilde{f}-f_\theta)^2 \le \bar{R}(\theta)$ 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 $\mathbb{E}_{\mathbb{P}}\Big\{\big[f^*(X)-\tilde{f}(X)\big]\big[\tilde{f}(X)-f_\theta(X)\big]\Big\} \ge 0$. 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

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

shows that aggregating regression procedures is more efficient than randomizing and that the difference is measured by $\mathbb{E}_{\mathbb{P}}\mathrm{Var}_{\rho(d\theta)}f_\theta(X)$. 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

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

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

Lemma 7.3. *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 $\lambda \geq 0$,

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

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

7.2.2. *Second step : Legendre formula.* Introduce $V = \mathbb{E}_{\mathbb{P}} \mathrm{Var}_{\hat{\rho}(d\theta)} f_\theta$ and $\bar{V} = \mathbb{E}_{\bar{\mathbb{P}}} \mathrm{Var}_{\hat{\rho}(d\theta)} f_\theta$. For any $\epsilon > 0$ and $\beta > 0$, the event

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

is equal to

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

which is included in the event

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

This last event can be written successively as

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

$$\left\{ \log \mathbb{E}_{\pi\otimes\pi(d\theta,d\theta')} e^{\frac{N\beta}{B^2} \left(\mathbb{E}_{\bar{\mathbb{P}}} Z_{\theta,\theta'} - [1+\beta g(\beta)] \mathbb{E}_{\mathbb{P}} Z_{\theta,\theta'} \right) - \log(\epsilon^{-1})} > 0 \right\},$$
$$\left\{ \mathbb{E}_{\pi\otimes\pi(d\theta,d\theta')} e^{\frac{N\beta}{B^2} \left(\mathbb{E}_{\bar{\mathbb{P}}} Z_{\theta,\theta'} - [1+\beta g(\beta)] \mathbb{E}_{\mathbb{P}} Z_{\theta,\theta'} \right) - \log(\epsilon^{-1})} > 1 \right\}.$$

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

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^{\otimes N}}\mathbb{E}_{\pi\otimes\pi(d\theta,d\theta')}e^{\frac{N\beta}{B^2}\left(\mathbb{E}_{\bar{\mathbb{P}}}Z_{\theta,\theta'}-[1+\beta g(\beta)]\mathbb{E}_{\mathbb{P}}Z_{\theta,\theta'}\right)-\log(\epsilon^{-1})} \\
&=\epsilon\mathbb{E}_{\pi\otimes\pi(d\theta,d\theta')}\mathbb{E}_{\mathbb{P}^{\otimes N}}e^{\frac{N\beta}{B^2}\left(\mathbb{E}_{\bar{\mathbb{P}}}Z_{\theta,\theta'}-\mathbb{E}_{\mathbb{P}}Z_{\theta,\theta'}-\beta g(\beta)\mathbb{E}_{\mathbb{P}}Z_{\theta,\theta'}\right)} && \text{(by Fubini's theorem)} \\
&\le\epsilon\mathbb{E}_{\pi}\left[e^{-\frac{N\beta^2 g(\beta)\mathbb{E}_{\mathbb{P}}Z_{\theta,\theta'}}{B^2}}\left(\mathbb{E}_{\mathbb{P}}\,e^{\frac{\beta}{B^2}(Z_{\theta,\theta'}-\mathbb{E}_{\mathbb{P}}Z_{\theta,\theta'})}\right)^N\right] && \text{(i.i.d. training sample)} \\
&\le\epsilon && \text{(from Lemma 7.3)}
\end{aligned}$$

7.3. **Proof of Lemma 4.4.** We will take the following parameter families

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

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 $\lambda_{\min}$ and $\lambda_{\max}$ (resp. $\beta_{\min}$ and $\beta_{\max}$) such that the constant η (resp. ζ) is large (in order that the bound is not significantly affected by the union bound term $\log[(\eta\epsilon)^{-1}]$ (resp. $\log[(\zeta\epsilon)^{-1}]$)). 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

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

In general, the quantity $\bar{V}(\tilde{\rho})=\mathbb{E}_{\bar{\mathbb{P}}}\mathrm{Var}_{\tilde{\rho}(d\theta)}f_\theta$ 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

$$\left\{\begin{array}{lcl}\lambda_{\max} & = & \kappa_1\\ \beta_{\max} & = & \kappa_2\end{array}\right.,$$

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 $0<x\le 1$, $1-x<\frac{1}{1+x}\le 1-\frac{x}{2}$, we have for any $0<\beta\le\beta_{\max}$,

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

The last RHS is minimum when

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

since $\epsilon \leq \frac{1}{2}$ and $\bar{V}(\tilde{\rho}) \leq \frac{B^2}{4}$ according to Assumption (2.1). Therefore, let us take

$$\beta_{\min} \triangleq \sqrt{\frac{2\beta_{\max}\log 2}{N}} \wedge \beta_{\max}.$$

Let us define the event

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

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

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

Particular case: $(E_1)^c$ occurs

Then, for $j = 0$, we have

$$\begin{aligned}
&-\tfrac{\bar{V}(\tilde{\rho})}{1+\beta_j g(\beta_j)} + \tfrac{B^2}{2N}\tfrac{2K(\tilde{\rho},\pi)+\log[(\zeta\epsilon)^{-1}]}{\beta_j[1+\beta_j g(\beta_j)]}\\
&\quad= -\tfrac{\bar{V}(\tilde{\rho})}{2} + \tfrac{B^2}{4N}\tfrac{2K(\tilde{\rho},\pi)+\log[(\zeta\epsilon)^{-1}]}{\beta_{\max}}.
\end{aligned}$$

Besides, we have

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

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

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

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

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

The last RHS is minimum when

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

Therefore, let us take $\lambda_{\min} \triangleq 2\sqrt{\frac{\lambda_{\max}\log 2}{N}} \wedge \lambda_{\max}$. Introduce the event

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

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 $\lambda_{\text{opt}} \le \lambda_{\max}$ So there exists an integer $0 \le i \le p$ such that $\lambda_i \le \lambda_{\text{opt}} < 2\lambda_i$. For this integer i, we have

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

Particular case: $(E_2)^c$ occurs

For $i = 0$, we have

$$\begin{aligned}
&\tfrac{\bar{V}(\tilde{\rho})}{1-\lambda_i G(\lambda_i)} + \tfrac{B^2}{N}\tfrac{K(\tilde{\rho},\pi)+\log[(\eta\epsilon)^{-1}]}{\lambda_i[1-\lambda_i G(\lambda_i)]} \\
&\qquad = 2\bar{V}(\tilde{\rho}) + \tfrac{2B^2}{N}\tfrac{K(\tilde{\rho},\pi)+\log[(\eta\epsilon)^{-1}]}{\lambda_{\max}}
\end{aligned}$$

and

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

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

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

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

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

hence

$$\left\{ \begin{array}{l} (\eta)^{-1} = \left\lfloor \dfrac{\log \frac{4\lambda_{\max}}{\lambda_{\min}}}{\log 2} \right\rfloor \le L_1 \\ (\zeta)^{-1} = \left\lfloor \dfrac{\log \frac{4\beta_{\max}}{\beta_{\min}}}{\log 2} \right\rfloor \le L_2 \end{array} \right. .$$

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

7.4. **Proof of Theorem 4.5.** The result mainly comes from Lemma 4.4 and Corollary 4.3 since an aggregating procedure minimizing

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

wrt the probability distribution ρ is such that

$$\mathbb{B}\big(\hat{\rho}, (\lambda_i), (\eta_i), (\beta_j), (\zeta_j)\big) \le \mathbb{B}\big(\tilde{\rho}, (\lambda_i), (\eta_i), (\beta_j), (\zeta_j)\big). \tag{7.11}$$

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

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

Now, recall that to prove this inequality, we put ourselves on a subset of the event: for any distribution $\rho \in \mathcal{M}_+^1(\Theta)$, for any β in the grid introduced in Section 7.3, we have

$$-V \leq -\frac{\bar{V}}{1+\beta g(\beta)} + \frac{B^2}{2N}\frac{2K(\rho,\pi)+\log(L_2\epsilon^{-1})}{\beta[1+\beta g(\beta)]}.$$

Taking $\beta = \beta_{\max}$, we obtain $\bar{V} \leq 2V + \frac{B^2}{2N}\frac{2K(\rho,\pi)+\log(L_2\epsilon^{-1})}{\beta_{\max}}$, which leads to the desired inequality.

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 w.r.t. μ. Inversely, consider ρ such that $\mathrm{supp}(\rho)$ is strictly included $\mathrm{supp}(\mu)$. Let $A \triangleq \mathrm{supp}(\mu) - \mathrm{supp}(\rho)$. 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 $\mathrm{supp}(\rho') = \mathrm{supp}(\mu)$. 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

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

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 $\mathbb{E}_{\mu_{\langle z,h\rangle}}h = \mathbb{E}_\rho h$. With this end in view, we introduce

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

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)[7]. So, for any $z \in \mathbb{R}^N$, $\rho(\langle z,h\rangle - \mathbb{E}_\rho\langle z,h\rangle > 0) > 0$, hence $\mu(\langle z,h\rangle - \mathbb{E}_\rho\langle z,h\rangle > 0) > 0$. 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 $\eta_{\frac{1}{n}}(u_n) \geq \frac{1}{n}$. 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

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

which is absurd. For this real β, we have

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

[7] For $h = \mathrm{Cst}$ μ–a.s., the result is trivial

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 $\mathbb{E}_{\mu_{\langle z,h\rangle}} h - \mathbb{E}_\rho h = \bigtriangledown\chi_\rho(z) = 0$. Hence,

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

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

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

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

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, $p_{\emptyset^\perp} \triangleq \mathrm{Id}_{\mathbb{R}^N}$). 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

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

for any $l \in \{1,\dots,L\}$ and $z_n \in \mathrm{Span}(v_1,\dots,v_L)$. 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 $b_n = \mathrm{o}(a_n)$. Even if it means to consider a subsequence of $(z_n)_{n\in\mathbb{N}}$, one can assume that for any $l \in \{1,\dots,L\}$, $\lambda_{n,l} \underset{n\to+\infty}{\longrightarrow} \lambda_l \in \mathbb{R}_+ \cup \{+\infty\}$. Let $\lambda_0 \triangleq +\infty$ and $L' \triangleq \max\big\{l \in \{0,\dots,L\} : \lambda_l = +\infty\big\}$. Introduce the following family of subsets of Θ:

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

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 $\limsup_{n\to+\infty} K(\mu_{\langle z_n,h\rangle},\mu) < +\infty$) 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, $\bigtriangledown\varphi(z) = \mathbb{V}\mathrm{ar}_{\mu_{\langle z,h\rangle}} h(\mathbb{E}_{\mu_{\langle z,h\rangle}} h + z)$, where $\mathbb{V}\mathrm{ar}_{\mu_{\langle z,h\rangle}} h$ 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 $\mathbb{V}\mathrm{ar}_{\mu_{\langle z,h\rangle}} h$ 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 $\bar{\rho} = \mu_{-\langle \mathbb{E}_{\bar{\rho}} h, h\rangle}$:

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

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

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

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

$$\nabla_r \bar{\varphi}(w^l) = ac\mathbb{V}\mathrm{ar}_{\pi^{w^l}} f(X)\,\big|_r z^l$$

hence

$$\begin{array}{rcl} \bar{\varphi}(w^{l+1}) - \bar{\varphi}(w^l) & = & \langle w^{l+1} - w^l, \nabla \bar{\varphi}(w^l)\rangle + o(\|w^{l+1} - w^l\|) \\ & = & -ac\alpha(z^l)'\mathbb{V}\mathrm{ar}_{\pi^{w^l}} f(X)\,\big|_r z^l + o(\alpha). \end{array}$$

The covariance matrix $\mathbb{V}\mathrm{ar}_{\pi^{w^l}} f(X)\,\big|_r$ 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 the 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 7.4. *Let Z be a positive random variable. If $\mathbb{E}\, e^{\alpha\sqrt{Z}} \le M'$ for some $\alpha > 0$ and $M' > 0$, then, for any $\eta \ge 0$ and $A \ge \big(\frac{2}{\alpha}\big)^2$,*

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

Proof. For any $A \ge \big(\frac{2}{\alpha}\big)^2$,

$$\begin{aligned} \mathbb{E}Z - Z &\le \mathbb{E}(Z\mathbb{1}_{Z\ge A}) + \mathbb{E}(Z\mathbb{1}_{Z<A}) - Z\mathbb{1}_{Z<A} \\ &\le \mathbb{E}\big(e^{\alpha\sqrt{Z}} \sup_{u\ge A} u e^{-\alpha\sqrt{u}}\big) + \mathbb{E}(Z\mathbb{1}_{Z<A}) - Z\mathbb{1}_{Z<A} \\ &\le M' A e^{-\alpha\sqrt{A}} + \mathbb{E}(Z\mathbb{1}_{Z<A}) - Z\mathbb{1}_{Z<A} \end{aligned}$$

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

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

□

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

Lemma 7.5. *For any $\epsilon \ge e^{-\kappa_3 N}$, we have*

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

where

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

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 $\mu \in \mathbb{R}$,

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

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

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

From the previous lemma, we get for any $A \ge \big(\frac{2}{\alpha}\big)^2$,

$$\mathbb{P}^{\otimes N}(R(\tilde{f}) - r(\tilde{f}) > \mu) \le \exp\Big\{ - N\lambda\mu + N\lambda M' A e^{-\alpha\sqrt{A}} + N\tfrac{\lambda^2}{2} A R(\tilde{f})\Big\} \le \epsilon,$$

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

$$\log\Big(M e^{\alpha B + 1} \sqrt{\frac{N}{2\log(\epsilon^{-1})\alpha^2 R(\tilde{f})}}\Big) \ge 3,$$

equivalently, $M^2e^{2\alpha B}\frac{N}{2\log(\epsilon^{-1})\alpha^2R(\tilde{f})} \geq e^4$ and $\frac{M^2e^{2\alpha B-4}}{2\alpha^2R(\tilde{f})}N \geq \log(\epsilon^{-1})$. 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 $\epsilon \geq e^{-\kappa_3 N}$, we have $(\tilde{L}-1)^2 \geq 4$ as required. □

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

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

Then, using Lemma 7.5, we obtain that with probability at least $1-3\epsilon$,

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

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 $c = \tilde{L}^2\sqrt{\frac{\log(\epsilon^{-1})}{2\alpha^2N}}$, we get

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

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

General case: $R(\tilde{f}) \geq \frac{4}{\kappa_1}\frac{\log(\epsilon^{-1})}{N}B^2$ 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

$$\tilde{L} \leq \log\Big(\frac{Me^{\alpha B+1}}{\alpha B}\sqrt{\frac{\kappa_1}{8}}\frac{N}{\log(\epsilon^{-1})}\Big),$$

hence

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

where $\mathcal{L} \triangleq \frac{1}{\sqrt{2}\alpha}\Big[\log\Big(\kappa_4\frac{N}{\log(\epsilon^{-1})}\Big)\Big]^2$ and $\kappa_4 \triangleq \frac{Me^{\alpha B+1}}{\alpha B}\sqrt{\frac{\kappa_1}{8}}$. This leads to the desired result.

Particular case: $R(\tilde{f}) < \frac{4}{\kappa_1}\frac{\log(\epsilon^{-1})}{N}B^2$ occurs

From Corollary 4.3, with probability at least $1-2\epsilon$, we have

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

The announced inequality is also true in this case.

Model selection and Remark 4.3

We recall here the convergence rate of model selection. The target of model selection is to find a procedure doing as well as the best function among d prediction functions $f_1, \dots, f_d$, up to a remainder term called the convergence rate of model selection.

In [4], Catoni proves that by using a progressive Gibbs mixture $\hat{f}$, for any probability satisfying the assumptions (2.1) and (2.2), we have

$$\mathbb{P}^{\otimes N} R(\hat{f}) - \min_{i \in \{1,\dots,d\}} R(f_i) \le C \tfrac{\log d}{N}.$$

Theorem 3.1 provides a different result which is weaker as far as model selection is concerned. However, it allows to prove that the empirical risk minimizer $\hat{f}_{\mathrm{ERM}}$ on the second sample over the functions (built on the first sample) associated with the $(\lambda, \beta)-$grid, which will be denoted $\mathcal{G}$, satisfies with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,

$$R(\hat{f}_{\mathrm{ERM}}) - \min_{(\lambda,\beta)\in\mathcal{G}} R\big(\mathbb{E}_{\hat{\rho}_{\lambda,\beta}(d\theta)} f_\theta\big) \le C\Big\{\sqrt{\tilde{\mathcal{C}}(\epsilon) V(\tilde{\rho})} \vee \tilde{\mathcal{C}}(\epsilon)\Big\},$$

where $\tilde{\mathcal{C}}(\epsilon) \triangleq \frac{K(\tilde{\rho},\pi)+\log[\log(3N)\epsilon^{-1}]}{N}$.

Proof. Let $\mathcal{R}_2 \triangleq \big\{\mathbb{E}_{\hat{\rho}_{\lambda,\beta}(d\theta)} f_\theta : (\lambda,\beta) \in \mathcal{G}\big\}$, and let N_1 and N_2 be the respective sizes of the first and second sample. Let $\check{f} \in \mathrm{argmin}_{f\in\mathcal{R}_2} R(f)$. The set $\mathcal{R}_2 \subset \mathcal{C}(\mathcal{R})$ is interesting since its cardinal is small: $|\mathcal{R}_2| = |\mathcal{G}| \le L_1 L_2$ (with $N \leftarrow N_1$) and from Theorem 4.5, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,

$$R(\check{f}) \le R(\tilde{f}) + \gamma'_{N\leftarrow N_1}(\epsilon/2), \tag{.15}$$

where we recall that $\tilde{f} = \mathrm{argmin}_{f\in\mathcal{C}(\mathcal{R})} R(f)$.

Introduce $\tilde{f}_2$ the best convex combination of functions in $\mathcal{R}_2$. Since $\mathcal{R}_2 \subset \mathcal{R}$, we have $R(\check{f}) \ge R(\tilde{f}_2) \ge R(\tilde{f})$. Let r_2 denote the empirical risk on the second sample. Define $\lambda_0 \in]0; \frac{\alpha B}{2}[$ as $\lambda_0 G(\lambda_0) = \frac{1}{2}$. Taking $\lambda = \lambda_0$, Theorem 3.1 applied to a uniform prior distribution on $\mathcal{R}_2$ gives

$$R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f}_2) \le 2\big[r_2(\check{f}) - r_2(\tilde{f}_2)\big] + \tfrac{2B^2}{\lambda_0 N_2}\big[\log|\mathcal{R}_2| + \log(\epsilon^{-1})\big]. \tag{.16}$$

Since Lemma 7.2 still holds when $Z_\theta \leftarrow -Z_\theta$, for any $\pi_2 \in \mathcal{M}^1_+(\mathcal{R}_2)$, with probability at least $1-\epsilon$ wrt the second sample distribution, for any $\rho_2 \in \mathcal{M}^1_+(\mathcal{R}_2)$, we have the same kind of formula as in Theorem 3.1:

$$\mathbb{E}_{\rho_2} r_2 - r_2(\tilde{f}_2) \le \big[1+\lambda G(\lambda)\big]\big[\mathbb{E}_{\rho_2} R - R(\tilde{f}_2)\big] + \tfrac{B^2[K(\rho_2,\pi_2)+\log(\epsilon^{-1})]}{\lambda N_2}.$$

Taking $\lambda = \lambda_0$ and $\pi_2 = \rho_2 = \delta_{\check{f}}$, we obtain

$$r_2(\check{f}) - r_2(\tilde{f}_2) \le \tfrac{3}{2}\big[R(\check{f}) - R(\tilde{f}_2)\big] + \tfrac{B^2\log(\epsilon^{-1})}{\lambda_0 N_2}. \tag{.17}$$

From inequalities (.15), (.16) and (.17), we obtain that with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,

$$\begin{array}{rcl} R(\hat{f}_{\mathrm{ERM}}) & \le & 3R(\check{f}) - 2R(\tilde{f}_2) + \frac{2B^2}{\lambda_0 N_2}\log(L_1L_2\epsilon^{-2}) \\ & \le & R(\tilde{f}) + 3\gamma'_{N\leftarrow N_1}(\epsilon/2) + \frac{2B^2}{\lambda_0 N_2}\log(L_1L_2\epsilon^{-2}) \\ & \le & R(\tilde{f}) + C\Big\{\sqrt{\tilde{\mathcal{C}}(\epsilon)V(\tilde{\rho})} \vee \tilde{\mathcal{C}}(\epsilon)\Big\} \end{array}$$

provided that N_1 and N_2 has the order of N. □

Remark .6. Since the procedure is independent from the confidence level, we may integrate the deviations to obtain $\mathbb{P}^{\otimes N} R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f}) \leq C\Big\{\sqrt{\tilde{\mathcal{C}}(1)V(\tilde{\rho})} \vee \tilde{\mathcal{C}}(1)\Big\}$ for an appropriate different constant $C > 0$.

References

1. A. Barron and Y. Yang, *Information-theoretic determination of minimax rates of convergence*, Ann. Stat. **27** (1999), no. 5, 1564–1599.
2. L. Breiman, *Bagging predictors*, Mach. Learn. **24** (1996), no. 2, 123–140.
3. ———, *Arcing classifiers*, Ann. Stat. **26** (1998), no. 3, 801–849.
4. O. Catoni, *Statistical learning theory and stochastic optimization,* Lecture notes, Saint-Flour summer school on Probability Theory, 2001, Springer, to be published.
5. Y. Freund and R. E. Schapire, *Experiments with a new boosting algorithm*, Machine Learning: Proceedings of the Thirteenth International Conference (1996), 148–156.
6. J. Friedman, T. Hastie, and R. Tibshirani, *Additive logistic regression: a statistical view of boosting*, (1998), Technical Report, Dept. of Statistics, Stanford University.
7. A. Juditsky and A. Nemirovski, *Functional aggregation for nonparametric estimation*, Ann. Stat. **28** (2000), 681–712.
8. E. Mammen and A.B. Tsybakov, *Smooth discrimination analysis*, Ann. Stat. **27** (1999), 1808–1829.
9. D. A. McAllester, *PAC-Bayesian stochastic model selection*, Mach. Learn. **51** (2003), no. 1, 5–21.
10. A. Nemirovski, *Lectures on probability theory and statistics. Part II: topics in non-parametric statistics*, Springer-Verlag, Probability summer school, Saint Flour 1998.
11. G. Rätsch, M. Warmuth, S. Mika, T. Onoda, S. Lemm, and K.-R. Müller, *Barrier boosting*, Proceedings of the 13th annual conference on Computational Learning Theory (N. Cesa-Bianchi and S. A. Goldman, eds.), Morgan Kaufmann, 2000, pp. 170–179.
12. R. E. Schapire and Y. Singer, *Improved boosting algorithms using confidence-rated predictions*, Mach. Learn. **37** (1999), no. 3, 297–336.
13. A.B Tsybakov, *Optimal rates of aggregation*, Computational Learning Theory and Kernel Machines, Lecture Notes in Artificial Intelligence (B.Scholkopf and M.Warmuth, eds.), vol. 2777, Springer, Heidelberg, 2003, pp. 303–313.
14. A.B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Ann. Stat. **32** (2004), no. 1.
15. Y. Yang, *Aggregating regression procedures for a better performance*, Bernoulli **10** (2004), 25–47.

A BETTER VARIANCE CONTROL FOR PAC-BAYESIAN CLASSIFICATION

J.-Y. AUDIBERT
jyaudibe@ccr.jussieu.fr

Université Paris VI and CREST, France

Abstract. The common method to understand and improve classification rules is to prove bounds on the generalization error. Here we provide localized data-based PAC-bounds for the difference between the risk of any two randomized estimators. We derive from these bounds two types of algorithms: the first one uses combinatorial technics and is related to compression schemes whereas the second one involves Gibbs estimators.

We also recover some of the results of the Vapnik-Chervonenkis theory and improve them by taking into account the variance term measured by the pseudo-distance $(f_1, f_2) \mapsto \mathbb{P}[f_1(X) \neq f_2(X)]$.

Finally, we present different ways of localizing the results in order to improve the bounds and make them less dependent on the choice of the prior. For some classes of functions (such as VC-classes), this will lead to gain a logarithmic factor without using the chaining technique (see [1] for more details).

Contents

2000 *Mathematics Subject Classification.* 62H30, 94A17, 68Q32.

Key words and phrases. Statistical learning theory, compression schemes, Gibbs classifiers, error bounds, adaptive estimator, oracle inequalities, VC theory.

I am very grateful to my PhD advisor, Professor Olivier Catoni, for his constant, kind and generous support during the working out of this paper, which has been much inspired by his recent research on PAC-Bayesian classification ([7]).

1. Setup and notations

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 $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}}) \triangleq (\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Y}})$, $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ is a measurable space called the pattern space, $\mathcal{Y} = \{1, \dots, |\mathcal{Y}|\}$ is the (finite) label space and $\mathcal{B}_{\mathcal{Y}}$ is the sigma algebra of all subsets of $\mathcal{Y}$. Let $\mathbb{P}(dY|X)$ denote a regular version of the conditional probabilities (which we will use in the following without further mention).

Let $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ denote the set of all measurable functions mapping $\mathcal{X}$ into $\mathcal{Y}$. The aim of a classification procedure is to build a function $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ from the learning sample such that $f(X)$ well predicts the label Y associated with X. The quality of the prediction is measured by the expected risk

$$R(f) \triangleq \mathbb{P}[Y \neq f(X)].$$

A function f^* such that for any $x \in \mathcal{X}$,

$$f^*(x) \in \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbb{P}(Y = y|X = x),$$

minimizes the expected risk. This function is not necessarily unique. We assume that there exists one which is measurable. We will once for all fix it and refer to it as the Bayes classifier. The regression function will be denoted

$$\eta^*(x) \triangleq \mathbb{P}(Y|X = x).$$

Since we have no prior information about the distribution $\mathbb{P}$ of (X, Y), the regression function and the Bayes classifier are unknown.

It is well known that there is generally no measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$ such that

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

So we have to work with a prescribed set of classification functions $\mathcal{F}$, called the model. This set is just some subset of the set of all measurable functions $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. Let us denote $\tilde{f}$ the best function in the model, i.e. a function minimizing the expected risk:

$$\tilde{f} \in \operatorname*{argmin}_{\mathcal{F}} R.$$

For sake of simplicity, we assume that it exists[1]. The empirical risk

$$r(f) \triangleq \bar{\mathbb{P}}[Y \neq f(X)],$$

where

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

gives an estimate of the expected risk. An estimator which minimizes the empirical risk

$$\hat{f}_{\text{ERM}} \in \operatorname*{argmin}_{\mathcal{F}} r$$

[1] Otherwise we would have to introduce some small positive real β and consider $\tilde{f}$ as an estimator minimizing the expected risk up to β. This real β would then appear in all the equations related to this function and make things needlessly messy.

is called an ERM[2]-classifier.

Since we will study randomized estimators, we assume that we have a $\sigma-$algebra $\mathcal{T}$ such that $(\mathcal{F},\mathcal{T})$ is a measurable space containing the sets $\{f\}$ for any $f \in \mathcal{F}$ and such that the function

$$\begin{array}{lcl} \mathcal{F} \times \mathcal{X} & \to & \mathcal{Y} \\ (f,x) & \mapsto & f(x) \end{array}$$

is measurable. A randomized estimator consists in drawing a function in $\mathcal{F}$ according to some random distribution $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}_+^1(\mathcal{F})$, where $\mathcal{M}_+^1(\mathcal{F})$ is the set of probability distributions on the measurable space $(\mathcal{F},\mathcal{T})$.

To shorten notations, we will use μh to denote the expectation of the random variable h under the probability distribution μ: $\mu h \triangleq \int h(x)d\mu(x)$. The symbol C will denote a positive universal constant whose value may differ from line to line. We define

$$\pi_h \triangleq \frac{\exp(h)}{\pi \exp(h)} \cdot \pi$$

for any measurable real function h such that $\exp(h)$ is π-integrable. Most of the posterior distributions encountered in this paper will have this form. The randomized estimators associated with the posterior distributions π_{-Cr} will be called the standard Gibbs estimators with temperature $\frac{1}{C}$.

Let us recall some basic properties of the Kullback-Leibler divergence defined as

$$K(\mu,\nu) \triangleq \begin{cases} \mu \log\left(\frac{\mu}{\nu}\right) & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise,} \end{cases}$$

where ν and μ are two probability distributions on a measurable set $(A,\mathcal{A})$. The Legendre transform of the convex function $\mu \mapsto K(\mu,\nu)$ is given by the following formula: for any measurable function $h : A \mapsto \mathbb{R}$,

$$\sup_{\mu \in \mathcal{M}_+^1(A)} \big\{\mu h - K(\mu,\nu)\big\} = \log \nu \exp(h), \tag{1.1}$$

where, by convention:

$$\begin{cases} \mu h \triangleq \sup_{H \in \mathbb{R}} \mu(H \wedge h) \\ \mu h - K(\mu,\nu) = -\infty \text{ if } K(\mu,\nu) = +\infty \end{cases}.$$

Moreover, when the measurable function $\exp(h)$ is ν-integrable, the probability distribution ν_h achieves the supremum.

In this paper, we will consider prior distributions which may depend on the data. Most of them will depend on the data in an almost exchangeable way according to the following definition.

Definition 1.1. A function Q on $\mathcal{Z}^{2N}$ is said to be almost exchangeable iff it satisfies: for any permutation σ such that for any $i \in \{1,\dots,N\}$, we have $\{\sigma(i),\sigma(N+i)\} = \{i, N+i\}$, the following equality holds

$$Q_{Z_{\sigma(1)},\dots,Z_{\sigma(2N)}} = Q_{Z_1,\dots,Z_{2N}}.$$

To shorten, we will sometimes write Q for $Q_{Z_1^{2N}}$.

[2] ERM = Empirical Risk Minimization

Finally, to circumvent some measurability problems, we will consider inner and outer expectations. Let $(A, \mathcal{A}, \mu)$ be a measure space and $\mathcal{C}(A; \mathbb{R})$ be the class of real measurable functions. For any (measurable or not) function f, its inner and outer expectation wrt μ are respectively $\mu_*(h) \triangleq \sup\Big\{\mu(g) : g \in \mathcal{C}(A; \mathbb{R}), g \leq h\Big\}$ and $\mu^*(h) \triangleq \inf\Big\{\mu(g) : g \in \mathcal{C}(A; \mathbb{R}), g \geq h\Big\}$. Naturally, for any set $B \subset A$, $\mu_*(B)$ and $\mu^*(B)$ are defined by $\mu_*(B) = \mu_*(\mathbb{1}_B)$ and $\mu^*(B) = \mu^*(\mathbb{1}_B)$. Note that μ_* and μ^* are not measures but satisfy $\mu^*(B) + \mu_*(B^c) = 1$ and $\mu^*(B_1 \cup B_2) \leq \mu^*(B_1) + \mu^*(B_2)$. Besides, if $\mu^*(h) < +\infty$, then there exists a random variable h^* such that $\mu^*(h) = \mu(h^*)$. For more details on properties of inner and outer expectations, see [20].

The paper is organized as follows. The next section is an introduction to generalization error bounds. Section 3 provides new classification rules which can be used for preventing a given classifier to overfit the data, choosing an algorithm among a family of algorithms and choosing the temperature of a Gibbs estimator. For all these algorithms, we give a guarantee on their efficiency. In particular, we prove that it is possible to empirically choose the Gibbs temperature such that under some Tsybakov's type assumptions the Gibbs classifier has the optimal convergence rate. The remainder of the paper, except Section 7, is dedicated to prove these generalization error bounds. Since some of the intermediate results are interesting by themselves, we produce them in separate sections. Sections 4 and 5 present relative data-dependent bounds in respectively the PAC-Bayesian and compression schemes frameworks. Section 6 proposes a tight bracketing of the efficiency of Gibbs estimators. Section 7 is just here to illustrate the sharpness of our bounds in the well-known setting of Vapnik-Chervonenkis theory. Finally, the unavoidable toolbox to prove the results of this paper is given in the self-contained Section 8. The PAC-bounds provided there are given in a general context such that it can be used for other loss functions than the classification one: $L[Y, f(X)] = \mathbb{1}_{Y \neq f(X)}$.

2. The different types of generalization error bounds in classification

To understand the tightness and the originality of the bounds presented in this paper, we need first to give some global vision on generalization error bounds. The concepts presented in this section are not specific to classification problems. It is similar for the other risks $R(f) = \mathbb{P}L[Y, f(X)]$ and $r(f) = \bar{\mathbb{P}}L[Y, f(X)]$ obtained for other loss functions - in particular for the L^2-risks for which $L[Y, f(X)] = [Y - f(X)]^2$.

2.1. First PAC-bounds.

The first PAC-bounds which have appeared in the literature are uniform deviation inequalities of the empirical risk: for any $\eta > 0$,

$$\mathbb{P}^{\otimes N}\big[\sup_{\mathcal{F}}\{R - r\} \geq \eta\big] \leq \psi_{\mathcal{F}}(\eta), \tag{2.1}$$

where $\psi_{\mathcal{F}}$ is some increasing function of η (which highly depends on the size -called *complexity* or *capacity*- of the model). This result is in general equivalent to the following assertions

- for any estimator $\hat{f}$ and $\eta > 0$,

$$\mathbb{P}^{\otimes N}\big[R(\hat{f}) - r(\hat{f}) \geq \eta\big] \leq \psi_{\mathcal{F}}(\eta). \tag{2.2}$$

- for any estimator $\hat{f}$ and $\epsilon > 0$,

$$\mathbb{P}^{\otimes N}\big[R(\hat{f}) - r(\hat{f}) \geq \gamma^{\mathcal{F}}(\epsilon)\big] \leq \epsilon, \tag{2.3}$$

  where $\gamma^{\mathcal{F}} = \psi_{\mathcal{F}}^{-1}$.
- for any $\epsilon > 0$,

$$\mathbb{P}^{\otimes N}\big[\sup_{\mathcal{F}}\{R - r\} \geq \gamma^{\mathcal{F}}(\epsilon)\big] \leq \epsilon. \tag{2.4}$$

Another way of presenting Inequality (2.4) is to say that for any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, for any function $f \in \mathcal{F}$, we have[3]

$$R(f) \leq r(f) + \gamma^{\mathcal{F}}(\epsilon).$$

For this kind of bounds, the best guarantee on the generalization ability of some classification procedure is obtained for the ERM-algorithm. For this estimator, we obtained that for any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$,

$$R(\hat{f}_{\text{ERM}}) \leq r(\hat{f}_{\text{ERM}}) + \gamma^{\mathcal{F}}(\epsilon).$$

This leads to

- an upper bound on the quantile of $R(\hat{f}_{\text{ERM}}) - R(\tilde{f})$: for any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, we have[4]

$$R(\hat{f}_{\text{ERM}}) \leq R(\tilde{f}) + \gamma^{\mathcal{F}}(\epsilon) + \sqrt{\frac{\log(\epsilon^{-1})}{2N}}$$

- an upper bound on the expected value of $R(\hat{f}_{\text{ERM}}) - R(\tilde{f})$:

$$\mathbb{P}^{\otimes N} R(\hat{f}_{\text{ERM}}) - R(\tilde{f}) \leq \int_0^1 \psi_{\mathcal{F}}(\eta) d\eta.$$

Besides, for any estimator, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, we have

$$R(\hat{f}) - R(\tilde{f}) \leq r(\hat{f}) - r(\tilde{f}) + \gamma^{\mathcal{F}}(\epsilon) + \sqrt{\frac{\log(\epsilon^{-1})}{2N}}.$$

For a large model, the complexity term can be so large that we prefer to look for the best function in a smaller model in order to get a better guarantee on the generalization error of our procedure. To fix the size of this smaller model, we first build a collection of embedded models $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ such that the union of the collection of models is equal to $\mathcal{F}$. Let $\hat{f}_{\text{ERM},\mathcal{F}_k}$ denote the ERM-algorithm relative to the model $\mathcal{F}_k$. The SRM[5]-algorithm is to use $\hat{f}_{\text{ERM},\mathcal{F}_{\hat{k}}}$ where

$$\hat{k} \triangleq \operatorname*{argmin}_{k \in \{1,2,\dots\}} r\big(\hat{f}_{\text{ERM},\mathcal{F}_k}\big) + \gamma^{\mathcal{F}_k}(\alpha_k \epsilon),$$

and where α_k are positive reals summing to one[6]. The real α_k is the weight given to the model $\mathcal{F}_k$. By using a union bound with these weights, we obtain that for

[3] this formulation justifies the prefix "PAC"(probably approximately correctly) given to this kind of bound.

[4] since, by using Hoeffding's inequality, we obtain $r(\tilde{f}) \leq R(\tilde{f}) + \sqrt{\frac{\log(\epsilon^{-1})}{2N}}$ with $\mathbb{P}^{\otimes N}$-probability $1 - \epsilon$.

[5] SRM = Structural Risk Minimization

[6] Once more, we do not bother with the existence of the argmin. Note that practitioners seem to skip the α_k when using the SRM principle.

any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$R\big(\hat{f}_{\mathrm{ERM},\mathcal{F}_{\hat{k}}}\big) \le r\big(\hat{f}_{\mathrm{ERM},\mathcal{F}_{\hat{k}}}\big) + \gamma^{\mathcal{F}}(\alpha_{\hat{k}}\epsilon).$$

To sum up, this type of bounds gives us a model selection algorithm and generalization error bounds for any estimator, which are minimized for the ERM-classifier.

For relatively small models (VC-classes for instance), the bounds are of order[7] $1/\sqrt{N}$ and are known to be suboptimal for some kind of probability distributions $\mathbb{P}$. In particular, when the unknown probability distribution is such that $R(\tilde{f})$ is small (i.e. has the order of $1/N^{\beta}$ with $\beta > 0$), the bound is known to be suboptimal ($\triangleq$ problem $\{1\}$). In this type of bounds, the deviations of the empirical risk of any function in the model is treated similarly without taking into account the relevance of the function to predict labels. From the central limit theorem, we know that the deviations of the empirical risk for the function f has the order of $\sqrt{\frac{R(f)[1-R(f)]}{N}}$. Therefore, when f is a good predictor (i.e. when the quantity $R(f)$ is small), the deviations are much smaller than when f is a poor classifier. This remark explains the suboptimality of this kind of bounds.

2.2. **First improvements.** To correct this last drawback, we have to allow $\gamma(\epsilon)$ to depend on f. Specifically, we now consider bounds of the following form : for any $\epsilon > 0$,

$$\mathbb{P}^{\otimes N}\big[\sup_{f\in\mathcal{F}}\{R(f) - r(f) - \gamma(f,\epsilon)\} \ge 0\big] \le \epsilon, \tag{2.5}$$

or in general equivalently, for any estimator $\hat{f}$ and $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$R(\hat{f}) \le r(\hat{f}) + \gamma(\hat{f},\epsilon). \tag{2.6}$$

From the previous discussion, we also see that we would like to take $\gamma(f,\epsilon)$ of the following form $\sqrt{R(f)}\gamma'(\epsilon)$. With this form, Inequality (2.6) can be written as

$$R(\hat{f}) \le \left(\sqrt{r(\hat{f}) + \frac{[\gamma'(\epsilon)]^2}{4}} + \gamma'(\epsilon)\right)^2.$$

This kind of bounds solves in general the problem $\{1\}$. For instance, in [22, 23], Vapnik and Chervonenkis obtained that for any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$\gamma'(\epsilon) = 2\sqrt{\frac{\log(\epsilon^{-1}) + \log[4\mathbb{S}^{\mathcal{F}}(2N)]}{N}}.$$

Therefore, when the model has a finite VC-dimension and when the minimum of the empirical risk has the order of $1/N^{\beta}$ for some $\beta \in \mathbb{R}_+ \cup \{+\infty\}$, the bound on $R(\hat{f})$ has the order of $\frac{1}{N^{\frac{1+\beta}{2}\wedge 1}}$.

However, in noisy classification tasks, we still have not $o(1/\sqrt{N})$-bounds for the relative expected risk $R(\hat{f}) - R(\tilde{f})$ when the probability distribution $\mathbb{P}$ has some

[7] In [21], Vapnik and Chervonenkis obtained $\gamma^{\mathcal{F}}(\epsilon) = \sqrt{8\frac{\log(\epsilon^{-1})+\log[4\mathbb{S}^{\mathcal{F}}(2N)]}{N}}$ where the shatter coefficient $\mathbb{S}^{\mathcal{F}}(N)$ is the maximal number of different sets $\{(f(x_1),\dots,f(x_N)) : f \in \mathcal{F}\}$ among all the possible input sets $(x_1,\dots,x_N)$ of size N. For VC-classes, there exists an integer h called the VC-dimension such that $\log[\mathbb{S}^{\mathcal{F}}(N)] \le h\log(eN/h)$.

particular form. Indeed, for any estimator $\hat{f}$, with $\mathbb{P}^{\otimes N}$-probability at least $1-2\epsilon$, we only have

$$R(\hat{f}) - R(\tilde{f}) \leq \left(\sqrt{r(\hat{f}) + \frac{[\gamma'(\epsilon)]^2}{4}} + \gamma'(\epsilon) \right)^2 - r(\tilde{f}) + \sqrt{\frac{\log(\epsilon^{-1})}{2N}}$$

since we separately deal with the deviations of $r(\tilde{f})$ and those of $r(\hat{f})$ and we cannot expect to have much better than with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, $r(\tilde{f}) \leq R(\tilde{f}) + \sqrt{\frac{\log(\epsilon^{-1})}{2N}}$ in noisy classification. Note that we cannot expect to obtain $o(1/\sqrt{N})$-bounds for any probability distribution $\mathbb{P}$ since in [22, 8] it has been proven that, when the model $\mathcal{F}$ has a VC-dimension $h \geq 2$ and when $N \geq 14h$, for any estimator $\hat{f}$, there exists some probability such that

$$\mathbb{P}^{\otimes N} R(\hat{f}) - R(\tilde{f}) \geq 10^{-5} \sqrt{\frac{h-1}{N}}.$$

So the next target is to find some kind of oracle inequalities which show that the estimators minimizing the bounds adapt themselves to the unknown distribution.

2.3. **Relative PAC-bounds.** One way of improving the previous bounds is to deal simultaneously with both the deviations of the functions f and $\tilde{f}$. So far, we have been adding these deviations. There is hope that, for some models $\mathcal{F}$ and probability distributions $\mathbb{P}$, the first order deviation terms of $r(f)$ and $r(\tilde{f})$ compensate themselves and that finally the bounds are driven by second order terms. This kind of bounds has the following form: for any $\epsilon > 0$,

$$\mathbb{P}^{\otimes N} \left(\sup_{f \in \mathcal{F}} \{ R(f) - r(f) - R(\tilde{f}) + r(\tilde{f}) - \gamma(f,\epsilon) \} \geq 0 \right) \leq \epsilon. \tag{2.7}$$

Once more, the central limit theorem advises us to take $\gamma(f,\epsilon)$ as

$$\gamma(f,\epsilon) = \sqrt{\mathbb{V}\mathrm{ar}_{\mathbb{P}} \left[L[Y, f(X)] - L[Y, \tilde{f}(X)] \right]} \gamma'(\epsilon)$$

for an appropriate function γ'. Equation (2.7) can also be written as: for any measurable estimator $\hat{f}$ and any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$R(\hat{f}) - R(\tilde{f}) \leq r(\hat{f}) - r(\tilde{f}) + \gamma(\hat{f},\epsilon).$$

Once we have succeeded in obtaining such bounds, the last step is to get bounds in which the unknown distribution $\mathbb{P}$ does not appear. To obtain this, we have to succeed in replacing $\mathbb{P}$ by its empirical version $\bar{\mathbb{P}}$ in the variance term.

This strategy to get tight bounds has already been addressed in the literature ([11, 9, 2, 17, 6]). However these results present different drawbacks:

- the unknown probability distribution $\mathbb{P}$ appears in the bound ([6, 2][8]),
- in binary classification ($\mathcal{Y} = \{0;1\}$), the bounds only hold when we have the two following assumptions ([11, 9, 2, 17])
  - $f^* = \tilde{f}$, i.e. the model contains the Bayes classifier,
  - $\mathbb{P}[|\eta^*(X) - 1/2| \geq t] \leq \check{C} t^{\alpha}$ for some $\alpha > 0$ and $\check{C} > 0$ and any $t > 0$, which roughly means that the regression function $\eta^*(X)$ is not with too high-probability close to 1/2,

[8] In [2], Sections 6.3 which deal with sample-based bounds do not concern *relative* PAC-bounds in *classification*.

- the bounds are not localized: the global size of the model appears, the complexity is not only computed on the "best" part of the model ([11])[9].

This paper will provide localized sample-based relative PAC-bounds for classification which have not these drawbacks and from which we can derive the algorithms presented in the following section.

3. Classification using relative data-dependent bounds

In this section, we will give new algorithms improving the variance estimation by comparing the efficiency of various estimators. These algorithms will be first described in the transductive setting since it allows to have simpler formulae and proofs.

Our transductive setting is the following: we possess two samples of size N. The first sample is labeled: $\{(X_1, Y_1), \dots, (X_N, Y_N)\}$. The second one $\{X_{N+1}, \dots, X_{2N}\}$ has to be labeled: the outputs $\{Y_{N+1}, \dots, Y_{2N}\}$ are unknown.

We will use the following notations for the empirical distributions and empirical risks:

$$\begin{cases} \bar{\mathbb{P}} & \triangleq & \frac{1}{N}\sum_{i=1}^{N} \delta_{(X_i,Y_i)} \\ \bar{\mathbb{P}}' & \triangleq & \frac{1}{N}\sum_{i=N+1}^{2N} \delta_{(X_i,Y_i)} \\ \bar{\bar{\mathbb{P}}} & \triangleq & \frac{1}{2N}\sum_{i=1}^{2N} \delta_{(X_i,Y_i)} \\ r(f) & \triangleq & \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{Y_i \neq f(X_i)} = \bar{\mathbb{P}}[Y \neq f(X)] \\ r'(f) & \triangleq & \frac{1}{N}\sum_{i=N+1}^{2N} \mathbb{1}_{Y_i \neq f(X_i)} = \bar{\mathbb{P}}'[Y \neq f(X)] \end{cases}$$

The variance terms in concentration inequalities will have the following pseudo-distances appeared

$$\begin{cases} \bar{\bar{\mathbb{P}}}_{f_1,f_2} & \triangleq & \bar{\bar{\mathbb{P}}}[f_1(X) \neq f_2(X)] \\ \bar{\mathbb{P}}_{f_1,f_2} & \triangleq & \bar{\mathbb{P}}[f_1(X) \neq f_2(X)] \\ \bar{\mathbb{P}}'_{f_1,f_2} & \triangleq & \bar{\mathbb{P}}'[f_1(X) \neq f_2(X)] \\ \mathbb{P}_{f_1,f_2} & \triangleq & \mathbb{P}[f_1(X) \neq f_2(X)] \end{cases}.$$

3.1. Compression schemes complexity.

Consider an algorithm

$$\hat{f} : \bigcup_{n \in \mathbb{N}^*} \mathcal{Z}^n \times \mathcal{X} \to \mathcal{Y}$$

which produces for any $n \geq 1$ and any training set z_1^n the prediction function $\hat{f}_{z_1^n} : \mathcal{X} \to \mathcal{Y}$. Assume that the algorithm is exchangeable: for any n and any permutation σ of $\{1, \dots, n\}$, we have $\hat{f}_{z_1^n} = \hat{f}_{z_{\sigma(1)},\dots,z_{\sigma(n)}}$.

Let $\hat{\mathcal{F}}_h \triangleq \{\hat{f}_{(X_{i_j},y_i)_{j=1}^h} : (i_1, \dots, i_h) \in \{1, \dots, 2N\}^h, y_1^h \in \mathcal{Y}^h\}$. A natural exchangeable model associated with the algorithm and the data X_1^{2N} is $\hat{\mathcal{F}} \triangleq \bigcup_{2 \leq h \leq N} \hat{\mathcal{F}}_h$. For any function $f \in \hat{\mathcal{F}}$, let $h(f)$ be the smallest integer $2 \leq h \leq N$ such that $f \in \hat{\mathcal{F}}_h$. Let $\alpha \in]0;1[$. Define $\mathcal{C}(f) \triangleq h(f) \log\left(\frac{2N|\mathcal{Y}|}{\alpha}\right)$ the complexity of the function f. Finally, introduce $L \triangleq \log[(1-\alpha)^{-2}\alpha^4\epsilon^{-1}]$ and

$$S(f_1, f_2) \triangleq \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{f_1,f_2}[\mathcal{C}(f_1)+\mathcal{C}(f_2)+L]}{N}}.$$

[9] In [2, 9, 17], the model is localized via the variance $\mathbb{V}\mathrm{ar}_{\mathbb{P}}\big(L[Y, f(X)] - L[Y, \tilde{f}(X)]\big)$ to the extent that the complexity of the model is measured on a subset of functions with low variance. In classification, small variance implies small probability $\mathbb{P}_{f,\tilde{f}}$, hence $R(f)$ close to $R(\tilde{f})$. Note that the converse is not true in general: "f classifies well" does not imply small variance. But it holds under the previous margin assumption.

The following procedure gives a way of using the initial algorithm $\hat{f}$ to produce a classifier with a good guarantee of efficiency.

Algorithm 3.1. *Let $f_0 \in \hat{\mathcal{F}}_2$. For any $k \geq 1$, define $f_k \in \hat{\mathcal{F}}$ as a function with the smallest complexity such that $r(f_k) - r(f_{k-1}) + S(f_{k-1}, f_k) < 0$. Classify using the function obtained at the last iteration.*

The following theorem guarantees the efficiency of this procedure.

Theorem 3.1. *The iterative scheme is not infinite: there exists $K \in \mathbb{N}$ such that f_K exists but not f_{K+1}. With $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, for any $k \in \{1, \dots, K\}$, we have*

- *$r(f_k) < r(f_{k-1})$ and $r'(f_k) < r'(f_{k-1})$,*
- *$\mathcal{C}(f_k) \geq \mathcal{C}(f_{k-1})$,*
- *defining for any $f \in \hat{\mathcal{F}}$ the integer $k(f) \triangleq \max\big\{0 \leq k \leq K; \mathcal{C}(f_k) \leq \mathcal{C}(f)\big\}$,*

$$r'(f_K) \leq \min_{f \in \hat{\mathcal{F}}} \big\{r'(f) + 2S(f_{k(f)}, f)\big\}. \tag{3.1}$$

- $$r'(f_K) \leq \inf_{f \in \hat{\mathcal{F}}} \sup_{g \in \hat{\mathcal{F}} : \mathcal{C}(g) \leq \mathcal{C}(f)} \Big\{2r'(f) - r'(g) + 8\sqrt{\tfrac{2\bar{\bar{\mathbb{P}}}_{f,g}[2\mathcal{C}(f)+L]}{N}}\Big\}. \tag{3.2}$$

Proof. See Section 9.1. □

Remark 3.1. From the second assertion of the previous theorem, we are allowed to search f_k in $\underset{h(f_{k-1}) \leq h \leq N}{\cup} \hat{\mathcal{F}}_h$.

Remark 3.2. In Inequality (3.1), the variance term $\bar{\bar{\mathbb{P}}}_{f_{k(f)},f}$ depends on the functions $f_k, 0 \leq k \leq K$. To get rid of it, we can weaken the bound and obtain the following oracle inequality

$$r'(f_K) \leq \min_{f \in \hat{\mathcal{F}}} \Big\{r'(f) + 8\sqrt{\tfrac{2\mathcal{C}(f)+L}{2N}}\Big\}.$$

Inequality (3.2) provides a smarter way of taking care of the variance term.

Remark 3.3. In our algorithm, there are several possible choices for the function f_k. Only the set $\hat{\mathcal{F}}_{h_k}$, in which the function f_k is, is well determined. A natural choice consists in taking the minimizer of $r(f_k) - r(f_{k-1}) + S(f_{k-1}, f_k)$ in the set $\hat{\mathcal{F}}_{h_k}$. This function is not necessarily the ERM in $\hat{\mathcal{F}}_{h_k}$. However we can prove that the theoretical guarantee associated with this function is not more than $\sqrt{2}$ smaller than the one associated with the ERM on $\hat{\mathcal{F}}_h$. In other words, for any $2 \leq h \leq N$ we can restrict our search to the functions minimizing the empirical risk on $\hat{\mathcal{F}}_h$.

Remark 3.4. The parameter α essentially influences the constants in the bound. Taking $\frac{1}{2}$ or $\frac{3}{4}$ for α will not in general modify drastically the final classifier.

This compression scheme will be useful when the initial algorithm $\hat{f}$ tends to overfit the data (for instance, the 1-Nearest Neighbor algorithm, non pruned trees, Support Vector Machine in the separable case[10] when errors are heavily penalized, lowly regularized boosting methods such as Adaboost, ...). Besides, contrary to other compression schemes, our procedure takes into account the variance term

[10] It is in particular the case when we use the gaussian kernel and when the input data X_i are pairwise distinct.

so that we can expect much better results than for other compression schemes (specially in noisy classification tasks).

Since to scan all the possible subsets $\big\{(x_i,y_i)_{i=1}^h : x_1^h \subset X_1^{2N}, y_1^h \in \mathcal{Y}^h\big\}$ is not computationally tractable, we can use some suboptimal heuristics such as the following one.

Detailed algorithm 3.1. The function f_0 is chosen as the function in $\hat{\mathcal{F}}_2$ minimizing the empirical risk. Let $z_1^2 \in \mathcal{Z}^2$ such that $f_0 = \hat{f}_{z_1^2}$.

We repeat for any $k \geq 3$,

$$z_k = \underset{z_k \in \text{misclassified points in } Z_1^N - z_1^2}{\operatorname{argmin}} \Big\{ r(\hat{f}_{z_1^k}) - r(\hat{f}_{z_1^2}) + S(\hat{f}_{z_1^k}, \hat{f}_{z_1^2}) \Big\}.$$

until the minimum is negative (or until we have no more point to add). When the minimum is negative, we define $f_1 = \hat{f}_{z_1^{k_1}}$. To define f_2, we repeat for any $k \geq k_1 + 1$,

$$z_k = \underset{z_k \in \text{misclassified points in } Z_1^N - z_1^{k_1}}{\operatorname{argmin}} \Big\{ r(\hat{f}_{z_1^k}) - r(\hat{f}_{z_1^{k_1}}) + S(\hat{f}_{z_1^k}, \hat{f}_{z_1^{k_1}}) \Big\}.$$

until the minimum is negative (or until we have no more point to add), and so on. A less costly alternative is to stop when adding one more point increases the criterion (i.e. when the growth of complexity is no longer compensated by the diminution of the empirical risk). At the end, we classify using the function denoted f_K obtained for the last negative minimum.

Let $\bar{\mathcal{I}} \subset \mathcal{I}$ be the set of compression sets considered in the previous heuristics and define for any $f \in \bar{\mathcal{F}} \triangleq \big\{\hat{f}_I ; I \in \bar{\mathcal{I}}\big\}$ the integer

$$k(f) \triangleq \max\big\{0 \leq k \leq K ; \mathcal{C}(f_k) \leq \mathcal{C}(f)\big\}.$$

We have the following guarantee:

Theorem 3.2. *With $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, for any $k \in \{1,\dots,K\}$, we have*

- $r(f_k) < r(f_{k-1})$ *and* $r'(f_k) < r'(f_{k-1})$,
- $r'(f_K) \leq \min_{f \in \bar{\mathcal{F}}} \big\{ r'(f) + 2S(f_{k(f)}, f) \big\}$.
- $r'(f_K) \leq \inf\limits_{f \in \bar{\mathcal{F}}} \sup\limits_{g \in \bar{\mathcal{F}} : \mathcal{C}(g) \leq \mathcal{C}(f)} \Big\{ 2r'(f) - r'(g) + 8\sqrt{\frac{2\bar{\bar{\mathbb{P}}}_{f,g}[2\mathcal{C}(f)+L]}{N}} \Big\}$.

Proof. The proof is similar to the one of Theorem 3.1. □

3.2. **PAC-Bayesian complexity.**

3.2.1. *Kullback-Leibler complexity.* In this section, the complexity of a randomized estimator is measured by the KL-divergence between the posterior distribution and a prior distribution π which is introduced in order to put a structure on the model. This approach pioneered by McAllester [16] has been developed in [5, 18, 7] among others.

For any $\epsilon > 0$, $\lambda > 0$ and $\rho', \rho'' \in \mathcal{M}_+^1(\mathcal{F})$, let

$$\begin{cases} L & \triangleq & \log[\log(eN)\epsilon^{-1}] \\ \tilde{\mathcal{K}}_{\rho',\rho''} & \triangleq & K(\rho',\pi) + K(\rho'',\pi) + L \\ S_\lambda(\rho',\rho'') & \triangleq & \frac{2\lambda}{N}(\rho' \otimes \rho'')\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\sqrt{e}}{\lambda}\tilde{\mathcal{K}}_{\rho',\rho''} \\ S(\rho',\rho'') & \triangleq & \min\limits_{\lambda \in [\sqrt{N};N]} S_\lambda(\rho',\rho'') \end{cases}$$

Algorithm 3.2. *Let $\rho_0 = \pi_{-\lambda_0 r}$. For any $k \geq 1$, define ρ_k as the distribution with the smallest complexity $K(\rho_k, \pi)$ such that $\rho_k r - \rho_{k-1} r + S(\rho_{k-1}, \rho_k) \leq 0$. Classify using a function drawn according to the posterior distribution obtained at the last iteration.*

The following result guarantees the efficiency of the randomized estimator.

Theorem 3.3. *Let*

$$\mathbb{G}(\lambda) \triangleq -\tfrac{1}{\lambda} \log \pi \exp\big(-\lambda r'\big) + \tfrac{1}{2\lambda} \log \pi_{-\lambda r'} \exp\Big(\tfrac{72\sqrt{e}\lambda^2}{N} \pi_{-\lambda r'} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + \tfrac{L}{2\lambda}. \tag{3.3}$$

The iterative scheme is not infinite: there exists $K \in \mathbb{N}$ such that ρ_K exists but not ρ_{K+1}. With $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, for any $k \in \{1, \ldots, K\}$, we have

- $\rho_k r - \rho_{k-1} r + S(\rho_k, \rho_{k-1}) = 0$,
- $\rho_k r < \rho_{k-1} r$ *and* $\rho_k r' \leq \rho_{k-1} r'$,
- $K(\rho_k, \pi) \geq K(\rho_{k-1}, \pi)$,
- $\rho_K r' \leq \min_{\frac{\sqrt{N}}{6\sqrt{e}} \leq \lambda \leq \frac{N}{6\sqrt{e}} : K(\pi_{-\lambda r'}, \pi) \geq K(\rho_0, \pi)} \mathbb{G}(\lambda)$.

Proof. See Section 9.2. □

Let us explain why we believe that the guarantee on the generalization ability of our procedure is tight and satisfactory. First, consider a prior distribution $\pi_{\mathcal{U}(X_1^{2N})}$ which is uniform on one of the smallest set $\mathcal{S}$ of functions such that for any $f \in \mathcal{F}$, there exists $f' \in \mathcal{S}$ equal to f on $\{X_1, \ldots, X_{2N}\}$. Using this prior distribution, we have

$$\mathbb{G}\bigg(\frac{\sqrt{N}}{6\sqrt{e}}\bigg) \leq r'(\tilde{f}') + C\sqrt{\frac{h \log\big(\frac{2eN}{h}\big) + \log(\epsilon^{-1})}{N}},$$

hence our randomized estimator achieves the optimal convergence rate for VC classes (up to the logarithmic factor).

Secondly, consider the following complexity and margin assumptions which will be refered to as (CM) assumptions:

- there exists $C' > 0$ and $0 < q < 1$ such that the covering entropy of the model $\mathcal{F}$ for the distance $\mathbb{P}_{\cdot,\cdot}$ satisfies for any $u > 0$, $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) \leq C' u^{-q}$,
- there exist $c'', C'' > 0$ and $\kappa \geq 1$ such that for any function $f \in \mathcal{F}$,

$$c''\big[R(f) - R(\tilde{f})\big]^{\frac{1}{\kappa}} \leq \mathbb{P}_{f,\tilde{f}} \leq C''\big[R(f) - R(\tilde{f})\big]^{\frac{1}{\kappa}},$$

where we recall that by definition $\tilde{f} \in \operatorname{argmin}_{\mathcal{F}} R$. Under (CM) assumptions, one can prove[11] that with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1 - \epsilon$,

$$\mathbb{G}(\lambda) \leq r'(\tilde{f}) + \log(e\epsilon^{-1}) \mathrm{O}\Big(N^{-\frac{\kappa}{2\kappa-1+q}}\Big)$$

provided that $\lambda_0 = 0$, $\lambda = N^{\frac{\kappa}{2\kappa-1+q}} (\in [\sqrt{N}; N])$ and π is taken independent from the data and such that

$$\pi\Big(\mathbb{P}_{\cdot,\tilde{f}} \leq \check{C}_1 N^{-\frac{1}{2\kappa-1+q}}\Big) \geq \exp\Big(-\check{C}_2 N^{-\frac{q}{2\kappa-1+q}}\Big) \tag{3.4}$$

for some constants $\check{C}_1$ and $\check{C}_2$. The convergence rate $N^{-\frac{\kappa}{2\kappa-1+q}}$ is known to be optimal in this situation (see [14, 19] for original results and [1] for more details on the assumptions and their implications).

[11] See Appendix B for the main lines.

Remark 3.5. Let us describe approximatively the quantity $\mathbb{G}$. It is made of three terms.

- The first term is a decreasing function wrt the parameter λ with limit equal to 0 when $\lambda \to +\infty$. It is linked to the error on the second sample associated with the randomized distributions $\pi_{-Cr'}$ through:
$$-\tfrac{1}{\lambda}\log \pi \exp\left(-\lambda r'\right) = \textstyle\int_0^1 \pi_{-\gamma\lambda r'} r' d\gamma.$$
- By Jensen's inequality, the second term is upperbounded with $36\sqrt{e}\frac{\lambda}{N}(\pi_{-\lambda[r'-72\sqrt{e}\frac{\lambda}{N}\pi_{-\lambda r'}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]} \otimes \pi_{-\lambda r'})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$,and lower bounded with
$$36\sqrt{e}\tfrac{\lambda}{N}(\pi_{-\lambda r'} \otimes \pi_{-\lambda r'})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}.$$
So it can be seen as a variance term.
- The last term roughly behaves as $\frac{1}{\lambda}$ (we neglect the $\log\log N$ factor).

Remark 3.6. Let us explain why the condition
$$\tfrac{\sqrt{N}}{6\sqrt{e}} \le \lambda \le \tfrac{N}{6\sqrt{e}} : K(\pi_{-\lambda r'}, \pi) \le K(\pi_{-\lambda_0 r}, \pi)$$
in the last assertion of Theorem 3.3 is not harmful.

