The Annals of Statistics
2020, Vol. 48, No. 6, 3699–3720
https://doi.org/10.1214/20-AOS1948

ROBUST BAYES-LIKE ESTIMATION: RHO-BAYES ESTIMATION

BY YANNICK BARAUD[1] AND LUCIEN BIRGÉ[2]

[1]*Department of Mathematics, University of Luxembourg, yannick.baraud@uni.lu*
[2]*CNRS, Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, lucien.birge@upmc.fr*

We observe n independent random variables with joint distribution $\mathbf{P}$ and pretend that they are i.i.d. with some common density s (with respect to a known measure μ) that we wish to estimate. We consider a density model $\overline{S}$ for s that we endow with a prior distribution π (with support in $\overline{S}$) and build a robust alternative to the classical Bayes posterior distribution which possesses similar concentration properties around s whenever the data are truly i.i.d. and their density s belongs to the model $\overline{S}$. Furthermore, in this case, the Hellinger distance between the classical and the robust posterior distributions tends to 0, as the number of observations tends to infinity, under suitable assumptions on the model and the prior. However, unlike what happens with the classical Bayes posterior distribution, we show that the concentration properties of this new posterior distribution are still preserved when the model is misspecified or when the data are not i.i.d. but the marginal densities of their joint distribution are close enough in Hellinger distance to the model $\overline{S}$.

1. Introduction. The purpose of this paper is to define and study a robust substitute to the classical posterior distribution in the Bayesian framework. It is known that the posterior is not robust with respect to misspecifications of the model. More precisely, if the true distribution P of an n-sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ does not belong to the support $\mathscr{P}$ of the prior and even if it is close to this support in total variation or Hellinger distance, the posterior may concentrate around a point of this support which is quite far from the truth. A simple example is the following one.

Let P_t be the uniform distribution on $[0, t]$ with $t \in \overline{S} = (0, +\infty)$ and, given $a > 0$ and $\alpha > 1$, let π be the prior with density $Ct^{-\alpha}\mathbb{1}_{[a,+\infty)}(t)$, $C = (\alpha - 1)^{-1}a^{1-\alpha}$, with respect to the Lebesgue measure on $\mathbb{R}_+$. Given a n-sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ with distribution P_{t_0}, the posterior distribution function writes as

$$t \mapsto G^L(t|\boldsymbol{X}) = \left[1 - \left(\frac{a \vee X_{(n)}}{t}\right)^{n+\alpha-1}\right]\mathbb{1}_{[a\vee X_{(n)},+\infty)}(t) \tag{1}$$

and, for $t_0 > a$, we see that this posterior is highly concentrated on intervals of the form $[a \vee X_{(n)}, (1 + cn^{-1})(a \vee X_{(n)})]$ with $c > 0$ large enough. Now assume that the true distribution has been contaminated and is rather

$$P = (1 - n^{-1})\mathcal{U}([0, t_0]) + n^{-1}\mathcal{U}([t_0 + 100, t_0 + 100 + n^{-1}]).$$

Although it is quite close to the initial distribution P_{t_0} in variation distance (their distance is $1/n$), on an event of probability $1 - (1 - n^{-1})^n > 1/2$, $t_0 + 100 < X_{(n)} < t_0 + 100 + n^{-1}$ and the posterior distribution is therefore concentrated around $t_0 + 100$ according to (1). The same problem would occur if we were using the maximum likelihood estimator (MLE for short) as an estimator of t.

Received November 2017; revised October 2019.
MSC2020 subject classifications. 62G35, 62F15, 62G05, 62G07, 62C20, 62F99.
Key words and phrases. Bayesian estimation, rho-Bayes estimation, robust estimation, density estimation, statistical models, metric dimension, VC-classes.

In the literature, most results about the behaviour of the posterior do not say anything about misspecification. Some papers like Kleijn and van der Vaart (2006, 2012) and Panov and Spokoiny (2015) address this problem but their results involve the behaviour of the Kullback–Leibler divergence between P and the distributions in $\mathscr{P}$, as is also often the case when studying the MLE; see, for instance, Massart (2007). However, two distributions may be very close in Hellinger distance and, therefore, indistinguishable with our sample $\boldsymbol{X}$, but have a large Kullback–Leibler divergence.

Even when the model is exact, the Kullback divergence is used to analyze the properties of the Bayes posterior. It is known mainly from the work of van der Vaart and co-authors (see in particular Ghosal, Ghosh and van der Vaart (2000)) that the posterior distribution concentrates around $P \in \mathscr{P}$ as n goes to infinity but those general results require that the prior puts enough mass on neighbourhoods of $P \in \mathscr{P}$ of the form $\mathcal{K}(P,\varepsilon) = \{P' \in \mathscr{P}, K(P,P') < \varepsilon\}$ where ε is a positive number and $K(P,P')$ the Kullback–Leibler divergence between P and P'. Unfortunately, such neighbourhoods may be empty (and consequently the condition unsatisfied) when the probabilities in $\mathscr{P}$ are not equivalent, which is, for example, the case for the translation model of the uniform distribution on $[0,1]$, even though the Bayes method may work well in such cases.

As already mentioned, the lack of robustness is not specific to the Bayesian framework but has also been noticed for the MLE. Alternatives to the MLE that remedy this lack of robustness have been considered many years ago by Le Cam (1973, 1975, 1986) and Birgé (1983, 1984, 2006b) but have some limitations. A new recent approach leading to what we called *ρ-estimators* and described in Baraud, Birgé and Sart (2017) (hereafter BBS for short), and Baraud and Birgé (2018) (hereafter BB) corrects a large part of these limitations. It also improves over the previous constructions since it recovers some of the nice properties of the MLE, like efficiency, under suitably strong regularity assumptions.

The aim of this paper is to extend the theory developed in BBS and BB to a Bayesian paradigm in view of designing a robust substitute to the classical Bayes posterior distribution. To be somewhat more precise, let us consider a classical Bayesian framework of density estimation from n i.i.d. observations, although other situations could be considered as well. We observe $\boldsymbol{X} = (X_1,\ldots,X_n)$ where the X_i belong to some measurable space $(\mathscr{X},\mathscr{A})$ with an unknown distribution P on $\mathscr{X}$. We have at disposal a family $\mathscr{P} = \{P_t, t \in \overline{S}\}$ of possible distributions on $\mathscr{X}$, which is dominated by a σ-finite measure μ with respective densities $f(x|t) = (dP_t/d\mu)(x)$. We set $f(\boldsymbol{X}|t) = \prod_{i=1}^n f(X_i|t)$ for the likelihood of t. Assuming that $\overline{S}$ is a measurable space endowed with a σ-algebra $\mathscr{S}$, we choose a prior distribution π on $\overline{S}$, which leads to a posterior $\pi_{\boldsymbol{X}}^L$ that is absolutely continuous with respect to π with density $g^L(t|\boldsymbol{X}) = (d\pi_{\boldsymbol{X}}^L/d\pi)(t)$. Following this notation, the log-likelihood function and log-likelihood ratios write respectively as $L(\boldsymbol{X}|t) = \log(f(\boldsymbol{X}|t)) = \sum_{i=1}^n \log(f(X_i|t))$ and $\mathbf{L}(\boldsymbol{X},t,t') = L(\boldsymbol{X}|t') - L(\boldsymbol{X}|t)$ so that the density $g^L(t|\boldsymbol{X})$ of the posterior distribution $\pi_{\boldsymbol{X}}^L$ with respect to π is given by

$$\frac{\exp[L(\boldsymbol{X}|t)]}{\int_{\overline{S}}\exp[L(\boldsymbol{X}|t)]\,d\pi(t)} = \frac{\exp[L(\boldsymbol{X}|t) - \sup_{t'\in\overline{S}} L(\boldsymbol{X}|t')]}{\int_{\overline{S}}\exp[L(\boldsymbol{X}|t) - \sup_{t'\in\overline{S}} L(\boldsymbol{X}|t')]\,d\pi(t)}$$

and consequently,

$$g^L(t|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|t)}{\int_{\overline{S}} f(\boldsymbol{X}|t)\,d\pi(t)} = \frac{\exp[-\sup_{t'\in\overline{S}}\mathbf{L}(\boldsymbol{X},t,t')]}{\int_{\overline{S}}\exp[-\sup_{t'\in\overline{S}}\mathbf{L}(\boldsymbol{X},t,t')]\,d\pi(t)}. \tag{2}$$

Note that, if the MLE $\widehat{t}(\boldsymbol{X})$ exists,

$$\sup_{t'\in\overline{S}}\mathbf{L}\big(\boldsymbol{X},t,t'\big) = L\big(\boldsymbol{X}|\widehat{t}(\boldsymbol{X})\big) - L(\boldsymbol{X}|t)$$

and that we could as well consider, for all $\beta > 0$ the distributions

$$g_\beta^L(t|\boldsymbol{X}) \cdot \pi \quad \text{with } g_\beta^L(t|\boldsymbol{X}) = \frac{\exp[\beta L(\boldsymbol{X}|t)]}{\int_{\overline{S}} \exp[\beta L(\boldsymbol{X}|t)]\, d\pi(t)}.$$

The posterior corresponds to $\beta = 1$ and when β goes to infinity the distribution $g_\beta^L(t|\boldsymbol{X}) \cdot \pi$ converges weakly, under mild assumptions, to the Dirac measure located at the MLE. All values of $\beta \in (1, +\infty)$ will then lead to interpolations between the posterior and the Dirac at the MLE.

Most problems connected with the maximum likelihood or Bayes estimators are due to the fact that the log-likelihood ratios $\mathbf{L}(\boldsymbol{X}, t, t')$ involve the logarithmic function which is unbounded. As a result, we may have

$$\mathbb{E}_t\big[\mathbf{L}(\boldsymbol{X}, t, t')\big] = -n\mathbb{E}_t\big[\log(dP_t/dP_{t'})(X_1)\big] = -\infty,$$

the situation being even more delicate when the true distribution of the X_i is different (even slightly) from P_t.

In BBS and BB, we offered an alternative to the MLE by replacing the logarithmic function in the log-likehood ratios by other ones. One possibility being the function $\varphi(x)$ defined by

$$\varphi(x) = 4\frac{\sqrt{x} - 1}{\sqrt{x} + 1} \quad \text{for all } x \geq 0,$$

so that, for $x > 0$,

$$\varphi'(x) = \frac{4}{(1 + \sqrt{x})^2\sqrt{x}} > 0 \quad \text{and} \quad \varphi''(x) = -\frac{2(1 + 3\sqrt{x})}{(1 + \sqrt{x})^3 x^{3/2}} < 0.$$

Like the log function, $\varphi(x)$ is increasing, concave and satisfies $\varphi(1/x) = -\varphi(x)$. In fact, these two functions coincide at $x = 1$, their first and second derivatives as well and for all $x \in [1/2, 2]$

$$0.99 < \frac{\varphi(x)}{\log x} \leq 1 \quad \text{and} \quad \big|\varphi(x) - \log x\big| \leq 0.055|x - 1|^3. \tag{3}$$

The main advantage of the function φ as compared to the log function lies in its boundedness. It can also be extended to $[0, +\infty]$ by continuity by setting $\varphi(+\infty) = 4$. As a consequence, the quantity $\varphi(t'(X)/t(X))$ is well-defined (with the convention $a/0 = +\infty$ for $a > 0$ and $0/0 = 1$) and bounded and we can use it as a surrogate for $\log(t'(X)/t(X))$. This suggests the replacement of $\mathbf{L}(\boldsymbol{X}, t, t')$ by $4\boldsymbol{\Psi}(\boldsymbol{X}, t, t')$ where the function $\boldsymbol{\Psi}$ is defined as

$$\boldsymbol{\Psi}(\mathbf{x}, t, t') = \sum_{i=1}^n \psi\left(\sqrt{\frac{t'(x_i)}{t(x_i)}}\right) \quad \text{for all } \mathbf{x} \in \mathscr{X}^n \text{ and } (t, t') \in \overline{S}^2, \tag{4}$$

with the conventions $0/0 = 1$, $a/0 = +\infty$ for $a > 0$ and

$$\psi(x) = \begin{cases} \dfrac{x - 1}{x + 1} & \text{for } 0 \leq x < +\infty, \\ 1 & \text{for } x = +\infty, \end{cases} \tag{5}$$

so that $\varphi(x) = 4\psi(\sqrt{x})$. Note that ψ is Lipschitz with Lipschitz constant 2. The important point here is that we have already studied in details in BB the behaviour and properties of a process which is closely related to $(t, t') \mapsto \boldsymbol{\Psi}(\boldsymbol{X}, t, t')$.

