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 viii), which is the number of errors made on the sample, and the expected error rate $R(\widehat{\theta})$, defined by equation (0.2, page ix), 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 $\Omega = \big(\mathcal{X} \times \mathcal{Y}\big)^N$, 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 $\mathcal{K}(\rho,\pi) < \infty$,*

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

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

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

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

$$\mathbb{P}\Bigl\{\exp\bigl\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \rho[U_\lambda(\theta)] - \mathcal{K}(\rho,\pi)\bigr\}\Bigr\} \le 1,$$

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,*

$$\mathbb{P}\Biggl\{\exp\Biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda\Bigl[\Phi_{\frac{\lambda}{N}}\bigl[\rho(R)\bigr] - \rho(r)\Bigr] - \mathcal{K}(\rho,\pi)\Biggr]\Biggr\}$$
$$\le \mathbb{P}\Biggl\{\exp\Biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda\Bigl[\rho\bigl(\Phi_{\frac{\lambda}{N}}\circ R\bigr) - \rho(r)\Bigr] - \mathcal{K}(\rho,\pi)\Biggr]\Biggr\} \le 1.$$

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,

$$\mathbb{P}\Biggl\{\exp\Biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda\Bigl[\rho\bigl[U_\lambda(\theta)\bigr] - \mathcal{K}(\rho,\pi)\Bigr]\Biggr]\Biggr\}$$
$$= \sup_{\rho:\Omega\to\mathcal{M}_+^1(\Theta)} \mathbb{P}\Biggl\{\exp\Bigl[\lambda\bigl[\rho\bigl[U_\lambda(\theta)\bigr] - \mathcal{K}(\rho,\pi)\bigr]\Bigr]\Biggr\},$$

and more formally,

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

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 $1 \leq \frac{\lambda}{N\left[1-\exp(-\frac{\lambda}{N})\right]} \leq \left[1 - \frac{\lambda}{2N}\right]^{-1}$ 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 4), 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 $x \in \{X_i : i = 1,\dots,N\}$). 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 144). 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

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

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

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

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

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

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 6): 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 $\mathbb{P}\big[\rho(r)\big]$, the average of the empirical risk, and of $\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]$, which controls overfitting.

Let the bound be written as

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

We see that

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

Thus, the optimal value for λ is such that

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

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 $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

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

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

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

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

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

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:

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

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

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

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

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

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

$$\begin{aligned}\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta &\leq \lambda\big(\operatorname*{ess\,inf}_\pi R + \eta\big)\\ &\quad + \int_0^\lambda \left[\frac{d_\eta}{\beta} \wedge (1 - \operatorname*{ess\,inf}_\pi R - \eta)\right] d\beta\end{aligned}$$

$$= \lambda\big(\operatorname*{ess\,inf}_{\pi} R + \eta\big) + d_\eta(\Theta, R) \log\left[\frac{e\lambda}{d_\eta(\Theta,R)}\big(1 - \operatorname*{ess\,inf}_{\pi} R - \eta\big)\right].$$

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 $q \in (0,1($,

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

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 4), 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 $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

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

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 3) and integrate it with respect to the prior distribution π to get for any real constant λ

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

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

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

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 $1-\epsilon$,*

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

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,

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

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,

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

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 $d(\rho,\epsilon) = \mathcal{K}(\rho,\pi) - \log(\epsilon)$,*

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

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 $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

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

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 6), 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 10), and if we choose $\rho = \pi_{\exp(-\lambda r)}$ to optimize this bound, we will be inclined to choose some π such that

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

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 7) (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

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

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

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

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 6), 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 $r \mapsto \log\{\pi[\exp(-\beta r)]\}$ as in equation (1.4) on page 10, we conclude that

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

This inequality has an interest of its own, since it provides a lower bound for $\mathbb{P}[\rho(R)]$. Moreover we can plug it into Theorem 1.2.1 (page 6) applied to the prior distribution $\pi_{\exp(-\beta R)}$ and obtain for any posterior distribution ρ and any positive parameter λ that

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

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

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

This is a convex function of p, moreover

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

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

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

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 $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

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

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

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

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* $\beta \in (0, N($,

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

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 6) when the posterior is a Gibbs measure $\rho = \pi_{\exp(-\beta r)}$. We see that we have got rid of the entropy term $\mathcal{K}\big[\pi_{\exp(-\beta r)}, \pi\big]$, 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* $0 \leq \beta < N[1-\exp(-\frac{\lambda}{N})]$,

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

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 13, 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 13).

