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

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

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

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

[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

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

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 $(\mathcal{F},\mathcal{T})$.

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

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

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 $h:A\mapsto\mathbb{R}$,

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

where, by convention:

$$\left\{\begin{array}{l}\mu h \triangleq \sup\limits_{H\in\mathbb{R}}\mu(H\wedge h)\\ \mu h - K(\mu,\nu) = -\infty \text{ if } K(\mu,\nu)=+\infty\end{array}\right. .$$

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 $\gamma^{\mathcal{F}} = \psi^{-1}$.
- 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 $1 - \epsilon$,

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

This leads to

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

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

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

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

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 $\hat{f}_{\mathrm{ERM},\mathcal{F}_k}$ denote the ERM-algorithm relative to the model $\mathcal{F}_k$. The SRM[5]-algorithm is to use $\hat{f}_{\mathrm{ERM},\mathcal{F}_{\hat{k}}}$ where

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

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

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

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 $\epsilon > 0$,

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

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

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

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

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

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

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

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 $\gamma^{\mathcal{F}}(\epsilon) = \sqrt{8\frac{\log(\epsilon^{-1})+\log[4\$^{\mathcal{F}}(2N)]}{N}}$ where the shatter coefficient $\$^{\mathcal{F}}(N)$ 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 $\log[\$^{\mathcal{F}}(N)] \leq h\log(eN/h)$.

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 $\epsilon > 0$,

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

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

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

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:

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

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

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

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 $\hat{\mathcal{F}}_h \triangleq \left\{\hat{f}_{(X_{i_j},y_i)_{j=1}^h} : (i_1,\dots,i_h)\in\{1,\dots,2N\}^h, y_1^h\in\mathcal{Y}^h\right\}$. 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 $\left(\mathbb{P}^{\otimes 2N}\right)_*$-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 $k(f) \triangleq \max\left\{0 \leq k \leq K; \mathcal{C}(f_k) \leq \mathcal{C}(f)\right\}$,*

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

-

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

Proof. See Section 9.1. □

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

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

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

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

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

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

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*

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

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 $k \in \{1, \dots, K\}$, 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 $\{X_1, \dots, X_{2N}\}$. Using this prior distribution, we have

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

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:
$$-\tfrac{1}{\lambda}\log \pi \exp\left(-\lambda r'\right) = \int_0^1 \pi_{-\gamma\lambda r'} r' d\gamma.$$
- 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* $\pi_{-\lambda_{u(k)} r} r' \le \pi_{-\lambda_{u(k-1)} r} r'$,
- $\pi_{-\lambda_{u(K)} r} r' \le \min_{1\le j\le \log N} \mathbb{G}_{\mathrm{loc}}(j)$.

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

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

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 $\log \pi_{-\lambda r} \exp\left(\frac{\lambda^2}{N} \pi_{-\lambda r} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right)$, we can approximate this quantity by

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

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

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

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 $\pi_0(I, y_1^{2N}, d\theta) \geq \frac{1-\alpha}{\alpha^2} \left(\frac{\alpha}{2N|\mathcal{Y}|}\right)^h \pi_1(d\theta)$ when $y_i = 0$ for $i > h$. The model is defined as

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

The prior distribution on the model is given by: for any measurable set $A \subset \hat{\mathcal{F}}$, $\pi(A) \triangleq \pi_0\big\{(I, y_1^{2N}, \theta) \in \mathcal{I} \times \mathcal{Y}^{2N} \times \Theta : \hat{F}_{(X_{i_1},y_1),\ldots,(X_{i_h},y_h),\theta} \in A\big\}$. 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*

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

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

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

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 $h \in \mathbb{N}^*$, $\mathcal{I}_h \triangleq \{1,\ldots,N\}^h$. Any $I \in \mathcal{I}_h$ can be written as $I = \{i_1,\ldots,i_h\}$. Define $I^c \triangleq \{1,\ldots,N\} - \{i_1,\ldots,i_h\}$ and $Z_I \triangleq \big(Z_{i_1},\ldots,Z_{i_h}\big)$. The law of the random variable Z_I will be denoted $\mathbb{P}^I$. For any $J \subset \{1,\ldots,N\}$, 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 $\mathcal{I} \triangleq \bigcup_{2\leq h\leq N-1} \mathcal{I}_h$ and $\theta, \theta_1, \theta_2$ in Θ, introduce

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

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

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

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 $k \geq 1$, define $I_k \in \bigcup_{2\leq h\leq N-1} \mathcal{I}_h$ and $\theta_k \in \Theta$ such that*

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

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 $(I,\theta) \in \mathcal{I}\times\Theta$, $k(I,\theta) \triangleq \max\left\{0 \leq k \leq K; \mathcal{C}(I_k,\theta_k) \leq \mathcal{C}(I,\theta)\right\}$. 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 $\left(\mathbb{P}^{\otimes N}\right)_*$-probability at least $1-2\epsilon$, for any $k \in \{1,\ldots,K\}$, we have*

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

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

and consequently

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

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 $\epsilon > 0$,

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

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

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

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)

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

The iterative scheme is not infinite: there exists $K \in \mathbb{N}$ such that $u(K)$ exists but not $u(K+1)$. With $\big(\mathbb{P}^{\otimes 2N}\big)_*$-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* $\pi_{-\lambda_{u(k)} r} R \le \pi_{-\lambda_{u(k-1)} r} R$,
- $\pi_{\lambda_{u(K)} r} R \le \min_{1\le j\le \log N} \mathbb{G}_{\mathrm{loc}}(j)$.

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 $\rho r' - r'(\tilde{f}) \leq \rho r - r(\tilde{f}) + \min_{\lambda\in\Lambda}\big\{\frac{2\lambda}{N}\rho\bar{\bar{\mathbb{P}}}_{\cdot,\tilde{f}} + \frac{\mathcal{K}}{\lambda}\big\}$,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 $\rho_2r'-\rho_1r'+\rho_1r-\rho_2r$.

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*

$$\rho_2r'-\rho_1r'+\rho_1r-\rho_2r\leq\min_{\lambda\in[\sqrt{N};N]}\left\{\frac{2\lambda}{N}(\rho_1\otimes\rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}+\zeta\frac{\mathcal{K}''_{1,2}}{\lambda}\right\}. \tag{4.4}$$

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 $\left(\pi_{-C[r+r']}\right)$ 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 $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-\epsilon$, for any $\rho \in \mathcal{M}_+^1(\mathcal{F})$, we have*

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

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 $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-3\epsilon$, for any distributions $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$, we have*

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

where

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

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 $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-3\epsilon$, for any $(\lambda, \lambda', \lambda'') \in \Lambda^3$, we have*

$$\pi_{-\lambda'' r} r' - \pi_{-\lambda' r} r' + \pi_{-\lambda' r} r - \pi_{-\lambda'' r} r \leq \frac{2\lambda}{N}(\pi_{-\lambda' r} \otimes \pi_{-\lambda'' r})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\bar{\bar{\mathcal{K}}}}{\lambda}, \tag{4.7}$$

where

$$\bar{\bar{\mathcal{K}}} \triangleq 2\log \pi_{-\lambda' r} \exp\left(\tfrac{\lambda'^2}{N}\pi_{-\lambda' r}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) + 2\log \pi_{-\lambda'' r} \exp\left(\tfrac{\lambda''^2}{N}\pi_{-\lambda'' r}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) + 3\log\left(|\Lambda|^3\epsilon^{-1}\right).$$

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 $\xi \geq 0$, we have*

- *when $\xi < 1$, with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-2\epsilon$, for any randomizing distribution $\rho \in \mathcal{M}_+^1(\mathcal{F})$,*

$$\rho r' - \breve{\rho} r' \quad \leq \quad \rho r - \breve{\rho} r + \tfrac{1+\xi}{1-\xi}\tfrac{2\lambda}{N}\big(\rho \otimes \breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{K(\rho, \pi_{-2\xi\lambda r}) + (1+\xi)\log(\epsilon^{-1})}{(1-\xi)\lambda}. \tag{4.8}$$

- *with* $\big(\mathbb{P}^{\otimes 2N}\big)_*$*-probability at least* $1-2\epsilon$*, for any randomizing distribution* $\rho \in \mathcal{M}_+^1(\mathcal{F})$*,*

$$\breve{\rho} r' - \rho r' \quad \leq \quad \breve{\rho} r - \rho r + \tfrac{2\lambda}{N}\big(\rho \otimes \breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{K(\rho,\pi_{-2\xi\lambda r})+(1+\xi)\log(\epsilon^{-1})}{(1+\xi)\lambda}.$$

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 $\Lambda \triangleq \big\{\sqrt{N}\zeta^k; 0 \leq k \leq \frac{\log N}{2\log\zeta}\big\}$ and $\Xi \triangleq \big\{\alpha^{-k}; 1 \leq k \leq \frac{\log(\alpha N)}{\log\alpha}\big\}$ 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 $|\Lambda||\Xi| \leq \frac{\log(\zeta^2 N)\log(\alpha N)}{2\log(\zeta)\log(\alpha)}$.