We get a pseudo-posterior density with respect to π by replacing in (2) the quantity $\sup_{t' \in \overline{S}} \mathbf{L}(\boldsymbol{X}, t, t')$ by $4 \sup_{t' \in \overline{S}} \boldsymbol{\Psi}(\boldsymbol{X}, t, t')$. This pseudo-posterior density can therefore be written

$$g(t|\boldsymbol{X}) = \frac{\exp[-4 \sup_{t' \in \overline{S}} \boldsymbol{\Psi}(\boldsymbol{X}, t, t')]}{\int_{\overline{S}} \exp[-4 \sup_{t' \in \overline{S}} \boldsymbol{\Psi}(\boldsymbol{X}, t, t')]\, d\pi(t)}.$$

More generally, we may consider, for $\beta > 0$, the random distribution $\pi_{\boldsymbol{X}}$ given by

$$\frac{d\pi_{\boldsymbol{X}}}{d\pi}(t) = \frac{\exp[-\beta \sup_{t' \in \overline{S}} \Psi(\boldsymbol{X}, t, t')]}{\int_{\overline{S}} \exp[-\beta \sup_{t' \in \overline{S}} \Psi(\boldsymbol{X}, t, t')]\, d\pi(t)}. \tag{6}$$

This will be the starting point for our study of this *Bayes-like* framework with a *posterior-like* distribution $\pi_{\boldsymbol{X}}$ defined by (6) that will play a similar role as the posterior distribution in the classical Bayesian paradigm except for the fact that a random variable with distribution $\pi_{\boldsymbol{X}}$ (conditionally to our sample $\boldsymbol{X}$) will possess robustness properties with respect to the hypothesis that P belongs to $\mathscr{P}$. We shall call it *ρ-posterior* by analogy with our construction of ρ-estimators as described in BBS and BB.

To conclude this Introduction, let us emphasize the specific properties of our method that distinguish it from classical Bayesian procedures.

- Contrary to the classical Bayesian framework, concentration properties of the ρ-Bayes method do not involve the Kullback–Leibler divergence but only the Hellinger distance.
- Our results are nonasymptotic and given in the form of large deviations of the pseudo-posterior distribution from the true density for a given value n of the number of observations.
- The method is robust to Hellinger deviations: even if the true distribution is at some positive Hellinger distance of the support of the prior, the posterior will behave almost as well as if this were not the case provided that this distance is small.
- Due to the just mentioned robustness properties, we may work with an approximate model for the true density. In particular, when the density is assumed to belong to a nonparametric set $\mathcal{S}$, it is actually enough to apply our ρ-Bayes procedure on a parametric set $\overline{S}$ possessing good approximation properties with respect to the elements of $\mathcal{S}$. Besides, starting from a continuous prior on a continuous model, we can discretize both of them without losing much provided that our discretization scale is small enough.
- The ρ-posterior also possesses robustness properties with respect to the assumption that the data are i.i.d. provided that the densities of the X_i are close enough to the model $\overline{S}$.

Substituting another function to the log-likelihood in the expression of the posterior distribution, as we do here, is not new in the literature. It has often been motivated by the will of replacing the Kullback–Leibler loss, which is naturally associated to the likelihood-function, by other losses that are more specifically associated to the problem that needs to be solved (estimation of a mean, classification, etc.) or to deal with the problem of misspecification. This approach leads to *quasi-posterior distributions* which properties have been studied by many authors among which Chernozhukov and Hong (2003) and Bissiri *et al.* (2016) (see also the references therein). These results do not include robustness but Chernozhukov and Hong (2003) proved some analogues of the Bernstein–von Mises theorem under suitable assumptions on the model and loss function. The use of *fractional likelihoods* by Jiang and Tanner (2008) was motivated by the problem of misspecification. In a sparse parametric framework (the true parameter $\theta \in \mathbb{R}^d$ has a small number of nonzero components), Atchadé (2017) replaces the joint density $f_{n,\theta}$ of the observations by a suitable function $q_{n,\theta}$. Together with a prior that forces sparsity, this results in tractable and consistent procedures for high-dimensional parametric problems. All the cited results are of an asymptotic nature contrary to the next one. Bhattacharya, Pati and Yang (2019) investigate the replacement, in the definition of the posterior, of the likelihood by a fractional one, also considering the case of misspecified models, but use what they call α-divergences instead of the KL one (but which may also be infinite) to evaluate the amount of misspecification.

Closer to our approach is the PAC-Bayesian one that has been developed by Olivier Catoni (2007) and our parameter β in the definition of the ρ-posterior (6) refers to the (inverse) of

the so-called temperature parameter in the definition of the Gibbs measure. This parameter essentially plays no role in our results.

The paper is organized as follows. In Section 2, we describe our framework and state our main assumption that allows to solve the measurability issues that are inherent to the construction of the posterior. An account of what can be achieved with a ρ-posterior distribution is presented and commented in Section 3 in the density and regression frameworks (with a random design). Our main result can be found in Section 4 where we present the concentration properties of our ρ-posterior distribution. These properties involve two quantities, one which depends on the choice of the prior while the other is independent of it but depends on the model and the true density. We show how one can control these quantities in Sections 6 and 5, respectively, giving there illustrative examples as well as general theorems that can be applied to many parametric models of interest. Our results on the connection between the classical Bayes posterior and the ρ-one are presented in Section 7. We show that under suitable assumptions on the density model and the prior, the Hellinger distance between these two distributions tends to 0 at rate $n^{-1/4}(\log n)^{3/4}$ as the sample size n tends to infinity. In particular, this result shows that under suitable assumptions our ρ-Bayes posterior satisfies a Bernstein–von Mises theorem. The problem of a hierarchical prior or, equivalently, that of model selection is handled in Section 8. The proofs and discussions about measurability issues are to be found in the Supplementary Material (Baraud and Birgé (2020)) while additional results and examples can be found in the original version of this paper, Baraud and Birgé (2017).

2. Framework, notation and basic assumptions.

2.1. *The framework and the basic notation.* We actually want to deal with more general situations than the one we presented in the Introduction, namely the case of independent but possibly non-i.i.d. observations, even though the statistician assumes them to be i.i.d. By doing so, our aim is to emphasize the robustness property of our ρ-posterior distribution with respect to the assumption that the data are i.i.d. This generalization leads to the following statistical framework. For $n \in \mathbb{N}^\star = \mathbb{N} \setminus \{0\}$, we observe a random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ defined on (Ω, Ξ), where the X_i are independent with values in a measurable space $(\mathscr{X}, \mathscr{A})$ endowed with a σ-finite mesure μ. We denote by $\mathscr{L}$ the set of all probability densities u with respect to μ (which means that u is a nonnegative measurable function on $\mathscr{X}$ such that $\int_{\mathscr{X}} u(x)\,d\mu(x) = 1$) and by $P_u = u \cdot \mu$ the probability on $(\mathscr{X}, \mathscr{A})$ with density $u \in \mathscr{L}$. We assume that for each $i \in \{1, \ldots, n\}$, X_i admits a density with respect to μ, that is, has distribution $P_{s_i} = s_i \cdot \mu$ with $s_i \in \mathscr{L}$. We set $\mathbf{s} = (s_1, \ldots, s_n)$ and denote by $\mathbb{P}_{\mathbf{s}}$ the probability on (Ω, Ξ) that gives $\boldsymbol{X}$ the distribution $\mathbf{P}_{\mathbf{s}} = \bigotimes_{i=1}^n P_{s_i}$ on $\mathscr{X}^n$ and by $\mathbb{E}_{\mathbf{s}}$ the corresponding expectation. We shall abusively refer to $\mathbf{s}$ as the (true) density of $\boldsymbol{X}$.

We denote by $|A|$ the cardinality of a finite set A and use the word *countable* for *finite or countable*. Parametric models will be indexed by some subset $\boldsymbol{\Theta}$ of $\mathbb{R}^d$ and $|\cdot|$ will denote the Euclidean norm on $\mathbb{R}^d$. Finally, we shall often use the inequalities

$$2ab \le \alpha a^2 + \alpha^{-1} b^2 \quad \text{and} \quad (a+b)^2 \le (1+\alpha)a^2 + (1+\alpha^{-1})b^2 \quad \text{for all } \alpha > 0. \tag{7}$$

2.2. *Hellinger type metrics.* For all $t, t' \in \mathscr{L}$, we shall write $h(t, t')$ and $\rho(t, t')$ for the Hellinger distance and affinity between P_t and $P_{t'}$. We recall that the Hellinger distance and affinity between two probabilities P, Q on a measurable space $(\mathscr{X}, \mathscr{A})$ are given respectively by

$$h(P,Q) = \left[\frac{1}{2}\int_{\mathscr{X}} \left(\sqrt{\frac{dP}{d\nu}} - \sqrt{\frac{dQ}{d\nu}}\right)^2 d\nu\right]^{1/2} \quad \text{and} \quad \rho(P,Q) = \int_{\mathscr{X}} \sqrt{\frac{dP}{d\nu}\frac{dQ}{d\nu}}\,d\nu,$$

where ν denotes an arbitrary measure which dominates both P and Q, the result being independent of the choice of ν. It is well known since Le Cam (1973) that $0 \le \rho(P,Q) = 1 - h^2(P,Q)$ and that the Hellinger distance is related to the total variation distance by the following inequalities:

$$h^2(P,Q) \le \sup_{A\in\mathcal{A}} \big|P(A) - Q(A)\big| \le h(P,Q)\sqrt{2 - h^2(P,Q)} \le \sqrt{2}h(P,Q). \tag{8}$$

Therefore, robustness with respect to the Hellinger distance implies robustness with respect to the total variation distance.

The Hellinger closed ball centred at $t \in \mathcal{L}$ with radius $r > 0$ is denoted $\mathcal{B}(t,r)$ and, for $\mathbf{s} \in \mathcal{L}^n$, we define

$$\mathcal{B}(\mathbf{s},r) = \big\{t \in \mathcal{L}, h^2(\mathbf{s},t) \le r^2\big\} \quad \text{with } h^2(\mathbf{s},t) = \frac{1}{n}\sum_{i=1}^n h^2(s_i,t) \le 1.$$

Then, for $S \subset \mathcal{L}$, we set $\mathcal{B}^S(t,r) = S \cap \mathcal{B}(t,r)$ and $\mathcal{B}^S(\mathbf{s},r) = S \cap \mathcal{B}(\mathbf{s},r)$. If the X_i are truly i.i.d. with density s, $\mathbf{s} = (s,\dots,s)$ and $h^2(\mathbf{s},t) = h^2(s,t)$ for all $t \in \mathcal{L}$; hence $\mathcal{B}^S(\mathbf{s},r) = \mathcal{B}^S(s,r)$.

Note that although h is a genuine distance on the space of all probabilities on $\mathcal{X}$, therefore, on $\{P_t, t \in \mathcal{L}\}$, it is only a *pseudo-distance* on $\mathcal{L}$ itself since $h(t,t') = 0$ if $t \ne t'$ but $t = t'$ μ-a.e. For simplicity, we shall nevertheless still call h a distance on $\mathcal{L}$ and set $h(t,A) = \inf_{u\in A} h(t,u)$ for the distance of a point $t \in \mathcal{L}$ to the subset A of $\mathcal{L}$. Similarly, $h(\mathbf{s},A) = \inf_{t\in A} h(\mathbf{s},t)$. We recall that a pseudo-distance d satisfies the axioms of a distance apart from the fact that one may have $d(x,y) = 0$ with $x \ne y$.