Since we have $\bar{\bar{\mathbb{P}}}_{f_1,f_2} \ge \frac{1}{2}\bar{\mathbb{P}}'_{f_1,f_2} \ge \frac{r'(f_1)-r'(f_2)}{2}$, the second term in the quantity $\mathbb{G}$ is very loosely lower bounded by
$$\inf_{\eta>0}\Big\{\tfrac{1}{2\lambda}\log\Big(\pi(r' - \min_{\mathcal{F}} r' \ge \eta)\exp\big[-\lambda + \tfrac{36\sqrt{e}\eta\lambda^2}{N}\pi_{-\lambda r'}(r' - \min_{\mathcal{F}} \le \tfrac{\eta}{2})\big]\Big)\Big\}.$$
When $\lambda > N^{1+\beta}$ with $\beta > 0$, it is reasonable to believe that in general there will be a fixed $\eta > 0$ such that $\pi_{-\lambda r'}(r' - \min_{\mathcal{F}} \le \frac{\eta}{2}) \approx 1$ and $\pi(r' - \min_{\mathcal{F}} r' \ge \eta) \ge \frac{1}{2}$ so that the previous lower bound ensures that $\frac{1}{2\lambda}\log \pi_{-\lambda r'} \exp\Big(\frac{72\sqrt{e}\lambda^2}{N}\pi_{-\lambda r'}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)$ is at least of order $C\frac{\lambda}{N}$ (when $\lambda > N^{1+\beta}$ with $\beta > 0$). Therefore the condition $\lambda \le \frac{N}{6\sqrt{e}}$ can be disregarded. Let $\lambda'_{\min} \triangleq \frac{\sqrt{N}}{6\sqrt{e}}$. For any $\lambda \le \lambda'_{\min}$, we have
$$\begin{array}{rcl}\mathbb{G}(\lambda'_{\min}) & \le & -\frac{1}{\lambda'_{\min}}\log \pi \exp\big(-\lambda'_{\min} r'\big) + \frac{C}{\sqrt{N}} \\ & \le & -\frac{1}{\lambda}\log \pi \exp\big(-\lambda r'\big) + \frac{C}{\sqrt{N}}\end{array}$$
hence $\mathbb{G}(\lambda'_{\min}) - r'(\tilde{f}') = \mathrm{O}\big(\mathbb{G}(\lambda) - r'(\tilde{f}')\big)$. So the condition $\lambda \ge \lambda'_{\min}$ is not harmful wrt the order of the convergence rate. Note that the optimality of the procedure under (CM) assumptions also justifies to have restricted ourselves to Gibbs distribution with temperature in $\big[\frac{C}{N}; \frac{C}{\sqrt{N}}\big]$.

So the only strong constraint on λ is that $K(\pi_{-\lambda r'}, \pi) \ge K(\pi_{-\lambda_0 r}, \pi)$. Taking $\lambda_0 = 0$ solves this problem. However if we are not pleased with a poor starting distribution, a tempting choice is to take λ_0 of order $\sqrt{N}$ since it is very likely that $K(\pi_{-C\sqrt{N}r'}, \pi) \ge K(\pi_{-\frac{C\sqrt{N}}{2}r}, \pi)$[12].

[12] In fact, this assertion is not as trivial as it may seem. By symmetry and from the inequality $K(\pi_{-C\sqrt{N}r'}, \pi) \ge K(\pi_{-\frac{C\sqrt{N}}{2}r}, \pi)$, the assertion holds with $\mathbb{P}^{\otimes 2N}$-probability at least $\frac{1}{2}$. To prove that the inequality holds with high probability (up to unimportant additive quantities depending on the confidence level) requires most of the technical tools developed in this paper. The proof is left to highly determined readers. Naturally, the factor 2 in the inequality has no fundamental meaning: it can be replaced with any constant greater than 1 at the price that the confidence level term explodes when the constant goes to 1.

Now one can argue that the previous algorithm is hard to implement. Fortunately, if we search the posterior distribution only among the standard Gibbs distributions $\pi_{-\lambda r}$ of inverse temperature parameter λ belonging to a finite geometric grid of $[\sqrt{N}; N]$, we can prove[13] a similar guarantee as in Theorem 3.3 and shows its optimality for VC classes or under (CM) assumptions.

3.2.2. *Localized complexity.* Here we use localized complexities to choose the temperature of a standard Gibbs estimator in a finite grid. Specifically, we arbitrarily use the grid $\Lambda \triangleq \big\{\lambda_j \triangleq \sqrt{N} e^{\frac{j}{2}}; 0 \le j \le \log N\big\}$. Consider the randomized estimator associated with the posterior distribution $\pi_{-\lambda_j r}$. For any $0 \le j \le \log N$, its complexity is defined as $\mathcal{C}(j) \triangleq \log \pi_{-\lambda_j r} \exp\big(\frac{\lambda_j^2}{N} \pi_{-\lambda_j r} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)$. For any $0 \le i < j \le \log N$ and $\epsilon > 0$, we introduce $L \triangleq \log[\log^2(eN)\epsilon^{-1}]$ and

$$S(i,j) \triangleq \tfrac{2\lambda_j}{N} \big(\pi_{-\lambda_i r} \otimes \pi_{-\lambda_j r}\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{2\mathcal{C}(i)+2\mathcal{C}(j)+3L}{\lambda_j}.$$

The following algorithm appropriately chooses the integer $0 \le j \le \log N$ such that the associated Gibbs classifier satisfies a localized version of the guarantee in Theorem 3.3.

Algorithm 3.3. *Let $u(0) = 0$. For any $k \ge 1$, define $u(k)$ as the smallest integer $j \in]u(k-1); \log N]$ such that $\pi_{-\lambda_j r} r - \pi_{-\lambda_{u(k-1)} r} r + S\big(u(k-1), j\big) \le 0$. Classify using a function drawn according to the posterior distribution associated with the last $u(k)$.*

Theorem 3.4. *Let*
(3.5)

$$\mathbb{G}_{\mathrm{loc}}(j) \triangleq \pi_{-\lambda_{j-1} r'} r' + \frac{\sup_{0 \le i \le j} \Big\{ \log \pi_{-\lambda_i r'} \otimes \pi_{-\lambda_i r'} \exp\Big(\frac{C\lambda_i^2}{N} \bar{\mathbb{P}}'_{\cdot,\cdot}\Big)\Big\}}{\lambda_j} + C\frac{\log[\log(eN)\epsilon^{-1}]}{\lambda_j}$$

for an appropriate constant $C > 0$. The iterative scheme is not infinite: there exists $K \in \mathbb{N}$ such that $u(K)$ exists but not $u(K+1)$. For any $\epsilon > 0$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, , for any $k \in \{1, \dots, K\}$, we have

- $\pi_{-\lambda_{u(k)} r} r < \pi_{-\lambda_{u(k-1)} r} r$ *and* $\pi_{-\lambda_{u(k)} r} r' \le \pi_{-\lambda_{u(k-1)} r} r'$,
- $\pi_{-\lambda_{u(K)} r} r' \le \min\limits_{1 \le j \le \log N} \mathbb{G}_{\mathrm{loc}}(j)$.

Proof. See Section 9.3. □

Remark 3.7. The localized guarantee (3.5) has the same form as the non localized one (see (3.3)). The first term is localized since

$$\pi_{-\lambda r'} r' \le \int_0^1 \pi_{-\gamma\lambda r'} r' d\gamma = -\tfrac{1}{\lambda} \log \pi \exp(-\lambda r').$$

The second term seems to be worse than in the non localized bound since the supremum appears. In fact, this supremum has no effect since when we upper bound this term in order to recover the known convergence rates (either under Vapnik's entropy condition or under (CM) assumptions), the bound increases with the parameter λ. Besides, the discretization of the parameter λ does not influence the convergence rates under these assumptions, and in general will not be harmful.

[13] We do not provide the proof of it since in Section 3.2.2 we give a more difficult-to-prove guarantee in the case of localized complexities.

Detailed algorithm 3.2. This is a possible implementation of Algorithms 3.3 and 3.6. Set M depending on the computer resources available and the required accuracy of approximation.

```
$j' := 0$
Simulate $M$ functions $f_{j',m}$, $m = 1, \dots, M$ under the distribution $\pi_{-\lambda_{j'} r}$
While $j' \leq \log N$ do
        $j := j'$
        Repeat
                $j' := j' + 1$
                If $j' \leq \log N$ Then
                        Simulate $M$ functions $f_{j',m}$, $m = 1, \dots, M$ under $\pi_{-\lambda_{j'} r}$
                        Using $f_{j,m}$ and $f_{j',m}$, estimate $\pi_{-\lambda_{j'} r} r - \pi_{-\lambda_j r} + S(j, j')$
                End If
        until $j' > \log N$ or $\pi_{-\lambda_{j'} r} r - \pi_{-\lambda_j r} + S(j, j') \leq 0$
        End Repeat
End While
```

Classify using $f_{j,1}$ or to follow the lines of boosting methods classify using

$$x \mapsto \mathrm{argmax}_{y \in \mathcal{Y}} \sum_{m \in [1;M]} \mathbb{1}_{f_{j,m}(x) = y}.$$

Remark 3.8. To simulate under the Gibbs distributions $\pi_{-\lambda' r}$ and $\pi_{-\lambda'' r}$, we may use the Metropolis algorithm. To avoid numerical troubles due to the exponential in $\log \pi_{-\lambda r} \exp\big(\frac{\lambda^2}{N} \pi_{-\lambda r} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)$, we can approximate this quantity by

$$\tfrac{\lambda^2}{N} \big(\pi_{-\lambda r} \otimes \pi_{-\lambda r + \frac{\lambda^2}{2N} \pi_{-\lambda r} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}}\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \quad \text{or} \quad \tfrac{\lambda^2}{N} (\pi \otimes \pi)_{-\lambda r(f_1) - \lambda r(f_2) + \frac{\lambda^2}{2N} \bar{\bar{\mathbb{P}}}_{f_1,f_2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$$

since it is lower bounded with $\frac{\lambda^2}{N}\big(\pi_{-\lambda r} \otimes \pi_{-\lambda r}\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$ and upper bounded with

$$\tfrac{\lambda^2}{N} \big(\pi_{-\lambda r} \otimes \pi_{-\lambda r + \frac{\lambda^2}{N} \pi_{-\lambda r} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}}\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \wedge \tfrac{\lambda^2}{N} (\pi \otimes \pi)_{-\lambda r(f_1) - \lambda r(f_2) + \frac{\lambda^2}{N} \bar{\bar{\mathbb{P}}}_{f_1,f_2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}.$$

3.3. **Mixing both complexities.**

This section explains that, by rewriting the algorithms given in Section 3.2 for an appropriate prior distribution, we obtain an algorithm combining the compression scheme approach (Section 3.1) and the usual PAC-Bayesian approach (Section 3.2).

Consider a "family of algorithms":

$$\hat{F} : \cup_{n=0}^{+\infty} \mathcal{Z}^n \times \Theta \times \mathcal{X} \to \mathcal{Y}.$$

For any $\theta \in \Theta$, $\hat{F}_\theta$ is an algorithm to the extent that, with any training set Z_1^N, it associates a prediction function $\hat{F}_{Z_1^N, \theta} : \mathcal{X} \to \mathcal{Y}$. In this sense, the parameter θ "indexes" the algorithms. We assume that these algorithms $\hat{F}_\theta$ are almost exchangeable.

Let $\mathcal{I} \triangleq \underset{2 \leq h \leq 2N}{\cup} \{1, \dots, 2N\}^h$. Any $I \in \mathcal{I}$ can be written as $I = \{i_1, \dots, i_h\}$ with $2 \leq h \leq 2N$. Let $\alpha \in]0;1[$ and π_1 be a prior distribution on the set Θ (possibly depending on Z_1^{2N} in an almost exchangeable way). Consider on the set $\mathcal{I} \times \mathcal{Y}^{2N} \times \Theta$ a distribution such that $\pi_0(I, y_1^{2N}, d\theta) \geq \frac{1-\alpha}{\alpha^2} \big(\frac{\alpha}{2N|\mathcal{Y}|}\big)^h \pi_1(d\theta)$ when $y_i = 0$ for $i > h$. The model is defined as

$$\hat{\mathcal{F}} \triangleq \Big\{ \hat{F}_{z_1^h, \theta} : 2 \leq h \leq 2N, x_i \in \{X_1, \dots, X_{2N}\}, \theta \in \Theta \Big\}.$$

The prior distribution on the model is given by: for any measurable set $A \subset \hat{\mathcal{F}}$, $\pi(A) \triangleq \pi_0\big\{(I, y_1^{2N}, \theta) \in \mathcal{I} \times \mathcal{Y}^{2N} \times \Theta : \hat{F}_{(X_{i_1}, y_1), \dots, (X_{i_h}, y_h), \theta} \in A\big\}$. Since the algorithms $\hat{F}_\theta$ are almost exchangeable, the distribution π is also almost exchangeable so that we can apply Algorithms 3.2 and 3.3 introduced in Section 3.2.

Remark 3.9. When the family of classification rules is just a family of functions (i.e. when the function $\hat{F}_{z_1^n, \theta}$ does not depend on the training set z_1^n), we recover the algorithm described in Section 3.2.

Such a procedure can be useful to choose the similarity measure on the input data, and in particular to choose the kernel (its type and its parameter) of a SVM. It is an alternative to the commonly used cross-validation procedure which has the benefit to be theoretically justified. When $|\Theta|$ is countable, we can also give the following non randomized version of the algorithm.

Algorithm 3.4. *Let us take* $\theta_0 \in argmax_{\theta \in \Theta} \pi_1(\theta)$ *and* $f_0 \triangleq \hat{F}_{X_1, X_2, \theta_0}$. *For any function* $\hat{f} \in \hat{\mathcal{F}}$, *define its complexity as*

$$\mathcal{C}(\hat{f}) \triangleq \min_{(I, y_1^{2N}, \theta) \in \mathcal{I} \times \mathcal{Y}^{2N} \times \Theta : \hat{f} \triangleq \hat{F}_{(X_{i_1}, y_1), \dots, (X_{i_h}, y_h), \theta}} \Big\{ h \log\big(\tfrac{2N|\mathcal{Y}|}{\alpha}\big) + \log \pi_1^{-1}(\theta) \Big\}.$$

For any $k \geq 1$, *define* f_k *as a function with the smallest complexity such that*

$$r(f_k) - r(f_{k-1}) + \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{f_{k-1}, f_k}\{\mathcal{C}(f_{k-1}) + \mathcal{C}(f_k) + \log[(1-\alpha)^{-2}\alpha^4\epsilon^{-1}]\}}{N}} \leq 0.$$

Classify using the function obtained at the last iteration.

From the arguments used in Section 3.1, one can prove a guarantee for this algorithm similar to the last assertion in Theorem 3.1.

Remark 3.10. When $|\Theta| = 1$, we recover the algorithm described in Section 3.1.

3.4. **Similar algorithms in the inductive setting.** In the inductive setting, new difficulties arise and the adaptation of the previous results requires i.i.d. compression schemes similar to the ones developed in [18, 7].

In this section, we only describe an algorithm using a mixed complexities when the set of primary algorithms is countable. When this set is not countable, we will give the algorithm without compression scheme and for a localized complexity (and obtain results of the same nature as the ones in Section 3.2.2).

Remark 3.11. We could have described a general algorithm from which these two algorithms would have been derived up to some variations. We will not give it since notations become quite messy and the practical utility of the resulting classification rule is not obvious since to choose both the algorithm θ and the compression set I is computationally expensive for "huge" set Θ.

3.4.1. *Mixed complexities.* In this section, we consider a family of algorithms:

$$\hat{F} : \cup_{n=0}^{+\infty} \mathcal{Z}^n \times \Theta \times \mathcal{X} \to \mathcal{Y}.$$

Introduce for any $h \in \mathbb{N}^*$, $\mathcal{I}_h \triangleq \{1, \dots, N\}^h$. Any $I \in \mathcal{I}_h$ can be written as $I = \{i_1, \dots, i_h\}$. Define $I^c \triangleq \{1, \dots, N\} - \{i_1, \dots, i_h\}$ and $Z_I \triangleq \big(Z_{i_1}, \dots, Z_{i_h}\big)$. The law of the random variable Z_I will be denoted $\mathbb{P}^I$. For any $J \subset \{1, \dots, N\}$, introduce $\bar{\mathbb{P}}^J \triangleq \frac{1}{|J|} \sum_{i \in J} \delta_{Z_i}$.

Finally, for any I, I_1, I_2 in $\mathcal{I} \triangleq \underset{2 \le h \le N-1}{\cup} \mathcal{I}_h$ and $\theta, \theta_1, \theta_2$ in Θ, introduce

$$\begin{cases} R(I,\theta) & \triangleq & \mathbb{P}[Y \neq \hat{F}_{Z_I,\theta}(X)] \\ r(I,\theta) & \triangleq & \bar{\mathbb{P}}^{I^c}[Y \neq \hat{F}_{Z_I,\theta}(X)] \\ \mathbb{P}(I_1,\theta_1,I_2,\theta_2) & \triangleq & \mathbb{P}[\hat{F}_{Z_{I_1},\theta_1}(X) \neq \hat{F}_{Z_{I_2},\theta_2}(X)] \\ \bar{\mathbb{P}}(I_1,\theta_1,I_2,\theta_2) & \triangleq & \bar{\mathbb{P}}^{(I_1 \cup I_2)^c}[\hat{F}_{Z_{I_1},\theta_1}(X) \neq \hat{F}_{Z_{I_2},\theta_2}(X)] \end{cases}$$

Let $\pi : \cup_{n=0}^{+\infty} \mathcal{Z}^n \to \mathcal{M}_+^1(\Theta)$ associate a prior distribution on the set Θ with any training sample Z_I. For any $\theta \in \Theta$ and any $I \in \mathcal{I}_h$, the complexity of the estimator $\hat{F}_{Z_I,\theta}$ is defined as $\mathcal{C}(I,\theta) \triangleq \log \pi_{Z_I}^{-1}(\theta) + h \log\left(\frac{N}{\alpha}\right)$. To shorten the formulae, introduce $C_{1,2} \triangleq \frac{\mathcal{C}(I_1,\theta_1)+\mathcal{C}(I_2,\theta_2)+\log[(1-\alpha)^{-2}\alpha^4\epsilon^{-1}]}{|(I_1 \cup I_2)^c|}$. For any $(I_1,\theta_1,I_2,\theta_2) \in \mathcal{I}\times\Theta\times\mathcal{I}\times\Theta$, define

$$S(I_1,\theta_1,I_2,\theta_2) \triangleq \sqrt{2C_{1,2}\bar{\mathbb{P}}(I_1,\theta_1,I_2,\theta_2) + C_{1,2}^2} + \frac{4C_{1,2}}{3}.$$

The following algorithm appropriately chooses the primary algorithm $\theta \in \Theta$ and the compression set I.

Algorithm 3.5. *Let $I_0 \in \mathcal{I}_2$ and $\theta_0 \in argmax_{\theta\in\Theta}\pi_{Z_{I_0}}(\theta)$. For any $k \ge 1$, define $I_k \in \underset{2 \le h \le N-1}{\cup} \mathcal{I}_h$ and $\theta_k \in \Theta$ such that*

$$(I_k,\theta_k) \in \underset{(I,\theta): r(I,\theta) - r(I_{k-1},\theta_{k-1}) + S(I,\theta,I_{k-1},\theta_{k-1}) \le 0}{argmin} \mathcal{C}(I,\theta).$$

Classify using the function $\hat{F}_{Z_{I_K},\theta_K}$ where (I_K,θ_K) is the compression set and algorithm obtained at the last iteration.

Define for any $(I,\theta) \in \mathcal{I}\times\Theta$, $k(I,\theta) \triangleq \max\{0 \le k \le K; \mathcal{C}(I_k,\theta_k) \le \mathcal{C}(I,\theta)\}$. The following theorem guarantees the efficiency of this procedure.

Theorem 3.5. *The iterative scheme is not infinite: there exists $K \in \mathbb{N}$ such that (I_K,θ_K) exists but not (I_{K+1},θ_{K+1}). With $(\mathbb{P}^{\otimes N})_*$-probability at least $1-2\epsilon$, for any $k \in \{1,\dots,K\}$, we have*

- *$r(I_k,\theta_k) < r(I_{k-1},\theta_{k-1})$ and $R(I_k,\theta_k) \le R(I_{k-1},\theta_{k-1})$,*
- *$\mathcal{C}(I_k,\theta_k) \ge \mathcal{C}(I_{k-1},\theta_{k-1})$,*
-

$$R(I_K,\theta_K) \le \inf_{(I,\theta)\in\mathcal{I}\times\Theta} \{R(I,\theta) + 2S(I_{k(I,\theta)},\theta_{k(I,\theta)},I,\theta)\}, \tag{3.6}$$

and consequently

$$R(I_K,\theta_K) \le \inf_{\substack{(I,\theta)\in\mathcal{I}\times\Theta \\ \xi \ge 0}} \sup_{\substack{(I',\theta')\in\mathcal{I}\times\Theta: \\ \mathcal{C}(I',\theta') \le \mathcal{C}(I,\theta)}} \Big\{(1+\xi)R(I,\theta) - \xi R(I',\theta') + 2(1+\xi)S(I',\theta',I,\theta)\Big\}. \tag{3.7}$$

Proof. See Section 9.4. □

Remark 3.12. Define $\Theta_{Z_I} \triangleq \mathrm{argmin}_{\theta\in\Theta}\bar{\mathbb{P}}^I[Y \neq \hat{F}_{Z_I,\theta}(X)]$ and let ν be a prior distribution on Θ independent from the data. A natural choice for the prior distributions is to take $\pi_{Z_I}(\theta) \triangleq \frac{\mathbb{1}_{\theta\in\Theta_{Z_I}}}{\nu(\Theta_{Z_I})} \cdot \nu(\theta)$ so that for each compression set I, we consider only the algorithms which minimizes the empirical risk on I. The resulting classifier is based on the ERM principle but does not overfit the data thanks to the compression scheme regularization.

3.4.2. *PAC-Bayesian complexities.* In this section, we consider a model $\mathcal{F}$ which is structured by a prior distribution $\pi \in \mathcal{M}_+^1(\mathcal{F})$ *independent from the data.* Introduce for any $0 \le j \le \log N$ and $\epsilon > 0$,

$$\begin{cases} \lambda_j & \triangleq & 0.19\sqrt{N}e^{\frac{j}{2}} \\ \mathcal{C}(j) & \triangleq & \log \pi_{-\lambda_j r} \exp\left(\frac{\lambda_j^2}{N}\pi_{-\lambda_j r}\bar{\mathbb{P}}_{\cdot,\cdot}\right) \\ g(u) & \triangleq & \frac{\exp(u)-1-u}{u^2} \\ \bar{a}(\lambda) & \triangleq & \frac{\lambda}{N}g(\frac{\lambda}{N})\left(1+\frac{\lambda}{2N}\right) \\ \bar{b}(\lambda) & \triangleq & \frac{1}{\lambda}\left[1+\frac{\lambda}{N}g(\frac{\lambda}{N})\left(1+\frac{\lambda}{2N}\right)^2\right] \\ L & \triangleq & \log[2\log^2(eN)\epsilon^{-1}] \end{cases}$$

and for any $0 \le i < j \le \log N$ and $\epsilon > 0$,

$$S(i,j) \triangleq \bar{a}(\lambda_j)\left(\pi_{-\lambda_i r} \otimes \pi_{-\lambda_j r}\right)\bar{\mathbb{P}}_{\cdot,\cdot} + \bar{b}(\lambda_j)\left[2\mathcal{C}(i) + 2\mathcal{C}(j) + 3L\right].$$

The following localized algorithm gives a way of choosing the standard Gibbs temperature which ensures to get the optimal convergence rate under (CM) assumptions.

Algorithm 3.6. *Let $u(0) = 0$. For any $k \ge 1$, define $u(k)$ as the smallest integer $j \in]u(k-1); \log N]$ such that $\pi_{-\lambda_j r}r - \pi_{-\lambda_{u(k-1)} r}r + S\big(u(k-1), j\big) \le 0$. Classify using a function drawn according to the posterior distribution associated with the last $u(k)$.*

Theorem 3.6. *Let*

(3.8)

$$\mathbb{G}_{\text{loc}}(j) \triangleq \pi_{-\lambda_{j-1} R}R + \frac{\sup_{0\le i\le j}\left\{\log \pi_{-\lambda_i R}\otimes\pi_{-\lambda_i R}\exp\left(\frac{C\lambda_i^2}{N}\mathbb{P}_{\cdot,\cdot}\right)\right\}}{\lambda_j} + C\frac{\log[\log(eN)\epsilon^{-1}]}{\lambda_j}.$$

The iterative scheme is not infinite: there exists $K \in \mathbb{N}$ such that $u(K)$ exists but not $u(K+1)$. With $\left(\mathbb{P}^{\otimes N}\right)_*$-probability at least $1-\epsilon$, for any $k \in \{1,\dots,K\}$, we have

- $\pi_{-\lambda_{u(k)} r}r < \pi_{-\lambda_{u(k-1)} r}r$ *and* $\pi_{-\lambda_{u(k)} r}R \le \pi_{-\lambda_{u(k-1)} r}R$,
- $\pi_{-\lambda_{u(K)} r}R \le \min_{1\le j\le \log N} \mathbb{G}_{\text{loc}}(j)$.

Proof. See Section 9.5. □

An implementation of this procedure is presented in Algorithm 3.2.

Remark 3.13. The algorithms presented in this section are based on the same principle since they all consist in "ranking" the functions in the model by increasing complexity, picking the "first" function in this list and taking at each step the function of smallest complexity such that its generalization error is smaller than the one at the previous step. Note that this section has indirectly emphasized the benefit of *relative* data-dependent bounds.

4. Comparison between the errors of any two randomized estimators

We start with the transductive setting which provides simpler formulae and in which the variance term is directly observable. Results for the inductive setting are collected in Section 4.4.

4.1. **Basic result.** Let π_1 and π_2: $\mathcal{Z}^{2N} \to \mathcal{M}_+^1(\mathcal{F})$ denote two almost exchangeable functions. Let us introduce $\mathcal{K}_{1,2} \triangleq K(\rho_1, \pi_1) + K(\rho_2, \pi_2) + \log(\epsilon^{-1})$.

Theorem 4.1. *For any $\epsilon > 0$, $\lambda > 0$, with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1 - \epsilon$, for any distributions $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$, we have*

$$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\mathcal{K}_{1,2}}{\lambda}. \tag{4.1}$$

Proof. Using Theorem 8.4 for $\mathcal{G} = \mathcal{F} \times \mathcal{F}$, $\mathcal{W}\big[(f_1, f_2), Z\big] = \mathbb{1}_{Y \neq f_2(X)} - \mathbb{1}_{Y \neq f_1(X)}$ and $(\mu, \nu) = (\rho_1 \otimes \rho_2, \pi_1 \otimes \pi_2)$, we obtain Inequality (4.1). □

The bound consists in a variance term $\frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$ and a complexity term $\frac{\mathcal{K}_{1,2}}{\lambda}$. The variance term will be small when the distributions ρ_1 and ρ_2 are concentrated around the same function. The complexity of a randomized estimator is measured by the Kullback-Leibler divergence of its posterior distribution wrt the prior distribution.

Since the variance term $\frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} = \frac{2\lambda}{N}\mathbb{E}_{\rho_1(df_1)}\mathbb{E}_{\rho_2(f_2)}\bar{\bar{\mathbb{P}}}_{f_1,f_2}$ is to be large when the distributions ρ_1 and ρ_2 are close and not concentrated, we might want to improve this term by coupling. This is done in Appendix A.

Remark 4.1. Since the labels $Y_{N+1}, \dots, Y_{2N}$ are unknown, the prior distributions will only be observable when they do not depend on the labels.

4.2. **Optimizing the result wrt the parameter λ.** First let us show how to optimize the free parameter in Theorem 4.1. Let $\Lambda \subset \mathbb{R}_+^*$ be a finite set and $\mathcal{K}'_{1,2} \triangleq K(\rho_1, \pi_1) + K(\rho_2, \pi_2) + \log(|\Lambda|\epsilon^{-1})$.

Theorem 4.2. *For any $\epsilon > 0$, with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1 - \epsilon$, we have for any $\lambda \in \Lambda$, $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$*

$$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \min_{\lambda \in \Lambda}\left\{\frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\mathcal{K}'_{1,2}}{\lambda}\right\}. \tag{4.2}$$

Proof. The result just comes from a union bound and Theorem 4.1. □

Remark 4.2. Let us take $\rho_1 = \pi_1 = \delta_{\tilde{f}}$. To shorten notations, introduce $\rho = \rho_2$ and $\pi = \pi_2$. We get $\rho r' - r'(\tilde{f}) \leq \rho r - r(\tilde{f}) + \min_{\lambda \in \Lambda}\left\{\frac{2\lambda}{N}\rho\bar{\bar{\mathbb{P}}}_{\cdot,\tilde{f}} + \frac{\mathcal{K}}{\lambda}\right\}$,where

$$\mathcal{K} \triangleq K(\rho, \pi) + \log(|\Lambda|\epsilon^{-1}).$$

Then the previous results compare the generalization errors of a ρ-randomized estimator and a reference function $\tilde{f}$. To understand well the bounds of this paper, it is important to keep in mind that we are interested in bounds having the order of $1/N^\beta$ where $\beta \in]0; 1]$. The power β of the bound appears to be closely linked to both the complexity of the model and the order of $\mathbb{P}_{f,\tilde{f}}$ when f gets close to the reference classifier. This idea already appears in [14, 19, 4] which assume $\left(\mathbb{P}_{f,\tilde{f}}\right)^\kappa \leq R(f) - R(\tilde{f})$ for some $\kappa \geq 1$ and then deduce the convergence rate of the ERM-algorithm. In this paper, we obtain empirical bounds in which the same kind of trade-off $\big($here between $\bar{\bar{\mathbb{P}}}_{f,\tilde{f}}$ and $r(f) - r(\tilde{f})\big)$ takes place. When the posterior distribution ρ is fixed, the optimal parameter λ has the order of $\sqrt{N\mathcal{K}/(\rho\bar{\bar{\mathbb{P}}}_{\cdot,\tilde{f}})}$ and for this parameter, $\frac{2\lambda}{N}\rho\bar{\bar{\mathbb{P}}}_{\cdot,\tilde{f}} + \frac{\mathcal{K}}{\lambda}$ has the order of $\sqrt{\mathcal{K}\rho\bar{\bar{\mathbb{P}}}_{\cdot,\tilde{f}}/N}$.

Remark 4.3. There is a simple way to recover non relative results in which the deviations of the functions f and $\tilde{f}$ were added (as explained in Section 2). It consists in upper bounding $\mathbb{1}_{f(X)\neq\tilde{f}(X)}$ by $\mathbb{1}_{Y\neq f(X)}+\mathbb{1}_{Y\neq\tilde{f}(X)}$. This inequality implies that $2\rho\bar{\bar{\mathbb{P}}}_{\cdot,\tilde{f}}\leq\rho r+\rho r'+r(\tilde{f})+r'(\tilde{f})$. Replacing $\rho\bar{\bar{\mathbb{P}}}_{\cdot,\tilde{f}}$ by its upper bound, we find inequalities to which non relative PAC-Bayesian bounds lead to.

Another way of recovering non relative PAC-Bayesian bounds is to use the results of Section 8 (such as Theorem 8.4) with $\mathcal{W}(f,Z)=\mathbb{1}_{Y\neq f(X)}$ instead of $\mathcal{W}\big[(f_1,f_2),Z\big]=\mathbb{1}_{Y\neq f_2(X)}-\mathbb{1}_{Y\neq f_1(X)}$.

In non relative bounds, the optimal randomizing distributions (i.e. the ones minimizing the bounds) are standard Gibbs distributions. In relative bounds, $\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$-terms appear but, finally, the form of the optimal distribution is not very different: the relative approach just really improves the bounds in noisy situations and leads to a less conservative choice of the temperature (i.e. to larger λ).

The optimal parameter λ in Inequality (4.1) is $\sqrt{\frac{N\mathcal{K}_{1,2}}{2(\rho_1\otimes\rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}}}\geq\sqrt{\frac{N\log(\epsilon^{-1})}{2}}$. Besides, for $\lambda\geq N$, the bound is greater than $2(\rho_1\otimes\rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$ which is a trivial upper bound on $\rho_2 r'-\rho_1 r'+\rho_1 r-\rho_2 r$.

So values of the parameter smaller than $\sqrt{N}$ or greater than N can be disregarded. Then a good set of parameters is

$$\Lambda\triangleq\left\{\sqrt{N}\zeta^k;0\leq k\leq\frac{\log N}{2\log\zeta}\right\}\tag{4.3}$$

where $\zeta>1$. Using this family, we obtain the following continuously uniform bound wrt λ:

Theorem 4.3. *Let $\epsilon>0$ and $\mathcal{K}''_{1,2}\triangleq K(\rho_1,\pi_1)+K(\rho_2,\pi_2)+\log\left[\frac{\log(\zeta^2N)}{2\log\zeta}\epsilon^{-1}\right]$. With $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-\epsilon$, for any $\rho_1,\rho_2\in\mathcal{M}^1_+(\mathcal{F})$, we have*

$$\rho_2 r'-\rho_1 r'+\rho_1 r-\rho_2 r\leq\min_{\lambda\in[\sqrt{N};N]}\left\{\frac{2\lambda}{N}(\rho_1\otimes\rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}+\zeta\frac{\mathcal{K}''_{1,2}}{\lambda}\right\}.\tag{4.4}$$

To conclude, it does not cost much (just a $\log\log N$ factor[14]) to gain uniformity in the parameter λ. We have shown how to get this uniformity. The same tools can be used to write uniform versions in real parameters of results claimed in this paper.

4.3. **Localization.**

4.3.1. *Localizing both KL-divergences.* In Theorem 4.1, the global size of the model appears in the Kullback-Leibler divergence. The complexity term $K(\rho,\pi)$ can be large and will be all the more substantial as we had in the model irrelevant functions for our classification task. This is clearly a drawback that we want to correct. By replacing the prior distribution π by a suitable almost exchangeable Gibbs distribution $\big(\pi_{-C[r+r']}\big)$ and by managing smartly the inequalities in order to recover an observable upper bound, we can correct it. We will use the following lemma.

[14] Note that $\log\log N\leq 4$ for $N\leq 10^{23}$!

Lemma 4.4. *For any $\epsilon > 0$, $\lambda > 0$ and $\xi \in]0;1[$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, for any $\rho \in \mathcal{M}^1_+(\mathcal{F})$, we have*

(4.5)
$$K\big(\rho, \pi_{-\frac{\lambda}{2}[r+r']}\big) \quad \le \quad \tfrac{1}{1-\xi}\Big[K\big(\rho, \pi_{-\lambda r}\big) + \log \pi_{-\lambda r} \exp\big(\tfrac{\lambda^2}{2\xi N}\rho \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) + \xi \log(\epsilon^{-1})\Big].$$

Proof. See Section 9.6. □

Combining Theorem 4.1 for prior distributions $(\pi_1)_{-\frac{1}{2}\lambda_1[r+r']}$ and $(\pi_2)_{-\frac{1}{2}\lambda_2[r+r']}$ (where π_1 and π_2 do not depend on the labels to be observable), and Lemma 4.4, we obtain the following localized inequality.

Theorem 4.5. *For any $\epsilon > 0$, $\xi \in]0;1[$ and $\lambda, \lambda_1, \lambda_2 > 0$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-3\epsilon$, for any distributions $\rho_1, \rho_2 \in \mathcal{M}^1_+(\mathcal{F})$, we have*

$$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \le \frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\mathcal{K}^{\mathrm{loc}}_{1,2}}{(1-\xi)\lambda}, \tag{4.6}$$

where

$$\begin{aligned}\mathcal{K}^{\mathrm{loc}}_{1,2} \triangleq K\big(\rho_1, (\pi_1)_{-\lambda_1 r}\big) + K\big(\rho_2, (\pi_2)_{-\lambda_2 r}\big) + \log(\pi_1)_{-\lambda_1 r} \exp\big(\tfrac{\lambda_1^2}{2\xi N}\rho_1 \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) \\ + \log(\pi_2)_{-\lambda_2 r} \exp\big(\tfrac{\lambda_2^2}{2\xi N}\rho_2 \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) + (1+\xi)\log(\epsilon^{-1}).\end{aligned}$$

For $\lambda_1 = \lambda_2 = \xi \to 0$, we recover the non localized inequality. As a special case of Theorem 4.5, for an almost exchangeable prior π, we have

Corollary 4.6. *For any $\epsilon > 0$ and any finite set $\Lambda \subset \mathbb{R}^*_+$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-3\epsilon$, for any $(\lambda, \lambda', \lambda'') \in \Lambda^3$, we have*

$$\pi_{-\lambda'' r} r' - \pi_{-\lambda' r} r' + \pi_{-\lambda' r} r - \pi_{-\lambda'' r} r \le \frac{2\lambda}{N}(\pi_{-\lambda' r} \otimes \pi_{-\lambda'' r})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\bar{\bar{\mathcal{K}}}}{\lambda}, \tag{4.7}$$

where

$$\begin{aligned}\bar{\bar{\mathcal{K}}} \triangleq 2\log \pi_{-\lambda' r} \exp\Big(\tfrac{\lambda'^2}{N}\pi_{-\lambda' r}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + 2\log \pi_{-\lambda'' r} \exp\Big(\tfrac{\lambda''^2}{N}\pi_{-\lambda'' r}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) \\ + 3\log\big(|\Lambda|^3 \epsilon^{-1}\big).\end{aligned}$$

Proof. Use the previous theorem with $\xi = \frac{1}{2}$, $(\rho_1, \pi_1, \rho_2, \pi_2) = (\pi_{-\lambda' r}, \pi, \pi_{-\lambda'' r}, \pi)$, $\lambda_1 = \lambda'$, $\lambda_2 = \lambda''$, and make a union bound on the parameters λ, λ' and λ''. □

To conclude this section, localization leads to smaller complexity terms and smaller influence of the choice of the prior distribution. Corollary 4.6 also shows that the complexity term can be seen as a variance term since the quantities $\log \pi_{-\lambda r} \exp\big(\frac{\lambda^2}{N}\pi_{-\lambda r}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)$ are roughly approximated with $\frac{\lambda^2}{N}(\pi_{-\lambda r} \otimes \pi_{-\lambda r})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$ (at least for small enough λ).

4.3.2. *Localizing one KL-divergence.* When we want to localize just one of the two KL-divergences, we can obtain a simpler result (without terms of the form $\log \pi_{-\lambda r} \exp\big\{C\frac{\lambda^2}{N}\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big\}$) by using a more direct proof:

Theorem 4.7. *Let $\breve{\rho}$ be an almost exchangeable prior distribution (for instance $\pi_{-C(r+r')}$ or $\delta_{\tilde{f}}$). For any $\epsilon > 0$, $\lambda > 0$ and $\xi \ge 0$, we have*

- *when $\xi < 1$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-2\epsilon$, for any randomizing distribution $\rho \in \mathcal{M}^1_+(\mathcal{F})$,*

$$\rho r' - \breve{\rho} r' \quad \le \quad \rho r - \breve{\rho} r + \tfrac{1+\xi}{1-\xi}\tfrac{2\lambda}{N}\big(\rho \otimes \breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{K(\rho, \pi_{-2\xi\lambda r}) + (1+\xi)\log(\epsilon^{-1})}{(1-\xi)\lambda}. \tag{4.8}$$

- *with* $\big(\mathbb{P}^{\otimes 2N}\big)_*$*-probability at least* $1-2\epsilon$*, for any randomizing distribution* $\rho \in \mathcal{M}^1_+(\mathcal{F})$*,*

$$\breve{\rho}r' - \rho r' \quad \le \quad \breve{\rho}r - \rho r + \tfrac{2\lambda}{N}\big(\rho\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{K(\rho,\pi_{-2\xi\lambda r})+(1+\xi)\log(\epsilon^{-1})}{(1+\xi)\lambda}.$$

Proof. See Section 9.7. □

For $\xi = 0$, we recover the non localized bound. We can also give uniform results in both parameters λ and ξ as the following remark shows.

Remark 4.4. Let $\Lambda \subset [\sqrt{N}; N]$ and $\Xi \subset [0;1[$. The previous bound holds uniformly in $\lambda \in \Lambda$ and $\xi \in \Xi$ by replacing the term $\log(\epsilon^{-1})$ by $\log(|\Lambda||\Xi|\epsilon^{-1})$.

Again, good sets of parameters have the following form $\Lambda \triangleq \big\{\sqrt{N}\zeta^k; 0 \le k \le \frac{\log N}{2\log\zeta}\big\}$ and $\Xi \triangleq \big\{\alpha^{-k}; 1 \le k \le \frac{\log(\alpha N)}{\log\alpha}\big\}$ where $\alpha > 1$, $\zeta > 1$. Using these sets, we can obtain continuously uniform version of the previous results. The union bound just introduces $\log\log N$ terms since $|\Lambda||\Xi| \le \frac{\log(\zeta^2 N)\log(\alpha N)}{2\log(\zeta)\log(\alpha)}$.

4.4. **In the inductive setting.** We can adapt all the methods developed in the transductive setting to the inductive setting when the prior distribution is *independent* from the data. The only extra difficulty comes from the variance term (since we have to transform $\mathbb{P}_{\cdot,\cdot}$ into $\bar{\mathbb{P}}_{\cdot,\cdot}$ when we want an observable bound and $\bar{\mathbb{P}}_{\cdot,\cdot}$ into $\mathbb{P}_{\cdot,\cdot}$ when we want theoretical bounds) but this problem is solved by using Theorem 8.1 with $\mathcal{W}(f_1,f_2,Z) = -\mathbb{1}_{f_1(X)\neq f_2(X)}$ and $\mathcal{W}(f_1,f_2,Z) = \mathbb{1}_{f_1(X)\neq f_2(X)}$.

Theorem 4.8. *For any* $\lambda > 0$, $\pi_1,\pi_2 \in \mathcal{M}^1_+(\mathcal{F})$, $\epsilon > 0$*, we have*

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any* $\rho_1,\rho_2 \in \mathcal{M}^1_+(\mathcal{F})$*,*

$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \le \tfrac{\lambda}{N} g\big(\tfrac{\lambda}{N}\big)(\rho_1\otimes\rho_2)\mathbb{P}_{\cdot,\cdot} + \tfrac{\mathcal{K}_{1,2}}{\lambda}$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any* $\rho_1,\rho_2 \in \mathcal{M}^1_+(\mathcal{F})$*,*

$$(\rho_1\otimes\rho_2)\mathbb{P}_{\cdot,\cdot} \le \big(1+\tfrac{\lambda}{2N}\big)(\rho_1\otimes\rho_2)\bar{\mathbb{P}}_{\cdot,\cdot} + \tfrac{(1+\frac{\lambda}{2N})^2\mathcal{K}_{1,2}}{\lambda} \ ,$$

where $\mathcal{K}_{1,2} \triangleq K(\rho_1,\pi_1) + K(\rho_2,\pi_2) + \log(\epsilon^{-1})$ *and* $g(u) \triangleq \frac{\exp(u)-1-u}{u^2}$.

Proof. Apply Theorem 8.1 for $\mathcal{G} = \mathcal{F}\times\mathcal{F}$, $\mu = \rho_1\otimes\rho_2$, $\nu = \pi_1\otimes\pi_2$ and successively for $\mathcal{W}(f_1,f_2,Z) = \mathbb{1}_{Y\neq f_1(X)} - \mathbb{1}_{Y\neq f_2(X)}$ and $\mathcal{W}(f_1,f_2,Z) = -\mathbb{1}_{f_1(X)\neq f_2(X)}$. For the second inequality, we change the parameter $\lambda \leftarrow \frac{\lambda}{1-\frac{\lambda}{2N}}$ to obtain the desired formulation. □

As a consequence, we have:

Corollary 4.9. *For any* $\lambda > 0$, $\pi_1,\pi_2 \in \mathcal{M}^1_+(\mathcal{F})$, $\epsilon > 0$*, with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-2\epsilon$*, for any* $\rho_1,\rho_2 \in \mathcal{M}^1_+(\mathcal{F})$*, we have*

$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \le \bar{a}(\lambda)(\rho_1\otimes\rho_2)\bar{\mathbb{P}}_{\cdot,\cdot} + \bar{b}(\lambda)\mathcal{K}_{1,2} \tag{4.9}$$

where $\bar{a}(\lambda) \triangleq \frac{\lambda}{N}g(\frac{\lambda}{N})\big(1+\frac{\lambda}{2N}\big)$ *and* $\bar{b}(\lambda) \triangleq \frac{1}{\lambda}\Big[1+\frac{\lambda}{N}g(\frac{\lambda}{N})\big(1+\frac{\lambda}{2N}\big)^2\Big]$.

Remark 4.5. To recover a simple formulation, it suffices to note that $\bar{a}(\lambda) \le 1.1\frac{\lambda}{N}$ and $\bar{b}(\lambda) \le \frac{2.7}{\lambda}$ for any $0 < \lambda \le N$.

To localize the KL-terms, we can prove the following result which is similar to Lemma 4.4 and which is used to justify Algorithm 3.6.

Lemma 4.10. *For any $\epsilon > 0$, $\xi \in]0;1[$ and $0 < \lambda \leq 0.39\,\xi N$, with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-2\epsilon$, for any $\rho \in \mathcal{M}^1_+(\mathcal{F})$, we have*

(4.10)
$$K\big(\rho, \pi_{-\lambda R}\big) \quad \leq \quad \tfrac{1}{1-\xi}\Big[K\big(\rho, \pi_{-\lambda r}\big) + \log \pi_{-\lambda r} \exp\big(\tfrac{2\lambda^2}{\xi N}\rho\bar{\mathbb{P}}_{\cdot,\cdot}\big) + \xi \log(\epsilon^{-1})\Big].$$

Proof. See Section 9.8. □

5. Compression schemes

5.1. **In the transductive setting.** The compression schemes were introduced by Littlestone and Warmuth ([12]). The results presented here are directly inspired from [7, Chapter 3.1]. The notations are the same as the ones used in Section 3.1. We have an exchangeable algorithm

$$\hat{f} : \underset{n\in\mathbb{N}^*}{\cup} \mathcal{Z}^n \times \mathcal{X} \to \mathcal{Y}$$

which produces for any training set $\mathcal{L}$ the prediction function $\hat{f}_{\mathcal{L}} : \mathcal{X} \to \mathcal{Y}$. Let $\hat{\mathcal{F}}_h \triangleq \big\{\hat{f}_{(X_{i_j},y_i)_{j=1}^h} : (i_1, \dots, i_h) \in \{1, \dots, 2N\}^h, y_1^h \in \mathcal{Y}^h\big\}$. We consider the data-dependent model $\hat{\mathcal{F}} \triangleq \underset{2\leq h\leq N}{\cup} \hat{\mathcal{F}}_h$.

Theorem 5.1. *Let $\epsilon > 0$, $\alpha \in]0;1[$ and $L \triangleq \log[(1-\alpha)^{-2}\alpha^4\epsilon^{-1}]$. With $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, for any $f_1, f_2 \in \hat{\mathcal{F}}$, we have*

$$r'(f_2) - r'(f_1) \leq r(f_2) - r(f_1) + \sqrt{\tfrac{8\bar{\bar{\mathbb{P}}}_{f_1,f_2}[h_1 \log(2N|\mathcal{Y}|/\alpha) + h_2 \log(2N|\mathcal{Y}|/\alpha) + L]}{N}},$$

where the integers h_1 and h_2 satisfy $f_1 \in \hat{\mathcal{F}}_{h_1}$ and $f_2 \in \hat{\mathcal{F}}_{h_2}$.

Proof. Let π be a prior distribution such that it is uniform on each $\hat{\mathcal{F}}_h$ and $\pi(\hat{\mathcal{F}}_h) \geq (1-\alpha)\alpha^{h-2}$. We have $\log|\hat{\mathcal{F}}_h| = \log\big[(2N)^h|\mathcal{Y}|^h\big] = h\log\big(2N|\mathcal{Y}|\big)$. The result comes from Inequality (8.7) in which we take $\mathcal{W}[(f_1,f_2),Z] = \mathbb{1}_{Y\neq f_2(X)} - \mathbb{1}_{Y\neq f_1(X)}$ and $\nu = \pi \otimes \pi$. □

Remark 5.1. This compression scheme can be extended to a family of algorithms $\hat{F} : \cup_{n=0}^{+\infty}\mathcal{Z}^n \times \Theta \times \mathcal{X} \to \mathcal{Y}$. In the inductive setting, we will directly give the result for this family.

5.2. **In the inductive setting.** Compression schemes in the inductive learning are *not* a direct consequence of the one in the transductive learning. Here we adapt the ideas developed in [7, Chapter 4]. The notations are the one introduced in Section 3.4.1. Let $\ddot{\pi} : \mathcal{Z}^N \to \mathcal{M}^1_+(\mathcal{I}\times\Theta\times\mathcal{I}\times\Theta)$ be some regular conditional probability measure such that

- $\ddot{\pi}_{Z_1^N}(I_1, I_2)$ is independent from Z_1^N,
- $\ddot{\pi}_{Z_1^N}(d\theta_1, d\theta_2|I_1, I_2)$ depends only on Z_{I_1} and Z_{I_2} $\big($and so will be denoted $\ddot{\pi}_{Z_{I_1},Z_{I_2}}(d\theta_1, d\theta_2)\big)$.

Theorem 5.2. *We still use $g(u) \triangleq \frac{\exp(u)-1-u}{u^2}$. Introduce $N_{1,2} \triangleq |(I_1 \cup I_2)^c|$ and $\mathcal{K}_{1,2} \triangleq K(\rho_1 \otimes \rho_2, \ddot{\pi}) + \log(\epsilon^{-1})$. For any $\epsilon > 0$, $\lambda > 0$, we have*

- *with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-\epsilon$, for any $\rho_1, \rho_2 \in \mathcal{M}^1_+(\mathcal{I}\times\Theta)$,*

$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \leq (\rho_1 \otimes \rho_2)\Big[\tfrac{\lambda}{N_{1,2}} g\big(\tfrac{\lambda}{N_{1,2}}\big)\mathbb{P}(I_1,\theta_1,I_2,\theta_2)\Big] + \tfrac{\mathcal{K}_{1,2}}{\lambda},$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any* $\rho_1, \rho_2 \in \mathcal{M}^1_+(\mathcal{I}\times\Theta)$*,*

$$\big(\rho_1\otimes\rho_2\big)\Big[\big(1-\tfrac{\lambda}{2N_{1,2}}\big)\mathbb{P}(I_1,\theta_1,I_2,\theta_2)\Big] \le \big(\rho_1\otimes\rho_2\big)\bar{\mathbb{P}}(I_1,\theta_1,I_2,\theta_2) + \tfrac{\mathcal{K}_{1,2}}{\lambda}.$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any* $I_1, I_2 \in \mathcal{I}$ *and* $\theta_1, \theta_2 \in \Theta$*,*

$$\begin{array}{ll} & R(I_2,\theta_2) - R(I_1,\theta_1) + r(I_1,\theta_1) - r(I_2,\theta_2) \\ \le & \sqrt{\frac{2[\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)+\log(\epsilon^{-1})]\mathbb{P}(I_1,\theta_1,I_2,\theta_2)}{N_{1,2}}} + \frac{\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)+\log(\epsilon^{-1})}{3N_{1,2}}, \end{array}$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any* $I_1, I_2 \in \mathcal{I}$ *and* $\theta_1, \theta_2 \in \Theta$*,*

$$\mathbb{P}(I_1,\theta_1,I_2,\theta_2) \le \Bigg(\sqrt{\bar{\mathbb{P}}(I_1,\theta_1,I_2,\theta_2) + \frac{\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)+\log(\epsilon^{-1})}{2N_{1,2}}} \\ + \sqrt{\frac{\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)+\log(\epsilon^{-1})}{2N_{1,2}}}\Bigg)^2.$$

Proof. Apply Theorem 8.7 successively with

$$\hat{G} : \big(Z_1^n, Z_1^{n'}, (\theta,\theta'), (x,y)\big) \mapsto \mathbb{1}_{y\neq\hat{F}_{Z_1^n,\theta}(x)} - \mathbb{1}_{y\neq\hat{F}_{Z_1^{n'},\theta'}(x)}$$

and $\hat{G} : \big(Z_1^n, Z_1^{n'}, (\theta,\theta'), (x,y)\big) \mapsto -\mathbb{1}_{\hat{F}_{Z_1^n,\theta}(x)\neq\hat{F}_{Z_1^{n'},\theta'}(x)}$. Then take

$$\left\{\begin{array}{lll} \mu\big(I_1,I_2,d(\theta_1,\theta_2)\big) & = & \rho_1(I_1,d\theta_1)\otimes\rho_2(I_2,d\theta_2) \\ \nu\big(I_1,I_2,d(\theta_1,\theta_2)\big) & = & \ddot{\pi}(I_1,d\theta_1,I_2,d\theta_2) \end{array}\right. .$$

□

6. Some properties of Gibbs estimators

6.1. **Concentration of Gibbs estimators.**

So far, we have looked for controlling the risk $\hat{\rho}r'$ and $\hat{\rho}R$ in respectively the transductive and inductive setting. One can ask whether the randomizing distribution $\hat{\rho}$ is enough concentrated so that, by drawing a function f according to this distribution $\hat{\rho}$, the resulting risk $r'(f)$ or $R(f)$ has the same order as $\hat{\rho}r'$ or $\hat{\rho}R$. In the transductive learning, the following theorem tends to say that this property holds to the extent that it holds for the risk $r+r'$.

Theorem 6.1. *Let* π *and* $\breve{\rho}$ *be almost exchangeable distributions. For any* $\epsilon>0$*,* $\lambda>0$*, with* $\mathbb{P}^{\otimes 2N}$*-probability at least* $1-\epsilon$ *and* $\pi_{-2\lambda r}$*-probability at least* $1-\epsilon$*, we have*

$$(r+r') - \breve{\rho}(r+r') \le \frac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{-\log\pi\exp\big\{-2\lambda[r-\breve{\rho}r]\big\} + 2\log(\epsilon^{-1})}{\lambda}. \tag{6.1}$$

Proof. See Section 9.9. □

Inequality (6.1) is to be compared with

$$\pi_{-2\lambda r}(r+r') - \breve{\rho}(r+r') \le \tfrac{2\lambda}{N}\big(\pi_{-2\lambda r}\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{-\log\pi\exp\{-2\lambda[r-\breve{\rho}r]\}+\log(\epsilon^{-1})}{\lambda},$$

which directly comes from Theorem 4.1 $\big($with $(\rho_2,\pi_2,\rho_1,\pi_1) = (\pi_{-2\lambda r},\pi,\breve{\rho},\breve{\rho})\big)$. Theorem 6.1 implies that Inequality (6.1) holds with probability at least $1-2\epsilon$ wrt randomness.

Remark 6.1. For sake of simplicity, the result has been given for the distribution $\pi_{-2\lambda r}$. We can adapt the proof to take into account other Gibbs distributions in which the variance term $\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$ appears.

In the inductive setting, when the prior distribution π is *independent* from the data, the previous theorem becomes

Theorem 6.2. *For any $\epsilon > 0$, $\lambda > 0$, $\pi \in \mathcal{M}_+^1(\mathcal{F})$ and $\tilde{\rho} \in \mathcal{M}_+^1(\mathcal{F})$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$ and $\pi_{-\lambda r}$-probability at least $1-\epsilon$, we have*

$$R - \tilde{\rho}R \leq \tfrac{\lambda}{N} g\big(\tfrac{\lambda}{N}\big)\tilde{\rho}\mathbb{P}_{\cdot,\cdot} + \tfrac{-\log \pi \exp\{-\lambda[r-\tilde{\rho}r]\}+2\log(\epsilon^{-1})}{\lambda}. \tag{6.2}$$

Proof. See Section 9.10. □

This result has to be compared with

$$\pi_{-\lambda r}R - \tilde{\rho}R \leq \tfrac{\lambda}{N} g\big(\tfrac{\lambda}{N}\big)\big(\pi_{-\lambda r} \otimes \tilde{\rho}\big)\mathbb{P}_{\cdot,\cdot} + \tfrac{-\log \pi \exp\{-\lambda[r-\tilde{\rho}r]\}+\log(\epsilon^{-1})}{\lambda},$$

which comes from Theorem 4.8.

6.2. **Bracketing on the efficiency of standard Gibbs estimators.** The following theorem brackets the efficiency of a standard Gibbs estimator in the transductive setting.

Theorem 6.3. *For any $\lambda > 0$,*

- *for any $0 \leq \xi < 1$, we have*

$$\begin{aligned} \pi_{-\lambda r}r' &\leq -\tfrac{\log \pi_{-\xi\lambda r'} \exp\{-(1-\xi)\lambda r'\}}{(1-\xi)\lambda} + \tfrac{K(\pi_{-\lambda r},\pi_{-\lambda r'})}{(1-\xi)\lambda} \\ &\leq \pi_{-\xi\lambda r'}r' + \tfrac{K(\pi_{-\lambda r},\pi_{-\lambda r'})}{(1-\xi)\lambda} \end{aligned} \tag{6.3}$$

- *for any $\chi > 0$, we have*

$$\begin{aligned} \pi_{-\lambda r}r' &\geq -\tfrac{\log \pi_{-\lambda r'} \exp(-\chi\lambda r')}{\chi\lambda} - \tfrac{K(\pi_{-\lambda r},\pi_{-\lambda r'})}{\chi\lambda} \\ &\geq \pi_{-(1+\chi)\lambda r'}r' - \tfrac{K(\pi_{-\lambda r},\pi_{-\lambda r'})}{\chi\lambda} \end{aligned} \tag{6.4}$$

These inequalities are completed by the following one: for any $\epsilon > 0$ and $0 < \gamma < 1$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have

$$\begin{aligned} K(\pi_{-\lambda r},\pi_{-\lambda r'}) \leq \tfrac{1}{1-\gamma} \log \pi_{-\lambda r'} \otimes \pi_{-\lambda r'} \exp\big(\tfrac{10\lambda^2}{\gamma N}\bar{\mathbb{P}}'_{\cdot,\cdot}\big) \\ +\big(35 + \tfrac{375\lambda^2}{\gamma^2 N^2}\big)\tfrac{\gamma}{1-\gamma}\log(8\epsilon^{-1}). \end{aligned} \tag{6.5}$$

Proof. The first two results come from the Legendre transform of the function $\rho \mapsto \mathcal{K}(\rho, \pi_{-\lambda r'})$ and Jensen's inequality. The last one is proved in Section 9.11. □

Remark 6.2. The constants are not very satisfactory since too many concentration inequalities are piled in the proof. With this respect, the intermediate step

$$K(\pi_{-\lambda r},\pi_{-\lambda r'}) \leq \tfrac{5}{1-\gamma} \log \pi_{-\lambda\frac{r+r'}{2}} \exp\Big(\tfrac{\lambda^2}{\gamma N}\pi_{-\lambda\frac{r+r'}{2}}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + \tfrac{20\gamma}{1-\gamma}\log(4\epsilon^{-1}).$$

was tighter. The parameter γ is here to balance the two terms of the RHS. For instance, for small enough λ $\big($at least for $\lambda = \mathrm{o}(\sqrt{N})\big)$, the optimal γ is $\mathrm{o}(1)$.

In the inductive setting, we have

Theorem 6.4. *For any $\lambda > 0$,*

- *for any $0 \leq \xi < 1$, we have*

$$\begin{aligned} \pi_{-\lambda r}R &\leq -\tfrac{\log \pi_{-\xi\lambda R} \exp\{-(1-\xi)\lambda R\}}{(1-\xi)\lambda} + \tfrac{K(\pi_{-\lambda r},\pi_{-\lambda R})}{(1-\xi)\lambda} \\ &\leq \pi_{-\xi\lambda R}R + \tfrac{K(\pi_{-\lambda r},\pi_{-\lambda R})}{(1-\xi)\lambda} \end{aligned} \tag{6.6}$$

- *for any $\chi > 0$, we have*

$$(6.7)\qquad \begin{array}{rcl} \pi_{-\lambda r}R & \geq & -\frac{\log \pi_{-\lambda R} \exp(-\chi\lambda R)}{\chi\lambda} - \frac{K(\pi_{-\lambda r},\pi_{-\lambda R})}{\chi\lambda} \\ & \geq & \pi_{-(1+\chi)\lambda R}R - \frac{K(\pi_{-\lambda r},\pi_{-\lambda R})}{\chi\lambda} \end{array}$$

These inequalities are completed by the following one: for any $\epsilon > 0$, $0 < \gamma < 1$ and $0 < \lambda \leq 0.39\,\gamma N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$(6.8)\quad K(\pi_{-\lambda r},\pi_{-\lambda R}) \quad\leq\quad \tfrac{4}{1-\gamma}\log \pi_{-\lambda R}\exp\left(\tfrac{4.1\,\lambda^2}{\gamma N}\pi_{-\lambda R}\mathbb{P}_{\cdot,\cdot}\right) + \tfrac{5\gamma}{1-\gamma}\log(4\epsilon^{-1}).$$

Proof. The first two results come from the Legendre transform of the function $\rho \mapsto \mathcal{K}(\rho,\pi_{-\lambda R})$ and Jensen's inequality. The last one is proved in Section 9.12. □

7. Vapnik's type bounds

To illustrate the relative data-dependent bounds developed in this paper, we can use them to recover and improve classical bounds of Vapnik and Chervonenkis theory. In particular, we will prove VC-bounds involving the pseudo-distance $\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$ and localize them. We start with the transductive inference in which results are much simpler. In Section 7.4, similar bounds are given for the inductive learning.

Let $\mathbb{X} \triangleq X_1^{2N}$ and $\mathcal{A}(\mathbb{X})$ be the partition of the model $\mathcal{F}$ defined by

$$\mathcal{A}(\mathbb{X}) \triangleq \Big\{\{f \in \mathcal{F} : f(X_i) = \sigma_i \text{ for any } i = 1,\dots,2N\};\sigma_1^{2N} \in \{0;1\}^{2N}\Big\}.$$

Let $N(\mathbb{X}) \triangleq \big|\mathcal{A}(\mathbb{X})\big| = \big|\big\{[f(X_k)]_{k=1}^{2N} : f \in \mathcal{F}\big\}\big|$ be the number of ways of shattering $\mathbb{X}$ using functions in the model and let $\pi_{\mathcal{U}(\mathbb{X})}$ denotes an exchangeable distribution uniform on $\mathcal{A}(\mathbb{X})$ to the extent that $\pi_{\mathcal{U}(\mathbb{X})}(A) = \frac{1}{N(\mathbb{X})}$ for any $A \in \mathcal{A}(\mathbb{X})$.

