PAC-BAYESIAN SUPERVISED CLASSIFICATION

(The thermodynamics of statistical learning)

Olivier Catoni

August 10, 2007

Abstract: This monograph deals with adaptive supervised classification, using tools borrowed from statistical mechanics and information theory, stemming from the PAC-Bayesian approach pioneered by David McAllester and applied to a conception of statistical learning theory forged by Vladimir Vapnik. Using convex analysis on the set of posterior probability measures, we show how to get local measures of the complexity of the classification model involving the relative entropy of posterior distributions with respect to Gibbs posterior measures. We then discuss relative bounds, comparing the generalization error of two classification rules, showing how the margin assumption of Mammen and Tsybakov can be replaced with some empirical measure of the covariance structure of the classification model. We show how to associate to any posterior distribution an *effective temperature* relating it to the Gibbs prior distribution with the same level of expected error rate, and how to estimate this effective temperature from data, resulting in an estimator whose expected error rate converges according to the best possible power of the sample size adaptively under any margin and parametric complexity assumptions. We describe and study an alternative selection scheme based on relative bounds between estimators, and present a two step localization technique which can handle the selection of a parametric model from a family of those. We show how to extend systematically all the results obtained in the inductive setting to transductive learning, and use this to improve Vapnik's generalization bounds, extending them to the case when the sample is made of independent non-identically distributed pairs of patterns and labels. Finally we review briefly the construction of Support Vector Machines and show how to derive generalization bounds for them, measuring the complexity either through the number of support vectors or through the value of the transductive or inductive margin.

2000 Mathematics Subject Classification: 62H30, 68T05, 62B10.

Keywords: adaptive statistics, compression schemes, effective temperature, Gibbs posterior distributions, local complexity bounds, Mammen and Tsybakov margin assumption, margin bounds, PAC-Bayesian theorems, pattern recognition, randomized estimators, statistical learning theory, Support Vector Machines, transductive inference, Vapnik–Cervonenkis dimension.

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

Contents

Introduction

Among the possible approaches to pattern recognition, statistical learning theory has received a lot of attention in the last few years. Although a realistic pattern recognition scheme involves data pre-processing and post-processing that need a theory of their own, a central role is often played by some kind of supervised learning algorithm. This central building block is the subject we are going to analyse in these notes.

Accordingly, we assume that we have prepared in some way or another a *sample* of N labelled patterns $(X_i, Y_i)_{i=1}^N$, where X_i ranges in some pattern space $\mathcal{X}$ and Y_i ranges in some finite label set $\mathcal{Y}$. We also assume that we have devised our experiment in such a way that the couples of random variables (X_i, Y_i) are independent (but not necessarily equidistributed). Here, randomness should be understood to come from the way the statistician has planned his experiment. He may for instance have drawn the X_is at random from some larger population of patterns the algorithm is meant to be applied to in a second stage. The labels Y_i may have been set with the help of some external expertise (which may itself be faulty or contain some amount of randomness, so we do not assume that Y_i is a function of X_i, and allow the couple of random variables (X_i, Y_i) to follow any kind of joint distribution). In practice, patterns will be extracted from some high dimensional and highly structured data, such as digital images, speech signals, DNA sequences, etc. We will not discuss this pre-processing stage here, although it poses crucial problems dealing with segmentation and the choice of a representation. The aim of supervised classification is to choose some classification rule $f : \mathcal{X} \to \mathcal{Y}$ which predicts Y from X making as few mistakes as possible on average.

The choice of f will be driven by a suitable use of the information provided by the sample $(X_i, Y_i)_{i=1}^N$ on the joint distribution of X and Y. Moreover, considering all the possible measurable functions f from $\mathcal{X}$ to $\mathcal{Y}$ would not be feasible in practice and maybe more importantly not well founded from a statistical point of view, at least as soon as the pattern space $\mathcal{X}$ is large and little is known in advance about the joint distribution of patterns X and labels Y. Therefore, we will consider parametrized subsets of classification rules $\{f_\theta : \mathcal{X} \to \mathcal{Y}\,;\, \theta \in \Theta_m\}$, $m \in M$, which may be grouped to form a big parameter set $\Theta = \bigcup_{m \in M} \Theta_m$.

The subject of this monograph is to introduce to statistical learning the-

ory, and more precisely to the theory of supervised classification, a number of technical tools akin to statistical mechanics and information theory, dealing with the concepts of entropy and temperature. A central task will in particular be to control the mutual information between an estimated parameter and the observed sample. The focus will not be directly on the description of the data to be classified, but on the description of the classification rules. As we want to deal with high dimensional data, we will be bound to consider high dimensional sets of candidate classification rules, and will analyse them with tools very similar to those used in statistical mechanics to describe particle systems with many degrees of freedom. More specifically, the sets of classification rules will be described by Gibbs measures defined on parameter sets and depending on the observed sample value. A Gibbs measure is the special kind of probability measure used in statistical mechanics to describe the state of a particle system driven by a given energy function at some given temperature. Here, Gibbs measures will emerge as minimizers of the average loss value under entropy (or mutual information) constraints. Entropy itself, more precisely the Kullback divergence function between probability measures, will emerge in conjunction with the use of exponential deviation inequalities: indeed, the log-Laplace transform may be seen as the Legendre transform of the Kullback divergence function, as will be stated in Lemma 1.1.3 (page 16).

To fix notation, let $(X_i, Y_i)_{i=1}^N$ be the canonical process on $\Omega = (\mathcal{X} \times \mathcal{Y})^N$ (which means the coordinate process). Let the pattern space be provided with a sigma-algebra $\mathcal{B}$ turning it into a measurable space $(\mathcal{X}, \mathcal{B})$. On the finite label space $\mathcal{Y}$, we will consider the trivial algebra $\mathcal{B}'$ made of all its subsets. Let $\mathcal{M}_+^1\big[(\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\big]$ be our notation for the set of probability measures (i.e. of positive measures of total mass equal to 1) on the measurable space $\big[(\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B} \times \mathcal{B}')^{\otimes N}\big]$. Once some probability distribution $\mathbb{P} \in \mathcal{M}_+^1\big[(\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\big]$ is chosen, it turns $(X_i, Y_i)_{i=1}^N$ into the canonical realization of a stochastic process modelling the observed sample (also called the training set). We will assume that $\mathbb{P} = \bigotimes_{i=1}^N P_i$, where for each $i = 1, \ldots, N$, $P_i \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{B}')$, to reflect the assumption that we observe independent pairs of patterns and labels. We will also assume that we are provided with some indexed set of possible classification rules

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

where $(\Theta, \mathcal{T})$ is some measurable index set. Assuming some indexation of the classification rules is just a matter of presentation. Although it leads to heavier notation, it allows us to integrate over the space of classification rules as well as over Ω, using the usual formalism of multiple integrals. For this matter, we will assume that $(\theta, x) \mapsto f_\theta(x) : (\Theta \times \mathcal{X}, \mathcal{B} \otimes \mathcal{T}) \to (\mathcal{Y}, \mathcal{B}')$ is a measurable function.

In many cases, as already mentioned, $\Theta = \bigcup_{m \in M} \Theta_m$ will be a finite (or more generally countable) union of subspaces, dividing the classification model $\mathcal{R}_\Theta = \bigcup_{m \in M} \mathcal{R}_{\Theta_m}$ into a union of sub-models. The importance of

introducing such a structure has been put forward by V. Vapnik, as a way to avoid making strong hypotheses on the distribution $\mathbb{P}$ of the sample. If neither the distribution of the sample nor the set of classification rules were constrained, it is well known that no kind of statistical inference would be possible. Considering a family of sub-models is a way to provide for adaptive classification where the choice of the model depends on the observed sample. Restricting the set of classification rules is more realistic than restricting the distribution of patterns, since the classification rules are a processing tool left to the choice of the statistician, whereas the distribution of the patterns is not fully under his control, except for some planning of the learning experiment which may enforce some weak properties like independence, but not the precise shapes of the marginal distributions P_i which are as a rule unknown distributions on some high dimensional space.

In these notes, we will concentrate on general issues concerned with a natural measure of risk, namely the *expected error rate* of each classification rule f_θ, expressed as

$$R(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\big[f_\theta(X_i) \neq Y_i\big]. \tag{0.1}$$

As this quantity is unobserved, we will be led to work with the corresponding *empirical error rate*

$$r(\theta, \omega) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big]. \tag{0.2}$$

This does not mean that practical learning algorithms will always try to minimize this criterion. They often on the contrary try to minimize some other criterion which is linked with the structure of the problem and has some nice additional properties (like smoothness and convexity, for example). Nevertheless, and independently of the precise form of the estimator $\widehat{\theta} : \Omega \to \Theta$ under study, the analysis of $R(\widehat{\theta})$ is a natural question, and often corresponds to what is required in practice.

Answering this question is not straightforward because, although $R(\theta)$ is the expectation of $r(\theta)$, a sum of independent Bernoulli random variables, $R(\widehat{\theta})$ is not the expectation of $r(\widehat{\theta})$, because of the dependence of $\widehat{\theta}$ on the sample, and neither is $r(\widehat{\theta})$ a sum of independent random variables. To circumvent this unfortunate situation, some uniform control over the deviations of r from R is needed.

We will follow the PAC-Bayesian approach to this problem, originated in the machine learning community and pioneered by McAllester (1998, 1999). It can be seen as some variant of the more classical approach of M-estimators relying on empirical process theory — as described for instance in Van de Geer (2000).

It is built on some general principles:

- One idea is to embed the set of estimators of the type $\widehat{\theta} : \Omega \to \Theta$ into the larger set of regular conditional probability measures $\rho : \big(\Omega, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\big) \to \mathcal{M}_+^1(\Theta, \mathcal{T})$. We will call these conditional probability measures *posterior distributions*, to follow standard terminology.
- A second idea is to measure the fluctuations of ρ with respect to the sample, using some prior distribution $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$, and the Kullback divergence function $\mathcal{K}(\rho, \pi)$. The expectation $\mathbb{P}\big\{\mathcal{K}(\rho, \pi)\big\}$ measures the randomness of ρ. The optimal choice of π would be $\mathbb{P}(\rho)$, resulting in a measure of the randomness of ρ equal to the mutual information between the sample and the estimated parameter drawn from ρ. Anyhow, since $\mathbb{P}(\rho)$ is usually not better known than $\mathbb{P}$, we will have to be content with some less concentrated prior distribution π, resulting in some looser measure of randomness, as shown by the identity $\mathbb{P}\big[\mathcal{K}(\rho, \pi)\big] = \mathbb{P}\big\{\mathcal{K}\big[\rho, \mathbb{P}(\rho)\big]\big\} + \mathcal{K}\big[\mathbb{P}(\rho), \pi\big]$.
- A third idea is to analyse the fluctuations of the random process $\theta \mapsto r(\theta)$ from its mean process $\theta \mapsto R(\theta)$ through the log-Laplace transform

$$-\frac{1}{\lambda} \log \left\{ \iint \exp\big[-\lambda r(\theta, \omega)\big] \pi(d\theta) \mathbb{P}(d\omega) \right\},$$

  as would be done in statistical mechanics, where this is called the free energy. This transform is well suited to relate $\min_{\theta \in \Theta} r(\theta)$ to $\inf_{\theta \in \Theta} R(\theta)$, since for large enough values of the parameter λ, corresponding to low enough values of the temperature, the system has small fluctuations around its ground state.
- A fourth idea deals with localization. It consists of considering a prior distribution $\overline{\pi}$ depending on the unknown expected error rate function R. Thus some central result of the theory will consist in an empirical upper bound for $\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big]$, where $\pi_{\exp(-\beta R)}$, defined by its density

$$\frac{d}{d\pi}\big[\pi_{\exp(-\beta R)}\big] = \frac{\exp(-\beta R)}{\pi\big[\exp(-\beta R)\big]},$$

  is a Gibbs distribution built from a known prior distribution $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$, some inverse temperature parameter $\beta \in \mathbb{R}_+$ and the expected error rate R. This bound will in particular be used when ρ is a posterior Gibbs distribution, of the form $\pi_{\exp(-\beta r)}$. The general idea will be to show that in the case when ρ is not too random, in the sense that it is possible to find a prior (that is non-random) distribution $\overline{\pi}$ such that $\mathcal{K}(\rho, \overline{\pi})$ is small, then $\rho(r)$ can be reliably taken for a good approximation of $\rho(R)$.

This monograph is divided into four chapters. The first deals with the inductive setting presented in these lines. The second is devoted to relative bounds. It shows that it is possible to obtain a tighter estimate of the mutual information between the sample and the estimated parameter by comparing

prior and posterior Gibbs distributions. It shows how to use this idea to obtain adaptive model selection schemes under very weak hypotheses.

The third chapter introduces the *transductive* setting of V. Vapnik (Vapnik, 1998), which consists in comparing the performance of classification rules on the learning sample with their performance on a test sample instead of their average performance. The fourth one is a fast introduction to Support Vector Machines. It is the occasion to show the implications of the general results discussed in the three first chapters when some particular choice is made about the structure of the classification rules.

In the first chapter, two types of bounds are shown. *Empirical bounds* are useful to build, compare and select estimators. *Non random bounds* are useful to assess the speed of convergence of estimators, relating this speed to the behaviour of the Gibbs prior expected error rate $\beta \mapsto \pi_{\exp(-\beta R)}(R)$ and to covariance factors related to the margin assumption of Mammen and Tsybakov when a finer analysis is performed. We will proceed from the most straightforward bounds towards more elaborate ones, built to achieve a better asymptotic behaviour. In this course towards more sophisticated inequalities, we will introduce *local bounds* and *relative bounds*.

The study of relative bounds is expanded in the third chapter, where tighter comparisons between prior and posterior Gibbs distributions are proved. Theorems 2.1.3 (page 72) and 2.2.4 (page 93) present two ways of selecting some nearly optimal classification rule. They are both proved to be adaptive in all the parameters under Mammen and Tsybakov margin assumptions and parametric complexity assumptions. This is done in Corollary 2.1.17 (page 86) of Theorem 2.1.15 (page 85) and in Theorem 2.2.11 (page 110). In the first approach, the performance of a randomized estimator modelled by a posterior distribution is compared with the performance of a prior Gibbs distribution. In the second approach posterior distributions are directly compared between themselves (and leads to slightly stronger results, to the price of using a more complex algorithm). When there are more than one parametric model, it is appropriate to use also some *doubly localized scheme*: two step localization is presented for both approaches, in Theorems 2.3.2 (page 116) and 2.3.9 (page 131) and provides bounds with a decreased influence of the number of empirically inefficient models included in the selection scheme.

We would not like to induce the reader into thinking that the most sophisticated results presented in these first two chapters are necessarily the most useful ones, they are as a rule only more efficient *asymptotically*, whereas, being more involved, they use looser constants leading to less precision for small sample sizes. In practice whether a sample is to be considered small is a question of the ratio between the number of examples and the complexity (roughly speaking the number of parameters) of the model used for classification. Since our aim here is to describe methods appropriate for complex data (images, speech, DNA, ...), we suspect that practitioners wanting to make use of our proposals will often be confronted with small sample sizes; thus

we would advise them to try the simplest bounds first and only afterwards see whether the asymptotically better ones can bring some improvement.

We would also like to point out that the results of the first two chapters are not of a purely theoretical nature: posterior parameter distributions can indeed be computed effectively, using Monte Carlo techniques, and there is well-established know-how about these computations in Bayesian statistics. Moreover, non-randomized estimators of the classical form $\widehat{\theta} : \Omega \to \Theta$ can be efficiently approximated by posterior distributions $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ supported by a fairly narrow neighbourhood of $\widehat{\theta}$, more precisely a neighbourhood of the size of the typical fluctuations of $\widehat{\theta}$, so that this randomized approximation of $\widehat{\theta}$ will most of the time provide the same classification as $\widehat{\theta}$ itself, except for a small amount of dubious examples for which the classification provided by $\widehat{\theta}$ would anyway be unreliable. This is explained on page 20.

As already mentioned, the third chapter is about the *transductive setting*, that is about comparing the performance of estimators on a training set and on a test set. We show first that this comparison can be based on a set of exponential deviation inequalities which parallels the one used in the inductive case. This gives the opportunity to *transport* all the results obtained in the inductive case in a systematic way. In the transductive setting, the use of prior distributions can be extended to the use of *partially exchangeable posterior distributions* depending on the union of training and test patterns, bringing increased possibilities to adapt to the data and giving rise to such crucial notions of complexity as the Vapnik–Cervonenkis dimension.

Having done so, we more specifically focus on the *small sample case*, where local and relative bounds are not expected to be of great help. Introducing a fictitious (that is unobserved) shadow sample, we study Vapnik-type generalization bounds, showing how to tighten and extend them with some original ideas, like making no Gaussian approximation to the log-Laplace transform of Bernoulli random variables, using a shadow sample of arbitrary size. shrinking from the use of any symmetrization trick, and using a suitable subset of the group of permutations to cover the case of independent non-identically distributed data. The culminating result of the third chapter is Theorem 3.3.3 (page 153), subsequent bounds showing the separate influence of the above ideas and providing an easier comparison with Vapnik's original results. Vapnik-type generalization bounds have a broad applicability, not only through the concept of Vapnik–Cervonenkis dimension, but also through the use of compression schemes (Little et al., 1986), which are briefly described on page 144.

The beginning of the fourth chapter introduces Support Vector Machines, both in the separable and in the non-separable case (using the box constraint). We then describe different types of bounds. We start with compression scheme bounds, to proceed with margin bounds. We begin with transductive margin bounds, recalling on this occasion in Theorem 4.2.2

(page 174) the growth bound for a family of classification rules with given Vapnik–Cervonenkis dimension. In Theorem 4.2.4 (page 176) we give the usual estimate of the Vapnik–Cervonenkis dimension of a family of separating hyperplanes with a given transductive margin (we mean by this that the margin is computed on the union of the training and test sets). We present an original probabilistic proof inspired by a similar one from Cristianini et al. (2000), whereas other proofs available usually rely on the informal claim that the simplex is the worst case. We end this short review of Support Vector Machines with a discussion of inductive margin bounds. Here the margin is computed on the training set only, and a more involved combinatorial lemma, due to Alon et al. (1997) and recalled in Lemma 4.2.6 (page 180) is used. We use this lemma and the results of the third chapter to establish a bound depending on the margin of the training set alone.

In appendix, we finally discuss the textbook example of classification by thresholding: in this setting, each classification rule is built by thresholding a series of measurements and taking a decision based on these thresholded values. This relatively simple example (which can be considered as an introduction to the more technical case of classification trees) can be used to give more flesh to the results of the first three chapters.

It is a pleasure to end this introduction with my greatest thanks to Anthony Davison, for his careful reading of the manuscript and his numerous suggestions.

Chapter 1

Inductive PAC-Bayesian learning

The setting of inductive inference (as opposed to transductive inference to be discussed later) is the one described in the introduction.

When we will have to take the expectation of a random variable $Z : \Omega \to \mathbb{R}$ as well as of a function of the parameter $h : \Theta \to \mathbb{R}$ with respect to some probability measure, we will as a rule use short functional notation instead of resorting to the integral sign: thus we will write $\mathbb{P}(Z)$ for $\int_\Omega Z(\omega)\mathbb{P}(d\omega)$ and $\pi(h)$ for $\int_\Theta h(\theta)\pi(d\theta)$.

A more traditional statistical approach would focus on estimators $\widehat{\theta} : \Omega \to \Theta$ of the parameter θ and be interested on the relationship between the *empirical error rate* $r(\widehat{\theta})$, defined by equation (0.1, page 7), which is the number of errors made on the sample, and the expected error rate $R(\widehat{\theta})$, defined by equation (0.2, page 7), which is the expected probability of error on new instances of patterns. The PAC-Bayesian approach instead chooses a broader perspective and allows the estimator $\widehat{\theta}$ to be drawn at random using some auxiliary source of randomness to smooth the dependence of $\widehat{\theta}$ on the sample. One way of representing the supplementary randomness allowed in the choice of $\widehat{\theta}$, is to consider what it is usual to call *posterior distributions* on the parameter space, that is probability measures $\rho : \Omega \to \mathcal{M}_+^1(\Theta, \mathcal{T})$, depending on the sample, or from a technical perspective, regular conditional (or transition) probability measures. Let us recall that we use the model described in the introduction: the training sample is modelled by the canonical process $(X_i, Y_i)_{i=1}^N$ on $\Omega = \left(\mathcal{X} \times \mathcal{Y}\right)^N$, and a product probability measure $\mathbb{P} = \bigotimes_{i=1}^N P_i$ on Ω is considered to reflect the assumption that the training sample is made of independent pairs of patterns and labels. The transition probability measure ρ, along with $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, defines a probability distribution on $\Omega \times \Theta$ and describes the conditional distribution of the estimated parameter $\widehat{\theta}$ knowing the sample $(X_i, Y_i)_{i=1}^N$.

The main subject of this broadened theory becomes to investigate the relationship between $\rho(r)$, the average error rate of $\widehat{\theta}$ on the training sample,

and $\rho(R)$, the expected error rate of $\widehat{\theta}$ on new samples. The first step towards using some kind of thermodynamics to tackle this question, is to consider the Laplace transform of $\rho(R) - \rho(r)$, a well known provider of non-asymptotic deviation bounds. This transform takes the form

$$\mathbb{P}\Big\{\exp\Big[\lambda\big[\rho(R) - \rho(r)\big]\Big]\Big\},$$

where some inverse temperature parameter $\lambda \in \mathbb{R}_+$, as a physicist would call it, is introduced. This Laplace transform would be easy to bound if ρ did not depend on $\omega \in \Omega$ (namely on the sample), because $\rho(R)$ would then be non-random, and

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

would be a sum of independent random variables. It turns out, and this will be the subject of the next section, that this annoying dependence of ρ on ω can be quantified, using the inequality

$$\rho(R) - \rho(r) \leq \lambda^{-1} \log\Big\{\pi\Big[\exp\big[\lambda(R - r)\big]\Big]\Big\} + \lambda^{-1}\mathcal{K}(\rho, \pi),$$

which holds for any probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ on the parameter space; for our purpose it will be appropriate to consider a *prior* distribution π that is non-random, as opposed to ρ, which depends on the sample. Here, $\mathcal{K}(\rho, \pi)$ is the Kullback divergence of ρ from π, whose definition will be recalled when we will come to technicalities; it can be seen as an upper bound for the mutual information between the $(X_i, Y_i)_{i=1}^N$ and the estimated parameter $\widehat{\theta}$. This inequality will allow us to relate the *penalized* difference $\rho(R) - \rho(r) - \lambda^{-1}\mathcal{K}(\rho, \pi)$ with the Laplace transform of sums of independent random variables.

1.1. Basic inequality

Let us now come to the details of the investigation sketched above. The first thing we will do is to study the Laplace transform of $R(\theta) - r(\theta)$, as a starting point for the more general study of $\rho(R) - \rho(r)$: it corresponds to the simple case where $\widehat{\theta}$ is not random at all, and therefore where ρ is a Dirac mass at some deterministic parameter value θ.

In the setting described in the introduction, let us consider the Bernoulli random variables $\sigma_i(\theta) = \mathbb{1}\big[Y_i \neq f_\theta(X_i)\big]$, which indicates whether the classification rule f_θ made an error on the ith component of the training sample. Using independence and the concavity of the logarithm function, it is readily seen that for any real constant λ

$$\log\Big\{\mathbb{P}\big\{\exp\big[-\lambda r(\theta)\big]\big\}\Big\} = \sum_{i=1}^{N} \log\Big\{\mathbb{P}\Big[\exp\big(-\tfrac{\lambda}{N}\sigma_i\big)\Big]\Big\}$$
$$\leq N \log\Big\{\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}\Big[\exp\big(-\tfrac{\lambda}{N}\sigma_i\big)\Big]\Big\}.$$

The right-hand side of this inequality is the log-Laplace transform of a Bernoulli distribution with parameter $\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}(\sigma_i) = R(\theta)$. As any Bernoulli distribution is fully defined by its parameter, this log-Laplace transform is necessarily a function of $R(\theta)$. It can be expressed with the help of the family of functions

$$\Phi_a(p) = -a^{-1}\log\big\{1 - \big[1-\exp(-a)\big]p\big\}, \quad a \in \mathbb{R}, p \in (0,1). \tag{1.1}$$

It is immediately seen that Φ_a is an increasing one-to-one mapping of the unit interval onto itself, and that it is convex when $a > 0$, concave when $a < 0$ and can be defined by continuity to be the identity when $a = 0$. Moreover the inverse of Φ_a is given by the formula

$$\Phi_a^{-1}(q) = \frac{1-\exp(-aq)}{1-\exp(-a)}, \qquad a \in \mathbb{R}, q \in (0,1).$$

This formula may be used to extend Φ_a^{-1} to $q \in \mathbb{R}$, and we will use this extension without further notice when required.

Using this notation, the previous inequality becomes

$$\log\Big\{\mathbb{P}\big\{\exp\big[-\lambda r(\theta)\big]\big\}\Big\} \leq -\lambda\Phi_{\frac{\lambda}{N}}\big[R(\theta)\big], \quad \text{proving}$$

Lemma 1.1.1. *For any real constant λ and any parameter $\theta \in \Theta$,*

$$\mathbb{P}\Big\{\exp\Big\{\lambda\Big[\Phi_{\frac{\lambda}{N}}\big[R(\theta)\big] - r(\theta)\Big]\Big\}\Big\} \leq 1.$$

In previous versions of this study, we had used some Bernstein bound, instead of this lemma. Anyhow, as it will turn out, keeping the log-Laplace transform of a Bernoulli instead of approximating it provides simpler and tighter results.

Lemma 1.1.1 implies that for any constants $\lambda \in \mathbb{R}_+$ and $\epsilon \in]0,1)$,

$$\mathbb{P}\left[\Phi_{\frac{\lambda}{N}}\big[R(\theta)\big] + \frac{\log(\epsilon)}{\lambda} \leq r(\theta)\right] \geq 1-\epsilon.$$

Choosing $\overline{\lambda} \in \arg\max_{\mathbb{R}_+} \Phi_{\frac{\lambda}{N}}\big[R(\theta)\big] + \frac{\log(\epsilon)}{\lambda}$, we deduce

Lemma 1.1.2. *For any $\epsilon \in]0,1)$, any $\theta \in \Theta$,*

$$\mathbb{P}\left\{R(\theta) \leq \inf_{\lambda\in\mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1}\left[r(\theta) - \frac{\log(\epsilon)}{\lambda}\right]\right\} \geq 1-\epsilon.$$

We will illustrate throughout these notes the bounds we prove with a small numerical example: in the case where $N = 1000$, $\epsilon = 0.01$ and $r(\theta) = 0.2$, we get with a confidence level of 0.99 that $R(\theta) \leq .2402$, this being obtained for $\lambda = 234$.

Now, to proceed towards the analysis of posterior distributions, let us put $U_\lambda(\theta,\omega) = \lambda\Big[\Phi_{\frac{\lambda}{N}}\big[R(\theta)\big] - r(\theta,\omega)\Big]$ for short, and let us consider some prior probability distribution $\pi \in \mathcal{M}_+^1(\Theta,\mathcal{T})$. A proper choice of π will be an important question, underlying much of the material presented in this monograph, so for the time being, let us only say that we will let this choice be as open as possible by writing inequalities which hold for *any* choice of π . Let us insist on the fact that when we say that π is a prior distribution, we mean that it *does not* depend on the training sample $(X_i,Y_i)_{i=1}^N$. The quantity of interest to obtain the bound we are looking for is $\log\Big\{\mathbb{P}\Big[\pi\big[\exp(U_\lambda)\big]\Big]\Big\}$. Using Fubini's theorem for non-negative functions, we see that

$$\log\Big\{\mathbb{P}\Big[\pi\big[\exp(U_\lambda)\big]\Big]\Big\} = \log\Big\{\pi\Big[\mathbb{P}\big[\exp(U_\lambda)\big]\Big]\Big\} \leq 0.$$

To relate this quantity to the expectation $\rho(U_\lambda)$ with respect to any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, we will use the properties of the Kullback divergence $\mathcal{K}(\rho,\pi)$ of ρ with respect to π, which is defined as

$$\mathcal{K}(\rho,\pi) = \begin{cases} \int \log(\frac{d\rho}{d\pi})d\rho, & \text{when } \rho \text{ is absolutely continuous} \\ & \text{with respect to } \pi, \\ +\infty, & \text{otherwise.} \end{cases}$$

The following lemma shows in which sense the Kullback divergence function can be thought of as the dual of the log-Laplace transform.

Lemma 1.1.3. *For any bounded measurable function $h : \Theta \to \mathbb{R}$, and any probability distribution $\rho \in \mathcal{M}_+^1(\Theta)$ such that $\mathcal{K}(\rho,\pi) < \infty$,*

$$\log\big\{\pi\big[\exp(h)\big]\big\} = \rho(h) - \mathcal{K}(\rho,\pi) + \mathcal{K}(\rho,\pi_{\exp(h)}),$$

where by definition $\dfrac{d\pi_{\exp(h)}}{d\pi} = \dfrac{\exp[h(\theta)]}{\pi[\exp(h)]}$. *Consequently*

$$\log\big\{\pi\big[\exp(h)\big]\big\} = \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho(h) - \mathcal{K}(\rho,\pi).$$

The proof is just a matter of writing down the definition of the quantities involved and using the fact that the Kullback divergence function is non-negative, and can be found in Catoni (2004, page 160). In the duality between measurable functions and probability measures, we thus see that the log-Laplace transform with respect to π is the Legendre transform of the Kullback divergence function with respect to π. Using this, we get

$$\mathbb{P}\Big\{\exp\big\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho[U_\lambda(\theta)] - \mathcal{K}(\rho,\pi)\big\}\Big\} \leq 1,$$

which, combined with the convexity of $\lambda\Phi_{\frac{\lambda}{N}}$, proves the basic inequality we were looking for.

THEOREM 1.1.4. *For any real constant λ,*

$$\mathbb{P}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\lambda\Big[\Phi_{\frac{\lambda}{N}}\big[\rho(R)\big]-\rho(r)\Big]-\mathcal{K}(\rho,\pi)\right]\right\}$$
$$\leq\mathbb{P}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\lambda\Big[\rho\big(\Phi_{\frac{\lambda}{N}}\circ R\big)-\rho(r)\Big]-\mathcal{K}(\rho,\pi)\right]\right\}\leq 1.$$

We insist on the fact that in this theorem, we take a supremum in $\rho\in\mathcal{M}_+^1(\Theta)$ *inside* the expectation with respect to $\mathbb{P}$, the sample distribution. This means that the proved inequality holds for any ρ depending on the training sample, that is for any posterior distribution: indeed, measurability questions set aside,

$$\mathbb{P}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\lambda\Big[\rho\big[U_\lambda(\theta)\big]-\mathcal{K}(\rho,\pi)\Big]\right]\right\}$$
$$=\sup_{\rho:\Omega\to\mathcal{M}_+^1(\Theta)}\mathbb{P}\left\{\exp\left[\lambda\Big[\rho\big[U_\lambda(\theta)\big]-\mathcal{K}(\rho,\pi)\Big]\right]\right\},$$

and more formally,

$$\sup_{\rho:\Omega\to\mathcal{M}_+^1(\Theta)}\mathbb{P}\left\{\exp\left[\lambda\Big[\rho\big[U_\lambda(\theta)\big]-\mathcal{K}(\rho,\pi)\Big]\right]\right\}$$
$$\leq\mathbb{P}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\lambda\Big[\rho\big[U_\lambda(\theta)\big]-\mathcal{K}(\rho,\pi)\Big]\right]\right\},$$

where the supremum in ρ taken in the left-hand side is restricted to *regular* conditional probability distributions.

The following sections will show how to use this theorem.

1.2. Non local bounds

At least three sorts of bounds can be deduced from Theorem 1.1.4.

The most interesting ones with which to build estimators and tune parameters, as well as the first that have been considered in the development of the PAC-Bayesian approach, are deviation bounds. They provide an empirical upper bound for $\rho(R)$ — that is a bound which can be computed from observed data — with some probability $1-\epsilon$, where ϵ is a presumably small and tunable parameter setting the desired confidence level.

Anyhow, most of the results about the convergence speed of estimators to be found in the statistical literature are concerned with the expectation

$\mathbb{P}\big[\rho(R)\big]$, therefore it is also enlightening to bound this quantity. In order to know at which rate it may be approaching $\inf_\Theta R$, a non-random upper bound is required, which will relate the average of the expected risk $\mathbb{P}\big[\rho(R)\big]$ with the properties of the contrast function $\theta \mapsto R(\theta)$.

Since the values of constants do matter a lot when a bound is to be used to select between various estimators using classification models of various complexities, a third kind of bound, related to the first, may be considered for the sake of its hopefully better constants: we will call them *unbiased empirical bounds*, to stress the fact that they provide some empirical quantity whose expectation under $\mathbb{P}$ can be proved to be an upper bound for $\mathbb{P}\big[\rho(R)\big]$, the average expected risk. The price to pay for these better constants is of course the lack of formal guarantee given by the bound: two random variables whose expectations are ordered in a certain way may very well be ordered in the reverse way with a large probability, so that basing the estimation of parameters or the selection of an estimator on some unbiased empirical bound is a hazardous business. Anyhow, since it is common practice to use the inequalities provided by mathematical statistical theory while replacing the proven constants with smaller values showing a better practical efficiency, considering unbiased empirical bounds as well as deviation bounds provides an indication about how much the constants may be decreased while not violating the theory too much.

1.2.1. Unbiased empirical bounds. Let $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ be some fixed (and arbitrary) posterior distribution, describing some randomized estimator $\widehat{\theta} : \Omega \to \Theta$. As we already mentioned, in these notes a posterior distribution will always be a regular conditional probability measure. By this we mean that

- for any $A \in \mathcal{T}$, the map $\omega \mapsto \rho(\omega, A) : \big(\Omega, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\big) \to \mathbb{R}_+$ is assumed to be measurable;
- for any $\omega \in \Omega$, the map $A \mapsto \rho(\omega, A) : \mathcal{T} \to \mathbb{R}_+$ is assumed to be a probability measure.

We will also assume without further notice that the σ-algebras we deal with are always countably generated. The technical implications of these assumptions are standard and discussed for instance in Catoni (2004, pages 50-54), where, among other things, a detailed proof of the decomposition of the Kullback Liebler divergence is given.

Let us restrict to the case when the constant λ is positive. We get from Theorem 1.1.4 that

$$\exp\left[\lambda\Big\{\Phi_{\frac{\lambda}{N}}\Big[\mathbb{P}\big[\rho(R)\big]\Big] - \mathbb{P}\big[\rho(r)\big]\Big\} - \mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]\right] \leq 1, \tag{1.2}$$

where we have used the convexity of the exp function and of $\Phi_{\frac{\lambda}{N}}$. Since we have restricted our attention to positive values of the constant λ, equation

(1.2) can also be written

$$\mathbb{P}\big[\rho(R)\big] \leq \Phi^{-1}_{\frac{\lambda}{N}}\Big\{\mathbb{P}\big[\rho(r) + \lambda^{-1}\mathcal{K}(\rho,\pi)\big]\Big\},$$

leading to

THEOREM 1.2.1. *For any posterior distribution $\rho : \Omega \to \mathcal{M}^1_+(\Theta)$, for any positive parameter λ,*

$$\begin{aligned}\mathbb{P}\big[\rho(R)\big] &\leq \frac{1 - \exp\Big[-N^{-1}\mathbb{P}\big[\lambda\rho(r) + \mathcal{K}(\rho,\pi)\big]\Big]}{1 - \exp(-\frac{\lambda}{N})} \\ &\leq \mathbb{P}\left\{\frac{\lambda}{N\big[1-\exp(-\frac{\lambda}{N})\big]}\left[\rho(r) + \frac{\mathcal{K}(\rho,\pi)}{\lambda}\right]\right\}.\end{aligned}$$

The last inequality provides the *unbiased empirical upper bound* for $\rho(R)$ we were looking for, meaning that the expectation of $\frac{\lambda}{N\left[1-\exp(-\frac{\lambda}{N})\right]}\left[\rho(r) + \frac{\mathcal{K}(\rho,\pi)}{\lambda}\right]$ is larger than the expectation of $\rho(R)$. Let us notice that $1 \leq \frac{\lambda}{N\left[1-\exp(-\frac{\lambda}{N})\right]} \leq \big[1 - \frac{\lambda}{2N}\big]^{-1}$ and therefore that this coefficient is close to 1 when λ is significantly smaller than N.

If we are ready to believe in this bound (although this belief is not mathematically well founded, as we already mentioned), we can use it to optimize λ and to choose ρ. While the optimal choice of ρ when λ is fixed is, according to Lemma 1.1.3 (page 16), to take it equal to $\pi_{\exp(-\lambda r)}$, a Gibbs posterior distribution, as it is sometimes called, we may for computational reasons be more interested in choosing ρ in some other class of posterior distributions.

For instance, our real interest may be to select some non-randomized estimator from a family $\widehat{\theta}_m : \Omega \to \Theta_m$, $m \in M$, of possible ones, where Θ_m are measurable subsets of Θ and where M is an arbitrary (non necessarily countable) index set. We may for instance think of the case when $\widehat{\theta}_m \in \arg\min_{\Theta_m} r$. We may slightly randomize the estimators to start with, considering for any $\theta \in \Theta_m$ and any $m \in M$,

$$\Delta_m(\theta) = \Big\{\theta' \in \Theta_m : \big[f_{\theta'}(X_i)\big]_{i=1}^N = \big[f_\theta(X_i)\big]_{i=1}^N\Big\},$$

and defining ρ_m by the formula

$$\frac{d\rho_m}{d\pi}(\theta) = \frac{\mathbb{1}\big[\theta \in \Delta_m(\widehat{\theta}_m)\big]}{\pi\big[\Delta_m(\widehat{\theta}_m)\big]}.$$

Our posterior minimizes $\mathcal{K}(\rho,\pi)$ among those distributions whose support is restricted to the values of θ in Θ_m for which the classification rule f_θ is identical to the estimated one $f_{\widehat{\theta}_m}$ on the observed sample. Presumably, in many practical situations, $f_\theta(x)$ will be ρ_m almost surely identical to $f_{\widehat{\theta}_m}(x)$

when θ is drawn from ρ_m, for the vast majority of the values of $x \in \mathcal{X}$ and all the sub-models Θ_m not plagued with too much overfitting (since this is by construction the case when $x \in \{X_i : i = 1, \ldots, N\}$). Therefore replacing $\widehat{\theta}_m$ with ρ_m can be expected to be a minor change in many situations. This change by the way can be estimated in the (admittedly not so common) case when the distribution of the patterns $(X_i)_{i=1}^N$ is known. Indeed, introducing the pseudo distance

$$D(\theta, \theta') = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\big[f_\theta(X_i) \neq f_{\theta'}(X_i)\big], \qquad \theta, \theta' \in \Theta, \tag{1.3}$$

one immediately sees that $R(\theta') \leq R(\theta) + D(\theta, \theta')$, for any $\theta, \theta' \in \Theta$, and therefore that

$$R(\widehat{\theta}_m) \leq \rho_m(R) + \rho_m\big[D(\cdot, \widehat{\theta}_m)\big].$$

Let us notice also that in the case where $\Theta_m \subset \mathbb{R}^{d_m}$, and R happens to be convex on $\Delta_m(\widehat{\theta}_m)$, then $\rho_m(R) \geq R\big[\int \theta \rho_m(d\theta)\big]$, and we can replace $\widehat{\theta}_m$ with $\widetilde{\theta}_m = \int \theta \rho_m(d\theta)$, and obtain bounds for $R(\widetilde{\theta}_m)$. This is not a very heavy assumption about R, in the case where we consider $\widehat{\theta}_m \in \arg\min_{\Theta_m} r$. Indeed, $\widehat{\theta}_m$, and therefore $\Delta_m(\widehat{\theta}_m)$, will presumably be close to $\arg\min_{\Theta_m} R$, and requiring a function to be convex in the neighbourhood of its minima is not a very strong assumption.

Since $r(\widehat{\theta}_m) = \rho_m(r)$, and $\mathcal{K}(\rho_m, \pi) = -\log\big\{\pi\big[\Delta_m(\widehat{\theta}_m)\big]\big\}$, our unbiased empirical upper bound in this context reads as

$$\frac{\lambda}{N\big[1 - \exp(-\frac{\lambda}{N})\big]} \left\{ r(\widehat{\theta}_m) - \frac{\log\big\{\pi\big[\Delta_m(\widehat{\theta}_m)\big]\big\}}{\lambda} \right\}.$$

Let us notice that we obtain a complexity factor $-\log\big\{\pi\big[\Delta_m(\widehat{\theta}_m)\big]\big\}$ which may be compared with the Vapnik–Cervonenkis dimension. Indeed, in the case of binary classification, when using a classification model with Vapnik–Cervonenkis dimension not greater than h_m, that is when any subset of $\mathcal{X}$ which can be split in any arbitrary way by some classification rule f_θ of the model Θ_m has at most h_m points, then

$$\big\{\Delta_m(\theta) : \theta \in \Theta_m\big\}$$

is a partition of Θ_m with at most $\left(\frac{eN}{h_m}\right)^{h_m}$ components: these facts, if not already familiar to the reader, will be proved in Theorems 4.2.2 and 4.2.3 (page 174). Therefore

$$\inf_{\theta \in \Theta_m} -\log\big\{\pi\big[\Delta_m(\theta)\big]\big\} \leq h_m \log\left(\frac{eN}{h_m}\right) - \log\big[\pi(\Theta_m)\big].$$

Thus, if the model and prior distribution are well suited to the classification task, in the sense that there is more "room" (where room is measured with

π) between the two clusters defined by $\widehat{\theta}_m$ than between other partitions of the sample of patterns $(X_i)_{i=1}^N$, then we will have

$$-\log\{\pi[\Delta_m(\widehat{\theta})]\} \leq h_m \log\left(\frac{eN}{h_m}\right) - \log[\pi(\Theta_m)].$$

An optimal value $\widehat{m}$ may be selected so that

$$\widehat{m} \in \arg\min_{m\in M}\left\{\inf_{\lambda\in\mathbb{R}_+} \frac{\lambda}{N[1-\exp(-\frac{\lambda}{N})]}\left(r(\widehat{\theta}_m) - \frac{\log\{\pi[\Delta_m(\widehat{\theta}_m)]\}}{\lambda}\right)\right\}.$$

Since $\rho_{\widehat{m}}$ is still another posterior distribution, we can be sure that

$$\begin{aligned}\mathbb{P}\Big\{R(\widehat{\theta}_{\widehat{m}}) - \rho_{\widehat{m}}[D(\cdot,\widehat{\theta}_{\widehat{m}})]\Big\} &\leq \mathbb{P}[\rho_{\widehat{m}}(R)]\\ &\leq \inf_{\lambda\in\mathbb{R}_+}\mathbb{P}\left\{\frac{\lambda}{N[1-\exp(-\frac{\lambda}{N})]}\left(r(\widehat{\theta}_{\widehat{m}}) - \frac{\log\{\pi[\Delta_{\widehat{m}}(\widehat{\theta}_{\widehat{m}})]\}}{\lambda}\right)\right\}.\end{aligned}$$

Taking the infimum in λ inside the expectation with respect to $\mathbb{P}$ would be possible at the price of some supplementary technicalities and a slight increase of the bound that we prefer to postpone to the discussion of deviation bounds, since they are the only ones to provide a rigorous mathematical foundation to the adaptive selection of estimators.

1.2.2. Optimizing explicitly the exponential parameter λ. In this section we address some technical issues we think helpful to the understanding of Theorem 1.2.1 (page 19): namely to investigate how the upper bound it provides could be optimized, or at least approximately optimized, in λ. It turns out that this can be done quite explicitly.

So we will consider in this discussion the posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ to be fixed, and our aim will be to eliminate the constant λ from the bound by choosing its value in some nearly optimal way as a function of $\mathbb{P}[\rho(r)]$, the average of the empirical risk, and of $\mathbb{P}[\mathcal{K}(\rho,\pi)]$, which controls overfitting.

Let the bound be written as

$$\varphi(\lambda) = [1-\exp(-\tfrac{\lambda}{N})]^{-1}\left\{1-\exp\left[-\tfrac{\lambda}{N}\mathbb{P}[\rho(r)] - N^{-1}\mathbb{P}[\mathcal{K}(\rho,\pi)]\right]\right\}.$$

We see that

$$N\frac{\partial}{\partial\lambda}\log[\varphi(\lambda)] = \frac{\mathbb{P}[\rho(r)]}{\exp\left[\frac{\lambda}{N}\mathbb{P}[\rho(r)] + N^{-1}\mathbb{P}[\mathcal{K}(\rho,\pi)]\right] - 1} - \frac{1}{\exp(\frac{\lambda}{N})-1}.$$

Thus, the optimal value for λ is such that

$$[\exp(\tfrac{\lambda}{N})-1]\mathbb{P}[\rho(r)] = \exp\left[\tfrac{\lambda}{N}\mathbb{P}[\rho(r)] + N^{-1}\mathbb{P}[\mathcal{K}(\rho,\pi)]\right] - 1.$$

Assuming that $1 \gg \frac{\lambda}{N}\mathbb{P}\big[\rho(r)\big] \gg \frac{\mathbb{P}[\mathcal{K}(\rho,\pi)]}{N}$, and keeping only higher order terms, we are led to choose

$$\lambda = \sqrt{\frac{2N\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]}{\mathbb{P}\big[\rho(r)\big]\big\{1-\mathbb{P}\big[\rho(r)\big]\big\}}},$$

obtaining

THEOREM 1.2.2. *For any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\big[\rho(R)\big] \leq \frac{1-\exp\left\{-\sqrt{\frac{2\mathbb{P}[\mathcal{K}(\rho,\pi)]\mathbb{P}[\rho(r)]}{N\{1-\mathbb{P}[\rho(r)]\}}} - \frac{\mathbb{P}[\mathcal{K}(\rho,\pi)]}{N}\right\}}{1-\exp\left\{-\sqrt{\frac{2\mathbb{P}[\mathcal{K}(\rho,\pi)]}{N\mathbb{P}[\rho(r)]\{1-\mathbb{P}[\rho(r)]\}}}\right\}}.$$

This result of course is not very useful in itself, since neither of the two quantities $\mathbb{P}\big[\rho(r)\big]$ and $\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]$ are easy to evaluate. Anyhow it gives a hint that replacing them boldly with $\rho(r)$ and $\mathcal{K}(\rho,\pi)$ could produce something close to a legitimate empirical upper bound for $\rho(R)$. We will see in the subsection about deviation bounds that this is indeed essentially true.

Let us remark that in the third chapter of this monograph, we will see another way of bounding

$$\inf_{\lambda\in\mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1}\left(q+\frac{d}{\lambda}\right), \text{ leading to}$$

THEOREM 1.2.3. *For any prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\big[\rho(R)\big] \leq \left(1+\frac{2\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]}{N}\right)^{-1}\left\{\mathbb{P}\big[\rho(r)\big] + \frac{\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]}{N} + \sqrt{\frac{2\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]\mathbb{P}\big[\rho(r)\big]\big\{1-\mathbb{P}\big[\rho(r)\big]\big\}}{N} + \frac{\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]^2}{N^2}}\right\},$$

as soon as $\mathbb{P}\big[\rho(r)\big] + \sqrt{\frac{\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]}{2N}} \leq \frac{1}{2}$,

and $\mathbb{P}\big[\rho(R)\big] \leq \mathbb{P}\big[\rho(r)\big] + \sqrt{\frac{\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]}{2N}}$ *otherwise.*

This theorem enlightens the influence of three terms on the average expected risk:

• the average empirical risk, $\mathbb{P}\big[\rho(r)\big]$, which as a rule will decrease as the size of the classification model increases, acts as a *bias* term, grasping the ability of the model to account for the observed sample itself;

• a *variance* term $\frac{1}{N}\mathbb{P}\big[\rho(r)\big]\big\{1-\mathbb{P}\big[\rho(r)\big]\big\}$ is due to the random fluctuations of $\rho(r)$;

• a *complexity* term $\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]$, which as a rule will increase with the size of the classification model, eventually acts as a multiplier of the variance term.

We observed numerically that the bound provided by Theorem 1.2.2 is better than the more classical Vapnik-like bound of Theorem 1.2.3. For instance, when $N = 1000$, $\mathbb{P}\big[\rho(r)\big] = 0.2$ and $\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big] = 10$, Theorem 1.2.2 gives a bound lower than 0.2604, whereas the more classical Vapnik-like approximation of Theorem 1.2.3 gives a bound larger than 0.2622. Numerical simulations tend to suggest the two bounds are always ordered in the same way, although this could be a little tedious to prove mathematically.

1.2.3. Non random bounds. It is time now to come to less tentative results and see how far is the average expected error rate $\mathbb{P}\big[\rho(R)\big]$ from its best possible value $\inf_\Theta R$.

Let us notice first that

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

Let us remark moreover that $r \mapsto \log\Big[\pi\big[\exp(-\lambda r)\big]\Big]$ is a convex functional, a property which from a technical point of view can be dealt with in the following way:

$$\begin{aligned}
\mathbb{P}\Big\{\log\Big[\pi\big[\exp(-\lambda r)\big]\Big]\Big\} &= \mathbb{P}\Big\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)} -\lambda\rho(r) - \mathcal{K}(\rho,\pi)\Big\} \\
&\geq \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \mathbb{P}\Big\{-\lambda\rho(r) - \mathcal{K}(\rho,\pi)\Big\} = \sup_{\rho\in\mathcal{M}_+^1(\Theta)} -\lambda\rho(R) - \mathcal{K}(\rho,\pi) \\
&= \log\Big\{\pi\big[\exp(-\lambda R)\big]\Big\} = -\textstyle\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta. \quad (1.4)
\end{aligned}$$

These remarks applied to Theorem 1.2.1 lead to

Theorem 1.2.4. *For any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, for any positive parameter λ,*

$$\begin{aligned}
\mathbb{P}\big[\rho(R)\big] &\leq \frac{1 - \exp\Big\{-\frac{1}{N}\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta - \frac{1}{N}\mathbb{P}\big[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\big]\Big\}}{1 - \exp(-\frac{\lambda}{N})} \\
&\leq \frac{1}{N\big[1 - \exp(-\frac{\lambda}{N})\big]}\Big\{\textstyle\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta + \mathbb{P}\big[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\big]\Big\}.
\end{aligned}$$

This theorem is particularly well suited to the case of the Gibbs posterior distribution $\rho = \pi_{\exp(-\lambda r)}$, where the entropy factor cancels and where $\mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big]$ is shown to get close to $\inf_\Theta R$ when N goes to $+\infty$, as soon as λ/N goes to 0 while λ goes to $+\infty$.

We can elaborate on Theorem 1.2.4 and define a notion of dimension of (Θ, R), with margin $\eta \geq 0$ putting

$$d_\eta(\Theta, R) = \sup_{\beta \in \mathbb{R}_+} \beta\big[\pi_{\exp(-\beta R)}(R) - \operatorname{ess\,inf}_\pi R - \eta\big] \\ \leq -\log\Big\{\pi\big[R \leq \operatorname{ess\,inf}_\pi R + \eta\big]\Big\}. \quad (1.5)$$

This last inequality can be established by the chain of inequalities:

$$\beta\pi_{\exp(-\beta R)}(R) \leq \textstyle\int_0^\beta \pi_{\exp(-\gamma R)}(R)d\gamma = -\log\Big\{\pi\big[\exp(-\beta R)\big]\Big\} \\ \leq \beta\Big(\operatorname{ess\,inf}_\pi R + \eta\Big) - \log\Big[\pi\big(R \leq \operatorname{ess\,inf}_\pi R + \eta\big)\Big],$$

where we have used successively the fact that $\lambda \mapsto \pi_{\exp(-\lambda R)}(R)$ is decreasing (because it is the derivative of the concave function $\lambda \mapsto -\log\big\{\pi\big[\exp(-\lambda R)\big]\big\}$) and the fact that the exponential function takes positive values.

In typical "parametric" situations $d_0(\Theta, R)$ will be finite, and in all circumstances $d_\eta(\Theta, R)$ will be finite for any $\eta > 0$ (this is a direct consequence of the definition of the essential infimum). Using this notion of dimension, we see that

$$\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta \leq \lambda\big(\operatorname{ess\,inf}_\pi R + \eta\big) \\ + \int_0^\lambda \left[\frac{d_\eta}{\beta} \wedge (1 - \operatorname{ess\,inf}_\pi R - \eta)\right] d\beta \\ = \lambda\big(\operatorname{ess\,inf}_\pi R + \eta\big) + d_\eta(\Theta, R)\log\left[\frac{e\lambda}{d_\eta(\Theta, R)}\big(1 - \operatorname{ess\,inf}_\pi R - \eta\big)\right].$$

This leads to

Corollary 1.2.5 *With the above notation, for any margin $\eta \in \mathbb{R}_+$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\big[\rho(R)\big] \leq \inf_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1}\left[\operatorname{ess\,inf}_\pi R + \eta + \frac{d_\eta}{\lambda}\log\left(\frac{e\lambda}{d_\eta}\right) + \frac{\mathbb{P}\big\{\mathcal{K}\big[\rho, \pi_{\exp(-\lambda r)}\big]\big\}}{\lambda}\right].$$

If one wants a posterior distribution with a small support, the theorem can also be applied to the case when ρ is obtained by truncating $\pi_{\exp(-\lambda r)}$ to some level set to reduce its support: let $\Theta_p = \{\theta \in \Theta : r(\theta) \leq p\}$, and let us define for any $q \in)0, 1)$ the level $p_q = \inf\{p : \pi_{\exp(-\lambda r)}(\Theta_p) \geq q\}$, let us then define ρ_q by its density

$$\frac{d\rho_q}{d\pi_{\exp(-\lambda r)}}(\theta) = \frac{\mathbb{1}(\theta \in \Theta_{p_q})}{\pi_{\exp(-\lambda r)}(\Theta_{p_q})},$$

then $\rho_0 = \pi_{\exp(-\lambda r)}$ and for any $q \in (0,1($,

$$\mathbb{P}\big[\rho_q(R)\big] \leq \frac{1-\exp\left\{-\frac{1}{N}\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta - \frac{\log(q)}{N}\right\}}{1-\exp(-\frac{\lambda}{N})}$$
$$\leq \frac{1}{N\big[1-\exp(-\frac{\lambda}{N})\big]}\Big\{\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta - \log(q)\Big\}.$$

1.2.4. Deviation bounds. They provide results holding under the distribution $\mathbb{P}$ of the sample with probability at least $1-\epsilon$, for any given confidence level, set by the choice of $\epsilon \in]0,1($. Using them is the only way to be quite (i.e. with probability $1-\epsilon$) sure to do the right thing, although this right thing may be over-pessimistic, since deviation upper bounds are larger than corresponding non-biased bounds.

Starting again from Theorem 1.1.4 (page 17), and using Markov's inequality $\mathbb{P}\big[\exp(h) \geq 1\big] \leq \mathbb{P}\big[\exp(h)\big]$, we obtain

Theorem 1.2.6. *For any positive parameter λ, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\rho(R) \leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{\rho(r) + \frac{\mathcal{K}(\rho,\pi) - \log(\epsilon)}{\lambda}\right\}$$
$$= \frac{1-\exp\left\{-\frac{\lambda\rho(r)}{N} - \frac{\mathcal{K}(\rho,\pi)-\log(\epsilon)}{N}\right\}}{1-\exp\big(-\frac{\lambda}{N}\big)}$$
$$\leq \frac{\lambda}{N\left[1-\exp\left(-\frac{\lambda}{N}\right)\right]}\left[\rho(r) + \frac{\mathcal{K}(\rho,\pi)-\log(\epsilon)}{\lambda}\right].$$

We see that for a fixed value of the parameter λ, the upper bound is optimized when the posterior is chosen to be the Gibbs distribution $\rho = \pi_{\exp(-\lambda r)}$.

In this theorem, we have bounded $\rho(R)$, the average expected risk of an estimator $\widehat{\theta}$ drawn from the posterior ρ. This is what we will do most of the time in this study. This is the error rate we will get if we classify a large number of test patterns, drawing a new $\widehat{\theta}$ for each one. However, we can also be interested in the error rate we get if we draw only one $\widehat{\theta}$ from ρ and use this single draw of $\widehat{\theta}$ to classify a large number of test patterns. This error rate is $R(\widehat{\theta})$. To state a result about its deviations, we can start back from Lemma 1.1.1 (page 15) and integrate it with respect to the prior distribution π to get for any real constant λ

$$\mathbb{P}\Big\{\pi\Big[\exp\Big\{\lambda\Big[\Phi_{\frac{\lambda}{N}}\big(R\big) - r\Big]\Big\}\Big]\Big\} \leq 1.$$

For any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, this can be rewritten as

$$\mathbb{P}\Big\{\rho\Big[\exp\Big\{\lambda\Big[\Phi_{\frac{\lambda}{N}}\big(R\big) - r\Big] - \log\big(\tfrac{d\rho}{d\pi}\big) + \log(\epsilon)\Big]\Big\}\Big]\Big\} \leq \epsilon,$$

proving

THEOREM 1.2.7 *For any positive real parameter λ, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}\rho$ probability at least $1-\epsilon$,*

$$R(\widehat{\theta}\,) \leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{ r(\widehat{\theta}\,) + \lambda^{-1}\log\left(\epsilon^{-1}\frac{d\rho}{d\pi}\right)\right\}$$
$$\leq \frac{\lambda}{N\big[1-\exp(-\frac{\lambda}{N})\big]}\left[r(\widehat{\theta}\,) + \lambda^{-1}\log\left(\epsilon^{-1}\frac{d\rho}{d\pi}\right)\right].$$

Let us remark that the bound provided here is the exact counterpart of the bound of Theorem 1.2.6, since $\log\big(\frac{d\rho}{d\pi}\big)$ appears as a *disintegrated* version of the divergence $\mathcal{K}(\rho,\pi)$. The parallel between the two theorems is particularly striking in the special case when $\rho = \pi_{\exp(-\lambda r)}$. Indeed Theorem 1.2.6 proves that with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\pi_{\exp(-\lambda r)}(R) \leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{ -\frac{\log\big\{\pi\big[\exp\big(-\lambda r\big)\big]\big\} + \log(\epsilon)}{\lambda}\right\},$$

whereas Theorem 1.2.7 proves that with $\mathbb{P}\pi_{\exp(-\lambda r)}$ probability at least $1-\epsilon$

$$R(\widehat{\theta}\,) \leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{ -\frac{\log\big\{\pi\big[\exp\big(-\lambda r\big)\big]\big\} + \log(\epsilon)}{\lambda}\right\},$$

showing that we get the same deviation bound for $\pi_{\exp(-\lambda r)}(R)$ under $\mathbb{P}$ and for $\widehat{\theta}$ under $\mathbb{P}\pi_{\exp(-\lambda r)}$.

We would like to show now how to optimize with respect to λ the bound given by Theorem 1.2.6 (the same discussion would apply to Theorem 1.2.7). Let us notice first that values of λ less than 1 are not interesting (because they provide a bound larger than one, at least as soon as $\epsilon \leq \exp(-1)$). Let us consider some real parameter $\alpha > 1$, and the set $\Lambda = \{\alpha^k; k \in \mathbb{N}\}$, on which we put the probability measure $\nu(\alpha^k) = [(k+1)(k+2)]^{-1}$. Applying Theorem 1.2.6 to $\lambda = \alpha^k$ at confidence level $1 - \frac{\epsilon}{(k+1)(k+2)}$, and using a union bound, we see that with probability at least $1-\epsilon$, for any posterior distribution ρ,

$$\rho(R) \leq \inf_{\lambda' \in \Lambda} \Phi^{-1}_{\frac{\lambda'}{N}}\left\{ \rho(r) + \frac{\mathcal{K}(\rho,\pi) - \log(\epsilon) + 2\log\Big[\frac{\log(\alpha^2\lambda')}{\log(\alpha)}\Big]}{\lambda'}\right\}.$$

Now we can remark that for any $\lambda \in (1, +\infty($, there is $\lambda' \in \Lambda$ such that $\alpha^{-1}\lambda \leq \lambda' \leq \lambda$. Moreover, for any $q \in (0,1)$, $\beta \mapsto \Phi_\beta^{-1}(q)$ is increasing on $\mathbb{R}_+$. Thus with probability at least $1-\epsilon$, for any posterior distribution ρ,

$$\rho(R) \leq \inf_{\lambda \in (1,\infty(} \Phi^{-1}_{\frac{\lambda}{N}}\left\{ \rho(r) + \frac{\alpha}{\lambda}\left[\mathcal{K}(\rho,\pi) - \log(\epsilon) + 2\log\Big(\frac{\log(\alpha^2\lambda)}{\log(\alpha)}\Big)\right]\right\}$$

$$= \inf_{\lambda\in(1,\infty(} \frac{1-\exp\left\{-\frac{\lambda}{N}\rho(r) - \frac{\alpha}{N}\left[\mathcal{K}(\rho,\pi) - \log(\epsilon) + 2\log\left(\frac{\log(\alpha^2\lambda)}{\log(\alpha)}\right)\right]\right\}}{1-\exp(-\frac{\lambda}{N})}.$$

Taking the approximately optimal value

$$\lambda = \sqrt{\frac{2N\alpha\left[\mathcal{K}(\rho,\pi) - \log(\epsilon)\right]}{\rho(r)[1-\rho(r)]}},$$

we obtain

THEOREM 1.2.8. *With probability* $1-\epsilon$*, for any posterior distribution* $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$*, putting* $d(\rho,\epsilon) = \mathcal{K}(\rho,\pi) - \log(\epsilon)$*,*

$$\rho(R) \leq \inf_{k\in\mathbb{N}} \frac{1-\exp\left\{-\frac{\alpha^k}{N}\rho(r) - \frac{1}{N}\Big[d(\rho,\epsilon) + \log\big[(k+1)(k+2)\big]\Big]\right\}}{1-\exp\left(-\frac{\alpha^k}{N}\right)}$$

$$\leq \frac{1-\exp\left\{-\sqrt{\frac{2\alpha\rho(r)d(\rho,\epsilon)}{N[1-\rho(r)]}} - \frac{\alpha}{N}\left[d(\rho,\epsilon) + 2\log\left(\frac{\log\left(\alpha^2\sqrt{\frac{2N\alpha d(\rho,\epsilon)}{\rho(r)[1-\rho(r)]}}\right)}{\log(\alpha)}\right)\right]\right\}}{1-\exp\left[-\sqrt{\frac{2\alpha d(\rho,\epsilon)}{N\rho(r)[1-\rho(r)]}}\right]}.$$

Moreover with probability at least $1-\epsilon$*, for any posterior distribution* ρ *such that* $\rho(r) = 0$*,*

$$\rho(R) \leq 1 - \exp\left[-\frac{\mathcal{K}(\rho,\pi) - \log(\epsilon)}{N}\right].$$

We can also elaborate on the results in an other direction by introducing the *empirical dimension*

$$d_e = \sup_{\beta\in\mathbb{R}_+} \beta\big[\pi_{\exp(-\beta r)}(r) - \operatorname{ess\,inf}_\pi r\big] \leq -\log\big[\pi\big(r = \operatorname{ess\,inf}_\pi r\big)\big]. \tag{1.6}$$

There is no need to introduce a margin in this definition, since r takes at most N values, and therefore $\pi\big(r = \operatorname{ess\,inf}_\pi r\big)$ is strictly positive. This leads to

COROLLARY 1.2.9. *For any positive real constant* λ*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any posterior distribution* $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$*,*

$$\rho(R) \leq \Phi^{-1}_{\frac{\lambda}{N}}\left[\operatorname{ess\,inf}_\pi r + \frac{d_e}{\lambda}\log\left(\frac{e\lambda}{d_e}\right) + \frac{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big] - \log(\epsilon)}{\lambda}\right].$$

We could then make the bound uniform in λ and optimize this parameter in a way similar to what was done to obtain Theorem 1.2.8.

1.3. Local bounds

In this section, better bounds will be achieved through a better choice of the prior distribution. This better prior distribution turns out to depend on the unknown sample distribution $\mathbb{P}$, and some work is required to circumvent this and obtain empirical bounds.

1.3.1. Choice of the prior. As mentioned in the introduction, if one is willing to minimize the bound in expectation provided by Theorem 1.2.1 (page 19), one is led to consider the optimal choice $\pi = \mathbb{P}(\rho)$. However, this is only an ideal choice, since $\mathbb{P}$ is in all conceivable situations unknown. Nevertheless it shows that it is possible through Theorem 1.2.1 to measure the *complexity* of the classification model with $\mathbb{P}\big\{\mathcal{K}\big[\rho, \mathbb{P}(\rho)\big]\big\}$, which is nothing but the *mutual information* between the random sample $(X_i, Y_i)_{i=1}^N$ and the estimated parameter $\hat{\theta}$, under the joint distribution $\mathbb{P}\rho$.

In practice, since we cannot choose $\pi = \mathbb{P}(\rho)$, we have to be content with a *flat* prior π, resulting in a bound measuring complexity according to $\mathbb{P}\big[\mathcal{K}(\rho, \pi)\big] = \mathbb{P}\big\{\mathcal{K}\big[\rho, \mathbb{P}(\rho)\big]\big\} + \mathcal{K}\big[\mathbb{P}(\rho), \pi\big]$ larger by the entropy factor $\mathcal{K}\big[\mathbb{P}(\rho), \pi\big]$ than the optimal one (we are still commenting on Theorem 1.2.1).

If we want to base the choice of π on Theorem 1.2.4 (page 23), and if we choose $\rho = \pi_{\exp(-\lambda r)}$ to optimize this bound, we will be inclined to choose some π such that

$$\frac{1}{\lambda}\textstyle\int_0^\lambda \pi_{\exp(-\beta R)}(R)\, d\beta = -\frac{1}{\lambda}\log\Big\{\pi\big[\exp(-\lambda R)\big]\Big\}$$

is as far as possible close to $\inf_{\theta\in\Theta} R(\theta)$ in all circumstances. To give a more specific example, in the case when the distribution of the design $(X_i)_{i=1}^N$ is known, one can introduce on the parameter space Θ the metric D already defined by equation (1.3, page 20) (or some available upper bound for this distance). In view of the fact that $R(\theta) - R(\theta') \le D(\theta, \theta')$, for any θ, $\theta' \in \Theta$, it can be meaningful, at least theoretically, to choose π as

$$\pi = \sum_{k=1}^{\infty} \frac{1}{k(k+1)}\pi_k,$$

where π_k is the uniform measure on some minimal (or close to minimal) 2^{-k}-net $\mathcal{N}(\Theta, D, 2^{-k})$ of the metric space (Θ, D). With this choice

$$-\frac{1}{\lambda}\log\Big\{\pi\big[\exp(-\lambda R)\big]\Big\} \le \inf_{\theta\in\Theta} R(\theta) + \inf_k \left\{2^{-k} + \frac{\log(|\mathcal{N}(\Theta, D, 2^{-k})|) + \log[k(k+1)]}{\lambda}\right\}.$$

Another possibility, when we have to deal with real valued parameters, meaning that $\Theta \subset \mathbb{R}^d$, is to code each real component $\theta_i \in \mathbb{R}$ of $\theta = (\theta_i)_{i=1}^d$

to some precision and to use a prior μ which is atomic on dyadic numbers. More precisely let us parametrize the set of dyadic real numbers as

$$\mathcal{D} = \left\{ r\big[s, m, p, (b_j)_{j=1}^p\big] = s2^m \left(1 + \sum_{j=1}^p b_j 2^{-j}\right) \right. \\ \left. : \, s \in \{-1, +1\}, m \in \mathbb{Z}, p \in \mathbb{N}, b_j \in \{0, 1\} \right\},$$

where, as can be seen, s codes the sign, m the order of magnitude, p the precision and $(b_j)_{j=1}^p$ the binary representation of the dyadic number $r\big[s, m, p, (b_j)_{j=1}^p\big]$. We can for instance consider on $\mathcal{D}$ the probability distribution

$$\mu\Big\{r\big[s, m, p, (b_j)_{j=1}^p\big]\Big\} = \Big[3(|m|+1)(|m|+2)(p+1)(p+2)2^p\Big]^{-1}, \quad (1.7)$$

and define $\pi \in \mathcal{M}_+^1(\mathbb{R}^d)$ as $\pi = \mu^{\otimes d}$. This kind of "coding" prior distribution can be used also to define a prior on the integers (by renormalizing the restriction of μ to integers to get a probability distribution). Using μ is somehow equivalent to picking up a representative of each dyadic interval, and makes it possible to restrict to the case when the posterior ρ is a Dirac mass without losing too much (when $\Theta = (0, 1)$, this approach is somewhat equivalent to considering as prior distribution the Lebesgue measure and using as posterior distributions the uniform probability measures on dyadic intervals, with the advantage of obtaining non-randomized estimators). When one uses in this way an atomic prior and Dirac masses as posterior distributions, the bounds proven so far can be obtained through a simpler union bound argument. This is so true that some of the detractors of the PAC-Bayesian approach (which, as a newcomer, has sometimes received a suspicious greeting among statisticians) have argued that it cannot bring anything that elementary union bound arguments could not essentially provide. We do not share of course this derogatory opinion, and while we think that allowing for non atomic priors and posteriors is worthwhile, we also would like to stress that the upcoming local and relative bounds could hardly be obtained with the only help of union bounds.

Although the choice of a *flat* prior seems at first glance to be the only alternative when nothing is known about the sample distribution $\mathbb{P}$, the previous discussion shows that this type of choice is lacking proper localisation, and namely that we loose a factor $\mathcal{K}\big\{\mathbb{P}\big[\pi_{\exp(-\lambda r)}\big], \pi\big\}$, the divergence between the bound-optimal prior $\mathbb{P}\big[\pi_{\exp(-\lambda r)}\big]$, which is concentrated near the minima of R in favourable situations, and the flat prior π. Fortunately, there are technical ways to get around this difficulty and to obtain more local empirical bounds.

1.3.2. Unbiased local empirical bounds. The idea is to start with some flat prior $\pi \in \mathcal{M}_+^1(\Theta)$, and the posterior distribution $\rho = \pi_{\exp(-\lambda r)}$ minimizing the bound of Theorem 1.2.1 (page 19), when π is used as a prior. To improve the bound, we would like to use $\mathbb{P}\big[\pi_{\exp(-\lambda r)}\big]$ instead of π, and we are going to make the guess that we could approximate it with $\pi_{\exp(-\beta R)}$ (we have replaced the parameter λ with some distinct parameter β to give some more freedom to our investigation, and also because, intuitively, $\mathbb{P}\big[\pi_{\exp(-\lambda r)}\big]$ may be expected to be less concentrated than each of the $\pi_{\exp(-\lambda r)}$ it is mixing, which suggests that the best approximation of $\mathbb{P}\big[\pi_{\exp(-\lambda r)}\big]$ by some $\pi_{\exp(-\beta R)}$ may be obtained for some parameter $\beta < \lambda$). We are then led to look for some empirical upper bound of $\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big]$. This is happily provided by the following computation

$$\begin{aligned}\mathbb{P}\Big\{\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big]\Big\} &= \mathbb{P}\big[\mathcal{K}(\rho,\pi)\big] + \beta\mathbb{P}\big[\rho(R)\big] + \log\Big\{\pi\big[\exp(-\beta R)\big]\Big\}\\ &= \mathbb{P}\Big\{\mathcal{K}\big[\rho, \pi_{\exp(-\beta r)}\big]\Big\} + \beta\mathbb{P}\big[\rho(R-r)\big]\\ &\qquad + \log\Big\{\pi\big[\exp(-\beta R)\big]\Big\} - \mathbb{P}\Big\{\log\pi\big[\exp(-\beta r)\big]\Big\}.\end{aligned}$$

Using the convexity of $r \mapsto \log\big\{\pi\big[\exp(-\beta r)\big]\big\}$ as in equation (1.4) on page 23, we conclude that

$$0 \le \mathbb{P}\Big\{\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big]\Big\} \le \beta\mathbb{P}\big[\rho(R-r)\big] + \mathbb{P}\Big\{\mathcal{K}\big[\rho, \pi_{\exp(-\beta r)}\big]\Big\}.$$

This inequality has an interest of its own, since it provides a lower bound for $\mathbb{P}\big[\rho(R)\big]$. Moreover we can plug it into Theorem 1.2.1 (page 19) applied to the prior distribution $\pi_{\exp(-\beta R)}$ and obtain for any posterior distribution ρ and any positive parameter λ that

$$\Phi_{\frac{\lambda}{N}}\Big\{\mathbb{P}\big[\rho(R)\big]\Big\} \le \mathbb{P}\bigg\{\rho(r) + \frac{\beta}{\lambda}\rho(R-r) + \frac{1}{\lambda}\mathbb{P}\Big\{\mathcal{K}\big[\rho, \pi_{\exp(-\beta r)}\big]\Big\}\bigg\}.$$

In view of this, it it convenient to introduce the function

$$\begin{aligned}\widetilde{\Phi}_{a,b}(p) &= (1-b)^{-1}\big[\Phi_a(p) - bp\big]\\ &= -(1-b)^{-1}\Big\{a^{-1}\log\big\{1 - p\big[1-\exp(-a)\big]\big\} + bp\Big\},\\ &\qquad\qquad p \in (0,1), a \in]0,\infty[, b \in (0,1(.\end{aligned}$$

This is a convex function of p, moreover

$$\widetilde{\Phi}'_{a,b}(0) = \Big\{a^{-1}\big[1-\exp(-a)\big] - b\Big\}(1-b)^{-1},$$

showing that it is an increasing one to one convex map of the unit interval unto itself as soon as $b \le a^{-1}\big[1-\exp(-a)\big]$. Its convexity, combined with the value of its derivative at the origin, shows that

$$\widetilde{\Phi}_{a,b}(p) \ge \frac{a^{-1}\big[1-\exp(-a)\big] - b}{1-b}p.$$

Using this notation and remarks, we can state

THEOREM 1.3.1. *For any positive real constants β and λ such that $0 \leq \beta < N[1-\exp(-\frac{\lambda}{N})]$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}\mathbb{P}\left\{\rho(r) - \frac{\mathcal{K}\left[\rho, \pi_{\exp(-\beta r)}\right]}{\beta}\right\} &\leq \mathbb{P}\left[\rho(R)\right] \\ &\leq \widetilde{\Phi}^{-1}_{\frac{\lambda}{N},\frac{\beta}{\lambda}}\left\{\mathbb{P}\left[\rho(r) + \frac{\mathcal{K}\left[\rho, \pi_{\exp(-\beta r)}\right]}{\lambda - \beta}\right]\right\} \\ &\leq \frac{\lambda - \beta}{N[1-\exp(-\frac{\lambda}{N})] - \beta}\mathbb{P}\left[\rho(r) + \frac{\mathcal{K}\left[\rho, \pi_{\exp(-\beta r)}\right]}{\lambda - \beta}\right].\end{aligned}$$

Thus (taking $\lambda = 2\beta$), for any β such that $0 \leq \beta < \frac{N}{2}$,

$$\mathbb{P}\left[\rho(R)\right] \leq \frac{1}{1-\frac{2\beta}{N}}\mathbb{P}\left\{\rho(r) + \frac{\mathcal{K}\left[\rho, \pi_{\exp(-\beta r)}\right]}{\beta}\right\}.$$

Note that the last inequality is obtained using the fact that $1 - \exp(-x) \geq x - \frac{x^2}{2}$, $x \in \mathbb{R}_+$.