2.3. *Models and main assumptions.* We consider a *density model* $\overline{S}$, that is, a subset of $\mathcal{L}$, acting as if the data were i.i.d., and our aim is to estimate the n-uple $\mathbf{s} = (s_1,\dots,s_n)$ from the observation of $\boldsymbol{X}$ on the basis of this model. Adopting the Bayesian paradigm, we endow $\overline{S}$ with a σ-algebra $\mathcal{S}$ as well as a prior π on $(\overline{S},\mathcal{S})$. There is no reason for $t \mapsto \boldsymbol{\Psi}(\boldsymbol{X},t,t')$ defined by (4) and $t \mapsto \sup_{t'\in\overline{S}} \boldsymbol{\Psi}(\boldsymbol{X},t,t')$ to be measurable functions of t on $(\overline{S},\mathcal{S})$ and the function $\omega \mapsto \sup_{t'\in\overline{S}} \boldsymbol{\Psi}(\boldsymbol{X}(\omega),t,t')$ to be a random variable on (Ω,Ξ). Therefore, our ρ-posterior distribution $\pi_{\boldsymbol{X}}$, as given by (6), might not be well-defined. In order to overcome these difficulties, we introduce the following assumption and also slightly modify the definition of our ρ-posterior distribution that was originally given by (6) in the density framework. The following assumption ensures that the sets and random variables that we shall introduce later are suitably measurable. We refer the reader to the Supplementary Material (Baraud and Birgé (2020)) for a discussion about Assumption 1 and how it can be checked on examples.

ASSUMPTION 1.

(i) The function $(x,t) \to t(x)$ on $\mathcal{X} \times \overline{S}$ is measurable with respect to the σ-algebra $\mathcal{A} \otimes \mathcal{S}$.

(ii) There exists a countable subset S of $\overline{S}$ and, given $t \in \overline{S}$ and $t' \in S$, one can find a sequence $(t_k)_{k\ge0}$ in S such that, for all $x \in \mathcal{X}$,

$$\lim_{k\to+\infty} t_k(x) = t(x) \quad \text{and} \quad \lim_{k\to+\infty} \psi\left(\sqrt{\frac{t'(x)}{t_k(x)}}\right) = \psi\left(\sqrt{\frac{t'(x)}{t(x)}}\right). \tag{9}$$

Note that it follows from Proposition 2 in the Supplementary Material (Baraud and Birgé (2020)) that S is dense in $\overline{S}$ with respect to the distance h. Of course, when $\overline{S}$ is countable,

we shall set $S = \overline{S}$ without further notice and Assumption 1(ii) will be automatically satisfied with the σ-algebra $\mathscr{S}$ gathering all the subsets of $\overline{S}$. In the sequel, we shall always assume that the set S associated to the model $\overline{S}$ has been fixed once and for all.

The following proposition (to be proven in the Supplementary Material) ensures that the measurability properties required for a proper definition of the posterior distribution hold.

PROPOSITION 1. *Under Assumption 1, given $t' \in S$ and $\mathbf{\Psi}(\mathbf{x}, t, t')$ defined by (4), the functions*

$$(\mathbf{x}, t) \mapsto \mathbf{\Psi}\big(\mathbf{x}, t, t'\big) \quad \text{and} \quad (\mathbf{x}, t) \mapsto \mathbf{\Psi}(\mathbf{x}, t) = \sup_{u \in S} \mathbf{\Psi}(\mathbf{x}, t, u)$$

are measurable with respect to the σ-algebra $\mathscr{A} \otimes \mathscr{S}$. Hence the function

$$\mathbf{x} \mapsto \int_{\overline{S}} \exp\big[-\beta \mathbf{\Psi}(\mathbf{x}, t)\big]\, d\pi(t)$$

is measurable with respect to $\mathscr{A}$ and the function $t \mapsto h(t, s)$ is measurable with respect to $\mathscr{S}$ whatever $s \in \mathscr{L}$.

2.4. *The ρ-posterior distribution $\pi_{\boldsymbol{X}}$.* Let S be the countable subset of $\overline{S}$ provided by Assumption 1. For $\omega \in \Omega$ and $\beta > 0$, we define the distribution $\pi_{\boldsymbol{X}(\omega)}$ on $\overline{S}$ by its density with respect to the prior π:

$$\frac{d\pi_{\boldsymbol{X}(\omega)}}{d\pi}(t) = g\big(t|\boldsymbol{X}(\omega)\big) = \frac{\exp[-\beta \mathbf{\Psi}(\boldsymbol{X}(\omega), t)]}{\int_{\overline{S}} \exp[-\beta \mathbf{\Psi}(\boldsymbol{X}(\omega), t')]\, d\pi(t')}. \tag{10}$$

Proposition 1 implies that the function $(\omega, t) \mapsto g(t|\boldsymbol{X}(\omega))$ is measurable with respect to the σ-algebra $\Xi \otimes \mathscr{S}$. We recall that the choice of $\beta = 4$ leads to an analogue of the classical Bayes posterior since the function $x \mapsto 4\psi(\sqrt{x})$ is close to $\log x$ as soon as x is not far from one. Throughout the paper, the parameter β will remain fixed and part of our results will depend on it.

DEFINITION 1. The method that leads from the set $\overline{S}$ and the prior π on $\overline{S}$ to the distribution $\pi_{\boldsymbol{X}}$ (and all related estimators) will be called *ρ-Bayes estimation* and $\pi_{\boldsymbol{X}}$ is the *ρ-posterior* distribution.

3. A flavour of what a ρ-Bayes procedure can achieve. Throughout this section, we take $\beta = 4$, the value for which the ρ-posterior distribution is the analogue of the classical Bayes posterior.

3.1. *The density framework.* Let $\overline{S}$ be a density model for the supposed common density of our observations $X_1, \dots, X_n$ and consider the following entropy condition.

ASSUMPTION 2. There exists a nonincreasing function H from $(0, 1]$ to $[3, +\infty)$ such that, for any $\varepsilon \in (0, 1]$, there exists a subset S_ε of $\overline{S}$ with cardinality not larger than $\exp[H(\varepsilon)]$ and such that $h(t, S_\varepsilon) \le \varepsilon$ for all $t \in \overline{S}$.

PROPOSITION 2. *Let $\overline{S}$ satisfy Assumption 2 and ε_n be such that*

$$\varepsilon_n \ge 1/(2\sqrt{n}) \quad \text{and} \quad H(\varepsilon_n) \le \big(4 \cdot 10^{-6}\big) n \varepsilon_n^2. \tag{11}$$

There exists a prior π on $\overline{S}$ (depending on ε_n only) such that, whatever the true density $\mathbf{s} = (s_1, \dots, s_n)$ and $\xi > 0$, there exists a measurable subset Ω_ξ of Ω satisfying $\mathbb{P}_{\mathbf{s}}(\Omega_\xi) \ge 1 - e^{-\xi}$ and for all $\omega \in \Omega_\xi$,

$$\pi_{\boldsymbol{X}(\omega)}\big(\big\{t \in \overline{S}, h(\mathbf{s}, t) \le C\overline{r}_n\big\}\big) \ge 1 - e^{-\xi'} \quad \text{for all } \xi' > 0,$$

with C a positive universal constant and

$$\overline{r}_n = h(\mathbf{s}, \overline{S}) + \varepsilon_n + \sqrt{\frac{\xi + \xi'}{n}}.$$

In particular, if $X_1, \ldots, X_n$ are truly i.i.d. with density $s \in \mathscr{L}$,

$$\pi_{\boldsymbol{X}(\omega)}\big(\mathscr{B}^{\overline{S}}(s, C\overline{r}_n)\big) \geq 1 - e^{-\xi'} \quad \text{with } \overline{r}_n = h(s, \overline{S}) + \varepsilon_n + \sqrt{\frac{\xi + \xi'}{n}}.$$

This result shows that with probability close to 1, the ρ-posterior distribution concentrates around points t in the density model $\overline{S}$ which satisfy

$$\left[\frac{1}{n}\sum_{i=1}^{n} h^2(s_i, t)\right]^{1/2} \leq C\overline{r}_n \quad \text{with } \overline{r}_n \text{ of order } h(\mathbf{s}, \overline{S}) + \varepsilon_n.$$

The quantity ε_n corresponds to the concentration rate we get when the X_i are truly i.i.d. with density in $\overline{S}$. For instance, when $H(\varepsilon) = A\varepsilon^{-V}$ for all $\varepsilon > 0$ and some constants $A, V > 0$, ε_n is of order $n^{-1/(V+2)}$. This concentration rate remains of the same order as long as $h(\mathbf{s}, \overline{S})$ is small enough compared to ε_n, which is actually possible even when none of the densities s_i belongs to $\overline{S}$. This stability result accounts for the robustness property of our procedure.

It is well known (see, for instance, Birgé (1983) and (1986)) that, in many cases, the smallest value of ε_n, which satisfies (11) corresponds to the minimax rate of estimation (with respect to n) over $\overline{S}$. Here are two typical illustrations for densities with respect to the Lebesgue measure:

(i) Assume that $\overline{S}$ is the set of all nonincreasing densities on $[0, 1]$ which are bounded by $M < +\infty$. Of course, if $s \in \overline{S}$, $\sqrt{s}$ is also nonincreasing and is bounded by $\sqrt{M}$ and the Hellinger entropy of $\overline{S}$ corresponds to the $\mathbb{L}_2$-entropy of the set $\{\sqrt{s}, s \in \overline{S}\}$ which is known from van de Geer (2000) to be bounded by $A\varepsilon^{-1}$ leading to an ε_n of order $n^{-1/3}$ which is known to be the minimax rate for this problem.

(ii) If $\overline{S}$ is the set of α-Hölderian densities on $[0, 1]^d$ with $\alpha > 0$, its Hellinger entropy is known from Birgé (1986) to be of order $\varepsilon^{-2d/\alpha}$ leading to a convergence rate with respect to n of order $n^{-\alpha/2(\alpha+d)}$. All details can be found in Birgé (1986) (see in particular his Corollary 3.2) where it is also proved that this rate is minimax (see his Proposition 4.3).

Assumption 2 can actually be replaced by the more general one that $\overline{S}$ admits a metric dimension D (according to Definition 3 below) in which case the same conclusion holds with $\varepsilon_n \geq 1/(2\sqrt{n})$ satisfying $D(\varepsilon_n) \leq 10^{-6} n\varepsilon_n^2$.

3.2. *The regression framework.* We observe i.i.d. pairs $X_i = (W_i, Y_i)$ with values in $\mathscr{W} \times \mathbb{R}$ drawn from the regression model

$$Y_i = f^\star(W_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n.$$

We assume that the regression function $f^\star$ is bounded in supnorm (denoted $\|\cdot\|_\infty$) by some known number $B > 0$, that the W_i are i.i.d. with unknown distribution P_W on $\mathscr{W}$ and the ε_i are i.i.d. with unknown density p with respect to the Lebesgue measure λ on $\mathbb{R}$.

We consider a model $\overline{\mathcal{F}}$ for $f^\star$ which is a set of functions on $\mathscr{W}$ satisfying the following property.

ASSUMPTION 3. For all $f \in \overline{\mathcal{F}}$, $\|f\|_\infty \leq B$ and there exists a nonincreasing function H on $[3, +\infty)$ such that for all $\varepsilon > 0$, one can find a subset $\mathcal{F}_\varepsilon \subset \overline{\mathcal{F}}$ with cardinality not larger than $\exp[H(\varepsilon)]$ which satisfies $\inf_{g \in \mathcal{F}_\varepsilon} \|f - g\|_\infty \leq \varepsilon$ for all $f \in \overline{\mathcal{F}}$.