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* $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$*,*

$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \leq \tfrac{\lambda}{N} g\big(\tfrac{\lambda}{N}\big)(\rho_1 \otimes \rho_2)\mathbb{P}_{\cdot,\cdot} + \tfrac{\mathcal{K}_{1,2}}{\lambda}$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any* $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$*,*

$$(\rho_1 \otimes \rho_2)\mathbb{P}_{\cdot,\cdot} \leq \big(1 + \tfrac{\lambda}{2N}\big)(\rho_1 \otimes \rho_2)\bar{\mathbb{P}}_{\cdot,\cdot} + \tfrac{(1+\frac{\lambda}{2N})^2\mathcal{K}_{1,2}}{\lambda} \ ,$$

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*

$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \leq \bar{a}(\lambda)(\rho_1 \otimes \rho_2)\bar{\mathbb{P}}_{\cdot,\cdot} + \bar{b}(\lambda)\mathcal{K}_{1,2} \tag{4.9}$$

where $\bar{a}(\lambda) \triangleq \frac{\lambda}{N} g\big(\frac{\lambda}{N}\big)\big(1+\frac{\lambda}{2N}\big)$ *and* $\bar{b}(\lambda) \triangleq \frac{1}{\lambda}\Big[1 + \frac{\lambda}{N} g\big(\frac{\lambda}{N}\big)\big(1+\frac{\lambda}{2N}\big)^2\Big]$.

Remark 4.5. To recover a simple formulation, it suffices to note that $\bar{a}(\lambda) \leq 1.1\frac{\lambda}{N}$ and $\bar{b}(\lambda) \leq \frac{2.7}{\lambda}$ for any $0 < \lambda \leq N$.

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 $\left(\mathbb{P}^{\otimes N}\right)_*$-probability at least $1-2\epsilon$, for any $\rho \in \mathcal{M}_+^1(\mathcal{F})$, we have*

(4.10)
$$K\big(\rho, \pi_{-\lambda R}\big) \quad \leq \quad \tfrac{1}{1-\xi}\Big[K\big(\rho, \pi_{-\lambda r}\big) + \log \pi_{-\lambda r} \exp\big(\tfrac{2\lambda^2}{\xi N}\rho\overline{\mathbb{P}}_{\cdot,\cdot}\big) + \xi \log(\epsilon^{-1})\Big].$$

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

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

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 $\hat{\mathcal{F}} \triangleq \bigcup_{2 \leq h \leq N} \hat{\mathcal{F}}_h$.

Theorem 5.1. *Let $\epsilon > 0$, $\alpha \in]0;1[$ and $L \triangleq \log[(1-\alpha)^{-2}\alpha^4\epsilon^{-1}]$. With $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-\epsilon$, for any $f_1, f_2 \in \hat{\mathcal{F}}$, we have*

$$r'(f_2) - r'(f_1) \leq r(f_2) - r(f_1) + \sqrt{\tfrac{8\overline{\mathbb{P}}_{f_1,f_2}[h_1 \log(2N|\mathcal{Y}|/\alpha) + h_2 \log(2N|\mathcal{Y}|/\alpha) + L]}{N}},$$

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} (and so will be denoted $\ddot{\pi}_{Z_{I_1}, Z_{I_2}}(d\theta_1, d\theta_2)$).

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 $\left(\mathbb{P}^{\otimes N}\right)_*$-probability at least $1-\epsilon$, for any $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{I} \times \Theta)$,*

$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \leq \big(\rho_1 \otimes \rho_2\big)\Big[\tfrac{\lambda}{N_{1,2}} g\big(\tfrac{\lambda}{N_{1,2}}\big)\mathbb{P}(I_1, \theta_1, I_2, \theta_2)\Big] + \tfrac{\mathcal{K}_{1,2}}{\lambda},$$

- *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* $\theta_1,\theta_2\in\Theta$*,*

$$\begin{aligned}&R(I_2,\theta_2)-R(I_1,\theta_1)+r(I_1,\theta_1)-r(I_2,\theta_2)\\ \le\ &\sqrt{\tfrac{2[\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)+\log(\epsilon^{-1})]\mathbb{P}(I_1,\theta_1,I_2,\theta_2)}{N_{1,2}}}+\tfrac{\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)+\log(\epsilon^{-1})}{3N_{1,2}},\end{aligned}$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$*, for any* $I_1,I_2\in\mathcal{I}$ *and* $\theta_1,\theta_2\in\Theta$*,*

$$\begin{aligned}\mathbb{P}(I_1,\theta_1,I_2,\theta_2)\le\Big(&\sqrt{\bar{\mathbb{P}}(I_1,\theta_1,I_2,\theta_2)+\tfrac{\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)+\log(\epsilon^{-1})}{2N_{1,2}}}\\ &+\sqrt{\tfrac{\log\ddot{\pi}^{-1}(I_1,\theta_1,I_2,\theta_2)+\log(\epsilon^{-1})}{2N_{1,2}}}\Big)^2.\end{aligned}$$

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

$$\left\{\begin{array}{rcl}\mu\big(I_1,I_2,d(\theta_1,\theta_2)\big)&=&\rho_1(I_1,d\theta_1)\otimes\rho_2(I_2,d\theta_2)\\ \nu\big(I_1,I_2,d(\theta_1,\theta_2)\big)&=&\ddot{\pi}(I_1,d\theta_1,I_2,d\theta_2)\end{array}\right..$$

□

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 $\lambda = o(\sqrt{N})\big)$, the optimal γ is $o(1)$.

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*

$$\begin{aligned} \pi_{-\lambda r} R &\geq -\frac{\log \pi_{-\lambda R} \exp\left(-\chi\lambda R\right)}{\chi\lambda} - \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \\ &\geq \pi_{-(1+\chi)\lambda R} R - \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \end{aligned} \tag{6.7}$$

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

$$K(\pi_{-\lambda r}, \pi_{-\lambda R}) \leq \tfrac{4}{1-\gamma} \log \pi_{-\lambda R} \exp\left(\tfrac{4.1\,\lambda^2}{\gamma N} \pi_{-\lambda R} \mathbb{P}_{\cdot,\cdot}\right) + \tfrac{5\gamma}{1-\gamma} \log(4\epsilon^{-1}). \tag{6.8}$$

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

$$\mathcal{A}(\mathbb{X}) \triangleq \Big\{ \big\{ f \in \mathcal{F} : f(X_i) = \sigma_i \text{ for any } i = 1, \dots, 2N \big\}; \sigma_1^{2N} \in \{0;1\}^{2N} \Big\}.$$

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 $\big(\mathbb{P}^{\otimes 2N}\big)_*$-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 $\tilde{f}' \triangleq \operatorname{argmin}_{\mathcal{F}} r'$, we obtain

$$r'(\hat{f}_{ERM}) - r'(\tilde{f}') \leq r(\hat{f}_{ERM}) - r(\tilde{f}') + \sqrt{\frac{8 \bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM},\tilde{f}'}[2 \log N(\mathbb{X}) + \log(\epsilon^{-1})]}{N}}. \tag{7.1}$$

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 $\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, \tilde{f}'}$ 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 $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-\epsilon$, we have

$$r'(\hat{f}_{ERM}) - r'(\tilde{f}') \leq r(\hat{f}_{ERM}) - r(\tilde{f}') + \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, \tilde{f}'}[\mathcal{C}(\hat{f}_{ERM}, \tilde{f}') + \log(\epsilon^{-1})]}{N}}. \tag{7.2}$$

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 $\lambda = 0$ (i.e. by using that $\mathcal{C}(f,g) \leq \mathcal{C}_0(f) + \mathcal{C}_0(g)$), 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 $\lambda \geq 0$

$$\begin{aligned} c_\lambda \triangleq \textstyle\sum_{A \in \mathcal{A}(\mathbb{X})} \exp\big(-\lambda[r_A + r'_A]\big) &\leq 1 + 2\textstyle\sum_{k=1}^{N} \exp\big(-k\frac{\lambda}{N}\big) \\ &= 1 + 2\exp\big(-\tfrac{\lambda}{N}\big)\tfrac{1-\exp(-\lambda)}{1-\exp\left(-\frac{\lambda}{N}\right)}. \end{aligned}$$

Let $\hat{r} \triangleq r'(\hat{f}_{\mathrm{ERM}}) + r(\tilde{f}')$. Inequality (7.2) gives $\hat{r} \leq \min_{\lambda \geq 0} \sqrt{\frac{4\hat{r}[2\log c_\lambda + \lambda\hat{r} + \log(\epsilon^{-1})]}{N}}$. 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 $\hat{r} \leq \frac{8\log(2N+1)+4\log(\epsilon^{-1})}{N}$ 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 $N \leq 200$, 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*

$$\bar{\mathcal{F}} \triangleq \left\{ f \in \mathcal{F} : r(f) \leq r(\hat{f}_{ERM}) + \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM},f}[2\log N(\mathbb{X})+\log(\epsilon^{-1})]}{N}} \right\}.$$

