Universités de Paris 6 & Paris 7 - CNRS (UMR 7599)

PRÉPUBLICATIONS DU LABORATOIRE

DE PROBABILITÉS & MODÈLES ALÉATOIRES

4, place Jussieu - Case 188 - 75 252 Paris cedex 05

http://www.proba.jussieu.fr

A PAC-Bayesian approach to adaptive classification

O. CATONI

SEPTEMBRE 2003

Prépublication n^o 840

Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599,
Université Paris VI & Université Paris VII,
4, place Jussieu, Case 188, F-75252 Paris Cedex 05.

A PAC-Bayesian approach to adaptive classification

Olivier Catoni

August 8, 2003

CNRS – Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6 (site Chevaleret), 4 place Jussieu – Case 188, 75 252 Paris Cedex 05.

Abstract. This is meant to be a self-contained presentation of adaptive classification seen from the PAC-Bayesian point of view. Although most of the results are original, some review materials about the VC dimension and support vector machines are also included. This study falls in the field of statistical learning theory, where complex data have to be analyzed from a limited amount of informations, drawn from a finite sample. It relies on non asymptotic deviation inequalities, where the complexity of models is captured through the use of prior measures. The main improvements brought here are more *localized* bounds and the use of *exchangeable* prior distributions. Interesting consequences are drawn for the generalization properties of support vector machines and the design of new classification algorithms.

2000 Mathematics Subject Classification: 62H30, 68T05, 62B10.
Keywords: Statistical learning theory, adaptive statistics, pattern recognition, PAC-Bayesian theorems, VC dimension, fat shattering dimension, localized complexity bounds, randomized estimators, Support Vector Machines, compression schemes, margin bounds.

Contents

CHAPTER 1

A PAC-Bayesian approach to adaptive inference

1. Foreword

In this paper, we will prove what could be called *localized PAC-Bayesian learning theorems* and illustrate their use to solve classification problems. The setting will be the one of *statistical learning theory* : 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.

Our aim is to give a self contained description of statistical classification from the PAC-Bayesian point of view. Although the bulk of the presented results are new, we have also included some expository materials whose proofs we wanted to adapt to our purpose and taste. We hope this additions will be convenient for the reader. We thus give a presentation of the VC dimension and of support vector machines, which come as natural applications of the PAC-Bayesian approach. As for support vector machines, we made two choices which may be considered a matter of taste: we deliberately avoided using the Kuhn-Tucker and Mercer theorems. We prefered to replace the Kuhn-Tucker theorem by a more geometrical approach, exploiting simple properties of the orthogonal projection on a convex set, with the hope of giving a more intuitive idea of what is going on in the computation of the canonical hyperplane. We did not mention Mercer's theorem, because the fact that it is not really needed brings some more generality.

2. Introduction

The idea of *PAC-Bayesian* learning theorems, as introduced by D. McAllester, [**23, 24**] 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!). 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 [**34**].

3. 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 $\mathcal{Y}$. (In some chapters, we will relax this hypothesis, replacing $P^{\otimes N}$ with an exchangeable distribution and considering a test set $(X_i, Y_i)_{i=N+1}^{2N}$ of the same size as the training set $(X_i, Y_i)_{i=1}^N$.)

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

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

where $(\Theta, \mathcal{T})$ is some measurable parameter set and

$$(\theta, x) \mapsto f_\theta(x) : \big(\Theta \times \mathcal{X}, \mathcal{T} \otimes \mathcal{B}\big) \to \big(\mathcal{Y}, \mathcal{B}'\big)$$

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, leading to a non asymptotic level of confidence for this risk. The risk of $f_\theta : \mathcal{X} \to \mathcal{Y}$ will be measured as its error rate

$$R(\theta) = P\big[Y \neq f_\theta(X)\big].$$

Let us mention here that throughout these lectures 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 minimum of the empirical risk

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

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 $\pi \in \mathcal{M}_+^1\big(\Theta, \mathcal{T}\big)$ defined on the parameter space.

More precisely, we cannot minimize $R(\theta)$ with respect to 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

$P\big[r(\theta)\big] = R(\theta)$, 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 regularizing 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 $P\Big[\inf_\theta \big[r(\theta)+\gamma_N(\theta)\big]\Big]$ 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 $\pi \in \mathcal{M}_+^1\big(\Theta, \mathcal{T}\big)$ on the parameter space. This is an indirect way to make the size of Θ come into play. Although we will not explicitely 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

$$q_\alpha(r) = \inf\Big\{\mu : \pi\big[r(\theta) \le \mu\big] > \alpha\Big\}.$$

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

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

This generalization is of practical interest to us, because, whereas $\operatorname{ess\,inf}_{\pi(d\theta)} r(\theta)$ 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 situation for the true infimum is of course even worse, since it is sensitive to a lower deviation of the empirical risk at a single value of the parameter).

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

$$d\rho(\theta) = d\pi_{\exp(-\beta r)}(\theta) = \frac{\exp\big(-\beta r(\theta)\big)}{\pi\big[\exp\big(-\beta r(\theta)\big)\big]} d\pi(\theta). \tag{3.1}$$

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 $\rho\big[R(\theta)\big]$, 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 estimators.

In McAllester's papers, the fluctuations of $\rho\big[r(\theta)\big]$ with respect to $\rho\big[R(\theta)\big]$ 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.

Then we will show how Vapnik's statistical learning theory can be proved and strengthened using the PAC-Bayesian approach : the idea is to replace the use of a deterministic prior with the use of a data dependent one.

Finally, we will deduce results for Vapnik's support vector machines, one of the most efficient and still promising classification algorithm. We will also study the wider class of compression schemes [**22**], a framework into which many practical algorithms can be designed and which covers generalization bounds for support vector machines computed from the number of support vectors.

4. 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 $\big[(\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\big]$ 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 $\mathcal{R} = \{f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}$ is at our disposal, where $(\theta, x) \mapsto f_\theta(x) : (\Theta \times X, \mathcal{T} \times \mathcal{B}) \to (\mathcal{Y}, \mathcal{B}')$ is measurable. We will not make any "low-noise" assumption, but it will just turn out that the theoretical bounds derived in this section will be of the sharpest possible order of magnitude with respect to the sample size N only when the optimal error rate $\inf_{\theta \in \Theta} R(\theta)$ is small (meaning that it is of order N^{-1}).(The situation is different for purely empirical bounds providing a level of confidence for the error rate, indeed

when the error rate is of order one — say 1/5 for example — independently of N, due to a noisy training set, all the empirical error rates have fluctuations of order $N^{-1/2}$ and it is impossible to derive a confidence level for the true error rate with a better accuracy, althouhg it is possible to derive an estimator approaching the optimal one at a higher speed !).

4.1. A reminder of non-asymptotic deviation techniques: Bernstein's inequality and the Legendre transform of the Kullback divergence function. We will 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 [**25**, p 203-204].

THEOREM 4.1. *Let $(\sigma_1, \ldots, \sigma_N)$ be independent real valued random variables and $\mathbb{P}$ their joint distribution. Let us assume that*

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

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

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

its renormalized variance. Let us introduce the increasing function of a real parameter

$$g(x) = x^{-2}\big[\exp(x) - 1 - x\big].$$

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

$$\mathbb{P}(S - m \geq \eta) \leq \mathbb{P}\Big[\exp\Big(-\lambda\eta + \lambda(S - m)\Big)\Big] \tag{4.1}$$

$$\leq \exp\left(-\eta\lambda + g\left(\frac{b\lambda}{N}\right)\frac{V}{N}\lambda^2\right), \tag{4.2}$$

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

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

PROOF. (Given for the sake of completeness.) It can be easily checked that the function $x \mapsto g(x) : \mathbb{R} \to \mathbb{R}$ is increasing. Moreover, it is clearly enough to prove

the theorem when $\mathbb{P}(\sigma_i) = 0$, for any $i = 1, \dots, N$. Reminding that $\log(1+x) \le x$, we see that

$$\begin{aligned}
\mathbb{E}\big[\exp(\lambda S)\big] &= \exp\left\{\sum_{i=1}^N \log\left[\mathbb{E}\left[\exp\left(\frac{\lambda}{N}\sigma_i\right)\right]\right] - \mathbb{E}\left[\frac{\lambda}{N}\sigma_i\right]\right\} \\
&\le \exp\left\{\sum_{i=1}^N \mathbb{E}\left[\exp\left(\frac{\lambda}{N}\sigma_i\right) - \frac{\lambda}{N}\sigma_i - 1\right]\right\} \\
&= \exp\left\{\sum_{i=1}^N \mathbb{E}\left[g\left(\frac{\lambda}{N}\sigma_i\right)\sigma_i^2\frac{\lambda^2}{N^2}\right]\right\} \\
&\le \exp\left\{g\left(\frac{b\lambda}{N}\right)\sum_{i=1}^N \mathbb{E}\big[\sigma_i^2\big]\frac{\lambda^2}{N^2}\right\} \\
&= \exp\left\{g\left(\frac{b\lambda}{N}\right)\frac{V}{N}\lambda^2\right\}.
\end{aligned}$$

This proves (4.2). The last statement of the theorem can be rewritten after a suitable change of variables as

$$-\eta\lambda + \lambda^2 g(\lambda) \le -\frac{3\eta^2}{6+2\eta} \quad \text{when} \quad \lambda = \log(1+\eta).$$

This is equivalent to

$$(1+\eta)\log(1+\eta) - \eta \ge \frac{3\eta^2}{6+2\eta},$$

and therefore to

$$(6+8\eta+2\eta^2)\log(1+\eta) - 5\eta^2 - 6\eta \ge 0.$$

This last inequality, which holds true when $\eta = 0$, can be checked to hold true for any positive value of η by differentiating twice its left-hand side. □

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

LEMMA 4.2. *Let us recall that for any measurable function $h : \Theta \to \mathbb{R}$,*

$$\log\left\{\pi\Big\{\exp\big[h(\theta)\big]\Big\}\right\} = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho\big[h(\theta)\big] - \mathcal{K}(\rho, \pi), \tag{4.3}$$

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

$$\rho\big[h(\theta)\big] \stackrel{def}{=} \sup_{B \in \mathbb{R}} \rho\Big[\min\big\{B, h(\theta)\big\}\Big], \tag{4.4}$$

and where it is also understood that

$$\infty - \infty = \sup_{B \in \mathbb{R}}(B) - \infty = \sup_{B \in \mathbb{R}}(B - \infty) = -\infty. \tag{4.5}$$

(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 $\rho \in \mathcal{M}_+^1(\Theta, \mathcal{T})$,

$$\log\left\{\pi\Big[\exp\big[h(\theta)\big]\Big]\right\} + \mathcal{K}(\rho, \pi) - \rho\big[h(\theta)\big] = \mathcal{K}(\rho, \nu), \tag{4.6}$$

where $d\nu(\theta) = \dfrac{\exp\bigl[h(\theta)\bigr]}{\pi\Bigl\{\exp\bigl[h(\theta)\bigr]\Bigr\}} d\pi(\theta)$. *(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* $-\rho\bigl[h(\theta)\bigr] < \infty$ *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 (4.6) 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

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

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

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

This is precisely (4.6).

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 $\min\bigl\{B, h(\theta)\bigr\} = B \wedge h(\theta)$,

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

□

4.2. A non localized learning theorem for low-noise classification. We will apply the second inequality (4.2) of Bernstein's theorem 4.1 successively to

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

and to

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

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 [**23, 24**], which were derived from the weaker Hoeffding's inequality.

LEMMA 4.3. *For any positive real parameter $\lambda \in \mathbb{R}_+^*$, any lower-bounded real valued measurable function $\eta : \Theta \to \mathbb{R}_+$, any prior probability distribution $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$,*

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

In the same way

$$P^{\otimes N}\left\{\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\rho\big[r(\theta)\big] - \lambda\rho\big[R(\theta)\big] - \rho\big[\eta(\theta)\big] - \mathcal{K}(\rho,\pi) \geq 0\right\} \tag{4.7}$$
$$\leq \pi\left\{\exp\left[\frac{\lambda^2}{N} g\left(\frac{\lambda}{N}\right) R(\theta)\big[1 - R(\theta)\big] - \eta(\theta)\right]\right\}.$$

PROOF. According to lemma 4.2,

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

Thus

$$P^{\otimes N}\left\{\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\rho\big[R(\theta)\big] - \lambda\rho\big[r(\theta)\big] - \rho\big[\eta(\theta)\big] - \mathcal{K}(\rho,\pi) \geq 0\right\} \tag{4.8}$$
$$= P^{\otimes N}\left\{\pi\left\{\exp\Big[\lambda\big[R(\theta) - r(\theta)\big] - \eta(\theta)\Big]\right\} \geq 1\right\} \tag{4.9}$$
$$\leq P^{\otimes N}\left\{\pi\left\{\exp\Big[\lambda\big[R(\theta) - r(\theta)\big] - \eta(\theta)\Big]\right\}\right\} \tag{4.10}$$
$$= \pi\left\{P^{\otimes N}\left\{\exp\Big[\lambda\big[R(\theta) - r(\theta)\big] - \eta(\theta)\Big]\right\}\right\} \tag{4.11}$$
$$\leq \pi\left\{\exp\left[\frac{\lambda^2}{N} g\left(\frac{\lambda R(\theta)}{N}\right) R(\theta)\big[1 - R(\theta)\big] - \eta(\theta)\right]\right\}. \tag{4.12}$$

Equality (4.11) is obtained by applying the Fubini theorem to the positive function $(\theta, X_1, Y_1, \ldots, X_N, Y_N) \mapsto \exp\Big\{\lambda\big[R(\theta) - r(\theta)\big] - \eta(\theta)\Big\}$. Inequality (4.12) is obtained by applying inequality (4.2) of Bernstein's theorem 4.1 for each value of the parameter θ.

The proof of the reverse inequality (4.7) is similar and is left to the reader. □

REMARK 4.1. The last step of the proof (4.12) 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 noticed that

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

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

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

This is a simple application of equality (4.6) in another context.

5. A short users guide to empirical bounds

5.1. Building an estimator. 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 the above 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 4.3 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 5.1. *For any $\lambda\in\mathbb{R}_+$ such that $1-g(\frac{\lambda}{N})\frac{\lambda}{N}>0$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any posterior distribution $\rho\in\mathcal{M}_+^1(\Theta)$,*

$$\rho\big[R(\theta)\big]\leq\left[1-g\left(\frac{\lambda}{N}\right)\frac{\lambda}{N}\right]^{-1}\Big\{\rho\big[r(\theta)\big]+\frac{1}{\lambda}\big[\mathcal{K}(\rho,\pi)+\log(\epsilon^{-1})\big]\Big\}. \tag{5.1}$$

The above inequality is the kind of non-asymptotic empirical bound we will be after throughout this 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. Although we will not use it in the following discussion, it may be interesting to notice here that $\hat\rho$ can be explicited: namely it is the Gibbs posterior distribution $\hat\rho=\pi_{\exp(-\lambda r)}$ (where we have used the notation introduced in (3.1)). Its risk has an upper confidence bound $B(\hat\rho,\epsilon)$ at level ϵ, where, putting $\kappa=g\left(\frac{\lambda}{N}\right)\simeq\frac{1}{2}$ for short,

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

In other words,

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

This is satisfactory from the practical point of view, since $B(\hat\rho,\epsilon)$ is computable from the observed sample $(X_i,Y_i)_{i=1}^N$ and thus provides a confidence level.

5.2. Deriving a theoretical bound. 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

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

It is straightforward to estimate its deviations. We can for instance deduce from Bernstein's theorem 4.1 that

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

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 $1 - \epsilon$,

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

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, (and remembering that $\hat{\rho} = \pi_{\exp(-\beta r)}$), we get

PROPOSITION 5.2. *For any $\lambda \in \mathbb{R}_+$ such that $1 - \kappa \frac{\lambda}{N} > 0$, with $P^{\otimes N}$ probability at least $1 - \epsilon$,*

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

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

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

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

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

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,

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

A slight improvement is achieved if we come back to (4.10) and (4.12). With a proper choice of parameters, we get

$$P^{\otimes N}\left\{ \pi\left[\exp\Big[\lambda\big(1 - \kappa \tfrac{\lambda}{N}\big) R(\theta) - \lambda r(\theta) \Big] \right] \right\}$$

$$= P^{\otimes N}\Big\{\exp\Big[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda\big(1-\kappa\tfrac{\lambda}{N}\big)\rho\big[R(\theta)\big] - \lambda\rho\big[r(\theta)\big] - \mathcal{K}(\rho,\pi)\Big]\Big\} \leq 1.$$

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

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

Therefore

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

PROPOSITION 5.3. *For any $\lambda\in\mathbb{R}_+$ such that $\kappa\frac{\lambda}{N}<1$,*

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

REMARK 5.1. The theoretical bounds provided by propositions 5.2 and 5.3 are helpful to answer the question of the choice of the prior distribution π: it should be chosen to make $-\frac{1}{\lambda}\log\Big\{\pi\Big[\exp\big[-\lambda R(\theta)\big]\Big]\Big\}$ as close to $\inf_\theta R(\theta)$ as possible. Of course, if $\theta\mapsto R(\theta)$ were a known function, we would choose for π the Dirac mass at some parameter θ where $\inf_\theta R(\theta)$ is reached. As we do not have this information, and are as a rule faced with the perspective that this infimum may be reached at any given location, we have to spread π over the whole parameter space Θ. In which way it should be spread depends on the kind of goal we want to achieve and on the kind of informations we may have on P and therefore on $\theta\mapsto R(\theta)$. For instance, if $\Theta=[0,1]^d$ and it is known that $\theta\mapsto R(\theta)$ is twice differentiable and

$$\mathrm{Det}\left[\frac{\partial^2}{\partial\theta^2}R(\theta)\right] \leq H(\theta) \tag{5.2}$$

at each point θ where $R(\theta)$ reaches its minimum, then it makes sense to choose for π the probability distribution whose density with respect to the Lebesgue measure is proportional to $\sqrt{H(\theta)}$, because this is the choice that makes

$$-\frac{1}{\lambda}\log\Big\{\pi\Big[\exp\big[-\lambda R(\theta)\big]\Big]\Big\} - \inf_\theta R(\theta)$$

asymptotically constant (i.e. independent of the point where $\inf_\theta R(\theta)$ is reached) when λ tends to infinity and the bound (5.2) is reached (notice that the optimal value of λ is at least of order $\sqrt{N}$). It is clear anyhow that such a bound as (5.2) implies that some information on P — or at least on the marginal distribution

of the pattern X under P — is known. The introduction of exchangeable prior distributions in chapter 2 (which indeed are not proper prior distributions, since they are allowed to depend on the patterns) will lead to more obvious ways to choose the prior, and to some kind of *automated* adaptation of the prior to the pattern distribution.

5.3. Optimizing the parameter λ. Another important remark is to notice that corollary 5.1 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. Defining $\hat{\lambda}$ to be the minimizer in $\lambda \in \Lambda$ of

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

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

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

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

PROPOSITION 5.4. *With $P^{\otimes N}$ probability at least $1-\epsilon$, for any posterior distributions $\mu \in \mathcal{M}_+^1(\mathbb{R}_+)$ such that $\frac{\mu(\kappa\lambda^2)}{\mu(\lambda)N} < 1$ and any $\rho \in \mathcal{M}_+^1(\Theta)$,*

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

(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 5.5. *If $\hat{\mu}$ and $\hat{\rho} = \pi_{\exp(-\hat{\mu}(\lambda)r)}$ are the optimizers of the empirical bound of proposition 5.4 at level of confidence ϵ, then with $P^{\otimes N}$ probability at least $1-\epsilon$*

$$\begin{aligned}\hat{\rho}\big[R(\theta)\big] &\leq \inf_{\mu,\rho}\left\{ \left(1-\frac{\mu\big(\kappa\lambda^2\big)}{\mu(\lambda)N}\right)^{-1} \left\{ \left(1+g\big(\tfrac{\mu(\lambda)}{N}\big)\frac{\mu(\lambda)}{N}\right)\rho\big[R(\theta)\big] \right.\right. \\ &\qquad \left.\left. + \frac{1}{\mu(\lambda)}\Big[\mathcal{K}(\rho,\pi) + \mathcal{K}(\mu,\nu) + 2\log(\tfrac{2}{\epsilon})\Big] \right\} \right\},\end{aligned}$$

where the infimum in μ is taken over all the distributions of $\mathcal{M}_+^1(\mathbb{R}_+)$ such that $\frac{\mu(\kappa\lambda^2)}{\mu(\lambda)N} < 1$, and the infimum in ρ is taken over $\mathcal{M}_+^1(\Theta)$.

