

LOCALIZED EMPIRICAL COMPLEXITY BOUNDS AND RANDOMIZED ESTIMATORS

OLIVIER CATONI

CNRS – Université Paris 6

ABSTRACT. Within the general framework of statistical learning theory, where a choice between a large number of crudely approximate models of some complex data has to be made, we present an alternative to the penalized maximum likelihood approach based on PAC-Bayesian learning theorems. This approach uses the fact that it is possible to produce non asymptotic universal bounds for the quantiles of the likelihood function under any given prior distribution on the union of the parameter sets of all the candidate models for the data. This is a technical justification for the use of *randomized estimators*, either as practical estimation schemes, or as a mathematical tool in the study of more classical estimators. We achieve here some further step in the localization of estimators, whose assessed performance is no longer bound to depend on the global size of the parameter space, but only on the structure of the quantiles of the likelihood function under a prior distribution (which can be chosen in any arbitrary way to perform some kind of *structural risk minimization*). This leads to a new empirical measure of the complexity of models belonging to the family of *empirical* Vapnik's entropy functions.

1. A PAC-BAYESIAN APPROACH TO ADAPTIVE INFERENCE

1.1. Introduction. In this paper, we will prove what could be called *localized PAC-Bayesian learning theorems*. We will illustrate their use to tackle classification problems. The general purpose of our work is to bring a contribution to *statistical learning theory* : in this setting, complex data have to be analyzed (e.g. images, speech, natural language, DNA, . . .), about which very little is known beforehand and some crudely approximate classification model has to be picked-up among a possibly huge number of candidates through some kind of robust and automated model selection mechanism.

The idea of *PAC-Bayesian* learning theorems, as introduced by D. McAllester, [20, 21] is to measure the complexity of models, and thereby their ability to generalize from observed examples to unknown situations, with the help of some prior probability measure defined on the parameter space. Here, we use for simplicity the term parameter space in a rather loose and unusual way, to talk about the union of all the parameters of all the models we envision (maybe the term model space would be more accurate : these parameters may be of finite or infinite dimension and we do not restrict the number of models, therefore we are definitely not describing a parametric statistical framework, but rather a non-parametric one!).

Date: January 14, 2003.

1991 Mathematics Subject Classification. Primary: 62F35, secondary: 62J02, 94A17, 62G08.

Key words and phrases. Model selection, pattern recognition, oracle inequalities, deviation inequalities, pseudo-Bayesian methods.

The status of the prior measure has not to be misunderstood either : it does not represent the frequency according to which we expect to observe data produced by different probability distributions, nor does it stand for the belief we put in the accuracy of different possible distributions or different possible models. It is somehow equivalent to the choice of some representation of the parameter space (since it is possible to derive some coding scheme from a probability distribution, according to coding theory), and therefore is related to the Minimum Description Length approach of Rissanen and to the structural risk minimization approach of Vapnik. On a more technical level, it is meant to produce non asymptotic *worst case* bounds, (as opposed to a Bayesian study of the mean risk under the prior). It shares some common features with the use of mixture codes in lossless data compression theory [28], which was one of our main sources of inspiration at the beginning of our interest in statistical learning theory.

1.2. Mathematical framework. Let us now sketch the mathematical framework of our study. We consider a product space $\mathcal{X} \times \mathcal{Y}$, where $(\mathcal{X}, \mathcal{B})$ is a measurable space and where \mathcal{Y} is a finite set. In a classification application, the set \mathcal{X} has to be thought of as the *pattern* space and \mathcal{Y} as the *label* space. Patterns in \mathcal{X} may be described by a combination of continuous and discrete parameters, however, except when it comes down to giving examples, we will capture the structure of \mathcal{X} only through the use of a family of classification functions defined on \mathcal{X} , we will come back to this later.

The observation is made of an i.i.d. sample $(X_i, Y_i)_{i=1}^N$, drawn according to some product distribution $P^{\otimes N}$, where P is a probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \times \mathcal{B}')$, \mathcal{B}' being the algebra $\{0, 1\}^{\mathcal{Y}}$ of all the subsets of \mathcal{Y} .

The relations between X and Y will be analyzed with the help of some prescribed set of classification rules

$$\mathcal{R} = \{f_\theta : \mathcal{X} \rightarrow \mathcal{Y}; \theta \in \Theta\},$$

where (Θ, \mathcal{J}) is some measurable parameter set and

$$(\theta, x) \mapsto f_\theta(x) : (\Theta \times \mathcal{X}, \mathcal{J} \times \mathcal{B}) \rightarrow (\mathcal{Y}, \mathcal{B}')$$

is assumed to be measurable. As we have already explained, the set \mathcal{R} will in general not be a single parametric model, but rather the union of a large number of parametric models. From the technical point of view, our aim will be to produce *non asymptotic bounds* for the risk of properly designed estimators of Y given X . The risk of $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ will be measured as its error rate

$$R(\theta) = P(Y \neq f_\theta(X))$$

Let us mention here that throughout the paper the short notation $P(W)$ will be used for the expectation of the random variable W under the distribution P .

The PAC Bayesian approach could roughly be explained as follows: instead of bounding the supremum of the empirical risk

$$r(\theta) = \frac{1}{N} \sum_{k=1}^N \mathbb{1}[Y_k \neq f_\theta(X_k)],$$

with respect to the parameter $\theta \in \Theta$, we study the deviations of the quantiles of $r(\theta)$ with respect to some prior probability measure $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{J})$ defined on the parameter space.

More precisely, as it is well known, we cannot minimize $R(\theta)$ in θ as we would like to do, because $R(\theta)$ is not observable: it depends on the unknown distribution P . The next sensible attempt is to minimize $r(\theta)$ instead. Unfortunately, although $P(r(\theta)) = R(\theta)$, the fluctuations of the random process $r(\theta) : \theta \in \Theta$ may be strong enough to make the solutions of the two minimization problems quite different, and even in many cases completely unrelated. An intensively studied way to get some control on this situation is to add a penalty term $\gamma_N(\theta)$ and study the relations between $\inf_{\theta} R(\theta) + \gamma_N(\theta)$ and $\inf_{\theta} r(\theta) + \gamma_N(\theta)$. The penalty $\gamma_N(\theta)$ has a regularization effect: it shrinks the size of the set of values of θ where $\inf_{\theta} r(\theta) + \gamma_N(\theta)$ is likely to be achieved and therefore provides a way to control the gap between $P\left[\inf_{\theta} [r(\theta) + \gamma_N(\theta)]\right]$ and $\inf_{\theta} R(\theta) + \gamma_N(\theta)$. The difficulty of this approach comes from the choice of $\gamma_N(\theta)$, which has to depend on the “size” of the parameter space Θ , measured in a suitable way.

In the PAC-Bayesian approach, we circumvent this difficulty by measuring weights under some prior distribution $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$ on the parameter space. This is an indirect way to make the size of Θ come into play. Although we will not explicitly manipulate quantiles in the technical part of our study, we will introduce here the role of the prior π with the help of this familiar concept which gives us an opportunity to make a link with the maximum likelihood approach. Let us define the α quantile of the empirical risk $r(\theta)$ as

$$q_{\alpha}(r) = \inf \left\{ \mu : \pi[r(\theta) \leq \mu] > \alpha \right\}.$$

It can be viewed as a probabilistic generalization of the essential infimum of $r(\theta)$ under π , since

$$\text{ess inf}_{\pi(d\theta)} r(\theta) = q_0(r).$$

This generalization is of practical interest to us, because, whereas $\text{ess inf}_{\pi(d\theta)} r(\theta)$ has fluctuations depending on the “size” (or more accurately the complexity) of the parameter space Θ , the fluctuations of the quantile $q_{\alpha}(r)$ can be evaluated as a function of α only, as long as $\alpha > 0$. The reason is that a quantile with positive parameter α is separating two sets of parameters with positive π -weights, unlike the essential infimum which may separate a single point of null π -weight from the rest of the parameter space: to produce a random deviation of the quantile $q_{\alpha}(r)$, the values of $r(\theta)$ for a given proportion (α , namely) of the parameters have to deviate from their typical values, whereas a lower deviation of the essential infimum may be the consequence of the behavior of the empirical risk on a set of parameters of arbitrarily small π -weight (and the behavior of the empirical risk at a single value of the parameter may of course be responsible for a lower deviation of the true infimum).

As shown by D. McAllester in his pioneering papers on the subject, the “hard threshold” vision of quantiles we explained above can be generalized to smoother objects, and indeed to any “posterior distribution” $\rho \in \mathcal{M}_+^1(\Theta)$ on the parameter space. A posterior distribution here is simply a probability measure $\rho \in \mathcal{M}_+^1(\Theta, \mathcal{T})$ on the parameter space which may depend on the observations $(X_i, Y_i)_{i=1}^N$ (therefore it is a random measure).

The random measures depending on the empirical risk $r(\theta)$ are a special case of posterior distributions. More precisely, we will make a heavy use of *Gibbs posterior*

distributions of the form

$$d\rho(\theta) = d\pi_{\exp(-\beta r)}(\theta) = \frac{\exp(-\beta r(\theta))}{\pi[\exp(-\beta r(\theta))]} d\pi(\theta).$$

The introduction of these posterior distributions, viewed as random objects whose fluctuations are easily manageable, leads us to consider *randomized* estimators : instead of picking some parameter $\hat{\theta}$ as a deterministic function of the observations $(X_i, Y_i)_{i=1}^N$, we choose it at random according to the posterior distribution ρ (which itself depends on the observations). The resulting risk of this randomized estimation scheme is $\rho[R(\theta)]$, which plays the same role as $R(\hat{\theta})$ in the deterministic setting. Although it depends on the unknown and deterministic risk function R , it is still a random variable, due to the randomness of the posterior measure ρ , in the same way as $R(\hat{\theta})$ is a random variable due to the dependence of $\hat{\theta}$ on the observations. In some situations, it is natural to use randomized estimators, in others the support of ρ will be concentrated around some deterministic estimator $\hat{\theta}$ in some sensible way and the introduction of randomized estimators should more likely be viewed as a technical steps in the study of more conventional estimation schemes.

In McAllester's papers and previous studies of ours too, the fluctuations of $\rho[r(\theta)]$ with respect to $\rho[R(\theta)]$ are controlled by some function of $\mathcal{K}(\rho, \pi)$, the Kullback divergence of the (random) posterior measure ρ with respect to the (fixed) prior measure π , defined as

$$\mathcal{K}(\rho, \pi) = \begin{cases} \rho \left[\log \left(\frac{d\rho}{d\pi} \right) \right], & \text{when } \rho \ll \pi, \\ +\infty, & \text{otherwise.} \end{cases}$$

In the present study, we will make an important step towards sharper bounds by replacing $\mathcal{K}(\rho, \pi)$ with $\mathcal{K}(\rho, \pi_{\exp(-\beta r)})$, where $\pi_{\exp(-\beta r)} \in \mathcal{M}_+^1(\Theta)$ is the Gibbs posterior built from π and r we already mentioned a few lines above.

We will start with simple PAC-Bayesian learning theorems, explain how they can be used, and introduce further improvements only in subsequent sections. We will also show how Vapnik's statistical learning theory can be proved and improved using the PAC-Bayesian approach : the idea is to replace the use of a deterministic prior with the use of a data dependent prior.

2. LOW NOISE PATTERN CLASSIFICATION

We will be interested here in the most favorable case of pattern recognition: the case when an i.i.d. sample $(X_i, Y_i)_{i=1}^N$ of classified patterns is observed, where the conditional distribution of the label Y given the pattern X is highly peaked on one label (which will of course be considered as the "true" label for pattern X). As already explained, $(X_i, Y_i)_{i=1}^N$ will be the canonical process on some space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{B}')$ endowed with a product measure $P^{\otimes N}$, where $P \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{B}')$. A set of classification rules $\mathcal{R} = \{f_\theta : \mathcal{X} \rightarrow \mathcal{Y}, \theta \in \Theta\}$ is at our disposal, where $(\theta, x) \mapsto f_\theta(x) : (\Theta \times \mathcal{X}, \mathcal{T} \times \mathcal{B}) \rightarrow (\mathcal{Y}, \mathcal{B}')$ is measurable. We will not make any "low-noise" assumption, but it will just turn out that the bounds derived in this section will be sharp only when the empirical risk

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

is such that $\inf_{\theta \in \Theta} r(\theta)$ is small with a high probability.

2.1. A reminder of non-asymptotic deviation techniques: Bernstein's inequality and the Legendre transform of the Kullback divergence function. We need a non-asymptotic deviation inequality for sums of independent random variables. For this purpose, a detailed formulation of Bernstein's inequality is useful. It can be found in [22, p 203-204].

Theorem 2.1. *Let $(\sigma_1, \dots, \sigma_N)$ be independent real valued random variables and \mathbb{P} their joint distribution. Let us assume that*

$$\sigma_i - \mathbb{P}(\sigma_i) \leq b, \quad i = 1, \dots, N.$$

Let

$$S = \frac{1}{N} \sum_{i=1}^N \sigma_i$$

be their normalized sum,

$$m = \mathbb{P}(S) = \frac{1}{N} \sum_{i=1}^N \mathbb{P}(\sigma_i)$$

its expectation and

$$V = N\mathbb{P} \left[(S - \mathbb{P}(S))^2 \right] = \frac{1}{N} \sum_{i=1}^N \mathbb{P} \left[(\sigma_i - \mathbb{P}(\sigma_i))^2 \right]$$

its renormalized variance. Let us introduce the increasing function

$$g(x) = \frac{1}{x^2} (e^x - 1 - x).$$

The deviations of S are bounded, for any $\lambda \in \mathbb{R}_+$, any $\eta \in \mathbb{R}_+$, by

$$\begin{aligned} (1) \quad \mathbb{P}(S - m \geq \eta) &\leq \mathbb{P} \left[\exp \left(-\lambda\eta + \lambda(S - m) \right) \right] \\ (2) \quad &\leq \exp \left(-\eta\lambda + g \left(\frac{b\lambda}{N} \right) \frac{V}{N} \lambda^2 \right), \end{aligned}$$

moreover when λ is chosen to be

$$\lambda = \frac{N}{b} \log \left(1 + \frac{b\eta}{V} \right),$$

the right-hand side of the previous equation is itself bounded by

$$\exp \left(-\eta\lambda + g \left(\frac{b\lambda}{N} \right) \frac{V}{N} \lambda^2 \right) \leq \exp \left(-\frac{3N\eta^2}{6V + 2b\eta} \right).$$

Some background on the Legendre transform of the convex function $\rho \mapsto \mathcal{K}(\rho, \pi)$ is also needed.