COROLLARY 1.3.2. *For any $\beta \in (0, N($,*

$$\begin{aligned}\mathbb{P}\left[\pi_{\exp(-\beta r)}(r)\right] &\leq \mathbb{P}\left[\pi_{\exp(-\beta r)}(R)\right] \\ &\leq \inf_{\lambda \in (-N\log(1-\frac{\beta}{N}),\infty(} \frac{\lambda - \beta}{N[1-\exp(-\frac{\lambda}{N})] - \beta}\mathbb{P}\left[\pi_{\exp(-\beta r)}(r)\right] \\ &\leq \frac{1}{1-\frac{2\beta}{N}}\mathbb{P}\left[\pi_{\exp(-\beta r)}(r)\right],\end{aligned}$$

the last inequality holding only when $\beta < \frac{N}{2}$.

It is interesting to compare the upper bound provided by this corollary with Theorem 1.2.1 (page 19) when the posterior is a Gibbs measure $\rho = \pi_{\exp(-\beta r)}$. We see that we have got rid of the entropy term $\mathcal{K}\left[\pi_{\exp(-\beta r)}, \pi\right]$, but at the price of an increase of the multiplicative factor, which for small values of $\frac{\beta}{N}$ grows from $(1 - \frac{\beta}{2N})^{-1}$ (when we take $\lambda = \beta$ in Theorem 1.2.1), to $(1 - \frac{2\beta}{N})^{-1}$. Therefore non-localized bounds have an interest of their own, and are superseded by localized bounds only in favourable circumstances (presumably when the sample is large enough when compared with the complexity of the classification model).

Corollary 1.3.2 shows that when $\frac{2\beta}{N}$ is small, $\pi_{\exp(-\beta r)}(r)$ is a tight approximation of $\pi_{\exp(-\beta r)}(R)$ in the mean (since we have an upper bound and a lower bound which are close together).

Another corollary is obtained by optimizing the bound given by Theorem 1.3.1 in ρ, which is done by taking $\rho = \pi_{\exp(-\lambda r)}$.

COROLLARY 1.3.3. *For any positive real constants β and λ such that $0 \leq \beta < N[1-\exp(-\frac{\lambda}{N})]$,*

$$\mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big] \leq \widetilde{\Phi}^{-1}_{\frac{\lambda}{N},\frac{\beta}{\lambda}}\left\{\mathbb{P}\left[\frac{1}{\lambda-\beta}\int_\beta^\lambda \pi_{\exp(-\gamma r)}(r)d\gamma\right]\right\}$$
$$\leq \frac{1}{N[1-\exp(-\frac{\lambda}{N})]-\beta}\mathbb{P}\Big[\textstyle\int_\beta^\lambda \pi_{\exp(-\gamma r)}(r)d\gamma\Big].$$

Although this inequality gives by construction a better upper bound for $\inf_{\lambda\in\mathbb{R}_+}\mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big]$ than Corollary 1.3.2, it is not easy to tell which one of the two inequalities is the best to bound $\mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big]$ for a fixed (and possibly suboptimal) value of λ, because in this case, one factor is improved while the other is worsened.

Using the *empirical dimension* d_e defined by equation (1.6) on page 27, we see that

$$\frac{1}{\lambda-\beta}\int_\beta^\lambda \pi_{\exp(-\gamma r)}(r)d\gamma \leq \operatorname*{ess\,inf}_\pi r + d_e\log\left(\frac{\lambda}{\beta}\right).$$

Therefore, in the case when we keep the ratio $\frac{\lambda}{\beta}$ bounded, we get a better dependence on the empirical dimension d_e than in Corollary 1.2.9 (page 27).

1.3.3. NON RANDOM LOCAL BOUNDS. Let us come now to the localization of the non-random upper bound given by Theorem 1.2.4 (page 23). According to Theorem 1.2.1 (page 19) applied to the localized prior $\pi_{\exp(-\beta R)}$,

$$\lambda\Phi_{\frac{\lambda}{N}}\big\{\mathbb{P}\big[\rho(R)\big]\big\} \leq \mathbb{P}\Big\{\lambda\rho(r)+\mathcal{K}(\rho,\pi)+\beta\rho(R)\Big\}+\log\big\{\pi\big[\exp(-\beta R)\big]\big\}$$
$$= \mathbb{P}\Big\{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]-\log\big\{\pi\big[\exp(-\lambda r)\big]\big\}+\beta\rho(R)\Big\}+\log\big\{\pi\big[\exp(-\beta R)\big]\big\}$$
$$\leq \mathbb{P}\Big\{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]+\beta\rho(R)\Big\}-\log\big\{\pi\big[\exp(-\lambda R)\big]\big\}+\log\big\{\pi\big[\exp(-\beta R)\big]\big\},$$

where we have used as previously inequality (1.4) (page 23). This proves

THEOREM 1.3.4. *For any posterior distribution $\rho:\Omega\to\mathcal{M}^1_+(\Theta)$, for any real parameters β and λ such that $0\leq\beta<N\big[1-\exp(-\frac{\lambda}{N})\big]$,*

$$\mathbb{P}\big[\rho(R)\big] \leq \widetilde{\Phi}^{-1}_{\frac{\lambda}{N},\frac{\beta}{\lambda}}\left\{\frac{1}{\lambda-\beta}\int_\beta^\lambda \pi_{\exp(-\gamma R)}(R)d\gamma+\mathbb{P}\left[\frac{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]}{\lambda-\beta}\right]\right\}$$
$$\leq \frac{1}{N\big[1-\exp(-\frac{\lambda}{N})\big]-\beta}\left\{\int_\beta^\lambda \pi_{\exp(-\gamma R)}(R)d\gamma+\mathbb{P}\Big\{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]\Big\}\right\}.$$

Let us notice in particular that this theorem contains Theorem 1.2.4 (page 23) which corresponds to the case $\beta=0$. As a corollary, we see also, taking

$\rho = \pi_{\exp(-\lambda r)}$ and $\lambda = 2\beta$, and noticing that $\gamma \mapsto \pi_{\exp(-\gamma R)}(R)$ is decreasing, that

$$\begin{aligned}\mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big] &\leq \inf_{\beta,\beta<N[1-\exp(-\frac{\lambda}{N})]} \frac{\beta}{N\big[1-\exp(-\frac{\lambda}{N})\big]-\beta}\pi_{\exp(-\beta R)}(R)\\ &\leq \frac{1}{1-\frac{\lambda}{N}}\pi_{\exp(-\frac{\lambda}{2}R)}(R).\end{aligned}$$

We can use this inequality in conjunction with the notion of dimension with margin η introduced by equation (1.5) on page 24, to see that the Gibbs posterior achieves for a proper choice of λ and any margin parameter $\eta \geq 0$ (which can be chosen to be equal to zero in parametric situations)

$$\begin{aligned}\inf_{\lambda}\mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big] \leq \operatorname{ess\,inf}_{\pi} R + \eta + \frac{4d_\eta}{N}\\ + 2\sqrt{\frac{2d_\eta\big(\operatorname{ess\,inf}_\pi R+\eta\big)}{N}+\frac{4d_\eta^2}{N^2}}.\end{aligned} \quad (1.8)$$

Deviation bounds to come next will show that the optimal λ can be estimated from empirical data.

Let us propose a little numerical example as an illustration: assuming that $d_0 = 10$, $N = 1000$ and $\operatorname{ess\,inf}_\pi R = 0.2$, we obtain from equation (1.8) that $\inf_\lambda \mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big] \leq 0.373$.

1.3.4. Local deviation bounds. When it comes to deviation bounds, for technical reasons we will choose a slightly more involved change of prior distribution and apply Theorem 1.2.6 (page 25) to the prior $\pi_{\exp[-\beta\Phi_{-\frac{\beta}{N}}\circ R]}$. The advantage of tweaking R with the nonlinear function $\Phi_{-\frac{\beta}{N}}$ will appear in the search for an empirical upper bound of the local entropy term. Theorem 1.1.4 (page 17), used with the above-mentioned local prior, shows that

$$\mathbb{P}\left\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda\Big\{\rho\big(\Phi_{\frac{\lambda}{N}}\circ R\big)-\rho(r)\Big\} - \mathcal{K}\big[\rho,\pi_{\exp(-\beta\Phi_{-\frac{\beta}{N}}\circ R)}\big]\right\} \leq 1. \quad (1.9)$$

Moreover

$$\begin{aligned}\mathcal{K}\big[\rho,\pi_{\exp[-\beta\Phi_{-\frac{\beta}{N}}\circ R]}\big] = \mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big] + \beta\rho\Big[\Phi_{-\frac{\beta}{N}}\circ R - r\Big]\\ + \log\Big\{\pi\Big[\exp\big(-\beta\Phi_{-\frac{\beta}{N}}\circ R\big)\Big]\Big\} - \log\Big\{\pi\Big[\exp(-\beta r)\Big]\Big\},\end{aligned} \quad (1.10)$$

which is an invitation to find an upper bound for $\log\Big\{\pi\Big[\exp\big[-\beta\Phi_{-\frac{\lambda}{N}}\circ R\big]\Big]\Big\} - \log\Big\{\pi\big[\exp(-\beta r)\big]\Big\}$. For conciseness, let us call our localized prior distribution $\overline{\pi}$, thus defined by its density

$$\frac{d\overline{\pi}}{d\pi}(\theta) = \frac{\exp\Big\{-\beta\Phi_{-\frac{\beta}{N}}\big[R(\theta)\big]\Big\}}{\pi\Big\{\exp\big[-\beta\Phi_{-\frac{\beta}{N}}\circ R\big]\Big\}}.$$

Applying once again Theorem 1.1.4 (page 17), but this time to $-\beta$, we see that

$$\begin{aligned}
&\mathbb{P}\biggl\{\exp\biggl[\log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr]\Bigr\}-\log\Bigl\{\pi\bigl[\exp(-\beta r)\bigr]\Bigr\}\biggr]\biggr\}\\
&\quad=\mathbb{P}\biggl\{\exp\biggl[\log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr]\Bigr\}+\inf_{\rho\in\mathcal{M}_+^1(\Theta)}\beta\rho(r)+\mathcal{K}(\rho,\pi)\biggr]\biggr\}\\
&\qquad\le\mathbb{P}\biggl\{\exp\biggl[\log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr]\Bigr\}+\beta\overline{\pi}(r)+\mathcal{K}(\overline{\pi},\pi)\biggr]\biggr\}\\
&\qquad\quad=\mathbb{P}\biggl\{\exp\Bigl[\beta\Bigl[\overline{\pi}(r)-\overline{\pi}\bigl(\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr]+\mathcal{K}(\overline{\pi},\overline{\pi})\Bigr]\biggr\}\le 1. \quad (1.11)
\end{aligned}$$

Combining equations (1.10) and (1.11) and using the concavity of $\Phi_{-\frac{\beta}{N}}$, we see that with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,

$$0\le\mathcal{K}(\rho,\overline{\pi})\le\mathcal{K}\bigl[\rho,\pi_{\exp(-\beta r)}\bigr]+\beta\Bigl[\Phi_{-\frac{\beta}{N}}\bigl[\rho(R)\bigr]-\rho(r)\Bigr]-\log(\epsilon).$$

We have proved a lower deviation bound:

THEOREM 1.3.5 *For any positive real constant β, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\frac{\exp\biggl\{\dfrac{\beta}{N}\biggl[\rho(r)-\dfrac{\mathcal{K}[\rho,\pi_{\exp(-\beta r)}]-\log(\epsilon)}{\beta}\biggr]\biggr\}-1}{\exp\bigl(\frac{\beta}{N}\bigr)-1}\le\rho(R).$$

We can also obtain a lower deviation bound for $\widehat{\theta}$. Indeed equation (1.11) can also be written as

$$\mathbb{P}\biggl\{\pi_{\exp(-\beta r)}\Bigl[\exp\Bigl\{\beta\Bigl[r-\Phi_{-\frac{\beta}{N}}\circ R\Bigr]\Bigr\}\Bigr]\biggr\}\le 1.$$

This means that for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\Bigl\{\rho\Bigl[\exp\bigl\{\beta\bigl[r-\Phi_{-\frac{\beta}{N}}\circ R\bigr]-\log\bigl(\tfrac{d\rho}{d\pi_{\exp(-\beta r)}}\bigr)\bigr\}\Bigr]\Bigr\}\le 1.$$

We have proved

THEOREM 1.3.6 *For any positive real constant β, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$, with $\mathbb{P}\rho$ probability at least $1-\epsilon$,*

$$\begin{aligned}
R(\widehat{\theta}\,)&\ge\Phi^{-1}_{-\frac{\beta}{N}}\biggl[r(\widehat{\theta}\,)-\frac{\log\bigl(\frac{d\rho}{d\pi_{\exp(-\beta r)}}\bigr)-\log(\epsilon)}{\beta}\biggr]\\
&=\frac{\exp\biggl\{\dfrac{\beta}{N}\biggl[r(\widehat{\theta}\,)-\dfrac{\log\bigl(\frac{d\rho}{d\pi_{\exp(-\beta r)}}\bigr)-\log(\epsilon)}{\beta}\biggr]\biggr\}-1}{\exp\biggl(\dfrac{\beta}{N}\biggr)-1}.
\end{aligned}$$

Let us now resume our investigation of the upper deviations of $\rho(R)$. Using the Cauchy-Schwarz inequality to combine equations (1.9, page 33) and (1.11, page 34), we obtain

$$\begin{aligned}
&\mathbb{P}\Big\{\exp\Big[\frac{1}{2}\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\lambda\rho\big(\Phi_{\frac{\lambda}{N}}\circ R\big)-\beta\rho\big(\Phi_{-\frac{\beta}{N}}\circ R\big)-(\lambda-\beta)\rho(r)-\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]\Big]\Big\}\\
&\qquad=\mathbb{P}\Big\{\exp\Big[\tfrac{1}{2}\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Big(\lambda\Big\{\rho\big(\Phi_{\frac{\lambda}{N}}\circ R\big)-\rho(r)\Big\}-\mathcal{K}(\rho,\overline{\pi})\Big)\Big]\\
&\qquad\times\exp\Big[\tfrac{1}{2}\Big(\log\Big\{\pi\Big[\exp\big(-\beta\Phi_{-\frac{\beta}{N}}\circ R\big)\Big]\Big\}-\log\Big\{\pi\Big[\exp(-\beta r)\Big]\Big\}\Big)\Big]\Big\}\\
&\qquad\le\mathbb{P}\Big\{\exp\Big[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Big(\lambda\Big\{\rho\big(\Phi_{\frac{\lambda}{N}}\circ R\big)-\rho(r)\Big\}-\mathcal{K}(\rho,\overline{\pi})\Big)\Big]\Big\}^{1/2}\\
&\times\mathbb{P}\Big\{\exp\Big[\Big(\log\Big\{\pi\Big[\exp\big(-\beta\Phi_{-\frac{\beta}{N}}\circ R\big)\Big]\Big\}-\log\Big\{\pi\Big[\exp(-\beta r)\Big]\Big\}\Big)\Big]\Big\}^{1/2}\le 1.
\end{aligned}\tag{1.12}$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution ρ,

$$\begin{aligned}
&\lambda\Phi_{\frac{\lambda}{N}}\big[\rho(R)\big]-\beta\Phi_{-\frac{\beta}{N}}\big[\rho(R)\big]\\
&\qquad\le\lambda\rho\big(\Phi_{\frac{\lambda}{N}}\circ R\big)-\beta\rho\big(\Phi_{-\frac{\beta}{N}}\circ R\big)\\
&\qquad\le(\lambda-\beta)\rho(r)+\mathcal{K}(\rho,\pi_{\exp(-\beta r)})-2\log(\epsilon).
\end{aligned}$$

(It would have been more straightforward to use a union bound on deviation inequalities instead of the Cauchy-Schwarz inequality on exponential moments, anyhow, this would have led to replace $-2\log(\epsilon)$ with the worse factor $2\log(\frac{2}{\epsilon})$.) Let us now recall that

$$\begin{aligned}
\lambda\Phi_{\frac{\lambda}{N}}(p)-\beta\Phi_{-\frac{\beta}{N}}(p)=&-N\log\Big\{1-\big[1-\exp\big(-\tfrac{\lambda}{N}\big)\big]p\Big\}\\
&-N\log\Big\{1+\big[\exp\big(\tfrac{\beta}{N}\big)-1\big]p\Big\},
\end{aligned}$$

and let us put

$$\begin{aligned}
B&=(\lambda-\beta)\rho(r)+\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]-2\log(\epsilon)\\
&=\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]+\textstyle\int_\beta^\lambda\pi_{\exp(-\xi r)}(r)d\xi-2\log(\epsilon).
\end{aligned}$$

Let us consider moreover the change of variables $\alpha=1-\exp(-\frac{\lambda}{N})$ and $\gamma=\exp(\frac{\beta}{N})-1$. We obtain $\big[1-\alpha\rho(R)\big]\big[1+\gamma\rho(R)\big]\ge\exp(-\frac{B}{N})$, leading to

THEOREM 1.3.7. *For any positive constants α, γ, such that $0\le\gamma<\alpha<1$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$, the bound*

$$M(\rho)=-\frac{\log\big[(1-\alpha)(1+\gamma)\big]}{\alpha-\gamma}\rho(r)+\frac{\mathcal{K}(\rho,\pi_{\exp[-N\log(1+\gamma)r]})-2\log(\epsilon)}{N(\alpha-\gamma)}$$

$$= \frac{\mathcal{K}\big[\rho, \pi_{\exp[N\log(1-\alpha)r]}\big] + \displaystyle\int_{N\log(1+\gamma)}^{-N\log(1-\alpha)} \pi_{\exp(-\xi r)}(r) d\xi - 2\log(\epsilon)}{N(\alpha-\gamma)},$$

is such that

$$\rho(R) \leq \frac{\alpha-\gamma}{2\alpha\gamma}\left(\sqrt{1+\frac{4\alpha\gamma}{(\alpha-\gamma)^2}\big\{1-\exp\big[-(\alpha-\gamma)M(\rho)\big]\big\}}-1\right) \leq M(\rho),$$

Let us now give an upper bound for $R(\widehat{\theta}\,)$. Equation (1.12 page 35) can also be written as

$$\mathbb{P}\left\{\left[\pi_{\exp(-\beta r)}\Big\{\exp\Big[\lambda\Phi_{\frac{\lambda}{N}}\circ R - \beta\Phi_{-\frac{\beta}{N}}\circ R - (\lambda-\beta)r\Big]\Big\}\right]^{\frac{1}{2}}\right\} \leq 1.$$

This means that for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\left\{\left[\rho\Big\{\exp\Big[\lambda\Phi_{\frac{\lambda}{N}}\circ R - \beta\Phi_{-\frac{\beta}{N}}\circ R - (\lambda-\beta)r - \log\big(\tfrac{d\rho}{d\pi_{\exp(-\beta r)}}\big)\Big]\Big\}\right]^{\frac{1}{2}}\right\} \leq 1.$$

Using the concavity of the square root function, this inequality can be weakened to

$$\mathbb{P}\left\{\rho\left[\exp\Big\{\tfrac{1}{2}\Big[\lambda\Phi_{\frac{\lambda}{N}}\circ R - \beta\Phi_{-\frac{\beta}{N}}\circ R - (\lambda-\beta)r - \log\big(\tfrac{d\rho}{d\pi_{\exp(-\beta r)}}\big)\Big]\Big\}\right]\right\} \leq 1.$$

We have proved

THEOREM 1.3.8. *For any positive real constants λ and β and for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}\rho$ probability at least $1-\epsilon$,*

$$\lambda\Phi_{\frac{\lambda}{N}}\big[R(\widehat{\theta}\,)\big] - \beta\Phi_{-\frac{\beta}{N}}\big[R(\widehat{\theta}\,)\big] \leq (\lambda-\beta)\, r(\widehat{\theta}\,) + \log\Big[\tfrac{d\rho}{d\pi_{\exp(-\beta r)}}(\widehat{\theta}\,)\Big] - 2\log(\epsilon).$$

Putting $\alpha = 1-\exp\big(-\frac{\lambda}{N}\big)$, $\gamma = \exp\big(\frac{\beta}{N}\big) - 1$ and

$$M(\theta) = -\frac{\log\big[(1-\alpha)(1+\gamma)\big]}{\alpha-\gamma} r(\theta) + \frac{\log\Big[\tfrac{d\rho}{d\pi_{\exp[-N\log(1+\gamma)r]}}(\theta)\Big] - 2\log(\epsilon)}{N(\alpha-\gamma)}$$

$$= \frac{\log\Big[\tfrac{d\rho}{d\pi_{\exp[N\log(1-\alpha)r]}}(\theta)\Big] + \displaystyle\int_{N\log(1+\gamma)}^{-N\log(1-\alpha)} \pi_{\exp(-\xi r)}(r)\, d\xi - 2\log(\epsilon)}{N(\alpha-\gamma)},$$

we can also, in the case when $\gamma < \alpha$, write this inequality as

$$R(\widehat{\theta}\,) \leq \frac{\alpha-\gamma}{2\alpha\gamma}\left(\sqrt{1+\frac{4\alpha\gamma}{(\alpha-\gamma)^2}\Big\{1-\exp\Big[-(\alpha-\gamma)M(\widehat{\theta}\,)\Big]\Big\}}-1\right) \leq M(\widehat{\theta}).$$

It may be enlightening to introduce the *empirical dimension* d_e defined by equation (1.6) on page 27. It provides the upper bound

$$\int_\beta^\lambda \pi_{\exp(-\xi r)}(r)d\xi \leq (\lambda-\beta)\operatorname*{ess\,inf}_\pi r + d_e \log\left(\frac{\lambda}{\beta}\right),$$

which shows that in Theorem 1.3.7 (page 35),

$$M(\rho) \leq \frac{\log\big[(1+\gamma)(1-\alpha)\big]}{\gamma-\alpha}\operatorname*{ess\,inf}_\pi r + \frac{d_e \log\left[\frac{-\log(1-\alpha)}{\log(1+\gamma)}\right] + \mathcal{K}\big[\rho, \pi_{\exp[N\log(1-\alpha)r]}\big] - 2\log(\epsilon)}{N(\alpha-\gamma)}.$$

Similarly, in Theorem 1.3.8 above,

$$M(\theta) \leq \frac{\log\big[(1+\gamma)(1-\alpha)\big]}{\gamma-\alpha}\operatorname*{ess\,inf}_\pi r + \frac{d_e \log\left[\frac{-\log(1-\alpha)}{\log(1+\gamma)}\right] + \log\left[\frac{d\rho}{d\pi_{\exp[N\log(1-\alpha)r]}}(\theta)\right] - 2\log(\epsilon)}{N(\alpha-\gamma)}$$

Let us give a little numerical illustration: assuming that $d_e = 10$, $N = 1000$, and $\operatorname{ess\,inf}_\pi r = 0.2$, taking $\epsilon = 0.01$, $\alpha = 0.5$ and $\gamma = 0.1$, we obtain from Theorem 1.3.7 $\pi_{\exp[N\log(1-\alpha)r]}(R) \simeq \pi_{\exp(-693r)}(R) \leq 0.332 \leq 0.372$, where we have given respectively the non-linear and the linear bound. This shows the practical interest of keeping the non-linearity. Optimizing the values of the parameters α and γ would not have yielded a significantly lower bound.

The following corollary is obtained by taking $\lambda = 2\beta$ and keeping only the linear bound; we give it for the sake of its simplicity:

Corollary 1.3.9. *For any positive real constant β such that* $\exp(\frac{\beta}{N}) + \exp(-\frac{2\beta}{N}) < 2$, *which is the case when $\beta < 0.48N$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\rho(R) \leq \frac{\beta\rho(r) + \mathcal{K}\big[\rho, \pi_{\exp(-\beta r)}\big] - 2\log(\epsilon)}{N\big[2 - \exp\big(\frac{\beta}{N}\big) - \exp\big(-\frac{2\beta}{N}\big)\big]} = \frac{\int_\beta^{2\beta} \pi_{\exp(-\xi r)}(r)d\xi + \mathcal{K}\big[\rho, \pi_{\exp(-2\beta r)}\big] - 2\log(\epsilon)}{N\big[2 - \exp(\frac{\beta}{N}) - \exp(-\frac{2\beta}{N})\big]}.$$

Let us mention that this corollary applied to the above numerical example gives $\pi_{\exp(-200r)}(R) \leq 0.475$ (when we take $\beta = 100$, consistently with the choice $\gamma = 0.1$).

1.3.5. Partially local bounds. Local bounds are suitable when the lowest values of the empirical error rate r are reached only on a small part of the parameter set Θ. When Θ is the disjoint union of sub-models of different complexities, the minimum of r will as a rule not be "localized" in a way that calls for the use of local bounds. Just think for instance of the case when $\Theta = \bigsqcup_{m=1}^{M} \Theta_m$, where the sets $\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta_M$ are nested. In this case we will have $\inf_{\Theta_1} r \geq \inf_{\Theta_2} r \geq \cdots \geq \inf_{\Theta_M} r$, although Θ_M may be too large to be the right model to use. In this situation, we do not want to localize the bound completely. Let us make a more specific fanciful but typical pseudo computation. Just imagine we have a countable collection $(\Theta_m)_{m\in M}$ of sub-models. Let us assume we are interested in choosing between the estimators $\widehat{\theta}_m \in \arg\min_{\Theta_m} r$, maybe randomizing them (e.g. replacing them with $\pi^m_{\exp(-\lambda r)}$). Let us imagine moreover that we are in a typically parametric situation, where, for some priors $\pi^m \in \mathcal{M}^1_+(\Theta_m)$, $m \in M$, there is a "dimension" d_m such that $\lambda\big[\pi^m_{\exp(-\lambda r)}(r) - r(\widehat{\theta}_m)\big] \simeq d_m$. Let $\mu \in \mathcal{M}^1_+(M)$ be some distribution on the index set M. It is easy to see that $(\mu\pi)_{\exp(-\lambda r)}$ will typically not be properly local, in the sense that typically

$$
\begin{aligned}
(\mu\pi)_{\exp(-\lambda r)}(r) &= \frac{\mu\Big\{\pi_{\exp(-\lambda r)}(r)\pi\big[\exp(-\lambda r)\big]\Big\}}{\mu\Big\{\pi\big[\exp(-\lambda r)\big]\Big\}} \\
&\simeq \frac{\displaystyle\sum_{m\in M}\big[(\inf_{\Theta_m} r) + \tfrac{d_m}{\lambda}\big]\exp\big[-\lambda(\inf_{\Theta_m} r) - d_m\log\big(\tfrac{e\lambda}{d_m}\big)\big]\mu(m)}{\displaystyle\sum_{m\in M}\exp\Big[-\lambda(\inf_{\Theta_m} r) - d_m\log\big(\tfrac{e\lambda}{d_m}\big)\Big]\mu(m)} \\
&\simeq \left\{\inf_{m\in M}(\inf_{\Theta_m} r) + \tfrac{d_m}{\lambda}\log\big(\tfrac{e\lambda}{d_m}\big) - \tfrac{1}{\lambda}\log[\mu(m)]\right\} \\
&\qquad + \log\left\{\sum_{m\in M}\exp\big[-d_m\log(\tfrac{\lambda}{d_m})\big]\mu(m)\right\}.
\end{aligned}
$$

where we have used the approximations

$$
\begin{aligned}
-\log\Big\{\pi\big[\exp(-\lambda r)\big]\Big\} &= \int_0^\lambda \pi_{\exp(-\beta r)}(r)d\beta \\
&\simeq \int_0^\lambda (\inf_{\Theta_m} r) + \big[\tfrac{d_m}{\beta}\wedge 1\big]d\beta \simeq \lambda(\inf_{\Theta_m} r) + d_m\big[\log\big(\tfrac{\lambda}{d_m}\big) + 1\big],
\end{aligned}
$$

and $\dfrac{\sum_m h(m)\exp[-h(m)]\nu(m)}{\sum_m \exp[-h(m)]\nu(m)} \simeq \inf_m h(m) - \log[\nu(m)]$, $\nu \in \mathcal{M}^1_+(M)$, taking $\nu(m) = \dfrac{\mu(m)\exp\big[-d_m\log\big(\frac{\lambda}{d_m}\big)\big]}{\sum_{m'}\mu(m')\exp\big[-d_{m'}\log\big(\frac{\lambda}{d_{m'}}\big)\big]}$.

These approximations have no pretension to be rigorous or very accurate, but they nevertheless give the best order of magnitude we can expect in

typical situations, and show that this order of magnitude is not what we are looking for: mixing different models with the help of μ spoils the localization, introducing a multiplier $\log\bigl(\frac{\lambda}{d_m}\bigr)$ to the dimension d_m which is precisely what we would have got if we had not localized the bound at all. What we would really like to do in such situations is to use a *partially localized* posterior distribution, such as $\pi^{\widehat{m}}_{\exp(-\lambda r)}$, where $\widehat{m}$ is an estimator of the best sub-model to be used. While the most straightforward way to do this is to use a union bound on results obtained for each sub-model Θ_m, here we are going to show how to allow arbitrary posterior distributions on the index set (corresponding to a randomization of the choice of $\widehat{m}$).

Let us consider the framework we just mentioned: let the measurable parameter set $(\Theta, \mathcal{T})$ be a union of measurable sub-models, $\Theta = \bigcup_{m\in M} \Theta_m$. Let the index set $(M, \mathcal{M})$ be some measurable space (most of the time it will be a countable set). Let $\mu \in \mathcal{M}_+^1(M)$ be a prior probability distribution on $(M, \mathcal{M})$. Let $\pi : M \to \mathcal{M}_+^1(\Theta)$ be a regular conditional probability measure such that $\pi(m, \Theta_m) = 1$, for any $m \in M$. Let $\mu\pi \in \mathcal{M}_+^1(M \times \Theta)$ be the product probability measure defined for any bounded measurable function $h : M \times \Theta \to \mathbb{R}$ by

$$\mu\pi(h) = \int_{m\in M} \left(\int_{\theta\in\Theta} h(m,\theta)\pi(m, d\theta) \right) \mu(dm).$$

For any bounded measurable function $h : \Omega \times M \times \Theta \to \mathbb{R}$, let $\pi_{\exp(h)} : \Omega \times M \to \mathcal{M}_+^1(\Theta)$ be the regular conditional posterior probability measure defined by

$$\frac{d\pi_{\exp(h)}}{d\pi}(m,\theta) = \frac{\exp\bigl[h(m,\theta)\bigr]}{\pi\bigl[m, \exp(h)\bigr]},$$

where consistently with previous notation $\pi(m,h) = \int_\Theta h(m,\theta)\pi(m,d\theta)$ (we will also often use the less explicit notation $\pi(h)$). For short, let

$$U(\theta,\omega) = \lambda\Phi_{\frac{\lambda}{N}}\bigl[R(\theta)\bigr] - \beta\Phi_{-\frac{\beta}{N}}\bigl[R(\theta)\bigr] - (\lambda-\beta)r(\theta,\omega).$$

Integrating with respect to μ equation (1.12, page 35), written in each sub-model Θ_m using the prior distribution $\pi(m,\cdot)$, we see that

$$\begin{aligned}
&\mathbb{P}\left\{ \exp\left[\sup_{\nu\in\mathcal{M}_+^1(M)} \sup_{\rho : M\to\mathcal{M}_+^1(\Theta)} \frac{1}{2}\Bigl[(\nu\rho)(U) - \nu\bigl\{\mathcal{K}\bigl(\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr)\bigr\}\Bigr] - \mathcal{K}(\nu,\mu)\right]\right\} \\
&\leq \mathbb{P}\left\{ \exp\left[\sup_{\nu\in\mathcal{M}_+^1(M)} \frac{1}{2}\nu\left(\sup_{\rho : M\to\mathcal{M}_+^1(\Theta)} \rho(U) - \mathcal{K}(\rho, \pi_{\exp(-\beta r)})\right) - \mathcal{K}(\nu,\mu)\right]\right\} \\
&\qquad = \mathbb{P}\left\{ \mu\left[\exp\left\{ \tfrac{1}{2} \sup_{\rho : M\to\mathcal{M}_+^1(\Theta)} \Bigl[\rho(U) - \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\Bigr]\right\}\right]\right\} \\
&\qquad\qquad = \mu\left\{ \mathbb{P}\left[\exp\left\{ \tfrac{1}{2} \sup_{\rho : M\to\mathcal{M}_+^1(\Theta)} \Bigl[\rho(U) - \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\Bigr]\right\}\right]\right\} \leq 1.
\end{aligned}$$

This proves that

$$\mathbb{P}\Bigg\{\exp\Bigg[\frac{1}{2}\sup_{\nu\in\mathcal{M}_+^1(M)}\sup_{\rho:M\to\mathcal{M}_+^1(\Theta)}\nu\rho\big[\lambda\Phi_{\frac{\lambda}{N}}(R)-\beta\Phi_{-\frac{\beta}{N}}(R)\big] \\ -(\lambda-\beta)\nu\rho(r)-2\mathcal{K}(\nu,\mu)-\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]\big\}\Bigg]\Bigg\}\leq 1. \quad (1.13)$$

Introducing the optimal value of r on each sub-model $r^\star(m)=\operatorname{ess\,inf}_{\pi(m,\cdot)} r$ and the empirical dimensions

$$d_e(m)=\sup_{\xi\in\mathbb{R}_+}\xi\big[\pi_{\exp(-\xi r)}(m,r)-r^\star(m)\big],$$

we can thus state

THEOREM 1.3.10. *For any positive real constants $\beta<\lambda$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu:\Omega\to\mathcal{M}_+^1(M)$, for any conditional posterior distribution $\rho:\Omega\times M\to\mathcal{M}_+^1(\Theta)$,*

$$\nu\rho\big[\lambda\Phi_{\frac{\lambda}{N}}(R)-\beta\Phi_{-\frac{\beta}{N}}(R)\big]\leq\lambda\Phi_{\frac{\lambda}{N}}\big[\nu\rho(R)\big]-\beta\Phi_{-\frac{\beta}{N}}\big[\nu\rho(R)\big]\leq B_1(\nu,\rho),$$

where $B_1(\nu,\rho)=(\lambda-\beta)\nu\rho(r)+2\mathcal{K}(\nu,\mu)+\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]\big\}-2\log(\epsilon)$

$$=\nu\bigg[\int_\beta^\lambda\pi_{\exp(-\alpha r)}(r)d\alpha\bigg]+2\mathcal{K}(\nu,\mu)+\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]\big\}-2\log(\epsilon)$$

$$=-2\log\Bigg\{\mu\bigg[\exp\bigg(-\frac{1}{2}\int_\beta^\lambda\pi_{\exp(-\alpha r)}(r)d\alpha\bigg)\bigg]\Bigg\} \\ +2\mathcal{K}\big[\nu,\mu_{\left(\frac{\pi[\exp(-\lambda r)]}{\pi[\exp(-\beta r)]}\right)^{1/2}}\big]+\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]\big\}-2\log(\epsilon),$$

and therefore $B_1(\nu,\rho)\leq\nu\Big[(\lambda-\beta)r^\star+\log\Big(\frac{\lambda}{\beta}\Big)d_e\Big]+2\mathcal{K}(\nu,\mu)$

$$+\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]\big\}-2\log(\epsilon),$$

as well as $B_1(\nu,\rho)\leq-2\log\Bigg\{\mu\bigg[\exp\bigg(-\frac{(\lambda-\beta)}{2}r^\star-\frac{1}{2}\log\big(\frac{\lambda}{\beta}\big)d_e\bigg)\bigg]\Bigg\}$

$$+2\mathcal{K}\big[\nu,\mu_{\left(\frac{\pi[\exp(-\lambda r)]}{\pi[\exp(-\beta r)]}\right)^{1/2}}\big]+\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]-2\log(\epsilon).$$

Thus, for any real constants α and γ such that $0\leq\gamma<\alpha<1$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu:\Omega\to\mathcal{M}_+^1(M)$ and any conditional posterior distribution $\rho:\Omega\times M\to\mathcal{M}_+^1(\Theta)$, the bound

$$B_2(\nu,\rho)=-\frac{\log\big[(1-\alpha)(1+\gamma)\big]}{\alpha-\gamma}\nu\rho(r)+\frac{2\mathcal{K}(\nu,\mu)+\nu\big\{\mathcal{K}\big[\rho,\pi_{(1+\gamma)^{-Nr}}\big]\big\}-2\log(\epsilon)}{N(\alpha-\gamma)}$$

$$=\frac{1}{N(\alpha-\gamma)}\Bigg\{2\mathcal{K}\bigg[\nu,\mu_{\left(\frac{\pi[(1-\alpha)^{Nr}]}{\pi[(1+\gamma)^{-Nr}]}\right)^{1/2}}\bigg]+\nu\Big\{\mathcal{K}\big[\rho,\pi_{(1-\alpha)^{Nr}}\big]\Big\}\Bigg\}$$

$$- \frac{2}{N(\alpha-\gamma)} \log\left\{ \mu\left[\exp\left[-\frac{1}{2} \int_{N\log(1+\gamma)}^{-N\log(1-\alpha)} \pi_{\exp(-\xi r)}(\cdot, r) d\xi \right] \right] \right\} - \frac{2\log(\epsilon)}{N(\alpha-\gamma)}$$

satisfies

$$\nu\rho(R) \leq \frac{\alpha-\gamma}{2\alpha\gamma} \left(\sqrt{1 + \frac{4\alpha\gamma}{(\alpha-\gamma)^2} \left\{ 1 - \exp\left[-(\alpha-\gamma) B_2(\nu,\rho) \right] \right\}} - 1 \right) \leq B_2(\nu,\rho).$$

If one is willing to bound the deviations with respect to $\mathbb{P}\nu\rho$, it is enough to remark that the equation preceding equation (1.13, page 40) can also be written as

$$\mathbb{P}\left\{ \mu\left[\left\{ \pi_{\exp(-\beta r)}\left[\exp\left\{ \lambda\Phi_{\frac{\lambda}{N}} \circ R - \beta\Phi_{-\frac{\beta}{N}} \circ R - (\lambda-\beta) r \right\} \right] \right\}^{1/2} \right] \right\} \leq 1.$$

Thus for any posterior distributions $\nu : \Omega \to \mathcal{M}_+^1(M)$ and $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\left\{ \nu\left[\left\{ \rho\Big[\exp\big\{ \lambda\Phi_{\frac{\lambda}{N}} \circ R - \beta\Phi_{-\frac{\beta}{N}} \circ R - (\lambda-\beta) r - 2\log\big(\tfrac{d\nu}{d\mu}\big) - \log\big(\tfrac{d\rho}{d\pi_{\exp(-\beta r)}}\big) \big\} \Big] \right\}^{1/2} \right] \right\} \leq 1.$$

Using the concavity of the square root function to pull the integration with respect to ρ out of the square root, we get

$$\mathbb{P}\nu\rho\left\{ \exp\left[\frac{1}{2} \left\{ \lambda\Phi_{\frac{\lambda}{N}} \circ R - \beta\Phi_{-\frac{\beta}{N}} \circ R - (\lambda-\beta) r - 2\log\big(\tfrac{d\nu}{d\pi}\big) - \log\big(\tfrac{d\rho}{d\pi_{\exp(-\beta r)}}\big) \right\} \right] \right\} \leq 1.$$

This leads to

THEOREM 1.3.11. *For any positive real constants $\beta < \lambda$, for any posterior distributions $\nu : \Omega \to \mathcal{M}_+^1(M)$ and $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}\nu\rho$ probability at least $1-\epsilon$,*

$$\lambda\Phi_{\frac{\lambda}{N}}\left[R(\widehat{m}, \widehat{\theta}) \right] - \beta\Phi_{-\frac{\beta}{N}}\left[R(\widehat{m}, \widehat{\theta}) \right] \leq (\lambda-\beta) r(\widehat{m}, \widehat{\theta}) + 2\log\left[\tfrac{d\nu}{d\mu}(\widehat{m}) \right] + \log\left[\tfrac{d\rho}{d\pi_{\exp(-\beta r)}}(\widehat{m}, \widehat{\theta}) \right] - 2\log(\epsilon) = \int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha$$

$$+ 2\log\big[\tfrac{d\nu}{d\mu}(\widehat{m})\big] + \log\big[\tfrac{d\rho}{d\pi_{\exp(-\lambda r)}}(\widehat{m},\widehat{\theta}\,)\big] - 2\log(\epsilon)$$
$$= 2\log\biggl\{\mu\biggl[\exp\biggl(-\frac{1}{2}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)d\alpha\biggr)\biggr]\biggr\}$$

$$+ 2\log\big[\tfrac{d\nu}{d\mu\left(\frac{\pi[\exp(-\lambda r)]}{\pi[\exp(-\beta r)]}\right)^{1/2}}(\widehat{m})\big] + \log\big[\tfrac{d\rho}{d\pi_{\exp(-\lambda r)}}(\widehat{m},\widehat{\theta}\,)\big] - 2\log(\epsilon).$$

Another way to state the same inequality is to say that for any real constants α and γ such that $0 \le \gamma < \alpha < 1$, with $\mathbb{P}\nu\rho$ probability at least $1-\epsilon$,

$$R(\widehat{m},\widehat{\theta}) \le \frac{\alpha-\gamma}{2\alpha\gamma}\biggl(\sqrt{1+\frac{4\alpha\gamma}{(\alpha-\gamma)^2}\Bigl\{1-\exp\bigl[-(\alpha-\gamma)B(\widehat{m},\widehat{\theta})\bigr]\Bigr\}}-1\biggr) \le B(\widehat{m},\widehat{\theta}),$$

where

$$B(\widehat{m},\widehat{\theta}) = -\frac{\log\bigl[(1-\alpha)(1+\gamma)\bigr]}{\alpha-\gamma} r(\widehat{m},\widehat{\theta}) + \frac{2\log\Bigl[\frac{d\nu}{d\mu}(\widehat{m})\Bigr] + \log\Bigl[\frac{d\rho}{d\pi_{(1+\gamma)^{-Nr}}}(\widehat{m},\widehat{\theta})\Bigr] - 2\log(\epsilon)}{N(\alpha-\gamma)}$$
$$= \frac{2}{N(\alpha-\gamma)}\log\Biggl[\frac{d\nu}{d\mu\left(\frac{\pi[(1-\alpha)^{Nr}]}{\pi[(1+\gamma)^{-Nr}]}\right)^{1/2}}(\widehat{m})\Biggr] + \frac{\log\Bigl[\frac{d\rho}{d\pi_{(1-\alpha)^{Nr}}}(\widehat{m},\widehat{\theta})\Bigr] - 2\log(\epsilon)}{N(\alpha-\gamma)} + \frac{2}{N(\alpha-\gamma)}\log\biggl\{\mu\biggl[\exp\biggl(-\frac{1}{2}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)d\alpha\biggr)\biggr]\biggr\}.$$

Let us remark that in the case when $\nu = \mu_{\left(\frac{\pi[(1-\alpha)^{Nr}]}{\pi[(1+\gamma)^{-Nr}]}\right)^{1/2}}$ and $\rho = \pi_{(1-\alpha)^{Nr}}$, we get as desired a bound that is adaptively local in all the Θ_m (at least when M is countable and μ is atomic):

$$B(\nu,\rho) \le -\tfrac{2}{N(\alpha-\gamma)}\log\biggl\{\mu\biggl\{\exp\biggl[\tfrac{N}{2}\log\bigl[(1+\gamma)(1-\alpha)\bigr]r^\star - \log\Bigl(\tfrac{-\log(1-\alpha)}{\log(1+\gamma)}\Bigr)\tfrac{d_e}{2}\biggr]\biggr\}\biggr\} - \frac{2\log(\epsilon)}{N(\alpha-\gamma)}$$
$$\le \inf_{m\in M}\biggl\{-\tfrac{\log\bigl[(1-\alpha)(1+\gamma)\bigr]}{\alpha-\gamma}r^\star(m)$$

$$+\log\Big(\tfrac{-\log(1-\alpha)}{\log(1+\gamma)}\Big)\,\tfrac{d_e(m)}{N(\alpha-\gamma)}-2\tfrac{\log\left[\epsilon\mu(m)\right]}{N(\alpha-\gamma)}\Big\}.$$

The penalization by the *empirical dimension* $d_e(m)$ in each sub-model is as desired linear in $d_e(m)$. Non random partially local bounds could be obtained in a way that is easy to imagine. We leave this investigation to the reader.

1.3.6. Two step localization. We have seen that the bound optimal choice of the posterior distribution ν on the index set in Theorem 1.3.10 (page 40) is such that

$$\frac{d\nu}{d\mu}(m)\sim\left(\frac{\pi\big[\exp\big(-\lambda r(m,\cdot)\big)\big]}{\pi\big[\exp\big(-\beta r(m,\cdot)\big)\big]}\right)^{\frac{1}{2}}=\exp\left[-\frac{1}{2}\int_\beta^\lambda\pi_{\exp(-\alpha r)}(m,r)d\alpha\right].$$

This suggests replacing the prior distribution μ with $\overline{\mu}$ defined by its density

$$\frac{d\overline{\mu}}{d\mu}(m)=\frac{\exp\big[-h(m)\big]}{\mu\big[\exp(-h)\big]},$$
$$\text{where } h(m)=-\xi\int_\beta^\gamma\pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\big[\Phi_{-\frac{\eta}{N}}\circ R(m,\cdot)\big]d\alpha. \quad (1.14)$$

The use of $\Phi_{-\frac{\eta}{N}}\circ R$ instead of R is motivated by technical reasons which will appear in subsequent computations. Indeed, we will need to bound

$$\nu\left[\int_\beta^\lambda\pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\big(\Phi_{-\frac{\eta}{N}}\circ R\big)d\alpha\right]$$

in order to handle $\mathcal{K}(\nu,\overline{\mu})$. In the spirit of equation (1.9, page 33), starting back from Theorem 1.1.4 (page 17), applied in each sub-model Θ_m to the prior distribution $\pi_{\exp(-\gamma\Phi_{-\frac{\eta}{N}}\circ R)}$ and integrated with respect to $\overline{\mu}$, we see that for any positive real constants λ, γ and η, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu:\Omega\to\mathcal{M}_+^1(M)$ on the index set and any conditional posterior distribution $\rho:\Omega\times M\to\mathcal{M}_+^1(\Theta)$,

$$\nu\rho\big(\lambda\Phi_{\frac{\lambda}{N}}\circ R-\gamma\Phi_{-\frac{\eta}{N}}\circ R\big)\le\lambda\nu\rho(r)$$
$$+\nu\mathcal{K}(\rho,\pi)+\mathcal{K}(\nu,\overline{\mu})+\nu\Big\{\log\Big[\pi\big[\exp\big(-\gamma\Phi_{-\frac{\eta}{N}}\circ R\big)\big]\Big]\Big\}-\log(\epsilon). \quad (1.15)$$

Since $x\mapsto f(x)\overset{\text{def}}{=}\lambda\Phi_{\frac{\lambda}{N}}-\gamma\Phi_{-\frac{\eta}{N}}(x)$ is a convex function, it is such that

$$f(x)\ge xf'(0)=xN\Big\{\big[1-\exp(-\tfrac{\lambda}{N})\big]+\tfrac{\gamma}{\eta}\big[\exp(\tfrac{\eta}{N})-1\big]\Big\}.$$

Thus if we put

$$\gamma=\frac{\eta\big[1-\exp(-\frac{\lambda}{N})\big]}{\exp(\frac{\eta}{N})-1}, \quad (1.16)$$

we obtain that $f(x) \geq 0$, $x \in \mathbb{R}$, and therefore that the left-hand side of equation (1.15) is non-negative. We can moreover introduce the prior conditional distribution $\overline{\pi}$ defined by

$$\frac{d\overline{\pi}}{d\pi}(m,\theta) = \frac{\exp\bigl[-\beta\Phi_{-\frac{\eta}{N}} \circ R(\theta)\bigr]}{\pi\bigl\{m, \exp\bigl[-\beta\Phi_{-\frac{\eta}{N}} \circ R\bigr]\bigr\}}.$$

With $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\nu : \Omega \to \mathcal{M}_+^1(M)$ and $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}
\beta\nu\rho(r) + \nu\bigl[\mathcal{K}(\rho,\pi)\bigr] &= \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} - \nu\biggl[\log\Bigl\{\pi\bigl[\exp(-\beta r)\bigr]\Bigr\}\biggr] \\
&\leq \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} + \beta\nu\overline{\pi}(r) + \nu\bigl[\mathcal{K}(\overline{\pi},\pi)\bigr] \\
&\leq \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} + \beta\nu\overline{\pi}\bigl(\Phi_{-\frac{\eta}{N}} \circ R\bigr) \\
&\qquad + \tfrac{\beta}{\eta}\bigl[\mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon)\bigr] + \nu\bigl[\mathcal{K}(\overline{\pi},\pi)\bigr] \\
&= \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} - \nu\Bigl\{\log\Bigl[\pi\bigl[\exp\bigl(-\beta\Phi_{-\frac{\eta}{N}} \circ R\bigr)\bigr]\Bigr]\Bigr\} \\
&\qquad + \tfrac{\beta}{\eta}\bigl[\mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon)\bigr].
\end{aligned}$$

Thus, coming back to equation (1.15), we see that under condition (1.16), with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\begin{aligned}
0 \leq{}& (\lambda-\beta)\nu\rho(r) + \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} \\
&- \nu\biggl[\int_\beta^\gamma \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}} \circ R)}\bigl(\Phi_{-\frac{\eta}{N}} \circ R\bigr) d\alpha\biggr] + (1+\tfrac{\beta}{\eta})\bigl[\mathcal{K}(\nu,\overline{\mu}) + \log(\tfrac{2}{\epsilon})\bigr].
\end{aligned}$$

Noticing moreover that

$$\begin{aligned}
&(\lambda-\beta)\nu\rho(r) + \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} \\
&\qquad = \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr]\bigr\} + \nu\biggl[\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha\biggr],
\end{aligned}$$

and choosing $\rho = \pi_{\exp(-\lambda r)}$, we have proved

THEOREM 1.3.12. *For any positive real constants β, γ and η, such that $\gamma < \eta\bigl[\exp(\frac{\eta}{N}) - 1\bigr]^{-1}$, defining λ by condition (1.16), so that $\lambda = -N\log\Bigl\{1 - \frac{\gamma}{\eta}\bigl[\exp(\frac{\eta}{N}) - 1\bigr]\Bigr\}$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$, any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
&\nu\biggl[\int_\beta^\gamma \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}} \circ R)}\bigl(\Phi_{-\frac{\eta}{N}} \circ R\bigr) d\alpha\biggr] \\
&\qquad \leq \nu\biggl[\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha\biggr] + \bigl(1+\tfrac{\beta}{\eta}\bigr)\bigl[\mathcal{K}(\nu,\overline{\mu}) + \log\bigl(\tfrac{2}{\epsilon}\bigr)\bigr].
\end{aligned}$$

Let us remark that this theorem does not require that $\beta < \gamma$, and thus provides both an upper and a lower bound for the quantity of interest:

COROLLARY 1.3.13. *For any positive real constants β, γ and η such that $\max\{\beta,\gamma\} < \eta\big[\exp(\frac{\eta}{N}) - 1\big]^{-1}$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\nu : \Omega \to \mathcal{M}_+^1(M)$ and $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
&\nu\left[\int_{-N\log\{1-\frac{\beta}{N}[\exp(\frac{\eta}{N})-1]\}}^{\gamma} \pi_{\exp(-\alpha r)}(r)d\alpha\right] - \big(1+\tfrac{\gamma}{\eta}\big)\big[\mathcal{K}(\nu,\overline{\mu}) + \log\big(\tfrac{3}{\epsilon}\big)\big] \\
&\quad\le \nu\left[\int_{\beta}^{\gamma} \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\big(\Phi_{-\frac{\eta}{N}}\circ R\big)d\alpha\right] \\
&\qquad\le \nu\left[\int_{\beta}^{-N\log\{1-\frac{\gamma}{\eta}[\exp(\frac{\eta}{N})-1]\}} \pi_{\exp(-\alpha r)}(r)d\alpha\right] \\
&\qquad\qquad + \big(1+\tfrac{\beta}{\eta}\big)\big[\mathcal{K}(\nu,\overline{\mu}) + \log\big(\tfrac{3}{\epsilon}\big)\big].
\end{aligned}$$

We can then remember that

$$\mathcal{K}(\nu,\overline{\mu}) = \xi\big(\nu-\overline{\mu}\big)\left[\int_{\beta}^{\gamma} \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\big(\Phi_{-\frac{\eta}{N}}\circ R\big)d\alpha\right] + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu),$$

to conclude that, putting

$$G_\eta(\alpha) = -N\log\big\{1-\tfrac{\alpha}{\eta}\big[\exp\big(\tfrac{\eta}{N}\big)-1\big]\big\} \ge \alpha, \qquad \alpha\in\mathbb{R}_+, \tag{1.17}$$

and

$$\frac{d\widehat{\nu}}{d\mu}(m) \stackrel{\text{def}}{=} \frac{\exp\big[-h(m)\big]}{\mu\big[\exp(-h)\big]} \text{ where } h(m) = \xi\int_{G_\eta(\beta)}^{\gamma} \pi_{\exp(-\alpha r)}(m,r)d\alpha, \tag{1.18}$$

the divergence of ν with respect to the local prior $\overline{\mu}$ is bounded by

$$\begin{aligned}
&\big[1-\xi\big(1+\tfrac{\beta}{\eta}\big)\big]\mathcal{K}(\nu,\overline{\mu}) \\
&\quad\le \xi\nu\left[\int_{\beta}^{G_\eta(\gamma)} \pi_{\exp(-\alpha r)}(r)d\alpha\right] - \xi\overline{\mu}\left[\int_{G_\eta(\beta)}^{\gamma} \pi_{\exp(-\alpha r)}(r)d\alpha\right] \\
&\qquad\qquad + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu) + \xi\big(2+\tfrac{\beta+\gamma}{\eta}\big)\log\big(\tfrac{3}{\epsilon}\big) \\
&\quad\le \xi\nu\left[\int_{\beta}^{G_\eta(\gamma)} \pi_{\exp(-\alpha r)}(r)d\alpha\right] + \mathcal{K}(\nu,\mu) \\
&\qquad\qquad + \log\left\{\mu\left[\exp\left(-\xi\int_{G_\eta(\beta)}^{\gamma} \pi_{\exp(-\alpha r)}(r)d\alpha\right)\right]\right\} \\
&\qquad\qquad\qquad + \xi\big(2+\tfrac{\beta+\gamma}{\eta}\big)\log\big(\tfrac{3}{\epsilon}\big) \\
&\quad= \mathcal{K}(\nu,\widehat{\nu}) + \xi\nu\left[\left(\int_{\beta}^{G_\eta(\beta)} + \int_{\gamma}^{G_\eta(\gamma)}\right)\pi_{\exp(-\alpha r)}(r)d\alpha\right] \\
&\qquad\qquad\qquad + \xi\big(2+\tfrac{\beta+\gamma}{\eta}\big)\log\big(\tfrac{3}{\epsilon}\big).
\end{aligned}$$

We have proved

THEOREM 1.3.14. *For any positive constants β, γ and η such that $\max\{\beta,\gamma\} < \eta\big[\exp(\frac{\eta}{N}) - 1\big]^{-1}$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$ and any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
\mathcal{K}(\nu,\overline{\mu}) &\leq \Big[1-\xi\Big(1+\frac{\beta}{\eta}\Big)\Big]^{-1}\Bigg\{\mathcal{K}(\nu,\widehat{\nu}) \\
&\qquad + \xi\nu\Bigg[\bigg(\int_\beta^{G_\eta(\beta)} + \int_\gamma^{G_\eta(\gamma)}\bigg)\pi_{\exp(-\alpha r)}(r)d\alpha\Bigg] \\
&\qquad\qquad + \xi\big(2+\tfrac{\beta+\gamma}{\eta}\big)\log\big(\tfrac{3}{\epsilon}\big)\Bigg\} \\
&\leq \Big[1-\xi\Big(1+\frac{\beta}{\eta}\Big)\Big]^{-1}\Bigg\{\mathcal{K}(\nu,\widehat{\nu}) \\
&\qquad + \xi\nu\Bigg[\big[G_\eta(\gamma)-\gamma+G_\eta(\beta)-\beta\big]r^\star + \log\bigg(\frac{G_\eta(\beta)G_\eta(\gamma)}{\beta\gamma}\bigg)d_e\Bigg] \\
&\qquad\qquad + \xi\big(2+\tfrac{\beta+\gamma}{\eta}\big)\log\big(\tfrac{3}{\epsilon}\big)\Bigg\},
\end{aligned}$$

where the local prior $\overline{\mu}$ is defined by equation (1.14, page 43) and the local posterior $\widehat{\nu}$ and the function G_η are defined by equation (1.18, page 45).

We can then use this theorem to give a local version of Theorem 1.3.10 (page 40). To get something pleasing to read, we can apply Theorem 1.3.14 with constants β', γ' and η chosen so that $\frac{2\xi}{1-\xi(1+\frac{\beta'}{\eta})} = 1$, $G_\eta(\beta') = \beta$ and $\gamma' = \lambda$, where β and λ are the constants appearing in Theorem 1.3.10. This gives

THEOREM 1.3.15. *For any positive real constants $\beta < \lambda$ and η such that $\lambda < \eta\big[\exp(\frac{\eta}{N}) - 1\big]^{-1}$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$, for any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$\nu\rho\big[\lambda\Phi_{\frac{\lambda}{N}}(R) - \beta\Phi_{-\frac{\beta}{N}}(R)\big] \leq \lambda\Phi_{\frac{\lambda}{N}}\big[\nu\rho(R)\big] - \beta\Phi_{-\frac{\beta}{N}}\big[\nu\rho(R)\big] \leq B_3(\nu,\rho),$$

where $B_3(\nu,\rho) = \nu\Bigg[\displaystyle\int_{G_\eta^{-1}(\beta)}^{G_\eta(\lambda)} \pi_{\exp(-\alpha r)}(r)d\alpha\Bigg]$

$$\begin{aligned}
&\qquad + \Big(3+\tfrac{G_\eta^{-1}(\beta)}{\eta}\Big)\mathcal{K}\Big[\nu, \mu_{\exp\left[-\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)d\alpha\right]}\Big] \\
&\qquad\qquad + \nu\big\{\mathcal{K}(\rho,\pi_{\exp(-\lambda r)}\big]\big\} + \Big(4+\tfrac{G_\eta^{-1}(\beta)+\lambda}{\eta}\Big)\log\big(\tfrac{4}{\epsilon}\big) \\
&\leq \nu\Big[\big[G_\eta(\lambda) - G_\eta^{-1}(\beta)\big]r^\star + \log\Big(\tfrac{G_\eta(\lambda)}{G_\eta^{-1}(\beta)}\Big)d_e\Big] \\
&\qquad + \Big(3+\tfrac{G_\eta^{-1}(\beta)}{\eta}\Big)\mathcal{K}\Big[\nu, \mu_{\exp\left[-\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)d\alpha\right]}\Big] \\
&\qquad\qquad + \nu\big\{\mathcal{K}(\rho,\pi_{\exp(-\lambda r)}\big]\big\} + \Big(4+\tfrac{G_\eta^{-1}(\beta)+\lambda}{\eta}\Big)\log\big(\tfrac{4}{\epsilon}\big),
\end{aligned}$$

and where the function G_η is defined by equation (1.17, page 45).

A first remark: if we had the stamina to use Cauchy Schwarz inequalities (or more generally Hölder inequalities) on exponential moments instead of using weighted union bounds on deviation inequalities, we could have replaced $\log(\frac{4}{\epsilon})$ with $-\log(\epsilon)$ in the above inequalities.

We see that we have achieved the desired kind of localization of Theorem 1.3.10 (page 40), since the new empirical entropy term

$$\mathcal{K}[\nu, \mu_{\exp[-\xi \int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha]}]$$

cancels for a value of the posterior distribution on the index set ν which is of the same form as the one minimizing the bound $B_1(\nu, \rho)$ of Theorem 1.3.10 (with a decreased constant, as could be expected). In a typical parametric setting, we will have

$$\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha \simeq (\lambda - \beta) r^\star(m) + \log\left(\tfrac{\lambda}{\beta}\right) d_e(m),$$

and therefore, if we choose for ν the Dirac mass at

$$\widehat{m} \in \arg\min_{m \in M} r^\star(m) + \tfrac{\log(\frac{\lambda}{\beta})}{\lambda - \beta} d_e(m),$$

and $\rho(m, \cdot) = \pi_{\exp(-\lambda r)}(m, \cdot)$, we will get, in the case when the index set M is countable,

$$\begin{aligned} B_3(\nu, \rho) \lesssim \max \left\{ \big[G_\eta(\lambda) - G_\eta^{-1}(\beta)\big], (\lambda - \beta) \tfrac{\log\left[\frac{G_\eta(\lambda)}{G_\eta^{-1}(\beta)}\right]}{\log(\frac{\lambda}{\beta})} \right\} \\ \times \Big[r^\star(\widehat{m}) + \tfrac{\log(\frac{\lambda}{\beta})}{\lambda - \beta} d_e(\widehat{m}) \Big] \\ + \Big(3 + \tfrac{G_\eta^{-1}(\beta)}{\eta}\Big) \log \Bigg\{ \sum_{m \in M} \frac{\mu(m)}{\mu(\widehat{m})} \exp \Bigg[- \Big(3 + \tfrac{G_\eta^{-1}(\beta)}{\eta}\Big)^{-1} \\ \times \Big\{ (\lambda - \beta) \big[r^\star(m) - r^\star(\widehat{m}) \big] + \log\big(\tfrac{\lambda}{\beta}\big) \big[d_e(m) - d_e(\widehat{m}) \big] \Big\} \Bigg] \Bigg\} \\ + \Big(4 + \tfrac{G_\eta^{-1}(\beta) + \lambda}{\eta}\Big) \log\big(\tfrac{4}{\epsilon}\big). \end{aligned}$$

This shows that the impact on the bound of the addition of supplementary models depends on their penalized minimum empirical risk $r^\star(m) + \frac{\log(\frac{\lambda}{\beta})}{\lambda - \beta} d_e(m)$. More precisely the adaptive and local complexity factor

$$\begin{aligned} \log \Bigg\{ \sum_{m \in M} \frac{\mu(m)}{\mu(\widehat{m})} \exp \Bigg[- \Big(3 + \tfrac{G_\eta^{-1}(\beta)}{\eta}\Big)^{-1} \\ \times \Big\{ (\lambda - \beta) \big[r^\star(m) - r^\star(\widehat{m}) \big] + \log\big(\tfrac{\lambda}{\beta}\big) \big[d_e(m) - d_e(\widehat{m}) \big] \Big\} \Bigg] \Bigg\} \end{aligned}$$

replaces in this bound the non local factor

$$\mathcal{K}(\nu,\mu) = -\log\bigl[\mu(\widehat{m})\bigr] = \log\left[\sum_{m\in M}\frac{\mu(m)}{\mu(\widehat{m})}\right]$$

which appears when applying Theorem 1.3.10 (page 40) to the Dirac mass $\nu = \delta_{\widehat{m}}$. Thus in the local bound, the influence of models decreases exponentially fast when their penalized empirical risk increases.

One can deduce a result about the deviations with respect to the posterior $\nu\rho$ from Theorem 1.3.15 (page 46) without much supplementary work: it is enough for that purpose to remark that with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$,

$$\begin{aligned}
\nu\left[\log\Bigl\{\pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{\lambda\Phi_{\frac{\lambda}{N}}(R) - \beta\Phi_{-\frac{\beta}{N}}(R)\bigr\}\Bigr]\Bigr\}\right] \\
- \nu\left(\int_{G_\eta^{-1}(\beta)}^{G_\eta(\lambda)} \pi_{\exp(-\alpha r)}(r)\,d\alpha\right) \\
- \Bigl(3 + \tfrac{G_\eta^{-1}(\beta)}{\eta}\Bigr)\mathcal{K}\Bigl[\nu, \mu_{\exp\left[-\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)\,d\alpha\right]}\Bigr] \\
- \Bigl(4 + \tfrac{G_\eta^{-1}(\beta)+\lambda}{\eta}\Bigr)\log\Bigl(\tfrac{4}{\epsilon}\Bigr) \le 0,
\end{aligned}$$

this inequality being obtained by taking a supremum in ρ in Theorem 1.3.15 (page 46). One can then take a supremum in ν, to get, still with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\begin{aligned}
\log\Biggl\{\mu_{\exp\left[-\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)\,d\alpha\right]}\Biggl[\\
\Bigl\{\pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{\lambda\Phi_{\frac{\lambda}{N}}(R) - \beta\Phi_{-\frac{\beta}{N}}(R)\bigr\}\Bigr]\Bigr\}^{\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}} \\
\times \exp\left(-\Bigl(3 + \tfrac{G_\eta^{-1}(\beta)}{\eta}\Bigr)^{-1}\int_{G_\eta^{-1}(\beta)}^{G_\eta(\lambda)} \pi_{\exp(-\alpha r)}(r)\,d\alpha\right)\Biggr]\Biggr\} \\
\le \frac{4 + \frac{G_\eta^{-1}(\beta)+\lambda}{\eta}}{3 + \frac{G_\eta^{-1}(\beta)}{\eta}}\log\bigl(\tfrac{4}{\epsilon}\bigr).
\end{aligned}$$

Using the fact that $x \mapsto x^\alpha$ is concave when $\alpha = \bigl(3 + \frac{G_\eta^{-1}(\beta)}{\eta}\bigr)^{-1} < 1$, we get for any posterior conditional distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

$$\mu_{\exp\left[-\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)\,d\alpha\right]}\rho\Biggl\{$$

$$\exp\left[\left(3+\tfrac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\left(\lambda\Phi_{\frac{\lambda}{N}}(R)-\beta\Phi_{-\frac{\beta}{N}}(R)-\int_{G_\eta^{-1}(\beta)}^{G_\eta(\lambda)}\pi_{\exp(-\alpha r)}(r)d\alpha\right.\right.$$
$$\left.\left.+\log\left[\frac{d\rho}{d\pi_{\exp(-\lambda r)}}(\widehat{m},\widehat{\theta})\right]\right)\right]\Bigg\}$$
$$\leq\exp\left(\frac{4+\frac{G_\eta^{-1}(\beta)+\lambda}{\eta}}{3+\frac{G_\eta^{-1}(\beta)}{\eta}}\log\left(\tfrac{4}{\epsilon}\right)\right).$$

We can thus state

THEOREM 1.3.16. *For any $\epsilon\in]0,1[$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu:\Omega\to\mathcal{M}_+^1(M)$ and conditional posterior distribution $\rho:\Omega\times M\to\mathcal{M}_+^1(\Theta)$, for any $\xi\in]0,1[$, with $\nu\rho$ probability at least $1-\xi$,*

$$\lambda\Phi_{\frac{\lambda}{N}}(R)-\beta\Phi_{-\frac{\beta}{N}}(R)\leq\int_{G_\eta^{-1}(\beta)}^{G_\eta(\lambda)}\pi_{\exp(-\alpha r)}(r)d\alpha$$
$$+\left(3+\tfrac{G_\eta^{-1}(\beta)}{\eta}\right)\log\left[\frac{d\nu}{d\mu}_{\exp\left[-\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\int_\beta^\lambda\pi_{\exp(-\alpha r)}(r)d\alpha\right]}(\widehat{m})\right]$$
$$+\log\left[\frac{d\rho}{d\pi_{\exp(-\lambda r)}}(\widehat{m},\widehat{\theta})\right]+\left(4+\tfrac{G_\eta^{-1}(\beta)+\lambda}{\eta}\right)\log\left(\tfrac{4}{\epsilon}\right)-\left(3+\tfrac{G_\eta^{-1}(\beta)}{\eta}\right)\log(\xi).$$

Note that the given bound consequently holds with $\mathbb{P}\nu\rho$ probability at least $(1-\epsilon)(1-\xi)\geq 1-\epsilon-\xi$.

1.4. Relative bounds

The behaviour of the minimum of the empirical process $\theta\mapsto r(\theta)$ is known to depend on the covariances between pairs $\big[r(\theta),r(\theta')\big]$, $\theta,\theta'\in\Theta$. In this respect, our previous study, based on the analysis of the variance of $r(\theta)$ (or technically on some exponential moment playing quite the same role), loses some accuracy in some circumstances (namely when $\inf_\Theta R$ is not close enough to zero).

In this section, instead of bounding the expected risk $\rho(R)$ of any posterior distribution, we are going to upper bound the difference $\rho(R)-\inf_\Theta R$, and more generally $\rho(R)-R(\widetilde{\theta})$, where $\widetilde{\theta}\in\Theta$ is some fixed parameter value.

In the next section we will analyse $\rho(R)-\pi_{\exp(-\beta R)}(R)$, allowing us to compare the expected error rate of a posterior distribution ρ with the error rate of a Gibbs prior distribution. We will also analyse $\rho_1(R)-\rho_2(R)$, where ρ_1 and ρ_2 are two arbitrary posterior distributions, using comparison with a Gibbs prior distribution as a tool, and in particular as a tool to establish the required Kullback divergence bounds.

Relative bounds do not provide the same kind of results as direct bounds on the error rate: it is not possible to estimate $\rho(R)$ with an order of precision higher than $(\rho(R)/N)^{1/2}$, so that relative bounds cannot of course achieve that, but they provide a way to reach a faster rate for $\rho(R) - \inf_\Theta R$, that is for the relative performance of the estimator within a restricted model.

The study of PAC-Bayesian relative bounds was initiated in the second and third parts of J.-Y. Audibert's dissertation (Audibert, 2004b).

In this section and the next, we will suggest a series of possible uses of relative bounds. As usual, we will start with the simplest inequalities and proceed towards more sophisticated techniques with better theoretical properties, but at the same time less precise constants, so that which one is the more fitted will depend on the size of the training sample.

The first thing we will do is to compute for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ a relative performance bound bearing on $\rho(R) - \inf_\Theta R$. We will also compare the classification model indexed by Θ with a sub-model indexed by one of its measurable subsets $\Theta_1 \subset \Theta$. For this purpose we will form the difference $\rho(R) - R(\widetilde{\theta})$, where $\widetilde{\theta} \in \Theta_1$ is some possibly unobservable value of the parameter in the sub-model defined by Θ_1, typically chosen in $\arg\min_{\Theta_1} R$. If this is so and $\rho(R) - R(\widetilde{\theta}) = \rho(R) - \inf_{\Theta_1} R$, a negative upper bound indicates that it is definitely worth using a randomized estimator ρ supported by the larger parameter set Θ instead of using only the classification model defined by the smaller set Θ_1.