The density p being unknown, we consider a candidate density q for p. Denoting by q_δ the translated density $q_\delta(\cdot) = q(\cdot - \delta)$ for $\delta \in \mathbb{R}$, we assume that q is of order $\alpha \in (-1, 1]$, that is, satisfies, for some constant $a \geq 1$,

$$a^{-1}\big[|\delta|^{1+\alpha} \wedge a^{-1}\big] \leq h^2(q_\delta, q) \leq a\big[|\delta|^{1+\alpha} \wedge a^{-1}\big] \quad \text{for all } \delta \in \mathbb{R}. \tag{12}$$

Note that the mapping $\delta \mapsto q_\delta$ is one-to-one.

For $f \in \overline{\mathcal{F}}$, we denote by q_f the density of X_1 (with respect to $\mu = P_W \otimes \lambda$) when $p = q$ and $f^\star = f$ which is given by $q_f(w, y) = q(y - f(w))$. The set $\overline{S} = \{q_f,\ f \in \overline{\mathcal{F}}\}$ is a density model for the true density s of X_1. A prior π' on $\overline{\mathcal{F}}$ induces a prior π on $\overline{S}$ by taking the image of π' by the mapping $f \mapsto q_f$. In turn, the ρ-posterior $\pi_{\boldsymbol{X}}$ on $(\overline{S}, \pi)$ induces a ρ-posterior distribution $\pi'_{\boldsymbol{X}}$ on $\overline{\mathcal{F}}$ which is the image of $\pi_{\boldsymbol{X}}$ by the reciprocal mapping $q_f \mapsto f$. Let us choose as our loss function on $\overline{\mathcal{F}}$

$$\|f^\star - f\|_{1+\alpha} = \left(\int_{\mathscr{W}} |f^\star - f|^{1+\alpha}\, dP_W\right)^{1/(1+\alpha)} \quad \text{for } f \in \overline{\mathcal{F}}.$$

PROPOSITION 3. *Let Assumption 3 hold, q satisfy (12) for some $a \geq 1$ and $\alpha \in (-1, 1]$ and let ε_n satisfy*

$$\varepsilon_n \geq 1/(2\sqrt{n}) \quad \text{and} \quad H\big[(\varepsilon_n^2/a)^{1/(1+\alpha)}\big] \leq (4 \cdot 10^{-6}) n \varepsilon_n^2. \tag{13}$$

There exists a prior π' on $\overline{\mathcal{F}}$ (which only depends on ε_n) such that, whatever the function $f^\star$ bounded by B, whatever the distribution P_W, whatever the density p and the positive number ξ, there exists a measurable subset Ω_ξ of Ω satisfying $\mathbb{P}_{\mathbf{s}}(\Omega_\xi) \geq 1 - e^{-\xi}$ and for all $\omega \in \Omega_\xi$,

$$\pi'_{\boldsymbol{X}(\omega)}\big(\big\{f \in \overline{\mathcal{F}}, \|f^\star - f\|_{1+\alpha} \leq C\overline{r}_n^{2/(1+\alpha)}\big\}\big) \geq 1 - e^{-\xi'} \quad \text{for all } \xi' > 0,$$

where

$$\overline{r}_n^2 = h^2(p, q) + \inf_{f \in \overline{\mathcal{F}}} \|f^\star - f\|_\infty^{1+\alpha} + \varepsilon_n^2 + \frac{\xi + \xi'}{n}, \tag{14}$$

for some constant $C > 0$ depending on a, B and α only.

Let us first emphasize the fact that neither the prior nor the construction of the posterior requires the knowledge of the distribution of the design P_W or any assumption about it. The result shows that with probability close to 1 the ρ-posterior on $\overline{\mathcal{F}}$ concentrates around functions $f \in \overline{\mathcal{F}}$ for which $\|f - f^\star\|_{1+\alpha}$ is of order $[h^2(p, q) + \inf_{f \in \overline{\mathcal{F}}} \|f^\star - f\|_\infty^{1+\alpha} + \varepsilon_n^2]^{1/(1+\alpha)}$. The quantity $\varepsilon_n^{2/(1+\alpha)}$ corresponds to the concentration rate we get when p is equal to q and $f^\star$ belongs to $\overline{\mathcal{F}}$ while the terms $\inf_{f \in \overline{\mathcal{F}}} \|f^\star - f\|_\infty^{1+\alpha}$ and $h^2(p, q)$ account for the robustness of the procedure with respect to a misspecification of the class $\overline{\mathcal{F}}$ of the regression functions and the noise distribution, respectively. The loss and the quantity ε_n depend on the specific features of the chosen density q.

When $H(\varepsilon) = A\varepsilon^{-V}$ for some constants $A, V > 0$, then

$$\varepsilon_n^{2/(1+\alpha)} = C' n^{-1/(V+1+\alpha)},$$

where $C' > 0$ depends on A, a, α and V only. We refer to Ibragimov and Has'minskiĭ (1981) Chapter VI, page 281 for sufficient conditions on the density q to be of order α. For illustration, when q is Gaussian, $\alpha = 1$, the loss corresponds to the $\mathbb{L}_2(P_W)$-norm and $\varepsilon_n^{2/(1+\alpha)} = \varepsilon_n$ is of order $n^{-1/(V+2)}$; when q is the uniform density on an interval, $\alpha = 0$, the loss corresponds to the $\mathbb{L}_1(P_W)$-norm and $\varepsilon_n^{2/(1+\alpha)} = \varepsilon_n^2$ is of order $n^{-1/(V+1)}$.

When $\overline{\mathcal{F}}$ is a subset of the $\mathbb{L}_\infty(P_W)$-ball with radius B and center 0 of a linear space with dimension $d \geq 1$, a classical result on the entropy of balls in a finite dimensional linear space

implies that Assumption 3 is satisfied with $H(\varepsilon) = d\log(1+[2B/\varepsilon])$ which leads to an upper bound for $\varepsilon_n^{2/(1+\alpha)}$ of order $[d\log(nB/d)/n]^{1/(1+\alpha)}$. Note that this rate is faster than the usual parametric rate $1/\sqrt{n}$ when $\alpha \in (-1,1)$.

Choosing a specific density q and a single model $\overline{\mathcal{F}}$ for $f^\star$ is usually not enough for many applications. It is however possible to mix up several choices of q and $\overline{\mathcal{F}}$ by using a hierarchical prior as we shall show in Section 8 and by arguing as in BBS, Sections 7.2 and 7.3.

4. Our main results. Our main results and definitions involve some numerical constants that we list below for further reference.

$$\begin{cases} c_0 = 10^3; \quad c_1 = 15; \quad c_2 = 16; \quad c_3 = 0.62; \\ c_4 = 3.5\max\{375;\beta^{-1/2}\}; \quad c_5 = 16\times 10^{-3}; \quad c_6 = 7\times 10^4; \\ c_7 = 4.01; \quad c_8 = 0.365; \quad c_9 = c_8^{-1}[(2c_6)\vee\beta^{-1}]; \\ \overline{c}_n = 1+[(\log 2)/\log(en)]; \quad \gamma = \beta/8. \end{cases} \tag{15}$$

The properties of $\pi_{\boldsymbol{X}}$ actually depend on two quantities, namely $\varepsilon_n^{\overline{S}}(\mathbf{s})$ and $\eta_n^{\overline{S},\pi}(t)$ for $t\in\overline{S}$, that we shall now define. The former only depends on $\overline{S}$ via S and also possibly on $\mathbf{s}$ while the latter depends on the choice of the prior π but not on $\mathbf{s}$.

4.1. *The quantity* $\varepsilon_n^{\overline{S}}(\mathbf{s})$. Given $\boldsymbol{X}$ with distribution $\mathbf{P}_{\mathbf{s}}$ and $y>0$, we set

$$\boldsymbol{Z}(\boldsymbol{X},t,t') = \boldsymbol{\Psi}(\boldsymbol{X},t,t') - \mathbb{E}_{\mathbf{s}}[\boldsymbol{\Psi}(\boldsymbol{X},t,t')],$$

and

$$\mathbf{w}^{\overline{S}}(\mathbf{s},y) = \mathbb{E}_{\mathbf{s}}\Big[\sup_{t,t'\in\mathscr{B}^S(\mathbf{s},y)} |\boldsymbol{Z}(\boldsymbol{X},t,t')|\Big] \quad \text{with the convention } \sup_{\varnothing} = 0.$$

Note that $\mathbf{w}^{\overline{S}}(\mathbf{s},y) = \mathbf{w}^{\overline{S}}(\mathbf{s},1)$ for $y>1$. We then define $\varepsilon_n^{\overline{S}}(\mathbf{s})$ as

$$\varepsilon_n^{\overline{S}}(\mathbf{s}) = \sup\{y>0 | \mathbf{w}^{\overline{S}}(\mathbf{s},y) > 6c_0^{-1}ny^2\} \vee \frac{1}{\sqrt{n}} \quad \text{with } \sup\varnothing = 0. \tag{16}$$

Since the function ψ is bounded by 1, $\mathbf{w}^{\overline{S}}(\mathbf{s},y)$ is not larger than $2n$; hence $\varepsilon_n^{\overline{S}}(\mathbf{s})$ is not larger than $(c_0/3)^{1/2}$. The quantity $\varepsilon_n^{\overline{S}}(\mathbf{s})$ measures in some sense the massiveness of the set S. In particular, if $S\subset S'$, $\varepsilon_n^{\overline{S}}(\mathbf{s}) \le \varepsilon_n^{\overline{S}'}(\mathbf{s})$.

4.2. *The quantity* $\eta_n^{\overline{S},\pi}(t)$.

DEFINITION 2. Let $\gamma = \beta/8$. Given the prior π on the model $\overline{S}$, we define the function $\eta_n^{\overline{S},\pi}$ on $\overline{S}$ by

$$\eta_n^{\overline{S},\pi}(t) = \sup\{\eta\in(0,1] | \pi(\mathscr{B}^{\overline{S}}(t,2\eta)) > \exp[\gamma n\eta^2]\pi(\mathscr{B}^{\overline{S}}(t,\eta))\},$$

with the convention $\sup\varnothing = 0$.

Note that $\eta_n^{\overline{S},\pi}(t)\le 1$ since $\pi(\mathscr{B}^{\overline{S}}(t,r)) = \pi(\overline{S}) = 1$ for $r\ge 1$ and that

$$\pi(\mathscr{B}^{\overline{S}}(t,2r)) \le \exp[\gamma nr^2]\pi(\mathscr{B}^{\overline{S}}(t,r)) \quad \text{for all } r\in[\eta_n^{\overline{S},\pi}(t),1]. \tag{17}$$

This inequality indeed holds by definition for $r > \eta_n^{\overline{S},\pi}(t)$, which implies by monotonicity that it also holds for $r = \eta_n^{\overline{S},\pi}(t)$. Then, if $0<\eta\le 1$ and

$$\pi(\mathscr{B}^{\overline{S}}(t,2r)) \le \exp[\gamma nr^2]\pi(\mathscr{B}^{\overline{S}}(t,r)) \quad \text{for all } r\in[\eta,1], \tag{18}$$

it follows from (17) that $\eta_n^{\overline{S},\pi}(t)\le\eta$.

The quantity $\eta_n^{\overline{S},\pi}(t)$ corresponds to some critical radius over which the π-probability of balls centred at t does not increase too quickly. In particular, if the prior puts enough mass on a small neighbourhood of t, $\eta_n^{\overline{S},\pi}(t)$ is small. Indeed, since $\pi(\mathscr{B}^{\overline{S}}(t,2r)) \le 1$ for all $r > 0$, the inequality

$$\pi\big(\mathscr{B}^{\overline{S}}(t,\eta)\big) \ge \exp\big[-\gamma n\eta^2\big] \quad \text{for some } \eta \in (0,1] \tag{19}$$

implies that, for $1 \ge r \ge \eta$,

$$\pi\big(\mathscr{B}^{\overline{S}}(t,r)\big) \ge \exp\big[-\gamma nr^2\big] \ge \pi\big(\mathscr{B}^{\overline{S}}(t,2r)\big)\exp\big[-\gamma nr^2\big],$$

hence that $\eta_n^{\overline{S},\pi}(t) \le \eta$. However, the upper bounds on $\eta_n^{\overline{S},\pi}(t)$ that are derived from (19) are usually less accurate than those derived from (18).