Lemma 2.1. *Let us recall that for any measurable function $h : \Theta \rightarrow \mathbb{R}$,*

$$(3) \quad \log \left\{ \pi \left\{ \exp[h(\theta)] \right\} \right\} = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho[h(\theta)] - \mathcal{K}(\rho, \pi),$$

where the value of $\rho[h(\theta)]$ is defined by convention as

$$(4) \quad \rho[h(\theta)] \stackrel{\text{def}}{=} \sup_{B \in \mathbb{R}} \rho \left[\min \{ B, h(\theta) \} \right],$$

and where it is also understood that

$$(5) \quad \infty - \infty = \sup_{B \in \mathbb{R}}(B) - \infty = \sup_{B \in \mathbb{R}}(B - \infty) = -\infty.$$

(In other words a priority is given to $-\infty$ in ambiguous cases : the expectation of a function whose negative part is not integrable will be assumed to be $-\infty$, even when its positive part integrates to $+\infty$.)

Moreover, when h is upper bounded, for any $\rho \in \mathcal{M}_+^1(\Theta, \mathcal{F})$,

$$(6) \quad \log \left\{ \pi \left[\exp[h(\theta)] \right] \right\} + \mathcal{K}(\rho, \pi) - \rho[h(\theta)] = \mathcal{K}(\rho, \nu),$$

where $d\nu(\theta) = \frac{\exp[h(\theta)]}{\pi \left\{ \pi[h(\theta)] \right\}} d\pi(\theta)$. (Equality is meant to hold in $\mathbb{R} \cup \{\infty\}$, meaning

that $\mathcal{K}(\rho, \nu) < \infty$ if and only if $\mathcal{K}(\rho, \pi) < \infty$ and $-\rho[h(\theta)] < \infty$ and that in this case equality holds in \mathbb{R} .)

Proof. Let us give for the sake of completeness a short proof of this well known result. The second part of the lemma is a straightforward computation. Let us remark first that ρ is absolutely continuous with respect to π if and only if it is absolutely continuous with respect to ν , because π and ν have the same negligible measurable sets. Therefore if ρ is singular with respect to π , then both members of (6) are equal to ∞ . Let us assume now that ρ is absolutely continuous with respect to π , and write from the definition of the divergence function

$$\mathcal{K}(\rho, \nu) = \rho \left\{ \log \left(\frac{d\rho}{d\pi} \right) - h(\theta) \right\} + \log \left\{ \pi \left[\exp[h(\theta)] \right] \right\}.$$

Remark that the negative part of $\log \left(\frac{d\rho}{d\pi} \right)$ is in $L^1(\rho)$, because $\frac{d\rho}{d\pi} \left[\log \left(\frac{d\rho}{d\pi} \right) \right]_-$ is bounded and therefore in $L^1(\pi)$. As $-h$ is lower bounded, we can thus write in $\mathbb{R} \cup \{\infty\}$ that

$$\rho \left\{ \log \left(\frac{d\rho}{d\pi} \right) - h(\theta) \right\} = \rho \left\{ \log \left(\frac{d\rho}{d\pi} \right) \right\} - \rho[h(\theta)].$$

This is precisely (6).

In the case when h is upper bounded, the first part of the lemma is a consequence of its second part, which shows moreover that the maximum in ρ is attained when $\rho = \nu$. In the general case, we can write the following chain of equalities, where we have used the notation $\min\{B, h(\theta)\} = B \wedge h(\theta)$,

$$\begin{aligned} \log \left\{ \pi \left\{ \exp[h(\theta)] \right\} \right\} &= \sup_{B \in \mathbb{R}} \log \left\{ \pi \left\{ \exp[B \wedge h(\theta)] \right\} \right\} \\ &= \sup_{B \in \mathbb{R}} \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[B \wedge h(\theta)] - \mathcal{K}(\rho, \pi) \right\} \\ &= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \sup_{B \in \mathbb{R}} \left\{ \rho[B \wedge h(\theta)] - \mathcal{K}(\rho, \pi) \right\} \\ &= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \sup_{B \in \mathbb{R}} \left\{ \rho[B \wedge h(\theta)] \right\} - \mathcal{K}(\rho, \pi) \\ &= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho[h(\theta)] - \mathcal{K}(\rho, \pi). \end{aligned}$$

□

2.2. A non localized learning theorem for low-noise classification. We will apply the second inequality (2) of Bernstein's theorem 2.1 successively to

$$\sigma_i \stackrel{\text{def}}{=} -\mathbf{1}(Y_i \neq f_\theta(X_i))$$

and to

$$\sigma_i \stackrel{\text{def}}{=} \mathbf{1}(Y_i \neq f_\theta(X_i)).$$

We will integrate both sides of the resulting inequality with respect to some prior $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$, to obtain a "learning" lemma which improves on the PAC-Bayesian bounds in [20, 21], which were derived from the weaker Hoeffding's inequality.

Lemma 2.2. *For any positive real parameter $\lambda \in \mathbb{R}_+^*$, any non negative real valued measurable function $\eta : \Theta \rightarrow \mathbb{R}_+$, any prior probability distribution $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$,*

$$\begin{aligned} P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[R(\theta)] - \lambda \rho[r(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi) \geq 0 \right\} \\ \leq \pi \left\{ \exp \left[\frac{\lambda^2}{N} g \left(\frac{\lambda R(\theta)}{N} \right) R(\theta) [1 - R(\theta)] - \eta(\theta) \right] \right\}. \end{aligned}$$

In the same way

$$\begin{aligned} (7) \quad P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[r(\theta)] - \lambda \rho[R(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi) \geq 0 \right\} \\ \leq \pi \left\{ \exp \left[\frac{\lambda^2}{N} g \left(\frac{\lambda}{N} \right) R(\theta) [1 - R(\theta)] - \eta(\theta) \right] \right\}. \end{aligned}$$

Proof. According to lemma 2.1,

$$\begin{aligned} \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho[\lambda R(\theta) - \lambda r(\theta) - \eta(\theta)] - \mathcal{K}(\rho, \pi) \\ = \log \left\{ \pi \left\{ \exp \left[\lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right\} \right\}. \end{aligned}$$

Thus

$$(8) \quad P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[R(\theta)] - \lambda \rho[r(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi) \geq 0 \right\}$$

$$(9) \quad = P^{\otimes N} \left\{ \pi \left\{ \exp \left[\lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right\} \geq 1 \right\}$$

$$(10) \quad \leq P^{\otimes N} \left\{ \pi \left\{ \exp \left[\lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right\} \right\}$$

$$(11) \quad = \pi \left\{ P^{\otimes N} \left\{ \exp \left[\lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right\} \right\}$$

$$(12) \quad \leq \pi \left\{ \exp \left[\frac{\lambda^2}{N} g \left(\frac{\lambda R(\theta)}{N} \right) R(\theta) [1 - R(\theta)] - \eta(\theta) \right] \right\}.$$

Equality (11) is obtained by applying the Fubini theorem to the positive function $(\theta, X_1, Y_1, \dots, X_N, Y_N) \mapsto \exp \left\{ \lambda [R(\theta) - r(\theta)] - \eta(\theta) \right\}$. Inequality (12) is obtained

by applying inequality (2) of Bernstein's theorem 2.1 for each value of the parameter θ .

The proof of the reverse inequality (7) is similar and is left to the reader. \square

Remark 2.1. The last step of the proof (12) can be replaced with an equality depending on the unknown distribution P , which is of a less practical interest but may bring some further understanding of the situation: indeed, it could be written that

$$\pi \left\{ P^{\otimes N} \left[\exp \left[\lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right] \right\} = \pi \left\{ \exp \left\{ N \mathcal{K} \left[P, P_{\exp(\frac{\lambda}{N} \sigma)} \right] - \eta(\theta) \right\} \right\},$$

where $\sigma(\theta, X, Y) = -\mathbf{1}[Y \neq f_\theta(X)]$ and for any positive measurable function $h: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+^*$ we have introduced the notation

$$dP_{h(X, Y)} = P[h(X, Y)]^{-1} dP(X, Y).$$

This is a simple application of equality (6) in another context.

In the sequel of this paper, we will state a series of more sophisticated learning lemmas. Therefore it may be of some help to stop for a moment and see what use can be made of this type of result and how it can be compared with more classical statistical theorems. The easiest way to build an estimator and estimate its performance using lemma 2.2 is to apply it choosing $\eta(\theta) = \log(\epsilon^{-1}) - \frac{\lambda^2}{N} g\left(\frac{\lambda}{N}\right) R(\theta)$, to get

Corollary 2.1. *With $P^{\otimes N}$ probability at least $1 - \epsilon$, for any posterior distribution $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$(13) \quad \rho[R(\theta)] \leq \left[1 - \frac{\lambda}{N} g\left(\frac{\lambda}{N}\right) \right]^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} [\mathcal{K}(\rho, \pi) + \log(\epsilon^{-1})] \right\}.$$

The above inequality is the kind of non-asymptotic empirical bound we will be hunting after in the whole paper. Let us show here that it provides in a natural way an estimator with a given level of confidence. Building a randomized estimator from an empirical bound is straightforward: it is obtained by minimizing the bound with respect to the posterior distribution ρ . Let $\hat{\rho}$ be this minimizing posterior. Its risk has an upper confidence bound $B(\hat{\rho}, \epsilon)$ at level ϵ , where

$$B(\rho, \epsilon) = \left(1 - \kappa \frac{\lambda}{N} \right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) + \frac{\log(\epsilon^{-1})}{\lambda} \right\},$$

where we have put $\kappa = g\left(\frac{\lambda}{N}\right) \simeq \frac{1}{2}$ for short. In other words,

$$P^{\otimes N} \left\{ \hat{\rho}[R(\theta)] \geq B(\hat{\rho}, \epsilon) \right\} \leq \epsilon.$$

This is satisfactory from the practical point of view, since $B(\hat{\rho}, \epsilon)$ is computable from the observed sample $(X_i, Y_i)_{i=1}^N$. However, from a theoretical point of view, the reader may wonder about the performance of the estimator, that is about the link between $B(\hat{\rho}, \epsilon)$ and $\inf_{\theta \in \Theta} R(\theta)$. There is a standard way to deal with this question. Let us explain it here as a motivation for the following. For any fixed

distribution $\rho \in \mathcal{M}_+^1(\Theta, \mathcal{T})$, the empirical (i.e. random) bound $B(\rho, \epsilon)$ is up to some constant a sum of i.i.d. random variables, with mean $\bar{B}(\rho, \epsilon)$ given by

$$\bar{B}(\rho, \epsilon) = \left(1 - \kappa \frac{\lambda}{N}\right)^{-1} \left\{ \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) + \frac{\log(\epsilon^{-1})}{\lambda} \right\}.$$

It is straightforward to estimate its deviations. We can for instance write that

$$P^{\otimes N} \left\{ \rho[r(\theta)] \geq \left(1 + \kappa \frac{\lambda}{N}\right) \rho[R(\theta)] + \frac{\log(\epsilon^{-1})}{\lambda} \right\} \leq \epsilon,$$

Moreover, from the construction of $\hat{\rho}$, $B(\hat{\rho}, \epsilon) \leq B(\rho, \epsilon)$. Thus for any $\rho \in \mathcal{M}_+^1(\Theta, \mathcal{T})$, with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$B(\hat{\rho}, \epsilon) \leq B(\rho, \epsilon) \leq \left(1 - \kappa \frac{\lambda}{N}\right)^{-1} \left\{ \left(1 + \kappa \frac{\lambda}{N}\right) \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) + \frac{2 \log(\epsilon^{-1})}{\lambda} \right\}.$$

However, the right-hand side of this last inequality is non random, and therefore can legitimately be optimized in ρ . Weakening a little the result to make it more readable, we have proved that with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$\hat{\rho}[R(\theta)] \leq \frac{1 + \kappa \frac{\lambda}{N}}{1 - \kappa \frac{\lambda}{N}} \left\{ \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\} + \frac{2 \log(\frac{2}{\epsilon})}{(1 - \kappa \frac{\lambda}{N}) \lambda}.$$

It is also possible to bound the mean risk $P^{\otimes N} \left\{ \hat{\rho}[R(\theta)] \right\}$. One standard way to achieve this is to start from inequality

$$P^{\otimes N} \left\{ \hat{\rho}[R(\theta)] \geq B(\rho, \epsilon) \right\} \leq \epsilon,$$

where we have kept ρ non random, and to rewrite it as

$$P^{\otimes N}(U \geq \alpha) \leq \exp(-\lambda \alpha),$$

where we have introduced the random variable

$$U = \hat{\rho}[R(\theta)] - \left(1 - \kappa \frac{\lambda}{N}\right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\}.$$

We have

$$P^{\otimes N}(U) \leq \int_0^{+\infty} P^{\otimes N}(U \geq \alpha) d\alpha \leq \frac{1}{\lambda} \left(1 - \kappa \frac{\lambda}{N}\right)^{-1}.$$

In other words,

$$P^{\otimes N} \left\{ \hat{\rho}[R(\theta)] \right\} \leq \left(1 - \kappa \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) + \frac{1}{\lambda} \right\}.$$