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 $r'(\hat{f}_{\mathrm{ERM}}) - r'(\tilde{f}') \geq 0$. □

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*

$$r'(\hat{f}_{ERM}) - r'(\tilde{f}') \leq \sup_{f\in\bar{\mathcal{F}}} \left\{ r(\hat{f}_{ERM}) - r(f) + \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM},f}[2\log N(\mathbb{X})+\log(\epsilon^{-1})]}{N}} \right\}$$

To simplify, we can weaken the previous inequality into

$$r'(\hat{f}_{\mathrm{ERM}}) - r'(\tilde{f}') \leq \sqrt{\frac{8\sup_{\bar{\mathcal{F}}} \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\cdot}[2\log N(\mathbb{X})+\log(\epsilon^{-1})]}{N}}.$$

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*

$$R(f_2) - R(f_1) \leq r(f_2) - r(f_1) + \sqrt{\frac{8\mathbb{P}^{\otimes 2N}[\bar{\bar{\mathbb{P}}}_{f_1,f_2}|Z_1^N]\left\{2(\mathbb{P}^{\otimes 2N})^*[\log N(\mathbb{X})|X_1^N]+\log(\epsilon^{-1})\right\}}{N}}.$$

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*

$$\left(\mathbb{P}^{\otimes 2N}\right)^* \log N(X_1^{2N}) + \tfrac{(\log 2)\log(\epsilon^{-1})}{3}\Big(1+\sqrt{1+\tfrac{18(\mathbb{P}^{\otimes 2N})^*\log N(X_1^{2N})}{(\log 2)\log(\epsilon^{-1})}}\Big). \tag{7.6}$$

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* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$*, for any functions* $f_1, f_2 \in \mathcal{F}$*,*

$$\bar{\mathbb{P}}_{f_1,f_2} \le \mathbb{P}_{f_1,f_2} + 2\sqrt{\tfrac{\mathcal{K}}{N}\left(\mathbb{P}_{f_1,f_2} + \tfrac{\mathcal{K}}{4N}\right)} + \tfrac{\mathcal{K}}{N} \tag{7.8}$$

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

$$\mathbb{P}_{f_1,f_2} - \bar{\mathbb{P}}_{f_1,f_2} - \left(\mathbb{P}^{\otimes 2N}\right)^* \sqrt{\tfrac{4\bar{\mathbb{P}}_{f_1,f_2}[2\log N(\mathbb{X}) + \log(\epsilon^{-1})]}{N}} \le 0,$$

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 $C\left[R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f})\right]^{\frac{1}{\kappa}}$ 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* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$*, we have* $R(\hat{f}_{ERM}) - R(\tilde{f}) \le C\log(e\epsilon^{-1})\left(\frac{h}{N}\log N\right)^{\frac{\kappa}{2\kappa-1}}$.

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 $Z_1, \ldots, Z_N$ 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* $\epsilon > 0$*,* $\lambda > 0$*,* $B \triangleq \sup_{\mathcal{G}\times\mathcal{Z}} \mathcal{W}$*,* $g(u) \triangleq \frac{\exp(u)-1-u}{u^2}$*,* $a_c(\lambda) \triangleq \frac{\lambda}{N} g\left(\frac{\lambda}{N}c\right)$ *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* $\left(\mathbb{P}^{\otimes N}\right)_*$*-probability at least* $1-\epsilon$*, for any function* $f \in \mathcal{G}$*,*

$$\begin{aligned}\bar{\mathbb{P}}\mathcal{W}(f,\cdot) - \mathbb{P}\mathcal{W}(f,\cdot) &\\ \leq \inf_{x>0}\left\{a_B(x)\mathbb{P}\mathcal{W}(f,\cdot)^2 + \tfrac{\log\nu^{-1}(f)+\log(\epsilon^{-1})}{x}\right\} &\\ \leq \sqrt{\tfrac{2[\log\nu^{-1}(f)+\log(\epsilon^{-1})]\mathbb{P}\mathcal{W}(f,\cdot)^2}{N}} + (B\vee 0)\tfrac{\log\nu^{-1}(f)+\log(\epsilon^{-1})}{3N}.&\end{aligned} \tag{8.2}$$

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*

$$\log \mathbb{E}\exp\left\{\eta(W - \mathbb{E}W)\right\} \leq \eta^2 \mathbb{E}W^2 g(\eta b).$$

Proof. We have

$$\exp(\eta W) = 1 + \eta W + \eta^2 W^2 g(\eta W).$$

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

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

which is the desired result. □

Now let us prove Theorem 8.1. We have

(8.3)

$$\begin{aligned}\mathbb{P}^{\otimes N}\Bigg(\sup_{\mu\in\mathcal{M}_+^1(\mathcal{G})}\left\{\mu\big[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2\big] - \tfrac{K(\mu,\nu)+\log(\epsilon^{-1})}{\lambda}\right\} > 0\Bigg)&\\ = \mathbb{P}^{\otimes N}\Big(\tfrac{1}{\lambda}\log\Big[\epsilon\nu\exp\left\{\lambda[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2]\right\}\Big] > 0\Big)&\\ = \mathbb{P}^{\otimes N}\Big(\epsilon\nu\exp\left\{\lambda[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2]\right\} > 1\Big)&\\ \leq \mathbb{P}^{\otimes N}\Big(\epsilon\nu\exp\left\{\lambda[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2]\right\}\Big)&\\ = \epsilon\nu\mathbb{P}^{\otimes N}\exp\left\{\lambda[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2]\right\}&\\ = \epsilon\nu\exp\left\{-\lambda a_B(\lambda)\mathbb{P}\mathcal{W}^2\right\}\Big(\mathbb{P}\exp\left\{\tfrac{\lambda}{N}[\mathcal{W} - \mathbb{P}\mathcal{W}]\right\}\Big)^N&\\ \leq \epsilon,&\end{aligned}$$

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

$$\mu\big\{\lambda\big[\bar{\mathbb{P}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_B(\lambda)\mathbb{P}\mathcal{W}^2\big]\big\} \leq K(\mu,\nu) + \log(\epsilon^{-1}).$$

Taking $\mu = \delta_f$, we obtain

$$\bar{\mathbb{P}}\mathcal{W}(f,\cdot) - \mathbb{P}\mathcal{W}(f,\cdot) \leq a_B[\lambda(f)]\mathbb{P}\mathcal{W}(f,\cdot)^2 + \tfrac{\log[\nu^{-1}(f)]+\log(\epsilon^{-1})}{\lambda(f)}.$$

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 $A \geq 0$, we have[16] $\inf_{x>0}\big\{xg(x) + \frac{A^2}{2x}\big\} \leq A + \frac{A^2}{6}$ and $\inf_{x>0}\big\{\frac{x}{2} + \frac{A^2}{2x}\big\} \leq A$. The last inequality is used when $B \leq 0$ since $g(u) \leq \frac{1}{2}$ for $u \leq 0$.

---

[16]Proof: we have $\inf_{x>0}\big\{xg(x) + \frac{A^2}{2x}\big\} \leq \log(1+A)g[\log(1+A)] + \frac{A^2}{2\log(1+A)} = A + \frac{A^2}{6} - \frac{1+A+\frac{A^2}{6}}{\log(1+A)}k(A)$ where $k(A) \triangleq \log(1+A) - \frac{A+\frac{A^2}{2}}{1+A+\frac{A^2}{6}}$. Since $k(0) = 0$ and $k'(A) = \frac{A^4}{36(1+A)(1+A+A^2/6)^2} \geq 0$, we get $k(A) \geq 0$, 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 $\log \mathbb{P}\exp\big\{\eta(W-\mathbb{P}W)\big\} \leq \eta^2 \mathbb{V}\mathrm{ar}_{\mathbb{P}} W g(\eta b')$.