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 10). According to Theorem 1.2.1 (page 6) applied to the localized prior $\pi_{\exp(-\beta R)}$,

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

where we have used as previously inequality (1.4) (page 10). 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 $0\leq\beta<N\big[1-\exp(-\frac{\lambda}{N})\big]$,*

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

Let us notice in particular that this theorem contains Theorem 1.2.4 (page 10) 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 10, 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)

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

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 11) 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 4), used with the above-mentioned local prior, shows that

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

Moreover

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

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 4), but this time to $-\beta$, we see that

$$\begin{aligned}&\mathbb{P}\bigg\{\exp\Big[\log\Big\{\pi\Big[\exp\big(-\beta\Phi_{-\frac{\beta}{N}}\circ R\big)\Big]\Big\}-\log\Big\{\pi\big[\exp(-\beta r)\big]\Big\}\Big]\bigg\}\\&=\mathbb{P}\bigg\{\exp\Big[\log\Big\{\pi\Big[\exp\big(-\beta\Phi_{-\frac{\beta}{N}}\circ R)\big)\Big]\Big\}+\inf_{\rho\in\mathcal{M}_+^1(\Theta)}\beta\rho(r)+\mathcal{K}(\rho,\pi)\Big]\bigg\}\\&\le\mathbb{P}\bigg\{\exp\Big[\log\Big\{\pi\Big[\exp\big(-\beta\Phi_{-\frac{\beta}{N}}\circ R)\big)\Big]\Big\}+\beta\overline{\pi}(r)+\mathcal{K}(\overline{\pi},\pi)\Big]\bigg\}\\&=\mathbb{P}\bigg\{\exp\Big[\beta\Big[\overline{\pi}(r)-\overline{\pi}\big(\Phi_{-\frac{\beta}{N}}\circ R\big)\Big]+\mathcal{K}(\overline{\pi},\overline{\pi})\Big]\bigg\}\le 1.\end{aligned}\tag{1.11}$$

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 $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,

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

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 $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

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

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

$$\mathbb{P}\bigg\{\pi_{\exp(-\beta r)}\Big[\exp\Big\{\beta\Big[r-\Phi_{-\frac{\beta}{N}}\circ R\Big]\Big\}\Big]\bigg\}\le 1.$$

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

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

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 $1-\epsilon$,*

$$R(\widehat{\theta}) \geq \Phi^{-1}_{-\frac{\beta}{N}}\left[r(\widehat{\theta}) - \frac{\log\left(\frac{d\rho}{d\pi_{\exp(-\beta r)}}\right) - \log(\epsilon)}{\beta}\right]$$
$$= \frac{\exp\left\{\frac{\beta}{N}\left[r(\widehat{\theta}) - \frac{\log\left(\frac{d\rho}{d\pi_{\exp(-\beta r)}}\right) - \log(\epsilon)}{\beta}\right]\right\} - 1}{\exp\left(\frac{\beta}{N}\right) - 1}.$$

Let us now resume our investigation of the upper deviations of $\rho(R)$. Using the Cauchy-Schwarz inequality to combine equations (1.9, page 18) and (1.11, page 19), we obtain

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

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

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

(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

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

and let us put

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

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

THEOREM 1.3.7. *For any positive constants α, γ, such that $0 \leq \gamma < \alpha < 1$, 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 20) can also be written as

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

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

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

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

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

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 $1-\epsilon$,*

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

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

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

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 13. It provides the upper bound