A slight improvement is achieved if we come back to (10) and (12). With a proper choice of parameters, we get

$$\begin{aligned} & P^{\otimes N} \left\{ \pi \left[\exp \left[\lambda \left(1 - \kappa \frac{\lambda}{N}\right) R(\theta) - \lambda r(\theta) \right] \right] \right\} \\ &= P^{\otimes N} \left\{ \exp \left[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \left(1 - \kappa \frac{\lambda}{N}\right) \rho[R(\theta)] - \lambda \rho[r(\theta)] - \mathcal{K}(\rho, \pi) \right] \right\} \\ &\leq 1. \end{aligned}$$

Using Jensen's inequality for the (convex) exponential function, we see that

$$P^{\otimes N} \left\{ \lambda \left(1 - \kappa \frac{\lambda}{N}\right) \hat{\rho}[R(\theta)] - \lambda \hat{\rho}[r(\theta)] - \mathcal{K}(\hat{\rho}, \pi) \right\} \leq 0.$$

Therefore

$$\begin{aligned} P^{\otimes N} \left\{ \hat{\rho}[R(\theta)] \right\} &\leq \left(1 - \kappa \frac{\lambda}{N}\right)^{-1} P^{\otimes N} \left\{ \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\} \\ &\leq \left(1 - \kappa \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_+^1(\Theta)} P^{\otimes N} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\} \\ &= \left(1 - \kappa \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\}. \end{aligned}$$

Another important remark is to notice that corollary 2.1 can also be optimized in λ . A simple way to do this is to consider a countable (possibly dense) family $\Lambda \subset \mathbb{R}$ and some probability measure ν on Λ . Then defining $\hat{\lambda}$ to be the minimizer in λ of

$$\left(1 - \kappa \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) + \frac{\log[\nu(\lambda)^{-1}]}{\lambda} \right\},$$

we get some estimator $\hat{\rho}_{\hat{\lambda}}$ satisfying with $P^{\otimes N}$ probability at least $1 - \epsilon$

$$\begin{aligned} &\hat{\rho}_{\hat{\lambda}}[R(\theta)] \\ &\leq \inf_{\lambda \in \Lambda, \rho \in \mathcal{M}_+^1(\Theta)} \left(1 - \kappa \frac{\lambda}{N}\right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) + \frac{1}{\lambda} \log(\epsilon^{-1} \nu(\lambda)^{-1}) \right\}. \end{aligned}$$

A more sophisticated way to optimize in λ is to establish a learning lemma uniform in both λ and ρ . Let $\nu \in \mathcal{M}(\mathbb{R}_+, \mathcal{B})$ be some prior on the positive real line equipped with the Borel sigma algebra. Similarly to what has been proved before

Corollary 2.2. *With $P^{\otimes N}$ probability at least $1 - \epsilon$ for any posterior distributions $\mu \in \mathcal{M}_+^1(\mathbb{R}_+)$ and $\rho \in \mathcal{M}_+^1(\Theta)$*

$$\rho[R(\theta)] \leq \left(1 - \frac{\mu(\kappa \lambda^2)}{\mu(\lambda)N}\right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\mu(\lambda)} \left[\mathcal{K}(\rho, \pi) + \mathcal{K}(\mu, \nu) + \log(\epsilon^{-1}) \right] \right\}.$$

(The union bound approach is the special case of this last inequality where ν has a countable support and μ is a Dirac mass).

Of course, the link previously made between empirical and theoretical bounds can be carried over to the empirical bounds optimized in λ :

Corollary 2.3. *If $\hat{\mu}$ and $\hat{\rho}$ are the optimizers of the empirical bound at level of confidence ϵ , then with $P^{\otimes N}$ probability at least $1 - \epsilon$*

$$\begin{aligned} &\hat{\rho}[R(\theta)] \\ &\leq \inf_{\mu \in \mathcal{M}_+^1(\mathbb{R}_+), \rho \in \mathcal{M}_+^1(\Theta)} \left\{ \left(1 - \frac{\mu(\kappa \lambda^2)}{\mu(\lambda)N}\right)^{-1} \left\{ \left(1 + g\left(\frac{\mu(\lambda)}{N}\right) \frac{\mu(\lambda)}{N}\right) \rho[R(\theta)] \right. \right. \\ &\quad \left. \left. + \frac{1}{\mu(\lambda)} \left[\mathcal{K}(\rho, \pi) + \mathcal{K}(\mu, \nu) + 2 \log\left(\frac{2}{\epsilon}\right) \right] \right\} \right\}. \end{aligned}$$

Let us show eventually how corollary 2.2 can be used. Consider the prior

$$\nu(d\lambda) = \alpha \lambda^{-(\alpha+1)} \mathbf{1}(\lambda \geq 1) d\lambda$$

and the posteriors

$$\mu_\beta(d\lambda) = \mathbf{1}(\beta \leq \lambda \leq 2\beta) \frac{\nu(d\lambda)}{\nu([\beta, 2\beta])}.$$

As $\mathcal{K}(\mu_\beta, \nu) = -\log[\nu([\beta, 2\beta])] \leq \frac{\alpha}{2}(2\beta)^{-\alpha}$, choosing $\alpha = \frac{1}{\log(N)}$, we get

Corollary 2.4. *For any $N \leq 2 \cdot 10^{21}$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\rho[R(\theta)] \leq \inf_{\beta \in [1, N/3]} \left(1 - \frac{3\beta}{N}\right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\beta} \left[\mathcal{K}(\rho, \pi) + 7 + \log(\epsilon^{-1}) \right] \right\}.$$

Note that the numerical constants in 2.4 are far from being optimized, we gave this corollary to show that optimization in λ is rather harmless. Note also that it is possible to get an explicit form for the optimal ρ for a fixed value of β in corollary 2.4. Namely

$$d\hat{\rho}_\beta(\theta) = \exp[-\beta r(\theta)] \frac{d\pi(\theta)}{\pi\{\exp[-\beta r(\theta)]\}}.$$

Corollary 2.5. *For any $N \leq 2 \cdot 10^{21}$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\beta \in [1, N/3]$,*

$$\hat{\rho}_\beta[R(\theta)] \leq \left(1 - \frac{3\beta}{N}\right)^{-1} \left\{ -\frac{1}{\beta} \log\{\pi\{\exp[-\beta r(\theta)]\}\} + \frac{7 + \log(\epsilon^{-1})}{\beta} \right\}.$$

Let us note also that the upper deviations of the risk $R(\theta)$ under the optimizer $\hat{\rho}_\beta$ of the bound given in corollary 2.4 can easily be bounded. Indeed with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\beta \in [1, N/3]$, any $\lambda \in \mathbb{R}_+$,

$$\begin{aligned} \log\left\{ \hat{\rho}_\beta\left\{ \exp[\lambda R(\theta)] \right\} \right\} &= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[R(\theta)] - \mathcal{K}(\rho, \hat{\rho}_\beta) \\ &= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[R(\theta)] - \mathcal{K}(\rho, \pi) \\ &\quad - \beta \rho[r(\theta)] - \log\left\{ \pi\left[\exp[-\beta r(\theta)] \right] \right\} \\ &\leq \lambda \left(1 - \frac{3\beta}{N}\right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\beta} \left[\mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}) + 7 \right] \right\} \\ &\quad - \mathcal{K}(\rho, \pi) - \beta \rho[r(\theta)] - \log\left\{ \pi\left[\exp[-\beta r(\theta)] \right] \right\}. \end{aligned}$$

Thus choosing $\lambda = \beta \left(1 - \frac{3\beta}{N}\right)$, we get

$$\begin{aligned} \log\left\{ \hat{\rho}_\beta\left[\exp\left[\beta \left(1 - \frac{3\beta}{N}\right) R(\theta)\right] \right] \right\} \\ \leq -\log\left\{ \pi\left[\exp[-\beta r(\theta)] \right] \right\} + \log(\epsilon^{-1}) + 7. \end{aligned}$$

This proves

Corollary 2.6. *For any integer $N \leq 2 \cdot 10^{21}$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\beta \in [1, N/3]$, with $\hat{\rho}_\beta$ probability at least $1 - \epsilon$,*

$$R(\theta) \leq \left(1 - \frac{3\beta}{N}\right)^{-1} \left\{ -\frac{1}{\beta} \log \left\{ \pi \left[\exp[-\beta r(\theta)] \right] \right\} + \frac{1}{\beta} \left[2 \log(\epsilon^{-1}) + 7 \right] \right\}$$

The same kind of deviation upper bounds with respect to the minimizing posterior can be derived from the bounds presented in the sequel of this paper. Details are straightforward and will be left to the reader.

2.3. Example. Before delving into improvements, let us illustrate the use of these simple bounds to build aggregated classifiers.

Let $\{f_\theta : \mathcal{X} \rightarrow \{-1, +1\}; \theta \in \Theta\}$ be some family of classification rules in a two classes pattern recognition problem. Here the label space is equal to $\mathcal{Y} = \{-1, +1\}$. For any probability measure $\rho \in \mathcal{M}_+^1(\Theta)$, we consider the aggregated classifier

$$f_\rho(x) = \text{sign}(\rho[f_\theta(x)]).$$

If P is as previously the joint distribution of the patterns and labels, then the error rate of f_ρ is

$$R(\rho) = P\{Y\rho[f_\theta(X)] < 0\},$$

and the corresponding empirical risk is

$$r(\rho) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{Y_i \rho[f_\theta(X_i)] < 0\}.$$

In this problem, the space of parameters is $\Theta' = \mathcal{M}_+^1(\Theta)$ and we need to consider some reference measure on this space to apply our method. One way to do this is to consider the mapping

$$\begin{aligned} \Psi : \Theta^M &\rightarrow \Theta' \\ \theta_1^M &\mapsto \frac{1}{M} \sum_{i=1}^M \delta_{\theta_i}. \end{aligned}$$

Consider some reference probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ and build a prior π' belonging to $\mathcal{M}_+^1(\Theta')$ from the formula

$$\pi' = \pi^{\otimes M} \circ \Psi^{-1}.$$

Lemma 2.3. *For any probability measure $\rho \in \mathcal{M}_+^1(\Theta^M)$, the posterior distribution $\rho' = \rho \circ \Psi^{-1}$ on $\mathcal{M}_+^1(\Theta')$ is such that*

$$\mathcal{K}(\rho', \pi') \leq \mathcal{K}(\rho, \pi^{\otimes M}).$$

Proof. This is a consequence of the decomposition of the Kullback divergence function :

$$\mathcal{K}(\rho, \pi^{\otimes M}) = \mathcal{K}(\rho', \pi') + \rho \left\{ \mathcal{K} \left[\rho [d\theta_1^M | \Psi(\theta_1^M)], \pi^{\otimes M} [d\theta_1^M | \Psi(\theta_1^M)] \right] \right\}.$$

Note that equality holds when ρ is a product measure. □

From corollary 2.4, for any $N \leq 2 \cdot 10^{21}$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\beta \in [1, N/3]$, any $\rho \in \mathcal{M}_+^1(\Theta^M)$,

$$\rho \left\{ R[\Psi(\theta_1^M)] \right\} \leq \left(1 - \frac{3\beta}{N} \right)^{-1} \left\{ \rho \left\{ r[\Psi(\theta_1^M)] \right\} + \frac{1}{\beta} \left[\mathcal{K}(\rho, \pi^{\otimes M}) + 7 + \log(\epsilon^{-1}) \right] \right\}.$$

Optimizing the right-hand side of this empirical inequality in ρ gives a posterior $\hat{\rho}_\beta$ defined by

$$d\hat{\rho}_\beta(\theta_1^M) = \exp \left\{ -\beta \frac{1}{N} \sum_{i=1}^N \mathbb{1} \left[Y_i \frac{1}{M} \sum_{j=1}^M f_{\theta_j}(X_i) < 0 \right] \right\} d\pi^{\otimes M}(\theta_1^M).$$

Theorem 2.2. *For any $N < 2 \cdot 10^{21}$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any choice of temperature $\beta \in [1, N/3]$,*

$$\begin{aligned} & \hat{\rho}_\beta \left\{ R[\Psi(\theta_1^M)] \right\} \\ & \leq \left(1 - \frac{3\beta}{N} \right)^{-1} \left\{ -\frac{1}{\beta} \log \left\{ \pi^{\otimes M} \left[\exp \left[-\beta r[\Psi(\theta_1^M)] \right] \right] \right\} + \frac{7 + \log(\epsilon^{-1})}{\beta} \right\} \end{aligned}$$

A union bound can furthermore be used to optimize the value of M .

The posterior $\hat{\rho}_\beta$ can be simulated using a Metropolis algorithm at temperature β . A simulated annealing scheme can be useful to compute an approximation of the right-hand side. Indeed we can write

$$\begin{aligned} -\log \left\{ \pi^{\otimes M} \left[\exp \left[-\beta r[\Psi(\theta_1^M)] \right] \right] \right\} &= \int_0^\beta \hat{\rho}_\gamma \left\{ r[\Psi(\theta_1^M)] \right\} d\gamma \\ &\leq \sum_{j=0}^{K-1} (\gamma_{j+1} - \gamma_j) \hat{\rho}_{\gamma_j} \left\{ r[\Psi(\theta_1^M)] \right\} \end{aligned}$$

for any sequence of temperatures $\gamma_0 = 0 < \gamma_1 < \dots < \gamma_K = \beta$, where we have used the fact that $\gamma \mapsto \hat{\rho}_\gamma \left\{ r[\Psi(\theta_1^M)] \right\}$ is decreasing (its derivative being the opposite of a variance). This leads to the following computation scheme : estimate $\hat{\rho}_{\gamma_j} \left\{ r[\Psi(\theta_1^M)] \right\}$ for increasing values of γ_j and compute the bound for $\hat{\rho}_{\gamma_j} \left\{ R[\Psi(\theta_1^M)] \right\}$. Keep the temperature with the best bound. If we do not trust the constants in the bound, we can first spend some more effort sharpening them and then also keep the highest temperature for which the bound is not more than a certain level above its minimum value. This could lead to less regularized estimators while keeping some warranty against over fitting.