1.4.1. Basic inequalities. Relative bounds in this section are based on the control of $r(\theta) - r(\widetilde{\theta})$, where $\theta, \widetilde{\theta} \in \Theta$. These differences are related to the random variables

$$\psi_i(\theta, \widetilde{\theta}) = \sigma_i(\theta) - \sigma_i(\widetilde{\theta}) = \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big] - \mathbb{1}\big[f_{\widetilde{\theta}}(X_i) \neq Y_i\big].$$

Some supplementary technical difficulties, as compared to the previous sections, come from the fact that $\psi_i(\theta, \widetilde{\theta})$ takes three values, whereas $\sigma_i(\theta)$ takes only two. Let

$$r'(\theta, \widetilde{\theta}) = r(\theta) - r(\widetilde{\theta}) = \frac{1}{N} \sum_{i=1}^N \psi_i(\theta, \widetilde{\theta}), \quad \theta, \widetilde{\theta} \in \Theta, \tag{1.19}$$

and $R'(\theta, \widetilde{\theta}) = R(\theta) - R(\widetilde{\theta}) = \mathbb{P}\big[r'(\theta, \widetilde{\theta})\big]$. We have as usual from independence that

$$\begin{aligned}\log\Big\{\mathbb{P}\Big[\exp\big[-\lambda r'(\theta, \widetilde{\theta})\big]\Big]\Big\} &= \sum_{i=1}^N \log\Big\{\mathbb{P}\Big[\exp\big[-\tfrac{\lambda}{N}\psi_i(\theta, \widetilde{\theta})\big]\Big]\Big\} \\ &\leq N \log\bigg\{\frac{1}{N}\sum_{i=1}^N \mathbb{P}\Big\{\exp\Big[-\frac{\lambda}{N}\psi_i(\theta, \widetilde{\theta})\Big]\Big\}\bigg\}.\end{aligned}$$

Let C_i be the distribution of $\psi_i(\theta,\widetilde{\theta})$ under $\mathbb{P}$ and let $\bar{C} = \frac{1}{N}\sum_{i=1}^{N} C_i \in \mathcal{M}_+^1\big(\{-1,0,1\}\big)$. With this notation

$$\log\Big\{\mathbb{P}\Big[\exp\big[-\lambda r'(\theta,\widetilde{\theta})\big]\Big]\Big\} \le N\log\Big\{\int_{\psi\in\{-1,0,1\}} \exp\Big(-\frac{\lambda}{N}\psi\Big)\bar{C}(d\psi)\Big\}. \tag{1.20}$$

The right-hand side of this inequality is a function of $\bar{C}$. On the other hand, $\bar{C}$ being a probability measure on a three point set, is defined by two parameters, that we may take equal to $\int \psi\bar{C}(d\psi)$ and $\int \psi^2\bar{C}(d\psi)$. To this purpose, let us introduce

$$M'(\theta,\widetilde{\theta}) = \int \psi^2\bar{C}(d\psi) = \bar{C}(+1) + \bar{C}(-1) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}\big[\psi_i^2(\theta,\widetilde{\theta})\big], \quad \theta,\widetilde{\theta}\in\Theta.$$

It is a pseudo distance (meaning that it is symmetric and satisfies the triangle inequality), since it can also be written as

$$M'(\theta,\widetilde{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}\Big\{\Big|\mathbb{1}\big[f_\theta(X_i)\neq Y_i\big] - \mathbb{1}\big[f_{\widetilde{\theta}}(X_i)\neq Y_i\big]\Big|\Big\}, \quad \theta,\widetilde{\theta}\in\Theta.$$

It is readily seen that

$$N\log\left\{\int \exp\left(-\frac{\lambda}{N}\psi\right)\bar{C}(d\psi)\right\} = -\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\theta,\widetilde{\theta}), M'(\theta,\widetilde{\theta})\big],$$

where

$$\begin{aligned}\Psi_a(p,m) &= -a^{-1}\log\Big[(1-m) + \frac{m+p}{2}\exp(-a) + \frac{m-p}{2}\exp(a)\Big]\\ &= -a^{-1}\log\Big\{1-\sinh(a)\big[p - m\tanh(\tfrac{a}{2})\big]\Big\}.\end{aligned} \tag{1.21}$$

Thus plugging this equality into inequality (1.20, page 51) we get

THEOREM 1.4.1. *For any real parameter λ,*

$$\log\Big\{\mathbb{P}\Big[\exp\big[-\lambda r'(\theta,\widetilde{\theta})\big]\Big]\Big\} \le -\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\theta,\widetilde{\theta}), M'(\theta,\widetilde{\theta})\big], \quad \theta,\widetilde{\theta}\in\Theta,$$

where r' is defined by equation (1.19, page 50) and Ψ and M' are defined just above.

To make a link with previous work of Mammen and Tsybakov — see e.g. Mammen et al. (1999) and Tsybakov (2004) — we may consider the pseudo-distance D on Θ defined by equation (1.3, page 20). This distance only depends on the distribution of the patterns. It is often used to formulate margin assumptions, in the sense of Mammen and Tsybakov. Here we are going to work rather with M': as it is dominated by D in the sense that $M'(\theta,\widetilde{\theta}) \le D(\theta,\widetilde{\theta})$, $\theta,\widetilde{\theta}\in\Theta$, with equality in the important case of binary

classification, hypotheses formulated on D induce hypotheses on M', and working with M' may only sharpen the results when compared to working with D.

Using the same reasoning as in the previous section, we deduce

THEOREM 1.4.2. *For any real parameter λ, any $\widetilde{\theta} \in \Theta$, any prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda\Big[\rho\big\{\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}),M'(\cdot,\widetilde{\theta})\big]\big\}-\rho\big[r'(\cdot,\widetilde{\theta})\big]\Big]-\mathcal{K}(\rho,\pi)\right]\right\}\leq 1.$$

We are now going to derive some other type of relative exponential inequality. In Theorem 1.4.2 we obtained an inequality comparing one observed quantity $\rho\big[r'(\cdot,\widetilde{\theta})\big]$ with two unobserved ones, $\rho\big[R'(\cdot,\widetilde{\theta})\big]$ and $\rho\big[M'(\cdot,\widetilde{\theta})\big]$, — indeed, because of the convexity of the function $\lambda\Psi_{\frac{\lambda}{N}}$,

$$\lambda\rho\big\{\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}),M'(\cdot,\widetilde{\theta})\big]\big\}\geq\lambda\Psi_{\frac{\lambda}{N}}\big\{\rho\big[R'(\cdot,\widetilde{\theta})\big],\rho\big[M'(\cdot,\widetilde{\theta})\big]\big\}.$$

This may be inconvenient when looking for an empirical bound for $\rho\big[R'(\cdot,\widetilde{\theta})\big]$, and we are going now to seek an inequality comparing $\rho\big[R'(\cdot,\widetilde{\theta})\big]$ with empirical quantities only.

This is possible by considering the log-Laplace transform of some modified random variable $\chi_i(\theta,\widetilde{\theta})$. We may consider more precisely the change of variable defined by the equation

$$\exp\left(-\frac{\lambda}{N}\chi_i\right)=1-\frac{\lambda}{N}\psi_i,$$

which is possible when $\frac{\lambda}{N}\in\,]-1,1[$ and leads to define

$$\chi_i=-\frac{N}{\lambda}\log\left(1-\frac{\lambda}{N}\psi_i\right).$$

We may then work on the log-Laplace transform

$$\begin{aligned}\log\left\{\mathbb{P}\left[\exp\left\{-\frac{\lambda}{N}\sum_{i=1}^N\chi_i(\theta,\widetilde{\theta})\right\}\right]\right\}&=\log\left\{\mathbb{P}\left[\prod_{i=1}^N\left(1-\frac{\lambda}{N}\psi_i(\theta,\widetilde{\theta})\right)\right]\right\}\\&=\log\left\{\mathbb{P}\left[\exp\left\{\sum_{i=1}^N\log\left[1-\frac{\lambda}{N}\psi_i(\theta,\widetilde{\theta})\right]\right\}\right]\right\}.\end{aligned}$$

We may now follow the same route as previously, writing

$$\log\left\{\mathbb{P}\left[\exp\left\{\sum_{i=1}^N\log\left[1-\frac{\lambda}{N}\psi_i(\theta,\widetilde{\theta})\right]\right\}\right]\right\}$$

$$= \sum_{i=1}^{N} \log\Big[1 - \frac{\lambda}{N}\mathbb{P}\big[\psi_i(\theta,\widetilde{\theta})\big]\Big] \le N \log\Big[1 - \frac{\lambda}{N} R'(\theta,\widetilde{\theta}\,)\Big].$$

Let us also introduce the random pseudo distance

$$\begin{aligned} m'(\theta,\widetilde{\theta}) &= \frac{1}{N}\sum_{i=1}^{N} \psi_i(\theta,\widetilde{\theta})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} \Big| \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big] - \mathbb{1}\big[f_{\widetilde{\theta}}(X_i) \neq Y_i\big]\Big|, \quad \theta,\widetilde{\theta} \in \Theta. \quad (1.22) \end{aligned}$$

This is the empirical counterpart of M', implying that $\mathbb{P}(m') = M'$. Let us notice that

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \log\big[1 - \tfrac{\lambda}{N}\psi_i(\theta,\widetilde{\theta})\big] &= \frac{\log(1-\frac{\lambda}{N}) - \log(1+\frac{\lambda}{N})}{2} r'(\theta,\widetilde{\theta}) \\ &\qquad + \frac{\log(1-\frac{\lambda}{N}) + \log(1+\frac{\lambda}{N})}{2} m'(\theta,\widetilde{\theta}) \\ &= \frac{1}{2}\log\left(\frac{1-\frac{\lambda}{N}}{1+\frac{\lambda}{N}}\right) r'\big(\theta,\widetilde{\theta}\,\big) + \frac{1}{2}\log\big(1-\tfrac{\lambda^2}{N^2}\big) m'\big(\theta,\widetilde{\theta}\,\big). \end{aligned}$$

Let us put $\gamma = \frac{N}{2}\log\Big(\frac{1+\frac{\lambda}{N}}{1-\frac{\lambda}{N}}\Big)$, so that

$$\lambda = N\tanh\big(\tfrac{\gamma}{N}\big) \text{ and } \tfrac{N}{2}\log\Big(1-\tfrac{\lambda^2}{N^2}\Big) = -N\log\big[\cosh(\tfrac{\gamma}{N})\big].$$

With this notation, we can conveniently write the previous inequality as

$$\begin{aligned} \mathbb{P}\Big\{\exp\Big[-N\log\big[1-\tanh\big(\tfrac{\gamma}{N}\big)R'(\theta,\widetilde{\theta})\big] \\ - \gamma r'\big(\theta,\widetilde{\theta}\,\big) - N\log\big[\cosh(\tfrac{\gamma}{N})\big] m'\big(\theta,\widetilde{\theta}\,\big)\Big]\Big\} \le 1. \end{aligned}$$

Integrating with respect to a prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$, we obtain

THEOREM 1.4.3. *For any real parameter γ, for any $\widetilde{\theta} \in \Theta$, for any prior probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned} \mathbb{P}\Bigg\{\exp\Bigg[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \Big\{-N\rho\Big\{\log\big[1-\tanh\big(\tfrac{\gamma}{N}\big)R'(\cdot,\widetilde{\theta}\,)\big]\Big\} \\ - \gamma\rho\big[r'(\cdot,\widetilde{\theta}\,)\big] - N\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho\big[m'(\cdot,\widetilde{\theta}\,)\big] - \mathcal{K}(\rho,\pi)\Big\}\Bigg]\Bigg\} \le 1. \end{aligned}$$

1.4.2. Non random bounds. Let us first deduce a non-random bound from Theorem 1.4.2 (page 52). This theorem can be conveniently taken advantage of by throwing the non-linearity into a localized prior, considering the prior probability measure μ defined by its density

$$\frac{d\mu}{d\pi}(\theta) = \frac{\exp\big\{-\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\theta,\widetilde{\theta}),M'(\theta,\widetilde{\theta})\big]+\beta R'(\theta,\widetilde{\theta})\big\}}{\pi\Big\{\exp\big\{-\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}),M'(\cdot,\widetilde{\theta})\big]+\beta R'(\cdot,\widetilde{\theta})\big\}\Big\}}.$$

Indeed, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,

$$\mathcal{K}(\rho,\mu) = \mathcal{K}(\rho,\pi) + \lambda\rho\Big\{\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}),M'(\cdot,\widetilde{\theta})\big]\Big\} - \beta\rho\big[R'(\cdot,\widetilde{\theta})\big] + \log\Big\{\pi\Big[\exp\big\{-\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}),M'(\cdot,\widetilde{\theta})\big]+\beta R'(\cdot,\widetilde{\theta})\big]\big\}\Big]\Big\}.$$

Plugging this into Theorem 1.4.2 (page 52) and using the convexity of the exponential function, we see that for any posterior probability distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,

$$\beta\mathbb{P}\big\{\rho\big[R'(\cdot,\widetilde{\theta})\big]\big\} \le \lambda\mathbb{P}\big\{\rho\big[r'(\cdot,\widetilde{\theta})\big]\big\} + \mathbb{P}\big[\mathcal{K}(\rho,\pi)\big] + \log\Big\{\pi\Big[\exp\big\{-\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}),M'(\cdot,\widetilde{\theta})\big]+\beta R'(\cdot,\widetilde{\theta})\big]\big\}\Big]\Big\}.$$

We can then recall that

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

and notice moreover that

$$-\mathbb{P}\Big\{\log\Big\{\pi\Big[\exp\big[-\lambda r'(\cdot,\widetilde{\theta})\big]\Big]\Big\}\Big\} \le -\log\Big\{\pi\Big[\exp\big[-\lambda R'(\cdot,\widetilde{\theta})\big]\Big]\Big\},$$

since $R'=\mathbb{P}(r')$ and $h\mapsto\log\Big\{\pi\big[\exp(h)\big]\Big\}$ is a convex functional. Putting these two remarks together, we obtain

Theorem 1.4.4. *For any real positive parameter λ, for any prior distribution $\pi\in\mathcal{M}_+^1(\Theta)$, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}\mathbb{P}\big\{\rho\big[R'(\cdot,\widetilde{\theta})\big]\big\} &\le \frac{1}{\beta}\mathbb{P}\big[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\big] \\ &\quad+\frac{1}{\beta}\log\Big\{\pi\Big[\exp\big\{-\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}),M'(\cdot,\widetilde{\theta})\big]+\beta R'(\cdot,\widetilde{\theta})\big]\big\}\Big]\Big\} \\ &\qquad\qquad-\frac{1}{\beta}\log\Big\{\pi\Big[\exp\big[-\lambda R'(\cdot,\widetilde{\theta})\big]\Big]\Big\} \\ &\le \frac{1}{\beta}\mathbb{P}\big[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\big]\end{aligned}$$

$$+\frac{1}{\beta}\log\Big\{\pi\Big[\exp\big\{-\big[N\sinh(\tfrac{\lambda}{N})-\beta\big]R'(\cdot,\widetilde{\theta}\,)+2N\sinh(\tfrac{\lambda}{2N})^2M'(\cdot,\widetilde{\theta}\,)\big\}\Big]\Big\}-\frac{1}{\beta}\log\Big\{\pi\Big[\exp\big[-\lambda R'(\cdot,\widetilde{\theta}\,)\big]\Big]\Big\}.$$

It may be interesting to derive some more suggestive (but slightly weaker) bound in the important case when $\Theta_1=\Theta$ and $R(\widetilde{\theta})=\inf_\Theta R$. In this case, it is convenient to introduce the *expected margin function*

$$\varphi(x)=\sup_{\theta\in\Theta}M'(\theta,\widetilde{\theta})-xR'(\theta,\widetilde{\theta}),\quad x\in\mathbb{R}_+. \tag{1.23}$$

We see that φ is convex and non-negative on $\mathbb{R}_+$. Using the bound $M'(\theta,\widetilde{\theta}\,)\le xR'(\theta,\widetilde{\theta}\,)+\varphi(x)$, we obtain

$$\mathbb{P}\big\{\rho\big[R'(\cdot,\widetilde{\theta}\,)\big]\big\}\le\frac{1}{\beta}\mathbb{P}\big[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\big]+\frac{1}{\beta}\log\Big\{\pi\Big[\exp\Big\{-\big\{N\sinh(\tfrac{\lambda}{N})\big[1-x\tanh(\tfrac{\lambda}{2N})\big]-\beta\big\}R'(\cdot,\widetilde{\theta}\,)\Big\}\Big]\Big\}+\frac{N\sinh(\frac{\lambda}{N})\tanh(\frac{\lambda}{2N})}{\beta}\varphi(x)-\frac{1}{\beta}\log\Big\{\pi\Big[\exp\big[-\lambda R'(\cdot,\widetilde{\theta}\,)\big]\Big]\Big\}.$$

Let us make the change of variable $\gamma=N\sinh(\frac{\lambda}{N})\big[1-x\tanh(\frac{\lambda}{2N})\big]-\beta$ to obtain

Corollary 1.4.5. *For any real positive parameters x, γ and λ such that $x\le\tanh(\frac{\lambda}{2N})^{-1}$ and $0\le\gamma<N\sinh(\frac{\lambda}{N})\big[1-x\tanh(\frac{\lambda}{2N})\big]$,*

$$\mathbb{P}\big[\rho(R)\big]-\inf_\Theta R\le\Big\{N\sinh(\tfrac{\lambda}{N})\big[1-x\tanh(\tfrac{\lambda}{2N})\big]-\gamma\Big\}^{-1}\times\Big\{\int_\gamma^\lambda\big[\pi_{\exp(-\alpha R)}(R)-\inf_\Theta R\big]d\alpha+N\sinh\big(\tfrac{\lambda}{N}\big)\tanh\big(\tfrac{\lambda}{2N}\big)\varphi(x)+\mathbb{P}\big[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\big]\Big\}.$$

Let us remark that these results, although well suited to study Mammen and Tsybakov's margin assumptions, hold in the general case: introducing the convex *expected margin function* φ is a substitute for making hypotheses about the relations between R and D.

Using the fact that $R'(\theta,\widetilde{\theta}\,)\ge0$, $\theta\in\Theta$ and that $\varphi(x)\ge0$, $x\in\mathbb{R}_+$, we can weaken and simplify the preceding corollary even more to get

Corollary 1.4.6. *For any real parameters β, λ and x such that $x\ge0$ and $0\le\beta<\lambda-x\frac{\lambda^2}{2N}$, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\big[\rho(R)\big] \leq \inf_{\Theta} R + \Big[\lambda - x\tfrac{\lambda^2}{2N} - \beta\Big]^{-1}\bigg\{\int_\beta^\lambda \big[\pi_{\exp(-\alpha R)}(R) - \inf_{\Theta} R\big]d\alpha + \mathbb{P}\big\{\mathcal{K}\big[\rho, \pi_{\exp(-\lambda r)}\big]\big\} + \varphi(x)\frac{\lambda^2}{2N}\bigg\}.$$

Let us apply this bound under the *margin assumption* first considered by Mammen and Tsybakov (Mammen et al., 1999; Tsybakov, 2004), which says that for some real positive constant c and some real exponent $\kappa \geq 1$,

$$R'(\theta, \widetilde{\theta}) \geq cD(\theta, \widetilde{\theta})^\kappa, \qquad \theta \in \Theta. \tag{1.24}$$

In the case when $\kappa = 1$, then $\varphi(c^{-1}) = 0$, proving that

$$\begin{aligned}\mathbb{P}\big\{\pi_{\exp(-\lambda r)}\big[R'(\cdot, \widetilde{\theta}\,)\big]\big\} &\leq \frac{\int_\beta^\lambda \pi_{\exp(-\gamma R)}\big[R'(\cdot, \widetilde{\theta}\,)\big]d\gamma}{N\sinh(\frac{\lambda}{N})\big[1 - c^{-1}\tanh(\frac{\lambda}{2N})\big] - \beta} \\ &\leq \frac{\int_\beta^\lambda \pi_{\exp(-\gamma R)}\big[R'(\cdot, \widetilde{\theta}\,)\big]d\gamma}{\lambda - \frac{\lambda^2}{2cN} - \beta}.\end{aligned}$$

Taking for example $\lambda = \frac{cN}{2}$, $\beta = \frac{\lambda}{2} = \frac{cN}{4}$, we obtain

$$\begin{aligned}\mathbb{P}\big[\pi_{\exp(-2^{-1}cNr)}(R)\big] &\leq \inf R + \frac{8}{cN}\int_{\frac{cN}{4}}^{\frac{cN}{2}} \pi_{\exp(-\gamma R)}\big[R'(\cdot, \widetilde{\theta})\big]d\gamma \\ &\leq \inf R + 2\pi_{\exp(-\frac{cN}{4}R)}\big[R'(\cdot, \widetilde{\theta}\,)\big].\end{aligned}$$

If moreover the behaviour of the prior distribution π is parametric, meaning that $\pi_{\exp(-\beta R)}\big[R'(\cdot, \widetilde{\theta}\,)\big] \leq \frac{d}{\beta}$, for some positive real constant d linked with the dimension of the classification model, then

$$\mathbb{P}\big[\pi_{\exp(-\frac{cN}{2}r)}(R)\big] \leq \inf R + \frac{8\log(2)d}{cN} \leq \inf R + \frac{5.55\,d}{cN}.$$

In the case when $\kappa > 1$,

$$\varphi(x) \leq (\kappa - 1)\kappa^{-\frac{\kappa}{\kappa-1}}(cx)^{-\frac{1}{\kappa-1}} = (1 - \kappa^{-1})(\kappa cx)^{-\frac{1}{\kappa-1}},$$

thus $\mathbb{P}\big\{\pi_{\exp(-\lambda r)}\big[R'(\cdot, \widetilde{\theta}\,)\big]\big\}$

$$\leq \frac{\int_\beta^\lambda \pi_{\exp(-\gamma R)}\big[R'(\cdot, \widetilde{\theta}\,)\big]d\gamma + (1 - \kappa^{-1})(\kappa cx)^{-\frac{1}{\kappa-1}}\frac{\lambda^2}{2N}}{\lambda - \frac{x\lambda^2}{2N} - \beta}.$$

Taking for instance $\beta = \frac{\lambda}{2}$, $x = \frac{N}{2\lambda}$, and putting $b = (1 - \kappa^{-1})(c\kappa)^{-\frac{1}{\kappa-1}}$, we obtain

$$\mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big] - \inf R \leq \frac{4}{\lambda}\int_{\lambda/2}^\lambda \pi_{\exp(-\gamma R)}\big[R'(\cdot, \widetilde{\theta}\,)\big]d\gamma + b\left(\frac{2\lambda}{N}\right)^{\frac{\kappa}{\kappa-1}}.$$

In the *parametric* case when $\pi_{\exp(-\gamma R)}\big[R'(\cdot,\widetilde{\theta}\,)\big] \leq \frac{d}{\gamma}$, we get

$$\mathbb{P}\big[\pi_{\exp(-\lambda r)}(R)\big] - \inf R \leq \frac{4\log(2)d}{\lambda} + b\left(\frac{2\lambda}{N}\right)^{\frac{\kappa}{\kappa-1}}.$$

Taking

$$\overline{\lambda} = 2^{-1}\big[8\log(2)d\big]^{\frac{\kappa-1}{2\kappa-1}}(\kappa c)^{\frac{1}{2\kappa-1}}N^{\frac{\kappa}{2\kappa-1}},$$

we obtain

$$\mathbb{P}\big[\pi_{\exp(-\overline{\lambda} r)}(R)\big] - \inf R \leq (2-\kappa^{-1})(\kappa c)^{-\frac{1}{2\kappa-1}}\left(\frac{8\log(2)d}{N}\right)^{\frac{\kappa}{2\kappa-1}}.$$

We see that this formula coincides with the result for $\kappa = 1$. We can thus reduce the two cases to a single one and state

Corollary 1.4.7. *Let us assume that for some $\widetilde{\theta} \in \Theta$, some positive real constant c, some real exponent $\kappa \geq 1$ and for any $\theta \in \Theta$, $R(\theta) \geq R(\widetilde{\theta}) + cD(\theta,\widetilde{\theta})^{\kappa}$. Let us also assume that for some positive real constant d and any positive real parameter γ, $\pi_{\exp(-\gamma R)}(R) - \inf R \leq \frac{d}{\gamma}$. Then*

$$\mathbb{P}\Big[\pi_{\exp\left\{-2^{-1}[8\log(2)d]^{\frac{\kappa-1}{2\kappa-1}}(\kappa c)^{\frac{1}{2\kappa-1}}N^{\frac{\kappa}{2\kappa-1}}r\right\}}(R)\Big]$$
$$\leq \inf R + (2-\kappa^{-1})(\kappa c)^{-\frac{1}{2\kappa-1}}\left(\frac{8\log(2)d}{N}\right)^{\frac{\kappa}{2\kappa-1}}.$$

Let us remark that the exponent of N in this corollary is known to be the minimax exponent under these assumptions: it is unimprovable, whatever estimator is used in place of the Gibbs posterior shown here (at least in the worst case compatible with the hypotheses). The interest of the corollary is to show not only the minimax exponent in N, but also an explicit non-asymptotic bound with reasonable and simple constants. It is also clear that we could have got slightly better constants if we had kept the full strength of Theorem 1.4.4 (page 54) instead of using the weaker Corollary 1.4.6 (page 55).

We will prove in the following empirical bounds showing how the constant λ can be estimated from the data instead of being chosen according to some margin and complexity assumptions.

1.4.3. Unbiased empirical bounds. We are going to define an empirical counterpart for the *expected margin function* φ. It will appear in empirical bounds having otherwise the same structure as the non-random bound we just proved. Anyhow, we will not launch into trying to compare the behaviour of our proposed *empirical margin function* with the *expected margin function*, since the margin function involves taking a supremum which is not straightforward to handle. When we will touch the issue of building *provably* adaptive estimators, we will instead formulate another type of bounds

based on integrated quantities, rather than try to analyse the properties of the empirical margin function.

Let us start as in the previous subsection with the inequality

$$\beta\mathbb{P}\Big\{\rho\big[R'(\cdot,\widetilde{\theta}\,)\big]\Big\} \leq \mathbb{P}\Big\{\lambda\rho\big[r'(\cdot,\widetilde{\theta}\,)\big] + \mathcal{K}(\rho,\pi)\Big\} + \log\Big\{\pi\Big[\exp\big\{-\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}\,),M'(\cdot,\widetilde{\theta}\,)\big] + \beta R'(\cdot,\widetilde{\theta}\,)\big\}\Big]\Big\}.$$

We have already defined by equation (1.22, page 53) the empirical pseudo-distance

$$m'(\theta,\widetilde{\theta}\,) = \frac{1}{N}\sum_{i=1}^{N}\psi_i(\theta,\widetilde{\theta}\,)^2.$$

Recalling that $\mathbb{P}\big[m'(\theta,\widetilde{\theta}\,)\big] = M'(\theta,\widetilde{\theta}\,)$, and using the convexity of $h \mapsto \log\Big\{\pi\big[\exp(h)\big]\Big\}$, leads to the following inequalities:

$$\begin{aligned}
&\log\Big\{\pi\Big[\exp\big\{-\lambda\Psi_{\frac{\lambda}{N}}\big[R'(\cdot,\widetilde{\theta}\,),M'(\cdot,\widetilde{\theta}\,)\big] + \beta R'(\cdot,\widetilde{\theta}\,)\big\}\Big]\Big\}\\
&\quad\leq \log\Big\{\pi\Big[\exp\big\{-N\sinh(\tfrac{\lambda}{N})R'(\cdot,\widetilde{\theta}\,)\\
&\qquad\qquad + N\sinh(\tfrac{\lambda}{N})\tanh(\tfrac{\lambda}{2N})M'(\cdot,\widetilde{\theta}\,) + \beta R'(\cdot,\widetilde{\theta}\,)\big]\big\}\Big]\Big\}\\
&\quad\leq \mathbb{P}\Big\{\log\Big\{\pi\Big[\exp\big\{-\big[N\sinh(\tfrac{\lambda}{N}) - \beta\big]r'(\cdot,\widetilde{\theta}\,)\\
&\qquad\qquad + N\sinh(\tfrac{\lambda}{N})\tanh(\tfrac{\lambda}{2N})m'(\cdot,\widetilde{\theta}\,)\big\}\Big]\Big\}\Big\}.
\end{aligned}$$

We may moreover remark that

$$\begin{aligned}
\lambda\rho\big[r'(\cdot,\widetilde{\theta}\,)\big] + \mathcal{K}(\rho,\pi) &= \big[\beta - N\sinh(\tfrac{\lambda}{N}) + \lambda\big]\rho\big[r'(\cdot,\widetilde{\theta}\,)\big]\\
&\quad + \mathcal{K}\big[\rho,\pi_{\exp\{-[N\sinh(\frac{\lambda}{N})-\beta]r\}}\big]\\
&\quad - \log\Big\{\pi\Big[\exp\big\{-\big[N\sinh(\tfrac{\lambda}{N}) - \beta\big]r'(\cdot,\widetilde{\theta}\,)\big\}\Big]\Big\}.
\end{aligned}$$

This establishes

Theorem 1.4.8. *For any positive real parameters β and λ, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
\mathbb{P}\big\{\rho\big[R'(\cdot,\widetilde{\theta}\,)\big]\big\} \leq \mathbb{P}\Big\{&\Big[1 - \frac{N\sinh(\frac{\lambda}{N}) - \lambda}{\beta}\Big]\rho\big[r'(\cdot,\widetilde{\theta}\,)\big]\\
&+ \frac{\mathcal{K}\big[\rho,\pi_{\exp\{-[N\sinh(\frac{\lambda}{N})-\beta]r\}}\big]}{\beta}\\
&+ \beta^{-1}\log\Big\{\pi_{\exp\{-[N\sinh(\frac{\lambda}{N})-\beta]r\}}\Big[\exp\big[N\sinh(\tfrac{\lambda}{N})\tanh(\tfrac{\lambda}{2N})m'(\cdot,\widetilde{\theta}\,)\big]\Big]\Big\}\Big\}.
\end{aligned}$$

Taking $\beta = \frac{N}{2}\sinh(\frac{\lambda}{N})$, using the fact that $\sinh(a) \geq a$, $a \geq 0$ and expressing $\tanh(\frac{a}{2}) = a^{-1}\big[\sqrt{1+\sinh(a)^2} - 1\big]$ and $a = \log\big[\sqrt{1+\sinh(a)^2} + \sinh(a)\big]$, we deduce

Corollary 1.4.9. *For any positive real constant β and any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\big\{\rho\big[R'(\cdot,\widetilde{\theta}\,)\big]\big\} \leq \mathbb{P}\Bigg\{\underbrace{\bigg[\frac{N}{\beta}\log\Big(\sqrt{1+\tfrac{4\beta^2}{N^2}}+\tfrac{2\beta}{N}\Big)-1\bigg]}_{\leq 1}\rho\big[r'(\cdot,\widetilde{\theta}\,)\big] + \frac{1}{\beta}\Bigg\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big] + \log\bigg[\pi_{\exp(-\beta r)}\Big\{\exp\Big[N\Big(\sqrt{1+\tfrac{4\beta^2}{N^2}}-1\Big)m'(\cdot,\widetilde{\theta}\,)\Big]\Big\}\bigg]\Bigg\}\Bigg\}.$$

This theorem and its corollary are really analogous to Theorem 1.4.4 (page 54), and it could easily be proved that under Mammen and Tsybakov margin assumptions we obtain an upper bound of the same order as Corollary 1.4.7 (page 57). Anyhow, in order to obtain an empirical bound, we are now going to take a supremum over all possible values of $\widetilde{\theta}$, that is over Θ_1. Although we believe that taking this supremum will not spoil the bound in cases when over-fitting remains under control, we will not try to investigate precisely if and when this is actually true, and provide our empirical bound as such. Let us say only that on qualitative grounds, the values of the margin function quantify the steepness of the contrast function R or its empirical counterpart r, and that the definition of the empirical margin function is obtained by substituting $\mathbb{P}$, the true sample distribution, with $\overline{\mathbb{P}} = \big(\frac{1}{N}\sum_{i=1}^N \delta_{(X_i,Y_i)}\big)^{\otimes N}$, the empirical sample distribution, in the definition of the expected margin function. Therefore, on qualitative grounds, it seems hopeless to presume that R is steep when r is not, or in other words that a classification model that would be inefficient at estimating a bootstrapped sample according to our non-random bound would be by some miracle efficient at estimating the true sample distribution according to the same bound. To this extent, we feel that our empirical bounds bring a satisfactory counterpart of our non-random bounds. Anyhow, we will also produce estimators which can be proved to be adaptive using PAC-Bayesian tools in the next section, at the price of a more sophisticated construction involving comparisons between a posterior distribution and a Gibbs prior distribution or between two posterior distributions.

Let us now restrict discussion to the important case when $\widetilde{\theta} \in \arg\min_{\Theta_1} R$. To obtain an observable bound, let $\widehat{\theta} \in \arg\min_{\theta\in\Theta} r(\theta)$ and let us introduce the *empirical margin functions*

$$\overline{\varphi}(x) = \sup_{\theta\in\Theta} m'(\theta,\widehat{\theta}) - x\big[r(\theta) - r(\widehat{\theta})\big], \quad x \in \mathbb{R}_+,$$

$$\widetilde{\varphi}(x) = \sup_{\theta\in\Theta_1} m'(\theta,\widehat{\theta}) - x\big[r(\theta) - r(\widehat{\theta})\big], \quad x \in \mathbb{R}_+.$$

Using the fact that $m'(\theta,\widetilde{\theta}) \le m'(\theta,\widehat{\theta}) + m'(\widehat{\theta},\widetilde{\theta})$, we get

COROLLARY 1.4.10. *For any positive real parameters β and λ, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
\mathbb{P}\big[\rho(R)\big] - \inf_{\Theta_1} R \le \mathbb{P}\Bigg\{ & \Big[1 - \tfrac{N\sinh(\frac{\lambda}{N}) - \lambda}{\beta}\Big]\big[\rho(r) - r(\widehat{\theta})\big] \\
& + \frac{\mathcal{K}\big[\rho, \pi_{\exp\{-[N\sinh(\frac{\lambda}{N})-\beta]r\}}\big]}{\beta} \\
& + \beta^{-1}\log\Big\{\pi_{\exp\{-[N\sinh(\frac{\lambda}{N})-\beta]r\}}\Big[\exp\big[N\sinh\big(\tfrac{\lambda}{N}\big)\tanh\big(\tfrac{\lambda}{2N}\big)m'(\cdot,\widehat{\theta})\big]\Big]\Big\} \\
+\beta^{-1}N\sinh(\tfrac{\lambda}{N})\tanh(\tfrac{\lambda}{2N})\widetilde{\varphi}\Bigg[& \frac{\beta}{N\sinh(\frac{\lambda}{N})\tanh(\frac{\lambda}{2N})}\left(1 - \frac{N\sinh(\frac{\lambda}{N}) - \lambda}{\beta}\right)\Bigg]\Bigg\}.
\end{aligned}$$

Taking $\beta = \frac{N}{2}\sinh(\frac{\lambda}{N})$, we also obtain

$$\begin{aligned}
\mathbb{P}\big[\rho(R)\big] - \inf_{\Theta_1} R \le \mathbb{P}\Bigg\{ & \underbrace{\Big[\tfrac{N}{\beta}\log\Big(\sqrt{1+\tfrac{4\beta^2}{N^2}} + \tfrac{2\beta}{N}\Big) - 1\Big]}_{\le 1}\big[\rho(r) - r(\widehat{\theta})\big] \\
& + \frac{1}{\beta}\Bigg\{\mathcal{K}\big[\rho, \pi_{\exp(-\beta r)}\big] \\
& \quad + \log\Big[\pi_{\exp(-\beta r)}\Big\{\exp\Big[N\Big(\sqrt{1+\tfrac{4\beta^2}{N^2}} - 1\Big)m'(\cdot,\widehat{\theta})\Big]\Big\}\Big]\Bigg\} \\
& \quad + \frac{N}{\beta}\Big(\sqrt{1+\tfrac{4\beta^2}{N^2}} - 1\Big)\widetilde{\varphi}\Bigg[\frac{\log\Big(\sqrt{1+\tfrac{4\beta^2}{N^2}} + \tfrac{2\beta}{N}\Big) - \tfrac{\beta}{N}}{\Big(\sqrt{1+\tfrac{4\beta^2}{N^2}} - 1\Big)}\Bigg]\Bigg\}.
\end{aligned}$$

Note that we could also use the upper bound $m'(\theta,\widehat{\theta}) \le x\big[r(\theta) - r(\widehat{\theta})\big] + \overline{\varphi}(x)$ and put $\alpha = N\sinh(\frac{\lambda}{N})\big[1 - x\tanh(\frac{\lambda}{2N})\big] - \beta$, to obtain

COROLLARY 1.4.11. *For any non-negative real parameters x, α and λ, such that $\alpha < N\sinh(\frac{\lambda}{N})\big[1 - x\tanh(\frac{\lambda}{2N})\big]$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
&\mathbb{P}\big[\rho(R)\big] - \inf_{\Theta_1} R \\
&\quad \le \mathbb{P}\Bigg\{\Bigg[1 - \frac{N\sinh(\frac{\lambda}{N})\big[1 - x\tanh(\frac{\lambda}{2N})\big] - \lambda}{N\sinh(\frac{\lambda}{N})\big[1 - x\tanh(\frac{\lambda}{2N})\big] - \alpha}\Bigg]\big[\rho(r) - r(\widehat{\theta})\big] \\
&\qquad + \frac{\mathcal{K}\big[\rho, \pi_{\exp(-\alpha r)}\big]}{N\sinh(\frac{\lambda}{N})\big[1 - x\tanh(\frac{\lambda}{2N})\big] - \alpha}
\end{aligned}$$

$$+\frac{N\sinh(\frac{\lambda}{N})\tanh(\frac{\lambda}{2N})}{N\sinh(\frac{\lambda}{N})\big[1-x\tanh(\frac{\lambda}{2N})\big]-\alpha}\times\left[\overline{\varphi}(x)+\widetilde{\varphi}\left(\frac{\lambda-\alpha}{N\sinh(\frac{\lambda}{N})\tanh(\frac{\lambda}{2N})}\right)\right]\Bigg\}.$$

Let us notice that in the case when $\Theta_1=\Theta$, the upper bound provided by this corollary has the same general form as the upper bound provided by Corollary 1.4.5 (page 55), with the sample distribution $\mathbb{P}$ replaced with the empirical distribution of the sample $\overline{\mathbb{P}}=\big(\frac{1}{N}\sum_{i=1}^N\delta_{(X_i,Y_i)}\big)^{\otimes N}$. Therefore, our empirical bound can be of a larger order of magnitude than our non-random bound only in the case when our non-random bound applied to the bootstrapped sample distribution $\overline{\mathbb{P}}$ would be of a larger order of magnitude than when applied to the true sample distribution $\mathbb{P}$. In other words, we can say that our empirical bound is close to our non-random bound in every situation where the bootstrapped sample distribution $\overline{\mathbb{P}}$ is not harder to bound than the true sample distribution $\mathbb{P}$. Although this does not prove that our empirical bound is always of the same order as our non-random bound, this is a good qualitative hint that this will be the case in most practical situations of interest, since in situations of "under-fitting", if they exist, it is likely that the choice of the classification model is inappropriate to the data and should be modified.

Another reassuring remark is that the empirical margin functions $\overline{\varphi}$ and $\widetilde{\varphi}$ behave well in the case when $\inf_\Theta r=0$. Indeed in this case $m'(\theta,\widehat{\theta})=r'(\theta,\widehat{\theta})=r(\theta)$, $\theta\in\Theta$, and thus $\overline{\varphi}(1)=\widetilde{\varphi}(1)=0$, and

$$\widetilde{\varphi}(x)\leq-(x-1)\inf_{\Theta_1}r,\ x\geq1.$$

This shows that in this case we recover the same accuracy as with non-relative local empirical bounds. Thus the bound of Corollary 1.4.11 does not collapse in presence of massive over-fitting in the larger model, causing $r(\widehat{\theta})=0$, which is another hint that this may be an accurate bound in many situations.

1.4.4. Relative empirical deviation bounds. It is natural to make use of Theorem 1.4.3 (page 53) to obtain empirical deviation bounds, since this theorem provides an empirical variance term.

Theorem 1.4.3 is written in a way which exploits the fact that ψ_i takes only the three values -1, 0 and $+1$. However, it will be more convenient for the following computations to use it in its more general form, which only makes use of the fact that $\psi_i\in\ (-1,1)$. With notation to be explained hereafter, it can indeed also be written as

$$\mathbb{P}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\left\{-N\rho\Big\{\log\Big[1-\lambda P(\psi)\Big]\Big\}+N\rho\Big\{\overline{P}\Big[\log(1-\lambda\psi)\Big]\Big\}-\mathcal{K}(\rho,\pi)\right\}\right]\right\}\leq1. \quad (1.25)$$

We have used the following notation in this inequality. We have put

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

so that $\overline{P}$ is our notation for the empirical distribution of the process $(X_i, Y_i)_{i=1}^N$. Moreover we have also used

$$P = \mathbb{P}(\overline{P}) = \frac{1}{N}\sum_{i=1}^{N} P_i,$$

where it should be remembered that the joint distribution of the process $(X_i, Y_i)_{i=1}^N$ is $\mathbb{P} = \bigotimes_{i=1}^N P_i$. We have considered $\psi(\theta, \widetilde{\theta})$ as a function defined on $\mathcal{X} \times \mathcal{Y}$ as $\psi(\theta, \widetilde{\theta})(x, y) = \mathbb{1}\big[y \neq f_\theta(x)\big] - \mathbb{1}\big[y \neq f_{\widetilde{\theta}}(x)\big]$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$ so that it should be understood that

$$\begin{aligned} P(\psi) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}\big[\psi_i(\theta, \widetilde{\theta})\big] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}\Big\{\mathbb{1}\big[Y_i \neq f_\theta(X_i)\big] - \mathbb{1}\big[Y_i \neq f_{\widetilde{\theta}}(X_i)\big]\Big\} = R'(\theta, \widetilde{\theta}). \end{aligned}$$

In the same way

$$\overline{P}\Big[\log(1 - \lambda\psi)\Big] = \frac{1}{N}\sum_{i=1}^{N} \log\big[1 - \lambda\psi_i(\theta, \widetilde{\theta})\big].$$

Moreover integration with respect to ρ bears on the index θ, so that

$$\begin{aligned} \rho\Big\{\log\Big[1 - \lambda P(\psi)\Big]\Big\} &= \int_{\theta\in\Theta} \log\Big\{1 - \frac{\lambda}{N}\sum_{i=1}^{N} \mathbb{P}\big[\psi_i(\theta, \widetilde{\theta})\big]\Big\}\rho(d\theta), \\ \rho\Big\{\overline{P}\Big[\log(1 - \lambda\psi)\Big]\Big\} &= \int_{\theta\in\Theta} \Big\{\frac{1}{N}\sum_{i=1}^{N} \log\big[1 - \lambda\psi_i(\theta, \widetilde{\theta})\big]\Big\}\rho(d\theta). \end{aligned}$$

We have chosen concise notation, as we did throughout these notes, in order to make the computations easier to follow.

To get an alternate version of empirical relative deviation bounds, we need to find some convenient way to localize the choice of the prior distribution π in equation (1.25, page 61). Here we propose replacing π with $\mu = \pi_{\exp\{-N\log[1+\beta P(\psi)]\}}$, which can also be written $\pi_{\exp\{-N\log[1+\beta R'(\cdot,\widetilde{\theta})]\}}$. Indeed we see that

$$\begin{aligned} \mathcal{K}(\rho, \mu) = N\rho\Big\{\log\big[1 + \beta P(\psi)\big]\Big\} + \mathcal{K}(\rho, \pi) \\ + \log\Big\{\pi\Big[\exp\big\{-N\log\big[1 + \beta P(\psi)\big]\big\}\Big]\Big\}. \end{aligned}$$

Moreover, we deduce from our deviation inequality applied to $-\psi$, that (as long as $\beta > -1$),

$$\mathbb{P}\Big\{\exp\Big[N\mu\Big\{\overline{P}\big[\log(1+\beta\psi)\big]\Big\} - N\mu\Big\{\log\big[1+\beta P(\psi)\big]\Big\}\Big]\Big\} \leq 1.$$

Thus

$$\begin{aligned}
\mathbb{P}\Big\{\exp\Big[\log\Big\{\pi\Big[\exp\big\{-N\log\big[1+\beta P(\psi)\big]\big\}\Big]\Big\} \\
- \log\Big\{\pi\Big[\exp\big\{-N\overline{P}\big[\log(1+\beta\psi)\big]\big\}\Big]\Big\}\Big]\Big\} \\
\leq \mathbb{P}\Big\{\exp\Big[-N\mu\Big\{\log\big[1+\beta P(\psi)\big]\Big\} - \mathcal{K}(\mu,\pi) \\
+ N\mu\Big\{\overline{P}\big[\log(1+\beta\psi)\big]\Big\} + \mathcal{K}(\mu,\pi)\Big]\Big\} \leq 1.
\end{aligned}$$

This can be used to handle $\mathcal{K}(\rho,\mu)$, making use of the Cauchy–Schwarz inequality as follows

$$\begin{aligned}
&\mathbb{P}\Big\{\exp\Big[\frac{1}{2}\Big[-N\log\Big\{\Big(1-\lambda\rho\big[P(\psi)\big]\Big)\Big(1+\beta\rho\big[P(\psi)\big]\Big)\Big\} \\
&\qquad +N\rho\Big\{\overline{P}\Big[\log(1-\lambda\psi)\Big]\Big\} \\
&\qquad - \mathcal{K}(\rho,\pi) - \log\Big\{\pi\Big[\exp\big\{-N\overline{P}\big[\log(1+\beta\psi)\big]\big\}\Big]\Big\}\Big]\Big]\Big\} \\
&\leq \mathbb{P}\Big\{\exp\Big[-N\log\Big\{\Big(1-\lambda\rho\big[P(\psi)\big]\Big)\Big\} \\
&\qquad + N\rho\Big\{\overline{P}\Big[\log(1-\lambda\psi)\Big]\Big\} - \mathcal{K}(\rho,\mu)\Big]\Big\}^{1/2} \\
&\times \mathbb{P}\Big\{\exp\Big[\log\Big\{\pi\Big[\exp\big\{-N\log\big[1+\beta P(\psi)\big]\big\}\Big]\Big\} \\
&\qquad - \log\Big\{\pi\Big[\exp\big\{-N\overline{P}\big[\log(1+\beta\psi)\big]\big\}\Big]\Big\}\Big]\Big\}^{1/2} \leq 1.
\end{aligned}$$

This implies that with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\begin{aligned}
&- N\log\Big\{\Big(1-\lambda\rho\big[P(\psi)\big]\Big)\Big(1+\beta\rho\big[P(\psi)\big]\Big)\Big\} \\
&\quad \leq -N\rho\Big\{\overline{P}\Big[\log(1-\lambda\psi)\Big]\Big\} \\
&\qquad + \mathcal{K}(\rho,\pi) + \log\Big\{\pi\Big[\exp\big\{-N\overline{P}\big[\log(1+\beta\psi)\big]\big\}\Big]\Big\} - 2\log(\epsilon).
\end{aligned}$$

It is now convenient to remember that

$$\overline{P}\Big[\log(1-\lambda\psi)\Big] = \frac{1}{2}\log\left(\frac{1-\lambda}{1+\lambda}\right) r'(\theta,\widetilde{\theta}) + \frac{1}{2}\log(1-\lambda^2) m'(\theta,\widetilde{\theta}).$$

We thus can write the previous inequality as

$$\begin{aligned}
&- N\log\Big\{\Big(1-\lambda\rho\big[R'(\cdot,\widetilde{\theta})\big]\Big)\Big(1+\beta\rho\big[R'(\cdot,\widetilde{\theta})\big]\Big)\Big\} \\
&\quad\le \frac{N}{2}\log\left(\frac{1+\lambda}{1-\lambda}\right)\rho\big[r'(\cdot,\widetilde{\theta})\big] - \frac{N}{2}\log(1-\lambda^2)\rho\big[m'(\cdot,\widetilde{\theta})\big] + \mathcal{K}(\rho,\pi) \\
&\qquad\qquad + \log\Big\{\pi\Big[\exp\Big\{-\frac{N}{2}\log\Big(\frac{1+\beta}{1-\beta}\Big) r'(\cdot,\widetilde{\theta}) \\
&\qquad\qquad\qquad\qquad - \frac{N}{2}\log(1-\beta^2) m'(\cdot,\widetilde{\theta})\Big\}\Big]\Big\} - 2\log(\epsilon).
\end{aligned}$$

Let us assume now that $\widetilde{\theta} \in \arg\min_{\Theta_1} R$. Let us introduce $\widehat{\theta} \in \arg\min_{\Theta} r$. Decomposing $r'(\theta,\widetilde{\theta}) = r'(\theta,\widehat{\theta}) + r'(\widehat{\theta},\widetilde{\theta})$ and considering that

$$m'(\theta,\widetilde{\theta}) \le m'(\theta,\widehat{\theta}) + m'(\widehat{\theta},\widetilde{\theta}),$$

we see that with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}
&- N\log\Big\{\Big(1-\lambda\rho\big[R'(\cdot,\widetilde{\theta})\big]\Big)\Big(1+\beta\rho\big[R'(\cdot,\widetilde{\theta})\big]\Big)\Big\} \\
&\quad\le \frac{N}{2}\log\Big(\frac{1+\lambda}{1-\lambda}\Big)\rho\big[r'(\cdot,\widehat{\theta})\big] - \frac{N}{2}\log(1-\lambda^2)\rho\big[m'(\cdot,\widehat{\theta})\big] + \mathcal{K}(\rho,\pi) \\
&\quad + \log\Big\{\pi\Big[\exp\Big\{-\tfrac{N}{2}\log\Big(\tfrac{1+\beta}{1-\beta}\Big)\big[r'(\cdot,\widehat{\theta}\,)\big] - \tfrac{N}{2}\log(1-\beta^2) m'(\cdot,\widehat{\theta}\,)\Big\}\Big]\Big\} \\
&\qquad\qquad + \tfrac{N}{2}\log\Big[\tfrac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\Big]\big[r(\widehat{\theta}\,) - r(\widetilde{\theta})\big] \\
&\qquad\qquad\qquad - \tfrac{N}{2}\log\big[(1-\lambda^2)(1-\beta^2)\big] m'(\widehat{\theta}\,,\widetilde{\theta}) - 2\log(\epsilon).
\end{aligned}$$

Let us now define for simplicity the posterior $\nu : \Omega \to \mathcal{M}_+^1(\Theta)$ by the identity

$$\frac{d\nu}{d\pi}(\theta) = \frac{\exp\Big\{-\frac{N}{2}\log\Big(\frac{1+\lambda}{1-\lambda}\Big) r'(\theta,\widehat{\theta}) + \frac{N}{2}\log(1-\lambda^2) m'(\theta,\widehat{\theta})\Big\}}{\pi\Big[\exp\Big\{-\frac{N}{2}\log\Big(\frac{1+\lambda}{1-\lambda}\Big) r'(\cdot,\widehat{\theta}) + \frac{N}{2}\log(1-\lambda^2) m'(\cdot,\widehat{\theta})\Big\}\Big]}.$$

Let us also introduce the random bound

$$\begin{aligned}
B = \frac{1}{N}\log\Big\{\nu\Big[\exp\Big[\tfrac{N}{2}\log\Big[\tfrac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\Big] r'(\cdot,\widehat{\theta}) \\
- \tfrac{N}{2}\log\big[(1-\lambda^2)(1-\beta^2)\big] m'(\cdot,\widehat{\theta}\,)\Big]\Big]\Big\}
\end{aligned}$$

$$+ \sup_{\theta\in\Theta_1} \frac{1}{2}\log\Big[\tfrac{(1-\lambda)(1+\beta)}{(1+\lambda)(1-\beta)}\Big] r'(\theta,\widehat{\theta}\,) - \frac{1}{2}\log\big[(1-\lambda^2)(1-\beta^2)\big] m'(\theta,\widehat{\theta}\,) - \frac{2}{N}\log(\epsilon).$$

THEOREM 1.4.12. *Using the above notation, for any real constants* $0 \le \beta < \lambda < 1$, *for any prior distribution* $\pi \in \mathcal{M}_+^1(\Theta)$, *for any subset* $\Theta_1 \subset \Theta$, *with* $\mathbb{P}$ *probability at least* $1-\epsilon$, *for any posterior distribution* $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$-\log\Big\{\Big(1-\lambda\big[\rho(R)-\inf_{\Theta_1} R\big]\Big)\Big(1+\beta\big[\rho(R)-\inf_{\Theta_1} R\big]\Big)\Big\} \le \frac{\mathcal{K}(\rho,\nu)}{N} + B.$$

Therefore,

$$\rho(R) - \inf_{\Theta_1} R \le \frac{\lambda-\beta}{2\lambda\beta}\left(\sqrt{1+4\frac{\lambda\beta}{(\lambda-\beta)^2}\left[1-\exp\left(-B-\frac{\mathcal{K}(\rho,\nu)}{N}\right)\right]}-1\right) \le \frac{1}{\lambda-\beta}\left(B+\frac{\mathcal{K}(\rho,\nu)}{N}\right).$$

Let us define the posterior $\widehat{\nu}$ by the identity

$$\frac{d\widehat{\nu}}{d\pi}(\theta) = \frac{\exp\Big[-\frac{N}{2}\log\Big(\frac{1+\beta}{1-\beta}\Big) r'(\theta,\widehat{\theta}) - \frac{N}{2}\log(1-\beta^2) m'(\theta,\widehat{\theta})\Big]}{\pi\Big\{\exp\Big[-\frac{N}{2}\log\Big(\frac{1+\beta}{1-\beta}\Big) r'(\cdot,\widehat{\theta}) - \frac{N}{2}\log(1-\beta^2) m'(\cdot,\widehat{\theta})\Big]\Big\}}.$$

It is useful to remark that

$$\frac{1}{N}\log\Big\{\nu\Big[\exp\Big[\frac{N}{2}\log\Big(\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\Big) r'(\cdot,\widehat{\theta}) - \frac{N}{2}\log\big[(1-\lambda^2)(1-\beta^2)\big] m'(\cdot,\widehat{\theta})\Big]\Big]\Big\}$$

$$\le \widehat{\nu}\Big\{\frac{1}{2}\log\Big(\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\Big) r'(\cdot,\widehat{\theta}) - \frac{1}{2}\log\big[(1-\lambda^2)(1-\beta^2)\big] m'(\cdot,\widehat{\theta})\Big\}.$$

This inequality is a special case of

$$\log\Big\{\pi\big[\exp(g)\big]\Big\} - \log\Big\{\pi\big[\exp(h)\big]\Big\} = \int_{\alpha=0}^{1} \pi_{\exp[h+\alpha(g-h)]}(g-h)\,d\alpha \le \pi_{\exp(g)}(g-h),$$

which is a consequence of the convexity of $\alpha \mapsto \log\Big\{\pi\Big[\exp\big[h+\alpha(g-h)\big]\Big]\Big\}$.

Let us introduce as previously $\overline{\varphi}(x) = \sup_{\theta\in\Theta} m'(\theta,\widehat{\theta}) - x\, r'(\theta,\widehat{\theta})$, $x \in \mathbb{R}_+$. Let us moreover consider $\widetilde{\varphi}(x) = \sup_{\theta\in\Theta_1} m'(\theta,\widehat{\theta}) - x\, r'(\theta,\widehat{\theta})$, $x \in \mathbb{R}_+$. These functions can be used to produce a result which is slightly weaker, but maybe easier to read and understand. Indeed, we see that, for any $x \in \mathbb{R}_+$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution ρ,

$$
\begin{aligned}
&-N\log\Big\{\Big(1-\lambda\rho\big[R'(\cdot,\widetilde{\theta})\big]\Big)\Big(1+\beta\rho\big[R'(\cdot,\widetilde{\theta})\big]\Big)\Big\}\\
&\quad\le \frac{N}{2}\log\left[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}\right]\rho\big[r'(\cdot,\widehat{\theta})\big]\\
&\qquad -\frac{N}{2}\log\big[(1-\lambda^2)(1-\beta^2)\big]\overline{\varphi}(x)+\mathcal{K}(\rho,\pi)\\
&\qquad +\log\left\{\pi\left[\exp\left\{-\tfrac{N}{2}\log\left[\tfrac{(1+\beta)}{(1-\beta)(1-\beta^2)^x}\right]r'(\cdot,\widehat{\theta})\right\}\right]\right\}\\
&\qquad -\frac{N}{2}\log\big[(1-\lambda^2)(1-\beta^2)\big]\widetilde{\varphi}\left(\frac{\log\left[\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\right]}{-\log\left[(1-\lambda^2)(1-\beta^2)\right]}\right)\\
&\qquad\qquad -2\log(\epsilon)\\
&\quad=\int_{\frac{N}{2}\log\left[\frac{(1+\beta)}{(1-\beta)(1-\beta^2)^x}\right]}^{\frac{N}{2}\log\left[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}\right]}\pi_{\exp(-\alpha r)}\big[r'(\cdot,\widehat{\theta})\big]d\alpha\\
&\qquad\qquad +\mathcal{K}(\rho,\pi_{\exp\{-\frac{N}{2}\log[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}]r\}})-2\log(\epsilon)\\
&\qquad -\frac{N}{2}\log\big[(1-\lambda^2)(1-\beta^2)\big]\left[\overline{\varphi}(x)+\widetilde{\varphi}\left(\frac{\log\left[\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\right]}{-\log[(1-\lambda^2)(1-\beta^2)]}\right)\right].
\end{aligned}
$$

Theorem 1.4.13. *With the previous notation, for any real constants $0 \le \beta < \lambda < 1$, for any positive real constant x, for any prior probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$, for any subset $\Theta_1 \subset \Theta$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, putting*

$$
\begin{aligned}
B(\rho) &= \frac{1}{N(\lambda-\beta)}\int_{\frac{N}{2}\log\left[\frac{(1+\beta)}{(1-\beta)(1-\beta^2)^x}\right]}^{\frac{N}{2}\log\left[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}\right]}\pi_{\exp(-\alpha r)}\big[r'(\cdot,\widehat{\theta})\big]d\alpha\\
&\qquad +\frac{\mathcal{K}(\rho,\pi_{\exp\{-\frac{N}{2}\log[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}]r\}})-2\log(\epsilon)}{N(\lambda-\beta)}\\
&\quad -\frac{1}{2(\lambda-\beta)}\log\big[(1-\lambda^2)(1-\beta^2)\big]\left[\overline{\varphi}(x)+\widetilde{\varphi}\left(\frac{\log\left[\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\right]}{-\log[(1-\lambda^2)(1-\beta^2)]}\right)\right]\\
&\le \frac{1}{N(\lambda-\beta)}d_e\log\left(\frac{\log\left[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}\right]}{\log\left(\frac{(1+\beta)}{(1-\beta)(1-\beta^2)^x}\right)}\right)
\end{aligned}
$$

$$+\frac{\mathcal{K}(\rho,\pi_{\exp\{-\frac{N}{2}\log[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}]r\}})-2\log(\epsilon)}{N(\lambda-\beta)}$$

$$-\frac{1}{2(\lambda-\beta)}\log\big[(1-\lambda^2)(1-\beta^2)\big]\left[\overline{\varphi}(x)+\widetilde{\varphi}\left(\frac{\log\left[\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\right]}{-\log[(1-\lambda^2)(1-\beta^2)]}\right)\right],$$

the following bounds hold true:

$$\rho(R)-\inf_{\Theta_1}R$$
$$\leq\frac{\lambda-\beta}{2\lambda\beta}\left(\sqrt{1+\frac{4\lambda\beta}{(\lambda-\beta)^2}\Big\{1-\exp\big[-(\lambda-\beta)B(\rho)\big]\Big\}}-1\right)$$
$$\leq B(\rho).$$

Let us remark that this alternative way of handling relative deviation bounds made it possible to carry on with non-linear bounds up to the final result. For instance, if $\lambda=0.5$, $\beta=0.2$ and $B(\rho)=0.1$, the non-linear bound gives $\rho(R)-\inf_{\Theta_1}R\leq 0.096$.

Chapter 2

Comparing posterior distributions to Gibbs priors

2.1. Bounds relative to a Gibbs distribution

We now come to an approach to relative bounds whose performance can be analysed with PAC-Bayesian tools.

The empirical bounds at the end of the previous chapter involve taking suprema in $\theta \in \Theta$, and replacing the *expected margin function* φ with some empirical counterparts $\overline{\varphi}$ or $\widetilde{\varphi}$, which may prove unsafe when using very complex classification models.

We are now going to focus on the control of the divergence $\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big]$. It is already obvious, we hope, that controlling this divergence is the crux of the matter, and that it is a way to upper bound the mutual information between the training sample and the parameter, which can be expressed as $\mathcal{K}\big[\rho, \mathbb{P}(\rho)\big] = \mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big] - \mathcal{K}\big[\mathbb{P}(\rho), \pi_{\exp(-\beta R)}\big]$, as explained on page 28.

Through the identity

$$\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big] = \beta\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] + \mathcal{K}(\rho, \pi) - \mathcal{K}\big[\pi_{\exp(-\beta R)}, \pi\big], \quad (2.1)$$

we see that the control of this divergence is related to the control of the difference $\rho(R) - \pi_{\exp(-\beta R)}(R)$. This is the route we will follow first.

Thus comparing any posterior distribution with a Gibbs prior distribution will provide a first way to build an estimator which can be proved to reach adaptively the best possible asymptotic error rate under Mammen and Tsybakov margin assumptions and parametric complexity assumptions (at least as long as orders of magnitude are concerned, we will not discuss the question of asymptotically optimal constants).

Then we will provide an empirical bound for the Kullback divergence $\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big]$ itself. This will serve to address the question of model selection, which will be achieved by comparing the performance of two posterior

distributions possibly supported by two different models. This will also provide a second way to build estimators which can be proved to be adaptive under Mammen and Tsybakov margin assumptions and parametric complexity assumptions (somewhat weaker than with the first method).

Finally, we will present two-step localization strategies, in which the performance of the posterior distribution to be analysed is compared with a *two-step* Gibbs prior.

2.1.1. Comparing a posterior distribution with a Gibbs prior. Similarly to Theorem 1.4.3 (page 53) we can prove that for any prior distribution $\widetilde{\pi} \in \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\biggl\{\widetilde{\pi}\otimes\widetilde{\pi}\biggl\{\exp\biggl[-N\log(1-N\tanh\bigl(\tfrac{\gamma}{N}\bigr)R') - \gamma r' - N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]m'\biggr]\biggr\}\biggr\} \leq 1. \quad (2.2)$$

Replacing $\widetilde{\pi}$ with $\pi_{\exp(-\beta R)}$ and considering the posterior distribution $\rho \otimes \pi_{\exp(-\beta R)}$, provides a starting point in the comparison of ρ with $\pi_{\exp(-\beta R)}$; we can indeed state with $\mathbb{P}$ probability at least $1-\epsilon$ that

$$\begin{aligned} &-N\log\Bigl\{1-\tanh\bigl(\tfrac{\gamma}{N}\bigr)\Bigl[\rho(R)-\pi_{\exp(-\beta R)}(R)\Bigr]\Bigr\} \\ &\quad\leq \gamma\bigl[\rho(r)-\pi_{\exp(-\beta R)}(r)\bigr] + N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\bigl[\rho\otimes\pi_{\exp(-\beta R)}\bigr](m') \\ &\qquad + \mathcal{K}\bigl[\rho,\pi_{\exp(-\beta R)}\bigr] - \log(\epsilon). \quad (2.3) \end{aligned}$$

Using equation (2.1, page 69) to handle the entropy term, we get

$$\begin{aligned} &-N\log\Bigl\{1-\tanh\bigl(\tfrac{\gamma}{N}\bigr)\Bigl[\rho(R)-\pi_{\exp(-\beta R)}(R)\Bigr]\Bigr\} - \beta\bigl[\rho(R)-\pi_{\exp(-\beta R)}(R)\bigr] \\ &\quad\leq \gamma\bigl[\rho(r)-\pi_{\exp(-\beta R)}(r)\bigr] + N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\rho\otimes\pi_{\exp(-\beta R)}(m') \\ &\qquad + \mathcal{K}(\rho,\pi) - \mathcal{K}\bigl[\pi_{\exp(-\beta R)},\pi\bigr] - \log(\epsilon). \quad (2.4) \end{aligned}$$

We can then decompose in the right-hand side $\gamma\bigl[\rho(r)-\pi_{\exp(-\beta R)}(r)\bigr]$ into $(\gamma-\lambda)\bigl[\rho(r)-\pi_{\exp(-\beta R)}(r)\bigr] + \lambda\bigl[\rho(r)-\pi_{\exp(-\beta R)}(r)\bigr]$ for some parameter λ to be set later on and use the fact that

$$\begin{aligned} &\lambda\bigl[\rho(r)-\pi_{\exp(-\beta R)}(r)\bigr] + N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\rho\otimes\pi_{\exp(-\beta R)}(m') \\ &\qquad + \mathcal{K}(\rho,\pi) - \mathcal{K}\bigl[\pi_{\exp(-\beta R)},\pi\bigr] \\ &\quad\leq \lambda\rho(r) + \mathcal{K}(\rho,\pi) + \log\Bigl\{\pi\Bigl[\exp\bigl\{-\lambda r + N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\rho(m')\bigr\}\Bigr]\Bigr\} \\ &\quad= \mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr] + \log\Bigl\{\pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\rho(m')\bigr\}\Bigr]\Bigr\}, \end{aligned}$$

to get rid of the appearance of the unobserved Gibbs prior $\pi_{\exp(-\beta R)}$ in most places of the right-hand side of our inequality, leading to

THEOREM 2.1.1. *For any real constants β and γ, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, for any real constant λ,*

$$\begin{aligned}
&\big[N\tanh(\tfrac{\gamma}{N}) - \beta\big]\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] \\
&\quad\leq -N\log\Big\{1 - \tanh(\tfrac{\gamma}{N})\Big[\rho(R) - \pi_{\exp(-\beta R)}(R)\Big]\Big\} \\
&\qquad\qquad - \beta\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] \\
&\quad\leq (\gamma-\lambda)\big[\rho(r) - \pi_{\exp(-\beta R)}(r)\big] + \mathcal{K}\big[\rho, \pi_{\exp(-\lambda r)}\big] \\
&\qquad + \log\Big\{\pi_{\exp(-\lambda r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho(m')\big\}\Big]\Big\} - \log(\epsilon) \\
&\quad= \mathcal{K}\big[\rho, \pi_{\exp(-\gamma r)}\big] \\
&\qquad + \log\Big\{\pi_{\exp(-\gamma r)}\Big[\exp\big\{(\gamma-\lambda)r + N\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho(m')\big\}\Big]\Big\} \\
&\qquad\qquad - (\gamma-\lambda)\pi_{\exp(-\beta R)}(r) - \log(\epsilon).
\end{aligned}$$

We would like to have a fully empirical upper bound even in the case when $\lambda \neq \gamma$. This can be done by using the theorem twice. We will need a lemma.

LEMMA 2.1.2 *For any probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$, for any bounded measurable functions $g, h : \Theta \to \mathbb{R}$,*

$$\pi_{\exp(-g)}(g) - \pi_{\exp(-h)}(g) \leq \pi_{\exp(-g)}(h) - \pi_{\exp(-h)}(h).$$

PROOF. Let us notice that

$$\begin{aligned}
0 \leq \mathcal{K}(\pi_{\exp(-g)}, \pi_{\exp(-h)}) &= \pi_{\exp(-g)}(h) + \log\big\{\pi\big[\exp(-h)\big]\big\} + \mathcal{K}(\pi_{\exp(-g)}, \pi) \\
&= \pi_{\exp(-g)}(h) - \pi_{\exp(-h)}(h) - \mathcal{K}(\pi_{\exp(-h)}, \pi) + \mathcal{K}(\pi_{\exp(-g)}, \pi) \\
&= \pi_{\exp(-g)}(h) - \pi_{\exp(-h)}(h) - \mathcal{K}(\pi_{\exp(-h)}, \pi) - \pi_{\exp(-g)}(g) - \log\big\{\pi\big[\exp(-g)\big]\big\}.
\end{aligned}$$

Moreover

$$-\log\big\{\pi\big[\exp(-g)\big]\big\} \leq \pi_{\exp(-h)}(g) + \mathcal{K}(\pi_{\exp(-h)}, \pi),$$

which ends the proof. □

For any positive real constants β and λ, we can then apply Theorem 2.1.1 to $\rho = \pi_{\exp(-\lambda r)}$, and use the inequality

$$\frac{\lambda}{\beta}\big[\pi_{\exp(-\lambda r)}(r) - \pi_{\exp(-\beta R)}(r)\big] \leq \pi_{\exp(-\lambda r)}(R) - \pi_{\exp(-\beta R)}(R) \tag{2.5}$$

provided by the previous lemma. We thus obtain with $\mathbb{P}$ probability at least $1-\epsilon$

$$\begin{aligned}
&-N\log\Big\{1 - \tanh(\tfrac{\gamma}{N})\tfrac{\lambda}{\beta}\Big[\pi_{\exp(-\lambda r)}(r) - \pi_{\exp(-\beta R)}(r)\Big]\Big\} \\
&\qquad\qquad - \gamma\big[\pi_{\exp(-\lambda r)}(r) - \pi_{\exp(-\beta R)}(r)\big]
\end{aligned}$$

$$\leq \log\Big\{\pi_{\exp(-\lambda r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda r)}(m')\big\}\Big]\Big\} - \log(\epsilon).$$

Let us introduce the convex function

$$F_{\gamma,\alpha}(x) = -N\log\big[1-\tanh(\tfrac{\gamma}{N})x\big] - \alpha x \geq \big[N\tanh(\tfrac{\gamma}{N}) - \alpha\big]x.$$

With $\mathbb{P}$ probability at least $1-\epsilon$,

$$\begin{aligned}-\pi_{\exp(-\beta R)}(r) \leq \inf_{\lambda\in\mathbb{R}_+^*}\Bigg\{&-\pi_{\exp(-\lambda r)}(r)\\ &+\frac{\beta}{\lambda}F^{-1}_{\gamma,\frac{\beta\gamma}{\lambda}}\Bigg[\log\Big\{\pi_{\exp(-\lambda r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda r)}(m')\big\}\Big]\Big\}\\ &\qquad\qquad -\log(\epsilon)\Bigg]\Bigg\}.\end{aligned}$$

Since Theorem 2.1.1 holds uniformly for any posterior distribution ρ, we can apply it again to some arbitrary posterior distribution ρ. We can moreover make the result uniform in β and γ by considering some atomic measure $\nu \in \mathcal{M}_+^1(\mathbb{R})$ on the real line and using a union bound. This leads to

THEOREM 2.1.3. *For any atomic probability distribution on the positive real line $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, for any positive real constants β and γ,*

$$\begin{aligned}&\big[N\tanh(\tfrac{\gamma}{N}) - \beta\big]\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big]\\ &\qquad\qquad \leq F_{\gamma,\beta}\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] \leq B(\rho,\beta,\gamma), \text{ where}\\ B(\rho,\beta,\gamma) = \inf_{\substack{\lambda_1\in\mathbb{R}_+,\lambda_1\leq\gamma\\ \lambda_2\in\mathbb{R},\lambda_2>\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1}}}\Bigg\{&\mathcal{K}\big[\rho,\pi_{\exp(-\lambda_1 r)}\big]\\ &+(\gamma-\lambda_1)\big[\rho(r)-\pi_{\exp(-\lambda_2 r)}(r)\big]\\ &+\log\Big\{\pi_{\exp(-\lambda_1 r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho(m')\big\}\Big]\Big\} - \log\big[\epsilon\nu(\beta)\nu(\gamma)\big]\\ &+(\gamma-\lambda_1)\frac{\beta}{\lambda_2}F^{-1}_{\gamma,\frac{\beta\gamma}{\lambda_2}}\Bigg[\log\Big\{\\ &\qquad\pi_{\exp(-\lambda_2 r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda_2 r)}(m')\big\}\Big]\Big\}\\ &\qquad\qquad -\log\big[\epsilon\nu(\beta)\nu(\gamma)\big]\Bigg]\Bigg\}\\ \leq \inf_{\substack{\lambda_1\in\mathbb{R}_+,\lambda_1\leq\gamma\\ \lambda_2\in\mathbb{R},\lambda_2>\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1}}}\Bigg\{&\mathcal{K}\big[\rho,\pi_{\exp(-\lambda_1 r)}\big]\\ &+(\gamma-\lambda_1)\big[\rho(r)-\pi_{\exp(-\lambda_2 r)}(r)\big]\\ &+\log\Big\{\pi_{\exp(-\lambda_1 r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho(m')\big\}\Big]\Big\}\end{aligned}$$

$$+\frac{\beta}{\lambda_2}\frac{(1-\frac{\lambda_1}{\gamma})}{\left[\frac{N}{\gamma}\tanh(\frac{\gamma}{N})-\frac{\beta}{\lambda_2}\right]}\log\Big\{\pi_{\exp(-\lambda_2 r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda_2 r)}(m')\big\}\Big]\Big\}$$
$$-\Big\{1+\frac{\beta}{\lambda_2}\frac{(1-\frac{\lambda_1}{\gamma})}{[\frac{N}{\gamma}\tanh(\frac{\gamma}{N})-\frac{\beta}{\lambda_2}]}\Big\}\log\big[\epsilon\nu(\beta)\nu(\gamma)\big]\Big\},$$

where we have written for short $\nu(\beta)$ *and* $\nu(\gamma)$ *instead of* $\nu(\{\beta\})$ *and* $\nu(\{\gamma\})$.

Let us notice that $B(\rho,\beta,\gamma)=+\infty$ when $\nu(\beta)=0$ or $\nu(\gamma)=0$, the uniformity in β and γ of the theorem therefore necessarily bears on a countable number of values of these parameters. We can typically choose distributions for ν such as the one used in Theorem 1.2.8 (page 27): namely we can put for some positive real ratio $\alpha>1$

$$\nu(\alpha^k)=\frac{1}{(k+1)(k+2)},\qquad k\in\mathbb{N},$$

or alternatively, since we are interested in values of the parameters less than N, we can prefer

$$\nu(\alpha^k)=\frac{\log(\alpha)}{\log(\alpha N)},\qquad 0\le k<\frac{\log(N)}{\log(\alpha)}.$$

We can also use such a coding distribution on dyadic numbers as the one defined by equation (1.7, page 29).