7.1. Basic bound.

Theorem 7.1. *With $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-\epsilon$, for any $f_1, f_2 \in \mathcal{F}$, we have*

$$r'(f_2) - r'(f_1) \leq r(f_2) - r(f_1) + \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{f_1,f_2}\big[2\log N(\mathbb{X}) + \log(\epsilon^{-1})\big]}{N}}.$$

In particular, introducing $\tilde{f}' \triangleq \text{argmin}_{\mathcal{F}}\, r'$, we obtain

$$(7.1)\qquad r'(\hat{f}_{ERM}) - r'(\tilde{f}') \leq r(\hat{f}_{ERM}) - r(\tilde{f}') + \sqrt{\tfrac{8\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM},\tilde{f}'}[2\log N(\mathbb{X})+\log(\epsilon^{-1})]}{N}}.$$

Proof. Let $\nu[(df_1,df_2)] \triangleq \pi_{\mathcal{U}(\mathbb{X})}(df_1)\pi_{\mathcal{U}(\mathbb{X})}(df_2)$. By taking $\pi_{\mathcal{U}(\mathbb{X})}$ such that it put masses on only one function in each set of the partition $\mathcal{A}(\mathbb{X})$, for any functions $f_1, f_2 \in \mathcal{F}$, there exist functions $f_1', f_2' \in \mathcal{F}$ such that

- f_1' and f_1 are in the same set of the partition,
- f_2' and f_2 are in the same set of the partition,
- $\nu[(f_1',f_2')] = \frac{1}{[N(\mathbb{X})]^2}$.

The result then follows from Inequality (8.7) applied to $\mathcal{W}\big[(f_1,f_2),Z\big] = \mathbb{1}_{Y\neq f_2(X)} - \mathbb{1}_{Y\neq f_1(X)}$. □

In particular, when $\mathcal{Y} = \{0;1\}$, introduce the local VC-dimension

$$h_{\mathbb{X}} \triangleq \max\big\{|A| : A \subset \mathbb{X} \text{ and } |\{A \cap f^{-1}(1) : f \in \mathcal{F}\}| = 2^{|A|}\big\}.$$

Since $\log N(\mathbb{X}) \leq h_{\mathbb{X}} \log\left(\frac{2eN}{h_{\mathbb{X}}}\right)$, we get

$$r'(\hat{f}_{\mathrm{ERM}}) - \min_{\mathcal{F}} r' \leq 4\sqrt{\frac{2h_{\mathbb{X}} \log\left(\frac{2eN}{h_{\mathbb{X}}}\right) + \log(\epsilon^{-1})}{2N}}.$$

Note that this last bound is very rough since we expect the variance term $\bar{\bar{\bar{\mathbb{P}}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}'}$ to be much smaller than 1. In Section 7.3, we propose an observable upper bound of this quantity, and more generally a way of empirically bounding any quantity depending on $\tilde{f}'$.

7.2. **Localized VC-bound.** For any $A \in \mathcal{A}(\mathbb{X})$, the empirical risks r and r' are constant on the set A. Let r_A and r'_A denote these values and $(r+r')_A \triangleq r_A + r'_A$.

Theorem 7.2. *For any $\lambda \geq 0$, define*

$$\mathcal{C}_\lambda(f) \triangleq \log \textstyle\sum_{A \in \mathcal{A}(\mathbb{X})} \exp\Big\{ -\lambda\big[(r+r')_A - (r+r')(f)\big]\Big\}.$$

Let $\mathcal{C}(f,g) \triangleq \min_{\lambda \geq 0} \big\{\mathcal{C}_\lambda(f) + \mathcal{C}_\lambda(g)\big\}$. For any $\epsilon > 0$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have

$$r'(\hat{f}_{ERM}) - r'(\tilde{f}') \leq r(\hat{f}_{ERM}) - r(\tilde{f}') + \sqrt{\frac{8\bar{\bar{\bar{\mathbb{P}}}}_{\hat{f}_{ERM},\tilde{f}'}[\mathcal{C}(\hat{f}_{ERM},\tilde{f}') + \log(\epsilon^{-1})]}{N}}. \tag{7.2}$$

Proof. The proof is similar to the one of Theorem 7.1. The difference comes from the choice of the prior distribution. Let $r'' \triangleq r + r'$ and $\lambda : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a real-valued function possibly depending on the data Z_1^{2N} in an exchangeable way. We take the exchangeable prior distribution

$$\nu(df_1, df_2) \triangleq \frac{\exp\big\{-\lambda(f_1,f_2)[r''(f_1)+r''(f_2)]\big\}}{\pi_{\mathcal{U}(\mathbb{X})} \otimes \pi_{\mathcal{U}(\mathbb{X})} \exp\{-\lambda(f_1,f_2)[r''(f_1)+r''(f_2)]\}} \cdot \pi_{\mathcal{U}(\mathbb{X})} \otimes \pi_{\mathcal{U}(\mathbb{X})}(df_1, df_2).$$

So for any functions $f, g \in \mathcal{F}$ such that $\pi_{\mathcal{U}(\mathbb{X})}(f) = \pi_{\mathcal{U}(\mathbb{X})}(g) = \frac{1}{N(\mathbb{X})}$, we have $\log \nu^{-1}(f,g) = \mathcal{C}_{\lambda(f,g)}(f) + \mathcal{C}_{\lambda(f,g)}(g)$. Since the parameter minimizing $\mathcal{C}_\lambda(f) + \mathcal{C}_\lambda(g)$ (at some small positive constant if the minimum does not exist) depends on the data in an exchangeable way, we can choose $\lambda(f,g)$ equal to this parameter. □

For $\lambda = 0$ $\big($i.e. by using that $\mathcal{C}(f,g) \leq \mathcal{C}_0(f) + \mathcal{C}_0(g)\big)$, we recover Inequality (7.1). By appropriately choosing the parameter λ, we may expect to have $\mathcal{C}(\hat{f}_{\mathrm{ERM}})$ and $\mathcal{C}(\tilde{f}')$ much smaller than $\log N(\mathbb{X})$.

Remark 7.1. To illustrate this assertion, consider the toy example in which we have $\mathcal{X} = [0;1]$, $\mathcal{F} = \{\mathbb{1}_{[\theta;1]}; \theta \in [0;1]\}$, $Y = \mathbb{1}_{X \geq \tilde{\theta}}$ for some $\tilde{\theta} \in [0;1]$ and $\mathbb{P}(dX)$ absolutely continuous wrt Lebesgue measure. Then we almost surely have $N(\mathbb{X}) = 2N+1$ and for any $\lambda \geq 0$

$$\begin{aligned} c_\lambda \triangleq \textstyle\sum_{A \in \mathcal{A}(\mathbb{X})} \exp\big(-\lambda[r_A + r'_A]\big) &\leq 1 + 2\textstyle\sum_{k=1}^{N} \exp\big(-k\frac{\lambda}{N}\big) \\ &= 1 + 2\exp\big(-\frac{\lambda}{N}\big)\frac{1-\exp(-\lambda)}{1-\exp\big(-\frac{\lambda}{N}\big)}. \end{aligned}$$

Let $\hat{r} \triangleq r'(\hat{f}_{\mathrm{ERM}}) + r(\tilde{f}')$. Inequality (7.2) gives $\hat{r} \leq \min\limits_{\lambda \geq 0} \sqrt{\frac{4\hat{r}[2\log c_\lambda + \lambda\hat{r} + \log(\epsilon^{-1})]}{N}}$. Taking $\lambda = \frac{N}{20}$, we obtain

$$\hat{r} \leq \min_{\lambda \geq 0} \left\{ \frac{8\log\big\{1 + \frac{2\exp(-\lambda/N)[1-\exp(-\lambda)]}{1-\exp(-\lambda/N)}\big\} + 4\log(\epsilon^{-1})}{N - 4\lambda} \right\} \leq \frac{37 + 5\log(\epsilon^{-1})}{N},$$

which has to be compared with $\hat{r} \le \frac{8\log(2N+1)+4\log(\epsilon^{-1})}{N}$ obtained for $\lambda = 0$, i.e. from the non localized bound. So localizing allows to have sharper bounds and in particular to get rid of the $\log N$ which appears in classical VC-bounds. However, numerically, since the previous minimum does not differ much from its value at $\lambda = 0$ for $N \le 200$, this improvement is not significant for small training samples.

7.3. **Empirical VC-bound taking into account the variance term.** This section proposes a way of locating the best function $\tilde{f}'$ in the model in a small subset containing the empirical risk minimizer. This can be useful to give observable bounds of any quantity depending on $\tilde{f}'$, and in particular to upper bound $\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}'}$.

Lemma 7.3. *Let* $\epsilon > 0$ *and*

$$\bar{\mathcal{F}} \triangleq \left\{ f \in \mathcal{F} : r(f) \le r(\hat{f}_{ERM}) + \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM},f}[2\log N(\mathbb{X})+\log(\epsilon^{-1})]}{N}} \right\}.$$

With $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$*, we have* $\tilde{f}' \in \bar{\mathcal{F}}$*.*

Proof. It directly comes from Inequality (7.1) and $r'(\hat{f}_{\mathrm{ERM}}) - r'(\tilde{f}') \ge 0$. □

As a consequence, Inequality (7.1) leads to

Theorem 7.4. *For any* $\epsilon > 0$*, with* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$*, we have*

$$r'(\hat{f}_{ERM}) - r'(\tilde{f}') \le \sup_{f\in\bar{\mathcal{F}}} \left\{ r(\hat{f}_{ERM}) - r(f) + \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM},f}[2\log N(\mathbb{X})+\log(\epsilon^{-1})]}{N}} \right\}$$

To simplify, we can weaken the previous inequality into

$$r'(\hat{f}_{\mathrm{ERM}}) - r'(\tilde{f}') \le \sqrt{\frac{8\sup_{\bar{\mathcal{F}}} \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\cdot}[2\log N(\mathbb{X})+\log(\epsilon^{-1})]}{N}}.$$

7.4. **In the inductive learning.** The following theorem is Theorem 7.1 adapted to the inductive inference.

Theorem 7.5. *With* $\left(\mathbb{P}^{\otimes 2N}\right)_*$*-probability at least* $1-\epsilon$*, for any functions* $f_1, f_2 \in \mathcal{F}$*, we have*

$$R(f_2) - R(f_1) \le r(f_2) - r(f_1) + \sqrt{\frac{8\mathbb{P}^{\otimes 2N}[\bar{\bar{\mathbb{P}}}_{f_1,f_2}|Z_1^N]\left\{2(\mathbb{P}^{\otimes 2N})^*[\log N(\mathbb{X})|X_1^N]+\log(\epsilon^{-1})\right\}}{N}}.$$

Proof. The result is similar to the one of Theorem 7.1 except that we use Inequality (8.9) instead of Inequality (8.7), and we conclude by using Cauchy-Schwarz inequality. □

In the inductive setting, the variance term $\mathbb{P}^{\otimes 2N}[\bar{\bar{\mathbb{P}}}_{f_1,f_2}|X_1^N] = \frac{\mathbb{P}_{f_1,f_2}+\bar{\mathbb{P}}_{f_1,f_2}}{2}$ and the complexity term $\left(\mathbb{P}^{\otimes 2N}\right)^*[\log N(\mathbb{X})|X_1^N]$ are not observable and we need extra concentration inequalities to convert them into observable quantities.

7.4.1. *Complexity term.* For the complexity term, the following lemma proposes theoretical and empirical bounds of it.

Lemma 7.6. *The conditional expectation* $\left(\mathbb{P}^{\otimes 2N}\right)^* \left[\log N(\mathbb{X}) \,\middle|\, X_1^N\right]$ *can be upper bounded*

- *by*

$$\sup_{x_{N+1}^{2N}\in\mathcal{X}^N} \log N(X_1^N, x_{N+1}^{2N}), \tag{7.3}$$

- *with* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$*, by*

$$2\log N(X_1^N) + (\log 2)\log(\epsilon^{-1})\Big(1+\sqrt{1+\tfrac{2\log N(X_1^N)}{(\log 2)\log(\epsilon^{-1})}}\Big), \tag{7.4}$$

- *with* $\left(\mathbb{P}^{\otimes 2N}\right)_*$*-probability at least* $1-2\epsilon$*, by*

$$\log N(X_1^{2N}) + 2(\log 2)\log(\epsilon^{-1})\Big(\tfrac{6}{5}+\sqrt{1+\tfrac{2\log N(X_1^{2N})}{(\log 2)\log(\epsilon^{-1})}}\Big), \tag{7.5}$$

- *with* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$*, by*

$$\left(\mathbb{P}^{\otimes 2N}\right)^* \log N(X_1^{2N}) + \tfrac{(\log 2)\log(\epsilon^{-1})}{3}\Big(1+\sqrt{1+\tfrac{18\left(\mathbb{P}^{\otimes 2N}\right)^*\log N(X_1^{2N})}{(\log 2)\log(\epsilon^{-1})}}\Big). \tag{7.6}$$

Proof. The first bound is trivial. For Inequalities (7.4), (7.5) and (7.6), we use fine concentration inequalities due to Boucheron, Lugosi and Massart ([3]). Let $\log_2$ denote the binary logarithm: $\log_2 x \triangleq \frac{\log x}{\log 2}$ for any $x > 0$. The quantities $\log_2 N(X_1^N)$, $\log_2 N(X_1^{2N})$ and $\left(\mathbb{P}^{\otimes 2N}\right)^*\left[\log_2 N(X_1^{2N})|X_1^N\right]$ are self-bounded quantities in the sense given in [13, p.23][15]. By Theorem 15 in [13, p.40] and some computations, any self-bounded variable Z satisfy

- with probability at least $1-\epsilon$, $Z \le \mathbb{E}Z + \frac{\log(\epsilon^{-1})}{3}\Big(1+\sqrt{1+\frac{18\mathbb{E}Z}{\log(\epsilon^{-1})}}\Big)$,
- with probability at least $1-\epsilon$, $\mathbb{E}Z \le Z + \log(\epsilon^{-1})\Big(1+\sqrt{1+\frac{2Z}{\log(\epsilon^{-1})}}\Big)$.

From the inequality $\log N(X_1^{2N}) \le \log N(X_1^N) + \log N(X_{N+1}^{2N})$ and bounding the expectation of $\log_2 N(X_{N+1}^{2N})$ using the previous inequality, we obtain (7.4). Using both previous concentration inequalities, we link $\left(\mathbb{P}^{\otimes 2N}\right)^*\left[\log_2 N(X_1^{2N})|X_1^N\right]$ with $\left(\mathbb{P}^{\otimes 2N}\right)^*\left[\log_2 N(X_1^{2N})\right]$ and $\left(\mathbb{P}^{\otimes 2N}\right)^*\left[\log_2 N(X_1^{2N})\right]$ with $\log_2 N(X_1^{2N})$. After some computations, we get Inequality (7.5). Inequality (7.6) directly comes from the first of the two concentration inequalities. □

Remark 7.2. Bound (7.5) is useful only if the user possesses N extra input points $X_{2N+1},\dots,X_{2N}$ drawn independently according to the distribution $\mathbb{P}(dX)$. Contrarily to the transductive setting, these points are not necessarily (the) points to be classified. In the absence of these extra points, we should use Inequality (7.4) to give an empirical bound of the complexity term.

7.4.2. *Variance term.* Let $\mathcal{K} \triangleq \mathbb{P}^{\otimes 2N}[2\log N(\mathbb{X})|Z_1^N]+\log(\epsilon^{-1})$. We have just seen how to bound $\mathcal{K}$ with an observable or theoretical bound. To deal with the variance term, we can use the following lemma:

Lemma 7.7. *For any* $\epsilon > 0$*, we have*

- *with* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$*, for any functions* $f_1, f_2 \in \mathcal{F}$*,*

$$\mathbb{P}_{f_1,f_2} \le \bar{\mathbb{P}}_{f_1,f_2} + 2\sqrt{\tfrac{\mathcal{K}}{N}\big(\bar{\mathbb{P}}_{f_1,f_2}+\tfrac{\mathcal{K}}{4N}\big)} + \tfrac{\mathcal{K}}{N} \tag{7.7}$$

[15] For the self-boundedness of the quantity $\left(\mathbb{P}^{\otimes 2N}\right)^*\left[\log_2 N(X_1^{2N})|X_1^N\right]$, we first prove that for any $x_{N+1}^{2N} \in \mathcal{X}^N$, the quantity $\log_2 N(X_1^N, x_{N+1}^{2N})$ is self-bounded. This can be done by introducing the quantities $\log_2 N(X_1,\dots,X_{i-1},X_{i+1},\dots,X_N,x_{N+1}^{2N})$ for any $1 \le i \le N$ and slightly modifying Han's inequality ([13, p.31]). Then we take the outer expectations.

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any functions* $f_1, f_2 \in \mathcal{F}$*,*

$$\bar{\mathbb{P}}_{f_1,f_2} \le \mathbb{P}_{f_1,f_2} + 2\sqrt{\tfrac{\mathcal{K}}{N}\big(\mathbb{P}_{f_1,f_2} + \tfrac{\mathcal{K}}{4N}\big)} + \tfrac{\mathcal{K}}{N} \tag{7.8}$$

Proof. By using the same prior distribution as in the proof of Theorem 7.1 and by applying Inequality (8.9) to $\mathcal{W}[(f_1,f_2),X] = \mathbb{1}_{f_1(X)\neq f_2(X)}$, we obtain

$$\mathbb{P}_{f_1,f_2} - \bar{\mathbb{P}}_{f_1,f_2} - \big(\mathbb{P}^{\otimes 2N}\big)^* \sqrt{\tfrac{4\bar{\bar{\mathbb{P}}}_{f_1,f_2}[2\log N(\mathbb{X}) + \log(\epsilon^{-1})]}{N}} \le 0,$$

hence, setting $P = \sqrt{\mathbb{P}_{f_1,f_2} + \bar{\mathbb{P}}_{f_1,f_2}}$ and using Cauchy-Scwarz inequality, we obtain

$$P^2 \le 2\bar{\mathbb{P}}_{f_1,f_2} + P\sqrt{\tfrac{2\mathcal{K}}{N}}.$$

Solving this quadratic equation leads to the first assertion of the theorem.

For the second inequality, it suffices to take $\mathcal{W}[(f_1,f_2),X] = -\mathbb{1}_{f_1(X)\neq f_2(X)}$ instead of $\mathcal{W}[(f_1,f_2),X] = \mathbb{1}_{f_1(X)\neq f_2(X)}$. □

7.4.3. *Conclusion.* Let $\tilde{f} \in \operatorname{argmin}_{\mathcal{F}} R$. Combining Theorem 7.5, Lemma 7.6 and Lemma 7.7, we obtain an empirical bound of $R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f})$ except for the $\bar{\mathbb{P}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}$ quantity. This last quantity can be bounded using a locating scheme as the one given in Section 7.3.

Combining the three previous results, we can also give a theoretical bound of $R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f})$ except for the $\mathbb{P}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}$ quantity. Under Tsybakov's margin assumption, this quantity can be bounded with $C\big[R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f})\big]^{\frac{1}{\kappa}}$ for some $\kappa \ge 1$. This leads to the following satisfactory theoretical bound:

Theorem 7.8. *When* $\mathcal{F}$ *is a VC-class of dimension* h*, with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, we have* $R(\hat{f}_{ERM}) - R(\tilde{f}) \le C\log(e\epsilon^{-1})\big(\frac{h}{N}\log N\big)^{\frac{\kappa}{2\kappa-1}}$.

This is the known optimal convergence rate in this situation up to possibly the logarithmic factor (see [15, Corollary 2.2] and [1] for more details).

8. General PAC-Bayesian bounds

Let $Z_1, \dots, Z_N$ be N i.i.d. random variables distributed according to a probability distribution $\mathbb{P}$ on a measurable space $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$. Let $(\mathcal{G}, \mathcal{B}_{\mathcal{G}})$ be a measurable space and $\mathcal{M}_+^1(\mathcal{G})$ be the set of probability distributions on this space. Let $\mathcal{B}_{\mathbb{R}}$ denote the Borel σ-algebra on $\mathbb{R}$.

8.1. A basic PAC-Bayesian bound.

Theorem 8.1. *Let* $\mathcal{W} : (\mathcal{G}\times\mathcal{Z}, \mathcal{B}_{\mathcal{G}}\otimes\mathcal{B}_{\mathcal{Z}}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ *be a measurable function. Let* $\epsilon > 0$*,* $\lambda > 0$*,* $B \triangleq \sup_{\mathcal{G}\times\mathcal{Z}} \mathcal{W}$*,* $g(u) \triangleq \frac{\exp(u)-1-u}{u^2}$*,* $a_c(\lambda) \triangleq \frac{\lambda}{N} g\big(\frac{\lambda}{N}c\big)$ *and* $\nu \in \mathcal{M}_+^1(\mathcal{G})$*. We have*

- *with* $\mathbb{P}^{\otimes N}$*-probability at least* $1-\epsilon$*, for any distribution* $\mu \in \mathcal{M}_+^1(\mathcal{G})$*,*

$$\mu\bar{\mathbb{P}}\mathcal{W} - \mu\mathbb{P}\mathcal{W} \le a_B(\lambda)\mu\mathbb{P}\mathcal{W}^2 + \frac{K(\mu,\nu) + \log(\epsilon^{-1})}{\lambda}, \tag{8.1}$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any function* $f \in \mathcal{G}$*,*

$$\begin{aligned}
&\bar{\mathbb{P}}\mathcal{W}(f,\cdot) - \mathbb{P}\mathcal{W}(f,\cdot)\\
&\qquad\qquad \le \inf_{x>0}\Big\{a_B(x)\mathbb{P}\mathcal{W}(f,\cdot)^2 + \tfrac{\log \nu^{-1}(f)+\log(\epsilon^{-1})}{x}\Big\}\\
&\qquad\qquad \le \sqrt{\tfrac{2[\log\nu^{-1}(f)+\log(\epsilon^{-1})]\mathbb{P}\mathcal{W}(f,\cdot)^2}{N}} + (B\vee 0)\tfrac{\log\nu^{-1}(f)+\log(\epsilon^{-1})}{3N}.
\end{aligned}\tag{8.2}$$

The proof relies on the following lemma and on Legendre transform.

Lemma 8.2. *Let* W *be a random variable bounded by* $b \in \mathbb{R}$*. Then for any* $\eta > 0$*, we have*

$$\log \mathbb{E}\exp\big\{\eta(W - \mathbb{E}W)\big\} \le \eta^2 \mathbb{E}W^2 g(\eta b).$$

Proof. We have

$$\exp(\eta W) = 1 + \eta W + \eta^2 W^2 g(\eta W).$$

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

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

which is the desired result. □

Now let us prove Theorem 8.1. We have

$$\begin{aligned}
&\mathbb{P}^{\otimes N}\Bigg(\sup_{\mu\in\mathcal{M}_+^1(\mathcal{G})}\Big\{\mu\big[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2\big] - \tfrac{K(\mu,\nu)+\log(\epsilon^{-1})}{\lambda}\Big\} > 0\Bigg)\\
&\qquad = \mathbb{P}^{\otimes N}\Big(\tfrac{1}{\lambda}\log\Big[\epsilon\nu\exp\big\{\lambda[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2]\big\}\Big] > 0\Big)\\
&\qquad = \mathbb{P}^{\otimes N}\Big(\epsilon\nu\exp\big\{\lambda[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2]\big\} > 1\Big)\\
&\qquad \le \mathbb{P}^{\otimes N}\Big(\epsilon\nu\exp\big\{\lambda[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2]\big\}\Big)\\
&\qquad = \epsilon\nu\mathbb{P}^{\otimes N}\exp\big\{\lambda[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2]\big\}\\
&\qquad = \epsilon\nu\exp\big\{-\lambda a_B(\lambda)\mathbb{P}\mathcal{W}^2\big\}\Big(\mathbb{P}\exp\big\{\tfrac{\lambda}{N}[\mathcal{W} - \mathbb{P}\mathcal{W}]\big\}\Big)^N\\
&\qquad \le \epsilon,
\end{aligned}\tag{8.3}$$

where at the last step we use Lemma 8.2.

To prove Inequality (8.2), it suffices to note that when we allow the parameter λ to depend on f, we get

$$\mu\big\{\lambda\big[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2\big]\big\} \le K(\mu,\nu) + \log(\epsilon^{-1}).$$

Taking $\mu = \delta_f$, we obtain

$$\bar{\mathbb{P}}\mathcal{W}(f,\cdot) - \mathbb{P}\mathcal{W}(f,\cdot) \le a_B[\lambda(f)]\mathbb{P}\mathcal{W}(f,\cdot)^2 + \tfrac{\log[\nu^{-1}(f)]+\log(\epsilon^{-1})}{\lambda(f)}.$$

Choosing $\lambda(f)$ appropriately, we obtain the first part of Inequality (8.2). To prove the second part, it suffices to note that for any $A \ge 0$, we have[16] $\inf_{x>0}\big\{xg(x)+\frac{A^2}{2x}\big\} \le A+\frac{A^2}{6}$ and $\inf_{x>0}\big\{\frac{x}{2}+\frac{A^2}{2x}\big\} \le A$. The last inequality is used when $B \le 0$ since $g(u) \le \frac{1}{2}$ for $u \le 0$.

[16] Proof: we have $\inf_{x>0}\{xg(x) + \frac{A^2}{2x}\} \le \log(1+A)g[\log(1+A)] + \frac{A^2}{2\log(1+A)} = A + \frac{A^2}{6} - \frac{1+A+\frac{A^2}{6}}{\log(1+A)}k(A)$ where $k(A) \triangleq \log(1+A) - \frac{A+\frac{A^2}{2}}{1+A+\frac{A^2}{6}}$. Since $k(0) = 0$ and $k'(A) = \frac{A^4}{36(1+A)(1+A+A^2/6)^2} \ge 0$, we get $k(A) \ge 0$, hence the result.

Remark 8.1. In Inequality (8.1), we can replace $\mathbb{P}\mathcal{W}^2$ with $\mathbb{V}\mathrm{ar}_{\mathbb{P}}\mathcal{W}$ provided that $B \triangleq \sup_{\mathcal{G}\times\mathcal{Z}} \mathcal{W}$ is replaced with $B' \triangleq B - \mathbb{P}\mathcal{W}$. To obtain this result, it suffices to substitute Lemma 8.2 with: for any random variable W such that $b' \triangleq \sup W - \mathbb{P}W$, we have $\log \mathbb{P} \exp\big\{\eta(W - \mathbb{P}W)\big\} \le \eta^2 \mathbb{V}\mathrm{ar}_{\mathbb{P}} W g(\eta b')$.

Remark 8.2. Inequalities (8.2) can also be proven using a Bennett's type inequality: for any i.i.d. random variables W_i upper bounded by B, we have

$$\mathbb{P}^{\otimes N}\Big(\tfrac{\sum_{i=1}^N (W_i - \mathbb{P}W)}{N} > \inf_{x>0}\Big\{ug(uB)\mathbb{P}W^2 + \tfrac{\log(\epsilon^{-1})}{Nu}\Big\}\Big) \le \epsilon,$$

and a union bound. The link between both inequalities in (8.2) is similar to the one between Bennett's and Bernstein's inequality (see for instance [10, p.124]).

8.2. **Concentration of partition functions.** The following result is in particular useful for localizing and for getting theoretical bounds from data-dependent bounds and vice versa. We use the same notations as in Theorem 8.1. Let us introduce $A \triangleq -\inf_{\mathcal{G}\times\mathcal{Z}} \mathcal{W}$.

Theorem 8.3. *For any $\epsilon > 0$, $\lambda > 0$ and any probability distribution $\nu \in \mathcal{M}_+^1(\mathcal{G})$,*

- *for any $\lambda' > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have*

$$\log \nu \exp\big\{-\lambda \bar{\mathbb{P}}\mathcal{W}\big\} \ge \log \nu \exp\big\{-\lambda[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2]\big\} - \frac{\lambda}{\lambda'}\log(\epsilon^{-1}), \tag{8.4}$$

- *for any $\lambda' \ge \lambda$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have*

$$\log \nu \exp\big\{-\lambda \bar{\mathbb{P}}\mathcal{W}\big\} \le \log \nu \exp\big\{-\lambda[\mathbb{P}\mathcal{W} - a_A(\lambda')\mathbb{P}\mathcal{W}^2]\big\} + \frac{\lambda}{\lambda'}\log(\epsilon^{-1}). \tag{8.5}$$

Remark 8.3. Recall that $a_c(\lambda) \triangleq \frac{\lambda}{N} g\big(\frac{\lambda}{N}c\big)$ and $g : u \mapsto \frac{\exp(u)-1-u}{u^2}$ is a positive convex increasing function such that $g(0) = \frac{1}{2}$ by continuity. Theorems 8.1 and 8.3 trivially hold when A, B and $\mathbb{P}\mathcal{W}^2$ are replaced with respective upper bounds.

Proof. For the lower bound of $\log \nu \exp\big\{-\lambda \bar{\mathbb{P}}\mathcal{W}\big\}$, the proof is inspired from [6, Section 3]. Let $\mu' \triangleq \nu_{-\lambda[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2]}$. Applying Theorem 8.1 to $\mathcal{W}$ and the pair of distributions (μ', μ'), we get, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,

$$-\mu'[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2] \le -\mu'\bar{\mathbb{P}}\mathcal{W} + \frac{\log(\epsilon^{-1})}{\lambda'}.$$

So we have

$$\begin{aligned}
\log \nu \exp\big\{-\lambda[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2]\big\}& \\
&= -\lambda\mu'[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2] - K(\mu', \nu)\\
&\le -\lambda\mu'\bar{\mathbb{P}}\mathcal{W} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}) - K(\mu', \nu)\\
&\le \sup_{\mu \in \mathcal{M}_+^1(\mathcal{G})} \Big\{-\lambda\mu\bar{\mathbb{P}}\mathcal{W} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}) - K(\mu, \nu)\Big\}\\
&= \log \nu \exp\big\{-\lambda\bar{\mathbb{P}}\mathcal{W}\big\} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}).
\end{aligned}$$

For the upper bound of $\log \nu \exp\big\{-\lambda \bar{\mathbb{P}}\mathcal{W}\big\}$, introduce $\nu' \triangleq \nu_{-\lambda[\mathbb{P}\mathcal{W} - a_A(\lambda')\mathbb{P}\mathcal{W}^2]}$. We have

$$\begin{aligned}\mathbb{P}^{\otimes N}\big[\log \nu \exp\big\{-\lambda \bar{\mathbb{P}}\mathcal{W}\big\} > \log \nu \exp\big\{-\lambda[\mathbb{P}\mathcal{W} - a_A(\lambda')\mathbb{P}\mathcal{W}^2]\big\} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1})\big]&\\ = \mathbb{P}^{\otimes N}\Big(\nu' \exp\big\{\lambda[\mathbb{P}\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - a_A(\lambda')\mathbb{P}\mathcal{W}^2]\big\} > \epsilon^{-\frac{\lambda}{\lambda'}}\Big)&\\ = \mathbb{P}^{\otimes N}\Big(\epsilon\big[\nu' \exp\big\{\lambda[\mathbb{P}\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - a_A(\lambda')\mathbb{P}\mathcal{W}^2]\big\}\big]^{\frac{\lambda'}{\lambda}} > 1\Big)&\\ \le \epsilon\mathbb{P}^{\otimes N}\Big(\big[\nu' \exp\big\{\lambda[\mathbb{P}\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - a_A(\lambda')\mathbb{P}\mathcal{W}^2]\big\}\big]^{\frac{\lambda'}{\lambda}}\Big)&\\ \le \epsilon,&\end{aligned}$$

where at the last step we use Jensen's inequality, Fubini's theorem and $\mathbb{P}^{\otimes N}\exp\big\{\lambda'[\mathbb{P}\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - a_A(\lambda')\mathbb{P}\mathcal{W}^2]\big\} \le 1$. □

8.3. **PAC-Bayesian bounds with almost exchangeable prior.**

8.3.1. *Basic bound.* We still use the same notations as in Theorem 8.1. However in this section, $\mathcal{W}$ are allowed to depend on the data Z_1^{2N} in an exchangeable way. Introduce ν: $\mathcal{Z}^{2N} \to \mathcal{M}_+^1(\mathcal{G})$ an almost exchangeable (not necessarily ${\mathcal{B}_{\mathcal{Z}}}^{\otimes 2N}$-measurable) function (see Definition 1.1). We define the distributions $\bar{\mathbb{P}}' \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{Z_i}$ and $\bar{\bar{\mathbb{P}}} \triangleq \frac{1}{2N}\sum_{i=1}^{2N}\delta_{Z_i}$.

Theorem 8.4. *Let* $\mathbb{W} \triangleq \frac{\sum_{i=1}^N [\mathcal{W}(\cdot,Z_i) - \mathcal{W}(\cdot,Z_{N+i})]^2}{N}$. *For any* $\epsilon > 0$ *and* $\lambda > 0$, *we have*

- *with* $\big(\mathbb{P}^{\otimes 2N}\big)_*$*-probability at least* $1-\epsilon$, *for any distribution* $\mu \in \mathcal{M}_+^1(\mathcal{G})$,

$$\mu\bar{\mathbb{P}}'\mathcal{W} - \mu\bar{\mathbb{P}}\mathcal{W} \le \frac{\lambda}{2N}\mu\mathbb{W} + \frac{K(\mu, \nu_{Z_1^{2N}}) + \log(\epsilon^{-1})}{\lambda}. \tag{8.6}$$

- *with* $\big(\mathbb{P}^{\otimes 2N}\big)_*$*-probability at least* $1-\epsilon$, *for any function* $f \in \mathcal{G}$,

$$\bar{\mathbb{P}}'\mathcal{W}(f,\cdot) - \bar{\mathbb{P}}\mathcal{W}(f,\cdot) \le \sqrt{\frac{2\mathbb{W}(f)\big\{\log\big[\nu_{Z_1^{2N}}^{-1}(f)\big] + \log(\epsilon^{-1})\big\}}{N}} \tag{8.7}$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$, *we have*

$$\big(\mathbb{P}^{\otimes 2N}\big)^*\left\{\sup_{\mu\in\mathcal{M}_+^1(\mathcal{G})}\left[\mu\bar{\mathbb{P}}'\mathcal{W} - \mu\bar{\mathbb{P}}\mathcal{W} - \frac{\lambda}{2N}\mu\mathbb{W} - \frac{K(\mu, \nu_{Z_1^{2N}}) + \log(\epsilon^{-1})}{\lambda}\right] \,\middle|\, Z_1^N\right\} \le 0. \tag{8.8}$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$, *we have*

$$\big(\mathbb{P}^{\otimes 2N}\big)^*\left\{\sup_{f\in\mathcal{F}}\left[\bar{\mathbb{P}}'\mathcal{W}(f,\cdot) - \bar{\mathbb{P}}\mathcal{W}(f,\cdot) - \sqrt{\frac{2\mathbb{W}(f)\big[\log \nu_{Z_1^{2N}}^{-1}(f) + \log(\epsilon^{-1})\big]}{N}}\right] \,\middle|\, Z_1^N\right\} \le 0. \tag{8.9}$$

Note that we have $\mathbb{W} \le 4\bar{\bar{\mathbb{P}}}\mathcal{W}^2$ *(and even* $\mathbb{W} \le 2\bar{\bar{\mathbb{P}}}\mathcal{W}^2$ *when* $\mathcal{W}$ *is either positive or negative).*

Remark 8.4. To understand how the quantity $\mathbb{W}$ behaves, we can compute its expectation $\mathbb{P}^{\otimes 2N}\mathbb{W} = 2\mathbb{V}\mathrm{ar}_{\mathbb{P}}\mathcal{W}$ and note that, according to Corollary 8.5 with $Z_i \leftarrow (Z_i, Z_{N+i})$ and $\mathcal{W}(g,Z) \leftarrow \mathcal{W}\big(g,(Z,Z')\big) \triangleq \big[\mathcal{W}(g,Z) - \mathcal{W}(g,Z')\big]^2$, the quantity $\mu\mathbb{W}$ is concentrated around its expectation.

Proof. • Let $\mathcal{F}(\mathcal{G};\mathbb{R})$ be the set of real-valued functions over $\mathcal{G}$. Introduce an almost exchangeable function[17] $\eta : \mathcal{Z}^{2N} \to \mathcal{F}(\mathcal{G};\mathbb{R})$ such that for any $Z_1^{2N} \in \mathcal{Z}^{2N}$, the function $\eta(Z_1^{2N})$ is $\mathcal{B}_{\mathcal{G}}$-measurable.

Let us prove the first inequation. To shorten the inequalities, we introduce $S_i(g) \triangleq \mathcal{W}(g, Z_{N+i}) - \mathcal{W}(g, Z_i)$ for any $(g, Z_1^{2N}, i) \in \mathcal{G} \times \mathcal{Z}^{2N} \times \{1, \dots, N\}$. For any $\lambda > 0$, we have

$$\begin{aligned}
& \left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N}) + \lambda(\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W})\right]\Big\} \\
= \quad & \left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N}) + \tfrac{\lambda}{N} \textstyle\sum_{i=1}^N S_i\right]\Big\} \\
= \quad & \left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N})\right]\Pi_{i=1}^N \cosh\left(\tfrac{\lambda}{N} S_i\right)\Big\} \\
\leq \quad & \left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N}) + \tfrac{\lambda^2}{2N^2} \textstyle\sum_{i=1}^N S_i^2\right]\Big\},
\end{aligned}$$

where, at the last step, we use $\cosh x \leq \exp\left\{\frac{x^2}{2}\right\}$. Taking the exchangeable function $\eta(Z_1^{2N}) \triangleq -\frac{\lambda^2}{2N}\mathbb{W} - \log(\epsilon^{-1})$, we obtain

$$\left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N}) + \lambda(\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W})\right]\Big\} \leq \epsilon,$$

hence $\left(\mathbb{P}^{\otimes 2N}\right)^* \Big(\log \nu_{Z_1^{2N}}\big\{ \exp\left[\eta(Z_1^{2N}) + \lambda(\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W})\right]\big\} \geq 0\Big) \leq \epsilon$, Introducing

$$U \triangleq \sup_{\mu \in \mathcal{M}_+^1(\mathcal{G})} \big\{\mu\eta(Z_1^{2N}) + \lambda\mu(\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W}) - K(\mu, \nu_{Z_1^{2N}})\big\},$$

we have proved $\left(\mathbb{P}^{\otimes 2N}\right)^* (U \geq 0) \leq \epsilon$. Therefore, we get Inequality (8.6).

• The second assertion is deduced from the first one by using the same trick as for Inequality (8.2) and by noting that

$$\inf_{x>0} \left\{ \tfrac{x}{2N}\mathbb{W}(f) + \tfrac{\log[\nu^{-1}(f)] + \log(\epsilon^{-1})}{x} \right\} = \sqrt{\tfrac{2\mathbb{W}(f)\left\{ \log[\nu^{-1}(f)] + \log(\epsilon^{-1})\right\}}{N}}.$$

• We have seen that $\left(\mathbb{P}^{\otimes 2N}\right)^* \exp(U) \leq \epsilon$. By Jensen's inequality, we obtain[18] $\left(\mathbb{P}^{\otimes N}\right)^* \exp\left\{(\mathbb{P}^{\otimes 2N})^*(U|Z_1^N)\right\} \leq \epsilon$, hence $\left(\mathbb{P}^{\otimes N}\right)^* \Big\{ \left(\mathbb{P}^{\otimes 2N}\right)^* \left(U|Z_1^N\right) \geq 0 \Big\} \leq \epsilon$, which leads to Inequality (8.8).

• We obtain Inequality (8.9) by using the same argument as for Inequality (8.8). □

The following corollary shows the interest of Inequality (8.8).

Corollary 8.5. *Assume that the function $\mathcal{W}$ does not depend on the data Z_1^{2N}. For any $\epsilon > 0$ and $\lambda > 0$, with $\left(\mathbb{P}^{\otimes N}\right)_*$-probability at least $1 - \epsilon$, for any $\mu \in \mathcal{M}_+^1(\mathcal{G})$, we have*

(8.10)
$$\mu\mathbb{P}\mathcal{W} - \mu\bar{\mathbb{P}}\mathcal{W} \quad \leq \quad \tfrac{\lambda}{2N}\mu\mathbb{P}^{\otimes 2N}[\mathbb{W}|Z_1^N] + \frac{\left(\mathbb{P}^{\otimes 2N}\right)^* \left[K\left(\mu, \nu_{Z_1^{2N}}\right)|Z_1^N\right] + \log(\epsilon^{-1})}{\lambda}.$$

with $\mathbb{P}^{\otimes 2N}[\mathbb{W}|Z_1^N] = \mathbb{P}\mathcal{W}^2 + \bar{\mathbb{P}}\mathcal{W}^2 - 2\mathbb{P}\mathcal{W}\bar{\mathbb{P}}\mathcal{W} \leq 2(\mathbb{P}\mathcal{W}^2 + \bar{\mathbb{P}}\mathcal{W}^2)$. *(This last factor 2 can be omitted when $\mathcal{W}$ is either positive or negative).*

[17] to the extent that we have

$$\eta(Z_{\sigma(1)}, \dots, Z_{\sigma(2N)}, \cdot) = \eta(Z_1, \dots, Z_{2N}, \cdot)$$

for any $Z_1^{2N} \in \mathcal{Z}^{2N}$ and any permutation σ of $\{1, \dots, 2N\}$ satisfying $\{\sigma(i), \sigma(N+i)\} = \{i, N+i\}$ for any $i \in \{1, \dots, N\}$.

[18] Naturally, $(\mathbb{P}^{\otimes 2N})^*(U|Z_1^N)$ should be understood as $\left[\mathbb{P}^{\otimes 2N}(\cdot|Z_1^N)\right]^* U$.

Proof. We use Inequality (8.8) and note that $\big(\mathbb{P}^{\otimes 2N}\big)^* \{\mu\bar{\mathbb{P}}'\mathcal{W}|Z_1^N\} = \mu\mathbb{P}\mathcal{W}$ and $\big(\mathbb{P}^{\otimes 2N}\big)^* \{\mu\bar{\mathbb{P}}\mathcal{W}|Z_1^N\} = \mu\bar{\mathbb{P}}\mathcal{W}$. □

Remark 8.5. When the prior distribution $\nu_{Z_1^{2N}}$ puts masses on a finite set of points (chosen in an exchangeable way) and when we are in the inductive setting, the previous corollary is very limited since in general the posterior distribution which is taken using only the first sample will not be absolutely continuous wrt $\nu_{Z_1^{2N}}$. This happens in particular for the ERM-algorithm on an uncountable model. However, with nets and using differently (8.8), we can also deal with this case.

8.3.2. *Concentration of partition functions.* The following result is an adaptation of Theorem 8.3 to the exchangeable setting. We use the same notations as in Theorem 8.4.

Theorem 8.6. *For any $\epsilon > 0$ and $\lambda > 0$,*

- *for any $\lambda' > 0$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have*

$$\log \nu \exp\big\{-2\lambda\bar{\mathbb{P}}\mathcal{W}\big\} \geq \log \nu \exp\big\{-2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W} + \tfrac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]\big\} - \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}),$$

- *for any $\lambda' \geq \lambda$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have*

$$\log \nu \exp\big\{-2\lambda\bar{\mathbb{P}}\mathcal{W}\big\} \leq \log \nu \exp\big\{-2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W} - \tfrac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]\big\} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}).$$

Proof. For the lower bound, let $\mu' \triangleq \nu_{-2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W}+\frac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]}$. Applying Theorem 8.4 to $\mathcal{W}$ and the pair of probability distributions (μ', μ'), we get, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$,

$$-\mu'[\bar{\mathbb{P}}'\mathcal{W} + \frac{2\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2] \leq -\mu'\bar{\mathbb{P}}\mathcal{W} + \frac{\log(\epsilon^{-1})}{\lambda'}.$$

So we have

$$\begin{aligned}
&\log \nu \exp\big\{-2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W} + \tfrac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]\big\} \\
&\qquad = -2\lambda\mu'[\bar{\bar{\mathbb{P}}}\mathcal{W} + \tfrac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2] - K(\mu', \nu) \\
&\qquad \leq -2\lambda\mu'\bar{\mathbb{P}}\mathcal{W} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}) - K(\mu', \nu) \\
&\qquad \leq \sup_{\mu \in \mathcal{M}_+^1(\mathcal{G})} \Big\{-2\lambda\mu\bar{\mathbb{P}}\mathcal{W} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}) - K(\mu, \nu)\Big\} \\
&\qquad = \log \nu \exp\big\{-2\lambda\bar{\mathbb{P}}\mathcal{W}\big\} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}).
\end{aligned}$$

For the upper bound of $\log \nu \exp\big\{-2\lambda\bar{\mathbb{P}}\mathcal{W}\big\}$, introduce $\nu' \triangleq \nu_{-2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W}-\frac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]}$. We have

$$\begin{aligned}
&\mathbb{P}^{\otimes 2N}\big[\log \nu \exp\big\{-2\lambda\bar{\mathbb{P}}\mathcal{W}\big\} > \log \nu \exp\big\{-2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W} - \tfrac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]\big\} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1})\big] \\
&\qquad = \mathbb{P}^{\otimes 2N}\big(\nu' \exp\big\{\lambda[\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - \tfrac{2\lambda'}{N}\mathbb{P}\mathcal{W}^2]\big\} > \epsilon^{-\frac{\lambda}{\lambda'}}\big) \\
&\qquad = \mathbb{P}^{\otimes 2N}\Big(\epsilon\big[\nu' \exp\big\{\lambda[\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - \tfrac{2\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]\big\}\big]^{\frac{\lambda'}{\lambda}} > 1\Big) \\
&\qquad \leq \epsilon\mathbb{P}^{\otimes 2N}\Big(\big[\nu' \exp\big\{\lambda[\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - \tfrac{2\lambda'}{N}\mathbb{P}\mathcal{W}^2]\big\}\big]^{\frac{\lambda'}{\lambda}}\Big) \\
&\qquad \leq \epsilon,
\end{aligned}$$

where at the last step we use Jensen's inequality and

$$\mathbb{P}^{\otimes 2N}\nu' \exp\left\{\lambda'\left[\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - \frac{2\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2\right]\right\} \leq 1.$$

□

8.3.3. *Comparison between Theorem 8.4 and Theorem 8.1.* For comparison purposes, Theorem 8.1 leads to

$$\mu\mathbb{P}\mathcal{W} - \mu\bar{\mathbb{P}}\mathcal{W} \le \frac{\lambda}{N} g\bigg(\frac{\lambda}{N}\sup_{\mathcal{G}\times\mathcal{Z}}(-\mathcal{W})\bigg)\mu\mathbb{P}\mathcal{W}^2 + \frac{K(\mu,\nu)+\log(\epsilon^{-1})}{\lambda}.$$

We see that, thanks to the symmetrization argument, we can deal with unbounded variables $\mathcal{W}$. Inequality (8.10) is meaningful when the RHS is not infinite which is not a strong constraint on the unboundedness of $\mathcal{W}$.

The cost of taking an exchangeable prior is that, since $g(x) \underset{x\to 0}{\longrightarrow} \frac{1}{2}$, we roughly lose a factor 4 in the first term of the upper bound.

If $\mathcal{W}$ is either everywhere positive or everywhere negative, we just lose a factor 2. Otherwise, we can apply (8.10) to show that $\mathbb{P}^{\otimes 2N}[\mathbb{W}|Z_1^N]$ is concentrated around its expectation $2\mathbb{V}\mathrm{ar}_{\mathbb{P}}\mathcal{W}$. So even in this case, we lose a factor 2.

This factor 2 comes from the step in which we "take the conditional expectation" in (8.6) to obtain (8.8). In fact, we believe that Inequality (8.6) is tight since to some extent the difference $\bar{\mathbb{P}}\mathcal{W} - \bar{\mathbb{P}}'\mathcal{W}$ contains *twice* the deviations of $\mathcal{W}$ around its expectation.

8.4. **Compression schemes in the inductive learning.** The compression schemes in the inductive learning was recently developed in [18, 7]. Let $\hat{G}$ be a measurable real-valued function defined on $\cup_{n=1}^{+\infty}\mathcal{Z}^n \times \cup_{n=1}^{+\infty}\mathcal{Z}^n \times \mathcal{G} \times \mathcal{Z}$ upper bounded by a non negative constant B.

Introduce for any $h \in \mathbb{N}^*$, $\mathcal{I}_h \triangleq \{1,\dots,N\}^h$. Any set $I \in \mathcal{I}_h$ can be written as $I = \{i_1,\dots,i_h\}$. Define $I^c \triangleq \{1,\dots,N\} - \{i_1,\dots,i_h\}$ and $Z_I \triangleq \big(Z_{i_1},\dots,Z_{i_h}\big)$. The law of the random variable Z_I will be denoted $\mathbb{P}^I$.

Let $\mathcal{I} \triangleq \underset{2\le h\le N-1}{\cup} \mathcal{I}_h$ and $\nu : \mathcal{Z}^N \to \mathcal{M}_+^1(\mathcal{I}\times\mathcal{I}\times\mathcal{G})$ be some regular conditional probability measure such that

- $\nu_{Z_1^N}(I_1,I_2)$ is independent from Z_1^N,
- $\nu_{Z_1^N}(df|I_1,I_2)$ depends only on Z_{I_1} and Z_{I_2} $\big($and so will be denoted $\nu_{Z_{I_1},Z_{I_2}}(df)\big)$.

For any $J \subset \{1,\dots,N\}$, introduce $\bar{\mathbb{P}}^J \triangleq \frac{1}{|J|}\sum_{i\in J}\delta_{Z_i}$. Let $\mathcal{W}$ be the measurable real-valued function defined on $\mathcal{Z}^N \times \mathcal{I}\times\mathcal{I}\times\mathcal{G}\times\mathcal{Z}$ as

$$\mathcal{W}(Z_1^N, I_1, I_2, g, Z) = \hat{G}(Z_{I_1}, Z_{I_2}, g, Z).$$

Finally, for any sets I_1 and I_2 in $\mathcal{I}$, introduce $I_{1,2} \triangleq (I_1 \cup I_2)^c$.

Theorem 8.7. *Let $\epsilon > 0$, $\lambda > 0$ and for any $n \in \mathbb{N}^*$, $a_{c,n}(\lambda) \triangleq \frac{\lambda}{n} g\big(\frac{\lambda}{n}c\big)$. We have*

- *with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any $\mu \in \mathcal{M}_+^1(\mathcal{I}\times\mathcal{I}\times\mathcal{G})$,*

$$\mu\bar{\mathbb{P}}^{I_{1,2}}\mathcal{W} - \mu\mathbb{P}\mathcal{W} \le \mu\big[a_{B,|I_{1,2}|}(\lambda)\mathbb{P}\mathcal{W}^2\big] + \frac{K(\mu,\nu)+\log(\epsilon^{-1})}{\lambda}, \tag{8.11}$$

- *with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-\epsilon$, for any $I_1, I_2 \in \mathcal{I}$ and $f \in \mathcal{G}$,*

$$\begin{aligned}
&\bar{\mathbb{P}}^{I_{1,2}}\hat{G}(Z_{I_1},Z_{I_2},f,\cdot) - \mathbb{P}\hat{G}(Z_{I_1},Z_{I_2},f,\cdot)\\
&\quad\le \min_{x>0}\Big\{a_{B,|I_{1,2}|}(x)\mathbb{P}\hat{G}^2(Z_{I_1},Z_{I_2},f,\cdot) + \tfrac{\log\nu^{-1}(I_1,I_2,f)+\log(\epsilon^{-1})}{x}\Big\}\\
&\quad\le \sqrt{\tfrac{2[\log\nu^{-1}(I_1,I_2,f)+\log(\epsilon^{-1})]\mathbb{P}\hat{G}^2(Z_{I_1},Z_{I_2},f,\cdot)}{|I_{1,2}|}} + B\tfrac{\log\nu^{-1}(I_1,I_2,f)+\log(\epsilon^{-1})}{3|I_{1,2}|}.
\end{aligned} \tag{8.12}$$

Proof. • It suffices to modify the proof of Theorem 8.1. Specifically, in Inequalities (8.3), we can no longer use Fubini's theorem to swap $\mathbb{P}^{\otimes N}$ and ν. However, we have

$$\mathbb{P}^{\otimes N}\nu_{Z_1^N}(dI_1, dI_2, df) = \nu(dI_1, dI_2)\mathbb{P}^{I_1 \cup I_2}(dZ_{I_1 \cup I_2})\nu_{Z_{I_1}, Z_{I_2}}(df)\mathbb{P}^{I_{1,2}}(dZ_{I_{1,2}}),$$

which is sufficient to get the result, since for any $(I_1, I_2, f) \in \mathcal{I} \times \mathcal{I} \times \mathcal{G}$ we have

$$\mathbb{P}^{I_{1,2}} \exp\left\{\lambda[\bar{\mathbb{P}}^{I_{1,2}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_{B,|I_{1,2}|}(\lambda)\mathbb{P}\mathcal{W}^2]\right\} \leq 1.$$

• We use the same trick as for Inequality (8.2) by considering a parameter λ depending on (I_1, I_2, f). □

9. Proofs

9.1. Proof of Theorem 3.1.

In this proof, we put ourselves in the event

$$\Big\{\text{for any } f_1, f_2 \in \hat{\mathcal{F}},\ r'(f_2) - r'(f_1) \leq r(f_2) - r(f_1) + S(f_1, f_2)\Big\}.$$

From Theorem 5.1, with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1 - \epsilon$, this event holds.

• Since we have $S(f_{k-1}, f_k) \geq 0$ and $r(f_k) - r(f_{k-1}) + S(f_{k-1}, f_k) < 0$, we obtain $r(f_k) < r(f_{k-1})$. As a consequence, the iterative is not infinite: there exists $0 \leq K \leq N$ such that f_K exists but not f_{K+1}.

We have

$$r'(f_k) - r'(f_{k-1}) \leq r(f_k) - r(f_{k-1}) + S(f_{k-1}, f_k).$$

From the definition of f_k, we obtain $r'(f_k) < r'(f_{k-1})$.

• Let us prove the second item by induction. Since f_0 has been taken in the set of smallest complexity, we have necessarily $\mathcal{C}(f_1) \geq \mathcal{C}(f_0)$. When $\mathcal{C}(f_{k-1}) \geq \mathcal{C}(f_{k-2})$, we will prove that $\mathcal{C}(f_k) \geq \mathcal{C}(f_{k-1})$ by contradiction. We have

$$r(f_{k-1}) - r(f_{k-2}) + S(f_{k-1}, f_{k-2}) < 0,$$

and

$$r(f_k) - r(f_{k-1}) + S(f_k, f_{k-1}) < 0.$$

Assume that $\mathcal{C}(f_k) < \mathcal{C}(f_{k-1})$, then, by definition of f_{k-1}, we also have

$$r(f_k) - r(f_{k-2}) + S(f_k, f_{k-2}) \geq 0$$

and we get

$$S(f_k, f_{k-2}) > S(f_k, f_{k-1}) + S(f_{k-1}, f_{k-2}).$$

Since we have $\bar{\bar{\mathbb{P}}}_{f_k, f_{k-2}} \leq \bar{\bar{\mathbb{P}}}_{f_k, f_{k-1}} + \bar{\bar{\mathbb{P}}}_{f_{k-1}, f_{k-2}}$, for any $a, b > 0$ $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $\mathcal{C}(f_{k-2}) \leq \mathcal{C}(f_{k-1})$ and $\bar{\bar{\mathbb{P}}}_{f_{k-2}, f_{k-1}} \neq 0$, we obtain that $\mathcal{C}(f_k) > \mathcal{C}(f_{k-1})$, hence the contradiction. This concludes the induction.

• For any $f \in \hat{\mathcal{F}}$, we have $r'(f_K) \leq r'(f_{k(f)}) \leq r'(f) + r(f_{k(f)}) - r(f) + S(f_{k(f)}, f)$. Since by definition of $k(f)$ we have $r(f) - r(f_{k(f)}) + S(f_{k(f)}, f) \geq 0$, we obtain $r'(f_K) \leq r'(f) + 2S(f_{k(f)}, f)$.

• We have just seen that for any $f \in \hat{\mathcal{F}}$, $r'(f_{k(f)}) \leq r'(f) + 2S(f_{k(f)}, f)$, hence

$$r'(f_{k(f)}) \leq 2r'(f) - r'(f_{k(f)}) + 8\sqrt{\frac{2\bar{\bar{\mathbb{P}}}_{f, f_{k(f)}}[C(f) + C(f_{k(f)}) + L]}{N}}.$$

Therefore we have

$$r'(f_K) \leq \sup_{g \in \hat{\mathcal{F}} : h(g) \leq h(f)} \left\{2r'(f) - r'(g) + 8\sqrt{\frac{2\bar{\bar{\mathbb{P}}}_{f,g}[C(f) + C(g) + L]}{N}}\right\}$$

9.2. **Proof of Theorem 3.3.** • The first assertion holds by continuity.

• Since $S_\lambda(\rho,\rho_{k-1})>0$, we have $\rho_k r<\rho_{k-1}r$, hence $\rho_k r\le\rho_{k-1}r-\frac{1}{N}$. So the iterative scheme ends at some step $K\le N$.

Consider the event on which Inequality (4.4) holds for $\zeta=\sqrt{e}$. Theorem 4.3 ensures that it has a $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$. In the remainder of the proof, we put ourselves on this event. Consequently, for any $k\in\{1,\dots,K\}$, we have

$$\rho_k r'-\rho_{k-1}r'\le\rho_k r-\rho_{k-1}r+S(\rho_k,\rho_{k-1})\le 0.$$

• By definition of ρ_0, we have $\rho_0 r+\frac{K(\rho_0,\pi)}{\lambda_0}\le\rho_1 r+\frac{K(\rho_1,\pi)}{\lambda_0}$. Since we have $\rho_1 r\le\rho_0 r-S(\rho_0,\rho_1)$, we obtain $\frac{K(\rho_0,\pi)}{\lambda_0}+S(\rho_0,\rho_1)\le\frac{K(\rho_1,\pi)}{\lambda_0}$, and consequently $K(\rho_1,\pi)>K(\rho_0,\pi)$.

For any $k\in\{2,\dots,K\}$, by definition of ρ_{k-1}, for any $\rho\in\mathcal{M}^1_+(\mathcal{F})$, we have either $K(\rho,\pi)\ge K(\rho_{k-1},\pi)$ or

$$\rho r-\rho_{k-2}r+S(\rho,\rho_{k-2})\ge 0.$$

This last inequality implies that $\rho r+S(\rho,\rho_{k-2})\ge\rho_{k-1}r+S(\rho_{k-1},\rho_{k-2})$.

Let us prove the inequality $K(\rho_k,\pi)\ge K(\rho_{k-1},\pi)$ by induction and contradiction. Assume that the inequalities $K(\rho_k,\pi)<K(\rho_{k-1},\pi)$ and $K(\rho_{k-1},\pi)\ge K(\rho_{k-2},\pi)$ hold. Then we have

$$\begin{cases}\rho_k r+S(\rho_{k-1},\rho_k)\le\rho_{k-1}r\\ \rho_{k-1}r+S(\rho_{k-2},\rho_{k-1})\le\rho_k r+S(\rho_{k-2},\rho_k),\end{cases}$$

hence $S(\rho_{k-1},\rho_k)+S(\rho_{k-2},\rho_{k-1})\le S(\rho_{k-2},\rho_k)$. Define $\lambda_{k'}\in[\sqrt{N};N]$ such that $S_{\lambda_{k'}}(\rho_{k'-1},\rho_{k'})=S(\rho_{k'-1},\rho_{k'})$. Let $\lambda=\lambda_{k-1}\wedge\lambda_k$. We have

$$S_{\lambda_k}(\rho_{k-1},\rho_k)+S_{\lambda_{k-1}}(\rho_{k-2},\rho_{k-1})\le S_\lambda(\rho_{k-2},\rho_k).$$

From the inequality $\rho_k\otimes\rho_{k-2}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\le\rho_k\otimes\rho_{k-1}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}+\rho_{k-1}\otimes\rho_{k-2}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$, we get

$$\frac{\tilde{\mathcal{K}}_{\rho_k,\rho_{k-1}}}{\lambda_k}+\frac{\tilde{\mathcal{K}}_{\rho_{k-1},\rho_{k-2}}}{\lambda_{k-1}}\le\frac{\tilde{\mathcal{K}}_{\rho_k,\rho_{k-2}}}{\lambda_k\wedge\lambda_{k-1}}.$$

Since we have $K(\rho_{k-1},\pi)\ge K(\rho_{k-2},\pi)$, we obtain successively $\lambda_k\ge\lambda_{k-1}$ and $K(\rho_k,\pi)\ge K(\rho_{k-1},\pi)$. So the result is proved by induction and contradiction.

• Let $\eta>0$. Consider $\tilde\lambda>0$ such that we have $\frac{\sqrt{N}}{2(\eta+2)\sqrt{e}}\le\tilde\lambda\le\frac{N}{2(\eta+2)\sqrt{e}}$ and $K(\pi_{-\tilde\lambda r'},\pi)\ge K(\rho_0,\pi)$. Define $\tilde\rho\triangleq\pi_{-\tilde\lambda r'}$. Introduce the largest integer $\tilde k$ such that $K(\rho_{\tilde k},\pi)\le K(\tilde\rho,\pi)$. We have $\min\limits_{\lambda\in[\sqrt{N};N]}\big\{\tilde\rho r-\rho_{\tilde k}r+S_\lambda(\tilde\rho,\rho_{\tilde k})\big\}>0$, hence for any $\lambda\in[\sqrt{N};N]$ and $\eta>0$,

$$\begin{aligned}\rho_{\tilde k}r'-\tilde\rho r'&\le 2S_\lambda(\tilde\rho,\rho_{\tilde k})\\ &\le \tfrac{4\lambda}{N}(\tilde\rho\otimes\rho_{\tilde k})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}+\tfrac{2\sqrt{e}}{\lambda}\big[(2+\eta)K(\tilde\rho,\pi)+L\big]-\tfrac{2\sqrt{e}\eta}{\lambda}K(\rho_{\tilde k},\pi),\end{aligned}$$

where $L\triangleq\log[\log(eN)\epsilon^{-1}]$. Now let us take $\lambda=2(\eta+2)\sqrt{e}\tilde\lambda\in[\sqrt{N};N]$ and introduce $\xi\in[0;1[$. By Legendre transform, we get

$$\begin{aligned}(1-\xi)\rho_{\tilde k}r'&\le -\xi\rho_{\tilde k}r'+\tfrac{4\lambda}{N}(\tilde\rho\otimes\rho_{\tilde k})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}-\tfrac{2\sqrt{e}\eta}{\lambda}K(\rho_{\tilde k},\pi)\\ &\qquad\qquad +\tilde\rho r'+\tfrac{1}{\tilde\lambda}K(\tilde\rho,\pi)+\tfrac{2\sqrt{e}}{\lambda}L\\ &\le \tfrac{\eta}{(\eta+2)\tilde\lambda}\log\pi\exp\Big\{-\tfrac{\tilde\lambda(\eta+2)}{\eta}\xi r'+\tfrac{8\tilde\lambda^2(\eta+2)^2\sqrt{e}}{\eta N}\tilde\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big\}\\ &\qquad\qquad -\tfrac{1}{\tilde\lambda}\log\pi\exp\big(-\tilde\lambda r'\big)+\tfrac{L}{(\eta+2)\tilde\lambda}.\end{aligned}$$

A natural choice for the parameter ξ is $\xi = \frac{\eta}{\eta+2}$ such that we obtain

$$\rho_{\tilde{k}} r' \quad \leq \quad \tfrac{\eta}{2\tilde{\lambda}} \log \pi_{-\tilde{\lambda} r'} \exp\Big\{ \tfrac{8\tilde{\lambda}^2(\eta+2)^2\sqrt{e}}{\eta N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} - \tfrac{1}{\tilde{\lambda}} \log \pi \exp\big(- \tilde{\lambda} r' \big) + \tfrac{L}{2\tilde{\lambda}}.$$

Taking $\eta = 1$ to simplify the result, we obtain the last assertion of the theorem since $\rho_K r' \leq \rho_{\tilde{k}} r'$.