2.4. Comments. The results of this section have at least two weaknesses:

- the penalty $\mathcal{K}(\rho, \pi)$ is not as local as it could be;
- noisy samples are not handled properly.

We would also like to make some connection between the penalty terms we present here and Vapnik's entropy. This is to be the subject of the three following sections.

3. LOCALIZED LEARNING LEMMAS

The loss of localization in the use we made so far of lemma 2.2 came from the choice of $\eta(\theta)$: it was chosen to make the contribution of each θ in the level of confidence equal to ϵ , whereas close to optimal values of θ may be expected to play a more critical role than others.

Better localization is achieved by choosing

$$\eta(\theta) = \frac{\lambda^2}{N} g\left(\frac{\lambda}{N}\right) R(\theta) + \beta R(\theta) + \log \left\{ \pi \left[\exp[-\beta R(\theta)] \right] \right\} + \log(\epsilon^{-1}),$$

leading to

$$(14) \quad P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left(\lambda - \beta - \kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] - \lambda \rho[r(\theta)] \right. \\ \left. - \mathcal{K}(\rho, \pi) - \log \left\{ \pi \left[\exp[-\beta R(\theta)] \right] \right\} \geq \log(\epsilon^{-1}) \right\} \leq \epsilon,$$

where we have put as usual $\kappa = g\left(\frac{\lambda}{N}\right)$ for short. With the same choice of parameters, the reverse inequality reads as

$$(15) \quad P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[r(\theta)] - \left(\lambda + \beta + \kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] \right. \\ \left. - \mathcal{K}(\rho, \pi) - \log \left\{ \pi \left[\exp[-\beta R(\theta)] \right] \right\} \geq \log(\epsilon^{-1}) \right\} \leq \epsilon,$$

To exploit these inequalities, we need an empirical upper bound for $\log \left\{ \pi \left[\exp[-\beta R(\theta)] \right] \right\}$. This is where the reverse inequality (15) comes into play: with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$\log \left\{ \pi \left[\exp[-\beta R(\theta)] \right] \right\} = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} -\beta \rho[R(\theta)] - \mathcal{K}(\rho, \pi) \\ \leq \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \beta \left(\lambda + \beta + \kappa \frac{\lambda^2}{N} \right)^{-1} \left\{ -\lambda \rho[r(\theta)] + \mathcal{K}(\rho, \pi) \right. \\ \left. + \log \left\{ \pi \left[\exp[-\beta R(\theta)] \right] \right\} + \log(\epsilon^{-1}) \right\} - \mathcal{K}(\rho, \pi)$$

Putting $\xi = \frac{\beta}{\lambda + \kappa \frac{\lambda^2}{N}}$, this can be rewritten as

$$\log \left\{ \pi \left[\exp[-\beta R(\theta)] \right] \right\} \leq \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ -\xi \lambda \rho[r(\theta)] - \mathcal{K}(\rho, \pi) \right\} + \xi \log(\epsilon^{-1}) \\ (16) \quad = \log \left\{ \pi \left[\exp[-\xi \lambda r(\theta)] \right] \right\} + \xi \log(\epsilon^{-1}).$$

Combining this result with (14), we get

Lemma 3.1 (localized learning lemma, first form). *For any $\lambda \in \mathbb{R}_+$ and $\xi \in [0, 1[$*

$$P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left((1 - \xi)\lambda - (1 + \xi)\kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] - \lambda \rho[r(\theta)] \right. \\ \left. - \mathcal{K}(\rho, \pi) - \log \left\{ \pi \left[\exp[-\xi \lambda r(\theta)] \right] \right\} \geq (1 + \xi) \log\left(\frac{2}{\epsilon}\right) \right\} \leq \epsilon$$

Another way to write this localized learning lemma is to remark that for any $\rho \in \mathcal{M}_+^1$,

$$\mathcal{K}(\rho, \pi) + \log \left\{ \pi \left[\exp[-\xi \lambda r(\theta)] \right] \right\} = \mathcal{K} \left(\rho, \pi_{\exp[-\xi \lambda r(\theta)]} \right) - \xi \lambda \rho[r(\theta)],$$

where

$$d\pi_{\exp[-\xi \lambda r(\theta)]}(\theta) = \frac{\exp[-\xi \lambda r(\theta)]}{\pi \left[\exp[-\xi \lambda r(\theta)] \right]} d\pi(\theta).$$

Lemma 3.2 (localized learning lemma, second form). *For any $\lambda \in \mathbb{R}_+$ and $\xi \in [0, 1[$*

$$P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left((1 - \xi)\lambda - (1 + \xi)\kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] - (1 - \xi)\lambda \rho[r(\theta)] \right. \\ \left. - \mathcal{K} \left(\rho, \pi_{\exp[-\xi \lambda r(\theta)]} \right) \geq (1 + \xi) \log\left(\frac{2}{\epsilon}\right) \right\} \leq \epsilon$$

Note that the newly introduced parameter ξ controls the level of localization of the bound : the value $\xi = 0$ corresponds to the non localized learning lemma (up to some minor loss in the confidence level).

Applying the first form of the localized learning lemma, we see that the optimal posterior $\hat{\rho}_\lambda$ is of the same form as in the non localized case :

$$d\hat{\rho}_\lambda(\theta) = \frac{\exp[-\lambda r(\theta)]}{\pi \left\{ \exp[-\lambda r(\theta)] \right\}} d\pi(\theta).$$

It satisfies

Corollary 3.1. *With $P^{\otimes N}$ probability at least $1 - \epsilon$,*

$$\hat{\rho}_\lambda[R(\theta)] \leq \left[(1 - \xi)\lambda - (1 + \xi)\kappa \frac{\lambda^2}{N} \right]^{-1} \left\{ \log \left\{ \pi \left[\exp[-\lambda r(\theta)] \right] \right\} \right. \\ \left. + \log \left\{ \pi \left[\exp[-\xi \lambda r(\theta)] \right] \right\} + (1 + \xi) \log\left(\frac{2}{\epsilon}\right) \right\} \\ = \left(1 - \frac{1 + \xi}{1 - \xi} \kappa \frac{\lambda}{N} \right)^{-1} \left\{ \frac{1}{1 - \xi} \int_\xi^1 \hat{\rho}_{\beta\lambda}[r(\theta)] d\beta + \frac{(1 + \xi)}{(1 - \xi)\lambda} \log\left(\frac{2}{\epsilon}\right) \right\} \\ \leq \left(1 - \frac{1 + \xi}{1 - \xi} \kappa \frac{\lambda}{N} \right)^{-1} \left\{ \hat{\rho}_{\xi\lambda}[r(\theta)] + \frac{(1 + \xi)}{(1 - \xi)\lambda} \log\left(\frac{2}{\epsilon}\right) \right\}$$

Let us remark that this theorem is quite satisfactory from the point of view of localization. It says that the performance of the Gibbs randomized estimator on the observed sample used for training can be trusted to be the same as it will be on previously unseen patterns, up to some penalty factors which do not depend on

the size of the model and some increase of the temperature from $\frac{1}{\lambda}$ to $\frac{1}{\xi\lambda}$: it can be said that the complexity of the model is taken into account by the Gibbs estimator in an automated way.

To get the corresponding theoretical bound, we can come back to (14) to see that for any fixed probability measure $\rho \in \mathcal{M}_+^1(\Theta)$, with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$\hat{\rho}_\lambda[R(\theta)] \leq \left[\lambda - \beta - \kappa \frac{\lambda^2}{N} \right]^{-1} \left\{ \lambda \rho[r(\theta)] + \mathcal{K}(\rho, \pi) \right. \\ \left. + \log \left\{ \pi \left[\exp[-\beta R(\theta)] \right] \right\} + \log(\epsilon^{-1}) \right\}$$

Moreover, from Bernstein's inequality (2), with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$\lambda \rho[r(\theta)] \leq \left(\lambda + \kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] + \log(\epsilon^{-1}).$$

Thus, putting $\beta = \xi\lambda$, with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$\hat{\rho}_\lambda[R(\theta)] \leq \left[(1 - \xi)\lambda - \kappa \frac{\lambda^2}{N} \right]^{-1} \left\{ \left(\lambda + \kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] \right. \\ \left. + \mathcal{K}(\rho, \pi) + \log \left\{ \pi \left[\exp[-\xi\lambda R(\theta)] \right] \right\} + 2 \log \left(\frac{2}{\epsilon} \right) \right\}$$

As explained in the case of non localized bounds, the right-hand side being non random can be optimized in ρ , leading to

Corollary 3.2. *With $P^{\otimes N}$ probability at least $1 - \epsilon$,*

$$\hat{\rho}_\lambda[R(\theta)] \leq \inf_{\xi \in [0, 1[} \left(1 - \frac{\kappa}{1 - \xi} \frac{\lambda}{N} \right)^{-1} \left\{ \frac{1}{1 - \xi} \int_{\xi}^{1 + \frac{\kappa}{N}} \pi_{\exp[-\beta\lambda R(\theta)]} [R(\theta)] d\beta \right. \\ \left. + \frac{2}{(1 - \xi)\lambda} \log \left(\frac{2}{\epsilon} \right) \right\} \\ \leq \inf_{\xi \in [0, 1[} \left(1 - \frac{1}{1 - \xi} \kappa \frac{\lambda}{N} \right)^{-1} \left\{ \left(1 + \frac{\kappa}{1 - \xi} \frac{\lambda}{N} \right) \pi_{\exp[-\xi\lambda R(\theta)]} [R(\theta)] \right. \\ \left. + \frac{2}{(1 - \xi)\lambda} \log \left(\frac{2}{\epsilon} \right) \right\}.$$

Let us show now how to make corollary 3.1 uniform in λ and ξ , as it is desirable to optimize these two constants.

Let us first use a union bound on λ for a fixed value of ξ . Let ζ be some constant in $[\xi, 1[$, (we can for instance choose $\zeta = \max\{\xi, \frac{1}{2}\}$) and let

$$\Lambda = \left\{ 2N\zeta^k, 0 \leq k < \frac{\log(2N)}{\log(\zeta^{-1})} \right\}.$$

For any $\lambda \in [1, 2N]$, let $\lambda' \in \Lambda$ be such that $\zeta\lambda' \leq \lambda \leq \lambda'$. From lemma 3.1 we deduce that with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\lambda \in [1, 2N]$,

$$\begin{aligned} \hat{\rho}_\lambda[R(\theta)] \leq & \left[(1 - \xi)\lambda' - (1 + \xi)g\left(\frac{\lambda'}{N}\right)\frac{\lambda'^2}{N} \right]^{-1} \left\{ \lambda' \hat{\rho}_{\lambda'}[r(\theta)] + \mathcal{K}(\hat{\rho}_{\lambda'}, \pi) \right. \\ & \left. + \log \left\{ \pi \left[\exp[-\xi\lambda' r(\theta)] \right] \right\} + (1 + \xi) \log \left(\frac{2 \log(2N)}{\epsilon \log(\zeta^{-1})} \right) \right\} \end{aligned}$$

We can now use the fact that

$$\mathcal{K}(\hat{\rho}_\lambda, \pi) = -\log \left\{ \pi \left[\exp[-\lambda r(\theta)] \right] \right\} - \lambda \hat{\rho}_\lambda[r(\theta)]$$

to get

$$\begin{aligned} \hat{\rho}_\lambda[R(\theta)] \leq & \left(1 - \xi - (1 + \xi)g\left(\frac{\lambda'}{N}\right)\frac{\lambda'}{N} \right)^{-1} \left\{ \left(1 - \frac{\lambda}{\lambda'} \right) \hat{\rho}_\lambda[r(\theta)] \right. \\ & \left. + \int_\xi^{\frac{\lambda}{\lambda'}} \hat{\rho}_{\beta\lambda'}[r(\theta)] d\beta + \frac{(1 + \xi)}{\lambda'} \log \left(\frac{2 \log(2N)}{\epsilon \log(\zeta^{-1})} \right) \right\}. \end{aligned}$$

Let us remark now that

$$\begin{aligned} & \int_\xi^{\frac{\lambda}{\lambda'}} \hat{\rho}_{\beta\lambda'}[r(\theta)] d\beta + \left(1 - \frac{\lambda}{\lambda'} \right) \hat{\rho}_\lambda[r(\theta)] \\ & \leq \int_\xi^{\lambda/\lambda'} \hat{\rho}_{\beta\lambda}[r(\theta)] d\beta + \int_{\lambda/\lambda'}^1 \hat{\rho}_{\lambda\beta}[r(\theta)] d\beta \\ & = \int_\xi^1 \hat{\rho}_{\beta\lambda}[r(\theta)] d\beta. \end{aligned}$$

We have proved

Corollary 3.3. *For any $\xi \in [0, 1[$, any $\zeta \in [\xi, 1[$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\lambda \in [1, 2N]$,*

$$\begin{aligned} \hat{\rho}_\lambda[R(\theta)] \leq & \frac{\frac{1}{1 - \xi} \int_\xi^1 \hat{\rho}_{\beta\lambda}[r(\theta)] d\beta + \frac{(1 + \xi)}{(1 - \xi)\lambda} \log \left(\frac{2 \log(2N)}{\epsilon \log(\zeta^{-1})} \right)}{1 - \frac{1 + \xi}{1 - \xi} g\left(\frac{\lambda}{\zeta N}\right) \frac{\lambda}{\zeta N}} \\ & \leq \frac{\hat{\rho}_{\xi\lambda}[r(\theta)] + \frac{(1 + \xi)}{(1 - \xi)\lambda} \log \left(\frac{2 \log(2N)}{\epsilon \log(\zeta^{-1})} \right)}{1 - \frac{1 + \xi}{1 - \xi} g\left(\frac{\lambda}{\zeta N}\right) \frac{\lambda}{\zeta N}}. \end{aligned}$$

Comparing this results with corollary 3.1 shows that gaining uniformity in λ is quite harmless to the quality of the bound. We can of course now go further by using a union bound for different values of ξ . Since the bound explodes when $\xi = 1$ and the degree of localization is linked with the order of magnitude of ξ , we would suggest a discretization set for ξ of the form

$$\left\{ \alpha^k, 1 \leq k \leq \frac{\log(N)}{\log(\alpha^{-1})} \right\}.$$

If we want to choose α as a function of N and still avoid introducing $\log(N)$ factors in the bound, we can for instance choose $\alpha = 1 - \frac{1}{\log(N)}$.