Anyhow, we mentioned proposition 5.4 and its corollary rather as a curiosity, using the union bound on a grid (which is by the way a special case of this proposition) being quite sufficient in practice. Let us make this more explicit. Consider

for some real parameter $\alpha > 1$ the grid

$$\Lambda = \left\{ N\alpha^{-k} \ : k \in \mathbb{N}, 0 \leq k \leq \frac{\log(2N)}{\log(\alpha)} \right\}.$$

Let us consider the uniform probability distribution on Λ, in other words, let us choose

$$\nu = \frac{1}{|\Lambda|} \sum_{\lambda' \in \Lambda} \delta_{\lambda'}.$$

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

$$\begin{aligned}
\rho\big[R(\theta)\big] &\leq \inf_{\lambda' \in \Lambda} \left(1 - g(\tfrac{\lambda'}{N})\frac{\lambda'}{N}\right)^{-1} \Bigg\{ \rho\big[r(\theta)\big] \\
&\qquad + \frac{1}{\lambda'}\Big[\mathcal{K}(\rho,\pi) + \log\big[\tfrac{\log(2N)}{\log(\alpha)} + 1\big] + \log(\epsilon^{-1})\Big]\Bigg\} \\
&\leq \inf_{\lambda \in [1,N]} \left(1 - g(\tfrac{\alpha\lambda}{N})\frac{\alpha\lambda}{N}\right)^{-1} \Bigg\{ \rho\big[r(\theta)\big] \\
&\qquad + \frac{1}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left[\tfrac{\log(2N)}{\log(\alpha)} + 1\right] + \log(\epsilon^{-1})\right]\Bigg\},
\end{aligned}$$

where the infima in λ' and λ are restricted to those values for which $g(\frac{\lambda'}{N})\frac{\lambda'}{N} < 1$ and $g(\frac{\alpha\lambda}{N})\frac{\alpha\lambda}{N} < 1$ respectively.

The second inequality is obtained by considering that for any $\lambda \in [1, N]$ there is $\lambda' \in \Lambda$ such that $\lambda \leq \lambda' \leq \alpha\lambda$. Moreover the right-hand side of it is minimized for any fixed value of λ by the Gibbs distribution $\pi_{\exp(-\lambda r)}$ (see (3.1)), which we will write for short as $\hat{\rho}_\lambda = \pi_{\exp(-\lambda r)}$.

COROLLARY 5.6. *For any $\alpha > 1$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\lambda \in [1, 2N]$ such that $g(\frac{\alpha\lambda}{N})\frac{\alpha\lambda}{N} < 1$,*

$$\begin{aligned}
\hat{\rho}_\lambda\big[R(\theta)\big] \leq \left(1 - g(\tfrac{\alpha\lambda}{N})\frac{\alpha\lambda}{N}\right)^{-1} \frac{1}{\lambda}\Bigg\{ &- \log\Big\{\pi\Big[\exp\big[-\lambda r(\theta)\big]\Big]\Big\} \\
&+ \log\left[\tfrac{\log(2N)}{\log(\alpha)} + 1\right] + \log(\epsilon^{-1})\Bigg\}.
\end{aligned}$$

Let us remark that it can be useful for computing $\log\Big\{\pi\Big[\exp\big[-\lambda r(\theta)\big]\Big]\Big\}$ to use the identity

$$-\log\Big\{\pi\Big[\exp\big[-\lambda r(\theta)\big]\Big]\Big\} = \int_0^\lambda \hat{\rho}_\beta\big[r(\theta)\big]d\beta,$$

and to notice that for any sequence $\beta_0 = 0 < \beta_1 < \cdots < \beta_m = \lambda$,

$$\sum_{k=1}^m (\beta_k - \beta_{k-1})\hat{\rho}_{\beta_k}\big[r(\theta)\big] \leq \int_0^\lambda \hat{\rho}_\beta d\beta \leq \sum_{k=1}^m (\beta_k - \beta_{k-1})\hat{\rho}_{\beta_{k-1}}\big[r(\theta)\big].$$

(This comes from the fact that $\beta \mapsto \hat{\rho}_\beta\big[r(\theta)\big]$ is decreasing, its derivative being the opposite of the variance of $r(\theta)$ under $\hat{\rho}_\beta$.)

5.4. Bounding the deviations under the posterior. Let us note also that the upper deviations of the risk $R(\theta)$ under the Gibbs posterior $\hat\rho_\lambda$ can easily be bounded. This is important, since it proves that the posterior can be used to pick some parameter $\hat\theta$ once for all and then use $f_{\hat\theta}$ to classify all forthcoming data.

Indeed with $P^{\otimes N}$ probability at least $1-\epsilon$, for any $\lambda \in [1,N]$, any $\beta \in \mathbb{R}_+$,

$$\begin{aligned}
\log\Big\{\hat\rho_\lambda\Big\{\exp\big[\beta R(\theta)\big]\Big\}\Big\} &= \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \beta\rho\big[R(\theta)\big] - \mathcal{K}(\rho,\hat\rho_\lambda) \\
&= \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \beta\rho\big[R(\theta)\big] - \mathcal{K}(\rho,\pi) \\
&\qquad - \lambda\rho\big[r(\theta)\big] - \log\Big\{\pi\Big[\exp\big[-\lambda r(\theta)\big]\Big]\Big\} \\
&\le \beta\left(1 - g(\tfrac{\alpha\lambda}{N})\frac{\alpha\lambda}{N}\right)^{-1}\Big\{\rho\big[r(\theta)\big] \\
&\qquad + \frac{1}{\lambda}\Big\{\mathcal{K}(\rho,\pi) + \log(\epsilon^{-1}) + \log\Big[\tfrac{\log(N)}{\log(\alpha)}\Big]\Big\}\Big\} \\
&\qquad - \mathcal{K}(\rho,\pi) - \lambda\rho\big[r(\theta)\big] - \log\Big\{\pi\Big[\exp\big[-\lambda r(\theta)\big]\Big]\Big\}.
\end{aligned}$$

Thus choosing $\beta = \lambda\big(1 - g(\frac{\alpha\lambda}{N})\frac{\alpha\lambda}{N}\big)$, we get

$$\log\Big\{\hat\rho_\lambda\Big[\exp\big[\beta R(\theta)\big]\Big]\Big\} \le -\log\Big\{\pi\Big[\exp\big[-\lambda r(\theta)\big]\Big]\Big\} + \log(\epsilon^{-1}) + \log\Big[\tfrac{\log(N)}{\log(\alpha)} + 1\Big].$$

This proves

COROLLARY 5.7. *For any real constant $\alpha > 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any $\lambda \in [1,2N]$ such that $g(\frac{\alpha\lambda}{N})\frac{\alpha\lambda}{N} < 1$, with $\hat\rho_\lambda$ probability at least $1-\eta$,*

$$\begin{aligned}
R(\theta) \le \left(1 - g(\tfrac{\alpha\lambda}{N})\frac{\alpha\lambda}{N}\right)^{-1}\frac{1}{\lambda}\Big\{&-\log\Big\{\pi\Big[\exp\big[-\lambda r(\theta)\big]\Big]\Big\} \\
&- \log(\epsilon\eta) + \log\Big[\tfrac{\log(2N)}{\log(\alpha)} + 1\Big]\Big\}.
\end{aligned}$$

Although we will not mention it any further, the same kind of upper deviation bounds with respect to the Gibbs posterior can be derived from the results presented in the sequel of this article.

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

Let $\{f_\theta : \mathcal{X} \to \{-1,+1\}; \theta\in\Theta\}$ 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 $\nu \in \mathcal{M}_+^1(\Theta)$, we consider the aggregated classifier

$$f_\nu(x) = \operatorname{sign}\Big(\nu\big[f_\theta(x)\big]\Big).$$

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

$$R(\nu) = P\Big\{Y \ne \operatorname{sign}\Big[\nu\big[f_\theta(X)\big]\Big]\Big\},$$

and the corresponding empirical risk is

$$r(\nu) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big\{Y_i \neq \operatorname{sign}\Big[\nu\big[f_\theta(X_i)\big]\Big]\Big\}.$$

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

$$\Psi : \Theta^M \to \Theta'$$

$$\theta_1^M \mapsto \frac{1}{M}\sum_{i=1}^{M} \delta_{\theta_i}.$$

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 5.8. *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 :

$$\mathcal{K}(\rho, \pi^{\otimes M}) = \mathcal{K}(\rho', \pi') + \rho\Big\{\mathcal{K}\Big[\rho\big[d\theta_1^M \mid \Psi(\theta_1^M)\big], \pi^{\otimes M}\big[d\theta_1^M \mid \Psi(\theta_1^M)\big]\Big]\Big\}.$$

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

From corollary 5.6, a real parameter $\alpha > 1$ being chosen, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any $\lambda \in [1, 2N]$ such that $g(\frac{\alpha\lambda}{N})\frac{\alpha\lambda}{N} < 1$, any $\rho \in \mathcal{M}_+^1(\Theta^M)$,

$$\rho\Big\{R\big[\Psi(\theta_1^M)\big]\Big\} \leq \left(1 - g(\tfrac{\alpha\lambda}{N})\frac{\alpha\lambda}{N}\right)^{-1} \Big\{\rho\Big\{r\big[\Psi(\theta_1^M)\big]\Big\} + \frac{1}{\lambda}\Big[\mathcal{K}(\rho, \pi^{\otimes M}) + \log(\epsilon^{-1}) + \log\Big[\tfrac{\log(2N)}{\log(\alpha)} + 1\Big]\Big]\Big\}.$$

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

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

THEOREM 5.9. *For any real parameter $\alpha > 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any choice of inverse temperature $\lambda \in [1, 2N]$ such that $g(\frac{\alpha\lambda}{N})\frac{\alpha\lambda}{N} < 1$,*

$$\hat{\rho}_\lambda\Big\{R\big[\Psi(\theta_1^M)\big]\Big\} \leq \left(1 - g(\tfrac{\alpha\lambda}{N})\frac{\alpha\lambda}{N}\right)^{-1} \frac{1}{\lambda}\Big\{-\log\Big\{\pi^{\otimes M}\Big[\exp\Big[-\lambda r\big[\Psi(\theta_1^M)\big]\Big]\Big]\Big\} + \log(\epsilon^{-1}) + \log\Big[\tfrac{\log(2N)}{\log(\alpha)} + 1\Big]\Big\}.$$

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

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

$$-\log\left\{\pi^{\otimes M}\left[\exp\left[-\lambda r\left[\Psi(\theta_1^M)\right]\right]\right]\right\} = \int_0^\lambda \hat{\rho}_\gamma\left\{r\left[\Psi(\theta_1^M)\right]\right\}d\gamma$$
$$\leq \sum_{k=1}^m (\beta_k - \beta_{k-1})\hat{\rho}_{\beta_{k-1}}\left\{r\left[\Psi(\theta_1^M)\right]\right\}$$

for any sequence of inverse temperatures $\beta_0 = 0 < \beta_1 < \cdots < \beta_m = \lambda$. This leads to the following computation scheme : estimate $\hat{\rho}_{\beta_k}\left\{r\left[\Psi(\theta_1^M)\right]\right\}$ for increasing values of β_k and compute the bound for $\hat{\rho}_{\beta_k}\left\{R\left[\Psi(\theta_1^M)\right]\right\}$. Keep the temperature with the best bound. If we do not trust the constants in the bound, we can 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.

In practice, one of the most successful method for aggregating classification rules is the boosting algorithm. We refer to [**18**] for more informations on this topic, and to [**19**] which explains how the PAC-Bayesian approach can be used to study classification rules of the boosting type. Another (presumably more powerful, although it would not be so easy to prove it mathematically) approach for aggregating classifiers using support vector machines will be described in chapter 5.

5.6. 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, at least as far as theoretical bounds are concerned (the situation for the computation of an empirical level of confidence for the error rate being different, as already mentioned).

We have also in mind to make some connection between the penalty terms presented here and Vapnik's entropy. This is to be the subject of the following sections and chapters.

6. Localized learning lemmas

The loss of localization in the use we made so far of lemma 4.3 comes from the choice of $\eta(\theta)$: the level of confidence, i.e. the right-hand side of the learning inequality, appears as the expectation under the prior π of contributions coming from each possible value of the parameter θ. We used to make these contributions equal to ϵ for each θ. Another choice stems from the remark that we are only interested in the relationship between the points where the minimum of $R(\theta)$ is reached and the points where the minimum of $r(\theta)$ is reached. This means that for values of $R(\theta)$ away from $\inf_\theta R(\theta)$, we only want to make sure that $r(\theta)$ is sufficiently above $\inf_\theta r(\theta)$: therefore we can afford larger confidence intervals for these values, resulting in higher confidence levels.

As the lower bound of the confidence interval for $r(\theta)$ (we are not interested in the upper bound, which we can take to be infinite) is $R(\theta) - \eta(\theta)$, a better

localization may be expected from choosing a larger value for $\eta(\theta)$. A natural choice (indexed by some positive parameter β tuning the strength of the localization) is

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

This leads to

$$P^{\otimes N}\Bigg\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\left(\lambda-\beta-\kappa\frac{\lambda^2}{N}\right)\rho\big[R(\theta)\big]-\lambda\rho\big[r(\theta)\big] - \mathcal{K}(\rho,\pi) - \log\Big\{\pi\Big[\exp\big[-\beta R(\theta)\big]\Big]\Big\} \geq \log(\epsilon^{-1})\Bigg\} \leq \epsilon, \tag{6.1}$$

where we have put as usual $\kappa = g(\frac{\lambda}{N})$ for short and where we assume that we have chosen parameters such that $\lambda-\beta-\kappa\frac{\lambda^2}{N} > 0$. With the same choice of parameters, the reverse inequality reads as

$$P^{\otimes N}\Bigg\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\lambda\rho\big[r(\theta)\big]-\left(\lambda+\beta+\kappa\frac{\lambda^2}{N}\right)\rho\big[R(\theta)\big] - \mathcal{K}(\rho,\pi) - \log\Big\{\pi\Big[\exp\big[-\beta R(\theta)\big]\Big]\Big\} \geq \log(\epsilon^{-1})\Bigg\} \leq \epsilon. \tag{6.2}$$

To exploit these inequalities, we need an empirical upper bound for $\log\Big\{\pi\Big[\exp\big[-\beta R(\theta)\big]\Big]\Big\}$. This is where the reverse inequality (6.2) comes into play: with $P^{\otimes N}$ probability at least $1-\epsilon$,

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

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

$$\begin{aligned}
\log\Big\{\pi\Big[\exp\big[-\beta R(\theta)\big]\Big]\Big\} &\leq \sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Big\{-\xi\lambda\rho\big[r(\theta)\big]-\mathcal{K}(\rho,\pi)\Big\} + \xi\log(\epsilon^{-1}) \\
&= \log\Big\{\pi\Big[\exp\big[-\xi\lambda r(\theta)\big]\Big]\Big\} + \xi\log(\epsilon^{-1}). \qquad (6.3)
\end{aligned}$$

Combining this result with (6.1), we get

LEMMA 6.1 (localized learning lemma, first form). *For any $\lambda\in\mathbb{R}_+$ and $\xi\in[0,1[$ such that $(1-\xi)-(1+\xi)\kappa\frac{\lambda}{N} > 0$,*

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

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

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

where

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

LEMMA 6.2 (localized learning lemma, second form). *For any $\lambda \in \mathbb{R}_+$ and $\xi \in [0,1[$ such that $(1-\xi) - (1+\xi)\kappa\frac{\lambda}{N} > 0$,*

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

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 :

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

It satisfies

COROLLARY 6.3. *For any $\lambda \in \mathbb{R}_+$ and $\xi \in [0,1]$ such that $(1-\xi) - (1+\xi)\kappa\frac{\lambda}{N} > 0$, with $P^{\otimes N}$ probability at least $1-\epsilon$,*

$$\begin{aligned}
\hat{\rho}_\lambda\big[R(\theta)\big] &\leq \left[(1-\xi)\lambda - (1+\xi)\kappa\frac{\lambda^2}{N}\right]^{-1} \Big\{\log\Big\{\pi\Big[\exp\big[-\lambda r(\theta)\big]\Big]\Big\} \\
&\qquad + \log\Big\{\pi\Big[\exp\big[-\xi\lambda r(\theta)\big]\Big]\Big\} + (1+\xi)\log(\tfrac{2}{\epsilon})\Big\} \\
&= \left(1 - \frac{1+\xi}{1-\xi}\kappa\frac{\lambda}{N}\right)^{-1} \left\{\frac{1}{1-\xi}\int_\xi^1 \hat{\rho}_{\beta\lambda}\big[r(\theta)\big]d\beta + \frac{(1+\xi)}{(1-\xi)\lambda}\log\left(\tfrac{2}{\epsilon}\right)\right\} \\
&\leq \left(1 - \frac{1+\xi}{1-\xi}\kappa\frac{\lambda}{N}\right)^{-1} \left\{\hat{\rho}_{\xi\lambda}\big[r(\theta)\big] + \frac{(1+\xi)}{(1-\xi)\lambda}\log(\tfrac{2}{\epsilon})\right\}
\end{aligned}$$

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 forthcoming 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.

Another interesting — although suboptimal — choice of the posterior distribution ρ in lemma 6.2 is to cancel the divergence term by considering $\rho = \hat{\rho}_{\xi\lambda}$. This is handy as it provides a way to compare $\hat{\rho}_\lambda\big[R(\theta)\big]$ with $\hat{\rho}_\lambda\big[r(\theta)\big]$.

COROLLARY 6.4. *For any $\lambda > 0$, any $\xi \in [0,1[$, such that $g(\frac{\lambda}{\xi N})\frac{(1+\xi)\lambda}{\xi(1-\xi)N} < 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$,*

$$\hat\rho_\lambda\big[R(\theta)\big] \leq \left(1 - g\left(\tfrac{\lambda}{\xi N}\right)\frac{(1+\xi)\lambda}{\xi(1-\xi)N}\right)^{-1}\left\{\hat\rho_\lambda\big[r(\theta)\big] + \frac{\xi(1+\xi)}{(1-\xi)\lambda}\log(\tfrac{2}{\epsilon})\right\}.$$

To get a theoretical localized bound, we can come back to (6.1) to see that for any fixed probability measure $\rho \in \mathcal{M}_+^1(\Theta)$, with $P^{\otimes N}$ probability at least $1-\epsilon$,

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

Moreover, from Bernstein's inequality (4.2), with $P^{\otimes N}$ probability at least $1-\epsilon$,

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

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

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

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

COROLLARY 6.5. *For any $\lambda > 0$ and $\xi \in [0,1]$ such that $1 - \xi - \kappa\frac{\lambda}{N} > 0$, with $P^{\otimes N}$ probability at least $1-\epsilon$,*

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

Let us show now how to make corollary 6.3 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

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

For any $\lambda \in [1, 2N]$, let $\lambda' \in \Lambda$ be such that $\zeta\lambda' \leq \lambda \leq \lambda'$. From lemma 6.1 we deduce that with $P^{\otimes N}$ probability at least $1-\epsilon$, for any $\lambda \in [1, 2N]$,

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

We can now use the fact that

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

to get

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

Let us remark now that

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

We have proved

THEOREM 6.6. *For any $\xi \in [0,1[$, any $\zeta \in [\xi,1[$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any $\lambda \in [1,2N]$ such that $1-\xi-(1+\xi)g(\frac{\lambda}{\zeta N})\frac{\lambda}{\zeta N} > 0$,*

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

Comparing this results with corollary 6.3 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 $\alpha = 1 - \frac{1}{\log(N)}$.

7. 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 $f_{\widetilde{\theta}}$ 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 $f_{\widetilde{\theta}}$. The most favorable case is when $R(\widetilde{\theta}) = \inf_{\theta\in\Theta} R(\theta)$, but this condition will not be strictly imposed here. The case when $\widetilde{\theta} \notin \Theta$ makes no difference : it is covered by adding $\widetilde{\theta}$ to the parameter set Θ and extending the prior π putting $\pi(\widetilde{\theta}) = 0$. Of course $\widetilde{\theta}$, 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 $\mathbb{V}\mathrm{ar}_P$ denotes the variance with respect to P:

$$
\begin{aligned}
\overline{R}(\theta) &= P\big[Y \neq f_\theta(X)\big] - P\big[Y \neq f_{\widetilde{\theta}}(X)\big] \\
\overline{r}(\theta) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big[Y_i \neq f_\theta(X_i)\big] - \mathbb{1}\big[Y_i \neq f_{\widetilde{\theta}}(X_i)\big] \\
\overline{V}(\theta) &= \mathbb{V}\mathrm{ar}_P\Big\{\mathbb{1}\big[Y \neq f_\theta(X)\big] - \mathbb{1}\big[Y \neq f_{\widetilde{\theta}}(X)\big]\Big\} \\
\overline{R}(\theta \,|X) &= P\big[Y \neq f_\theta(X) \,|X\big] - P\big[Y \neq f_{\widetilde{\theta}}(X) \,|X\big] \\
\overline{V}(\theta \,|X) &= \mathbb{V}\mathrm{ar}_P\Big\{\mathbb{1}\big[Y \neq f_\theta(X)\big] - \mathbb{1}\big[Y \neq f_{\widetilde{\theta}}(X)\big] \,|X\Big\}.
\end{aligned}
$$

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

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

(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 $f_{\widetilde{\theta}}(x)$ 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 :

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

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[1]

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

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 4.1 in a way similar to what has already been done to establish lemma 4.3 in the previous section, we get some non localized learning lemma

LEMMA 7.1. *For any $\lambda \in \mathbb{R}_+$, any lower-bounded measurable function $\eta : \Theta \to \mathbb{R}$,*

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

In the same way

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

[1] My PhD student Jean-Yves Audibert made the remark that the following inequalities could be improved to:

$$\begin{aligned}
\overline{V}(\theta) &\leq P\Big\{ \Big[\mathbb{1}\big[Y \neq f_\theta(X)\big] - \mathbb{1}\big[Y \neq f_{\widetilde{\theta}}(X)\big]\Big]^2 \Big\} \\
&\leq P\Big[\mathbb{1}\big[f_\theta(X) \neq f_{\widetilde{\theta}}(X)\big]\Big] \\
&\leq P\left[\frac{\overline{R}(\theta\,|X)}{\alpha}\mathbb{1}(X \notin \Omega_\alpha) + \mathbb{1}(\Omega_\alpha)\right] \\
&\leq \frac{1}{\alpha}\overline{R}(\theta) + \frac{1}{\alpha}P(\Omega_0) + P(\Omega_\alpha).
\end{aligned}$$

Therefore the constants a and b can be improved to $a = \frac{1}{\alpha}$ and $b = \frac{1}{\alpha}P(\Omega_0) + P(\Omega_\alpha)$. Empirical bounds without margin assumptions (i.e. where the variance $\overline{V}(\theta)$ is bounded by some empirical quantity, whereas b depends on the unknown sample distribution P) are to be found in a forthcoming paper of Jean-Yves Audibert.