Following the same route as for Theorem 1.3.15 (page 46), we can also prove the following result about the deviations under any posterior distribution ρ:

THEOREM 2.1.4 *For any* $\epsilon\in]0,1[$, *with* $\mathbb{P}$ *probability at least* $1-\epsilon$, *for any posterior distribution* $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$, *with* ρ *probability at least* $1-\xi$,

$$F_{\gamma,\beta}\big[R(\widehat{\theta})-\pi_{\exp(-\beta R)}(R)\big]\le\inf_{\substack{\lambda_1\in\mathbb{R}_+,\lambda_1\le\gamma,\\ \lambda_2\in\mathbb{R},\lambda_2>\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1}}}\Bigg\{\log\left[\frac{d\rho}{d\pi_{\exp(-\lambda_1 r)}}(\widehat{\theta})\right]$$
$$+(\gamma-\lambda_1)\big[r(\widehat{\theta})-\pi_{\exp(-\lambda_2 r)}(r)\big]$$
$$+\log\Big\{\pi_{\exp(-\lambda_1 r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]m'(\cdot,\widehat{\theta})\big\}\Big]\Big\}-\log\big[\epsilon\xi\nu(\beta)\nu(\gamma)\big]$$
$$+(\gamma-\lambda_1)\frac{\beta}{\lambda_2}F^{-1}_{\gamma,\frac{\beta\gamma}{\lambda_2}}\Big[\log\Big\{\pi_{\exp(-\lambda_2 r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda_2 r)}(m')\big\}\Big]\Big\}$$
$$-\log\big[\epsilon\nu(\beta)\nu(\gamma)\big]\Big]\Bigg\}.$$

The only tricky point is to justify that we can still take an infimum in λ_1 without using a union bound. To justify this, we have to notice that the following variant of Theorem 2.1.1 (page 71) holds: with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, for any real constant λ,

$$\rho\Big\{F_{\gamma,\beta}\big[R-\pi_{\exp(-\beta R)}(R)\big]\Big\} \le \mathcal{K}\big[\rho,\pi_{\exp(-\gamma r)}\big] + \rho\Bigg[\inf_{\lambda\in\mathbb{R}} \log\Big\{\pi_{\exp(-\gamma r)}\Big[\exp\big\{(\gamma-\lambda)r + N\log\big[\cosh(\tfrac{\gamma}{N})\big]m'(\cdot,\widehat{\theta}\,)\big\}\Big]\Big\} - (\gamma-\lambda)\pi_{\exp(-\beta R)}(r)\Bigg] - \log(\epsilon).$$

We leave the details as an exercise.

2.1.2. The effective temperature of a posterior distribution. Using the parametric approximation $\pi_{\exp(-\alpha r)}(r) - \inf_\Theta r \simeq \frac{d_e}{\alpha}$, we get as an order of magnitude

$$\begin{aligned} B(\pi_{\exp(-\lambda_1 r)},\beta,\gamma) &\lesssim -(\gamma-\lambda_1)d_e\big[\lambda_2^{-1}-\lambda_1^{-1}\big] \\ &\quad + 2d_e\log\frac{\lambda_1}{\lambda_1 - N\log\big[\cosh(\frac{\gamma}{N})\big]x} \\ &\quad + 2\frac{\beta}{\lambda_2}\frac{(1-\frac{\lambda_1}{\gamma})}{\big[\frac{N}{\gamma}\tanh(\frac{\gamma}{N})-\frac{\beta}{\lambda_2}\big]}d_e\log\left(\frac{\lambda_2}{\lambda_2 - N\log\big[\cosh(\frac{\gamma}{N})\big]x}\right) \\ &\quad + 2N\log\big[\cosh(\tfrac{\gamma}{N})\big]\left[1+\frac{\beta}{\lambda_2}\frac{(1-\frac{\lambda_1}{\gamma})}{\big[\frac{N}{\gamma}\tanh(\frac{\gamma}{N})-\frac{\beta}{\lambda_2}\big]}\right]\widetilde{\varphi}(x) \\ &\quad - \left\{1+\frac{\beta}{\lambda_2}\frac{(1-\frac{\lambda_1}{\gamma})}{\big[\frac{N}{\gamma}\tanh(\frac{\gamma}{N})-\frac{\beta}{\lambda_2}\big]}\right\}\log\big[\nu(\beta)\nu(\gamma)\epsilon\big]. \end{aligned}$$

Therefore, if the empirical dimension d_e stays bounded when N increases, we are going to obtain a negative upper bound for any values of the constants $\lambda_1 > \lambda_2 > \beta$, as soon as γ and $\frac{N}{\gamma}$ are chosen to be large enough. This ability to obtain negative values for the bound $B(\pi_{\exp(-\lambda_1 r)},\gamma,\beta)$, and more generally $B(\rho,\gamma,\beta)$, leads the way to introducing the new concept of the *effective temperature* of an estimator.

Definition 2.1.1 For any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ we define the *effective temperature* $T(\rho) \in \mathbb{R}\cup\{-\infty,+\infty\}$ of ρ by the equation

$$\rho(R) = \pi_{\exp(-\frac{R}{T(\rho)})}(R).$$

Note that $\beta \mapsto \pi_{\exp(-\beta R)}(R) : \mathbb{R}\cup\{-\infty,+\infty\} \to (0,1)$ is continuous and strictly decreasing from $\operatorname{ess\,sup}_\pi R$ to $\operatorname{ess\,inf}_\pi R$ (as soon as these two bounds

do not coincide). This shows that the effective temperature $T(\rho)$ is a well-defined random variable.

Theorem 2.1.3 provides a bound for $T(\rho)$, indeed:

PROPOSITION 2.1.5. *Let*

$$\widehat{\beta}(\rho) = \sup\Big\{\beta \in \mathbb{R}; \inf_{\gamma, N\tanh(\frac{\gamma}{N})>\beta} B(\rho,\beta,\gamma) \leq 0\Big\},$$

where $B(\rho,\beta,\gamma)$ is as in Theorem 2.1.3 (page 72). Then with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, $T(\rho) \leq \widehat{\beta}(\rho)^{-1}$, or equivalently $\rho(R) \leq \pi_{\exp[-\widehat{\beta}(\rho)R]}(R)$.

This notion of *effective temperature* of a (randomized) estimator ρ is interesting for two reasons:

• the difference $\rho(R) - \pi_{\exp(-\beta R)}(R)$ can be estimated with better accuracy than $\rho(R)$ itself, due to the use of relative deviation inequalities, leading to convergence rates up to $1/N$ in favourable situations, even when $\inf_\Theta R$ is not close to zero;

• and of course $\pi_{\exp(-\beta R)}(R)$ is a decreasing function of β, thus being able to estimate $\rho(R) - \pi_{\exp(-\beta R)}(R)$ with some given accuracy, means being able to discriminate between values of $\rho(R)$ with the same accuracy, although doing so through the parametrization $\beta \mapsto \pi_{\exp(-\beta R)}(R)$, which can neither be observed nor estimated with the same precision!

2.1.3. ANALYSIS OF AN EMPIRICAL BOUND FOR THE EFFECTIVE TEMPERATURE. We are now going to launch into a mathematically rigorous analysis of the bound $B(\pi_{\exp(-\lambda_1 r),\beta,\gamma})$ provided by Theorem 2.1.3 (page 72), to show that $\inf_{\rho\in\mathcal{M}_+^1(\Theta)} \pi_{\exp[-\widehat{\beta}(\rho)R]}(R)$ converges indeed to $\inf_\Theta R$ at some optimal rate in favourable situations.

It is more convenient for this purpose to use deviation inequalities involving M' rather than m'. It is straightforward to extend Theorem 1.4.2 (page 52) to

THEOREM 2.1.6. *For any real constants β and γ, for any prior distributions $\pi, \mu \in \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\gamma\rho \otimes \pi_{\exp(-\beta R)}\big[\Psi_{\frac{\gamma}{N}}(R', M')\big] \leq \gamma\rho \otimes \pi_{\exp(-\beta R)}(r') + \mathcal{K}(\rho,\mu) - \log(\eta).$$

In order to transform the left-hand side into a linear expression and in the same time localize this theorem, let us choose μ defined by its density

$$\frac{d\mu}{d\pi}(\theta_1) = C^{-1}\exp\Bigg[-\beta R(\theta_1) - \gamma\int_\Theta \Big\{\Psi_{\frac{\gamma}{N}}\big[R'(\theta_1,\theta_2), M'(\theta_1,\theta_2)\big] - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'(\theta_1,\theta_2)\Big\}\pi_{\exp(-\beta R)}(d\theta_2)\Bigg],$$

where C is such that $\mu(\Theta) = 1$. We get

$$\begin{aligned}
\mathcal{K}(\rho,\mu) &= \beta\rho(R) + \gamma\rho\otimes\pi_{\exp(-\beta R)}\big[\Psi_{\frac{\gamma}{N}}(R',M') - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'\big] + \mathcal{K}(\rho,\pi) \\
&\quad + \log\Big\{\int_\Theta \exp\Big[-\beta R(\theta_1) \\
&\qquad\qquad - \gamma\int_\Theta \Big\{\Psi_{\frac{\gamma}{N}}\big[R'(\theta_1,\theta_2),M'(\theta_1,\theta_2)\big] \\
&\qquad\qquad\quad - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'(\theta_1,\theta_2)\Big\}\pi_{\exp(-\beta R)}(d\theta_2)\Big]\pi(d\theta_1)\Big\} \\
&= \beta\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] \\
&\qquad + \gamma\rho\otimes\pi_{\exp(-\beta R)}\big[\Psi_{\frac{\gamma}{N}}(R',M') - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'\big] \\
&\qquad\qquad + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi) \\
&\quad + \log\Big\{\int_\Theta \exp\Big[-\gamma\int_\Theta \Big\{\Psi_{\frac{\gamma}{N}}\big[R'(\theta_1,\theta_2),M'(\theta_1,\theta_2)\big] \\
&\qquad\qquad - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'(\theta_1,\theta_2)\Big\}\pi_{\exp(-\beta R)}(d\theta_2)\Big]\pi_{\exp(-\beta R)}(d\theta_1)\Big\}.
\end{aligned}$$

Thus with $\mathbb{P}$ probability at least $1-\eta$,

$$\begin{aligned}
&\big[N\sinh(\tfrac{\gamma}{N}) - \beta\big]\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] \\
&\quad \le \gamma\big[\rho(r) - \pi_{\exp(-\beta R)}(r)\big] + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi) - \log(\eta) + C(\beta,\gamma) \\
&\text{where } C(\beta,\gamma) = \log\Big\{\int_\Theta \exp\Big[-\gamma\int_\Theta \Big\{\Psi_{\frac{\gamma}{N}}\big[R'(\theta_1,\theta_2),M'(\theta_1,\theta_2)\big] \\
&\qquad\qquad - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'(\theta_1,\theta_2)\Big\}\pi_{\exp(-\beta R)}(d\theta_2)\Big]\pi_{\exp(-\beta R)}(d\theta_1)\Big\}. \qquad (2.6)
\end{aligned}$$

Remarking that

$$\mathcal{K}\big[\rho,\pi_{\exp(-\beta R)}\big] = \beta\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi),$$

we deduce from the previous inequality

THEOREM 2.1.7. *For any real constants β and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
N\sinh(\tfrac{\gamma}{N})\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] &\le \gamma\big[\rho(r) - \pi_{\exp(-\beta R)}(r)\big] \\
&\quad + \mathcal{K}\big[\rho,\pi_{\exp(-\beta R)}\big] - \log(\eta) + C(\beta,\gamma).
\end{aligned}$$

We can also go into a slightly different direction, starting back again from equation (2.6, page 76) and remarking that for any real constant λ,

$$\begin{aligned}
&\lambda\big[\rho(r) - \pi_{\exp(-\beta R)}(r)\big] + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi) \\
&\qquad \le \lambda\rho(r) + \mathcal{K}(\rho,\pi) + \log\big\{\pi\big[\exp(-\lambda r)\big]\big\} = \mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big].
\end{aligned}$$

This leads to

THEOREM 2.1.8. *For any real constants β and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any real constant λ,*

$$\begin{aligned}\big[N\sinh(\tfrac{\gamma}{N})-\beta\big]\big[\rho(R)-\pi_{\exp(-\beta R)}(R)\big]\\ \leq(\gamma-\lambda)\big[\rho(r)-\pi_{\exp(-\beta R)}(r)\big]+\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]-\log(\eta)+C(\beta,\gamma),\end{aligned}$$

where the definition of $C(\beta,\gamma)$ is given by equation (2.6, page 76).

We can now use this inequality in the case when $\rho=\pi_{\exp(-\lambda r)}$ and combine it with Inequality (2.5, page 71) to obtain

THEOREM 2.1.9 *For any real constants β and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any real constant λ,*

$$\big[\tfrac{N\lambda}{\beta}\sinh(\tfrac{\gamma}{N})-\gamma\big]\big[\pi_{\exp(-\lambda r)}(r)-\pi_{\exp(-\beta R)}(r)\big]\leq C(\beta,\gamma)-\log(\eta).$$

We deduce from this theorem

PROPOSITION 2.1.10 *For any real positive constants β_1, β_2 and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any real constants λ_1 and λ_2, such that $\lambda_2<\beta_2\frac{\gamma}{N}\sinh(\frac{\gamma}{N})^{-1}$ and $\lambda_1>\beta_1\frac{\gamma}{N}\sinh(\frac{\gamma}{N})^{-1}$,*

$$\begin{aligned}\pi_{\exp(-\lambda_1 r)}(r)-\pi_{\exp(-\lambda_2 r)}(r)&\leq\pi_{\exp(-\beta_1 R)}(r)-\pi_{\exp(-\beta_2 R)}(r)\\ &\quad+\frac{C(\beta_1,\gamma)+\log(2/\eta)}{\frac{N\lambda_1}{\beta_1}\sinh(\frac{\gamma}{N})-\gamma}+\frac{C(\beta_2,\gamma)+\log(2/\eta)}{\gamma-\frac{N\lambda_2}{\beta_2}\sinh(\frac{\gamma}{N})}.\end{aligned}$$

Moreover, $\pi_{\exp(-\beta_1 R)}$ and $\pi_{\exp(-\beta_2 R)}$ being prior distributions, with $\mathbb{P}$ probability at least $1-\eta$,

$$\begin{aligned}\gamma\big[\pi_{\exp(-\beta_1 R)}(r)-\pi_{\exp(-\beta_2 R)}(r)\big]\\ \leq\gamma\pi_{\exp(-\beta_1 R)}\otimes\pi_{\exp(-\beta_2 R)}\big[\Psi_{-\frac{\gamma}{N}}(R',M')\big]-\log(\eta).\end{aligned}$$

Hence

PROPOSITION 2.1.11 *For any positive real constants β_1, β_2 and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any positive real constants λ_1 and λ_2 such that $\lambda_2<\beta_2\frac{\gamma}{N}\sinh(\frac{\gamma}{N})^{-1}$ and $\lambda_1>\beta_1\frac{\gamma}{N}\sinh(\frac{\gamma}{N})^{-1}$,*

$$\begin{aligned}\pi_{\exp(-\lambda_1 r)}(r)-\pi_{\exp(-\lambda_2 r)}(r)\\ \leq\pi_{\exp(-\beta_1 R)}\otimes\pi_{\exp(-\beta_2 R)}\big[\Psi_{-\frac{\gamma}{N}}(R',M')\big]\\ +\frac{\log(\frac{3}{\eta})}{\gamma}+\frac{C(\beta_1,\gamma)+\log(\frac{3}{\eta})}{\frac{N\lambda_1}{\beta_1}\sinh(\frac{\gamma}{N})-\gamma}+\frac{C(\beta_2,\gamma)+\log(\frac{3}{\eta})}{\gamma-\frac{N\lambda_2}{\beta_2}\sinh(\frac{\gamma}{N})}.\end{aligned}$$

In order to achieve the analysis of the bound $B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma)$ given by Theorem 2.1.3 (page 72), it now remains to bound quantities of the general form

$$\log\Big\{\pi_{\exp(-\lambda r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda r)}(m')\big\}\Big]\Big\} = \sup_{\rho\in\mathcal{M}_+^1(\Theta)} N\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho\otimes\pi_{\exp(-\lambda)}(m') - \mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big].$$

Let us consider the prior distribution $\mu \in \mathcal{M}_+^1(\Theta\times\Theta)$ on couples of parameters defined by the density

$$\frac{d\mu}{d(\pi\otimes\pi)}(\theta_1,\theta_2) = C^{-1}\exp\Big\{-\beta R(\theta_1) - \beta R(\theta_2) + \alpha\Phi_{-\frac{\alpha}{N}}\big[M'(\theta_1,\theta_2)\big]\Big\},$$

where the normalizing constant C is such that $\mu(\Theta\times\Theta) = 1$. Since for fixed values of the parameters θ and $\theta' \in \Theta$, $m'(\theta,\theta')$, like $r(\theta)$, is a sum of independent Bernoulli random variables, we can easily adapt the proof of Theorem 1.1.4 on page 17, to establish that with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution ρ and any real constant λ,

$$\begin{aligned}\alpha\rho\otimes\pi_{\exp(-\lambda r)}(m') &\le \alpha\rho\otimes\pi_{\exp(-\lambda r)}\big[\Phi_{-\frac{\alpha}{N}}(M')\big] \\ &\qquad + \mathcal{K}(\rho\otimes\pi_{\exp(-\lambda r)},\mu) - \log(\eta)\\ &= \mathcal{K}\big[\rho,\pi_{\exp(-\beta R)}\big] + \mathcal{K}\big[\pi_{\exp(-\lambda r)},\pi_{\exp(-\beta R)}\big]\\ &\qquad + \log\Big\{\pi_{\exp(-\beta R)}\otimes\pi_{\exp(-\beta R)}\Big[\exp\big(\alpha\Phi_{-\frac{\alpha}{N}}\circ M'\big)\Big]\Big\} - \log(\eta).\end{aligned}$$

Thus for any real constant β and any positive real constants α and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any real constant λ,

$$\begin{aligned}&\log\Big\{\pi_{\exp(-\lambda r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda r)}(m')\big\}\Big]\Big\}\\ &\le \sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Bigg(\tfrac{N}{\alpha}\log\big[\cosh(\tfrac{\gamma}{N})\big]\Big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta R)}\big] + \mathcal{K}\big[\pi_{\exp(-\lambda r)},\pi_{\exp(-\beta R)}\big]\\ &\qquad + \log\big\{\pi_{\exp(-\beta R)}\otimes\pi_{\exp(-\beta R)}\big[\exp(\alpha\Phi_{-\frac{\alpha}{N}}\circ M')\big]\big\}\\ &\qquad\qquad - \log(\eta)\Big\} - \mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big]\Bigg). \quad (2.7)\end{aligned}$$

To finish, we need some appropriate upper bound for the entropy $\mathcal{K}\big[\rho,\pi_{\exp(-\beta R)}\big]$. This question can be handled in the following way: using Theorem 2.1.7 (page 76), we see that for any positive real constants γ and β, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution ρ,

$$\begin{aligned}\mathcal{K}\big[\rho,\pi_{\exp(-\beta R)}\big] &= \beta\big[\rho(R) - \pi_{\exp(-\beta R)}(R)\big] + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi)\\ &\le \frac{\beta}{N\sinh(\frac{\gamma}{N})}\Big[\gamma\big[\rho(r) - \pi_{\exp(-\beta R)}(r)\big]\end{aligned}$$

$$
\begin{aligned}
&\quad + \mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big] - \log(\eta) + C(\beta,\gamma)\Big] \\
&\qquad\qquad + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi) \\
&\le \mathcal{K}\big[\rho, \pi_{\exp(-\frac{\beta\gamma}{N\sinh(\frac{\gamma}{N})}r)}\big] \\
&\qquad + \frac{\beta}{N\sinh(\frac{\gamma}{N})}\Big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta R)}\big] + C(\beta,\gamma) - \log(\eta)\Big\}.
\end{aligned}
$$

In other words,

THEOREM 2.1.12. *For any positive real constants β and γ such that $\beta < N\sinh(\frac{\gamma}{N})$, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$
\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big] \le \frac{\mathcal{K}\big[\rho, \pi_{\exp[-\beta\frac{\gamma}{N}\sinh(\frac{\gamma}{N})^{-1}r]}\big]}{1 - \dfrac{\beta}{N\sinh(\frac{\gamma}{N})}} + \frac{C(\beta,\gamma) - \log(\eta)}{\dfrac{N\sinh(\frac{\gamma}{N})}{\beta} - 1},
$$

where the quantity $C(\beta,\gamma)$ is defined by equation (2.6, page 76). Equivalently, it will be in some cases more convenient to use this result in the form: for any positive real constants λ and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$
\mathcal{K}\big[\rho, \pi_{\exp[-\lambda\frac{N}{\gamma}\sinh(\frac{\gamma}{N})R]}\big] \le \frac{\mathcal{K}\big[\rho, \pi_{\exp(-\lambda r)}\big]}{1 - \frac{\lambda}{\gamma}} + \frac{C(\lambda\frac{N}{\gamma}\sinh(\frac{\gamma}{N}),\gamma) - \log(\eta)}{\frac{\lambda}{\beta} - 1}.
$$

Choosing in equation (2.7, page 78) $\alpha = \dfrac{N\log\big[\cosh(\frac{\gamma}{N})\big]}{1 - \frac{\beta}{N\sinh(\frac{\gamma}{N})}}$ and $\beta = \lambda\frac{N}{\gamma}\sinh(\frac{\gamma}{N})$, so that $\alpha = \dfrac{N\log\big[\cosh(\frac{\gamma}{N})\big]}{1-\frac{\lambda}{\gamma}}$, we obtain with $\mathbb{P}$ probability at least $1-\eta$,

$$
\begin{aligned}
&\log\Big\{\pi_{\exp(-\lambda r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda r)}(m')\big\}\Big]\Big\} \\
&\quad \le \tfrac{2\lambda}{\gamma}\big[C(\beta,\gamma) + \log(\tfrac{2}{\eta})\big] \\
&\qquad + \Big(1 - \tfrac{\lambda}{\gamma}\Big)\Big[\log\Big\{\pi_{\exp(-\beta R)} \otimes \pi_{\exp(-\beta R)}\big[\exp(\alpha\Phi_{-\frac{\alpha}{N}}\circ M')\big]\Big\} \\
&\qquad\qquad + \log(\tfrac{2}{\eta})\Big].
\end{aligned}
$$

This proves

PROPOSITION 2.1.13. *For any positive real constants $\lambda < \gamma$, with $\mathbb{P}$ probability at least $1-\eta$,*

$$\log\Big\{\pi_{\exp(-\lambda r)}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi_{\exp(-\lambda r)}(m')\big\}\Big]\Big\}$$
$$\leq \frac{2\lambda}{\gamma}\big[C(\tfrac{N\lambda}{\gamma}\sinh(\tfrac{\gamma}{N}),\gamma)+\log(\tfrac{2}{\eta})\big]$$
$$+\Big(1-\tfrac{\lambda}{\gamma}\Big)\log\Big\{\pi^{\otimes 2}_{\exp[-\frac{N\lambda}{\gamma}\sinh(\frac{\gamma}{N})R]}\Big[$$
$$\exp\Big(\frac{N\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda}{\gamma}}\Phi_{-\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda}{\gamma}}}\circ M'\Big)\Big]\Big\}$$
$$+\Big(1-\tfrac{\lambda}{\gamma}\Big)\log(\tfrac{2}{\eta}).$$

We are now ready to analyse the bound $B(\pi_{\exp(-\lambda_1 r)},\beta,\gamma)$ of Theorem 2.1.3 (page 72).

THEOREM 2.1.14. *For any positive real constants λ_1, λ_2, β_1, β_2, β and γ, such that*

$$\lambda_1<\gamma, \qquad \beta_1<\tfrac{N\lambda_1}{\gamma}\sinh(\tfrac{\gamma}{N}),$$
$$\lambda_2<\gamma, \qquad \beta_2>\tfrac{N\lambda_2}{\gamma}\sinh(\tfrac{\gamma}{N}),$$
$$\beta<\tfrac{N\lambda_2}{\gamma}\tanh(\tfrac{\gamma}{N}),$$

with $\mathbb{P}$ probability $1-\eta$, the bound $B(\pi_{\exp(-\lambda_1 r)},\beta,\gamma)$ of Theorem 2.1.3 (page 72) satisfies

$$B(\pi_{\exp(-\lambda_1 r)},\beta,\gamma)$$
$$\leq(\gamma-\lambda_1)\Big\{\pi_{\exp(-\beta_1 R)}\otimes\pi_{\exp(-\beta_2 R)}\big[\Psi_{-\frac{\gamma}{N}}(R',M')\big]+\frac{\log(\frac{7}{\eta})}{\gamma}$$
$$+\frac{C(\beta_1,\gamma)+\log(\frac{7}{\eta})}{\frac{N\lambda_1}{\beta_1}\sinh(\frac{\gamma}{N})-\gamma}+\frac{C(\beta_2,\gamma)+\log(\frac{7}{\eta})}{\gamma-\frac{N\lambda_2}{\beta_2}\sinh(\frac{\gamma}{N})}\Big\}$$
$$+\frac{2\lambda_1}{\gamma}\Big[C\big(\tfrac{N\lambda_1}{\gamma}\sinh(\tfrac{\gamma}{N}),\gamma\big)+\log\big(\tfrac{7}{\eta}\big)\Big]$$
$$+\Big(1-\tfrac{\lambda_1}{\gamma}\Big)\log\Big\{\pi^{\otimes 2}_{\exp[-\frac{N\lambda_1}{\gamma}\sinh(\frac{\gamma}{N})R]}\Big[$$
$$\exp\Big(\frac{N\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_1}{\gamma}}\Phi_{-\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_1}{\gamma}}}\circ M'\Big)\Big]\Big\}$$
$$+\Big(1-\tfrac{\lambda_1}{\gamma}\Big)\log(\tfrac{7}{\eta})-\log\big[\nu(\{\beta\})\nu(\{\gamma\})\epsilon\big]$$
$$+(\gamma-\lambda_1)\tfrac{\beta}{\lambda_2}F^{-1}_{\gamma,\frac{\beta\gamma}{\lambda_2}}\Big\{\frac{2\lambda_2}{\gamma}\Big[C\big(\tfrac{N\lambda_2}{\gamma}\sinh(\tfrac{\gamma}{N}),\gamma\big)+\log\big(\tfrac{7}{\eta}\big)\Big]$$
$$+\Big(1-\tfrac{\lambda_2}{\gamma}\Big)\log\Big\{\pi^{\otimes 2}_{\exp[-\frac{N\lambda_2}{\gamma}\sinh(\frac{\gamma}{N})R]}\Big[$$

$$\exp\Bigg(\frac{N\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_2}{\gamma}}\Phi_{-\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_2}{\gamma}}}\circ M'\Bigg)\Bigg]\Bigg\}$$
$$+\Big(1-\tfrac{\lambda_2}{\gamma}\Big)\log\big(\tfrac{7}{\eta}\big)-\log\big[\nu(\{\beta\})\nu(\{\gamma\})\epsilon\big]\Bigg\},$$

where the function $C(\beta,\gamma)$ is defined by equation (2.6, page 76).

2.1.4. Adaptation to parametric and margin assumptions. To help understand the previous theorem, it may be useful to give linear upper-bounds to the factors appearing in the right-hand side of the previous inequality. Introducing $\widetilde{\theta}$ such that $R(\widetilde{\theta}) = \inf_\Theta R$ (assuming that such a parameter exists) and remembering that

$$\begin{aligned}
\Psi_{-a}(p,m) &\le a^{-1}\sinh(a)p + 2a^{-1}\sinh(\tfrac{a}{2})^2 m, && a\in\mathbb{R}_+,\\
\Phi_{-a}(p) &\le a^{-1}\big[\exp(a)-1\big]p, && a\in\mathbb{R}_+,\\
\Psi_{a}(p,m) &\ge a^{-1}\sinh(a)p - 2a^{-1}\sinh(\tfrac{a}{2})^2 m, && a\in\mathbb{R}_+,\\
M'(\theta_1,\theta_2) &\le M'(\theta_1,\widetilde{\theta}) + M'(\theta_2,\widetilde{\theta}), && \theta_1,\theta_2\in\Theta,\\
M'(\theta_1,\widetilde{\theta}) &\le xR'(\theta_1,\widetilde{\theta}) + \varphi(x), && x\in\mathbb{R}_+,\theta_1\in\Theta,
\end{aligned}$$

the last inequality being rather a consequence of the definition of φ than a property of M', we easily see that

$$\begin{aligned}
&\pi_{\exp(-\beta_1 R)}\otimes\pi_{\exp(-\beta_2 R)}\big[\Psi_{-\frac{\gamma}{N}}(R',M')\big]\\
&\quad\le \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\big[\pi_{\exp(-\beta_1 R)}(R)-\pi_{\exp(-\beta_2 R)}(R)\big]\\
&\qquad\qquad + \tfrac{2N}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\pi_{\exp(-\beta_1 R)}\otimes\pi_{\exp(-\beta_2 R)}(M')\\
&\quad\le \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\big[\pi_{\exp(-\beta_1 R)}(R)-\pi_{\exp(-\beta_2 R)}(R)\big]\\
&\qquad + \frac{2xN}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\Big\{\pi_{\exp(-\beta_1 R)}\big[R'(\cdot,\widetilde{\theta})\big]+\pi_{\exp(-\beta_2 R)}\big[R'(\cdot,\widetilde{\theta})\big]\Big\}\\
&\qquad\qquad + \frac{4N}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\varphi(x),
\end{aligned}$$

that

$$\begin{aligned}
C(\beta,\gamma) &\le \log\Big\{\pi_{\exp(-\beta R)}\Big\{\exp\Big[2N\sinh\big(\tfrac{\gamma}{2N}\big)^2\pi_{\exp(-\beta R)}(M')\Big]\Big\}\Big\}\\
&\le \log\Big\{\pi_{\exp(-\beta R)}\Big\{\exp\Big[2N\sinh\big(\tfrac{\gamma}{2N}\big)^2 M'(\cdot,\widetilde{\theta})\Big]\Big\}\Big\}\\
&\qquad\qquad + 2N\sinh(\tfrac{\gamma}{2N})^2\pi_{\exp(-\beta R)}\big[M'(\cdot,\widetilde{\theta})\big]\\
&\le \log\Big\{\pi_{\exp(-\beta R)}\Big\{\exp\Big[2xN\sinh(\tfrac{\gamma}{2N})^2 R'(\cdot,\widetilde{\theta})\Big]\Big\}\Big\}\\
&\qquad + 2xN\sinh(\tfrac{\gamma}{2N})^2\pi_{\exp(-\beta R)}\big[R'(\cdot,\widetilde{\theta})\big] + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\beta - 2xN\sinh(\frac{\gamma}{2N})^2}^{\beta} \pi_{\exp(-\alpha R)}\big[R'(\cdot,\widetilde{\theta})\big]\,d\alpha \\
&\qquad + 2xN\sinh(\tfrac{\gamma}{2N})^2 \pi_{\exp(-\beta R)}\big[R'(\cdot,\widetilde{\theta})\big] + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x) \\
&\le 4xN\sinh(\tfrac{\gamma}{2N})^2 \pi_{\exp[-(\beta - 2xN\sinh(\frac{\gamma}{2N})^2)R]}\big[R'(\cdot,\widetilde{\theta})\big] \\
&\qquad\qquad + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x),
\end{aligned}
$$

and that

$$
\begin{aligned}
&\log\Big\{\pi^{\otimes 2}_{\exp(-\beta R)}\Big[\exp\Big(N\alpha\Phi_{-\alpha}\circ M'\Big)\Big]\Big\} \\
&\qquad \le 2\log\Big\{\pi_{\exp(-\beta R)}\Big[\exp\Big(N\big[\exp(\alpha)-1\big]M'(\cdot,\widetilde{\theta})\Big)\Big]\Big\} \\
&\qquad \le 2xN\big[\exp(\alpha)-1\big]\pi_{\exp[-(\beta - xN[\exp(\alpha)-1])R]}\big[R'(\cdot,\widetilde{\theta})\big] \\
&\qquad\qquad + 2xN\big[\exp(\alpha)-1\big]\varphi(x).
\end{aligned}
$$

Let us push further the investigation under the parametric assumption that for some positive real constant d

$$
\lim_{\beta\to+\infty} \beta\pi_{\exp(-\beta R)}\big[R'(\cdot,\widetilde{\theta})\big] = d, \tag{2.8}
$$

This assumption will for instance hold true with $d = \frac{n}{2}$ when $R : \Theta \to (0,1)$ is a smooth function defined on a compact subset Θ of $\mathbb{R}^n$ that reaches its minimum value on a finite number of non-degenerate (i.e. with a positive definite Hessian) interior points of Θ, and π is absolutely continuous with respect to the Lebesgue measure on Θ and has a smooth density.

In case of assumption (2.8), if we restrict ourselves to sufficiently large values of the constants β, β_1, β_2, λ_1, λ_2 and γ (the smaller of which is as a rule β, as we will see), we can use the fact that for some (small) positive constant δ, and some (large) positive constant A,

$$
\frac{d}{\alpha}(1-\delta) \le \pi_{\exp(-\alpha R)}\big[R'(\cdot,\widetilde{\theta})\big] \le \frac{d}{\alpha}(1+\delta), \qquad \alpha \ge A. \tag{2.9}
$$

Under this assumption,

$$
\begin{aligned}
&\pi_{\exp(-\beta_1 R)} \otimes \pi_{\exp(-\beta_2 R)}\big[\Psi_{-\frac{\gamma}{N}}(R', M')\big] \\
&\qquad \le \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\big[\tfrac{d}{\beta_1}(1+\delta) - \tfrac{d}{\beta_2}(1-\delta)\big] \\
&\qquad\quad + \tfrac{2xN}{\gamma}\sinh(\tfrac{\gamma}{2N})^2(1+\delta)\big[\tfrac{d}{\beta_1} + \tfrac{d}{\beta_2}\big] + \tfrac{4N}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\varphi(x). \\
&C(\beta,\gamma) \le d(1+\delta)\log\Big(\tfrac{\beta}{\beta - 2xN\sinh(\frac{\gamma}{2N})^2}\Big) \\
&\qquad\qquad + 2xN\sinh(\tfrac{\gamma}{2N})^2\tfrac{(1+\delta)d}{\beta} + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x). \\
&\log\Big\{\pi^{\otimes 2}_{\exp(-\beta R)}\Big[\exp\Big(N\alpha\Phi_{-\alpha}\circ M'\Big)\Big]\Big\} \\
&\qquad \le 2xN\big[\exp(\alpha)-1\big]\frac{d(1+\delta)}{\beta - xN[\exp(\alpha)-1]} + 2N\big[\exp(\alpha)-1\big]\varphi(x).
\end{aligned}
$$

Thus with $\mathbb{P}$ probability at least $1-\eta$,

$$
\begin{aligned}
&B(\pi_{\exp(-\lambda_1 r)},\beta,\gamma) \le -(\gamma-\lambda_1)\tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\tfrac{d}{\beta_2}(1-\delta)\\
&+(\gamma-\lambda_1)\Bigg\{\tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\tfrac{(1+\delta)d}{\beta_1}\\
&\qquad+\tfrac{2xN}{\gamma}\sinh(\tfrac{\gamma}{2N})^2(1+\delta)\big[\tfrac{d}{\beta_1}+\tfrac{d}{\beta_2}\big]+\tfrac{4N}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\varphi(x)+\frac{\log(\frac{7}{\eta})}{\gamma}\\
&\quad+\frac{4xN\sinh(\frac{\gamma}{2N})^2\frac{(1+\delta)d}{\beta_1-2xN\sinh(\frac{\gamma}{2N})^2}+4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x)+\log(\frac{7}{\eta})}{\frac{N\lambda_1}{\beta_1}\sinh(\frac{\gamma}{N})-\gamma}\\
&\quad+\frac{4xN\sinh(\frac{\gamma}{2N})^2\frac{(1+\delta)d}{\beta_2-2xN\sinh(\frac{\gamma}{2N})^2}+4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x)+\log(\frac{7}{\eta})}{\gamma-\frac{N\lambda_2}{\beta_2}\sinh(\frac{\gamma}{N})}\Bigg\}\\
&+\frac{2\lambda_1}{\gamma}\Bigg\{4xN\sinh(\tfrac{\gamma}{2N})^2\frac{(1+\delta)d}{\frac{N\lambda_1}{\gamma}\sinh(\frac{\gamma}{N})-2xN\sinh(\frac{\gamma}{2N})^2}\\
&\qquad+4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x)+\log(\tfrac{7}{\eta})\Bigg\}\\
&+\Big(1-\frac{\lambda_1}{\gamma}\Big)\Bigg\{2d(1+\delta)\Bigg(\frac{\lambda_1\sinh\big(\frac{\gamma}{N}\big)}{x\gamma\Big[\exp\Big(\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_1}{\gamma}}\Big)-1\Big]}-1\Bigg)^{-1}\\
&\qquad+2N\Big[\exp\Big(\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_1}{\gamma}}\Big)-1\Big]\varphi(x)\Bigg\}\\
&\quad+\Big(1-\frac{\lambda_1}{\gamma}\Big)\log(\tfrac{7}{\eta})-\log\big[\nu(\{\beta\})\nu(\{\gamma\})\epsilon\big]\\
&+\frac{1-\frac{\lambda_1}{\gamma}}{\frac{N\lambda_2}{\beta\gamma}\tanh(\frac{\gamma}{N})-1}\Bigg\{\frac{2\lambda_2}{\gamma}\Bigg\{4xN\sinh(\tfrac{\gamma}{2N})^2\frac{(1+\delta)d}{\frac{N\lambda_2}{\gamma}\sinh(\frac{\gamma}{N})-2xN\sinh(\frac{\gamma}{2N})^2}\\
&\qquad+4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x)+\log(\tfrac{7}{\eta})\Bigg\}\\
&+\Big(1-\frac{\lambda_2}{\gamma}\Big)\Bigg[2d(1+\delta)\Bigg(\frac{\lambda_2\sinh\big(\frac{\gamma}{N}\big)}{x\gamma\Big[\exp\Big(\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_2}{\gamma}}\Big)-1\Big]}-1\Bigg)^{-1}\\
&\qquad+2N\Big[\exp\Big(\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_2}{\gamma}}\Big)-1\Big]\varphi(x)\Bigg]\\
&\qquad+\Big(1-\frac{\lambda_2}{\gamma}\Big)\log(\tfrac{7}{\eta})-\log\big[\nu(\beta)\nu(\gamma)\epsilon\big]\Bigg\}.
\end{aligned}
$$

Now let us choose for simplicity $\beta_2 = 2\lambda_2 = 4\beta$, $\beta_1 = \lambda_1/2 = \gamma/4$, and let us introduce the notation

$$C_1 = \frac{N}{\gamma}\sinh(\frac{\gamma}{N}),$$

$$C_2 = \frac{N}{\gamma}\tanh(\frac{\gamma}{N}),$$
$$C_3 = \frac{N^2}{\gamma^2}\big[\exp(\frac{\gamma^2}{N^2}) - 1\big]$$
$$\text{and} \quad C_4 = \frac{2N^2(1-\frac{2\beta}{\gamma})}{\gamma^2}\Big[\exp\Big(\frac{\gamma^2}{2N^2(1-\frac{2\beta}{\gamma})}\Big) - 1\Big],$$

to obtain

$$\begin{aligned}
B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma) \le &-\frac{C_1\gamma}{8\beta}(1-\delta)d \\
&+ \frac{C_1\gamma}{2}\Big\{\tfrac{4(1+\delta)d}{\gamma} + x\tfrac{\gamma}{2N}(1+\delta)\big[\tfrac{4d}{\gamma} + \tfrac{d}{4\beta}\big] + \tfrac{\gamma}{N}\varphi(x)\Big\} + \tfrac{1}{2}\log\big(\tfrac{7}{\eta}\big) \\
&+ \frac{1}{2C_1 - 1}\Big[(1+\delta)d\Big(\tfrac{N}{2xC_1\gamma} - 1\Big)^{-1} + C_1\frac{\gamma^2}{2N}\varphi(x) + \tfrac{1}{2}\log(\tfrac{7}{\eta})\Big] \\
&+ \frac{1}{2 - C_1}\Big[2(1+\delta)d\Big(\tfrac{8N\beta}{xC_1\gamma^2} - 1\Big)^{-1} + C_1\frac{\gamma^2}{N}\varphi(x) + \log(\tfrac{7}{\eta})\Big] \\
&+ \frac{2x\gamma(1+\delta)d}{N - x\gamma} + C_1\tfrac{\gamma^2}{N}\varphi(x) + \log(\tfrac{7}{\eta}) \\
&+ d(1+\delta)\frac{x\gamma}{N}\left(\frac{C_1}{2C_3} - \frac{x\gamma}{N}\right)^{-1} + \frac{\gamma^2}{N}C_3\varphi(x) + \frac{\log(\frac{7}{\eta})}{2} - \log\big[\nu(\beta)\nu(\gamma)\epsilon\big] \\
&+ \Big(4C_2 - 2\Big)^{-1}\Big\{\frac{4\beta}{\gamma}\Big\{x\frac{\gamma^2}{N}C_1(1+\delta)d\Big(2\beta C_1 - xC_1\frac{\gamma^2}{2N}\Big)^{-1} \\
&\qquad + \tfrac{\gamma^2}{N}\varphi(x) + \log(\tfrac{7}{\eta})\Big\} \\
&+ \Big(1 - \frac{2\beta}{\gamma}\Big)\Big\{2d(1+\delta)\frac{x\gamma}{N}\left[\frac{4\beta C_1}{\gamma C_4}\left(1 - \frac{2\beta}{\gamma}\right) - \frac{x\gamma}{N}\right]^{-1} \\
&\qquad + \frac{\gamma^2}{N(1-\frac{2\beta}{\gamma})}C_4\varphi(x)\Big\} \\
&\qquad + \Big(1 - \tfrac{2\beta}{\gamma}\Big)\log(\tfrac{7}{\eta}) - \log\big[\nu(\beta)\nu(\gamma)\epsilon\big]\Big\}.
\end{aligned}$$

This simplifies to

$$\begin{aligned}
B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma) \le &-\frac{C_1}{8}(1-\delta)d\frac{\gamma}{\beta} \\
&+ 2C_1(1+\delta)d + \log(\tfrac{7}{\eta})\left[2 + \tfrac{3C_1}{(4C_1-2)(2-C_1)} + \frac{1 + \frac{2\beta}{\gamma}}{4C_2 - 2}\right] \\
&\qquad - \big(1 + \tfrac{1}{4C_2-2}\big)\log\big[\nu(\beta)\nu(\gamma)\epsilon\big] \\
&+ \frac{(1+\delta)dx\gamma}{N}\Big\{C_1 + \tfrac{1}{2C_1-1}\Big(\tfrac{1}{2C_1} - \tfrac{\gamma x}{N}\Big)^{-1}
\end{aligned}$$

$$+ 2\Big(1 - \tfrac{\gamma x}{N}\Big)^{-1} + \Big(\tfrac{C_1}{2C_3} - \tfrac{\gamma x}{N}\Big)^{-1} + \tfrac{4C_1\beta}{\gamma(4C_2-2)}\Big\}$$
$$+ \frac{(1+\delta)dx\gamma^2}{N\beta}\Big\{\tfrac{C_1}{16} + \tfrac{2}{2-C_1}\Big(\tfrac{8}{C_1} - \tfrac{x\gamma^2}{N\beta}\Big)^{-1}$$
$$+ \Big(1 - \tfrac{2\beta}{\gamma}\Big)\tfrac{1}{2C_2-1}\Big[\tfrac{4C_1}{C_4}\Big(1 - \tfrac{2\beta}{\gamma}\Big) - \tfrac{\gamma^2 x}{\beta N}\Big]^{-1}\Big\}$$
$$+ \frac{\gamma^2}{N}\varphi(x)\Big\{\tfrac{3C_1}{2} + \tfrac{C_1}{4C_1-2} + \tfrac{C_1}{2-C_1} + C_3 + \tfrac{4\beta}{\gamma(4C_2-2)} + \tfrac{C_4}{4C_2-2}\Big\}.$$

This shows that there exist universal positive real constants A_1, A_2, B_1, B_2, B_3, and B_4 such that as soon as $\frac{\gamma \max\{x,1\}}{N} \leq A_1\frac{\beta}{\gamma} \leq A_2$,

$$B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma) \leq -B_1(1-\delta)d\frac{\gamma}{\beta} + B_2(1+\delta)d$$
$$- B_3 \log\big[\nu(\beta)\nu(\gamma)\epsilon\,\eta\big] + B_4\frac{\gamma^2}{N}\varphi(x).$$

Thus $\pi_{\exp(-\lambda_1 r)}(R) \leq \pi_{\exp(-\beta R)}(R) \leq \inf_\Theta R + \frac{(1+\delta)d}{\beta}$ as soon as

$$\frac{\beta}{\gamma} \leq \frac{B_1}{B_2\frac{(1+\delta)}{(1-\delta)} + \frac{B_4\frac{\gamma^2}{N}\varphi(x) - B_3\log[\nu(\beta)\nu(\gamma)\epsilon\eta]}{(1-\delta)d}}.$$

Choosing some real ratio $\alpha > 1$, we can now make the above result uniform for any

$$\beta, \gamma \in \Lambda_\alpha \stackrel{\text{def}}{=} \Big\{\alpha^k; k \in \mathbb{N}, 0 \leq k < \tfrac{\log(N)}{\log(\alpha)}\Big\}, \tag{2.10}$$

by substituting $\nu(\beta)$ and $\nu(\gamma)$ with $\frac{\log(\alpha)}{\log(\alpha N)}$ and $-\log(\eta)$ with $-\log(\eta) + 2\log\Big[\frac{\log(\alpha N)}{\log(\alpha)}\Big]$.

Taking $\eta = \epsilon$ for simplicity, we can summarize our result in

THEOREM 2.1.15. *There exist positive real universal constants A, B_1, B_2, B_3 and B_4 such that for any positive real constants $\alpha > 1$, d and δ, for any prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\beta, \gamma \in \Lambda_\alpha$ (where Λ_α is defined by equation (2.10) above) such that*

$$\sup_{\beta' \in \mathbb{R}, \beta' \geq \beta}\Big|\frac{\beta'}{d}\big[\pi_{\exp(-\beta' R)}(R) - \inf_\Theta R\big] - 1\Big| \leq \delta$$

and such that also for some positive real parameter x

$$\frac{\gamma\max\{x,1\}}{N} \leq \frac{A\beta}{\gamma} \text{ and } \frac{\beta}{\gamma} \leq \frac{B_1}{B_2\frac{(1+\delta)}{(1-\delta)} + \frac{B_4\frac{\gamma^2}{N}\varphi(x) - 2B_3\log(\epsilon) + 4B_3\log\left[\frac{\log(N)}{\log(\alpha)}\right]}{(1-\delta)d}},$$

the bound $B(\pi_{\exp(-\frac{\gamma}{2}r)},\beta,\gamma)$ *given by Theorem 2.1.3 on page 72 in the case where we have chosen* ν *to be the uniform probability measure on* Λ_α*, satisfies* $B(\pi_{\exp(-\frac{\gamma}{2}r)},\beta,\gamma)\leq 0$*, proving that* $\widehat{\beta}(\pi_{\exp(-\frac{\gamma}{2}r)})\geq\beta$ *and therefore that*

$$\pi_{\exp(-\gamma\frac{r}{2})}(R)\leq\pi_{\exp(-\beta R)}(R)\leq\inf_\Theta R+\frac{(1+\delta)d}{\beta}.$$

What is important in this result is that we do not only bound $\pi_{\exp(-\frac{\gamma}{2}r)}(R)$, but also $B(\pi_{\exp(-\frac{\gamma}{2}r)},\beta,\gamma)$, and that we do it uniformly on a grid of values of β and γ, showing that we can indeed set the constants β and γ adaptively using the empirical bound $B(\pi_{\exp(-\frac{\gamma}{2}r)},\beta,\gamma)$.

Let us see what we get under the margin assumption (1.24, page 56). When $\kappa=1$, we have $\varphi(c^{-1})\leq 0$, leading to

Corollary 2.1.16. *Assuming that the margin assumption (1.24, page 56) is satisfied for* $\kappa=1$*, that* $R:\Theta\to(0,1)$ *is independent of* N *(which is the case for instance when* $\mathbb{P}=P^{\otimes N}$*), and is such that*

$$\lim_{\beta'\to+\infty}\beta'\big[\pi_{\exp(-\beta' R)}(R)-\inf_\Theta R\big]=d,$$

there are universal positive real constants B_5 *and* B_6 *and* $N_1\in\mathbb{N}$ *such that for any* $N\geq N_1$*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$

$$\pi_{\exp(-\widehat{\gamma}\frac{r}{2})}(R)\leq\inf_\Theta R+\frac{B_5 d}{cN}\left[1+\frac{B_6}{d}\log\left(\frac{\log(N)}{\epsilon}\right)\right]^2,$$

where $\widehat{\gamma}\in\arg\max_{\gamma\in\Lambda_2}\max\big\{\beta\in\Lambda_2;B(\pi_{\exp(-\gamma\frac{r}{2})},\beta,\gamma)\leq 0\big\}$*, where* Λ_2 *is defined by equation (2.10, page 85), and* B *is the bound of Theorem 2.1.3 (page 72).*

When $\kappa>1$, $\varphi(x)\leq(1-\kappa^{-1})\big(\kappa c x\big)^{-\frac{1}{\kappa-1}}$, and we can choose γ and x such that $\frac{\gamma^2}{N}\varphi(x)\simeq d$ to prove

Corollary 2.1.17. *Assuming that the margin assumption (1.24, page 56) is satisfied for some exponent* $\kappa>1$*, that* $R:\Theta\to(0,1)$ *is independent of* N *(which is for instance the case when* $\mathbb{P}=P^{\otimes N}$*), and is such that*

$$\lim_{\beta'\to+\infty}\beta'\big[\pi_{\exp(-\beta' R)}(R)-\inf_\Theta R\big]=d,$$

there are universal positive constants B_7 *and* B_8 *and* $N_1\in\mathbb{N}$ *such that for any* $N\geq N_1$*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*,*

$$\pi_{\exp(-\widehat{\gamma}\frac{r}{2})}(R)\leq\inf_\Theta R+B_7 c^{-\frac{1}{2\kappa-1}}\left[1+\frac{B_8}{d}\log\left(\frac{\log(N)}{\epsilon}\right)\right]^{\frac{2\kappa}{2\kappa-1}}\left(\frac{d}{N}\right)^{\frac{\kappa}{2\kappa-1}},$$

where $\widehat{\gamma}\in\arg\max_{\gamma\in\Lambda_2}\max\big\{\beta\in\Lambda_2;B(\pi_{\exp(-\gamma\frac{r}{2})},\beta,\gamma)\leq 0\big\}$*,* Λ_2 *being defined by equation (2.10, page 85) and* B *by Theorem 2.1.3 (page 72).*

We find the same rate of convergence as in Corollary 1.4.7 (page 57), but this time, we were able to provide an empirical posterior distribution $\pi_{\exp(-\widehat{\gamma}\frac{r}{2})}$ which achieves this rate adaptively in all the parameters (meaning in particular that we do not need to know d, c or κ). Moreover, as already mentioned, the power of N in this rate of convergence is known to be optimal in the worst case (see Mammen et al. (1999); Tsybakov (2004); Tsybakov et al. (2005), and more specifically in Audibert (2004b) — downloadable from its author's web page — Theorem 3.3, page 132).

2.1.5. Estimating the divergence of a posterior with respect to a Gibbs prior. Another interesting question is to estimate $\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big]$ using relative deviation inequalities. We follow here an idea to be found first in (Audibert, 2004b, page 93). Indeed, combining equation (2.3, page 70) with equation (2.1, page 69), we see that for any positive real parameters β and λ, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}
\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big] &\leq \frac{\beta}{N\tanh(\frac{\gamma}{N})}\Big\{\gamma\big[\rho(r) - \pi_{\exp(-\beta R)}(r)\big] \\
&\qquad + N\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho\otimes\pi_{\exp(-\beta R)}(m') \\
&\qquad + \mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big] - \log(\epsilon)\Big\} + \mathcal{K}(\rho,\pi) - \mathcal{K}\big[\pi_{\exp(-\beta R)}, \pi\big] \\
&\leq \mathcal{K}\big[\rho, \pi_{\exp[-\frac{\beta\gamma}{N\tanh(\frac{\gamma}{N})}r]}\big] + \frac{\beta}{N\tanh(\frac{\gamma}{N})}\Big\{\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big] - \log(\epsilon)\Big\} \\
&\qquad + \log\Big[\pi_{\exp[-\frac{\beta\gamma}{N\tanh(\frac{\gamma}{N})}r]}\Big\{\exp\Big[\frac{\beta}{\tanh(\frac{\gamma}{N})}\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho(m')\Big]\Big\}\Big].
\end{aligned}$$

We thus obtain

Theorem 2.1.18. *For any positive real constants β and γ such that $\beta < N\tanh(\frac{\gamma}{N})$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
\mathcal{K}\big[\rho, \pi_{\exp(-\beta R)}\big] &\leq \left(1 - \frac{\beta}{N}\tanh\left(\frac{\gamma}{N}\right)^{-1}\right)^{-1} \\
&\qquad \times \left\{\mathcal{K}\big[\rho, \pi_{\exp[-\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1}r]}\big] - \frac{\beta}{N\tanh(\frac{\gamma}{N})}\log(\epsilon)\right. \\
&\qquad \left. + \log\Big\{\pi_{\exp[-\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1}r]}\Big[\exp\big\{\beta\tanh(\tfrac{\gamma}{N})^{-1}\log[\cosh(\tfrac{\gamma}{N})]\rho(m')\big\}\Big]\Big\}\right\}.
\end{aligned}$$

This theorem provides another way of measuring over-fitting, since it gives an upper bound for $\mathcal{K}\big[\pi_{\exp[-\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1}r]}, \pi_{\exp(-\beta R)}\big]$. It may be used in

combination with Theorem 1.2.6 (page 25) as an alternative to Theorem 1.3.7 (page 35). It will also be used in the next section.

An alternative parametrization of the same result providing a simpler right-hand side is also useful:

COROLLARY 2.1.19. *For any positive real constants β and γ such that $\beta < \gamma$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathcal{K}\big[\rho, \pi_{\exp[-N\frac{\beta}{\gamma}\tanh(\frac{\gamma}{N})R]}\big] \leq \left(1-\frac{\beta}{\gamma}\right)^{-1}\left\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big] - \frac{\beta}{\gamma}\log(\epsilon) + \log\left\{\pi_{\exp(-\beta r)}\left[\exp\left\{N\tfrac{\beta}{\gamma}\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho(m')\right\}\right]\right\}\right\}.$$

2.2. Playing with two posterior and two local prior distributions

2.2.1. COMPARING TWO POSTERIOR DISTRIBUTIONS. Estimating the effective temperature of an estimator provides an efficient way to tune parameters in a model with parametric behaviour. On the other hand, it will not be fitted to choose between different models, especially when they are nested, because as we already saw in the case when Θ is a union of nested models, the prior distribution $\pi_{\exp(-\beta R)}$ does not provide an efficient localization of the parameter in this case, in the sense that $\pi_{\exp(-\beta R)}(R)$ does not go down to $\inf_\Theta R$ at the desired rate when β goes to $+\infty$, requiring a resort to partial localization.

Once some estimator (in the form of a posterior distribution) has been chosen in each sub-model, these estimators can be compared between themselves with the help of the relative bounds that we will establish in this section. It is also possible to choose several estimators in each sub-model, to tune parameters in the same time (like the inverse temperature parameter if we decide to use Gibbs posterior distributions in each sub-model).

From equation (2.2 page 70) (slightly modified by replacing $\pi \otimes \pi$ with $\pi^1 \otimes \pi^2$), we easily obtain

THEOREM 2.2.1. *For any positive real constant λ, for any prior distributions $\pi^1, \pi^2 \in \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions ρ_1 and $\rho_2 : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$-N\log\Big\{1-\tanh\big(\tfrac{\lambda}{N}\big)\Big[\rho_2(R)-\rho_1(R)\Big]\Big\} \leq \lambda\big[\rho_2(r)-\rho_1(r)\big] + N\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big]\rho_1\otimes\rho_2(m') + \mathcal{K}\big(\rho_1,\pi^1\big) + \mathcal{K}\big(\rho_2,\pi^2\big) - \log(\epsilon).$$

This is where the entropy bound of the previous section enters into the game, providing a localized version of Theorem 2.2.1 (page 88). We will use the notation

$$\Xi_a(q) = \tanh(a)^{-1}\big[1-\exp(-aq)\big] \leq \frac{a}{\tanh(a)}q, \qquad a,q\in\mathbb{R}. \tag{2.11}$$

THEOREM 2.2.2. *For any $\epsilon\in]0,1[$, any sequence of prior distributions $(\pi^i)_{i\in\mathbb{N}}\in\mathcal{M}_+^1(\Theta)^{\mathbb{N}}$, any probability distribution μ on $\mathbb{N}$, any atomic probability distribution ν on $\mathbb{R}_+$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\rho_1,\rho_2:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\rho_2(R)-\rho_1(R)\leq B(\rho_1,\rho_2),\ \textit{where}$$

$$\begin{aligned}
B(\rho_1,\rho_2) = \inf_{\lambda,\beta_1<\gamma_1,\beta_2<\gamma_2\in\mathbb{R}_+,i,j\in\mathbb{N}} \Xi_{\frac{\lambda}{N}}\Bigg\{ &\big[\rho_2(r)-\rho_1(r)\big] \\
&\qquad + \tfrac{N}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\rho_1\otimes\rho_2(m') \\
+\frac{1}{\lambda\Big(1-\frac{\beta_1}{\gamma_1}\Big)}\Bigg\{ &\mathcal{K}\big[\rho_1,\pi^i_{\exp(-\beta_1 r)}\big] \\
&\quad + \log\Big\{\pi^i_{\exp(-\beta_1 r)}\Big[\exp\big\{\beta_1\tfrac{N}{\gamma_1}\log\big[\cosh(\tfrac{\gamma_1}{N})\big]\rho_1(m')\big\}\Big]\Big\} \\
&\qquad\qquad - \frac{\beta_1}{\gamma_1}\log\big[\nu(\gamma_1)\big]\Bigg\} \\
+\frac{1}{\lambda\Big(1-\frac{\beta_2}{\gamma_2}\Big)}\Bigg\{ &\mathcal{K}\big[\rho_2,\pi^j_{\exp(-\beta_2 r)}\big] \\
&\quad + \log\Big\{\pi^j_{\exp(-\beta_2 r)}\Big[\exp\big\{\beta_2\tfrac{N}{\gamma_2}\log\big[\cosh(\tfrac{\gamma_2}{N})\big]\rho_2(m')\big\}\Big]\Big\} \\
&\qquad\qquad - \frac{\beta_2}{\gamma_2}\log\big[\nu(\gamma_2)\big]\Bigg\} \\
-\Big[\big(\tfrac{\gamma_1}{\beta_1}-1\big)^{-1}+\big(\tfrac{\gamma_2}{\beta_2}-1\big)^{-1}+1\Big]&\frac{\log\big[3^{-1}\nu(\beta_1)\nu(\beta_2)\nu(\lambda)\mu(i)\mu(j)\epsilon\big]}{\lambda}\Bigg\}.
\end{aligned}$$

The sequence of prior distributions $(\pi^i)_{i\in\mathbb{N}}$ should be understood to be typically supported by subsets of Θ corresponding to parametric sub-models, that is sub-models for which it is reasonable to expect that

$$\lim_{\beta\to+\infty}\beta\big[\pi^i_{\exp(-\beta R)}(R)-\operatorname*{ess\,inf}_{\pi^i}R\big]$$

exists and is positive and finite. As there is no reason why the bound $B(\rho_1,\rho_2)$ provided by the previous theorem should be sub-additive (in the sense that $B(\rho_1,\rho_3)\leq B(\rho_1,\rho_2)+B(\rho_2,\rho_3)$), it is adequate to consider some workable subset $\mathcal{P}$ of posterior distributions (for instance the distributions of the form $\pi^i_{\exp(-\beta r)}$, $i\in\mathbb{N}$, $\beta\in\mathbb{R}_+$), and to define the sub-additive chained bound

$$\widetilde{B}(\rho,\rho') = \inf\left\{\sum_{k=0}^{n-1} B(\rho_k,\rho_{k+1});\ n\in\mathbb{N}^*, (\rho_k)_{k=0}^n \in \mathcal{P}^{n+1},\right.$$
$$\left.\rho_0=\rho, \rho_n=\rho'\right\},\quad \rho,\rho'\in\mathcal{P}. \quad (2.12)$$

Proposition 2.2.3. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\rho_1,\rho_2\in\mathcal{P}$, $\rho_2(R)-\rho_1(R)\le\widetilde{B}(\rho_1,\rho_2)$. Moreover for any posterior distribution $\rho_1\in\mathcal{P}$, any posterior distribution $\rho_2\in\mathcal{P}$ such that $\widetilde{B}(\rho_1,\rho_2)=\inf_{\rho_3\in\mathcal{P}}\widetilde{B}(\rho_1,\rho_3)$ is unimprovable with the help of $\widetilde{B}$ in $\mathcal{P}$ in the sense that $\inf_{\rho_3\in\mathcal{P}}\widetilde{B}(\rho_2,\rho_3)\ge 0$.*

Proof. The first assertion is a direct consequence of the previous theorem, so only the second assertion requires a proof: for any $\rho_3\in\mathcal{P}$, we deduce from the optimality of ρ_2 and the sub-additivity of $\widetilde{B}$ that

$$\widetilde{B}(\rho_1,\rho_2)\le\widetilde{B}(\rho_1,\rho_3)\le\widetilde{B}(\rho_1,\rho_2)+\widetilde{B}(\rho_2,\rho_3).$$

□

This proposition provides a way to improve a posterior distribution $\rho_1\in\mathcal{P}$ by choosing $\rho_2\in\arg\min_{\rho\in\mathcal{P}}\widetilde{B}(\rho_1,\rho)$ whenever $\widetilde{B}(\rho_1,\rho_2)<0$. This improvement is proved by Proposition 2.2.3 to be one-step: the obtained improved posterior ρ_2 cannot be improved again using the same technique.

Let us give some examples of possible starting distributions ρ_1 for this improvement scheme: ρ_1 may be chosen as the best posterior Gibbs distribution according to Proposition 2.1.5 (page 75). More precisely, we may build from the prior distributions π^i, $i\in\mathbb{N}$, a global prior $\pi=\sum_{i\in\mathbb{N}}\mu(i)\pi^i$. We can then define the estimator of the inverse effective temperature as in Proposition 2.1.5 (page 75) and choose $\rho_1\in\arg\min_{\rho\in\mathcal{P}}\widehat{\beta}(\rho)$, where $\mathcal{P}$ is as suggested above the set of posterior distributions

$$\mathcal{P}=\left\{\pi^i_{\exp(-\beta r)};\ i\in\mathbb{N},\beta\in\mathbb{R}_+\right\}.$$

This starting point ρ_1 should already be pretty good, at least in an asymptotic perspective, the only gain in the rate of convergence to be expected bearing on spurious $\log(N)$ factors.

2.2.2. Elaborate uses of relative bounds between posteriors. More elaborate uses of relative bounds are described in the third section of the second chapter of Audibert (2004b), where an algorithm is proposed and analysed, which allows one to use relative bounds between two posterior distributions as a stand-alone estimation tool.

Let us give here some alternative way to address this issue. We will assume for simplicity and without great loss of generality that the working set of posterior distributions $\mathcal{P}$ is finite (so that among other things any ordering of it has a first element).

It is natural to define the estimated complexity of any given posterior distribution $\rho \in \mathcal{P}$ in our working set as the bound for $\inf_{i\in\mathbb{N}} \mathcal{K}(\rho, \pi^i)$ used in Theorem 2.2.1 (page 88). This leads to set (given some confidence level $1-\epsilon$)

$$\begin{aligned}\mathcal{C}(\rho) = \inf_{\beta<\gamma\in\mathbb{R}_+, i\in\mathbb{N}} \left(1-\frac{\beta}{\gamma}\right)^{-1} \Bigg\{ & \mathcal{K}\big[\rho, \pi^i_{\exp(-\beta r)}\big] \\ & + \log\Big\{\pi^i_{\exp(-\beta r)}\Big[\exp\big\{\beta\tfrac{N}{\gamma}\log\big[\cosh(\tfrac{\gamma}{N})\big]\rho(m')\big\}\Big]\Big\} \\ & - \frac{\beta}{\gamma}\log\big[3^{-1}\nu(\gamma)\nu(\beta)\mu(i)\epsilon\big]\Bigg\}.\end{aligned}$$

Let us moreover call $\gamma(\rho)$, $\beta(\rho)$ and $i(\rho)$ the values achieving this infimum, or nearly achieving it, which requires a slight change of the definition of $\mathcal{C}(\rho)$ to take this modification into account. For the sake of simplicity, we can assume without substantial loss of generality that the supports of ν and μ are large but finite, and thus that the minimum is reached.

To understand how this notion of complexity comes into play, it may be interesting to keep in mind that for any posterior distributions ρ and ρ' we can write the bound in Theorem 2.2.2 (page 89) as

$$B(\rho, \rho') = \inf_{\lambda\in\mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\big[\rho'(r) - \rho(r) + S_\lambda(\rho, \rho')\big], \tag{2.13}$$

where

$$\begin{aligned}S_\lambda(\rho, \rho') = S_\lambda(\rho', \rho) \leq \frac{N}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\rho\otimes\rho'(m') + \frac{\mathcal{C}(\rho)+\mathcal{C}(\rho')}{\lambda} - \frac{\log(3^{-1}\epsilon)}{\lambda} \\ - \frac{\log\big\{\nu\big[\beta(\rho)\big]\mu\big[i(\rho)\big]\big\}}{\lambda\big(1-\frac{\beta(\rho')}{\gamma(\rho')}\big)} - \frac{\log\big\{\nu\big[\beta(\rho')\big]\mu\big[i(\rho')\big]\big\}}{\lambda\big(1-\frac{\beta(\rho)}{\gamma(\rho)}\big)} \\ - \Big[\big(\tfrac{\gamma(\rho)}{\beta(\rho)}-1\big)^{-1} + \big(\tfrac{\gamma(\rho')}{\beta(\rho')}-1\big)^{-1} + 1\Big]\frac{\log\big[\nu(\lambda)\big]}{\lambda}.\end{aligned}$$

(Let us recall that the function Ξ is defined by equation (2.11, page 89).) Thus for any ρ, ρ' such that $B(\rho', \rho) > 0$, we can deduce from the monotonicity of $\Xi_{\frac{\lambda}{N}}$ that

$$\rho'(r) - \rho(r) \leq \inf_{\lambda\in\mathbb{R}_+} S_\lambda(\rho, \rho'),$$

proving that the left-hand side is small, and consequently that $B(\rho, \rho')$ and its chained counterpart defined by equation (2.12, page 90) are small:

$$\widetilde{B}(\rho, \rho') \leq B(\rho, \rho') \leq \inf_{\lambda\in\mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\big[2S_\lambda(\rho, \rho')\big].$$

It is also worth noticing that $B(\rho, \rho')$ and $\widetilde{B}(\rho, \rho')$ are upper bounded in terms of variance and complexity only.

The presence of the ratios $\frac{\gamma(\rho)}{\beta(\rho)}$ should not be obnoxious, since their values should be automatically tamed by the fact that $\beta(\rho)$ and $\gamma(\rho)$ should make the estimate of the complexity of ρ optimal.

As an alternative, it is possible to restrict to set of parameter values β and γ such that, for some fixed constant $\zeta > 1$, the ratio $\frac{\gamma}{\beta}$ is bounded away from 1 by the inequality $\frac{\gamma}{\beta} \geq \zeta$. This leads to an alternative definition of $\mathcal{C}(\rho)$:

$$\mathcal{C}(\rho) = \inf_{\gamma \geq \zeta\beta \in \mathbb{R}_+, i \in \mathbb{N}} \left(1 - \frac{\beta}{\gamma}\right)^{-1} \Bigg\{ \mathcal{K}\big[\rho, \pi^i_{\exp(-\beta r)}\big] + \log\Big\{ \pi^i_{\exp(-\beta r)} \Big[\exp\big\{ \beta \tfrac{N}{\gamma} \log\big[\cosh(\tfrac{\gamma}{N})\big] \rho(m') \big\} \Big] \Big\} - \frac{\beta}{\gamma} \log\big[3^{-1} \nu(\gamma) \nu(\beta) \mu(i) \epsilon\big] \Bigg\} - \frac{\log\big[\nu(\beta)\mu(i)\big]}{(1 - \zeta^{-1})} - \frac{\log(3^{-1}\epsilon)}{2}.$$

We can even push simplification a step further, postponing the optimization of the ratio $\frac{\gamma}{\beta}$, and setting it to the fixed value ζ. This leads us to adopt the definition

$$\mathcal{C}(\rho) = \inf_{\beta \in \mathbb{R}_+, i \in \mathbb{N}} \big(1 - \zeta^{-1}\big)^{-1} \Bigg\{ \mathcal{K}\big[\rho, \pi^i_{\exp(-\beta r)}\big] + \log\Big\{ \pi^i_{\exp(-\beta r)} \Big[\exp\big\{ \tfrac{N}{\zeta} \log\big[\cosh(\tfrac{\zeta\beta}{N})\big] \rho(m') \big\} \Big] \Big\} \Bigg\} - \frac{\zeta + 1}{\zeta - 1} \Bigg\{ \log\big[\nu(\beta)\mu(i)\big] + 2^{-1} \log(3^{-1}\epsilon) \Bigg\}. \quad (2.14)$$

With either of these modified definitions of the complexity $\mathcal{C}(\rho)$, we get the upper bound

$$S_\lambda(\rho, \rho') \leq \widetilde{S}_\lambda(\rho, \rho') \stackrel{\text{def}}{=} \frac{N}{\lambda} \log\big[\cosh(\tfrac{\lambda}{N})\big] \rho \otimes \rho'(m') + \frac{1}{\lambda} \Bigg\{ \mathcal{C}(\rho) + \mathcal{C}(\rho') - \frac{\zeta + 1}{\zeta - 1} \log\big[\nu(\lambda)\big] \Bigg\}. \quad (2.15)$$

With these definitions, we have for any posterior distributions ρ and ρ'

$$B(\rho, \rho') \leq \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}} \Big\{ \rho'(r) - \rho(r) + \widetilde{S}_\lambda(\rho, \rho') \Big\}.$$

Consequently in the case when $B(\rho', \rho) > 0$, we get

$$\widetilde{B}(\rho, \rho') \leq B(\rho, \rho') \leq \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}} \big[2\widetilde{S}_\lambda(\rho, \rho') \big].$$

To select some nearly optimal posterior distribution in $\mathcal{P}$, it is appropriate to order the posterior distributions of $\mathcal{P}$ according to increasing values of their complexity $\mathcal{C}(\rho)$ and consider some indexation $\mathcal{P} = \{\rho_1, \ldots, \rho_M\}$, where $\mathcal{C}(\rho_k) \leq \mathcal{C}(\rho_{k+1})$, $1 \leq k < M$.

Let us now consider for each $\rho_k \in \mathcal{P}$ the first posterior distribution in $\mathcal{P}$ which cannot be proved to be worse than ρ_k according to the bound $\widetilde{B}$:

$$t(k) = \min\Big\{ j \in \{1, \dots M\} \,:\, \widetilde{B}(\rho_j, \rho_k) > 0 \Big\}. \tag{2.16}$$

In this definition, which uses the chained bound defined by equation (2.12, page 90), it is appropriate to assume by convention that $\widetilde{B}(\rho, \rho) = 0$, for any posterior distribution ρ. Let us now define our estimated best $\rho \in \mathcal{P}$ as $\rho_{\widehat{k}}$, where

$$\widehat{k} = \min(\arg\max t). \tag{2.17}$$

Thus we take the posterior with smallest complexity which can be proved to be better than the largest starting interval of $\mathcal{P}$ in terms of estimated relative classification error.

The following theorem is a simple consequence of the chosen optimisation scheme. It is valid for any arbitrary choice of the complexity function $\rho \mapsto \mathcal{C}(\rho)$.