4. NOISY PATTERN RECOGNITION

The mathematical setting is the same as previously, and we assume without further notice that there exists a regular version of the conditional probability measure $P(Y|X)$. In this section, we are going to bring further improvements in the case when $\inf_{\theta \in \Theta} R(\theta) > 0$. This can result from various causes:

- The observed sample may be “noisy” in the sense that it is drawn according to a joint distribution P for which the best achievable error rate for pattern x , $\inf_{y \in \mathcal{Y}} P(Y \neq y | X = x)$ is large for many patterns. This noise may come either from an inherently ambiguous classification task or from errors made in labeling the training examples.
- Even if the sample is not noisy, the best available classification rule may be poor.

The theoretical bounds in the previous section were at best of order $\inf_{\theta} R(\theta) + \sqrt{\frac{R(\theta)}{N}} + \frac{c}{N}$, leading to a convergence speed not faster than $\frac{1}{\sqrt{N}}$ in the case of a noisy sample. We will improve this rate in the case when some classification rule $f_{\tilde{\theta}}$ produces the most likely label among all the available rules for a strong majority of patterns.

To formulate this we will consider some distinguished classification rule $f_{\tilde{\theta}}$. The most favorable case is when $R(\tilde{\theta}) = \inf_{\theta \in \Theta} R(\theta)$, but this condition will not be strictly imposed here. The case when $\tilde{\theta} \notin \Theta$ makes no difference : it is covered by adding $\tilde{\theta}$ to the parameter set Θ and extending the prior π putting $\pi(\tilde{\theta}) = 0$. Of course $\tilde{\theta}$, whose clever choice is bound to depend on P , is *not* assumed to be known by the statistician !

Let us introduce the following relative quantities, where Var_P denotes the variance with respect to P :

$$\begin{aligned} \overline{R}(\theta) &= P[Y \neq f_{\theta}(X)] - P[Y \neq f_{\tilde{\theta}}(X)] \\ \overline{r}(\theta) &= \frac{1}{N} \sum_{i=1}^N \mathbb{1}[Y_i \neq f_{\theta}(X_i)] - \mathbb{1}[Y_i \neq f_{\tilde{\theta}}(X_i)] \\ \overline{V}(\theta) &= \text{Var}_P \left\{ \mathbb{1}[Y \neq f_{\theta}(X)] - \mathbb{1}[Y \neq f_{\tilde{\theta}}(X)] \right\} \\ \overline{R}(\theta|X) &= P[Y \neq f_{\theta}(X) | X] - P[Y \neq f_{\tilde{\theta}}(X) | X] \\ \overline{V}(\theta|X) &= \text{Var}_P \left\{ \mathbb{1}[Y \neq f_{\theta}(X)] - \mathbb{1}[Y \neq f_{\tilde{\theta}}(X)] | X \right\}. \end{aligned}$$

Let us define for any pattern $x \in \mathcal{X}$ the margin $\alpha(x)$ of success of $f_{\tilde{\theta}}(x)$ as

$$\alpha(x) = \min \left\{ \overline{R}(\theta|x), \theta \in \Theta, f_{\theta}(x) \neq f_{\tilde{\theta}}(x) \right\}.$$

(In this formula we assume that some realization of the conditional expectations has been chosen once for all). Note that $\alpha(x)$ may be negative in the case when $f_{\tilde{\theta}}(x)$ is not the most likely label for pattern x .

Thresholding the margin $\alpha(x)$ at level α defines some exceptional set Ω_α of “ α -ambiguous” patterns :

$$\Omega_\alpha \stackrel{\text{def}}{=} \{x \in \mathcal{X} : \alpha(x) < \alpha\}.$$

We introduce this notion of ambiguity to control the variance $\overline{V}(\theta)$ by the mean $\overline{R}(\theta)$ of the relative error rate. Indeed

$$\begin{aligned} \overline{V}(\theta) &= P[\overline{V}(\theta|X)] + \text{Var}[\overline{R}(\theta|X)] \\ &\leq P\left[\frac{\overline{R}(\theta|X)}{\alpha}\mathbb{1}(X \notin \Omega_\alpha) + \mathbb{1}(\Omega_\alpha)\right] + P[\overline{R}(\theta|X)^2] \\ &\leq \frac{1}{\alpha}[\overline{R}(\theta) + P(\Omega_0)] + P(\Omega_\alpha) + \overline{R}(\theta) + 2P(\Omega_0) \\ &= a\overline{R}(\theta) + b, \end{aligned}$$

where we have put

$$\begin{aligned} a &= \left(\frac{1}{\alpha} + 1\right), \\ b &= \left(\frac{1}{\alpha} + 2\right)P(\Omega_0) + P(\Omega_\alpha). \end{aligned}$$

Applying Bernstein’s theorem 2.1 in a way similar to what has already been done to establish lemma 2.2 in the previous section, we get some non localized learning lemma

Lemma 4.1. *For any $\lambda \in \mathbb{R}_+$, any measurable function $\eta : \Theta \rightarrow \mathbb{R}$,*

$$\begin{aligned} P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1} \lambda \rho[\overline{R}(\theta)] - \lambda \rho[\overline{r}(\theta)] - \mathcal{K}(\rho, \pi) - \eta(\theta) \geq 0 \right\} \\ \leq \pi \left\{ \exp \left[g \left(\frac{[1 + \overline{R}(\theta)]\lambda}{N} \right) (a\overline{R}(\theta) + b) \frac{\lambda^2}{N} - \eta(\theta) \right] \right\}. \end{aligned}$$

In the same way

$$\begin{aligned} P^{\otimes N} \left\{ \sup_{\rho \in \mathcal{M}_+^1} \lambda \rho[\overline{r}(\theta)] - \lambda \rho[\overline{R}(\theta)] - \mathcal{K}(\rho, \pi) - \eta(\theta) \geq 0 \right\} \\ \leq \pi \left\{ \exp \left[g \left(\frac{[1 - \overline{R}(\theta)]\lambda}{N} \right) (a\overline{R}(\theta) + b) \frac{\lambda^2}{N} - \eta(\theta) \right] \right\}. \end{aligned}$$

4.1. Non localized results. Putting $\kappa = g\left(\frac{2\lambda}{N}\right)$ and taking

$$\eta(\theta) = \kappa[a\overline{R}(\theta) + b] \frac{\lambda^2}{N} + \log(\epsilon^{-1}),$$

we get

Corollary 4.1. *With $P^{\otimes N}$ probability at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$(17) \quad \rho[R(\theta)] \leq R(\tilde{\theta}) + \left(1 - \kappa a \frac{\lambda}{N}\right)^{-1} \left\{ \rho[r(\theta)] - r(\tilde{\theta}) + \frac{1}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}) \right] + \kappa b \frac{\lambda}{N} \right\}.$$

Note that the right-hand side of inequality (17) is not an observable quantity. Anyhow, it defers from an observable quantity by an additive term independent of ρ , thus it is still possible to optimize it in ρ from empirical observations. It is also possible to get an empirical bound for $\rho[R(\theta)] - \rho[R(\tilde{\theta})]$, the defect of optimality of the randomized estimator built from ρ , using the trivial bound $-r(\tilde{\theta}) \leq -\inf_{\theta \in \Theta} r(\theta)$. The optimal posterior for this bound is as before the Gibbs posterior $\hat{\rho}_\lambda$. It satisfies

Corollary 4.2. *With $P^{\otimes N}$ probability at least $1 - \epsilon$,*

$$(18) \quad \hat{\rho}_\lambda[R(\theta)] \leq R(\tilde{\theta}) + \left(1 - \kappa a \frac{\lambda}{N}\right)^{-1} \left\{ \frac{1}{\lambda} \int_0^\lambda \hat{\rho}_\beta[r(\theta)] d\beta - r(\tilde{\theta}) + \frac{\log(\epsilon^{-1})}{\lambda} + \kappa b \frac{\lambda}{N} \right\}.$$

Moreover

$$P^{\otimes N} \left\{ \hat{\rho}_\lambda[R(\theta)] \right\} \leq R(\tilde{\theta}) + \left(1 - \kappa a \frac{\lambda}{N}\right)^{-1} \left\{ \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\} - R(\tilde{\theta}) + \kappa b \frac{\lambda}{N} \right\}.$$

Note that in the case when $\pi(\{\tilde{\theta}\}) > 0$ (that is presumably when Θ is countable), and moreover when $b = 0$, we get

$$P^{\otimes N} \left\{ \hat{\rho}_\lambda[R(\theta)] \right\} \leq R(\tilde{\theta}) + \frac{\log[\pi(\{\tilde{\theta}\})^{-1}]}{\lambda \left(1 - \kappa a \frac{\lambda}{N}\right)}.$$

Choosing $\lambda = \frac{N}{2a}$ and noticing that for this value of λ , $\kappa = g\left(\frac{2\lambda}{N}\right) \leq g(0.5) \leq 1$, we get

$$P^{\otimes N} \left\{ \hat{\rho}_{\frac{N}{2a}}[R(\theta)] \right\} \leq R(\tilde{\theta}) + \frac{4a \log[\pi(\{\tilde{\theta}\})^{-1}]}{N}.$$

Therefore, we achieve a rate of convergence of $1/N$ whatever the order of magnitude of $R(\tilde{\theta})$ may be, as requested.

Moreover getting uniform results in λ can be achieved as explained before. Using a grid $\Lambda = \{N2^{-k} : 0 \leq k \leq \log(N)/\log(2)\}$, a union bound for this grid, and comparing values of $\lambda \in [1, N]$ with the next value in the grid, we get

Corollary 4.3. *With $P^{\otimes N}$ at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\rho[R(\theta)] \leq R(\tilde{\theta}) + \inf_{\lambda \in [1, N]} \left(1 - \kappa a \frac{2\lambda}{N}\right)^{-1} \left\{ \rho[r(\theta)] - r(\tilde{\theta}) + \frac{1}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{\log(N)}{2\epsilon}\right) \right] + \kappa b \frac{\lambda}{N} \right\}.$$

Note that to perform the optimization in λ from empirical data, we need first to apply the empirical bound $-r(\tilde{\theta}) \leq -\inf_{\theta \in \Theta} r(\theta)$.

4.2. Localized results. A localized learning lemma can be established exactly as explained in the previous section. It requires to choose

$$\eta(\theta) = \kappa[a\bar{R}(\theta) + b] \frac{\lambda^2}{N} + \beta\bar{R}(\theta) + \log\left\{\pi\left[\exp[-\beta\bar{R}(\theta)]\right]\right\} + \log(\epsilon^{-1}),$$

where $\kappa = g(\frac{2\lambda}{N})$.

Lemma 4.2. *With $P^{\otimes N}$ probability at least $1 - \epsilon$,*

$$\begin{aligned} \rho[\bar{R}(\theta)] &\leq \left(\lambda - \beta - \kappa a \frac{\lambda^2}{N}\right)^{-1} \left\{ \lambda\rho[\bar{r}(\theta)] + \mathcal{K}(\rho, \pi) \right. \\ &\quad \left. + \log\left\{\pi\left[\exp[-\beta\bar{R}(\theta)]\right]\right\} + \log(\epsilon^{-1}) + \kappa b \frac{\lambda^2}{N} \right\}. \end{aligned}$$

In the same way,

$$\begin{aligned} -\rho[\bar{R}(\theta)] &\leq \left(\lambda + \beta + \kappa a \frac{\lambda^2}{N}\right)^{-1} \left\{ -\lambda\rho[\bar{r}(\theta)] + \mathcal{K}(\rho, \pi) \right. \\ &\quad \left. + \log\left\{\pi\left[\exp[-\beta\bar{R}(\theta)]\right]\right\} + \log(\epsilon^{-1}) + \kappa b \frac{\lambda^2}{N} \right\}. \end{aligned}$$

Putting $\xi = \frac{\beta}{\lambda(1+a\kappa\frac{\lambda^2}{N})}$, we see that with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$\begin{aligned} \log\left\{\pi\left[\exp[-\beta\bar{R}(\theta)]\right]\right\} &= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[-\beta\rho[\bar{R}(\theta)] - \mathcal{K}(\rho, \pi) \right] \\ &\leq \sup_{\rho \in \mathcal{M}_+^1} \frac{\xi}{1 + \xi} \left\{ -\lambda\rho[\bar{r}(\theta)] + \mathcal{K}(\rho, \pi) \right. \\ &\quad \left. + \log\left\{\pi\left[\exp[-\beta\bar{R}(\theta)]\right]\right\} + \log(\epsilon^{-1}) + b\kappa \frac{\lambda^2}{N} \right\} \\ &\quad - \mathcal{K}(\rho, \pi), \end{aligned}$$

which can be rewritten as

$$\begin{aligned} \log\left\{\pi\left[\exp[-\beta\bar{R}(\theta)]\right]\right\} &\leq \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[-\xi\lambda\bar{r}(\theta) - \mathcal{K}(\rho, \pi) \right] + \xi \log(\epsilon^{-1}) + \xi b\kappa \frac{\lambda^2}{N} \\ &= \log\left\{\pi\left[\exp[-\xi\lambda\bar{r}(\theta)]\right]\right\} + \xi \log(\epsilon^{-1}) + \xi b\kappa \frac{\lambda^2}{N}. \end{aligned}$$