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

which shows that in Theorem 1.3.7 (page 21),

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

Similarly, in Theorem 1.3.8 above,

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

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 $\pi_{\exp[N\log(1-\alpha)r]}(R) \simeq \pi_{\exp(-693r)}(R) \le 0.332 \le 0.372$, 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 $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

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

Let us mention that this corollary applied to the above numerical example gives $\pi_{\exp(-200r)}(R) \le 0.475$ (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 $\inf_{\Theta_1} r \ge \inf_{\Theta_2} r \ge$

$\cdots \geq \inf_{\Theta_M} r$, 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

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

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 $\dfrac{\sum_m h(m)\exp[-h(m)]\nu(m)}{\sum_m \exp[-h(m)]\nu(m)} \simeq \inf\limits_m h(m) - \log[\nu(m)]$, $\nu \in \mathcal{M}^1_+(M)$, 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 $\log\big(\frac{\lambda}{d_m}\big)$ 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

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

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

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

Integrating with respect to μ equation (1.12, page 20), written in each sub-model Θ_m using the prior distribution $\pi(m, \cdot)$, we see that

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

This proves that

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

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 $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

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

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

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

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

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

as well as $B_1(\nu,\rho) \le -2\log\left\{\mu\left[\exp\left(-\frac{(\lambda-\beta)}{2}r^\star - \frac{1}{2}\log\big(\frac{\lambda}{\beta}\big)d_e\right)\right]\right\}$
$$+ 2\mathcal{K}\big[\nu,\mu_{\left(\frac{\pi[\exp(-\lambda r)]}{\pi[\exp(-\beta r)]}\right)^{1/2}}\big] + \nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\lambda r)}\big] - 2\log(\epsilon).$$

Thus, for any real constants α *and* γ *such that* $0 \le \gamma < \alpha < 1$*, 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*

$$B_2(\nu,\rho) = -\frac{\log\big[(1-\alpha)(1+\gamma)\big]}{\alpha-\gamma}\nu\rho(r) + \frac{2\mathcal{K}(\nu,\mu)+\nu\big\{\mathcal{K}\big[\rho,\pi_{(1+\gamma)^{-Nr}}\big]\big\}-2\log(\epsilon)}{N(\alpha-\gamma)}$$
$$= \frac{1}{N(\alpha-\gamma)}\left\{2\mathcal{K}\left[\nu,\mu_{\left(\frac{\pi[(1-\alpha)^{Nr}]}{\pi[(1+\gamma)^{-Nr}]}\right)^{1/2}}\right] + \nu\Big\{\mathcal{K}\big[\rho,\pi_{(1-\alpha)^{Nr}}\big]\Big\}\right\}$$
$$- \frac{2}{N(\alpha-\gamma)}\log\left\{\mu\left[\exp\left[-\frac{1}{2}\int_{N\log(1+\gamma)}^{-N\log(1-\alpha)} \pi_{\exp(-\xi r)}(\cdot,r)d\xi\right]\right]\right\}$$
$$- \frac{2\log(\epsilon)}{N(\alpha-\gamma)}$$