9.3. **Proof of Theorem 3.4.**

9.3.1. *Preliminary lemma.* We will need the following technical lemma.

Lemma 9.1. *Let $\tilde{\pi} \in \mathcal{M}_+^1(\mathcal{F})$ possibly depending on Z_1^{2N} in an exchangeable way. Let $\epsilon > 0$, $\lambda' \geq \lambda > 0$, $\lambda'' > 0$ and $\alpha > 0$. We have*

- *with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1 - 2\epsilon$,*

$$\begin{aligned} &\log \pi_{-\lambda r} \exp\big(\alpha \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) \\ &\qquad \leq \log \pi_{-\lambda \frac{r+r'}{2}} \exp\Big\{ \Big(\alpha + \tfrac{\lambda\lambda'}{2N} + \tfrac{\lambda\lambda''}{2N}\Big) \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} + \big(\tfrac{1}{\lambda'} + \tfrac{1}{\lambda''}\big) \lambda \log(\epsilon^{-1}), \end{aligned} \tag{9.1}$$

- *for $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1 - 4\epsilon$,*

$$\begin{aligned} &\log\big(\pi_{-\lambda \frac{r+r'}{2}} \otimes \pi_{-\lambda \frac{r+r'}{2}}\big) \exp\big(\alpha\lambda \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) \\ &\qquad \leq \tfrac{2}{q} \log\big(\pi_{-\lambda \frac{r+r'}{2}} \otimes \pi_{-\lambda \frac{r+r'}{2}}\big) \exp\Big(\tfrac{2q+1}{2N} \big[\lambda' + \lambda''(1+\alpha^2)\big] \lambda \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big) \\ &\qquad\quad + \tfrac{1}{p} \log\big(\pi_{-\lambda r} \otimes \pi_{-\lambda r}\big) \exp\big(p\alpha\lambda \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) + \tfrac{q+2}{q} \big(\tfrac{\lambda}{\lambda'} + \tfrac{\lambda}{\lambda''}\big) \log(\epsilon^{-1}). \end{aligned} \tag{9.2}$$

Proof. • Let $\mathcal{W}(f, Z) = \mathbb{1}_{Y \neq f(X)} - \tilde{\pi} \mathbb{1}_{Y \neq \cdot(X)}$. We have

$$\log \pi_{-\lambda r} \exp\big(\alpha \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) = \log \pi \exp\big(- \lambda \bar{\mathbb{P}} \mathcal{W} + \alpha \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) - \log \pi \exp\big(- \lambda \bar{\mathbb{P}} \mathcal{W}\big).$$

By using Theorem 8.6 for appropriate prior distributions, we obtain

$$\begin{aligned} \log \pi_{-\lambda r} \exp\big(\alpha \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) \quad &\leq \quad \log \pi \exp\Big\{ - \lambda \bar{\bar{\mathbb{P}}} \mathcal{W} + \Big(\alpha + \tfrac{\lambda\lambda'}{2N}\Big) \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big) + \tfrac{\lambda}{\lambda'} \log(\epsilon^{-1}) \\ &\qquad\qquad - \log \pi \exp\big(- \lambda \bar{\bar{\mathbb{P}}} \mathcal{W} - \tfrac{\lambda\lambda''}{2N} \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) + \tfrac{\lambda}{\lambda''} \log(\epsilon^{-1}) \\ &= \quad \log \pi_{-\lambda \frac{r+r'}{2} - \frac{\lambda\lambda''}{2N} \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}} \exp\Big\{ \Big(\alpha + \tfrac{\lambda\lambda''}{2N} + \tfrac{\lambda\lambda''}{2N}\Big) \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} \\ &\qquad\qquad + \big(\tfrac{\lambda}{\lambda'} + \tfrac{\lambda}{\lambda''}\big) \log(\epsilon^{-1}) \\ &\leq \quad \log \pi_{-\lambda \frac{r+r'}{2}} \exp\Big\{ \Big(\alpha + \tfrac{\lambda\lambda''}{2N} + \tfrac{\lambda\lambda''}{2N}\Big) \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} \\ &\qquad\qquad + \big(\tfrac{\lambda}{\lambda'} + \tfrac{\lambda}{\lambda''}\big) \log(\epsilon^{-1}), \end{aligned}$$

where we used at the last step that

(1) $\pi_{-\lambda \frac{r+r'}{2} - \frac{\lambda\lambda''}{2N} \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}} = \big(\pi_{-\lambda \frac{r+r'}{2}}\big)_{- \frac{\lambda\lambda''}{2N} \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}}$

(2) for any $a > 0$, $\mathbb{E}\big(\frac{\exp(-X)}{\mathbb{E} \exp(-X)} \exp(aX)\big) \leq \mathbb{E} \exp(aX)$ $\big($since we have $\mathbb{C}\text{ov}\big(\exp(aX), \exp(-X)\big) \leq 0$ $\big)$.

• Let us introduce $\mathcal{W}'\big((f_1, f_2), Z\big) = \mathbb{1}_{Y \neq f_1(X)} + \mathbb{1}_{Y \neq f_2(X)} - 2\tilde{\pi} \mathbb{1}_{Y \neq \cdot(X)}$ and $\mathcal{W}''\big((f_1, f_2), Z\big) = \mathbb{1}_{Y \neq f_1(X)} + \mathbb{1}_{Y \neq f_2(X)} - 2\tilde{\pi} \mathbb{1}_{Y \neq \cdot(X)} - \alpha \mathbb{1}_{f_1(X) \neq f_2(X)}$. We have

$$\begin{aligned} \log\big(\pi_{-\lambda \frac{r+r'}{2}} \otimes \pi_{-\lambda \frac{r+r'}{2}}\big) \exp\big(\alpha\lambda \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) = \log \pi \otimes \pi \exp\big(- \lambda \bar{\bar{\mathbb{P}}} \mathcal{W}''\big) \\ - \log \pi \otimes \pi \exp\big(- \lambda \bar{\bar{\mathbb{P}}} \mathcal{W}'\big). \end{aligned}$$

From Theorem 8.6 and the inequalities

$$\begin{cases} \bar{\bar{\mathbb{P}}} \mathcal{W}'^2 \leq 2\alpha^2 \bar{\bar{\mathbb{P}}}_{f_1,f_2} + 2\tilde{\pi} \bar{\bar{\mathbb{P}}}_{f_1,\cdot} + 2\tilde{\pi} \bar{\bar{\mathbb{P}}}_{f_2,\cdot} \leq 2(1+\alpha^2) \tilde{\pi} \big(\bar{\bar{\mathbb{P}}}_{f_1,\cdot} + \bar{\bar{\mathbb{P}}}_{f_2,\cdot}\big) \\ \bar{\bar{\mathbb{P}}} \mathcal{W}''^2 \leq 2\tilde{\pi} \big(\bar{\bar{\mathbb{P}}}_{f_1,\cdot} + \bar{\bar{\mathbb{P}}}_{f_2,\cdot}\big) \end{cases},$$

we have

$$\begin{aligned}
&\log\big(\pi_{-\lambda\frac{r+r'}{2}}\otimes\pi_{-\lambda\frac{r+r'}{2}}\big)\exp\big(\alpha\lambda\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)\\
&\quad\le\log\pi\otimes\pi\exp\big(-\lambda\bar{\mathbb{P}}\mathcal{W}''+\tfrac{\lambda\lambda''}{2N}\bar{\bar{\mathbb{P}}}\mathcal{W}''^2\big)+\tfrac{\lambda}{\lambda''}\log(\epsilon^{-1})\\
&\qquad-\log\pi\otimes\pi\exp\big(-\lambda\bar{\bar{\mathbb{P}}}\mathcal{W}'-\tfrac{\lambda\lambda'}{2N}\bar{\bar{\mathbb{P}}}\mathcal{W}'^2\big).+\tfrac{\lambda}{\lambda'}\log(\epsilon^{-1})\\
&\quad\le\log\pi\otimes\pi\exp\Big\{-\lambda\big[r(f_1)+r(f_2)-\alpha\bar{\mathbb{P}}_{f_1,f_2}\big]+\tfrac{\lambda\lambda''(1+\alpha^2)}{N}\tilde{\pi}\big(\bar{\bar{\mathbb{P}}}_{f_1,\cdot}+\bar{\bar{\mathbb{P}}}_{f_2,\cdot}\big)\Big\}\\
&\qquad-\log\pi\otimes\pi\exp\Big\{-\lambda\big[r(f_1)+r(f_2)\big]-\tfrac{\lambda\lambda'}{N}\tilde{\pi}\big(\bar{\bar{\mathbb{P}}}_{f_1,\cdot}+\bar{\bar{\mathbb{P}}}_{f_2,\cdot}\big)\Big\}\\
&\qquad+\big(\tfrac{\lambda}{\lambda'}+\tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1})\\
&\quad\le\log\pi_{-\lambda r}\otimes\pi_{-\lambda r}\exp\Big\{\alpha\lambda\bar{\mathbb{P}}_{f_1,f_2}+\tfrac{\lambda\lambda''(1+\alpha^2)}{N}\tilde{\pi}\big(\bar{\bar{\mathbb{P}}}_{f_1,\cdot}+\bar{\bar{\mathbb{P}}}_{f_2,\cdot}\big)\Big\}\\
&\qquad-\log\pi_{-\lambda r}\otimes\pi_{-\lambda r}\exp\big\{-\tfrac{\lambda\lambda'}{N}\tilde{\pi}\big(\bar{\bar{\mathbb{P}}}_{f_1,\cdot}+\bar{\bar{\mathbb{P}}}_{f_2,\cdot}\big)\big\}+\big(\tfrac{\lambda}{\lambda'}+\tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1}).
\end{aligned}$$

From Hölder's inequality and Jensen's inequality, we get

$$\begin{aligned}
&\log\big(\pi_{-\lambda\frac{r+r'}{2}}\otimes\pi_{-\lambda\frac{r+r'}{2}}\big)\exp\big(\alpha\lambda\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)\\
&\quad\le\tfrac{1}{p}\log\pi_{-\lambda r}\otimes\pi_{-\lambda r}\exp\big(p\alpha\lambda\bar{\mathbb{P}}_{\cdot,\cdot}\big)\\
&\qquad+\tfrac{1}{q}\log\pi_{-\lambda r}\otimes\pi_{-\lambda r}\exp\Big\{\tfrac{q\lambda\lambda''(1+\alpha^2)}{N}\tilde{\pi}\big(\bar{\bar{\mathbb{P}}}_{f_1,\cdot}+\bar{\bar{\mathbb{P}}}_{f_2,\cdot}\big)\Big\}\\
&\qquad-2\log\pi_{-\lambda r}\exp\big(-\tfrac{\lambda\lambda'}{N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)+\big(\tfrac{\lambda}{\lambda'}+\tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1})\\
&\quad\le\tfrac{1}{p}\log\pi_{-\lambda r}\otimes\pi_{-\lambda r}\exp\big(p\alpha\lambda\bar{\mathbb{P}}_{\cdot,\cdot}\big)\\
&\qquad+\tfrac{2}{q}\log\pi_{-\lambda r}\exp\big(\tfrac{q\lambda\lambda''(1+\alpha^2)}{N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)-\tfrac{2}{q}\log\pi_{-\lambda r}\exp\big(-\tfrac{q\lambda\lambda'}{N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)\\
&\qquad+\big(\tfrac{\lambda}{\lambda'}+\tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1})\\
&\quad\le\tfrac{1}{p}\log\pi_{-\lambda r}\otimes\pi_{-\lambda r}\exp\big(p\alpha\lambda\bar{\mathbb{P}}_{\cdot,\cdot}\big)\\
&\qquad+\tfrac{2}{q}\log\pi_{-\lambda r}\exp\Big\{\tfrac{q\lambda}{N}\big[\lambda'+\lambda''(1+\alpha^2)\big]\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big\}+\big(\tfrac{\lambda}{\lambda'}+\tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1}).
\end{aligned}$$

Now from Inequality (9.1), we have

$$\begin{aligned}
&\log\pi_{-\lambda r}\exp\Big\{\tfrac{q\lambda}{N}\big[\lambda'+\lambda''(1+\alpha^2)\big]\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big\}\\
&\quad\le\log\pi_{-\lambda\frac{r+r'}{2}}\exp\Big\{\tfrac{q\lambda}{N}\big[\lambda'+\lambda''(1+\alpha^2)\big]\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}+\big[\tfrac{\lambda\lambda'}{2N}+\tfrac{\lambda\lambda''}{2N}\big]\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big\}\\
&\qquad+\big(\tfrac{\lambda}{\lambda'}+\tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1})\\
&\quad\le\log\pi_{-\lambda\frac{r+r'}{2}}\exp\Big\{\tfrac{(2q+1)\lambda}{2N}\big[\lambda'+\lambda''(1+\alpha^2)\big]\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big\}+\big(\tfrac{\lambda}{\lambda'}+\tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1}).
\end{aligned}$$

Taking $\tilde{\pi}=\pi_{-\lambda\frac{r+r'}{2}}$ and using Jensen's inequality, we obtain Inequality (9.2). □

Remark 9.1. Since

- we want the first term in the RHS to be more than compensated by the LHS,
- the smallest λ' we are allowed to take is λ,
- we can take $\lambda''>0$ as small as necessary (when we do not concentrate on the confidence level term),

the last assertion of Lemma 9.1 will be interesting when either

$$\left\{\begin{array}{l} q\le 2\\ \big(2+\frac{1}{q}\big)\frac{\lambda}{N}<\alpha\end{array}\right. \qquad\text{or}\qquad \left\{\begin{array}{l} q>2\\ \big(q+\frac{1}{2}\big)\frac{\lambda}{N}<\alpha\end{array}\right. .$$

So Inequality (9.2) asserts that for α large enough, with high probability, we have

$$\begin{aligned}
\log\big(\pi_{-\lambda\frac{r+r'}{2}}\otimes\pi_{-\lambda\frac{r+r'}{2}}\big)\exp\big(\alpha\lambda\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)\le\log\big(\pi_{-\lambda r}\otimes\pi_{-\lambda r}\big)\exp\big(C\alpha\lambda\bar{\mathbb{P}}_{\cdot,\cdot}\big)&\\
+\text{ confidence level term.}&
\end{aligned}$$

9.3.2. *Proof.* • Let us define for any $0 \le j \le \log N$, $\rho_j \triangleq \pi_{-\lambda_j r}$. From Lemma 4.4 and Theorem 4.1 applied to prior distributions of the form $\pi_{-\lambda_j \frac{r+r'}{2}}$, $0 \le j \le \log N$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we simultaneously have

$$\left\{\begin{array}{ll} \forall\, 0 \le j \le \log N, & K\big(\rho_j, \pi_{\lambda_j \frac{r+r'}{2}}\big) \le 2 \log \rho_j \exp\left(\frac{\lambda_j^2}{N} \rho_j \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) + L \\ \forall\, 0 \le i \ne j \le \log N, & \rho_i r' - \rho_j r' \le \rho_i r - \rho_j r + S(i \wedge j, i \vee j) \end{array}\right. ,$$

and in particular $\rho_{u(k)} r' - \rho_{u(k-1)} r' \le \rho_{u(k)} r - \rho_{u(k-1)} r + S\big(u(k-1), u(k)\big)$. Now we have $S\big(u(k), u(k-1)\big) > 0$ and $\rho_{u(k)} r - \rho_{u(k-1)} r + S\big(u(k-1), u(k)\big) \le 0$. So we obtain $\rho_{u(k)} r - \rho_{u(k-1)} r < 0$ and $\rho_{u(k)} r' - \rho_{u(k-1)} r' \le 0$.

• For any $0 \le j \le \log N$, there exists k such that $u(k) \le j$. To simplify the formulae, we will not be too careful on constants. If $j = u(k)$, then we trivially have $\rho_{u(k)} r' \le \rho_j r'$. Otherwise, by contradiction, we prove $\rho_j r - \rho_{u(k)} r + S\big(u(k), j\big) > 0$, hence

$$\begin{array}{rcl} \rho_{u(k)} r' - \rho_j r' & \le & \rho_{u(k)} r - \rho_j r + S\big(u(k), j\big) \\ & < & 3S\big(u(k), j\big) - \rho_{u(k)} r + \rho_j r \\ & \le & \frac{6\lambda_j}{N} \big(\rho_{u(k)} \otimes \rho_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{6\mathcal{C}[u(k)] + 6\mathcal{C}(j) + 9L}{\lambda_j} - \rho_{u(k)} r + \rho_j r. \end{array}$$

Let $\tilde{\mathcal{C}}(j) \triangleq \sup_{0 \le i \le j} \mathcal{C}(i)$ and $\tilde{\pi}_j = \pi_{-\lambda_j \frac{r+r'}{2}}$.

Since we have

$$\big(\rho_{u(k)} \otimes \rho_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \le \big(\rho_{u(k)} \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \big(\tilde{\pi}_j \otimes \rho_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$$

and

$$-\rho_{u(k)} r + \rho_j r = \frac{-K(\rho_{u(k)}, \rho_j) + K(\rho_{u(k)}, \pi) - K(\rho_j, \pi)}{\lambda_j} \le -\frac{K(\rho_{u(k)}, \rho_j)}{\lambda_j},$$

we obtain

$$\begin{array}{rl} \rho_{u(k)} r' - \rho_j r' & \le \frac{6\lambda_j}{N} \big(\rho_{u(k)} \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{6\lambda_j}{N} \big(\rho_j \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{12\tilde{\mathcal{C}}(j) + 9L}{\lambda_j} - \frac{K(\rho_{u(k)}, \rho_j)}{\lambda_j} \\ & \le \sup_{\rho \in \mathcal{M}_+^1(\mathcal{F})} \left\{ \frac{6\lambda_j}{N} \big(\rho \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} - \frac{K(\rho, \rho_j)}{\lambda_j} \right\} + \frac{6\lambda_j}{N} \big(\rho_j \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \\ & \qquad + \frac{12\tilde{\mathcal{C}}(j) + 9L}{\lambda_j}. \end{array}$$

By Jensen's inequality, we get

$$\rho_{u(k)} r' - \rho_j r' \le \frac{2}{\lambda_j} \log \rho_j \exp\left(\frac{6\lambda_j^2}{N} \tilde{\pi}_j \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) + \frac{12\tilde{\mathcal{C}}(j) + 9L}{\lambda_j}.$$

Now, from the inequality $\rho_i \bar{\bar{\mathbb{P}}}_{\cdot,f} \le \big(\rho_i \otimes \tilde{\pi}_i\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,f}$ which holds for any function $f \in \mathcal{F}$ and using once more Jensen's inequality, we have

$$\mathcal{C}(i) \le 2 \log \rho_i \exp\left(\frac{\lambda_i^2}{N} \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) \le \frac{1}{3} \log \rho_i \exp\left(\frac{6\lambda_i^2}{N} \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right).$$

We obtain

$$\rho_{u(k)} r' - \rho_j r' \le \frac{6}{\lambda_j} \sup_{0 \le i \le j} \left\{ \log \rho_i \exp\left(\frac{6\lambda_i^2}{N} \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) \right\} + \frac{9L}{\lambda_j}.$$

Let $\rho_i' \triangleq \pi_{-\lambda_i r'}$. It remains to prove that the quantity $\log \rho_i \exp\left(\frac{6\lambda_i^2}{N} \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right)$ behaves like the quantity $\log \rho_i' \exp\left(\frac{C\lambda_i^2}{N} \rho_i' \bar{\mathbb{P}}'_{\cdot,\cdot}\right)$ for an appropriate constant C. To simplify, let us forget the index "i" for a while. From Inequality (9.1) with $\big(\alpha, \lambda', \lambda''\big) = \big(\frac{6\lambda^2}{N}, \lambda, \lambda\big)$, we have

$$\log \pi_{-\lambda r} \exp\left(\frac{6\lambda^2}{N} \pi_{-\lambda \frac{r+r'}{2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) \le \log \pi_{-\lambda \frac{r+r'}{2}} \exp\left(\frac{7\lambda^2}{N} \pi_{-\lambda \frac{r+r'}{2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) + 2L.$$

From Inequality (9.2) with $\big(\alpha,\lambda',\lambda'',p,q\big)=\Big(\frac{7\lambda}{N},\lambda,\frac{\lambda}{1+(\frac{7\lambda}{N})^2},\frac{3}{2},3\Big)$, we have

$$\begin{aligned}\log \pi_{-\lambda\frac{r+r'}{2}}\otimes\pi_{-\lambda\frac{r+r'}{2}}\exp\Big(\tfrac{7\lambda^2}{N}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)&\\ \le \tfrac{2}{3}\log\pi_{-\lambda\frac{r+r'}{2}}\otimes\pi_{-\lambda\frac{r+r'}{2}}\exp\Big(\tfrac{7\lambda^2}{N}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)&\\ +\tfrac{2}{3}\log\pi_{-\lambda r'}\otimes\pi_{-\lambda r'}\exp\Big(\tfrac{21\lambda^2}{2N}\bar{\mathbb{P}}'_{\cdot,\cdot}\Big)+\tfrac{5}{3}\Big(2+\tfrac{49\lambda^2}{N^2}\Big)L,&\end{aligned}$$

hence

$$\begin{aligned}\log \pi_{-\lambda\frac{r+r'}{2}}\exp\Big(\tfrac{7\lambda^2}{N}\pi_{-\lambda\frac{r+r'}{2}}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)&\\ \le\log\pi_{-\lambda r'}\otimes\pi_{-\lambda r'}\exp\Big(\tfrac{21\lambda^2}{N}\bar{\mathbb{P}}'_{\cdot,\cdot}\Big)+5\Big(2+\tfrac{49\lambda^2}{N^2}\Big)L.&\end{aligned}$$

Therefore with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-6\frac{\epsilon}{\log^2(eN)}$, we have

$$\log\rho_i\exp\Big(\tfrac{6\lambda_i^2}{N}\tilde{\pi}_i\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)\le\log\rho_i'\otimes\rho_i'\exp\Big(\tfrac{21\lambda_i^2}{N}\bar{\mathbb{P}}'_{\cdot,\cdot}\Big)+\Big(12+\tfrac{245\lambda_i^2}{N^2}\Big)L.$$

Introducing $\mathcal{C}'(j)\triangleq\sup_{0\le i\le j}\log\rho_i'\otimes\rho_i'\exp\Big(\frac{21\lambda_i^2}{N}\bar{\mathbb{P}}'_{\cdot,\cdot}\Big)$. With $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-6|\Lambda|\frac{\epsilon}{\log^2(eN)}$, for any $0\le j\le\log N$, we have

$$\sup_{0\le i\le j}\Big\{\log\rho_i\exp\Big(\tfrac{6\lambda_i^2}{N}\tilde{\pi}_i\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)\Big\}\le\mathcal{C}'(j)+257L.$$

Therefore, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-(|\Lambda|^2+6|\Lambda|)\frac{\epsilon}{\log^2(eN)}$, we have

$$\rho_{u(K)}r'\le\rho_{u(k)}r'\le\rho_j r'+6\frac{\mathcal{C}'(j)}{\lambda_j}+1551\frac{L}{\lambda_j}.$$

To finish the proof, we use Theorem 6.3 to replace $\pi_{-\lambda_j r}r'$ with $\pi_{-\lambda_{j-1}r'}r'$. At last, by counting the number of deviation inequalities we used, we obtain that all the previous inequalities hold with probability at least $1-\frac{(|\Lambda|^2+14|\Lambda|)\epsilon}{\log^2(eN)}\ge 1-15\epsilon$. Setting $\epsilon\leftarrow 15\epsilon$, we get rid of this factor 15 by putting it in the constant of the last term of the bound.

9.4. **Proof of Theorem 3.5.** By construction, we have $r(I_k,\theta_k)<r(I_{k-1},\theta_{k-1})$. Let $\pi_0\in\mathcal{M}_+^1(\mathcal{I})$ satisfy for any $2\le h\le N-1$ and $I\in\mathcal{I}_h$

$$\pi_0(I)\ge\tfrac{(1-\alpha)\alpha^{h-2}}{N^h}.$$

Let $\ddot{\pi}:\mathcal{Z}^N\to\mathcal{M}_+^1(\mathcal{I}\times\Theta\times\mathcal{I}\times\Theta)$ be defined as

$$\ddot{\pi}(I_1,\theta_1,I_2,\theta_2)\triangleq\pi_0(I_1)\pi_0(I_2)\pi_{Z_{I_1}}(d\theta_1)\pi_{Z_{I_2}}(d\theta_2).$$

By applying the last two inequalities in Theorem 5.2, since we have

$$\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)\le\mathcal{C}(I_1,\theta_1)+\mathcal{C}(I_2,\theta_2)+\log[(1-\alpha)^{-2}\alpha^4],$$

we obtain that with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-2\epsilon$, for any $I_1,I_2\in\mathcal{I}$ and $\theta_1,\theta_2\in\Theta$, we have

$$\begin{aligned}R(I_2,\theta_2)-R(I_1,\theta_1)+r(I_1,\theta_1)-r(I_2,\theta_2)&\\ \le\sqrt{2C_{1,2}\mathbb{P}(I_1,\theta_1,I_2,\theta_2)}+\tfrac{C_{1,2}}{3}&\\ \le\sqrt{2C_{1,2}}\Big(\sqrt{\bar{\mathbb{P}}(I_1,\theta_1,I_2,\theta_2)+C_{1,2}/2}+\sqrt{C_{1,2}/2}\Big)+\tfrac{C_{1,2}}{3}&\\ \le S(I_1,\theta_1,I_2,\theta_2).&\end{aligned}$$

By definition of (I_k,θ_k), we get $R(I_k,\theta_k)\le R(I_{k-1},\theta_{k-1})$.

- The proofs are similar to the ones of Inequalities (3.1) and (3.2).

9.5. **Proof of Theorem 3.6.** • For any $0 \le i < j \le \log N$ we have $S(i,j) \ge 0$. By definition of $u(k)$, the first inequality holds.

Let $\mathcal{J} = \{0 \le j \le \log N\}$. The second inequality comes from Corollary 4.9 applied $|\mathcal{J}|^2 - |\mathcal{J}|$ times for pairs of standard Gibbs estimators $(\pi_{-\lambda_i r}, \pi_{-\lambda_j r})$ with $i \neq j$ and appropriate prior distributions, Lemma 4.10 applied $|\mathcal{J}|$ times and the definition of $u(k)$.

• We need the following technical lemma.

Lemma 9.2. *Let* $\tilde{\pi} \in \mathcal{M}^1_+(\mathcal{F})$ independent *from the data. Let* $\epsilon > 0$, $\lambda' \ge \lambda > 0$, $\lambda'' > 0$ *and* $\alpha > 0$. *Define* $a_c(\lambda) \triangleq \frac{\lambda}{N} g\big(c\frac{\lambda}{N}\big)$ *and* $\tilde{\alpha} \triangleq \alpha + a_1(\lambda') + 2(1+\alpha^2)a_{1+\alpha}(\lambda'')$. *With* $\mathbb{P}^{\otimes N}$*-probability at least* $1 - 2\epsilon$, *we have*

$$\log \pi_{-\lambda r} \exp\big(\alpha\lambda\tilde{\pi}\bar{\mathbb{P}}_{\cdot,\cdot}\big) \le \log \pi_{-\lambda R} \exp\big(\tilde{\alpha}\lambda\tilde{\pi}\mathbb{P}_{\cdot,\cdot}\big) + \big(\tfrac{1}{\lambda'} + \tfrac{1}{\lambda''}\big)\lambda \log(\epsilon^{-1}). \tag{9.3}$$

Proof. Let $\mathcal{W}'(f,Z) \triangleq \mathbb{1}_{Y \neq f(X)} - \tilde{\pi}\mathbb{1}_{Y \neq \cdot(X)}$ and $\mathcal{W}''(f,Z) \triangleq \mathbb{1}_{Y \neq f(X)} - \tilde{\pi}\mathbb{1}_{Y \neq \cdot(X)} - \alpha\tilde{\pi}\mathbb{1}_{f(X) \neq \cdot(X)}$. From Theorem 8.3, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, we have

$$\begin{aligned}
\log \pi_{-\lambda r} \exp\big(\alpha\lambda\tilde{\pi}\bar{\mathbb{P}}_{\cdot,\cdot}\big) &= \log \pi \exp\big(-\lambda\bar{\mathbb{P}}\mathcal{W}''\big) - \log \pi \exp\big(-\lambda\bar{\mathbb{P}}\mathcal{W}'\big) \\
&\le \log \pi \exp\big\{-\lambda\mathbb{P}\mathcal{W}'' + \lambda a_{1+\alpha}(\lambda'')\mathbb{P}\big(\mathcal{W}''^2\big)\big\} \\
&\qquad - \log \pi \exp\big\{-\lambda\mathbb{P}\mathcal{W}' - \lambda a_1(\lambda')\mathbb{P}\big(\mathcal{W}'^2\big)\big\} \\
&\qquad + \big(\tfrac{\lambda}{\lambda'} + \tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1}) \\
&\le \log \pi \exp\big\{-\lambda R + \alpha\lambda\tilde{\pi}\mathbb{P}_{\cdot,\cdot} + 2(1+\alpha^2)\lambda a_{1+\alpha}(\lambda'')\tilde{\pi}\mathbb{P}_{\cdot,\cdot}\big\} \\
&\qquad - \log \pi \exp\big\{-\lambda R - \lambda a_1(\lambda')\tilde{\pi}\mathbb{P}_{\cdot,\cdot}\big\} \\
&\qquad + \big(\tfrac{\lambda}{\lambda'} + \tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1}) \\
&\le \log \pi_{-\lambda R - \lambda a_1(\lambda')\tilde{\pi}\mathbb{P}_{\cdot,\cdot}} \exp\big(\lambda\tilde{\alpha}\tilde{\pi}\mathbb{P}_{\cdot,\cdot}\big) \\
&\qquad + \big(\tfrac{\lambda}{\lambda'} + \tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1}) \\
&\le \log \pi_{-\lambda R} \exp\big(\lambda\tilde{\alpha}\tilde{\pi}\mathbb{P}_{\cdot,\cdot}\big) + \big(\tfrac{\lambda}{\lambda'} + \tfrac{\lambda}{\lambda''}\big)\log(\epsilon^{-1}).
\end{aligned}$$

□

Since we will use the same ideas as in the proof of Theorem 3.4, we will just give the main lines of the proof. For any $0 \le j \le \log N$, there exists k such that $u(k) \le j$. To shorten the formulae, introduce $a_i \triangleq \bar{a}(\lambda_i)$, $b_j \triangleq \bar{b}(\lambda_j)$ and $\tilde{\pi}_i \triangleq \pi_{-\lambda_i R}$. We have

$$\begin{aligned}
\rho_{u(k)}R - \rho_j R &\le \rho_{u(k)}r - \rho_j r + S\big(u(k), j\big) \\
&\le 3S\big(u(k), j\big) - \rho_{u(k)}r + \rho_j r \\
&\le 3S\big(u(k), j\big) - \tfrac{K(\rho_{u(k)}, \rho_j)}{\lambda_j} \\
&\le 3a_j\big(\rho_j \otimes \tilde{\pi}_j\big)\bar{\mathbb{P}}_{\cdot,\cdot} + 6b_j\mathcal{C}[u(k)] + 6b_j\mathcal{C}(j) + 9b_j L \\
&\qquad\qquad + 3a_j\big(\rho_{u(k)} \otimes \tilde{\pi}_j\big)\bar{\mathbb{P}}_{\cdot,\cdot} - \tfrac{K(\rho_{u(k)}, \rho_j)}{\lambda_j} \\
&\le \tfrac{2}{\lambda_j}\log \rho_j \exp\big(3a_j\lambda_j\tilde{\pi}_j\bar{\mathbb{P}}_{\cdot,\cdot}\big) + 12b_j \sup_{0 \le i \le j}\mathcal{C}(i) + 9b_j L
\end{aligned}$$

For any $0 \le i \le \log N$, we have $0.5\frac{\lambda_i}{N} \le a_i \le 0.6\frac{\lambda_i}{N}$ and $\frac{1}{\lambda_i} \le b_i \le \frac{1.2}{\lambda_i}$. By Jensen's inequality, we get

$$\mathcal{C}(i) \le 2\log \rho_i \exp\big(\tfrac{\lambda_i^2}{N}\tilde{\pi}_i\bar{\mathbb{P}}_{\cdot,\cdot}\big) \le \tfrac{4}{3}\log \rho_i \exp\big(3a_i\lambda_i\tilde{\pi}_i\bar{\mathbb{P}}_{\cdot,\cdot}\big).$$

Therefore we have

$$\rho_{u(k)}R - \rho_j R \le \tfrac{21.2}{\lambda_j}\sup_{0 \le i \le j}\log \rho_i \exp\big(1.8\tfrac{\lambda_i^2}{N}\tilde{\pi}_i\bar{\mathbb{P}}_{\cdot,\cdot}\big) + 10.8L.$$

Then it remains to use Lemma 9.2 to convert the quantities $\log \rho_i \exp\big(1.8\frac{\lambda_i^2}{N}\tilde{\pi}_i\bar{\mathbb{P}}_{\cdot,\cdot}\big)$ into $\log \tilde{\pi}_i \exp\big(C\frac{\lambda_i^2}{N}\tilde{\pi}_i\mathbb{P}_{\cdot,\cdot}\big)$ and Theorem 6.4 to replace $\pi_{-\lambda_j r}R$ with $\pi_{-\lambda_{j-1}R}R$.

Then it remains to count the number of concentration inequalities we used, to check that with probability at least $1 - C\epsilon$, all the previous results hold.

9.6. **Proof of Lemma 4.4.** Introduce $\tilde{\rho} \triangleq \pi_{-\frac{\lambda}{2}[r+r']}$. We have

$$K(\rho,\tilde{\rho}) = K(\rho,\pi) - K(\tilde{\rho},\pi) + \tfrac{\lambda}{2}\big[\rho r + \rho r' - \tilde{\rho} r - \tilde{\rho} r'\big].$$

Now, from Theorem 4.1, for any $\xi \in]0;1[$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have $\rho r' - \tilde{\rho} r' \le \rho r - \tilde{\rho} r + \frac{\lambda}{\xi N}(\rho \otimes \tilde{\rho})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{2\xi}{\lambda} K(\rho,\tilde{\rho}) + \frac{2\xi}{\lambda}\log(\epsilon^{-1})$. We get that

$$\begin{aligned}(1-\xi)K(\rho,\tilde{\rho}) &\le K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1}) - \lambda\tilde{\rho} r + \tfrac{\lambda^2}{2\xi N}(\rho\otimes\tilde{\rho})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} - K(\tilde{\rho},\pi)\\ &\le K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1})\\ &\qquad + \sup_{\rho'\in\mathcal{M}^1_+(\mathcal{F})}\big\{ -\lambda\rho' r + \tfrac{\lambda^2}{2\xi N}(\rho'\otimes\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} - K(\rho',\pi)\big\}\\ &= K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1}) + \log\pi\exp\big\{-\lambda[r - \tfrac{\lambda}{2\xi N}\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]\big\}\\ &= K(\rho,\pi_{-\lambda r}) + \log\pi_{-\lambda r}\exp\big\{\tfrac{\lambda^2}{2\xi N}\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big\} + \xi\log(\epsilon^{-1}).\end{aligned}$$

9.7. **Proof of Theorem 4.7.** • Let $\xi \in [0;1[$. Define $\tilde{\rho} \triangleq \pi_{-\xi\lambda[r+r'+\frac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]}$. Apply Theorem 8.4 for $\mathcal{W}(f,Z) = -\mathbb{1}_{Y\neq f(X)} + \breve{\rho}\mathbb{1}_{Y\neq\cdot(X)}$ with $(\mu,\nu) = (\rho,\tilde{\rho})$ and for $\mathcal{W}(f,Z) = \mathbb{1}_{Y\neq f(X)} - \breve{\rho}\mathbb{1}_{Y\neq\cdot(X)}$ with $(\mu,\nu) = (\tilde{\rho},\tilde{\rho})$, we obtain that with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-2\epsilon$, we have

$$\rho r' - \breve{\rho} r' \le \rho r - \breve{\rho} r + \frac{2\lambda}{N}\big(\rho\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{K(\rho,\tilde{\rho}) + \log(\epsilon^{-1})}{\lambda} \tag{9.4}$$

and

$$\breve{\rho} r' - \tilde{\rho} r' \le \breve{\rho} r - \tilde{\rho} r + \frac{2\lambda}{N}\big(\tilde{\rho}\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\log(\epsilon^{-1})}{\lambda}. \tag{9.5}$$

From this last inequality, we have

$$\begin{aligned} &\log\pi\exp\big\{-\xi\lambda[r - \breve{\rho} r + r' - \breve{\rho} r' + \tfrac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]\big\}\\ =\;& -\xi\lambda\tilde{\rho}[r - \breve{\rho} r + r' - \breve{\rho} r' + \tfrac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}] - K(\tilde{\rho},\pi)\\ \le\;& -2\xi\lambda[\tilde{\rho} r - \breve{\rho} r] + \xi\log(\epsilon^{-1}) - K(\tilde{\rho},\pi)\\ \le\;& \xi\log(\epsilon^{-1}) + \log\pi\exp\big\{-2\xi\lambda[r - \breve{\rho} r]\big\}.\end{aligned} \tag{9.6}$$

Now from Inequality (9.4), we have

$$\begin{aligned}\rho r' - \breve{\rho} r' \le \rho r - \breve{\rho} r + \tfrac{2\lambda}{N}\big(\rho\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \xi\rho[r + r' + \tfrac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]&\\ + \tfrac{K(\rho,\pi) + \log\pi\exp\{-\xi\lambda[r+r'+\frac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]\} + \log(\epsilon^{-1})}{\lambda}&,\end{aligned}$$

hence

$$\begin{aligned}(1-\xi)[\rho r' - \breve{\rho} r'] &\le (1+\xi)[\rho r - \breve{\rho} r] + (1+\xi)\tfrac{2\lambda}{N}\big(\rho\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\\ &\qquad + \tfrac{K(\rho,\pi) + \log\pi\exp\{-\xi\lambda[r+r'+\frac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]\} + \log(\epsilon^{-1})}{\lambda}\\ &\le (1-\xi)[\rho r - \breve{\rho} r] + (1+\xi)\tfrac{2\lambda}{N}\big(\rho\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\\ &\qquad + \tfrac{K(\rho,\pi_{-2\xi\lambda r}) + (1+\xi)\log(\epsilon^{-1})}{\lambda},\end{aligned}$$

where, at the last step, we have injected Inequality (9.6).

• For the second inequality, we use the same ideas. Here are the main lines of the proof. From Theorem 8.6 applied to $\mathcal{W}(f,Z) = \mathbb{1}_{Y\neq f(X)} - \breve{\rho}\mathbb{1}_{Y\neq\cdot(X)}$, we have

$$\log\pi\exp\big(-\xi\lambda\big[r + r' + \tfrac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big]\big) \le \log\pi\exp\big(-2\xi\lambda r\big) + \xi\lambda\breve{\rho}(r - r') + \xi\log(\epsilon^{-1}). \tag{9.7}$$

Introduce $\tilde\rho \triangleq \pi_{-\xi\lambda[r+r'+\frac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]}$. We have successively

$$\breve\rho r' - \rho r' \leq \breve\rho r - \rho r + \tfrac{2\lambda}{N}(\rho\otimes\breve\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{K(\rho,\tilde\rho)+\log(\epsilon^{-1})}{\lambda},$$

$$\begin{aligned} \breve\rho r' - \rho r' \quad &\leq \quad \breve\rho r - \rho r + \xi\rho(r+r') + (1+\xi)\tfrac{2\lambda}{N}(\rho\otimes\breve\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \\ &\qquad + \tfrac{K(\rho,\pi)+\log\pi\exp\left(-\xi\lambda[r+r'+\frac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]\right)+\log(\epsilon^{-1})}{\lambda} \\ &\leq \quad \breve\rho r - \rho r + \xi\bigl(\rho r + \rho r' + \breve\rho r - \breve\rho r'\bigr) + (1+\xi)\tfrac{2\lambda}{N}(\rho\otimes\breve\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \\ &\qquad + \tfrac{K(\rho,\pi)+\log\pi\exp(-2\xi\lambda r)+(1+\xi)\log(\epsilon^{-1})}{\lambda}, \end{aligned}$$

$$\begin{aligned} (1+\xi)(\breve\rho r' - \rho r') \leq (1+\xi)(\breve\rho r - \rho r) &+ (1+\xi)\tfrac{2\lambda}{N}(\rho\otimes\breve\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \\ &+ \tfrac{K(\rho,\pi_{-2\xi\lambda r})+(1+\xi)\log(\epsilon^{-1})}{\lambda}. \end{aligned}$$

9.8. **Proof of Lemma 4.10.** A numerical studies of the function $\bar b$ shows that it decreases on $]0;x_{\min}]$ and increases on $[x_{\min},+\infty[$ with $0.82N < x_{\min} < 0.83N$. We obtain that $[\frac{2.56}{N};+\infty[\subset \bar b(]0;0.77N])$. Hence for any $\lambda\in]0;0.39\,\xi N]$, there exists $0<\lambda'\leq 0.77N$ such that $\lambda \triangleq \frac{\xi}{\bar b(\lambda')}$. Introduce $\tilde\rho \triangleq \pi_{-\lambda R}$. We have $K(\rho,\tilde\rho) = K(\rho,\pi) - K(\tilde\rho,\pi) + \lambda[\rho R - \tilde\rho R]$. Now, with $\bigl(\mathbb{P}^{\otimes N}\bigr)_*$-probability at least $1-2\epsilon$, we have $\rho R - \tilde\rho R \leq \rho r - \tilde\rho r + \bar a(\lambda')(\rho\otimes\tilde\rho)\bar{\mathbb{P}}_{\cdot,\cdot} + \bar b(\lambda')K(\rho,\tilde\rho) + \bar b(\lambda')\log(\epsilon^{-1})$. We get that

$$\begin{aligned} (1-\xi)K(\rho,\tilde\rho) \quad &\leq K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1}) - \lambda\tilde\rho r + \lambda\bar a(\lambda')(\rho\otimes\tilde\rho)\bar{\mathbb{P}}_{\cdot,\cdot} - K(\tilde\rho,\pi) \\ &\leq K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1}) \\ &\qquad + \sup_{\rho'\in\mathcal{M}^1_+(\mathcal{F})}\bigl\{-\lambda\rho' r + \lambda\bar a(\lambda')(\rho'\otimes\rho)\bar{\mathbb{P}}_{\cdot,\cdot} - K(\rho',\pi)\bigr\} \\ &= K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1}) + \log\pi\exp\bigl\{-\lambda\bigl[r - \bar a(\lambda')\rho\bar{\mathbb{P}}_{\cdot,\cdot}\bigr]\bigr\} \\ &= K(\rho,\pi_{-\lambda r}) + \log\pi_{-\lambda r}\exp\bigl\{\lambda\bar a(\lambda')\rho\bar{\mathbb{P}}_{\cdot,\cdot}\bigr\} + \xi\log(\epsilon^{-1}). \end{aligned}$$

[This upper bound can also be written $K(\rho,\pi_{-\lambda[r-\bar a(\lambda')\rho\bar{\mathbb{P}}_{\cdot,\cdot}]}) + \lambda\bar a(\lambda')(\rho\otimes\rho)\bar{\mathbb{P}}_{\cdot,\cdot} + \xi\log(\epsilon^{-1})$.] Since $0<\lambda'\leq 0.77N$, we have $\bar a(\lambda') \leq \frac{\lambda'}{N} \leq \frac{2}{N\bar b(\lambda')} \leq \frac{2\lambda}{\xi N}$.

9.9. **Proof of Theorem 6.1.** Let us apply Theorem 8.6 to the random variable $\mathcal{W} = \mathbb{1}_{Y\neq f(X)} - \breve\rho\mathbb{1}_{Y\neq\cdot(X)}$ and the exchangeable distribution $\nu = \pi_{2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W}-\frac{\lambda}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]}$. We obtain that with $\bigl(\mathbb{P}^{\otimes 2N}\bigr)_*$-probability at least $1-\epsilon$,

$$\begin{aligned} &\log\pi_{\lambda[(r+r')-\breve\rho(r+r')-\frac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]}\exp\bigl\{-2\lambda[r-\breve\rho r]\bigr\} \\ &\qquad\qquad \leq -\log\pi\exp\Bigl\{\lambda\Bigl[(r+r') - \breve\rho(r+r') - \tfrac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Bigr]\Bigr\} + \log(\epsilon^{-1}), \end{aligned}$$

hence

$$\begin{aligned} \log\pi_{-2\lambda[r-\breve\rho r]}\exp\bigl\{\lambda\bigl[(r+r') - \breve\rho(r+r') - \tfrac{2\lambda}{N}\breve\rho\mathbb{P}_{\cdot,\cdot}\bigr]\bigr\}& \\ \leq -\log\pi\exp\bigl\{-2\lambda\bigl[r-\breve\rho r\bigr]\bigr\} + \log(\epsilon^{-1}).& \end{aligned} \tag{9.8}$$

By Markov's inequality, with $\bigl(\mathbb{P}^{\otimes 2N}\bigr)_*$-probability at least $1-\epsilon$, we have

$$\begin{aligned} &\pi_{-2\lambda r}\left((r+r') - \breve\rho(r+r') > \tfrac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{-\log\pi\exp\{-2\lambda[r-\breve\rho r]\}+2\log(\epsilon^{-1})}{\lambda}\right) \\ \leq \quad &\pi_{-2\lambda r}\exp\left\{\lambda\left[(r+r') - \breve\rho(r+r') - \tfrac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{\log\pi\exp\{-2\lambda[r-\breve\rho r]\}+2\log\epsilon}{\lambda}\right]\right\} \\ = \quad &\epsilon^2\pi\exp\bigl\{-2\lambda[r-\breve\rho r]\bigr\}\pi_{-2\lambda[r-\breve\rho r]}\exp\Bigl\{\lambda\Bigl[(r+r') - \breve\rho(r+r') - \tfrac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Bigr]\Bigr\} \\ \leq \quad &\epsilon, \end{aligned}$$

where the last step uses Inequality (9.8).

9.10. **Proof of Theorem 6.2.** The proof is similar to the one of Theorem 6.1. Let us apply Theorem 8.3 to the random variable $\mathcal{W} = \mathbb{1}_{Y \neq f(X)} - \tilde{\rho}\mathbb{1}_{Y \neq \cdot(X)}$ and the probability distribution $\nu = \pi_{\lambda\left[\mathbb{P}\mathcal{W} - \frac{\lambda}{N}g\left(\frac{\lambda}{N}\right)\mathbb{P}\mathcal{W}^2\right]}$. We obtain that with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,

$$\log \pi_{\lambda\left[R - \tilde{\rho}R - \frac{\lambda}{N}g\left(\frac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot}\right]} \exp\left\{-\lambda[r - \tilde{\rho}r]\right\} \\ \leq -\log \pi \exp\left\{\lambda\left[R - \tilde{\rho}R - \frac{\lambda}{N}g\left(\frac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot}\right]\right\} + \log(\epsilon^{-1}),$$

hence

$$\log \pi_{-\lambda[r - \tilde{\rho}r]} \exp\left\{\lambda\left[R - \tilde{\rho}R - \frac{\lambda}{N}g\left(\frac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot}\right]\right\} \\ \leq -\log \pi \exp\left\{-\lambda\left[r - \tilde{\rho}r\right]\right\} + \log(\epsilon^{-1}). \tag{9.9}$$

By Markov's inequality, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$\begin{aligned} & \pi_{-\frac{r}{b(\lambda)}}\left(R - \tilde{\rho}R > \frac{\lambda}{N}g\left(\frac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot} + \frac{-\log \pi \exp\{-\lambda[r-\tilde{\rho}r]\} + 2\log(\epsilon^{-1})}{\lambda}\right) \\ \leq \; & \pi_{-\lambda r} \exp\left\{\lambda\left[R - \tilde{\rho}R - \frac{\lambda}{N}g\left(\frac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot} + \frac{\log \pi \exp\{-\lambda[r-\tilde{\rho}r]\} + 2\log\epsilon}{\lambda}\right]\right\} \\ = \; & \epsilon^2 \pi \exp\left\{-\lambda\left[r - \tilde{\rho}r\right]\right\} \pi_{-\lambda[r-\tilde{\rho}r]} \exp\left\{\lambda\left[R - \tilde{\rho}R - \frac{\lambda}{N}g\left(\frac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot}\right]\right\} \\ \leq \; & \epsilon, \end{aligned}$$

where the last step uses Inequality (9.9).

9.11. **Proof of Inequality** (6.5)**.** This is the most technical proof. The basic idea of the proof is that to go from quantities depending on the first sample to quantities depending on the second sample, it suffices to know how to go from first sample quantities to exchangeable quantities. Symbolically, we have $\bar{\mathbb{P}}\mathcal{W} \to \bar{\bar{\mathbb{P}}}\mathcal{W} \to \bar{\mathbb{P}}'\mathcal{W}$.

So we write the KL-divergence as

$$K(\pi_{-\lambda r}, \pi_{-\lambda r'}) = \log \pi \exp(-\lambda r) + \log \pi \exp(-\lambda r') - 2\log \pi \exp\left(-\lambda\frac{r+r'}{2}\right) \\ +2K(\pi_{-\lambda r}, \pi_{-\lambda\frac{r+r'}{2}}).$$

Then we use the following lemma.

Lemma 9.3. *Let $\epsilon > 0$, $0 < \gamma \leq 1$ and $\lambda > 0$. Introducing $\tilde{\pi} \triangleq \pi_{-\lambda\frac{r+r'}{2}}$, we have*

- *with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-2\epsilon$,*

$$\log \pi \exp(-\lambda r) + \log \pi \exp(-\lambda r') - 2\log \pi \exp\left(-\lambda\frac{r+r'}{2}\right) \\ \leq 2\log \tilde{\pi} \exp\left(\frac{\lambda^2}{2\gamma N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) + 2\gamma\log(\epsilon^{-1}), \tag{9.10}$$

- *with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-2\epsilon$,*

$$K(\pi_{-\lambda r}, \tilde{\pi}) \leq \frac{2}{1-\gamma}\log \tilde{\pi} \exp\left(\frac{\lambda^2}{\gamma N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) + 9\frac{\gamma}{1-\gamma}\log(\epsilon^{-1}) \tag{9.11}$$

Proof. • From Theorem 8.6 applied to $\mathcal{W}(f,Z) = \mathbb{1}_{Y \neq f(X)} - \tilde{\pi}\mathbb{1}_{Y \neq \cdot(X)}$, for any $\lambda' \geq \lambda$, with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-2\epsilon$, we have

$$\begin{cases} \log \pi \exp\left\{-\lambda(r - \tilde{\pi}r)\right\} & \leq \log \pi \exp\left\{-\lambda\left(\frac{r+r'}{2} - \tilde{\pi}\frac{r+r'}{2} - \frac{\lambda'}{2N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right)\right\} \\ & \qquad +\frac{\lambda}{\lambda'}\log(\epsilon^{-1}) \\ \log \pi \exp\left\{-\lambda(r' - \tilde{\pi}r')\right\} & \leq \log \pi \exp\left\{-\lambda\left(\frac{r+r'}{2} - \tilde{\pi}\frac{r+r'}{2} - \frac{\lambda'}{2N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right)\right\} \\ & \qquad +\frac{\lambda}{\lambda'}\log(\epsilon^{-1}) \end{cases},$$

The first assertion then follows by taking $\lambda' = \frac{\lambda}{\gamma}$.

• To prove (9.11), we start with the empirical bound of the KL-divergence $K\big(\pi_{-\lambda r}, \pi_{-\frac{\lambda}{2}[r+r']}\big)$ given by Lemma 4.4:

$$K\big(\pi_{-\lambda r}, \pi_{-\frac{\lambda}{2}[r+r']}\big) \leq \frac{1}{1-\xi} \log \pi_{-\lambda r} \exp\left\{ \frac{\lambda^2}{2\xi N} \pi_{-\lambda r} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \right\} + \frac{\xi}{1-\xi} \log(\epsilon^{-1}).$$

Let us introduce $\bar{\rho} \triangleq \pi_{-\lambda r}$ and $\tilde{\rho} \triangleq \pi_{-\lambda \frac{r+r'}{2}}$. For any $f_1, f_2, f_3 \in \mathcal{F}$, we have $\bar{\bar{\mathbb{P}}}_{f_1,f_2} \leq \bar{\bar{\mathbb{P}}}_{f_1,f_3} + \bar{\bar{\mathbb{P}}}_{f_3,f_2}$, hence $\bar{\bar{\mathbb{P}}}_{f_1,f_2} \leq \tilde{\rho} \bar{\bar{\mathbb{P}}}_{f_1,\cdot} + \tilde{\rho} \bar{\bar{\mathbb{P}}}_{f_2,\cdot}$. We get

$$\log \bar{\rho} \exp\left\{ \frac{\lambda^2}{2\xi N} \bar{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \right\} \leq \log \bar{\rho}_{(df_1)} \exp\left\{ \frac{\lambda^2}{2\xi N} \bar{\rho}_{(df_2)} [\tilde{\rho} \bar{\bar{\mathbb{P}}}_{f_1,\cdot} + \tilde{\rho} \bar{\bar{\mathbb{P}}}_{f_2,\cdot}] \right\}.$$

By Jensen's inequality, we obtain $\log \bar{\rho} \exp\big\{ \frac{\lambda^2}{2\xi N} \bar{\rho} \bar{\mathbb{P}}_{\cdot,\cdot} \big\} \leq 2 \log \bar{\rho} \exp\big\{ \frac{\lambda^2}{2\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\}$. Introducing

$$\left\{ \begin{array}{lcl} \mathcal{L}' & \triangleq & \log \pi \exp\big\{ -\lambda[r - \tilde{\rho} r] + \frac{\lambda^2}{2\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\} \\ \mathcal{L}'' & \triangleq & \log \pi \exp\big\{ -\lambda[r - \tilde{\rho} r] \big\} \end{array} \right. ,$$

we have $\log \bar{\rho} \exp\big\{ \frac{\lambda^2}{2\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\} = \mathcal{L}' - \mathcal{L}''$. These two quantities can be bounded using Theorem 8.6 for

$$\mathcal{W}(f, Z) = \tfrac{1}{2} \big[\mathbb{1}_{Y \neq f(X)} - \tilde{\rho}_{(df')} \mathbb{1}_{Y \neq f'(X)} \big].$$

(We use here that Theorem 8.6 still holds when the quantity $\mathcal{W}(f, Z)$ depends on the data Z_1^{2N} in an exchangeable way). For any $\lambda'' \geq \lambda$ and $\lambda''' > 0$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1 - \epsilon$, we have

$$\mathcal{L}' \leq \log \pi \exp\Big\{ -\tfrac{\lambda}{2} \Big[(r + r') - \tilde{\rho}(r + r') - \tfrac{\lambda''}{N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big] + \tfrac{\lambda^2}{2\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} + \tfrac{\lambda}{\lambda''} \log(\epsilon^{-1})$$

and

$$-\mathcal{L}'' \leq -\log \pi \exp\Big\{ -\tfrac{\lambda}{2} \Big[(r + r') - \tilde{\rho}(r + r') + \tfrac{\lambda'''}{N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big] \Big\} + \tfrac{\lambda}{\lambda'''} \log(\epsilon^{-1}).$$

Choosing $\lambda'' = \lambda''' = \frac{\lambda}{2\xi}$, we obtain

$$\begin{aligned} &\log \bar{\rho} \exp\big\{ \tfrac{\lambda^2}{2\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\} \\ &\qquad \leq \log \pi_{-\frac{\lambda}{2}[(r+r') - \tilde{\rho}(r+r') + \frac{\lambda}{2\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]} \exp\big\{ \tfrac{\lambda^2}{\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\} + 4\xi \log(\epsilon^{-1}) \\ &\qquad \leq \log \pi_{-\frac{\lambda}{2}[(r+r') - \tilde{\rho}(r+r')]} \exp\big\{ \tfrac{\lambda^2}{\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\} + 4\xi \log(\epsilon^{-1}) \\ &\qquad = \log \tilde{\rho} \exp\big\{ \tfrac{\lambda^2}{\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\} + 4\xi \log(\epsilon^{-1}). \end{aligned}$$

Putting the previous results together, we get

$$\begin{aligned} K\big(\pi_{-\lambda r}, \pi_{-\frac{\lambda}{2}[r+r']}\big) &\leq \tfrac{1}{1-\xi} \log \pi_{-\lambda r} \exp\Big\{ \tfrac{\lambda^2}{2\xi N} \pi_{-\lambda r} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} + \tfrac{\xi}{1-\xi} \log(\epsilon^{-1}) \\ &\leq \tfrac{2}{1-\xi} \log \pi_{-\lambda r} \exp\Big\{ \tfrac{\lambda^2}{2\xi N} \pi_{-\lambda \frac{r+r'}{2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} + \tfrac{\xi}{1-\xi} \log(\epsilon^{-1}) \\ &\leq \tfrac{2}{1-\xi} \log \pi_{-\lambda \frac{r+r'}{2}} \exp\Big\{ \tfrac{\lambda^2}{\xi N} \pi_{-\lambda \frac{r+r'}{2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} + 9 \tfrac{\xi}{1-\xi} \log(\epsilon^{-1}). \end{aligned}$$

□

We obtain that for any $0 < \gamma < 1$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1 - 4\epsilon$,

$$\begin{aligned} K(\pi_{-\lambda r}, \pi_{-\lambda r'}) &\le 2 \log \tilde{\pi} \exp\Big(\tfrac{\lambda^2}{2\gamma N} \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + \tfrac{4}{1-\gamma} \log \tilde{\pi} \exp\Big(\tfrac{\lambda^2}{\gamma N} \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + \tfrac{20\gamma \log(\epsilon^{-1})}{1-\gamma} \\ &\le \tfrac{5-\gamma}{1-\gamma} \log \tilde{\pi} \exp\Big(\tfrac{\lambda^2}{\gamma N} \tilde{\pi} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + \tfrac{20\gamma \log(\epsilon^{-1})}{1-\gamma} \\ &\le \tfrac{1}{1-\gamma} \log \tilde{\pi} \otimes \tilde{\pi} \exp\Big(\tfrac{5\lambda^2}{\gamma N} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + \tfrac{20\gamma \log(\epsilon^{-1})}{1-\gamma} \end{aligned}$$

From Inequality (9.2) with

$$\big(\alpha, \lambda', \lambda'', p, q\big) = \Big(\tfrac{5\lambda}{\gamma N}, \tfrac{\lambda}{\gamma}, \tfrac{\lambda}{9\gamma\big(1+\frac{25\lambda^2}{\gamma^2 N^2}\big)}, \tfrac{4}{3}, 4\Big),$$

with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1 - 4\epsilon$, we have

$$\begin{aligned} \log \tilde{\pi} \otimes \tilde{\pi} \exp\big(\tfrac{5\lambda^2}{\gamma N} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) \le \tfrac{3}{2} \log \pi_{-\lambda r'} \otimes \pi_{-\lambda r'} &\exp\big(\tfrac{20\lambda^2}{3\gamma N} \bar{\mathbb{P}}'_{\cdot,\cdot}\big) \\ &+ 15\gamma\big(1 + \tfrac{25\lambda^2}{\gamma^2 N^2}\big)\log(\epsilon^{-1}) \end{aligned}$$

To conclude, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1 - 8\epsilon$, we have

$$K(\pi_{-\lambda r}, \pi_{-\lambda r'}) \le \tfrac{1}{1-\gamma} \log \pi_{-\lambda r'} \otimes \pi_{-\lambda r'} \exp\big(\tfrac{10\lambda^2}{\gamma N} \bar{\mathbb{P}}'_{\cdot,\cdot}\big) + \big(35 + \tfrac{375\lambda^2}{\gamma^2 N^2}\big) \tfrac{\gamma}{1-\gamma} \log(\epsilon^{-1}).$$

9.12. **Proof of Inequality** (6.8)**.** The proof is just slightly different from the one of Inequality (9.11). We start with the empirical bound of the KL-divergence given by Lemma 4.10. Let $\bar{\rho} \triangleq \pi_{-\lambda r}$ and $\tilde{\rho} \triangleq \pi_{-\lambda R}$. For any $\epsilon > 0$, $\xi \in]0;1[$ and $0 < \lambda \le 0.39\,\xi N$, with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1 - 2\epsilon$, we have

$$K\big(\bar{\rho}, \tilde{\rho}\big) \quad \le \quad \tfrac{1}{1-\xi} \Big[\log \bar{\rho} \exp\big\{ \tfrac{2\lambda^2}{\xi N} \bar{\rho} \bar{\mathbb{P}}_{\cdot,\cdot} \big\} + \xi \log(\epsilon^{-1}) \Big].$$

Inequality (6.8) is then a consequence of the following lemma.