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

$$\mathbb{P}^{\otimes N}\Big(\tfrac{\sum_{i=1}^N (W_i - \mathbb{P}W)}{N} > \inf_{x>0}\Big\{ u g(uB)\mathbb{P}W^2 + \tfrac{\log(\epsilon^{-1})}{Nu}\Big\}\Big) \leq \epsilon,$$

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*

$$\log \nu \exp\big\{ -\lambda \bar{\mathbb{P}}\mathcal{W}\big\} \geq \log \nu \exp\big\{ -\lambda[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2]\big\} - \frac{\lambda}{\lambda'}\log(\epsilon^{-1}), \tag{8.4}$$

- *for any $\lambda' \geq \lambda$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have*

$$\log \nu \exp\big\{ -\lambda \bar{\mathbb{P}}\mathcal{W}\big\} \leq \log \nu \exp\big\{ -\lambda[\mathbb{P}\mathcal{W} - a_A(\lambda')\mathbb{P}\mathcal{W}^2]\big\} + \frac{\lambda}{\lambda'}\log(\epsilon^{-1}). \tag{8.5}$$

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 $1-\epsilon$,

$$-\mu'[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2] \leq -\mu'\bar{\mathbb{P}}\mathcal{W} + \frac{\log(\epsilon^{-1})}{\lambda'}.$$

So we have

$$\begin{aligned}
&\log \nu \exp\big\{ -\lambda[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2]\big\} \\
&\qquad = -\lambda\mu'[\mathbb{P}\mathcal{W} + a_B(\lambda')\mathbb{P}\mathcal{W}^2] - K(\mu', \nu) \\
&\qquad \leq -\lambda\mu'\bar{\mathbb{P}}\mathcal{W} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}) - K(\mu', \nu) \\
&\qquad \leq \sup_{\mu \in \mathcal{M}_+^1(\mathcal{G})} \Big\{ -\lambda\mu\bar{\mathbb{P}}\mathcal{W} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}) - K(\mu, \nu)\Big\} \\
&\qquad = \log \nu \exp\big\{ -\lambda \bar{\mathbb{P}}\mathcal{W}\big\} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}).
\end{aligned}$$

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 $\nu\colon\ \mathcal{Z}^{2N} \to \mathcal{M}_+^1(\mathcal{G})$ 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*

(8.8)
$$\big(\mathbb{P}^{\otimes 2N}\big)^*\left\{\sup_{\mu\in\mathcal{M}_+^1(\mathcal{G})}\left[\mu\bar{\mathbb{P}}'\mathcal{W} - \mu\bar{\mathbb{P}}\mathcal{W} - \frac{\lambda}{2N}\mu\mathbb{W} - \frac{K(\mu,\nu_{Z_1^{2N}}) + \log(\epsilon^{-1})}{\lambda}\right] \,\middle|\, Z_1^N\right\} \le 0.$$

- *with* $\big(\mathbb{P}^{\otimes N}\big)_*$*-probability at least* $1-\epsilon$, *we have*

(8.9)
$$\big(\mathbb{P}^{\otimes 2N}\big)^*\left\{\sup_{f\in\mathcal{F}}\left[\bar{\mathbb{P}}'\mathcal{W}(f,\cdot) - \bar{\mathbb{P}}\mathcal{W}(f,\cdot) - \sqrt{\frac{2\mathbb{W}(f)\big[\log \nu_{Z_1^{2N}}^{-1}(f) + \log(\epsilon^{-1})\big]}{N}}\right] \,\middle|\, Z_1^N\right\} \le 0.$$

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

$$\begin{aligned}
& \left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N}) + \lambda(\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W})\right]\Big\} \\
= & \left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N}) + \tfrac{\lambda}{N}\textstyle\sum_{i=1}^N S_i\right]\Big\} \\
= & \left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N})\right]\Pi_{i=1}^N \cosh\left(\tfrac{\lambda}{N}S_i\right)\Big\} \\
\le & \left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N}) + \tfrac{\lambda^2}{2N^2}\textstyle\sum_{i=1}^N S_i^2\right]\Big\},
\end{aligned}$$

where, at the last step, we use $\cosh x \le \exp\left\{\frac{x^2}{2}\right\}$. Taking the exchangeable function $\eta(Z_1^{2N}) \triangleq -\frac{\lambda^2}{2N}\mathbb{W} - \log(\epsilon^{-1})$, we obtain

$$\left(\mathbb{P}^{\otimes 2N}\right)^* \nu_{Z_1^{2N}}\Big\{ \exp\left[\eta(Z_1^{2N}) + \lambda(\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W})\right]\Big\} \le \epsilon,$$

hence $\left(\mathbb{P}^{\otimes 2N}\right)^*\Big(\log \nu_{Z_1^{2N}}\big\{ \exp\left[\eta(Z_1^{2N}) + \lambda(\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W})\right]\big\} \ge 0\Big) \le \epsilon$, 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 $\left(\mathbb{P}^{\otimes 2N}\right)^*(U \ge 0) \le \epsilon$. 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 $\left(\mathbb{P}^{\otimes 2N}\right)^* \exp(U) \le \epsilon$. By Jensen's inequality, we obtain[18] $\left(\mathbb{P}^{\otimes N}\right)^* \exp\left\{(\mathbb{P}^{\otimes 2N})^*(U|Z_1^N)\right\} \le \epsilon$, hence $\left(\mathbb{P}^{\otimes N}\right)^*\Big\{\left(\mathbb{P}^{\otimes 2N}\right)^*\left(U|Z_1^N\right) \ge 0\Big\} \le \epsilon$, 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)

$$\mu\mathbb{P}\mathcal{W} - \mu\bar{\mathbb{P}}\mathcal{W} \quad \le \quad \tfrac{\lambda}{2N}\mu\mathbb{P}^{\otimes 2N}[\mathbb{W}|Z_1^N] + \frac{\left(\mathbb{P}^{\otimes 2N}\right)^*\left[K\left(\mu, \nu_{Z_1^{2N}}\right)|Z_1^N\right] + \log(\epsilon^{-1})}{\lambda}.$$

with $\mathbb{P}^{\otimes 2N}[\mathbb{W}|Z_1^N] = \mathbb{P}\mathcal{W}^2 + \bar{\mathbb{P}}\mathcal{W}^2 - 2\mathbb{P}\mathcal{W}\bar{\mathbb{P}}\mathcal{W} \le 2(\mathbb{P}\mathcal{W}^2 + \bar{\mathbb{P}}\mathcal{W}^2)$. (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 $\big(\mathbb{P}^{\otimes 2N}\big)^*\big\{\mu\bar{\mathbb{P}}'\mathcal{W}|Z_1^N\big\} = \mu\mathbb{P}\mathcal{W}$ and $\big(\mathbb{P}^{\otimes 2N}\big)^*\big\{\mu\bar{\mathbb{P}}\mathcal{W}|Z_1^N\big\} = \mu\bar{\mathbb{P}}\mathcal{W}$. □

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 $\mu' \triangleq \nu_{-2\lambda[\bar{\mathbb{P}}\mathcal{W} + \frac{\lambda'}{N}\bar{\mathbb{P}}\mathcal{W}^2]}$. 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

$$\begin{aligned}
&\log\nu\exp\big\{-2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W} + \tfrac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]\big\}\\
&\quad= -2\lambda\mu'[\bar{\bar{\mathbb{P}}}\mathcal{W} + \tfrac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2] - K(\mu',\nu)\\
&\quad\leq -2\lambda\mu'\bar{\mathbb{P}}\mathcal{W} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}) - K(\mu',\nu)\\
&\quad\leq \sup_{\mu\in\mathcal{M}_+^1(\mathcal{G})}\Big\{-2\lambda\mu\bar{\mathbb{P}}\mathcal{W} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}) - K(\mu,\nu)\Big\}\\
&\quad= \log\nu\exp\big\{-2\lambda\bar{\mathbb{P}}\mathcal{W}\big\} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1}).
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}^{\otimes 2N}\big[\log\nu\exp\big\{-2\lambda\bar{\mathbb{P}}\mathcal{W}\big\} > \log\nu\exp\big\{-2\lambda[\bar{\bar{\mathbb{P}}}\mathcal{W} - \tfrac{\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]\big\} + \tfrac{\lambda}{\lambda'}\log(\epsilon^{-1})\big]\\
&\quad= \mathbb{P}^{\otimes 2N}\big(\nu'\exp\big\{\lambda[\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - \tfrac{2\lambda'}{N}\mathbb{P}\mathcal{W}^2]\big\} > \epsilon^{-\frac{\lambda}{\lambda'}}\big)\\
&\quad= \mathbb{P}^{\otimes 2N}\Big(\epsilon\big[\nu'\exp\big\{\lambda[\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - \tfrac{2\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2]\big\}\big]^{\frac{\lambda'}{\lambda}} > 1\Big)\\
&\quad\leq \epsilon\mathbb{P}^{\otimes 2N}\Big(\big[\nu'\exp\big\{\lambda[\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - \tfrac{2\lambda'}{N}\mathbb{P}\mathcal{W}^2]\big\}\big]^{\frac{\lambda'}{\lambda}}\Big)\\
&\quad\leq \epsilon,
\end{aligned}$$

where at the last step we use Jensen's inequality and

$$\mathbb{P}^{\otimes 2N}\nu'\exp\Big\{\lambda'\Big[\bar{\mathbb{P}}'\mathcal{W} - \bar{\mathbb{P}}\mathcal{W} - \frac{2\lambda'}{N}\bar{\bar{\mathbb{P}}}\mathcal{W}^2\Big]\Big\} \leq 1.$$

□

8.3.3. *Comparison between Theorem 8.4 and Theorem 8.1.* For comparison purposes, Theorem 8.1 leads to

$$\mu\mathbb{P}\mathcal{W} - \mu\bar{\mathbb{P}}\mathcal{W} \leq \frac{\lambda}{N} g\left(\frac{\lambda}{N}\sup_{\mathcal{G}\times\mathcal{Z}}(-\mathcal{W})\right)\mu\mathbb{P}\mathcal{W}^2 + \frac{K(\mu,\nu)+\log(\epsilon^{-1})}{\lambda}.$$