THEOREM 2.2.4. *Let us put $\widehat{t} = t(\widehat{k})$, where t is defined by equation* (2.16) *and $\widehat{k}$ is defined by equation* (2.17). *With $\mathbb{P}$ probability at least $1 - \epsilon$,*

$$\rho_{\widehat{k}}(R) \leq \rho_j(R) + \begin{cases} 0, & 1 \leq j < \widehat{t}, \\ \widetilde{B}(\rho_j, \rho_{t(j)}), & \widehat{t} \leq j < \widehat{k}, \\ \widetilde{B}(\rho_j, \rho_{\widehat{t}}) + \widetilde{B}(\rho_{\widehat{t}}, \rho_{\widehat{k}}), & j \in (\arg\max t), \\ \widetilde{B}(\rho_j, \rho_{\widehat{k}}), & j \in \big\{\widehat{k}+1, \dots, M\big\} \setminus (\arg\max t), \end{cases}$$

where the chained bound $\widetilde{B}$ is defined from the bound of Theorem 2.2.2 (page 89) by equation (2.12, page 90). In the mean time, for any j such that $\widehat{t} \leq j < \widehat{k}$, $t(j) < \widehat{t} = \max t$, because $j \notin (\arg\max t)$. Thus

$$\rho_{\widehat{k}}(R) \leq \rho_{t(j)}(R) \leq \rho_j(R) + \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\big[2 S_\lambda(\rho_j, \rho_{t(j)})\big]$$

while $\rho_{t(j)}(r) \leq \rho_j(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_j, \rho_{t(j)})$,

where the function Ξ is defined by equation (2.11, page 89) and S_λ is defined by equation (2.13, page 91). For any $j \in (\arg\max t)$, (including notably $\widehat{k}$),

$$B(\rho_{\widehat{t}}, \rho_j) \geq \widetilde{B}(\rho_{\widehat{t}}, \rho_j) > 0,$$
$$B(\rho_j, \rho_{\widehat{t}}) \geq \widetilde{B}(\rho_j, \rho_{\widehat{t}}) > 0,$$

so in this case

$$\rho_{\widehat{k}}(R) \leq \rho_j(R) + \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\Big[S_\lambda(\rho_j, \rho_{\widehat{t}}) + S_\lambda(\rho_{\widehat{t}}, \rho_{\widehat{k}}) + S_\lambda(\rho_j, \rho_{\widehat{k}})\Big],$$

while $\rho_{\widehat{t}}(r) \leq \rho_j(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_j, \rho_{\widehat{t}})$,

$$\rho_{\widehat{k}}(r) \leq \rho_{\widehat{t}}(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_{\widehat{t}}, \rho_{\widehat{k}}),$$
$$\text{and } \rho_{\widehat{t}}(R) \leq \rho_j(R) + \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\left[2S_\lambda(\rho_j, \rho_{\widehat{t}})\right].$$

Finally in the case when $j \in \left\{\widehat{k}+1, \dots, M\right\} \setminus (\arg\max t)$*, due to the fact that in particular* $j \notin (\arg\max t)$*,*

$$B(\rho_{\widehat{k}}, \rho_j) \geq \widetilde{B}(\rho_{\widehat{k}}, \rho_j) > 0.$$

Thus in this last case

$$\rho_{\widehat{k}}(R) \leq \rho_j(R) + \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\left[2S_\lambda(\rho_j, \rho_{\widehat{k}})\right],$$
$$\text{while } \rho_{\widehat{k}}(r) \leq \rho_j(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_j, \rho_{\widehat{k}}).$$

Thus for any $j = 1, \dots, M$*,* $\rho_{\widehat{k}}(R) - \rho_j(R)$ *is bounded from above by an empirical quantity involving only variance and entropy terms of posterior distributions* ρ_ℓ *such that* $\ell \leq j$*, and therefore such that* $\mathcal{C}(\rho_\ell) \leq \mathcal{C}(\rho_j)$*. Moreover, these distributions* ρ_ℓ *are such that* $\rho_\ell(r) - \rho_j(r)$ *and* $\rho_\ell(R) - \rho_j(R)$ *have an empirical upper bound of the same order as the bound stated for* $\rho_{\widehat{k}}(R) - \rho_j(R)$ *— namely the bound for* $\rho_\ell(r) - \rho_j(r)$ *is in all circumstances not greater than* $\Xi^{-1}_{\frac{\lambda}{N}}$ *applied to the bound stated for* $\rho_{\widehat{k}}(R) - \rho_j(R)$*, whereas the bound for* $\rho_\ell(R) - \rho_j(R)$ *is always smaller than two times the bound stated for* $\rho_{\widehat{k}}(R) - \rho_j(R)$*. This shows that variance terms are between posterior distributions whose empirical as well as expected error rates cannot be much larger than those of* ρ_j*.*

Let us remark that the estimation scheme described in this theorem is very general, the same method can be used as soon as some *confidence interval* for the relative expected risks

$$-B(\rho_2, \rho_1) \leq \rho_2(R) - \rho_1(R) \leq B(\rho_1, \rho_2) \text{ with } \mathbb{P} \text{ probability at least } 1 - \epsilon,$$

is available. The definition of the complexity is arbitrary, and could in an abstract context be chosen as

$$\mathcal{C}(\rho_1) = \inf_{\rho_2 \neq \rho_1} B(\rho_1, \rho_2) + B(\rho_2, \rho_1).$$

Proof. The case when $1 \leq j < \widehat{t}$ is straightforward from the definitions: when $j < \widehat{t}$, $\widetilde{B}(\rho_j, \rho_{\widehat{k}}) \leq 0$ and therefore $\rho_{\widehat{k}}(R) \leq \rho_j(R)$.

In the second case, that is when $\widehat{t} \leq j < \widehat{k}$, j cannot be in $\arg\max t$, because of the special choice of $\widehat{k}$ in $\arg\max t$. Thus $t(j) < \widehat{t}$ and we deduce from the first case that

$$\rho_{\widehat{k}}(R) \leq \rho_{t(j)}(R) \leq \rho_j(R) + \widetilde{B}(\rho_j, \rho_{t(j)}).$$

Moreover, we see from the defintion of t that $\widetilde{B}(\rho_{t(j)}, \rho_j) > 0$, implying

$$\rho_{t(j)}(r) \leq \rho_j(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_j, \rho_{t(j)}),$$

and therefore that

$$\rho_{\widehat{k}}(R) \leq \rho_j(R) + \inf_{\lambda} \Xi_{\frac{\lambda}{N}}\big[2S_\lambda(\rho_j, \rho_{t(j)})\big].$$

In the third case j belongs to $\arg\max t$. In this case, we are not sure that $\widetilde{B}(\rho_{\widehat{k}}, \rho_j) > 0$, and it is appropriate to involve $\widehat{t}$, which is the index of the first posterior distribution which cannot be improved by $\rho_{\widehat{k}}$, implying notably that $\widetilde{B}(\rho_{\widehat{t}}, \rho_k) > 0$ for any $k \in \arg\max t$. On the other hand, $\rho_{\widehat{t}}$ cannot either improve any posterior distribution ρ_k with $k \in (\arg\max t)$, because this would imply for any $\ell < \widehat{t}$ that $\widetilde{B}(\rho_\ell, \rho_{\widehat{t}}) \leq \widetilde{B}(\rho_\ell, \rho_k) + \widetilde{B}(\rho_k, \rho_{\widehat{t}}) \leq 0$, and therefore that $t(\widehat{t}) \geq \widehat{t} + 1$, in contradiction of the fact that $\widehat{t} = \max t$. Thus $\widetilde{B}(\rho_k, \rho_{\widehat{t}}) > 0$, and these two remarks imply that

$$\begin{aligned}
\rho_{\widehat{t}}(r) &\leq \rho_j(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_j, \rho_{\widehat{t}}), \\
\rho_{\widehat{k}}(r) &\leq \rho_{\widehat{t}}(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_{\widehat{t}}, \rho_{\widehat{k}}) \\
&\leq \rho_j(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_j, \rho_{\widehat{t}}) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_{\widehat{t}}, \rho_{\widehat{k}}),
\end{aligned}$$

and consequently also that

$$\begin{aligned}
\rho_{\widehat{k}}(R) &\leq \rho_j(R) + \widetilde{B}(\rho_j, \rho_{\widehat{k}}) \\
&\leq \rho_j(R) + \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\Big[S_\lambda(\rho_j, \rho_{\widehat{t}}) + S_\lambda(\rho_{\widehat{t}}, \rho_{\widehat{k}}) + S_\lambda(\rho_j, \rho_{\widehat{k}})\Big]
\end{aligned}$$

and that

$$\rho_{\widehat{t}}(R) \leq \rho_j(R) + \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\big[2S_\lambda(\rho_j, \rho_{\widehat{t}})\big] \leq \rho_j(R) + 2 \inf_{\lambda \in \mathbb{R}_+} 2\Xi_{\frac{\lambda}{N}}\big[S_\lambda(\rho_j, \rho_{\widehat{t}})\big],$$

the last inequality being due to the fact that $\Xi_{\frac{\lambda}{N}}$ is a concave function. Let us notice that it may be the case that $\widehat{k} < \widehat{t}$, but that only the case when $j \geq \widehat{t}$ is to be considered, since otherwise we already know that $\rho_{\widehat{k}}(R) \leq \rho_j(R)$.

In the fourth case, j is greater than $\widehat{k}$, and the complexity of ρ_j is larger than the complexity of $\rho_{\widehat{k}}$. Moreover, j is not in $\arg\max t$, and thus $\widetilde{B}(\rho_{\widehat{k}}, \rho_j) > 0$, because otherwise, the sub-additivity of $\widetilde{B}$ would imply that $\widetilde{B}(\rho_\ell, \rho_j) \leq 0$ for any $\ell \leq \widehat{t}$ and therefore that $t(j) \geq \widehat{t} = \max t$. Therefore

$$\rho_{\widehat{k}}(r) \leq \rho_j(r) + \inf_{\lambda \in \mathbb{R}_+} S_\lambda(\rho_j, \rho_{\widehat{k}}),$$

and

$$\rho_{\widehat{k}}(R) \leq \rho_j(R) + \widetilde{B}(\rho_j, \rho_{\widehat{k}}) \leq \rho_j(R) + \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}}\big[2S_\lambda(\rho_j, \rho_{\widehat{k}})\big].$$

□

2.2.3. Analysis of relative bounds. Let us start our investigation of the theoretical properties of the algorithm described in Theorem 2.2.4 (page 93) by computing some non-random upper bounds for $B(\rho, \rho')$, the bound of Theorem 2.2.2 (page 89), and $\mathcal{C}(\rho)$, the complexity factor defined by equation (2.14, page 92), for any $\rho, \rho' \in \mathcal{P}$.

This analysis will be done in the case when

$$\mathcal{P} = \Big\{ \pi^i_{\exp(-\beta r)} \,:\, \nu(\beta) > 0, \mu(i) > 0 \Big\},$$

in which it will be possible to get some control on the randomness of any $\rho \in \mathcal{P}$, in addition to controlling the other random expressions appearing in the definition of $B(\rho, \rho')$, $\rho, \rho' \in \mathcal{P}$. We will also use a simpler choice of complexity function, removing from equation (2.14 page 92) the optimization in i and β and using instead the definition

$$\mathcal{C}(\pi^i_{\exp(-\beta r)}) \stackrel{\text{def}}{=} \big(1 - \zeta^{-1}\big)^{-1} \log\Big\{ \pi^i_{\exp(-\beta r)} \Big[\exp\Big\{ \tfrac{N}{\zeta} \log\big[\cosh\big(\tfrac{\zeta\beta}{N}\big)\big] \pi^i_{\exp(-\beta r)}(m') \Big\} \Big] \Big\} + \frac{\zeta + 1}{\zeta - 1} \log\big[\nu(\beta)\mu(i)\big]. \quad (2.18)$$

With this definition,

$$S_\lambda(\pi^i_{\exp(-\beta r)}, \pi^j_{\exp(-\beta' r)}) \le \frac{N}{\lambda} \log\big[\cosh(\tfrac{\lambda}{N})\big] \pi^i_{\exp(-\beta r)} \otimes \pi^j_{\exp(-\beta' r)}(m') + \frac{\mathcal{C}\big[\pi^i_{\exp(-\beta r)}\big] + \mathcal{C}\big[\pi^j_{\exp(-\beta' r)}\big]}{\lambda} + \frac{(\zeta + 1)}{(\zeta - 1)\lambda} \log\big[3^{-1}\nu(\lambda)\epsilon\big],$$

where S_λ is defined by equation (2.13, page 91), so that

$$B\big[\pi^i_{\exp(-\beta r)}, \pi^j_{\exp(-\beta' r)}\big] = \inf_{\lambda \in \mathbb{R}_+} \Xi_{\frac{\lambda}{N}} \Big\{ \pi^j_{\exp(-\beta' r)}(r) - \pi^i_{\exp(-\beta r)}(r) + S_\lambda\big[\pi^i_{\exp(-\beta r)}, \pi^j_{\exp(-\beta r)}\big] \Big\}.$$

Let us successively bound the various random factors entering into the definition of $B\big[\pi^i_{\exp(-\beta r)}, \pi^j_{\exp(-\beta' r)}\big]$. The quantity $\pi^j_{\exp(-\beta' r)}(r) - \pi^i_{\exp(-\beta r)}(r)$ can be bounded using a slight adaptation of Proposition 2.1.11 (page 77).

Proposition 2.2.5. *For any positive real constants λ, λ' and γ, with $\mathbb{P}$ probability at least $1 - \eta$, for any positive real constants β, β' such that $\beta < \lambda \frac{\gamma}{N} \sinh(\frac{\gamma}{N})^{-1}$ and $\beta' > \lambda' \frac{\gamma}{N} \sinh(\frac{\gamma}{N})^{-1}$,*

$$\pi^j_{\exp(-\beta' r)}(r) - \pi^i_{\exp(-\beta r)}(r)$$
$$\leq \pi^j_{\exp(-\lambda' R)} \otimes \pi^i_{\exp(-\lambda R)}\big[\Psi_{-\frac{\gamma}{N}}(R', M')\big] + \frac{\log\big(\frac{3}{\eta}\big)}{\gamma} + \frac{C^j(\lambda', \gamma) + \log(\frac{3}{\eta})}{\frac{N\beta'}{\lambda'}\sinh(\frac{\gamma}{N}) - \gamma} + \frac{C^i(\lambda, \gamma) + \log(\frac{3}{\eta})}{\gamma - \frac{N\beta}{\lambda}\sinh(\frac{\gamma}{N})},$$

where

$$C^i(\lambda, \gamma) \stackrel{\text{def}}{=} \log\bigg\{\int_\Theta \exp\bigg[-\gamma \int_\Theta \Big\{\Psi_{\frac{\gamma}{N}}\big[R'(\theta_1, \theta_2), M'(\theta_1, \theta_2)\big] - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N}) R'(\theta_1, \theta_2)\Big\} \pi^i_{\exp(-\lambda R)}(d\theta_2)\bigg] \pi^i_{\exp(-\lambda R)}(d\theta_1)\bigg\}$$
$$\leq \log\bigg\{\pi^i_{\exp(-\lambda R)}\bigg[\exp\Big\{2N\sinh\big(\tfrac{\gamma}{2N}\big)^2 \pi^i_{\exp(-\lambda R)}(M')\Big\}\bigg]\bigg\}.$$

As for $\pi^i_{\exp(-\beta r)} \otimes \pi^j_{\exp(-\beta' r)}(m')$, we can write with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distributions ρ and $\rho' : \Omega \to \mathcal{M}^1_+(\Theta)$,

$$\gamma\rho \otimes \rho'(m') \leq \log\Big[\pi^i_{\exp(-\lambda R)} \otimes \pi^j_{\exp(-\lambda' R)}\big\{\exp\big[\gamma\Phi_{-\frac{\gamma}{N}}(M')\big]\big\}\Big] + \mathcal{K}\big[\rho, \pi^i_{\exp(-\lambda R)}\big] + \mathcal{K}\big[\rho', \pi^j_{\exp(-\lambda' R)}\big] - \log(\eta).$$

We can then replace λ with $\beta\frac{N}{\lambda}\sinh(\frac{\lambda}{N})$ and use Theorem 2.1.12 (page 79) to get

Proposition 2.2.6. *For any positive real constants γ, λ, λ', β and β', with $\mathbb{P}$ probability $1-\eta$,*

$$\gamma\rho \otimes \rho'(m') \leq \log\Big[\pi^i_{\exp[-\beta\frac{N}{\lambda}\sinh(\frac{\lambda}{N})R]} \otimes \pi^j_{\exp[-\beta'\frac{N}{\lambda'}\sinh(\frac{\lambda'}{N})R]}\big\{\exp\big[\gamma\Phi_{-\frac{\gamma}{N}}(M')\big]\big\}\Big] + \frac{\mathcal{K}\big[\rho, \pi^i_{\exp(-\beta r)}\big]}{1-\frac{\beta}{\lambda}} + \frac{C^i\big[\beta\frac{N}{\lambda}\sinh(\frac{\lambda}{N}), \lambda\big] - \log(\frac{\eta}{3})}{\frac{\lambda}{\beta} - 1}$$
$$+ \frac{\mathcal{K}\big[\rho', \pi^j_{\exp(-\beta' r)}\big]}{1-\frac{\beta'}{\lambda'}} + \frac{C^j\big[\beta'\frac{N}{\lambda'}\sinh(\frac{\lambda'}{N}), \lambda'\big] - \log(\frac{\eta}{3})}{\frac{\lambda}{\beta'} - 1} - \log(\tfrac{\eta}{3}).$$

The last random factor in $B(\rho, \rho')$ that we need to upper bound is

$$\log\Big\{\pi^i_{\exp(-\beta r)}\Big[\exp\big\{\beta\tfrac{N}{\gamma}\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi^i_{\exp(-\beta r)}(m')\big\}\Big]\Big\}.$$

A slight adaptation of Proposition 2.1.13 (page 79) shows that with $\mathbb{P}$ probability at least $1-\eta$,

$$\log\Big\{\pi^i_{\exp(-\beta r)}\Big[\exp\big\{\beta\tfrac{N}{\gamma}\log\big[\cosh(\tfrac{\gamma}{N})\big]\pi^i_{\exp(-\beta r)}(m')\big\}\Big]\Big\}$$
$$\leq \frac{2\beta}{\gamma}C^i\big[\tfrac{N\beta}{\gamma}\sinh(\tfrac{\gamma}{N}),\gamma\big] + \big(1-\tfrac{\beta}{\gamma}\big)\log\Big\{\Big(\pi^i_{\exp[-\frac{N\beta}{\gamma}\sinh(\frac{\gamma}{N})R]}\Big)^{\otimes 2}\Big[\exp\Big(\frac{N\log\big[\cosh(\frac{\gamma}{N})\big]}{\frac{\gamma}{\beta}-1}\Phi_{-\frac{\log[\cosh(\frac{\gamma}{N})]}{\frac{\gamma}{\beta}-1}}\circ M'\Big)\Big]\Big\} + \big(1+\tfrac{\beta}{\gamma}\big)\log(\tfrac{2}{\eta}),$$

where as usual Φ is the function defined by equation (1.1, page 15). This leads us to define for any $i,j\in\mathbb{N}$, any $\beta,\beta'\in\mathbb{R}_+$,

$$\overline{\mathcal{C}}(i,\beta)\stackrel{\text{def}}{=}\frac{2}{\zeta-1}C^i\Big[\tfrac{N}{\zeta}\sinh(\tfrac{\zeta\beta}{N}),\zeta\beta\Big] + \log\Big\{\big(\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}\big)^{\otimes 2}\Big[\exp\Big(\frac{N\log\big[\cosh(\frac{\zeta\beta}{N})\big]}{\zeta-1}\Phi_{-\frac{\log[\cosh(\frac{\zeta\beta}{N})]}{\zeta-1}}\circ M'\Big)\Big]\Big\} - \frac{\zeta+1}{\zeta-1}\Big\{2\log\big[\nu(\beta)\mu(i)\big]+\log\big(\tfrac{\eta}{2}\big)\Big\}. \quad (2.19)$$

Recall that the definition of $C^i(\lambda,\gamma)$ is to be found in Proposition 2.2.5, page 96. Let us remark that, since

$$\exp\big[Na\Phi_{-a}(p)\big]=\exp\Big\{N\log\Big[1+\big[\exp(a)-1\big]p\Big]\Big\} \leq \exp\Big\{N\big[\exp(a)-1\big]p\Big\},\quad p\in(0,1),a\in\mathbb{R},$$

we have

$$\overline{\mathcal{C}}(i,\beta)\leq\frac{2}{\zeta-1}\log\Big\{\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}\Big[\exp\Big\{2N\sinh\big(\tfrac{\zeta\beta}{2N}\big)^2\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}\big(M'\big)\Big\}\Big]\Big\} + \log\Big\{\Big(\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}\Big)^{\otimes 2}\Big[\exp\Big\{N\Big[\exp\big\{(\zeta-1)^{-1}\log\big[\cosh\big(\tfrac{\zeta\beta}{N}\big)\big]\big\}-1\Big]M'\Big\}\Big]\Big\} - \frac{\zeta+1}{\zeta-1}\Big\{2\log\big[\nu(\beta)\mu(i)\big]+\log\big(\tfrac{\eta}{2}\big)\Big\}.$$

Let us put

$$\overline{S}_\lambda\big[(i,\beta),(j,\beta')\big]\stackrel{\text{def}}{=}\frac{N}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\inf_{\gamma\in\mathbb{R}_+}\gamma^{-1}\Big\{

$$\log\Big[\Big(\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}\otimes\pi^j_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta'}{N})R]}\Big)\Big\{\exp\big[\gamma\Phi_{-\frac{\gamma}{N}}(M')\big]\Big\}\Big]$$
$$+\frac{C^i\big[\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N}),\zeta\beta\big]-\log(\frac{\overline{\eta}}{3})}{\zeta-1}$$
$$+\frac{C^j\big[\frac{N}{\zeta}\sinh(\frac{\zeta\beta'}{N}),\zeta\beta'\big]-\log(\frac{\overline{\eta}}{3})}{\zeta-1}-\log(\tfrac{\overline{\eta}}{3})\Big\}$$
$$+\frac{1}{\lambda}\Big[\overline{\mathcal{C}}(i,\beta)+\overline{\mathcal{C}}(j,\beta')-\frac{\zeta+1}{\zeta-1}\log\big[3^{-1}\nu(\lambda)\epsilon\big]\Big],$$

where

$$\overline{\eta}=\nu(\gamma)\nu(\beta)\nu(\beta')\mu(i)\mu(j)\eta.$$

Let us remark that

$$\overline{S}_\lambda\big[(i,\beta),(j,\beta')\big]\leq\inf_{\gamma\in\mathbb{R}_+}$$
$$\frac{\lambda}{2N\gamma}\log\Big[\Big(\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}\otimes\pi^j_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta'}{N})R]}\Big)\Big\{$$
$$\exp\Big[N\big[\exp(\tfrac{\gamma}{N})-1\big]M'\Big]\Big\}\Big]$$
$$+\Big(\frac{\lambda}{2N\gamma(\zeta-1)}+\frac{2}{\lambda(\zeta-1)}\Big)\log\Big\{\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}\Big[$$
$$\exp\Big\{2N\sinh\big(\tfrac{\zeta\beta}{2N}\big)^2\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}(M')\Big\}\Big]\Big\}$$
$$+\lambda^{-1}\log\Big\{\Big(\pi^i_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta}{N})R]}\Big)^{\otimes2}\Big[$$
$$\exp\Big\{N\Big[\exp\big\{(\zeta-1)^{-1}\log\big[\cosh\big(\tfrac{\zeta\beta}{N}\big)\big]\big\}-1\Big]M'\Big\}\Big]\Big\}$$
$$+\Big(\frac{\lambda}{2N\gamma(\zeta-1)}+\frac{2}{\lambda(\zeta-1)}\Big)\log\Big\{\pi^j_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta'}{N})R]}\Big[$$
$$\exp\Big\{2N\sinh\big(\tfrac{\zeta\beta'}{2N}\big)^2\pi^j_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta'}{N})R]}(M')\Big\}\Big]\Big\}$$
$$+\lambda^{-1}\log\Big\{\Big(\pi^j_{\exp[-\frac{N}{\zeta}\sinh(\frac{\zeta\beta'}{N})R]}\Big)^{\otimes2}\Big[$$
$$\exp\Big\{N\Big[\exp\big\{(\zeta-1)^{-1}\log\big[\cosh\big(\tfrac{\zeta\beta'}{N}\big)\big]\big\}-1\Big]M'\Big\}\Big]\Big\}$$
$$-\frac{(\zeta+1)\lambda}{2N(\zeta-1)\gamma}\log\big[3^{-1}\nu(\gamma)\nu(\beta)\nu(\beta')\mu(i)\mu(j)\eta\big]$$
$$-\frac{(\zeta+1)}{(\zeta-1)\lambda}\Big(2\log\big[2^{-1}\nu(\beta)\nu(\beta')\mu(i)\mu(j)\eta\big]+\log\big[3^{-1}\nu(\lambda)\epsilon\big]\Big).$$

Let us define accordingly

$$\overline{B}\big[(i,\beta),(j,\beta')\big] \stackrel{\text{def}}{=} \inf_\lambda \Xi_{\frac{\lambda}{N}} \Bigg\{ \inf_{\alpha,\gamma,\alpha',\gamma'} \Bigg[\pi^j_{\exp(-\alpha' R)} \otimes \pi^i_{\exp(-\alpha R)} \big[\Psi_{-\frac{\lambda}{N}}(R',M')\big] - \frac{\log\big(\frac{\widetilde{\eta}}{3}\big)}{\lambda} + \frac{C^j(\alpha',\gamma') - \log\big(\frac{\widetilde{\eta}}{3}\big)}{\frac{N\beta'}{\alpha'}\sinh(\frac{\gamma'}{N}) - \gamma'} + \frac{C^i(\alpha,\gamma) - \log\big(\frac{\widetilde{\eta}}{3}\big)}{\gamma - \frac{N\beta}{\alpha}\sinh(\frac{\gamma}{N})} \Bigg] + \overline{S}_\lambda\big[(i,\beta),(j,\beta')\big] \Bigg\},$$

where

$$\widetilde{\eta} = \nu(\lambda)\nu(\alpha)\nu(\gamma)\nu(\beta)\nu(\alpha')\nu(\gamma')\nu(\beta')\mu(i)\mu(j)\eta.$$

Proposition 2.2.7.

- *With* $\mathbb{P}$ *probability at least* $1-\eta$*, for any* $\beta \in \mathbb{R}_+$ *and* $i \in \mathbb{N}$, $\mathcal{C}(\pi^i_{\exp(-\beta r)}) \le \overline{\mathcal{C}}(i,\beta)$;
- *With* $\mathbb{P}$ *probability at least* $1-3\eta$*, for any* $\lambda,\beta,\beta' \in \mathbb{R}_+$*, any* $i,j \in \mathbb{N}$, $S_\lambda\big[(i,\beta),(j,\beta')\big] \le \overline{S}_\lambda\big[(i,\beta),(j,\beta')\big]$;
- *With* $\mathbb{P}$ *probability at least* $1-4\eta$*, for any* $i,j \in \mathbb{N}$*, any* $\beta,\beta' \in \mathbb{R}_+$, $B(\pi^i_{\exp(-\beta r)}, \pi^j_{\exp(-\beta' r)}) \le \overline{B}\big[(i,\beta),(j,\beta')\big]$.

It is also interesting to find a non-random lower bound for $\mathcal{C}(\pi^i_{\exp(-\beta r)})$. Let us start from the fact that with $\mathbb{P}$ probability at least $1-\eta$,

$$\pi^i_{\exp(-\alpha R)} \otimes \pi^i_{\exp(-\alpha R)}\big[\Phi_{\frac{\gamma'}{N}}(M')\big] \le \pi^i_{\exp(-\alpha R)} \otimes \pi^i_{\exp(-\alpha R)}(m') - \frac{\log(\eta)}{\gamma'}.$$

On the other hand, we already proved that with $\mathbb{P}$ probability at least $1-\eta$,

$$0 \le \Big(1 - \frac{\alpha}{N\tanh(\frac{\lambda}{N})}\Big)\mathcal{K}\big[\rho, \pi^i_{\exp(-\alpha R)}\big] \le \frac{\alpha}{N\tanh(\frac{\lambda}{N})}\Big\{\lambda\big[\rho(r) - \pi^i_{\exp(\alpha R)}(r)\big] + N\log\big[\cosh(\tfrac{\lambda}{N})\big]\rho\otimes\pi^i_{\exp(-\alpha R)}(m') - \log(\eta)\Big\} + \mathcal{K}\big(\rho,\pi^i\big) - \mathcal{K}\big(\pi^i_{\exp(-\alpha R)},\pi^i\big).$$

Thus for any $\xi > 0$, putting $\beta = \frac{\alpha\lambda}{N\tanh(\frac{\lambda}{N})}$, with $\mathbb{P}$ probability at least $1-\eta$,

$$\xi\pi^i_{\exp(-\alpha R)} \otimes \pi^i_{\exp(-\alpha R)}\big[\Phi_{\frac{\gamma'}{N}}(M')\big] \le \pi^i_{\exp(-\alpha R)}\Big\{\log\Big[\pi^i_{\exp(-\beta r)}\Big\{\exp\Big[\beta\tfrac{N}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\pi^i_{\exp(-\beta r)}(m') + \xi m'\Big]\Big\}\Big]\Big\}$$

$$
\begin{aligned}
&\qquad\qquad - \left(\frac{\beta}{\lambda} + \frac{\xi}{\gamma'}\right) \log\left(\frac{\eta}{2}\right) \\
&\leq \log\Big\{ \pi^i_{\exp(-\beta r)} \Big[\exp\Big\{ \beta \tfrac{N}{\lambda} \log\big[\cosh(\tfrac{\lambda}{N})\big] \pi^i_{\exp(-\beta r)}(m') \Big\} \\
&\qquad \times \pi^i_{\exp(-\beta r)} \Big\{ \exp\Big[\beta \tfrac{N}{\lambda} \log\big[\cosh(\tfrac{\lambda}{N})\big] \pi^i_{\exp(-\beta r)}(m') + \xi m' \Big] \Big\} \Big] \Big\} \\
&\qquad\qquad - \left(2\frac{\beta}{\lambda} + \frac{\xi}{\gamma'}\right) \log\left(\frac{\eta}{2}\right) \\
&\leq 2 \log\Big\{ \pi^i_{\exp(-\beta r)} \Big[\exp\Big\{ \Big[\xi + \beta \tfrac{N}{\lambda} \log\big[\cosh(\tfrac{\lambda}{N})\big] \Big] \pi^i_{\exp(-\beta r)}(m') \Big\} \Big] \Big\} \\
&\qquad\qquad - \left(2\frac{\beta}{\lambda} + \frac{\xi}{\gamma'}\right) \log\left(\frac{\eta}{2}\right) \\
&\leq 2 \log\Big\{ \pi^i_{\exp(-\beta r)} \Big[\exp\Big\{ \Big[\xi + \tfrac{\beta\lambda}{2N} \Big] \pi^i_{\exp(-\beta r)}(m') \Big\} \Big] \Big\} \\
&\qquad\qquad - \left(\frac{2\beta}{\lambda} + \frac{\xi}{\gamma'}\right) \log\left(\frac{\eta}{2}\right).
\end{aligned}
$$

Taking $\xi = \frac{\beta\lambda}{2N}$, we get with $\mathbb{P}$ probability at least $1 - \eta$

$$
\begin{aligned}
&\frac{\beta\lambda}{4N} \Big(\pi^i_{\exp[-\beta \frac{N}{\lambda} \tanh(\frac{\lambda}{N}) R]} \Big)^{\otimes 2} \Big[\Phi_{\frac{\gamma'}{N}}(M') \Big] \\
&\qquad \leq \log\Big\{ \pi^i_{\exp(-\beta r)} \Big[\exp\Big\{ \frac{\beta\lambda}{N} \pi^i_{\exp(-\beta r)}(m') \Big\} \Big] \Big\} \\
&\qquad\qquad - \left(\frac{2\beta}{\lambda} + \frac{\beta\lambda}{2N\gamma'}\right) \log\left(\frac{\eta}{2}\right).
\end{aligned}
$$

Putting

$$
\begin{aligned}
\lambda &= \frac{N^2}{\gamma} \log\big[\cosh(\tfrac{\gamma}{N})\big] \\
\text{and } \Upsilon(\gamma) &\stackrel{\text{def}}{=} \frac{\gamma \tanh\big\{ \frac{N}{\gamma} \log\big[\cosh(\frac{\gamma}{N})\big] \big\}}{N \log\big[\cosh(\frac{\gamma}{N})\big]} \underset{\gamma \to 0}{\sim} 1,
\end{aligned}
$$

this can be rewritten as

$$
\begin{aligned}
&\frac{\beta N}{4\gamma} \log\big[\cosh(\tfrac{\gamma}{N})\big] \Big(\pi^i_{\exp(-\beta \Upsilon(\gamma) R)} \Big)^{\otimes 2} \Big[\Phi_{\frac{\gamma'}{N}}(M') \Big] \\
&\qquad \leq \log\Big\{ \pi^i_{\exp(-\beta r)} \Big[\exp\Big\{ \beta \tfrac{N}{\gamma} \log\big[\cosh(\tfrac{\gamma}{N})\big] \pi^i_{\exp(-\beta r)}(m') \Big\} \Big] \Big\} \\
&\qquad\qquad - \left(\frac{2\beta\gamma}{N^2 \log\big[\cosh(\frac{\gamma}{N})\big]} + \frac{\beta N \log\big[\cosh(\frac{\gamma}{N})\big]}{2\gamma\gamma'} \right) \log\left(\frac{\eta}{2}\right).
\end{aligned}
$$

It is now tempting to simplify the picture a little bit by setting $\gamma' = \gamma$, leading to

PROPOSITION 2.2.8. *With $\mathbb{P}$ probability at least $1-\eta$, for any $i \in \mathbb{N}$, any $\beta \in \mathbb{R}_+$,*

$$\begin{aligned}\mathcal{C}\big[\pi^i_{\exp(-\beta r)}\big] &\geq \underline{\mathcal{C}}(i,\beta)\\ &\stackrel{\text{def}}{=} \frac{1}{\zeta-1}\Bigg\{\frac{N}{4}\log\big[\cosh(\tfrac{\zeta\beta}{N})\big]\Big(\pi^i_{\exp(-\beta\Upsilon(\zeta\beta)R)}\Big)^{\otimes 2}\Big[\Phi_{\frac{\zeta\beta}{N}}\big(M'\big)\Big]\\ &\quad+\left(\frac{2\zeta^2\beta^2}{N^2\log\big[\cosh(\frac{\zeta\beta}{N})\big]}+\frac{N\log\big[\cosh(\frac{\zeta\beta}{N})\big]}{2\zeta\beta}\right)\log\big[2^{-1}\nu(\beta)\mu(i)\eta\big]\\ &\qquad\qquad-(\zeta+1)\Big\{\log\big[\nu(\beta)\mu(i)\big]+2^{-1}\log\big(3^{-1}\epsilon\big)\Big\}\Bigg\},\end{aligned}$$

where $\mathcal{C}\big[\pi^i_{\exp(-\beta r)}\big]$ is defined by equation (2.18, page 96).

We are now going to analyse Theorem 2.2.4 (page 93). For this, we will also need an upper bound for $S_\lambda(\rho,\rho')$, defined by equation (2.13, page 91), using M' and empirical complexities, because of the special relations between empirical complexities induced by the selection algorithm. To this purpose, a useful alternative to Proposition 2.2.6 (page 97) is to write, with $\mathbb{P}$ probability at least $1-\eta$,

$$\begin{aligned}\gamma\rho\otimes\rho'(m') &\leq \gamma\rho\otimes\rho'\big[\Phi_{-\frac{\gamma}{N}}\big(M'\big)\big]\\ &\qquad+\mathcal{K}\big[\rho,\pi^i_{\exp(-\lambda R)}\big]+\mathcal{K}\big[\rho',\pi^j_{\exp(-\lambda' R)}\big]-\log(\eta),\end{aligned}$$

and thus at least with $\mathbb{P}$ probability $1-3\eta$,

$$\begin{aligned}\gamma\rho\otimes\rho'(m') &\leq \gamma\rho\otimes\rho'\big[\Phi_{-\frac{\gamma}{N}}\big(M'\big)\big]\\ &\qquad+(1-\zeta^{-1})^{-1}\Big\{\mathcal{K}\big[\rho,\pi^i_{\exp(-\beta r)}\big]\\ &\quad+\log\Big\{\pi^i_{\exp(-\beta r)}\Big[\exp\big\{\tfrac{N}{\zeta}\log\big[\cosh\big(\tfrac{\zeta\beta}{N}\big)\big]\rho(m')\big\}\Big]\Big\}-\zeta^{-1}\log(\eta)\Big\}\\ &\qquad+(1-\zeta^{-1})^{-1}\Big\{\mathcal{K}\big[\rho,\pi^j_{\exp(-\beta' r)}\big]\\ &\quad+\log\Big\{\pi^j_{\exp(-\beta' r)}\Big[\exp\big\{\tfrac{N}{\zeta}\log\big[\cosh\big(\tfrac{\zeta\beta'}{N}\big)\big]\rho(m')\big\}\Big]\Big\}-\zeta^{-1}\log(\eta)\Big\}\\ &\qquad\qquad-\log(\eta).\end{aligned}$$

When $\rho=\pi^i_{\exp(-\beta r)}$ and $\rho'=\pi^j_{\exp(-\beta' r)}$, we get with $\mathbb{P}$ probability at least $1-\eta$, for any β, β', $\gamma\in\mathbb{R}_+$, any i, $j\in\mathbb{N}$,

$$\begin{aligned}\gamma\rho\otimes\rho'(m') &\leq \gamma\rho\otimes\rho'\big[\Phi_{-\frac{\gamma}{N}}\big[\big(M'\big)\big]\\ &\qquad+\mathcal{C}(\rho)+\mathcal{C}(\rho')-\frac{\zeta+1}{\zeta-1}\Big[\log\big[3^{-1}\nu(\gamma)\eta\big]\Big].\end{aligned}$$

PROPOSITION 2.2.9. *With* $\mathbb{P}$ *probability at least* $1-\eta$*, for any* $\rho = \pi^i_{\exp(-\beta r)}$*, any* $\rho' = \pi^j_{\exp(-\beta' r)} \in \mathcal{P}$*,*

$$S_\lambda(\rho,\rho') \leq \frac{N}{\lambda}\log\bigl[\cosh(\tfrac{\lambda}{N})\bigr]\rho\otimes\rho'\bigl[\Phi_{-\frac{\gamma}{N}}\bigl(M'\bigr)\bigr] + \frac{1+\frac{N}{\gamma}\log\bigl[\cosh(\frac{\lambda}{N})\bigr]}{\lambda}\bigl[\mathcal{C}(\rho)+\mathcal{C}(\rho')\bigr] - \frac{(\zeta+1)}{(\zeta-1)\lambda}\biggl\{\log\bigl[3^{-1}\nu(\lambda)\epsilon\bigr]+\tfrac{N}{\gamma}\log\bigl[\cosh\bigl(\tfrac{\lambda}{N}\bigr)\bigr]\log\bigl[3^{-1}\nu(\gamma)\eta\bigr]\biggr\}.$$

In order to analyse Theorem 2.2.4 (page 93), we need to index $\mathcal{P} = \bigl\{\rho_1,\dots,\rho_M\bigr\}$ in order of increasing empirical complexity $\mathcal{C}(\rho)$. To deal in a convenient way with this indexation, we will write $\overline{\mathcal{C}}(i,\beta)$ as $\overline{\mathcal{C}}\bigl[\pi^i_{\exp(-\beta r)}\bigr]$, $\underline{\mathcal{C}}(i,\beta)$ as $\underline{\mathcal{C}}\bigl[\pi^i_{\exp(-\beta r)}\bigr]$, and $\overline{S}\bigl[(i,\beta),(j,\beta')\bigr]$ as $\overline{S}\bigl[\pi^i_{\exp(-\beta r)},\pi^j_{\exp(-\beta' r)}\bigr]$.

With $\mathbb{P}$ probability at least $1-\epsilon$, when $\widehat{t}\leq j<\widehat{k}$, as we already saw,

$$\rho_{\widehat{k}}(R)\leq\rho_i(R)\leq\rho_j(R)+\inf_{\lambda\in\mathbb{R}_+}\Xi_{\frac{\lambda}{N}}\bigl[2S_\lambda(\rho_j,\rho_i)\bigr],$$

where $i=t(j)<\widehat{t}$. Therefore, with $\mathbb{P}$ probability at least $1-\epsilon-\eta$,

$$\rho_i(R)\leq\rho_j(R)+\inf_{\lambda\in\mathbb{R}_+}\Xi_{\frac{\lambda}{N}}\biggl\{2\frac{N}{\lambda}\log\bigl[\cosh\bigl(\tfrac{\lambda}{N}\bigr)\bigr]\rho_j\otimes\rho_i\bigl[\Phi_{-\frac{\gamma}{N}}\bigl(M'\bigr)\bigr] + 4\frac{1+\frac{N}{\gamma}\log\bigl[\cosh\bigl(\frac{\lambda}{N}\bigr)\bigr]}{\lambda}\mathcal{C}(\rho_j) - \frac{(\zeta+1)}{(\zeta-1)\lambda}\biggl\{\log\bigl[3^{-1}\nu(\lambda)\epsilon\bigr]+\tfrac{N}{\gamma}\log\bigl[\cosh\bigl(\tfrac{\lambda}{N}\bigr)\bigr]\log\bigl[3^{-1}\nu(\gamma)\eta\bigr]\biggr\}\biggr\}.$$

We can now remark that

$$\Xi_a(p+q)\leq\Xi_a(p)+q\Xi'_a(p)q\leq\Xi_a(p)+\Xi'_a(0)q=\Xi_a(p)+\frac{a}{\tanh(a)}q$$

and that

$$\Phi_{-a}(p+q)\leq\Phi_{-a}(p)+\Phi'_{-a}(0)q=\Phi_{-a}(p)+\frac{\exp(a)-1}{a}q.$$

Moreover, assuming as usual without substantial loss of generality that there exists $\widetilde{\theta}\in\arg\min_\Theta R$, we can split $M'(\theta,\theta')\leq M'(\theta,\widetilde{\theta})+M'(\widetilde{\theta},\theta')$. Let us then consider the *expected margin function* defined by

$$\varphi(y)=\sup_{\theta\in\Theta}M'(\theta,\widetilde{\theta})-yR'(\theta,\widetilde{\theta}),\quad y\in\mathbb{R}_+,$$

and let us write for any $y\in\mathbb{R}_+$,

$$\rho_j \otimes \rho_i\big[\Phi_{-\frac{\lambda}{N}}\big(M'\big)\big] \le \rho_j \otimes \rho_i\big\{\Phi_{-\frac{\gamma}{N}}\big[M'(.,\widetilde{\theta}) + yR'(.,\widetilde{\theta}) + \varphi(y)\big]\big\}$$
$$\le \rho_j\big\{\Phi_{-\frac{\lambda}{N}}\big[M'(.,\widetilde{\theta}) + \varphi(y)\big]\big\} + \frac{Ny\big[\exp(\frac{\gamma}{N}) - 1\big]}{\gamma}\big[\rho_i(R) - R(\widetilde{\theta})\big]$$

and

$$\left(1 - \frac{2yN\big[\exp(\frac{\gamma}{N}) - 1\big]\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\gamma\tanh\big(\frac{\lambda}{N}\big)}\right)\big[\rho_i(R) - R(\widetilde{\theta})\big]$$
$$\le \big[\rho_j(R) - R(\widetilde{\theta})\big] + \Xi_{\frac{\lambda}{N}}\left\{\frac{2N}{\lambda}\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]\rho_j\big\{\Phi_{-\frac{\gamma}{N}}\big[M'(.,\widetilde{\theta}) + \varphi(y)\big]\big\}\right.$$
$$+ 4\frac{1 + \frac{N}{\gamma}\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\lambda}\mathcal{C}(\rho_j)$$
$$\left.- \frac{2(\zeta+1)}{(\zeta-1)\lambda}\Big\{\log\big[3^{-1}\nu(\lambda)\epsilon\big] + \tfrac{N}{\gamma}\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big]\log\big[3^{-1}\nu(\gamma)\eta\big]\Big\}\right\}.$$

With $\mathbb{P}$ probability at least $1 - \epsilon - \eta$, for any λ, γ, x, $y \in \mathbb{R}_+$, any $j \in \big\{\widehat{t}, \ldots, \widehat{k} - 1\big\}$,

$$\rho_{\widehat{k}}(R) - R(\widetilde{\theta}) \le \rho_i(R) - R(\widetilde{\theta})$$
$$\le \left(1 - \frac{2yN\big[\exp(\frac{\gamma}{N}) - 1\big]\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\gamma\tanh\big(\frac{\lambda}{N}\big)}\right)^{-1}\left\{\right.$$
$$\left(1 + \frac{2xN\big[\exp(\frac{\gamma}{N}) - 1\big]\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\gamma\tanh\big(\frac{\lambda}{N}\big)}\right)\big[\rho_j(R) - R(\widetilde{\theta})\big]$$
$$+ \Xi_{\frac{\lambda}{N}}\left\{\frac{2N}{\lambda}\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]\Phi_{-\frac{\gamma}{N}}\big[\varphi(x) + \varphi(y)\big]\right.$$
$$+ 4\frac{1 + \frac{N}{\gamma}\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\lambda}\overline{\mathcal{C}}(\rho_j)$$
$$\left.\left.- \frac{2(\zeta+1)}{(\zeta-1)\lambda}\Big\{\log\big[3^{-1}\nu(\lambda)\epsilon\big] + \tfrac{N}{\gamma}\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big]\log\big[3^{-1}\nu(\gamma)\eta\big]\Big\}\right\}\right\}.$$

Now we have to get an upper bound for $\rho_j(R)$. We can write $\rho_j = \pi^{\ell}_{\exp(-\beta' r)}$, as we assumed that all the posterior distributions in $\mathcal{P}$ are of this special form. Moreover, we already know from Theorem 2.1.8 (page 77) that with $\mathbb{P}$ probability at least $1 - \eta$,

$$\big[N\sinh\big(\tfrac{\beta'}{N}\big) - \beta'\zeta^{-1}\big]\big[\pi^{\ell}_{\exp(-\beta' r)}(R) - \pi^{\ell}_{\exp(-\beta'\zeta^{-1}R)}(R)\big]$$
$$\le C^{\ell}(\beta'\zeta^{-1}, \beta') - \log\big[\nu(\beta')\mu(\ell)\eta\big].$$

This proves that with $\mathbb{P}$ probability at least $1 - \epsilon - 2\eta$,

$$
\begin{aligned}
\rho_{\widehat{k}}(R) &\le R(\widetilde{\theta}) \\
&+ \left(1 - \frac{2yN\big[\exp\big(\frac{\gamma}{N}\big) - 1\big]\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\gamma\tanh\big(\frac{\lambda}{N}\big)}\right)^{-1}\Bigg\{ \\
&\left(1 + \frac{2xN\big[\exp\big(\frac{\gamma}{N}\big) - 1\big]\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\gamma\tanh\big(\frac{\lambda}{N}\big)}\right) \\
&\quad\times\left(\pi^{\ell}_{\exp(-\zeta^{-1}\beta' R)}(R) - R(\widetilde{\theta}) + \frac{C^{\ell}(\zeta^{-1}\beta',\beta') - \log\big[\nu(\beta')\mu(\ell)\eta\big]}{N\sinh(\frac{\beta'}{N}) - \zeta^{-1}\beta'}\right) \\
&\qquad + \Xi_{\frac{\lambda}{N}}\Bigg\{\frac{2N}{\lambda}\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big]\Phi_{-\frac{\gamma}{N}}\big[\varphi(x) + \varphi(y)\big] \\
&\qquad\qquad + 4\frac{1 + \frac{N}{\gamma}\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\lambda}\overline{\mathcal{C}}(\ell,\beta') \\
&\qquad - \frac{2(\zeta+1)}{(\zeta-1)\lambda}\Big\{\log\big[3^{-1}\nu(\lambda)\epsilon\big] + \tfrac{N}{\gamma}\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big]\log\big[3^{-1}\nu(\gamma)\eta\big]\Big\}\Bigg\}\Bigg\}.
\end{aligned}
$$

The case when $j \in \big\{\widehat{k}+1,\dots,M\big\} \setminus (\arg\max t)$ is dealt with exactly in the same way, with $i = t(j)$ replaced directly with $\widehat{k}$ itself, leading to the same inequality.

The case when $j \in (\arg\max t)$ is dealt with bounding first $\rho_{\widehat{k}}(R) - R(\widetilde{\theta})$ in terms of $\rho_{\widehat{t}}(R) - R(\widetilde{\theta})$, and this latter in terms of $\rho_j(R) - R(\widetilde{\theta})$. Let us put

$$
\begin{aligned}
A(\lambda,\gamma) &= \left(1 - \frac{2xN\big[\exp\big(\frac{\gamma}{N}\big) - 1\big]\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\gamma\tanh\big(\frac{\lambda}{N}\big)}\right), \\
B(\lambda,\gamma) &= 1 + \frac{2yN\big[\exp\big(\frac{\gamma}{N}\big) - 1\big]\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\gamma\tanh\big(\frac{\lambda}{N}\big)}, \\
D(\lambda,\gamma,\rho_j) &= \Xi_{\frac{\lambda}{N}}\Bigg\{\frac{2N}{\lambda}\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big]\Phi_{-\frac{\gamma}{N}}\big[\varphi(x) + \varphi(y)\big] \\
&\qquad\qquad + 4\frac{1 + \frac{N}{\gamma}\log\big[\cosh\big(\frac{\lambda}{N}\big)\big]}{\lambda}\overline{\mathcal{C}}(\rho_j) \\
&\qquad - \frac{2(\zeta+1)}{(\zeta-1)\lambda}\Big\{\log\big[3^{-1}\nu(\lambda)\epsilon\big] \\
&\qquad\qquad + \frac{N}{\gamma}\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big]\log\big[3^{-1}\nu(\gamma)\eta\big]\Big\}\Bigg\},
\end{aligned}
\tag{2.20}
$$

where $\overline{\mathcal{C}}(\rho_j) = \overline{\mathcal{C}}(\ell,\beta')$ is defined, when $\rho_j = \pi^{\ell}_{\exp(-\beta' r)}$, by equation (2.19, page 98). We obtain, still with $\mathbb{P}$ probability $1 - \epsilon - 2\eta$,

$$
\begin{aligned}
\rho_{\widehat{k}}(R) - R(\widetilde{\theta}) &\le \frac{B(\lambda,\gamma)}{A(\lambda,\gamma)}\big[\rho_{\widehat{t}}(R) - R(\widetilde{\theta})\big] + \frac{D(\lambda,\gamma,\rho_j)}{A(\lambda,\gamma)}, \\
\rho_{\widehat{t}}(R) - R(\widetilde{\theta}) &\le \frac{B(\lambda,\gamma)}{A(\lambda,\gamma)}\big[\rho_j(R) - R(\widetilde{\theta})\big] + \frac{D(\lambda,\gamma,\rho_j)}{A(\lambda,\gamma)}.
\end{aligned}
$$

The use of the factor $D(\lambda,\gamma,\rho_j)$ in the first of these two inequalities, instead of $D(\lambda,\gamma,\rho_{\widehat{t}})$, is justified by the fact that $\mathcal{C}(\rho_{\widehat{t}}) \leq \mathcal{C}(\rho_j)$. Combining the two we get

$$\rho_{\widehat{k}}(R) \leq R(\widetilde{\theta}) + \frac{B(\lambda,\gamma)^2}{A(\lambda,\gamma)^2}\big[\rho_j(R) - R(\widetilde{\theta})\big] + \left[\frac{B(\lambda,\gamma)}{A(\lambda,\gamma)} + 1\right]\frac{D(\lambda,\gamma,\rho_j)}{A(\lambda,\gamma)}.$$

Since it is the worst bound of all cases, it holds for any value of j, proving

THEOREM 2.2.10. *With $\mathbb{P}$ probability at least $1-\epsilon-2\eta$,*

$$\begin{aligned}
\rho_{\widehat{k}}(R) \leq R(\widetilde{\theta}) + \inf_{i,\beta,\lambda,\gamma,x,y} \Bigg\{ & \\
& \frac{B(\lambda,\gamma)^2}{A(\lambda,\gamma)^2}\Big[\pi^i_{\exp(-\beta r)}(R) - R(\widetilde{\theta})\Big] + \left[\frac{B(\lambda,\gamma)}{A(\lambda,\gamma)} + 1\right]\frac{D(\lambda,\gamma,\pi^i_{\exp(-\beta r)})}{A(\lambda,\gamma)} \Bigg\} \\
& \leq R(\widetilde{\theta}) + \inf_{i,\beta,\lambda,\gamma,x,y} \Bigg\{ \\
& \frac{B(\lambda,\gamma)^2}{A(\lambda,\gamma)^2}\left(\pi^i_{\exp(-\zeta^{-1}\beta R)}(R) - R(\widetilde{\theta}) + \frac{C^i(\zeta^{-1}\beta,\beta) - \log\big[\nu(\beta)\mu(i)\eta\big]}{N\sinh\big(\frac{\beta}{N}\big) - \zeta^{-1}\beta}\right) \\
& \qquad + \left[\frac{B(\lambda,\gamma)}{A(\lambda,\gamma)} + 1\right]\frac{D(\lambda,\gamma,\pi^i_{\exp(-\beta r)})}{A(\lambda,\gamma)} \Bigg\},
\end{aligned}$$

where the notation $A(\lambda,\gamma)$, $B(\lambda,\gamma)$ and $D(\lambda,\gamma,\rho)$ is defined by equation (2.20 page 105) and where the notation $C^i(\beta,\gamma)$ is defined in Proposition 2.2.5 (page 96).

The bound is a little involved, but as we will prove next, it gives the same rate as Theorem 2.1.15 (page 85) and its corollaries, when we work with a single model (meaning that the support of μ is reduced to one point) and the goal is to choose adaptively the temperature of the Gibbs posterior, except for the appearance of the union bound factor $-\log\big[\nu(\beta)\big]$ which can be made of order $\log\big[\log(N)\big]$ without spoiling the order of magnitude of the bound.

We will encompass the case when one must choose between possibly several parametric models. Let us assume that each π^i is supported by some measurable parameter subset Θ_i (meaning that $\pi^i(\Theta_i) = 1$), let us also assume that the behaviour of π^i is parametric in the sense that there exists a dimension $d_i \in \mathbb{R}_+$ such that

$$\sup_{\beta\in\mathbb{R}_+} \beta\big[\pi^i_{\exp(-\beta R)}(R) - \inf_{\Theta_i} R\big] \leq d_i. \tag{2.21}$$

Then

$$C^i(\lambda,\gamma) \leq \log\left\{\pi^i_{\exp(-\lambda R)}\left[\exp\left\{2N\sinh\big(\tfrac{\gamma}{2N}\big)^2 M'(.,\widetilde{\theta})\right\}\right]\right\}$$

$$
\begin{aligned}
&\quad + 2N \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 \pi^i_{\exp(-\lambda R)}\bigl[M'(.,\widetilde{\theta})\bigr] \\
&\le \log\biggl\{\pi^i_{\exp(-\lambda R)}\biggl[\exp 2xN \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2\bigl[R - R(\widetilde{\theta})\bigr]\biggr\}\biggr] \\
&\quad + 2xN \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 \pi^i_{\exp(-\lambda R)}\bigl[R - R(\widetilde{\theta})\bigr] \\
&\quad + 4N \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 \varphi(x) \\
&\le 2xN \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 \pi^i_{\exp\{-[\lambda - 2xN\sinh(\frac{\gamma}{2N})^2]R\}}\bigl[R - R(\widetilde{\theta})\bigr] \\
&\quad + 2xN \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 \pi^i_{\exp(-\lambda R)}\bigl[R - R(\widetilde{\theta})\bigr] \\
&\quad + 4N \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 \varphi(x).
\end{aligned}
$$

Thus

$$
\begin{aligned}
C^i(\lambda,\gamma) \le 4N \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 \Biggl(& x\bigl[\inf_{\Theta_i} R - R(\widetilde{\theta})\bigr] + \varphi(x) \\
& + \frac{xd_i}{2\lambda} + \frac{xd_i}{2\lambda - 4xN\sinh\bigl(\frac{\gamma}{2N}\bigr)^2}\Biggr).
\end{aligned}
$$

In the same way,

$$
\begin{aligned}
\overline{\mathcal{C}}(i,\beta) \le \tfrac{8N}{\zeta-1} \sinh\bigl(\tfrac{\zeta\beta}{2N}\bigr)^2 \Biggl[& x\bigl[\inf_{\Theta_i} R - R(\widetilde{\theta})\bigr] + \varphi(x) \\
& + \frac{\zeta x d_i}{2N\sinh\bigl(\frac{\zeta\beta}{N}\bigr)}\biggl(1 + \frac{1}{1 - x\zeta\tanh\bigl(\frac{\zeta\beta}{2N}\bigr)}\biggr)\Biggr] \\
& + 2N\Bigl[\exp\Bigl(\tfrac{\zeta^2\beta^2}{2N^2(\zeta-1)}\Bigr) - 1\Bigr]\Biggl(\varphi(x) + x\bigl[\inf_{\Theta_i} R - R(\widetilde{\theta})\bigr] \\
& + \frac{x\zeta d_i}{N\sinh\bigl(\frac{\zeta\beta}{N}\bigr) - x\zeta N\bigl[\exp\bigl(\frac{\zeta^2\beta^2}{2N^2(\zeta-1)}\bigr) - 1\bigr]}\Biggr) \\
& - \frac{(\zeta+1)}{(\zeta-1)}\Bigl[2\log\bigl[\nu(\beta)\mu(i)\bigr] + \log\bigl(\tfrac{\eta}{2}\bigr)\Bigr].
\end{aligned}
$$

In order to keep the right order of magnitude while simplifying the bound, let us consider

$$
\begin{aligned}
C_1 = \max\Biggl\{ & \zeta - 1, \Bigl(\tfrac{2N}{\zeta\beta_{\max}}\Bigr)^2 \sinh\Bigl(\tfrac{\zeta\beta_{\max}}{2N}\Bigr)^2, \\
& \tfrac{2N^2(\zeta-1)}{\zeta^2\beta^2_{\max}}\Bigl[\exp\Bigl(\tfrac{\zeta^2\beta^2_{\max}}{2N^2(\zeta-1)}\Bigr) - 1\Bigr]\Biggr\}. \qquad (2.22)
\end{aligned}
$$

Then, for any $\beta \in (0, \beta_{\max})$,

$$\overline{\mathcal{C}}(i,\beta) \le \inf_{y\in\mathbb{R}_+} \frac{3C_1\zeta^2\beta^2}{(\zeta-1)N}\left[y\big[\inf_{\Theta_i} R - R(\widetilde{\theta})\big] + \varphi(y) + \frac{yd_i}{\beta\big[1-\frac{yC_1\zeta^2\beta}{2(\zeta-1)N}\big]}\right] - \frac{(\zeta+1)}{(\zeta-1)}\left[2\log\big[\nu(\beta)\mu(i)\big] + \log\big(\tfrac{\eta}{2}\big)\right].$$

Thus

$$D\big[\lambda,\gamma,\pi^i_{\exp(-\beta r)}\big] \le \frac{\lambda}{N\tanh\big(\frac{\lambda}{N}\big)}\left\{\frac{\lambda\big[\exp\big(\frac{\gamma}{N}\big)-1\big]}{\gamma}\big[\varphi(x)+\varphi(y)\big] + 4\frac{1+\frac{\lambda^2}{2N\gamma}}{\lambda}\left[\frac{3C_1\zeta^2\beta^2}{(\zeta-1)N}\left(z\big[\inf_{\Theta_i} R - R(\widetilde{\theta})\big] + \varphi(z) + \frac{zd_i}{\beta\big[1-\frac{zC_1\zeta^2\beta}{2(\zeta-1)N}\big]}\right) - \frac{(\zeta+1)}{(\zeta-1)}\left[2\log\big[\nu(\beta)\mu(i)\big] + \log\big(\tfrac{\eta}{2}\big)\right]\right] - \frac{2(\zeta+1)}{(\zeta-1)\lambda}\left[\log\big[3^{-1}\nu(\lambda)\epsilon\big] + \frac{\lambda^2}{2N\gamma}\log\big[3^{-1}\nu(\gamma)\eta\big]\right]\right\}$$

If we are not seeking tight constants, we can take for the sake of simplicity $\lambda=\gamma=\beta$, $x=y$ and $\zeta=2$.

Let us put

$$C_2 = \max\left\{C_1, \frac{N\big[\exp\big(\frac{\beta_{\max}}{N}\big)-1\big]}{\beta_{\max}}, \frac{2N\log\big[\cosh\big(\frac{\beta_{\max}}{N}\big)\big]}{\beta_{\max}\tanh\big(\frac{\beta_{\max}}{N}\big)}, \frac{\beta_{\max}}{N\tanh\big(\frac{\beta_{\max}}{N}\big)}\right\}, \quad (2.23)$$

so that

$$A(\beta,\beta)^{-1} \le \left(1-\frac{C_2x\beta}{N}\right)^{-1},$$
$$B(\beta,\beta) \le 1+\frac{C_2x\beta}{N},$$

$$D\big[\beta,\beta,\pi^i_{\exp(-\beta r)}\big] \le C_2^2\frac{2\beta}{N}\varphi(x) + \Big(4+\tfrac{2\beta}{N}\Big)\frac{C_2}{\beta}\left[\frac{12C_1\beta^2}{N}\left(z\big[\inf_{\Theta_i} R - R(\widetilde{\theta})\big] + \varphi(z) + \frac{zd_i}{\beta\big[1-\frac{2zC_1\beta}{N}\big]}\right) - 6\log\big[\nu(\beta)\mu(i)\big] - 3\log\big(\tfrac{\eta}{2}\big)\right] - \frac{6C_2}{\beta}\left[\log\big[3^{-1}\nu(\beta)\epsilon\big] + \frac{\beta}{2N}\log\big[3^{-1}\nu(\beta)\eta\big]\right]$$

and

$$C^i(\zeta^{-1}\beta,\beta) \le \frac{C_1\beta^2}{N}\biggl(x\bigl[\inf_{\Theta_i} R - R(\widetilde{\theta})\bigr] + \varphi(x) + \frac{2xd_i}{\beta\bigl[1-\frac{x\beta}{N}\bigr]}\biggr).$$

This leads to

$$\begin{aligned}\rho_{\widehat{k}}(R) \le R(\widetilde{\theta}) + \inf_{i,\beta}\biggl(\frac{1+\frac{C_2x\beta}{N}}{1-\frac{C_2x\beta}{N}}\biggr)^2\biggl\{&\frac{2d_i}{\beta} + \inf_{\Theta_i} R - R(\widetilde{\theta})\\ &+ \frac{2}{\beta}\biggl[\frac{C_1\beta^2}{N}\biggl(x\bigl[\inf_{\Theta_i} R - R(\widetilde{\theta}\bigr] + \varphi(x) + \frac{2xd_i}{\beta\bigl(1-\frac{x\beta}{N}\bigr)}\biggr)\\ &\qquad - \log\bigl[\nu(\beta)\mu(i)\eta\bigr]\biggr]\biggr\}\\ &+ \frac{2}{\bigl(1-\frac{C_2x\beta}{N}\bigr)^2}\biggl\{C_2^2\frac{2\beta}{N}\varphi(x)\\ &+ \Bigl(4+\tfrac{2\beta}{N}\Bigr)\frac{C_2}{\beta}\biggl[\frac{12C_1\beta^2}{N}\biggl(x\bigl[\inf_{\Theta_i} R - R(\widetilde{\theta})\bigr] + \varphi(x) + \frac{xd_i}{\beta\bigl[1-\frac{2xC_1\beta}{N}\bigr]}\biggr)\\ &\qquad - 6\log\bigl[\nu(\beta)\mu(i)\bigr] - 3\log\bigl(\tfrac{\eta}{2}\bigr)\biggr]\\ &\qquad - \frac{6C_2}{\beta}\biggl[\log\bigl[3^{-1}\nu(\beta)\epsilon\bigr] + \frac{\beta}{2N}\log\bigl[3^{-1}\nu(\beta)\eta\bigr]\biggr]\biggr\}.\end{aligned}$$

We see in this expression that, in order to balance the various factors depending on x it is advisable to choose x such that

$$\inf_{\Theta_i} R - R(\widetilde{\theta}) = \frac{\varphi(x)}{x},$$

as long as $x \le \frac{N}{4C_2\beta}$.

Following Mammen and Tsybakov, let us assume that the usual margin assumption holds: for some real constants $c > 0$ and $\kappa \ge 1$,

$$R(\theta) - R(\widetilde{\theta}) \ge c\bigl[D(\theta,\widetilde{\theta})\bigr]^\kappa.$$

As $D(\theta,\widetilde{\theta}\,) \ge M'(\theta,\widetilde{\theta}\,)$, this also implies the weaker assumption

$$R(\theta) - R(\widetilde{\theta}\,) \ge c\bigl[M'(\theta,\widetilde{\theta})\bigr]^\kappa, \qquad \theta \in \Theta,$$

which we will really need and use. Let us take $\beta_{\max} = N$ and

$$\nu = \frac{1}{\lceil\log_2(N)\rceil}\sum_{k=1}^{\lceil\log_2(N)\rceil}\delta_{2^k}.$$

Then, as we have already seen, $\varphi(x) \le (1-\kappa^{-1})\big(\kappa c x\big)^{-\frac{1}{\kappa-1}}$. Thus $\varphi(x)/x \le bx^{-\frac{\kappa}{\kappa-1}}$, where $b = (1-\kappa^{-1})\big(\kappa c\big)^{-\frac{1}{\kappa-1}}$. Let us choose accordingly

$$x = \min\left\{ x_1 \stackrel{\text{def}}{=} \left(\frac{\inf_{\Theta_i} R - R(\widetilde{\theta})}{b} \right)^{-\frac{\kappa-1}{\kappa}}, x_2 \stackrel{\text{def}}{=} \frac{N}{4C_2\beta} \right\}.$$

Using the fact that when $r \in (0, \frac{1}{2})$, $\big(\frac{1+r}{1-r}\big)^2 \le 1 + 16r \le 9$, we get with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\beta \in \operatorname{supp}\nu$, in the case when $x = x_1 \le x_2$,

$$\begin{aligned}
&\rho_{\widehat{k}}(R) \le \inf_{\Theta_i} R + 538\, C_2^2 \frac{\beta}{N} b^{\frac{\kappa-1}{\kappa}} \big[\inf_{\Theta_i} R - R(\widetilde{\theta})\big]^{\frac{1}{\kappa}} \\
&+ \frac{C_2}{\beta}\left[138\, d_i + 166 \log\big[1+\log_2(N)\big] - 134 \log\big[\mu(i)\big] - 102 \log(\epsilon) + 724\right],
\end{aligned}$$

and in the case when $x = x_2 \le x_1$,

$$\begin{aligned}
&\rho_{\widehat{k}}(R) \le \inf_{\Theta_i} R + 68 C_1 \big[\inf_{\Theta_i} R - R(\widetilde{\theta})\big] + 269\, C_2^2 \frac{\beta}{N}\varphi(x) \\
&+ \frac{C_2}{\beta}\left[138\, d_i + 166 \log\big[1+\log_2(N)\big] - 134 \log\big[\mu(i)\big] - 102 \log(\epsilon) + 724\right] \\
&\le \inf_{\Theta_i} R + 541 C_2^2 \frac{\beta}{N}\varphi(x) \\
&+ \frac{C_2}{\beta}\left[138\, d_i + 166 \log\big[1+\log_2(N)\big] - 134 \log\big[\mu(i)\big] - 102 \log(\epsilon) + 724\right].
\end{aligned}$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\begin{aligned}
\rho_{\widehat{k}}(R) \le \inf_{\Theta_i} R + \inf_{\beta\in(1,N)} 1082\, C_2^2 \frac{\beta}{N} \max\Bigg\{ & b^{\frac{\kappa-1}{\kappa}} \big[\inf_{\Theta_i} R - R(\widetilde{\theta})\big]^{\frac{1}{\kappa}}, \\
& b\left(\frac{4C_2\beta}{N}\right)^{\frac{1}{\kappa-1}} \Bigg\} \\
+ \frac{C_2}{\beta}\Big[& 138\, d_i + 166 \log\big[1+\log_2(N)\big] \\
& - 134 \log\big[\mu(i)\big] - 102 \log(\epsilon) + 724\Big].
\end{aligned}$$

THEOREM 2.2.11. *With probability at least $1-\epsilon$, for any $i \in \mathbb{N}$,*

$$\begin{aligned}
&\rho_{\widehat{k}}(R) \le \inf_{\Theta_i} R \\
&+ \max\left\{ 847 C_2^{\frac{3}{2}} \sqrt{\frac{b^{\frac{\kappa-1}{\kappa}} \big[\inf_{\Theta_i} R - R(\widetilde{\theta})\big]^{\frac{1}{\kappa}} \Big\{ d_i + \log\Big(\frac{1+\log_2(N)}{\epsilon\mu(i)}\Big) + 5 \Big\}}{N}}, \right.
\end{aligned}$$

$$2C_2\big[1082\,b\big]^{\frac{\kappa-1}{2\kappa-1}}4^{\frac{1}{2\kappa-1}}\left\{\frac{166C_2\Big[d_i+\log\Big(\frac{1+\log_2(N)}{\epsilon\mu(i)}\Big)+5\Big]}{N}\right\}^{\frac{\kappa}{2\kappa-1}}\Bigg\},$$

where C_2, given by equation (2.23 page 108), will in most cases be close to 1*, and in any case less than* 3.2*.*

This result gives a bound of the same form as that given in Theorem 2.1.15 (page 85) in the special case when there is only one model — that is when μ is a Dirac mass, for instance $\mu(1) = 1$, implying that $R(\widetilde{\theta}_1) - R(\widetilde{\theta}) = 0$. Morover the parametric complexity assumption we made for this theorem, given by equation (2.21 page 106), is weaker than the one used in Theorem 2.1.15 and described by equation (2.8, page 82). When there is more than one model, the bound shows that the estimator makes a trade-off between model accuracy, represented by $\inf_{\Theta_i} R - R(\widetilde{\theta})$, and dimension, represented by d_i, and that for optimal parametric sub-models, meaning those for which $\inf_{\Theta_i} R = \inf_{\Theta} R$, the estimator does at least as well as the minimax optimal convergence speed in the best of these.

Another point is that we obtain more explicit constants than in Theorem 2.1.15. It is also clear that a more careful choice of parameters could have brought some improvement in the value of these constants.

These results show that the selection scheme described in this section is a good candidate to perform temperature selection of a Gibbs posterior distribution built within a single parametric model in a rate optimal way, as well as a proposal with proven performance bound for model selection.