Lemma 9.4. *For any* $\epsilon > 0$, $\xi \in]0;1[$ *and* $0 < \lambda \le 0.39\,\xi N$, *with* $\mathbb{P}^{\otimes N}$*-probability at least* $1 - 2\epsilon$, *we have*

$$\log \pi_{-\lambda r} \exp\left(\frac{2\lambda^2}{\xi N} \pi_{-\lambda r} \bar{\mathbb{P}}_{\cdot,\cdot} \right) \le 4 \log \pi_{-\lambda R} \exp\left(\frac{4.1\lambda^2}{\xi N} \pi_{-\lambda R} \mathbb{P}_{\cdot,\cdot} \right) + 4\xi \log(\epsilon^{-1}).$$

Proof. Let $\tilde{r}$, $\tilde{R}$, $\bar{\mathbb{P}}_{\cdot,\sim}$ and $\mathbb{P}_{\cdot,\sim}$ respectively denote $\tilde{\rho} r$, $\tilde{\rho} R$, $\tilde{\rho}_{(df')} \bar{\mathbb{P}}_{f',\cdot}$ and $\tilde{\rho}_{(df')} \mathbb{P}_{f',\cdot}$. Let $\alpha \triangleq \frac{2\lambda}{\xi N} \in]0;0.78]$. For any $f_1, f_2 \in \mathcal{F}$, we have $\bar{\mathbb{P}}_{f_1,f_2} \le \bar{\mathbb{P}}_{f_1,\sim} + \bar{\mathbb{P}}_{f_2,\sim}$. We get

$$\log \bar{\rho} \exp\big\{ \alpha\lambda \bar{\rho} \bar{\mathbb{P}}_{\cdot,\cdot} \big\} \le \log \bar{\rho}_{(df_1)} \exp\big\{ \alpha\lambda \bar{\rho}_{(df_2)} \big[\bar{\mathbb{P}}_{f_1,\sim} + \bar{\mathbb{P}}_{f_2,\sim} \big] \big\}.$$

By Jensen's inequality, we obtain $\log \bar{\rho} \exp\big(\alpha\lambda\bar{\rho}\bar{\mathbb{P}}_{\cdot,\cdot}\big) \le 2 \log \bar{\rho} \exp\big(\alpha\lambda\bar{\mathbb{P}}_{\cdot,\sim}\big)$. Now, we have $\log \bar{\rho} \exp\big(\alpha\lambda\bar{\mathbb{P}}_{\cdot,\sim}\big) = \mathcal{L}' - \mathcal{L}''$, where

$$\left\{ \begin{array}{lcl} \mathcal{L}' & \triangleq & \log \pi \exp\big(-\lambda\big[r - \tilde{r} - \alpha\bar{\mathbb{P}}_{\cdot,\sim}\big]\big) \\ \mathcal{L}'' & \triangleq & \log \pi \exp\big\{ -\lambda(r - \tilde{r}) \big\} \end{array} \right. .$$

These two quantities can be bounded using Theorem 8.3 for

$$\left\{ \begin{array}{lcl} \mathcal{W}'(f,Z) & \triangleq & \mathbb{1}_{Y \ne f(X)} - \tilde{\rho}_{(df')} \mathbb{1}_{Y \ne f'(X)} - \tilde{\rho}_{(df')} \alpha \mathbb{1}_{f(X) \ne f'(X)} \in [-(1+\alpha);1] \\ \mathcal{W}''(f,Z) & \triangleq & \mathbb{1}_{Y \ne f(X)} - \tilde{\rho}_{(df')} \mathbb{1}_{Y \ne f'(X)} \in [-1;1] \end{array} \right. .$$

Since $\mathbb{P}[(\mathcal{W}')^2] \le (1+\alpha)^2 \mathbb{P}_{\cdot,\sim}$ and $\mathbb{P}[(\mathcal{W}'')^2] \le \mathbb{P}_{\cdot,\sim}$, for any $\lambda'' \ge \lambda$ and $\lambda''' > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, we have

$$\mathcal{L}' \le \log \pi \exp\Big\{ -\lambda\big[R - \tilde{R}\big] + \lambda\big[\alpha + (1+\alpha)^2 a_{1+\alpha}(\lambda'')\big] \mathbb{P}_{\cdot,\sim} \Big\} + \tfrac{\lambda}{\lambda''} \log(\epsilon^{-1})$$

and

$$-\mathcal{L}'' \leq -\log \pi \exp\big\{ -\lambda\big[R-\tilde{R}\big] - \lambda a_1(\lambda''')\mathbb{P}_{\cdot,\sim}\big\} + \tfrac{\lambda}{\lambda'''}\log(\epsilon^{-1}).$$

Choosing $\lambda'' = \lambda''' = \frac{\lambda}{\xi}$, we obtain

$$\begin{aligned}\log\bar{\rho}\exp\big(\alpha\lambda\bar{\mathbb{P}}_{\cdot,\sim}\big) \quad &\leq \quad \log\tilde{\rho}\exp\big\{\lambda\big[\alpha+(1+\alpha)^2 a_{1+\alpha}(\lambda/\xi)\big]\mathbb{P}_{\cdot,\sim}\big\} \\ &\qquad\qquad - \log\tilde{\rho}\exp\big\{ -\lambda a_1(\lambda/\xi)\mathbb{P}_{\cdot,\sim}\big\} + 2\xi\log(\epsilon^{-1}) \\ &\leq \quad \log\tilde{\rho}\exp\big\{\lambda\big[\alpha+(1+\alpha)^2 a_{1+\alpha}(\lambda/\xi)\big]\mathbb{P}_{\cdot,\sim}\big\} \\ &\qquad\qquad + \log\tilde{\rho}\exp\big\{\lambda a_1(\lambda/\xi)\mathbb{P}_{\cdot,\sim}\big\} + 2\xi\log(\epsilon^{-1}) \\ &\leq \quad 2\log\tilde{\rho}\exp\big\{\lambda\big[\alpha+(1+\alpha)^2 a_{1+\alpha}(\lambda/\xi)\big]\mathbb{P}_{\cdot,\sim}\big\} + 2\xi\log(\epsilon^{-1}),\end{aligned}$$

which leads to the desired inequality. □

Appendix A. Optimal coupling

One drawback of the variance term $\frac{2\lambda}{N}(\rho_1\otimes\rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$ in Theorem 4.1 is to be large when ρ_1 and ρ_2 are close and not concentrated around a particular function. This problem can be solved by coupling.

Let us start with some new notations. For any p_1, p_2 in $[0;1]$, define

$$K(p_1,p_2) \triangleq p_1\log\big(\tfrac{p_1}{p_2}\big) + (1-p_1)\log\big(\tfrac{1-p_1}{1-p_2}\big)$$

the Kullback-Leibler divergence between two Bernouilli distributions of respective parameters p_1 and p_2.

Let $\pi \in \mathcal{M}_+^1(\mathcal{F})$. Introduce π_Δ the associated distribution on the diagonal of $\mathcal{F}\times\mathcal{F}$: $\pi_\Delta(df_1, df_2) \triangleq \pi(df_1)\delta_{f_1}(df_2)$, where δ_f denote the Dirac distribution on the function f. In other words, π_Δ is the distribution in $\mathcal{M}_+^1(\mathcal{F}\times\mathcal{F})$ such that $\pi_\Delta(f_1 = f_2) = 1$ and $\pi_\Delta(df_1) = \pi(df_1)$.

Let ρ_1 and ρ_2 be absolutely continuous distributions wrt π. Define the positive measures $\rho_1\wedge\rho_2 \triangleq \big(\frac{\rho_1}{\pi}\wedge\frac{\rho_2}{\pi}\big)\cdot\pi$, $|\rho_1-\rho_2| \triangleq \big|\frac{\rho_1}{\pi}-\frac{\rho_2}{\pi}\big|\cdot\pi$ and $(\rho_1-\rho_2)_+ \triangleq \big(\frac{\rho_1}{\pi}-\frac{\rho_2}{\pi}\big)_+\cdot\pi$. Let $m_{1,2} \triangleq (\rho_2-\rho_1)_+(\mathcal{F})$. Then the positive measures $\frac{(\rho_2-\rho_1)_+}{m_{1,2}}$, $\frac{(\rho_1-\rho_2)_+}{m_{1,2}}$ and $\frac{\rho_1\wedge\rho_2}{1-m_{1,2}}$ are probability distributions. An optimal coupling of ρ_1 and ρ_2 is defined as

$$\rho_1\odot\rho_2 \triangleq (1-m_{1,2})\bigg(\frac{\rho_1\wedge\rho_2}{1-m_{1,2}}\bigg)_\Delta + m_{1,2}\bigg(\frac{(\rho_1-\rho_2)_+}{m_{1,2}}\bigg)\otimes\bigg(\frac{(\rho_2-\rho_1)_+}{m_{1,2}}\bigg).$$

We obtain

Theorem A.1. *For any $\epsilon > 0$, $\lambda > 0$ and $\pi_{1,2} \in \mathcal{M}_+^1(\mathcal{F}\times\mathcal{F})$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have for any $\rho_1,\rho_2 \in \mathcal{M}_+^1(\mathcal{F})$*

$$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \frac{2\lambda}{N}(\rho_1\odot\rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\dot{\mathcal{K}}_{1,2}}{\lambda}$$

where $\dot{\mathcal{K}}_{1,2} \triangleq K(\rho_1\odot\rho_2, \pi_{1,2}) + \log(\epsilon^{-1})$.

Proof. It suffices to modify the proof of Theorem 4.1 by taking $(\mu,\nu) = (\rho_1\odot\rho_2, \pi_{1,2})$ instead of $(\mu,\nu) = (\rho_1\otimes\rho_2, \pi_1\otimes\pi_2)$. Then it remains to notice that the marginals of $\rho_1\odot\rho_2$ are respectively ρ_1 and ρ_2. □

Corollary A.2. *For any $\lambda > 0$, $\pi \in \mathcal{M}_+^1(\mathcal{F})$, $\epsilon > 0$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have for any $\rho_1,\rho_2 \in \mathcal{M}_+^1(\mathcal{F})$*

$$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \frac{2\lambda}{N}(\rho_1\odot\rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\dot{\mathcal{K}}}{\lambda}$$

where

$$\begin{aligned}\dot{\mathcal{K}} \quad &\triangleq \quad m_{1,2}K\big(\tfrac{(\rho_1-\rho_2)_+}{m_{1,2}},\pi\big)+m_{1,2}K\big(\tfrac{(\rho_2-\rho_1)_+}{m_{1,2}},\pi\big)\\ &\qquad +(1-m_{1,2})K\big(\tfrac{\rho_1\wedge\rho_2}{1-m_{1,2}},\pi\big)+K(m_{1,2},\tfrac12)+\log(\epsilon^{-1}).\end{aligned}$$

Proof. Take $\pi_{1,2}\triangleq\frac12\pi\otimes\pi+\frac12\pi_\Delta$ in the previous theorem. The result follows from (A.1)

$$\begin{aligned}K(\rho_1\odot\rho_2,\pi_{1,2}) \quad &= \quad \rho_1\odot\rho_2\log\tfrac{\rho_1\odot\rho_2}{\pi_{1,2}}\\ &\leq \quad \big(\tfrac{\rho_1}{\pi}\wedge\tfrac{\rho_2}{\pi}\big)\cdot\pi\log\big(2\tfrac{\rho_1}{\pi}\wedge\tfrac{\rho_2}{\pi}\big)+m_{1,2}\big(\tfrac{(\rho_1-\rho_2)_+}{m_{1,2}}\big)\otimes\big(\tfrac{(\rho_2-\rho_1)_+}{m_{1,2}}\big)\\ &\qquad \log\left(2\frac{\big(\frac{\rho_1}{\pi}(f_1)-\frac{\rho_2}{\pi}(f_1)\big)_+}{m_{1,2}}\frac{\big(\frac{\rho_2}{\pi}(f_2)-\frac{\rho_1}{\pi}(f_2)\big)_+}{m_{1,2}}\right)\\ &= \quad K(m_{1,2},\tfrac12)+(1-m_{1,2})K\big(\tfrac{\rho_1\wedge\rho_2}{1-m_{1,2}},\pi\big)\\ &\qquad +m_{1,2}K\big(\tfrac{(\rho_1-\rho_2)_+}{m_{1,2}},\pi\big)+m_{1,2}K\big(\tfrac{(\rho_2-\rho_1)_+}{m_{1,2}},\pi\big).\end{aligned}$$

Inequality (A.1) is an equality when π_Δ and $\pi\otimes\pi$ are mutually singular (i.e. π diffuse). □

The interest of coupling is to reduce significantly the variance term involving $\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}$ at least when ρ_1 and ρ_2 are close to each other. From the last corollary, we see the impact in the Kullback-Leibler term.

In the worst case (i.e. when ρ_1 and ρ_2 are mutually singular, equivalently when $\rho_1\odot\rho_2=\rho_1\otimes\rho_2$), we just lose an additive term $\log 2$ in the Kullback-Leibler term since we get $\dot{\mathcal{K}}=K(\rho_1\otimes\rho_2,\pi\otimes\pi)+\log 2$ in this case. On the contrary, when $\rho_1=\rho_2=\rho$, we have $\dot{\mathcal{K}}=K(\rho,\pi)+\log 2=\frac12K(\rho_1\otimes\rho_2,\pi\otimes\pi)+\log 2$. Naturally, $\rho_1=\rho_2$ is not an interesting case since Inequality (4.1) is useless in this situation. But to look at the Kullback-Leibler term when $\rho_1=\rho_2$ gives an idea of how it behaves when ρ_2 is close to ρ_1.

To conclude this section, we see that the basic Inequality (4.1) can be improved to deal with close posterior distributions which are not concentrated[19]. However, the inequalities become less readable and less tractable both for theory and practice.

Appendix B. Optimality of Algorithm 3.2 under (CM) assumptions

We recall that C denotes a positive constant which value may differ from line to line. By using the same ideas as in the proofs of Lemmas 9.1 and 9.2, we can upper bound $-\log\pi\exp\big\{-\lambda[r'-r'(\tilde f)]\big\}$ and $\log\pi_{-\lambda r'}\exp\big(C\frac{\lambda^2}{N}\pi_{-\lambda r'}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)$ by similar theoretical quantities. Indeed, schematically, by intensively using Theorems 8.6 and 8.3 and Jensen's inequality, with $\mathbb{P}^{\otimes 2N}$-high probability, for any $\lambda\leq cN$ for a small enough universal constant $c>0$ and any prior distribution π independent from the data, we have

$$\begin{aligned}-\log\pi\exp\big\{-\lambda[r'-r'(\tilde f)]\big\} \quad &\leq \quad -\log\pi\exp\big(-\lambda[R-R(\tilde f)]-C\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde f}\big)+\dots\\ &\leq \quad -\log\pi\exp\big\{-\lambda[R-R(\tilde f)]\big\}\\ &\qquad +\log\pi_{-\lambda R}\exp\big(C\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde f}\big)+\dots\end{aligned}$$

[19] When they are concentrated and close, the variance term is already small.

and

$$\begin{aligned}
\log \pi_{-\lambda r'} \exp\big(C\tfrac{\lambda^2}{N}\pi_{-\lambda r'}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) &\leq \log \pi_{-\lambda r'} \exp\big(C\tfrac{\lambda^2}{N}\pi_{-\lambda \frac{r+r'}{2}}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) + \dots \\
&\leq \log \pi_{-\lambda \frac{r+r'}{2}} \exp\big(C\tfrac{\lambda^2}{N}\pi_{-\lambda \frac{r+r'}{2}}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) + \dots \\
&\leq \log \pi_{-\lambda r} \exp\big(C\tfrac{\lambda^2}{N}\pi_{-\lambda r}\bar{\mathbb{P}}_{\cdot,\cdot}\big) + \dots \\
&\leq \log \pi_{-\lambda R} \exp\big(C\tfrac{\lambda^2}{N}\pi_{-\lambda R}\mathbb{P}_{\cdot,\cdot}\big) + \dots \\
&\leq \log \pi_{-\lambda R} \exp\big(C\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\big) + \dots
\end{aligned}$$

Let

$$\mathbb{G}^{\text{th}}_C(\lambda) \triangleq -\tfrac{1}{\lambda}\log \pi \exp\big\{-\lambda[R - R(\tilde{f})]\big\} + \tfrac{1}{\lambda}\log \pi_{-\lambda R}\exp\Big(C\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\Big) + \frac{C\log[\log(eN)\epsilon^{-1}]}{\lambda}$$

and $\Lambda \triangleq \big\{\sqrt{N}e^{\frac{j}{2}}; 0 \leq j \leq \log N\big\}$. The precise result is that for any $\epsilon > 0$ and $\lambda \leq cN$, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have $\mathbb{G}(\lambda) - r'(\tilde{f}) \leq \mathbb{G}^{\text{th}}_C(\lambda)$, hence with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, for any $\lambda \in \Lambda$, we have

$$\mathbb{G}(\lambda) - r'(\tilde{f}) \leq \mathbb{G}^{\text{th}}_C(\lambda).$$

Then it remains to check that for a parameter $\lambda \in \Lambda$ close to $N^{\frac{\kappa}{2\kappa-1+q}}$ and a prior distribution satisfying[20]

$$\pi\Big(\mathbb{P}_{\cdot,\tilde{f}} \leq \check{C}_1 N^{-\frac{1}{2\kappa-1+q}}\Big) \geq \exp\Big(-\check{C}_2 N^{-\frac{q}{2\kappa-1+q}}\Big),$$

we have $\mathbb{G}^{\text{th}}_C(\lambda) \leq C\log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}$.

References

1. J.-Y. Audibert, *Classification using Gibbs estimators under complexity and margin assumptions*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2004.
2. P.L. Bartlett, O. Bousquet, and S. Mendelson, *Localized rademacher complexities*, Proceedings of the 15th annual conference on Computational Learning Theory, Lecture Notes in Computer Science (K. Kivinen, ed.), vol. 2375, Springer-Verlag, 2002.
3. S. Boucheron, G. Lugosi, and P. Massart, *A sharp concentration inequality with applications*, Random Struct. Algorithms (2000), 277–292.
4. O. Bousquet, *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*, Ph.D. thesis, Department of Applied Mathematics, Ecole Polytechnique, 2002.
5. O. Catoni, *Statistical learning theory and stochastic optimization,* Lecture notes, Saint-Flour summer school on Probability Theory, 2001, Springer, to be published.
6. ———, *Localized empirical complexity bounds and randomized estimators*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2003.
7. ———, *A PAC-Bayesian approach to adaptive classification*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2003.
8. L. Devroye and G. Lugosi, *Lower bounds in pattern recognition and learning*, Pattern recognition **28** (1995), 1011–1018.
9. M. Kohler, *Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression*, J. Stat. Plann. Inference **89** (2000), no. 1-2, 1–23.
10. L. Györfi L. Devroye and G. Lugosi, *A probabilistic theory of pattern recognition*, Springer-Verlag, 1996.
11. W.S Lee, P.L. Bartlett, and R.C. Williamson, *Efficient agnostic learning of neural network with bounded fan-in*, IEEE Trans. Inform. Theory **42** (1996), no. 6, 2118–2132.

[20] From the complexity assumption in (CM), such a prior distribution exists. We can even choose it independently from the parameter κ so that the Gibbs classifier proposed in Algorithms 3.2, 3.3 and 3.6 are adaptive wrt the margin parameter (see [1] for more details).

12. N. Littlestone and M. Warmuth, *Relating data compression and learnability*, Technical report, University of California, Santa Cruz, 1986.
13. G. Lugosi, *Concentration-of-measure inequalities*, 2003, Lecture notes, Machine Learning Summer School, Canberra.
14. E. Mammen and A.B. Tsybakov, *Smooth discrimination analysis*, Ann. Stat. **27** (1999), 1808–1829.
15. P. Massart and E. Nédélec, *Risk bounds for statistical learning*, Available from http://www.math.u-psud.fr/∼massart/margin.pdf, 2003.
16. D. A. McAllester, *PAC-Bayesian model averaging*, Proceedings of the 12th annual conference on Computational Learning Theory, Morgan Kaufmann, 1999.
17. O.Bousquet, V. Koltchinskii, and D. Panchenko, *Some local measures of complexity of convex hulls and generalization bounds*, Proceedings of the 15th annual conference on Computational Learning Theory, Lecture Notes in Computer Science (K. Kivinen, ed.), vol. 2375, Springer-Verlag, 2002.
18. M. Seeger, *PAC-Bayesian generalization error bounds for gaussian process classification*, Informatics report series EDI-INF-RR-0094, Division of Informatics, University of Edinburgh, 2002.
19. A.B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Ann. Stat. **32** (2004), no. 1.
20. A. van der Vaart and J. Wellner, *Weak convergence and empirical processes with application to statistics*, John Wiley & Sons, New York, 1996.
21. V. Vapnik and A. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory Probab. Appl. **16** (1971), 264–280.
22. ______, *Theory of pattern recognition*, **16** (1974), 264–280, Nauka, Moscow (in Russian).
23. ______, *Theorie der zeichenerkennung*, (1979), Berlin, (german translation of the previous paper).

CLASSIFICATION UNDER POLYNOMIAL ENTROPY AND MARGIN ASSUMPTIONS AND RANDOMIZED ESTIMATORS

J.-Y. AUDIBERT

Université Paris VI and CREST

Abstract. The aim of this paper is two-fold. First we want to develop the PAC-Bayesian point of view [13, 3, 4, 1] and show how the efficiency of a Gibbs estimator relies on the weights given by the prior distribution to the balls centered at the best function in the model and associated with the pseudo-distance $(f_1, f_2) \mapsto \mathbb{P}[f_1(X) \neq f_2(X)]$.

Secondly, we show how to recover and improve results under empirical and non empirical polynomial entropy assumptions and Tsybakov's margin assumption. We also study the links between empirical and non empirical nets and give an observable version of the integral entropy [6, 9, 14].

Contents

2000 *Mathematics Subject Classification.* Primary 62H30, Secondary 68Q32.

Key words and phrases. Gibbs classifiers, entropy assumptions, margin assumptions, PAC-Bayesian bounds, chaining, oracle inequalities, VC theory.

I would like to thank very deeply Professor Olivier Catoni, my PhD advisor, for the numerous hours he spent with me. I am also very grateful to Olivier Bousquet and Professor Alexandre Tsybakov for the numerous discussions which have fed this work. I would also like to thank Professor Bernhard Schölkopf who invited me twice at the Max Planck Institute of Tuebingen. Meeting people there has been a key point for widening my scope of interest during my PhD.

1. Setup and notations

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 called the pattern space, $\mathcal{Y} = \{1, \dots, |\mathcal{Y}|\}$ is the (finite) label space and $\mathcal{B}_{\mathcal{Y}}$ is the sigma algebra of all subsets of $\mathcal{Y}$. Let $\mathbb{P}(dY|X)$ denote a regular version of the conditional probabilities (which we will use in the following without further mention).

Let $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ denote the set of all measurable functions mapping $\mathcal{X}$ into $\mathcal{Y}$. The aim of a classification procedure is to build a function $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ from the learning sample such that $f(X)$ predicts the label Y associated with X. The quality of the prediction is measured by the expected risk

$$R(f) \triangleq \mathbb{P}[Y \neq f(X)].$$

A function $f^*_{\mathbb{P}}$ such that for any $x \in \mathcal{X}$,

$$f^*_{\mathbb{P}}(x) \in \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}(Y = y | X = x)$$

minimizes the expected risk. This function is not necessarily unique. We assume that there exists a measurable one. We will once for all fix it, refer to it as the Bayes classifier and often denote it f^* to shorten. Since we have no prior information about the distribution $\mathbb{P}$ of (X, Y), this classifier is unknown.

Since there is generally no measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$ such that

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

we have to work with a prescribed set of classification functions $\mathcal{F}$, called the model. This set is just some subset of the set of all measurable functions $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. Let us denote $\tilde{f}$ the best function in the model, i.e. a function minimizing the expected risk:

$$\tilde{f} \in \underset{\mathcal{F}}{\operatorname{argmin}}\, R.$$

For sake of simplicity, we assume that it exists[1]. Let

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

be the empirical distribution. The empirical risk

$$r(f) \triangleq \bar{\mathbb{P}}[Y \neq f(X)]$$

gives an estimate of the expected risk : from the law of large numbers, for any measurable function, it tends to the expected risk almost surely. An estimator which minimizes the empirical risk

$$\hat{f}_{\mathrm{ERM}} \in \underset{\mathcal{F}}{\operatorname{argmin}}\, r$$

is called an ERM[2]-classifier. The regression function will be denoted

$$\eta^*(x) \triangleq \mathbb{P}(Y | X = x).$$

In the binary classification setting ($\mathcal{Y} = \{0; 1\}$), we have $\eta^*(x) = \mathbb{P}(Y = 1 | X = x)$.

Since we will study randomized estimators, we assume that we have a $\sigma-$algebra $\mathcal{T}$ such that $(\mathcal{F}, \mathcal{T})$ is a measurable space containing the sets $\{f\}$ for any $f \in \mathcal{F}$ and such that the function

$$\begin{array}{ccc} \mathcal{F} \times \mathcal{X} & \to & \mathcal{Y} \\ (f, x) & \mapsto & f(x) \end{array}$$

is measurable. A randomized estimator consists in drawing a function in $\mathcal{F}$ according to some random distribution $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}^1_+(\mathcal{F})$, where $\mathcal{M}^1_+(\mathcal{F})$ is the set of probability distributions on the measurable space $(\mathcal{F}, \mathcal{T})$.

[1] Otherwise we would have to introduce some small positive real β and consider $\tilde{f}$ as an estimator minimizing the expected risk up to β. This real β would then appear in all the equations related to this function and make things needlessly messy.

[2] ERM = Empirical Risk Minimization

To shorten, we will use μh to denote the expectation of the random variable h under the probability distribution μ: $\mu h \triangleq \int h(x) d\mu(x)$. The Kullback-Leibler divergence between two probability distributions is defined as $K(\mu,\nu) = \mu \log \frac{d\mu}{d\nu}$ when μ is absolutely continuous with respect to ν and $K(\mu,\nu) = +\infty$ otherwise.

The symbol C will denote a positive universal constant whose value may differ from line to line whereas the symbol $\breve{C}$ will denote a positive constant whose value depends on other constants and may also differ from line to line.

We define

$$\pi_h \triangleq \frac{\exp(h)}{\pi \exp(h)} \cdot \pi$$

for any measurable real function h such that $\exp(h)$ is π-integrable. The randomized estimators associated with the posterior distributions π_{-Cr} will be called the standard Gibbs estimators with temperature $\frac{1}{C}$.

1.1. **Measurability.** Finally, to circumvent some measurability problems, we will consider inner and outer expectations. Let $(A, \mathcal{A}, \mu)$ be a measure space and $\mathcal{C}(A;\mathbb{R})$ be the class of real measurable functions. For any (measurable or not) function f, its inner and outer expectation wrt μ are respectively $\mu_*(h) \triangleq \sup\big\{\mu(g) : g \in \mathcal{C}(A;\mathbb{R}), g \leq h\big\}$ and $\mu^*(h) \triangleq \inf\big\{\mu(g) : g \in \mathcal{C}(A;\mathbb{R}), g \geq h\big\}$. Naturally, for any set $B \subset A$, $\mu_*(B)$ and $\mu^*(B)$ are defined by $\mu_*(B) = \mu_*(\mathbb{1}_B)$ and $\mu^*(B) = \mu^*(\mathbb{1}_B)$. Note that μ_* and μ^* are not measures but satisfy $\mu^*(B) + \mu_*(B^c) = 1$ and $\mu^*(B_1 \cup B_2) \leq \mu^*(B_1) + \mu^*(B_2)$. Besides, if $\mu^*(h) < +\infty$, then there exists a random variable h^* such that $\mu^*(h) = \mu(h^*)$. For more details on properties of inner and outer expectations, see [17].

1.2. **Covering, packing and bracketing nets and entropies.** Let $\mathbb{Q}$ denote a probability distribution on the measurable space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$. The mapping $\mathbb{Q}_{\cdot,\cdot}$ from $\mathcal{F} \times \mathcal{F}$ into $\mathbb{R}_+$ defined as

$$\mathbb{Q}_{f_1,f_2} \triangleq \mathbb{Q}[f_1(X) \neq f_2(X)] \qquad \text{for any } f_1, f_2 \in \mathcal{F}$$

is a pseudo-distance. For any $u \geq 0$, a set of measurable functions $\mathcal{G} \subset \mathcal{F}(\mathcal{X},\mathcal{Y})$ such that

$$\sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{G}} \mathbb{Q}_{f,g} \leq u$$

is called a $u-$covering net of the set $\mathcal{F}$ wrt the pseudo-distance $\mathbb{Q}$.

The log-cardinal $H(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot})$ of the smallest $u-$covering net (possibly infinite) is called the $u-$covering entropy. A $u-$covering net with log-cardinal equal to $H(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot})$ is called a minimal $u-$covering net[3].

In bracketing nets, we require in addition that any function in $\mathcal{F}$ can be encapsulated by two functions of the net. Specifically, for any $u \geq 0$, a set of measurable functions $\mathcal{G} \subset \mathcal{F}(\mathcal{X},\mathcal{Y})$ such that for any function $f \in \mathcal{F}$, there exist $f_L, f_U \in \mathcal{G}$ satisfying $f_L \leq f \leq f_U$ and $\mathbb{Q}_{f_L,f_U} \leq u$, is called a $u-$bracketing net of the set $\mathcal{F}$ wrt the pseudo-distance $\mathbb{Q}$. The log-cardinal $H^{[]}(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot})$ of the smallest $u-$bracketing net (possibly infinite) is called the $u-$bracketing entropy. A $u-$bracketing net with log-cardinal equal to $H^{[]}(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot})$ is called a minimal $u-$bracketing net.

[3] Here the functions in the net can be taken outside $\mathcal{F}$. This is not so important since it is well-known that a $2u-$covering net with functions in $\mathcal{F}$ can be constructed from any $u-$covering net.

Packing nets are covering nets such that for any functions f_1, f_2 in the net, we have $\mathbb{Q}_{f_1,f_2} > u$. The packing entropy $H_{\mathrm{p}}(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot})$ is the log-cardinal of a minimal packing net.

We have $H(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot}) \leq H_{\mathrm{p}}(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot}) \leq H(\frac{u}{2}, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot})$. Any $u-$bracketing net is a $u-$covering net. The converse is false since it is easy to find a set $\mathcal{F}$ with finite $u-$covering entropy and infinite $u-$bracketing entropy.

Finally, we will say that a family of u_N-nets, $N \in \mathbb{N}$, is almost minimal when the log-cardinal of the size of the u_N-net has the same order as the $(u_N, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-entropy.

The paper is organized as follows. Section 2 recalls some PAC-Bayesian concentration inequalities which are extracted from [1]. In Section 3, we assume that we have Tsybakov's margin assumption and that the $\mathbb{P}_{\cdot,\cdot}$-entropies are polynomial. In this setting, we study the convergence rate of standard Gibbs estimators and classifiers minimizing the empirical risk on $\mathbb{P}_{\cdot,\cdot}$-covering nets. In particular, we stresses on the influence of the chaining trick and the differences between bracketing and covering entropy assumptions. Section 4 tries to answer the questions: what happens when we relieve the polynomial $\mathbb{P}_{\cdot,\cdot}$-entropy assumption? Can we give an empirical equivalent (i.e. with $\bar{\mathbb{P}}_{\cdot,\cdot}$-entropies) of the previous results? Section 5 gives a version of Assouad's lemma dedicated to classification. The proofs are gathered in Section 6.

2. Known PAC-Bayesian bounds

In this section, we recall some results of [1] which will be useful in this paper.

Theorem 2.1. *Let* $g(u) \triangleq \frac{\exp(u)-1-u}{u^2}$ *for any* $u > 0$. *For any* $\lambda > 0$, $\epsilon > 0$ *and* $\pi_1, \pi_2 \in \mathcal{M}_+^1(\mathcal{F})$, *with* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$, *for any* $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$, *we have*

(2.1)
$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \leq \tfrac{\lambda}{N} g\big(\tfrac{\lambda}{N}\big)(\rho_1 \otimes \rho_2)\mathbb{P}_{\cdot,\cdot} + \tfrac{K(\rho_1,\pi_1)+K(\rho_2,\pi_2)+\log(\epsilon^{-1})}{\lambda}$$

As a consequence, with $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$, *for any* $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$,

(2.2)
$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \leq \min_{\lambda \in [\sqrt{N};N]} \Big\{ 0.8 \tfrac{\lambda}{N}(\rho_1 \otimes \rho_2)\mathbb{P}_{\cdot,\cdot} + 1.7 \tfrac{K(\rho_1,\pi_1)+K(\rho_2,\pi_2)+\log[\log(eN)\epsilon^{-1}]}{\lambda} \Big\}.$$

Besides, let $\mathcal{S}_1$ *and* $\mathcal{S}_2$ *be finite subsets of* $\mathcal{F}$, *with* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$, *for any* $(f_1, f_2) \in \mathcal{S}_1 \times \mathcal{S}_2$, *we have*

$$R(f_2) - R(f_1) + r(f_1) - r(f_2) \leq \sqrt{\tfrac{2\log(|\mathcal{S}_1||\mathcal{S}_2|\epsilon^{-1})\mathbb{P}_{f_1,f_2}}{N}} + \tfrac{\log(|\mathcal{S}_1||\mathcal{S}_2|\epsilon^{-1})}{3N}. \tag{2.3}$$

Proof. The first part comes from Theorem 4.8 in [1]. Then the second part is obtained by a union bound on the set of parameters $\Lambda \triangleq \big\{\sqrt{N}e^{k/2}; 0 \leq k \leq \log N\big\}$ (see Section 4.2 in [1] for details). The third part comes from Theorem 8.1 in [1] applied to $\mathcal{W}\big[(f_1, f_2), Z\big] = \mathbb{1}_{Y \neq f_2(X)} - \mathbb{1}_{Y \neq f_1(X)}$ and ν equal to the uniform measure on $\mathcal{S}_1 \times \mathcal{S}_2$. □

The following theorem ([1, Theorem 6.4]) brackets the efficiency of a standard Gibbs classifier

Theorem 2.2. *For any $\lambda > 0$ and $0 < \chi \leq 1$, we have*

$$\pi_{-(1+\chi)\lambda R} R - \tfrac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \leq \pi_{-\lambda r} R \leq \pi_{-(1-\chi)\lambda R} R + \tfrac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda},$$

and for any $\epsilon > 0$, $0 < \gamma < \frac{1}{2}$ and $0 < \lambda \leq 0.39\,\gamma N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$\begin{aligned} K(\pi_{-\lambda r}, \pi_{-\lambda R}) &\leq \tfrac{4}{1-\gamma} \log \pi_{-\lambda R} \exp\left(\tfrac{4.1\,\lambda^2}{\gamma N} \pi_{-\lambda R} \mathbb{P}_{\cdot,\cdot}\right) + \tfrac{5\gamma}{1-\gamma} \log(4\epsilon^{-1}) \\ &\leq 16 \log \pi_{-\lambda R} \exp\left(\tfrac{4.1\,\lambda^2}{\gamma N} \mathbb{P}_{\cdot,\tilde{f}}\right) + 10\gamma \log(4\epsilon^{-1}). \end{aligned} \tag{2.4}$$

3. Convergence rate of classifiers under complexity and margin assumptions

3.1. **Complexity and margin assumptions.** The following assumptions have the same form as the one used in the pioneering work of Mammen and Tsybakov ([11]). The margin assumption appears to be the key assumption to obtain fast rates of convergence (i.e. $N^{-\beta}$ with $\beta > \frac{1}{2}$).

3.1.1. *Complexity assumptions.* Let $q \geq 0$. Define

$$h_q(u) \triangleq \begin{cases} \log(eu^{-1}) & \text{when} \quad q = 0 \\ u^{-q} & \text{when} \quad q > 0 \end{cases}.$$

We will alternatively use the following complexity assumptions.
(CA1) : there exists $C' > 0$ such that the covering entropy of the model $\mathcal{F}$ for the distance $\mathbb{P}_{\cdot,\cdot}$ satisfies for any $u > 0$, $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) \leq C' h_q(u)$.
(CA2) : there exists $C' > 0$ such that the bracketing entropy of the model $\mathcal{F}$ for the distance $\mathbb{P}_{\cdot,\cdot}$ satisfies for any $u > 0$, $H^{[]}(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) \leq C' h_q(u)$.
(CA3) : there exist $C' > 0$ and $\pi \in \mathcal{M}^1_+(\mathcal{F})$ such that for any $t > 0$, for any $f' \in \mathcal{F}$, we have $\pi\big(\mathbb{P}_{\cdot,f'} \leq t\big) \geq \exp[-C' h_q(t)]$.

We have[4]: *(CA2) ⇒ (CA1) ⇔ (CA3).* Let t and C' be positive reals. We will say that a probability distribution π satisfies (t, C')-*(CA3)* when we have

$$\pi\big(\mathbb{P}_{\cdot,\tilde{f}} \leq t\big) \geq \exp[-C' h_q(t)].$$

Note that this last assumption is, unlike the others, a local complexity assumption.

3.1.2. *Margin assumptions.* We will consider variants of Tsybakov's margin assumption ([11, 15]). Let $\alpha \in \mathbb{R}_+ \cup \{+\infty\}$ and $\kappa \in [1; +\infty]$. We define

$$\Delta R(f) \triangleq R(f) - R(\tilde{f}).$$

(MA1) : $\mathcal{Y} = \{0; 1\}$ and there exists $C'' > 0$ such that for any $t > 0$,

$$\mathbb{P}\big(0 < |\eta^*(X) - 1/2| \leq t\big) \leq C'' t^{\alpha}.$$

(MA2) : there exists $C'' > 0$ such that for any function $f \in \mathcal{F}$,

$$\mathbb{P}_{f,\tilde{f}} \leq C'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}}.$$

(MA3) : there exist $c'', C'' > 0$ such that for any function $f \in \mathcal{F}$,

$$c'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}} \leq \mathbb{P}_{f,\tilde{f}} \leq C'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}}. \tag{3.1}$$

[4] To prove *(CA1) ⇒ (CA3)*: for any $k \in \mathbb{N}^*$, introduce π_k the uniform distribution on a $(2^{-k}, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})-$minimal covering net. The prior distribution $\pi \triangleq \sum_{k \geq 1} \frac{\pi_k}{k(k+1)}$ satisfies the claim.

(MA4) : there exist $c'', C'' > 0$ such that $\mathbb{P}_{\cdot,\tilde{f}} \le C''[\Delta R]^{\frac{1}{\kappa}}$, and for any $t > 0$, $\pi(\Delta R \le t) \ge c''\pi\big(\mathbb{P}_{\cdot,\tilde{f}} \le C'' t^{\frac{1}{\kappa}}\big)$[5].

This last assumption makes sense only in the bayesian context where a prior distribution π is put on the model. It is easy to check the following implications: *(MA3)* $\Rightarrow$ *(MA4)* $\Rightarrow$ *(MA2)*. Besides when $f^* \in \mathcal{F}$ (= no bias assumption), we have: *(MA1)* $\Rightarrow$ *(MA2)* for $\kappa = \frac{1+\alpha}{\alpha}$. When $\kappa = +\infty$, Assumption *(MA2)* is empty[6] and Assumptions *(MA3)* and *(MA4)* are not satisfied by non-trivial models. The margin Assumptions *(MA1)* and *(MA2)* are all the stronger as κ is small. When $\kappa = 1$, the lower bound in inequality (3.1) holds trivially for $c'' = 1$ and we have: *(MA3)* $\Leftrightarrow$ *(MA4)* $\Leftrightarrow$ *(MA2)*.

Remark 3.1. For sake of simplicity, we have assumed that there exists a function $\tilde{f} \in \mathcal{F}$ such that $R(\tilde{f}) = \inf_{\mathcal{F}} R$. Then, under Assumption *(MA2)*, this function needs to be unique. In fact this is not more necessary than the existence of the minimum. To be more specific, the results in this paper under Assumption *(MA2)* will still hold when this assumption is replaced with: there exists $k \in \mathbb{N}^*$ such that for any $\beta > 0$, there exists $f_1, \dots, f_k \in \mathcal{F}$

$$\forall f \in \mathcal{F},\ \exists i \in \{1, \dots, k\},\ \mathbb{P}_{f,f_i} \le C''\big[R(f) - \inf_{\mathcal{F}} R\big]^{\frac{1}{\kappa}} + \beta.$$

Note that this implies that for any $i \in \{1, \dots, k\}$, $R(f_i) - \inf_{\mathcal{F}} R \le \beta$. Similarly, we can give weakened versions of Assumptions *(MA3)* and *(MA4)*. Naturally, the value of k will influence the value of the constants in the results under Assumption *(MA2)*.

3.2. **Gibbs classifier.**

3.2.1. *Under Assumptions* (MA4) *and* (CA3) *for* $q > 0$. In this paper, we will often consider prior distributions $\pi^{(N)}$ which may depend on N. To shorten, we will simply write it π. The following lemma guarantees the efficiency of the standard Gibbs estimator for a temperature appropriately chosen.

Lemma 3.1. *Let π be a probability distribution such that*

$$\pi\big[\Delta R \le \check{C}_1 N^{-\frac{\kappa}{2\kappa-1+q}}\big] \ge e^{-\check{C}_2 N^{\frac{q}{2\kappa-1+q}}} \tag{3.2}$$

and λ_N have the same order as $N^{\frac{\kappa+q}{2\kappa-1+q}}$, i.e. such that

$$\check{C}_3 N^{\frac{\kappa+q}{2\kappa-1+q}} \le \lambda_N \le \check{C}_4 N^{\frac{\kappa+q}{2\kappa-1+q}} \tag{3.3}$$

for some positive constants $\check{C}_i, i = 1, \dots, 4$. Then, under Assumption (MA2)*, the standard Gibbs classifier in which the prediction function is drawn according to the posterior distribution $\pi_{-\lambda_N r}$ has the convergence rate $N^{-\frac{\kappa}{2\kappa-1+q}}$ to the extent that*

$$\mathbb{P}^{\otimes N} \pi_{-\lambda_N r} R - R(\tilde{f}) \le \check{C} N^{-\frac{\kappa}{2\kappa-1+q}}$$

for some constant $\check{C} > 0$ (depending only on c'', C'', κ and $\check{C}_i, i = 1, \dots, 4$).

More precisely, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, with $\pi_{-\lambda_N r}$-probability at least $1 - \epsilon$, we have

$$R - R(\tilde{f}) \le \check{C} \log(e\epsilon^{-1}) N^{-\frac{\kappa}{2\kappa-1+q}}, \tag{3.4}$$

for some constant $\check{C} > 0$ (depending on C'', κ and $\check{C}_i, i = 1, \dots, 4$).

[5] As a consequence, $\pi(\Delta R \le t)$ has the same order as $\pi(\mathbb{P}_{\cdot,\tilde{f}} \le C'' t^{\frac{1}{\kappa}})$.

[6] since the inequality trivially holds for $C'' = 1$

Proof. See Section 6.1. □

Theorem 3.2. *Let π be a distribution satisfying Assumptions* (MA4) *and* $\big(\check{C}_1 N^{-\frac{1}{2\kappa-1+q}}, \check{C}_2\big)$-(CA3) *for $q > 0$ and λ_N be a real satisfying inequality* (3.3) *for given positive constants $\check{C}_i, i = 1, \dots, 4$. Then we have*

$$\mathbb{P}^{\otimes N} \pi_{-\lambda_N r} R - R(\tilde{f}) \leq \check{C} N^{-\frac{\kappa}{2\kappa-1+q}}$$

for some constant $\check{C} > 0$ (depending only on c'', C'', κ and $\check{C}_i, i = 1, \dots, 4$).

Proof. It suffices to check that, under these assumptions, we can apply Lemma 3.1. □

Remark 3.2. To make the link with previous works about non randomized sieve estimators, one can choose π as the uniform distribution on an almost minimal $\big(\check{C}_1 N^{-\frac{1}{2\kappa-1+q}}, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}\big)$-covering net. Then Assumption *(CA1)* implies that the distribution π satisfies Assumption $\big(\check{C}_1 N^{-\frac{1}{2\kappa-1+q}}, \check{C}_2\big)$-*(CA3)* for some constant $\check{C}_2 > 0$ (depending on $\check{C}_1$, on the almost minimality constant and on the constant C' involved in *(CA1)*). Note that, as in Mammen and Tsybakov's work ([11, 15]), the computation of the estimator requires that, without knowing $\mathbb{P}(dX)$ exactly, one can construct a $(t, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-net with log-cardinality of order $H(t, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$.

The convergence rate of the standard Gibbs estimator in Theorem 3.2 is optimal since the following lower bound holds.

Theorem 3.3. *Let $q \geq 0$ and $\kappa \in [1; +\infty]$. There exist an input space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, a model $\mathcal{F}$ and a set $\mathcal{P}$ be the set of probability distributions satisfying*

- *for any $\mathbb{P} \in \mathcal{P}$, $f^*_{\mathbb{P}} \in \mathcal{F}$*
- *Assumptions* (CA2), (MA3) *and* (MA1) *with $\alpha = \frac{1}{\kappa-1} \in [0; +\infty]$*

such that for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$,

$$\sup_{\mathbb{P} \in \mathcal{P}} \big\{ \mathbb{P}^{\otimes N} R(\hat{f}) - R(\tilde{f}) \big\} \geq C N^{-\frac{\kappa}{2\kappa-1+q}}.$$

Proof. See Section 6.2. □

Remark 3.3. In [15], the same result is proved (using the classes of boundary fragments) for the set of probability distributions such that the Bayes classifier is in the model and Assumptions *(CA2)* and *(MA1)* with $\alpha = \frac{1}{\kappa-1}$ hold.

Remark 3.4. The previous theorem is stronger than what is required to prove that the convergence rate obtained in Theorem 3.2 is optimal since the set $\mathcal{P}$ in Theorem 3.3 is smaller than the set of probability distributions $\mathbb{P}$ such that there exists a distribution π satisfying Assumptions *(MA4)* and $\big(\check{C}_1 N^{-\frac{1}{2\kappa-1+q}}, \check{C}_2\big)$-*(CA3)*.

3.2.2. *Under Assumptions* (MA2) *and* (CA3) *for $q = 0$.* Using the same tools as in the previous section, we can prove

Theorem 3.4. *Let π be a distribution satisfying Assumption* $\big(\check{C}_1 N^{-\frac{\kappa}{2\kappa-1}}, \check{C}_2\big)$-(CA3) *for $q = 0$ and λ_N be a real satisfying*

$$\check{C}_3 N^{\frac{\kappa}{2\kappa-1}} \leq \lambda_N \leq \check{C}_4 N^{\frac{\kappa}{2\kappa-1}} \tag{3.5}$$

for given positive constants $\check{C}_i, i = 1, \dots, 4$. Under Assumption (MA2)*, we have*

$$\mathbb{P}^{\otimes N} \pi_{-\lambda_N r} R - R(\tilde{f}) \leq \check{C} (\log N) N^{-\frac{\kappa}{2\kappa-1}}$$

for some constant $\check{C} > 0$ (depending only on C'', κ and $\check{C}_i, i = 1, \dots, 4$).

Proof. We use Lemma 6.1 and the inequalities

$$\begin{cases} T_2(\pi) & \le & \breve{C}\lambda\big(\frac{\lambda}{N}\big)^{\frac{\kappa}{\kappa-1}} \\ T_1(\pi) & \le & -\log\big[\pi\big(\Delta R \le \breve{C}_1 N^{-\frac{\kappa}{2\kappa-1}}\big)\big] + \lambda\breve{C}_1 N^{-\frac{\kappa}{2\kappa-1}} \\ & \le & -\log\big[\pi\big(\mathbb{P}_{\cdot,\tilde{f}} \le \breve{C}_1 N^{-\frac{\kappa}{2\kappa-1}}\big)\big] + \lambda\breve{C}_1 N^{-\frac{\kappa}{2\kappa-1}} \end{cases}.$$

□

From Theorem 3.3, since we have *(CA2) ⇒ (CA3)* and *(MA3) ⇒ (MA2)*, this convergence rate is optimal up to the $\log N$ factor.

3.2.3. *Under Assumption* (MA2) *and a local complexity assumption.* The following theorem considers a local complexity assumption and its first and second parts respectively complete Theorem 3.4 and Lemma 3.1.

Theorem 3.5. *Let $\epsilon > 0$, $s \ge 0$, $C' > 0$, $C''' \in \mathbb{R}$ and $1 \le \kappa \le +\infty$. Consider λ depending on N such that $\lambda \underset{N\to+\infty}{\longrightarrow} +\infty$ and that Assumption* (MA2) *holds.*

First, assume that $\log \pi^{-1}\big\{R - R(\tilde{f}) \le x\big\} = -C'\log x + C''' + \underset{x\to 0}{\mathrm{o}}(x^s)$. Then we have

- *for $\lambda = \underset{N\to+\infty}{\mathrm{o}}\big(N^{\frac{\kappa}{2\kappa-1}}\big)$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,*

$$\pi_{-\lambda r}R = \tfrac{C'}{\lambda} + \tfrac{1}{\lambda}\Big\{\underset{N\to+\infty}{\mathrm{o}}\big(\lambda^{-\frac{s}{2}}\big) + \underset{N\to+\infty}{\mathrm{O}}\Big(\lambda^{\frac{(2\kappa-1)C'}{2(2\kappa C'+\kappa-1)}} N^{-\frac{\kappa C'}{2(2\kappa C'+\kappa-1)}}\Big)\log(e\epsilon^{-1})\Big\}$$

- *when $\lambda = cN^{\frac{\kappa}{2\kappa-1}}$: for any $\beta > 0$, there exist $c > 0$ and $N_0 > 0$ such that for any $N > N_0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, :*

$$\tfrac{C'-\beta}{\lambda} \le \pi_{-\lambda r}R \le \tfrac{C'+\beta}{\lambda}.$$

Secondly, assume that $\log \pi^{-1}\big\{R - R(\tilde{f}) \le x\big\} = C'x^{-\frac{q}{\kappa}} + C''' + \underset{x\to 0}{\mathrm{o}}(1)$ with $q > 0$ and $\kappa \ne +\infty$. Then we have

- *for $\lambda = \underset{N\to+\infty}{\mathrm{o}}\big(N^{-\frac{\kappa+q}{2\kappa-1+q}}\big)$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,*

$$\pi_{-\lambda r}R = \big(\tfrac{qC'+\mathrm{o}(1)}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}} + \underset{N\to+\infty}{\mathrm{O}}(1)\tfrac{\log(\epsilon^{-1})}{\lambda}$$

- *when $\lambda = cN^{-\frac{\kappa+q}{2\kappa-1+q}}$: for any $0 < \beta \le qC'$, there exist $c > 0$ and $N_0 > 0$ such that for any $N > N_0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, :*

$$\big(\tfrac{qC'-\beta}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}} \le \pi_{-\lambda r}R \le \big(\tfrac{qC'+\beta}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}.$$

Proof. See Section 6.3. □

It is interesting to note that this asymptotic behaviour only depends on the local complexity given by the weight of the sets $\big\{f \in \mathcal{F} : R(f) - R(\tilde{f}) \le x\big\}$ when $x \to 0$. Had we had $\mathbb{P}_{f,\tilde{f}} \underset{x\to 0}{\sim} C''\big[R - R(\tilde{f})\big]^{\frac{1}{\kappa}}$ on these sets, the complexity assumption would be similar to the ones introduced in Section 3.1.1 to the extent that we would have $\log \pi^{-1}\big(\mathbb{P}_{\cdot,\tilde{f}} \le x\big) \underset{x\to 0}{\sim} \breve{C}h_q(x)$.

In Theorems 3.2 and 3.4, we have seen how to choose the parameter λ depending on N such that the Gibbs classifier has the optimal convergence rate. The previous result shows that for λ smaller than these “optimal” parameters and a slightly modified complexity assumption, we can tightly bracket the efficiency of standard

Gibbs classifiers. For larger λ, the picture is not clear: it seems that the KL-divergence term in Theorem 2.2 becomes the leading term. This KL-divergence will in general explode for $\lambda \gg N$, and finally we just know that

$$\pi_{-\lambda r} R \underset{\lambda \to +\infty}{\longrightarrow} \pi|_{r=\min_{\mathcal{F}} r} R \triangleq \frac{\int_{\mathcal{F}} \mathbb{1}_{r(f)=\min_{\mathcal{F}} r} R(f) d\pi(f)}{\int_{\mathcal{F}} \mathbb{1}_{r(f)=\min_{\mathcal{F}} r} d\pi(f)}.$$

Remark 3.5. The confidence level ϵ does not appear in the main terms of the expansions of $\pi_{-\lambda r} R$, hence the asymptotic orders of $\pi_{-\lambda r} R$ hold with exponential probability.

3.2.4. *Adaptive choice of the temperature.* Here we consider that Assumption *(MA3)* holds for an unknown margin parameter κ and we prove that under assumption *(CA3)* a standard Gibbs classifier with an appropriately chosen temperature is adaptive wrt this parameter, i.e. without prior knowledge of κ, the generalization error of the randomized estimator is upper bounded by $\breve{C} N^{\frac{\kappa}{2\kappa-1+q}}$ when $q > 0$ and by $\breve{C}(\log N) N^{\frac{\kappa}{2\kappa-1}}$ when $q = 0$. The adaptation to the margin problem has also been studied in [15, 16]. In particular, in [16], Tsybakov and van de Geer proposed an adaptive penalized classifier using wavelets.

Theorem 3.6. *Under Assumptions* (MA3) *and* (CA3)*, the algorithm given in Section* 3.4.2 *of* [1] *achieves an adaptive choice of the temperature of the standard Gibbs classifier wrt the margin parameter* κ*.*

Proof. See Section 6.4. □

3.3. **Empirical risk minimization on nets.**

3.3.1. *Under Assumptions* (MA3) *and* (CA1) *for* $q > 0$*.* This section shows that, by using inequality (2.1), we can recover results on sieve estimators given in [11, 15]. These results have to be compared with the ones in Section 3.2.1 (recall that *(MA3)* $\Rightarrow$ *(MA4)* and *(CA1)* $\Leftrightarrow$ *(CA3)*).

Theorem 3.7. *Under Assumptions* (MA3) *and* (CA1) *for* $q > 0$*, for any classifier* $\hat{f}$ *minimizing the empirical risk among a* u_N*-covering net* $\mathcal{N}_{u_N}$ *such that*

$$\breve{C}_1 N^{-\frac{1}{2\kappa-1+q}} \le u_N \le \breve{C}_2 N^{-\frac{1}{2\kappa-1+q}} \tag{3.6}$$

and

$$\log \left|\mathcal{N}_{u_N}\right| \le \breve{C}_3 h_q(u_N) \tag{3.7}$$

for some positive constants $\breve{C}_i, i = 1, \dots, 3$*, we have*

$$\mathbb{P}^{\otimes N}\left[R(\hat{f}) - R(\tilde{f})\right] \le \breve{C} N^{-\frac{\kappa}{2\kappa-1+q}}$$

for some constant $\breve{C} > 0$ *(depending only on* C', c'', C'' *and* $\breve{C}_i, i = 1, \dots, 3$*).*

Proof. See Section 6.5. □

Remark 3.6. Inequality (3.7) just says that the net $\mathcal{N}_{u_N}$ is almost minimal.

3.3.2. *Under Assumptions* (MA2) *and* (CA1) *for* $q = 0$. Since *(CA1)* $\Leftrightarrow$ *(CA3)*, this section gives results to be compared with the ones in Section 3.2.2.

Theorem 3.8. *Under Assumptions* (MA2) *and* (CA1) *for* $q = 0$, *for any classifier* $\hat{f}$ *minimizing the empirical risk among a* u_N*-covering net* $\mathcal{N}_{u_N}$ *such that*

$$N^{-\check{C}_1} \le u_N \le \check{C}_2 (\log N) N^{-\frac{\kappa}{2\kappa-1}} \tag{3.8}$$

and

$$\log \left|\mathcal{N}_{u_N}\right| \le \check{C}_3 h_0(u_N) \tag{3.9}$$

for some positive constants $\check{C}_i, i = 1, \dots, 3$, *we have*

$$\mathbb{P}^{\otimes N}\big[R(\hat{f}) - R(\tilde{f})\big] \le \check{C} (\log N) N^{-\frac{\kappa}{2\kappa-1}}$$

for some constant $\check{C} > 0$ *(depending only on* c'', C'' *and* $\check{C}_i, i = 1, \dots, 3$*).*

Proof. It follows the lines of Section 6.5. This time, we take $\left(\frac{\lambda}{N}\right)^{\frac{\kappa}{\kappa-1}}$ and $\frac{\log(eu^{-1})}{\lambda}$ of the same order and greater than u. This is realized when inequality (3.8) is satisfied and $\lambda = N^{\frac{\kappa}{2\kappa-1}}$. □

3.4. **Chaining.** When a class of functions has a polynomial entropy, there is a trick called the chaining ([6]) which allows us to improve the previous results. This technique is used to get tighter upper bounds of the difference $R(f_1) - R(f_2)$ between the expected risk at two different functions f_1 and f_2. It is based on finer and finer approximations of these functions. The advantage of considering rough approximation of these functions is that the set of all possible rough approximations is small (in other words, has a small complexity). On the contrary, the set of fine approximations is big, but the distance between the fine approximation and the function approximated is small. So there is a kind of bias/variance trade-off and for polynomial entropy classes of functions, it is interesting to have this trade-off on a sequence of links and not directly on the big link $f_1 \cdots f_2$.

Let us give some results due to this technique. Consider the context of Theorem 3.7. Let us see what happens if we replace the margin Assumption *(MA3)* with Assumption *(MA2)*. Then we can no longer upper bound ΔR with Cst $\mathbb{P}^{\kappa}_{\cdot,\tilde{f}}$ (inequality which is used to obtain (6.8)). We only have $\Delta R \le \mathbb{P}_{\cdot,\tilde{f}}$. This leads to the convergence rate $N^{-\frac{\kappa}{2\kappa-1+q\kappa}}$ instead of $N^{-\frac{\kappa}{2\kappa-1+q}}$. Using the chaining trick, we will prove (see Theorem 3.10) that this rate is suboptimal and that, by minimizing the empirical risk on well chosen nets, we can still reach the rate $N^{-\frac{\kappa}{2\kappa-1+q}}$ when $0 < q < 1$ and the rate $N^{-\frac{1}{1+q}}$ when $q > 1$.

Remark 3.7. The convergence rate $N^{-\frac{\kappa}{2\kappa-1+q\kappa}}$ is optimal under Assumption *(MA2)* and the complexity assumption $H(u_N) \le C' h_q(u_N)$ for the radius $u_N = N^{-\frac{\kappa}{2\kappa-1+q\kappa}}$. The lower bound comes from Lemma 5.1 applied to a $\big(N^{\frac{q\kappa}{2\kappa-1+q\kappa}}, N^{-\frac{1+q\kappa}{2\kappa-1+q\kappa}}, \frac{1}{2}N^{-\frac{\kappa-1}{2\kappa-1+q\kappa}}\big)$-constant hypercube. By slightly modifying the proof of Theorem 3.7, we can obtain that, under the previous margin and complexity assumptions, any classifier $\hat{f}$ minimizing the empirical risk among a u_N-almost minimal net satisfies

$$\mathbb{P}^{\otimes N}\big[R(\hat{f}) - R(\tilde{f})\big] \le \check{C} N^{-\frac{\kappa}{2\kappa-1+q\kappa}}.$$

The chaining technique appears to be the only tool which allows to take into account an entropy assumption which holds for any radius such as *(CA1)* and *(CA2)*.

The chaining trick may also be used to prove that the empirical risk can be minimized on tighter nets (provided that they are still minimal, or at least almost minimal, under polynomial entropy assumptions).

Before giving results concerning nets, one can illustrate the chaining technique by considering randomized posteriors concentrated on small balls of fixed radius. For any $u > 0$, introduce $B_{f,u} \triangleq \{f' \in \mathcal{F} : \mathbb{P}_{f,f'} \le u\}$ and $\pi_{f,u} \triangleq \frac{\mathbb{1}_{B_{f,u}}}{\pi(B_{f,u})} \cdot \pi$. Define $h(v) \triangleq \sup_{f \in \mathcal{F}} \log \pi^{-1}(B_{f,v})$ and $h_+(v) \triangleq h(v) \vee 1$.

Theorem 3.9. *Let $u > 0$, $L \triangleq \frac{\log(2/u)}{\log 2}$, $C_1 \triangleq \sqrt{\frac{4h_+(u)}{3Nu}}$ and $C_0 \triangleq 2\sqrt{3}[1 + 2g(C_1)]$. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, for any $f_1, f_2 \in \mathcal{F}$ such that $\mathbb{P}_{f_1,f_2} > u$, we have*

$$\begin{aligned}\pi_{f_2,u}R - \pi_{f_1,u}R + \pi_{f_1,u}r - \pi_{f_2,u}r \\ &\le \frac{C_0}{\sqrt{N}} \sum_{k \in \mathbb{N} : u2^k < \mathbb{P}_{f_1,f_2}} \sqrt{u2^k h_+(u2^k)} + 6\sqrt{\frac{\mathbb{P}_{f_1,f_2}}{N}} \log[L\epsilon^{-1}] \\ &\le \frac{2C_0}{\sqrt{N}} \int_{u/2}^{\mathbb{P}_{f_1,f_2}} \sqrt{\frac{h_+(v)}{v}} dv + 6\sqrt{\frac{\mathbb{P}_{f_1,f_2}}{N}} \log[L\epsilon^{-1}].\end{aligned}$$

Proof. See Section 6.6. □

Had we not chained inequality (2.1), we would have obtained

$$\pi_{f_2,u}R - \pi_{f_1,u}R + \pi_{f_1,u}r - \pi_{f_2,u}r \le \frac{\lambda}{N} g\big(\frac{\lambda}{N}\big)\big(\mathbb{P}_{f_1,f_2} + 2u\big) + \frac{2h(u) + \log(\epsilon^{-1})}{\lambda}$$

This upper bound is greater than $\inf_{\lambda > 0} \big\{\frac{\lambda}{2N}\mathbb{P}_{f_1,f_2} + \frac{2h(u)}{\lambda}\big\} = 2\sqrt{\frac{\mathbb{P}_{f_1,f_2} h(u)}{N}}$, which is much bigger than the chained bound for polynomial entropies $h(u) \approx u^{-q}$, $q > 0$ when[7] $N^{-\frac{1}{1+q}} \le u \ll \mathbb{P}_{f_1,f_2}$.

The following result is an extension of Theorems 3.7 and 3.8.

Theorem 3.10. *We assume that Assumptions* (MA2) *and* (CA1) *hold. When Assumption* (MA3) *also holds, we define*

$$(v_N, a_N) \triangleq \begin{cases} \Big(\big[\frac{\log N}{N}\big]^{\frac{\kappa}{2\kappa-1}}, \exp\Big\{-\check{C}_1(\log N)^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}}\Big\}\Big) & \textit{for } q = 0 \\ \Big(N^{-\frac{\kappa}{2\kappa-1+q}}, \check{C}_1 N^{-\frac{(\kappa-1)\mathbb{1}_{q\le 1}+q}{q(2\kappa-1+q)}}\Big) & \textit{for } q > 0 \end{cases}$$

and $b_N \triangleq \check{C}_2(v_N)^{\frac{1}{\kappa}}$.

When Assumption (MA3) *does not hold, we define*

$$(v_N, a_N) \triangleq \begin{cases} \Big(\big[\frac{\log(eN^{1/\kappa})}{N}\big]^{\frac{\kappa}{2\kappa-1}}, \exp\Big\{-\check{C}_1(\log[eN^{1/\kappa}])^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}}\Big\}\Big) & \textit{for } q = 0 \\ \Big(N^{-\frac{\kappa}{2\kappa-1+q}}, \check{C}_1 N^{-\frac{\kappa-1+q}{q(2\kappa-1+q)}}\Big) & \textit{for } 0 < q < 1 \\ \Big((\log N)N^{-\frac{1}{2}}, \check{C}_1(\log N)^{-\frac{1}{2}} N^{-\frac{1}{2}}\Big) & \textit{for } q = 1 \\ \Big(N^{-\frac{1}{1+q}}, \check{C}_1 N^{-\frac{1}{1+q}}\Big) & \textit{for } q > 1 \end{cases}$$

and $b_N \triangleq \check{C}_2 v_N$.