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 $\mathcal{I} \triangleq \underset{2\leq h\leq N-1}{\cup}\mathcal{I}_h$ 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 $\mu \in \mathcal{M}_+^1(\mathcal{I}\times\mathcal{I}\times\mathcal{G})$,*

$$\mu\bar{\mathbb{P}}^{I_{1,2}}\mathcal{W} - \mu\mathbb{P}\mathcal{W} \leq \mu\big[a_{B,|I_{1,2}|}(\lambda)\mathbb{P}\mathcal{W}^2\big] + \frac{K(\mu,\nu)+\log(\epsilon^{-1})}{\lambda}, \tag{8.11}$$

- *with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-\epsilon$, for any $I_1, I_2 \in \mathcal{I}$ and $f \in \mathcal{G}$,*

$$\begin{aligned}
&\bar{\mathbb{P}}^{I_{1,2}}\hat{G}(Z_{I_1},Z_{I_2},f,\cdot) - \mathbb{P}\hat{G}(Z_{I_1},Z_{I_2},f,\cdot)\\
&\quad\leq \min_{x>0}\Big\{a_{B,|I_{1,2}|}(x)\mathbb{P}\hat{G}^2(Z_{I_1},Z_{I_2},f,\cdot) + \tfrac{\log\nu^{-1}(I_1,I_2,f)+\log(\epsilon^{-1})}{x}\Big\}\\
&\quad\leq \sqrt{\tfrac{2[\log\nu^{-1}(I_1,I_2,f)+\log(\epsilon^{-1})]\mathbb{P}\hat{G}^2(Z_{I_1},Z_{I_2},f,\cdot)}{|I_{1,2}|}} + B\tfrac{\log\nu^{-1}(I_1,I_2,f)+\log(\epsilon^{-1})}{3|I_{1,2}|}.
\end{aligned} \tag{8.12}$$

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

$$\mathbb{P}^{I_{1,2}}\exp\left\{\lambda[\bar{\mathbb{P}}^{I_{1,2}}\mathcal{W} - \mathbb{P}\mathcal{W} - a_{B,|I_{1,2}|}(\lambda)\mathbb{P}\mathcal{W}^2]\right\} \le 1.$$

• 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

$$\left\{\text{for any } f_1, f_2 \in \hat{\mathcal{F}},\ r'(f_2) - r'(f_1) \le r(f_2) - r(f_1) + S(f_1, f_2)\right\}.$$

From Theorem 5.1, with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-\epsilon$, this event holds.

• Since we have $S(f_{k-1}, f_k) \ge 0$ 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 $0 \le K \le N$ such that f_K exists but not f_{K+1}.

We have

$$r'(f_k) - r'(f_{k-1}) \le r(f_k) - r(f_{k-1}) + S(f_{k-1}, f_k).$$

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 $\mathcal{C}(f_1) \ge \mathcal{C}(f_0)$. When $\mathcal{C}(f_{k-1}) \ge \mathcal{C}(f_{k-2})$, we will prove that $\mathcal{C}(f_k) \ge \mathcal{C}(f_{k-1})$ 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

$$r(f_k) - r(f_{k-2}) + S(f_k, f_{k-2}) \ge 0$$

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 $\bar{\bar{\mathbb{P}}}_{f_k,f_{k-2}} \le \bar{\bar{\mathbb{P}}}_{f_k,f_{k-1}} + \bar{\bar{\mathbb{P}}}_{f_{k-1},f_{k-2}}$, for any $a, b > 0$ $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, $\mathcal{C}(f_{k-2}) \le \mathcal{C}(f_{k-1})$ and $\bar{\bar{\mathbb{P}}}_{f_{k-2},f_{k-1}} \ne 0$, 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 $r'(f_K) \le r'(f_{k(f)}) \le r'(f)+r(f_{k(f)})-r(f)+S(f_{k(f)}, f)$. Since by definition of $k(f)$ we have $r(f) - r(f_{k(f)}) + S(f_{k(f)}, f) \ge 0$, we obtain $r'(f_K) \le r'(f) + 2S(f_{k(f)}, f)$.

• We have just seen that for any $f \in \hat{\mathcal{F}}$, $r'(f_{k(f)}) \le r'(f) + 2S(f_{k(f)}, f)$, hence

$$r'(f_{k(f)}) \le 2r'(f) - r'(f_{k(f)}) + 8\sqrt{\frac{2\bar{\bar{\mathbb{P}}}_{f,f_{k(f)}}[C(f)+C(f_{k(f)})+L]}{N}}.$$

Therefore we have

$$r'(f_K) \le \sup_{g\in\hat{\mathcal{F}}:h(g)\le h(f)}\left\{2r'(f) - r'(g) + 8\sqrt{\frac{2\bar{\bar{\mathbb{P}}}_{f,g}[C(f)+C(g)+L]}{N}}\right\}$$

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 $\left(\mathbb{P}^{\otimes 2N}\right)_*$-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

$$\left\{\begin{array}{l}\rho_k r+S(\rho_{k-1},\rho_k)\le \rho_{k-1}r\\ \rho_{k-1}r+S(\rho_{k-2},\rho_{k-1})\le \rho_k r+S(\rho_{k-2},\rho_k),\end{array}\right.$$

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 $\min_{\lambda\in[\sqrt{N};N]}\left\{\tilde\rho r-\rho_{\tilde k}r+S_\lambda(\tilde\rho,\rho_{\tilde k})\right\}>0$, hence for any $\lambda\in[\sqrt{N};N]$ and $\eta>0$,

$$\begin{array}{rcl}\rho_{\tilde k}r'-\tilde\rho r' & \le & 2S_\lambda(\tilde\rho,\rho_{\tilde k})\\ & \le & \frac{4\lambda}{N}(\tilde\rho\otimes\rho_{\tilde k})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}+\frac{2\sqrt{e}}{\lambda}\big[(2+\eta)K(\tilde\rho,\pi)+L\big]-\frac{2\sqrt{e}\eta}{\lambda}K(\rho_{\tilde k},\pi),\end{array}$$

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

$$\begin{array}{rcl}(1-\xi)\rho_{\tilde k}r' & \le & -\xi\rho_{\tilde k}r'+\frac{4\lambda}{N}(\tilde\rho\otimes\rho_{\tilde k})\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}-\frac{2\sqrt{e}\eta}{\lambda}K(\rho_{\tilde k},\pi)\\ & & \qquad+\tilde\rho r'+\frac{1}{\tilde\lambda}K(\tilde\rho,\pi)+\frac{2\sqrt{e}}{\lambda}L\\ & \le & \frac{\eta}{(\eta+2)\tilde\lambda}\log\pi\exp\Big\{-\frac{\tilde\lambda(\eta+2)}{\eta}\xi r'+\frac{8\tilde\lambda^2(\eta+2)^2\sqrt{e}}{\eta N}\tilde\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big\}\\ & & \qquad-\frac{1}{\tilde\lambda}\log\pi\exp\big(-\tilde\lambda r'\big)+\frac{L}{(\eta+2)\tilde\lambda}.\end{array}$$

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 $\mathbb{C}\mathrm{ov}\big(\exp(aX), \exp(-X)\big) \leq 0\big)$.

• 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

$$\begin{cases} q\le 2\\ \big(2+\frac{1}{q}\big)\frac{\lambda}{N}<\alpha\end{cases}\quad\text{or}\quad\begin{cases} q>2\\ \big(q+\frac{1}{2}\big)\frac{\lambda}{N}<\alpha\end{cases}.$$

So Inequality (9.2) asserts that for α large enough, with high probability, we have

$$\log\big(\pi_{-\lambda\frac{r+r'}{2}}\otimes\pi_{-\lambda\frac{r+r'}{2}}\big)\exp\big(\alpha\lambda\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big)\le\log\big(\pi_{-\lambda r}\otimes\pi_{-\lambda r}\big)\exp\big(C\alpha\lambda\bar{\mathbb{P}}_{\cdot,\cdot}\big)$$
$$+\text{ confidence level term.}$$

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

$$\left\{ \begin{array}{ll} \forall\, 0 \le j \le \log N, & K\big(\rho_j, \pi_{\lambda_j \frac{r+r'}{2}}\big) \le 2 \log \rho_j \exp\Big(\frac{\lambda_j^2}{N} \rho_j \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + L \\ \forall\, 0 \le i \ne j \le \log N, & \rho_i r' - \rho_j r' \le \rho_i r - \rho_j r + S(i \wedge j, i \vee j) \end{array} \right. ,$$