4.3. *Our main theorem.* The concentration properties of the ρ-posterior distribution $\pi_{\boldsymbol{X}}$ are given by the following theorem.

THEOREM 1. *Let Assumption 1 be satisfied. Then, whatever the true density* $\mathbf{s}$ *of* $\boldsymbol{X}$ *and* $\xi > 0$, *there exists a measurable subset* Ω_ξ *of* Ω *with* $\mathbb{P}_{\mathbf{s}}(\Omega_\xi) \ge 1 - e^{-\xi}$ *such that*

$$\pi_{\boldsymbol{X}(\omega)}\big(\mathscr{B}^{\overline{S}}(\mathbf{s},r)\big) \ge 1 - e^{-\xi'} \quad \textit{for all } \omega \in \Omega_\xi, \xi' > 0 \textit{ and } r \ge \overline{r}_n \tag{20}$$

with

$$\overline{r}_n = \inf_{t\in\overline{S}}\big[c_1 h(\mathbf{s},t) + c_2\eta_n^{\overline{S},\pi}(t)\big] + c_3\varepsilon_n^{\overline{S}}(\mathbf{s}) + c_4\sqrt{\frac{\xi+\xi'+2.61}{n}}. \tag{21}$$

The constants c_j, $1 \le j \le 4$ *are given in* (15) *and actually universal as soon as* $\beta \ge 7.2 \times 10^{-6}$.

In the favorable situation where the observations $X_1, \dots, X_n$ are truly i.i.d. so that $\mathbf{s} = (s, \dots, s)$, (20) can be reformulated equivalently as

$$\pi_{\boldsymbol{X}(\omega)}\big(\mathscr{B}^{\overline{S}}(s,r)\big) \ge 1 - e^{-\xi'} \quad \text{for all } \omega \in \Omega_\xi, \xi' > 0 \text{ and } r \ge \overline{r}_n$$

with

$$\overline{r}_n = \inf_{t\in\overline{S}}\big[c_1 h(s,t) + c_2\eta_n^{\overline{S},\pi}(t)\big] + c_3\varepsilon_n^{\overline{S}}(\mathbf{s}) + c_4\sqrt{\frac{\xi+\xi'+2.61}{n}}, \tag{22}$$

which measures the concentration of the ρ-posterior distribution $\pi_{\boldsymbol{X}}$ around the true density s of our i.i.d. observations $X_1, \dots, X_n$. It involves three main terms: $h(s,t)$, $\eta_n^{\overline{S},\pi}(t)$ and $\varepsilon_n^{\overline{S}}(\mathbf{s})$. For many models $\overline{S}$ of interest, as we shall see in Section 5, it is possible to show an upper bound of the form

$$\varepsilon_n^{\overline{S}}(\mathbf{s}) \le v_n(\overline{S}) \quad \text{for all } \mathbf{s} \in \mathscr{L}^n, \tag{23}$$

where $v_n(\overline{S})$ is of the order of the minimax rate of estimation on $\overline{S}$ (up to possible logarithmic factors), that is, the rate one would expect by using a frequentist or a classical Bayes estimator provided that the true density s does belong to the model $\overline{S}$ and the prior distribution puts enough mass around s. Under (23), if s does belong to $\overline{S}$, we deduce from (22) that

$$\overline{r}_n \le (c_2 + c_3)\max\big\{\eta_n^{\overline{S},\pi}(s); v_n(\overline{S})\big\} + c_4\sqrt{\frac{\xi+\xi'+2.61}{n}}. \tag{24}$$

In many cases, the quantity $\eta_n^{\overline{S},\pi}(s)$ turns out to be of the same order or smaller than $v_n(\overline{S})$ provided that the prior π puts enough mass around s. In (22), the term $\inf_{t\in\overline{S}}[c_1 h(s,t) +$

$c_2\eta_n^{\overline{S},\pi}(t)]$ expresses some robustness with respect to this ideal situation: if π puts too little mass around s, possibly zero mass when s does not belong to the model, but if s is close enough to some point $t \in \overline{S}$ around which π puts enough mass, the previous situation does not deteriorate too much. When s does not belong to the model, one may think of t as a best approximation point $\bar{t}$ of s in $\overline{S}$ when $\eta_n^{\overline{S},\pi}(\bar{t})$ is not too large or alternatively to some point t that may be slightly further away from s but for which $\eta_n^{\overline{S},\pi}(t)$ is smaller than $\eta_n^{\overline{S},\pi}(\bar{t})$ in order to minimize the function $t' \mapsto c_1 h(s,t') + c_2\eta_n^{\overline{S},\pi}(t')$ over $\overline{S}$.

If $X_1,\dots,X_n$ are not truly i.i.d. but are independent and close to being drawn from a common density $s_0 \in \overline{S}$, that is, $\mathbf{s} = (s_1,\dots,s_n)$ with $h(s_i,s_0) \le \varepsilon$ for some small $\varepsilon > 0$ and all $i \in \{1,\dots,n\}$, then $h(\mathbf{s},s_0) \le \varepsilon$ and $\mathscr{B}^{\overline{S}}(\mathbf{s},r) \subset \mathscr{B}^{\overline{S}}(s_0,\varepsilon + r)$. We therefore deduce from (20) and (21) with $t = s_0$ that, if (23) holds, the posterior distribution concentrates on Hellinger balls around s_0 with radius not larger than

$$\varepsilon + \overline{r}_n \le (1+c_1)\varepsilon + (c_2+c_3)\max\{\eta_n^{\overline{S},\pi}(s_0); v_n(\overline{S})\} + c_4\sqrt{\frac{\xi+\xi'+2.61}{n}},$$

which is similar to (24) with $s = s_0$ except for the additional term $(1+c_1)\varepsilon$ which expresses the fact that our procedure is robust with respect to a possible departure from the assumption of equidistribution.

5. Upper bounds for $\varepsilon_n^{\overline{S}}(\mathbf{s})$.

5.1. *Case of a finite set $\overline{S}$.* There are many situations for which it is natural, in view of the robustness properties of the ρ-Bayes posterior, to choose for $\overline{S}$ a finite set, in which case we take $S = \overline{S}$ and the quantity $\varepsilon_n^{\overline{S}}(\mathbf{s})$ can then be bounded from above as follows.

PROPOSITION 4. *If $\overline{S}$ is a finite set and $S = \overline{S}$,*

$$\varepsilon_n^{\overline{S}}(\mathbf{s}) < \bigl(\sqrt{c_0/3}\bigr)\min\bigl\{\sqrt{\sqrt{2}c_0 n^{-1}\log(2|\overline{S}|^2)}, 1\bigr\}.$$

An important example of such a finite set $\overline{S}$ is that of an ε-net for a totally bounded set. We recall that, if $\widetilde{S}$ is a subset of some pseudo-metric space M endowed with a pseudo-distance d and $\varepsilon > 0$, a subset S_ε of M is an ε-net for $\widetilde{S}$ if, for all $t \in \widetilde{S}$, one can find $t' \in S_\varepsilon$ such that $d(t,t') \le \varepsilon$. When $\widetilde{S}$ is totally bounded one can find a finite ε-net for $\widetilde{S}$ whatever $\varepsilon > 0$. This applies in particular to totally bounded subsets $\widetilde{S}$ of $(\mathscr{L},h)$. The smallest possible size of such nets depends on the metric properties of $(\widetilde{S},h)$ and the following notion of metric dimension, as introduced in Birgé (2006a) (Definition 6, p. 293) turns out to be a central tool.

DEFINITION 3. Let D be a function from $(0,1]$ to $[3/4,+\infty)$ which is right-continuous. A model $\widetilde{S} \subset \mathscr{L}$ admits a metric dimension bounded by D if, for all $\varepsilon \in (0,1]$, there exists an ε-net S_ε for $\widetilde{S}$ such that, for any s in $\mathscr{L}$,

$$\bigl|\{t \in S_\varepsilon, h(s,t) \le r\}\bigr| \le \exp\bigl[D(\varepsilon)(r/\varepsilon)^2\bigr] \quad \text{for all } r \ge 2\varepsilon. \tag{25}$$

Note that this implies that S_ε is finite and that one can always take $D(1) = 3/4$ since h is bounded by 1. The following result shows how a bound D for the metric dimension can be used to bound $\varepsilon_n^{\overline{S}}(\mathbf{s})$ for a model $\overline{S}$ which is an ε-net for $\widetilde{S}$ which satisfies (25).

PROPOSITION 5. *Let $\widetilde{S}$ be a totally bounded subset of $(\mathscr{L},h)$ with metric dimension bounded by D and let ε be a positive number satisfying*

$$\varepsilon \ge 1/(2\sqrt{n}) \quad \text{and} \quad D(\varepsilon) \le n(\varepsilon/c_0)^2. \tag{26}$$

If S_ε is an ε-net for $\widetilde{S}$ satisfying (25) and $S = \overline{S} = S_\varepsilon$, then $\varepsilon_n^{\overline{S}}(\mathbf{s}) \le 2\varepsilon$ whatever $\mathbf{s} \in \mathscr{L}^n$.

Starting from a classical statistical model $\widetilde{S}$ with metric dimension bounded by D we may therefore replace it by a suitable ε-net $\overline{S}$ in order to build a ρ-Bayes posterior based on some prior distribution on $\overline{S}$. The robustness of the procedure, as shown by Theorem 1, implies that the replacement of $\widetilde{S}$ by $\overline{S}$ will only entail an additional bias term of order ε.w.

5.2. *Weak VC-major classes.*

DEFINITION 4. A class of real-valued functions $\mathscr{F}$ on a set $\mathscr{X}$ is said to be weak VC-major with dimension not larger than $d \in \mathbb{N}$ if, for all $u \in \mathbb{R}$, the class of sets

$$\mathscr{C}_u(\mathscr{F}) = \{\{f > u\}, f \in \mathscr{F}\}$$

is VC on $\mathscr{X}$, with VC-dimension not larger than d. The weak VC-major dimension of $\mathscr{F}$ is the smallest such integer d.

For details on the definition and properties of VC-classes, we refer to van der Vaart and Wellner (1996) and for weak VC-major classes to Baraud (2016). One major point about weak VC-major classes is the fact that if $\mathscr{F}$ is weak VC-major with dimension not larger than $d \in \mathbb{N}$, the same holds for any subset $\mathscr{F}'$ of $\mathscr{F}$.

PROPOSITION 6. *Let $\mathscr{F}$ be the class of functions on $\mathscr{X}$ given by*

$$\mathscr{F} = \left\{\psi\left(\sqrt{\frac{t'}{t}}\right), (t,t') \in \overline{S}^2\right\}. \tag{27}$$

If it is weak VC-major with dimension not larger than $d \geq 1$, then, whatever the density $\mathbf{s} \in \mathscr{L}^n$,

$$\varepsilon_n^{\overline{S}}(\mathbf{s}) \leq \frac{11c_0}{4}\sqrt{\frac{\overline{c}_n(d \wedge n)}{n}}\left[\log\left(\frac{en}{d \wedge n}\right)\right]^{3/2} \quad \text{with } \overline{c}_n = 1 + \frac{\log 2}{\log(en)}. \tag{28}$$

5.3. *Examples.* We provide below examples of parametric models indexed by some subset $\mathbf{\Theta}$ of a Euclidean space putting on our models the σ-algebra induced by the Borel one on $\mathbf{\Theta}$. Since our results are in terms of VC-dimensions, they hold for all submodels of those described below.