[7] The quantity C_0 behaves as a constant only when $\frac{h_+(u)}{Nu} \le C$, so when $u \ge CN^{-\frac{1}{1+q}}$.

For any classifier minimizing the empirical risk among a u_N-covering net $\mathcal{N}_{u_N}$ such that

$$a_N \le u_N \le b_N \tag{3.10}$$

and

$$\log \big|\mathcal{N}_{u_N}\big| \le \check{C}_3 h_q(u_N) \tag{3.11}$$

for some positive constants $\check{C}_i, i = 1, \dots, 3$, we have

$$\mathbb{P}^{\otimes N}\big[R(\hat{f}) - R(\tilde{f})\big] \le C v_N$$

for some constant $C > 0$ $\big($depending on C'', $\check{C}_i, i = 1, \dots, 3$ [and also on c'' under Assumption (MA3)]$\big)$.

Proof. See Section 6.7. □

Remark 3.8. When $q = 0$ and $\kappa = +\infty$ (i.e. no margin assumption), the $\log N$ factor in $\log(eN^{1/\kappa})$ disappears. The suppression of the logarithmic factor, obtained by chaining, is similar to what occurs for VC classes (see Corollary 4.6). The difference is just that the complexity assumption concerns $\mathbb{P}$-nets here instead of empirical nets.

From Theorems 3.3 and the following theorem, these convergence rates are optimal (up to the logarithmic factor when we have $q \in \{0; 1\}$).

Theorem 3.11. *There exist an input space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, a model $\mathcal{F}$ and a set $\mathcal{P}$ of probability distributions satisfying Assumptions* (CA2) *and* (MA2) *such that for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$,*

$$\sup_{\mathbb{P} \in \mathcal{P}} \big\{\mathbb{P}^{\otimes N} R(\hat{f}) - R(\tilde{f})\big\} \ge C N^{-\frac{1}{1+q}}.$$

Proof. Apply Lemma 5.1 for a set $\mathcal{P}$ equal to a $\big(N^{\frac{q}{1+q}}, N^{-1}, \frac{1}{2}\big)$-constant hypercube and take $\mathcal{F} \triangleq \big\{f^*_{\mathbb{P}} : \mathbb{P} \in \mathcal{P}\big\}$. □

In Theorem 3.10, we consider classifiers which minimize the empirical risk on an almost minimal net $\mathcal{N}$. The following result just asserts that the same convergence rate holds for randomized estimators which “roughly” minimizes the empirical risk.

Theorem 3.12. *For any randomized classifier $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}^1_+(\mathcal{N}_{u_N})$ such that there exists a function $\breve{f} \in \mathcal{F}$ satisfying*

- $\mathbb{P}_{\breve{f},\tilde{f}} \le C u_N$,
- *for any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, $\hat{\rho} r \le r(\breve{f}) + C \log(e\epsilon^{-1}) v_N$,*

we have

$$\mathbb{P}^{\otimes N} \hat{\rho} R - R(\tilde{f}) \le \check{C} v_N.$$

Proof. It suffices to modify slightly the proof of Theorem 3.10. Let $\tilde{f}_{\mathcal{N}}$ be the nearest neighbour of $\breve{f}$ in $\mathcal{N}_{u_N}$. We have $\mathbb{P}_{\tilde{f}_{\mathcal{N}},\breve{f}} \le \check{C} u_N$. From inequality (2.3) with $\mathcal{S}_1 = \{\breve{f}\}$ and $\mathcal{S}_2 = \{\tilde{f}_{\mathcal{N}}\}$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, we have

$$r(\breve{f}) \le r(\tilde{f}_{\mathcal{N}}) + \check{C}\sqrt{\tfrac{u_N \log(\epsilon^{-1})}{N}} + \check{C}\tfrac{\log(\epsilon^{-1})}{N} + \sup_{\mathbb{P}_{\cdot,\breve{f}} \le \check{C} u_N} \Delta R,$$

hence $r(\breve{f}) \le r(\tilde{f}_{\mathcal{N}}) + \breve{C}\log(e\epsilon^{-1})v_N$. We obtain that $\hat{\rho}r \le r(\tilde{f}_{\mathcal{N}}) + \breve{C}\log(e\epsilon^{-1})v_N$. From this inequality and by using the last bound in Corollary 6.6, we obtain a new version of inequality (6.12) from which the convergence rate follows. □

As a consequence, the Gibbs estimators $\pi_{-\lambda r}$ in which the prior distribution is the uniform distribution on a net $\mathcal{N}$ perform as well as an (ERM, $\mathcal{N}$)-algorithm (i.e. a classifier which minimizes the empirical risk on the net $\mathcal{N}$) as soon as the inverse temperature parameter λ is sufficiently large. This is not surprising to the extent that the Gibbs estimator $\pi_{-\lambda r}$ when $\lambda \to +\infty$ classifies, roughly speaking, as an (ERM, $\mathcal{N}$)-algorithm.

The following theorem completes Theorem 3.10.

Theorem 3.13. *Let $\mathcal{N}_{u_N}$ be a u_N-covering net such that inequations* (3.10) *and* (3.11) *hold, let $\lambda_N \ge \breve{C}_4 \frac{h_q(u_N)}{v_N}$, and let π be a probability distribution on the net $\mathcal{N}_{u_N}$ satisfying $(u_N, \breve{C}_5)$-*(CA3) *for some positive constants $\breve{C}_i, i = 1, \dots, 5$. Then we have*

$$\mathbb{P}^{\otimes N}\big[\pi_{-\lambda_N r}R - R(\tilde{f})\big] \le \breve{C}v_N$$

for some constant $\breve{C} > 0$ $\big($depending on C'', $\breve{C}_i, i = 1, \dots, 5$ [and also on c'' under Assumption (MA3)$]\big)$.

Proof. Introduce the function $\tilde{f}_{\mathcal{N}}$ in the net $\mathcal{N}_{u_N}$ such that $\mathbb{P}_{\tilde{f},\tilde{f}_{\mathcal{N}}} \le u_N$. By Assumption $(u_N, \breve{C}_5)$-*(CA3)* and inequality (3.11), we can choose the function $\tilde{f}_{\mathcal{N}}$ such that we also have $\pi(\{\tilde{f}_{\mathcal{N}}\}) \ge e^{-(\breve{C}_5+\breve{C}_3)h_q(u_N)}$. So we have

$$\pi_{-\lambda_N r}r - r(\tilde{f}_{\mathcal{N}}) \le \frac{\log[\pi(\tilde{f}_{\mathcal{N}})^{-1})]}{\lambda_N} \le \breve{C}\frac{h_q(u_N)}{\lambda_N} \le \breve{C}v_N.$$

The result then follows from Theorem 3.12. □

3.5. **Bracketing entropy.** To minimize the empirical risk over all the model $\mathcal{F}$ can lead to inconsistency even for models with small covering entropy. For instance, define the set $\mathcal{X} = [0;1]$, the functions $f_0 \equiv 0$ and $f_1 \equiv 1$, and the probability distribution $\mathbb{P}$ such that $\mathbb{P}(dX) = \mathcal{U}([0;1])(dX)$ (uniform law over $\mathcal{X}$) and $Y = \mathbb{1}_{X \ge \frac{3}{4}}$. Consider the model formed by f_1 and all the functions equal to f_0 except on a finite number of points. For any $u < 1$, we have $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) = \log 2$. However, in general, the ERM-algorithm will classify poorly[8]. (On the contrary, the classifier based on the ERM-principle over a $(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-net for small u is efficient). This phenomenon occurs since the covering entropy does not suitably measures the complexity of models. In this section, we will see that the bracketing entropy does not suffer from this drawback.

Under polynomial bracketing entropy conditions, the empirical data contain what happens in expectation to the extent that two functions close for the distance $\mathbb{P}_{\cdot,\cdot}$ are also close for the distance $\bar{\mathbb{P}}_{\cdot,\cdot}$.

Recall that if $\mathcal{G}$ is a u-bracketing net of the set $\mathcal{F}$, then for any function $f \in \mathcal{F}$, there exist $f_L, f_U \in \mathcal{G}$ satisfying $f_L \le f \le f_U$ and $\mathbb{P}_{f_L,f_U} \le u$. Let us define the mappings $n_L, n_U : \mathcal{F} \to \mathcal{G}$ such that $n_L(f) = f_L$ and $n_U(f) = f_U$ (from the axiom of choice, they exist).

The following theorem, to be compared with Theorem 3.10, shows the influence of considering bracketing entropy assumptions instead of covering ones.

[8] That is why, in Theorem 3.10, we need to consider almost *minimal* nets (inequality (3.7)).

Theorem 3.14. *Let us define*

$$w_N \triangleq \begin{cases} \left[\frac{\log(eN^{1/\kappa})}{N}\right]^{\frac{\kappa}{2\kappa-1}} & \textit{under Assumptions } \text{(MA2)+(CA2)} \textit{ for } q=0 \\ N^{-\frac{\kappa}{2\kappa-1+q}} & \textit{under Assumptions } \text{(MA2)+(CA2)} \textit{ for } 0<q<1 \\ (\log N)N^{-\frac{1}{2}} & \textit{under Assumptions } \text{(MA2)+(CA2)} \textit{ for } q=1 \\ N^{-\frac{1}{1+q}} & \textit{under Assumptions } \text{(MA2)+(CA2)} \textit{ for } q>1 \end{cases}.$$

For any classifier $\hat{f}_{ERM,\mathcal{N}}$ minimizing the empirical risk in a $u_N \triangleq \check{C}_1 w_N$-covering net $\mathcal{N}$ for some positive constant $\check{C}_1$, we have

$$\mathbb{P}^{\otimes N}\big[R(\hat{f}_{ERM,\mathcal{N}}) - R(\tilde{f})\big] \le \check{C} w_N$$

for some constant $\check{C} > 0$ $\big($depending on C', C'' and $\check{C}_1\big)$.

Proof. Let $\mathcal{N}'$ be a u_N-minimal bracketing net of the net $\mathcal{N}$. Let $\tilde{f}_{\mathcal{N}'}$ be the nearest neighbour of the function $\tilde{f}$ in the net $\mathcal{N}'$. By definition of the set $\mathcal{N}'$,

- we have $\log|\mathcal{N}'| \le C' h_q(u_N)$,
- there exists a function $f_{\mathcal{N}}$ such that $n_L(f_{\mathcal{N}}) = \tilde{f}_{\mathcal{N}'}$ or $n_U(f_{\mathcal{N}}) = \tilde{f}_{\mathcal{N}'}$; consequently, we have $r(f_{\mathcal{N}}) \le r(\tilde{f}_{\mathcal{N}'}) + u_N$,
- there exists a classifier $\hat{f}_{\mathcal{N}'} : \mathcal{Z}^N \to \mathcal{N}'$ $\big(\hat{f}_{\mathcal{N}'} \triangleq n_L(\hat{f}_{\text{ERM},\mathcal{N}})$ for instance$\big)$ such that we have $r(\hat{f}_{\mathcal{N}'}) \le r(\hat{f}_{\text{ERM},\mathcal{N}}) + u_N$ and

$$R(\hat{f}_{\text{ERM},\mathcal{N}}) \le R(\hat{f}_{\mathcal{N}'}) + u_N. \tag{3.12}$$

So the estimator $\hat{f}_{\mathcal{N}'} : \mathcal{Z}^N \to \mathcal{N}'$ satisfies

$$r(\hat{f}_{\mathcal{N}'}) \le r(\hat{f}_{\text{ERM},\mathcal{N}}) + u_N \le r(f_{\mathcal{N}}) + u_N \le r(\tilde{f}_{\mathcal{N}'}) + 2u_N.$$

Then the result follows from Theorem 3.12 and inequality (3.12). □

Remark 3.9. Since we have *(CA2) ⇒ (CA1)*, Theorem 3.10 can be applied when the assumptions of Theorem 3.14 hold. We see that, under bracketing entropy assumptions, the ERM on nets containing a huge (possibly infinite) number of functions has also the optimal convergence rate. This was not the case under covering entropy assumptions.

Remark 3.10. The same convergence rate holds for classifiers minimizing the empirical risk up to an additive factor Cw_N.

The following theorem completes the previous one.

Theorem 3.15. *Let $\lambda_N \ge \check{C}_1 \frac{h_q(w_N)}{w_N}$ and π be a probability distribution satisfying $(\check{C}_2 w_N, \check{C}_3)$-(CA3) for some positive constants $\check{C}_i, i = 1, \dots, 3$. Then we have*

$$\mathbb{P}^{\otimes N}\big[\pi_{-\lambda_N r} R - R(\tilde{f})\big] \le \check{C} w_N$$

for some constant $\check{C} > 0$ (depending on C'', $\check{C}_i, i = 1, \dots, 3$).

Proof. See Section 6.8. □

Remark 3.11. From the previous theorem, the inverse temperature parameter λ_N should be taken as

$$\lambda_N \ge C \begin{cases} (\log N)N^{\frac{\kappa}{2\kappa-1}} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q=0 \\ N^{\frac{\kappa(1+q)}{2\kappa-1+q}} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } 0<q<1 \\ \frac{N}{(\log N)^2} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q=1 \\ N & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q>1 \end{cases}.$$

The threshold value is all the smaller as the model is small[9] (i.e for small q) and the margin assumption is weak[10] (i.e for large κ).

Finally, the following theorem shows that, under polynomial bracketing entropy assumption, with high probability, the empirical covering nets are similar to the covering nets wrt the pseudo-distance $\mathbb{P}(dX)$.

Theorem 3.16. *Let $\check{C}$ be positive constant and define*

$$(\alpha_q, \beta_q) = \begin{cases} \big(\frac{1}{N}, \frac{\log N}{N}\big) & \textit{when } q = 0 \\ \big(\exp\big\{-N^{\frac{q}{1+q}}\big\}, N^{-\frac{1}{1+q}}\big) & \textit{when } q > 0 \end{cases} .$$

With $\mathbb{P}^{\otimes N}$-probability at least $1 - (\alpha_q)^{\check{C}}$, there exists $\check{C}_1, \check{C}_2, \check{C}_3, \check{C}_4 > 0$ such that for any $u \geq \check{C}_1 \beta_q$,

- *a $(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-covering net is a $(\check{C}_3 u, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})$-covering net,*
- *a $(u, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})$-covering net is a $(\check{C}_2 u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-covering net,*
- *$H(u, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot}) \leq \check{C}_4 h_q(u)$.*

Proof. See Section 6.9. □

Therefore under polynomial bracketing entropy assumption, we can classify optimally by using the minimizer of the empirical risk on an *empirical* net of radius less than Cw_N. Note that another way of proving this result consists in saying that this classifier minimizes the empirical risk on the set $\mathcal{F}$ up to an additive Cw_N factor.

4. Classification under empirical complexity assumptions

In this section, we will see that if we replace the complexity assumption concerning $\mathbb{P}$-entropies with a similar assumption on the empirical entropies, the same kind of convergence rates appear. VC-classes are a special case in which for any $u > 0$ and any training set, we have $H(u, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot}) \leq CVh_0(u)$ where V is the VC-dimension of $\mathcal{F}$.

4.1. **Concentration of the empirical entropies.** In general, the link between the $\mathbb{P}$-entropies and $\bar{\mathbb{P}}$-entropies is not known. However, thanks to recent work by Boucheron, Bousquet, Lugosi and Massart, we are able to prove that the empirical entropies are concentrated.

Theorem 4.1. *For any $\epsilon > 0$ and $u \geq 0$*

- *with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, we have*

(4.1)
$$H(u, \mathcal{F}, \bar{\mathbb{P}}) \leq \mathbb{P}^{\otimes N} H(u, \mathcal{F}, \bar{\mathbb{P}}) + \frac{(\log 2)\log(\epsilon^{-1})}{3}\left(1 + \sqrt{1 + \frac{18\mathbb{P}^{\otimes N} H(u, \mathcal{F}, \bar{\mathbb{P}})}{(\log 2)\log(\epsilon^{-1})}}\right)$$

[9] This might be explained by looking at the size of the sets $\big\{f \in \mathcal{F} : r(f) - \min_{\mathcal{F}} r = \frac{k}{N}\big\}$. Indeed, when the model becomes larger and larger, the weight on these sets increases much more for small k than for very small k, hence we need to have larger λ to get rid of functions having a not-so-small empirical risk.

[10] This is not surprising since the stronger the margin assumption is, the smaller the optimal convergence rate is, and consequently the more selective we need to be.

equivalently

$$\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}) \geq H(u,\mathcal{F},\bar{\mathbb{P}}) + \frac{2(\log 2)\log(\epsilon^{-1})}{3}\left(1-\sqrt{1+\frac{9H(u,\mathcal{F},\bar{\mathbb{P}})}{2(\log 2)\log(\epsilon^{-1})}}\right)$$

- *with* $\mathbb{P}^{\otimes N}$*-probability at least* $1-\epsilon$*,*

$$H(u,\mathcal{F},\bar{\mathbb{P}}) \geq \mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}) - \sqrt{2(\log 2)\log(\epsilon^{-1})\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}})}$$

equivalently

$$\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}) \leq H(u,\mathcal{F},\bar{\mathbb{P}}) + (\log 2)\log(\epsilon^{-1})\left(1+\sqrt{1+\frac{2H(u,\mathcal{F},\bar{\mathbb{P}})}{(\log 2)\log(\epsilon^{-1})}}\right) \tag{4.2}$$

- *with* $\mathbb{P}^{\otimes 2N}$*-probability at least* $1-2\epsilon$*,*

$$H(u,\mathcal{F},\bar{\mathbb{P}}') \leq H(u,\mathcal{F},\bar{\mathbb{P}}) + 2(\log 2)\log(\epsilon^{-1})\left(\frac{6}{5}+\sqrt{1+\frac{2H(u,\mathcal{F},\bar{\mathbb{P}})}{(\log 2)\log(\epsilon^{-1})}}\right) \tag{4.3}$$

Proof. See Section 6.10. □

The previous result shows that the empirical entropies behave with high probability as the non empirical quantity $\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}})$. Specifically, by using a union bound on the different possible radius, we obtain that for any $\breve{C}' > 0$ there exists $\breve{C} > 0$ such that with probability at least $1 - \frac{1}{N^{\breve{C}'}}$, for any $u > 0$, we have[11]

$$\left\{\begin{array}{lcl} \mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) & \leq & \breve{C}[H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + \log N] \\ H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) & \leq & \breve{C}[\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + \log N] \\ H(u,\mathcal{F},\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}) & \leq & \breve{C}[H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + \log N] \\ H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) & \leq & H(u/2,\mathcal{F},\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}) \end{array}\right. .$$

4.2. **Chaining empirical quantities...**

4.2.1. *...in the transductive learning.* In this section, we assume that we possess two samples of size N. The first sample is labeled: $\{(X_1,Y_1),\dots,(X_N,Y_N)\}$. The second one $\{X_{N+1},\dots,X_{2N}\}$ has to be labeled: the outputs $\{Y_{N+1},\dots,Y_{2N}\}$ are unknown. We will use the following notations:

$$\left\{\begin{array}{lcl} \bar{\mathbb{P}} & \triangleq & \frac{1}{N}\sum_{i=1}^{N}\delta_{(X_i,Y_i)} \\ \bar{\mathbb{P}}' & \triangleq & \frac{1}{N}\sum_{i=N+1}^{2N}\delta_{(X_i,Y_i)} \\ \bar{\bar{\mathbb{P}}} & \triangleq & \frac{1}{2N}\sum_{i=1}^{2N}\delta_{(X_i,Y_i)} \\ r(f) & \triangleq & \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{Y_i\neq f(X_i)} = \bar{\mathbb{P}}[Y\neq f(X)] \\ r'(f) & \triangleq & \frac{1}{N}\sum_{i=N+1}^{2N}\mathbb{1}_{Y_i\neq f(X_i)} = \bar{\mathbb{P}}'[Y\neq f(X)] \end{array}\right.$$

Let us start with a basic result which is not “chained”.

Lemma 4.2. *Let* $\mathcal{S}_1$ *and* $\mathcal{S}_2$ *be two finite sets of functions from* $\mathcal{X}$ *into* $\mathcal{Y}$ *possibly depending on the data* Z_1^{2N} *in an exchangeable way. For any* $\epsilon > 0$*, with* $\mathbb{P}^{\otimes 2N}$*-probability at least* $1-\epsilon$*, for any functions* $f_1 \in \mathcal{S}_1$ *and* $f_2 \in \mathcal{S}_2$*, we have*

$$r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \leq \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{f_1,f_2}\log(|\mathcal{S}_1||\mathcal{S}_2|\epsilon^{-1})}{N}}$$

[11] For the third inequality, we use the inequality $H(u,\mathcal{F},\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}) \leq H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + H(u,\mathcal{F},\bar{\mathbb{P}}'_{\cdot,\cdot})$ and inequality (4.3). The fourth inequality always holds.

Proof. The result comes from inequality (8.7) in [1] in which we take ν equal to the uniform distribution on $\mathcal{S}_1 \times \mathcal{S}_2$ and $\mathcal{W}[(f_1, f_2), Z] = \mathbb{1}_{Y \neq f_2(X)} - \mathbb{1}_{Y \neq f_1(X)}$. □

By chaining this inequality, we obtain:

Theorem 4.3. *Let $U \in \mathbb{N}^*$ and $u \triangleq 2^{-U}$. Let $\mathcal{N}$ be an u-minimal covering net. For any $k \in \mathbb{N}^*$, let H_k be an upper bound of $H(2^{-k}, \mathcal{F}, \bar{\bar{\mathbb{P}}}) \vee 1$. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$, for any $f_1, f_2 \in \mathcal{N}$,*

$$\begin{aligned} r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \\ \leq \sum_{k \in \mathbb{N}^* : u \leq 2^{-k} \leq \bar{\bar{\mathbb{P}}}_{f_1, f_2} \vee u} 4\sqrt{\tfrac{6 \times 2^{-k}\{2H_k + \log[3k(k+1)] + \log(\epsilon^{-1})\}}{N}}. \end{aligned}$$

Proof. See Section 6.11. □

Remark 4.1. The previous result can also be written in terms of integral. For instance, for $H_k = H(2^{-k}, \mathcal{F}, \bar{\bar{\mathbb{P}}}) \vee 1$, the previous RHS is upper bounded by[12]

$$\tfrac{28}{\sqrt{N}} \int_{\frac{u}{2}}^{(\bar{\bar{\mathbb{P}}}_{f_1, f_2} \vee u) \wedge \frac{1}{2}} \left(\sqrt{\tfrac{H(x, \mathcal{F}, \bar{\bar{\mathbb{P}}}) \vee 1}{x}} + \sqrt{\tfrac{\log(4 \log x^{-1})}{x}} \right) dx + 34\sqrt{\tfrac{(\bar{\bar{\mathbb{P}}}_{f_1, f_2} \vee u) \log(3\epsilon^{-1})}{N}}.$$

4.2.2. *...in the inductive learning.* The empirical bound for the inductive learning is derived from the one for the transductive learning and from the concentration properties of the pseudo-distances and the empirical entropies.

Theorem 4.4. *Let $\epsilon > 0$ and $H_\infty \triangleq 16 \log N(X_1^N) + 20 \log(5 \log N) + 12\log(\epsilon^{-1})$. With $\mathbb{P}^{\otimes 2N}$-probability $1 - 3\epsilon$, for any functions f_1 and f_2 in the set $\mathcal{F}$, we have*

$$\begin{aligned} r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \\ \leq \tfrac{10}{\sqrt{N}} \sum_{\substack{k \in \mathbb{N}^* : \\ \frac{1}{2N} \leq 2^{-k} \leq \frac{5}{4} \bar{\mathbb{P}}_{f_1, f_2} + \frac{H_\infty}{N}}} \sqrt{8H(2^{-k}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot}) + 6 \log[k(k+1)\epsilon^{-1}] + 1}\,. \end{aligned}$$

Proof. See Section 6.12. □

The previous theorem gives, for instance, a guarantee of misclassification rate of the ERM-classifier on N new input data to classify. We recall that the leading term in the square root is generally the entropy one. Once more, we can upper bound the associated sum with the integral entropy

$$\tfrac{C}{\sqrt{N}} \int_{\frac{1}{4N}}^{\frac{5}{4} \bar{\mathbb{P}}_{f_1, f_2} + \frac{H_\infty}{N}} \sqrt{\tfrac{H(x, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})}{x}} dx$$

Note that this result is less general than the one for transductive learning since the integral starts from $\frac{1}{4N}$, which means that the largest complexity terms are taken into account. In Section 3, we have seen that for polynomial entropies with $q \geq 1$, the optimal convergence rate (which was of order $N^{-\frac{1}{1+q}}$ up to the logarithmic factor) was proved since the largest complexities were not in the integral entropy.

On the contrary, for $q < 1$, we can recover the same convergence rates under the assumption $H(u, \mathcal{F}, \bar{\mathbb{P}}) \leq C' h_q(u)$ for any $u > 0$, as under the polynomial bracketing entropy assumption. The following section deals with a special case of the case $q = 0$.

[12] Proof at the end of Section 6.11.

4.3. **Application to VC-classes.** At first sight, it is not obvious that Theorem 4.3 gives a tighter bound than Lemma 4.2 applied to $(\mathcal{S}_1, \mathcal{S}_2) = (\mathcal{N}, \mathcal{N})$. We will see in this section that for VC-classes, the two bounds gives the same convergence rate for the ERM-classifier, except when we have no margin assumption. In this last case, the chained result allows to get rid of a logarithmic factor.

Let us consider the binary classification setting: $\mathcal{Y} = \{0; 1\}$. Introduce the shattering number $N(X_1^{2N}) \triangleq \big|\big\{[f(X_k)]_{k=1}^{2N} : f \in \mathcal{F}\big\}\big| = H(u, \mathcal{F}, \bar{\bar{\mathbb{P}}})$ for any $u < \frac{1}{2N}$. Let V be the VC-dimension of the set $\mathcal{F}$

$$V \triangleq \max\big\{|A| : A \in \mathcal{X}^{2N} \text{ such that } |\{A \cap f^{-1}(1) : f \in \mathcal{F}\}| = 2^{|A|}\big\}.$$

The empirical entropies satisfy[13]

$$H(u, \mathcal{F}, \bar{\bar{\mathbb{P}}}) \le \left\{ \begin{array}{l} V \log\left(\frac{2Ne}{V}\right) \\ V \log\left(\frac{4e}{u}\right) \end{array} \right. .$$

Let $\hat{f}_{\mathrm{ERM}}$ be the minimizer of the empirical risk on the set $\mathcal{F}$ and $\tilde{f}'$ be the minimizer on $\mathcal{F}$ of either r' or R. From Lemma 4.2, with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$, we have

$$r'(\hat{f}_{\mathrm{ERM}}) \le \inf_{f \in \mathcal{F}} \left\{ r'(f) + 4\sqrt{\frac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, f}\left[V \log\left(\frac{2eN}{V}\right) + \frac{1}{2}\log(\epsilon^{-1})\right]}{N}} \right\},$$

and consequently, after some standard computations:

$$\mathbb{P}^{\otimes N} R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f}) \le 4\sqrt{\frac{V \mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, \tilde{f}}}{N} \log\left(\frac{2eN}{V}\right)} + 2\sqrt{\frac{2\mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, \tilde{f}}}{N}}. \tag{4.4}$$

To compare, from Theorem 4.3, we obtain

Corollary 4.5. *For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$, we have*

$$r'(\hat{f}_{ERM}) \le \inf_{f \in \mathcal{F}} \left\{ r'(f) + 47\sqrt{\frac{(V+1)\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, f}}{N} \log\left(\frac{8e}{\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, f}}\right)} + 34\sqrt{\frac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, f} \log(\epsilon^{-1})}{N}} \right\}$$

and, consequently,

$$\begin{aligned} \mathbb{P}^{\otimes N} R(\hat{f}_{ERM}) - R(\tilde{f}) \quad \le \quad & 47\sqrt{\frac{(V+1)\mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, \tilde{f}}}{N} \log\left(\frac{8e}{\mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, \tilde{f}}}\right)} \\ & + 34\sqrt{\frac{\mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, \tilde{f}}}{N}}. \end{aligned} \tag{4.5}$$

Proof. See Section 6.13. □

As a consequence, we obtain

Corollary 4.6. *Under assumption* (MA2)*, for any set $\mathcal{F}$ of VC-dimension V, the ERM-classifier satisfies*

$$\mathbb{P}^{\otimes N} R(\hat{f}_{ERM}) - R(\tilde{f}) \le \breve{C} \left\{ \begin{array}{ll} \left(\frac{V}{N} \log N\right)^{\frac{\kappa}{2\kappa - 1}} & \textit{when } 1 \le \kappa < +\infty \\ \sqrt{\frac{V}{N}} & \textit{when } \kappa = +\infty \end{array} \right. .$$

Proof. See section 6.14. □

[13] The first inequality is well-known consequence of Sauer's lemma; the second one comes from Haussler's formula ([7]), which asserts that for any $u > 0$, $H(u, \mathcal{F}, \bar{\bar{\mathbb{P}}}) \le V \log\left(\frac{2e}{u}\right) + \log[e(V+1)]$.

Remark 4.2. This is an improvement of Massart and Nédélec results [12, Corollary 2.2] to the extent we do not have an extra additive term $R(\tilde{f}) - R(f^*)$, where f^* is the Bayes classifier. The second part of the corollary is a well-known result which is given with a simple proof in [5, p.31].

Remark 4.3. By comparing Inequalities (4.4) and (4.5), we see that the constants in chained inequalities are not satisfactory. The gap between the upper bound (6.23) and the lower bound (see Theorem 5.2) is the factor 8 × 83 ! We do not know how to chain inequalities with significantly tighter constants.

5. Assouad's lemma

Definition 5.1. Let $m \in \mathbb{N}^*$, $w \in]0;1]$, $b \in]0;1]$ and $b' \in]0;1]$. A (m, w, b, b')-*hypercube* of probability distributions is a family

$$\left\{ \mathbb{P}_{\vec{\sigma}} \in \mathcal{M}_+^1(\mathcal{Z}) : \vec{\sigma} \triangleq (\sigma_1, \dots, \sigma_m) \in \{-1;+1\}^m \right\}$$

of 2^m probability distributions having the same first marginal:

$$\mathbb{P}_{\vec{\sigma}}(dX) = \mathbb{P}_{(+1,\dots,+1)}(dX) \triangleq \mu \text{ for any } \vec{\sigma} \in \{-1;+1\}^m,$$

and such that there exists a partition $\mathcal{X}_0, \dots, \mathcal{X}_m$ of $\mathcal{X}$ satisfying

- for any $j \in \{1, \dots, m\}$, we have $\mu(\mathcal{X}_j) = w$
- for any $j \in \{0, \dots, m\}$, for any $X \in \mathcal{X}_j$, we have

$$\mathbb{P}_{\vec{\sigma}}(Y = 1|X) = \tfrac{1+\sigma_j \xi(X)}{2} = 1 - \mathbb{P}_{\vec{\sigma}}(Y = 0|X),$$

where $\sigma_0 \triangleq 1$ and $\xi : \mathcal{X} \to [0;1]$ is such that for any $j \in \{1, \dots, m\}$,

$$\begin{cases} b &= \sqrt{1 - \left(\mu\left[\sqrt{1-\xi^2(X)}\,\middle|X \in \mathcal{X}_j\right]\right)^2} \\ b' &= \mu\left[\xi(X)|X \in \mathcal{X}_j\right] \end{cases}.$$

When ξ is constant on $\mathcal{X}_j, j = 1, \dots, m$ (which implies $\xi \equiv b' = b$ on $\mathcal{X} - \mathcal{X}_0$), the hypercube will be said a (m, w, b)-constant hypercube. The hypercube will be said noiseless when $\xi \equiv 1$ on $\mathcal{X}_0$.

The following lemma is Assouad's lemma adapted to the classification framework.

Lemma 5.1. *If a set $\mathcal{P}$ of probability distributions contains a (m, w, b, b')-hypercube, then for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, we have*

$$\sup_{\mathbb{P} \in \mathcal{P}} \left\{ \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \right\} \geq \tfrac{1 - b\sqrt{Nw}}{2} mwb'.$$

Proof. See Section 6.15. □

Lemma 5.1 gives a very simple strategy to obtain a lower bound for a given set $\mathcal{P}$ of probability distributions: it consists in looking for the (m, w, b, b')-hypercube which is contained in the set $\mathcal{P}$ and for which $\frac{1-b\sqrt{Nw}}{2} mwb'$ is maximized.

In general, the order of the bound is given by the quantity mwb' and w, b are taken such that $\sqrt{Nw}b = \text{Cst} < 1$. To obtain this order, we do not need the sophisticated computations detailed in the proof of the lemma. We can use two well-known lemmas instead (Birgé's lemma and Huber's lemma) as it is proved in Appendix E.

Lemma 5.1 implies lower bounds for VC-classes with decent constants. The following result is to be compared with Theorems 14.1 and 14.5 in [5].

Theorem 5.2. *For any model $\mathcal{F}$, define $\mathcal{P}_L$ as the set of probability distributions such that* $\inf_{f\in\mathcal{F}} R_{\mathbb{P}}(f) = L$ *for a fixed* $L \in [0; 1/2]$.

- *When* $L = 0$*:*

for any classification model $\mathcal{F}$ of VC-dimension $V \geq 2$, for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_0} \Big\{ \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \Big\} \geq \begin{cases} \frac{V-1}{2e(N+1)} \text{ when } N \geq (V-2) \vee 1 \\ \frac{1}{2}\big(1 - \frac{1}{V-1}\big)^N \end{cases}.$$

- *When* $0 < L \leq 1/2$*:*

for any classification model $\mathcal{F}$ of VC-dimension $V \geq 2$, for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_L} \Big\{ \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \Big\} \geq \begin{cases} \sqrt{\frac{L(V-1)}{32N}} \vee \frac{V-1}{16N} \text{ when } \frac{(1-2L)^2 N}{V-1} \geq \frac{1}{4} \\ (\frac{1}{2} - L)\sqrt{\frac{L}{2}} \text{ otherwise} \end{cases}.$$

Proof. See Appendix F. □

Remark 5.1. It is a well known result that, when $\inf_{f\in\mathcal{F}} R_{\mathbb{P}}(f)$ is of order $1/N$ and when the complexity of the class is not too high, there exists an estimator such that $\mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - \inf_{f\in\mathcal{F}} R_{\mathbb{P}}(f) = \mathrm{O}\big(\frac{1}{N}\big)$. The previous theorem gives a corresponding lower bound.

6. Proofs

6.1. Proof of Lemma 3.1.

Let $T_1(\pi) \triangleq -\log \pi \exp(-\lambda\Delta R)$ and

$$T_2(\pi) \triangleq 0 \vee \log \pi \exp\Big(8.2\frac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}} - \lambda\Delta R\Big).$$

We start with the following lemma.

Lemma 6.1. *For any $\epsilon > 0$ and $0 < \lambda \leq 0.19N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have $\pi_{-\lambda r} R \leq \frac{C}{\lambda}\big[T_1(\pi) + T_2(\pi) + \log(4\epsilon^{-1})\big]$.*

Proof. Taking $\chi = \frac{1}{2}$ and $\gamma = \frac{1}{2}$ in Theorem 2.2, we get

$$\begin{aligned} \pi_{-\lambda r}\Delta R &\leq \pi_{-\frac{\lambda}{2}R}\Delta R + \tfrac{2}{\lambda}\Big[16 \log \pi_{-\lambda R} \exp\Big(\tfrac{8.2\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\Big) + 5\log(4\epsilon^{-1})\Big] \\ &\leq -\tfrac{2}{\lambda}\log \pi \exp\big(-\tfrac{\lambda}{2}\Delta R\big) + \tfrac{32}{\lambda}\log \pi_{-\lambda R} \exp\Big(\tfrac{8.2\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\Big) + \tfrac{10}{\lambda}\log(4\epsilon^{-1}) \\ &\leq -\tfrac{34}{\lambda}\log \pi \exp(-\lambda\Delta R) + \tfrac{32}{\lambda}\log \pi \exp\Big(-\lambda\Delta R + \tfrac{8.2\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\Big) \\ &\qquad + \tfrac{10}{\lambda}\log(4\epsilon^{-1}). \end{aligned}$$

□

Remark 6.1. In order to explain the assumptions used in Lemma 3.1, let us give upper bounds for the quantities T_1 and T_2 using the strong complexity and margin Assumptions *(CA1)* and *(MA3)* for a well chosen distribution π. Under Assumption *(CA1)* (which is equivalent to Assumption *(CA3)*), there exists a distribution $\pi^{(t)}$ such that for any $f' \in \mathcal{F}$, $\pi^{(t)}\big(\mathbb{P}_{\cdot,f'} \leq t\big) \geq e^{-C't^{-q}}$.

For any $0 < t < 1$, we have

$$\begin{aligned} T_1\Big[\pi^{(c''t^{1/\kappa})}\Big] &\leq -\log\big[\pi^{(c''t^{1/\kappa})}(\Delta R \leq t)e^{-\lambda t}\big] && \text{(by Markov's inequality)} \\ &\leq -\log\big[\pi^{(c''t^{1/\kappa})}(\mathbb{P}_{\cdot,\tilde{f}} \leq c''t^{\frac{1}{\kappa}})\big] + \lambda t && \big(\text{according to } (MA3)\big) \\ &\leq C'c''^{-q}t^{-\frac{q}{\kappa}} + \lambda t && \big(\text{by definition of } \pi^{(t)}\big). \end{aligned}$$

Assumption *(MA2)* $\big($recall that *(MA3)* $\Rightarrow$ *(MA2)*$\big)$ implies that for any $\lambda > 0$

$$\begin{array}{rcl} 8.2\frac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}} - \lambda\Delta R & \leq & 8.2C''\frac{\lambda^2}{N}(\Delta R)^{\frac{1}{\kappa}} - \lambda\Delta R \\ & \leq & \lambda\sup\limits_{x\geq 0}\Big\{8.2C''\frac{\lambda}{N}x^{\frac{1}{\kappa}} - x\Big\} \\ & = & (\kappa-1)\lambda\Big(\frac{8.2C''\lambda}{\kappa N}\Big)^{\frac{\kappa}{\kappa-1}}, \end{array}$$

hence $T_2(\pi) \leq \breve{C}\lambda\big(\frac{\lambda}{N}\big)^{\frac{\kappa}{\kappa-1}}$ for any distribution π and a constant $\breve{C} > 0$ depending on C'' and κ. $\big($Note that for the limit case $\kappa = 1$, we have $T_2 = 0$ for any $\lambda \leq \frac{N}{8.2C''}$.$\big)$ Therefore, with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1 - 4\epsilon$, we have

$$\big\{\pi^{\left(c''t^{1/\kappa}\right)}\big\}_{-\lambda r}R - R(\tilde{f}) \leq \breve{C}\bigg[t + \frac{t^{-\frac{q}{\kappa}} + \log(\epsilon^{-1})}{\lambda} + \Big(\frac{\lambda}{N}\Big)^{\frac{\kappa}{\kappa-1}}\bigg],$$

where the constant $\breve{C} > 0$ depends on C', c'' and κ. The sum $t + \frac{t^{-\frac{q}{\kappa}}}{\lambda} + \big(\frac{\lambda}{N}\big)^{\frac{\kappa}{\kappa-1}}$ has the minimal order $N^{-\frac{\kappa}{2\kappa-1+q}}$ when λ has the order of $N^{\frac{\kappa+q}{2\kappa-1+q}}$ and t has the order of $N^{-\frac{\kappa}{2\kappa-1+q}}$. This computation explains the choice of Assumptions (3.2) and (3.3).

From inequality (3.2), we have $T_1(\pi) \leq \breve{C}N^{\frac{q}{2\kappa-1+q}} + \lambda_N N^{-\frac{\kappa}{2\kappa-1+q}}$. From Assumption *(MA2)*, we have seen in the previous remark that $T_2(\pi) \leq \breve{C}\lambda_N\big(\frac{\lambda_N}{N}\big)^{\frac{\kappa}{\kappa-1}}$. From inequality (3.3), we obtain the desired convergence rate.

Now let us prove the sharper result: inequality (3.4). Let $a(\lambda) \triangleq \frac{\lambda}{N}g\big(\frac{\lambda}{N}\big)$. From Theorem 6.2 in [1] and the same computations as for the quantity $\mathcal{L}''$ in Section 9.12 of [1] to upper bound $-\log\pi\exp\big\{-\lambda[r - r(\tilde{f})]\big\}$, we obtain :

Lemma 6.2. *For any $\epsilon > 0$, $\lambda > 0$ and $\xi > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, with $\pi_{-\lambda r}$-probability at least $1 - \epsilon$, we have*

$$\Delta R \leq a(\lambda)\mathbb{P}_{\cdot,\tilde{f}} + \frac{-\log\pi\exp\big\{-\lambda\Delta R - \lambda a\big(\frac{\lambda}{\xi}\big)\mathbb{P}_{\cdot,\tilde{f}}\big\} + (2+\xi)\log(\epsilon^{-1})}{\lambda}.$$

Taking $\xi = 1$ and $\lambda = \lambda_N$, using the margin assumption $\mathbb{P}_{\cdot,\tilde{f}} \leq C''(\Delta R)^{\frac{1}{\kappa}}$ and noting that $a(\lambda_N) \leq g(\breve{C}_4)\frac{\lambda}{N}$, we get

$$\Delta R \leq \breve{C}\tfrac{\lambda_N}{N}(\Delta R)^{\frac{1}{\kappa}} + \frac{3\log(\epsilon^{-1})}{\lambda_N} + \sup_{x\geq 0}\big\{\breve{C}\tfrac{\lambda_N}{N}x^{\frac{1}{\kappa}} - x\big\} - \frac{\log\pi\exp(-2\lambda_N\Delta R)}{\lambda_N},$$

where the constant $\breve{C} > 0$ only depends on C'' and $\breve{C}_4$. Now from the same computations as in Remark 6.1, when the Inequalities (3.2) and (3.3) hold, we get

$$\Delta R \leq \breve{C}\bigg[\frac{\lambda_N}{N}(\Delta R)^{\frac{1}{\kappa}} + N^{-\frac{\kappa}{2\kappa-1+q}}\bigg] + \frac{3\log(\epsilon^{-1})}{\lambda_N}.$$

We obtain successively

$$\Delta R \leq \breve{C}\big[N^{\frac{1-\kappa}{2\kappa-1+q}}(\Delta R)^{\frac{1}{\kappa}} + \log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}\big]$$

and

$$\Delta R \leq \breve{C}\log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}.$$

6.2. **Proof of Theorem 3.3.** A standard idea to prove lower bounds is to consider an adequate hypercube of probability distributions and to use Assouad's lemma (see Section 5 for the definition of the hypercube of distributions).

Consider a $\big(N^{\frac{q}{2\kappa-1+q}}, N^{-\frac{1+q}{2\kappa-1+q}}, aN^{-\frac{\kappa-1}{2\kappa-1+q}}\big)$-constant noiseless hypercube of probability distributions $\big\{\mathbb{P}_{\vec{\sigma}} : \vec{\sigma} \in \{-1;+1\}^m\big\}$, where $a > 0$ is a constant which will be chosen later.

Had we replaced Assumption *(MA3)* with Assumption *(MA2)* in Theorem 3.3, the result would have been a direct consequence of Lemma 5.1 applied to this hypercube with $a = \frac{1}{2}$.

In this proof, we will not apply Assouad's lemma but Fano's lemma since Assumption *(MA3)* is not satisfied by the whole hypercube. First let us state the following classical result on the hypercube which is a refined version of Varshamov-Gilbert bound (1962).

Lemma 6.3 (Huber,[8, p.256])**.** *Let $\delta(\Sigma,\Sigma')$ denote the Hamming distance between Σ and Σ' in $\{-1,1\}^m$: $\delta(\Sigma,\Sigma') \triangleq \sum_{i=1}^m \mathbb{1}_{\Sigma_i \neq \Sigma'_i}$. There exists a subset $\mathcal{S}$ of the hypercube $\{-1,1\}^m$ such that*

- *for any $\Sigma \neq \Sigma'$ in $\mathcal{S}$, we have $\delta(\Sigma,\Sigma') \geq \frac{m}{4}$*
- *$\log|\mathcal{S}| \geq \frac{m}{8}$.*

Proof. It suffices to upper bound the number of points in the ball centered at a point σ of the hypercube and of radius $\frac{m}{4}$. Consider the uniform distribution $\nu(d\Sigma)$ on the hypercube $\{-1,1\}^m$. Specifically, we have

$$\nu\bigg(\delta(\Sigma,\sigma) \leq \frac{m}{4}\bigg) \leq \nu e^{\frac{m}{4}-\delta(\Sigma,\sigma)} = e^{\frac{m}{4}}\bigg(\nu e^{-\mathbb{1}_{\Sigma_i \neq \sigma_i}}\bigg)^m = \bigg(\frac{e^{\frac{1}{4}}(1+e^{-1})}{2}\bigg)^m \leq e^{-\frac{m}{8}},$$

which leads to the desired result. □

Let $\mathcal{S} \subset \{-1;+1\}^m$ such that $|\mathcal{S}| = \lfloor e^{\frac{m}{8}} \rfloor$ and for any $\Sigma \neq \Sigma'$ in $\mathcal{S}$, $\delta(\Sigma,\Sigma') \geq \frac{m}{4}$. From inequality (5.1) in [2], Birgé's version of Fano's lemma can be stated as

Lemma 6.4. *Given a non-trivial (i.e. cardinal ≥ 2) finite family $\mathcal{D}$ of probability measures on some measurable set (E,ξ) and a random variable $\bar{X}$ with an unknown distribution in the family, we have*

$$\inf_{\hat{T}} \sup_{\mathbb{P}\in\mathcal{D}} \mathbb{P}\big[\hat{T}(\bar{X}) \neq \mathbb{P}\big] \geq 0.36 \wedge \bigg(1 - \frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\bigg),$$

where $K_{\mathcal{D}} \triangleq \inf_{\mathbb{P}\in\mathcal{D}} \sum_{\mathbb{Q}\neq\mathbb{P}} K(\mathbb{Q},\mathbb{P})$ and the infimum is taken over all measurable (possibly randomized) estimators based on $\bar{X}$ with values in the finite set $\mathcal{D}$.

Define $\mathcal{D}' \triangleq \big\{\mathbb{P}_{\vec{\sigma}} : \vec{\sigma} \in \mathcal{S}\big\}$. Let us apply Birgé's lemma to the set of probability distributions

$$\mathcal{D} \triangleq \big\{\mathbb{P}^{\otimes N} : \mathbb{P} \in \mathcal{D}'\big\}.$$

With any estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X},\mathcal{Y})$, we can associate an estimator $\hat{T} : \mathcal{Z}^N \to \mathcal{D}$ defined as $\hat{T}(Z_1^N) = \mathbb{P}^{\otimes N}$, where $\mathbb{P} \in \mathcal{D}'$ minimizes $\mu[\xi(X)\mathbb{1}_{f^*_{\mathbb{P}}(X)\neq\hat{f}(Z_1^N)(X)}]$, where $f^*_{\mathbb{P}}$ denotes the Bayes classifier associated with the distribution $\mathbb{P}$.

By Birgé's lemma, we have $\sup_{\mathbb{Q}\in\mathcal{D}} \mathbb{Q}\big[\hat{T}(Z_1^N) \neq \mathbb{Q}\big] \geq 0.36 \wedge \Big(1 - \frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\Big)$. Now, when $\hat{T}(Z_1^N) \neq \mathbb{P}^{\otimes N}$, we have $\mu(f^*_{\mathbb{P}} \neq \hat{f}) \geq \frac{m}{8}w$, hence $R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \geq \frac{m}{8}w\beta$.

Therefore, we get

$$\sup_{\mathbb{P}^{\otimes N}\in\mathcal{D}} \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \geq \frac{m}{8} w\beta \Big[0.36 \wedge \Big(1 - \frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\Big)\Big]. \tag{6.1}$$

For any $\mathbb{P} \neq \mathbb{Q} \in \mathcal{D}$, we have $K(\mathbb{P},\mathbb{Q}) \leq Nmw\beta \log\big(\frac{1+\beta}{1-\beta}\big)$. Since we have $|\mathcal{D}| = \lfloor e^{\frac{m}{8}} \rfloor$, we obtain $\frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|} \leq \frac{1}{\log\lfloor e^{\frac{m}{8}}\rfloor} Nmw\beta \log\big(\frac{1+\beta}{1-\beta}\big) \leq 20Nw\beta^2$ for m large enough and β small enough. In our case, we have $m = N^{\frac{q}{2\kappa-1+q}}$, $w = N^{-\frac{1+q}{2\kappa-1+q}}$ and $\beta = aN^{-\frac{\kappa-1}{2\kappa-1+q}}$. So for N large enough, we have $\frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|} \leq 20a^2$. Let us choose a such that $20a^2 = 0.64$. We obtain $\sup_{\mathbb{P}\in\mathcal{D}'} \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \geq 0.008N^{-\frac{\kappa}{2\kappa-1+q}}$ for N sufficiently large.

Finally it remains to check that the set of distributions $\mathcal{D}'$ is included in $\mathcal{P}$. For any $\mathbb{P} \in \mathcal{D}'$, the complexity Assumption *(CA2)* is satisfied since

- for $u < mw$, $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) \leq \log|\mathcal{F}| \leq Cu^{-q}$ for some constant $C > 0$.
- for $u \geq mw$, $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) = 0 \leq Cu^{-q}$.

For any $\mathbb{P} \in \mathcal{D}'$, the margin Assumption *(MA3)* is satisfied since for any functions $f \in \mathcal{F} - \{\tilde{f}\}$, $\mathbb{P}_{f,\tilde{f}}$ has the order of $mw = N^{-\frac{1}{2\kappa-1+q}}$ and $\Delta R(f)$ has the order of $mw\beta = aN^{-\frac{\kappa}{2\kappa-1+q}}$. The margin Assumption *(MA1)* also holds since we have

$$\mathbb{P}\left(0 < |\eta^*(x) - \tfrac{1}{2}| \leq t\right) = \begin{cases} 0 & \text{when } t < \beta \\ mw & \text{when } \beta \leq t < \frac{1}{2} \end{cases}$$

Remark 6.2. The proof also holds when $q = 0$. In this case, we take $m = 1$, $w = N^{-\frac{1}{2\kappa-1}}$ and $\beta = \sqrt{\frac{0.64}{20}} N^{-\frac{\kappa-1}{2\kappa-1}}$.

6.3. **Proof of Theorem 3.5.**

6.3.1. *First case:* $\log \pi^{-1}(\Delta R \leq x) = -C' \log x + C''' + o(x^s)$. Since we have[14]

$$\pi_{-\lambda R}\Delta R = \frac{C' + \underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda}, \tag{6.2}$$

from Theorem 2.2, for any $0 \leq \chi < 1$, we get

$$\begin{cases} \pi_{-\lambda r} R & \leq & \frac{C' + \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}(\chi)}{\lambda} + \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \\ \pi_{-\lambda r} R & \geq & \frac{C' + \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}(\chi)}{\lambda} - \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \end{cases}.$$

Taking $\chi = \sqrt{K(\pi_{-\lambda r}, \pi_{-\lambda R})}$, we obtain

$$\lambda \pi_{-\lambda r} R = C' + \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}\Big(\sqrt{K(\pi_{-\lambda r}, \pi_{-\lambda R})}\Big). \tag{6.3}$$

First subcase: $\lambda = \mathrm{o}\big(N^{\frac{\kappa}{2\kappa-1}}\big)$. Assume that $\lambda = \underset{N\to+\infty}{\mathrm{o}}\big(N^{\frac{\kappa}{2\kappa-1}}\big)$. Then there exists $\gamma \in]0; \frac{1}{2}]$ such that $\gamma = \underset{N\to+\infty}{\mathrm{o}}(1)$ and $\lambda\big(\frac{\lambda}{\gamma N}\big)^{\frac{\kappa}{\kappa-1}} = \underset{N\to+\infty}{\mathrm{o}}(1)$. We have[15]

$$\log \pi_{-\lambda R} \exp\Big\{C\frac{\lambda^2}{\gamma N}(\Delta R)^{\frac{1}{\kappa}}\Big\} = \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}\bigg(\Big[\lambda\big(\tfrac{\lambda}{\gamma N}\big)^{\frac{\kappa}{\kappa-1}}\Big]^{\frac{(\kappa-1)C'}{\kappa C'+\kappa-1}}\bigg). \tag{6.4}$$

[14] See Appendix A.

[15] Proof in Appendix B.

Let $L \triangleq \log(e\epsilon^{-1})$. From Theorem 2.2, for any $0 < \gamma < \frac{1}{2}$ and $0 < \lambda \le 0.39\,\gamma N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$\begin{array}{rcl} K(\pi_{-\lambda r}, \pi_{-\lambda R}) & \le & \log \pi_{-\lambda R} \exp\Big\{C\frac{\lambda^2}{\gamma N}(\Delta R)^{\frac{1}{\kappa}}\Big\} + C\gamma L \\ & = & \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}\bigg(\Big[\lambda\big(\frac{\lambda}{N}\big)^{\frac{\kappa}{\kappa-1}}\Big]^{\frac{(\kappa-1)C'}{\kappa C'+\kappa-1}}\gamma^{-\frac{\kappa C'}{\kappa C'+\kappa-1}} + \gamma L\bigg). \end{array}$$

Taking $\gamma = \lambda^{\frac{(2\kappa-1)C'}{2\kappa C'+\kappa-1}} N^{-\frac{\kappa C'}{2\kappa C'+\kappa-1}} L^{-\frac{\kappa C'+\kappa-1}{2\kappa C'+\kappa-1}}$, we obtain
(6.5)

$$K(\pi_{-\lambda r}, \pi_{-\lambda R}) = \underset{N\to+\infty}{\mathrm{O}}\bigg(\lambda^{\frac{(2\kappa-1)C'}{2\kappa C'+\kappa-1}} N^{-\frac{\kappa C'}{2\kappa C'+\kappa-1}} L^{\frac{\kappa C'}{2\kappa C'+\kappa-1}}\bigg) + \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}),$$

which, combined with equality (6.3), gives the desired result.

Second subcase: $\lambda = cN^{\frac{\kappa}{2\kappa-1}}$ *for a small enough* c.

Then the previous computations can be adapted and we obtain that for any $\beta > 0$ there exist $c > 0$ and $N_0 > 0$ such that for any $N > N_0$ with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, :

$$\frac{C'-\beta}{\lambda} \le \pi_{-\lambda r} R \le \frac{C'+\beta}{\lambda}.$$

6.3.2. *Second case:* $\log \pi^{-1}(\Delta R \le x) = C' x^{-\frac{q}{\kappa}} + C''' + \mathrm{o}(1)$. Since we have[16]

$$\pi_{-\lambda R}\Delta R \underset{\lambda\to+\infty}{\sim} \big(\tfrac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}, \tag{6.6}$$

from Theorem 2.2, for any $0 \le \chi < 1$, we get

$$\left\{\begin{array}{rcl} \pi_{-\lambda r} R & \le & \big(\frac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}\big[1 + \underset{N\to+\infty}{\mathrm{o}}(1) + \underset{N\to+\infty}{\mathrm{O}}(\chi)\big] + \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \\ \pi_{-\lambda r} R & \ge & \big(\frac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}\big[1 + \underset{N\to+\infty}{\mathrm{o}}(1) + \underset{N\to+\infty}{\mathrm{O}}(\chi)\big] - \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \end{array}\right. .$$

Taking $\chi = \lambda^{-\frac{q}{2(\kappa+q)}}\sqrt{K(\pi_{-\lambda r}, \pi_{-\lambda R})}$, we obtain

$$\pi_{-\lambda r} R = \big(\tfrac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}\Big[1 + \underset{N\to+\infty}{\mathrm{o}}(1) + \underset{N\to+\infty}{\mathrm{O}}\Big(\lambda^{-\frac{q}{2(\kappa+q)}}\sqrt{K(\pi_{-\lambda r}, \pi_{-\lambda R})}\Big)\Big].$$

First subcase: $\lambda = \mathrm{o}\big(N^{-\frac{\kappa+q}{2\kappa-1+q}}\big)$. From Theorem 2.2, for any $0 < \lambda \le 0.19\,N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$K(\pi_{-\lambda r}, \pi_{-\lambda R}) \le \log \pi_{-\lambda R} \exp\Big\{C\frac{\lambda^2}{N}(\Delta R)^{\frac{1}{\kappa}}\Big\} + C\log(e\epsilon^{-1}).$$

We can prove[17] that for any $\alpha \le \breve{c}\lambda^{-\frac{\kappa-1}{\kappa+q}}$ for $\breve{c}$ small enough, we have

$$\log \pi_{-\lambda R} \exp\Big\{\lambda\alpha(\Delta R)^{\frac{1}{\kappa}}\Big\} = \underset{\lambda\to+\infty}{\mathrm{O}}\Big(\lambda^{\frac{q}{\kappa+q}}\alpha\lambda^{\frac{\kappa-1}{\kappa+q}}\Big), \tag{6.7}$$

Let us assume that $\lambda = \underset{N\to+\infty}{\mathrm{o}}\Big(N^{-\frac{\kappa+q}{2\kappa-1+q}}\Big)$. Then we have $\frac{\lambda}{N}\lambda^{\frac{\kappa-1}{\kappa+q}} = \underset{N\to+\infty}{\mathrm{o}}(1)$, hence

$$K(\pi_{-\lambda r}, \pi_{-\lambda R}) \le \underset{N\to+\infty}{\mathrm{o}}\big(\lambda^{\frac{q}{\kappa+q}}\big) + C\log(e\epsilon^{-1}).$$

[16] See Appendix C.

[17] See Appendix D.

So we obtain that for $\lambda = \underset{N\to+\infty}{\mathrm{o}}\Big(N^{-\frac{\kappa+q}{2\kappa-1+q}}\Big)$,

$$\pi_{-\lambda r}R = \left(\tfrac{qC'}{\kappa\lambda}\right)^{\frac{\kappa}{\kappa+q}}\Big[1 + \underset{N\to+\infty}{\mathrm{o}}(1)\Big].$$

Second subcase: $\lambda = cN^{-\frac{\kappa+q}{2\kappa-1+q}}$ *for a small enough* c.

Once more, the previous computations can be adapted in order to obtain that for any $\beta > 0$ there exist $c > 0$ and $N_0 > 0$ such that for any $N > N_0$ with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, :

$$\left(\tfrac{qC'-\beta}{\kappa\lambda}\right)^{\frac{\kappa}{\kappa+q}} \le \pi_{-\lambda r}R \le \left(\tfrac{qC'+\beta}{\kappa\lambda}\right)^{\frac{\kappa}{\kappa+q}}.$$

6.4. **Proof of Theorem 3.6.** For any $0 \le j \le \log N$, introduce $\lambda_j \triangleq 0.19\sqrt{N}e^{\frac{j}{2}}$. Define $L \triangleq \log[\log(eN)\epsilon^{-1}]$. In [1, Section 3.4.2], an algorithm is proposed to choose the temperature of the standard Gibbs classifier. The associated generalization error is bounded by

$$\mathbb{G} \triangleq \min_{1\le j\le\log N}\left\{\pi_{-\lambda_{j-1}R}R + \frac{\sup_{0\le i\le j}\Big\{\log \pi_{-\lambda_i R}\otimes\pi_{-\lambda_i R}\exp\Big(\frac{C\lambda_i^2}{N}\mathbb{P}_{\cdot,\cdot}\Big)\Big\}}{\lambda_j} + C\frac{L}{\lambda_j}\right\}.$$

Under Assumptions *(MA3)* and *(CA1)*, for any $1 \le j \le \log N$ and $t > 0$, by Jensen's inequality, we have

$$\begin{aligned}
\mathbb{G} \;\le\;& -\frac{\log \pi\exp(-\lambda_{j-1}R)}{\lambda_{j-1}} + \frac{\sup_{0\le i\le j}\Big\{\log \pi_{-\lambda_i R}\exp\Big(\frac{C\lambda_i^2}{N}\mathbb{P}_{\cdot,\tilde f}\Big)\Big\}}{\lambda_j} + C\frac{L}{\lambda_j}\\
\le\;& -\frac{\log \pi\exp(-\lambda_{j-1}R)}{\lambda_{j-1}} + \frac{\sup_{0\le i\le j}\Big\{\log \pi\exp\Big(-\lambda_i\Delta R+\frac{C\lambda_i^2}{N}(\Delta R)^{\frac{1}{\kappa}}\Big)\Big\}}{\lambda_j}\\
&\qquad + \sup_{0\le i\le j}\left\{-\frac{\log\pi\exp(-\lambda_i\Delta R)}{\lambda_j}\right\} + C\frac{L}{\lambda_j}\\
\le\;& R(\tilde f) - 2\sqrt{e}\frac{\log\pi\exp(-\lambda_j R)}{\lambda_j} + \frac{\sup_{0\le i\le j;x\ge 0}\Big\{-\lambda_i x+\frac{C\lambda_i^2}{N}x^{\frac{1}{\kappa}}\Big\}}{\lambda_j} + C\frac{L}{\lambda_j}\\
\le\;& R(\tilde f) - 2\sqrt{e}\frac{\log[\pi(\Delta R\le t)\exp(-\lambda_j t)]}{\lambda_j} + C\Big(\frac{\lambda_j}{N}\Big)^{\frac{\kappa}{\kappa-1}} + C\frac{L}{\lambda_j}\\
\le\;& R(\tilde f) + C\frac{h_q(t^{1/\kappa})}{\lambda_j} + Ct + C\frac{L}{\lambda_j}.
\end{aligned}$$

Taking j such that λ_j is of order $N^{\frac{\kappa+q}{2\kappa-1+q}}$ and t minimizing $C\frac{h_q(t^{1/\kappa})}{\lambda_j} + Ct$, we obtain the desired rates (the ones given in Theorems 3.2 and 3.4). So the algorithm is adaptive wrt the margin parameter κ.