2.3. Two step localization

2.3.1. Two step localization of bounds relative to a Gibbs prior. Let us reconsider the case where we want to choose adaptively among a family of parametric models. Let us thus assume that the parameter set is a disjoint union of measurable sub-models, so that we can write $\Theta = \sqcup_{m\in M}\Theta_m$, where M is some measurable index set. Let us choose some prior probability distribution on the index set $\mu \in \mathcal{M}_+^1(M)$, and some regular conditional prior distribution $\pi : M \to \mathcal{M}_+^1(\Theta)$, such that $\pi(i, \Theta_i) = 1$, $i \in M$. Let us then study some arbitrary posterior distributions $\nu : \Omega \to \mathcal{M}_+^1(M)$ and $\rho : \Omega \times M :\to \mathcal{M}_+^1(\Theta)$, such that $\rho(\omega, i, \Theta_i) = 1$, $\omega \in \Omega$, $i \in M$. We would like to compare $\nu\rho(R)$ with some doubly localized prior distribution $\mu_{\exp[-\frac{\beta}{1+\zeta_2}\pi_{\exp(-\beta R)}(R)]}\big[\pi_{\exp(-\beta R)}\big](R)$ (where ζ_2 is a positive parameter to be set as needed later on). To ease notation we will define two prior distributions (one being more precisely a conditional distribution) depending on the positive real parameters β and ζ_2, putting

$$\overline{\pi} = \pi_{\exp(-\beta R)} \text{ and } \overline{\mu} = \mu_{\exp[-\frac{\beta}{1+\zeta_2}\overline{\pi}(R)]}. \tag{2.24}$$

Similarly to Theorem 1.4.3 on page 53 we can write for any positive real constants β and γ

$$\mathbb{P}\biggl\{(\overline{\mu}\,\overline{\pi})\otimes(\overline{\mu}\,\overline{\pi})\biggl[\exp\Bigl[-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})R'\bigr] - \gamma r' - N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]m'\Bigr]\biggr]\biggr\}\leq 1,$$

and deduce, using Lemma 1.1.3 on page 16, that

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\nu\in\mathcal{M}_+^1(M)}\sup_{\rho:M\to\mathcal{M}_+^1(\Theta)}\Bigl\{-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})(\nu\rho-\overline{\mu}\,\overline{\pi})(R)\bigr] - \gamma(\nu\rho-\overline{\mu}\,\overline{\pi})(r) - N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr](\nu\rho)\otimes(\overline{\mu}\,\overline{\pi})(m') - \mathcal{K}(\nu,\overline{\mu}) - \nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr]\Bigr\}\biggr]\biggr\}\leq 1. \quad (2.25)$$

This will be our starting point in comparing $\nu\rho(R)$ with $\overline{\mu}\,\overline{\pi}(R)$. However, obtaining an empirical bound will require some supplementary efforts. For each index of the model index set M, we can write in the same way

$$\mathbb{P}\biggl\{\overline{\pi}\otimes\overline{\pi}\biggl[\exp\Bigl[-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})R'\bigr]-\gamma r'-N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]m'\Bigr]\biggr]\biggr\}\leq 1.$$

Integrating this inequality with respect to $\overline{\mu}$ and using Fubini's lemma for positive functions, we get

$$\mathbb{P}\biggl\{\overline{\mu}(\overline{\pi}\otimes\overline{\pi})\biggl[\exp\Bigl[-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})R'\bigr]-\gamma r'-N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]m'\Bigr]\biggr]\biggr\}\leq 1.$$

Note that $\overline{\mu}(\overline{\pi}\otimes\overline{\pi})$ is a probability measure on $M\times\Theta\times\Theta$, whereas $(\overline{\mu}\,\overline{\pi})\otimes(\overline{\mu}\,\overline{\pi})$ considered previously is a probability measure on $(M\times\Theta)\times(M\times\Theta)$. We get as previously

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\nu\in\mathcal{M}_+^1(M)}\sup_{\rho:M\to\mathcal{M}_+^1(\Theta)}\Bigl\{-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})\nu(\rho-\overline{\pi})(R)\bigr] - \gamma\nu(\rho-\overline{\pi})(r) - N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\nu(\rho\otimes\overline{\pi})(m') - \mathcal{K}(\nu,\overline{\mu}) - \nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr]\Bigr\}\biggr]\biggr\}\leq 1. \quad (2.26)$$

Let us finally recall that

$$\mathcal{K}(\nu,\overline{\mu}) = \tfrac{\beta}{1+\zeta_2}(\nu-\overline{\mu})\overline{\pi}(R) + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu), \quad (2.27)$$

$$\mathcal{K}(\rho,\overline{\pi}) = \beta(\rho-\overline{\pi})(R) + \mathcal{K}(\rho,\pi) - \mathcal{K}(\overline{\pi},\pi). \quad (2.28)$$

From equations (2.25), (2.26) and (2.28) we deduce

PROPOSITION 2.3.1. *For any positive real constants β, γ and ζ_2, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$ and any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$
\begin{aligned}
-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})(\nu\rho-\overline{\mu}\,\overline{\pi})(R)\bigr] &- \beta\nu(\rho-\overline{\pi})(R)\\
&\leq \gamma(\nu\rho-\overline{\mu}\,\overline{\pi})(r) + N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr](\nu\rho)\otimes(\overline{\mu}\,\overline{\pi})(m')\\
&\qquad + \mathcal{K}(\nu,\overline{\mu}) + \nu\bigl[\mathcal{K}(\rho,\pi)\bigr] - \nu\bigl[\mathcal{K}(\overline{\pi},\pi)\bigr] + \log\bigl(\tfrac{2}{\epsilon}\bigr).
\end{aligned}
$$

and

$$
\begin{aligned}
-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})\nu(\rho-\overline{\pi})(R)\bigr]&\\
&\leq \gamma\nu(\rho-\overline{\pi})(r) + N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\nu(\rho\otimes\overline{\pi})(m')\\
&\qquad + \mathcal{K}(\nu,\overline{\mu}) + \nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr] + \log\bigl(\tfrac{2}{\epsilon}\bigr),
\end{aligned}
$$

where the prior distribution $\overline{\mu}\,\overline{\pi}$ is defined by equation (2.24) *on page 111 and depends on β and ζ_2.*

Let us put for short

$$
T = \tanh(\tfrac{\gamma}{N}) \text{ and } C = N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr].
$$

We will use an entropy compensation strategy for which we need a couple of entropy bounds. We have according to Proposition 2.3.1, with $\mathbb{P}$ probability at least $1-\epsilon$,

$$
\begin{aligned}
\nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr] &= \beta\nu(\rho-\overline{\pi})(R) + \nu\bigl[\mathcal{K}(\rho,\pi) - \mathcal{K}(\overline{\pi},\pi)\bigr]\\
&\leq \frac{\beta}{NT}\Bigl[\gamma\nu(\rho-\overline{\pi})(r) + C\nu(\rho\otimes\overline{\pi})(m')\\
&\qquad + \mathcal{K}(\nu,\overline{\mu}) + \nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr] + \log(\tfrac{2}{\epsilon})\Bigr]\\
&\qquad + \nu\bigl[\mathcal{K}(\rho,\pi) - \mathcal{K}(\overline{\pi},\pi)\bigr].
\end{aligned}
$$

Similarly

$$
\begin{aligned}
\mathcal{K}(\nu,\overline{\mu}) &= \frac{\beta}{1+\zeta_2}(\nu-\overline{\mu})\overline{\pi}(R) + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu)\\
&\leq \frac{\beta}{(1+\zeta_2)NT}\Bigl[\gamma(\nu-\overline{\mu})\overline{\pi}(r) + C(\nu\overline{\pi})\otimes(\overline{\mu}\,\overline{\pi})(m')\\
&\qquad + \mathcal{K}(\nu,\overline{\mu}) + \log(\tfrac{2}{\epsilon})\Bigr] + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu).
\end{aligned}
$$

Thus, for any positive real constants β, γ and ζ_i, $i = 1,\dots,5$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\nu, \nu_3 : \Omega \to \mathcal{M}_+^1(\Theta)$, any posterior conditional distributions $\rho, \rho_1, \rho_2, \rho_4, \rho_5 : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

$$
\begin{aligned}
&- N \log\big[1 - T(\nu\rho - \overline{\mu}\,\overline{\pi})(R)\big] - \beta\nu(\rho - \overline{\pi})(R) \\
&\qquad \le \gamma(\nu\rho - \overline{\mu}\,\overline{\pi})(r) + C(\nu\rho) \otimes (\overline{\mu}\,\overline{\pi})(m') \\
&\qquad\qquad + \mathcal{K}(\nu, \overline{\mu}) + \nu\big[\mathcal{K}(\rho, \pi) - \mathcal{K}(\overline{\pi}, \pi)\big] + \log(\tfrac{2}{\epsilon}), \\
&\zeta_1 \frac{NT}{\beta} \overline{\mu}\big[\mathcal{K}(\rho_1, \overline{\pi})\big] \le \zeta_1\gamma\overline{\mu}(\rho_1 - \overline{\pi})(r) + \zeta_1 C\overline{\mu}(\rho_1 \otimes \overline{\pi})(m') \\
&\qquad + \zeta_1\overline{\mu}\big[\mathcal{K}(\rho_1, \overline{\pi})\big] + \zeta_1 \log(\tfrac{2}{\epsilon}) + \zeta_1 \frac{NT}{\beta} \overline{\mu}\big[\mathcal{K}(\rho_1, \pi) - \mathcal{K}(\overline{\pi}, \pi)\big], \\
&\zeta_2 \frac{NT}{\beta} \nu\big[\mathcal{K}(\rho_2, \overline{\pi})\big] \le \zeta_2\gamma\nu(\rho_2 - \overline{\pi})(r) + \zeta_2 C\nu(\rho_2 \otimes \overline{\pi})(m') \\
&\qquad + \zeta_2\mathcal{K}(\nu, \overline{\mu}) + \zeta_2\nu\big[\mathcal{K}(\rho_2, \overline{\pi})\big] + \zeta_2 \log(\tfrac{2}{\epsilon}) \\
&\qquad\qquad + \zeta_2 \frac{NT}{\beta} \nu\big[\mathcal{K}(\rho_2, \pi) - \mathcal{K}(\overline{\pi}, \pi)\big], \\
&\zeta_3(1 + \zeta_2) \frac{NT}{\beta} \mathcal{K}(\nu_3, \overline{\mu}) \le \zeta_3\gamma(\nu_3 - \overline{\mu})\overline{\pi}(r) \\
&\quad + \zeta_3 C\big[(\nu_3\overline{\pi}) \otimes (\nu_3\rho_1) + (\nu_3\rho_1) \otimes (\overline{\mu}\,\overline{\pi})\big](m') + \zeta_3\mathcal{K}(\nu_3, \overline{\mu}) + \zeta_3 \log(\tfrac{2}{\epsilon}) \\
&\qquad\qquad + \zeta_3(1 + \zeta_2) \frac{NT}{\beta} \big[\mathcal{K}(\nu_3, \mu) - \mathcal{K}(\overline{\mu}, \mu)\big], \\
&\zeta_4 \frac{NT}{\beta} \nu_3\big[\mathcal{K}(\rho_4, \overline{\pi})\big] \le \zeta_4\gamma\nu_3(\rho_4 - \overline{\pi})(r) \\
&\quad + \zeta_4 C\nu_3(\rho_4 \otimes \overline{\pi})(m') + \zeta_4\mathcal{K}(\nu_3, \overline{\mu}) + \zeta_4\nu_3\big[\mathcal{K}(\rho_4, \overline{\pi})\big] + \zeta_4 \log(\tfrac{2}{\epsilon}) \\
&\qquad\qquad + \zeta_4 \frac{NT}{\beta} \nu_3\big[\mathcal{K}(\rho_4, \pi) - \mathcal{K}(\overline{\pi}, \pi)\big], \\
&\zeta_5 \frac{NT}{\beta} \overline{\mu}\big[\mathcal{K}(\rho_5, \overline{\pi})\big] \le \zeta_5\gamma\overline{\mu}(\rho_5 - \overline{\pi})(r) + \zeta_5 C\overline{\mu}(\rho_5 \otimes \overline{\pi})(m') \\
&\qquad + \zeta_5\overline{\mu}\big[\mathcal{K}(\rho_5, \overline{\pi})\big] + \zeta_5 \log(\tfrac{2}{\epsilon}) + \zeta_5 \frac{NT}{\beta} \overline{\mu}\big[\mathcal{K}(\rho_5, \pi) - \mathcal{K}(\overline{\pi}, \pi)\big].
\end{aligned}
$$

Adding these six inequalities and assuming that

$$
\zeta_4 \le \zeta_3\big[(1 + \zeta_2)\tfrac{NT}{\beta} - 1\big], \tag{2.29}
$$

we find

$$
\begin{aligned}
&- N \log\big[1 - T(\nu\rho - \overline{\mu}\,\overline{\pi})(R)\big] - \beta(\nu\rho - \overline{\mu}\,\overline{\pi})(R) \\
&\quad \le -N \log\big[1 - T(\nu\rho - \overline{\mu}\,\overline{\pi})(R)\big] - \beta(\nu\rho - \overline{\mu}\,\overline{\pi})(R) \\
&\qquad + \zeta_1\big(\tfrac{NT}{\beta} - 1\big)\overline{\mu}\big[\mathcal{K}(\rho_1, \overline{\pi})\big] + \zeta_2\big(\tfrac{NT}{\beta} - 1\big)\nu\big[\mathcal{K}(\rho_2, \overline{\pi})\big] \\
&\qquad\quad + \big[\zeta_3(1 + \zeta_2)\tfrac{NT}{\beta} - \zeta_3 - \zeta_4\big]\mathcal{K}(\nu_3, \overline{\mu}) \\
&\qquad\quad + \zeta_4\big(\tfrac{NT}{\beta} - 1\big)\nu_3\big[\mathcal{K}(\rho_4, \overline{\pi})\big] + \zeta_5\big(\tfrac{NT}{\beta} - 1\big)\overline{\mu}\big[\mathcal{K}(\rho_5, \overline{\pi})\big] \\
&\quad \le \gamma(\nu\rho - \overline{\mu}\,\overline{\pi})(r) + \zeta_1\gamma\overline{\mu}(\rho_1 - \overline{\pi})(r) + \zeta_2\gamma\nu(\rho_2 - \overline{\pi})(r) \\
&\qquad + \zeta_3\gamma(\nu_3 - \overline{\mu})\overline{\pi}(r) + \zeta_4\gamma\nu_3(\rho_4 - \overline{\pi})(r) + \zeta_5\gamma\overline{\mu}(\rho_5 - \overline{\pi})(r) \\
&\qquad\quad + C\big[(\nu\rho) \otimes (\overline{\mu}\,\overline{\pi}) + \zeta_1\overline{\mu}(\rho_1 \otimes \overline{\pi}) + \zeta_2\nu(\rho_2 \otimes \overline{\pi})
\end{aligned}
$$

$$
\begin{aligned}
&+ \zeta_3(\nu_3\overline{\pi}) \otimes (\nu_3\rho_1) + \zeta_3(\nu_3\rho_1) \otimes (\overline{\mu}\,\overline{\pi}) \\
&\qquad\qquad + \zeta_4\nu_3(\rho_4 \otimes \overline{\pi}) + \zeta_5\overline{\mu}(\rho_5 \otimes \overline{\pi})\big](m') \\
&+ (1+\zeta_2)\big[\mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu)\big] + \nu\big[\mathcal{K}(\rho,\pi) - \mathcal{K}(\overline{\pi},\pi)\big] \\
&\qquad + \zeta_1\tfrac{NT}{\beta}\overline{\mu}\big[\mathcal{K}(\rho_1,\pi) - \mathcal{K}(\overline{\pi},\pi)\big] + \zeta_2\tfrac{NT}{\beta}\nu\big[\mathcal{K}(\rho_2,\pi) - \mathcal{K}(\overline{\pi},\pi)\big] \\
&+ \zeta_3(1+\zeta_2)\tfrac{NT}{\beta}\big[\mathcal{K}(\nu_3,\mu) - \mathcal{K}(\overline{\mu},\mu)\big] + \zeta_4\tfrac{NT}{\beta}\nu_3\big[\mathcal{K}(\rho_4,\pi) - \mathcal{K}(\overline{\pi},\pi)\big] \\
&\quad + \zeta_5\tfrac{NT}{\beta}\overline{\mu}\big[\mathcal{K}(\rho_5,\pi) - \mathcal{K}(\overline{\pi},\pi)\big] + (1+\zeta_1+\zeta_2+\zeta_3+\zeta_4+\zeta_5)\log(\tfrac{2}{\epsilon}),
\end{aligned}
$$

where we have also used the fact (concerning the 11th line of the preceding inequalities) that

$$
\begin{aligned}
&-\beta(\nu\rho - \overline{\mu}\,\overline{\pi})(R) + \mathcal{K}(\nu,\overline{\mu}) + \nu\big[\mathcal{K}(\rho,\overline{\pi})\big] \\
&\qquad \le -\beta(\nu\rho - \overline{\mu}\,\overline{\pi})(R) + (1+\zeta_2)\mathcal{K}(\nu,\overline{\mu}) + \nu\big[\mathcal{K}(\rho,\overline{\pi})\big] \\
&\qquad\qquad = (1+\zeta_2)\big[\mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu)\big] + \nu\big[\mathcal{K}(\rho,\pi) - \mathcal{K}(\overline{\pi},\pi)\big].
\end{aligned}
$$

Let us now apply to $\overline{\pi}$ (we shall later do the same with $\overline{\mu}$) the following inequalities, holding for any random functions of the sample and the parameters $h : \Omega \times \Theta \to \mathbb{R}$ and $g : \Omega \times \Theta \to \mathbb{R}$,

$$
\begin{aligned}
\overline{\pi}(g-h) - \mathcal{K}(\overline{\pi},\pi) &\le \sup_{\rho:\Omega\times M\to\mathcal{M}_+^1(\Theta)} \rho(g-h) - \mathcal{K}(\rho,\pi) \\
&= \log\big\{\pi\big[\exp(g-h)\big]\big\} \\
&= \log\big\{\pi\big[\exp(-h)\big]\big\} + \log\big\{\pi_{\exp(-h)}\big[\exp(g)\big]\big\} \\
&= -\pi_{\exp(-h)}(h) - \mathcal{K}(\pi_{\exp(-h)},\pi) + \log\big\{\pi_{\exp(-h)}\big[\exp(g)\big]\big\}.
\end{aligned}
$$

When h and g are observable, and h is not too far from $\beta r \simeq \beta R$, this gives a way to replace $\overline{\pi}$ with a satisfactory empirical approximation. We will apply this method, choosing ρ_1 and ρ_5 such that $\overline{\mu}\,\overline{\pi}$ is replaced either with $\overline{\mu}\rho_1$, when it comes from the first two inequalities or with $\overline{\mu}\rho_5$ otherwise, choosing ρ_2 such that $\nu\overline{\pi}$ is replaced with $\nu\rho_2$ and ρ_4 such that $\nu_3\overline{\pi}$ is replaced with $\nu_3\rho_4$. We will do so because it leads to a lot of helpful cancellations. For those to happen, we need to choose $\rho_i = \pi_{\exp(-\lambda_i r)}$, $i = 1, 2, 4$, where λ_1, λ_2 and λ_4 are such that

$$
(1+\zeta_1)\gamma = \zeta_1\tfrac{NT}{\beta}\lambda_1, \tag{2.30}
$$

$$
\zeta_2\gamma = \big(1+\zeta_2\tfrac{NT}{\beta}\big)\lambda_2, \tag{2.31}
$$

$$
(\zeta_4-\zeta_3)\gamma = \zeta_4\frac{NT}{\beta}\lambda_4, \tag{2.32}
$$

$$
\zeta_3\gamma = \zeta_5\tfrac{NT}{\beta}\lambda_5, \tag{2.33}
$$

and to assume that

$$
\zeta_4 > \zeta_3. \tag{2.34}
$$

We obtain that with $\mathbb{P}$ probability at least $1-\epsilon$,

$$
\begin{aligned}
-N\log\bigl[1-T(\mu\rho-\overline{\mu}\,\overline{\pi})(R)\bigr]-\beta(\nu\rho-\overline{\mu}\,\overline{\pi})(R)&\\
\leq\gamma(\nu\rho-\overline{\mu}\,\rho_1)(r)+\zeta_3\gamma(\nu_3\rho_4-\overline{\mu}\rho_5)(r)&\\
+\zeta_1\tfrac{NT}{\beta}\overline{\mu}\biggl\{\log\biggl[\rho_1\biggl\{\exp\Bigl[C\tfrac{\beta}{NT\zeta_1}\bigl[\nu\rho+\zeta_1\rho_1\bigr](m')\Bigr]\biggr\}\biggr]\biggr\}&\\
+\bigl(1+\zeta_2\tfrac{NT}{\beta}\bigr)\nu\biggl\{\log\biggl\{\rho_2\biggl\{\exp\Bigl[\tfrac{C}{1+\zeta_2\frac{NT}{\beta}}\zeta_2\rho_2(m')\Bigr]\biggr\}\biggr]\biggr\}&\\
+\zeta_4\tfrac{NT}{\beta}\nu_3\biggl\{\log\biggl[\rho_4\biggl\{\exp\Bigl[C\tfrac{\beta}{NT\zeta_4}\bigl[\zeta_3\nu_3\rho_1+\zeta_4\rho_4\bigr](m')\Bigr]\biggr\}\biggr]\biggr\}&\\
+\zeta_5\tfrac{NT}{\beta}\overline{\mu}\biggl\{\log\biggl[\rho_5\biggl\{\exp\Bigl[C\tfrac{\beta}{NT\zeta_5}\bigl[\zeta_3\nu_3\rho_1+\zeta_5\rho_5\bigr](m')\Bigr]\biggr\}\biggr]\biggr\}&\\
+(1+\zeta_2)\bigl[\mathcal{K}(\nu,\mu)-\mathcal{K}(\overline{\mu},\mu)\bigr]+\nu\bigl[\mathcal{K}(\rho,\pi)-\mathcal{K}(\rho_2,\pi)\bigr]&\\
+\zeta_3(1+\zeta_2)\tfrac{NT}{\beta}\bigl[\mathcal{K}(\nu_3,\mu)-\mathcal{K}(\overline{\mu},\mu)\bigr]&\\
+\biggl(1+\sum_{i=1}^{5}\zeta_i\biggr)\log\bigl(\tfrac{2}{\epsilon}\bigr).&
\end{aligned}
$$

In order to obtain more cancellations while replacing $\overline{\mu}$ by some posterior distribution, we will choose the constants such that $\lambda_5=\lambda_4$, which can be done by choosing

$$\zeta_5=\frac{\zeta_3\zeta_4}{\zeta_4-\zeta_3}. \tag{2.35}$$

We can now replace $\overline{\mu}$ with $\mu_{\exp-\xi_1\rho_1(r)-\xi_4\rho_4(r)}$, where

$$\xi_1=\frac{\gamma}{(1+\zeta_2)\bigl(1+\frac{NT}{\beta}\zeta_3\bigr)}, \tag{2.36}$$

$$\xi_4=\frac{\gamma\zeta_3}{(1+\zeta_2)\bigl(1+\frac{NT}{\beta}\zeta_3\bigr)}. \tag{2.37}$$

Choosing moreover $\nu_3=\mu_{\exp-\xi_1\rho_1(r)-\xi_4\rho_4(r)}$, to induce some more cancellations, we get

THEOREM 2.3.2. *Let us use the notation introduced above. For any positive real constants satisfying equations (2.29, page 114), (2.30, page 115), (2.31, page 115), (2.32, page 115), (2.33, page 115), (2.34, page 115), (2.35, page 116), (2.36, page 116), (2.37, page 116), with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu:\Omega\to\mathcal{M}_+^1(M)$ and any conditional posterior distribution $\rho:\Omega\times M\to\mathcal{M}_+^1(\Theta)$,*

$$
\begin{aligned}
&-N\log\bigl[1-T(\nu\rho-\overline{\mu}\,\overline{\pi})(R)\bigr]-\beta(\nu\rho-\overline{\mu}\,\overline{\pi})(R)\leq B(\nu,\rho,\beta),\\
&\text{where } B(\nu,\rho,\beta)\stackrel{\text{def}}{=}\gamma(\nu\rho-\nu_3\rho_1)(r)\\
&\qquad+(1+\zeta_2)\bigl(1+\tfrac{NT}{\beta}\zeta_3\bigr)
\end{aligned}
$$

$$\times \log\Biggl\{\nu_3\Biggl[\rho_1\Biggl\{\exp\Bigl[C\tfrac{\beta}{NT\zeta_1}\bigl[\nu\rho+\zeta_1\rho_1\bigr](m')\Bigr]\Biggr\}^{\frac{\zeta_1 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}}$$
$$\times \rho_4\Biggl\{\exp\Bigl[C\tfrac{\beta}{NT\zeta_5}\bigl[\zeta_3\nu_3\rho_1+\zeta_5\rho_4\bigr](m')\Bigr]\Biggr\}^{\frac{\zeta_5 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}}\Biggr]\Biggr\}$$
$$+\bigl(1+\zeta_2\tfrac{NT}{\beta}\bigr)\nu\Biggl\{\log\Biggl\{\rho_2\Biggl\{\exp\Bigl[\tfrac{C}{1+\zeta_2\frac{NT}{\beta}}\zeta_2\rho_2(m')\Bigr]\Biggr\}\Biggr]\Biggr\}$$
$$+\zeta_4\tfrac{NT}{\beta}\nu_3\Biggl\{\log\Biggl[\rho_4\Biggl\{\exp\Bigl[C\tfrac{\beta}{NT\zeta_4}\bigl[\zeta_3\nu_3\rho_1+\zeta_4\rho_4\bigr](m')\Bigr]\Biggr\}\Biggr]\Biggr\}$$
$$+(1+\zeta_2)\bigl[\mathcal{K}(\nu,\mu)-\mathcal{K}(\nu_3,\mu)\bigr]$$
$$+\nu\bigl[\mathcal{K}(\rho,\pi)-\mathcal{K}(\rho_2,\pi)\bigr]+\Biggl(1+\sum_{i=1}^{5}\zeta_i\Biggr)\log\bigl(\tfrac{2}{\epsilon}\bigr).$$

This theorem can be used to find the largest value $\widehat{\beta}(\nu\rho)$ of β such that $B(\nu,\rho,\beta)\leq 0$, thus providing an estimator for $\beta(\nu\rho)$ defined as $\nu\rho(R)=\overline{\mu}_{\beta(\nu\rho)}\overline{\pi}_{\beta(\nu\rho)}(R)$, where we have mentioned explicitly the dependence of $\overline{\mu}$ and $\overline{\pi}$ in β, the constant ζ_2 staying fixed. The posterior distribution $\nu\rho$ may then be chosen to maximize $\widehat{\beta}(\nu\rho)$ within some manageable subset of posterior distributions $\mathcal{P}$, thus gaining the assurance that $\nu\rho(R)\leq\overline{\mu}_{\widehat{\beta}(\nu\rho)}\overline{\pi}_{\widehat{\beta}(\nu\rho)}(R)$, with the largest parameter $\widehat{\beta}(\nu\rho)$ that this approach can provide. Maximizing $\widehat{\beta}(\nu\rho)$ is supported by the fact that $\lim_{\beta\to+\infty}\overline{\mu}_\beta\overline{\pi}_\beta(R)=\operatorname{ess\,inf}_{\mu\pi}R$. Anyhow, there is no assurance (to our knowledge) that $\beta\mapsto\overline{\mu}_\beta\overline{\pi}_\beta(R)$ will be a decreasing function of β all the way, although this may be expected to be the case in many practical situations.

We can make the bound more explicit in several ways. One point of view is to put forward the optimal values of ρ and ν. We can thus remark that

$$\begin{aligned}
&\nu\bigl[\gamma\rho(r)+\mathcal{K}(\rho,\pi)-\mathcal{K}(\rho_2,\pi)\bigr]+(1+\zeta_2)\mathcal{K}(\nu,\mu)\\
&=\nu\Biggl[\mathcal{K}\bigl[\rho,\pi_{\exp(-\gamma r)}\bigr]+\lambda_2\rho_2(r)+\int_{\lambda^2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\Biggr]+(1+\zeta_2)\mathcal{K}(\nu,\mu)\\
&=\nu\bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\gamma r)}\bigr]\bigr\}+(1+\zeta_2)\mathcal{K}\bigl[\nu,\mu_{\exp\left(-\frac{\lambda_2\rho_2(r)}{1+\zeta_2}-\frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\right)}\bigr]\\
&\quad-(1+\zeta_2)\log\Biggl\{\mu\Biggl[\exp\Biggl\{-\frac{\lambda_2}{1+\zeta_2}\rho_2(r)-\frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\Biggr\}\Biggr]\Biggr\}.
\end{aligned}$$

Thus

$$\begin{aligned}
B(\nu,\rho,\beta)&=(1+\zeta_2)\Bigl[\xi_1\nu_3\rho_1(r)+\xi_4\nu_3\rho_4(r)\\
&\qquad+\log\bigl\{\mu\bigl[\exp\bigl(-\xi_1\rho_1(r)-\xi_4\rho_4(r)\bigr)\bigr]\bigr\}\Bigr]\\
&\quad-(1+\zeta_2)\log\Biggl\{\mu\Biggl[\exp\Biggl\{-\frac{\lambda_2}{1+\zeta_2}\rho_2(r)-\frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\Biggr\}\Biggr]\Biggr\}
\end{aligned}$$

$$
\begin{aligned}
&- \gamma\nu_3\rho_1(r) + (1+\zeta_2)\big(1+\tfrac{NT}{\beta}\zeta_3\big) \\
&\quad \times \log\Bigg\{\nu_3\Bigg[\rho_1\Bigg\{\exp\Bigg[C\tfrac{\beta}{NT\zeta_1}\big[\nu\rho+\zeta_1\rho_1\big](m')\Bigg]\Bigg\}^{\frac{\zeta_1 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}} \\
&\qquad \times \rho_4\Bigg\{\exp\Bigg[C\tfrac{\beta}{NT\zeta_5}\big[\zeta_3\nu_3\rho_1+\zeta_5\rho_4\big](m')\Bigg]\Bigg\}^{\frac{\zeta_5 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}}\Bigg] \\
&\qquad + \big(1+\zeta_2\tfrac{NT}{\beta}\big)\nu\Bigg\{\log\Bigg\{\rho_2\Bigg\{\exp\Bigg[\tfrac{C}{1+\zeta_2\frac{NT}{\beta}}\zeta_2\rho_2(m')\Bigg]\Bigg\}\Bigg]\Bigg\} \\
&\quad + \zeta_4\tfrac{NT}{\beta}\nu_3\Bigg\{\log\Bigg[\rho_4\Bigg\{\exp\Bigg[C\tfrac{\beta}{NT\zeta_4}\big[\zeta_3\nu_3\rho_1+\zeta_4\rho_4\big](m')\Bigg]\Bigg\}\Bigg]\Bigg\} \\
&+ \nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\gamma r)}\big]\big\} \\
&\qquad + (1+\zeta_2)\mathcal{K}\big[\nu,\mu_{\exp\left(-\frac{\lambda_2\rho_2(r)}{1+\zeta_2}-\frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\right)}\big] \\
&\qquad\qquad + \Bigg(1+\sum_{i=1}^{5}\zeta_i\Bigg)\log\big(\tfrac{2}{\epsilon}\big).
\end{aligned}
$$

This formula is better understood when thinking about the following upper bound for the two first lines in the expression of $B(\nu,\rho,\beta)$:

$$
\begin{aligned}
&(1+\zeta_2)\Big[\xi_1\nu_3\rho_1(r)+\xi_4\nu_3\rho_4(r)+\log\big\{\mu\big[\exp\big(-\xi_1\rho_1(r)-\xi_4\rho_4(r)\big)\big]\big\}\Big] \\
&\quad - (1+\zeta_2)\log\Bigg\{\mu\Bigg[\exp\Bigg\{-\frac{\lambda_2}{1+\zeta_2}\rho_2(r) \\
&\qquad\qquad - \frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\Bigg\}\Bigg]\Bigg\} - \gamma\nu_3\rho_1(r) \\
&\qquad\qquad \le \nu_3\Bigg[\lambda_2\rho_2(r)+\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha-\gamma\rho_1(r)\Bigg].
\end{aligned}
$$

Another approach to understanding Theorem 2.3.2 is to put forward $\rho_0 = \pi_{\exp(-\lambda_0 r)}$, for some positive real constant $\lambda_0 < \gamma$, noticing that

$$
\nu\big[\mathcal{K}(\rho_0,\pi)-\mathcal{K}(\rho_2,\pi)\big] = \lambda_0\nu(\rho_2-\rho_0)(r) - \nu\big[\mathcal{K}(\rho_2,\rho_0)\big].
$$

Thus

$$
\begin{aligned}
&B(\nu,\rho_0,\beta) \le \nu_3\big[(\gamma-\lambda_0)(\rho_0-\rho_1)(r)+\lambda_0(\rho_2-\rho_1)(r)\big] \\
&\quad + (1+\zeta_2)\big(1+\tfrac{NT}{\beta}\zeta_3\big) \\
&\qquad \times \log\Bigg\{\nu_3\Bigg[\rho_1\Bigg\{\exp\Bigg[C\tfrac{\beta}{NT\zeta_1}\big[\nu\rho_0+\zeta_1\rho_1\big](m')\Bigg]\Bigg\}^{\frac{\zeta_1 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}} \\
&\qquad\qquad \times \rho_4\Bigg\{\exp\Bigg[C\tfrac{\beta}{NT\zeta_5}\big[\zeta_3\nu_3\rho_1+\zeta_5\rho_4\big](m')\Bigg]\Bigg\}^{\frac{\zeta_5 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}}\Bigg]\Bigg\}
\end{aligned}
$$

$$+\big(1+\zeta_2\tfrac{NT}{\beta}\big)\nu\bigg\{\log\bigg\{\rho_2\bigg\{\exp\bigg[\tfrac{C}{1+\zeta_2\frac{NT}{\beta}}\zeta_2\rho_2(m')\bigg]\bigg\}\bigg]\bigg\}$$
$$+\zeta_4\tfrac{NT}{\beta}\nu_3\bigg\{\log\bigg[\rho_4\bigg\{\exp\bigg[C\tfrac{\beta}{NT\zeta_4}\big[\zeta_3\nu_3\rho_1+\zeta_4\rho_4\big](m')\bigg]\bigg\}\bigg]\bigg\}$$
$$+(1+\zeta_2)\mathcal{K}\Big[\nu,\mu_{\exp\left(-\frac{(\gamma-\lambda_0)\rho_0(r)+\lambda_0\rho_2(r)}{1+\zeta_2}\right)}\Big]$$
$$-\nu\big[\mathcal{K}(\rho_2,\rho_0)\big]+\bigg(1+\sum_{i=1}^{5}\zeta_i\bigg)\log\big(\tfrac{2}{\epsilon}\big).$$

In the case when we want to select a single model $\widehat{m}(\omega)$, and therefore to set $\nu=\delta_{\widehat{m}}$, the previous inequality engages us to take

$$\widehat{m}\in\arg\min_{m\in M}(\gamma-\lambda_0)\rho_0(m,r)+\lambda_0\rho_2(m,r).$$

In parametric situations where

$$\pi_{\exp(-\lambda r)}(r)\simeq r^\star(m)+\frac{d_e(m)}{\lambda},$$

we get

$$(\gamma-\lambda_0)\rho_0(m,r)-\lambda_0\rho_2(m,r)\simeq\gamma\big[r^\star(m)+d_e(m)\big(\tfrac{1}{\lambda_0}+\tfrac{\lambda_0-\lambda_2}{\gamma\lambda_2}\big)\big],$$

resulting in a linear penalization of the empirical dimension of the models.

2.3.2. Analysis of two step bounds relative to a Gibbs prior. We will not state a formal result, but will nevertheless give some hints about how to establish one. This is a rather technical section, which can be skipped at a first reading , since it will not be used below. We should start from Theorem 1.4.2 (page 52), which gives a deterministic variance term. From Theorem 1.4.2, after a change of prior distribution, we obtain for any positive constants α_1 and α_2, any prior distributions $\widetilde{\mu}_1$ and $\widetilde{\mu}_2\in\mathcal{M}_+^1(M)$, for any prior conditional distributions $\widetilde{\pi}_1$ and $\widetilde{\pi}_2: M\to\mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distributions $\nu_1\rho_1$ and $\nu_2\rho_2$,

$$\alpha_1(\nu_1\rho_1-\nu_2\rho_2)(R)\le\alpha_2(\nu_1\rho_1-\nu_2\rho_2)(r)$$
$$+\mathcal{K}\big[(\nu_1\rho_1)\otimes(\nu_2\rho_2),(\widetilde{\mu}_1\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_2\,\widetilde{\pi}_2)\big]$$
$$+\log\Big\{(\widetilde{\mu}_1\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_2\,\widetilde{\pi}_2)\Big[\exp\big\{-\alpha_2\Psi_{\frac{\alpha_2}{N}}(R',M')+\alpha_1R'\big\}\Big]\Big\}-\log(\eta).$$

Applying this to $\alpha_1=0$, we get that

$$(\nu\rho-\nu_3\rho_1)(r)\le\frac{1}{\alpha_2}\bigg[\mathcal{K}\big[(\nu\rho)\otimes(\nu_3\rho_1),(\widetilde{\mu}\,\widetilde{\pi})\otimes(\widetilde{\mu}_3\,\widetilde{\pi}_1)\big]$$
$$+\log\Big\{(\widetilde{\mu}\,\widetilde{\nu})\otimes(\widetilde{\mu}_3\,\widetilde{\pi}_1)\Big[\exp\big\{\alpha_2\Psi_{-\frac{\alpha_2}{N}}(R',M')\big\}\Big]\Big\}-\log(\eta)\bigg].$$

In the same way, to bound quantities of the form

$$\begin{aligned}
&\log\biggl\{\nu_3\biggl[\rho_1\Bigl\{\exp\Bigl[C_1(\nu\rho+\zeta_1\rho_1)(m')\Bigr]\Bigr\}^{p_1}\\
&\qquad\qquad\times\rho_4\Bigl\{\exp\Bigl[C_2\bigl[\zeta_3\nu_3\rho_1+\zeta_5\rho_4\bigr](m')\Bigr]\Bigr\}^{p_2}\biggr]\biggr\}\\
&\quad=\sup_{\nu_5}\biggl\{p_1\sup_{\rho_5}\Bigl\{C_1\bigl[(\nu\rho)\otimes(\nu_5\rho_5)+\zeta_1\nu_5(\rho_1\otimes\rho_5)\bigr](m')-\mathcal{K}(\rho_5,\rho_1)\Bigr\}\\
&\qquad+p_2\sup_{\rho_6}\Bigl\{C_2\bigl[\zeta_3(\nu_3\rho_1)\otimes(\nu_5\rho_6)\\
&\qquad\qquad+\zeta_5\nu_5(\rho_4\otimes\rho_6)\bigr](m')-\mathcal{K}(\rho_6,\rho_4)\Bigr\}-\mathcal{K}(\nu_5,\nu_3)\biggr\},
\end{aligned}$$

where C_1, C_2, p_1 and p_2 are positive constants, and similar terms, we need to use inequalities of the type: for any prior distributions $\widetilde{\mu}_i\,\widetilde{\pi}_i$, $i=1,2$, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distributions $\nu_i\rho_i$, $i=1,2$,

$$\begin{aligned}
\alpha_3(\nu_1\rho_1)\otimes(\nu_2\rho_2)(m')\le{}&\log\Bigl\{(\widetilde{\mu}_1\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_2\,\widetilde{\pi}_2)\exp\Bigl[\alpha_3\Phi_{\frac{-\alpha_3}{N}}(M')\Bigr]\Bigr\}\\
&+\mathcal{K}\bigl[(\nu_1\rho_1)\otimes(\nu_2\rho_2),(\widetilde{\mu}_1\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_2\,\widetilde{\pi}_2)\bigr]-\log(\eta).
\end{aligned}$$

We need also the variant: with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\nu_1:\Omega\to\mathcal{M}_+^1(M)$ and any conditional posterior distributions $\rho_1,\rho_2:\Omega\times M\to\mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}
\alpha_3\nu_1(\rho_1\otimes\rho_2)(m')\le{}&\log\Bigl\{\widetilde{\mu}_1\bigl(\widetilde{\pi}_1\otimes\widetilde{\pi}_2\bigr)\exp\Bigl[\alpha_3\Phi_{-\frac{\alpha_3}{N}}(M')\Bigr]\Bigr\}\\
&+\mathcal{K}(\nu_1,\widetilde{\mu}_1)+\nu_1\bigl\{\mathcal{K}\bigl[\rho_1\otimes\rho_2,\widetilde{\pi}_1\otimes\widetilde{\pi}_2\bigr]\bigr\}-\log(\eta).
\end{aligned}$$

We deduce that

$$\begin{aligned}
&\log\biggl\{\nu_3\biggl[\rho_1\Bigl\{\exp\Bigl[C_1(\nu\rho+\zeta_1\rho_1)(m')\Bigr]\Bigr\}^{p_1}\\
&\qquad\qquad\times\rho_4\Bigl\{\exp\Bigl[C_2\bigl[\zeta_3\nu_3\rho_1+\zeta_5\rho_4\bigr](m')\Bigr]\Bigr\}^{p_2}\biggr]\biggr\}\\
&\quad\le\sup_{\nu_5}\biggl\{p_1\sup_{\rho_5}\biggl[\frac{C_1}{\alpha_3}\biggl\{\log\Bigl\{(\widetilde{\mu}\,\widetilde{\pi})\otimes(\widetilde{\mu}_5\,\widetilde{\pi}_5)\exp\Bigl[\alpha_3\Phi_{-\frac{\alpha_3}{N}}(M')\Bigr]\Bigr\}\\
&\qquad+\mathcal{K}\bigl[(\nu\rho)\otimes(\nu_5\rho_5),(\widetilde{\mu}\,\widetilde{\pi}\otimes(\widetilde{\mu}_5\,\widetilde{\pi}_5)\bigr]+\log(\tfrac{2}{\eta})\\
&\qquad+\zeta_1\biggl[\log\Bigl\{\widetilde{\mu}_5\bigl(\widetilde{\pi}_1\otimes\widetilde{\pi}_5\bigr)\exp\Bigl[\alpha_3\Phi_{-\frac{\alpha_3}{N}}(M')\Bigr]\Bigr\}\\
&\quad+\mathcal{K}(\nu_5,\widetilde{\mu}_5)+\nu_5\bigl\{\mathcal{K}\bigl[\rho_1\otimes\rho_5,\widetilde{\pi}_1\otimes\widetilde{\pi}_5\bigr]\bigr\}+\log\bigl(\tfrac{2}{\eta}\bigr)\biggr]\biggr\}-\mathcal{K}(\rho_5,\rho_1)\biggr]
\end{aligned}$$

$$+ p_2 \sup_{\rho_6} \Bigg[\frac{C_1}{\alpha_3} \Big\{ \log\Big\{ (\widetilde{\mu}_3\, \widetilde{\pi}_1) \otimes (\widetilde{\mu}_5\, \widetilde{\pi}_6) \exp\Big[\alpha_3 \Phi_{-\frac{\alpha_3}{N}}(M') \Big] \Big\} + \mathcal{K}\big[(\nu_3\rho_1) \otimes (\nu_5\rho_6), (\widetilde{\mu}_3\, \widetilde{\pi}_1 \otimes (\widetilde{\mu}_5\, \widetilde{\pi}_6) \big] + \log(\tfrac{2}{\eta}) + \zeta_1 \Bigg[\log\Big\{ \widetilde{\mu}_5\big(\widetilde{\pi}_4 \otimes \widetilde{\pi}_6\big) \exp\Big[\alpha_3 \Phi_{-\frac{\alpha_3}{N}}(M') \Big] \Big\} + \mathcal{K}(\nu_5, \widetilde{\mu}_5) + \nu_5\big\{ \mathcal{K}\big[\rho_4 \otimes \rho_6, \widetilde{\pi}_4 \otimes \widetilde{\pi}_6 \big] \big\} + \log\big(\tfrac{2}{\eta}\big) \Bigg] \Big\} - \mathcal{K}(\rho_6, \rho_4) \Bigg] - \mathcal{K}(\nu_5, \nu_3) \Big\}.$$

We are then left with the need to bound entropy terms like $\mathcal{K}(\nu_3\rho_1, \widetilde{\mu}_3\widetilde{\pi}_1)$, where we have the choice of $\widetilde{\mu}_3$ and $\widetilde{\pi}_1$, to obtain a useful bound. As could be expected, we decompose it into

$$\mathcal{K}(\nu_3\rho_1, \widetilde{\mu}_3\widetilde{\pi}_1) = \mathcal{K}(\nu_3, \widetilde{\mu}_3) + \nu_3\big[\mathcal{K}(\rho_1, \widetilde{\pi}_1)\big].$$

Let us look after the second term first, choosing $\widetilde{\pi}_1 = \pi_{\exp(-\beta_1 R)}$:

$$\begin{aligned}
\nu_3\big[\mathcal{K}(\rho_1, \widetilde{\pi}_1)\big] &= \nu_3\big[\beta_1(\rho_1 - \widetilde{\pi}_1)(R) + \mathcal{K}(\rho_1, \pi) - \mathcal{K}(\widetilde{\pi}_1, \pi)\big] \\
&\leq \frac{\beta_1}{\alpha_1} \Bigg[\alpha_2 \nu_3(\rho_1 - \widetilde{\pi}_1)(r) + \mathcal{K}(\nu_3, \widetilde{\mu}_3) + \nu_3\big[\mathcal{K}(\rho_1, \widetilde{\pi}_1)\big] \\
&\quad + \log\Big\{ \widetilde{\mu}_3\big(\widetilde{\pi}_1^{\otimes 2}\big) \Big[\exp\big\{ -\alpha_2 \Psi_{\frac{\alpha_2}{N}}(R', M') + \alpha_1 R' \big\} \Big] \Big\} - \log(\eta) \Bigg] \\
&\quad + \nu_3\big[\mathcal{K}(\rho_1, \pi) - \mathcal{K}(\widetilde{\pi}_1, \pi)\big] \\
&\leq \frac{\beta_1}{\alpha_1} \Bigg[\mathcal{K}(\nu_3, \widetilde{\mu}_3) + \nu_3\big[\mathcal{K}(\rho_1, \widetilde{\pi}_1)\big] \\
&\quad + \log\Big\{ \widetilde{\mu}_3\big(\widetilde{\pi}_1^{\otimes 2}\big) \Big[\exp\big\{ -\alpha_2 \Psi_{\frac{\alpha_2}{N}}(R', M') + \alpha_1 R' \big\} \Big] \Big\} - \log(\eta) \Bigg] \\
&\quad + \nu_3\big\{ \mathcal{K}\big[\rho_1, \pi_{\exp(-\frac{\beta_1\alpha_2}{\alpha_1} r)}\big] \big\}.
\end{aligned}$$

Thus, when the constraint $\lambda_1 = \frac{\beta_1\alpha_2}{\alpha_1}$ is satisfied,

$$\begin{aligned}
\nu_3\big[\mathcal{K}(\rho_1, \widetilde{\pi}_1)\big] \leq \Big(1 - \frac{\beta_1}{\alpha_1}\Big)^{-1} \frac{\beta_1}{\alpha_1} \Bigg[& \mathcal{K}(\nu_3, \widetilde{\mu}_3) \\
& + \log\Big\{ \widetilde{\mu}_3\big(\widetilde{\pi}_1^{\otimes 2}\big) \Big[\exp\big\{ -\alpha_2 \Psi_{\frac{\alpha_2}{N}}(R', M') + \alpha_1 R' \big\} \Big] \Big\} - \log(\eta) \Bigg].
\end{aligned}$$

We can further specialize the constants, choosing $\alpha_1 = N \sinh(\frac{\alpha_2}{N})$, so that

$$-\alpha_2 \Psi_{\frac{\alpha_2}{N}}(R', M') + \alpha_1 R' \leq 2N \sinh\Big(\frac{\alpha_2}{2N}\Big)^2 M'.$$

We can for instance choose $\alpha_2 = \gamma$, $\alpha_1 = N \sinh(\frac{\gamma}{N})$ and $\beta_1 = \lambda_1 \frac{N}{\gamma} \sinh(\frac{\gamma}{N})$, leading to

PROPOSITION 2.3.3. *With the notation of Theorem 2.3.2, the constants being set as explained above, putting* $\widetilde{\pi}_1 = \pi_{\exp(-\lambda_1 \frac{N}{\gamma} \sinh(\frac{\gamma}{N})R)}$, *with* $\mathbb{P}$ *probability at least* $1-\eta$,

$$\nu_3\big[\mathcal{K}(\rho_1,\widetilde{\pi}_1)\big] \leq \Big(1-\frac{\lambda_1}{\gamma}\Big)^{-1}\frac{\lambda_1}{\gamma}\bigg[\mathcal{K}(\nu_3,\widetilde{\mu}_3) + \log\Big\{\widetilde{\mu}_3\big(\widetilde{\pi}_1^{\otimes 2}\big)\Big[\exp\big\{2N\sinh(\tfrac{\gamma}{2N})^2 M'\big\}\Big]\Big\} - \log(\eta)\bigg].$$

More generally

$$\nu_3\big[\mathcal{K}(\rho,\widetilde{\pi}_1)\big] \leq \Big(1-\frac{\lambda_1}{\gamma}\Big)^{-1}\frac{\lambda_1}{\gamma}\bigg[\mathcal{K}(\nu_3,\widetilde{\mu}_3) + \log\Big\{\widetilde{\mu}_3\big(\widetilde{\pi}_1^{\otimes 2}\big)\Big[\exp\big\{2N\sinh(\tfrac{\gamma}{2N})^2 M'\big\}\Big]\Big\} - \log(\eta)\bigg] + \Big(1-\frac{\lambda_1}{\gamma}\Big)^{-1}\nu_3\big[\mathcal{K}(\rho,\rho_1)\big].$$

In a similar way, let us now choose $\widetilde{\mu}_3 = \mu_{\exp[-\alpha_3\overline{\pi}(R)]}$. We can write

$$\begin{aligned}\mathcal{K}(\nu,\widetilde{\mu}_3) &= \alpha_3(\nu-\widetilde{\mu}_3)\overline{\pi}(R) + \mathcal{K}(\nu,\mu) - \mathcal{K}(\widetilde{\mu}_3,\mu)\\ &\leq \frac{\alpha_3}{\alpha_1}\bigg[\alpha_2(\nu-\widetilde{\mu}_3)\overline{\pi}(r) + \mathcal{K}(\nu,\widetilde{\mu}_3)\\ &\quad + \log\Big\{(\widetilde{\mu}_3\overline{\pi})\otimes(\widetilde{\mu}_3\overline{\pi})\Big[\exp\big\{-\alpha_2\Psi_{\frac{\alpha_2}{N}}(R',M') + \alpha_1 R'\big\}\Big]\Big\} - \log(\eta)\bigg]\\ &\quad + \mathcal{K}(\nu,\mu) - \mathcal{K}(\widetilde{\mu}_3,\mu).\end{aligned}$$

Let us choose $\alpha_2 = \gamma$, $\alpha_1 = N\sinh(\frac{\gamma}{N})$, and let us add some other entropy inequalities to get rid of $\overline{\pi}$ in a suitable way, the approach of entropy compensation being the same as that used to obtain the empirical bound of Theorem 2.3.2 (page 116). This results with $\mathbb{P}$ probability at least $1-\eta$ in

$$\begin{aligned}\Big(1-\frac{\alpha_3}{\alpha_1}\Big)\mathcal{K}(\nu,\widetilde{\mu}_3) &\leq \frac{\alpha_3}{\alpha_1}\bigg[\gamma(\nu-\widetilde{\mu}_3)\overline{\pi}(r)\\ &\quad + \log\Big\{(\widetilde{\mu}_3\overline{\pi})\otimes(\widetilde{\mu}_3\overline{\pi})\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M') + \alpha_1 R'\big\}\Big]\Big\} + \log(\tfrac{2}{\eta})\bigg]\\ &\quad + \mathcal{K}(\nu,\mu) - \mathcal{K}(\widetilde{\mu}_3,\mu),\\ \zeta_6\Big(1-\frac{\beta}{\alpha_1}\Big)\widetilde{\mu}_3\big[\mathcal{K}(\rho_6,\overline{\pi})\big] &\leq \zeta_6\frac{\beta}{\alpha_1}\bigg[\gamma\widetilde{\mu}_3(\rho_6-\overline{\pi})(r)\\ &\quad + \log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M') + \alpha_1 R'\big\}\Big]\Big\} + \log(\tfrac{2}{\eta})\bigg]\\ &\quad + \zeta_6\widetilde{\mu}_3\big[\mathcal{K}(\rho_6,\pi) - \mathcal{K}(\overline{\pi},\pi)\big],\\ \zeta_7\Big(1-\frac{\beta}{\alpha_1}\Big)\widetilde{\mu}_3\big[\mathcal{K}(\rho_7,\overline{\pi})\big] &\leq \zeta_7\frac{\beta}{\alpha_1}\bigg[\gamma\widetilde{\mu}_3(\rho_7-\overline{\pi})(r)\end{aligned}$$

$$
\begin{aligned}
&+\log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}+\log(\tfrac{2}{\eta})\Big]\\
&\qquad\qquad+\zeta_7\widetilde{\mu}_3\big[\mathcal{K}(\rho_7,\pi)-\mathcal{K}(\overline{\pi},\pi)\big],\\
\zeta_8\Big(1-\frac{\beta}{\alpha_1}\Big)\nu\big[\mathcal{K}(\rho_8,\overline{\pi})\big]&\le\zeta_8\frac{\beta}{\alpha_1}\Big[\gamma\nu(\rho_8-\overline{\pi})(r)+\mathcal{K}(\nu,\widetilde{\mu}_3)\\
&+\log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}+\log(\tfrac{2}{\eta})\Big]\\
&\qquad\qquad+\zeta_8\nu\big[\mathcal{K}(\rho_8,\pi)-\mathcal{K}(\overline{\pi},\pi)\big],\\
\zeta_9\Big(1-\frac{\beta}{\alpha_1}\Big)\nu\big[\mathcal{K}(\rho_9,\overline{\pi})\big]&\le\zeta_9\frac{\beta}{\alpha_1}\Big[\gamma\nu(\rho_9-\overline{\pi})(r)+\mathcal{K}(\nu,\widetilde{\mu}_3)\\
&+\log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}+\log(\tfrac{2}{\eta})\Big]\\
&\qquad\qquad+\zeta_9\nu\big[\mathcal{K}(\rho_9,\pi)-\mathcal{K}(\overline{\pi},\pi)\big],
\end{aligned}
$$

where we have introduced a bunch of constants, assumed to be positive, that we will more precisely set to

$$
\begin{aligned}
x_8+x_9&=1,\\
(\zeta_6\beta+x_8\alpha_3)\frac{\gamma}{\alpha_1}&=\lambda_6,\\
(\zeta_7\beta+x_9\alpha_3)\frac{\gamma}{\alpha_1}&=\lambda_7,\\
(\zeta_8\beta-x_8\alpha_3)\frac{\gamma}{\alpha_1}&=\lambda_8,\\
(\zeta_9\beta-x_9\alpha_3)\frac{\gamma}{\alpha_1}&=\lambda_9.
\end{aligned}
$$

We get with $\mathbb{P}$ probability at least $1-\eta$,

$$
\begin{aligned}
\Big(1-\frac{\alpha_3}{\alpha_1}-(\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\Big)\mathcal{K}(\nu,\widetilde{\mu}_3)&\le\\
\frac{\alpha_3}{\alpha_1}\Big[\gamma\big[\nu(x_8\rho_8+x_9\rho_9)(r)-\widetilde{\mu}_3(x_8\rho_6+x_9\rho_7)(r)\big]&\\
+\frac{\alpha_3}{\alpha_1}\log\Big\{(\widetilde{\mu}_3\overline{\pi})\otimes(\widetilde{\mu}_3\overline{\pi})\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}&\\
+(\zeta_6+\zeta_7+\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}&\\
+\mathcal{K}(\nu,\mu)-\mathcal{K}(\widetilde{\mu}_3,\mu)+\Big(\frac{\alpha_3}{\alpha_1}+(\zeta_6+\zeta_7+\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\Big)\log\big(\tfrac{2}{\eta}\big).&
\end{aligned}
$$

Let us choose the constants so that $\lambda_1=\lambda_7=\lambda_9$, $\lambda_4=\lambda_6=\lambda_8$, $\alpha_3 x_9\frac{\gamma}{\alpha_1}=\xi_1$ and $\alpha_3 x_8\frac{\gamma}{\alpha_1}=\xi_4$. This is done by setting

$$
x_8=\frac{\xi_4}{\xi_1+\xi_4},
$$

$$
\begin{aligned}
x_9 &= \frac{\xi_1}{\xi_1+\xi_4},\\
\alpha_3 &= \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})(\xi_1+\xi_4),\\
\zeta_6 &= \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\frac{(\lambda_4-\xi_4)}{\beta},\\
\zeta_7 &= \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\frac{(\lambda_1-\xi_1)}{\beta},\\
\zeta_8 &= \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\frac{(\lambda_4+\xi_4)}{\beta},\\
\zeta_9 &= \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\frac{(\lambda_1+\xi_1)}{\beta}.
\end{aligned}
$$

The inequality $\lambda_1 > \xi_1$ is always satisfied. The inequality $\lambda_4 > \xi_4$ is required for the above choice of constants, and will be satisfied for a suitable choice of ζ_3 and ζ_4.

Under these assumptions, we obtain with $\mathbb{P}$ probability at least $1-\eta$

$$
\begin{aligned}
\Big(1-\frac{\alpha_3}{\alpha_1}-(\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\Big)\mathcal{K}(\nu,\widetilde{\mu}_3) &\le (\nu-\widetilde{\mu}_3)(\xi_1\rho_1+\xi_4\rho_4)(r)\\
&\quad+\frac{\alpha_3}{\alpha_1}\log\Big\{(\widetilde{\mu}_3\overline{\pi})\otimes(\widetilde{\mu}_3\overline{\pi})\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}\\
&\quad+(\zeta_6+\zeta_7+\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}\\
&\quad+\mathcal{K}(\nu,\mu)-\mathcal{K}(\widetilde{\mu}_3,\mu)+\Big(\frac{\alpha_3}{\alpha_1}+(\zeta_6+\zeta_7+\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\Big)\log\big(\tfrac{2}{\eta}\big).
\end{aligned}
$$

This proves

Proposition 2.3.4. *The constants being set as explained above, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\nu:\Omega\to\mathcal{M}_+^1(M)$,*

$$
\begin{aligned}
\mathcal{K}(\nu,\widetilde{\mu}_3) &\le \Big(1-\frac{\alpha_3}{\alpha_1}-(\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\Big)^{-1}\Big[\mathcal{K}(\nu,\nu_3)\\
&\quad+\frac{\alpha_3}{\alpha_1}\log\Big\{(\widetilde{\mu}_3\overline{\pi})\otimes(\widetilde{\mu}_3\overline{\pi})\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}\\
&\quad+(\zeta_6+\zeta_7+\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}\\
&\quad+\Big(\frac{\alpha_3}{\alpha_1}+(\zeta_6+\zeta_7+\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\Big)\log\big(\tfrac{2}{\eta}\big)\Big].
\end{aligned}
$$

Thus

$$
\begin{aligned}
\mathcal{K}(\nu_3\rho_1,\widetilde{\mu}_3\,\widetilde{\pi}_1) &\le \frac{1+\big(1-\frac{\lambda_1}{\gamma}\big)^{-1}\frac{\lambda_1}{\gamma}}{1-\frac{\alpha_3}{\alpha_1}-(\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}}\\
&\quad\times\Big[\frac{\alpha_3}{\alpha_1}\log\Big\{(\widetilde{\mu}_3\overline{\pi}\otimes(\widetilde{\mu}_3\overline{\pi})\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}
\end{aligned}
$$

$$+ (\zeta_6+\zeta_7+\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R',M')+\alpha_1 R'\big\}\Big]\Big\}$$
$$+\Big(\frac{\alpha_3}{\alpha_1}+(\zeta_6+\zeta_7+\zeta_8+\zeta_9)\frac{\beta}{\alpha_1}\Big)\log\big(\tfrac{2}{\eta}\big)\Big]$$
$$+\Big(1-\frac{\lambda_1}{\gamma}\Big)^{-1}\frac{\lambda_1}{\gamma}\Big[\log\Big\{\widetilde{\mu}_3\big(\widetilde{\pi}_1^{\otimes 2}\big)\Big[\exp\big\{2N\sinh\big(\tfrac{\gamma}{2N}\big)^2 M'\big\}\Big]\Big\}-\log\big(\tfrac{2}{\eta}\big)\Big].$$

We will not go further, lest it may become tedious, but we hope we have given sufficient hints to state informally that the bound $B(\nu,\rho,\beta)$ of Theorem 2.3.2 (page 116) is upper bounded with $\mathbb{P}$ probability close to one by a bound of the same flavour where the empirical quantities r and m' have been replaced with their expectations R and M'.

2.3.3. Two step localization between posterior distributions. Here we work with a family of prior distributions described by a regular conditional prior distribution $\pi = M \to \mathcal{M}_+^1(\Theta)$, where M is some measurable index set. This family may typically describe a countable family of parametric models. In this case $M=\mathbb{N}$, and each of the prior distributions $\pi(i,.)$, $i\in\mathbb{N}$ satisfies some parametric complexity assumption of the type

$$\limsup_{\beta\to+\infty}\beta\big[\pi_{\exp(-\beta R)}(i,.)(R)-\operatorname*{ess\,inf}_{\pi(i,.)}R\big]=d_i<+\infty,\qquad i\in M.$$

Let us consider also a prior distribution $\mu\in\mathcal{M}_+^1(M)$ defined on the index set M.

Our aim here will be to compare the performance of two given posterior distributions $\nu_1\rho_1$ and $\nu_2\rho_2$, where $\nu_1,\nu_2:\Omega\to\mathcal{M}_+^1(M)$, and where $\rho_1,\rho_2:\Omega\times M\to\mathcal{M}_+^1(\Theta)$. More precisely, we would like to establish a bound for $(\nu_1\rho_1-\nu_2\rho_2)(R)$ which could be a starting point to implement a selection method similar to the one described in Theorem 2.2.4 (page 93). To this purpose, we can start with Theorem 2.2.1 (page 88), which says that with $\mathbb{P}$ probability at least $1-\epsilon$,

$$-N\log\Big\{1-\tanh\big(\tfrac{\lambda}{N}\big)\big(\nu_1\rho_1-\nu_2\rho_2\big)(R)\Big\}\le\lambda(\nu_1\rho_1-\nu_2\rho_2)(r)$$
$$+N\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big](\nu_1\rho_1)\otimes(\nu_2\rho_2)(m')+\mathcal{K}(\nu_1,\widetilde{\mu})+\mathcal{K}(\nu_2,\widetilde{\mu})$$
$$+\nu_1\big[\mathcal{K}(\rho_1,\widetilde{\pi})\big]+\nu_2\big[\mathcal{K}(\rho_2,\widetilde{\pi})\big]-\log(\epsilon),$$

where $\widetilde{\mu}\in\mathcal{M}_+^1(M)$ and $\widetilde{\pi}:M\to\mathcal{M}_+^1(\Theta)$ are suitably localized prior distributions to be chosen later on. To use these localized prior distributions, we need empirical bounds for the entropy terms $\mathcal{K}(\nu_i,\widetilde{\mu})$ and $\nu_i\big[\mathcal{K}(\rho_i,\widetilde{\pi})\big]$, $i=1,2$.

Bounding $\nu\big[\mathcal{K}(\rho,\widetilde{\pi})\big]$ can be done using the following generalization of Corollary 2.1.19 page 88:

Corollary 2.3.5. *For any positive real constants γ and λ such that $\gamma<\lambda$, for any prior distribution $\overline{\mu}\in\mathcal{M}_+^1(M)$ and any conditional prior distribution $\pi:M\to\mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior*

distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$*, and any conditional posterior distribution* $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$*,*

$$\nu\Big\{\mathcal{K}\big[\rho, \pi_{\exp[-N\frac{\gamma}{\lambda}\tanh(\frac{\lambda}{N})R]}\big]\Big\} \leq K'(\nu,\rho,\gamma,\lambda,\epsilon) + \frac{1}{\frac{\lambda}{\gamma}-1}\mathcal{K}(\nu,\overline{\mu}),$$

where

$$K'(\nu,\rho,\gamma,\lambda,\epsilon) \stackrel{\text{def}}{=} \Big(1-\frac{\gamma}{\lambda}\Big)^{-1}\Big\{\nu\big[\mathcal{K}(\rho,\pi_{\exp(-\gamma r)}\big] - \frac{\gamma}{\lambda}\log(\epsilon) + \nu\Big\{\log\Big[\pi_{\exp(-\gamma r)}\Big(\exp\big\{N\frac{\gamma}{\lambda}\log\big[\cosh(\frac{\lambda}{N})\big]\rho(m')\big\}\Big)\Big]\Big\}\Big\}.$$

To apply this corollary to our case, we have to set

$$\widetilde{\pi} = \pi_{\exp[-N\frac{\gamma}{\lambda}\tanh(\frac{\lambda}{N})R]}.$$

Let us also consider for some positive real constant β the conditional prior distribution

$$\overline{\pi} = \pi_{\exp(-\beta R)}$$

and the prior distribution

$$\overline{\mu} = \mu_{\exp[-\alpha\overline{\pi}(R)]}.$$

Let us see how we can bound, given any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$, the divergence $\mathcal{K}(\nu,\overline{\mu})$. We can see that

$$\mathcal{K}(\nu,\overline{\mu}) = \alpha(\nu-\overline{\mu})\overline{\pi}(R) + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu).$$

Now, let us introduce the conditional posterior distribution

$$\widehat{\pi} = \pi_{\exp(-\gamma r)}$$

and let us decompose

$$(\nu-\overline{\mu})\big[\overline{\pi}(R)\big] = \nu\big[\overline{\pi}(R)-\widehat{\pi}(R)\big] + (\nu-\overline{\mu})\big[\widehat{\pi}(R)\big] + \overline{\mu}\big[\widehat{\pi}(R)-\overline{\pi}(R)\big].$$

Starting from the exponential inequality

$$\mathbb{P}\left[\overline{\mu}\big[\overline{\pi}\otimes\overline{\pi}\big]\exp\Big\{-N\log\big[1-\tanh(\tfrac{\gamma}{N})R'\big] - \gamma r' - N\log\big[\cosh(\tfrac{\gamma}{N})\big]m'\Big\}\right] \leq 1,$$

and reasoning in the same way that led to Theorem 2.1.1 (page 71) in the simple case when we take in this theorem $\lambda = \gamma$, we get with $\mathbb{P}$ probability at least $1-\epsilon$, that

$$-N\log\big\{1-\tanh(\tfrac{\gamma}{N})\nu(\overline{\pi}-\widehat{\pi})(R)\big\} + \beta\nu(\overline{\pi}-\widehat{\pi})(R)$$

$$\leq \nu\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\widehat{\pi}(m')\big\}\Big]\Big\}\Big] + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon).$$

$$-N\log\big\{1-\tanh(\tfrac{\gamma}{N})\overline{\mu}(\widehat{\pi}-\overline{\pi})(R)\big\} - \beta\overline{\mu}(\widehat{\pi}-\overline{\pi})(R)$$
$$\leq \overline{\mu}\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\widehat{\pi}(m')\big\}\Big]\Big\}\Big] - \log(\epsilon).$$

In the meantime, using Theorem 2.2.1 (page 88) and Corollary 2.3.5 above, we see that with $\mathbb{P}$ probability at least $1-2\epsilon$, for any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}
&-N\log\Big\{1-\tanh(\tfrac{\lambda}{N})(\nu-\overline{\mu})\rho(R)\Big\} \leq \lambda(\nu-\overline{\mu})\rho(r)\\
&+N\log\big[\cosh(\tfrac{\lambda}{N})\big](\nu\rho)\otimes(\overline{\mu}\rho)(m') + (\nu+\overline{\mu})\mathcal{K}(\rho,\widetilde{\pi}) + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon)\\
&\leq \lambda(\nu-\overline{\mu})\rho(r) + N\log\big[\cosh(\tfrac{\lambda}{N})\big](\nu\rho)\otimes(\overline{\mu}\rho)(m') + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon)\\
&+\Big(1-\tfrac{\gamma}{\lambda}\Big)^{-1}(\nu+\overline{\mu})\Big\{\mathcal{K}\big(\rho,\widehat{\pi}\big) + \log\Big\{\widehat{\pi}\Big[\exp\big\{N\tfrac{\gamma}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\rho(m')\big\}\Big]\Big\}\Big\}\\
&\qquad + \Big(\tfrac{\lambda}{\gamma}-1\Big)^{-1}\big[\mathcal{K}(\nu,\overline{\mu}) - 2\log(\epsilon)\big].
\end{aligned}$$

Putting all this together, we see that with $\mathbb{P}$ probability at least $1-3\epsilon$, for any posterior distribution $\nu \in \mathcal{M}_+^1(M)$,

$$\begin{aligned}
&\Big[1 - \frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta} - \frac{\alpha}{N\tanh(\frac{\lambda}{N})(1-\frac{\gamma}{\lambda})}\Big]\mathcal{K}(\nu,\overline{\mu}) \leq\\
&\alpha\Big[N\tanh(\tfrac{\gamma}{N})+\beta\Big]^{-1}\Big\{\nu\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big] - \log(\epsilon)\Big\}\\
&+\alpha\Big[N\tanh(\tfrac{\gamma}{N})-\beta\Big]^{-1}\Big\{\overline{\mu}\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big] - \log(\epsilon)\Big\}\\
&\qquad + \alpha\big[N\tanh(\tfrac{\lambda}{N})\big]^{-1}\Big\{\\
&\qquad \lambda(\nu-\overline{\mu})\widehat{\pi}(r) + N\log\big[\cosh(\tfrac{\lambda}{N})\big](\nu\widehat{\pi})\otimes(\overline{\mu}\widehat{\pi})(m')\\
&\qquad +\Big(1-\tfrac{\gamma}{\lambda}\Big)^{-1}(\nu+\overline{\mu})\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{N\tfrac{\gamma}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big]\\
&\qquad\qquad - \frac{1+\frac{\gamma}{\lambda}}{1-\frac{\gamma}{\lambda}}\log(\epsilon)\Big\} + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu).
\end{aligned}$$

Replacing in the right-hand side of this inequality the unobserved prior distribution $\overline{\mu}$ with the worst possible posterior distribution, we obtain

THEOREM 2.3.6. *For any positive real constants α, β, γ and λ, using the notation,*

$$\overline{\pi} = \pi_{\exp(-\beta R)},$$

$$\overline{\mu} = \mu_{\exp[-\alpha\overline{\pi}(R)]},$$
$$\widehat{\pi} = \pi_{\exp(-\gamma r)},$$
$$\widehat{\mu} = \mu_{\exp[-\alpha\frac{\lambda}{N}\tanh(\frac{\lambda}{N})^{-1}\widehat{\pi}(r)]},$$

with $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any posterior distribution* $\nu : \Omega \to \mathcal{M}_+^1(M)$*,*

$$\begin{aligned}
&\Big[1 - \frac{\alpha}{N\tanh(\frac{\gamma}{N}) + \beta} - \frac{\alpha}{N\tanh(\frac{\lambda}{N})\big(1-\frac{\gamma}{\lambda}\big)}\Big]\mathcal{K}(\nu,\overline{\mu}) \le \mathcal{K}(\nu,\widehat{\mu}) \\
&\quad + \frac{\alpha}{N\tanh(\frac{\gamma}{N}) + \beta}\left\{\nu\left[\log\left\{\widehat{\pi}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\right\}\right]\right\} \\
&\quad + \frac{\alpha}{N\tanh(\frac{\lambda}{N})(1-\frac{\gamma}{\lambda})}\left\{\nu\left[\log\left\{\widehat{\pi}\Big[\exp\big\{N\tfrac{\gamma}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\widehat{\pi}(m')\big\}\Big]\right\}\right]\right\} \\
&\quad + \log\left\{\widehat{\mu}\left[\left[\widehat{\pi}\Big\{\exp\Big[N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\Big]\Big\}\right]^{\frac{\alpha}{N\tanh(\frac{\gamma}{N})-\beta}}\right.\right. \\
&\quad \times \left[\widehat{\pi}\Big\{\exp\Big[N\tfrac{\gamma}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\widehat{\pi}(m')\Big]\Big\}\right]^{\frac{\alpha}{N\tanh(\frac{\lambda}{N})(1-\frac{\gamma}{\lambda})}} \\
&\quad \left.\left. \times \exp\left[\frac{\alpha\log[\cosh(\frac{\lambda}{N})]}{\tanh(\frac{\lambda}{N})}(\nu\widehat{\pi})\otimes\widehat{\pi}(m')\right]\right]\right\} \\
&\quad + \left[\frac{1}{N\tanh(\frac{\gamma}{N})+\beta} + \frac{1}{N\tanh(\frac{\gamma}{N})-\beta} + \frac{1+\frac{\gamma}{\lambda}}{N\tanh(\frac{\lambda}{N})\big(1-\frac{\gamma}{\lambda}\big)}\right]\log\big(\tfrac{3}{\epsilon}\big).
\end{aligned}$$

This result is satisfactory, but in the same time hints at some possible improvement in the choice of the localized prior $\overline{\mu}$, which is here somewhat lacking a variance term. We will consider in the remainder of this section the use of

$$\overline{\mu} = \mu_{\exp[-\alpha\overline{\pi}(R) - \xi\widetilde{\pi}\otimes\widetilde{\pi}(M')}, \tag{2.38}$$

where ξ is some positive real constant and $\widetilde{\pi} = \pi_{\exp(-\widetilde{\beta}R)}$ is some appropriate conditional prior distribution with positive real parameter $\widetilde{\beta}$. With this new choice

$$\mathcal{K}(\nu,\overline{\mu}) = \alpha(\nu-\overline{\mu})\overline{\pi}(R) + \xi(\nu-\overline{\mu})(\widetilde{\pi}\otimes\widetilde{\pi})(M') + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu).$$

We already know how to deal with the first factor $\alpha(\nu-\overline{\mu})\overline{\pi}(R)$, since the computations we made to give it an empirical upper bound were valid for any choice of the localized prior distribution $\overline{\mu}$. Let us now deal with $\xi(\nu-\overline{\mu})(\widetilde{\pi}\otimes\widetilde{\pi})(M')$. Since $m'(\theta,\theta')$ is a sum of independent Bernoulli random variables, we can easily generalize the result of Theorem 1.1.4 (page 17) to prove that with $\mathbb{P}$ probability at least $1-\epsilon$

$$N\big[1-\exp(-\tfrac{\zeta}{N})\big]\nu(\widetilde{\pi}\otimes\widetilde{\pi})(M')$$

$$\leq \zeta \Phi_{\frac{\zeta}{N}}\big[\nu(\widetilde{\pi}\otimes\widetilde{\pi})(M')\big] \leq \zeta\nu(\widetilde{\pi}\otimes\widetilde{\pi})(m') + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon).$$

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

$$-N\big[\exp(\tfrac{\zeta}{N})-1\big]\overline{\mu}(\widetilde{\pi}\otimes\widetilde{\pi})(M') \leq -\zeta\Phi_{-\frac{\zeta}{N}}\big[\overline{\mu}(\widetilde{\pi}\otimes\widetilde{\pi})(M')\big] \leq -\zeta\overline{\mu}(\widetilde{\pi}\otimes\widetilde{\pi})(m') - \log(\epsilon).$$

We would like now to replace $(\widetilde{\pi}\otimes\widetilde{\pi})(m')$ with an empirical quantity. In order to do this, we will use an entropy bound. Indeed for any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}\nu\big[\mathcal{K}(\rho,\widetilde{\pi})\big] &= \widetilde{\beta}\nu(\rho-\widetilde{\pi})(R) + \nu\big[\mathcal{K}(\rho,\pi)-\mathcal{K}(\widetilde{\pi},\pi)\big]\\ &\leq \frac{\widetilde{\beta}}{N\tanh(\frac{\gamma}{N})}\Big\{\gamma\nu(\rho-\widetilde{\pi})(r) + N\log\big[\cosh(\tfrac{\gamma}{N})\big]\nu(\rho\otimes\widetilde{\pi})(m')\\ &\qquad + \mathcal{K}(\nu,\overline{\mu}) + \nu\big[\mathcal{K}(\rho,\widetilde{\pi})\big] - \log(\epsilon)\Big\} + \nu\big[\mathcal{K}(\rho,\pi)-\mathcal{K}(\widetilde{\pi},\pi)\big].\end{aligned}$$

Thus choosing $\widetilde{\beta} = N\tanh(\frac{\gamma}{N})$,

$$\gamma\nu(\widetilde{\pi}-\rho)(r) + \nu\big[\mathcal{K}(\widetilde{\pi},\pi)-\mathcal{K}(\rho,\pi)\big] \leq N\log\big[\cosh(\tfrac{\gamma}{N})\big]\nu(\rho\otimes\widetilde{\pi})(m') + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon).$$

Choosing $\rho = \widehat{\pi}$, we get

$$\nu\big[\mathcal{K}(\widetilde{\pi},\widehat{\pi})\big] \leq N\log\big[\cosh(\tfrac{\gamma}{N})\big]\nu(\widehat{\pi}\otimes\widetilde{\pi})(m') + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon).$$

This implies that

$$\begin{aligned}\xi\nu(\widehat{\pi}\otimes\widetilde{\pi})(m') &= \nu\Big\{\widetilde{\pi}\big[\xi\widehat{\pi}(m')\big] - \mathcal{K}(\widetilde{\pi},\widehat{\pi})\Big\} + \nu\big[\mathcal{K}(\widetilde{\pi},\widehat{\pi})\big]\\ &\leq \nu\Big\{\log\Big[\widehat{\pi}\big\{\exp\big[\xi\widehat{\pi}(m')\big]\big\}\Big]\Big\}\\ &\quad + N\log\big[\cosh(\tfrac{\gamma}{N})\big]\nu(\widehat{\pi}\otimes\widetilde{\pi})(m') + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon).\end{aligned}$$

Thus

$$\big\{\xi - N\log\big[\cosh(\tfrac{\gamma}{N})\big]\big\}\nu(\widehat{\pi}\otimes\widetilde{\pi})(m') \leq \nu\Big\{\log\Big[\widehat{\pi}\big\{\exp\big[\xi\widehat{\pi}(m')\big]\big\}\Big]\Big\} + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon)$$

and

$$\nu\big[\mathcal{K}(\widetilde{\pi},\widehat{\pi})\big] \leq \left(\frac{\xi}{N\log[\cosh(\frac{\gamma}{N})]} - 1\right)^{-1}\Big[\nu\Big\{\log\Big[\widehat{\pi}\big\{\exp\big[\xi\widehat{\pi}(m')\big]\big\}\Big]\Big\} + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon)\Big] + \mathcal{K}(\nu,\overline{\mu}) - \log(\epsilon).$$

Taking for simplicity $\xi = 2N \log\big[\cosh(\frac{\gamma}{N})\big]$ and noticing that

$$2N \log\big[\cosh(\tfrac{\gamma}{N})\big] = -N \log\big(1 - \tfrac{\widetilde{\beta}^2}{N^2}\big),$$

we get

THEOREM 2.3.7. *Let us put $\widetilde{\pi} = \pi_{\exp(-\widetilde{\beta}R)}$ and $\widehat{\pi} = \pi_{\exp(-\gamma r)}$, where γ is some arbitrary positive real constant and $\widetilde{\beta} = N \tanh(\frac{\gamma}{N})$, so that $\gamma = \frac{N}{2} \log\Big(\frac{1+\frac{\widetilde{\beta}}{N}}{1-\frac{\widetilde{\beta}}{N}}\Big)$. With $\mathbb{P}$ probability at least $1-\epsilon$,*

$$\nu\big[\mathcal{K}(\widetilde{\pi}, \widehat{\pi})\big] \leq \nu\bigg[\log\Big\{\widehat{\pi}\Big[\exp\big\{2N \log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\bigg] + 2\big[\mathcal{K}(\nu, \overline{\mu}) - \log(\epsilon)\big].$$

As a consequence

$$\begin{aligned}
\zeta\nu(\widetilde{\pi} \otimes \widetilde{\pi})(m') &= \zeta\nu(\widetilde{\pi} \otimes \widetilde{\pi})(m') - \nu\big[\mathcal{K}(\widetilde{\pi} \otimes \widetilde{\pi}, \widehat{\pi} \otimes \widehat{\pi})\big] + 2\nu\big[\mathcal{K}(\widetilde{\pi}, \widehat{\pi})\big] \\
&\leq \nu\Big\{\log\Big[\widehat{\pi} \otimes \widehat{\pi}\big[\exp(\zeta m')\big]\Big]\Big\} \\
&\quad + 2\nu\bigg[\log\Big\{\widehat{\pi}\Big[\exp\big\{2N \log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\bigg] + 4\big[\mathcal{K}(\nu, \overline{\mu}) - \log(\epsilon)\big].
\end{aligned}$$

Let us take for the sake of simplicity $\zeta = 2N \log\big[\cosh(\frac{\gamma}{N})\big]$, to get

$$\zeta\nu(\widetilde{\pi} \otimes \widetilde{\pi})(m') \leq 3\nu\Big\{\log\Big[\widehat{\pi} \otimes \widehat{\pi}\big[\exp(\zeta m')\big]\Big]\Big\} + 4\big[\mathcal{K}(\nu, \overline{\mu}) - \log(\epsilon)\big].$$

This proves

PROPOSITION 2.3.8. *Let us consider some arbitrary prior distribution $\overline{\mu} \in \mathcal{M}_+^1(M)$ and some arbitrary conditional prior distribution $\pi : M \to \mathcal{M}_+^1(\Theta)$. Let $\widetilde{\beta} < N$ be some positive real constant. Let us put $\widetilde{\pi} = \pi_{\exp(-\widetilde{\beta}R)}$ and $\widehat{\pi} = \pi_{\exp(-\gamma r)}$, with $\widetilde{\beta} = N \tanh(\frac{\gamma}{N})$. Moreover let us put $\zeta = 2N \log\big[\cosh(\frac{\gamma}{N})\big]$. With $\mathbb{P}$ probability at least $1-2\epsilon$, for any posterior distribution $\nu \in \mathcal{M}_+^1(M)$,*

$$\begin{aligned}
\nu(\widetilde{\pi} \otimes \widetilde{\pi})(M') &\leq \frac{3\nu\Big\{\log\Big[\widehat{\pi} \otimes \widehat{\pi}\big[\exp(\zeta m')\big]\Big]\Big\} + 5\big[\mathcal{K}(\nu, \overline{\mu}) - \log(\epsilon)\big]}{N\big[1 - \exp(-\frac{\zeta}{N})\big]} \\
&= \frac{1}{N \tanh(\frac{\gamma}{N})^2}\bigg\{3\nu\bigg[\log\Big\{\widehat{\pi} \otimes \widehat{\pi}\Big[\exp\big\{2N \log\big[\cosh(\tfrac{\gamma}{N})\big] m'\big\}\Big]\Big\}\bigg] \\
&\qquad\qquad + 5\big[\mathcal{K}(\nu, \overline{\mu}) - \log(\epsilon)\big]\bigg\}.
\end{aligned}$$

In the same way,

$$-\zeta\overline{\mu}(\widetilde{\pi} \otimes \widetilde{\pi})(m') \leq \overline{\mu}\Big\{\log\Big[\widehat{\pi} \otimes \widehat{\pi}\big[\exp(-\zeta m')\big]\Big]\Big\}$$

$$+ 2\overline{\mu}\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{2N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big] - 4\log(\epsilon)$$

and thus

$$-\overline{\mu}(\widetilde{\pi}\otimes\widetilde{\pi})(M') \leq \frac{1}{N\big[\exp(\frac{\zeta}{N})-1\big]}\Big\{\overline{\mu}\Big\{\log\Big[\widehat{\pi}\otimes\widehat{\pi}\big[\exp(-\zeta m')\big]\Big]\Big\} + 2\overline{\mu}\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{2N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big] - 5\log(\epsilon)\Big\}.$$

Here we have purposely kept ζ as an arbitrary positive real constant, to be tuned later (in order to be able to strengthen more or less the compensation of variance terms).