PROPOSITION 7. *Let $(g_j)_{1\leq j\leq J}$ with $J \geq 1$ be real-valued functions on a set $\mathscr{X}$.*

(a) *If the elements t of the model $\overline{S}$ are of the form*

$$t(x) = \exp\left[\theta_0 + \sum_{j=1}^{J} \theta_j g_j(x)\right] \quad \text{for all } x \in \mathscr{X} \tag{29}$$

with $\theta_0, \ldots, \theta_J \in \mathbb{R}$, then $\mathscr{F}$ defined by (27) is weak VC-major with dimension not larger than $d = J + 2$.

(b) *Let $\mathscr{J} = (I_i)_{i=1,\ldots,k}$ $(k \geq 2)$ be a partition of $\mathscr{X}$. If the elements t of the model $\overline{S}$ are of the form*

$$t(x) = \sum_{i=1}^{k} \exp\left[\sum_{j=1}^{J} \theta_{i,j} g_j(x)\right] \mathbb{1}_{I_i}(x) \quad \text{for all } x \in \mathscr{X} \tag{30}$$

with $\theta_{i,j} \in \mathbb{R}$ for $i = 1, \ldots, k$ and $j = 1, \ldots J$, then $\mathscr{F}$ defined by (27) is weak VC-major with dimension not larger than $d = k(J + 2)$.

If $\mathcal{X}$ is an interval of $\mathbb{R}$ (possibly $\mathbb{R}$ itself), the second part of the proposition extends to densities based on variable partitions of $\mathcal{X}$.

PROPOSITION 8. *Let $(g_j)_{1\le j\le J}$ $(J\ge 1)$ be real-valued functions on an interval $\mathbf{I}$ of $\mathbb{R}$. Let the elements t of the model $\overline{S}$ be of the form*

$$t(x)=\sum_{I\in\mathcal{J}(t)}\exp\left[\sum_{j=1}^{J}\theta_{I,j}g_j(x)\right]\mathbb{1}_I(x)\quad \text{for all } x\in\mathcal{X}, \tag{31}$$

where $\mathcal{J}(t)$ is a partition of $\mathbf{I}$ which may depend on t, into at most k intervals $(k\ge 2)$, and $(\theta_{I,j})_{j=1,\dots,J}\in\mathbb{R}^J$ for all $I\in\mathcal{J}(t)$. Then $\mathcal{F}$ defined by (27) is weak VC-major with dimension not larger than $d=\lceil 18.8k(J+2)\rceil$, which means the smallest integer $j\ge 18.8k(J+2)$.

If, for instance, $\overline{S}$ consists of all positive histograms defined on a bounded interval $\mathbf{I}$ of $\mathbb{R}$ with at most k pieces, then one may take $J=1$, $g_1\equiv 1$ and Proposition 8 implies that $\mathcal{F}$ is weak VC-major with dimension not larger than $56.4k$.

Note that the densities t given by (30) can be viewed as elements of a piecewise exponential family. Let us indeed consider a classical exponential family on the set $\mathcal{X}$ with densities (with respect to μ) of the form

$$t_{\boldsymbol{\theta}}(x)=\exp\left[\sum_{j=1}^{J}\theta_j T_j(x)-A(\boldsymbol{\theta})\right]\quad \text{for all } x\in\mathcal{X} \tag{32}$$

with $\boldsymbol{\theta}=(\theta_1,\dots,\theta_J)\in\boldsymbol{\Theta}\subset\mathbb{R}^J$. It leads to a model $\overline{S}$ of the form (29) with $g_j=T_j$ for $1\le j\le J$ and $\theta_0=-A(\boldsymbol{\theta})$. In particular, $\mathcal{F}$ is weak VC-major with dimension not larger than $d=J+2$ and we deduce from Proposition 7 that

$$\varepsilon_n^{\overline{S}}(\mathbf{s})\le(11/4)c_0\sqrt{\frac{\overline{c}_n(J+2)}{n}}\log^{3/2}(en)\quad \text{for all } \mathbf{s}\in\mathcal{L}^n. \tag{33}$$

If all elements of $\overline{S}$ are piecewise of the form (32) on some partition $\mathcal{J}=(I_i)_{i=1,\dots,k}$ of $\mathcal{X}$ into k subsets, $\mathcal{F}$ is then weak VC-major with dimension not larger than $k(J+3)$ and for some positive universal constant c',

$$\varepsilon_n^{\overline{S}}(\mathbf{s})\le c'\sqrt{\frac{kJ}{n}}\log^{3/2}(en)\quad \text{for all } \mathbf{s}\in\mathcal{L}^n. \tag{34}$$

When $\mathcal{X}=[0,1]$, one illustration of case $b)$ is provided by $\Theta_i=[-M,M]^J$ for $i\in\{1,\dots,k\}$ and $T_j(x)=x^{j-1}$ for $j\in\{1,\dots,J\}$. We may then apply Proposition 7 and the performance of the ρ-posterior distribution will depend on the approximation properties of the family of piecewise polynomials on the partition $\mathcal{J}$ with respect to the logarithm of the true density. Numerous results about such approximations can be found in DeVore and Lorentz (1993).

6. Upper bounds for $\eta_n^{\overline{S},\pi}(t)$.

6.1. *Uniform distribution on an ε-net.* We consider here the situation where $\widetilde{S}$ is a totally bounded subset of $(\mathcal{L},h)$ with metric dimension bounded by D, $\varepsilon\in(0,1]$, $\overline{S}=S_\varepsilon$ is an ε-net for $\widetilde{S}$ which satisfies (25) and we choose π as the uniform distribution on $\overline{S}$.

PROPOSITION 9. *If $D(\varepsilon)\le(\gamma/4)n\varepsilon^2$, then $\eta_n^{\overline{S},\pi}(t)\le\varepsilon$ for all $t\in\overline{S}$.*

PROOF. Let $t \in \overline{S}$. For all $r > 0$, $\pi(\mathscr{B}^{\overline{S}}(t,r)) \geq \pi(\{t\}) = |\overline{S}|^{-1}$. Using (25), we derive that

$$\frac{\pi(\mathscr{B}^{\overline{S}}(t,2r))}{\pi(\mathscr{B}^{\overline{S}}(t,r))} \leq |\overline{S}|\pi\big(\mathscr{B}^{\overline{S}}(t,2r)\big) = \big|\mathscr{B}^{\overline{S}}(t,2r)\big| \leq \exp\left[4D(\varepsilon)\left(\frac{r}{\varepsilon}\right)^2\right]$$

for all $r \geq \varepsilon$. The conclusion follows from the fact that $4D(\varepsilon)/\varepsilon^2 \leq \gamma n$. □

6.2. *Parametric models indexed by a bounded subset of* $\mathbb{R}^d$. In this section, we consider the situation where $\overline{S}$ is a parametric model $\{t_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ indexed by a measurable (with respect to the Borel σ-algebra) bounded subset $\boldsymbol{\Theta} \subset \mathbb{R}^d$ and we assume that the prior π is the image by the mapping $\boldsymbol{\theta} \mapsto t_{\boldsymbol{\theta}}$ of some probability ν on $\boldsymbol{\Theta}$. Besides, we assume that the Hellinger distance on $\overline{S}$ is related on $\boldsymbol{\Theta}$ to some norm $|\cdot|_*$ on $\mathbb{R}^d$ in the following way:

$$\underline{a}\big|\boldsymbol{\theta} - \boldsymbol{\theta}'\big|_*^{\alpha} \leq h(t_{\boldsymbol{\theta}}, t_{\boldsymbol{\theta}'}) \leq \overline{a}\big|\boldsymbol{\theta} - \boldsymbol{\theta}'\big|_*^{\alpha} \quad \text{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}, \tag{35}$$

where $\underline{a}$, $\overline{a}$ and α are positive numbers. Since h is bounded by 1, (35) implies that $\boldsymbol{\Theta}$ is necessarily bounded. Let us denote by $\mathcal{B}_*(\boldsymbol{\theta}, r)$ the closed ball (with respect to the norm $|\cdot|_*$) of center $\boldsymbol{\theta}$ and radius r in $\mathbb{R}^d$.

PROPOSITION 10. *Assume that* $\boldsymbol{\Theta}$ *is measurable and bounded in* $\mathbb{R}^d$, *that* (35) *holds and that* ν *satisfies*

$$\nu\big(\mathcal{B}_*(\boldsymbol{\theta}, 2x)\big) \leq \kappa_{\boldsymbol{\theta}}(x)\nu\big(\mathcal{B}_*(\boldsymbol{\theta}, x)\big) \quad \textit{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \textit{ and } x > 0, \tag{36}$$

where $\kappa_{\boldsymbol{\theta}}(x)$ *denotes some positive nonincreasing function on* $\mathbb{R}_+$. *Then, for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\eta_n^{\overline{S},\pi}(t_{\boldsymbol{\theta}}) \leq \inf\left\{\eta > 0 \,\middle|\, \eta^2 \geq \frac{\log(\kappa_{\boldsymbol{\theta}}([\eta/\overline{a}]^{1/\alpha}))}{\gamma n}\left[\frac{\log(2\overline{a}/\underline{a})}{\alpha \log 2} + 1\right]\right\}. \tag{37}$$

If $\kappa_{\boldsymbol{\theta}}(x) \equiv \kappa_0$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and* $x > 0$, *then*

$$\eta_n^{\overline{S},\pi}(t_{\boldsymbol{\theta}}) \leq \sqrt{\frac{\log \kappa_0}{\gamma n}\left[\frac{\log(2\overline{a}/\underline{a})}{\alpha \log 2} + 1\right]} \quad \textit{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{38}$$

In particular, if $\boldsymbol{\Theta}$ *is convex and* ν *admits a density* g *with respect to the Lebesgue measure* λ *on* $\mathbb{R}^d$ *which satisfies*

$$\underline{b} \leq g(\boldsymbol{\theta}) \leq \overline{b} \quad \textit{for } \lambda\textit{-almost all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \textit{ with } 0 < \underline{b} \leq \overline{b}, \tag{39}$$

then (36) *holds with* $\kappa_{\boldsymbol{\theta}}(x) \equiv \kappa_0 = 2^d(\overline{b}/\underline{b})$, *hence, for all* $t \in \overline{S}$,

$$\eta_n^{\overline{S},\pi}(t) \leq c\sqrt{\frac{d}{n}} \quad \textit{with } c^2 = \frac{\log(2[\overline{b}/\underline{b}]^{1/d})}{\gamma}\left[\frac{\log(2\overline{a}/\underline{a})}{\alpha \log 2} + 1\right]. \tag{40}$$

6.3. *Example*. Let us consider, in the density model with n i.i.d. observations on $\mathbb{R}$, the following translation family $t_\theta(x) = t(x - \theta)$ where t is the density of the Gamma$(2\alpha, 1)$ distribution, namely

$$t(x) = c(\alpha)x^{2\alpha-1}e^{-x}\mathbb{1}_{x\geq 0} \quad \text{with } 0 < \alpha < 1$$

and θ belongs to the interval $\Theta = [-1, 1]$. It is known from Example 1.3, page 287 of Ibragimov and Has'minskiĭ (1981) that, in this situation, (35) holds for $|\cdot|_*$ the absolute value and $\underline{a}$, $\overline{a}$ depending on α. Let us now derive upper bounds for $\eta_n^{\overline{S},\pi}(t_{\boldsymbol{\theta}})$ when ν has a density g with respect to the Lebesgue measure.

– If ν is uniform on Θ, then $\overline{b} = \underline{b}$ and (40) is satisfied for some constant c depending on α and γ only.
– If $g(z) = (\xi/2)|z|^{\xi-1}\mathbb{1}_{[-1,1]}(z)$ with $0 < \xi < 1$, in order to compute κ_0 one has to compare the ν-measures of the intervals $I_1 = [(\theta - x) \vee -1, (\theta + x) \wedge 1]$ and $I_2 = [(\theta - 2x) \vee -1, (\theta + 2x) \wedge 1]$ for $x > 0$.