6.5. **Proof of Theorem 3.7.** We will prove the result for a minimal net. It is easy to generalize it to almost minimal nets. Let $u > 0$. Let π be the uniform distribution on a minimal $(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-net denoted $\mathcal{N}_u$. Let $\tilde f_u$ be the the nearest neighbour of $\tilde f$ in the net $\mathcal{N}_u$. Define $a(\lambda) \triangleq \frac{\lambda}{N}g\big(\frac{\lambda}{N}\big)$. From inequality (2.1) for $\big(\rho_2, \pi_2, \rho_1, \pi_1\big) = \big(\delta_{\hat f}, \pi, \delta_{\tilde f_u}, \delta_{\tilde f_u}\big)$, with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-\epsilon$, we have

$$R(\hat f) - R(\tilde f_u) + r(\tilde f_u) - r(\hat f) \le a(\lambda)\mathbb{P}_{\hat f,\tilde f_u} + \frac{H(u,\mathcal{F},\mathbb{P}_{\cdot,\cdot}) + \log(\epsilon^{-1})}{\lambda},$$

When $\hat f = \hat f_{\mathrm{ERM},u}$ minimizes the empirical risk over the net $\mathcal{N}_u$, we obtain

$$R(\hat f) - R(\tilde f_u) \le a(\lambda)\mathbb{P}_{\hat f,\tilde f_u} + \frac{H(u,\mathcal{F},\mathbb{P}_{\cdot,\cdot}) + \log(\epsilon^{-1})}{\lambda},$$

hence

$$R(\hat f) - R(\tilde f) \le R(\tilde f_u) - R(\tilde f) + a(\lambda)\mathbb{P}_{\hat f,\tilde f} + a(\lambda)\mathbb{P}_{\tilde f,\tilde f_u} + \frac{H(u,\mathcal{F},\mathbb{P}_{\cdot,\cdot}) + \log(\epsilon^{-1})}{\lambda}.$$

Let $\breve C > 0$ denote a constant (possibly depending on c'', C'', C', $\breve C_1$ and $\breve C_2$) whose value may differ from line to line. For any $0 < \lambda \le N$, we have

$$\begin{array}{rcl} \Delta R(\hat f) & \le & \breve C\Big(\mathbb{P}^{\kappa}_{\tilde f,\tilde f_u} + \frac{\lambda}{N}\Delta R^{\frac{1}{\kappa}}(\hat f) + \frac{\lambda}{N}\mathbb{P}_{\tilde f,\tilde f_u} + \frac{u^{-q}+\log(\epsilon^{-1})}{\lambda}\Big) \\ & \le & \breve C\Big(\frac{\lambda}{N}\Delta R^{\frac{1}{\kappa}}(\hat f) + u^{\kappa} + \frac{\lambda}{N}u + \frac{u^{-q}+\log(\epsilon^{-1})}{\lambda}\Big). \end{array} \tag{6.8}$$

Let us take u and λ such that $\frac{\lambda}{N}u$, u^{κ} and $\frac{u^{-q}}{\lambda}$ have the same orders. This is realized when Inequalities (3.6) hold and $\lambda = N^{\frac{\kappa+q}{2\kappa-1+q}}$. We obtain

$$\Delta R(\hat f) \le \breve C\big[N^{-\frac{\kappa-1}{2\kappa-1+q}}\Delta R^{\frac{1}{\kappa}}(\hat f) + \log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}\big].$$

Simple computations lead to

$$\Delta R(\hat f) \le \breve C\log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}$$

and, then, to $\mathbb{P}^{\otimes N}\Delta R(\hat f) \le \breve C N^{-\frac{\kappa}{2\kappa-1+q}}$.

6.6. **Proof of Theorem 3.9.** The chaining idea comes from [6] and is well presented also, for instance, in [5, p.19-21]. Let $\partial(\rho_1,\rho_2) \triangleq \rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r$. Let $u_k = u2^k$. Let $c_k \triangleq h_+(u_k)$. To shorten, denote $\pi_{i,k} \triangleq \pi_{f_i,u_k}$. Let K be the nonnegative integer such that $\frac{\mathbb{P}_{f_1,f_2}}{2} \le u_K < \mathbb{P}_{f_1,f_2}$. The integer K exists as soon as $\mathbb{P}_{f_1,f_2} > u$. Let L' be the nonnegative integer such that $\frac{1}{2} \le u_{L'} < 1$. Let $\lambda_1,\dots,\lambda_{L'+1}$ be real positive parameters to be chosen. We apply inequality (2.1) for this $L'+1$ parameters and for $\pi_1 = \pi_2 = \pi$.

With $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-(L'+1)\epsilon$, we have

$$\begin{array}{rl} & \partial(\pi_{1,0},\pi_{2,0}) \\ = & \partial(\pi_{1,K},\pi_{2,K}) + \sum_{k=1}^{K}\big\{\partial(\pi_{1,k-1},\pi_{1,k}) + \partial(\pi_{2,k},\pi_{2,k-1})\big\} \\ \le & \frac{\lambda_{K+1}}{N}g\big(\frac{\lambda_{K+1}}{N}\big)4u_K + \frac{K(\pi_{1,K},\pi)+K(\pi_{2,K},\pi)+\log(\epsilon^{-1})}{\lambda_{K+1}} \\ & +\sum_{k=1}^{K}\Big\{2\frac{\lambda_k}{N}g\big(\frac{\lambda_k}{N}\big)(u_{k-1}+u_k) + \frac{2\log(\epsilon^{-1})+\sum_{i=1}^{2}\sum_{k'=k-1}^{k}K(\pi_{i,k'},\pi)}{\lambda_k}\Big\} \\ \le & \frac{\lambda_{K+1}}{N}g\big(\frac{\lambda_{K+1}}{N}\big)4u_K + \frac{2c_K+\log(\epsilon^{-1})}{\lambda_{K+1}} \\ & +\sum_{k=1}^{K}\Big\{6\frac{\lambda_k}{N}g\big(\frac{\lambda_k}{N}\big)u_{k-1} + \frac{2\log(\epsilon^{-1})+4c_{k-1}}{\lambda_k}\Big\} \\ \le & \sum_{k=1}^{K+1}\Big\{6\frac{\lambda_k}{N}g\big(\frac{\lambda_k}{N}\big)u_{k-1} + \frac{2\log(\epsilon^{-1})+4c_{k-1}}{\lambda_k}\Big\}. \end{array}$$

Let us choose the λ_k's such that they do not depend on ϵ and they roughly minimize the RHS of the last bound. Taking $\lambda_k = \sqrt{\frac{4Nc_{k-1}}{3u_{k-1}}}$ for $k \ge 1$, we obtain

$$\begin{array}{rcl} \partial(\pi_{1,0},\pi_{2,0}) & \le & \sum_{k=1}^{K+1}\big[1+2g\big(\frac{\lambda_k}{N}\big)\big]\sqrt{\frac{12c_{k-1}u_{k-1}}{N}} + \sum_{k=1}^{K+1}\frac{2\log(\epsilon^{-1})}{\lambda_k} \\ & \le & \sum_{k=1}^{K+1}\big[1+2g\big(\frac{\lambda_k}{N}\big)\big]\sqrt{\frac{12c_{k-1}u_{k-1}}{N}} + 2\log(\epsilon^{-1})\sqrt{\frac{3u}{4N}}\sum_{k=1}^{K+1}\sqrt{2}^{k-1}. \end{array}$$

For any $k \in \{1, \dots, L'+1\}$, we have $\frac{\lambda_k}{N} \leq \sqrt{\frac{4c_0}{3Nu}} \triangleq C_1$. Since $C_0 = 2\sqrt{3}[1+2g(C_1)]$, we obtain

$$\begin{aligned}\partial(\pi_{1,0}, \pi_{2,0}) &\leq \frac{2C_0}{\sqrt{N}} \sum_{k=1}^{K+1} (u_{k-1} - u_{k-2}) \sqrt{\frac{c_{k-1}}{u_{k-1}}} + \sqrt{\frac{6\mathbb{P}_{f_1,f_2}}{N}} \frac{\log(\epsilon^{-1})}{\sqrt{2}-1} \\ &\leq \frac{2C_0}{\sqrt{N}} \int_{u/2}^{\mathbb{P}_{f_1,f_2}} \sqrt{\frac{h_+(v)}{v}} dv + 6\sqrt{\frac{\mathbb{P}_{f_1,f_2}}{N}} \log(\epsilon^{-1}).\end{aligned}$$

Remark 6.3. We have used the "global bayesian entropy" $h(v) \triangleq \sup_{f\in\mathcal{F}} \log \pi^{-1}(B_{f,v})$ since it was convenient to have an (almost) optimal λ's which do not depend on the functions f_1 and f_2. Had we done a union bound on the parameters λ, we would have been able to make it depend on the functions f_1, f_2. Then the global bayesian entropy would have been replaced with the local ones $h(v, f_1)$ and $h(v, f_2)$ where $h(v, f) \triangleq \log \pi^{-1}(B_{f,v})$. In other words, the quantity $\partial(\pi_{1,0}, \pi_{2,0})$ is mainly driven by the two integrals $\int_{u/2}^{\mathbb{P}_{f_1,f_2}} \sqrt{\frac{h(v,f_1)}{vN}} dv$ and $\int_{u/2}^{\mathbb{P}_{f_1,f_2}} \sqrt{\frac{h(v,f_2)}{vN}} dv$.

6.7. **Proof of Theorem 3.10.**

6.7.1. *First step: upper bounds due to the chaining technique.* We start with the following chained result which is slightly different from Theorem 3.9 to the extent that we chained functions belonging to covering nets instead of chaining balls. Had we been interested in results for packing nets, Theorem 3.9 applied to an appropriate prior distribution[18] would have been sufficient. Let $H(u) \triangleq H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$.

Theorem 6.5. *Let $u > 0$, $\mathcal{N}$ a minimal $u-$covering net and $L \triangleq \frac{\log(2u^{-1})}{\log 2}$. We have*

- *for any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, for any $f_1, f_2 \in \mathcal{N}_u$,*

$$\begin{aligned}&R(f_2) - R(f_1) + r(f_1) - r(f_2) \\ &\qquad \leq 8\sqrt{\frac{3}{N}} \int_{u/2}^{\mathbb{P}_{f_1,f_2}\vee u} \sqrt{\frac{H(v)}{v}} dv + \frac{8}{3N} \int_{u/2}^{\mathbb{P}_{f_1,f_2}\vee u} \frac{H(v)}{v} dv \\ &\qquad\quad + 17\sqrt{\frac{\log(3L\epsilon^{-1})}{N}} \sqrt{\mathbb{P}_{f_1,f_2} \vee u} + \frac{2L\log(3L\epsilon^{-1})}{3N},\end{aligned} \tag{6.9}$$

- *for any $\epsilon > 0$, for any $f_1 \in \mathcal{N}_u$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, for any $f_2 \in \mathcal{N}_u$,*

$$\begin{aligned}&R(f_2) - R(f_1) + r(f_1) - r(f_2) \\ &\qquad \leq 4\sqrt{\frac{3}{N}} \int_{u/2}^{\mathbb{P}_{f_1,f_2}\vee u} \sqrt{\frac{H(v)}{v}} dv + \frac{4}{3N} \int_{u/2}^{\mathbb{P}_{f_1,f_2}\vee u} \frac{H(v)}{v} dv \\ &\qquad\quad + 8.5\sqrt{\frac{\log(2L\epsilon^{-1})}{N}} \sqrt{\mathbb{P}_{f_1,f_2} \vee u} + \frac{L\log(2L\epsilon^{-1})}{3N}\end{aligned}$$

Proof. The proof is similar to the one of Theorem 3.9. Instead of chaining balls, we will chain on covering nets. Let $\partial(f_1, f_2) \triangleq R(f_2) - R(f_1) + r(f_1) - r(f_2)$, $u_k = u2^k$ and $c_k \triangleq H_+(u_k)$. Introduce $P \triangleq \mathbb{P}_{f_1,f_2} \vee u$ and let $0 \leq K \leq L'$ be integers such that $\frac{\mathbb{P}_{f_1,f_2}}{2} < u_K \leq P$ and $\frac{1}{2} < u_{L'} \leq 1$.

Consider the family $\big(\mathcal{N}_k\big)_{k=\{0,\dots,L'\}}$ of minimal nets of radius u_k. For any $(j, k) \in \{1, 2\} \times \{0, \dots, L'\}$, introduce $f_{j,k} \in \operatorname{argmin}_{\mathcal{N}_{u2^k}} \mathbb{P}_{\cdot,f_j}$ a nearest neighbour of f_j in

[18] Let $\mathcal{N}_{\mathrm{p}}$ be a u-packing net. Using the notation of Section 6.6, an appropriate prior distribution is $\pi = \frac{1}{L'+1} \sum_{k=0}^{L'} \pi_k$, where π_k is the uniform distribution on a u_k-minimal packing net of the set $\mathcal{F}$ built using points in $\mathcal{N}_{\mathrm{p}}$. The log-cardinal of such a set is upper bounded by $H(u_{k-1}, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$, hence $h(u_k) \leq H(u_{k-1}, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) + \log(L' + 1)$.

$\mathcal{N}_{u2^k}$. Since $f_1, f_2 \in \mathcal{N}_u$, we have $f_{1,0} = f_1$ and $f_{2,0} = f_2$. Let π_k be the uniform distribution on the net $\mathcal{N}_k$.

Let $l \triangleq \log[(3L'+1)\epsilon^{-1}]$. By applying $3L'+1$ times inequality (2.3), we obtain that with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any functions f_1, f_2 in $\mathcal{N}_u$, we have

(6.10)

$$\begin{array}{rcl}
\partial(f_1,f_2) & = & \partial(f_{1,K},f_{2,K}) + \sum_{k=1}^{K} \big\{ \partial(f_{1,k-1},f_{1,k}) + \partial(f_{2,k},f_{2,k-1}) \big\} \\
& \leq & \sqrt{\frac{2[2H(u_K)+l]\mathbb{P}_{f_{1,K},f_{2,K}}}{N}} + \frac{2H(u_K)+l}{3N} \\
& & + \sum_{k=1}^{K} \Big\{ \sqrt{\frac{2[H(u_{k-1})+H(u_k)+l]\mathbb{P}_{f_{1,k-1},f_{1,k}}}{N}} + \frac{H(u_{k-1})+H(u_k)+l}{3N} \\
& & \qquad + \sqrt{\frac{2[H(u_{k-1})+H(u_k)+l]\mathbb{P}_{f_{2,k-1},f_{2,k}}}{N}} + \frac{H(u_{k-1})+H(u_k)+l}{3N} \Big\} \\
& \leq & 2 \sum_{k=1}^{K+1} \Big\{ \sqrt{\frac{6[2H(u_{k-1})+l]u_{k-1}}{N}} + \frac{2H(u_{k-1})+l}{3N} \Big\} \\
& \leq & 2 \sum_{k=1}^{K+1} \Big\{ \sqrt{\frac{12H(u_{k-1})u_{k-1}}{N}} + \sqrt{\frac{6lu_{k-1}}{N}} + \frac{2H(u_{k-1})}{3N} \Big\} + \frac{2(K+1)l}{3N} \\
& \leq & 2\sqrt{\frac{6lu}{N}} \frac{\sqrt{2}^{K+1}}{\sqrt{2}-1} + \frac{2(K+1)l}{3N} + 2\sum_{k=1}^{K+1} \Big\{ \sqrt{\frac{12H(u_{k-1})u_{k-1}}{N}} + \frac{2H(u_{k-1})}{3N} \Big\}
\end{array}$$

Now the last sum can be upper bounded using integrals since the function $v \mapsto H(v)$ is non increasing on $\mathbb{R}_+^*$. We obtain

$$\partial(f_1,f_2) \leq \frac{4}{\sqrt{2}-1}\sqrt{\frac{3lP}{N}} + \frac{2l}{3N}\frac{\log(\frac{2P}{u})}{\log 2} + 8\sqrt{\frac{3}{N}} \int_{u/2}^{P} \sqrt{\frac{H(v)}{v}} dv + \frac{8}{3N}\int_{u/2}^{P} \frac{H(v)}{v} dv.$$

For the second part of Theorem 6.5, it suffices to modify slightly the previous argument. This time, the functions $f_{2,k}$ are defined as previously. The functions $f_{1,k}$ are defined as $f_{1,k} \triangleq f_1$. Therefore we have $\partial(f_{1,k-1}, f_{1,k}) = 0$, hence the modification of the constants. □

Consider that Assumption *(CA1)* holds. Let $c_q > 0$ such that for any $0 < u \leq 1$, $\sum_{k=0}^{L'} 3e^{-c_q h_q(u_k)} \leq 1$. In the previous proof, we used a uniform union bound over the $3L'+1$ inequalities coming from (2.3). If we are just interested in the order of the bounds, we can weight the inequalities associated with $\partial(f_{1,k-1}, f_{1,k})$ and $\partial(f_{2,k}, f_{2,k-1})$ with $e^{-c_q h_q(u_{k-1})}$ and those corresponding to $\partial(f_{2,k}, f_{2,k-1})$ with at least weight $e^{-c_q h_q(u_k)}$.

Then, in Inequalities (6.10), we may replace $2H(u_K)+l$ and $H(u_{k-1})+H(u_k)+l$ with respectively $2H(u_K)+c_q h_q(u_K)+\log(\epsilon^{-1})$ and $H(u_{k-1})+H(u_k)+c_q h_q(u_{k-1})+\log(\epsilon^{-1})$, so that we obtain

(6.11)

$$\begin{array}{rcl}
\partial(f_1,f_2) & \leq & \breve{C} \sum_{k=1}^{K+1} \Big\{ \sqrt{\frac{[h_q(u_{k-1})+\log(\epsilon^{-1})]u_{k-1}}{N}} + \frac{h_q(u_{k-1})+\log(\epsilon^{-1})}{N} \Big\} \\
& \leq & \breve{C}\sqrt{\frac{P\log(\epsilon^{-1})}{N}} + \breve{C}\frac{\log(eu^{-1})\log(\epsilon^{-1})}{N} + \breve{C}\int_{u/2}^{P} \Big(\sqrt{\frac{h_q(v)}{Nv}} + \frac{h_q(v)}{Nv} \Big) dv.
\end{array}$$

Corollary 6.6. *Let $\mathcal{N}$ denote a minimal $u-$net, where u is a positive real. Define $\mathcal{U} \triangleq \sup_{f:\mathbb{P}_{f,\tilde{f}} \leq u} \big\{ R(f) - R(\tilde{f}) \big\}$. Introduce a function $\tilde{f}_{\mathcal{N}} \in \mathcal{N}$ such that $\mathbb{P}_{\tilde{f}_{\mathcal{N}},\tilde{f}} \leq u$. Let $\gamma_u :]u;1] \to \mathbb{R}$ and $\Gamma_u :]u;1] \to \mathbb{R}$ be non decreasing concave functions respectively upper bounding the functions $\int_{\frac{u}{2}}^{\cdot} \sqrt{\frac{H(v)}{v}} dv$ and $\int_{\frac{u}{2}}^{\cdot} \frac{H(v)}{v} dv$..*

For any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any function $f \in \mathcal{N}$, we have

$$\begin{aligned} R(f) - R(\tilde{f}) \quad \le \quad & r(f) - r(\tilde{f}_{\mathcal{N}}) + \tfrac{\breve{C}}{\sqrt{N}}\Big\{\gamma_u(\mathbb{P}_{f,\tilde{f}} + u) + \sqrt{(\mathbb{P}_{f,\tilde{f}} + u)\log(\epsilon^{-1})}\Big\} \\ & + \tfrac{\breve{C}}{N}\Big\{\Gamma_u\big(\mathbb{P}_{f,\tilde{f}} + u\big) + \log(eu^{-1})\log(\epsilon^{-1})\Big\} + \mathcal{U} \end{aligned}$$

Consequently, for any $\rho \in \mathcal{M}^1_+(\mathcal{N})$, we have

$$\begin{aligned} \rho R - R(\tilde{f}) \quad \le \quad & \rho r - r(\tilde{f}_{\mathcal{N}}) + \tfrac{\breve{C}}{\sqrt{N}}\Big\{\gamma_u(\rho\mathbb{P}_{\cdot,\tilde{f}} + u) + \sqrt{(\rho\mathbb{P}_{\cdot,\tilde{f}} + u)\log(\epsilon^{-1})}\Big\} \\ & + \tfrac{\breve{C}}{N}\Big\{\Gamma_u\big(\rho\mathbb{P}_{\cdot,\tilde{f}} + u\big) + \log(eu^{-1})\log(\epsilon^{-1})\Big\} + \mathcal{U} \end{aligned}$$

Proof. The first inequality comes mainly from inequalities (6.11) and the decomposition: $R(f) - R(\tilde{f}) = R(f) - R(\tilde{f}_{\mathcal{N}}) + R(\tilde{f}_{\mathcal{N}}) - R(\tilde{f})$. The second inequality is then deduced from Jensen's inequality. □

6.7.2. *Second step: determining the radius of the nets.* Corollary 6.6 implies that for any $\epsilon > 0$, for any classifier $\hat{f}$ minimizing the empirical risk over the net $\mathcal{N}$, with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-\epsilon$, we have

$$\begin{aligned} \Delta R(\hat{f}) \quad \le \quad & \mathcal{U} + \tfrac{\breve{C}}{\sqrt{N}}\Big\{\gamma_u\big(\mathbb{P}_{\hat{f},\tilde{f}} + u\big) + \sqrt{\big(\mathbb{P}_{\hat{f},\tilde{f}} + u\big)\log(\epsilon^{-1})}\Big\} \\ & + \tfrac{\breve{C}}{N}\Big\{\Gamma_u\big(\mathbb{P}_{\hat{f},\tilde{f}} + u\big) + \log(eu^{-1})\log(\epsilon^{-1})\Big\}. \end{aligned} \tag{6.12}$$

Now we have

$$\mathcal{U} \le \begin{cases} \breve{C}u^{\kappa} & \text{under Assumption } \textit{(MA3)} \\ 2u & \text{in any case} \end{cases},$$

and we can take

$$\gamma_u(x) \triangleq \begin{cases} \breve{C}\sqrt{\log(e^2x^{-1})x} & \text{under Assumption } \textit{(CA1)} \text{ for } q = 0 \\ \breve{C}x^{\frac{1-q}{2}} & \text{under Assumption } \textit{(CA1)} \text{ for } 0 < q < 1 \\ \breve{C}\log\left(\frac{2x}{u}\right) & \text{under Assumption } \textit{(CA1)} \text{ for } q = 1 \\ \breve{C}u^{\frac{1-q}{2}} & \text{under Assumption } \textit{(CA1)} \text{ for } q > 1 \end{cases}$$

and

$$\Gamma_u(x) \triangleq \begin{cases} \breve{C}[\log(eu^{-1})]^2 & \text{under Assumption } \textit{(CA1)} \text{ for } q = 0 \\ \breve{C}u^{-q} & \text{under Assumption } \textit{(CA1)} \text{ for } q > 0 \end{cases}.$$

Then we have eight cases corresponding to the different complexity and margin assumptions. When we have $q > 0$, inequality (6.12) implies

$$\Delta R(\hat{f}) \le \mathcal{U} + \breve{C}\tfrac{\log(e\epsilon^{-1})}{\sqrt{N}}\gamma_u\big(\mathbb{P}_{\hat{f},\tilde{f}} + u\big) + \breve{C}\tfrac{\log(e\epsilon^{-1})}{N}u^{-q}.$$

Under Assumptions (MA2) *and* (CA1) *for $q = 0$*

Let $\Delta \triangleq \Delta R(\hat{f})$ to shorten. Inequality (6.12) becomes

$$\Delta \le \breve{C}\Big[\log(e\epsilon^{-1})N^{-\frac{1}{2}}\Big(\sqrt{\log(e^2\Delta^{-\frac{1}{\kappa}})}\Delta^{\frac{1}{2\kappa}} + \sqrt{\log(e^2u^{-1})u}\Big) + u + \tfrac{[\log(eu^{-1})]^2}{N}\Big].$$

We obtain $\Delta \le \breve{C}\log(e\epsilon^{-1})(\log[eN^{1/\kappa}])^{\frac{\kappa}{2\kappa-1}}N^{-\frac{\kappa}{2\kappa-1}}$ when[19]

$$\sqrt{\tfrac{\log(e^2u^{-1})u}{N}} + u + \tfrac{[\log(eu^{-1})]^2}{N} \le \breve{C}(\log[eN^{1/\kappa}])^{\frac{\kappa}{2\kappa-1}}N^{-\frac{\kappa}{2\kappa-1}},$$

[19] We use $\log[eN^{1/\kappa}]$ since the logarithmic factor disappears for $\kappa = +\infty$. For $\kappa < +\infty$, the factor $\log[eN^{1/\kappa}]$ can be simplified into $\log N$ for $N \ge 2$.

hence when there exists $\check{C}_1, \check{C}_2 > 0$ such that

$$\exp\Big\{ -\check{C}_1 (\log[eN^{1/\kappa}])^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}} \Big\} \le u \le \check{C}_2 (\log[eN^{1/\kappa}])^{\frac{\kappa}{2\kappa-1}} N^{-\frac{\kappa}{2\kappa-1}}.$$

Under Assumptions (MA3) *and* (CA1) *for* $q = 0$ *and* $\kappa < +\infty$

Inequality (6.12) gives

$$\Delta \le \check{C}\Big[\log(e\epsilon^{-1}) N^{-\frac{1}{2}} \Big(\sqrt{\log(e^2\Delta^{-\frac{1}{\kappa}})}\Delta^{\frac{1}{2\kappa}} + \sqrt{\log(e^2u^{-1})u} \Big) + u^{\kappa} + \tfrac{[\log(eu^{-1})]^2}{N} \Big].$$

We obtain $\Delta \le \check{C} \log(e\epsilon^{-1})(\log N)^{\frac{\kappa}{2\kappa-1}} N^{-\frac{\kappa}{2\kappa-1}}$ when

$$\sqrt{\tfrac{\log(e^2u^{-1})u}{N}} + u^{\kappa} + \tfrac{[\log(eu^{-1})]^2}{N} \le \check{C}(\log N)^{\frac{\kappa}{2\kappa-1}} N^{-\frac{\kappa}{2\kappa-1}},$$

so when there exists $\check{C}_1, \check{C}_2 > 0$ such that

$$\exp\Big\{ -\check{C}_1 (\log N)^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}} \Big\} \le u \le \check{C}_2 (\log N)^{\frac{1}{2\kappa-1}} N^{-\frac{1}{2\kappa-1}}.$$

Under Assumptions (MA2) *and* (CA1) *for* $0 < q < 1$

Inequality (6.12) becomes

$$\Delta \le \check{C}\Big[u + \log(e\epsilon^{-1}) N^{-\frac{1}{2}} \Big(\Delta^{\frac{1-q}{2\kappa}} + u^{\frac{1-q}{2}} \Big) + \check{C} \frac{\log(e\epsilon^{-1})}{N} u^{-q} \Big].$$

This leads to $\Delta \le \check{C} \log(e\epsilon^{-1}) N^{-\frac{\kappa}{2\kappa-1+q}}$ when the inequality

$$N^{-\frac{1}{2}} u^{\frac{1-q}{2}} + u + \check{C} \tfrac{\log(e\epsilon^{-1})}{N} u^{-q} \le \check{C} N^{-\frac{\kappa}{2\kappa-1+q}}$$

holds, hence when there exist $\check{C}_1, \check{C}_2$ such that $\check{C}_1 N^{-\frac{\kappa-1+q}{q(2\kappa-1+q)}} \le u \le \check{C}_2 N^{-\frac{\kappa}{2\kappa-1+q}}$.

Similarly, we deal with the five other cases. To finish the proof, we just have to notice that, when for any $\epsilon > 0$ and some real function ϕ, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have $\Delta \le \log(e\epsilon^{-1})\phi(N)$, then we have $\mathbb{P}^{\otimes N}\Delta \le 2\phi(N)$.

Remark 6.4. Once more, for sake of simplicity, we have done the proof for minimal nets without explicit values of the constants. It is easy to adapt the proof to almost minimal nets and to get an explicit constant $\check{C}$ in terms of the other constants of the problem.

6.8. **Proof of Theorem 3.15.** Let $u_N = \check{C}_2 w_N$. Let $\mathcal{N}'$ be a u_N-minimal bracketing net of the model $\mathcal{F}$. Let $A \triangleq \{f \in \mathcal{F} : \mathbb{P}_{f,\tilde{f}} \le u_N\}$. There exists a posterior distribution $\hat{\rho}_{\mathcal{N}'} : \mathcal{Z}^N \to \mathcal{M}_+^1(\mathcal{N}')$ (for instance, $\hat{\rho}_{\mathcal{N}'} \triangleq \pi_{-\lambda_N r} \circ n_L^{-1}$) such that we have $\hat{\rho}_{\mathcal{N}'} r \le \pi_{-\lambda_N r} r + u_N$ and

$$\pi_{-\lambda_N r} R \le \hat{\rho}_{\mathcal{N}'} R + u_N. \tag{6.13}$$

We have

$$\hat{\rho}_{\mathcal{N}'} r \le \pi_{-\lambda_N r} r + u_N \le \pi|_A r + \frac{K(\pi|_A, \pi)}{\lambda_N} + u_N \tag{6.14}$$

and

$$K(\pi|_A, \pi) = \log[\pi(A)^{-1}] \le \check{C} h_q(u_N) \le \check{C}\lambda_N u_N. \tag{6.15}$$

From inequality (2.1) for $\big(\rho_2, \pi_2, \rho_1, \pi_1\big) = \big(\delta_{\tilde{f}}, \delta_{\tilde{f}}, \pi|_A, \pi|_A\big)$ and $\lambda = N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have $\pi|_A r - r(\tilde{f}) \le \pi|_A R - R(\tilde{f}) + u_N + \frac{\log(\epsilon^{-1})}{N}$, hence

$$\pi|_A r - r(\tilde{f}) \le 2u_N + \frac{\log(\epsilon^{-1})}{N} \le C\log(e\epsilon^{-1})w_N. \tag{6.16}$$

Combining Inequalities (6.14), (6.15) and (6.16), with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we obtain $\hat{\rho}_{\mathcal{N}'} r \le r(\tilde{f}) + \breve{C}\log(e\epsilon^{-1})w_N$. The result follows from Theorem 3.12 and inequality (6.13).

6.9. **Proof of Theorem 3.16.** Let $\breve{C}_1 > 0$, $u \ge \breve{C}_1\beta_q$ and $\mathcal{N}$ be a $\big(u, \mathcal{F}, \mathbb{P}\big)$-minimal bracketing net. Let π be the uniform distribution on this net. From inequality (8.2) in [1] for $\mathcal{W}\big((f_1, f_2), X\big) = \mathbb{1}_{f_1(X)\neq f_2(X)}$ and $\nu = \pi\otimes\pi$, we obtain that with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any function f_1', f_2' in the net $\mathcal{N}$ and any $\lambda > 0$, we have

$$\bar{\mathbb{P}}_{f_1',f_2'} \le \big[1 + \tfrac{\lambda}{N}g\big(\tfrac{\lambda}{N}\big)\big]\mathbb{P}_{f_1',f_2'} + \tfrac{2\log|\mathcal{N}| + \log(\epsilon^{-1})}{\lambda}$$

Recall that $\breve{C}$ is a constant (possibly depending on the other constants of the problem) which value may differ from line to line. Taking $\epsilon = (\alpha_q)^C$ and $\lambda = N$, we obtain that with probability at least $1-(\alpha_q)^C$, for any functions f_1, f_2 in the set $\mathcal{F}$, we have

$$\begin{array}{rcl}
\bar{\mathbb{P}}_{f_1,f_2} & \le & \bar{\mathbb{P}}_{f_1,n_L(f_1)} + \bar{\mathbb{P}}_{n_L(f_1),n_L(f_2)} + \bar{\mathbb{P}}_{n_L(f_2),f_2} \\
& \le & \bar{\mathbb{P}}_{n_L(f_1),n_U(f_1)} + \bar{\mathbb{P}}_{n_L(f_1),n_L(f_2)} + \bar{\mathbb{P}}_{n_L(f_2),n_U(f_2)} \\
& \le & (e-1)\mathbb{P}_{n_L(f_1),n_L(f_2)} + \frac{6C'h_q(\breve{C}_1\beta_q) + 3C\log(\alpha_q^{-1})}{N} + 2u \\
& \le & (e-1)\mathbb{P}_{n_L(f_1),n_L(f_2)} + \breve{C}\beta_q + 2u \\
& \le & (e-1)\mathbb{P}_{f_1,f_2} + \breve{C}u.
\end{array}$$

By applying inequality (8.2) in [1] to $\mathcal{W}\big((f_1, f_2), X\big) = -\mathbb{1}_{f_1(X)\neq f_2(X)}$, we can similarly proved that with probability at least $1-(\alpha_q)^C$, for any functions f_1, f_2 in the set $\mathcal{F}$, we have $(3-e)\mathbb{P}_{f_1,f_2} \le \bar{\mathbb{P}}_{f_1,f_2} + \breve{C}u$. (The constants $e-1$ and $3-e$ have nothing fundamental and we can make them as close as 1 as we want provided that we change the other constants.) These two inequalities allows to prove respectively the first two items of the theorem for one radius $u \ge \breve{C}_1\beta_q$. To get a result uniform wrt the radius, it suffices to make a union bound for radius in a geometric grid of $[\breve{C}_1\beta_q; 1]$.

For the last item, when we have $u \ge \breve{C}_1\beta_q$ for a sufficiently large $\breve{C}_1$, there exists a small constant $\breve{C}'$ satisfying

$$H(u, \mathcal{F}, \bar{\mathbb{P}}) \le H_{\mathrm{p}}(u, \mathcal{F}, \bar{\mathbb{P}}) \le H_{\mathrm{p}}(\breve{C}'u, \mathcal{F}, \mathbb{P}) \le H(\breve{C}'u/2, \mathcal{F}, \mathbb{P}) \le \breve{C}h_q(u).$$

6.10. **Proof of Theorem 4.1.** The proof is adapted from the proof of the concentration of $N(X_1^N)$ [10, p.42]. First, we prove that for any $k \in \mathbb{N}$, the quantity $\frac{H(\frac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})}{\log 2}$ is a self-bounded quantity in the sense given in [10, p.23]. Let $\mathcal{N}_k$ be a $\big(\frac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}\big)$-minimal net. Define the probability distribution

$$\bar{\mathbb{P}}^{(i)} \triangleq \tfrac{\delta_{Z_1} + \cdots + \delta_{Z_{i-1}} + \delta_{Z_{i+1}} + \cdots + \delta_{Z_N}}{N-1}.$$

Let $H^{(i)}$ be the logarithm of the cardinal of the smallest $\big(\frac{k}{N-1}, \mathcal{F}, \bar{\mathbb{P}}^{(i)}\big)$-net using only functions in the net $\mathcal{N}_k$. We have

$$0 \le H(\tfrac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot}) - H^{(i)} \le H(\tfrac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot}) - H(\tfrac{k}{N-1}, \mathcal{F}, \bar{\mathbb{P}}^{(i)}) \le \log 2.$$

Let $\mathbb{V} = \big(f(X_1), \dots, f(X_N)\big)$ be the random vector induced by the uniform distribution on the net $\mathcal{N}_k$. The Shannon entropy of this vector is $\log|\mathcal{N}_k| = H(\frac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})$. Define $\mathbb{V}^{(i)} \triangleq \big(f(X_1), \dots, f(X_{i-1}), f(X_{i+1}), \dots, f(X_N)\big)$. Since the uniform distribution maximizes the entropy, we have $H^{(i)} \geq H(\mathbb{V}^{(i)})$. Then, using Han's inequality (see for instance [10, p.31]), we obtain that $\frac{H(\frac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})}{\log 2}$ is a self-bounded quantity.

Therefore, we can apply Corollary 5 in [10, p.43]. To shorten, we write until the end of the proof $H(k)$ for $H(\frac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})$. For any $\eta > 0$, we have

$$\mathbb{P}^{\otimes N}\Big[H(k) \geq \mathbb{P}^{\otimes N} H(k) + (\log 2)\eta\Big] \leq e^{-\frac{\eta^2}{\frac{2\mathbb{P}^{\otimes N} H(k)}{\log 2} + \frac{2\eta}{3}}} \tag{6.17}$$

and

$$\mathbb{P}^{\otimes N}\Big[H(k) \leq \mathbb{P}^{\otimes N} H(k) - (\log 2)\eta\Big] \leq e^{-\frac{\eta^2}{2\mathbb{P}^{\otimes N} H(k)}}. \tag{6.18}$$

Introducing $\epsilon = e^{-\frac{\eta^2}{\frac{2\mathbb{P}^{\otimes N} H(k)}{\log 2} + \frac{2\eta}{3}}}$, equivalently

$$\eta^2 - \bigg(\frac{2\mathbb{P}^{\otimes N} H(k)}{\log 2} + \frac{2\eta}{3}\bigg)\log(\epsilon^{-1}) = 0,$$

$$\eta = \frac{\log(\epsilon^{-1})}{3}\left(1 + \sqrt{1 + \frac{18\mathbb{P}^{\otimes N} H(k)}{(\log 2)\log(\epsilon^{-1})}}\right),$$

we obtain that for any $\epsilon > 0$,

$$\mathbb{P}^{\otimes N}\left\{H(k) \geq \mathbb{P}^{\otimes N} H(k) + \frac{(\log 2)\log(\epsilon^{-1})}{3}\left(1 + \sqrt{1 + \frac{18\mathbb{P}^{\otimes N} H(k)}{(\log 2)\log(\epsilon^{-1})}}\right)\right\} \leq \epsilon,$$

which is the first assertion of the lemma. The second inequality of the lemma is a direct consequence of inequality (6.18). The following two inequalities in the lemma comes similarly from inequality (6.17). Finally, inequality (4.3) comes from combining Inequalities (4.1) and (4.2).

6.11. **Proof of Theorem 4.3.** For any $k \in \{0, \dots, U\}$, let $\mathcal{N}_k$ be a 2^{-k}-minimal covering net of $\mathcal{F}$ for the pseudo-distance $\bar{\bar{\mathbb{P}}}$. For any $(i,k) \in \{1,2\} \times \{0, \dots, U\}$, let $f_{i,k}$ be a nearest neighbour of f_i in the set $\mathcal{N}_k$. Let $0 \leq K \leq U$ be the integer satisfying $\frac{\bar{\bar{\mathbb{P}}}_{f_1,f_2} \vee u}{2} < 2^{-K} \leq \bar{\bar{\mathbb{P}}}_{f_1,f_2} \vee u$.

Since we have

$$\partial(f_1; f_2) = \partial(f_{1,K}; f_{2,K}) + \textstyle\sum_{k=K+1}^{U} \Big\{\partial(f_{1,k}; f_{1,k-1}) + \partial(f_{2,k-1}; f_{2,k})\Big\},$$

we need to apply Lemma 4.2 to $(\mathcal{S}_1, \mathcal{S}_2) \in \cup_{1 \leq k \leq U}\big\{(\mathcal{N}_{k-1}, \mathcal{N}_k) \cup (\mathcal{N}_k, \mathcal{N}_{k-1}) \cup (\mathcal{N}_k, \mathcal{N}_k)\big\}$ and to do a union bound on the associated $3U$ inequalities. Let w_k, $k \in \mathbb{N}^*$ be positive integers such that $\sum_{k \geq 1} w_k = 1$. With probability at least $1 - \epsilon$, for any $k \in \mathbb{N}^*$, for any $(f_1', f_2') \in (\mathcal{N}_{k-1} \times \mathcal{N}_k) \cup (\mathcal{N}_k \times \mathcal{N}_{k-1}) \cup (\mathcal{N}_k \times \mathcal{N}_k)$, we have

$$\partial(f_1'; f_2') \leq \sqrt{\tfrac{8\bar{\bar{\mathbb{P}}}_{f_1', f_2'} \log(3|\mathcal{N}_k|^2 w_k^{-1} \epsilon^{-1})}{N}}$$

For any $(i,k) \in \{1,2\} \times \{1, \dots, U\}$, we have $\mathbb{P}_{f_{i,k-1}, f_{i,k}} \leq 3 \times 2^{-k}$. Denote

$$B_k \triangleq \sqrt{\tfrac{24 \times 2^{-k} \log(3|\mathcal{N}_k|^2 w_k^{-1} \epsilon^{-1})}{N}}$$

We have $\mathbb{P}_{f_{1,K},f_{2,K}} \leq 4 \times 2^{-K}$ and $f_{1,0} = f_{2,0}$. Chaining the inequalities, we obtain that, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$,

$$\partial(f_1; f_2) \leq 2B_K \mathbb{1}_{K>0} + 2\textstyle\sum_{k=K+1}^{U} B_k \leq 2\textstyle\sum_{k=K\vee 1}^{U} B_k,$$

hence

$$\partial(f_1; f_2) \leq 4\textstyle\sum_{k=K\vee 1}^{U} \sqrt{\frac{6\times 2^{-k}[2H_k+\log(3w_k^{-1})+\log(\epsilon^{-1})]}{N}}.$$

We want that the union bound term $\log(3w_k^{-1})$ remains negligible wrt the complexity term $2H_k$. This leads to choose, for instance, $w_k = \frac{1}{k(k+1)}$ since $\sum_{k\geq 1} \frac{1}{k(k+1)} = 1$ and for small classes of functions (i.e. VC-classes), the entropy H_k has the order of k, hence $\log(3w_k^{-1}) \ll H_k$. We obtain

$$\partial(f_1; f_2) \leq \sum_{k\in\mathbb{N}^*: u \leq 2^{-k} \leq \bar{\bar{\mathbb{P}}}_{f_1,f_2}\vee u} 4\sqrt{\frac{6\times 2^{-k}\{2H_k+\log[3k(k+1)]+\log(\epsilon^{-1})\}}{N}}.$$

Remark 6.5. The previous result can also be written in terms of integral. Introduce the set $E \triangleq \{k \in \mathbb{N} : 2^{-k} \leq (\bar{\bar{\mathbb{P}}}_{f_1,f_2} \vee u) \wedge \frac{1}{2}\}$ and take $H_k = H(2^{-k}, \mathcal{F}, \bar{\bar{\mathbb{P}}}) \vee 1$. We get

$$\begin{aligned}
\partial(f_1; f_2) \quad &\leq \quad 16\sqrt{\tfrac{3}{N}} \sum_{k\in E} \sqrt{\tfrac{H_k}{2^{-k}}}\big(2^{-k} - 2^{-k-1}\big) + 4\sqrt{\tfrac{6\log(3\epsilon^{-1})}{N}} \sum_{k\in E} (\sqrt{2})^{-k} \\
&\qquad\qquad + 4\sqrt{\tfrac{6}{N}} \sum_{k\in E} (\sqrt{2})^{-k} \sqrt{\log[k(k+1)]} \\
&\leq \quad 16\sqrt{\tfrac{3}{N}} \int_{\frac{u}{2}}^{(\bar{\bar{\mathbb{P}}}_{f_1,f_2}\vee u)\wedge\frac{1}{2}} \sqrt{\tfrac{H(x,\mathcal{F},\bar{\bar{\mathbb{P}}})\vee 1}{x}}\, dx \\
&\qquad + 4\sqrt{\tfrac{6\log(3\epsilon^{-1})}{N}} \tfrac{\sqrt{2}}{\sqrt{2}-1} \sqrt{\bar{\bar{\mathbb{P}}}_{f_1,f_2} \vee u} + 4\sqrt{\tfrac{6}{N}} \sum_{k\in E} (\sqrt{2})^{-k} \sqrt{2\log(k+1)}.
\end{aligned}$$

Let $\varphi(x) \triangleq \frac{1}{\sqrt{x}}\sqrt{\log\big(e\frac{\log x^{-1}}{\log 2}\big)}$ for any $0 < x \leq \frac{1}{2}$. The function φ is decreasing on $\big[0; \frac{1}{2}\big]$. The last term can be written as

$$8\sqrt{\tfrac{3}{N}} \sum_{k\in E} \frac{2\big(2^{-k}-2^{-k-1}\big)}{\sqrt{2^{-k}}} \sqrt{\log(k+1)} \leq 16\sqrt{\tfrac{3}{N}} \int_{\frac{u}{2}}^{(\bar{\bar{\mathbb{P}}}_{f_1,f_2}\vee u)\wedge\frac{1}{2}} \varphi(x)dx.$$

6.12. **Proof of Theorem 4.4.** Let us take $U \in \mathbb{N}$ such that $2^{-U} < \frac{1}{2N}$. From Theorem 4.3, for any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any functions f_1 and f_2 in the set $\mathcal{F}$,

(6.19)

$$\begin{aligned}
&r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \\
&\qquad\qquad \leq \sum_{k\in\mathbb{N}: \frac{1}{2N} \leq 2^{-k} \leq \bar{\bar{\mathbb{P}}}_{f_1,f_2}\wedge\frac{1}{2}} 4\sqrt{\frac{6\times 2^{-k}\{2H_k+\log[3k(k+1)]+\log(\epsilon^{-1})\}}{N}}.
\end{aligned}$$

Let $\mathcal{N}$ be a $(\frac{1}{3N}, \mathcal{F}, \bar{\bar{\mathbb{P}}}_{\cdot,\cdot})$-minimal covering net. From Theorem 8.4 in [1] apply to $\mathcal{W}\big((f_1, f_2), X\big) = \mathbb{1}_{f_1(X)\neq f_2(X)}$, we obtain that with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any function f_1, f_2 in the net $\mathcal{F}$, we have

$$\bar{\mathbb{P}}'_{f_1,f_2} \leq \bar{\mathbb{P}}_{f_1,f_2} + 2\sqrt{\frac{\bar{\bar{\mathbb{P}}}_{f_1,f_2}\log[N^2(X_1^{2N})\epsilon^{-1}]}{N}},$$

hence

$$\bar{\bar{\mathbb{P}}}_{f_1,f_2} \leq \bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\frac{\bar{\bar{\mathbb{P}}}_{f_1,f_2}\log[N^2(X_1^{2N})\epsilon^{-1}]}{N}}.$$

Let $\check{\mathcal{K}} \triangleq \frac{2\log N(X_1^{2N})+\log(\epsilon^{-1})}{N}$. By solving the previous inequation, we obtain

$$\bar{\bar{\mathbb{P}}}_{f_1,f_2} \le \bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\check{\mathcal{K}}\bar{\mathbb{P}}_{f_1,f_2} + \tfrac{\check{\mathcal{K}}^2}{4}} + \tfrac{\check{\mathcal{K}}}{2}. \tag{6.20}$$

From inequality (4.3), we have

$$\begin{array}{rcl} H(u,\mathcal{F},\bar{\mathbb{P}}') & \le & H(u,\mathcal{F},\bar{\mathbb{P}}) + (\log 2)\log(\epsilon^{-1})\left(\tfrac{12}{5} + 1 + \tfrac{2H(u,\mathcal{F},\bar{\mathbb{P}})}{(\log 2)\log(\epsilon^{-1})}\right) \\ & = & 3H(u,\mathcal{F},\bar{\mathbb{P}}) + \tfrac{17\log 2}{5}\log(\epsilon^{-1}), \end{array}$$

hence

$$H(u,\mathcal{F},\bar{\bar{\mathbb{P}}}) \le 4H(u,\mathcal{F},\bar{\mathbb{P}}) + \tfrac{17\log 2}{5}\log(\epsilon^{-1}).$$

Taking a union bound with weight $\frac{1}{k(k+1)}$, we obtain that with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any $k \ge 1$, we have

$$H(2^{-k},\mathcal{F},\bar{\bar{\mathbb{P}}}) \le 4H(2^{-k},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + 2.4\log[k(k+1)\epsilon^{-1}]. \tag{6.21}$$

Let $H'_k \triangleq 4H(2^{-k},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + 2.4\log[k(k+1)\epsilon^{-1}]$. Rigorously, we cannot apply Theorem 4.3 for $H_k = H'_k$ since H'_k is not always an upper bound of $H(2^{-k},\mathcal{F},\bar{\bar{\mathbb{P}}}) \vee 1$. However we can modify the proof of Theorem 4.3 to take into a "probably approximatively correct" inequality. Therefore, combining Inequalities (6.19), (6.20) and (6.21), letting $\bar{\mathcal{K}} \triangleq \frac{2H'_U+\log(\epsilon^{-1})}{N}$ and

$$\bar{E} \triangleq \left\{k \in \mathbb{N}^* : \tfrac{1}{2N} \le 2^{-k} \le \bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\bar{\mathcal{K}}\bar{\mathbb{P}}_{f_1,f_2} + \tfrac{\bar{\mathcal{K}}^2}{4}} + \tfrac{\bar{\mathcal{K}}}{2}\right\},$$

we obtain that with $\mathbb{P}^{\otimes 2N}$-probability at least $1-3\epsilon$, for any functions f_1, f_2 in $\mathcal{F}$, we have

$$r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \le \sum_{k\in\bar{E}} 4\sqrt{\tfrac{6\times 2^{-k}\{2H'_k+\log[3k(k+1)]+\log(\epsilon^{-1})\}}{N}}.$$

To obtain the announced result, we simplify this formula by using

$$\bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\bar{\mathcal{K}}\bar{\mathbb{P}}_{f_1,f_2} + \tfrac{\bar{\mathcal{K}}^2}{4}} + \tfrac{\bar{\mathcal{K}}}{2} \le \bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\bar{\mathcal{K}}\bar{\mathbb{P}}_{f_1,f_2}} + \bar{\mathcal{K}} \le \tfrac{5}{4}\bar{\mathbb{P}}_{f_1,f_2} + 2\bar{\mathcal{K}}, \tag{6.22}$$

$$\log[U(U+1)] \le \log\left[\left(2 + \tfrac{\log N}{\log 2}\right)\left(3 + \tfrac{\log N}{\log 2}\right)\right] \le \log 6 + 2\log\left(\tfrac{e}{2\log 2}\log N\right)$$

and

$$2H'_k + \log[3k(k+1)\epsilon^{-1}] \le 8H(2^{-k},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + 6\log[k(k+1)\epsilon^{-1}] + 1.$$

6.13. **Proof of Corollary 4.5.** Let $f \in \mathcal{F}$. If $\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f} = 0$, then we trivially have $r'(\hat{f}_{\mathrm{ERM}}) \le r'(f)$. Otherwise, we have $\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f} \ge \frac{1}{2N}$. Let $K \in \mathbb{N}$ such that $\frac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}{2} < 2^{-K} \le \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}$. From Theorem 4.3, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, we have

$$\begin{array}{rcl} r'(\hat{f}_{\mathrm{ERM}}) - r'(f) & \le & \sum_{k\ge K} 4\sqrt{\frac{6\times 2^{-k}\{2V\log(e2^{k+2})+\log[3k(k+1)]+\log(\epsilon^{-1})\}}{N}} \\ & \le & \sum_{k\ge K} 4\sqrt{\frac{6\times 2^{-k}\{(2V+1)\log(e2^{k+2})+\log(\epsilon^{-1})\}}{N}} \\ & \le & 4\sqrt{\frac{6(2V+1)}{N}}\sum_{k\ge K}\sqrt{2^{-k}\log(e2^{k+2})} + 4\sqrt{\frac{6\log(\epsilon^{-1})}{N}}\sum_{k\ge K}(\sqrt{2})^{-k} \end{array}$$

Now, for any $k \geq K \geq 0$ and $V \geq 1$, we have $\frac{\log(e2^{k+2})}{\log(e2^{K+2})} \leq \frac{(k-K)(\log 2)}{1+2\log 2}+1$. Therefore we get

$$\begin{aligned}
r'(\hat{f}_{\mathrm{ERM}}) - r'(f) \quad &\leq \quad 4\sqrt{\tfrac{6(2V+1)2^{-K}\log(e2^{K+2})}{N}}\sum_{k\geq 0}\sqrt{2^{-k}\left(\tfrac{k\log 2}{1+2\log 2}+1\right)} \\
&\qquad\qquad +4\sqrt{\tfrac{6\log(\epsilon^{-1})}{N}}(\sqrt{2})^{-K}\tfrac{\sqrt{2}}{\sqrt{2}-1} \\
&\leq \quad 4\sqrt{\tfrac{6(2V+1)2^{-K}\log(e2^{K+2})}{N}}\sum_{k\geq 0}\sqrt{2^{-k}\left(\tfrac{k\log 2}{1+2\log 2}+1\right)} \\
&\qquad\qquad +4\sqrt{\tfrac{6\log(\epsilon^{-1})}{N}}(\sqrt{2})^{-K}\tfrac{\sqrt{2}}{\sqrt{2}-1} \\
&\leq \quad 47\sqrt{\tfrac{V+1}{N}}\sqrt{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}\log\left(\tfrac{8e}{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}\right)}+34\sqrt{\tfrac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}\log(\epsilon^{-1})}{N}}
\end{aligned}$$

For the second assertion of the corollary, we use Jensen's inequality and the concavity of $x \mapsto \sqrt{x\log(8ex^{-1})}$ in order to obtain that for any function $f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{P}^{\otimes N}R(\hat{f}_{\mathrm{ERM}}) \quad &\leq \quad R(f) + 47\sqrt{\tfrac{(V+1)\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}{N}\log\left(\tfrac{8e}{\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}\right)} \\
&\qquad\qquad +34\sqrt{\tfrac{\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}{N}}.
\end{aligned}$$

6.14. **Proof of Corollary 4.6.** For $\kappa = +\infty$ (i.e. no margin assumption), the result comes from inequality (4.5) since the function $x \mapsto x\log(8e/x)$ is an increasing function on $[0;1]$, hence upper bounded by its value for $x = 1$. Specifically, we obtain

$$\mathbb{P}^{\otimes N}R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f}) \leq 83\sqrt{\tfrac{V+1}{N}} + \tfrac{34}{\sqrt{N}}. \tag{6.23}$$

Note that it is thanks to the chaining that we get rid of the $\log N$ factor. For $\kappa < +\infty$, chained and unchained results lead to the same convergence rate: $\left(\frac{V}{N}\log N\right)^{\frac{\kappa}{2\kappa-1}}$.

To obtain this rate from the previous bounds, we just need to link the variance term $\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}$ with $\mathbb{P}^{\otimes N}\mathbb{P}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}$ in order to use the margin assumption.

Combining Inequalities (6.20) and (6.22), we obtain

$$\begin{aligned}
\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} \quad &\leq \quad \tfrac{5}{4}\bar{\mathbb{P}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} + \tfrac{4\log N(X_1^{2N})+2\log(\epsilon^{-1})}{N} \\
&\leq \quad \tfrac{5}{4}\bar{\mathbb{P}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} + \tfrac{4V\log\left(\frac{2eN}{V}\right)+2\log(\epsilon^{-1})}{N},
\end{aligned}$$

hence

$$\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} \leq \tfrac{5}{4}\mathbb{P}^{\otimes N}\mathbb{P}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} + \tfrac{4V\log\left(\frac{2eN}{V}\right)+2}{N}. \tag{6.24}$$

Now, by the margin assumption and Jensen's inequality, we have

$$\mathbb{P}^{\otimes N}\mathbb{P}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} \leq C''\mathbb{P}^{\otimes N}\left(\Delta^{\frac{1}{\kappa}}\right) \leq C''\left(\mathbb{P}^{\otimes N}\Delta\right)^{\frac{1}{\kappa}}. \tag{6.25}$$

The convergence rate then follows from (6.24), (6.25) and either (4.4) or (4.5).

6.15. **Proof of Lemma 5.1.** Let $\vec{\sigma}_{j,r} \triangleq (\sigma_1,\ldots,\sigma_{j-1},r,\sigma_{j+1},\ldots,\sigma_m)$ for any $r \in \{-1,0,+1\}$. The distribution $\mathbb{P}_{\vec{\sigma}_{j,0}}$ is such that $\mathbb{P}_{\vec{\sigma}_{j,0}}(dX) = \mu(dX)$ and

$$\mathbb{P}_{\vec{\sigma}_{j,0}}(Y=1|X) = \left\{\begin{array}{l} \frac{1}{2} \text{ for any } X \in \mathcal{X}_j \\ \mathbb{P}_{\vec{\sigma}}(Y=1|X) \text{ otherwise} \end{array}\right. .$$

Introduce the quantities $\pi_{r,j} \triangleq \frac{\mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,r}}}{\mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,0}}}(Z_1^N) = \prod_{i=1}^N \big[1 + r\mathbb{1}_{X_i \in \mathcal{X}_j}(2Y_i - 1)\xi(X_i)\big]$ for any $r \in \{-1;+1\}$. Let ν denote the distribution of a Rademacher variable:

$$\nu(\sigma = +1) = \nu(\sigma = -1) = \frac{1}{2}.$$

The variational distance between two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as $V(\mathbb{P}_1, \mathbb{P}_2) \triangleq \sup_{A \text{ measurable set}} \big\{\mathbb{P}_1(A) - \mathbb{P}_2(A)\big\}$. When the distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ are absolutely continuous wrt a probability distribution $\mathbb{P}_0$, we have

$$V(\mathbb{P}_1, \mathbb{P}_2) = \tfrac{1}{2} \int \big|\tfrac{\mathbb{P}_1}{\mathbb{P}_0} - \tfrac{\mathbb{P}_2}{\mathbb{P}_0}\big| d\mathbb{P}_0 = 1 - \int \big(\tfrac{\mathbb{P}_1}{\mathbb{P}_0} \wedge \tfrac{\mathbb{P}_2}{\mathbb{P}_0}\big) d\mathbb{P}_0.$$

We have successively

$$\begin{aligned}
& \sup_{\mathbb{P} \in \mathcal{P}} \Big\{ \mathbb{P}^{\otimes N} \mathbb{P}[\hat{f}(X) \neq Y] - \mathbb{P}[f^*_{\mathbb{P}}(X) \neq Y) \Big\} \\
\geq \quad & \sup_{\vec{\sigma} \in \{-1;+1\}^m} \Big\{ \big(\mathbb{P}^{\otimes N}_{\vec{\sigma}}\big) \mathbb{P}_{\vec{\sigma}}[\hat{f}(X) \neq Y] - \mathbb{P}_{\vec{\sigma}}[f^*_{\mathbb{P}_{\vec{\sigma}}} \neq Y] \Big\} \\
= \quad & \sup_{\vec{\sigma} \in \{-1;+1\}^m} \Big\{ \big(\mathbb{P}^{\otimes N}_{\vec{\sigma}}\big) \mathbb{P}_{\vec{\sigma}} \Big(\xi(X) \mathbb{1}_{\hat{f}(X) \neq f^*_{\mathbb{P}_{\vec{\sigma}}}(X)} \Big) \Big\} \\
= \quad & \sup_{\vec{\sigma} \in \{-1;+1\}^m} \Big\{ \mathbb{P}^{\otimes N}_{\vec{\sigma}} \Big(\textstyle\sum_{j=1}^m \mu\big[\xi(X) \mathbb{1}_{\hat{f}(X) \neq \frac{1+\sigma_j}{2}; X \in \mathcal{X}_j}\big] \Big) \Big\} \\
\geq \quad & \mathbb{E}_{\nu^{\otimes m}} \textstyle\sum_{j=1}^m \mathbb{P}^{\otimes N}_{\vec{\sigma}} \Big[\mu\big[\xi(X) \mathbb{1}_{\hat{f}(X) \neq \frac{1+\sigma_j}{2}; X \in \mathcal{X}_j}\big] \Big] \\
= \quad & \mathbb{E}_{\nu^{\otimes m}} \textstyle\sum_{j=1}^m \mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,0}} \left(\frac{\mathbb{P}^{\otimes N}_{\vec{\sigma}}}{\mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,0}}} \mu\big[\xi(X) \mathbb{1}_{\hat{f}(X) \neq \frac{1+\sigma_j}{2}; X \in \mathcal{X}_j}\big] \right) \\
= \quad & \mathbb{E}_{\nu^{\otimes (m-1)}(d\sigma_1, \dots, d\sigma_{j-1}, d\sigma_{j+1}, \dots, d\sigma_m)} \textstyle\sum_{j=1}^m \mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,0}} \mathbb{E}_{\nu(d\sigma_j)} \\
& \qquad\qquad \left(\frac{\mathbb{P}^{\otimes N}_{\vec{\sigma}}}{\mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,0}}} \mu\big[\xi(X) \mathbb{1}_{\hat{f}(X) \neq \frac{1+\sigma_j}{2}; X \in \mathcal{X}_j}\big] \right) \\
\geq \quad & \mathbb{E}_{\nu^{\otimes (m-1)}(d\sigma_1, \dots, d\sigma_{j-1}, d\sigma_{j+1}, \dots, d\sigma_m)} \textstyle\sum_{j=1}^m \mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,0}} \\
& \qquad\qquad \Big[(\pi_{-1,j} \wedge \pi_{+1,j}) \mathbb{E}_{\nu(d\sigma_j)} \mu\big[\xi(X) \mathbb{1}_{\hat{f}(X) \neq \frac{1+\sigma_j}{2}; X \in \mathcal{X}_j}\big] \Big] \\
= \quad & \mathbb{E}_{\nu^{\otimes (m-1)}(d\sigma_1, \dots, d\sigma_{j-1}, d\sigma_{j+1}, \dots, d\sigma_m)} \textstyle\sum_{j=1}^m \\
& \qquad\qquad \tfrac{1}{2} \mu\big[\xi(X) \mathbb{1}_{X \in \mathcal{X}_j}\big] \Big[1 - V\Big(\mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,-1}}, \mathbb{P}^{\otimes N}_{\vec{\sigma}_{j,+1}} \Big) \Big] \\
= \quad & \tfrac{mw}{2} \mu\big[\xi(X) \,\big|\, X \in \mathcal{X}_j\big] \Big[1 - V\Big(\mathbb{P}^{\otimes N}_{-1,1,\dots,1}, \mathbb{P}^{\otimes N}_{1,1,\dots,1} \Big) \Big].
\end{aligned} \tag{6.26}$$

Now let us prove

$$V\Big(\mathbb{P}^{\otimes N}_{-1,1,\dots,1}, \mathbb{P}^{\otimes N}_{1,1,\dots,1}\Big) \leq b\sqrt{Nw}. \tag{6.27}$$

First, we have

$$V\Big(\mathbb{P}^{\otimes N}_{-1,1,\dots,1}, \mathbb{P}^{\otimes N}_{1,1,\dots,1}\Big) = \sum_{l=1}^N \binom{N}{l} w^l (1-w)^{N-l} \mathcal{V}_l, \tag{6.28}$$

where $\mathcal{V}_l \triangleq V\Big(\mathbb{P}^{\otimes l}_{-1}, \mathbb{P}^{\otimes l}_{+1}\Big)$ and $\mathbb{P}_\sigma \triangleq \mathbb{P}_{\sigma,1,\dots,1}(\cdot | X \in \mathcal{X}_1)$ for any $\sigma \in \{-1, +1\}$. By simple computations, we get $\mathcal{V}_1 = \mu[\xi(X) | X \in \mathcal{X}_j]$. From Jensen's inequality, by the concavity of $x \mapsto \sqrt{1 - x^2}$, we have

$$\sqrt{1 - b^2} = \mu\big[\sqrt{1 - \xi^2(X)} \,\big|\, X \in \mathcal{X}_j\big] \leq \sqrt{1 - \{\mu[\xi(X) | X \in \mathcal{X}_j]\}^2},$$

hence $\mathcal{V}_1 \leq b$.