We are now properly equipped to estimate the divergence with respect to $\overline{\mu}$, the choice of prior distribution made in equation (2.38, page 128). Indeed we can now write

$$\begin{aligned}
&\Big[1-\frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta}-\frac{\alpha}{N\tanh(\frac{\lambda}{N})(1-\frac{\gamma}{\lambda})}-\frac{5\xi}{N\tanh(\frac{\gamma}{N})^2}\Big]\mathcal{K}(\nu,\overline{\mu})\\
&\leq \frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta}\Big\{\nu\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big]-\log(\epsilon)\Big\}\\
&+\frac{\alpha}{N\tanh(\frac{\gamma}{N})-\beta}\Big\{\overline{\mu}\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big]-\log(\epsilon)\Big\}\\
&\qquad+\frac{\alpha}{N\tanh(\frac{\lambda}{N})}\Big\{\\
&\qquad\lambda(\nu-\overline{\mu})\widehat{\pi}(r)+N\log\big[\cosh(\tfrac{\lambda}{N})\big](\nu\widehat{\pi})\otimes(\overline{\mu}\widehat{\pi})(m')\\
&\qquad+\Big(1-\frac{\gamma}{\lambda}\Big)^{-1}(\nu+\overline{\mu})\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{N\tfrac{\gamma}{\lambda}\log\big[\cosh(\tfrac{\lambda}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big]\\
&\qquad-\frac{1+\frac{\gamma}{N}}{1-\frac{\gamma}{N}}\log(\epsilon)\Big\}\\
&+\frac{\xi}{N\tanh(\frac{\gamma}{N})^2}\Big\{3\nu\Big[\log\Big\{\widehat{\pi}\otimes\widehat{\pi}\Big[\exp\big\{2N\log\big[\cosh(\tfrac{\gamma}{N})\big]m'\big\}\Big]\Big\}\Big]-5\log(\epsilon)\Big\}\\
&\qquad+\frac{\xi}{N\big[\exp(\frac{\zeta}{N})-1\big]}\Big\{\overline{\mu}\Big\{\log\Big[\widehat{\pi}\otimes\widehat{\pi}\big[\exp(-\zeta m')\big]\Big]\Big\}\\
&\qquad+2\overline{\mu}\Big[\log\Big\{\widehat{\pi}\Big[\exp\big\{2N\log\big[\cosh(\tfrac{\gamma}{N})\big]\widehat{\pi}(m')\big\}\Big]\Big\}\Big]-5\log(\epsilon)\Big\}.\\
&\qquad+\mathcal{K}(\nu,\mu)-\mathcal{K}(\overline{\mu},\mu).
\end{aligned}$$

It remains now only to replace in the right-hand side of this inequality $\overline{\mu}$ with the worst possible posterior distribution to obtain

THEOREM 2.3.9. *Let $\lambda > \gamma > \beta$, ζ, α and ξ be arbitrary positive real constants. Let us use the notation $\overline{\pi} = \pi_{\exp(-\beta R)}$, $\widetilde{\pi} = \pi_{\exp(-N\tanh(\frac{\gamma}{N})R)}$,*

$\widehat{\pi} = \pi_{\exp(-\gamma r)}$, $\overline{\mu} = \mu_{\exp[-\alpha\overline{\pi}(R)-\xi\widetilde{\pi}\otimes\widetilde{\pi}(M')]}$ and let us define the posterior distribution $\widehat{\mu} : \Omega \to \mathcal{M}_+^1(M)$ by

$$\frac{d\widehat{\mu}}{d\mu} \sim \exp\biggl\{ -\frac{\alpha\lambda}{N\tanh(\frac{\lambda}{N})}\widehat{\pi}(r) + \frac{\xi}{N\bigl[\exp(\frac{\zeta}{N})-1\bigr]}\log\Bigl\{\widehat{\pi}\otimes\widehat{\pi}\bigl[\exp(-\zeta m')\bigr]\Bigr\}\biggr\}.$$

Let us assume moreover that

$$\frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta} + \frac{\alpha}{N\tanh(\frac{\lambda}{N})(1-\frac{\gamma}{\lambda})} + \frac{5\xi}{N\tanh(\frac{\gamma}{N})^2} < 1.$$

With $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$,

$$\begin{aligned}
\mathcal{K}(\nu,\overline{\mu}) \le \biggl[1 &- \frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta} \\
&- \frac{\alpha}{N\tanh(\frac{\lambda}{N})\bigl(1-\frac{\gamma}{\lambda}\bigr)} - \frac{5\xi}{N\tanh(\frac{\gamma}{N})^2}\biggr]^{-1}\biggl\{\mathcal{K}(\nu,\widehat{\mu}) \\
&+ \frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta}\biggl\{\nu\biggl[\log\Bigl\{\widehat{\pi}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\widehat{\pi}(m')\bigr\}\Bigr]\Bigr\}\biggr]\biggr\} \\
&+ \frac{\alpha}{N\tanh(\frac{\lambda}{N})\bigl(1-\frac{\gamma}{\lambda}\bigr)}\biggl\{\nu\biggl[\log\Bigl\{\widehat{\pi}\Bigl[\exp\bigl\{N\tfrac{\gamma}{\lambda}\log\bigl[\cosh(\tfrac{\lambda}{N})\bigr]\widehat{\pi}(m')\bigr\}\Bigr]\Bigr\}\biggr]\biggr\} \\
&+ \frac{\xi}{N\tanh(\frac{\gamma}{N})^2}\biggl\{3\nu\biggl[\log\Bigl\{\widehat{\pi}\otimes\widehat{\pi}\Bigl[\exp\bigl\{2N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]m'\bigr\}\Bigr]\Bigr\}\biggr]\biggr\} \\
&+ \frac{\xi}{N\bigl[\exp(\frac{\zeta}{N})-1\bigr]}\biggl\{\nu\Bigl\{\log\Bigl[\widehat{\pi}\otimes\widehat{\pi}\bigl[\exp(-\zeta m')\bigr]\Bigr]\Bigr\}\biggr\} \\
&+ \log\biggl\{\widehat{\mu}\biggl[\Bigl\{\widehat{\pi}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\widehat{\pi}(m')\bigr\}\Bigr]\Bigr\}^{\frac{\alpha}{N\tanh(\frac{\gamma}{N})-\beta}} \\
&\times\Bigl\{\widehat{\pi}\Bigl[\exp\bigl\{N\tfrac{\gamma}{\lambda}\log\bigl[\cosh(\tfrac{\lambda}{N})\bigr]\widehat{\pi}(m')\bigr\}\Bigr]\Bigr\}^{\frac{\alpha}{N\tanh(\frac{\lambda}{N})\left(1-\frac{\gamma}{\lambda}\right)}} \\
&\times\Bigl\{\widehat{\pi}\Bigl[\exp\bigl\{2N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\widehat{\pi}(m')\bigr\}\Bigr]\Bigr\}^{\frac{2\xi}{N\left[\exp(\frac{\zeta}{N})-1\right]}} \\
&\times\exp\Bigl\{N\log\bigl[\cosh(\tfrac{\lambda}{N})\bigr]\bigl[(\nu\widehat{\pi})\otimes\widehat{\pi}\bigr](m')\Bigr\}\biggr]\biggr\} \\
&+ \biggl[\frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta} + \frac{\alpha}{N\tanh(\frac{\gamma}{N})-\beta} + \frac{2\alpha\bigl(1+\frac{\gamma}{N}\bigr)}{N\tanh(\frac{\lambda}{N})\bigl(1-\frac{\gamma}{\lambda}\bigr)} \\
&\qquad + \frac{5\xi}{N\tanh(\frac{\gamma}{N})^2} + \frac{5\xi}{N\bigl[\exp(\frac{\zeta}{N})-1\bigr]}\biggr]\log\bigl(\tfrac{5}{\epsilon}\bigr)\biggr\}.
\end{aligned}$$

The interest of this theorem lies in the presence of a variance term in the localized posterior distribution $\widehat{\mu}$, which with a suitable choice of parameters seems to be an interesting option in the case when there are nested models: in this situation there may be a need to prevent integration with respect to $\widehat{\mu}$ in the right-hand side to put weight on wild oversized models with large variance terms. Moreover, the right-hand side being empirical, parameters can be, as usual, optimized from data using a union bound on a grid of candidate values.

If one is only interested in the general shape of the result, a simplified inequality as the one below may suffice:

COROLLARY 2.3.10. *For any positive real constants $\lambda > \gamma > \beta$, ζ, α and ξ, let us use the same notation as in Theorem 2.3.9 (page 131). Let us put moreover*

$$
\begin{aligned}
A_1 &= \frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta} + \frac{\alpha}{N\tanh(\frac{\lambda}{N})\big(1-\frac{\gamma}{\lambda}\big)} + \frac{5\xi}{N\tanh(\frac{\gamma}{N})^2},\\
A_2 &= \frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta} + \frac{\alpha}{N\tanh(\frac{\lambda}{N})\big(1-\frac{\gamma}{\lambda}\big)} + \frac{3\xi}{N\tanh(\frac{\gamma}{N})^2}\\
A_3 &= \frac{\xi}{N\big[\exp\big(\frac{\zeta}{N}\big)-1\big]}\\
A_4 &= \frac{\alpha}{N\tanh(\frac{\gamma}{N})-\beta} + \frac{\alpha}{N\tanh(\frac{\lambda}{N})(1-\frac{\gamma}{\lambda})} + \frac{2\xi}{N[\exp(\frac{\zeta}{N})-1]},\\
A_5 &= \frac{\alpha}{N\tanh(\frac{\gamma}{N})+\beta} + \frac{\alpha}{N\tanh(\frac{\gamma}{N})-\beta} + \frac{2\alpha\big(1+\frac{\gamma}{N}\big)}{N\tanh(\frac{\lambda}{N})\big(1-\frac{\gamma}{\lambda}\big)}\\
&\qquad + \frac{5\xi}{N\tanh(\frac{\gamma}{N})^2} + \frac{5\xi}{N\big[\exp(\frac{\zeta}{N})-1\big]},\\
C_1 &= 2N\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big],\\
C_2 &= N\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big].
\end{aligned}
$$

Let us assume that $A_1 < 1$. With $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$,

$$
\begin{aligned}
\mathcal{K}(\nu,\overline{\mu}) \le K(\nu,\alpha,\beta,\gamma,\lambda,\xi,\zeta,\epsilon) &\stackrel{\text{def}}{=} \big(1-A_1\big)^{-1}\Bigg\{\mathcal{K}(\nu,\widehat{\mu})\\
&+ A_2\nu\Big[\log\Big(\widehat{\pi}\otimes\widehat{\pi}\big[\exp\big(C_1 m'\big)\big]\Big)\Big] + A_3\nu\Big[\log\Big(\widehat{\pi}\otimes\widehat{\pi}\big[\exp\big(-\zeta m'\big)\big]\Big)\Big]\\
&+ \log\Bigg\{\widehat{\mu}\Bigg[\Big[\widehat{\pi}\Big(\exp\big[C_1\widehat{\pi}(m')\big]\Big)\Big]^{A_4}\exp\Big(C_2\big[(\nu\widehat{\pi})\otimes\widehat{\pi}\big](m')\Big)\Bigg]\Bigg\}\\
&+ A_5\log\big(\tfrac{5}{\epsilon}\big)\Bigg\}.
\end{aligned}
$$

Putting this corollary together with Corollary 2.3.5 (page 125), we obtain

THEOREM 2.3.11. *Let us consider the notation introduced in Corollary 2.3.5 (page 125) and in Theorem 2.3.9 (page 131) and its Corollary 2.3.10 (page 133). Let us consider real positive parameters* λ, $\gamma_1' < \lambda_1'$ *and* $\gamma_2' < \lambda_2'$. *Let us consider also two sets of parameters* $\alpha_i, \beta_i, \gamma_i, \lambda_i, \xi_i, \zeta_i$, *where* $i = 1, 2$, *both satisfying the conditions stated in Corollary 2.3.10 (page 133). With* $\mathbb{P}$ *probability at least* $1 - \epsilon$, *for any posterior distributions* $\nu_1, \nu_2 : \Omega \to \mathcal{M}_+^1(M)$, *any conditional posterior distributions* $\rho_1, \rho_2 : \Omega \times M \to \mathcal{M}_+^1\big(\Theta\big)$,

$$
\begin{aligned}
-N\log\Big\{1-\tanh\big(\tfrac{\lambda}{N}\big)\big(\nu_1\rho_1-\nu_2\rho_2\big)(R)\Big\} &\leq \lambda\big(\nu_1\rho_1-\nu_2\rho_2\big)(r)\\
&\quad + N\log\big[\cosh\big(\tfrac{\lambda}{N}\big)\big]\big(\nu_1\rho_1\big)\otimes\big(\nu_2\rho_2\big)\big(m'\big)\\
&\quad + K'\big(\nu_1,\rho_1,\gamma_1',\lambda_1',\tfrac{\epsilon}{5}\big) + K'\big(\nu_2,\rho_2,\gamma_2',\lambda_2',\tfrac{\epsilon}{5}\big)\\
&\quad + \frac{1}{1-\frac{\gamma_1'}{\lambda_1'}}K\big(\nu_1,\alpha_1,\beta_1,\gamma_1,\lambda_1,\xi_1,\zeta_1,\tfrac{\epsilon}{5}\big)\\
&\quad + \frac{1}{1-\frac{\gamma_2'}{\lambda_2'}}K\big(\nu_2,\alpha_2,\beta_2,\gamma_2,\lambda_2,\xi_2,\zeta_2,\tfrac{\epsilon}{5}\big) - \log\big(\tfrac{\epsilon}{5}\big).
\end{aligned}
$$

This theorem provides, using a union bound argument to further optimize the parameters, an empirical bound for $\nu_1\rho_1(R) - \nu_2\rho_2(R)$, which can serve to build a selection algorithm exactly in the same way as what was done in Theorem 2.2.4 (page 93). This represents the highest degree of sophistication that we will achieve in this monograph, as far as model selection is concerned: this theorem shows that it is indeed possible to derive a selection scheme in which localization is performed in two steps and in which the localization of the model selection itself, as opposed to the localization of the estimation in each model, includes a variance term as well as a bias term, so that it should be possible to localize the choice of nested models, something that would not have been feasible with the localization techniques exposed in the previous sections of this study. We should point out however that *more sophisticated* does not necessarily mean *more efficient*: as the reader may have noticed, sophistication comes at a price, in terms of the complexity of the estimation schemes, with some possible loss of accuracy in the constants that can mar the benefits of using an asymptotically more efficient method for small sample sizes.

We will do the hurried reader a favour: we will not launch into a study of the theoretical properties of this selection algorithm, although it is clear that all the tools needed are at hand !

We would like as a conclusion to this chapter, to put forward a simple idea: this approach of model selection revolves around entropy estimates concerned with the divergence of posterior distributions with respect to localized prior distributions. Moreover, this localization of the prior distribution is more effectively done in several steps in some situations, and it is worth mentioning that these situations include the typical case of selection from a family of parametric models. Finally, the whole story relies upon es-

timating the relative generalization error rate of one posterior distribution with respect to some local prior distribution as well as with respect to another posterior distribution, because these relative rates can be estimated more accurately than absolute generalization error rates, at least as soon as no classification model of reasonable size provides a good match to the training sample, meaning that the classification problem is either difficult or noisy.

Chapter 3

Transductive PAC-Bayesian learning

3.1. Basic inequalities

3.1.1. The transductive setting. In this chapter the *observed* sample $(X_i, Y_i)_{i=1}^N$ will be supplemented with a *test* or *shadow* sample $(X_i, Y_i)_{i=N+1}^{(k+1)N}$. This point of view, called *transductive classification*, has been introduced by V. Vapnik. It may be justified in different ways.

On the practical side, one interest of the transductive setting is that it is often a lot easier to collect examples than it is to label them, so that it is not unrealistic to assume that we indeed have two training samples, one labelled and one unlabelled. It also covers the case when a batch of patterns is to be classified and we are allowed to observe the whole batch before issuing the classification.

On the mathematical side, considering a shadow sample proves technically fruitful. Indeed, when introducing the Vapnik–Cervonenkis entropy and Vapnik–Cervonenkis dimension concepts, as well as when dealing with compression schemes, albeit the *inductive* setting is our final concern, the transductive setting is a useful detour. In this second scenario, intermediate technical results involving the shadow sample are integrated with respect to unobserved random variables in a second stage of the proofs.

Let us describe now the changes to be made to previous notation to adapt them to the transductive setting. The distribution $\mathbb{P}$ will be a probability measure on the canonical space $\Omega = (\mathcal{X} \times \mathcal{Y})^{(k+1)N}$, and $(X_i, Y_i)_{i=1}^{(k+1)N}$ will be the canonical process on this space (that is the coordinate process). Unless explicitly mentioned, the parameter k indicating the size of the shadow sample will remain fixed. Assuming the shadow sample size is a multiple of the training sample size is convenient without significantly restricting generality. For a while, we will use a weaker assumption than independence, assuming that $\mathbb{P}$ is *partially exchangeable*, since this is all we need in the proofs.

DEFINITION 3.1.1. For $i = 1, \ldots, N$, let $\tau_i : \Omega \to \Omega$ be defined for any $\omega = (\omega_j)_{j=1}^{(k+1)N} \in \Omega$ by

$$\begin{cases} \tau_i(\omega)_{i+jN} = \omega_{i+(j-1)N}, & j = 1, \ldots, k, \\ \tau_i(\omega)_i = \omega_{i+kN}, & \\ \text{and } \tau_i(\omega)_{m+jN} = \omega_{m+jN}, & m \neq i, m = 1, \ldots, N, j = 0, \ldots k. \end{cases}$$

Clearly, if we arrange the $(k+1)N$ samples in a $N \times (k+1)$ array, τ_i performs a circular permutation of $k+1$ entries on the ith row, leaving the other rows unchanged. Moreover, all the circular permutations of the ith row have the form τ_i^j, j ranging from 0 to k.

The probability distribution $\mathbb{P}$ is said to be partially exchangeable if for any $i = 1, \ldots, N$, $\mathbb{P} \circ \tau_i^{-1} = \mathbb{P}$.

This means equivalently that for any bounded measurable function $h : \Omega \to \mathbb{R}$, $\mathbb{P}(h \circ \tau_i) = \mathbb{P}(h)$.

In the same way a function h defined on Ω will be said to be partially exchangeable if $h \circ \tau_i = h$ for any $i = 1, \ldots, N$. Accordingly a posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta, \mathcal{T})$ will be said to be partially exchangeable when $\rho(\omega, A) = \rho\big[\tau_i(\omega), A\big]$, for any $\omega \in \Omega$, any $i = 1, \ldots, N$ and any $A \in \mathcal{T}$.

For any bounded measurable function h, let us define $T_i(h) = \frac{1}{k+1} \sum_{j=0}^{k} h \circ \tau_i^j$. Let $T(h) = T_N \circ \cdots \circ T_1(h)$. For any partially exchangeable probability distribution $\mathbb{P}$, and for any bounded measurable function h, $\mathbb{P}\big[T(h)\big] = \mathbb{P}(h)$. Let us put

$$\sigma_i(\theta) = \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big], \quad \text{indicating the success or failure of } f_\theta \text{ to predict } Y_i \text{ from } X_i,$$

$$r_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i(\theta), \quad \text{the empirical error rate of } f_\theta \text{ on the observed sample,}$$

$$r_2(\theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \sigma_i(\theta), \quad \text{the error rate of } f_\theta \text{ on the shadow sample,}$$

$$\overline{r}(\theta) = \frac{r_1(\theta) + k r_2(\theta)}{k+1} = \frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} \sigma_i(\theta), \quad \text{the global error rate of } f_\theta,$$

$$R_i(\theta) = \mathbb{P}\big[f_\theta(X_i) \neq Y_i\big], \quad \text{the expected error rate of } f_\theta \text{ on the } i\text{th input,}$$

$$R(\theta) = \frac{1}{N} \sum_{i=1}^{N} R_i(\theta) = \mathbb{P}\big[r_1(\theta)\big] = \mathbb{P}\big[r_2(\theta)\big], \quad \text{the average expected error rate of } f_\theta \text{ on all inputs.}$$

We will allow for posterior distributions $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ depending on the shadow sample. The most interesting ones will anyhow be independent of

the shadow labels $Y_{N+1}, \dots, Y_{(k+1)N}$. We will be interested in the conditional expected error rate of the randomized classification rule described by ρ on the shadow sample, given the observed sample, that is, $\mathbb{P}\big[\rho(r_2)|(X_i,Y_i)_{i=1}^N\big]$. This is a natural extension of the notion of *generalization error rate*: this is indeed the error rate to be expected when the randomized classification rule described by the posterior distribution ρ is applied to the shadow sample (which should in this case more purposefully be called the test sample).

To see the connection with the previously defined generalization error rate, let us comment on the case when $\mathbb{P}$ is invariant by any permutation of any row, meaning that

$$\mathbb{P}\big[h(\omega\circ s)\big]=\mathbb{P}\big[h(\omega)\big] \text{ for all } s\in\mathfrak{S}(\{i+jN; j=0,\dots,k\})$$

and all $i=1,\dots,N$, where $\mathfrak{S}(A)$ is the set of permutations of A, extended to $\{1,\dots,(k+1)N\}$ so as to be the identity outside of A. In other words, $\mathbb{P}$ is assumed to be invariant under any permutation which keeps the rows unchanged. In this case, if ρ is invariant by any permutation of any row of the shadow sample, meaning that $\rho(\omega\circ s)=\rho(\omega)\in\mathcal{M}_+^1(\Theta)$, $s\in\mathfrak{S}(\{i+jN; j=1,\dots,k\})$, $i=1,\dots,N$, then $\mathbb{P}\big[\rho(r_2)|(X_i,Y_i)_{i=1}^N\big]=\frac{1}{N}\sum_{i=1}^N\mathbb{P}\big[\rho(\sigma_{i+N})|(X_i,Y_i)_{i=1}^N\big]$, meaning that the expectation can be taken on a restricted shadow sample of the same size as the observed sample. If moreover the rows are equidistributed, meaning that their marginal distributions are equal, then

$$\mathbb{P}\big[\rho(r_2)|(X_i,Y_i)_{i=1}^N\big]=\mathbb{P}\big[\rho(\sigma_{N+1})|(X_i,Y_i)_{i=1}^N\big].$$

This means that under these quite commonly fulfilled assumptions, the expectation can be taken on a single new object to be classified, our study thus covers the case when only one of the patterns from the shadow sample is to be labelled and one is interested in the expected error rate of this single labelling. Of course, in the case when $\mathbb{P}$ is i.i.d. and ρ depends only on the training sample $(X_i,Y_i)_{i=1}^N$, we fall back on the usual criterion of performance $\mathbb{P}\big[\rho(r_2)|(Z_i)_{i=1}^N\big]=\rho(R)=\rho(R_1)$.

3.1.2. Absolute bound. Using an obvious factorization, and considering for the moment a fixed value of θ and any partially exchangeable positive real measurable function $\lambda:\Omega\to\mathbb{R}_+$, we can compute the log-Laplace transform of r_1 under T, which acts like a conditional probability distribution:

$$\begin{aligned}\log\Big\{T\big[\exp(-\lambda r_1)\big]\Big\}&=\sum_{i=1}^N\log\Big\{T_i\big[\exp(-\tfrac{\lambda}{N}\sigma_i)\big]\Big\}\\&\le N\log\Big\{\frac{1}{N}\sum_{i=1}^N T_i\Big[\exp\big(-\tfrac{\lambda}{N}\sigma_i\big)\Big]\Big\}=-\lambda\Phi_{\frac{\lambda}{N}}(\overline{r}),\end{aligned}$$

where the function $\Phi_{\frac{\lambda}{N}}$ was defined by equation (1.1, page 15). Remarking that $T\Big\{\exp\Big[\lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r})-r_1\big]\Big]\Big\}=\exp\big[\lambda\Phi_{\frac{\lambda}{N}}(\overline{r})\big]T\big[\exp(-\lambda r_1)\big]$ we obtain

LEMMA 3.1.1. *For any $\theta \in \Theta$ and any partially exchangeable positive real measurable function $\lambda : \Omega \to \mathbb{R}_+$,*

$$T\Big\{\exp\Big[\lambda\big\{\Phi_{\frac{\lambda}{N}}\big[\overline{r}(\theta)\big] - r_1(\theta)\big\}\Big]\Big\} \leq 1.$$

We deduce from this lemma a result analogous to the inductive case:

THEOREM 3.1.2. *For any partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, for any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\bigg\{\exp\Big[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho\Big[\lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1\big]\Big] - \mathcal{K}(\rho,\pi)\Big]\bigg\} \leq 1.$$

The proof is deduced from the previous lemma, using the fact that π is partially exchangeable:

$$\begin{aligned}
&\mathbb{P}\bigg\{\exp\Big[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho\Big[\lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1\big]\Big] - \mathcal{K}(\rho,\pi)\Big]\bigg\} \\
&\quad = \mathbb{P}\Big\{\pi\Big\{\exp\Big[\lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1\big]\Big]\Big\}\Big\} = \mathbb{P}\Big\{T\pi\Big\{\exp\Big[\lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1\big]\Big]\Big\}\Big\} \\
&\qquad\qquad = \mathbb{P}\Big\{\pi\Big\{T\exp\Big[\lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1\big]\Big]\Big\}\Big\} \leq 1.
\end{aligned}$$

3.1.3. RELATIVE BOUNDS. Introducing in the same way

$$m'(\theta,\theta') = \frac{1}{N}\sum_{i=1}^{N}\Big|\mathbb{1}\big[f_\theta(X_i) \neq Y_i\big] - \mathbb{1}\big[f_{\theta'}(X_i) \neq Y_i\big]\Big|$$

$$\text{and} \quad \overline{m}(\theta,\theta') = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\Big|\mathbb{1}\big[f_\theta(X_i) \neq Y_i\big] - \mathbb{1}\big[f_{\theta'}(X_i) \neq Y_i\big]\Big|,$$

we could prove along the same line of reasoning

THEOREM 3.1.3. *For any real parameter λ, any $\widetilde{\theta} \in \Theta$, any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}
\mathbb{P}\bigg\{\exp\Big[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\Big[\rho\big\{\Psi_{\frac{\lambda}{N}}\big[\overline{r}(\cdot) - \overline{r}(\widetilde{\theta}), \overline{m}(\cdot,\widetilde{\theta})\big]\big\}& \\
- \big[\rho(r_1) - r_1(\widetilde{\theta})\big]\Big] - \mathcal{K}(\rho,\pi)\Big]\bigg\} \leq 1,&
\end{aligned}$$

where the function $\Psi_{\frac{\lambda}{N}}$ was defined by equation (1.21, page 51).

THEOREM 3.1.4. *For any real constant γ, for any $\widetilde{\theta} \in \Theta$, for any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\Bigg\{\exp\Bigg[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Bigg\{-N\rho\Big\{\log\Big[1-\tanh\big(\tfrac{\gamma}{N}\big)\big[\overline{r}(\cdot)-\overline{r}(\widetilde{\theta})\big]\Big]\Big\}$$
$$-\gamma\big[\rho(r_1)-r_1(\widetilde{\theta})\big]-N\log\big[\cosh\big(\tfrac{\gamma}{N}\big)\big]\rho\big[m'(\cdot,\widetilde{\theta})\big]-\mathcal{K}(\rho,\pi)\Bigg\}\Bigg]\Bigg\}\leq 1.$$

This last theorem can be generalized to give

THEOREM 3.1.5. *For any real constant γ, for any partially exchangeable posterior distributions $\pi^1, \pi^2 : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\Bigg\{\exp\Bigg[\sup_{\rho_1,\rho_2\in\mathcal{M}_+^1(\Theta)}\Bigg\{-N\log\Big\{1-\tanh\big(\tfrac{\gamma}{N}\big)\big[\rho_1(\overline{r})-\rho_2(\overline{r})\big]\Big\}$$
$$-\gamma\big[\rho_1(r_1)-\rho_2(r_1)\big]-N\log\big[\cosh\big(\tfrac{\gamma}{N}\big)\big]\rho_1\otimes\rho_2(m')$$
$$-\mathcal{K}(\rho_1,\pi^1)-\mathcal{K}(\rho_2,\pi^2)\Bigg\}\Bigg]\Bigg\}\leq 1.$$

To conclude this section, we see that the basic theorems of transductive PAC-Bayesian classification have exactly the same form as the basic inequalities of inductive classification, Theorems 1.1.4 (page 17), 1.4.2 (page 52) and 1.4.3 (page 53) *with $R(\theta)$ replaced with $\overline{r}(\theta)$*, $r(\theta)$ replaced with $r_1(\theta)$ and $M'(\theta,\widetilde{\theta})$ replaced with $\overline{m}(\theta,\widetilde{\theta})$.

Thus all the results of the first two chapters remain true under the hypotheses of transductive classification, with $R(\theta)$ replaced with $\overline{r}(\theta)$, $r(\theta)$ replaced with $r_1(\theta)$ and $M'(\theta,\widetilde{\theta}\,)$ replaced with $\overline{m}(\theta,\widetilde{\theta})$.

Consequently, in the case when the unlabelled shadow sample is observed, it is possible to improve on the Vapnik bounds to be discussed hereafter by using an explicit partially exchangeable posterior distribution π and resorting to localized or to relative bounds (in the case at least of unlimited computing resources, which of course may still be unrealistic in many real world situations, and with the caveat, to be recalled in the conclusion of this study, that for small sample sizes and comparatively complex classification models, the improvement may not be so decisive).

Let us notice also that the transductive setting when experimentally available, has the advantage that

$$\overline{d}(\theta,\theta')=\frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\mathbb{1}\big[f_{\theta'}(X_i)\neq f_\theta(X_i)\big]$$
$$\geq\overline{m}(\theta,\theta')\geq\overline{r}(\theta)-\overline{r}(\theta'),\qquad\theta,\theta'\in\Theta,$$

is observable in this context, providing an empirical upper bound for the difference $\overline{r}(\widehat{\theta}) - \rho(\overline{r})$ for any non-randomized estimator $\widehat{\theta}$ and any posterior distribution ρ, namely

$$\overline{r}(\widehat{\theta}) \leq \rho(\overline{r}) + \rho\big[\,\overline{d}(\cdot, \widehat{\theta})\big].$$

Thus in the setting of transductive statistical experiments, the PAC-Bayesian framework provides fully empirical bounds for the error rate of non-randomized estimators $\widehat{\theta} : \Omega \to \Theta$, even when using a non-atomic prior π (or more generally a non-atomic partially exchangeable posterior distribution π), even when Θ is not a vector space and even when $\theta \mapsto R(\theta)$ cannot be proved to be convex on the support of some useful posterior distribution ρ.

3.2. Vapnik bounds for transductive classification

In this section, we will stick to plain unlocalized non-relative bounds. As we have already mentioned, (and as it was put forward by Vapnik himself in his seminal works), these bounds are not always superseded by the asymptotically better ones when the sample is of small size: they deserve all our attention for this reason. We will start with the general case of a shadow sample of arbitrary size. We will then discuss the case of a shadow sample of equal size to the training set and the case of a fully exchangeable sample distribution, showing how they can be taken advantage of to sharpen inequalities.

3.2.1. With a shadow sample of arbitrary size.

The great thing with the transductive setting is that we are manipulating only r_1 and $\overline{r}$ which can take only a finite number of values and therefore are piecewise constant on Θ. This makes it possible to derive inequalities that will hold uniformly for any value of the parameter $\theta \in \Theta$. To this purpose, let us consider for any value $\theta \in \Theta$ of the parameter the subset $\Delta(\theta) \subset \Theta$ of parameters θ' such that the classification rule $f_{\theta'}$ answers the same on the extended sample $(X_i)_{i=1}^{(k+1)N}$ as f_θ. Namely, let us put for any $\theta \in \Theta$

$$\Delta(\theta) = \big\{\theta' \in \Theta; f_{\theta'}(X_i) = f_\theta(X_i), i = 1, \ldots, (k+1)N\big\}.$$

We see immediately that $\Delta(\theta)$ is an exchangeable parameter subset on which r_1 and r_2 and therefore also $\overline{r}$ take constant values. Thus for any $\theta \in \Theta$ we may consider the posterior ρ_θ defined by

$$\frac{d\rho_\theta}{d\pi}(\theta') = \mathbb{1}\big[\theta' \in \Delta(\theta)\big]\pi\big[\Delta(\theta)\big]^{-1},$$

and use the fact that $\rho_\theta(r_1) = r_1(\theta)$ and $\rho_\theta(\overline{r}) = \overline{r}(\theta)$, to prove that

LEMMA 3.2.1. *For any partially exchangeable positive real measurable function* $\lambda : \Omega \times \Theta \to \mathbb{R}$ *such that*

$$\lambda(\omega, \theta') = \lambda(\omega, \theta), \quad \theta \in \Theta, \theta' \in \Delta(\theta), \omega \in \Omega, \tag{3.1}$$

and any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$*, with* $\mathbb{P}$ *probability at least* $1 - \epsilon$*, for any* $\theta \in \Theta$*,*

$$\Phi_{\frac{\lambda}{N}}\big[\overline{r}(\theta)\big] + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda(\theta)} \leq r_1(\theta).$$

We can then remark that for any value of λ independent of ω, the left-hand side of the previous inequality is a partially exchangeable function of $\omega \in \Omega$. Thus this left-hand side is maximized by some partially exchangeable function λ, namely

$$\arg\max_{\lambda} \left\{ \Phi_{\frac{\lambda}{N}}\big[\overline{r}(\theta)\big] + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda} \right\}$$

is partially exchangeable as depending only on partially exchangeable quantities. Moreover this choice of $\lambda(\omega, \theta)$ satisfies also condition (3.1) stated in the previous lemma of being constant on $\Delta(\theta)$, proving

LEMMA 3.2.2. *For any partially exchangeable posterior distribution* $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$*, with* $\mathbb{P}$ *probability at least* $1 - \epsilon$*, for any* $\theta \in \Theta$ *and any* $\lambda \in \mathbb{R}_+$*,*

$$\Phi_{\frac{\lambda}{N}}\big[\overline{r}(\theta)\big] + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda} \leq r_1(\theta).$$

Writing $\overline{r} = \frac{r_1 + k r_2}{k+1}$ and rearranging terms we obtain

THEOREM 3.2.3. *For any partially exchangeable posterior distribution* $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$*, with* $\mathbb{P}$ *probability at least* $1 - \epsilon$*, for any* $\theta \in \Theta$*,*

$$r_2(\theta) \leq \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{1 - \exp\left(-\frac{\lambda}{N} r_1(\theta) + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{N} \right)}{1 - \exp\big(-\frac{\lambda}{N}\big)} - \frac{r_1(\theta)}{k}.$$

If we have a set of binary classification rules $\{f_\theta; \theta \in \Theta\}$ whose Vapnik–Cervonenkis dimension is not greater than h, we can choose π such that $\pi\big[\Delta(\theta)\big]$ is independent of θ and not less than $\left(\frac{h}{e(k+1)N}\right)^h$, as will be proved further on in Theorem 4.2.2 (page 174).

Another important setting where the complexity term $-\log\big\{\pi\big[\Delta(\theta)\big]\big\}$ can easily be controlled is the case of *compression schemes*, introduced by

Little et al. (1986). It goes as follows: we are given for each labelled sub-sample $(X_i, Y_i)_{i\in J}$, $J \subset \{1, \dots, N\}$, an estimator of the parameter

$$\widehat{\theta}\big[(X_i, Y_i)_{i\in J}\big] = \widehat{\theta}_J, \quad J \subset \{1, \dots, N\}, |J| \leq h,$$

where

$$\widehat{\theta} : \bigsqcup_{k=1}^{N} \big(\mathcal{X} \times \mathcal{Y}\big)^k \to \Theta$$

is an exchangeable function providing estimators for sub-samples of arbitrary size. Let us assume that $\widehat{\theta}$ is exchangeable, meaning that for any $k = 1, \dots, N$ and any permutation σ of $\{1, \dots, k\}$

$$\widehat{\theta}\big[(x_i, y_i)_{i=1}^k\big] = \widehat{\theta}\big[(x_{\sigma(i)}, y_{\sigma(i)})_{i=1}^k\big], \qquad (x_i, y_i)_{i=1}^k \in \big(\mathcal{X} \times \mathcal{Y}\big)^k.$$

In this situation, we can introduce the exchangeable subset

$$\Big\{\widehat{\theta}_J; J \subset \{1, \dots, (k+1)N\}, |J| \leq h\Big\} \subset \Theta,$$

which is seen to contain at most

$$\sum_{j=0}^{h} \binom{(k+1)N}{j} \leq \left(\frac{e(k+1)N}{h}\right)^h$$

classification rules — as will be proved later on in Theorem 4.2.3 (page 174). Note that we had to extend the range of J to all the subsets of the extended sample, although we will use for estimation only those of the training sample, on which the labels are observed. Thus in this case also we can find a partially exchangeable posterior distribution π such that

$$\pi\big[\Delta(\widehat{\theta}_J)\big] \geq \left(\frac{h}{e(k+1)N}\right)^h.$$

We see that the size of the compression scheme plays the same role in this complexity bound as the Vapnik–Cervonenkis dimension for Vapnik–Cervonenkis classes.

In these two cases of binary classification with Vapnik–Cervonenkis dimension not greater than h and compression schemes depending on a compression set with at most h points, we get a bound of

$$r_2(\theta) \leq \frac{k+1}{k} \inf_{\lambda\in\mathbb{R}_+} \frac{1 - \exp\left(-\frac{\lambda}{N} r_1(\theta) - \frac{h \log\left(\frac{e(k+1)N}{h}\right) - \log(\epsilon)}{N}\right)}{1 - \exp\big(-\frac{\lambda}{N}\big)} - \frac{r_1(\theta)}{k}.$$

Let us make some numerical application: when $N = 1000, h = 10, \epsilon = 0.01$, and $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$, we find that $r_2(\widehat{\theta}) \leq 0.4093$, for k between 15 and 17, and values of λ equal respectively to 965, 968 and 971. For $k = 1$, we find only $r_2(\widehat{\theta}) \leq 0.539$, showing the interest of allowing k to be larger than 1.

3.2.2. When the shadow sample has the same size as the training sample. In the case when $k = 1$, we can improve Theorem 3.1.2 by taking advantage of the fact that $T_i(\sigma_i)$ can take only 3 values, namely 0, 0.5 and 1. We see thus that $T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\big[T_i(\sigma_i)\big]$ can take only two values, 0 and $\frac{1}{2} - \Phi_{\frac{\lambda}{N}}(\frac{1}{2})$, because $\Phi_{\frac{\lambda}{N}}(0) = 0$ and $\Phi_{\frac{\lambda}{N}}(1) = 1$. Thus

$$T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\big[T_i(\sigma_i)\big] = \big[1 - |1 - 2T_i(\sigma_i)|\big]\big[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\big].$$

This shows that in the case when $k = 1$,

$$\begin{aligned}\log\Big\{T\big[\exp(-\lambda r_1)\big]\Big\} &= -\lambda\overline{r} + \frac{\lambda}{N}\sum_{i=1}^N T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\big[T_i(\sigma_i)\big] \\ &= -\lambda\overline{r} + \frac{\lambda}{N}\sum_{i=1}^N \big[1 - |1 - 2T_i(\sigma_i)|\big]\big[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\big] \\ &\leq -\lambda\overline{r} + \lambda\big[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\big]\big[1 - |1 - 2\overline{r}|\big].\end{aligned}$$

Noticing that $\frac{1}{2} - \Phi_{\frac{\lambda}{N}}(\frac{1}{2}) = \frac{N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]$, we obtain

Theorem 3.2.4. *For any partially exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, for any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\Big\{\exp\Big[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho\Big[\lambda(\overline{r} - r_1) - N\log\big[\cosh(\tfrac{\lambda}{2N})\big]\big(1 - |1 - 2\overline{r}|\big)\Big] - \mathcal{K}(\rho,\pi)\Big]\Big\} \leq 1.$$

As a consequence, reasoning as previously, we deduce

Theorem 3.2.5. *In the case when $k = 1$, for any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$ and any $\lambda \in \mathbb{R}_+$,*

$$\overline{r}(\theta) - \tfrac{N}{\lambda}\log\big[\cosh(\tfrac{\lambda}{2N})\big]\big(1 - |1 - 2\overline{r}(\theta)|\big) + \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda} \leq r_1(\theta);$$

and consequently for any $\theta \in \Theta$,

$$r_2(\theta) \leq 2\inf_{\lambda\in\mathbb{R}_+} \frac{r_1(\theta) - \dfrac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda}}{1 - \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]} - r_1(\theta).$$

In the case of binary classification using a Vapnik–Cervonenkis class of Vapnik–Cervonenkis dimension not greater than h, we can choose π such that $-\log\{\pi[\Delta(\theta)]\} \leq h\log(\frac{2eN}{h})$ and obtain the following numerical illustration of this theorem: for $N = 1000$, $h = 10$, $\epsilon = 0.01$ and $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$, we find an upper bound $r_2(\widehat{\theta}) \leq 0.5033$, which improves on Theorem 3.2.3 but still is not under the significance level $\frac{1}{2}$ (achieved by blind random classification). This indicates that considering shadow samples of arbitrary sizes some noisy situations yields a significant improvement on bounds obtained with a shadow sample of the same size as the training sample.

3.2.3. When moreover the distribution of the augmented sample is exchangeable. When $k = 1$ and $\mathbb{P}$ is exchangeable meaning that for any bounded measurable function $h : \Omega \to \mathbb{R}$ and any permutation $s \in \mathfrak{S}\big(\{1,\dots,2N\}\big)$ $\mathbb{P}[h(\omega \circ s)] = \mathbb{P}[h(\omega)]$, then we can still improve the bound as follows. Let

$$T'(h) = \frac{1}{N!} \sum_{s\in\mathfrak{S}\left(\{N+1,\dots,2N\}\right)} h(\omega \circ s).$$

Then we can write

$$1 - |1 - 2T_i(\sigma_i)| = (\sigma_i - \sigma_{i+N})^2 = \sigma_i + \sigma_{i+N} - 2\sigma_i\sigma_{i+N}.$$

Using this identity, we get for any exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$,

$$T\left\{\exp\left[\lambda(\overline{r} - r_1) - \log\left[\cosh(\tfrac{\lambda}{2N})\right] \sum_{i=1}^{N}\big(\sigma_i + \sigma_{i+N} - 2\sigma_i\sigma_{i+N}\big)\right]\right\} \leq 1.$$

Let us put

$$A(\lambda) = \tfrac{2N}{\lambda}\log\left[\cosh(\tfrac{\lambda}{2N})\right], \tag{3.2}$$

$$v(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(\sigma_i + \sigma_{i+N} - 2\sigma_i\sigma_{i+N}). \tag{3.3}$$

With this notation

$$T\Big\{\exp\big\{\lambda\big[\overline{r} - r_1 - A(\lambda)v\big]\big\}\Big\} \leq 1.$$

Let us notice now that

$$T'\big[v(\theta)\big] = \overline{r}(\theta) - r_1(\theta)r_2(\theta).$$

Let $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$ be any given exchangeable posterior distribution. Using the exchangeability of $\mathbb{P}$ and π and the exchangeability of the exponential function, we get

$$\mathbb{P}\Big\{\pi\Big[\exp\big\{\lambda\big[\overline{r} - r_1 - A(\overline{r} - r_1r_2)\big]\big\}\Big]\Big\} = \mathbb{P}\Big\{\pi\Big[\exp\big\{\lambda\big[\overline{r} - r_1 - AT'(v)\big]\big\}\Big]\Big\}$$

$$\leq \mathbb{P}\Big\{\pi\Big[T'\exp\big\{\lambda\big[\overline{r}-r_1-Av\big]\big\}\Big]\Big\} = \mathbb{P}\Big\{T'\pi\Big[\exp\big\{\lambda\big[\overline{r}-r_1-Av\big]\big\}\Big]\Big\}$$
$$= \mathbb{P}\Big\{\pi\Big[\exp\big\{\lambda\big[\overline{r}-r_1-Av\big]\big\}\Big]\Big\} = \mathbb{P}\Big\{T\pi\Big[\exp\big\{\lambda\big[\overline{r}-r_1-Av\big]\big\}\Big]\Big\}$$
$$= \mathbb{P}\Big\{\pi\Big[T\exp\big\{\lambda\big[\overline{r}-r_1-Av\big]\big\}\Big]\Big\} \leq 1.$$

We are thus ready to state

THEOREM 3.2.6. *In the case when $k = 1$, for any exchangeable probability distribution $\mathbb{P}$, for any exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, for any exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$,*

$$\mathbb{P}\Big\{\exp\Big[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho\Big\{\lambda\big[\overline{r}-r_1-A(\lambda)(\overline{r}-r_1r_2)\big]\Big\} - \mathcal{K}(\rho,\pi)\Big]\Big\} \leq 1,$$

where $A(\lambda)$ is defined by equation (3.2, page 146).

We then deduce as previously

COROLLARY 3.2.7. *For any exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, for any exchangeable probability measure $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, for any measurable exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$\overline{r}(\theta) \leq r_1(\theta) + A(\lambda)\big[\overline{r}(\theta) - r_1(\theta)r_2(\theta)\big] - \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda},$$

where $A(\lambda)$ is defined by equation (3.2, page 146).

In order to deduce an empirical bound from this theorem, we have to make some choice for $\lambda(\omega,\theta)$. Fortunately, it is easy to show that the bound holds uniformly in λ, because the inequality can be rewritten as a function of only one non-exchangeable quantity, namely $r_1(\theta)$. Indeed, since $r_2 = 2\overline{r} - r_1$, we see that the inequality can be written as

$$\overline{r}(\theta) \leq r_1(\theta) + A(\lambda)\big[\overline{r}(\theta) - 2\overline{r}(\theta)r_1(\theta) + r_1(\theta)^2\big] - \frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]}{\lambda}.$$

It can be solved in $r_1(\theta)$, to get

$$r_1(\theta) \geq f\Big(\lambda, \overline{r}(\theta), -\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}\Big),$$

where

$$f(\lambda,\overline{r},d) = \big[2A(\lambda)\big]^{-1}\Big\{2\overline{r}A(\lambda) - 1 + \sqrt{\big[1-2\overline{r}A(\lambda)\big]^2 + 4A(\lambda)\Big\{\overline{r}\big[1-A(\lambda)\big] - \tfrac{d}{\lambda}\Big\}}\Big\}.$$

Thus we can find some exchangeable function $\lambda(\omega,\theta)$, such that

$$f\Big(\lambda(\omega,\theta),\overline{r}(\theta),-\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}\Big)=\sup_{\beta\in\mathbb{R}_+} f\Big(\beta,\overline{r}(\theta),-\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}\Big).$$

Applying Corollary 3.2.7 (page 147) to that choice of λ, we see that

THEOREM 3.2.8. *For any exchangeable probability measure $\mathbb{P}\in\mathcal{M}_+^1(\Omega)$, for any exchangeable posterior probability distribution $\pi:\Omega\to\mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$, for any $\lambda\in\mathbb{R}_+$,*

$$\overline{r}(\theta)\leq r_1(\theta)+A(\lambda)\big[\overline{r}(\theta)-r_1(\theta)r_2(\theta)\big]-\frac{\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda},$$

where $A(\lambda)$ is defined by equation (3.2, page 146).

Solving the previous inequality in $r_2(\theta)$, we get

COROLLARY 3.2.9. *Under the same assumptions as in the previous theorem, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$,*

$$r_2(\theta)\leq\inf_{\lambda\in\mathbb{R}_+}\frac{r_1(\theta)\Big\{1+\frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\Big\}-\dfrac{2\log\big\{\epsilon\pi\big[\Delta(\theta)\big]\big\}}{\lambda}}{1-\frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\big[1-2r_1(\theta)\big]}.$$

Applying this to our usual numerical example of a binary classification model with Vapnik–Cervonenkis dimension not greater than $h=10$, when $N=1000$, $\inf_\Theta r_1=r_1(\widehat{\theta})=10$ and $\epsilon=0.01$, we obtain that $r_2(\widehat{\theta})\leq 0.4450$.

3.3. VAPNIK BOUNDS FOR INDUCTIVE CLASSIFICATION

3.3.1. ARBITRARY SHADOW SAMPLE SIZE. We assume in this section that

$$\mathbb{P}=\left(\bigotimes_{i=1}^{N}P_i\right)^{\otimes\infty}\in\mathcal{M}_+^1\Big\{\big[(\mathcal{X}\times\mathcal{Y})^N\big]^{\mathbb{N}}\Big\},$$

where $P_i\in\mathcal{M}_+^1\big(\mathcal{X}\times\mathcal{Y}\big)$: we consider an infinite i.i.d. sequence of independent *non*-identically distributed samples of size N, the first one only being observed. More precisely, under $\mathbb{P}$ each sample $(X_{i+jN},Y_{i+jN})_{i=1}^N$ is distributed according to $\bigotimes_{i=1}^N P_i$, and they are all independent from each other. Only the first sample $(X_i,Y_i)_{i=1}^N$ is assumed to be observed. The shadow samples will only appear in the proofs. The aim of this section is to prove better Vapnik bounds, generalizing them in the same time to the independent non-i.i.d. setting, which to our knowledge has not been done before.

Let us introduce the notation $\mathbb{P}'\big[h(\omega)\big]=\mathbb{P}\big[h(\omega)\,|\,(X_i,Y_i)_{i=1}^N\big]$, where h may be any suitable (e.g. bounded) random variable, let us also put $\Omega=\big[(\mathcal{X}\times\mathcal{Y})^N\big]^{\mathbb{N}}$.

DEFINITION 3.3.1. For any subset $A \subset \mathbb{N}$ of integers, let $\mathfrak{C}(A)$ be the set of circular permutations of the totally ordered set A, extended to a permutation of $\mathbb{N}$ by taking it to be the identity on the complement $\mathbb{N} \setminus A$ of A. We will say that a random function $h : \Omega \to \mathbb{R}$ is k-partially exchangeable if

$$h(\omega \circ s) = h(\omega), \quad s \in \mathfrak{C}\big(\{i + jN\,;\, j = 0, \dots, k\}\big), i = 1, \dots, N.$$

In the same way, we will say that a posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$ is k-partially exchangeable if

$$\pi(\omega \circ s) = \pi(\omega) \in \mathcal{M}_+^1(\Theta), \quad s \in \mathfrak{C}\big(\{i + jN\,;\, j = 0, \dots, k\}\big), i = 1, \dots, N.$$

Note that $\mathbb{P}$ itself is k-partially exchangeable for any k in the sense that for any bounded measurable function $h : \Omega \to \mathbb{R}$

$$\mathbb{P}\big[h(\omega \circ s)\big] = \mathbb{P}\big[h(\omega)\big], \quad s \in \mathfrak{C}\big(\{i + jN\,;\, j = 0, \dots, k\}\big), i = 1, \dots, N.$$

Let $\Delta_k(\theta) = \Big\{\theta' \in \Theta\,;\, \big[f_{\theta'}(X_i)\big]_{i=1}^{(k+1)N} = \big[f_\theta(X_i)\big]_{i=1}^{(k+1)N}\Big\}$, $\theta \in \Theta, k \in \mathbb{N}^*$, and let also $\overline{r}_k(\theta) = \dfrac{1}{(k+1)N} \displaystyle\sum_{i=1}^{(k+1)N} \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big]$. Theorem 3.1.2 shows that for any positive real parameter λ and any k-partially exchangeable posterior distribution $\pi_k : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\bigg\{\exp\Big[\sup_{\theta\in\Theta} \lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1\big] + \log\big\{\epsilon\pi_k\big[\Delta_k(\theta)\big]\big\}\Big]\bigg\} \leq \epsilon.$$

Using the general fact that

$$\mathbb{P}\big[\exp(h)\big] = \mathbb{P}\Big\{\mathbb{P}'\big[\exp(h)\big]\Big\} \geq \mathbb{P}\Big\{\exp\big[\mathbb{P}'(h)\big]\Big\},$$

and the fact that the expectation of a supremum is larger than the supremum of an expectation, we see that with $\mathbb{P}$ probability at most $1-\epsilon$, for any $\theta \in \Theta$,

$$\mathbb{P}'\Big\{\Phi_{\frac{\lambda}{N}}\big[\overline{r}_k(\theta)\big]\Big\} \leq r_1(\theta) - \frac{\mathbb{P}'\Big\{\log\big\{\epsilon\pi_k\big[\Delta_k(\theta)\big]\big\}\Big\}}{\lambda}.$$

For short let us put

$$\begin{aligned}
\bar{d}_k(\theta) &= -\log\big\{\epsilon\pi_k\big[\Delta_k(\theta)\big]\big\},\\
d'_k(\theta) &= -\mathbb{P}'\Big\{\log\big\{\epsilon\pi_k\big[\Delta_k(\theta)\big]\big\}\Big\},\\
d_k(\theta) &= -\mathbb{P}\Big\{\log\big\{\epsilon\pi_k\big[\Delta_k(\theta)\big]\big\}\Big\}.
\end{aligned}$$

We can use the convexity of $\Phi_{\frac{\lambda}{N}}$ and the fact that $\mathbb{P}'(\overline{r}_k) = \frac{r_1 + kR}{k+1}$, to establish that

$$\mathbb{P}'\Big\{\Phi_{\frac{\lambda}{N}}\big[\overline{r}_k(\theta)\big]\Big\} \geq \Phi_{\frac{\lambda}{N}}\left[\frac{r_1(\theta) + kR(\theta)}{k+1}\right].$$

We have proved

THEOREM 3.3.1. *Using the above hypotheses and notation, for any sequence $\pi_k : \Omega \to \mathcal{M}_+^1(\Theta)$, where π_k is a k-partially exchangeable posterior distribution, for any positive real constant λ, any positive integer k, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$\Phi_{\frac{\lambda}{N}}\left[\frac{r_1(\theta)+kR(\theta)}{k+1}\right] \leq r_1(\theta) + \frac{d'_k(\theta)}{\lambda}.$$

We can make as we did with Theorem 1.2.6 (page 25) the result of this theorem uniform in $\lambda \in \{\alpha^j\,;\, j \in \mathbb{N}^*\}$ and $k \in \mathbb{N}^*$ (considering on k the prior $\frac{1}{k(k+1)}$ and on j the prior $\frac{1}{j(j+1)}$), and obtain

THEOREM 3.3.2. *For any real parameter $\alpha > 1$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{k\in\mathbb{N}^*, j\in\mathbb{N}^*} \frac{1-\exp\left\{-\frac{\alpha^j}{N}r_1(\theta) - \frac{1}{N}\left\{d'_k(\theta) + \log\left[k(k+1)j(j+1)\right]\right\}\right\}}{\frac{k}{k+1}\left[1-\exp\left(-\frac{\alpha^j}{N}\right)\right]} - \frac{r_1(\theta)}{k}.$$

As a special case we can choose π_k such that $\log\{\pi_k[\Delta_k(\theta)]\}$ is independent of θ and equal to $\log(\mathfrak{N}_k)$, where

$$\mathfrak{N}_k = \left|\left\{\left[f_\theta(X_i)\right]_{i=1}^{(k+1)N}\,;\,\theta\in\Theta\right\}\right|$$

is the size of the trace of the classification model on the extended sample of size $(k+1)N$. With this choice, we obtain a bound involving a new flavour of conditional Vapnik entropy, namely

$$d'_k(\theta) = \mathbb{P}\left[\log(\mathfrak{N}_k)\,|(Z_i)_{i=1}^N\right] - \log(\epsilon).$$

In the case of binary classification using a Vapnik–Cervonenkis class of Vapnik–Cervonenkis dimension not greater than $h = 10$, when $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$ and $\epsilon = 0.01$, choosing $\alpha = 1.1$, we obtain $R(\widehat{\theta}) \leq 0.4271$ (for an optimal value of $\lambda = 1071.8$, and an optimal value of $k = 16$).

3.3.2. A BETTER MINIMIZATION WITH RESPECT TO THE EXPONENTIAL PARAMETER. If we are not pleased with optimizing λ on a discrete subset of the real line, we can use a slightly different approach. From Theorem 3.1.2 (page 140), we see that for any positive integer k, for any k-partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$ satisfying equation (3.1, page 143) — with $\Delta(\theta)$ replaced with $\Delta_k(\theta)$ — for any $\epsilon \in]0,1)$ and $\eta \in]0,1)$,

$$\mathbb{P}\left\{\mathbb{P}'\left[\exp\left[\sup_\theta \lambda\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1\right] + \log\left\{\epsilon\eta\pi_k\left[\Delta_k(\theta)\right]\right\}\right]\right\} \leq \epsilon\eta,$$

therefore with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\mathbb{P}'\left\{\exp\Big[\sup_\theta \lambda\big[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1\big] + \log\big\{\epsilon\eta\pi_k\big[\Delta_k(\theta)\big]\big\}\Big] \right\} \leq \eta,$$

and consequently, with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\theta \in \Theta$,

$$\Phi_{\frac{\lambda}{N}}(\overline{r}_k) + \frac{\log\big\{\epsilon\eta\pi_k\big[\Delta_k(\theta)\big]\big\}}{\lambda} \leq r_1.$$

Now we are entitled to choose

$$\lambda(\omega,\theta) \in \arg\max_{\lambda' \in \mathbb{R}_+} \Phi_{\frac{\lambda'}{N}}(\overline{r}_k) + \frac{\log\big\{\epsilon\eta\pi_k\big[\Delta_k(\theta)\big]\big\}}{\lambda'}.$$

This shows that with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\theta \in \Theta$,

$$\sup_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}(\overline{r}_k) - \frac{\bar{d}_k(\theta) - \log(\eta)}{\lambda} \leq r_1,$$

which can also be written

$$\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda} \leq -\frac{\log(\eta)}{\lambda}, \quad \lambda \in \mathbb{R}_+.$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, any $\lambda \in \mathbb{R}_+$,

$$\mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda}\right] \leq -\frac{\log(\eta)}{\lambda} + \left[1 - r_1 + \frac{\log(\eta)}{\lambda}\right]\eta.$$

On the other hand, $\Phi_{\frac{\lambda}{N}}$ being a convex function,

$$\begin{aligned}\mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda}\right] &\geq \Phi_{\frac{\lambda}{N}}\big[\mathbb{P}'(\overline{r}_k)\big] - r_1 - \frac{d'_k}{\lambda} \\ &= \Phi_{\frac{\lambda}{N}}\left(\frac{kR + r_1}{k+1}\right) - r_1 - \frac{d'_k}{\lambda}.\end{aligned}$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,

$$\frac{kR + r_1}{k+1} \leq \inf_{\lambda \in \mathbb{R}_+} \Phi^{-1}_{\frac{\lambda}{N}}\left[r_1(1-\eta) + \eta + \frac{d'_k - \log(\eta)(1-\eta)}{\lambda}\right].$$

We can generalize this approach by considering a finite decreasing sequence $\eta_0 = 1 > \eta_1 > \eta_2 > \cdots > \eta_J > \eta_{J+1} = 0$, and the corresponding sequence of levels

$$L_j = -\frac{\log(\eta_j)}{\lambda}, 0 \leq j \leq J,$$

$$L_{J+1} = 1 - r_1 - \frac{\log(J) - \log(\epsilon)}{\lambda}.$$

Taking a union bound in j, we see that with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, for any $\lambda \in \mathbb{R}_+$,

$$\mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda} \geq L_j\right] \leq \eta_j, \quad j = 0, \ldots, J+1,$$

and consequently

$$\begin{aligned}
&\mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda}\right] \\
&\quad \leq \int_0^{L_{J+1}} \mathbb{P}'\left[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda} \geq \alpha\right] d\alpha \quad \leq \sum_{j=1}^{J+1} \eta_{j-1}(L_j - L_{j-1}) \\
&\quad = \eta_J\left[1 - r_1 - \frac{\log(J) - \log(\epsilon) - \log(\eta_J)}{\lambda}\right] - \frac{\log(\eta_1)}{\lambda} + \sum_{j=1}^{J-1} \frac{\eta_j}{\lambda} \log\left(\frac{\eta_j}{\eta_{j+1}}\right).
\end{aligned}$$

Let us put

$$\begin{aligned}
d_k''\left[\theta, (\eta_j)_{j=1}^J\right] = d_k'(\theta) + \log(J) - \log(\eta_1) \\
+ \sum_{j=1}^{J-1} \eta_j \log\left(\frac{\eta_j}{\eta_{j+1}}\right) + \log\left(\frac{\epsilon \eta_J}{J}\right) \eta_J.
\end{aligned}$$

We have proved that for any decreasing sequence $(\eta_j)_{j=1}^J$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,

$$\frac{kR + r_1}{k+1} \leq \inf_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1}\left[r_1(1-\eta_J) + \eta_J + \frac{d_k''\left[\theta, (\eta_j)_{j=1}^J\right]}{\lambda}\right].$$

Remark 3.3.1. We can for instance choose $J = 2$, $\eta_2 = \frac{1}{10N}$, $\eta_1 = \frac{1}{\log(10N)}$, resulting in

$$d_k'' = d_k' + \log(2) + \log\log(10N) + 1 - \frac{\log\log(10N)}{\log(10N)} - \frac{\log\left(\frac{20N}{\epsilon}\right)}{10N}.$$

In the case where $N = 1000$ and for any $\epsilon \in]0,1)$, we get $d_k'' \leq d_k' + 3.7$, in the case where $N = 10^6$, we get $d_k'' \leq d_k' + 4.4$, and in the case $N = 10^9$, we get $d_k'' \leq d_k' + 4.7$.

Therefore, for any practical purpose we could take $d_k'' = d_k' + 4.7$ and $\eta_J = \frac{1}{10N}$ in the above inequality.

Taking moreover a weighted union bound in k, we get

THEOREM 3.3.3. *For any* $\epsilon \in]0,1)$*, any sequence* $1 > \eta_1 > \cdots > \eta_J > 0$*, any sequence* $\pi_k : \Omega \to \mathcal{M}_+^1(\Theta)$*, where* π_k *is a* k*-partially exchangeable posterior distribution, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any* $\theta \in \Theta$*,*

$$R(\theta) \leq \inf_{k\in\mathbb{N}^*} \frac{k+1}{k} \inf_{\lambda\in\mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1}\Bigg[r_1(\theta) + \eta_J\big[1-r_1(\theta)\big] + \frac{d_k''\big[\theta,(\eta_j)_{j=1}^J\big] + \log\big[k(k+1)\big]}{\lambda}\Bigg] - \frac{r_1(\theta)}{k}.$$

COROLLARY 3.3.4. *For any* $\epsilon \in]0,1)$*, for any* $N \leq 10^9$*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any* $\theta \in \Theta$*,*

$$R(\theta) \leq \inf_{k\in\mathbb{N}^*} \inf_{\lambda\in\mathbb{R}_+} \frac{k+1}{k}\big[1-\exp(-\tfrac{\lambda}{N})\big]^{-1}\Bigg\{1-\exp\Bigg[-\tfrac{\lambda}{N}\big[r_1(\theta)+\tfrac{1}{10N}\big] - \frac{\mathbb{P}'\big[\log(\mathfrak{N}_k)\,|\,(Z_i)_{i=1}^N\big] - \log(\epsilon) + \log\big[k(k+1)\big] + 4.7}{N}\Bigg]\Bigg\} - \frac{r_1(\theta)}{k}.$$

Let us end this section with a numerical example: in the case of binary classification with a Vapnik–Cervonenkis class of dimension not greater than 10, when $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$ and $\epsilon = 0.01$, we get a bound $R(\widehat{\theta}) \leq 0.4211$ (for optimal values of $k = 15$ and of $\lambda = 1010$).

3.3.3. EQUAL SHADOW AND TRAINING SAMPLE SIZES. In the case when $k = 1$, we can use Theorem 3.2.5 (page 145) and replace $\Phi_{\frac{\lambda}{N}}^{-1}(q)$ with $\big\{1 - \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\big\}^{-1}q$, resulting in

THEOREM 3.3.5. *For any* $\epsilon \in]0,1)$*, any* $N \leq 10^9$*, any one-partially exchangeable posterior distribution* $\pi_1 : \Omega \to \mathcal{M}_+^1(\Theta)$*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any* $\theta \in \Theta$*,*

$$R(\theta) \leq \inf_{\lambda\in\mathbb{R}_+} \frac{\Big\{1 + \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\Big\}r_1(\theta) + \frac{1}{5N} + 2\frac{d_1'(\theta)+4.7}{\lambda}}{1 - \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]}.$$

3.3.4. IMPROVEMENT ON THE EQUAL SAMPLE SIZE BOUND IN THE I.I.D. CASE. Finally, in the case when $\mathbb{P}$ is i.i.d., meaning that all the P_i are equal, we can improve the previous bound. For any partially exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, we saw in the discussion preceding Theorem 3.2.6 (page 147) that

$$T\Big[\exp\big[\lambda(\overline{r}_k - r_1) - A(\lambda)v\big]\Big] \leq 1,$$

with the notation introduced therein. Thus for any partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$ satisfying equation (3.1,

page 143), any one-partially exchangeable posterior distribution $\pi_1 : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\Big\{\exp\Big[\sup_{\theta\in\Theta}\lambda\big[\overline{r}_k(\theta) - r_1(\theta) - A(\lambda)v(\theta)\big] + \log\big[\epsilon\pi_1\big[\Delta(\theta)\big]\Big]\Big\} \leq 1.$$

Therefore with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability $1-\eta$,

$$\overline{r}_k(\theta) \leq r_1(\theta) + A(\lambda)v(\theta) + \frac{1}{\lambda}\big[\bar{d}_1(\theta) - \log(\eta)\big].$$

We can then choose $\lambda(\omega,\theta) \in \arg\min_{\lambda'\in\mathbb{R}_+} A(\lambda')v(\theta) + \dfrac{\bar{d}_1(\theta) - \log(\eta)\big]}{\lambda'}$, which satisfies the required conditions, to show that with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\lambda\in\mathbb{R}_+$,

$$\overline{r}_k(\theta) \leq r_1(\theta) + A(\lambda)v(\theta) + \frac{\bar{d}_1(\theta) - \log(\eta)}{\lambda}.$$

We can then take a union bound on a decreasing sequence of J values $\eta_1 \geq \cdots \geq \eta_J$ of η. Weakening the order of quantifiers a little, we then obtain the following statement: with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$, for any $\lambda\in\mathbb{R}_+$, for any $j = 1,\ldots,J$

$$\mathbb{P}'\left[\overline{r}_k(\theta) - r_1(\theta) - A(\lambda)v(\theta) - \frac{\bar{d}_1(\theta) + \log(J)}{\lambda} \geq -\frac{\log(\eta_j)}{\lambda}\right] \leq \eta_j.$$

Consequently for any $\lambda\in\mathbb{R}_+$,

$$\begin{aligned}\mathbb{P}'\left[\overline{r}_k(\theta) - r_1(\theta) - A(\lambda)v(\theta) - \frac{\bar{d}_1(\theta) + \log(J)}{\lambda}\right]& \\ \leq -\frac{\log(\eta_1)}{\lambda} + \eta_J\left[1 - r_1(\theta) - \frac{\log(J) - \log(\epsilon) - \log(\eta_J)}{\lambda}\right]& \\ + \sum_{j=1}^{J-1}\frac{\eta_j}{\lambda}\log\left(\frac{\eta_j}{\eta_{j+1}}\right).&\end{aligned}$$

Moreover $\mathbb{P}'\big[v(\theta)\big] = \frac{r_1+R}{2} - r_1 R$, (this is where we need equidistribution) thus proving that

$$\frac{R - r_1}{2} \leq \frac{A(\lambda)}{2}\Big[R + r_1 - 2r_1R\Big] + \frac{d_1''\big[\theta,(\eta_j)_{j=1}^J\big]}{\lambda} + \eta_J\big[1 - r_1(\theta)\big].$$

Keeping track of quantifiers, we obtain

THEOREM 3.3.6. *For any decreasing sequence $(\eta_j)_{j=1}^J$, any $\epsilon\in)0,1)$, any one-partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$,*

$$R(\theta) \leq \inf_{\lambda\in\mathbb{R}_+} \frac{\Big\{1 + \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\Big\}r_1(\theta) + \dfrac{2d_1''\big[\theta,(\eta_j)_{j=1}^J\big]}{\lambda} + 2\eta_J\big[1 - r_1(\theta)\big]}{1 - \frac{2N}{\lambda}\log\big[\cosh(\frac{\lambda}{2N})\big]\big[1 - 2r_1(\theta)\big]}.$$

3.4. Gaussian approximation in Vapnik bounds

3.4.1. Gaussian upper bounds of variance terms. To obtain formulas which could be easily compared with original Vapnik bounds, we may replace $p - \Phi_a(p)$ with a Gaussian upper bound:

Lemma 3.4.1. *For any $p \in (0, \frac{1}{2})$, any $a \in \mathbb{R}_+$,*

$$p - \Phi_a(p) \leq \frac{a}{2} p(1-p).$$

For any $p \in (\frac{1}{2}, 1)$,

$$p - \Phi_a(p) \leq \frac{a}{8}.$$

Proof. Let us notice that for any $p \in (0,1)$,

$$\begin{aligned}
\frac{\partial}{\partial a}\big[-a\Phi_a(p)\big] &= -\frac{p\exp(-a)}{1-p+p\exp(-a)}, \\
\frac{\partial^2}{\partial^2 a}\big[-a\Phi_a(p)\big] &= \frac{p\exp(-a)}{1-p+p\exp(-a)}\left(1 - \frac{p\exp(-a)}{1-p+p\exp(-a)}\right) \\
&\leq \begin{cases} p(1-p) & p \in (0, \frac{1}{2}), \\ \frac{1}{4} & p \in (\frac{1}{2}, 1). \end{cases}
\end{aligned}$$

Thus taking a Taylor expansion of order one with integral remainder:

$$-a\Phi(a) \leq \begin{cases} -ap + \displaystyle\int_0^a p(1-p)(a-b)db & \\ \qquad\qquad = -ap + \dfrac{a^2}{2} p(1-p), & p \in (0, \frac{1}{2}), \\ -ap + \displaystyle\int_0^a \frac{1}{4}(a-b)db = -ap + \dfrac{a^2}{8}, & p \in (\frac{1}{2}, 1). \end{cases}$$

This ends the proof of our lemma. □

Lemma 3.4.2. *Let us consider the bound*

$$B(q,d) = \left(1 + \frac{2d}{N}\right)^{-1}\left[q + \frac{d}{N} + \sqrt{\frac{2dq(1-q)}{N} + \frac{d^2}{N^2}}\right], \quad q \in \mathbb{R}_+, d \in \mathbb{R}_+.$$

Let us also put

$$\bar{B}(q,d) = \begin{cases} B(q,d) & B(q,d) \leq \frac{1}{2}, \\ q + \sqrt{\frac{d}{2N}} & \text{otherwise.} \end{cases}$$

For any positive real parameters q and d

$$\inf_{\lambda \in \mathbb{R}_+} \Phi^{-1}_{\frac{\lambda}{N}}\left(q + \frac{d}{\lambda}\right) \leq \bar{B}(q,d).$$

PROOF. Let $p = \inf_{\lambda} \Phi^{-1}_{\frac{\lambda}{N}}\left(q + \frac{d}{\lambda}\right)$. For any $\lambda \in \mathbb{R}_+$,

$$p - \frac{\lambda}{2N}(p \wedge \tfrac{1}{2})\big[1 - (p \wedge \tfrac{1}{2})\big] \leq \Phi_{\frac{\lambda}{N}}(p) \leq q + \frac{d}{\lambda}.$$

Thus

$$p \leq q + \inf_{\lambda \in \mathbb{R}_+} \frac{\lambda}{2N}(p \wedge \tfrac{1}{2})\big[1 - (p \wedge \tfrac{1}{2})\big] + \frac{d}{\lambda}$$
$$= q + \sqrt{\frac{2d(p \wedge \frac{1}{2})\big[1 - (p \wedge \frac{1}{2})\big]}{N}} \leq q + \sqrt{\frac{d}{2N}}.$$

Then let us remark that $B(q,d) = \sup\left\{p' \in \mathbb{R}_+ \,;\, p' \leq q + \sqrt{\frac{2dp'(1-p')}{N}}\right\}$.