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

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

we get

COROLLARY 7.2. *For any $\lambda > 0$ such that $\kappa a\frac{\lambda}{N} < 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\rho\big[R(\theta)\big] \le R(\widetilde{\theta}) + \left(1 - \kappa a\frac{\lambda}{N}\right)^{-1}\left\{\rho\big[r(\theta)\big] - r(\widetilde{\theta}) + \frac{1}{\lambda}\Big[\mathcal{K}(\rho,\pi) + \log(\epsilon^{-1})\Big] + \kappa b\frac{\lambda}{N}\right\}. \tag{7.1}$$

Note that the right-hand side of inequality (7.1) is not an observable quantity. Anyhow, it differs 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 $\rho\big[R(\theta)\big] - R(\widetilde{\theta})$, the defect of optimality of the randomized estimator built from ρ, using the trivial bound $-r(\widetilde{\theta}) \le -\inf_{\theta\in\Theta} r(\theta)$. The optimal posterior for this bound is as before the Gibbs posterior $\hat{\rho}_\lambda$. It satisfies

COROLLARY 7.3. *For any $\lambda > 0$ such that $\kappa a\frac{\lambda}{N} < 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$,*

$$\hat{\rho}_\lambda\big[R(\theta)\big] \le R(\widetilde{\theta}) + \left(1 - \kappa a\frac{\lambda}{N}\right)^{-1}\left\{\frac{1}{\lambda}\int_0^\lambda \hat{\rho}_\beta\big[r(\theta)\big]d\beta - r(\widetilde{\theta}) + \frac{\log(\epsilon^{-1})}{\lambda} + \kappa b\frac{\lambda}{N}\right\}. \tag{7.2}$$

Moreover

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

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

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

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

$$P^{\otimes N}\Big\{\hat{\rho}_{\frac{N}{2a}}\big[R(\theta)\big]\Big\} \le R(\widetilde{\theta}) + \frac{4a\log\Big[\pi(\{\widetilde{\theta}\})^{-1}\Big]}{N}.$$

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

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

COROLLARY 7.4. *For any $\alpha > 1$, with $P^{\otimes N}$ at least $1-\epsilon$, for any posterior distribution $\rho \in \mathcal{M}_+^1(\Theta)$,*

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

where $\kappa = g(\frac{2\alpha\lambda}{N})$, and the infimum in λ is restricted to those values for which $\kappa a \frac{\alpha\lambda}{N} < 1$.

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

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

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

where $\kappa = g(\frac{2\lambda}{N})$ and the parameters are such that $\lambda - \beta - \kappa a \frac{\lambda^2}{N} > 0$.

LEMMA 7.5. *With $P^{\otimes N}$ probability at least $1-\epsilon$,*

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

In the same way,

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

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

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

which can be rewritten as

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

Coming back to lemma 7.5, we obtain

COROLLARY 7.6. *For any $\lambda > 0$ and any $\xi \in [0,1[$ such that $1 - \xi - (1+\xi)\kappa a \frac{\lambda}{N} > 0$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any posterior $\rho \in \mathcal{M}_+^1$,*

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

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

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

For a fixed value of ξ, getting a uniform result in λ is achieved as in the case of theorem 6.6:

COROLLARY 7.7. *For any $\xi \in [0,1[$, any $\zeta \in [\xi, 1[$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any $\lambda \in [1, 2N]$ such that $1 - \xi - (1+\xi) g(\frac{2\lambda}{\kappa N}) \kappa a \frac{\lambda}{\zeta N} > 0$,*

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

The same remarks which were made about theorem 6.6 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 $-r(\widetilde{\theta}) \leq -\inf_{\theta\in\Theta} r(\theta)$.

CHAPTER 2

Learning with an exchangeable prior

1. The Vapnik Cervonenkis dimension of a family of subsets

Let us consider some set X and some set $S \subset \{0,1\}^X$ of subsets of X. Let $h(S)$ be the VC dimension of S, defined as

$$h(S) = \max\{|A| : A \text{ finite and } A \cap S = \{0,1\}^A\},$$

where by definition $A \cap S = \{A \cap B : B \in S\}$. Let us notice that this definition does not depend on the choice of the reference set X. Indeed X can be chosen to be $\bigcup S$, the union of all the sets in S or any bigger set. Let us notice also that for any set B, $h(B \cap S) \leq h(S)$, the reason being that $A \cap (B \cap S) = B \cap (A \cap S)$.

This notion of VC dimension is useful because it can, as we will see about support vector machines, be computed in some important special cases. Let us prove here as an illustration that $h(S) = d + 1$ when $X = \mathbb{R}^d$ and S is made of all the half spaces :

$$S = \{A_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}, \text{ where } A_{w,b} = \{x \in X : \langle w, x\rangle \geq b\}.$$

PROPOSITION 1.1. *With the previous notations,* $h(S) = d + 1$.

PROOF. Let $(e_i)_{i=1}^{d+1}$ be the canonical base of $\mathbb{R}^{d+1}$, and let X be the affine subspace it generates, which can be identified with $\mathbb{R}^d$. For any $(\epsilon_i)_{i=1}^{d+1} \in \{-1,+1\}^{d+1}$, let $w = \sum_{i=1}^{d+1} \epsilon_i e_i$ and $b = 0$. The half space $A_{w,b} \cap X$ is such that $\{e_i\,; i = 1,\ldots,d+1\} \cap (A_{w,b} \cap X) = \{e_i\,; \epsilon_i = +1\}$. This proves that $h(S) \geq d + 1$.

To prove that $h(S) \leq d + 1$, we have to show that for any set $A \subset \mathbb{R}^d$ of size $|A| = d + 2$, there is $B \subset A$ such that $B \notin (A \cap S)$. This will obviously be the case if the convex hulls of B and $A \setminus B$ have a non empty intersection : indeed if a hyperplane separates two sets of points, it also separates their convex hulls. As $|A| > d + 1$, A is affine dependent : there is $(\lambda_x)_{x\in A} \in \mathbb{R}^{d+2} \setminus \{0\}$ such that $\sum_{x\in A} \lambda_x x = 0$ and $\sum_{x\in A} \lambda_x = 0$. The set $B = \{x \in A : \lambda_x > 0\}$ is non-empty, as well as its complement $A \setminus B$, because $\sum_{x\in A} \lambda_x = 0$ and $\lambda \neq 0$. Moreover $\sum_{x\in B} \lambda_x = \sum_{x\in A\setminus B} -\lambda_x > 0$. The relation

$$\frac{1}{\sum_{x\in B} \lambda_x} \sum_{x\in B} \lambda_x x = \frac{1}{\sum_{x\in B} \lambda_x} \sum_{x\in A\setminus B} -\lambda_x x$$

shows that the convex hulls of B and $A \setminus B$ have a non void intersection. □

Let us introduce the function of two integers

$$\Phi_n^h = \sum_{k=0}^{h} \binom{n}{k}$$

Let us notice that Φ can alternatively be defined by the relations :

$$\Phi_n^h = \begin{cases} 2^n & \text{when } n \leq h, \\ \Phi_{n-1}^{h-1} + \Phi_{n-1}^h & \text{when } n > h. \end{cases}$$

THEOREM 1.2. *Whenever $\bigcup S$ is finite,*

$$|S| \leq \Phi\left(\left|\bigcup S\right|, h(S)\right).$$

THEOREM 1.3. *For any $h \leq n$,*

$$\Phi_n^h \leq \exp\bigl(nH(\tfrac{h}{n})\bigr) \leq \exp\bigl[h\bigl(\log(\tfrac{n}{h}) + 1\bigr)\bigr],$$

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

PROOF OF THEOREM 1.2. Let us prove this theorem by induction on $|\bigcup S|$. It is easy to check that it holds true when $|\bigcup S| = 1$. Let $X = \bigcup S$, let $x \in X$ and $X' = X \setminus \{x\}$. Define ($\triangle$ denoting the symmetric difference of two sets)

$$S' = \{A \in S : A \triangle \{x\} \in S\},$$
$$S'' = \{A \in S : A \triangle \{x\} \notin S\}.$$

Clearly, $\sqcup$ denoting the disjoint union, $S = S' \sqcup S''$ and $S \cap X' = (S' \cap X') \sqcup (S'' \cap X')$. Moreover $|S'| = 2|S' \cap X'|$ and $|S''| = |S'' \cap X'|$. Thus $|S| = |S'| + |S''| = 2|S' \cap X'| + |S''| = |S \cap X'| + |S' \cap X'|$. Obviously $h(S \cap X') \leq h(S)$. Moreover $h(S' \cap X') = h(S') - 1$, because if $A \subset X'$ is shattered by S' (or equivalently by $S' \cap X'$), then $A \cup \{x\}$ is shattered by S' (we say that A is shattered by S when $S \cap A = \{0,1\}^A$). Using the induction hypothesis, we then see that $|S \cap X'| \leq \Phi_{|X'|}^{h(S)} + \Phi_{|X'|}^{h(S)-1}$. But as $|X'| = |X| - 1$, the righthand side of this inequality is equal to $\Phi_{|X|}^{h(S)}$, according to the recurrence equation satisfyied by Φ. □

PROOF OF THEOREM 1.3. This is the well known Chernoff bound for the deviation of sums of Bernoulli r.v.: let $(\sigma_1, \ldots, \sigma_n)$ be i.i.d. Bernoulli r.v. with parameter $1/2$. Let us notice that

$$\Phi_n^h = 2^n \mathbb{P}\left(\sum_{i=1}^n \sigma_i \leq h\right).$$

For any positive real number λ ,

$$\begin{aligned} \mathbb{P}(\sum_{i=1}^n \sigma_i \leq h) &\leq \exp(\lambda h)\mathbb{E}\left[\exp\left(-\lambda \sum_{i=1}^n \sigma_i\right)\right] \\ &= \exp\Bigl\{\lambda h + n\log\bigl\{\mathbb{E}\bigl[\exp\bigl(-\lambda\sigma_1\bigr)\bigr]\bigr\}\Bigr\}. \end{aligned}$$

Differentiating the right-hand side in λ shows that its minimal value is $\exp\bigl[-n\mathcal{K}(\frac{h}{n}, \frac{1}{2})\bigr]$, where $\mathcal{K}(p,q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ is the Kullback divergence function between two Bernoulli distributions of parameters p and q. The announced result then follows from the identity

$$\begin{aligned} H(p) &= \log(2) - \mathcal{K}(p, \tfrac{1}{2}) \\ &= p\log(p^{-1}) + (1-p)\log(1 + \frac{p}{1-p}) \leq p\bigl[\log(p^{-1}) + 1\bigr]. \end{aligned}$$

□

2. Non localized bounds

In this chapter 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 let as usual $(X_i,Y_i)_{i=1}^{2N}$ be the canonical process on $(\mathcal{X},\mathcal{Y})^{2N}$. We let also $(Z_i)_{i=1}^{2N}=(X_i,Y_i)_{i=1}^{2N}$ and $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$. When we require that P_{2N} is exchangeable, we mean that for any permutation $\sigma\in\mathfrak{S}_{2N}$, the distribution of $(X_{\sigma(i)},Y_{\sigma(i)})_{i=1}^{2N}$ under P_{2N} is the same as the distribuiton of $(X_i,Y_i)_{i=1}^{2N}$. 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 : we have at our disposal a training set and a test set of the same size. Taking into account a test set in the design of the classification rule is what V. Vapnik [**21**] calls *transductive* statistical inference. As mentioned by V. Vapnik and as it will be seen here, this is a very fruitful framework, allowing a mix of supervised and unsupervised learning.

Starting with a family $\{f_\theta:\mathcal{X}\to\mathcal{Y};\theta\in\Theta\}$ of classification rules, we would like to minimize the error rate on the test set

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

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

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

LEMMA 2.1. *For any $\lambda\in\mathbb{R}_+$, any lower-bounded measurable function $\eta:(\mathcal{X}\times\mathcal{Y})^{2N}\times\Theta\to\mathbb{R}$ which is exchangeable with respect to its $2\times 2N$ first arguments, any $\theta\in\Theta$,*

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

More precisely the exchangeability hypothesis on η is that for any permutation $\sigma\in\mathfrak{S}_{2N}$, for any $(x_i,y_i)_{i=1}^{2N}\in(\mathcal{X}\times\mathcal{Y})^{2N}$, any $\theta\in\Theta$,

$$\eta\big((x_i,y_i)_{i=1}^{2N},\theta\big)=\eta\big((x_{\sigma(i)},y_{\sigma(i)})_{i=1}^{2N},\theta\big).$$

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

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

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

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

$$\leq P_{2N}\left\{ \exp\left[\frac{\lambda^2}{2N^2} \sum_{k=1}^{N} \bigl[\sigma_{k+N}(\theta) - \sigma_k(\theta)\bigr]^2 - \eta(\theta) \right] \right\}$$

$$\leq P_{2N}\left\{ \exp\left[\frac{\lambda^2}{2N^2} \sum_{k=1}^{N} \bigl[\sigma_{k+N}(\theta) + \sigma_k(\theta)\bigr] - \eta(\theta) \right] \right\}$$

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

□

Let us now consider some exchangeable random probability measure $\pi : (\mathcal{X} \times \mathcal{Y})^{2N} \to \mathcal{M}_+^1(\Theta)$. (We will assume that $(\Theta, \mathcal{T})$ is a Polish space and that π is a regular conditional probability measure. Moreover, in practice, interesting exchangeable priors will depend only on $(X_1, \ldots, X_{2N})$, although the forthcoming bounds do not preclude them to depend also on $(Y_1, \ldots, Y_{2N})$.) What we mean here is that the *function* $Z_1^{2N} \mapsto \pi\bigl(Z_1^{2N}\bigr)$ is exchangeable — i.e. that for any permutation $\sigma \in \mathfrak{S}_{2N}$, $\pi\bigl(Z_1^{2N}\bigr) = \pi\bigl[(Z_{\sigma_i})_{i=1}^{2N}\bigr]$ in $\mathcal{M}_+^1(\Theta)$. Integrating the previous deviation lemma with respect to π, we get

LEMMA 2.2. *For any $\lambda \in \mathbb{R}_+$, any lower-bounded measurable function $\eta : (\mathcal{X} \times \mathcal{Y})^{2N} \times \Theta \to \mathbb{R}$ which is exchangeable with respect to its $2 \times 2N$ first arguments,*

$$P_{2N}\left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\rho\bigl[r_2(\theta)\bigr] - \lambda\rho\bigl[r_1(\theta)\bigr] - \rho\bigl[\eta(\theta)\bigr] - \mathcal{K}(\rho, \theta) \geq 0 \right\}$$

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

PROOF. For any $Z_1^{2N} \in (\mathcal{X} \times \mathcal{Y})^{2N}$, let us consider the exchangeable probability distribution

$$\bar{P}_{Z_1^{2N}} = \frac{1}{|\mathfrak{S}_{2N}|} \sum_{\sigma \in \mathfrak{S}_{2N}} \delta_{(Z_{\sigma(i)})_{i=1}^{2N}}.$$

The fact that P_{2N} is exchangeable is equivalent to the identity

$$P_{2N}(h) = P_{2N}\bigl[\bar{P}_{Z_1^{2N}}(h)\bigr], \text{ for any bounded measurable function } h : \mathcal{Z}^{2N} \to \mathbb{R}.$$

Using this identity, which shows that $\bar{P}_{Z_1^{2N}}$ is the probability distribution P_{2N} conditioned by $\sum_{k=1}^{2N} \delta_{Z_k}$, we obtain a decomposition of P_{2N} into exchangeable distributions for which π and η are almost surely constant. Let us detail the beginning of the proof:

$$P_{2N}\left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\rho\bigl[r_2(\theta)\bigr] - \lambda\rho\bigl[r_1(\theta)\bigr] - \rho\bigl[\eta(\theta)\bigr] - \mathcal{K}(\rho, \theta) \geq 0 \right\}$$

$$\leq P_{2N}\left\{ \pi\Bigl[\exp\bigl[\lambda r_2(\theta) - \lambda r_1(\theta) - \eta(\theta)\bigr] \Bigr] \right\}$$

$$= P_{2N}\left\{ \bar{P}_{Z_1^{2N}}\left[\pi\Bigl[\exp\bigl[\lambda r_2(\theta) - \lambda r_1(\theta) - \eta(\theta)\bigr] \Bigr] \right] \right\}$$

$$= P_{2N}\left\{ \pi\left[\bar{P}_{Z_1^{2N}}\Bigl[\exp\bigl[\lambda r_2(\theta) - \lambda r_1(\theta) - \eta(\theta)\bigr] \Bigr] \right] \right\}.$$

Now we can apply lemma 2.1 to $\bar{P}_{Z_1^{2N}}$ and exchange once again the expectations with respect to this measure and with respect to π — using Fubini's theorem — to prove what is claimed. □

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

COROLLARY 2.3. *For any $\lambda \in]0, 2N[$, with P^{2N} probability at least $1-\epsilon$, for any posterior distribution $\rho \in \mathcal{M}_+^1(\Theta)$,*

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

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

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

COROLLARY 2.4. *For any $\lambda \in]0, 2N[$, with P_{2N} probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

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

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

$$\log\big[N(X_1^{2N})\big] \leq 2NH(\tfrac{h}{2N}) \leq h\big[\log(\tfrac{2N}{h}) + 1\big],$$

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

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

is the VC dimension of the set $\big\{\{X_1, \ldots, X_{2N}\} \cap f_\theta^{-1}(1)\,; \theta \in \Theta\big\}$ of subsets of $\{X_1, \ldots, X_{2N}\}$.

PROOF. Let

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

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 $\theta' \in \Theta'$,

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

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 $r_2(\theta) = r_2(\theta')$. □

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

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

where we have put $Z_1^N = (X_k, Y_k)_{k=1}^N$ for short and assumed that there exists a regular version of the conditional probability distribution $P_{2N}\big(\cdot\big|Z_1^N\big)$.

To this purpose we can use a variant of lemma 2.2:

LEMMA 2.5. *For any $\lambda \in \mathbb{R}_+$,*

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

PROOF. Let

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

and

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

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

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

as required. □

Choosing in lemma 2.5

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

we get

COROLLARY 2.6. *For any $\lambda \in]0, 2N[$, 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 $\rho : \mathcal{X}^{2N} \times \mathcal{Y}^{2N} \to \mathcal{M}_+^1(\Theta)$,*

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

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 2.7. *For any $\lambda \in]0, 2N[$, with $P_{2N}(dZ_1^N)$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R_2(\theta) \leq \bigg(\lambda - \frac{\lambda^2}{2N}\bigg)^{-1}\bigg\{\bigg(\lambda+\frac{\lambda^2}{2N}\bigg)r_1(\theta) + P_{2N}\Big\{\log\big[N(X_1^{2N})\big]\,|Z_1^N\Big\} + \log(\epsilon^{-1})\bigg\}.$$

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

PROOF. This is an integrated variant of corollary 2.4. With the same notations, we can define θ' with values in Θ', such that $\Psi(\theta) = \Psi(\theta')$. The proof for θ' is a direct consequence of the preceding corollary, and the proof for θ comes from the fact that $r_1(\theta) = r_1(\theta')$ and $r_2(\theta) = r_2(\theta')$. □

Let us remark that theorem 2.7 can easily be made uniform in λ. Let us choose some parameter $\zeta > 1$ and consider the set of values

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

Using the fact that for any $\lambda \in [1, 2N]$ there is $\lambda' \in \Lambda$ such that $\lambda \leq \lambda' \leq \zeta\lambda$, we can establish that with $P_{2N}(dZ_1^N)$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,

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

A simple computation shows that for any positive constants a, b, c, d,

$$\inf_{\lambda \in]0, a^{-1}[} (1 - a\lambda)^{-1}(b + c\lambda + d\lambda^{-1}) = b + 2ad + \sqrt{d(ab + a^2 d + c)}, \tag{2.1}$$

this infimum being reached when $\lambda^{-1} = \lambda_*^{-1} = a + \sqrt{a^2 + \frac{ab+c}{d}}$. In our case $a = \frac{\zeta}{2N}$, $b = r_1(\theta) \leq 1$, $c = \frac{\zeta}{2N} r_1(\theta) = ab$ and

$$d = P_{2N}\left\{ \log\left[N(X_1^{2N})\right] | Z_1^N \right\} + \log\left[\epsilon^{-1}\left(\frac{\log(2N)}{\log(\zeta)} + 1\right)\right] \geq 1 \tag{2.2}$$

as soon as $\epsilon \leq e^{-1}$. Thus $\lambda_*^{-1} \leq 1$ as soon as $N \geq 4\zeta$ and $\epsilon \leq e^{-1}$.