satisfies

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

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 24) can also be written as

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

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

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

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

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

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 $1-\epsilon$,*

$$\begin{aligned}\lambda\Phi_{\frac{\lambda}{N}}\big[R(\widehat{m},\widehat{\theta}\,)\big]-\beta\Phi_{-\frac{\beta}{N}}\big[R(\widehat{m},\widehat{\theta}\,)\big]&\leq(\lambda-\beta)r(\widehat{m},\widehat{\theta})\\ &+2\log\big[\tfrac{d\nu}{d\mu}(\widehat{m}\,)\big]+\log\big[\tfrac{d\rho}{d\pi_{\exp(-\beta r)}}(\widehat{m},\widehat{\theta}\,)\big]-2\log(\epsilon)\\ &=\int_\beta^\lambda\pi_{\exp(-\alpha r)}(r)d\alpha\\ &+2\log\big[\tfrac{d\nu}{d\mu}(\widehat{m})\big]+\log\big[\tfrac{d\rho}{d\pi_{\exp(-\lambda r)}}(\widehat{m},\widehat{\theta}\,)\big]-2\log(\epsilon)\\ &=2\log\Big\{\mu\Big[\exp\Big(-\frac{1}{2}\int_\beta^\lambda\pi_{\exp(-\alpha r)}(r)d\alpha\Big)\Big]\Big\}\\ &+2\log\big[\tfrac{d\nu}{d\mu\left(\frac{\pi[\exp(-\lambda r)]}{\pi[\exp(-\beta r)]}\right)^{1/2}}(\widehat{m})\big]+\log\big[\tfrac{d\rho}{d\pi_{\exp(-\lambda r)}}(\widehat{m},\widehat{\theta}\,)\big]-2\log(\epsilon).\end{aligned}$$

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

$$\begin{aligned}&R(\widehat{m},\widehat{\theta})\\ &\quad\leq\frac{\alpha-\gamma}{2\alpha\gamma}\Bigg(\sqrt{1+\frac{4\alpha\gamma}{(\alpha-\gamma)^2}\Big\{1-\exp\big[-(\alpha-\gamma)B(\widehat{m},\widehat{\theta})\big]\Big\}}-1\Bigg)\\ &\quad\leq B(\widehat{m},\widehat{\theta}),\end{aligned}$$

where

$$\begin{aligned}B(\widehat{m},\widehat{\theta})&=-\frac{\log\big[(1-\alpha)(1+\gamma)\big]}{\alpha-\gamma}r(\widehat{m},\widehat{\theta})\\ &\qquad+\frac{2\log\Big[\frac{d\nu}{d\mu}(\widehat{m})\Big]+\log\Big[\frac{d\rho}{d\pi_{(1+\gamma)-Nr}}(\widehat{m},\widehat{\theta})\Big]-2\log(\epsilon)}{N(\alpha-\gamma)}\\ &=\frac{2}{N(\alpha-\gamma)}\log\Bigg[\frac{d\nu}{d\mu\left(\frac{\pi[(1-\alpha)^{Nr}]}{\pi[(1+\gamma)^{-Nr}]}\right)^{1/2}}(\widehat{m})\Bigg]\\ &\qquad+\frac{\log\Big[\frac{d\rho}{d\pi_{(1-\alpha)^{Nr}}}(\widehat{m},\widehat{\theta})\Big]-2\log(\epsilon)}{N(\alpha-\gamma)}\end{aligned}$$

$$+\frac{2}{N(\alpha-\gamma)}\log\biggl\{\mu\Bigl[\exp\Bigl(-\frac{1}{2}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)d\alpha\Bigr)\Bigr]\biggr\}.$$

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):

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

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 24) is such that

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

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

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

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

$$\nu\Bigl[\int_\beta^\lambda\pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\bigl(\Phi_{-\frac{\eta}{N}}\circ R\bigr)d\alpha\Bigr]$$

in order to handle $\mathcal{K}(\nu,\overline{\mu})$. In the spirit of equation (1.9, page 18), starting back from Theorem 1.1.4 (page 4), 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 $\rho:\Omega\times M\to\mathcal{M}_+^1(\Theta)$,

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

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

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

we obtain that $f(x)\ge 0$, $x\in\mathbb{R}$, 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

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

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

$$\begin{aligned}\beta\nu\rho(r)+\nu\big[\mathcal{K}(\rho,\pi)\big]&=\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]\big\}-\nu\Big[\log\Big\{\pi\big[\exp(-\beta r)\big]\Big\}\Big]\\&\le\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]\big\}+\beta\nu\overline{\pi}(r)+\nu\big[\mathcal{K}(\overline{\pi},\pi)\big]\\&\le\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]\big\}+\beta\nu\overline{\pi}\big(\Phi_{-\frac{\eta}{N}}\circ R\big)\\&\qquad+\tfrac{\beta}{\eta}\big[\mathcal{K}(\nu,\overline{\mu})-\log(\epsilon)\big]+\nu\big[\mathcal{K}(\overline{\pi},\pi)\big]\\&=\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]\big\}-\nu\Big\{\log\Big[\pi\big[\exp\big(-\beta\Phi_{-\frac{\eta}{N}}\circ R\big)\big]\Big]\Big\}\\&\qquad+\tfrac{\beta}{\eta}\big[\mathcal{K}(\nu,\overline{\mu})-\log(\epsilon)\big].\end{aligned}$$