PROPOSITION 11. *If in this example* $g(z) = (\xi/2)|z|^{\xi-1}\mathbb{1}_{[-1,1]}(z)$, (38) *holds since*

$$\nu(I_2) \le \kappa_0 \nu(I_1) \quad \text{with } \kappa_0 = 2^{1+\xi}\big(2^\xi - 1\big)^{-1}.$$

One should therefore note that if (39) is sufficient for $\kappa_{\boldsymbol{\theta}}(r)$ to be constant, it is by no means necessary.

– Let us now set $g(z) = c_\delta^{-1} \exp[-(2|z|^\delta)^{-1}]\mathbb{1}_{[-1,1]}(z)$ for some $\delta > 0$, which means that the prior puts very little mass around the point $\theta = 0$.

Then

PROPOSITION 12. *In this example,* $\eta_n^{\overline{S},\pi}(t_0) \le K n^{-\alpha/[2\alpha+\delta]}$, *for some* K *depending on* α, δ, $\overline{a}$, $\underline{a}$ *and* γ.

It is not difficult to check that in this situation the family $\mathscr{F}$ defined by (27) consists of elements f for which either f or $-f$ is unimodal. In particular, for $f \in \mathscr{F}$, the levels sets $\{f > u\}$ with $u \in \mathbb{R}$ consist of a union of at most two disjoint intervals. It follows from Lemma 1 of Baraud and Birgé (2016) that $\mathscr{F}$ is then weak-VC major with dimension not larger than 4 so that, as a consequence of Proposition 6, $\varepsilon_n^{\overline{S}}(\mathbf{s}) \le C(\log n)^{3/2}$ for some universal constant $C > 0$ and all densities $\mathbf{s} \in \mathscr{L}^n$. Applying Theorem 1 when the true parameter θ is 0 leads to a bound for (22) of the form

$$\overline{r}_n \le K\left[n^{-\alpha/(2\alpha+\delta)} + \sqrt{\frac{\log^3 n}{n}} + \sqrt{\frac{\xi + \xi' + 2.61}{n}}\right],$$

which is of the order of $n^{-\alpha/(2\alpha+\delta)}$ and clearly depends on the relative values of α and δ. In particular, if $\alpha = 1/2$, which corresponds to the exponential density, we get a bound for $\overline{r}_n$ of order $n^{-1/[2(1+\delta)]}$.

7. Connexion with classical Bayes estimators. Throughout this section, we assume that the data $X_1, \dots, X_n$ are i.i.d. with density s on the measured space $(\mathscr{X}, \mathscr{A}, \mu)$.

We consider a parametric set of real nonnegative functions $\{t_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ satisfying $\int_{\mathscr{X}} t_{\boldsymbol{\theta}}(x)\, d\mu(x) = 1$, indexed by some subset $\boldsymbol{\Theta}$ of $\mathbb{R}^d$ and such that the mapping $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}} = t_{\boldsymbol{\theta}} \cdot \mu$ is one-to-one so that our statistical model for s is $\overline{S} = \{t_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. We set $\|t\|_\infty = \sup_{x \in \mathscr{X}} |t(x)|$ for any function t on $\mathscr{X}$. Since the mapping $\boldsymbol{\theta} \mapsto t_{\boldsymbol{\theta}}$ is one-to-one, the Hellinger distance can be transferred to $\boldsymbol{\Theta}$ and we shall write $h(\boldsymbol{\theta}, \boldsymbol{\theta}')$ for $h(t_{\boldsymbol{\theta}}, t_{\boldsymbol{\theta}'}) = h(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}'})$.

We consider on $(\overline{S}, h)$ the Borel σ-algebra $\mathscr{S}$ and, given a prior π on $(\overline{S}, \mathscr{S})$, we consider both the usual Bayes posterior distribution $\pi_{\boldsymbol{X}}^L$ and our ρ-posterior distribution $\pi_{\boldsymbol{X}}$ given by (6) with $\beta = 4$. A natural question is whether these two distributions are similar or not, at least asymptotically when n tends to $+\infty$. This question is suggested by the fact, proven in Section 5.1 of BBS, that under suitable regularity assumptions, the maximum likelihood estimator is a ρ-estimator, at least asymptotically.

In order to show that the two distributions $\pi_{\boldsymbol{X}}^L$ and $\pi_{\boldsymbol{X}}$ are asymptotically close, we shall introduce the following assumptions that are certainly not minimal but at least lead to simpler proofs.

ASSUMPTION 4.

(i) The function $(x,\boldsymbol{\theta}) \mapsto t_{\boldsymbol{\theta}}(x)$ is measurable from $(\mathscr{X} \times \boldsymbol{\Theta}, \mathscr{A} \otimes \mathscr{G}(\boldsymbol{\Theta}))$ to $(\mathbb{R}_+, \mathscr{R})$ where $\mathscr{G}(\boldsymbol{\Theta})$ and $\mathscr{R}$ denote respectively the Borel σ-algebras on $\boldsymbol{\Theta} \subset \mathbb{R}^d$ and $\mathbb{R}_+$.

(ii) The parameter set $\boldsymbol{\Theta}$ is a compact and convex subset of $\boldsymbol{\Theta}' \subset \mathbb{R}^d$ and the true density $s = t_{\boldsymbol{\vartheta}}$ belongs to $\overline{S}$.

(iii) There exists a positive function A_2 on $\boldsymbol{\Theta}$ such that the following relationship between the Hellinger and Euclidean distances holds:

$$\frac{A_2(\boldsymbol{\theta}')}{2}|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}| \le h\left(\frac{t_{\overline{\boldsymbol{\theta}}} + t_{\boldsymbol{\theta}'}}{2}, \frac{t_{\boldsymbol{\theta}} + t_{\boldsymbol{\theta}'}}{2}\right) \quad \text{for all } \overline{\boldsymbol{\theta}}, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}. \tag{41}$$

(iv) Whatever $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the density $t_{\boldsymbol{\theta}}$ is positive on $\mathscr{X}$ and there exists a constant A_1 such that

$$\left\| \sqrt{\frac{t_{\boldsymbol{\theta}}}{t_{\boldsymbol{\theta}'}}} - \sqrt{\frac{t_{\overline{\boldsymbol{\theta}}}}{t_{\boldsymbol{\theta}'}}} \right\|_\infty \le A_1 |\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}| \quad \text{for all } \boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \text{ and } \boldsymbol{\theta}' \in \boldsymbol{\Theta}.$$

These assumptions imply that $(\overline{S}, h)$ is a metric space and that the function $t \mapsto t(x)$ from $(\overline{S}, h)$ to $(0, +\infty)$ is continuous whatever $x \in \mathscr{X}$. Furthermore, Assumption 4(iv) implies that the Hellinger distance on $\boldsymbol{\Theta}$ is controlled by the Euclidean one in the following way:

$$h^2(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) = \frac{1}{2}\int \left(\sqrt{\frac{t_{\boldsymbol{\theta}}}{t_{\boldsymbol{\theta}'}}} - \sqrt{\frac{t_{\overline{\boldsymbol{\theta}}}}{t_{\boldsymbol{\theta}'}}} \right)^2 t_{\boldsymbol{\theta}'}\, d\mu \le \frac{A_1^2}{2}|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}|^2.$$

Since the concavity of the square root implies that

$$h\left(\frac{t_{\overline{\boldsymbol{\theta}}} + t_{\boldsymbol{\theta}'}}{2}, \frac{t_{\boldsymbol{\theta}} + t_{\boldsymbol{\theta}'}}{2}\right) \le \frac{1}{2} h(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}), \tag{42}$$

we derive from (41) with $\boldsymbol{\theta}' = \boldsymbol{\vartheta}$ that $h(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}) \ge A_2(\boldsymbol{\vartheta})|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}|$. The Hellinger and Euclidean distances are therefore equivalent on $\boldsymbol{\Theta}$:

$$A_2|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}| \le h(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}) = h(t_{\overline{\boldsymbol{\theta}}}, t_{\boldsymbol{\theta}}) \le A_3|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}| \quad \text{for all } \overline{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{43}$$

with $A_2 = A_2(\boldsymbol{\vartheta}) < A_3 = A_1/\sqrt{2}$.

In particular, the mapping $t_{\boldsymbol{\theta}} \mapsto \boldsymbol{\theta}$ is continuous from $(\overline{S}, h)$ to $(\boldsymbol{\Theta}, |\cdot|)$, hence measurable from $(\overline{S}, \mathscr{S})$ to $(\boldsymbol{\Theta}, \mathscr{G}(\boldsymbol{\Theta}))$ and so are $f : (x, t_{\boldsymbol{\theta}}) \mapsto (x, \boldsymbol{\theta})$ from $(\mathscr{X} \times \overline{S}, \mathscr{A} \otimes \mathscr{S})$ to $(\mathscr{X} \times \boldsymbol{\Theta}, \mathscr{A} \otimes \mathscr{G}(\boldsymbol{\Theta}))$ and $(x, t_{\boldsymbol{\theta}}) \mapsto t_{\boldsymbol{\theta}}(x)$ from $(\mathscr{X} \times \overline{S}, \mathscr{A} \otimes \mathscr{S})$ to $(\mathbb{R}_+, \mathscr{R})$ as the composition of f with $(x, \boldsymbol{\theta}) \mapsto t_{\boldsymbol{\theta}}(x)$ which is measurable under Assumption 4(ii). Consequently Assumption 1(i) is satisfied and so is (9) if we take for S the image by the mapping $\boldsymbol{\theta} \mapsto t_{\boldsymbol{\theta}}$ of a countable and dense subset of $(\boldsymbol{\Theta}, |\cdot|)$ and use the fact that for all $x \in \mathscr{X}$, the function $t \mapsto t(x)$ is continuous and positive on $(\overline{S}, h)$.

We deduce from (41) and (42) that

$$\frac{A_2}{2}|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}| \le h\left(\frac{t_{\overline{\boldsymbol{\theta}}} + s}{2}, \frac{t_{\boldsymbol{\theta}} + s}{2}\right) \le \frac{A_3}{2}|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}| \quad \text{for all } \overline{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{44}$$

and since ψ is a Lipschitz function with Lipschitz constant 2, Assumption 4(iv) implies that

$$\left\| \psi\left(\sqrt{\frac{t_{\boldsymbol{\theta}}}{t_{\boldsymbol{\theta}'}}}\right) - \psi\left(\sqrt{\frac{t_{\overline{\boldsymbol{\theta}}}}{t_{\boldsymbol{\theta}'}}}\right) \right\|_\infty \le 2A_1|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}| \quad \text{for all } \boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \text{ and } \boldsymbol{\theta}' \in \boldsymbol{\Theta}.$$

If $\boldsymbol{\Theta}$ is a compact subset of an open set $\boldsymbol{\Theta}'$ and the the parametric family $\{t_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}'\}$ is regular with invertible Fisher Information matrix, the same holds for the family $\{[t_{\boldsymbol{\theta}} + t_{\boldsymbol{\theta}'}]/2, \boldsymbol{\theta} \in \boldsymbol{\Theta}'\}$ for each given $\boldsymbol{\theta}'$ in $\boldsymbol{\Theta}$, which implies that Assumption 4(iii) holds.