THEOREM 2.8. *For any $\zeta > 1$, any $\epsilon \leq e^{-1}$, any integer $N \geq 4\zeta$, with P_{2N} probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$R_2(\theta) \leq r_1(\theta) + \frac{\zeta d}{N}\left(1 + \sqrt{\frac{1}{4} + \frac{N r_1(\theta)}{\zeta d}}\right),$$

where d is defined by equation (2.2).

REMARK 2.1. This result is to be compared with Vapnik's one (see [**21**, page 138]), which, in the i.i.d. case when $P_{2N} = P^{\otimes 2N}$, says that with $P^{\otimes N}$ probability at least $1 - \epsilon$,

$$R(\theta) \leq r_1(\theta) + \frac{2d'}{N}\left(1 + \sqrt{1 + \frac{N r_1(\theta)}{d'}}\right),$$

where

$$d' = \log\left\{ P^{\otimes 2N}\left[N\left(X_1^{2N}\right)\right]\right\} + \log(4\epsilon^{-1}).$$

We see that we obtain better constants for a large range of values of N and the replacement of the annealed entropy by something closer to the VC entropy. The link with the VC entropy is enlightened by the simple inequality $\log\left[N(X_1^{2N})\right] \leq \log\left[N(X_1^N)\right] + \log\left[N(X_{N+1}^{2N})\right]$, which leads to

$$P_{2N}\left\{ \log\left[N(X_1^{2N}\right] | Z_1^N \right\} \leq \log\left[N(X_1^N)\right] + P_{2N}\left\{ \log\left[N(X_{N+1}^{2N})\right]\right\}.$$

Proving some variant of Vapnik's theory is not the only possible use of corollary 2.3. Its right-hand side can also be optimized choosing for ρ the Gibbs distribution

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

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

COROLLARY 2.9. *For any $\lambda \in]0, 2N[$, with P_{2N} probability at least $1-\epsilon$,*

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

3. 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.

3.1. Compression schemes. Let us explore first the idea of compression schemes, put forward by Littlestone and Warmuth [**22, 16**].

Let us consider some measurable training rule

$$\hat{f} : \bigcup_{n=1}^{+\infty} \left(\mathcal{X} \times \mathcal{Y}\right)^n \times \mathcal{X} \to \mathcal{Y},$$

which produces for any size of problem n and any training set $Z' = (x'_i, y'_i)_{i=1}^n \in \left(\mathcal{X} \times \mathcal{Y}\right)$ some classifier $\hat{f}_{Z'} : \mathcal{X} \to \mathcal{Y}$. Let us assume that $\hat{f}_{Z'}$ is invariant under any permutation of the indices of the training set, as it is usually the case with estimators designed for i.i.d. or exchangeable samples.

A sample $Z = (X_i, Y_i)_{i=1}^{2N}$ being given, it is natural to build in this case the following model: for any $h = 1, \ldots, N$

$$\mathcal{R}_h = \left\{ \hat{f}_{(x'_i, y'_i)_{i=1}^h} : \{x'_i : 1 \leq i \leq h\} \subset \{X_i : 1 \leq i \leq 2N\}, (y'_i)_{i=1}^h \in \mathcal{Y}^h \right\}.$$

We will consider the union of all these models : $\mathcal{R} = \bigcup_{h=1}^N \mathcal{R}_h$. Our exchangeable prior will be uniform on each $\mathcal{R}_h$, and such that for some parameter $\alpha \in]0,1[$, $\pi(\mathcal{R}_h) \geq (1-\alpha)\alpha^h$. It is easy to see that

$$\log(|\mathcal{R}_h|) = \log\left[\binom{2N}{h} |\mathcal{Y}|^h\right] \leq h\left[\log\left(\frac{2N}{h}\right) + 1 + \log\left(|\mathcal{Y}|\right)\right].$$

We are ready to apply corollary 2.3, in the case when we observe a training set $(X_i, Y_i)_{i=1}^N$ and consider a test set $(X_i, Y_i)_{i=N+1}^{2N}$ under the exchangeable distribution $P_{2N} \in \mathcal{M}_+^1\left[(\mathcal{X} \times \mathcal{Y})^{2N}\right]$. According to this corollary, for any $\lambda \in]0, 2N[$, with P_{2N} probability at least $1-\epsilon$, for any $h = 1, \ldots, N$, any $f \in \mathcal{R}_h$,

$$r_2(f) \leq \left(1 - \frac{\lambda}{2N}\right)^{-1} \left\{ \left(1 + \frac{\lambda}{2N}\right) r_1(f) \right.$$
$$\left. + \frac{1}{\lambda}\left[-\log(1-\alpha) + h\left[\log\left(\tfrac{2N}{h}\right) + 1 + \log\left(|\mathcal{Y}|\right) - \log(\alpha)\right] + \log(\epsilon^{-1}) \right] \right\},$$

where as usual

$$r_1(f) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big[f(X_i) \neq Y_i\big],$$

$$\text{and} \quad r_2(f) = \frac{1}{N}\sum_{i=N+1}^{2N} \mathbb{1}\big[f(X_i) \neq Y_i\big].$$

The next useful step is to make this statement uniform in $\lambda \in [1, 2N]$. As explained earlier in these lectures, this can be achieved by considering for some real parameter $\zeta > 1$ a grid of values $\Lambda = \big\{2N\zeta^{-k} : 0 \leq k \leq \frac{\log(2N)}{\log(\zeta)}\big\}$. Applying the previous inequality of any $\lambda \in \Lambda$, we get

PROPOSITION 3.1. *With P_{2N} probability at least $1-\epsilon$ for any $h = 1, \dots, 2N$, for any $f \in \mathcal{R}_h$*

$$r_2(f) \leq \inf_{\lambda \in [1, 2N[} B(\lambda, h, f),$$

where

$$B(\lambda, h, f) = \left(1 - \frac{\zeta\lambda}{2N}\right)^{-1} \left\{ \left(1 + \frac{\zeta\lambda}{2N}\right) r_1(f) \right. \tag{3.1}$$
$$\left. + \frac{1}{\lambda}\left[-\log(1-\alpha) + h\Big[\log\big(\tfrac{2N}{h}\big) + 1 + \log(|\mathcal{Y}|) - \log(\alpha)\Big] + \log(\epsilon^{-1}) + \log\Big[\tfrac{\log(2N)}{\log(\zeta)} + 1\Big]\right]\right\}.$$

(Note that the optimization in λ can easily be carried through explicitely using equation 2.1, note also that we do not need to observe the test set $(X_i, Y_i)_{i=N+1}^{2N}$ to compute this bound when $f = f_{Z'}$ with $Z' \subset \{X_i : 1 \leq i \leq N\}$.)

An adaptative estimator $\hat{f}_a$ can then be built by minimizing (3.1). Let us discuss the slightly less obvious case when the test set is not observed. Let $\hat{\mathcal{R}}_h$ be the observable part of $\mathcal{R}_h$, more precisely let

$$\hat{\mathcal{R}}_h = \big\{\hat{f}_{(x_i', y_i')_{i=1}^h} : \{x_i' : 1 \leq i \leq h\} \subset \{X_i : 1 \leq i \leq N\}, (y_i')_{i=1}^h \in \mathcal{Y}_h\big\}.$$

Let us choose

$$\hat{h} \in \arg\min_{h=1,\dots,N} \inf\big\{B(\lambda, h, f), \lambda \in [1, 2N], f \in \hat{\mathcal{R}}_h\big\}$$
$$\hat{f}_a \in \arg\min_{f \in \hat{\mathcal{R}}_{\hat{h}}} \inf_{\lambda \in [1, 2N[} B(\lambda, \hat{h}, f).$$

PROPOSITION 3.2. *With the previous notations*

$$r_2(\hat{f}_a) \leq \inf\big\{B(\lambda, h, f) : \lambda \in [1, 2N[, h \in [1, N], f \in \hat{\mathcal{R}}_h\big\}.$$

Note that this learning scheme is different from cross validation, since, although we restrict ourselves to choosing $\hat{f}$ as a function of $(x_i', y_i')_{i=1}^h$ only, *we are allowed to choose $(x_i', y_i')_{i=1}^h$ 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 suitable way.*

Compression schemes in the i.i.d. case can be handled using a clever idea of M. Seeger [**28**]. We will dedicate chapter 4 to this interesting approach leading to tighter bounds than the ones presented here above (through the use of Gibbs posterior distributions computable from the training set $(X_i, Y_i)_{i=1}^N$ only).

3.2. Pruning decision trees. One possible use of compression schemes 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 "critical") examples $(x'_1, \dots, x'_h)$ drawn from $(X_1, \dots, X_N)$, we may build a pruned decision tree by stopping to ask questions as soon as only one example in $(x'_i)_{i=1}^h$ matches the query. Using proposition 3.1 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 a deterministic prior on trees). But we can do better : we can also prune inner nodes by deciding to remove questions which do not split $(x'_i)_{i=1}^h$, 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 "compression" set $(x'_i)_{i=1}^h$ (for instance we can choose a set of questions leading to a balanced tree). We can also use the labels $(y'_i)_{i=1}^h$ 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 $(x'_i, y'_i)_{i=1}^h$ 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'_i, y'_i) (using proposition 3.1). This gives a theoretical framework to guide the implementation of many algorithmic ideas in a data driven way.

3.3. Some boosting algorithm for compression schemes. Finding out the minimum of the bound $B(\lambda, h, f)$ may be difficult. We can instead use some suboptimal heuristic mimicking the boosting algorithm. It goes in the following way : find out first $f_{Z'_2} \in \hat{\mathcal{R}}_2$ minimizing $\inf_{\lambda \in [1,2N[} B(\lambda, 2, f)$ on $\hat{\mathcal{R}}_2$. As Z'_2 has only two elements, the complexity of this search is at worst proportional to N^2. Moreover, this boils down to minimizing $r(f)$ on the given set. Then add to the compression set Z'_2 some example to form a compression set Z'_3 of size three which minimizes $\inf_{\lambda \in [1,2N[} B(\lambda, 3, f_{Z'_3})$. Here again, this is equivalent to minimizing $r(f)$. More generally, Z'_h being formed, form Z'_{h+1} by adding to Z'_h the example which minimizes $r(f)$, and therefore also $\inf_{\lambda \in [1,2N[} B(\lambda, h+1, f_{Z'_{h+1}})$. The search can be restricted further by considering only to add examples which were misclassified at the previous step. Doing so, only a limited number of examples is search at each step. The algorithm may be stopped as soon as the criterion increases from one step to the next. (A more costly alternative is to scan all the possible values of h and to retain in the end the optimal value of h).

3.4. The transductive case. If we are in the so called "transductive learning" situation, where $(X_i)_{i=1}^{2N}$ and $(Y_i)_{i=1}^N$ are observed (the estimator has to be applied to a known batch of test examples at least of the same size as the training set), we can use corollary 2.9 instead of corollary 2.4. Indeed, in this situation, for any h, the whole model $\mathcal{R}_h$ is observable, and therefore the Gibbs posterior distributions $\hat{\rho}_\beta$ can be computed.

These Gibbs distributions can in particular be approximated using some Metropolis algorithm (see [**15**] for more details) where the coordinates of $(x'_i)_{i=1}^h$ and $(y'_i)_{i=1}^h$ are moved one at a time, and where additions and deletion of coordinates are also allowed to move from one $\mathcal{R}_h$ to $\mathcal{R}_{h-1}$ and $\mathcal{R}_{h+1}$.

4. Localization

We can localize our results for exchangeable priors as we had done in previously encountered situations. To achieve this, let us apply lemma 2.2 with

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

where the positive parameters λ and β are chosen in such a way that $\lambda - \beta - \frac{\lambda^2}{2N} > 0$.

LEMMA 4.1. *For any positive parameters λ and β such that $\lambda - \beta - \frac{\lambda^2}{2N} > 0$, with P_{2N} probability at least $1-\epsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,*

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

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

$$\log\left\{\pi\left[\exp\Big[-\beta\big[r_1(\theta) + r_2(\theta)\big]\Big]\right]\right\} = \sup_{\rho\in\mathcal{M}_+^1} -\beta\rho\big[r_1(\theta)\big] - \beta\rho\big[r_2(\theta)\big] - \mathcal{K}(\rho,\pi)$$

$$\le \sup_{\rho\in\mathcal{M}_+^1(\Theta)} -\beta\rho\big[r_1(\theta)\big] - \mathcal{K}(\rho,\pi) - \beta\left(\lambda + \beta + \frac{\lambda^2}{2N}\right)^{-1}\left\{\left(\lambda - \beta - \frac{\lambda^2}{2N}\right)\rho\big[r_1(\theta)\big] - \log\left\{\pi\Big[\exp\big\{-\beta\big[r_1(\theta) + r_2(\theta)\big]\big\}\Big]\right\} - \mathcal{K}(\rho,\pi) - \log(\epsilon^{-1})\right\}.$$

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

$$\log\left\{\pi\left[\exp\Big[-\beta\big[r_1(\theta) + r_2(\theta)\big]\Big]\right]\right\} \le \sup_{\rho\in\mathcal{M}_+^1(\Theta)} -2\xi\lambda\rho\big[r_1(\theta)\big] - \mathcal{K}(\rho,\pi) + \xi\log(\epsilon^{-1}) = \log\left\{\pi\Big[\exp\big[-2\xi\lambda r_1(\theta)\big]\Big]\right\} + \xi\log(\epsilon^{-1}). \tag{4.1}$$

This leads to

LEMMA 4.2. *For any positive parameter λ and any $\xi \in [0,1[$ such that $(1-\xi)\lambda - (1+\xi)\frac{\lambda^2}{2N} > 0$, with P_{2N} probability at least $1-\epsilon$, for any $\rho \in \mathcal{M}_+^1$,*

$$\rho\big[r_2(\theta)\big] \le \left[(1-\xi)\lambda - (1+\xi)\frac{\lambda^2}{2N}\right]^{-1}\left\{(1+\xi)\lambda\left(1 + \frac{\lambda}{2N}\right)\rho\big[r_1(\theta)\big] + \mathcal{K}(\rho,\pi) + \log\left\{\pi\Big[\exp\big[-2\xi\lambda r_1(\theta)\big]\Big]\right\} + (1+\xi)\log\big(\tfrac{2}{\epsilon}\big)\right\}$$

$$= \left[(1-\xi)\lambda - (1+\xi)\frac{\lambda^2}{2N}\right]^{-1}\left\{\left[(1-\xi)\lambda + (1+\xi)\frac{\lambda^2}{2N}\right]\rho\big[r_1(\theta)\big] + \mathcal{K}(\rho,\hat\rho_{2\xi\lambda}) + (1+\xi)\log\big(\tfrac{2}{\epsilon}\big)\right\}.$$

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

THEOREM 4.3. *For any $\lambda > 0$ and $\xi \in [0,1[$ such that $(1-\xi)\lambda - (1+\xi)\frac{\lambda^2}{2N} > 0$, with P_{2N} probability at least $1-\epsilon$,*

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

In the same way, the following suboptimal but simpler inequality obtained from cancelling the localized divergence term also holds as soon as $\frac{(1+\xi)\lambda}{4\xi(1-\xi)N} < 1$, with P_{2N} probability at least $1-\epsilon$:

$$\hat{\rho}_\lambda\big[r_2(\theta)\big] \leq \frac{\left[1 + \dfrac{(1+\xi)\lambda}{4\xi(1-\xi)N}\right]\hat{\rho}_\lambda\big[r_1(\theta)\big] + \dfrac{2\xi(1+\xi)}{(1-\xi)\lambda}\log\Big(\dfrac{2}{\epsilon}\Big)}{1 - \dfrac{(1+\xi)\lambda}{4\xi(1-\xi)N}}.$$

Choosing $\xi = 8^{-1/2}$ we can eventually weaken this last inequality to[1]

$$\hat{\rho}_\lambda\big[r_2(\theta)\big] \leq \frac{\left(1 + \dfrac{3\lambda}{2N}\right)\hat{\rho}_\lambda\big[r_1(\theta)\big] + \dfrac{3}{2\lambda}\log\Big(\dfrac{2}{\epsilon}\Big)}{1 - \dfrac{3\lambda}{2N}}. \tag{4.2}$$

[1]The constant 3/2 could be replaced with the slightly better constant $\frac{9\sqrt{2}+8}{14} \simeq 1.48$ in the three places where it appears in equation (4.2).

CHAPTER 3

Noisy classification with an exchangeable prior

1. Non localized bound

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 $\big((\mathcal{X}\times\mathcal{Y})^{2N}, (\mathcal{B}\otimes\mathcal{B}')^{\otimes 2N}\big)$. Similarly to what has been done in section 7 we consider some fixed (and unknown) parameter $\widetilde{\theta}\in\Theta$ and define

$$\begin{aligned}
\sigma_k(\theta) &= \mathbb{1}\big[Y_k \neq f_\theta(X_k)\big]\\
\overline{r}_1(\theta) &= \frac{1}{N}\sum_{k=1}^{N} \sigma_k(\theta) - \sigma_k(\widetilde{\theta})\\
\overline{r}_2(\theta) &= \frac{1}{N}\sum_{k=N+1}^{2N} \sigma_k(\theta) - \sigma_k(\widetilde{\theta})\\
\overline{R}(\theta\,|X_k) &= P\big[\sigma_k(\theta) - \sigma_k(\widetilde{\theta})\,|X_k\big]\\
r_1'(\theta) &= \frac{1}{N}\sum_{k=1}^{N} \overline{R}(\theta\,|X_k),\\
r_2'(\theta) &= \frac{1}{N}\sum_{k=N+1}^{2N} \overline{R}(\theta\,|X_k).
\end{aligned}$$

For the sake of simplicity, we will assume that the rule $f_{\widetilde{\theta}}$ 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 $x\in\mathcal{X}$,

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

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

$$P^{\otimes 2N}\Big\{\exp\big[\lambda\overline{r}_2(\theta)-\beta\overline{r}_1(\theta)\big]\,\big|X_1^{2N}\Big\} \leq \exp\Big[(\lambda+\kappa\tfrac{\lambda^2}{N})r_2'(\theta)-(\beta-\kappa\tfrac{\beta^2}{N})r_1'(\theta)\Big].$$

Moreover, putting

$$\lambda' = \lambda+\kappa\frac{\lambda^2}{N}$$

$$\beta' = \beta - \kappa \frac{\beta^2}{N},$$

we obtain

$$P^{\otimes 2N}\Big\{ \exp\Big[\lambda' r'_2(\theta) - \beta' r'_1(\theta)\Big] \, | \sum_{k=1}^{2N} \delta_{X_k} \Big\}$$
$$\leq P^{\otimes 2N}\Big\{ \exp\Big\{ \big[\tfrac{1}{2N}(\tfrac{\lambda'+\beta'}{2})^2 - \tfrac{\beta'-\lambda'}{2}\big]\big[r'_1(\theta) + r'_2(\theta)\big]\Big\} \, | \sum_{k=1}^{2N} \delta_{X_k} \Big\}.$$

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

LEMMA 1.1. *For the choice of parameters explained above, with $P^{\otimes 2N}$ probability at least $1-\epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\lambda \rho\big[\overline{r}_2(\theta)\big] \leq \beta \rho\big[\overline{r}_1(\theta)\big] + \mathcal{K}(\rho,\pi)$$
$$+ \log\Big\{ \pi\Big[\exp\Big\{ - \underbrace{\big[(\tfrac{\beta'-\lambda'}{2} - \tfrac{1}{2N}(\tfrac{\beta'+\lambda'}{2})^2\big]}_{\stackrel{\text{def}}{=}\beta''} \big[r'_1(\theta) + r'_2(\theta)\big]\Big\}\Big]\Big\} + \log(\epsilon^{-1}).$$

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

$$\lambda' \rho\big[r'_2(\theta)\big] \leq \beta' \rho\big[r'_1(\theta)\big] + \mathcal{K}(\rho,\pi)$$
$$+ \log\Big\{ \pi\Big[\pi\Big\{ \exp\Big[-\beta''\big[r'_1(\theta) + r'_2(\theta)\big]\Big]\Big\}\Big]\Big\} + \log(\epsilon^{-1}),$$