Coming back to lemma 4.2, we obtain

Corollary 4.4. *With $P^{\otimes N}$ probability at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_+^1$,*

$$\begin{aligned} \rho[R(\theta)] - R(\tilde{\theta}) &\leq \left(1 - \frac{1 + \xi}{1 - \xi} \kappa a \frac{\lambda}{N}\right)^{-1} \left\{ \rho[r(\theta)] - r(\tilde{\theta}) \right. \\ &\quad \left. + \frac{1}{(1 - \xi)\lambda} \left[\mathcal{K}(\rho, \hat{\rho}_{\xi\lambda}) + (1 + \xi) \log\left(\frac{2}{\epsilon}\right) \right] + \kappa \frac{1 + \xi}{1 - \xi} b \frac{\lambda}{N} \right\} \\ &= \left(1 - \frac{1 + \xi}{1 - \xi} \kappa a \frac{\lambda}{N}\right)^{-1} \left\{ \frac{1}{(1 - \xi)\lambda} \left[\lambda\rho[r(\theta)] + \mathcal{K}(\rho, \pi) + \log\left\{\pi\left[\exp[-\xi\lambda r(\theta)]\right]\right\} \right] \right. \\ &\quad \left. - r(\tilde{\theta}) + \frac{1 + \xi}{1 - \xi} \left[\frac{\log\left(\frac{2}{\epsilon}\right)}{\lambda} + \kappa b \frac{\lambda}{N} \right] \right\}. \end{aligned}$$

The optimal posterior according to this bound is the Gibbs distribution $\hat{\rho}_\lambda$. It is such that

$$\hat{\rho}_\lambda[R(\theta)] - R(\tilde{\theta}) \leq \left(1 - \frac{1+\xi}{1-\xi} \kappa a \frac{\lambda}{N}\right)^{-1} \left\{ \underbrace{\frac{1}{1-\xi} \int_\xi^1 \hat{\rho}_{\beta\lambda}[r(\theta)] d\beta}_{\leq \hat{\rho}_{\epsilon\lambda}[r(\theta)]} - r(\tilde{\theta}) + \frac{1+\xi}{1-\xi} \left[\frac{\log(\frac{2}{\epsilon})}{\lambda} + \kappa b \frac{\lambda}{N} \right] \right\}.$$

For a fixed value of ξ , getting a uniform result in λ is achieved as in the case of corollary 3.3:

Corollary 4.5. For any $\xi \in [0, 1]$, any $\zeta \in [\xi, 1]$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\lambda \in [1, N]$,

$$\hat{\rho}_\lambda[R(\theta)] - R(\tilde{\theta}) \leq \left(1 - \frac{1+\xi}{1-\xi} \kappa a \frac{\lambda}{\zeta N}\right)^{-1} \left\{ \underbrace{\frac{1}{1-\xi} \int_\xi^1 \hat{\rho}_{\beta\lambda}[r(\theta)] d\beta}_{\leq \hat{\rho}_{\epsilon\lambda}[r(\theta)]} - r(\tilde{\theta}) + \frac{1+\xi}{1-\xi} \left[\frac{1}{\lambda} \log \left(\frac{2 \log(N)}{\epsilon \log(\zeta^{-1})} \right) + \kappa b \frac{\lambda}{\zeta N} \right] \right\}$$

The same remarks which were made about corollary 3.3 apply here : a union bound on different values of ξ can furthermore be performed. Let us also notice that optimizing the bound in λ from observations requires to use the empirical bound $-r(\tilde{\theta}) \leq -\inf_{\theta \in \Theta} r(\theta)$.

5. LEARNING WITH AN EXCHANGEABLE PRIOR

In this section we assume that P_{2N} is some exchangeable distribution on $(\mathcal{X} \times \mathcal{Y})^{2N}$, where $(\mathcal{X}, \mathcal{B})$ is as previously a measurable space of patterns and \mathcal{Y} a finite set of labels. We assume that we observe (X_1, \dots, X_N) , (Y_1, \dots, Y_N) and possibly also (X_{N+1}, \dots, X_{2N}) . In other words half of the patterns are labeled and half of the patterns have to be labeled. Starting with a family $\{f_\theta : \mathcal{X} \rightarrow \mathcal{Y}; \theta \in \Theta\}$ of classification rules, we would like to minimize

$$r_2(\theta) = \frac{1}{N} \sum_{k=N+1}^{2N} \mathbb{1}(Y_k \neq f_\theta(X_k)).$$

We can apply our PAC-Bayesian methodology in this situation, using an exchangeable prior. The interest of exchangeable priors is that they will provide a way to make a link between PAC-Bayesian theorems and Vapnik's theory.

Let us first prove a deviation lemma based on the fact that P_{2N} is exchangeable. Let

$$r_1(\theta) = \frac{1}{N} \sum_{k=1}^N \mathbb{1}[Y_k \neq f_\theta(X_k)]$$

$$r_2(\theta) = \frac{1}{N} \sum_{k=N+1}^{2N} \mathbb{1}[Y_k \neq f_\theta(X_k)]$$

Lemma 5.1. *For any exchangeable measurable function $\eta : (\mathcal{X} \times \mathcal{Y})^{2N} \times \Theta \rightarrow \mathbb{R}$, any $\theta \in \Theta$,*

$$P_{2N} \left\{ \exp[\lambda[r_2(\theta) - r_1(\theta)] - \eta(\theta)] \right\} \leq P_{2N} \left\{ \exp\left[\frac{\lambda^2}{2N}[r_1(\theta) + r_2(\theta)] - \eta(\theta)\right] \right\}.$$

Proof. Let us remember that $\log[\cosh(s)] \leq \frac{1}{2}s^2$ for any $s \in \mathbb{R}$. Let

$$\sigma_k = \mathbf{1}[Y_k \neq f_\theta(X_k)]$$

Using the fact that P_{2N} is assumed to be exchangeable, we get

$$\begin{aligned} & P_{2N} \left\{ \exp[\lambda[r_2(\theta) - r_1(\theta)] - \eta(\theta)] \right\} \\ &= P_{2N} \left\{ \exp\left[\frac{\lambda}{N} \sum_{k=1}^N [\sigma_{k+N}(\theta) - \sigma_k(\theta)] - \eta(\theta)\right] \right\} \\ &= P_{2N} \left\{ \exp\left[\sum_{k=1}^N \log\left\{\cosh\left[\frac{\lambda}{N}[\sigma_{k+N}(\theta) - \sigma_k(\theta)]\right]\right\} - \eta(\theta)\right] \right\} \\ &\leq P_{2N} \left\{ \exp\left[\frac{\lambda^2}{2N^2} \sum_{k=1}^N [\sigma_{k+N}(\theta) - \sigma_k(\theta)]^2 - \eta(\theta)\right] \right\} \\ &\leq P_{2N} \left\{ \exp\left[\frac{\lambda^2}{2N^2} \sum_{k=1}^N [\sigma_{k+N}(\theta) + \sigma_k(\theta)] - \eta(\theta)\right] \right\} \\ &= P_{2N} \left\{ \exp\left[\frac{\lambda^2}{2N}[r_1(\theta) + r_2(\theta)] - \eta(\theta)\right] \right\} \end{aligned}$$

□

Let us now consider some exchangeable random probability measure $\pi : (\mathcal{X} \times \mathcal{Y})^{2N} \rightarrow \mathcal{M}_+^1(\Theta)$. (We will assume that (Θ, \mathcal{T}) is a Polish space and that π is a regular conditional probability measure. Moreover, in practice, interesting exchangeable priors will depend only on (X_1, \dots, X_{2N}) , although the forthcoming bounds do not preclude them to depend also on (Y_1, \dots, Y_{2N}) .) Integrating the previous deviation lemma with respect to π , we get

Lemma 5.2. *For any $\lambda \in \mathbb{R}_+$,*

$$\begin{aligned} & P_{2N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[r_2(\theta)] - \lambda \rho[r_1(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \theta) \geq 0 \right\} \\ &\leq P_{2N} \left\{ \pi \left[\exp\left\{\frac{\lambda^2}{2N}[r_1(\theta) + r_2(\theta)] - \eta(\theta)\right\} \right] \right\}. \end{aligned}$$

As in the preceding section, we can deduce from this lemma non-localized or localized results. Let us start with a non localized result.

Choosing $\eta(\theta) = \frac{\lambda^2}{2N}[r_1(\theta) + r_2(\theta)] + \log(\epsilon^{-1})$ we obtain

Corollary 5.1. *With P^{2N} probability at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\rho[r_2(\theta)] \leq \left(\lambda - \frac{\lambda^2}{2N}\right)^{-1} \left\{ \left(\lambda + \frac{\lambda^2}{2N}\right) \rho[r_1(\theta)] + \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}) \right\}.$$

As a special case, we find a result similar to Vapnik's bounds. Let

$$N(X_1^{2N}) = |\{ [f_\theta(X_k)]_{k=1}^{2N} : \theta \in \Theta \}|.$$

Corollary 5.2. *With P_{2N} probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$r_2(\theta) \leq \left(\lambda - \frac{\lambda^2}{2N} \right)^{-1} \left\{ \left(\lambda + \frac{\lambda^2}{2N} \right) r_1(\theta) + \log[N(X_1^{2N})] + \log(\epsilon^{-1}) \right\}.$$

Note that this is an improvement on classical Vapnik's theory, since the complexity term $\log[N(X_1^{2N})]$ is observable. Note also that in the binary case when $|\mathcal{Y}| = 2$,

$$\log[N(X_1^{2N})] \leq 2NH\left(\frac{h}{2N}\right) \leq h[\log\left(\frac{2N}{h}\right) + 1],$$

where $H(p) = -p \log(p) - (1-p) \log(1-p)$ is the Shannon entropy of the Bernoulli distribution with parameter p and where

$$h = \max\{|A| : A \subset \{X_k : 1 \leq k \leq 2N\} \text{ and } |\{A \cap f_\theta^{-1}(1) : \theta \in \Theta\}| = 2^{|A|}\}$$

is the VC dimension of the set $\{X_1, \dots, X_{2N}\}$.

Proof. Let

$$\Psi : \theta \mapsto [f_\theta(X_k)]_{k=1}^{2N} \in \mathcal{Y}^{2N}.$$

For each $y \in \Psi(\Theta)$, let us choose $\theta(y) \in \Psi^{-1}(y)$ to form a finite set $\Theta' \subset \Theta$ of size $N(X_1^{2N})$, as the collection $\{\Psi^{-1}(y) : y \in \Psi(\Theta)\}$ is an exchangeable function of X_1^{2N} , the random set $\Theta'(X_1^{2N})$ can be chosen to be an exchangeable function of X_1^{2N} . Let π be the uniform measure on Θ' . Then considering as posterior distribution the Dirac mass at $\theta' \in \Theta'$, we see that with P_{2N} probability at least $1 - \epsilon$, for any $\theta' \in \Theta'$,

$$r_2(\theta') \leq \left(\lambda - \frac{\lambda^2}{2N} \right)^{-1} \left\{ \left(\lambda + \frac{\lambda^2}{2N} \right) r_1(\theta') + \log[N(X_1^{2N})] + \log(\epsilon^{-1}) \right\}.$$

We end the proof with the remark that for any $\theta \in \Theta$, there is $\theta' \in \Theta'$ such that $\Psi(\theta) = \Psi(\theta')$, and therefore such that $r_1(\theta) = r_1(\theta')$ and $r_2(\theta) = r_2(\theta')$. \square

It also makes sense to compare $r_1(\theta)$ with

$$R_2(\theta) = P_{2N}[Y_{N+1} \neq f_\theta(X_{N+1}) | Z_1^N],$$

where we have put $Z_1^N = (X_k, Y_k)_{k=1}^N$ for short.

To this purpose we can use a variant of lemma 5.2

Lemma 5.3. *For any $\lambda \in \mathbb{R}_+$,*

$$P_{2N} \left\{ P_{2N} \left[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[r_2(\theta)] - \lambda \rho[r_1(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi) | Z_1^N \right] \geq 0 \right\} \leq P_{2N} \left\{ \pi \left[\exp \left\{ \frac{\lambda^2}{2N} [r_1(\theta) + r_2(\theta)] - \eta(\theta) \right\} \right] \right\}.$$

Proof. Let

$$U = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[r_2(\theta)] - \lambda \rho[r_1(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi)$$

and

$$\epsilon = P_{2N} \left\{ \pi \left[\exp \left[\frac{\lambda^2}{2N} [r_1(\theta) + r_2(\theta)] - \eta(\theta) \right] \right] \right\}.$$

Then as already proved, $P_{2N}[\exp(U)] \leq \epsilon$. But in the same time, from the convexity of the exponential function,

$$P_{2N}\left\{\exp\left[P_{2N}(U | Z_1^N)\right]\right\} \leq P_{2N}[\exp(U)],$$

as required. \square

Choosing in lemma 5.3

$$\eta(\theta) = \frac{\lambda^2}{2N} [r_1(\theta) + r_2(\theta)] + \log(\epsilon^{-1})$$

we get

Corollary 5.3. *With $P_{2N}(dZ_1^N)$ (the distribution of Z_1^N under P_{2N}) probability at least $1 - \epsilon$, for any regular conditional probability distribution $\rho : \mathcal{X}^{2N} \times \mathcal{Y}^{2N} \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$P_{2N}\left\{\rho[r_2(\theta)] | Z_1^N\right\} \leq \left(\lambda - \frac{\lambda^2}{2N}\right)^{-1} \left\{ \left(\lambda + \frac{\lambda^2}{2N}\right) P_{2N}\left\{\rho[r_1(\theta)] | Z_1^N\right\} + P_{2N}\left\{\mathcal{K}(\rho, \pi) | Z_1^N\right\} + \log(\epsilon^{-1}) \right\}.$$

The interesting case is of course when ρ in fact does not depend on the non observed labels Y_{N+1}^{2N} . This may look cumbersome, but has a simple application.

Theorem 5.1. *With $P_{2N}(dZ_1^N)$ probability at least $1 - \epsilon$, for any estimator $\hat{\theta} : \mathcal{X}^N \times \mathcal{Y}^N \rightarrow \Theta$, assumed to be a measurable function,*

$$R_2(\hat{\theta}) \leq \left(\lambda - \frac{\lambda^2}{2N}\right)^{-1} \left\{ \left(\lambda + \frac{\lambda^2}{2N}\right) r_1(\hat{\theta}) + P_{2N}\left\{\log[N(X_1^{2N})] | Z_1^N\right\} + \log(\epsilon^{-1}) \right\}.$$

Note that in the independent case when $P_{2N} = P^{\otimes 2N}$, then $R_2(\hat{\theta}) = R(\hat{\theta})$ defined in the previous sections.