If moreover $\frac{1}{2} \geq B(q,d)$, then according to this remark $\frac{1}{2} \geq q + \sqrt{\frac{d}{2N}} \geq p$. Therefore $p \leq \frac{1}{2}$, and consequently $p \leq q + \sqrt{\frac{2dp(1-p)}{N}}$, implying that $p \leq B(q,d)$. □

3.4.2. ARBITRARY SHADOW SAMPLE SIZE. The previous lemma combined with Corollary 3.3.4 (page 153) implies

COROLLARY 3.4.3. *Let us use the notation introduced in Lemma 3.4.2 (page 155). For any $\epsilon \in]0,1)$, any integer $N \leq 10^9$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{k \in \mathbb{N}^*} \frac{k+1}{k}\Big\{\bar{B}\Big[r_1(\theta) + \frac{1}{10N}, d'_k(\theta) + \log\big[k(k+1)\big] + 4.7\Big]\Big\} - \frac{r_1(\theta)}{k}.$$

3.4.3. EQUAL SAMPLE SIZES IN THE I.I.D. CASE. To make a link with Vapnik's result, it is useful to state the Gaussian approximation to Theorem 3.3.6 (page 154). Indeed, using the upper bound $A(\lambda) \leq \frac{\lambda}{4N}$, where $A(\lambda)$ is defined by equation (3.2) on page 146, we get with $\mathbb{P}$ probability at least $1-\epsilon$

$$R - r_1 - 2\eta_J \leq \inf_{\lambda \in \mathbb{R}_+} \frac{\lambda}{4N}\big[R + r_1 - 2r_1 R\big] + \frac{2d''_1}{\lambda} = \sqrt{\frac{2d''_1(R + r_1 - 2r_1 R)}{N}},$$

which can be solved in R to obtain

COROLLARY 3.4.4. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq r_1(\theta) + \frac{d''_1(\theta)}{N}\big[1 - 2r_1(\theta)\big] + 2\eta_J$$
$$+\sqrt{\frac{4d''_1(\theta)\big[1 - r_1(\theta)\big]r_1(\theta)}{N} + \frac{d''_1(\theta)^2}{N^2}\big[1 - 2r_1(\theta)\big]^2 + \frac{4d''_1(\theta)}{N}\big[1 - 2r_1(\theta)\big]\eta_J}.$$

This is to be compared with Vapnik's result, as proved in Vapnik (1998, page 138):

THEOREM 3.4.5 (VAPNIK). *For any i.i.d. probability distribution* $\mathbb{P}$*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any* $\theta \in \Theta$*, putting*

$$d_V = \log\big[\mathbb{P}(\mathfrak{N}_1)\big] + \log(4/\epsilon),$$

$$R(\theta) \leq r_1(\theta) + \frac{2d_V}{N} + \sqrt{\frac{4d_V r_1(\theta)}{N} + \frac{4d_V^2}{N^2}}.$$

Recalling that we can choose $(\eta_j)_{j=1}^2$ such that $\eta_J = \eta_2 = \frac{1}{10N}$ (which brings a negligible contribution to the bound) and such that for any $N \leq 10^9$,

$$d_1''(\theta) \leq \mathbb{P}\big[\log(\mathfrak{N}_1) \,|\, (Z_i)_{i=1}^N\big] - \log(\epsilon) + 4.7,$$

we see that our complexity term is somehow more satisfactory than Vapnik's, since it is integrated outside the logarithm, with a slightly larger additional constant (remember that $\log 4 \simeq 1.4$, which is better than our 4.7, which could presumably be improved by working out a better sequence η_j, but not down to $\log(4)$). Our variance term is better, since we get $r_1(1-r_1)$, instead of r_1. We also have $\frac{d_1''}{N}$ instead of $2\frac{d_V}{N}$, because we use no symmetrization trick.

Let us illustrate these bounds on a numerical example, corresponding to a situation where the sample is noisy or the classification model is weak. Let us assume that $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$, that we are performing binary classification with a model with Vapnik–Cervonenkis dimension not greater than $h = 10$, and that we work at confidence level $\epsilon = 0.01$. Vapnik's theorem provides an upper bound for $R(\widehat{\theta})$ not smaller than 0.610, whereas Corollary 3.4.4 gives $R(\widehat{\theta}) \leq 0.461$ (using the bound $d_1'' \leq d_1' + 3.7$ when $N = 1000$). Now if we go for Theorem 3.3.6 and do not make a Gaussian approximation, we get $R(\widehat{\theta}) \leq 0.453$. It is interesting to remark that this bound is achieved for $\lambda = 1195 > N = 1000$. This explains why the Gaussian approximation in Vapnik's bound can be improved: for such a large value of λ, $\lambda r_1(\theta)$ does not behave like a Gaussian random variable.

Let us recall in conclusion that the best bound is provided by Theorem 3.3.3 (page 153), giving $R(\widehat{\theta}) \leq 0.4211$, (that is approximately 2/3 of Vapnik's bound), for optimal values of $k = 15$, and of $\lambda = 1010$. This bound can be seen to take advantage of the fact that Bernoulli random variables are not Gaussian (its Gaussian approximation, Corollary 3.4.3, gives a bound $R(\theta) \simeq 0.4325$, still with an optimal $k = 15$), and of the fact that the optimal size of the shadow sample is significantly larger than the size of the observed sample. Moreover, Theorem 3.3.3 does not assume that the sample is i.i.d., but only that it is independent, thus generalizing Vapnik's bounds to inhomogeneous data (this will presumably be the case when data are collected from different places where the experimental conditions may not be the same, although they may reasonably be assumed to be independent).

Our little numerical example was chosen to illustrate the case when it is non-trivial to decide whether the chosen classifier does better than the 0.5 error rate of blind random classification. This case is of interest to choose "weak learners" to be aggregated or combined in some appropriate way in a second stage to reach a better classification rate. This stage of feature selection is unavoidable in many real world classification tasks. Our little computations are meant to exemplify the fact that Vapnik's bounds, although asymptotically suboptimal, as is obvious by comparison with the first two chapters, can do the job when dealing with moderate sample sizes.

Chapter 4

Support Vector Machines

4.1. How to build them

4.1.1. The canonical hyperplane. Support Vector Machines, of wide use and renown, were conceived by V. Vapkik (Vapnik, 1998). Before introducing them, we will study as a prerequisite the separation of points by hyperplanes in a finite dimensional Euclidean space. Support Vector Machines perform the same kind of linear separation after an implicit change of pattern space. The preceding PAC-Bayesian results provide a fit framework to analyse their generalization properties.

In this section we deal with the classification of points in $\mathbb{R}^d$ 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 us split the set of indices $I = \{1, \ldots, N\}$ according to the labels into two subsets

$$I_+ = \{i \in I \, : y_i = +1\},$$
$$I_- = \{i \in I \, : y_i = -1\}.$$

Let us then consider the set of admissible separating directions

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

which 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 4.1.1. *When $A_Z \neq \varnothing$, $\inf\{\|w\|^2 \, : w \in A_Z\}$ is reached for only one value w_Z of w.*

PROOF. Let $w_0 \in A_Z$. The set $A_Z \cap \{w \in \mathbb{R}^d : \|w\| \leq \|w_0\|\}$ is a compact convex set and $w \mapsto \|w\|^2$ is strictly convex and therefore has a unique minimum on this set, which is also obviously its minimum on A_Z. $\square$

DEFINITION 4.1.1. 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.

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.

4.1.2. COMPUTATION OF THE CANONICAL HYPERPLANE. Let us consider the convex hulls X_+ and X_- of the positive and negative patterns:

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

Let us introduce the closed convex set

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

As $v \mapsto \|v\|^2$ is strictly convex, with compact lower level sets, there is a unique vector v^* such that

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

LEMMA 4.1.2. *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 $\frac{1}{2}\|v^*\|$.

This lemma proves that the distance between the convex hulls of the positive and negative patterns is equal to twice the margin of the canonical hyperplane.

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

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

so $\min_{i \in I_+} \langle w, x_i \rangle - \max_{i \in I_-} \langle w, x_i \rangle \leq 0$, 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 $w^* = 2v^*/\|v^*\|^2$. 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 $\|w^*\| = \inf\{\|w\| : w \in A_Z\}$, and therefore that $w^* = w_Z$ as claimed. □

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

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

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

PROPOSITION 4.1.3. *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

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

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 $F(\alpha^*) = \inf\{F(\alpha) : \alpha \in \mathcal{A}\}$, $w(\alpha^*) = \frac{2}{\|v^*\|^2} v^* = w_Z$. □

4.1.3. SUPPORT VECTORS.

DEFINITION 4.1.2. 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 4.1.4. *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 neighbourhood 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

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

□

4.1.4. The non-separable case. In the case when the training set $Z = (x_i, y_i)_{i=1}^N$ is not linearly separable, we can define a noisy canonical hyperplane as follows: we can choose $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ to minimize

$$C(w,b) = \sum_{i=1}^{N} \big[1 - \big(\langle w, x_i\rangle - b\big)y_i\big]_+ + \tfrac{1}{2}\|w\|^2, \tag{4.1}$$

where for any real number r, $r_+ = \max\{r, 0\}$ is the positive part of r.

Theorem 4.1.5. *Let us introduce the dual criterion*

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

and the domain $\mathcal{A}' = \left\{\alpha \in \mathbb{R}_+^N : \alpha_i \leq 1, i = 1, \ldots, N, \sum_{i=1}^{N} y_i\alpha_i = 0\right\}$. *Let* $\alpha^* \in \mathcal{A}'$ *be such that* $F(\alpha^*) = \sup_{\alpha\in\mathcal{A}'} F(\alpha)$. *Let* $w^* = \sum_{i=1}^N y_i\alpha_i^* x_i$. *There is a threshold* b^* *(whose construction will be detailed in the proof), such that*

$$C(w^*, b^*) = \inf_{w\in\mathbb{R}^d, b\in\mathbb{R}} C(w,b).$$

Corollary 4.1.6. (scaled criterion) *For any positive real parameter* λ *let us consider the criterion*

$$C_\lambda(w,b) = \lambda^2 \sum_{i=1}^{N} \big[1 - (\langle w, x_i\rangle - b)y_i\big]_+ + \tfrac{1}{2}\|w\|^2$$

and the domain $\mathcal{A}'_\lambda = \Big\{\alpha \in \mathbb{R}_+^N : \alpha_i \le \lambda^2, i = 1, \dots, N, \sum_{i=1}^N y_i\alpha_i = 0\Big\}$. *For any solution* α^* *of the minimization problem* $F(\alpha^*) = \sup_{\alpha\in\mathcal{A}'_\lambda} F(\alpha)$, *the vector* $w^* = \sum_{i=1}^N y_i\alpha_i^* x_i$ *is such that*

$$\inf_{b\in\mathbb{R}} C_\lambda(w^*, b) = \inf_{w\in\mathbb{R}^d, b\in\mathbb{R}} C_\lambda(w, b).$$

In the separable case, the scaled criterion is minimized by the canonical hyperplane for λ large enough. This extension of the canonical hyperplane computation in dual space is often called *the box constraint*, for obvious reasons.

Proof. The corollary is a straightforward consequence of the scale property $C_\lambda(w, b, x) = \lambda^2 C(\lambda^{-1}w, b, \lambda x)$, where we have made the dependence of the criterion in $x \in \mathbb{R}^{dN}$ explicit. Let us come now to the proof of the theorem.

The minimization of $C(w, b)$ can be performed in dual space extending the couple of parameters (w, b) to $\overline{w} = (w, b, \gamma) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+^N$ and introducing the dual multipliers $\alpha \in \mathbb{R}_+^N$ and the criterion

$$G(\alpha, \overline{w}) = \sum_{i=1}^N \gamma_i + \sum_{i=1}^N \alpha_i\big\{\big[1 - (\langle w, x_i\rangle - b)y_i\big] - \gamma_i\big\} + \tfrac{1}{2}\|w\|^2.$$

We see that

$$C(w, b) = \inf_{\gamma\in\mathbb{R}_+^N} \sup_{\alpha\in\mathbb{R}_+^N} G\big[\alpha, (w, b, \gamma)\big],$$

and therefore, putting $\overline{\mathcal{W}} = \big\{(w, b, \gamma) : w \in \mathbb{R}^d, b \in \mathbb{R}, \gamma \in \mathbb{R}_+^N\big\}$, we are led to solve the minimization problem

$$G(\alpha_*, \overline{w}_*) = \inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha, \overline{w}),$$

whose solution $\overline{w}_* = (w_*, b_*, \gamma_*)$ is such that $C(\overline{w}_*, b_*) = \inf_{(w,b)\in\mathbb{R}^{d+1}} C(w, b)$, according to the preceding identity. As for any value of $\alpha' \in \mathbb{R}_+^N$,

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha, \overline{w}) \ge \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha', \overline{w}),$$

it is immediately seen that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha, \overline{w}) \ge \sup_{\alpha\in\mathbb{R}_+^N} \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha, \overline{w}).$$

We are going to show that there is no duality gap, meaning that this inequality is indeed an equality. More importantly, we will do so by exhibiting a saddle point, which, solving the dual minimization problem will also solve the original one.

Let us first make explicit the solution of the dual problem (the interest of this dual problem precisely lies in the fact that it can more easily be solved explicitly). Introducing the admissible set of values of α,

$$\mathcal{A}' = \big\{\alpha \in \mathbb{R}^N : 0 \le \alpha_i \le 1, i = 1, \dots, N, \sum_{i=1}^{N} y_i \alpha_i = 0\big\},$$

it is elementary to check that

$$\inf_{\overline{w} \in \overline{\mathcal{W}}} G(\alpha, \overline{w}) = \begin{cases} \inf\limits_{w \in \mathbb{R}^d} G\big[\alpha, (w,0,0)\big], & \alpha \in \mathcal{A}', \\ -\infty, & \text{otherwise.} \end{cases}$$

As

$$G\big[\alpha, (w,0,0)\big] = \tfrac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \alpha_i \big(1 - \langle w, x_i \rangle y_i\big),$$

we see that $\inf_{w \in \mathbb{R}^d} G\big[\alpha, (w,0,0)\big]$ is reached at

$$w_\alpha = \sum_{i=1}^{N} y_i \alpha_i x_i.$$

This proves that

$$\inf_{\overline{w} \in \overline{\mathcal{W}}} G(\alpha, \overline{w}) = F(\alpha).$$

The continuous map $\alpha \mapsto \inf_{\overline{w} \in \overline{\mathcal{W}}} G(\alpha, \overline{w})$ reaches a maximum α^*, not necessarily unique, on the compact convex set $\mathcal{A}'$. We are now going to exhibit a choice of $\overline{w}^* \in \overline{\mathcal{W}}$ such that $(\alpha^*, \overline{w}^*)$ is a *saddle point*. This means that we are going to show that

$$G(\alpha^*, \overline{w}^*) = \inf_{\overline{w} \in \overline{\mathcal{W}}} G(\alpha^*, \overline{w}) = \sup_{\alpha \in \mathbb{R}^N_+} G(\alpha, \overline{w}^*).$$

It will imply that

$$\inf_{\overline{w} \in \overline{\mathcal{W}}} \sup_{\alpha \in \mathbb{R}^d_+} G(\alpha, \overline{w}) \le \sup_{\alpha \in \mathbb{R}^N_+} G(\alpha, \overline{w}^*) = G(\alpha^*, \overline{w}^*)$$

on the one hand and that

$$\inf_{\overline{w} \in \overline{\mathcal{W}}} \sup_{\alpha \in \mathbb{R}^d_+} G(\alpha, \overline{w}) \ge \inf_{\overline{w} \in \overline{\mathcal{W}}} G(\alpha^*, \overline{w}) = G(\alpha^*, \overline{w}^*)$$

on the other hand, proving that

$$G(\alpha^*, \overline{w}^*) = \inf_{\overline{w} \in \overline{\mathcal{W}}} \sup_{\alpha \in \mathbb{R}^N_+} G(\alpha, \overline{w})$$

as required.

Construction of $\overline{w}^*$.

- Let us put $w^* = w_{\alpha^*}$.
- If there is $j \in \{1, \dots, N\}$ such that $0 < \alpha_j^* < 1$, let us put

$$b^* = \langle x_j, w^* \rangle - y_j.$$

  Otherwise, let us put

$$b^* = \sup\{\langle x_i, w^* \rangle - 1 : \alpha_i^* > 0, y_i = +1, i = 1, \dots, N\}.$$

- Let us then put

$$\gamma_i^* = \begin{cases} 0, & \alpha_i^* < 1, \\ 1 - (\langle w^*, x_i \rangle - b^*) y_i, & \alpha_i^* = 1. \end{cases}$$

If we can prove that

$$1 - (\langle w^*, x_i \rangle - b^*) y_i \begin{cases} \leq 0, & \alpha_i^* = 0, \\ = 0, & 0 < \alpha_i^* < 1, \\ \geq 0, & \alpha_i^* = 1, \end{cases} \tag{4.2}$$

it will show that $\gamma^* \in \mathbb{R}_+^N$ and therefore that $\overline{w}^* = (w^*, b^*, \gamma^*) \in \overline{\mathcal{W}}$. It will also show that

$$G(\alpha, \overline{w}^*) = \sum_{i=1}^{N} \gamma_i^* + \sum_{i, \alpha_i^* = 0} \alpha_i \big[1 - (\langle \overline{w}^*, x_i \rangle - b^*) y_i\big] + \tfrac{1}{2} \|\overline{w}^*\|^2,$$

proving that $G(\alpha^*, \overline{w}^*) = \sup_{\alpha \in \mathbb{R}_+^N} G(\alpha, \overline{w}^*)$. As obviously $G(\alpha^*, \overline{w}^*) = G\big[\alpha^*, (w^*, 0, 0)\big]$, we already know that $G(\alpha^*, \overline{w}^*) = \inf_{\overline{w} \in \overline{\mathcal{W}}} G(\alpha^*, \overline{w})$. This will show that $(\alpha^*, \overline{w}^*)$ is the saddle point we were looking for, thus ending the proof of the theorem.

Proof of equation (4.2). Let us deal first with the case when there is $j \in \{1, \dots, N\}$ such that $0 < \alpha_j^* < 1$.

For any $i \in \{1, \dots, N\}$ such that $0 < \alpha_i^* < 1$, there is $\epsilon > 0$ such that for any $t \in (-\epsilon, \epsilon)$, $\alpha^* + t y_i e_i - t y_j e_j \in \mathcal{A}'$, where $(e_k)_{k=1}^N$ is the canonical base of $\mathbb{R}^N$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + t y_i e_i - t y_j e_j) = 0$. Computing this derivative, we obtain

$$\begin{aligned} \frac{\partial}{\partial t}_{|t=0} F(\alpha^* + t y_i e_i - t y_j e_j) &= y_i - \langle w^*, x_i \rangle + \langle w^*, x_j \rangle - y_j \\ &= y_i \big[1 - \big(\langle w, x_i \rangle - b^*\big) y_i\big]. \end{aligned}$$

Thus $1 - \big(\langle w, x_i \rangle - b^*\big) y_i = 0$, as required. This shows also that the definition of b^* does not depend on the choice of j such that $0 < \alpha_j^* < 1$.

For any $i \in \{1, \dots, N\}$ such that $\alpha_i^* = 0$, there is $\epsilon > 0$ such that for any $t \in (0, \epsilon)$, $\alpha^* + t e_i - t y_i y_j e_j \in \mathcal{A}'$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + t e_i - t y_i y_j e_j) \leq 0$, showing that $1 - \big(\langle w^*, x_i \rangle - b^*\big) y_i \leq 0$ as required.

For any $i \in \{1, \dots, N\}$ such that $\alpha_i^* = 1$, there is $\epsilon > 0$ such that $\alpha^* - te_i + ty_iy_je_j \in \mathcal{A}'$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* - te_i + ty_iy_je_j) \leq 0$, showing that $1 - \big(\langle w^*, x_i\rangle - b^*\big)y_i \geq 0$ as required. This shows that $(\alpha^*, \overline{w}^*)$ is a saddle point in this case.

Let us deal now with the case where $\alpha^* \in \{0,1\}^N$. If we are not in the trivial case where the vector $(y_i)_{i=1}^N$ is constant, the case $\alpha^* = 0$ is ruled out. Indeed, in this case, considering $\alpha^* + te_i + te_j$, where $y_iy_j = -1$, we would get the contradiction $2 = \frac{\partial}{\partial t}_{|t=0} F(\alpha^* + te_i + te_j) \leq 0$.

Thus there are values of j such that $\alpha_j^* = 1$, and since $\sum_{i=1}^N \alpha_i y_i = 0$, both classes are present in the set $\{j : \alpha_j^* = 1\}$.

Now for any $i, j \in \{1, \dots, N\}$ such that $\alpha_i^* = \alpha_j^* = 1$ and such that $y_i = +1$ and $y_j = -1$, $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* - te_i - te_j) = -2 + \langle w^*, x_i\rangle - \langle w^*, x_j\rangle \leq 0$. Thus

$$\sup\{\langle w^*, x_i\rangle - 1 : \alpha_i^* = 1, y_i = +1\} \leq \inf\{\langle w^*, x_j\rangle + 1 : \alpha_j^* = 1, y_j = -1\},$$

showing that

$$1 - \big(\langle w^*, x_k\rangle - b^*\big)y_k \geq 0, \alpha_k^* = 1.$$

Finally, for any i such that $\alpha_i^* = 0$, for any j such that $\alpha_j^* = 1$ and $y_j = y_i$, we have

$$\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + te_i - te_j) = y_i\langle w^*, x_i - x_j\rangle \leq 0,$$

showing that $1 - \big(\langle w^*, x_i\rangle - b^*\big)y_i \leq 0$. This shows that $(\alpha^*, \overline{w}^*)$ is always a saddle point.

4.1.5. Support Vector Machines.

Definition 4.1.3. 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 previously

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

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 4.1.4. Let K be a positive symmetric kernel. The training set Z is said to be K-separable if

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

LEMMA 4.1.7. *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 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 4.1.8. *Let K be a symmetric positive kernel and let $Z = (x_i, y_i)_{i=1}^N$ be some K-separable training set. Let $\alpha^* \in \mathcal{A}$ be such that $F(\alpha^*) = \inf\{F(\alpha) : \alpha \in \mathcal{A}\}$. Let*

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

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

An important consequence of this proposition is that the support vector machine defined by K and Z is also the support vector machine defined by K and $Z' = \{(x_i, y_i) : \alpha_i^* > 0, 1 \le i \le N\}$, since this restriction of the index set contains the value α^* where the minimum of F is reached.

PROOF. The independence of 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 notation $x_{N+1} = x$. Let M be the $(N+1)\times(N+1)$ symmetric semi-definite matrix defined by $M(i,j) = K(x_i, x_j)$, $i = 1, \dots, N+1$, $j = 1, \dots, N+1$. Let us consider the mapping $\Psi : \{x_i : i = 1, \dots, N+1\} \to \mathbb{R}^{N+1}$ defined by

$$\Psi(x_i) = \big[M^{1/2}(i,j)\big]_{j=1}^{N+1} \in \mathbb{R}^{N+1}. \tag{4.3}$$

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 proved. □

We can in the same way use the box constraint and show that any solution $\alpha^* \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i \le \lambda^2, i = 1, \dots, N\}$ minimizes

$$\inf_{b\in\mathbb{R}} \lambda^2 \sum_{i=1}^N \Big[1 - \Big(\sum_{j=1}^N y_j \alpha_j K(x_j, x_i) - b\Big) y_i\Big]_+ + \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j). \quad (4.4)$$

4.1.6. Building kernels. Except the last, the results of this section are drawn from Cristianini et al. (2000). 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 4.1.9. *Let K_1 and K_2 be positive symmetric kernels on $\mathcal{X}$. Then for any $a \in \mathbb{R}_+$*

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

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

$$\langle \alpha, (aK_1 + K_2)\alpha\rangle = a\langle \alpha, K_1\alpha\rangle + \langle \alpha, K_2\alpha\rangle \ge 0,$$
$$\langle \alpha, (K_1 \cdot K_2)\alpha\rangle = \sum_{i,j} \alpha_i K_1(i, j)K_2(i, j)\alpha_j$$

$$= \sum_{i,j,k} \alpha_i K_1^{1/2}(i,k) K_1^{1/2}(k,j) K_2(i,j) \alpha_j$$
$$= \sum_k \underbrace{\sum_{i,j} \big[K_1^{1/2}(k,i)\alpha_i\big] K_2(i,j) \big[K_1^{1/2}(k,j)\alpha_j\big]}_{\geq 0} \geq 0,$$

$$\langle \alpha, K_g \alpha \rangle = \sum_{i,j} \alpha_i g(i) g(j) \alpha_j = \left(\sum_i \alpha_i g(i) \right)^2 \geq 0.$$

□

Proposition 4.1.10. *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 comes from the fact that the exponential function is the pointwise limit of a sequence of polynomial functions with positive coefficients. The third is seen from the second 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 4.1.11. *With the notation 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 $g(x_i)$, $i = 1, \ldots, N$ 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 (4.3). It is enough to prove that $\Psi(x_1), \ldots, \Psi(x_N)$ are affine independent, since the simplex, and therefore any affine independent set of points, can be split in any arbitrary way by affine half-spaces. Let us assume that $(x_1, \ldots, x_N)$ are affine dependent; then for some $(\lambda_1, \ldots, \lambda_N) \neq 0$ 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. □

4.2. Bounds for Support Vector Machines

4.2.1. Compression scheme bounds. We can use Support Vector Machines in the framework of compression schemes and apply Theorem 3.3.3 (page 153). 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 described on page 144, 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 Theorem 3.3.3 in the case when the observed sample $(X_i, Y_i)_{i=1}^N$ is K-separable, a natural if perhaps sub-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 4.1.8 (page 168). If Z is not K-separable, we can train a Support Vector Machine with the box constraint, then remove all the errors to obtain a K-separable sub-sample $Z' = \{(X_i, Y_i) : \alpha_i^* < \lambda^2, 1 \le i \le N\}$, using the same notation as in equation (4.4) on page 169, and then consider its support vectors as the compression set. Still using the notation of page 169, this means

we have to compute successively $\alpha^* \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i \leq \lambda^2\}$, and $\alpha^{**} \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i = 0 \text{ when } \alpha_i^* = \lambda^2\}$, to keep the compression set indexed by $J = \{i : 1 \leq i \leq N, \alpha_i^{**} > 0\}$, and the corresponding Support Vector Machine $\widehat{f}_J$. Different values of λ can be used at this stage, producing different candidate compression sets: when λ increases, the number of errors should decrease, on the other hand when λ decreases, the margin $\|w\|^{-1}$ of the separable subset Z' increases, supporting the hope for a smaller set of support vectors, thus we can use λ to monitor the number of errors on the training set we accept from the compression scheme. As we can use whatever heuristic we want while selecting the compression set, we can also try to threshold in the previous construction α_i^{**} at different levels $\eta \geq 0$, to produce candidate compression sets $J_\eta = \{i : 1 \leq i \leq N, \alpha_i^{**} > \eta\}$ of various sizes.

As the size $|J|$ of the compression set is random in this construction, we must use a version of Theorem 3.3.3 (page 153) which handles compression sets of arbitrary sizes. This is done by choosing for each k a k-partially exchangeable posterior distribution π_k which weights the compression sets of all dimensions. We immediately see that we can choose π_k such that $-\log\bigl[\pi_k(\Delta_k(J))\bigr] \leq \log\bigl[|J|(|J|+1)\bigr] + |J| \log\Bigl[\frac{(k+1)eN}{|J|}\Bigr]$.

If we observe the shadow sample patterns, and if computer resources permit, we can of course use more elaborate bounds than Theorem 3.3.3, such as the transductive equivalent for Theorem 1.3.15 (page 46) (where we may consider the submodels made of all the compression sets of the same size). Theorems based on relative bounds, such as Theorem 2.2.4 (page 93) or Theorem 2.3.9 (page 131) can also be used. Gibbs distributions can be approximated by Monte Carlo techniques, where a Markov chain with the proper invariant measure consists in appropriate local perturbations of the compression set.

Let us mention also that the use of compression schemes based on Support Vector Machines can be tailored to perform some kind of *feature aggregation*. 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 $\bigl\{f_\theta : \mathcal{X} \to \mathbb{R}\,;\theta \in \Theta\bigr\}$ the kernel

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

In this setting, the Support Vector Machine applied to the training set $Z = (x_i, y_i)_{i=1}^N$ has the form

$$f_Z(x) = \operatorname{sign}\left(\int_{\theta\in\Theta} f_\theta(x) \sum_{i=1}^N y_i \alpha_i f_\theta(x_i) \pi(d\theta) - b \right)$$

and, if this is too burdensome to compute, we can replace it with some finite

approximation

$$\widetilde{f}_Z(x) = \operatorname{sign}\left(\frac{1}{m}\sum_{k=1}^{m} f_{\theta_k}(x) w_k - b\right),$$

where the set $\{\theta_k,\, k = 1, \dots, m\}$ and the weights $\{w_k,\, k = 1, \dots, m\}$ are computed in some suitable way from the set $Z' = (x_i, y_i)_{i,\alpha_i>0}$ of support vectors of f_Z. For instance, we can draw $\{\theta_k,\, k = 1, \dots, m\}$ at random according to the probability distribution proportional to

$$\left|\sum_{i=1}^{N} y_i \alpha_i f_\theta(x_i)\right| \pi(d\theta),$$

define the weights w_k by

$$w_k = \operatorname{sign}\left(\sum_{i=1}^{N} y_i \alpha_i f_{\theta_k}(x_i)\right) \int_{\theta\in\Theta} \left|\sum_{i=1}^{N} y_i \alpha_i f_\theta(x_i)\right| \pi(d\theta),$$

and choose the smallest value of m for which this approximation still classifies Z' without errors. Let us remark that we have built $\widetilde{f}_Z$ in such a way that

$$\lim_{m\to+\infty} \widetilde{f}_Z(x_i) = f_Z(x_i) = y_i, \quad \text{a.s.}$$

for any support index i such that $\alpha_i > 0$.

Alternatively, given Z', we can select a finite set of features $\Theta' \subset \Theta$ such that Z' 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 $f_{Z'}$ built with the kernel $K_{\Theta'}$. As soon as Θ' is chosen as a function of Z' only, Theorem 3.3.3 (page 153) applies and provides some level of confidence for the risk of $f_{Z'}$.

4.2.2. 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 Vapnik–Cervonenkis dimension of S, defined as

$$h(S) = \max\Big\{|A| : A \subset X, |A| < \infty \text{ and } A \cap S = \{0,1\}^A\Big\},$$

where by definition $A \cap S = \{A \cap B : B \in S\}$ and $|A|$ is the number of points in A. 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, $h(B \cap S) \le h(S)$, the reason being that $A \cap (B \cap S) = B \cap (A \cap S)$.

This notion of Vapnik–Cervonenkis dimension is useful because, as we will see for Support Vector Machines, it can 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:

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

PROPOSITION 4.2.1. *With the previous notation, $h(S) = d + 1$.*

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 $\{e_i \,; i = 1, \dots, d+1\} \cap (A_{w,b} \cap X) = \{e_i \,; \epsilon_i = +1\}$. 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 $B \notin (A \cap S)$. Obviously this will 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 $B = \{x \in A : \lambda_x > 0\}$ and its complement $A \setminus B$ are non-empty, 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

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

which 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 4.2.2. *Whenever $\bigcup S$ is finite,*

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

THEOREM 4.2.3. *For any $h \leq n$,*

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

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 4.2.2. 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 $A \cap S = \{0,1\}^A$). 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 right-hand side of this inequality is equal to $\Phi_{|X|}^{h(S)}$, according to the recurrence equation satisfied by Φ.

Proof of theorem 4.2.3: This is the well-known Chernoff bound for the deviation of sums of Bernoulli random variables: let $(\sigma_1, \ldots, \sigma_n)$ be i.i.d. Bernoulli random variables 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 ,

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

Differentiating the right-hand side in λ shows that its minimal value is $\exp\big[-n\mathcal{K}(\frac{h}{n}, \frac{1}{2})\big]$, 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 B_p and B_q of parameters p and q. Indeed the optimal value λ^* of λ is such that

$$h = n\frac{\mathbb{E}\big[\sigma_1 \exp(-\lambda^*\sigma_1)\big]}{\mathbb{E}\big[\exp(-\lambda^*\sigma_1)\big]} = nB_{h/n}(\sigma_1).$$

Therefore, using the fact that two Bernoulli distributions with the same expectations are equal,

$$\log\big\{\mathbb{E}\big[\exp(-\lambda^*\sigma_1)\big]\big\} = -\lambda^* B_{h/n}(\sigma_1) - \mathcal{K}(B_{h/n}, B_{1/2}) = -\lambda^*\tfrac{h}{n} - \mathcal{K}(\tfrac{h}{n}, \tfrac{1}{2}).$$

The announced result then follows from the identity

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

4.2.3. Vapnik–Cervonenkis dimension of linear rules with margin. The proof of the following theorem was suggested to us by a similar proof presented in Cristianini et al. (2000).

Theorem 4.2.4. *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 this notation,

$$\frac{\mathbb{V}\mathrm{ar}(x_1, \dots, x_n)}{\gamma^2} \geq \begin{cases} n-1 & \text{when } n \text{ is even,} \\ (n-1)\frac{n^2-1}{n^2} & \text{when } n \text{ is odd.} \end{cases} \tag{4.5}$$

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

$$\begin{aligned} \left\langle \sum_{i=1}^{n} s_{\sigma(i)} x_i, w \right\rangle &= \sum_{i=1}^{n} s_{\sigma(i)} \big(\langle x_i, w \rangle - b - b_i \big) + \sum_{i=1}^{n} s_{\sigma(i)} b_i \\ &\geq \sum_{i=1}^{n} \gamma |s_{\sigma(i)}| + s_{\sigma(i)} b_i. \end{aligned}$$

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

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

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

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

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 (4.5) 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). □

4.2.4. Application to Support Vector Machines. We are going to apply Theorem 4.2.4 (page 176) to Support Vector Machines in the transductive case. Let $(X_i, Y_i)_{i=1}^{(k+1)N}$ be distributed according to some partially exchangeable distribution $\mathbb{P}$ and assume that $(X_i)_{i=1}^{(k+1)N}$ 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 $Z' = (X_i, y'_i)_{i=1}^{(k+1)N}$, where $(y'_i)_{i=1}^{(k+1)N} \in \mathcal{Y}^{(k+1)N}$, let $\hat{f}_{Z'}$ be the Support Vector Machine defined by K and Z' and let $\gamma(Z')$ be its margin. Let

$$R^2 = \max_{i=1,\dots,(k+1)N} K(X_i, X_i) + \frac{1}{(k+1)^2N^2}\sum_{j=1}^{(k+1)N}\sum_{k=1}^{(k+1)N} K(X_j, X_k) - \frac{2}{(k+1)N}\sum_{j=1}^{(k+1)N} K(X_i, X_j).$$

This is an easily computable upper-bound for the radius of some ball containing the image of $(X_1, \dots, X_{(k+1)N})$ in feature space.

Let us define for any integer h the margins

$$\gamma_{2h} = (2h-1)^{-1/2} \text{ and } \gamma_{2h+1} = \left[2h\left(1 - \frac{1}{(2h+1)^2}\right)\right]^{-1/2}. \tag{4.6}$$

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

$$\mathcal{R}_h = \left\{\hat{f}_{Z'} : Z' = (X_i, y'_i)_{i=1}^{(k+1)N} \text{ is } K\text{-separable and } \gamma(Z') \geq R\gamma_h\right\}.$$

The family of models $\mathcal{R}_h$, $h = 1, \dots, N$ is nested, and we know from Theorem 4.2.4 (page 176) and Theorems 4.2.2 (page 174) and 4.2.3 (page 174) that

$$\log\bigl(|\mathcal{R}_h|\bigr) \leq h \log\bigl(\tfrac{(k+1)eN}{h}\bigr).$$

We can then consider on the large model $\mathcal{R} = \bigsqcup_{h=1}^{N} \mathcal{R}_h$ (the disjoint union of the sub-models) an exchangeable prior π which is uniform on each $\mathcal{R}_h$ and is such that $\pi(\mathcal{R}_h) \geq \frac{1}{h(h+1)}$. Applying Theorem 3.2.3 (page 143) we get

PROPOSITION 4.2.5. *With* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any* $h = 1, \dots, N$*, any Support Vector Machine* $f \in \mathcal{R}_h$*,*

$$r_2(f) \leq \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{1 - \exp\Bigl[-\frac{\lambda}{N} r_1(f) - \frac{h}{N}\log\Bigl(\frac{e(k+1)N}{h}\Bigr) - \frac{\log[h(h+1)] - \log(\epsilon)}{N}\Bigr]}{1 - \exp(-\frac{\lambda}{N})} - \frac{r_1(f)}{k}.$$

Searching the whole model $\mathcal{R}_h$ to optimize the bound may require more computer resources than are available, but any heuristic can be applied to choose f, since the bound is uniform. For instance, a Support Vector Machine f' using a box constraint can be trained from the training set $(X_i, Y_i)_{i=1}^{N}$ and then $(y'_i)_{i=1}^{(k+1)N}$ can be set to $y'_i = \operatorname{sign}(f'(X_i))$, $i = 1, \dots, (k+1)N$.

4.2.5. INDUCTIVE MARGIN BOUNDS FOR SUPPORT VECTOR MACHINES. In order to establish inductive margin bounds, we will need a different combinatorial lemma. It is due to Alon et al. (1997). We will reproduce their proof with some tiny improvements on the values of constants.

Let us consider the finite case when $\mathcal{X} = \{1, \dots, n\}$, $\mathcal{Y} = \{1, \dots, b\}$ and $b \geq 3$. The question we will study would be meaningless when $b \leq 2$. Assume as usual that we are dealing with a prescribed set of classification rules $\mathcal{R} = \bigl\{f : \mathcal{X} \to \mathcal{Y}\bigr\}$. Let us say that a pair (A, s), where $A \subset \mathcal{X}$ is a non-empty set of shapes and $s : A \to \{2, \dots, b-1\}$ 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 $\min_{x \in A} \sigma_x \bigl[f(x) - s(x)\bigr] \geq 1$.

DEFINITION 4.2.1. 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 4.2.6. *With the above notation, whenever the fat shattering dimension of $(\mathcal{X},\mathcal{R})$ is not greater than h,*

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

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, from its definition $t(m,n)$ is clearly 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 4.2.7. *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 $1 \leq y_1 < y_2 - 1 \leq b-1$, 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. Finally, 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). $\square$

Resuming the proof of lemma 4.2.6, 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 $n > r$,

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

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 $\mathbb{R}^d$ (the generalization to Support Vector Machines being straightforward). Assume that $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. For any sample $(X)_{i=1}^{(k+1)N}$, let

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

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 $\gamma = R(X_1^{(k+1)N})\gamma_h$, where γ_h is defined by equation (4.6, page 178).

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 4.2.1) of

$$\Big(X_1^{(k+1)N}, \big\{(G_{w,b} + 7)/2 : (w, b) \in \Theta\big\}\Big)$$

is not greater than h (according to Theorem 4.2.4, page 176), therefore there is some set $\mathcal{F}$ of functions from $X_1^{(k+1)N}$ to $\{-5, -3, -1, +1, +3, +5\}$ such

that

$$\log\big(|\mathcal{F}|\big) \leq \log\big[20(k+1)N\big]\left\{\frac{h}{\log(2)}\left[\log\left(\frac{4(k+1)N}{h}\right)+1\right]+1\right\}+\log(2).$$

and for any $(w,b)\in\Theta$, there is $f_{w,b}\in\mathcal{F}$ such that $\sup\big\{|f_{w,b}(X_i)-G_{w,b}(X_i)|\,:\, i=1,\ldots,(k+1)N\big\}\leq 2$. Moreover, the choice of $f_{w,b}$ may be required to depend on $(X_i)_{i=1}^{(k+1)N}$ in an exchangeable way. Similarly to Theorem 3.2.3 (page 143), it can be proved that for any partially exchangeable probability distribution $\mathbb{P}\in\mathcal{M}_+^1(\Omega)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $f_{w,b}\in\mathcal{F}$,

$$\begin{aligned}
&\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\mathbb{1}\big[f_{w,b}(X_i)Y_i\leq 1\big]\\
&\quad\leq\frac{k+1}{k}\inf_{\lambda\in\mathbb{R}_+}\big[1-\exp(-\tfrac{\lambda}{N})\big]^{-1}\Bigg\{1-\\
&\qquad\qquad\exp\Bigg[-\frac{\lambda}{N^2}\sum_{i=1}^{N}\mathbb{1}\big[f_{w,b}(X_i)Y_i\leq 1\big]-\frac{\log\big(|\mathcal{F}|\big)-\log(\epsilon)}{N}\Bigg]\Bigg\}\\
&\qquad\qquad\qquad\qquad-\frac{1}{kN}\sum_{i=1}^{N}\mathbb{1}\big[f_{w,b}(X_i)Y_i\leq 1\big].
\end{aligned}$$

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 4.2.8. *Let us consider the sequence $(\gamma_h)_{h\in\mathbb{N}^*}$ defined by equation (4.6, page 178). With $\mathbb{P}$ probability at least $1-\epsilon$, for any $(w,b)\in\Theta$,*

$$\begin{aligned}
&\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\mathbb{1}\Big\{2\mathbb{1}\big[g_{w,b}(X_i)\geq 0\big]-1\neq Y_i\Big\}\\
&\leq\frac{k+1}{k}\inf_{\lambda\in\mathbb{R}_+,h\in\mathbb{N}^*}\big[1-\exp(-\tfrac{\lambda}{N})\big]^{-1}\Bigg\{1-\\
&\quad\exp\Bigg[-\frac{\lambda}{N^2}\sum_{i=1}^{N}\mathbb{1}\big[g_{w,b}(X_i)Y_i\leq 4R\gamma_h\big]\\
&\qquad-\frac{\log\big[20(k+1)N\big]\Big\{\frac{h}{\log(2)}\log\Big(\frac{4e(k+1)N}{h}\Big)+1\Big\}+\log\Big[\frac{2h(h+1)}{\epsilon}\Big]}{N}\Bigg]\Bigg\}
\end{aligned}$$

$$- \frac{1}{kN} \sum_{i=1}^{N} \mathbb{1}\big[g_{w,b}(X_i)Y_i \leq 4R\gamma_h\big].$$

Properly speaking this theorem is not a margin bound, but more precisely a *margin quantile* bound, since it covers the case where some fraction of the training sample falls within the region defined by the margin parameter γ_h which optimizes the bound.

As a consequence though, we get a true (weaker) margin bound: with $\mathbb{P}$ probability at least $1-\epsilon$, for any $(w,b) \in \Theta$ such that

$$\gamma = \min_{i=1,\dots,N} g_{w,b}(X_i)Y_i > 0,$$

$$\begin{aligned}\frac{1}{kN} \sum_{i=N+1}^{(k+1)N} & \mathbb{1}\big[g_{w,b}(X_i)Y_i < 0\big] \\ & \leq \tfrac{k+1}{k}\left\{1 - \exp\left[-\tfrac{\log\left[20(k+1)N\right]}{N}\left\{\tfrac{16R^2+2\gamma^2}{\log(2)\gamma^2}\log\left(\tfrac{e(k+1)N\gamma^2}{4R^2}\right)+1\right\}\right.\right. \\ & \qquad \left.\left. + \frac{1}{N}\log(\tfrac{\epsilon}{2})\right]\right\}.\end{aligned}$$

This inequality compares favourably with similar inequalities in Cristianini et al. (2000), which moreover do not extend to the margin quantile case as this one.

Let us also mention that it is easy to circumvent the fact that R is not observed when the test set $X_{N+1}^{(k+1)N}$ is not observed.

Indeed, we can consider the sample obtained by projecting $X_1^{(k+1)N}$ on some ball of fixed radius $R_{\max}$, putting

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

We can further consider an atomic prior distribution $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$ bearing on $R_{\max}$, to obtain a uniform result through a union bound. As a consequence of the previous theorem, we have

Corollary 4.2.9. *For any atomic prior $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$, for any partially exchangeable probability measure $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $(w,b) \in \Theta$, any $R_{\max} \in \mathbb{R}_+$,*

$$\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\mathbb{1}\Big\{2\mathbb{1}\big[g_{w,b}\circ t_{R_{\max}}(X_i)\geq 0\big]-1\neq Y_i\Big\}$$
$$\leq \frac{k+1}{k}\inf_{\lambda\in\mathbb{R}_+,h\in\mathbb{N}^*}\big[1-\exp(-\tfrac{\lambda}{N})\big]^{-1}\Bigg\{1-$$
$$\exp\Bigg[-\frac{\lambda}{N^2}\sum_{i=1}^{N}\mathbb{1}\big[g_{w,b}\circ t_{R_{\max}}(X_i)Y_i\leq 4R_{\max}\gamma_h\big]$$
$$-\frac{\log\big[20(k+1)N\big]\Big\{\frac{h}{\log(2)}\log\Big(\frac{4e(k+1)N}{h}\Big)+1\Big\}+\log\Big[\frac{2h(h+1)}{\epsilon\nu(R_{\max})}\Big]}{N}\Bigg]\Bigg\}$$
$$-\frac{1}{kN}\sum_{i=1}^{N}\mathbb{1}\big[g_{w,b}\circ t_{R_{\max}}(X_i)Y_i\leq 4R_{\max}\gamma_h\big].$$

Let us remark that $t_{R_{\max}}(X_i)=X_i$, $i=N+1,\ldots,(k+1)N$, as soon as we consider only the values of $R_{\max}$ not smaller than $\max_{i=N+1,\ldots,(k+1)N}\|X_i\|$ in this corollary. Thus we obtain a bound on the transductive generalization error of the unthresholded classification rule $2\mathbb{1}\big[g_{w,b}(X_i)\geq 0\big]-1$, as well as some incitation to replace it with a thresholded rule when the value of $R_{\max}$ minimizing the bound falls below $\max_{i=N+1,\ldots,(k+1)N}\|X_i\|$.

Appendix: classification by thresholding

In this appendix, we show how the bounds given in the first section of this monograph can be computed in practice on a simple example: the case when the classification is performed by comparing a series of measurements to threshold values. Let us mention that our description covers the case when the same measurement is compared to several thresholds, since it is enough to repeat a measurement in the list of measurements describing a pattern to cover this case.

5.1. Description of the model

Let us assume that the patterns we want to classify are described through h real valued measurements normalized in the range $(0,1)$. In this setting the pattern space can thus be defined as $\mathcal{X} = (0,1)^h$.

Consider the threshold set $\mathcal{T} = (0,1)^h$ and the response set $\mathcal{R} = \mathcal{Y}^{\{0,1\}^h}$. For any $t \in (0,1)^h$ and any $a : \{0,1\}^h \to \mathcal{Y}$, let

$$f_{(t,a)}(x) = a\Big\{\big[\mathbb{1}(x^j \geq t_j)\big]_{j=1}^h\Big\}, \quad x \in \mathcal{X},$$

where x^j is the jth coordinate of $x \in \mathcal{X}$. Thus our parameter set here is $\Theta = \mathcal{T} \times \mathcal{R}$. Let us consider the Lebesgue measure L on $\mathcal{T}$ and the uniform probability distribution U on $\mathcal{R}$. Let our prior distribution be $\pi = L \otimes U$. Let us define for any threshold sequence $t \in \mathcal{T}$

$$\Delta_t = \Big\{t' \in \mathcal{T} : \overline{(t'_j, t_j)} \cap \{X_i^j; i = 1, \ldots, N\} = \varnothing, j = 1, \ldots, h\Big\},$$

where X_i^j is the jth coordinate of the sample pattern X_i, and where the interval $\overline{(t'_j, t_j)}$ of the real line is defined as the convex hull of the two point set $\{t'_j, t_j\}$, whether $t'_j \leq t_j$ or not. We see that Δ_t is the set of thresholds giving the same response as t on the training patterns. Let us consider for any $t \in \mathcal{T}$ the middle

$$m(\Delta_t) = \frac{\int_{\Delta_t} t' L(dt')}{L(\Delta_t)}$$

of Δ_t. The set Δ_t being a product of intervals, its middle is the point whose coordinates are the middle of these intervals. Let us introduce the finite set T composed of the middles of the cells Δ_t, which can be defined as

$$T = \{t \in \mathcal{T} : t = m(\Delta_t)\}.$$

It is easy to see that $|T| \leq (N+1)^h$ and that $|\mathcal{R}| = |\mathcal{Y}|^{2^h}$.

5.2. Computation of inductive bounds

For any parameter $(t, a) \in \mathcal{T} \times \mathcal{R} = \Theta$, let us consider the posterior distribution defined by its density

$$\frac{d\rho_{(t,a)}}{d\pi}(t', a') = \frac{\mathbb{1}\big(t' \in \Delta_t\big)\mathbb{1}\big(a' = a\big)}{\pi\big(\Delta_t \times \{a\}\big)}.$$

In fact we are considering a finite number of posterior distributions, since $\rho_{(t,a)} = \rho_{(m(\Delta_t),a)}$, where $m(\Delta_t) \in T$. Moreover, for any exchangeable sample distribution $\mathbb{P} \in \mathcal{M}_+^1\big[(\mathcal{X} \times \mathcal{Y})^{N+1}\big]$ and any thresholds $t \in \mathcal{T}$,

$$\mathbb{P}\Big[\overline{(X_{N+1}^j, t_j)} \cap \{X_i^j, i = 1, \ldots, N\} = \varnothing\Big] \leq \frac{2}{N+1}.$$

Thus, for any $(t, a) \in \Theta$,

$$\mathbb{P}\Big\{\rho_{(t,a)}\big[f_\cdot(X_{N+1})\big] \neq f_{(t,a)}(X_{N+1})\Big\} \leq \frac{2h}{N+1},$$

showing that the classification produced by $\rho_{(t,a)}$ on new examples is typically non-random; this result is only indicative, since it is concerned with a non-random choice of (t, a).

Let us compute the various quantities needed to apply the results of the first section, focussing our attention on Theorem 2.1.3 (page 72).

First note that $\rho_{(t,a)}(r) = r[(t, a)]$. The entropy term is such that

$$\mathcal{K}(\rho_{t,a}, \pi) = -\log\big[\pi\big(\Delta_t \times \{r\}\big)\big] = -\log\big[L(\Delta_t)\big] + 2^h \log\big(|\mathcal{Y}|\big).$$

Let us notice accordingly that

$$\min_{(t,a)\in\Theta} \mathcal{K}(\rho_{(t,a)}, \pi) \leq h \log(N+1) + 2^h \log\big(|\mathcal{Y}|\big).$$

Let us introduce the counters

$$b_y^t(c) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\Big\{Y_i = y \text{ and } \big[\mathbb{1}(X_i^j \geq t_j)\big]_{j=1}^h = c\Big\},$$

$$t \in T, c \in \{0,1\}^h, y \in \mathcal{Y},$$

$$b^t(c) = \sum_{y \in \mathcal{Y}} b_y^t(c) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\Big\{ \big[\mathbb{1}(X_i^j \geq t_j)\big]_{j=1}^h = c \Big\}, \qquad t \in T, c \in \{0,1\}^h.$$

Since

$$r[(t,a)] = \sum_{c \in \{0,1\}^h} \big[b^t(c) - b_{a(c)}^t(c) \big],$$

the partition function of the Gibbs estimator can be computed as

$$\begin{aligned}
\pi\big[\exp(-\lambda r)\big] &= \sum_{t \in T} L(\Delta_t) \sum_{a \in \mathcal{R}} \frac{1}{|\mathcal{Y}|^{2^h}} \exp\bigg[-\lambda \sum_{i=1}^{N} \mathbb{1}\big[Y_i \neq f_{(t,a)}(X_i) \big] \bigg] \\
&= \sum_{t \in T} L(\Delta_t) \sum_{a \in \mathcal{R}} \frac{1}{|\mathcal{Y}|^{2^h}} \exp\bigg[-\lambda \sum_{c \in \{0,1\}^h} \big[b^t(c) - b_{a(c)}^t(c) \big] \bigg] \\
&= \sum_{t \in T} L(\Delta_t) \prod_{c \in \{0,1\}^h} \bigg[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Big(-\lambda \big[b^t(c) - b_y^t(c) \big] \Big) \bigg].
\end{aligned}$$

We see that the number of operations needed to compute $\pi\big[\exp(-\lambda r)\big]$ is proportional to $|T| \times 2^h \times |\mathcal{Y}| \leq (N+1)^h 2^h |\mathcal{Y}|$. An exact computation will therefore be feasible only for small values of N and h. For higher values, a Monte Carlo approximation of this sum will have to be performed instead.

If we want to compute the bound provided by Theorem 2.1.3 (page 72) or by Theorem 2.2.2 (page 89), we need also to compute, for any fixed parameter $\theta \in \Theta$, quantities of the type

$$\pi_{\exp(-\lambda r)}\Big\{ \exp\big[\xi m'(\cdot, \theta)\big] \Big\} = \pi_{\exp(-\lambda r)}\Big\{ \exp\big[\xi \rho_\theta(m')\big] \Big\}, \quad \lambda, \xi \in \mathbb{R}_+.$$

We need to introduce

$$\overline{b}_y^t(\theta, c) = \frac{1}{N} \sum_{i=1}^{N} \Big| \mathbb{1}\big[f_\theta(X_i) \neq Y_i \big] - \mathbb{1}(y \neq Y_i) \Big| \mathbb{1}\big\{ \big[\mathbb{1}(X_i^j \geq t_j)\big]_{j=1}^h = c \big\}.$$

Similarly to what has been done previously, we obtain

$$\begin{aligned}
&\pi\big\{ \exp\big[-\lambda r + \xi m'(\cdot, \theta) \big] \big\} \\
&\qquad = \sum_{t \in T} L(\Delta_t) \prod_{c \in \{0,1\}^h} \bigg[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Big(-\lambda \big[b^t(c) - b_y^t(c) \big] + \xi \overline{b}_y^t(\theta, c) \Big) \bigg].
\end{aligned}$$

We can then compute

$$\begin{aligned}
\pi_{\exp(-\lambda r)}(r) &= -\frac{\partial}{\partial \lambda} \log\big\{ \pi\big[\exp(-\lambda r)\big] \big\}, \\
\pi_{\exp(-\lambda r)}\Big\{ \exp\big[\xi \rho_\theta(m')\big] \Big\} &= \frac{\pi\big\{ \exp\big[-\lambda r + \xi m'(\cdot, \theta) \big] \big\}}{\pi\big[\exp(-\lambda r)\big]},
\end{aligned}$$

$$\pi_{\exp(-\lambda r)}\big[m'(\cdot,\theta)\big] = \frac{\partial}{\partial\xi}_{|\xi=0} \log\Big[\pi\big\{\exp\big[-\lambda r + \xi m'(\cdot,\theta)\big]\big\}\Big].$$

This is all we need to compute $B(\rho_\theta,\beta,\gamma)$ (and also $B(\pi_{\exp(-\lambda r)},\beta,\gamma)$) in Theorem 2.1.3 (page 72), using the approximation

$$\begin{aligned}
&\log\Big\{\pi_{\exp(-\lambda_1 r)}\Big[\exp\big\{\xi\pi_{\exp(-\lambda_2 r)}(m')\big\}\Big]\Big\} \\
&\quad\le \log\Big\{\pi_{\exp(-\lambda_1 r)}\Big[\exp\big\{\xi m'(\cdot,\theta)\big\}\Big]\Big\} + \xi\pi_{\exp(-\lambda_2 r)}\big[m'(\cdot,\theta)\big], \quad \xi\ge 0.
\end{aligned}$$

Let us also explain how to apply the posterior distribution $\rho_{(t,a)}$, in other words our randomized estimated classification rule, to a new pattern X_{N+1}:

$$\begin{aligned}
\rho_{(t,a)}\big[f_\cdot(X_{N+1}) = y\big] &= L(\Delta_t)^{-1}\int_{\Delta_t} \mathbb{1}\Big[a\big\{\big[\mathbb{1}(X^j_{N+1}\ge t'_j)\big]_{j=1}^h\big\} = y\Big]L(dt') \\
&= L(\Delta_t)^{-1}\sum_{c\in\{0,1\}^h} L\Big(\Big\{t'\in\Delta_t : \big[\mathbb{1}(X^j_{N+1}\ge t'_j)\big]_{j=1}^h = c\Big\}\Big)\mathbb{1}\big[a(c)=y\big].
\end{aligned}$$

Let us define for short

$$\Delta_t(c) = \Big\{t'\in\Delta_t : \big[\mathbb{1}(X^j_{N+1}\ge t'_j)\big]_{j=1}^h = c\Big\}, \qquad c\in\{0,1\}^h.$$

With this notation

$$\rho_{(t,a)}\big[f_\cdot(X_{N+1}) = y\big] = L\big(\Delta_t\big)^{-1}\sum_{c\in\{0,1\}^h} L\big[\Delta_t(c)\big]\mathbb{1}\big[a(c)=y\big].$$

We can compute in the same way the probabilities for the label of the new pattern under the Gibbs posterior distribution:

$$\begin{aligned}
&\pi_{\exp(-\lambda r)}\big[f_\cdot(X_{N+1}) = y'\big] \\
&\quad= \Bigg\{\sum_{t\in T}\prod_{c\in\{0,1\}^h}\Big[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Big(-\lambda\big[b^t(c)-b^t_y(c)\big]\Big)\Big] \\
&\qquad\times\sum_{c\in\{0,1\}^h} L\big[\Delta_t(c)\big]\frac{\sum_{y\in\mathcal{Y}}\mathbb{1}(y=y')\exp\big\{-\lambda\big[b^t(c)-b^t_y(c)\big]\big\}}{\sum_{y\in\mathcal{Y}}\exp\big\{-\lambda\big[b^t(x)-b^t_y(c)\big]\big\}}\Bigg\} \\
&\qquad\times\Bigg\{\sum_{t\in T} L(\Delta_t)\prod_{c\in\{0,1\}^h}\Big[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Big(-\lambda\big[b^t(c)-b^t_y(c)\big]\Big)\Big]\Bigg\}^{-1}.
\end{aligned}$$

5.3. Transductive bounds

In the case when we observe the patterns of a shadow sample $(X_i)_{i=N+1}^{(k+1)N}$ on top of the training sample $(X_i,Y_i)_{i=1}^N$, we can introduce the set of thresholds responding as t on the extended sample $(X_i)_{i=1}^{(k+1)N}$

$$\overline{\Delta}_t = \Big\{t'\in\mathcal{T} : \overline{(t'_j,t_j)}\cap\big\{X^j_i; i=1,\dots,(k+1)N\big\} = \varnothing, j=1,\dots,h\Big\},$$

consider the set

$$\overline{T} = \big\{t \in \mathcal{T} : t = m(\overline{\Delta}_t)\big\},$$

of the middle points of the cells $\overline{\Delta}_t$, $t \in \mathcal{T}$, and replace the Lebesgue measure $L \in \mathcal{M}_+^1\big[(0,1)^h\big]$ of the previous section with the uniform probability measure $\overline{L}$ on $\overline{T}$. We can then consider $\pi = \overline{L} \otimes U$, where U is as previously the uniform probability measure on $\mathcal{R}$. This gives obviously an exchangeable posterior distribution and therefore qualifies π for transductive bounds. Let us notice that $|\overline{T}| \leq \big[(k+1)N+1\big]^h$, and therefore that $\pi(t,a) \geq \big[(k+1)N+1\big]^{-h} |\mathcal{Y}|^{-2^h}$, for any $(t,a) \in \overline{T} \times \mathcal{R}$.

For any $(t,a) \in \mathcal{T} \times \mathcal{R}$ we may similarly to the inductive case consider the posterior distribution $\rho_{(t,a)}$ defined by

$$\frac{d\rho_{(t,a)}}{d\pi}(t',a') = \frac{\mathbb{1}(t' \in \Delta_t)\mathbb{1}(a'=a)}{\pi\big(\Delta_t \times \{a\}\big)},$$

but we may also consider $\delta_{(m(\overline{\Delta}_t),a)}$, which is such that $r_i\{[m(\overline{\Delta}_t),a]\} = r_i[(t,a)]$, $i=1,2$, whereas only $\rho_{(t,a)}(r_1) = r_1[(t,a)]$, while

$$\rho_{(t,a)}(r_2) = \frac{1}{|\overline{T} \cap \Delta_t|} \sum_{t' \in \overline{T} \cap \Delta_t} r_2[(t',a)].$$

We get

$$\begin{aligned}\mathcal{K}(\rho_{(t,a)},\pi) &= -\log\big[\overline{L}(\Delta_t)\big] + 2^h \log\big(|\mathcal{Y}|\big) \\ &\leq \log\big(|\overline{T}|\big) + 2^h \log(|\mathcal{Y}|) = \mathcal{K}(\delta_{[m(\overline{\Delta}_t),a]},\pi) \\ &\leq h \log\big[(k+1)N+1\big] + 2^h \log(|\mathcal{Y}|),\end{aligned}$$

whereas we had no such uniform bound in the inductive case. Similarly to the inductive case

$$\pi\big[\exp(-\lambda r_1)\big] = \sum_{t \in T} \overline{L}(\Delta_t) \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Big(-\lambda\big[b^t(c) - b_y^t(c)\big]\Big)\right].$$

Moreover, for any $\theta \in \Theta$,

$$\begin{aligned}\pi\big\{\exp\big[-\lambda r_1 + \xi\rho_\theta(m')\big]\big\} &= \pi\big\{\exp\big[-\lambda r_1 + \xi m'(\cdot,\theta)\big]\big\} \\ &= \sum_{t \in T} \overline{L}(\Delta_t) \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Big(-\lambda\big[b^t(c) - b_y^t(c)\big] + \xi\bar{b}(\theta,c)\Big)\right].\end{aligned}$$

The bound for the transductive counterpart to Theorems 2.1.3 (page 72) or 2.2.2 (page 89), obtained as explained page 141, can be computed as in the inductive case, from these two partition functions and the above entropy computation.

Let us mention finally that, using the same notation as in the inductive case,

$$\begin{aligned}
&\pi_{\exp(-\lambda r_1)}\big[f_\cdot(X_{N+1}) = y'\big] \\
&\quad = \Bigg\{\sum_{t\in T}\prod_{c\in\{0,1\}^h}\Bigg[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Big(-\lambda\big[b^t(c) - b_y^t(c)\big]\Big)\Bigg] \\
&\qquad \times \sum_{c\in\{0,1\}^h}\overline{L}\big[\Delta_t(c)\big]\frac{\sum_{y\in\mathcal{Y}}\mathbb{1}(y=y')\exp\big\{-\lambda\big[b^t(c) - b_y^t(c)\big]\big\}}{\sum_{y\in\mathcal{Y}}\exp\big\{-\lambda\big[b^t(x) - b_y^t(c)\big]\big\}}\Bigg\} \\
&\qquad\quad \times \Bigg\{\sum_{t\in T}\overline{L}(\Delta_t)\prod_{c\in\{0,1\}^h}\Bigg[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Big(-\lambda\big[b^t(c) - b_y^t(c)\big]\Big)\Bigg]\Bigg\}^{-1}.
\end{aligned}$$

To conclude this appendix on classification by thresholding, note that similar factorized computations are feasible in the important case of *classification trees*. This can be achieved using some variant of the *context tree weighting method* discovered by Willems et al. (1995) and successfully used in lossless compression theory. The interested reader can find a description of this algorithm applied to classification trees in Catoni (2004, page 62).

Bibliography

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

J.-Y. Audibert, (2004a) Aggregated estimators and empirical complexity for least square regression, *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, **40**: 685-736.

J. Y. Audibert, (2004b) *PAC-Bayesian Statistical Learning Theory*, PhD, University Paris 6, June, 1–178, `http://cermics.enpc.fr/~audibert/`.

A. Barron (1987) Are Bayes rules consistent in information ? in *Open Problems in Communication and Computation,* T. M. Cover and B. Gopinath (Editors), Springer-Verlag, New York, 85-91.

A. Barron and Y. Yang (1999) Information-theoretic determination of minimax rates of convergence, *Annals of Statistics* **27**: 1564–1599.

A. Barron, L. Birgé and P. Massart, (1999) Risk bounds for model selection by penalization, *Probabability Theory and Related Fields*, **113**: 301–413.

G. Blanchard, (1999) The "progressive mixture" estimator for regression trees, *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, **35**: 793–820.

G. Blanchard, (2001) *Mixture and Aggregation of Estimators for Pattern Recognition. Application to Decision Trees* PhD, Université Paris 13, January, 1–158, `http://ida.first.fraunhofer.de/~blanchard/`.

G. Blanchard, (2004) Un algorithme accéléré d'échantillonnage Bayésien pour le modèle CART, *Revue d'Intelligence artificielle,* **18**: 383–410.

L. Birgé and P. Massart, (1997) From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam*, D. Pollard (Editor), Springer-Verlag, New York, 55-87.

L. Birgé and P. Massart, (1998) Minimum contrast estimators on sieves, *Bernoulli* **4**: 329–375.

L. Birgé and P. Massart, (2001a) A generalized C_p criterion for Gaussian model selection, *preprint*, 1–39, `http://www.math.u-psud.fr/~massart/`.

L. Birgé and P. Massart, (2001b) Gaussian model selection, *Journal of the European Mathematical Society*, **3**: 203-268.

A. Blum and J. Langford, (2003) PAC-MDL bounds, in *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, Lecture Notes in Computer Science **2777**, Springer, 344-357.

O. Catoni, (2002) Data compression and adaptive histograms, in *Foundations of Computational Mathematics, Proceedings of the Smalefest 2000*, F. Cucker and J.M. Rojas (Editors), World Scientific, 35–60.

O. Catoni, (2003) Laplace transform estimates and deviation inequalities, *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, **39**: 1–26.

O. Catoni, (2004) *Statistical Learning Theory and Stochastic Optimization, Ecole d'Été de Probabilités de Saint-Flour XXXI - 2001*, Lecture Notes in Mathematics, **1851**, Springer, 1–270.

N. Cristianini and J. Shawe Taylor, (2000) *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods,* Cambridge University Press, 1–189.

M. Feder and N. Merhav, (1996) Hierarchical universal coding, *IEEE Transactions on Informormation Theory*, **42**: 1354–1364.

T. Hastie, R. Tibshirani and J. Friedman, (2001) *The Elements of Statistical Learning*, Springer-Verlag, New York, 1–536.

J. Langford and D. McAllester, (2004) Computable shell decomposition bounds, *Journal of Machine Learning Research*, **5**: 529-547, (communicated at COLT 2000).

J. Langford and M. Seeger, (2001a) Bounds for averaging classifiers, *technical report CMU-CS-01-102*, Carnegie Mellon University, 1–13, `http://www.cs.cmu.edu/~jcl`.

J. Langford, M. Seeger and N. Megiddo, (2001b) An improved predictive accuracy bound for averaging classifiers, *International Conference on Machine Learning*, **18**, 290-297.

N. Littlestone and M. Warmuth, (1986) Relating data compression and learnability, *technical report,* University of California, Santa Cruz, 1–13, `http://www.soe.ucsc.edu/~manfred/pubs.html`.

D. A. McAllester, (1998) Some PAC-Bayesian theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, ACM, New York, 230–234 (electronic).

D. A. McAllester, (1999) PAC-Bayesian Model Averaging, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, ACM, New York, 164–170 (electronic).

C. McDiarmid, (1998) Concentration, in *Probabilistic Methods for Algorithmic Discrete Mathematics*, M. Habib , C. McDiarmid and B. Reed (Editors), Springer, 195–248.

E. Mammen and A. Tsybakov, (1999) Smooth discrimination analysis. *Annals of Statistics* **27**: 1808–1829.

B. Y. Ryabko, (1984) Twice-universal coding, *Problems of Information Transmission*, **20**: 24-28.

M. Seeger, (2002) PAC-Bayesian generalization error bounds for Gaussian process classification, *Journal of Machine Learning Research* **3**: 233–269.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson and M. Anthony, (1998) Structural risk minimization over data-dependent hierarchies, *IEEE Transactions on Information Theory*, **44**: 1926–1940.

J. Shawe-Taylor and N. Cristianini, (2002) On the generalization of soft margin algorithms, *IEEE Transactions on Information Theory,* **48**: 2721–2735.

A. Tsybakov, (2004) Optimal aggregation of classifiers in statistical learning, *Annals of Statistics* **32**: 135–166.

A. Tsybakov and S. Van de Geer, (2005) Square root penalty: adaptation to the margin in classification and in edge estimation, *Annals of Statistics*, **33**: 1203-1224.

S. Van de Geer, (2000) *Applications of Empirical Process Theory.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 1-286.

V. N. Vapnik, (1998) *Statistical Learning Theory*, Wiley, New York, 1–736.

J.-P. Vert, (2000) Double mixture and universal inference, *preprint*, 1–27, `http://cbio.ensmp.fr/~vert/publi/`.

J.-P. Vert, (2001a) Adaptive context trees and text clustering, *IEEE Transactions on Information Theory*, **47**: 1884–1901.

J.-P. Vert, (2001b) Text categorization using adaptive context trees, in *Proceedings of the CICLing-2001 conference*, A. Gelbukh (Editor), Lecture Notes in Computer Science **2004**, Springer-Verlag, 423–436.

F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, (1995) The context-tree weighting method: basic properties, *IEEE Transactions on Information Theory*, **41**: 653–664.

F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, (1996) Context weighting for general finite-context sources, *IEEE Transactions on Informormation Theory*, **42**: 1514–1520.

T. Zhang, (2006a) From epsilon-entropy to KL-entropy: analysis of minimum information complexity density estimation, *Annals of Statistics*, to appear, 1–37, `http://www-cs-students.stanford.edu/~tzhang/pubs.html`.

T. Zhang, (2006b) Information theoretical upper and lower bounds for statistical estimation, *IEEE Transactions on Information Theory*, to appear, 1–18, `http://www-cs-students.stanford.edu/~tzhang/pubs.html`.