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

To get a non localized learning theorem, we can choose for some parameter μ

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

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

THEOREM 1.2. *For the choice of parameters explained above, with $P^{\otimes 2N}$ probability at least $1-\epsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\rho\big[r_2(\theta)\big] \leq r_2(\widetilde{\theta}) + \frac{\mu + \frac{\mu^2}{2N} + \kappa\frac{\beta^2}{N}}{\mu - \frac{\mu^2}{2N} - \kappa\frac{\lambda^2}{N}}\Big[\rho\big[r_1(\theta)\big] - r_1(\widetilde{\theta})\Big]$$
$$+ \frac{1}{\mu - \frac{\mu^2}{2N} - \kappa\frac{\lambda^2}{N}}\Big\{\mathcal{K}(\rho,\pi) + \log(\epsilon^{-1})\Big\}.$$

2. Localized bound

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

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

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 assume that $\beta' > \lambda'$ and let us put

$$\gamma' = \beta'' \frac{\beta' + \lambda'}{\beta' - \beta''} = (\beta' - \lambda') \frac{1 - \frac{1}{N} \frac{(\lambda' + \beta')^2}{4(\beta' - \lambda')}}{1 + \frac{1}{2N} \frac{\lambda' + \beta'}{2}}$$

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

The same computation that led to (4.1) shows that

LEMMA 2.1. *For the choice of parameters explained above, with $P^{\otimes 2N}$ probability at least $1 - \epsilon$,*

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

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

$$P^{\otimes 2N}\left\{\exp\Big[\lambda\big[\overline{r}_1(\theta) - r'_1(\theta)\big] - \mu r'_1(\theta)\Big] \,\Big|\, X_1^N\right\} \leq \exp\Big[\big(\kappa \tfrac{\lambda^2}{N} - \mu\big) r'_1(\theta)\Big].$$

LEMMA 2.2. *For the choice of parameters explained above, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any posterior probability distribution $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\lambda \rho\big[\overline{r}_1(\theta)\big] \leq \left(\lambda + \gamma' + \kappa \frac{\lambda^2}{N}\right) \rho\big[r'_1(\theta)\big] + \log\left\{\pi\Big[\exp\big[-\gamma' r'_1(\theta)\big]\Big]\right\} + \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}).$$

Exactly as we derived (6.3), we can establish that with $P^{\otimes N}$ probability at least $1 - \epsilon$,

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

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

$$\rho\big[r_2(\theta)\big] \leq r_2(\widetilde{\theta}) + \frac{\beta}{\lambda}\Big[\rho\big[r_1(\theta)\big] - r_1(\widetilde{\theta})\Big] + \frac{1}{\lambda}\left\{\mathcal{K}(\rho, \pi) + \log\left\{\pi\Big[\exp\big[-(\beta - \lambda)\zeta\, \overline{r}_1(\theta)\big]\Big]\right\} + \left(1 + \tfrac{\beta' - \lambda'}{\beta' + \lambda'} + \tfrac{\beta' - \lambda'}{\lambda}\right) \log\big(\tfrac{3}{\epsilon}\big)\right\}$$
$$= r_2(\widetilde{\theta}) + \Big(\zeta + (1 - \zeta)\tfrac{\beta}{\lambda}\Big)\big\{\rho\big[r_1(\theta)\big] - r_1(\widetilde{\theta})\big\}$$

$$+ \frac{1}{\lambda}\Big\{\mathcal{K}(\rho, \hat{\rho}_{(\beta-\lambda)\zeta}) + \big(1 + \tfrac{\beta'-\lambda'}{\beta'+\lambda'} + \tfrac{\beta'-\lambda'}{\lambda}\big)\log\big(\tfrac{3}{\epsilon}\big)\Big\}.$$

As a special case,

$$\hat{\rho}_\beta\big[r_2(\theta)\big] \leq r_2(\widetilde{\theta}) + \frac{1}{\lambda}\int_{(\beta-\lambda)\zeta}^{\beta} \hat{\rho}_\gamma\big[\overline{r}_1(\theta)\big]d\gamma + \frac{1}{\lambda}\Big(1 + \tfrac{\beta'-\lambda'}{\beta'+\lambda'} + \tfrac{\beta'-\lambda'}{\lambda}\Big)\log\big(\tfrac{3}{\epsilon}\big).$$

CHAPTER 4

Compression schemes in the i.i.d. case

It is worth devoting a full chapter to expend the clever study of compression schemes made in M. Seeger's paper [**28**].

1. Non localized bound

Let us consider some distribution $P \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})$, where $(\mathcal{X},\mathcal{B})$ is a measurable space and $\mathcal{Y}$ a finite set. Let $(\Theta,\mathcal{T})$ be some measurable set of parameters. Assume that for each sample $Z' \in (\mathcal{X}\times\mathcal{Y})^h$ of arbitrary size h we are given some measurable set of extra parameters $\Theta(Z') \in \mathcal{T}$ and a set of classification rules $\big\{f_{Z',\theta} : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta(Z')\big\}$. Assume as usual that $(Z',\theta,x) \mapsto f_{Z',\theta}(x)$ is a measurable function from the measurable subset of $(\mathcal{X}\times\mathcal{Y})^h \times \Theta \times \mathcal{X}$ on which it is defined to $\mathcal{Y}$. Let $Z = Z_1^N$ be as usual the canonical process on $(\mathcal{X}\times\mathcal{Y})^N$. For any subsequence $I \in \{1,\dots,N\}^h$ of length h, we can define $Z_I = \{Z_{I(i)} : 1 \le i \le h\}$ and consider the set of classification rules $\{f_{Z_I,\theta}; \theta \in \Theta(Z_I)\}$ (this is a random set, since it depends on Z_I). The introduction of the parameter set Θ is a slight extension of compression schemes which allows to consider several estimators for each choice of the compression set Z_I. For instance, in the case of support vector machines, we may consider, as we will see in the sequel of this paper, for each possible choice of support vectors, different possible choices of kernels.

We can then, following M. Seeger, use the fact that, under the product distribution $P^{\otimes N}$ on $(\mathcal{X}\times\mathcal{Y})^N$, for a fixed choice of I, $\{f_{Z_I,\theta}; \theta \in \Theta(Z_I)\}$ is independent from Z_{I^c}, where $I^c = \{1,\dots,N\} \setminus I(\{1,\dots,h\})$.

For any $I \in \{1,\dots,N\}^h$ and any $\theta \in \Theta(Z_I)$, let us consider the empirical error rate

$$r(I,\theta) = \frac{1}{N-h}\sum_{j\in I^c} \mathbb{1}\big[Y_j \neq f_{Z_I,\theta}(X_j)\big].$$

This empirical error rate is to be compared with the expected error rate $R(I,\theta) = P\big[Y_{N+1} \neq f_{Z_I,\theta}(X_{N+1})\big]$, where $(X_{N+1},Y_{N+1}) \in (\mathcal{X}\times\mathcal{Y})$ is some extra test point. Let us remark that $R(I,\theta)$ is a random variable since it depends on Z_I, that it is defined[1] for any $\theta \in \Theta(Z_I)$, and is independent from Z_{I^c}. We can therefore apply Bernstein's theorem to get for any positive λ and any function $\eta : \big(\mathcal{X}\times\mathcal{Y}\big)^h \times \Theta \to \mathbb{R}$, any integer h, any $I \in \{1,\dots,N\}^h$, any $Z_I \in (\mathcal{X}\times\mathcal{Y})^h$, any $\theta \in \theta(Z_I)$,

$$P^{\otimes(N-h)}\Big\{\exp\Big[\lambda R(I,\theta) - \lambda r(I,\theta) - g(\tfrac{\lambda}{N-h})\frac{\lambda^2}{N-h}R(I,\theta) - \eta(Z_I,\theta)\Big]\Big\}$$

[1]The fact that the range of θ is a random set $\Theta(Z_I)$ may seem disturbing: a more rigorous but more combersome way of introducing $R(I,\theta)$ would be to consider the random set $\{R(I,\theta); \theta \in \Theta(Z_I)\}$, which is indeed the true mathematical object we are dealing with in a loose but hopefully more intuitive way.

$$\leq P^{\otimes(N-h)}\Big\{\exp\big[-\eta(Z_I,\theta)\big]\Big\},$$

where $P^{\otimes(N-h)}$ is the product measure on $(\mathcal{X}\times\mathcal{Y})^{I^c}$ and where as usual $g(s) = s^{-2}\big[\exp(s)-1-s\big]$. Let $\mathcal{I} = \big\{I \in \{1,\dots,N\}^h; 1 \leq h < N\big\}$ and let $\pi : (\mathcal{X}\times\mathcal{Y})^N \to \mathcal{M}_+^1(\mathcal{I}\times\Theta)$ be some regular conditional probability measure. Assume that for any $I \in \mathcal{I}$, $\pi(I \,|\, Z)$ is independent of $Z \in (\mathcal{X}\times\mathcal{Y})^N$. Assume also that $\pi(d\theta \,|\, Z, I)$ depends only on Z_I and satisfies $\pi\big(\Theta(Z_I) \,|\, Z, I\big) = 1$.

Integrating the previous inequality with respect to π, using Fubini's theorem and the Legendre transform (4.3) of chapter 1, we get

LEMMA 1.1. *For any positive parameter λ*

$$P^{\otimes N}\Bigg\{\sup_{\rho\in\mathcal{M}_+^1(\mathcal{I}\times\Theta)} \lambda\rho\Big\{\Big[1-g(\tfrac{\lambda}{N-h})\tfrac{\lambda}{N-h}\Big]R(I,\theta)\Big\} - \lambda\rho\Big[r(I,\theta)\Big] - \rho\big[\eta(Z_I,\theta)\big] - \mathcal{K}(\rho,\pi) \geq 0\Bigg\} \leq P^{\otimes N}\Bigg\{\pi\Big\{\exp\big[-\eta(Z_I,\theta)\big]\Big\}\Bigg\}.$$

In the same way

$$P^{\otimes N}\Bigg\{\sup_{\rho\in\mathcal{M}_+^1(\mathcal{I}\times\Theta)} -\lambda\rho\Big\{\Big[1+g(\tfrac{\lambda}{N-h})\tfrac{\lambda}{N-h}\Big]R(I,\theta)\Big\} + \lambda\rho\Big[r(I,\theta)\Big] - \rho\big[\eta(Z_I,\theta)\big] - \mathcal{K}(\rho,\pi) \geq 0\Bigg\} \leq P^{\otimes N}\Bigg\{\pi\Big\{\exp\big[-\eta(Z_I,\theta)\big]\Big\}\Bigg\}.$$

Let us notice that we use h here for the length of I, therefore h is not a constant but a function of the parameter I. Let us notice also that the two inequalities of the above lemma require formely $R(I,\theta)$ and $r(I,\theta)$ to be defined on the whole of $(\mathcal{I}\times\Theta)$, but that their left-hand and right-hand sides are independent from the way these error rates are defined when $\theta \notin \Theta(Z_I)$, therefore to be completely rigorous, we can put $r(I,\theta) = R(I,\theta) = 0$ for any $\theta \notin \Theta(Z_I)$.

Taking $\eta(Z_I,\theta) = \log(\epsilon^{-1})$ gives a non localized result. Let us put $\mathcal{I}_H = \big\{I \in \{1,\dots,N\}^h; 1 \leq h \leq H\big\}$.

COROLLARY 1.2. *For any integer $H < N$, any λ such that $g(\frac{\lambda}{N-H})\frac{\lambda}{N-H} \leq 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any $H' \leq H$, for any $\rho \in \mathcal{M}_+^1(\mathcal{I}\times\Theta)$ such that $\rho(\mathcal{I}_{H'}\times\Theta) = 1$,*

$$\rho\big[R(I,\theta)\big] \leq \Big[1-g(\tfrac{\lambda}{N-H'})\tfrac{\lambda}{N-H'}\Big]^{-1}\Big\{\rho\big[r(I,\theta)\big] + \frac{\mathcal{K}(\rho,\pi)+\log(\epsilon^{-1})}{\lambda}\Big\}.$$

When applied to the case when $\Theta(Z_I)$ is always a one point set, when the prior π is the uniform probability measure on $\mathcal{I}_H \backslash \mathcal{I}_{H-1}$, when $f_{Z_I,\theta} = f_{Z_I}$ is exchangeable (i.e. independent of the permutations of the subsample Z_I) and when the posterior ρ is the uniform measure on the set of all the permutations of some sequence I, this result, once properly optimized in λ through a union bound, is analogous to Seeger's one [**28**], with the difference that we use the slightly weaker Bernstein inequality, compared with the Chernoff one. The details are very analogous to the situation

described in previous chapters, so we will not bother the reader with them any further. Anyhow, the interest of working with Bernstein's inequality, apart from a more intuitive interpretation of the results in terms of bias and variance (but this is rather a matter of taste than anything else), is that it leads to localized bounds, which are the subject of the next section.

2. Localized bounds

Let us keep the same notations and setting as in the previous section. Consider some integer $H < N$ and assume that $\pi(\mathcal{I}_H \times \Theta) = 1$. For some positive parameter β, choose

$$\eta(Z_I, \theta) = \beta R(I,\theta) + \log\Big\{\pi\Big[\exp\big[-\beta R(I,\theta)\big]\Big]\Big\} + \log(\epsilon^{-1}).$$

Let us put for short $M = N - H$ and $\kappa = g(\frac{\lambda}{M})$. Then with $P^{\otimes N}$ probability at least $1-\epsilon$,

$$(\lambda - \beta - \kappa\tfrac{\lambda^2}{M})\rho\big[R(I,\theta)\big] \le \lambda\rho\big[r(I,\theta)\big] + \log\Big\{\pi\Big[\exp\big[-\beta R(I,\theta)\big]\Big]\Big\} + \mathcal{K}(\rho,\pi) + \log(\epsilon^{-1}).$$

In the same way, with $P^{\otimes N}$ probability at least $1-\epsilon$,

$$(-\lambda - \beta - \kappa\tfrac{\lambda^2}{M})\rho\big[R(I,\theta)\big] \le -\lambda\rho\big[r(I,\theta)\big] + \log\Big\{\pi\Big[\exp\big[-\beta R(I,\theta)\big]\Big]\Big\} + \mathcal{K}(\rho,\pi) + \log(\epsilon^{-1}).$$

From this last reverse inequality, with $P^{\otimes N}$ probability at least $1-\epsilon$,

$$\begin{aligned}\log\Big\{\pi\Big[\exp\big[-\beta R(I,\theta)\big]\Big]\Big\} &= \sup_{\rho\in\mathcal{M}_+^1(\mathcal{I}_H\times\Theta)} -\beta\rho\big[R(I,\theta)\big] - \mathcal{K}(\rho,\pi)\\ &\le \frac{\beta}{\lambda+\beta+\kappa\frac{\lambda^2}{M}}\Big\{-\lambda\rho\big[r(I,\theta)\big] + \log\Big\{\pi\Big[\exp\big[-\beta R(I,\theta)\big]\Big]\Big\} \\ &\qquad + \mathcal{K}(\rho,\pi) + \log(\epsilon^{-1})\Big\} - \mathcal{K}(\rho,\pi).\end{aligned}$$

Putting $\xi = \frac{\beta}{\lambda+\kappa\frac{\lambda^2}{M}}$, this leads with $P^{\otimes N}$ probability at least $1-\epsilon$ to

$$\log\Big\{\pi\Big[\exp\big[-\beta R(I,\theta)\big]\Big]\Big\} \le \log\Big\{\pi\Big[\exp\big[-\xi\lambda r(I,\theta)\big]\Big]\Big\} + \xi\log(\epsilon^{-1}).$$

THEOREM 2.1. *For any $\xi \in [0,1[$, any positive λ such that $\frac{1+\xi}{1-\xi}g(\frac{\lambda}{N-H})\frac{\lambda}{N-H} < 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$, for any posterior distribution $\rho \in \mathcal{M}_+^1(\mathcal{I}_H\times\Theta)$,*

$$\begin{aligned}\rho\big[R(I,\theta)\big] &\le \Big[(1-\xi) - (1+\xi)g(\tfrac{\lambda}{N-H})\tfrac{\lambda}{N-H}\Big]^{-1}\Big\{\rho\big[r(I,\theta)\big] \\ &\qquad + \frac{1}{\lambda}\Big[\log\Big\{\pi\Big[\exp\big[-\lambda\xi r(I,\theta)\big]\Big]\Big\} + \mathcal{K}(\rho,\pi) + (1+\xi)\log(\tfrac{2}{\epsilon})\Big]\Big\}\\ &= \Big[1-\tfrac{1+\xi}{1-\xi}g(\tfrac{\lambda}{N-H})\tfrac{\lambda}{N-H}\Big]^{-1}\Big\{\rho\big[r(I,\theta)\big] + \frac{1}{(1-\xi)\lambda}\Big[\mathcal{K}(\rho,\pi_{\exp(-\xi\lambda r)}) + (1+\xi)\log(\tfrac{2}{\epsilon})\Big]\Big\}.\end{aligned}$$

As a consequence, putting

$$d\hat{\rho}_\beta = d\pi_{\exp(-\beta r)} \stackrel{\text{def}}{=} \frac{\exp\big[-\beta r(I,\theta)\big]}{\pi\Big\{\exp\big[-\beta r(I,\theta)\big]\Big\}} d\pi(I,\theta),$$

and minimizing in ρ the right-hand side of the previous inequality, we get

COROLLARY 2.2. *For any $\xi \in [0,1[$, any positive λ such that $\frac{1+\xi}{1-\xi} g(\frac{\lambda}{N-H})\frac{\lambda}{N-H} < 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$,*

$$\begin{aligned}(2.1)\quad \hat{\rho}_\lambda\big[R(I,\theta)\big] &\le \Big[1 - \tfrac{1+\xi}{1-\xi} g(\tfrac{\lambda}{N-H})\tfrac{\lambda}{N-H}\Big]^{-1} \\ &\quad \times \left\{ \frac{1}{(1-\xi)\lambda} \int_{\xi\lambda}^{\lambda} \hat{\rho}_\beta\big[r(I,\theta)\big] d\beta + \tfrac{1+\xi}{(1-\xi)\lambda} \log(\tfrac{2}{\epsilon}) \right\} \\ &\le \Big[1 - \tfrac{1+\xi}{1-\xi} g(\tfrac{\lambda}{N-H})\tfrac{\lambda}{N-H}\Big]^{-1} \Big\{ \hat{\rho}_{\xi\lambda}\big[r(I,\theta)\big] + \tfrac{1+\xi}{(1-\xi)\lambda} \log(\tfrac{2}{\epsilon}) \Big\}.\end{aligned}$$

The following weaker but simpler inequality obtained by cancelling the localized divergence term also holds as soon as $\frac{1+\xi}{\xi(1-\xi)} g(\frac{\lambda}{\xi(N-H)})\frac{\lambda}{(N-H)} < 1$ with $P^{\otimes N}$ probability at least $1-\epsilon$:

$$(2.2)\qquad \hat{\rho}_\lambda\big[R(I,\theta)\big] \le \frac{\hat{\rho}_\lambda\big[r(I,\theta)\big] + \dfrac{\xi(1+\xi)}{(1-\xi)\lambda}\log\Big(\dfrac{2}{\epsilon}\Big)}{1 - \dfrac{(1+\xi)}{\xi(1-\xi)} g\Big(\dfrac{\lambda}{\xi(N-H)}\Big)\dfrac{\lambda}{(N-H)}}.$$

Let us note that it is possible to prove a theoretical bound analogous to corollary 6.5 of chapter 1, the proof being the same:

PROPOSITION 2.3. *For any positive λ such that $g(\frac{\lambda}{N-H})\frac{\lambda}{(1-\xi)(N-H)} < 1$, with $P^{\otimes N}$ probability at least $1-\epsilon$,*

$$\begin{aligned}\hat{\rho}_\lambda\big[R(I,\theta)\big] &\le \inf_{\xi\in[0,1[} \frac{\dfrac{1}{(1-\xi)\lambda}\displaystyle\int_{\lambda\xi}^{\lambda+g(\frac{\lambda}{N-H})\frac{\lambda^2}{N-H}} \pi_{\exp(-\beta R)}\big[R(I,\theta)\big] d\beta + \frac{2}{(1-\xi)\lambda}\log(\frac{2}{\epsilon})}{1-(1-\xi)^{-1} g(\frac{\lambda}{N-H})\frac{\lambda}{N-H}} \\ &\le \inf_{\xi\in[0,1[} \frac{\big[1+g(\frac{\lambda}{N-H})\frac{\lambda}{(1-\xi)(N-H)}\big]\pi_{\exp(-\xi\lambda R)}\big[R(I,\theta)\big] + \frac{2}{(1-\xi)\lambda}\log(\frac{2}{\epsilon})}{1-(1-\xi)^{-1} g(\frac{\lambda}{N-H})\frac{\lambda}{N-H}}.\end{aligned}$$

Gaining uniformity in λ is done exactly as in theorem 6.6 of chapter 1. We let the details to the reader.