Proof. This is an integrated variant of corollary 5.2. With the same notations, we can define an estimator $\hat{\theta}'$ with values in Θ' , such that $\Psi(\hat{\theta}) = \Psi(\hat{\theta}')$ everywhere. The proof for $\hat{\theta}'$ is a direct consequence of the preceding corollary, and the proof for $\hat{\theta}$ comes from the fact that $r_1(\hat{\theta}) = r_1(\hat{\theta}')$ and $r_2(\hat{\theta}) = r_2(\hat{\theta}')$ everywhere. \square

Proving some variant of Vapnik's theory is not the only possible use of corollary 5.1. Its right-hand side can also be optimized choosing for ρ the Gibbs distribution

$$d\hat{\rho}_\beta(\theta) = \frac{\exp[-\beta r_1(\theta)]}{\pi\left\{\exp[-\beta r_1(\theta)]\right\}} d\pi(\theta).$$

(Note that $\hat{\rho}$ depends not only on Z_1^N but also on X_{N+1}^{2N} through π .)

Corollary 5.4. *With P_{2N} probability at least $1 - \epsilon$,*

$$\begin{aligned} \hat{\rho}_{\lambda + \frac{\lambda^2}{2N}}[r_2(\theta)] &\leq \left(\lambda - \frac{\lambda^2}{2N}\right)^{-1} \left\{ -\log\left[\pi\left\{\exp\left[-\left(\lambda + \frac{\lambda^2}{2N}\right) r_1(\theta)\right]\right\}\right] + \log(\epsilon^{-1}) \right\} \\ &= \frac{1 + \frac{\lambda}{2N}}{1 - \frac{\lambda}{2N}} \left\{ \frac{1}{\lambda + \frac{\lambda^2}{2N}} \int_0^{\lambda + \frac{\lambda^2}{2N}} \hat{\rho}_\beta[r_1(\theta)] d\beta \right\} + \frac{\log(\epsilon^{-1})}{\lambda - \frac{\lambda^2}{2N}}. \end{aligned}$$

5.1. Some possible applications of learning with an exchangeable prior.

Before getting into more sophisticated bounds (localized or tailored for the noisy classification case), let us put forward that the choice of π as a function of $\sum_{k=1}^{2N} \delta_{X_k}$ opens interesting possibilities.

One example is to control the upper deviations of the error rate of a support vector machine as a function of the number of support vectors. There are indeed at most $\binom{2N}{s} (2^s - 2)$ hyperplanes with s support vectors. Therefore if we restrict to this class, by choosing π as the uniform probability measure on it, which is indeed exchangeable if the support vectors are drawn from (X_1, \dots, X_{2N}) , we see that $\log[N(X_1^{2N})] \leq s[\log(\frac{2N}{s}) + 1 + \log(2)]$. More generally, following the ideas of compression schemes put forward by Littlestone and Warmuth [19, 17], we obtain generalization bounds for any classification rule which depends only on a restricted number of examples. More precisely, imagine that we have some classification estimator which provides for any set of labeled examples $(X', Y') = (X'_1, Y'_1), \dots, (X'_s, Y'_s)$ some classification rule $\hat{f}_{(X', Y')} : \mathcal{X} \rightarrow \{0, 1\}$. Then, using an exchangeable prior, we can restrict to those classification rules built from X' drawn from (X_1, \dots, X_{2N}) and Y' arbitrary, and be sure that corollary 5.2 (as well as all the improvements we can make to it : localization, working with a Gibbs posterior etc.) will apply with $\log[N(X_1^{2N})] \leq s[\log(\frac{2N}{s}) + 1 + \log(2)]$. Indeed in the definition of $N(X_1^{2N})$, the parameter space Θ can be replaced with the support of π , i.e. any random set Θ' such that $\pi(\Theta') = 1$.

Note also that simulating the Gibbs posterior in such a setting is quite straightforward : we can use the Metropolis algorithm (see [16] for more details) and move the coordinates of X' and Y' one at a time. Note also that this learning scheme is different from cross validation, since, although we should restrict ourselves to choosing \hat{f} as a function of (X', Y') only, *we are allowed to choose (X', Y') as a function of the observed sample $(X_1, \dots, X_N), (Y_1, \dots, Y_N)$, and also if we wish of (X_{N+1}, \dots, X_{2N}) in any way we may think suitable.*

One possible use of this setting is to choose adaptively a (pruned) decision tree: given a set of questions (q_1, q_2, \dots, q_n) and a small set of (hopefully “typical”) patterns (X'_1, \dots, X'_s) drawn from (X_1, \dots, X_{2N}) , we may build a pruned decision tree by stopping to ask questions as soon as only one example in X' matches the query. Using corollary 5.2 in this context leads to penalize the risk with a penalty proportional to the number of nodes, something we could have achieved through a different approach (like considering a Galton Watson process as the prior on trees). But we can do better : we can also prune inner nodes by deciding to remove questions which do not split X' , and we can think about more clever strategies to choose the questions to be asked and the order in which they should be asked as a function of our “typical” set X' (for instance we can choose a set of questions leading to a balanced tree). We can also use the labels Y' to prune the tree and select questions: indeed we can choose the decision tree in any way we like, as long as we build it in a unique way as a function of (X', Y') only. Then we can compare the performance of the obtained classifiers on the whole training sample $(X_1, Y_1, \dots, X_N, Y_N)$ and retain the best typical compression set (X', Y') . All these adaptations to data seem difficult to perform in a theoretically justified context (i.e. in a context where we can derive explicit generalization bounds) using another approach (at least to our knowledge).

5.2. Localization. We can localize our results for exchangeable priors as we had done in previously encountered situations. To achieve this, let us apply lemma 5.2 with

$$\eta(\theta) = \left(\frac{\lambda^2}{2N} + \beta \right) [r_1(\theta) + r_2(\theta)] + \log \left\{ \pi \left[\exp \left[-\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\} + \log(\epsilon^{-1}).$$

We get

Lemma 5.4. *With P_{2N} probability at least $1 - \epsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned} \left(\lambda - \beta - \frac{\lambda^2}{2N} \right) \rho[r_2(\theta)] &\leq \left(\lambda + \beta + \frac{\lambda^2}{2N} \right) \rho[r_1(\theta)] \\ &\quad + \log \left\{ \pi \left[\exp \left[-\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\} + \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}). \end{aligned}$$

Moreover we get with P_{2N} probability at least $1 - \epsilon$,

$$\begin{aligned} \log \left\{ \pi \left[\exp \left[-\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\} &= \sup_{\rho \in \mathcal{M}_+^1} -\beta \rho[r_1(\theta)] - \beta \rho[r_2(\theta)] - \mathcal{K}(\rho, \pi) \\ &\leq \sup_{\rho \in \mathcal{M}_+^1(\Theta)} -\beta \rho[r_1(\theta)] - \mathcal{K}(\rho, \pi) - \beta \left(\lambda + \beta + \frac{\lambda^2}{2N} \right)^{-1} \left\{ \left(\lambda - \beta - \frac{\lambda^2}{2N} \right) \rho[r_1(\theta)] \right. \\ &\quad \left. - \log \left\{ \pi \left[\exp \left\{ -\beta [r_1(\theta) + r_2(\theta)] \right\} \right] \right\} - \mathcal{K}(\rho, \pi) - \log(\epsilon^{-1}) \right\}. \end{aligned}$$

Putting $\xi = \frac{\beta}{\lambda + \frac{\lambda^2}{2N}}$, this can also be written as

$$\begin{aligned} (19) \quad \log \left\{ \pi \left[\exp \left[-\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\} \\ \leq \sup_{\rho \in \mathcal{M}_+^1(\Theta)} -2\xi \lambda \rho[r_1(\theta)] - \mathcal{K}(\rho, \pi) + \xi \log(\epsilon^{-1}) \\ = \log \left\{ \pi \left[\exp \left[-2\xi \lambda r_1(\theta) \right] \right] \right\} + \xi \log(\epsilon^{-1}). \end{aligned}$$

This leads to

Lemma 5.5. *With P_{2N} probability at least $1 - \epsilon$, for any $\rho \in \mathcal{M}_+^1$,*

$$\begin{aligned} \rho[r_2(\theta)] &\leq \left[(1 - \xi)\lambda - (1 + \xi)\frac{\lambda^2}{2N} \right]^{-1} \left\{ (1 + \xi)\lambda \left(1 + \frac{\lambda}{2N} \right) \rho[r_1(\theta)] \right. \\ &\quad \left. + \mathcal{K}(\rho, \pi) + \log \left\{ \pi \left[\exp \left[-2\xi \lambda r_1(\theta) \right] \right] \right\} + (1 + \xi) \log\left(\frac{2}{\epsilon}\right) \right\} \\ &= \left[(1 - \xi)\lambda - (1 + \xi)\frac{\lambda^2}{2N} \right]^{-1} \left\{ \left[(1 - \xi)\lambda + (1 + \xi)\frac{\lambda^2}{2N} \right] \rho[r_1(\theta)] \right. \\ &\quad \left. + \mathcal{K}(\rho, \hat{\rho}_{2\xi\lambda}) + (1 + \xi) \log\left(\frac{2}{\epsilon}\right) \right\}. \end{aligned}$$

Here again the right-hand side is minimized by a Gibbs distribution, leading to

Corollary 5.5. *With P_{2N} probability at least $1 - \epsilon$,*

$$\begin{aligned} \hat{\rho}_{(1+\xi)\lambda(1+\frac{\lambda}{2N})}[r_2(\theta)] &\leq \left[(1-\xi)\lambda - (1+\xi)\frac{\lambda^2}{2N} \right]^{-1} \left\{ \int_{2\xi\lambda}^{(1+\xi)\lambda(1+\frac{\lambda}{2N})} \hat{\rho}_\beta[r_1(\theta)] d\beta \right. \\ &\quad \left. + (1+\xi)\log\left(\frac{2}{\epsilon}\right) \right\} \\ &\leq \left[(1-\xi)\lambda - (1+\xi)\frac{\lambda^2}{2N} \right]^{-1} \left\{ \left[(1-\xi)\lambda + (1+\xi)\frac{\lambda^2}{2N} \right] \hat{\rho}_{2\xi\lambda}[r_1(\theta)] \right. \\ &\quad \left. + (1+\xi)\log\left(\frac{2}{\epsilon}\right) \right\}. \end{aligned}$$

6. NOISY CLASSIFICATION WITH AN EXCHANGEABLE PRIOR

As in the case of deterministic priors treated before, we can derive bounds relative to a given reference classification rule which are sharper in the presence of noise. We will assume here that the distribution of patterns and labels is i.i.d. and therefore consider a product distribution $P^{\otimes 2N}$ on $((\mathcal{X} \times \mathcal{Y})^{2N}, (\mathcal{B} \otimes \mathcal{B}')^{\otimes 2N})$. Similarly to what has been done in section 4 we consider some fixed (and unknown) parameter $\tilde{\theta} \in \Theta$ and define

$$\begin{aligned} \sigma_k(\theta) &= \mathbb{1}[Y_k \neq f_\theta(X_k)] \\ \bar{r}_1(\theta) &= \frac{1}{N} \sum_{k=1}^N \sigma_k(\theta) - \sigma_k(\tilde{\theta}) \\ \bar{r}_2(\theta) &= \frac{1}{N} \sum_{k=N+1}^{2N} \sigma_k(\theta) - \sigma_k(\tilde{\theta}) \\ \bar{R}(\theta | X_k) &= P[\sigma_k(\theta) - \sigma_k(\tilde{\theta}) | X_k] \\ r'_1(\theta) &= \frac{1}{N} \sum_{k=1}^N \bar{R}(\theta | X_k), \\ r'_2(\theta) &= \frac{1}{N} \sum_{k=N+1}^{2N} \bar{R}(\theta | X_k). \end{aligned}$$

For the sake of simplicity, we will assume that the rule $f_{\tilde{\theta}}$ clearly outperforms the other rules for any pattern, in the sense that for some constant $\alpha > 0$ which will stay fixed in the remaining of this discussion, for any $x \in \mathcal{X}$,

$$\alpha(x) = \min\{\bar{R}(\theta | x), \theta \in \Theta, f_\theta(x) \neq f_{\tilde{\theta}}(x)\} \geq \alpha.$$

Let us consider two real numbers $\beta > \lambda > 0$ and put for short $\kappa = \frac{1}{\alpha}g(\frac{2\beta}{N})$. The following exponential inequalities will be helpful:

$$\begin{aligned} P^{\otimes 2N} \left\{ \exp[\lambda \bar{r}_2(\theta) - \beta \bar{r}_1(\theta)] | X_1^{2N} \right\} \\ \leq P^{\otimes 2N} \left\{ \exp\left[(\lambda + \kappa \frac{\lambda^2}{N}) r'_2(\theta) - (\beta - \kappa \frac{\beta^2}{N}) r'_1(\theta) \right] | X_1^{2N} \right\}. \end{aligned}$$

Moreover, putting

$$\begin{aligned}\lambda' &= \lambda + \kappa \frac{\lambda^2}{N} \\ \beta' &= \beta - \kappa \frac{\beta^2}{N},\end{aligned}$$

$$\begin{aligned}P^{\otimes 2N} \left\{ \exp \left[\lambda' r'_2(\theta) - \beta' r'_1(\theta) \right] \left| \sum_{k=1}^{2N} \delta_{X_k} \right. \right\} \\ \leq P^{\otimes 2N} \left\{ \exp \left\{ \left[\frac{1}{2N} (\lambda' + \beta')^2 - \frac{\beta' - \lambda'}{2} \right] [r'_1(\theta) + r'_2(\theta)] \right\} \left| \sum_{k=1}^{2N} \delta_{X_k} \right. \right\}.\end{aligned}$$

Integrating these inequalities with respect to a random exchangeable prior distribution $\pi : (\mathcal{X}^{2N}, \mathcal{B}^{\otimes 2N}) \rightarrow \mathcal{M}_+^1(\Theta)$ we get

Lemma 6.1. *With $P^{\otimes 2N}$ probability at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}\lambda \rho[\bar{r}_2(\theta)] &\leq \beta \rho[\bar{r}_1(\theta)] + \mathcal{K}(\rho, \pi) \\ &+ \log \left\{ \pi \left[\exp \left\{ - \underbrace{\left[\left(\frac{\beta' - \lambda'}{2} - \frac{1}{2N} (\frac{\beta' + \lambda'}{2})^2 \right) [r'_1(\theta) + r'_2(\theta)] \right]}_{\stackrel{\text{def}}{=} \beta''} \right\} \right] \right\} + \log(\epsilon^{-1}).\end{aligned}$$

Moreover, with $P^{\otimes 2N}$ probability at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}\lambda' \rho[r'_2(\theta)] &\leq \beta' \rho[r'_1(\theta)] + \mathcal{K}(\rho, \pi) \\ &+ \log \left\{ \pi \left[\pi \left\{ \exp \left[-\beta'' [r'_1(\theta) + r'_2(\theta)] \right] \right\} \right] \right\} + \log(\epsilon^{-1}),\end{aligned}$$

(where β'' is defined in the previous equation).