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

$$-\rho_{u(k)} r + \rho_j r = \tfrac{-K(\rho_{u(k)}, \rho_j) + K(\rho_{u(k)}, \pi) - K(\rho_j, \pi)}{\lambda_j} \le -\tfrac{K(\rho_{u(k)}, \rho_j)}{\lambda_j},$$

we obtain

$$\begin{array}{rcl} \rho_{u(k)} r' - \rho_j r' & \le & \frac{6\lambda_j}{N} \big(\rho_{u(k)} \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{6\lambda_j}{N} \big(\rho_j \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{12\tilde{\mathcal{C}}(j) + 9L}{\lambda_j} - \frac{K(\rho_{u(k)}, \rho_j)}{\lambda_j} \\ & \le & \sup_{\rho \in \mathcal{M}_+^1(\mathcal{F})} \Big\{ \frac{6\lambda_j}{N} \big(\rho \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} - \frac{K(\rho, \rho_j)}{\lambda_j} \Big\} + \frac{6\lambda_j}{N} \big(\rho_j \otimes \tilde{\pi}_j\big) \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \\ & & \qquad + \frac{12\tilde{\mathcal{C}}(j) + 9L}{\lambda_j}. \end{array}$$

By Jensen's inequality, we get

$$\rho_{u(k)} r' - \rho_j r' \le \tfrac{2}{\lambda_j} \log \rho_j \exp\Big(\tfrac{6\lambda_j^2}{N} \tilde{\pi}_j \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + \tfrac{12\tilde{\mathcal{C}}(j) + 9L}{\lambda_j}.$$

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

$$\mathcal{C}(i) \le 2 \log \rho_i \exp\Big(\tfrac{\lambda_i^2}{N} \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) \le \tfrac{1}{3} \log \rho_i \exp\Big(\tfrac{6\lambda_i^2}{N} \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big).$$

We obtain

$$\rho_{u(k)} r' - \rho_j r' \le \tfrac{6}{\lambda_j} \sup_{0 \le i \le j} \Big\{ \log \rho_i \exp\Big(\tfrac{6\lambda_i^2}{N} \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) \Big\} + \tfrac{9L}{\lambda_j}.$$

Let $\rho_i' \triangleq \pi_{-\lambda_i r'}$. It remains to prove that the quantity $\log \rho_i \exp\Big(\frac{6\lambda_i^2}{N} \tilde{\pi}_i \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)$ behaves like the quantity $\log \rho_i' \exp\Big(\frac{C\lambda_i^2}{N} \rho_i' \bar{\mathbb{P}}'_{\cdot,\cdot}\Big)$ 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

$$\log \pi_{-\lambda r} \exp\Big(\tfrac{6\lambda^2}{N} \pi_{-\lambda \frac{r+r'}{2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) \le \log \pi_{-\lambda \frac{r+r'}{2}} \exp\Big(\tfrac{7\lambda^2}{N} \pi_{-\lambda \frac{r+r'}{2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + 2L.$$

From Inequality (9.2) with $\big(\alpha,\lambda',\lambda'',p,q\big)=\Big(\frac{7\lambda}{N},\lambda,\frac{\lambda}{1+\left(\frac{7\lambda}{N}\right)^2},\frac{3}{2},3\Big)$, 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

$$\rho_{u(K)}r'\le\rho_{u(k)}r'\le\rho_jr'+6\tfrac{\mathcal{C}'(j)}{\lambda_j}+1551\tfrac{L}{\lambda_j}.$$

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.** $r(I_k,\theta_k)<r(I_{k-1},\theta_{k-1})$. 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

$$\begin{aligned}R(I_2,\theta_2)-R(I_1,\theta_1)+r(I_1,\theta_1)-r(I_2,\theta_2)&\\ \le\sqrt{2C_{1,2}\mathbb{P}(I_1,\theta_1,I_2,\theta_2)}+\tfrac{C_{1,2}}{3}&\\ \le\sqrt{2C_{1,2}}\Big(\sqrt{\mathbb{P}(I_1,\theta_1,I_2,\theta_2)+C_{1,2}/2}+\sqrt{C_{1,2}/2}\Big)+\tfrac{C_{1,2}}{3}&\\ \le S(I_1,\theta_1,I_2,\theta_2).&\end{aligned}$$

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 $i \ne j$ 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 $\mathcal{W}'(f,Z) \triangleq \mathbb{1}_{Y\ne f(X)} - \tilde\pi\mathbb{1}_{Y\ne\cdot(X)}$ and $\mathcal{W}''(f,Z) \triangleq \mathbb{1}_{Y\ne f(X)} - \tilde\pi\mathbb{1}_{Y\ne\cdot(X)} - \alpha\tilde\pi\mathbb{1}_{f(X)\ne\cdot(X)}$. 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

$$\begin{aligned}
\rho_{u(k)}R - \rho_j R &\le \rho_{u(k)}r - \rho_j r + S\big(u(k),j\big)\\
&\le 3S\big(u(k),j\big) - \rho_{u(k)}r + \rho_j r\\
&\le 3S\big(u(k),j\big) - \tfrac{K(\rho_{u(k)},\rho_j)}{\lambda_j}\\
&\le 3a_j\big(\rho_j\otimes\tilde\pi_j\big)\bar{\mathbb{P}}_{\cdot,\cdot} + 6b_j\mathcal{C}[u(k)] + 6b_j\mathcal{C}(j) + 9b_jL\\
&\qquad + 3a_j\big(\rho_{u(k)}\otimes\tilde\pi_j\big)\bar{\mathbb{P}}_{\cdot,\cdot} - \tfrac{K(\rho_{u(k)},\rho_j)}{\lambda_j}\\
&\le \tfrac{2}{\lambda_j}\log\rho_j\exp\big(3a_j\lambda_j\tilde\pi_j\bar{\mathbb{P}}_{\cdot,\cdot}\big) + 12b_j\sup_{0\le i\le j}\mathcal{C}(i) + 9b_jL
\end{aligned}$$

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

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

hence

$$\begin{aligned}
(1-\xi)[\rho r' - \breve{\rho} r'] &\le (1+\xi)[\rho r - \breve{\rho} r] + (1+\xi)\tfrac{2\lambda}{N}\big(\rho\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \\
&\qquad + \frac{K(\rho,\pi) + \log\pi\exp\big\{-\xi\lambda\big[r+r'+\frac{2\lambda}{N}\breve{\rho}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big]\big\} + \log(\epsilon^{-1})}{\lambda} \\
&\le (1-\xi)[\rho r - \breve{\rho} r] + (1+\xi)\tfrac{2\lambda}{N}\big(\rho\otimes\breve{\rho}\big)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \\
&\qquad + \frac{K(\rho,\pi_{-2\xi\lambda r}) + (1+\xi)\log(\epsilon^{-1})}{\lambda},
\end{aligned}$$

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

$$\begin{aligned}
\breve\rho r' - \rho r' &\le \breve\rho r - \rho r + \tfrac{2\lambda}{N}(\rho\otimes\breve\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{K(\rho,\tilde\rho)+\log(\epsilon^{-1})}{\lambda},\\
\breve\rho r' - \rho r' &\le \breve\rho r - \rho r + \xi\rho(r+r') + (1+\xi)\tfrac{2\lambda}{N}(\rho\otimes\breve\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\\
&\qquad + \tfrac{K(\rho,\pi)+\log\pi\exp\left(-\xi\lambda[r+r'+\frac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]\right)+\log(\epsilon^{-1})}{\lambda}\\
&\le \breve\rho r - \rho r + \xi\big(\rho r + \rho r' + \breve\rho r - \breve\rho r'\big) + (1+\xi)\tfrac{2\lambda}{N}(\rho\otimes\breve\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\\
&\qquad + \tfrac{K(\rho,\pi)+\log\pi\exp(-2\xi\lambda r)+(1+\xi)\log(\epsilon^{-1})}{\lambda},\\
(1+\xi)(\breve\rho r' - \rho r') &\le (1+\xi)(\breve\rho r - \rho r) + (1+\xi)\tfrac{2\lambda}{N}(\rho\otimes\breve\rho)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\\
&\qquad + \tfrac{K(\rho,\pi_{-2\xi\lambda r})+(1+\xi)\log(\epsilon^{-1})}{\lambda}.
\end{aligned}$$

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 $0<\lambda'\le 0.77N$ 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 $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-2\epsilon$, we have $\rho R - \tilde\rho R \le \rho r - \tilde\rho r + \bar a(\lambda')(\rho\otimes\tilde\rho)\bar{\mathbb{P}}_{\cdot,\cdot} + \bar b(\lambda')K(\rho,\tilde\rho) + \bar b(\lambda')\log(\epsilon^{-1})$. We get that

$$\begin{aligned}
(1-\xi)K(\rho,\tilde\rho) &\le K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1}) - \lambda\tilde\rho r + \lambda\bar a(\lambda')(\rho\otimes\tilde\rho)\bar{\mathbb{P}}_{\cdot,\cdot} - K(\tilde\rho,\pi)\\
&\le K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1})\\
&\qquad + \sup_{\rho'\in\mathcal{M}_+^1(\mathcal{F})}\big\{-\lambda\rho' r + \lambda\bar a(\lambda')(\rho'\otimes\rho)\bar{\mathbb{P}}_{\cdot,\cdot} - K(\rho',\pi)\big\}\\
&= K(\rho,\pi) + \lambda\rho r + \xi\log(\epsilon^{-1}) + \log\pi\exp\big\{-\lambda\big[r - \bar a(\lambda')\rho\bar{\mathbb{P}}_{\cdot,\cdot}\big]\big\}\\
&= K(\rho,\pi_{-\lambda r}) + \log\pi_{-\lambda r}\exp\big\{\lambda\bar a(\lambda')\rho\bar{\mathbb{P}}_{\cdot,\cdot}\big\} + \xi\log(\epsilon^{-1}).
\end{aligned}$$

[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 $0<\lambda'\le 0.77N$, we have $\bar a(\lambda') \le \frac{\lambda'}{N} \le \frac{2}{N\bar b(\lambda')} \le \frac{2\lambda}{\xi N}$.