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

$$0\le(\lambda-\beta)\nu\rho(r)+\nu\big\{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]\big\}-\nu\Big[\int_\beta^\gamma\pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\big(\Phi_{-\frac{\eta}{N}}\circ R\big)d\alpha\Big]+(1+\tfrac{\beta}{\eta})\big[\mathcal{K}(\nu,\overline{\mu})+\log(\tfrac{2}{\epsilon})\big].$$

Noticing moreover that

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

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 $\gamma<\eta\big[\exp(\frac{\eta}{N})-1\big]^{-1}$, defining λ by condition (1.16), so that*

$\lambda = -N \log\Big\{1 - \frac{\gamma}{\eta}\big[\exp(\frac{\eta}{N}) - 1\big]\Big\}$, *with* $\mathbb{P}$ *probability at least* $1-\epsilon$, *for any posterior distribution* $\nu : \Omega \to \mathcal{M}_+^1(M)$, *any conditional posterior distribution* $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

$$\nu\bigg[\int_\beta^\gamma \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}} \circ R)}\big(\Phi_{-\frac{\eta}{N}} \circ R\big) d\alpha\bigg] \leq \nu\bigg[\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha\bigg] + \big(1 + \tfrac{\beta}{\eta}\big)\big[\mathcal{K}(\nu, \overline{\mu}) + \log\big(\tfrac{2}{\epsilon}\big)\big].$$

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* $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

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

We can then remember that

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

to conclude that, putting

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

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

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

$$+\xi\big(2+\tfrac{\beta+\gamma}{\eta}\big)\log\big(\tfrac{3}{\epsilon}\big)$$

$$= \mathcal{K}(\nu,\widehat{\nu}) + \xi\nu\bigg[\Big(\int_\beta^{G_\eta(\beta)} + \int_\gamma^{G_\eta(\gamma)}\Big)\pi_{\exp(-\alpha r)}(r)d\alpha\bigg] + \xi\big(2+\tfrac{\beta+\gamma}{\eta}\big)\log\big(\tfrac{3}{\epsilon}\big).$$

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 $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

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

where the local prior $\overline{\mu}$ is defined by equation (1.14, page 27) and the local posterior $\widehat{\nu}$ and the function G_η are defined by equation (1.18, page 29).

We can then use this theorem to give a local version of Theorem 1.3.10 (page 24). 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 $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

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

$$\begin{aligned}
\textit{where } B_3(\nu,\rho) &= \nu\bigg[\int_{G_\eta^{-1}(\beta)}^{G_\eta(\lambda)} \pi_{\exp(-\alpha r)}(r)d\alpha\bigg] \\
&\quad + \Big(3+\tfrac{G_\eta^{-1}(\beta)}{\eta}\Big)\mathcal{K}\Big[\nu,\mu_{\exp\left[-\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)d\alpha\right]}\Big] \\
&\quad + \nu\big\{\mathcal{K}\big(\rho,\pi_{\exp(-\lambda r)}\big]\big\} + \Big(4+\frac{G_\eta^{-1}(\beta)+\lambda}{\eta}\Big)\log\big(\tfrac{4}{\epsilon}\big) \\
&\le \nu\Big[\big[G_\eta(\lambda) - G_\eta^{-1}(\beta)\big]r^\star + \log\Big(\frac{G_\eta(\lambda)}{G_\eta^{-1}(\beta)}\Big)d_e\Big]
\end{aligned}$$

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

and where the function G_η is defined by equation (1.17, page 29).