Since $\inf_{I\in\mathcal{I}_H,\theta\in\Theta(Z_I)} r(I,\theta)$ will as a rule not be much larger than $\inf_{I,\theta} R(I,\theta)$, (although it can be much smaller if H or $\Theta(Z_I)$ are large), $\hat{\rho}_{\xi\lambda}\big[r(I,\theta)\big]$ can be expected not to be much larger than $\hat{\rho}_{\xi\lambda}\big[R(I,\theta)\big]$, showing that the bounds given in corollary 2.2 may be expected to be in many situations quite tight (that is of the right order of magnitude). This would of course deserve to be checked by numerical experiments on benchmark classification problems and we will try to do this in the future. The main interest of this study of compression schemes in the i.i.d. case, when compared to what we did in previous sections using exchangeable priors, is that in this setting the Gibbs distribution can be computed without being given any test set, which is a usual practical situation.

Since it does not lead to confidence bounds, and therefore is of a least practical interest, we will let the reader imagine for himself how noisy classification bounds analogous to those derived in section 7 of chapter 1 could be proved for compression schemes in the i.i.d. case.

3. A toy example

Let us work out a toy example explicitly to show that the localized bound (2.2) above is sharper than any non localized bounds, such as those derived in this article, but also by Vapnik [**21**] or Seeger [**28**]. A similar computation could be done, although more painstakingly, for the other localized bounds presented in this paper.

Consider indeed the case when $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$ and assume that for some real parameter $\alpha \in]0,1[$, $Y = \mathbb{1}(X \geq \alpha)$, P almost surely. This is the most simple noiseless classification problem that could be imagined !

Using the same notations as in the previous sections of this chapter, let us define $\hat{f}_{\{i\}}(x) = \mathbb{1}(x \geq X_i)$, and consider on $\mathcal{I}_1$, the set of sequences of indices of length one, which we will identify with $\{1, \dots, N\}$ to shorten notations, the uniform prior probability measure $\pi(\{i\}) = \frac{1}{N}$ (in this simple example, $\Theta = \varnothing$: there is no additional parameter set). Let σ be the (random) permutation which sorts the sample $(X_i)_{i=1}^N$ in increasing order, so that $X_{\sigma(1)} \leq X_{\sigma(2)} \leq \cdots \leq X_{\sigma(N)}$. It is easy to see that

$$r\big[\sigma(i)\big] = \begin{cases} \frac{i - i_0}{N-1} & i \geq i_0, \\ \frac{i_0 - 1 - i}{N-1} & i < i_0, \end{cases}$$

where i_0 is defined by the property $X_{\sigma(i_0-1)} < \alpha \leq X_{\sigma(i_0)}$. We can also compute

$$\begin{aligned} \hat{\rho}_\lambda\big[r(i)\big] &= -\frac{\partial}{\partial \lambda} \log\left\{ \sum_{i=1}^N \exp\big[-\lambda r(i)\big] \right\} \\ &= -\frac{\partial}{\partial \lambda} \log \left\{ \frac{2 - \exp\Big[-\frac{\lambda(N-i_0)}{N-1}\Big] - \exp\Big[-\frac{\lambda(i_0-1)}{N-1}\Big]}{1 - \exp\Big(-\frac{\lambda}{N-1}\Big)} \right\} \\ &\leq \frac{1}{N-1} \left[\exp\Big(\frac{\lambda}{N-1}\Big) - 1 \right]^{-1}. \end{aligned}$$

We can now apply equation (2.2) with $\xi = \frac{2}{5}$ and $\lambda = \frac{N-1}{6}$ to get

$$\hat{\rho}_{\frac{N-1}{6}}\big[R(i)\big] \leq \frac{\big[\exp(\frac{1}{6}) - 1\big]^{-1} + \frac{28}{5}\log(2/\epsilon)}{\big[1 - \frac{35}{36} g(\frac{5}{12})\big](N-1)}.$$

Numerically, this gives

$$\hat{\rho}_{\frac{N-1}{6}}\big[R(i)\big] \leq \frac{12.6}{N-1} + \frac{12.8}{N-1}\log(\tfrac{2}{\epsilon}) \leq \frac{21.5}{N-1} + \frac{12.8\log(\epsilon^{-1})}{N-1}.$$

Therefore with $P^{\otimes N}$ probability at least $\frac{9}{10}$,

$$\hat{\rho}_{\frac{N-1}{6}}\big[R(i)\big] \leq \frac{50.1}{N-1}.$$

Moreover, integrating with respect to ϵ proves that

$$P^{\otimes N}\Big\{ \hat{\rho}_{\frac{N-1}{6}}\big[R(i)\big] \Big\} \leq \frac{34.2}{N-1}.$$

The conclusion is that we get on this simple example a bound of order $(N-1)^{-1}$, whereas non localized bounds (ours, Vapnik's or Seeger's) would be of order $\frac{\log(N)}{N}$. This shows that localization brings a qualitative improvement on previous results (at least when N is large enough). Even numerical constants are not shameful in this example, although clearly still suboptimal [2] (let us notice that for simplicity we did not start from the sharpest inequality of corollary 2.2). Let us notice also that we did not need to make any assumption on the distribution of the patterns X_i : adaptation to the design distribution is a typical feature shared by the exchangeable prior approach and the compression scheme approach. Indeed, if we had tried the PAC-Bayesian approach with a fixed prior on this example, choosing the family of classification rules $\{f_\theta(x) = \mathbb{1}(x \geq \theta) : x \in [0,1], \theta \in [0,1]\}$, the best results would have been obtained with a prior close to the distribution of the patterns X_i, whereas the compression scheme prior turns out conveniently to be related to the empirical distribution of the pattern sample $(X_i)_{i=1}^N$, which serves as an estimator of its true distribution.

[2] We can expect the constants to improve in the presence of noise and for a high dimensional problem. Indeed, the term involving ϵ, which has a strong impact on the bound in this toy example, is dimension free, and therefore should not be felt in a high dimensional problem. Moreover, the optimal value of λ decreases when the level of noise increases. For a fixed positive level of noise, the optimal value of λ is of order $N^{1/2}$, and we may conjecture that equation (2.2) gives a suboptimal constant by a factor of order $\sqrt{\frac{6}{1-\inf_{I,\theta} R(I,\theta)}} = \frac{2.45}{\sqrt{1-\inf_{I,\theta} R(I,\theta)}}$, for high dimensional problems, instead of being off from a factor 6, meaning that we get a bound lower than $\inf_{I,\theta} R(I,\theta) + \frac{c}{\sqrt{N}}$ with c no more than $\frac{2.45}{\sqrt{1-\inf_{I,\theta} R(I,\theta)}}$ times too big. The situation should be even better when using the sharper inequality (2.1). Anyhow, this footnote is only a conjecture, which should be taken with care and checked on simulations.

CHAPTER 5

Support Vector Machines

A prerequisite to the definition of support vector machines is to study the separation of points by hyperplanes in a finite dimensional Euclidean space. Support vector machines, introduced by V. Vapkik [**21**], are a fundamental classification algorithm and a natural framework to apply the preceding PAC-Bayesian results.

1. The canonical hyperplane

We will deal in this section with the classification of points of $\mathbb{R}^d$ in two classes. Let $Z = (x_i, y_i)_{i=1}^N \in \big(\mathbb{R}^d \times \{-1, +1\}\big)^N$ be some set of labelled examples (called the training set hereafter). Let

$$\begin{aligned} I &= \{1, \dots, N\}, \\ I_+ &= \{i \in I \, : y_i = +1\}, \\ I_- &= \{i \in I \, : y_i = -1\}, \end{aligned}$$

and consider

$$A_Z = \big\{ w \in \mathbb{R}^d \, : \sup_{b \in \mathbb{R}} \inf_{i \in I} (\langle w, x_i \rangle - b) y_i \geq 1 \big\}.$$

Let us remark that this set of admissible separating directions can also be written as

$$A_Z = \big\{ w \in \mathbb{R}^d \, : \max_{i \in I_-} \langle w, x_i \rangle + 2 \leq \min_{i \in I_+} \langle w, x_i \rangle \big\}.$$

As it is easily seen, the optimal value of b for a fixed value of w, in other words the value of b which maximizes $\inf_{i \in I} (\langle w, x_i \rangle - b) y_i$, is equal to

$$b_w = \frac{1}{2} \Big[\max_{i \in I_-} \langle w, x_i \rangle + \min_{i \in I_+} \langle w, x_i \rangle \Big].$$

LEMMA 1.1. *When $A_Z \neq \varnothing$, $\inf\{\|w\|^2 \, : w \in A_Z\}$ is reached for only one value w_Z of w.*

PROOF. The set A_Z is convex and $w \mapsto \|w\|^2$ is strictly convex. □

DEFINITION 1.1. When $A_Z \neq \varnothing$, the training set Z is said to be linearly separable. The hyperplane

$$H = \{x \in \mathbb{R}^d \, : \langle w_Z, x \rangle - b_Z = 0\},$$

where

$$\begin{aligned} w_Z &= \arg\min\{\|w\| \, : w \in A_Z\}, \\ b_Z &= b_{w_Z}, \end{aligned}$$

is called the canonical separating hyperplane of the training set Z. The quantity $\|w_Z\|^{-1}$ is called the margin of the canonical hyperplane.

Note that as $\min_{i\in I_+}\langle w_Z, x_i\rangle - \max_{i\in I_-}\langle w_Z, x_i\rangle = 2$, the margin is also equal to half the distance between the projections on the direction w_Z of the positive and negative patterns.

2. Computation of the canonical hyperplane

Let us consider the convex hulls X_+ and X_- of the positive and negative patterns:

$$\mathcal{X}_+ = \Big\{ \sum_{i\in I_+} \lambda_i x_i \ : (\lambda_i)_{i\in I_+} \in \mathbb{R}_+^{I_+}, \sum_{i\in I_+} \lambda_i = 1 \Big\},$$
$$\mathcal{X}_- = \Big\{ \sum_{i\in I_-} \lambda_i x_i \ : (\lambda_i)_{i\in I_-} \in \mathbb{R}_+^{I_-}, \sum_{i\in I_-} \lambda_i = 1 \Big\}.$$

Let us introduce the convex set

$$\mathcal{V} = \mathcal{X}_+ - \mathcal{X}_- = \big\{ x_+ - x_- \ : x_+ \in \mathcal{X}_+, x_- \in \mathcal{X}_- \big\}.$$

As $v \mapsto \|v\|^2$ is strictly convex, there is a unique vector v^* such that

$$\|v^*\|^2 = \inf_{v\in\mathcal{V}} \big\{ \|v\|^2 \ : v \in \mathcal{V} \big\}.$$

LEMMA 2.1. *The set A_Z is non empty (i.e. the training set Z is linearly separable) if and only if $v^* \neq 0$. In this case*

$$w_Z = \frac{2}{\|v^*\|^2} v^*,$$

and the margin of the canonical hyperplane is equal to $\frac{1}{2}\|v^*\|$.

PROOF. Let us assume first that $v^* = 0$, or equivalently that $\mathcal{X}_+ \cap \mathcal{X}_- \neq \varnothing$. As for any vector $w \in \mathbb{R}^d$,

$$\min_{i\in I_+} \langle w, x_i\rangle = \min_{x\in\mathcal{X}_+} \langle w, x\rangle,$$
$$\max_{i\in I_-} \langle w, x_i\rangle = \max_{x\in\mathcal{X}_-} \langle w, x\rangle,$$

we see that necessarily $\min_{i\in I_+}\langle w, x_i\rangle - \max_{i\in I_-}\langle w, x_i\rangle \leq 0$, which shows that w cannot be in A_Z and therefore that A_Z is empty.

Let us assume now that $v^* \neq 0$, or equivalently that $\mathcal{X}_+ \cap \mathcal{X}_- = \varnothing$. Let us put $w^* = \frac{2}{\|v^*\|^2} v^*$. Let us remark first that

$$\begin{aligned} \min_{i\in I_+} \langle w^*, x_i\rangle - \max_{i\in I_-} \langle w^*, x_i\rangle &= \inf_{x\in\mathcal{X}_+} \langle w^*, x\rangle - \sup_{x\in\mathcal{X}_-} \langle w^*, x\rangle \\ &= \inf_{x_+\in\mathcal{X}_+, x_-\in\mathcal{X}_-} \langle w^*, x_+ - x_-\rangle \\ &= \frac{2}{\|v^*\|^2} \inf_{v\in\mathcal{V}} \langle v, v^*\rangle. \end{aligned}$$

Let us now prove that $\inf_{v\in\mathcal{V}}\langle v, v^*\rangle = \|v^*\|^2$. Some arbitrary $v \in \mathcal{V}$ being fixed, consider the function

$$\beta \mapsto \|\beta v + (1-\beta)v^*\|^2 : [0,1] \to \mathbb{R}.$$

By definition of v^*, it reaches its minimum value for $\beta = 0$, and therefore has a non negative derivative at this point. Computing this derivative, we find that $\langle v - v^*, v^* \rangle \geq 0$, as claimed. We have proved that

$$\min_{i \in I_+} \langle w^*, x_i \rangle - \max_{i \in I_-} \langle w^*, x_i \rangle = 2,$$

and therefore that $w^* \in A_Z$. On the other hand, any $w \in A_Z$ is such that

$$2 \leq \min_{i \in I_+} \langle w, x_i \rangle - \max_{i \in I_-} \langle w, x_i \rangle = \inf_{v \in \mathcal{V}} \langle w, v \rangle \leq \|w\| \inf_{v \in \mathcal{V}} \|v\| = \|w\| \, \|v^*\|.$$

This proves that $\|w^*\| = \inf\big\{\|w\| \, : \, w \in A_Z\big\}$, and therefore that $w^* = w_Z$ as claimed. □

One way to compute w_Z would be therefore to compute v^* by minimizing

$$\big\{\|\sum_{i \in I} \lambda_i y_i x_i\|^2 \, : (\lambda_i)_{i \in I} \in \mathbb{R}_+^I, \sum_{i \in I} \lambda_i = 2, \sum_{i \in I} y_i \lambda_i = 0\big\}.$$

Although this is a tractable quadratic programming problem, a direct computation of w_Z through the following proposition is usually prefered.

PROPOSITION 2.2. *The canonical direction w_Z can be expressed as*

$$w_Z = \sum_{i=1}^N \alpha_i^* y_i x_i,$$

where $(\alpha_i^*)_{i=1}^N$ is obtained by minimizing

$$\inf\big\{F(\alpha) \, : \alpha \in \mathcal{A}\big\},$$

where

$$\mathcal{A} = \Big\{(\alpha_i)_{i \in I} \in \mathbb{R}_+^I, \sum_{i \in I} \alpha_i y_i = 0\Big\},$$

and

$$F(\alpha) = \Big\|\sum_{i \in I} \alpha_i y_i x_i\Big\|^2 - 2 \sum_{i \in I} \alpha_i.$$

PROOF. Let $w(\alpha) = \sum_{i \in I} \alpha_i y_i x_i$ and let $S(\alpha) = \frac{1}{2} \sum_{i \in I} \alpha_i$. We can express the function $F(\alpha)$ as $F(\alpha) = \|w(\alpha)\|^2 - 4S(\alpha)$. Moreover it is important to notice that for any $s \in \mathbb{R}_+$ $\{w(\alpha) \, : \, \alpha \in \mathcal{A}, S(\alpha) = s\} = s\mathcal{V}$. This shows that for any $s \in \mathbb{R}_+$, $\inf\{F(\alpha) : \alpha \in \mathcal{A}, S(\alpha) = s\}$ is reached and that for any $\alpha_s \in \{\alpha \in \mathcal{A} \, : \, S(\alpha) = s\}$ reaching this infimum, $w(\alpha_s) = sv^*$. As $s \mapsto s^2 \|v^*\|^2 - 4s : \mathbb{R}_+ \to \mathbb{R}$ reaches its infimum for only one value s^* of s, namely at $s^* = \frac{2}{\|v^*\|^2}$, this shows that $F(\alpha)$ reaches its infimum on $\mathcal{A}$, and that for any $\alpha^* \in \mathcal{A}$ such that $F(\alpha^*) = \inf\{F(\alpha) : \alpha \in \mathcal{A}\}$, $w(\alpha^*) = \frac{2}{\|v^*\|^2} v^* = w_Z$. □

3. Support vectors

DEFINITION 3.1. The set of support vectors $\mathcal{S}$ is defined by

$$\mathcal{S} = \{x_i \, : \langle w_Z, x_i \rangle - b_Z = y_i\}.$$

PROPOSITION 3.1. *Any α^* minimizing $F(\alpha)$ on $\mathcal{A}$ is such that*

$$\{x_i \, : \alpha_i^* > 0\} \subset \mathcal{S}.$$

This implies that the representation $w_Z = w(\alpha^*)$ involves in general only a limited number of non zero coefficients and that $w_Z = w_{Z'}$, where $Z' = \{(x_i, y_i) \, : x_i \in \mathcal{S}\}$.

PROOF. Let us consider any given $i \in I_+$ and $j \in I_-$, such that $\alpha_i^* > 0$ and $\alpha_j^* > 0$ (there exists at least one such index in each set I_- and I_+, since the sum of the components of α^* on each of these sets are equal and since $\sum_{k\in I} \alpha_k^* > 0$). For any $t \in \mathbb{R}$, consider

$$\alpha_k(t) = \alpha_k^* + t\mathbb{1}(k \in \{i,j\}), \quad k \in I.$$

The vector $\alpha(t)$ is in $\mathcal{A}$ for any value of t in some neighborhood of 0, therefore $\frac{\partial}{\partial t}_{|t=0} F\big[\alpha(t)\big] = 0$. Computing this derivative, we find that

$$y_i\langle w(\alpha^*), x_i\rangle + y_j\langle w(\alpha^*), x_j\rangle = 2.$$

As $y_i = -y_j$, this can also be written as

$$y_i\big[\langle w(\alpha^*), x_i\rangle - b_Z\big] + y_j\big[\langle w(\alpha^*), x_j\rangle - b_Z\big] = 2.$$

As $w(\alpha^*) \in A_Z$,

$$y_k\big[\langle w(\alpha^*), x_k\rangle - b_Z\big] \geq 1, \qquad k \in I,$$

which implies necessarily as claimed that

$$y_i\big[\langle w(\alpha^*), x_i\rangle - b_Z\big] = y_j\big[\langle w(\alpha^*), x_j\rangle - b_Z\big] = 1.$$

□

4. Support Vector Machines

DEFINITION 4.1. The symmetric measurable kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is said to be positive (or more precisely positive semi-definite) if for any $n \in \mathbb{N}$, any $(x_i)_{i=1}^n \in \mathcal{X}^n$,

$$\inf_{\alpha\in\mathbb{R}^n} \sum_{i=1}^n \sum_{j=1}^n \alpha_i K(x_i, x_j)\alpha_j \geq 0.$$

Let $Z = (x_i, y_i)_{i=1}^N$ be some training set. Let us consider as in the previous sections of this chapter

$$\mathcal{A} = \big\{\alpha \in \mathbb{R}_+^N : \sum_{i=1}^N \alpha_i y_i = 0\big\}.$$

Let

$$F(\alpha) = \sum_{i=1}^N \sum_{j=1}^N \alpha_i y_i K(x_i, x_j) y_j \alpha_j - 2\sum_{i=1}^N \alpha_i.$$

DEFINITION 4.2. Let K be a positive symmmetric kernel. The training set Z is said to be K-separable if

$$\inf\big\{F(\alpha) : \alpha \in \mathcal{A}\big\} > -\infty.$$

LEMMA 4.1. *When Z is K-separable,* $\inf\{F(\alpha) : \alpha \in \mathcal{A}\}$ *is reached.*

PROOF. Consider the training set $Z' = (x_i', y_i)_{i=1}^N$, where

$$x_i' = \left\{ \left[\Big\{ K(x_k, x_\ell) \Big\}_{k=1,\ell=1}^{N \quad N} \right]^{1/2} (i,j) \right\}_{j=1}^N \in \mathbb{R}^N.$$

We see that $F(\alpha) = \left\| \sum_{i=1}^N \alpha_i y_i x_i' \right\|^2 - 2\sum_{i=1}^N \alpha_i$. We have proved in the previous section that Z' is linearly separable if and only if $\inf\{F(\alpha) : \alpha \in \mathcal{A}\} > -\infty$, and that the infimum is reached in this case. □

PROPOSITION 4.2. *Let K be a symmetric positive kernel and let $(Z_i)_{i=1}^N$ be some K-separable training set. Let $\alpha^* \in \mathcal{A}$ be such that $F(\alpha^*) = \inf\{F(\alpha) : \alpha \in \mathcal{A}\}$. Let*

$$I_-^* = \{i \in \mathbb{N} : 1 \leq i \leq N, y_i = -1, \alpha_i^* > 0\}$$
$$I_+^* = \{i \in \mathbb{N} : 1 \leq i \leq N, y_i = +1, \alpha_i^* > 0\}$$
$$b^* = \frac{1}{2}\Big\{\sum_{j=1}^N \alpha_j^* y_j K(x_j, x_{i_-}) + \sum_{j=1}^N \alpha_j^* y_j K(x_j, x_{i_+})\Big\}, \qquad i_- \in I_-^*, i_+ \in I_+^*,$$

where the value of b^* does not depend on the choice of i_- and i_+. The classification rule $f : \mathcal{X} \to \mathcal{Y}$ defined by the formula

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^N \alpha_i^* y_i K(x_i, x) - b^*\right)$$

is independent of the choice of α^* and is called the support vector machine defined by K and Z. The set $\mathcal{S} = \{x_j : \sum_{i=1}^N \alpha_i^* y_i K(x_i, x_j) - b^* = y_j\}$ is called the set of support vectors. For any choice of α^*, $\{x_i : \alpha_i^* > 0\} \subset \mathcal{S}$.