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 $\nu = \pi_{2\lambda[\bar{\mathbb{P}}\mathcal{W} - \frac{\lambda}{N}\bar{\mathbb{P}}\mathcal{W}^2]}$. We obtain that with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$,

$$\begin{aligned}
&\log\pi_{\lambda[(r+r')-\breve\rho(r+r')-\frac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}]}\exp\big\{-2\lambda[r-\breve\rho r]\big\}\\
&\qquad\le -\log\pi\exp\Big\{\lambda\big[(r+r') - \breve\rho(r+r') - \tfrac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big]\Big\} + \log(\epsilon^{-1}),
\end{aligned}$$

hence

$$\begin{aligned}
&\log\pi_{-2\lambda[r-\breve\rho r]}\exp\big\{\lambda\big[(r+r') - \breve\rho(r+r') - \tfrac{2\lambda}{N}\breve\rho\mathbb{P}_{\cdot,\cdot}\big]\big\}\\
&\qquad\le -\log\pi\exp\big\{-2\lambda\big[r-\breve\rho r\big]\big\} + \log(\epsilon^{-1}).
\end{aligned}\tag{9.8}$$

By Markov's inequality, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, we have

$$\begin{aligned}
&\pi_{-2\lambda r}\left((r+r') - \breve\rho(r+r') > \tfrac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{-\log\pi\exp\{-2\lambda[r-\breve\rho r]\}+2\log(\epsilon^{-1})}{\lambda}\right)\\
&\le \pi_{-2\lambda r}\exp\left\{\lambda\left[(r+r') - \breve\rho(r+r') - \tfrac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \tfrac{\log\pi\exp\{-2\lambda[r-\breve\rho r]\}+2\log\epsilon}{\lambda}\right]\right\}\\
&= \epsilon^2\pi\exp\big\{-2\lambda[r-\breve\rho r]\big\}\pi_{-2\lambda[r-\breve\rho r]}\exp\Big\{\lambda\big[(r+r') - \breve\rho(r+r') - \tfrac{2\lambda}{N}\breve\rho\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big]\Big\}\\
&\le \epsilon,
\end{aligned}$$

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 $1-\epsilon$,

$$\log \pi_{\lambda\left[R-\tilde{\rho}R-\frac{\lambda}{N}g\left(\frac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot}\right]} \exp\big\{-\lambda[r-\tilde{\rho}r]\big\} \\ \leq -\log \pi \exp\Big\{\lambda\Big[R-\tilde{\rho}R-\tfrac{\lambda}{N}g\left(\tfrac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot}\Big]\Big\} + \log(\epsilon^{-1}),$$

hence

$$\log \pi_{-\lambda[r-\tilde{\rho}r]} \exp\Big\{\lambda\Big[R-\tilde{\rho}R-\tfrac{\lambda}{N}g\left(\tfrac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot}\Big]\Big\} \\ \leq -\log \pi \exp\big\{-\lambda\big[r-\tilde{\rho}r\big]\big\} + \log(\epsilon^{-1}). \tag{9.9}$$

By Markov's inequality, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$\begin{aligned} & \pi_{-\frac{r}{b(\lambda)}}\left(R-\tilde{\rho}R > \tfrac{\lambda}{N}g\left(\tfrac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot} + \tfrac{-\log\pi\exp\{-\lambda[r-\tilde{\rho}r]\}+2\log(\epsilon^{-1})}{\lambda}\right) \\ \leq\ & \pi_{-\lambda r}\exp\Big\{\lambda\Big[R-\tilde{\rho}R-\tfrac{\lambda}{N}g\left(\tfrac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot} + \tfrac{\log\pi\exp\{-\lambda[r-\tilde{\rho}r]\}+2\log\epsilon}{\lambda}\Big]\Big\} \\ =\ & \epsilon^2\pi\exp\big\{-\lambda\big[r-\tilde{\rho}r\big]\big\}\pi_{-\lambda[r-\tilde{\rho}r]}\exp\Big\{\lambda\Big[R-\tilde{\rho}R-\tfrac{\lambda}{N}g\left(\tfrac{\lambda}{N}\right)\tilde{\rho}\mathbb{P}_{\cdot,\cdot}\Big]\Big\} \\ \leq\ & \epsilon, \end{aligned}$$

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

$$K(\pi_{-\lambda r}, \pi_{-\lambda r'}) = \log\pi\exp(-\lambda r) + \log\pi\exp(-\lambda r') - 2\log\pi\exp\big(-\lambda\tfrac{r+r'}{2}\big) \\ + 2K(\pi_{-\lambda r}, \pi_{-\lambda\frac{r+r'}{2}}).$$

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 $1-2\epsilon$,*

$$\log\pi\exp(-\lambda r) + \log\pi\exp(-\lambda r') - 2\log\pi\exp\big(-\lambda\tfrac{r+r'}{2}\big) \\ \leq 2\log\tilde{\pi}\exp\Big(\tfrac{\lambda^2}{2\gamma N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + 2\gamma\log(\epsilon^{-1}), \tag{9.10}$$

- *with $\left(\mathbb{P}^{\otimes 2N}\right)_*$-probability at least $1-2\epsilon$,*

$$K(\pi_{-\lambda r}, \tilde{\pi}) \leq \tfrac{2}{1-\gamma}\log\tilde{\pi}\exp\Big(\tfrac{\lambda^2}{\gamma N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big) + 9\tfrac{\gamma}{1-\gamma}\log(\epsilon^{-1}) \tag{9.11}$$

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

$$\begin{cases} \log\pi\exp\big\{-\lambda(r-\tilde{\pi}r)\big\} \leq \log\pi\exp\Big\{-\lambda\Big(\frac{r+r'}{2} - \tilde{\pi}\frac{r+r'}{2} - \frac{\lambda'}{2N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)\Big\} \\ \qquad\qquad + \frac{\lambda}{\lambda'}\log(\epsilon^{-1}) \\ \log\pi\exp\big\{-\lambda(r'-\tilde{\pi}r')\big\} \leq \log\pi\exp\Big\{-\lambda\Big(\frac{r+r'}{2} - \tilde{\pi}\frac{r+r'}{2} - \frac{\lambda'}{2N}\tilde{\pi}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\Big)\Big\} \\ \qquad\qquad + \frac{\lambda}{\lambda'}\log(\epsilon^{-1}) \end{cases},$$

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 $\log \bar{\rho} \exp\big\{ \frac{\lambda^2}{2\xi N} \bar{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\} \leq 2 \log \bar{\rho} \exp\big\{ \frac{\lambda^2}{2\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \big\}$. Introducing

$$\left\{ \begin{array}{rcl} \mathcal{L}' & \triangleq & \log \pi \exp\Big\{ -\lambda[r - \tilde{\rho} r] + \frac{\lambda^2}{2\xi N} \tilde{\rho} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} \\ \mathcal{L}'' & \triangleq & \log \pi \exp\big\{ -\lambda[r - \tilde{\rho} r] \big\} \end{array} \right. ,$$

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

$$\begin{array}{rcl} K\big(\pi_{-\lambda r}, \pi_{-\frac{\lambda}{2}[r+r']}\big) & \leq & \frac{1}{1-\xi} \log \pi_{-\lambda r} \exp\Big\{ \frac{\lambda^2}{2\xi N} \pi_{-\lambda r} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} + \frac{\xi}{1-\xi} \log(\epsilon^{-1}) \\ & \leq & \frac{2}{1-\xi} \log \pi_{-\lambda r} \exp\Big\{ \frac{\lambda^2}{2\xi N} \pi_{-\lambda \frac{r+r'}{2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} + \frac{\xi}{1-\xi} \log(\epsilon^{-1}) \\ & \leq & \frac{2}{1-\xi} \log \pi_{-\lambda \frac{r+r'}{2}} \exp\Big\{ \frac{\lambda^2}{\xi N} \pi_{-\lambda \frac{r+r'}{2}} \bar{\bar{\mathbb{P}}}_{\cdot,\cdot} \Big\} + 9 \frac{\xi}{1-\xi} \log(\epsilon^{-1}). \end{array}$$