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 24), 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

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

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,

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

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

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

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 24) 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 30) 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 $\nu : \Omega \to \mathcal{M}_+^1(M)$,

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

this inequality being obtained by taking a supremum in ρ in Theorem 1.3.15 (page 30). One can then take a supremum in ν, to get, still with $\mathbb{P}$ probability at least $1-\epsilon$,

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

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 $\rho : \Omega\times M \to \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}
&\mu_{\exp\left[-\left(3+\frac{G_\eta^{-1}(\beta)}{\eta}\right)^{-1}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r)d\alpha\right]}\rho\Biggl\{\\
&\quad\exp\left[\Bigl(3+\tfrac{G_\eta^{-1}(\beta)}{\eta}\Bigr)^{-1}\left(\lambda\Phi_{\frac{\lambda}{N}}(R)-\beta\Phi_{-\frac{\beta}{N}}(R)-\int_{G_\eta^{-1}(\beta)}^{G_\eta(\lambda)}\pi_{\exp(-\alpha r)}(r)d\alpha\right.\right.\\
&\qquad\left.\left.+\log\left[\frac{d\rho}{d\pi_{\exp(-\lambda r)}}(\widehat{m},\widehat{\theta})\right]\right)\right]\Biggr\}
\end{aligned}$$

$$\leq \exp\left(\frac{4+\frac{G_\eta^{-1}(\beta)+\lambda}{\eta}}{3+\frac{G_\eta^{-1}(\beta)}{\eta}}\log\left(\tfrac{4}{\epsilon}\right)\right).$$

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 $1-\xi$,*

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

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

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

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

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

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 34) 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 34) 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 7). 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 $\pi\in\mathcal{M}_+^1(\Theta)$,*

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

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 $\lambda\Psi_{\frac{\lambda}{N}}$,

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

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

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

We may now follow the same route as previously, writing

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

Let us also introduce the random pseudo distance

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

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

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

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

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

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 $\pi \in \mathcal{M}_+^1(\Theta)$,*

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

1.4.2. NON RANDOM BOUNDS. Let us first deduce a non-random bound from Theorem 1.4.2 (page 35). 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 35) and using the convexity of the exponential function, we see that for any posterior probability distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

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

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

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

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 $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

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

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

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

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

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 $0 \le \gamma < N\sinh(\frac{\lambda}{N})\big[1-x\tanh(\frac{\lambda}{2N})\big]$,*

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

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 $R'(\theta,\widetilde{\theta}\,) \ge 0$, $\theta\in\Theta$ and that $\varphi(x)\ge 0$, $x\in\mathbb{R}_+$, 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 $x \ge 0$ 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

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

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

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

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 $\kappa > 1$,

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

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

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 37) instead of using the weaker Corollary 1.4.6 (page 38).

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 36) 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:

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

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 $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

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

Taking $\beta=\frac{N}{2}\sinh(\frac{\lambda}{N})$, using the fact that $\sinh(a)\ge a$, $a\ge 0$ 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 $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

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

This theorem and its corollary are really analogous to Theorem 1.4.4 (page 37), 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 40). 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 $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

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

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

$$\mathbb{P}\big[\rho(R)\big] - \inf_{\Theta_1} R \le \mathbb{P}\Bigg\{ \underbrace{\Big[\tfrac{N}{\beta}\log\Big(\sqrt{1+\tfrac{4\beta^2}{N^2}} + \tfrac{2\beta}{N}\Big) - 1\Big]}_{\le 1}\big[\rho(r) - r(\widehat{\theta})\big]$$