6.1. Non localized bound. To get a non localized learning theorem, we can choose for some parameter μ

$$\begin{aligned}\lambda' &= \mu - \frac{1}{2N} \mu^2 = \lambda + \kappa \frac{\lambda^2}{N}, \\ \beta' &= \mu + \frac{1}{2N} \mu^2 = \beta - \kappa \frac{\beta^2}{N},\end{aligned}$$

and take advantage of the fact that $r'_1(\theta)$ and $r'_2(\theta)$ are all positive random variables (since we assumed that $\tilde{\theta}$ was everywhere optimal).

Theorem 6.1. *With $P^{\otimes 2N}$ probability at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}\rho[r_2(\theta)] &\leq r_2(\tilde{\theta}) + \frac{\mu + \frac{\mu^2}{2N} + \kappa \frac{\beta^2}{N}}{\mu - \frac{\mu^2}{2N} - \kappa \frac{\lambda^2}{N}} \left[\rho[r_1(\theta)] - r_1(\tilde{\theta}) \right] \\ &+ \frac{1}{\mu - \frac{\mu^2}{2N} - \kappa \frac{\lambda^2}{N}} \left\{ \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}) \right\}.\end{aligned}$$

6.2. Localized bound. To get a localized learning theorem, we need an upper bound for

$$\log \left\{ \pi \left\{ \exp \left[-\beta'' [r'_1(\theta) + r'_2(\theta)] \right] \right\} \right\}.$$

We will achieve this in two steps. The first one is similar to the low noise case with an exchangeable prior, and compares the above quantity with $\log \{ \pi [\exp[-\gamma[r'_1(\theta)]] \}$ for a suitable choice of γ . Let us put

$$\begin{aligned} \gamma' &= \beta'' \frac{\beta' + \lambda'}{\beta' - \beta''} = (\beta' - \lambda') \frac{1 - \frac{1}{N} \frac{(\lambda' + \beta')^2}{4(\beta' - \lambda')}}{1 + \frac{1}{2N} \frac{\lambda' + \beta'}{2}} \\ \xi &= \frac{\beta''}{\beta' - \beta''} = \frac{\beta' - \lambda'}{\beta' + \lambda'} \frac{1 - \frac{1}{N} \frac{(\lambda' + \beta')^2}{4(\beta' - \lambda')}}{1 + \frac{1}{N} \frac{\lambda' + \beta'}{4}} \leq \frac{\beta' - \lambda'}{\beta' + \lambda'}. \end{aligned}$$

The same computation that led to (19) shows that

Lemma 6.2. *With $P^{\otimes 2N}$ probability at least $1 - \epsilon$,*

$$\log \left\{ \pi \left[\exp \left\{ -\beta'' [r'_1(\theta) + r'_2(\theta)] \right\} \right] \right\} \leq \log \left\{ \pi \left[\exp \left\{ -\gamma' [r'_1(\theta)] \right\} \right] \right\} + \xi \log(\epsilon^{-1}).$$

Now we need to compare $\log \{ \pi [\exp[-\gamma' r'_1(\theta)]] \}$ with $\log \{ \pi [\exp[-\gamma \bar{r}_1(\theta)]] \}$ for some suitable value of γ . To achieve this, we use another learning lemma, derived from the inequality

$$P^{\otimes 2N} \left\{ \exp \left[\lambda [\bar{r}_1(\theta) - r'_1(\theta)] - \mu r'_1(\theta) \right] \mid X_1^N \right\} \leq \exp \left[\left(\kappa \frac{\lambda^2}{N} - \mu \right) r'_1(\theta) \right].$$

Lemma 6.3. *With $P^{\otimes N}$ probability at least $1 - \epsilon$, for any posterior probability distribution $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned} \lambda \rho[\bar{r}_1(\theta)] &\leq \left(\lambda + \gamma' + \kappa \frac{\lambda^2}{N} \right) \rho[r'_1(\theta)] + \log \left\{ \pi \left[\exp[-\gamma' r'_1(\theta)] \right] \right\} \\ &\quad + \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}). \end{aligned}$$

Exactly as we derived (16), we can establish that with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$\log \left\{ \pi \left\{ \exp[-\gamma' r'_1(\theta)] \right\} \right\} \leq \log \left\{ \pi \left\{ \exp \left[-\frac{\gamma'}{1 + \kappa \frac{\lambda^2}{N}} \bar{r}_1(\theta) \right] \right\} \right\} + \frac{\gamma'}{\lambda + \kappa \frac{\lambda^2}{N}} \log(\epsilon^{-1}).$$

Putting all these things together leads to a localized learning theorem for noisy classification using an exchangeable prior. Let us put

$$\zeta = \frac{\left(1 - \kappa \frac{\lambda^2 + \beta^2}{N(\beta - \lambda)} \right) \left(1 - \frac{(\lambda' + \beta')^2}{4N(\beta' - \lambda')} \right)}{\left(1 - \kappa \frac{\lambda}{N} \right) \left(1 + \frac{\lambda' + \beta'}{4N} \right)}.$$

Theorem 6.2. *With the notations introduced in this section, with $P^{\otimes 2N}$ probability at least $1 - \epsilon$, for any posterior distribution $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned} \rho[r_2(\theta)] &\leq r_2(\tilde{\theta}) + \frac{\beta}{\lambda} [\rho[r_1(\theta)] - r_1(\tilde{\theta})] \\ &+ \frac{1}{\lambda} \left\{ \mathcal{K}(\rho, \pi) + \log \left\{ \pi \left[\exp[-(\beta - \lambda)\zeta \bar{r}_1(\theta)] \right] \right\} + \left(1 + \frac{\beta' - \lambda'}{\beta' + \lambda'} + \frac{\beta' - \lambda'}{\lambda} \right) \log \left(\frac{3}{\epsilon} \right) \right\} \\ &= r_2(\tilde{\theta}) + \left(\zeta + (1 - \zeta) \frac{\beta}{\lambda} \right) \{ \rho[r_1(\theta)] - r_1(\tilde{\theta}) \} \\ &\quad + \frac{1}{\lambda} \left\{ \mathcal{K}(\rho, \hat{\rho}_{(\beta - \lambda)\zeta}) + \left(1 + \frac{\beta' - \lambda'}{\beta' + \lambda'} + \frac{\beta' - \lambda'}{\lambda} \right) \log \left(\frac{3}{\epsilon} \right) \right\}. \end{aligned}$$

As a special case,

$$\hat{\rho}_\beta[r_2(\theta)] \leq r_2(\tilde{\theta}) + \frac{1}{\lambda} \int_{(\beta - \lambda)\zeta}^{\beta} \hat{\rho}_\gamma[\bar{r}_1(\theta)] d\gamma + \frac{1}{\lambda} \left(1 + \frac{\beta' - \lambda'}{\beta' + \lambda'} + \frac{\beta' - \lambda'}{\lambda} \right) \log \left(\frac{3}{\epsilon} \right).$$

CONCLUSION

We have shown that the PAC-Bayesian approach is powerful enough to provide some significant improvements to Vapnik's learning theory. Although we have treated only the case of classification, it is clear that the tools we used are generic and could be applied in various contexts where a non asymptotic deviation inequality is available for any value of the parameter θ (we previously studied the regression setting in [16], which was a source of inspiration to derive localized bounds). It is also clear that the use of exchangeable priors open the road to the conception and study of new ways to adapt a classification scheme to the data, still preserving good generalization properties, as we briefly mentioned. We hope to investigate further these possible generalizations and applications in forthcoming studies.

REFERENCES

- [1] A. Barron (1987) Are Bayes Rules Consistent in Information ? *Open Problems in Communication and Computation*, T. M. Cover and B. Gopinath Ed., Springer Verlag 1987.
- [2] A. Barron and Y. Yang (1995) Information Theoretic Determination of Minimax Rates of Convergence, *preprint of the Department of Statistics at Yale University*, <http://www.stat.yale.edu/Preprints>, submitted to the *Annals of Statistics*
- [3] A. Barron, L. Birgé and P. Massart, (1995) Risk bounds for model selection via penalization, *Probab. Theory Related Fields*, **113** (1999), no. 3, 301-413.
- [4] G. Blanchard, The "progressive mixture" estimator for regression trees, *Annales de l'I.H.P.*, **35**(6):793-820, 1999.
- [5] G. Blanchard, A new algorithm for Bayesian MCMC CART sampling, *preprint*, 2000.
- [6] G. Blanchard, *Mixture and aggregation of estimators for pattern recognition. Application to decision trees [Méthodes de mélange et d'agrégation d'estimateurs en reconnaissance de formes. Application aux arbres de décision.]* in English with an introduction in French, PhD dissertation, Université Paris XIII, January 2001.
- [7] L. Birgé and P. Massart (1997) From model selection to adaptive estimation, *Festschrift for Lucien Le Cam*, 55-87, Springer, New York.
- [8] L. Birgé and P. Massart (1995) Minimum contrast estimators on sieves, *Bernoulli* **4** (1998), no. 3, 329-375.
- [9] L. Birgé and P. Massart, A generalized C_p criterion for Gaussian model selection, *preprint*, 2001.
- [10] L. Birgé and P. Massart, Gaussian model selection, *J. Eur. Math. Soc.*, 2001. <http://www.springer.de/link>
- [11] O. Catoni (1997) A mixture approach to universal model selection, *preprint LMENS - 97 - 30*, <http://www.dmi.ens.fr/preprints>.

- [12] O. Catoni, Universal aggregation rules with sharp oracle inequalities, *to appear in the Annals of Statistics*, 2000, N.B.: this is a revised and extended version of "A mixture approach to universal model selection".
- [13] O. Catoni, Gibbs estimators, *preprint*, first draft 1998, last revision 2000, to appear in *Probab. Th. Rel. Fields*.
- [14] O. Catoni, Free energy estimates and deviation inequalities, *preprint, to appear in revised form, under the title Laplace transform estimates and deviation inequalities, in the Annales de l'I.H.P.*, 1999.
- [15] O. Catoni, Data compression and adaptive histograms, *to appear in the proceedings of the International Conference on Foundations of Computational Mathematics in honor of Professor Steve Smale's 70th Birthday, July 2000*.
- [16] O. Catoni, Statistical learning theory and stochastic optimization, *Lecture notes, Saint-Flour summer school on Probability Theory, 2001*, Springer, to appear.
- [17] N. Cristianini and J. Shawe Taylor, *An introduction to Support Vector Machines and other kernel based learning methods*, Cambridge University Press, 2000.
- [18] M. Feder and N. Merhav, Hierarchical Universal Coding, *IEEE Trans. Inform. Theory*, vol 42, no 5, Sept, 1996.
- [19] N. Littlestone and M. Warmuth, Relating data compression and learnability. *Technical report*, University of California, Santa Cruz, 1986.
- [20] D. A. McAllester, Some PAC-Bayesian Theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, 230–234 (electronic), ACM, New York, 1998;
- [21] D. A. McAllester, PAC-Bayesian Model Averaging, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, 164–170 (electronic), ACM, New York, 1999;
- [22] C. McDiarmid, Concentration *Probabilistic Methods for Algorithmic Discrete Mathematics*, Habib M., McDiarmid C. and Reed B. Eds., Springer, 1998.
- [23] B. Y. Ryabko, Twice-universal coding, *Probl. Inform. Transm.*, vol 20, no 3, pp. 24-28, July-Sept 1984.
- [24] J.-P. Vert, Double mixture and universal inference, *preprint*, 2000.
- [25] J.-P. Vert, Adaptive context trees and text clustering, *IEEE Trans. Inform. Theory* to appear.
- [26] J.-P. Vert, Text categorization using adaptive context trees, *preprint*, 2000.
- [27] E. T. Whittaker and G. N. Watson *A course of modern analysis*, Cambridge University Press, 1927.
- [28] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, The Context-Tree Weighting Method: Basic Properties, *IEEE Trans. Inform. Theory*, vol 41, no 3, May, 1995.
- [29] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, Context Weighting for General Finite-Context Sources, *IEEE Trans. Inform. Theory*, vol 42, no 5, Sept, 1996.

LABORATOIRE DE PROBABILITÉS ET MODÈLES ALÉATOIRES,, U.M.R. 7599 DU C.N.R.S. CASE 188, UNIVERSITÉ PARIS 6,, BUREAU 4 E 19 BÂT CHEVALERET, 16 RUE CLISSON,, F-75 013 PARIS, TEL: (33) (1) 44 27 85 15,, FAX: (33) (1) 44 27 72 23

E-mail address: catoni@ccr.jussieu.fr