□

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

$$\big(\alpha,\lambda',\lambda'',p,q\big)=\Big(\tfrac{5\lambda}{\gamma N},\tfrac{\lambda}{\gamma},\tfrac{\lambda}{9\gamma\big(1+\frac{25\lambda^2}{\gamma^2N^2}\big)},\tfrac{4}{3},4\Big),$$

with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-4\epsilon$, we have

$$\begin{aligned}\log\tilde\pi\otimes\tilde\pi\exp\big(\tfrac{5\lambda^2}{\gamma N}\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\big) \le \tfrac{3}{2}\log\pi_{-\lambda r'}\otimes\pi_{-\lambda r'}\exp\big(\tfrac{20\lambda^2}{3\gamma N}\bar{\mathbb{P}}'_{\cdot,\cdot}\big)\\ +15\gamma\big(1+\tfrac{25\lambda^2}{\gamma^2N^2}\big)\log(\epsilon^{-1})\end{aligned}$$

To conclude, with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-8\epsilon$, we have

$$K(\pi_{-\lambda r},\pi_{-\lambda r'})\le\tfrac{1}{1-\gamma}\log\pi_{-\lambda r'}\otimes\pi_{-\lambda r'}\exp\big(\tfrac{10\lambda^2}{\gamma N}\bar{\mathbb{P}}'_{\cdot,\cdot}\big)+\big(35+\tfrac{375\lambda^2}{\gamma^2N^2}\big)\tfrac{\gamma}{1-\gamma}\log(\epsilon^{-1}).$$

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*

$$\log\pi_{-\lambda r}\exp\Big(\frac{2\lambda^2}{\xi N}\pi_{-\lambda r}\bar{\mathbb{P}}_{\cdot,\cdot}\Big)\le 4\log\pi_{-\lambda R}\exp\Big(\frac{4.1\lambda^2}{\xi N}\pi_{-\lambda R}\mathbb{P}_{\cdot,\cdot}\Big)+4\xi\log(\epsilon^{-1}).$$

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

$$\left\{\begin{array}{lcl}\mathcal{L}' &\triangleq& \log\pi\exp\big(-\lambda\big[r-\tilde r-\alpha\bar{\mathbb{P}}_{\cdot,\sim}\big]\big)\\ \mathcal{L}'' &\triangleq& \log\pi\exp\big\{-\lambda(r-\tilde r)\big\}\end{array}\right.\quad.$$

These two quantities can be bounded using Theorem 8.3 for

$$\left\{\begin{array}{lcl}\mathcal{W}'(f,Z) &\triangleq& \mathbb{1}_{Y\neq f(X)}-\tilde\rho_{(df')}\mathbb{1}_{Y\neq f'(X)}-\tilde\rho_{(df')}\alpha\mathbb{1}_{f(X)\neq f'(X)}\in[-(1+\alpha);1]\\ \mathcal{W}''(f,Z) &\triangleq& \mathbb{1}_{Y\neq f(X)}-\tilde\rho_{(df')}\mathbb{1}_{Y\neq f'(X)}\in[-1;1]\end{array}\right.\quad.$$

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

$$\mathcal{L}'\le\log\pi\exp\Big\{-\lambda\big[R-\tilde R\big]+\lambda\big[\alpha+(1+\alpha)^2a_{1+\alpha}(\lambda'')\big]\mathbb{P}_{\cdot,\sim}\Big\}+\tfrac{\lambda}{\lambda'''}\log(\epsilon^{-1})$$

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

$$\begin{aligned}
\log \bar{\rho}\exp\big(\alpha\lambda\bar{\mathbb{P}}_{\cdot,\sim}\big) &\leq \log\tilde{\rho}\exp\big\{\lambda\big[\alpha+(1+\alpha)^2 a_{1+\alpha}(\lambda/\xi)\big]\mathbb{P}_{\cdot,\sim}\big\} \\
&\qquad - \log\tilde{\rho}\exp\big\{-\lambda a_1(\lambda/\xi)\mathbb{P}_{\cdot,\sim}\big\} + 2\xi\log(\epsilon^{-1}) \\
&\leq \log\tilde{\rho}\exp\big\{\lambda\big[\alpha+(1+\alpha)^2 a_{1+\alpha}(\lambda/\xi)\big]\mathbb{P}_{\cdot,\sim}\big\} \\
&\qquad + \log\tilde{\rho}\exp\big\{\lambda a_1(\lambda/\xi)\mathbb{P}_{\cdot,\sim}\big\} + 2\xi\log(\epsilon^{-1}) \\
&\leq 2\log\tilde{\rho}\exp\big\{\lambda\big[\alpha+(1+\alpha)^2 a_{1+\alpha}(\lambda/\xi)\big]\mathbb{P}_{\cdot,\sim}\big\} + 2\xi\log(\epsilon^{-1}),
\end{aligned}$$

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

$$\rho_1\odot\rho_2 \triangleq (1-m_{1,2})\left(\frac{\rho_1\wedge\rho_2}{1-m_{1,2}}\right)_\Delta + m_{1,2}\left(\frac{(\rho_1-\rho_2)_+}{m_{1,2}}\right)\otimes\left(\frac{(\rho_2-\rho_1)_+}{m_{1,2}}\right).$$

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

$$\dot{\mathcal{K}} \triangleq m_{1,2}K\big(\tfrac{(\rho_1-\rho_2)_+}{m_{1,2}},\pi\big)+m_{1,2}K\big(\tfrac{(\rho_2-\rho_1)_+}{m_{1,2}},\pi\big) +(1-m_{1,2})K\big(\tfrac{\rho_1\wedge\rho_2}{1-m_{1,2}},\pi\big)+K(m_{1,2},\tfrac{1}{2})+\log(\epsilon^{-1}).$$

Proof. Take $\pi_{1,2} \triangleq \frac{1}{2}\pi\otimes\pi+\frac{1}{2}\pi_\Delta$ in the previous theorem. The result follows from (A.1)

$$\begin{aligned} K(\rho_1\odot\rho_2,\pi_{1,2}) &= \rho_1\odot\rho_2\log\tfrac{\rho_1\odot\rho_2}{\pi_{1,2}} \\ &\le \big(\tfrac{\rho_1}{\pi}\wedge\tfrac{\rho_2}{\pi}\big)\cdot\pi\log\big(2\tfrac{\rho_1}{\pi}\wedge\tfrac{\rho_2}{\pi}\big)+m_{1,2}\big(\tfrac{(\rho_1-\rho_2)_+}{m_{1,2}}\big)\otimes\big(\tfrac{(\rho_2-\rho_1)_+}{m_{1,2}}\big) \\ &\qquad \log\left(2\frac{\big(\frac{\rho_1}{\pi}(f_1)-\frac{\rho_2}{\pi}(f_1)\big)_+}{m_{1,2}}\frac{\big(\frac{\rho_2}{\pi}(f_2)-\frac{\rho_1}{\pi}(f_2)\big)_+}{m_{1,2}}\right) \\ &= K(m_{1,2},\tfrac{1}{2})+(1-m_{1,2})K\big(\tfrac{\rho_1\wedge\rho_2}{1-m_{1,2}},\pi\big) \\ &\qquad +m_{1,2}K\big(\tfrac{(\rho_1-\rho_2)_+}{m_{1,2}},\pi\big)+m_{1,2}K\big(\tfrac{(\rho_2-\rho_1)_+}{m_{1,2}},\pi\big). \end{aligned}$$

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 $\dot{\mathcal{K}}=K(\rho,\pi)+\log 2=\frac{1}{2}K(\rho_1\otimes\rho_2,\pi\otimes\pi)+\log 2$. 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 $\lambda\le cN$ for a small enough universal constant $c>0$ and any prior distribution π independent from the data, we have

$$\begin{aligned} -\log\pi\exp\big\{-\lambda[r'-r'(\tilde{f})]\big\} &\le -\log\pi\exp\big(-\lambda[R-R(\tilde{f})]-C\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\big)+\dots \\ &\le -\log\pi\exp\big\{-\lambda[R-R(\tilde{f})]\big\} \\ &\qquad +\log\pi_{-\lambda R}\exp\big(C\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\big)+\dots \end{aligned}$$

[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

$$\mathbb{G}_C^{\mathrm{th}}(\lambda) \triangleq -\tfrac{1}{\lambda}\log \pi \exp\big\{-\lambda[R - R(\tilde{f})]\big\} + \tfrac{1}{\lambda}\log \pi_{-\lambda R}\exp\Big(C\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\Big) + \tfrac{C\log[\log(eN)\epsilon^{-1}]}{\lambda}$$

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 $\mathbb{G}(\lambda) - r'(\tilde{f}) \leq \mathbb{G}_C^{\mathrm{th}}(\lambda)$, hence with $\big(\mathbb{P}^{\otimes 2N}\big)_*$-probability at least $1-\epsilon$, for any $\lambda \in \Lambda$, we have

$$\mathbb{G}(\lambda) - r'(\tilde{f}) \leq \mathbb{G}_C^{\mathrm{th}}(\lambda).$$

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 $\mathbb{G}_C^{\mathrm{th}}(\lambda) \leq C\log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}$.

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).