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

Note that we could also use the upper bound $m'(\theta,\widehat{\theta}) \leq x\big[r(\theta)-r(\widehat{\theta})\big]+\overline{\varphi}(x)$ 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 $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\big[\rho(R)\big]-\inf_{\Theta_1} R \leq \mathbb{P}\Bigg\{\Bigg[1-\frac{N\sinh(\frac{\lambda}{N})\big[1-x\tanh(\frac{\lambda}{2N})\big]-\lambda}{N\sinh(\frac{\lambda}{N})\big[1-x\tanh(\frac{\lambda}{2N})\big]-\alpha}\Bigg]\big[\rho(r)-r(\widehat{\theta})\big] + \frac{\mathcal{K}\big[\rho,\pi_{\exp(-\alpha r)}\big]}{N\sinh(\frac{\lambda}{N})\big[1-x\tanh(\frac{\lambda}{2N})\big]-\alpha} + \frac{N\sinh(\frac{\lambda}{N})\tanh(\frac{\lambda}{2N})}{N\sinh(\frac{\lambda}{N})\big[1-x\tanh(\frac{\lambda}{2N})\big]-\alpha} \times\Bigg[\overline{\varphi}(x)+\widetilde{\varphi}\Bigg(\frac{\lambda-\alpha}{N\sinh(\frac{\lambda}{N})\tanh(\frac{\lambda}{2N})}\Bigg)\Bigg]\Bigg\}.$$

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 38), 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

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

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 37) 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

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\biggl\{-N\rho\Bigl\{\log\Bigl[1-\lambda P(\psi)\Bigr]\Bigr\} + N\rho\Bigl\{\overline{P}\Bigl[\log(1-\lambda\psi)\Bigr]\Bigr\} - \mathcal{K}(\rho,\pi)\biggr\}\biggr]\biggr\} \leq 1. \tag{1.25}$$

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 $\psi(\theta,\widetilde{\theta})(x,y) = \mathbb{1}\bigl[y \neq f_\theta(x)\bigr] - \mathbb{1}\bigl[y \neq f_{\widetilde{\theta}}(x)\bigr]$, $(x,y)\in\mathcal{X}\times\mathcal{Y}$ so that it should be understood that

$$P(\psi) = \frac{1}{N}\sum_{i=1}^N \mathbb{P}\bigl[\psi_i(\theta,\widetilde{\theta})\bigr] = \frac{1}{N}\sum_{i=1}^N \mathbb{P}\Bigl\{\mathbb{1}\bigl[Y_i \neq f_\theta(X_i)\bigr] - \mathbb{1}\bigl[Y_i \neq f_{\widetilde{\theta}}(X_i)\bigr]\Bigr\} = R'(\theta,\widetilde{\theta}).$$

In the same way

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

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

$$\rho\Bigl\{\log\Bigl[1-\lambda P(\psi)\Bigr]\Bigr\} = \int_{\theta\in\Theta}\log\Bigl\{1-\frac{\lambda}{N}\sum_{i=1}^N \mathbb{P}\bigl[\psi_i(\theta,\widetilde{\theta})\bigr]\Bigr\}\rho(d\theta),$$

$$\rho\Bigl\{\overline{P}\Bigl[\log(1-\lambda\psi)\Bigr]\Bigr\} = \int_{\theta\in\Theta}\Bigl\{\frac{1}{N}\sum_{i=1}^N \log\bigl[1-\lambda\psi_i(\theta,\widetilde{\theta})\bigr]\Bigr\}\rho(d\theta).$$

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 44). 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

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

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

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

Thus

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$- \frac{1}{2} \log\big[(1-\lambda^2)(1-\beta^2)\big] m'(\theta, \widehat{\theta}\,) - \frac{2}{N}\log(\epsilon).$$

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,

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

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*

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

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

the following bounds hold true:

$$\begin{aligned}\rho(R) - \inf_{\Theta_1} R &\\ &\leq \frac{\lambda-\beta}{2\lambda\beta}\left(\sqrt{1+\frac{4\lambda\beta}{(\lambda-\beta)^2}\Bigl\{1-\exp\bigl[-(\lambda-\beta)B(\rho)\bigr]\Bigr\}}-1\right)\\ &\leq B(\rho).\end{aligned}$$

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 $\rho(R) - \inf_{\Theta_1} R \leq 0.096$.