PROOF. The independence from the choice of α^*, which is not necessarily unique, is seen as follows. Let $(x_i)_{i=1}^N$ and $x \in \mathcal{X}$ be fixed. Let us put for ease of notations $x_{N+1} = x$. Let M be the $(N+1) \times (N+1)$ symmetric semi-definite matrix defined by $M(i,j) = K(x_i, x_j)$, $i = 1, \ldots, N+1$, $j = 1, \ldots, N+1$. Let us consider the mapping $\Psi : \{x_i : i = 1, \ldots, N+1\} \to \mathbb{R}^{N+1}$ defined by

$$\Psi(x_i) = \big[M^{1/2}(i,j)\big]_{j=1}^{N+1} \in \mathbb{R}^{N+1}. \tag{4.1}$$

Let us consider the training set $Z' = \big[\Psi(x_i), y_i\big]_{i=1}^N$. Then Z' is linearly separable,

$$F(\alpha) = \Big\|\sum_{i=1}^N \alpha_i y_i \Psi(x_i)\Big\|^2 - 2\sum_{i=1}^N \alpha_i,$$

and we have proved that for any choice of $\alpha^* \in \mathcal{A}$ minimizing $F(\alpha)$, $w_{Z'} = \sum_{i=1}^N \alpha_i^* y_i \Psi(x_i)$. Thus the support vector machine defined by K and Z can also be expressed by the formula

$$f(x) = \operatorname{sign}\Big[\langle w_{Z'}, \Psi(x)\rangle - b_{Z'}\Big]$$

which does not depend on α^*. The definition of $\mathcal{S}$ is such that $\Psi(\mathcal{S})$ is the set of support vectors defined in the linear case, where its stated property has already been prooved. □

5. Support vector machines seen as compression schemes

We can use support vector machines in the framework of compression schemes and apply proposition 3.1 of chapter 2. More precisely, given some positive symmetric kernel K on $\mathcal{X}$, we may consider for any training set $Z' = (x_i', y_i')_{i=1}^h$ the classifier $\hat{f}_{Z'} : \mathcal{X} \to \mathcal{Y}$ which is equal to the support vector machine defined by K and Z' whenever Z' is K-separable, and which is equal to some constant classification rule otherwise (we take this convention to stick to the framework of section 3.1, we will only use $\hat{f}_{Z'}$ in the K-separable case, so this extension of the definition is just a matter of presentation). In the application of proposition 3.1, in the case when the observed sample $(X_i, Y_i)_{i=1}^N$ is K-separable, a natural (if not always optimal) choice

of Z' is to choose for (x'_i) the set of support vectors defined by $Z = (X_i, Y_i)_{i=1}^N$ and to choose for (y'_i) the corresponding values of Y. This is justified by the fact that $\hat{f}_Z = \hat{f}_{Z'}$, as shown in proposition 4.2. In the case when Z is not K-separable, in theory, we can flip the smallest number of values of Y_i to define a K-separable training set $Z' = (X_i, y'_i)_{i=1}^N$, and then restrict this training set to its support vectors to see in which submodel this classification rule really lives. In practice, this may be time consuming, and various minimization criterion directly adapted to the non K-separable case are of common use. We suggest [**16**] as a further reading on this topic. Another possible suggestion is to use the kind of boosting algorithm described in section 3.3 of chapter 2. When applied to support vector machines it has the interesting property that only misclassified examples and examples with a margin lower than that of the current compression set Z'_h have to be considered as candidates to add to Z'_h to form Z'_{h+1}. Indeed, adding other examples would lead to $f_{Z'_{h+1}} = f_{Z'_h}$, which would clearly not decrease the criterion $\inf_{\lambda \in [1,2N]} B(\lambda, h, f)$ which is an increasing function of h for f fixed.

Using a compression scheme based on support vector machines can also be tailored to perform some feature extraction. Imagine that the kernel K is defined as the scalar product in $L_2(\pi)$, where $\pi \in \mathcal{M}_+^1(\Theta)$. More precisely let us consider for some set of soft classification rules $\left\{ f_\theta : \mathcal{X} \to \mathbb{R} \, ; \theta \in \Theta \right\}$ the kernel

$$K(x, x') = \int_{\theta \in \Theta} f_\theta(x) f_\theta(x') \pi(d\theta).$$

In this setting

$$f_{Z'_h}(x') = \int_{\theta \in \Theta} \sum_{(x,y) \in Z'_h} y \alpha(x) f_\theta(x) f_\theta(x') \pi(d\theta)$$

and we may replace it with some finite approximation

$$\widetilde{f}_{Z'_h}(x') = \sum_{k=1}^m w_k f_{\theta_k}(x'),$$

where the set $\{\theta_k,\, k = 1, \ldots, m\}$ and the weights $\{w_k,\, k = 1, \ldots, m\}$ are computed in some suitable way from Z'_h. For instance, we can draw $\{\theta_k,\, k = 1, \ldots, m\}$ at random according to the probability distribution proportional to

$$\left| \sum_{(x,y) \in Z'_h} y \alpha(x) f_\theta(x) \right| \pi(d\theta),$$

define the weights w_k by

$$w_k = \operatorname{sign} \left(\sum_{(x,y) \in Z'_h} y \alpha(x) f_{\theta_k}(x) \right) \int_{\theta \in \Theta} \left| \sum_{(x,y) \in Z'_h} y \alpha(x) f_\theta(x) \right| \pi(d\theta),$$

and choose the smallest value of m for which this approximation still classifies Z'_h without errors.

More generally, given Z'_h, we can select a finite set of features $\Theta' \subset \Theta$ such that Z'_h is $K_{\Theta'}$ separable, where $K_{\Theta'}(x, x') = \sum_{\theta \in \Theta'} f_\theta(x) f_\theta(x')$ and consider the support vector machines $f_{Z'_h}$ built with the kernel $K_{\Theta'}$. As soon as Θ' is chosen as a function of Z'_h only, proposition 3.1 applies and provides some level of confidence for the risk of $f_{Z'_h}$.

In the i.i.d. case, we can furthermore apply the results of chapter 4 to support vector machines at the price of simulating posterior distributions.

6. Building kernels

The results of this section (except the last one) are drawned from [**16**]. We have no reference for the last proposition of this section, although we believe it is well known. We include them for the convenience of the reader.

PROPOSITION 6.1. *Let K_1 and K_2 be positive symmetric kernels on $\mathcal{X}$. Then for any $a \in \mathbb{R}_+$*

$$(aK_1 + K_2)(x, x') \stackrel{def}{=} aK_1(x, x') + K_2(x, x')$$
$$\text{and } (K_1 \cdot K_2)(x, x') \stackrel{def}{=} K_1(x, x')K_2(x, x')$$

are also positive symmetric kernels. Moreover, for any measurable function $g : \mathcal{X} \to \mathbb{R}$, $K_g(x, x') \stackrel{def}{=} g(x)g(x')$ is also a positive symmetric kernel.

PROOF. It is enough to prove the proposition in the case when $\mathcal{X}$ is finite and kernels are just ordinary symmetric matrices. Thus we can assume without loss of generality that $\mathcal{X} = \{1, \dots, n\}$. Then for any $\alpha \in \mathbb{R}^N$, using usual matrix notations,

$$\begin{aligned}
\langle \alpha, (aK_1 + K_2)\alpha \rangle &= a\langle \alpha, K_1\alpha \rangle + \langle \alpha, K_2\alpha \rangle \geq 0, \\
\langle \alpha, (K_1 \cdot K_2)\alpha \rangle &= \sum_{i,j} \alpha_i K_1(i,j) K_2(i,j) \alpha_j \\
&= \sum_{i,j,k} \alpha_i K_1^{1/2}(i,k) K_1^{1/2}(k,j) K_2(i,j) \alpha_j \\
&= \sum_k \underbrace{\sum_{i,j} \big[K_1^{1/2}(k,i)\alpha_i\big] K_2(i,j) \big[K_1^{1/2}(k,j)\alpha_j\big]}_{\geq 0} \geq 0, \\
\langle \alpha, K_g\alpha \rangle &= \sum_{i,j} \alpha_i g(i) g(j) \alpha_j = \left(\sum_i \alpha_i g(i) \right)^2 \geq 0.
\end{aligned}$$

□

PROPOSITION 6.2. *Let K be some positive symmetric kernel on $\mathcal{X}$. Let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial with positive coefficients. Let $g : \mathcal{X} \to \mathbb{R}^d$ be a measurable function. Then*

$$p(K)(x, x') \stackrel{def}{=} p\big[K(x, x')\big],$$
$$\exp(K)(x, x') \stackrel{def}{=} \exp\big[K(x, x')\big]$$
$$\text{and } G_g(x, x') \stackrel{def}{=} \exp\big(-\|g(x) - g(x')\|^2\big)$$

are all positive symmetric kernels.

PROOF. The first assertion is a direct consequence of the previous proposition. The second one comes from the fact that the exponential function is the pointwise

limit of a sequence of polynomial functions with positive coefficients. The third one is seen from the second one and the decomposition

$$G_g(x,x') = \Big[\exp\big(-\|g(x)\|^2\big)\exp\big(-\|g(x')\|^2\big)\Big]\exp\big[2\langle g(x),g(x')\rangle\big]$$

□

PROPOSITION 6.3. *With the notations of the previous proposition,* any *training set* $Z = (x_i,y_i)_{i=1}^N \in \big(\mathcal{X}\times\{-1,+1\}\big)^N$ *is* G_g*-separable as soon as* $g(x_i)$, $i = 1,\dots,N$ *are distinct points of* $\mathbb{R}^d$.

PROOF. It is clearly enough to prove the case when $\mathcal{X} = \mathbb{R}^d$ and g is the identity. Let us consider some other generic point $x_{N+1} \in \mathbb{R}^d$ and define Ψ as in (4.1). It is enough to prove that $\Psi(x_1),\dots,\Psi(x_N)$ are affine independent, since the simplex, and therefore any affine independent set of points can be shattered by affine half-spaces. Let us assume that $(x_1,\dots,x_N)$ are affine dependent, this means that for some $(\lambda_1,\dots,\lambda_N) \neq 0$ such that $\sum_{i=1}^N \lambda_i = 0$,

$$\sum_{i=1}^N\sum_{j=1}^N \lambda_i G(x_i,x_j)\lambda_j = 0.$$

Thus, $(\lambda_i)_{i=1}^{N+1}$, where we have put $\lambda_{N+1} = 0$ is in the kernel of the symmetric positive semi-definite matrix $G(x_i,x_j)_{i,j\in\{1,\dots,N+1\}}$. Therefore

$$\sum_{i=1}^N \lambda_i G(x_i,x_{N+1}) = 0,$$

for any $x_{N+1} \in \mathbb{R}^d$. This would mean that the functions $x \mapsto \exp(-\|x-x_i\|^2)$ are linearly dependent, which can be easily proved to be false. Indeed, let $n \in \mathbb{R}^d$ be such that $\|n\| = 1$ and $\langle n,x_i\rangle$, $i = 1,\dots,N$ are distinct (such a vector exists, because it has to be outside the union of a finite number of hyperplanes, which is of zero Lebesgue measure on the sphere). Let us assume for a while that for some $(\lambda_i)_{i=1}^N \in \mathbb{R}^N$, for any $x \in \mathbb{R}^d$,

$$\sum_{i=1}^N \lambda_i \exp(-\|x-x_i\|^2) = 0.$$

Considering $x = tn$, for $t \in \mathbb{R}$, we would get

$$\sum_{i=1}^N \lambda_i \exp(2t\langle n,x_i\rangle - \|x_i\|^2) = 0, \qquad t \in \mathbb{R}.$$

Letting t go to infinity, we see that this is only possible if $\lambda_i = 0$ for all values of i. □

CHAPTER 6

VC dimension of linear rules with margin constraints

1. How far can subsets be linearly separated

The proof of the following theorem has been suggested to us by a similar proof presented in [**16**].

THEOREM 1.1. *Consider a family of points $(x_1, \dots, x_n)$ in some Euclidean vector space E and a family of affine functions*

$$\mathcal{H} = \big\{ g_{w,b} : E \to \mathbb{R} \,;\, w \in E, \|w\| = 1, b \in \mathbb{R} \big\},$$

where

$$g_{w,b}(x) = \langle w, x \rangle - b, \qquad x \in E.$$

Assume that there is a set of thresholds $(b_i)_{i=1}^n \in \mathbb{R}^n$ such that for any $(y_i)_{i=1}^n \in \{-1, +1\}^n$, there is $g_{w,b} \in \mathcal{H}$ such that

$$\inf_{i=1}^n \big(g_{w,b}(x_i) - b_i \big) y_i \geq \gamma.$$

Let us also introduce the empirical variance of $(x_i)_{i=1}^n$,

$$\mathbb{V}\mathrm{ar}(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n \left\| x_i - \frac{1}{n} \sum_{j=1}^n x_j \right\|^2.$$

In this case and with these notations,

$$\frac{\mathbb{V}\mathrm{ar}(x_1, \dots, x_n)}{\gamma^2} \geq \begin{cases} n-1 & \text{when } n \text{ is even,} \\ (n-1)\frac{n^2-1}{n^2} & \text{when } n \text{ is odd.} \end{cases} \tag{1.1}$$

Moreover, equality is reached when γ is optimal, $b_i = 0$, $i = 1, \dots, n$ and $(x_1, \dots, x_n)$ is a regular simplex (i.e. when 2γ is the minimum distance between the convex hulls of any two subsets of $\{x_1, \dots, x_n\}$ and $\|x_i - x_j\|$ does not depend on $i \neq j$).

PROOF. Let $(s_i)_{i=1}^n \in \mathbb{R}^n$ be such that $\sum_{i=1}^n s_i = 0$. Let σ be a uniformly distributed random variable with values in $\mathfrak{S}_n$, the set of permutations of the first n integers $\{1, \dots, n\}$. By assumption, for any value of σ, there is an affine function $g_{w,b} \in \mathcal{H}$ such that

$$\min_{i=1,\dots,n} \big[g_{w,b}(x_i) - b_i \big] \big[2\mathbb{1}(s_{\sigma(i)} > 0) - 1 \big] \geq \gamma.$$

As a consequence

$$\left\langle \sum_{i=1}^n s_{\sigma(i)} x_i, w \right\rangle = \sum_{i=1}^n s_{\sigma(i)} \big(\langle x_i, w \rangle - b - b_i \big) + \sum_{i=1}^n s_{\sigma(i)} b_i$$

$$\geq \sum_{i=1}^{n} \gamma |s_{\sigma(i)}| + s_{\sigma(i)} b_i .$$

Therefore, using the fact that the map $x \mapsto \Big(\max\big\{0, x\big\}\Big)^2$ is convex,

$$\begin{aligned}
\mathbb{E}\left(\left\| \sum_{i=1}^{n} s_{\sigma(i)} x_i \right\|^2 \right) &\geq \mathbb{E}\left[\left(\max\left\{ 0, \sum_{i=1}^{n} \gamma |s_{\sigma(i)}| + s_{\sigma(i)} b_i \right\} \right)^2 \right] \\
&\geq \left(\max\left\{ 0, \sum_{i=1}^{n} \gamma \mathbb{E}\big(|s_{\sigma(i)}|\big) + \mathbb{E}\big(s_{\sigma(i)}\big) b_i \right\} \right)^2 = \gamma^2 \left(\sum_{i=1}^{n} |s_i| \right)^2 ,
\end{aligned}$$

where $\mathbb{E}$ is the expectation with respect to the random permutation σ. On the other hand

$$\mathbb{E}\left(\left\| \sum_{i=1}^{n} s_{\sigma(i)} x_i \right\|^2 \right) = \sum_{i=1}^{n} \mathbb{E}(s_{\sigma(i)}^2) \|x_i\|^2 + \sum_{i \neq j} \mathbb{E}(s_{\sigma(i)} s_{\sigma(j)}) \langle x_i, x_j \rangle .$$

Moreover

$$\mathbb{E}(s_{\sigma(i)}^2) = \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^{n} s_{\sigma(i)}^2 \right) = \frac{1}{n} \sum_{i=1}^{n} s_i^2 .$$

In the same way, for any $i \neq j$,

$$\begin{aligned}
\mathbb{E}\left(s_{\sigma(i)} s_{\sigma(j)} \right) &= \frac{1}{n(n-1)} \mathbb{E}\left(\sum_{i \neq j} s_{\sigma(i)} s_{\sigma(j)} \right) \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} s_i s_j \\
&= \frac{1}{n(n-1)} \left[\underbrace{\left(\sum_{i=1}^{n} s_i \right)^2}_{=0} - \sum_{i=1}^{n} s_i^2 \right] \\
&= -\frac{1}{n(n-1)} \sum_{i=1}^{n} s_i^2 .
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}\left(\left\| \sum_{i=1}^{n} s_{\sigma(i)} x_i \right\|^2 \right) &= \left(\sum_{i=1}^{n} s_i^2 \right) \left[\frac{1}{n} \sum_{i=1}^{n} \|x_i\|^2 - \frac{1}{n(n-1)} \sum_{i \neq j} \langle x_i, x_j \rangle \right] \\
&= \left(\sum_{i=1}^{n} s_i^2 \right) \left[\left(\frac{1}{n} + \frac{1}{n(n-1)} \right) \sum_{i=1}^{n} \|x_i\|^2 \right. \\
&\qquad \left. - \frac{1}{n(n-1)} \left\| \sum_{i=1}^{n} x_i \right\|^2 \right] \\
&= \frac{n}{n-1} \left(\sum_{i=1}^{n} s_i^2 \right) \mathbb{V}\mathrm{ar}(x_1, \ldots, x_n).
\end{aligned}$$

We have proved that

$$\frac{\mathbb{V}\mathrm{ar}(x_1,\dots,x_n)}{\gamma^2} \geq \frac{(n-1)\Big(\sum_{i=1}^n |s_i|\Big)^2}{n\sum_{i=1}^n s_i^2}.$$

This can be used with $s_i = \mathbb{1}(i \leq \frac{n}{2}) - \mathbb{1}(i > \frac{n}{2})$ in the case when n is even and $s_i = \frac{2}{(n-1)}\mathbb{1}(i \leq \frac{n-1}{2}) - \frac{2}{n+1}\mathbb{1}(i > \frac{n-1}{2})$ in the case when n is odd to establish the first inequality (1.1) of the theorem.

Checking that equality is reached for the simplex is an easy computation when the simplex $(x_i)_{i=1}^n \in (\mathbb{R}^n)^n$ is parametrized in such a way that

$$x_i(j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed the distance between the convex hulls of any two subsets of the simplex is the distance between their mean values (i.e. centers of mass). □

2. Application to support vector machines

We are going to apply theorem 1.1 to support vector machines in the transductive case. So let us consider $(X_i, Y_i)_{i=1}^{2N}$ distributed according to some exchangeable distribution P_{2N} and assume that $(X_i)_{i=1}^{2N}$ and $(Y_i)_{i=1}^{N}$ are observed. Let us consider some positive kernel K on $\mathcal{X}$. For any K-separable training set of the form $Z' = (X_i, y_i')_{i=1}^{2N}$, where $(y_i')_{i=1}^{2N} \in \mathcal{Y}^{2N}$, let $\hat{f}_{Z'}$ be the support vector machine defined by K and Z' and let $\gamma(Z')$ be its margin. Let

$$R^2 = \max_{i=1,\dots,2N} K(x_i,x_i) + \frac{1}{4N^2}\sum_{j=1}^{2N}\sum_{k=1}^{2N} K(x_j,x_k) - \frac{1}{N}\sum_{j=1}^{2N} K(x_i,x_j).$$

(This is an easily computable upper-bound for the radius of some ball containing the image of $(X_1,\dots,X_{2N})$ in feature space.)