For $l \geq 2$, we upper bound the variational distance by the Hellinger distance. By definition, the Hellinger distance $H(\mathbb{P},\mathbb{Q})$ satisfies $1 - \frac{H^2(\mathbb{P},\mathbb{Q})}{2} = \int \sqrt{d\mathbb{P}}\sqrt{d\mathbb{Q}}$. Hence the tensorization equality is $1 - \frac{H^2(\mathbb{P}^{\otimes l},\mathbb{Q}^{\otimes l})}{2} = \big(1 - \frac{H^2(\mathbb{P},\mathbb{Q})}{2}\big)^l$. We have

$$\mathcal{V}_l \leq H\Big(\mathbb{P}_{-1}^{\otimes l}, \mathbb{P}_{+1}^{\otimes l}\Big) = \sqrt{2\Big(1 - \Big[1 - \tfrac{H^2(\mathbb{P}_{-1},\mathbb{P}_{+1})}{2}\Big]^l\Big)}$$

and $1 - \frac{H^2(\mathbb{P}_{-1},\mathbb{P}_{+1})}{2} = \mu\big[\sqrt{1-\xi^2(X)}|X \in \mathcal{X}_1\big] = \sqrt{1-b^2}$, by definition of b. Now, for any $l \geq 2$ and $x \geq 0$, $2\Big(1 - \big[1-x^2\big]^{\frac{l}{2}}\Big) \leq lx^2$. Finally, for any $l \geq 1$, we have $\mathcal{V}_l \leq b\sqrt{l}$. Putting this result in equality (6.28), we get

$$V\Big(\mathbb{P}_{-1,1,\dots,1}^{\otimes N}, \mathbb{P}_{1,1,\dots,1}^{\otimes N}\Big) \leq b \sum_{l=0}^{N} \mathbb{P}\Big(\sum_{i=1}^{N} \epsilon_i = l\Big)\sqrt{l},$$

where the ϵ_i are i.i.d. random variables such that $\mathbb{P}(\epsilon_i = 1) = w = 1 - \mathbb{P}(\epsilon_i = 0)$. So we have $V\Big(\mathbb{P}_{-1,1,\dots,1}^{\otimes N}, \mathbb{P}_{1,1,\dots,1}^{\otimes N}\Big) \leq b\mathbb{P}\sqrt{\sum_{i=1}^N \epsilon_i} \leq b\sqrt{\mathbb{P}\sum_{i=1}^N \epsilon_i} = b\sqrt{Nw}$.

Remark 6.6. The last inequality in (6.26) is an equality when for any $j \in \{1,\dots,m\}$, $\hat{f} = \underset{r\in\{-1;+1\}}{\text{argmax}}\, \mathbb{P}^{\otimes N}{}_{\vec{\sigma}_{j,r}}$ on $\mathcal{X}_j$, i.e. when $\hat{f}$ is the maximum likelyhood estimator on the set $\mathcal{X} - \mathcal{X}_0$.

Appendix A. Proof of inequality (6.2)

For any $r > 0$, define $\Gamma(r) \triangleq \int_0^{+\infty} u^{r-1}\exp(-u)du$. Integrating by parts, we obtain the well-known property $\Gamma(r+1) = r\Gamma(r)$.

- We have

$$\begin{array}{rcl} \pi\exp\big(-\lambda\Delta R\big) & = & \int_0^{+\infty} \pi\big\{\exp\big(-\lambda\Delta R\big) \geq u\big\}du \\ & = & \exp(-\lambda) + \int_{\exp(-\lambda)}^1 \pi\big\{\exp\big(-\lambda\Delta R\big) \geq u\big\}du \\ & = & \exp(-\lambda) + \int_0^1 \lambda\exp\big(-\lambda x\big)\pi\big(\Delta R \leq x\big)dx \end{array} \tag{A.1}$$

Let us introduce $A' \triangleq \exp(-C''')$. Since we have $\pi(\Delta R \leq x) = x^{C'}\big[A' + \eta(x)\big]$ with $\eta(x) = \underset{x\to 0}{\text{o}}(x^s)$ and $\eta(x) \leq x^{-C'}$ and $\int_0^{+\infty} \lambda\exp\big(-\lambda x\big)x^{C'}dx = \frac{\Gamma(C'+1)}{\lambda^{C'}}$, we get

$$\begin{aligned} &\Big|\pi\exp\big(-\lambda\Delta R\big) - A'\tfrac{\Gamma(C'+1)}{\lambda^{C'}} - \exp(-\lambda)\Big| \\ &\qquad = \int_0^1 \lambda\exp\big(-\lambda x\big)x^{C'}\eta(x)dx + A'\int_1^{+\infty} \lambda\exp\big(-\lambda x\big)x^{C'}dx \end{aligned}$$

Since we have

$$\begin{aligned} &\int_0^1 \lambda\exp\big(-\lambda x\big)x^{C'}\eta(x)dx \\ &\qquad = \int_0^{\frac{1}{\sqrt{\lambda}}} \lambda\exp\big(-\lambda x\big)x^{C'+s}\underset{x\to 0}{\text{o}}(1)dx + \int_{\frac{1}{\sqrt{\lambda}}}^1 \lambda\exp\big(-\lambda x\big)x^{C'}\eta(x)dx \\ &\qquad \leq \underset{\lambda\to+\infty}{\text{o}}\big(\lambda^{-(C'+s)}\big) + \int_{\frac{1}{\sqrt{\lambda}}}^1 \lambda\exp\big(-\lambda x\big)dx \\ &\qquad = \underset{\lambda\to+\infty}{\text{o}}\big(\lambda^{-(C'+s)}\big) \end{aligned}$$

and $\int_1^{+\infty} \lambda\exp\big(-\lambda x\big)x^{C'}dx = \underset{\lambda\to+\infty}{\text{o}}\big(\lambda^{-(C'+s)}\big)$, we obtain

$$\pi\exp\big(-\lambda\Delta R\big) = A'\frac{\Gamma(C'+1) + \underset{\lambda\to+\infty}{\text{o}}(\lambda^{-s})}{\lambda^{C'}} \tag{A.2}$$

- From equalities (A.1), we have

(A.3)
$$\pi\big[\Delta R \exp\big(-\lambda\Delta R\big)\big] = \exp(-\lambda) + \int_0^1 (\lambda x - 1)\exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx.$$

We have just seen that $\int_0^1 \exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx = A'\frac{\Gamma(C'+1)+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda^{C'+1}}$. Besides, from the same argument as above, we have

$$\int_0^1 \lambda x \exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx = A'\frac{\Gamma(C'+2)+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda^{C'+1}}.$$

Since $\Gamma(C'+2) = (C'+1)\Gamma(C'+1)$, we obtain

$$\pi\big[\Delta R \exp\big(-\lambda\Delta R\big)\big] = A'\frac{C'\Gamma(C'+1)+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda^{C'+1}}.$$

- Combining the previous results, we obtain $\pi_{-\lambda R}\Delta R = \frac{C'+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda}$.

Appendix B. Proof of inequality (6.4)

Let $\alpha > 0$ depend on λ such that $\lambda\alpha^{\frac{\kappa}{\kappa-1}} \underset{\lambda\to+\infty}{\longrightarrow} 0$. Then there exists $0 < \zeta < 1$ depending on λ such that $\zeta \underset{\lambda\to+\infty}{\longrightarrow} 1$ and $\lambda\big(\frac{\alpha}{1-\zeta}\big)^{\frac{\kappa}{\kappa-1}} \underset{\lambda\to+\infty}{\longrightarrow} 0$. Let $h_\alpha(x) \triangleq x - \alpha x^{\frac{1}{\kappa}}$ and $x_0 \triangleq \big(\frac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}$. The function h decreases on $[0;x_0]$ and increases on $[x_0;+\infty]$. We have

$$\begin{aligned}\pi\exp\big[-\lambda h_\alpha(\Delta R)\big] &= \pi\Big\{\exp\big[-\lambda h_\alpha(\Delta R)\big]\mathbb{1}_{h_\alpha(\Delta R)\le\zeta\Delta R}\Big\}\\ &\quad+\pi\Big\{\exp\big[-\lambda h_\alpha(\Delta R)\big]\mathbb{1}_{h_\alpha(\Delta R)>\zeta\Delta R}\Big\}\\ &\le \exp\big[-\lambda h_\alpha(x_0)\big]\pi\Big\{\Delta R \le \big(\tfrac{\alpha}{1-\zeta}\big)^{\frac{\kappa}{\kappa-1}}\Big\} + \pi\exp\big[-\lambda\zeta\Delta R\big]\\ &\le \exp\big[(\kappa-1)\lambda\big(\tfrac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}\big]\big(\tfrac{\alpha}{1-\zeta}\big)^{\frac{C'\kappa}{\kappa-1}}\Big[1+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]\\ &\quad+\pi\exp\big[-\lambda\zeta\Delta R\big]\end{aligned}$$

From equality (A.2), we get

$$\begin{aligned}&\pi_{-\lambda R}\exp\big(\lambda\alpha\Delta R\big)\\ &\quad= \exp\Big[(\kappa-1)\lambda\big(\tfrac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}\Big]\lambda^{C'}\big(\tfrac{\alpha}{1-\zeta}\big)^{\frac{C'\kappa}{\kappa-1}}\Big[\tfrac{1}{\Gamma(C'+1)} + \underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big] + \frac{1+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\zeta^{C'}}\\ &\quad= \underset{\lambda\to+\infty}{\mathrm{O}}\Big(\lambda^{C'}\big[\tfrac{\alpha}{1-\zeta}\big]^{\frac{\kappa C'}{\kappa-1}}\Big) + 1 + \underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{\lambda\to+\infty}{\mathrm{O}}\big(1-\zeta\big)\end{aligned}$$

Taking $\zeta = 1 - \big(\lambda\alpha^{\frac{\kappa}{\kappa-1}}\big)^{\frac{(\kappa-1)C'}{\kappa C'+\kappa-1}}$, we obtain

$$\log\pi_{-\lambda R}\exp\big(\lambda\alpha\Delta R\big) = \underset{\lambda\to+\infty}{\mathrm{O}}\Big(\big[\lambda\alpha^{\frac{\kappa}{\kappa-1}}\big]^{\frac{(\kappa-1)C'}{\kappa C'+\kappa-1}}\Big) + \underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s}).$$

Appendix C. Proof of inequality (6.6)

We start with the following lemma.

Lemma C.1. *Let $h : \mathbb{R}^* \to \mathbb{R}$ be a C^3 convex function such that there exists $u_0 > 0$ satisfying $h'(u_0) = 0$ and $h''(u_0) > 0$. Let $\phi : \mathbb{R}^* \to \mathbb{R}$ be a continuous non negative* *function such that $\phi(u_0 > 0)$ and $u \mapsto \phi(u)\exp(-t_0 u)$ integrable for some $t_0 > 0$. Then for any $A > u_0$, we have*

$$\int_0^A \phi(u)\exp\big[-th(u)\big]du \underset{t\to+\infty}{\sim} \phi(u_0)\exp\big[-th(u_0)\big]\sqrt{\tfrac{2\pi}{th''(u_0)}}.$$

Proof. • Since the function h'' is non negative, continuous, $h''(u_0) > 0$ and $h'(u_0) = 0$, there exists $c > 0$ such that for any $u \in [0;A]$, $h(u) - h(u_0) \geq c(u-u_0)^2$. Let $\alpha_t \triangleq t^{-p}$ with $\frac{1}{3} < p < \frac{1}{2}$. We have

$$\begin{aligned}\int_{[0;u_0-\alpha_t]\cup[u_0+\alpha_t;A]} &\phi(u)\exp\big[-th(u)\big]du \\ &\leq \exp\big[-th(u_0)\big]\int_{[0;u_0-\alpha_t]\cup[u_0+\alpha_t;A]}\phi(u)\exp\big[-tc(u-u_0)^2\big]du \\ &= \exp\big[-th(u_0)\big]\underset{t\to+\infty}{\mathrm{O}}\big(\exp\big[-ct{\alpha_t}^2\big]\big).\end{aligned}$$

• From Taylor's theorem, for any $u \in [u_0-\alpha_t;u_0+\alpha_t]$, there exists $u^* \in [u_0-\alpha_t;u_0+\alpha_t]$ such that

$$h(u) = h(u_0) + \tfrac{h''(u_0)}{2}(u-u_0)^2 + \tfrac{h'''(u^*)}{6}(u-u_0)^3$$

Let $A'' \triangleq \sup_{[\frac{u_0}{2};A]}\big|\frac{h'''(u)}{6}\big|$ and $I_t \triangleq \int_{[u_0-\alpha_t;u_0+\alpha_t]}\phi(u)\exp\big[-t\frac{h''(u_0)}{2}(u-u_0)^2\big]du$ We get

$$\int_{[u_0-\alpha_t;u_0+\alpha_t]}\phi(u)\exp\big[-th(u)\big]du \leq \exp\big[A''t\alpha_t^3\big]\exp\big[-th(u_0)\big]I_t$$

and

$$\int_{[u_0-\alpha_t;u_0+\alpha_t]}\phi(u)\exp\big[-th(u)\big]du \geq \exp\big[-A''t\alpha_t^3\big]\exp\big[-th(u_0)\big]I_t.$$

We have

$$\begin{aligned}\Big|I_t - \int_{-\infty}^{+\infty}&\phi(u_0)\exp\big[-t\tfrac{h''(u_0)}{2}(u-u_0)^2\big]du\Big| \\ &\leq \int_{[u_0-\alpha_t;u_0+\alpha_t]}|\phi(u)-\phi(u_0)|\exp\big[-t\tfrac{h''(u_0)}{2}(u-u_0)^2\big]du \\ &\quad + \int_{]-\infty;u_0-\alpha_t]\cup[u_0+\alpha_t;+\infty[}\phi(u_0)\exp\big[-t\tfrac{h''(u_0)}{2}(u-u_0)^2\big]du \\ &\leq \underset{t\to+\infty}{\mathrm{o}}\Big(\int_{-\infty}^{+\infty}\exp\big[-t\tfrac{h''(u_0)}{2}(u-u_0)^2\big]du\Big) \\ &\quad + \underset{t\to+\infty}{\mathrm{O}}\Big(\exp\big[-\tfrac{h''(u_0)}{2}t{\alpha_t}^2\big]du\Big)\end{aligned}$$

Since we have $\int_{-\infty}^{+\infty}\exp\big[-t\frac{h''(u_0)}{2}(u-u_0)^2\big]du = \sqrt{\frac{2\pi}{th''(u_0)}}$, we obtain

$$I_t = \big[\phi(u_0) + \underset{t\to+\infty}{\mathrm{o}}(1)\big]\sqrt{\tfrac{2\pi}{th''(u_0)}}.$$

• Combining the previous results, we obtain

$$\int_0^A \phi(u)\exp\big[-th(u)\big]du = \big[\phi(u_0) + \underset{t\to+\infty}{\mathrm{o}}(1)\big]\exp\big[-th(u_0)\big]\sqrt{\tfrac{2\pi}{th''(u_0)}}.$$

□

By assumption, we may write $\pi\big(\Delta R \leq x\big) = \exp\big(-C'x^{-\frac{q}{\kappa}} - C'''\big)[1+\eta(x)]$ with $\eta(x) = \underset{x\to0}{\mathrm{o}}(1)$. Let $A' \triangleq \exp(-C''')$, $u_0 \triangleq \mathrm{argmin}_{x>0}\big(x + C'x^{-\frac{q}{\kappa}}\big)$, $H \triangleq u_0 + C'{u_0}^{-\frac{q}{\kappa}}$ and $\theta \triangleq 2H\lambda^{-\frac{\kappa}{\kappa+q}}$.

From inequality (A.1), we have

$$\text{(C.1)}\qquad \begin{aligned}\pi\exp\big(-\lambda\Delta R\big) &= \exp(-\lambda) + \int_0^1\lambda\exp\big(-\lambda x\big)\pi\big(\Delta R\leq x\big)dx \\ &\leq \exp(-\lambda\theta) + \int_0^\theta\lambda\exp\big(-\lambda x\big)\pi\big(\Delta R\leq x\big)dx\end{aligned}$$

Besides, we have

$$\begin{aligned}&\textstyle\int_0^\theta \exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx\\ &\qquad = A' \textstyle\int_0^\theta \exp\big(-\lambda x - C' x^{-\frac{q}{\kappa}}\big)[1+\eta(x)]dx\\ &\qquad = A' \textstyle\int_0^{2H} \exp\big(-\lambda^{\frac{q}{\kappa+q}}\big[x + C' x^{-\frac{q}{\kappa}}\big]\big)\big[1+\eta\big(\lambda^{-\frac{\kappa}{\kappa+q}}x\big)\big]dx\end{aligned}$$

For any $\beta > 0$, there exists λ_0 such that for any $\lambda > \lambda_0$ and any $x \le \theta$, we have $\big|\eta\big(\lambda^{-\frac{\kappa}{\kappa+q}}x\big)\big| \le \beta$. We obtain

$$\left|\frac{\int_0^\theta \exp(-\lambda x)\pi(\Delta R \le x)dx}{A' \int_0^{2H} \exp\Big(-\lambda^{\frac{q}{\kappa+q}}\Big[x + C' x^{-\frac{q}{\kappa}}\Big]\Big)dx} - 1\right| \le \beta.$$

Using Lemma C.1, we get

(C.2)

$$\begin{aligned}\textstyle\int_0^\theta \exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx &\underset{\lambda\to+\infty}{\sim} A' \textstyle\int_0^{2H} \exp\big(-\lambda^{\frac{q}{\kappa+q}}\big[x + C' x^{-\frac{q}{\kappa}}\big]\big)dx\\ &\underset{\lambda\to+\infty}{\sim} \exp\big(-\lambda^{\frac{q}{\kappa+q}}H\big)\end{aligned}$$

So inequality (C.1) implies

$$\pi \exp\big(-\lambda\Delta R\big) \underset{\lambda\to+\infty}{\sim} \lambda \exp\big(-\lambda^{\frac{q}{\kappa+q}}H\big). \tag{C.3}$$

From equality (A.3), we have

$$\begin{aligned}\pi\big[\Delta R \exp\big(-\lambda\Delta R\big)\big] = \exp(-\lambda) - &\textstyle\int_0^1 \exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx\\ &+\lambda \textstyle\int_0^1 x \exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx.\end{aligned}$$

Using similar computations to the one used to prove (C.2) and from the equality

$$\begin{aligned}&\textstyle\int_0^\theta x \exp\big(-\lambda x - C' x^{-\frac{q}{\kappa}}\big)[1+\eta(x)]\\ &\qquad = \textstyle\int_0^{2H} \lambda^{-\frac{\kappa}{\kappa+q}} x \exp\big(-\lambda^{\frac{q}{\kappa+q}}\big[x + C' x^{-\frac{q}{\kappa}}\big]\big)\big[1+\eta\big(\lambda^{-\frac{\kappa}{\kappa+q}}x\big)\big]dx,\end{aligned}$$

we obtain

$$\pi\big[\Delta R \exp\big(-\lambda\Delta R\big)\big] \underset{\lambda\to+\infty}{\sim} \lambda^{\frac{q}{\kappa+q}} u_0 \exp\big(-\lambda^{\frac{q}{\kappa+q}}H\big).$$

Consequently, we have proved $\pi_{-\lambda R}\Delta R \underset{\lambda\to+\infty}{\sim} u_0 \lambda^{-\frac{\kappa}{\kappa+q}}$. By definition of u_0, we get $\pi_{-\lambda R}\Delta R \underset{\lambda\to+\infty}{\sim} \big(\frac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}$.

Appendix D. Proof of inequality (6.7)

Let $0 < \alpha \le \breve{c}\lambda^{-\frac{\kappa-1}{\kappa+q}}$ for some constant $\breve{c} > 0$ to be determined, and $\frac{1}{2} < \zeta < 1$. We use once more the function $h_\alpha(x) \triangleq x - \alpha x^{\frac{1}{\kappa}}$ which is minimum at $x_0 \triangleq \big(\frac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}$. Let $\nu \triangleq \eta\big\{\big(\frac{\alpha}{1-\zeta}\big)^{\frac{\kappa}{\kappa-1}}\big\}$. We have

$$\begin{aligned}\pi \exp\big[-\lambda h_\alpha(\Delta R)\big] = \pi\Big\{&\exp\big[-\lambda h_\alpha(\Delta R)\big]\mathbb{1}_{h_\alpha(\Delta R)\le\zeta\Delta R}\Big\}\\ &+\pi\Big\{\exp\big[-\lambda h_\alpha(\Delta R)\big]\mathbb{1}_{h_\alpha(\Delta R)>\zeta\Delta R}\Big\}\\ \le \exp&\big[-\lambda h_\alpha(x_0)\big]\pi\Big\{\Delta R \le \big(\tfrac{\alpha}{1-\zeta}\big)^{\frac{\kappa}{\kappa-1}}\Big\} + \pi \exp\big[-\lambda\zeta\Delta R\big]\\ \le \exp&\Big\{\lambda(\kappa-1)\big(\tfrac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}} - C'\big(\tfrac{1-\zeta}{\alpha}\big)^{\frac{q}{\kappa-1}} - C'''\Big\}(1+\nu)\\ &+\pi \exp\big[-\lambda\zeta\Delta R\big]\end{aligned}$$

Using (C.3), we obtain

$$\begin{aligned}&\pi_{-\lambda R}\exp\big(\lambda\alpha\Delta R\big)\\ &\qquad\le \lambda^{-1}\exp\Big(H\lambda^{\frac{q}{\kappa+q}}+\lambda(\kappa-1)\big(\tfrac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}-C'\big(\tfrac{1-\zeta}{\alpha}\big)^{\frac{q}{\kappa-1}}\Big)\Big[\breve{C}+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]\\ &\qquad\quad+\zeta\exp\Big(\lambda^{\frac{q}{\kappa+q}}H\big[1-\zeta^{\frac{q}{\kappa+q}}\big]\Big)\Big[1+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]\end{aligned}$$

Let $\zeta=1-\big(\frac{2H}{C'}\big)^{\frac{\kappa-1}{q}}\alpha\lambda^{\frac{\kappa-1}{\kappa+q}}$ so that $C'\big(\frac{1-\zeta}{\alpha}\big)^{\frac{q}{\kappa-1}}=2H\lambda^{\frac{q}{\kappa+q}}$ and let $\breve{c}>0$ such that $(\kappa-1)\big(\frac{\breve{c}}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}\le H$. Then we have

$$\begin{aligned}&\pi_{-\lambda R}\exp\big(\lambda\alpha\Delta R\big)\\ &\qquad\le \lambda^{-1}\Big[\breve{C}+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]+\exp\Big\{\lambda^{\frac{q}{\kappa+q}}\tfrac{q}{\kappa+q}H\mathrm{O}(1-\zeta)\Big\}\Big[1+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]\\ &\qquad=\underset{\lambda\to+\infty}{\mathrm{O}}\Big(\exp\Big\{\breve{C}\lambda^{\frac{q}{\kappa+q}}\alpha\lambda^{\frac{\kappa-1}{\kappa+q}}\Big\}\Big),\end{aligned}$$

hence

$$\log\pi_{-\lambda R}\exp\big(\lambda\alpha\Delta R\big)=\underset{\lambda\to+\infty}{\mathrm{O}}\Big(\lambda^{\frac{q}{\kappa+q}}\alpha\lambda^{\frac{\kappa-1}{\kappa+q}}\Big).$$

Appendix E. Another way of getting the right order

This section proves that by using well-known results, we can obtain a lower bound having the same spirit as Lemma 5.1 but without proper constants.

Applying Lemma 6.4 to the set of probability distributions

$$\mathcal{D}\triangleq\big\{\mathbb{P}^{\otimes N}:\mathbb{P}\in\mathcal{D}'\big\}$$

where $\mathcal{D}'\triangleq\big\{\mathbb{P}_{\sigma_1^m}:\sigma_1^m\in\mathcal{S}\subset\{-1;+1\}^m\big\}$ and $\mathcal{S}$ satisfies $\delta(\Sigma,\Sigma')\ge\frac{m}{4}$ for any $\Sigma\neq\Sigma'\in\mathcal{S}$ and $|\mathcal{S}|=\lfloor e^{\frac{m}{8}}\rfloor$. From Lemma 6.3, such a set $\mathcal{S}$ exists. With any estimator $\hat{f}:\mathcal{Z}^N\to\mathcal{F}(\mathcal{X},\mathcal{Y})$, we can associate an estimator $\hat{T}:\mathcal{Z}^N\to\mathcal{D}$ defined as $\hat{T}(Z_1^N)=\mathbb{P}^{\otimes N}$, where $\mathbb{P}\in\mathcal{D}'$ minimizes $\mu[\xi(X)\mathbb{1}_{f^*_{\mathbb{P}}(X)\neq\hat{f}(Z_1^N)(X)}]$.

By Birgé's lemma, we have $\sup_{\mathbb{P}\in\mathcal{D}'}\mathbb{P}^{\otimes N}\big[\hat{T}(Z_1^N)\neq\mathbb{P}\big]\ge 0.36\wedge\Big(1-\frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\Big)$. Now, when $\hat{T}(Z_1^N)\neq\mathbb{P}$, we have $R_{\mathbb{P}}(\hat{f})-R_{\mathbb{P}}(f^*_{\mathbb{P}})\ge\frac{m}{8}wb'$. Therefore, we get

$$\sup_{\mathbb{P}\in\mathcal{D}}\big\{\mathbb{P}^{\otimes N}R_{\mathbb{P}}(\hat{f})-R_{\mathbb{P}}(f^*_{\mathbb{P}})\big\}\ge\frac{m}{8}wb'\Big[0.36\wedge\Big(1-\frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\Big)\Big]. \tag{E.1}$$

For any $\mathbb{P}\neq\mathbb{Q}\in\mathcal{D}'$, we have

$$K(\mathbb{P},\mathbb{Q})\le N\mu\Big[\xi(X)\log\Big(\frac{1+\xi(X)}{1-\xi(X)}\Big)\mathbb{1}_{X\notin\mathcal{X}_0}\Big]\le Nmwb'\log\Big(\frac{1+B}{1-B}\Big),$$

where $B\triangleq\sup_{x\in\mathcal{X}-\mathcal{X}_0}\xi(x)$. If we assume that $B\le Cb<1$, we get

$$K(\mathbb{P},\mathbb{Q})\le CNmwb^2$$

for some constant $C>0$. Since we have $|\mathcal{D}|=\lfloor e^{\frac{m}{8}}\rfloor$, we obtain

$$\tfrac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\le\tfrac{C}{\log\lfloor e^{\frac{m}{8}}\rfloor}Nmwb^2\le C'Nwb^2$$

for m large enough and some constant $C'>0$. So we obtain the right order to the extent that when the quantity Nwb^2 is small enough, the order of the bound is given by the product mwb.

Appendix F. Proof of Theorem 5.2

• When $L = 0$: let $x_0, x_1, \dots, x_{V-1}$ denote the V points shattered by the model. Let us take

$$\begin{cases} m = V-1 \\ \mathcal{X}_0 = \mathcal{X} - \big\{x_1, \dots, x_{V-1}\big\} \\ \mathcal{X}_j = \{x_j\} \\ \mu(\mathcal{X}_j) = w \qquad \text{for any } j \in \{1, \dots, m\} \\ \mu(\{x_0\}) = 1 - mw \\ b = 1 \qquad (\xi \equiv 1) \end{cases},$$

where w is a free positive parameter which satisfies $mw \le 1$ (since μ is a probability distribution). By noticing that

$$1 - V\Big(\mathbb{P}^{\otimes N}_{-1,1,\dots,1}, \mathbb{P}^{\otimes N}_{1,1,\dots,1}\Big) = \mu^{\otimes N}\big(\text{for any } i \in \{1, \dots, N\}, X_i \notin \mathcal{X}_j\big) = (1-w)^N$$

and using inequality (6.26), we obtain

$$\sup_{\mathbb{P} \in \mathcal{P}} \Big\{ \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \Big\} \quad \ge \quad \tfrac{V-1}{2} \sup_{w \le \frac{1}{V-1}} \Big\{ w\Big(1-w\Big)^N \Big\}$$

This supremum is attained for $w = \frac{1}{N+1}$ when $N \ge (V-2) \vee 1$ and for $w = \frac{1}{V-1}$ otherwise.

• When $0 \le L \le \frac{1}{2}$: once more, $x_0, x_1, \dots, x_{V-1}$ denote the V points shattered by the model. This time, we take

$$\begin{cases} m = V-1 \\ \mathcal{X}_0 = \mathcal{X} - \big\{x_1, \dots, x_{V-1}\big\} \\ \mathcal{X}_j = \{x_j\} \\ \mu(\mathcal{X}_j) = w \qquad \text{for any } j \in \{1, \dots, m\} \\ \mu(\{x_0\}) = 1 - mw \\ \xi(x) = \begin{cases} b_0 \text{ when } x \in \mathcal{X}_0 \\ b \text{ otherwise} \end{cases} \end{cases},$$

where w is a free positive parameter which satisfies $mw \le 1$ (since μ is a probability distribution) and b and b_0 belong to $[0;1]$. Since we have

$$L = \frac{1}{2} mw(1-b) + \frac{1}{2}(1-mw)(1-b_0)$$

and $b_0 \in [0;1]$, the parameters m, w and b should satisfy

$$mw(1-b) \le 2L \le 1 - mwb.$$

Since this condition implies that $mw \le 1$, we have the following lower bound

$$\sup_{\mathbb{P} \in \mathcal{P}} \Big\{ \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*) \Big\} \quad \ge \quad \sup_{\substack{w \ge 0 \\ 0 \le b \le 1 \\ mw(1-b) \le 2L \le 1 - mw\beta}} \quad \tfrac{1}{2} mwb\big[1 - b\sqrt{Nw}\big].$$

From this lower bound, one can recover the first assertion of Theorem 5.2 with a constant slightly worsened (due to the upper bound (6.27)). We will now slightly weaken this result in order to get a simple lower bound. Introduce $x = b^2 wN$. The previous supremum can be written as

$$\begin{aligned}
&\sup_{\substack{x\geq 0\\ 0<b\leq 1\\ \frac{(V-1)x}{bN}\frac{1-b}{b}\leq 2L\\ \frac{(V-1)x}{bN}\leq 1-2L}} \tfrac{1}{2}\tfrac{(V-1)x}{bN}\big[1-\sqrt{x}\big]\\
&\qquad\geq \sup_{\substack{x\geq 0\\ 0<b<1\\ \frac{(V-1)x}{bN}\frac{1-b}{b}=2L\\ 2L\frac{b}{1-b}\leq 1-2L}} \tfrac{1}{2}\tfrac{(V-1)x}{bN}\big[1-\sqrt{x}\big]\\
&\qquad= \sup_{\substack{x>0\\ b=\frac{2}{1+\sqrt{1+\frac{8LN}{(V-1)x}}}\\ b\leq 1-2L}} \tfrac{1}{2}\tfrac{(V-1)x}{bN}\big[1-\sqrt{x}\big]\\
&\qquad= \sup_{0<x\leq\frac{(1-2L)^2N}{V-1}} \tfrac{(V-1)x}{4N}\Big(1+\sqrt{1+\tfrac{8LN}{(V-1)x}}\Big)\big[1-\sqrt{x}\big]\\
&\qquad> \sup_{0<x\leq\frac{(1-2L)^2N}{V-1}} \sqrt{\tfrac{L(V-1)x}{2N}}\big[1-\sqrt{x}\big] \qquad\qquad (A)\\
&\qquad= \begin{cases} \sqrt{\frac{L(V-1)}{32N}} & \text{when } \frac{(1-2L)^2N}{V-1}\geq\frac{1}{4}\\ \sqrt{\frac{L(1-2L)^2}{8}} & \text{otherwise.}\end{cases}
\end{aligned}$$

Note that the step (A) prevents us to have a good lower bound when $L=\mathrm{o}\big(\frac{V-1}{N}\big)$. In this last case, the lower bound (A) can be replaced with $\frac{V-1}{2N}x(1-\sqrt{x})$ which, by taking $x=\frac{1}{4}$, leads to the desired bound $\frac{V-1}{16N}$.

References

1. J.-Y. Audibert, *Data-dependent generalization error bounds for (noisy) classification: the PAC-Bayesian approach*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2004.
2. L. Birgé, *A new look at an old result: Fano's lemma*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2001.
3. O. Catoni, *Statistical learning theory and stochastic optimization,* Lecture notes, Saint-Flour summer school on Probability Theory, 2001, Springer, to be published.
4. ______, *A PAC-Bayesian approach to adaptive classification*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2003.
5. L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*, Springer-Verlag, 2000.
6. R.M. Dudley, *Central limit theorems for empirical measures*, Ann. Probab. **6** (1978), 899–929.
7. D. Haussler, *Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension*, Journal of Combinatorial Theory **69** (1995), 217–232, Series A.
8. C. Huber, *Lower bounds for function estimation*, Research paper in probability and statistics: Festschrift for Lucien Le Cam (1996), 245–258.
9. V.I. Koltchinskii, *On the central limit theorem for empirical measures*, Theory Probab. Math. Stat. **24** (1981), 71–82.
10. G. Lugosi, *Concentration-of-measure inequalities*, 2003, Lecture notes, Machine Learning Summer School, Canberra.
11. E. Mammen and A.B. Tsybakov, *Smooth discrimination analysis*, Ann. Stat. **27** (1999), 1808–1829.
12. P. Massart and E. Nédélec, *Risk bounds for statistical learning*, Available from http://www.math.u-psud.fr/∼massart/margin.pdf, 2003.
13. D. A. McAllester, *PAC-Bayesian model averaging*, Proceedings of the 12th annual conference on Computational Learning Theory, Morgan Kaufmann, 1999.

14. D. Pollard, *A central limit theorem for empirical measures*, J. Aust. Math. Soc., Ser. A **33** (1982), 235–248.
15. A.B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Ann. Stat. **32** (2004), no. 1.
16. A.B. Tsybakov and S. van de Geer, *Square root penalty: adaptation to the margin in classification and in edge estimation*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2004.
17. A. van der Vaart and J. Wellner, *Weak convergence and empirical processes with application to statistics*, John Wiley & Sons, New York, 1996.

PAC-Bayesian Generic Chaining

Jean-Yves Audibert [*]
Université Paris VI
Laboratoire de Probabilités et Modèles aléatoires
175 rue du Chevaleret
75013 Paris - France
jyaudibe@ccr.jussieu.fr

Olivier Bousquet
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
D-72076 Tübingen - Germany
olivier.bousquet@tuebingen.mpg.de

Abstract

There exist many different generalization error bounds for classification. Each of these bounds contains an improvement over the others for certain situations. Our goal is to combine these different improvements into a single bound. In particular we combine the PAC-Bayes approach introduced by McAllester [1], which is interesting for averaging classifiers, with the optimal union bound provided by the generic chaining technique developed by Fernique and Talagrand [2]. This combination is quite natural since the generic chaining is based on the notion of majorizing measures, which can be considered as priors on the set of classifiers, and such priors also arise in the PAC-bayesian setting.

1 Introduction

Since the first results of Vapnik and Chervonenkis on uniform laws of large numbers for classes of $\{0,1\}$-valued functions, there has been a considerable amount of work aiming at obtaining generalizations and refinements of these bounds. This work has been carried out by different communities. On the one hand, people developing empirical processes theory like Dudley and Talagrand (among others) obtained very interesting results concerning the behaviour of the suprema of empirical processes. On the other hand, people exploring learning theory tried to obtain refinements for specific algorithms with an emphasis on data-dependent bounds.
One crucial aspect of all the generalization error bounds is that they aim at controlling the behaviour of the function that is returned by the algorithm. This function is data-dependent and thus unknown before seeing the data. As a consequence, if one wants to make statements about its behaviour (e.g. the difference between its empirical error and true error), one has to be able to *predict* which function is likely to be chosen by the algorithm. But

[*] Secondary affiliation: CREST, Laboratoire de Finance et Assurance, Malakoff, France

since this cannot be done exactly, there is a need to provide guarantees that hold simultaneously for several candidate functions. This is known as the union bound. The way to perform this union bound optimally is now well mastered in the empirical processes community.

In the learning theory setting, one is interested in bounds that are as algorithm and data dependent as possible. This particular focus has made concentration inequalities (see e.g. [3]) popular as they allow to obtain data-dependent results in an effortless way. Another aspect that is of interest for learning is the case where the classifiers are randomized or averaged. McAllester [1, 4] has proposed a new type of bound that takes the randomization into account in a clever way.

Our goal is to combine several of these improvements, bringing together the power of the majorizing measures as an optimal union bound technique and the power of the PAC-Bayesian bounds that handle randomized predictions efficiently, and obtain a generalization of both that is suited for learning applications.

The paper is structured as follows. Next section introduces the notation and reviews the previous improved bounds that have been proposed. Then we give our main result and discuss its applications, showing in particular how to recover previously known results. Finally we give the proof of the presented results.

2 Previous results

We first introduce the notation and then give an overview of existing generalization error bounds. We consider an input space $\mathcal{X}$, an output space $\mathcal{Y}$ and a probability distribution P on the product space $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$. Let $Z \triangleq (X, Y)$ denote a pair of random variables distributed according to P and for a given integer n, let $Z_1, \ldots, Z_n$ and $Z'_1, \ldots, Z'_n$ be two independent samples of n independent copies of Z. We denote by P_n, P'_n and P_{2n} the empirical measures associated respectively to the first, the second and the union of both samples.

To each function $g : \mathcal{X} \to \mathcal{Y}$ we associate the corresponding loss function $f : \mathcal{Z} \to \mathbb{R}$ defined by $f(z) = L[g(x), y]$ where L is a loss function. In classification, the loss function is $L = \mathbb{I}_{g(x) \neq y}$ where $\mathbb{I}$ denotes the indicator function. $\mathcal{F}$ will denote a set of such functions. For such functions, we denote their expectation under P by Pf and their empirical expectation by $P_n f$ (i.e. $P_n f = n^{-1} \sum_{i=1}^n f(Z_i)$). $\mathbb{E}_n$, $\mathbb{E}'_n$ and $\mathbb{E}_{2n}$ denote the expectation with respect to the first, second and union of both training samples.

We consider the pseudo-distances $d^2(f_1, f_2) = P(f_1 - f_2)^2$ and similarly d_n, d'_n and d_{2n}. We define the covering number $N(\mathcal{F}, \epsilon, d)$ as the minimum number of balls of radius ϵ needed to cover $\mathcal{F}$ in the pseudo-distance d.

We denote by ρ and π two probability measures on the space $\mathcal{F}$, so that $\rho P f$ will actually mean the expectation of Pf when f is sampled according to the probability measure ρ. For two such measures, $K(\rho, \pi)$ will denote their Kullback-Leibler divergence ($K(\rho, \pi) = \rho \log \frac{d\rho}{d\pi}$ when ρ is absolutely continuous with respect to π and $K(\rho, \pi) = +\infty$ otherwise). Also, β denotes some positive real number while C is some positive constant (whose value may differ from line to line) and $\mathcal{M}^1_+(\mathcal{F})$ is the set of probability measures on $\mathcal{F}$. We assume that the functions in $\mathcal{F}$ have range in $[a, b]$.

Generalization error bounds give an upper bound on the difference between the true and empirical error of functions in a given class, which holds with high probability with respect to the sampling of the training set.

Single function. By Hoeffding's inequality one easily gets that for each fixed $f \in \mathcal{F}$, with probability at least $1 - \beta$,

$$Pf - P_n f \leq C \sqrt{\frac{\log 1/\beta}{n}} \,. \tag{1}$$

Finite union bound. It is easy to convert the above statement into one which is valid simultaneously for a finite set of functions $\mathcal{F}$. The simplest form of the union bound gives that with probability at least $1-\beta$,

$$\forall f \in \mathcal{F},\; Pf - P_n f \leq C\sqrt{\frac{\log|\mathcal{F}| + \log 1/\beta}{n}}\,. \tag{2}$$

Symmetrization. When $\mathcal{F}$ is infinite, the trick is to introduce the second sample $Z_1', \ldots, Z_n'$ and to consider the set of vectors formed by the values of each function in $\mathcal{F}$ on the double sample. When the functions have values in $\{0,1\}$, this is a finite set and the above union bound applies. This idea was first used by Vapnik and Chervonenkis [5] to obtain that with probability at least $1-\beta$,

$$\forall f \in \mathcal{F},\; Pf - P_n f \leq C\sqrt{\frac{\log \mathbb{E}_{2n} N(\mathcal{F}, 1/n, d_{2n}) + \log 1/\beta}{n}}\,. \tag{3}$$

Weighted union bound and localization. The finite union bound can be directly extended to the countable case by introducing a probability distribution π over $\mathcal{F}$ which weights each function and gives that with probability at least $1-\beta$,

$$\forall f \in \mathcal{F},\; Pf - P_n f \leq C\sqrt{\frac{\log 1/\pi(f) + \log 1/\beta}{n}}\,. \tag{4}$$

It is interesting to notice that now the bound depends on the actual function f being considered and not just on the set $\mathcal{F}$. This can thus be called a *localized* bound.

Variance. Since the deviations between Pf and $P_n f$ for a given function f actually depend on its variance (which is upper bounded by Pf^2/n or Pf/n when the functions are in $[0,1]$), one can refine (1) into

$$Pf - P_n f \leq C\left(\sqrt{\frac{Pf^2 \log 1/\beta}{n}} + \frac{\log 1/\beta}{n}\right)\,, \tag{5}$$

and combine this improvement with the above union bounds. This was done by Vapnik and Chervonenkis [5] (for functions in $\{0,1\}$).

Averaging. Consider a probability distribution ρ defined on a countable $\mathcal{F}$, take the expectation of (4) with respect to ρ and use Jensen's inequality. This gives with probability at least $1-\beta$,

$$\forall \rho,\; \rho(Pf - P_n f) \leq C\sqrt{\frac{K(\rho,\pi) + H(\rho) + \log 1/\beta}{n}}\,,$$

where $H(\rho)$ is the Shannon entropy. The l.h.s. is the difference between true and empirical error of a randomized classifier which uses ρ as weights for choosing the decision function (independently of the data). The PAC-Bayes bound [1] is a refined version of the above bound since it has the form (for possibly uncountable $\mathcal{F}$)

$$\forall \rho,\; \rho(Pf - P_n f) \leq C\sqrt{\frac{K(\rho,\pi) + \log n + \log 1/\beta}{n}}\,. \tag{6}$$

To some extent, one can consider that the PAC-Bayes bound is a refined union bound where the gain happens when ρ is not concentrated on a single function (or more precisely ρ has entropy larger than $\log n$).

Rademacher averages. The quantity $\mathbb{E}_n \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{n}\sum \sigma_i f(Z_i)$, where the σ_i are independent random signs ($+1, -1$ with probability $1/2$), called the Rademacher average for $\mathcal{F}$, is, up to a constant equal to $\mathbb{E}_n \sup_{f\in\mathcal{F}} Pf - P_n f$ which means that it best captures the complexity of $\mathcal{F}$. One has with probability $1-\beta$,

$$\forall f \in \mathcal{F},\; Pf - P_n f \leq C\left(\frac{1}{n}\mathbb{E}_n \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum \sigma_i f(Z_i) + \sqrt{\frac{\log 1/\beta}{n}}\right)\,. \tag{7}$$

Chaining. Another direction in which the union bound can be refined is by considering finite covers of the set of function at different scales. This is called the *chaining* technique, pioneered by Dudley (see e.g. [6]) since one constructs a chain of functions that approximate a given function more and more closely. The results involve the Koltchinskii-Pollard entropy integral as, for example in [7], with probability $1-\beta$,

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq C\left(\frac{1}{\sqrt{n}}\mathbb{E}_n \int_0^\infty \sqrt{\log N(\mathcal{F}, \epsilon, d_n)}d\epsilon + \sqrt{\frac{\log 1/\beta}{n}}\right). \tag{8}$$

Generic chaining. It has been noticed by Fernique and Talagrand that it is possible to capture the complexity in a better way than using minimal covers by considering majorizing measures (essentially optimal for Gaussian processes). Let $r > 0$ and $(\mathcal{A}_j)_{j\geq 1}$ be partitions of $\mathcal{F}$ of diameter r^{-j} w.r.t. the distance d_n such that $\mathcal{A}_{j+1}$ refines $\mathcal{A}_j$. Using (7) and techniques from [2] we obtain that with probability $1-\beta$, $\forall f \in \mathcal{F}$

$$Pf - P_n f \leq C\left(\frac{1}{\sqrt{n}}\mathbb{E}_n \inf_{\pi\in\mathcal{M}_+^1(\mathcal{F})} \sup_{f\in\mathcal{F}} \sum_{j=1}^{\infty} r^{-j}\sqrt{\log 1/\pi A_j(f)} + \sqrt{\frac{\log 1/\beta}{n}}\right). \tag{9}$$

If one takes partitions induced by minimal covers of $\mathcal{F}$ at radii r^{-j}, one recovers (8) up to a constant.

Concentration. Using concentration inequalities as in [3] for example, one can get rid of the expectation appearing in the r.h.s. of (3), (8), (7) or (9) and thus obtain a bound that can be computed from the data.

Refining the bound (7) is possible as one can localize it (see e.g. [8]) by computing the Rademacher average only on a small ball around the function of interest. So this comes close to combining all improvements. However it has not been combined with the PAC-Bayes improvement. Our goal is to try and combine all the above improvements.

3 Main results

Let $\mathcal{F}$ be as defined in section 2 with $a = 0, b = 1$ and $\pi \in \mathcal{M}_+^1(\mathcal{F})$. Instead of using partitions as in (9) we use approximating sets (which also induce partitions but are easier to handle here). Consider a sequence S_j of embedded finite subsets of $\mathcal{F}$: $\{f_0\} \triangleq S_0 \subset \cdots \subset S_{j-1} \subset S_j \subset \cdots$.
Let $p_j : \mathcal{F} \to S_j$ be maps (which can be thought of as projections) satisfying $p_j(f) = f$ for $f \in S_j$ and $p_{j-1} \circ p_j = p_{j-1}$.

The quantities π, S_j and p_j are allowed to depend on X_1^{2n} in an exchangeable way (i.e. exchanging X_i and X_i' does not affect their value). For a probability distribution ρ on $\mathcal{F}$, define its j-th projection as $\rho_j = \sum_{f\in S_j}\rho\{f' : p_j(f') = f\}\delta_f$, where δ_f denotes the Dirac measure on f. To shorten notations, we denote the average distance between two successive "projections" by $\rho d_j^2 \triangleq \rho d_{2n}^2[p_j(f), p_{j-1}(f)]$. Finally, let $\Delta_{n,j}(f) \triangleq P_n'[f - p_j(f)] - P_n[f - p_j(f)]$.

Theorem 1 *If the following condition holds*

$$\lim_{j\to+\infty}\sup_{f\in\mathcal{F}} \Delta_{n,j}(f) = 0, \qquad \textit{a.s.} \tag{10}$$

then for any $0 < \beta < 1/2$, *with probability at least* $1-\beta$, *for any distribution* ρ, *we have*

$$\rho P_n' f - P_n' f_0 \leq \rho P_n f - P_n f_0 + 5\sum_{j=1}^{+\infty}\sqrt{\frac{\rho d_j^2 K(\rho_j, \pi_j)}{n}} + \frac{1}{\sqrt{n}}\sum_{j=1}^{+\infty}\chi_j(\rho d_j^2),$$

where $\chi_j(x) = 4\sqrt{x \log\left(4j^2\beta^{-1}\log(e^2/x)\right)}$.

Remark 1 *Assumption* (10) *is not very restrictive. For instance, it is satisfied when* $\mathcal{F}$ *is finite, or when* $\lim_{j\to+\infty}\sup_{f\in\mathcal{F}} |f - p_j(f)| = 0$, *almost surely or also when the empirical process* $\left[f \mapsto Pf - P_n f\right]$ *is uniformly continuous (which happens for classes with finite* VC *dimension in particular) and* $\lim_{j\to+\infty}\sup_{f\in\mathcal{F}} d_{2n}(f, p_j(f)) = 0$.

Remark 2 *Let* $\mathcal{G}$ *be a model (i.e. a set of prediction functions). Let* $\tilde{g}$ *be a reference function (not necessarily in* $\mathcal{G}$*). Consider the class of functions* $\mathcal{F} = \left\{z \mapsto L[g(x), y] : g \in \mathcal{G} \cup \{\tilde{g}\}\right\}$. *Let* $f_0 = L[\tilde{g}(x), y]$. *The previous theorem compares the risk on the second sample of any (randomized) estimator with the risk on the second sample of the reference function* $\tilde{g}$.

Now let us give a version of the previous theorem in which the second sample does not appear.

Theorem 2 *If the following condition holds*

$$\lim_{j\to+\infty}\sup_{f\in\mathcal{F}} \mathbb{E}'_n\left[\Delta_{n,j}(f)\right] = 0, \qquad \textit{a.s.} \tag{11}$$

then for any $0 < \beta < 1/2$, *with probability at least* $1-\beta$, *for any distribution* ρ, *we have*

$$\rho Pf - Pf_0 \le \rho P_n f - P_n f_0 + 5\sum_{j=1}^{+\infty}\sqrt{\frac{\mathbb{E}'_n[\rho d_j^2]\mathbb{E}'_n[K(\rho_j, \pi_j)]}{n}} + \frac{1}{\sqrt{n}}\sum_{j=1}^{+\infty}\chi_j\left(\mathbb{E}'_n[\rho d_j^2]\right).$$

4 Discussion

We now discuss in which sense the result presented above combines several previous improvements in a single bound.

Notice that our bound is localized in the sense that it depends on the function of interest (or rather on the averaging distribution ρ) and does not involve a supremum over the class.

Also, the union bound is performed in an optimal way since, if one plugs in a distribution ρ concentrated on a single function, takes a supremum over $\mathcal{F}$ in the r.h.s., and upper bounds the squared distance by the diameter of the partition, one recovers a result similar to (9) up to logarithmic factors but which is localized. Also, when two successive projections are identical, they do not enter in the bound (which comes from the fact that the variance weights the complexity terms). Moreover Theorem 1 also includes the PAC-Bayesian improvement for averaging classifiers since if one considers the set $S_1 = \mathcal{F}$ one recovers a result similar to McAllester's (6) which in addition contains the variance improvement such as in [9].

Finally due to the power of the generic chaining, it is possible to upper bound our result by Rademacher averages, up to logarithmic factors (using the results of [10] and [11]).

As a remark, the choice of the sequence of sets S_j can generally be done by taking successive covers of the hypothesis space with geometrically decreasing radii.

However, the obtained bound is not completely empirical since it involves the expectation with respect to an extra sample. In the transduction setting, this is not an issue, it is even an advantage as one can use the unlabeled data in the computation of the bound. However, in the induction setting, this is a drawback. Future work will focus on using concentration inequalities to give a fully empirical bound.

5 Proofs

Proof of Theorem 1: The proof is inspired by previous works on PAC-bayesian bounds [12, 13] and on the generic chaining [2]. We first prove the following lemma.

Lemma 1 *For any $\beta > 0$, $\lambda > 0$, $j \in \mathbb{N}^*$ and any exchangeable function $\pi : \mathcal{X}^{2n} \to \mathcal{M}_+^1(\mathcal{F})$, with probability at least $1-\beta$, for any probability distribution $\rho \in \mathcal{M}_+^1(\mathcal{F})$, we have*

$$\rho\Big\{P'_n[p_j(f) - p_{j-1}(f)] - P_n[p_j(f) - p_{j-1}(f)]\Big\} \\ \leq \tfrac{2\lambda}{n}\rho d_{2n}^2[p_j(f), p_{j-1}(f)] + \tfrac{K(\rho,\pi)+\log(\beta^{-1})}{\lambda}.$$

Proof Let $\lambda > 0$ and let $\pi : \mathcal{X}^{2n} \to \mathcal{M}_+^1(\mathcal{F})$ be an exchangeable function. Introduce the quantity $\Delta_i \triangleq p_j(f)(Z_{n+i}) - p_{j-1}(f)(Z_{n+i}) + p_{j-1}(f)(Z_i) - p_j(f)(Z_i)$ and

$$h \triangleq \lambda P'_n\big[p_j(f) - p_{j-1}(f)\big] - \lambda P_n\big[p_j(f) - p_{j-1}(f)\big] - \frac{2\lambda^2}{n} d_{2n}\big[p_j(f), p_{j-1}(f)\big]. \quad (12)$$

By using the exchangeability of π, for any $\sigma \in \{-1; +1\}^n$, we have

$$\begin{aligned} \mathbb{E}_{2n}\pi e^h &= \mathbb{E}_{2n}\pi e^{-\frac{2\lambda^2}{n} d_{2n}[p_j(f),p_{j-1}(f)] + \frac{\lambda}{n}\sum_{i=1}^n \Delta_i} \\ &= \mathbb{E}_{2n}\pi e^{-\frac{2\lambda^2}{n} d_{2n}[p_j(f),p_{j-1}(f)] + \frac{\lambda}{n}\sum_{i=1}^n \sigma_i\Delta_i}. \end{aligned}$$

Now take the expectation wrt σ, where σ is a n-dimensional vector of Rademacher variables. We obtain

$$\begin{aligned} \mathbb{E}_{2n}\pi e^h &= \mathbb{E}_{2n}\pi e^{-\frac{2\lambda^2}{n} d_{2n}[p_j(f),p_{j-1}(f)]} \prod_{i=1}^n \cosh\left(\tfrac{\lambda}{n}\Delta_i\right) \\ &\leq \mathbb{E}_{2n}\pi e^{-\frac{2\lambda^2}{n} d_{2n}[p_j(f),p_{j-1}(f)]} e^{\sum_{i=1}^n \frac{\lambda^2}{2n^2}\Delta_i^2} \end{aligned}$$

where at the last step we use that $\cosh s \leq e^{\frac{s^2}{2}}$. Since

$$\Delta_i^2 \leq 2\big[p_j(f)(Z_{n+i}) - p_{j-1}(f)(Z_{n+i})\big]^2 + 2\big[p_j(f)(Z_i) - p_{j-1}(f)(Z_i)\big]^2,$$

we obtain that for any $\lambda > 0$, $\mathbb{E}_{2n}\pi e^h \leq 1$. Therefore, for any $\beta > 0$, we have

$$\mathbb{E}_{2n}\mathbb{I}_{\log \pi e^{h+\log\beta} > 0} = \mathbb{E}_{2n}\mathbb{I}_{\pi e^{h+\log\beta} > 1} \leq \mathbb{E}_{2n}\pi e^{h+\log\beta} \leq \beta, \quad (13)$$

On the event $\big\{\log \pi e^{h+\log\beta} \leq 0\big\}$, by the Legendre's transform, for any probability distribution $\rho \in \mathcal{M}_+^1(\mathcal{F})$, we have

$$\rho h + \log\beta \leq \log \pi e^{h+\log\beta} + K(\rho,\pi) \leq K(\rho,\pi), \quad (14)$$

which proves the lemma. ■

Now let us apply this result to the projected measures π_j and ρ_j. Since, by definition, π, S_j and p_j are exchangeable, π_j is also exchangeable. Since $p_j(f) = f$ for any $f \in S_j$, with probability at least $1-\beta$, uniformly in ρ, we have

$$\rho_j\Big\{P'_n[f - p_{j-1}(f)] - P_n[p_j(f) - p_{j-1}(f)]\Big\} \leq \frac{2\lambda}{n}\rho_j d_{2n}^2[f, p_{j-1}(f)] + \frac{K'_j}{\lambda},$$

where $K'_j \triangleq K(\rho_j, \pi_j) + \log(\beta^{-1})$. By definition of ρ_j, it implies that

$$\rho\Big\{P'_n[p_j(f) - p_{j-1}(f)] - P_n[p_j(f) - p_{j-1}(f)]\Big\} \leq \frac{2\lambda}{n}\rho d_{2n}^2[p_j(f), p_{j-1}(f)] + \frac{K'_j}{\lambda}. \quad (15)$$

To shorten notations, define $\rho d_j^2 \triangleq \rho d_{2n}^2[p_j(f), p_{j-1}(f)]$ and $\rho\Delta_j \triangleq \rho\big\{P'_n[p_j(f) - p_{j-1}(f)] - P_n[p_j(f) - p_{j-1}(f)]\big\}$. The parameter λ minimizing the RHS of the previous equation depends on ρ. Therefore, we need to get a version of this inequality which holds uniformly in λ.

First let us note that when $\rho d_j^2 = 0$, we have $\rho\Delta_j = 0$. When $\rho d_j^2 > 0$, let $m\sqrt{\frac{\log 2}{2n}}$ and $\lambda_k = me^{k/2}$ and let b be a function from $\mathbb{R}^*$ to $(0,1]$ such that $\sum_{k\geq 1} b(\lambda_k) \leq 1$. From the previous lemma and a union bound, we obtain that for any $\beta > 0$ and any integer j with probability at least $1-\beta$, for any $k \in \mathbb{N}^*$ and any distribution ρ, we have

$$\rho\Delta_j \leq \frac{2\lambda_k}{n}\rho d_j^2 + \frac{K(\rho_j,\pi_j) + \log\big([b(\lambda_k)]^{-1}\beta^{-1}\big)}{\lambda_k}.$$

Let us take the function b such that $\left[\lambda \mapsto \frac{\log\left([b(\lambda)]^{-1}\right)}{\lambda}\right]$ is continuous and decreasing. Then there exists a parameter $\lambda^* > 0$ such that $\frac{2\lambda^*}{n}\rho d_j^2 = \frac{K(\rho_j,\pi_j)+\log\left([b(\lambda^*)]^{-1}\beta^{-1}\right)}{\lambda^*}$. For any $\beta < 1/2$, we have $(\lambda^*)^2\rho d_j^2 \geq \frac{\log 2}{2}n$, hence $\lambda^* \geq m$. So there exists an integer $k \in \mathbb{N}^*$ such that $\lambda_k e^{-1/2} \leq \lambda^* \leq \lambda_k$. Then we have

$$\begin{array}{rcl}\rho\Delta_j & \leq & \frac{2\lambda^*}{n}\sqrt{e}\rho d_j^2 + \frac{K(\rho_j,\pi_j)+\log\left([b(\lambda_*)]^{-1}\beta^{-1}\right)}{\lambda_*} \\ & = & (1+\sqrt{e})\sqrt{\frac{2}{n}\rho d_j^2\Big[K(\rho_j,\pi_j) + \log\left([b(\lambda_*)]^{-1}\beta^{-1}\right)\Big]}.\end{array} \qquad (16)$$

To have an explicit bound, it remains to find an upperbound of $[b(\lambda^*)]^{-1}$. When b is decreasing, this comes down to upperbouding λ^*. Let us choose $b(\lambda) = \frac{1}{[\log(\frac{e^2\lambda}{m})]^2}$ when $\lambda \geq m$ and $b(\lambda) = 1/4$ otherwise. Since $b(\lambda_k) = \frac{4}{(k+4)^2}$, we have $\sum_{k\geq 1} b(\lambda_k) \leq 1$. Tedious computations give $\lambda^* \leq 7m\frac{\sqrt{K'_j}}{\rho d_j^2}$ which combined with (16), yield

$$\rho\Delta_j \leq 5\sqrt{\frac{\rho d_j^2 K(\rho_j,\pi_j)}{n}} + 3.75\sqrt{\frac{\rho d_j^2}{n}\log\left(2\beta^{-1}\log\left[\frac{e^2}{\rho d_j^2}\right]\right)}.$$

By simply using a union bound with weights taken proportional to $1/j^2$, we have that the previous inequation holds uniformly in $j \in \mathbb{N}^*$ provided that β^{-1} is replaced with $\frac{\pi^2}{6}j^2\beta^{-1}$ $\big($since $\sum_{j\in\mathbb{N}^*} 1/j^2 = \pi^2/6 \approx 1.64\big)$. Notice that

$$\rho\big[P'_n f - P'_n f_0 + P_n f_0 - P_n f\big] = \rho\Delta_{n,J}(f) + \sum_{j=1}^{J}\rho_j\big[(P'_n - P_n)f - (P'_n - P_n)p_{j-1}(f)\big]$$

because $p_{j-1} = p_{j-1} \circ p_j$. So, with probability at least $1-\beta$, for any distribution ρ, we have

$$\begin{array}{l}\rho\big[P'_n f - P'_n f_0 + P_n f_0 - P_n f\big] \leq \sup_{\mathcal{F}} \Delta_{n,J} + 5\sum_{j=1}^{J}\sqrt{\frac{\rho d_j^2 K(\rho_j,\pi_j)}{n}} \\ \qquad\qquad +3.75\sum_{j=1}^{J}\sqrt{\frac{\rho d_j^2}{n}\log\left(3.3j^2\beta^{-1}\log\left[\frac{e^2}{\rho d_j^2}\right]\right)}.\end{array}$$

Making $J \to +\infty$, we obtain theorem 1. □

Proof of Theorem 2: It suffices to modify slightly the proof of theorem 1. Introduce $U \triangleq \sup_\rho \big\{\rho h + \log\beta - K(\rho,\pi)\big\}$, where h is still defined as in equation (12). Inequations (14) implies that $\mathbb{E}_{2n}e^U \leq \beta$. By Jensen's inequality, we get $\mathbb{E}_n e^{\mathbb{E}'_n U} \leq \beta$, hence $\mathbb{E}_n\Big\{\mathbb{E}'_n U \geq 0\Big\} \leq \beta$. So with probability at least $1-\beta$, we have $\sup_\rho \mathbb{E}'_n\big\{\rho h + \log\beta - K(\rho,\pi)\big\} \leq \mathbb{E}'_n U \leq 0$. □

6 Conclusion

We have obtained a generalization error bound for randomized classifiers which combines several previous improvements. It contains an optimal union bound, both in the sense of optimally taking into account the metric structure of the set of functions (via the majorizing measure approach) and in the sense of taking into account the averaging distribution. We believe that this is a very natural way of combining these two aspects as the result relies on the comparison of a majorizing measure which can be thought of as a prior probability distribution and a randomization distribution which can be considered as a posterior distribution.

Future work will focus on giving a totally empirical bound (in the induction setting) and investigating possible constructions for the approximating sets S_j.

References

[1] D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 230–234. ACM Press, 1998.

[2] M. Talagrand. Majorizing measures: The generic chaining. *Annals of Probability*, 24(3):1049–1103, 1996.

[3] S. Boucheron, G. Lugosi, and S. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

[4] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*. ACM Press, 1999.

[5] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979).

[6] R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2–142, 1984.

[7] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer Verlag, New York, 2001.

[8] P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. Preprint, 2003.

[9] D. A. McAllester. Simplified pac-bayesian margin bounds. In *Proceedings of Computational Learning Theory (COLT)*, 2003.

[10] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer-Verlag, Berlin, 1991.

[11] M. Talagrand. The Glivenko-Cantelli problem. *Annals of Probability*, 6:837–870, 1987.

[12] O. Catoni. Localized empirical complexity bounds and randomized estimators, 2003. Preprint.

[13] J.-Y. Audibert. Data-dependent generalization error bounds for (noisy) classification: a PAC-bayesian approach. 2003. Work in progress.