Let us define for any integer h the margins

$$\gamma_{2h} = (2h-1)^{-1/2} \text{ and } \gamma_{2h+1} = \left[2h\left(1 - \frac{1}{(2h+1)^2}\right)\right]^{-1/2}. \tag{2.1}$$

Let us consider for any $h = 1,\dots,N$ the exchangeable model

$$\mathcal{R}_h = \big\{\hat{f}_{Z'} : Z' = (X_i, y_i')_{i=1}^{2N} \text{ is } K\text{-separable and } \gamma(Z') \geq R\gamma_h\big\}.$$

The family of models $\mathcal{R}_h$, $h = 1,\dots,N$ is nested, and we know from theorem 1.1 of this chapter and theorems 1.2 and 1.3 of chapter 2 that

$$|\mathcal{R}_h| \leq h\Big[\log\big(\tfrac{2N}{h}\big) + 1\Big].$$

We can then consider on the large model $\mathcal{R} = \bigsqcup_{h=1}^N \mathcal{R}_h$ (the disjoint union of the submodels) an exchangeable prior π which is uniform on each $\mathcal{R}_h$ and is such that $\pi(\mathcal{R}_h) \geq (1-\alpha)\alpha^h$ for some parameter $\alpha \in]0,1[$. Applying corollary 2.3 of chapter 2, and taking as posterior all the Dirac masses, we get

PROPOSITION 2.1. *For any $\lambda \in]0, 2N[$, with P_{2N} probability at least $1-\epsilon$, for any $h = 1, \ldots, N$, any support vector machine $f \in \mathcal{R}_h$,*

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

(This proposition could of course be made uniform in λ in the standard way many times explained in this paper.)

3. Non transductive margin bounds for support vector machines

In order to establish non transductive margin bounds, we will need a different combinatorial lemma. It is due to [**1**]. We will reproduce their proof with some tiny improvements on the values of constants.

Let us consider the finite case when $\mathcal{X} = \{1, \ldots, n\}$, $\mathcal{Y} = \{1, \ldots, b\}$ and $b \geq 3$ (the question we will study would be meaningless in the case when $b \leq 2$). Assume as usual that we are dealing with a prescribed set of classification rules $\mathcal{R} = \{f : \mathcal{X} \to \mathcal{Y}\}$. Let us say that a pair (A, s), where $A \subset \mathcal{X}$ is a non empty set of shapes and $s : A \to \{2, \ldots, b-1\}$ a threshold function, is *shattered* by the set of functions $F \subset \mathcal{R}$ if for any $(\sigma_x)_{x \in A} \in \{-1, +1\}^A$, there exists some $f \in F$ such that $\min_{x \in A} \sigma_x [f(x) - s(x)] \geq 1$.

DEFINITION 3.1. Let the *fat shattering dimension* of $(\mathcal{X}, \mathcal{R})$ be the maximal size $|A|$ of the first component of the pairs which are shattered by $\mathcal{R}$.

Let us say that a subset of classification rules $F \subset \mathcal{Y}^{\mathcal{X}}$ is *separated* whenever for any pair $(f, g) \in F^2$ such that $f \neq g$, $\|f - g\|_\infty = \max_{x \in \mathcal{X}} |f(x) - g(x)| \geq 2$. Let $\mathfrak{M}(\mathcal{R})$ be the maximum size $|F|$ of separated subsets F of $\mathcal{R}$. Note that if F is a separated subset of $\mathcal{R}$ such that $|F| = \mathfrak{M}(\mathcal{R})$, then it is a 1-net for the $\mathcal{L}_\infty$ distance: for any function $f \in \mathcal{R}$ there exists $g \in F$ such that $\|f - g\|_\infty \leq 1$ (otherwise f could be added to F to create a larger separated set).

LEMMA 3.1. *With the above notations, whenever the fat shattering dimension of $(\mathcal{X}, \mathcal{R})$ is not greater than h,*

$$\log\left[\mathfrak{M}(\mathcal{R})\right] < \log\left[(b-1)(b-2)n\right] \left\{ \frac{\log\left[\sum_{i=1}^h \binom{n}{i}(b-2)^i\right]}{\log(2)} + 1 \right\} + \log(2)$$
$$\leq \log\left[(b-1)(b-2)n\right] \left\{ \left[\log\left[\tfrac{(b-2)n}{h}\right] + 1\right] \frac{h}{\log(2)} + 1 \right\} + \log(2).$$

PROOF. For any set of functions $F \subset \mathcal{Y}^{\mathcal{X}}$, let $t(F)$ be the number of pairs (A, s) shattered by F. Let $t(m, n)$ be the minimum of $t(F)$ over all *separated* sets of functions $F \subset \mathcal{Y}^{\mathcal{X}}$ of size $|F| = m$ (n is here to recall that the shape space $\mathcal{X}$ is made of n shapes). For any m such that $t(m, n) > \sum_{i=1}^h \binom{n}{i}(b-2)^i$, it is clear that any separated set of functions of size $|F| \geq m$ shatters at least one pair (A, s) such that $|A| > h$. Indeed, $t(m, n)$ is clearly from its definition a non decreasing function of m, so that $t(|F|, n) > \sum_{i=1}^h \binom{n}{i}(b-2)^i$. Moreover there are only $\sum_{i=1}^h \binom{n}{i}(b-2)^i$ pairs (A, s) such that $|A| \leq h$. As a consequence, whenever the fat shattering dimension of $(\mathcal{X}, \mathcal{R})$ is not greater than h we have $\mathfrak{M}(\mathcal{R}) < m$.

It is clear that for any $n \geq 1$, $t(2,n) = 1$.

LEMMA 3.2. *For any $m \geq 1$, $t\big[mn(b-1)(b-2), n\big] \geq 2t\big[m, n-1\big]$, and therefore $t\big[2n(n-1)\dots(n-r+1)(b-1)^r(b-2)^r, n\big] \geq 2^r$.*

PROOF. Let $F = \{f_1, \dots, f_{mn(b-1)(b-2)}\}$ be some separated set of functions of size $mn(b-1)(b-2)$. For any pair (f_{2i-1}, f_{2i}), $i = 1, \dots, mn(b-1)(b-2)/2$, there is $x_i \in \mathcal{X}$ such that $|f_{2i-1}(x_i) - f_{2i}(x_i)| \geq 2$. Since $|\mathcal{X}| = n$, there is $x \in \mathcal{X}$ such that $\sum_{i=1}^{mn(b-1)(b-2)/2} \mathbb{1}(x_i = x) \geq m(b-1)(b-2)/2$. Let $I = \{i : x_i = x\}$. Since there are $(b-1)(b-2)/2$ pairs $(y_1, y_2) \in \mathcal{Y}^2$ such that $1 \leq y_1 < y_2 \leq b$, there is some pair (y_1, y_2), such that $1 \leq y_1 < y_2 \leq b$ and such that $\sum_{i \in I} \mathbb{1}(\{y_1, y_2\} = \{f_{2i-1}(x), f_{2i}(x)\}) \geq m$. Let $J = \big\{i \in I : \{f_{2i-1}(x), f_{2i}(x)\} = \{y_1, y_2\}\big\}$. Let

$$F_1 = \{f_{2i-1} : i \in J, f_{2i-1}(x) = y_1\} \cup \{f_{2i} : i \in J, f_{2i}(x) = y_1\},$$
$$F_2 = \{f_{2i-1} : i \in J, f_{2i-1}(x) = y_2\} \cup \{f_{2i} : i \in J, f_{2i}(x) = y_2\}.$$

Obviously $|F_1| = |F_2| = |J| = m$. Moreover the restrictions of the functions of F_1 to $\mathcal{X} \setminus \{x\}$ are separated, and it is the same with F_2. Thus F_1 strongly shatters at least $t(m, n-1)$ pairs (A, s) such that $A \subset \mathcal{X} \setminus \{x\}$ and it is the same with F_2. Eventually, if the pair (A, s) where $A \subset \mathcal{X} \setminus \{x\}$ is both shattered by F_1 and F_2, then $F_1 \cup F_2$ shatters also $(A \cup \{x\}, s')$ where $s'(x') = s(x')$ for any $x' \in A$ and $s'(x) = \lfloor \frac{y_1+y_2}{2} \rfloor$. Thus $F_1 \cup F_2$, and therefore F, shatters at least $2t(m, n-1)$ pairs (A, s). □

Resuming the proof of lemma 3.1, let us choose for r the smallest integer such that $2^r > \sum_{i=1}^h \binom{n}{i}(b-2)^i$, which is no greater than $\left\{ \frac{\log\left[\sum_{i=1}^h \binom{n}{i}(b-2)^i\right]}{\log(2)} + 1 \right\}$. In the case when $1 \leq n \leq r$,

$$\log(\mathfrak{M}(\mathcal{R})) < |\mathcal{X}| \log(|\mathcal{Y}|) = n \log(b) \leq r \log(b) \leq r \log\big[(b-1)(b-2)n\big] + \log(2),$$

which proves the lemma. In the remaining case $n > r$,

$$t\big[2n^r(b-1)^r(b-2)^r, n\big] \geq t\big[2n(n-1)\dots(n-r+1)(b-1)^r(b-2)^r, n\big] > \sum_{i=1}^h \binom{n}{i}(b-2)^i.$$

Thus $|\mathfrak{M}(\mathcal{R})| < 2\Big[(b-2)(b-1)n\Big]^r$ as claimed. □

In order to apply this combinatorial lemma to support vector machines, let us consider now the case of separating hyperplanes in $\mathbb{R}^d$. Assume that $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. For any sample X_1^{2N}, let

$$R(X_1^{2N}) = \max\{\|X_i\| : 1 \leq i \leq 2N\}.$$

Let us consider the set of parameters

$$\Theta = \big\{(w, b) \in \mathbb{R}^d \times \mathbb{R} : \|w\| = 1\big\}.$$

For any $(w, b) \in \Theta$, let $g_{w,b}(x) = \langle w, x \rangle - b$. Let h be some fixed integer and let $\gamma = R(X_1^{2N})\gamma_h$, where γ_h is defined by equation (2.1).

Let us define $\zeta : \mathbb{R} \to \mathbb{Z}$ by

$$\zeta(r) = \begin{cases} -5 & \text{when} \quad r \leq -4\gamma, \\ -3 & \text{when} \quad -4\gamma < r \leq -2\gamma, \\ -1 & \text{when} \quad -2\gamma < r \leq 0, \\ +1 & \text{when} \quad 0 < r \leq 2\gamma, \\ +3 & \text{when} \quad 2\gamma < r \leq 4\gamma, \\ +5 & \text{when} \quad 4\gamma < r. \end{cases}$$

Let $G_{w,b}(x) = \zeta\big[g_{w,b}(x)\big]$. The fat shattering dimension (as defined in 3.1) of

$$\Big(X_1^{2N}, \big\{(G_{w,b} + 7)/2 : (w,b) \in \Theta\big\}\Big)$$

is not greater than h (according to theorem 1.1), therefore there is some set $\mathcal{F}$ of functions from X_1^{2N} to $\{-3,-2,-1,+1,+2,+3\}$ such that

$$\log\big(|\mathcal{F}|\big) \leq \log(40N)\left\{\frac{h}{\log(2)}\left[\log\left(\frac{8N}{h}\right)+1\right]+1\right\}+\log(2).$$

and for any $(w,b) \in \Theta$, there is $f_{w,b} \in \mathcal{F}$ such that $\sup\big\{|f_{w,b}(X_i) - G_{w,b}(X_i)| : i = 1,\ldots,2N\big\} \leq 2$. Moreover, the choice of $f_{w,b}$ may be required to depend on X_1^{2N} in an exchangeable way. Very similarly to the proof of corollary 2.3 of chapter 2, it can be proved that for any exchangeable probability distribution $P_{2N} \in \mathcal{M}_+^1\big[(\mathcal{X}\times\mathcal{Y})^{2N}\big]$, for any $\lambda \in]0, 2N[$, with P_{2N} probability at least $1-\epsilon$, for any $f_{w,b} \in \mathcal{F}$,

$$\begin{aligned}\frac{1}{N}\sum_{i=N+1}^{2N} \mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big] \\ \leq \left(1-\frac{\lambda}{2N}\right)^{-1}\left\{\left(1+\frac{\lambda}{2N}\right)\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big] \right. \\ \left. + \frac{1}{\lambda}\Big[\log\big(|\mathcal{F}|\big) + \log(\epsilon^{-1})\Big]\right\}.\end{aligned}$$

Let us remark that

$$\mathbb{1}\Big\{2\mathbb{1}\big[g_{w,b}(X_i) \geq 0\big] - 1 \neq Y_i\Big\} = \mathbb{1}\big[G_{w,b}(X_i)Y_i < 0\big] \leq \mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big]$$

and

$$\mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big] \leq \mathbb{1}\big[G_{w,b}(X_i)Y_i \leq 3\big] \leq \mathbb{1}\big[g_{w,b}(X_i)Y_i \leq 4\gamma\big].$$

This proves the following theorem.

THEOREM 3.3. *For any $\lambda \in]0, 2N[$, any positive integer h, with P_{2N} probability at least $1-\epsilon$, for any $(w,b) \in \Theta$,*

$$\begin{aligned}\frac{1}{N}\sum_{i=N+1}^{2N} \mathbb{1}\Big\{2\mathbb{1}\big[g_{w,b}(X_i) \geq 0\big] - 1 \neq Y_i\Big\} \\ \leq \left(1-\frac{\lambda}{2N}\right)^{-1}\left\{\left(1+\frac{\lambda}{2N}\right)\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\big[g_{w,b}(X_i)Y_i \leq 4R(X_1^{2N})\gamma_h\big] \right. \\ \left. + \frac{1}{\lambda}\left[\log(40N)\Big\{\tfrac{h}{\log(2)}\big[\log\big(\tfrac{8N}{h}\big)+1\big]+1\Big\} + \log(2\epsilon^{-1})\right]\right\}.\end{aligned}$$

As usual, this result can be made uniform in h and λ. A simple consequence though, in the case when $R(X_1^{2N}) \leq R$ and

$$\sum_{i=1}^{N} \mathbb{1}\big[g_{w,b}(X_i)Y_i \leq \gamma\big] = 0,$$

is obtained by choosing $\lambda = N$ and h to be the smallest integer such that $\gamma \geq R\gamma_h$. This choice shows that with P_{2N} probability at least $1-\epsilon$, for any $(w,b) \in \Theta$,

$$\frac{1}{N}\sum_{i=N+1}^{2N} \mathbb{1}\big[g_{w,b}(X_i)Y_i < 0\big] \leq \frac{2}{N}\Bigg\{\log(40N)\Bigg\{\frac{1}{\log(2)}\left(\frac{16R^2}{\gamma^2}+2\right) \times \left[\log\left(\frac{N\gamma^2}{2R^2}\right)+1\right]+1\Bigg\} + \log\left(\frac{2}{\epsilon}\right)\Bigg\}.$$

This inequality compares favourably with similar inequalities in [**16**], which moreover do not extend to the margin quantile case as this one.

Let us also remark that it is easy to circonvent the fact that $R(X_1^{2N})$ is not observed when the test set X_{N+1}^{2N} is not observed.

Indeed, we can consider the sample obtained by projecting X_1^{2N} on some ball of fixed radius R, putting

$$X_i(R) = \min\left\{1, \frac{R}{\|X_i\|}\right\} X_i.$$

As a consequence of the previous theorem,

COROLLARY 3.4. *For any $\lambda \in]0, 2N[$ and any positive integer h, with P_{2N} probability at least $1-\epsilon$, for any $(w,b) \in \Theta$,*

$$\frac{1}{N}\sum_{i=N+1}^{2N} \mathbb{1}\Big\{2\mathbb{1}\Big[g_{w,b}\big[X_i(R)\big] \geq 0\Big] - 1 \neq Y_i\Big\} \leq \left(1-\frac{\lambda}{2N}\right)^{-1}\Bigg\{\left(1+\frac{\lambda}{2N}\right)\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\Big[g_{w,b}\big[X_i(R)\big]Y_i \leq 4R\gamma_h\Big] + \frac{1}{\lambda}\Bigg[\log(40N)\Big\{\tfrac{h}{\log(2)}\big[\log\big(\tfrac{8N}{h}\big)+1\big]+1\Big\} + \log(2\epsilon^{-1})\Bigg]\Bigg\}.$$

Choosing a sequence R_m, $m \in \mathbb{N}$ of values of R increasing to infinity, we can make this result uniform in λ, h and m (using a union bound), and optimize the right-hand side of the above inequality in λ, h and m.

Conclusion

It is our hope that this PAC-Bayesian study of adaptive classification will have provided some useful insights on an already powerful but in the same time still promising technique. The main improvements on previous works are localization and the use of exchangeable priors. It puts Vapnik's theory in a new perspective and opens the road for new mathematically justified ways of adapting classification algorithms to the data. The use of compression schemes should in particular be very flexible. We hope to investigate further applications in forthcoming papers. It is also to be mentioned that a substantial part of the techniques used here are not specific to classification and are relevant in any situation where a non asymptotic deviation inequality is available for each value of the parameter (in this respect, some materials about the use of PAC-Bayesian tools in the regression setting can be found in [**15**], this is where we felt for the first time the interest and possibility of getting localized bounds).

Bibliography

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale sensitive dimensions, uniform convergence and learnability, J. of ACM **44**(4):615-631, 1997.

[2] J. Y. Audibert, Aggregated estimators and empirical complexity for least square regression, *preprint at* `http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2003`, 2003.

[3] 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.

[4] A. Barron and Y. Yang (1995) Information Theoretic Determination of Minimax Rates of Convergence, *Ann. Statist.* **27** (1999), no. 5, 1564–1599.

[5] 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.

[6] G. Blanchard, The "progressive mixture" estimator for regression trees, *Annales de l'I.H.P.*, **35**(6):793-820, 1999.

[7] G. Blanchard, A new algorithm for Bayesian MCMC CART sampling, *preprint*, 2000.

[8] 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.

[9] L. Birgé and P. Massart (1997) From model selection to adaptive estimation, *Festschrift for Lucien Le Cam*, 55–87, Springer, New York.

[10] L. Birgé and P. Massart (1995) Minimum contrast estimators on sieves, Bernoulli **4** (1998), no. 3, 329–375.

[11] L. Birgé and P. Massart, A generalized C_p criterion for Gaussian model selection, *preprint*, 2001.

[12] L. Birgé and P. Massart, Gaussian model selection, *J. Eur. Math. Soc.*, 2001. `http://www.springer.de/link`

[13] O. Catoni, Laplace transform estimates and deviation inequalities *Ann. Inst. H. Poincaré Probab. Statist.* **39** (2003), no. 1, 1–26.

[14] O. Catoni, Data compression and adaptive histograms. In F. Cucker and J.M. Rojas, editors, *Foundations of Computational Mathematics, Proceedings of the Smalefest 2000*, pages 35–60. World Scientific, 2002.

[15] O. Catoni, Statistical learning theory and stochastic optimization, *Lecture notes, Saint-Flour summer school on Probability Theory, 2001*, Springer, to appear.

[16] N. Cristianini and J. Shawe Taylor, *An introduction to Support Vector Machines and other kernel based learning methods,* Cambridge University Press, 2000.

[17] M. Feder and N. Merhav, Hierarchical Universal Coding, *IEEE Trans. Inform. Theory*, vol 42, no 5, Sept, 1996.

[18] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning*, Springer, New York, 2001.

[19] J. Langford and M. Seeger, Bounds for Averaging Classifiers, *technical report CMU-CS-01-102*, Carnegie Mellon University, jan. 2001, `www.cs.cmu.edu/~jcl`.

[20] J. Langford, M. Seeger and N. Megiddo, An Improved Predictive Accuracy Bound for Averaging Classifiers, *International Conference on Machine Learning* **18** (2001), 290-297.

[21] V. N. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.

[22] N. Littlestone and M. Warmuth, Relating data compression and learnability. *Technical report,* University of California, Santa Cruz, 1986.

[23] 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;

[24] 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;

[25] C. McDiarmid, Concentration *Probabilistic Methods for Algorithmic Discrete Mathematics*, Habib M., McDiarmid C. and Reed B. Eds., Springer, 1998.

[26] B. Y. Ryabko, Twice-universal coding, *Probl. Inform. Transm.*, vol 20, no 3, pp. 24-28, July-Sept 1984.

[27] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, *Journal of Machine Learning Research* **3** (2002), 233–269.

[28] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, *Informatics Research Report, Division of Informatics, University of Edinburgh* EDI-INF-RR-0094 (2002), 1–42. `http://www.informatics.ed.ac.uk`

[29] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, *IEEE Trans. Inform. Theory* **44** (1998), no. 5, 1926–1940.

[30] J. Shawe-Taylor and N. Cristianini, On the generalization of soft margin algorithms, *IEEE Trans. Inform. Theory* **48** (2002), no. 10, 2721–2735.

[31] J.-P. Vert, Double mixture and universal inference, *preprint*, 2000, `http://cg.ensmp.fr/~vert/publi/` .

[32] J.-P. Vert, Adaptive context trees and text clustering, *IEEE Trans. Inform. Theory*, **47** (2001), no. 5, 1884-1901.

[33] J.-P. Vert, Text categorization using adaptive context trees, *Proceedings of the CICLing-2001 conference*, A. Gelbukh (Ed.), LNCS 2004, Springer-Verlag, pp. 423-436, 2001.

[34] 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.

[35] 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.