ASSUMPTION 5. The prior π on $(\overline{S}, \mathscr{S})$ is the image via the mapping $\boldsymbol{\theta} \mapsto t_{\boldsymbol{\theta}}$ of a probability ν on $(\boldsymbol{\Theta}, \mathscr{G}(\boldsymbol{\Theta}))$ that satisfies the following requirements for suitable constants $B \geq 1$ and $\overline{\gamma} \in [1, 4)$: if $\mathcal{B}(\boldsymbol{\theta}, r)$ denotes the closed Euclidean ball in $\boldsymbol{\Theta}$ with center $\boldsymbol{\theta}$ and radius r, whatever $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $r > 0$,

$$\nu\big[\mathcal{B}(\boldsymbol{\theta}, 2^k r)\big] \leq \exp\big[B\overline{\gamma}^k\big]\nu\big[\mathcal{B}(\boldsymbol{\theta}, r)\big] \quad \text{for all } k \in \mathbb{N}^\star. \tag{45}$$

The convexity of $\boldsymbol{\Theta}$ and the well-known formulas for the volume of Euclidean balls imply that this property holds for all probabilities which are absolutely continuous with respect to the Lebesgue measure with a density which is bounded from above and below but other situations are also possible. One simple example would be $\boldsymbol{\Theta} = [-1, 1]$ and ν with density $(1/2)(\alpha + 1)|x|^\alpha$, $\alpha > 0$ with respect to the Lebesgue measure.

THEOREM 2. *Under Assumptions 4 and 5, one can find two functions C and n_1 on $(0, +\infty)$, also depending on s and all the parameters involved in these assumptions but independent of n, such that, for all $n \geq n_1(z)$,*

$$\mathbb{P}_{\mathbf{s}}\left[h^2\big(\pi_{\boldsymbol{X}}^L, \pi_{\boldsymbol{X}}\big) \leq C(z)\frac{(\log n)^{3/2}}{\sqrt{n}}\right] \geq 1 - e^{-z} \quad \textit{for all } z > 0.$$

This means that, under suitably strong assumptions, the usual posterior and our ρ-posterior distributions are asymptotically the same which shows that our construction is a genuine generalization of the classical Bayesian approach. It also implies that the Bernstein–von Mises theorem also holds for $\pi_{\boldsymbol{X}}$ as shown by the following result.

COROLLARY 1. *Let Assumptions 4 and 5 hold, $(\widehat{\boldsymbol{\theta}}_n)$ be an asymptotically efficient sequence of estimators of the true parameter $\boldsymbol{\vartheta}$ and assume that the following version of the Bernstein–von Mises theorem is true*:

$$\big\|\pi_{\boldsymbol{X}}^L - \mathcal{N}\big(\widehat{\boldsymbol{\theta}}_n, [nI(\boldsymbol{\vartheta})]^{-1}\big)\big\|_{\mathrm{TV}} \xrightarrow[n\to+\infty]{\mathrm{P}} 0,$$

where I denotes the Fisher Information matrix and $\|\cdot\|_{\mathrm{TV}}$ the total variation norm. Then the ρ-posterior distribution also satisfies the same Bernstein–von Mises theorem, that is,

$$\big\|\pi_{\boldsymbol{X}} - \mathcal{N}\big(\widehat{\boldsymbol{\theta}}_n, [nI(\boldsymbol{\vartheta})]^{-1}\big)\big\|_{\mathrm{TV}} \xrightarrow[n\to+\infty]{\mathrm{P}} 0.$$

PROOF. It follows from the triangular inequality and the classical relationship between Hellinger and total variation distances given by (8). □

8. Combining different models.

8.1. *Priors and models.* In the case of simple parametric problems with parameter set $\boldsymbol{\Theta}$, such as those we considered in Section 7, $\overline{S}$ is the image of a subset of some Euclidean space $\mathbb{R}^d$ and one often chooses for π the image of a probability on $\boldsymbol{\Theta}$ which has a density with respect to the Lebesgue measure. The choice of a convenient prior π becomes more complex when $\overline{S}$ is a complicated function space which is very inhomogeneous with respect to the Hellinger distance. In such a case, it is often useful to introduce "models", that is to consider $\overline{S}$ as a countable union of more elementary and homogeneous disjunct subsets $\overline{S}_m$, $m \in \mathcal{M}$, and to choose a prior π_m on each $\overline{S}_m$ in such a way that Theorem 1 applies to each model $\overline{S}_m$ and leads to a nontrivial result. It remains to put all models together by choosing some prior ν on $\mathcal{M}$ and defining our final prior π on $\overline{S} = \bigcup_{m\in\mathcal{M}} \overline{S}_m$ as $\sum_{m\in\mathcal{M}} \nu(\{m\})\pi_m$. This corresponds to a hierarchical prior.

One can as well proceed in the opposite way, starting from a global prior π on $\overline{S}$ and partitioning $\overline{S}$ into subsets $\overline{S}_m$, $m \in \mathcal{M}$, of positive prior probability, then setting $\nu(\{m\}) = \pi(\overline{S}_m)$ and defining π_m as the conditional distribution of a random element $t \in \overline{S}$ when it belongs to $\overline{S}_m$. The two points of view are actually clearly equivalent, the important fact for us being that the pairs $(\overline{S}_m, \pi_m)$ are such that Theorem 1 can be applied to each of them.

Throughout this section, we work within the following framework. Given a countable sequence of disjunct probability spaces $(\overline{S}_m, \mathcal{S}_m, \pi_m)_{m\in\mathcal{M}}$ on $(\mathcal{X}, \mathcal{A})$, we consider $\overline{S} = \bigcup_{m\in\mathcal{M}} \overline{S}_m$ endowed with the σ-algebra $\mathcal{S}$ defined as

$$\mathcal{S} = \{A \subset \overline{S},\, A \cap \overline{S}_m \in \mathcal{S}_m \text{ for all } m \in \mathcal{M}\}.$$

In order to define our prior, we introduce a mapping pen from $\mathcal{M}$ into $\mathbb{R}_+$ that will also be involved in the definition of our ρ-posterior distribution. The prior π on $\overline{S}$ is given by

$$\pi(A) = \Delta \sum_{m\in\mathcal{M}} \int_{A\cap\overline{S}_m} \exp\bigl[-\beta \operatorname{pen}(m)\bigr]\, d\pi_m(t) \quad \text{for all } A \in \mathcal{S} \tag{46}$$

with

$$\Delta = \Biggl(\sum_{m\in\mathcal{M}} \int_{\overline{S}_m} \exp\bigl[-\beta \operatorname{pen}(m)\bigr]\, d\pi_m(t)\Biggr)^{-1},$$

so that π is a genuine prior. This amounts to put a prior weight proportional to $\exp[-\beta \operatorname{pen}(m)]$ on the model $\overline{S}_m$. We shall assume the following.

ASSUMPTION 6.

(i) For all $m \in \mathcal{M}$, the function $(x,t) \to t(x)$ on $\mathcal{X} \times \overline{S}_m$ is measurable with respect to the σ-algebra $\mathcal{A} \otimes \mathcal{S}_m$.

(ii) For all $m \in \mathcal{M}$, there exists a countable subset S_m of $\overline{S}_m$ with the following property: given $t \in \overline{S}_m$ and $t' \in S = \bigcup_{m'\in\mathcal{M}} S_{m'}$, one can find a sequence $(t_k)_{k\ge 0}$ in S_m such that (9) holds for all $x \in \mathcal{X}$.

(iii) There exists a mapping $m \mapsto \overline{\varepsilon}_m^2$ from $\mathcal{M}$ to $\mathbb{R}_+$ such that, whatever the density $\mathbf{s} \in \mathcal{L}^n$,

$$\varepsilon_n^{\overline{S}_m \cup \overline{S}_{m'}}(\mathbf{s}) \le \sqrt{\overline{\varepsilon}_m^2 + \overline{\varepsilon}_{m'}^2} \quad \text{for all } m, m' \in \mathcal{M}. \tag{47}$$

(iv) Given a set $\{L_m, m \in \mathcal{M}\}$ of nonnegative numbers satisfying

$$\sum_{m\in\mathcal{M}} \exp[-L_m] = 1, \tag{48}$$

the penalty function pen is lower bounded in the following way:

$$\operatorname{pen}(m) \ge c_5 n \overline{\varepsilon}_m^2 + \bigl(c_6 + \beta^{-1}\bigr) L_m \quad \text{for all } m \in \mathcal{M}, \tag{49}$$

with constants c_5 and c_6 defined in (15).

8.2. *The results.* We define the ρ-posterior distribution $\overline{\pi}_{\boldsymbol{X}}$ on $\overline{S}$ by its density with respect to the prior π given by (46) as follows:

$$\frac{d\overline{\pi}_{\boldsymbol{X}}}{d\pi}(t) = \frac{\exp[-\beta\overline{\boldsymbol{\Psi}}(\boldsymbol{X}, t)]}{\int_{\overline{S}} \exp[-\beta\overline{\boldsymbol{\Psi}}(\boldsymbol{X}, t')]\, d\pi(t')} \quad \text{for all } t \in \overline{S}, \tag{50}$$

with

$$\overline{\boldsymbol{\Psi}}(\boldsymbol{X}, t) = \sup_{m\in\mathcal{M}} \sup_{t'\in S_m} \bigl[\boldsymbol{\Psi}\bigl(\boldsymbol{X}, t, t'\bigr) - \operatorname{pen}(m)\bigr].$$

Note that if we choose $\beta = 1$ and replace $\Psi(X, t, t')$ by the difference of the log-likelihoods $\sum_{i=1}^n \log t'(X_i) - \sum_{i=1}^n \log t(X_i)$, $\overline{\pi}_X$ is the usual posterior distribution corresponding to the prior π. We finally, introduce a mapping $\overline{\eta}$ on $\overline{S}$ which associates to an element $t \in \overline{S}_m$ with $m \in \mathcal{M}$ the quantity $\overline{\eta}_n^2(t)$ given by

$$\overline{\eta}_n^2(t) = \inf_{r\in(0,1]}\left[c_7 r^2 + \frac{1}{2n\beta}\log\left(\frac{1}{\pi_m(\mathscr{B}^{\overline{S}_m}(t,r))}\right)\right] \quad \text{for all } t \in \overline{S}_m, \tag{51}$$

which only depends on the choice of the prior π_m on $\overline{S}_m$. Taking $r = 1$, we see that $\eta_n^2(t) \le c_7$ for all $t \in \overline{S}$. Moreover, if, for some $\eta \in (0, 1]$ and $\lambda > 0$,

$$\pi_m\big(\mathscr{B}^{\overline{S}_m}(t,r)\big) \ge \exp\big[-\lambda n r^2\big] \quad \text{for all } r \ge \eta, m \in \mathcal{M} \text{ and } t \in \overline{S}_m,$$

then

$$\overline{\eta}_n^2(t) \le \inf_{r\ge\eta}\left[c_7 r^2 + \frac{\lambda r^2}{2\beta}\right] = \left[c_7 + \frac{\lambda}{2\beta}\right]\eta^2,$$

a result which is similar to the one we derived for $\eta_n^{\overline{S},\pi}(t)$ in Section 4.2 under an analogous assumption.

THEOREM 3. *Let Assumption 6 hold. For all $\xi > 0$ and whatever the density $\mathbf{s} \in \mathscr{L}^n$ of X, there exists a set Ω_ξ with $\mathbb{P}_{\mathbf{s}}(\Omega_\xi) \ge 1 - e^{-\xi}$ and such that*

$$\overline{\pi}_{X(\omega)}\big(\mathscr{B}^{\overline{S}}(\mathbf{s}, r)\big) \ge 1 - e^{-\xi'} \quad \textit{for all } \omega \in \Omega_\xi, \xi' > 0 \textit{ and } r \ge \overline{r}_n$$

with

$$\overline{r}_n^2 = \inf_{m\in\mathcal{M}}\inf_{\overline{s}\in\overline{S}_m}\left[\frac{3c_7}{c_8}h^2(\mathbf{s},\overline{s}) - h^2(\mathbf{s},\overline{S}) + \frac{2}{c_8}\left(\frac{2\,\mathrm{pen}(m)}{n} + \overline{\eta}_n^2(\overline{s}) - \frac{L_m}{\beta n}\right)\right] + c_9\frac{\xi + \xi' + 2.4}{n}$$

and constants c_j, $7 \le j \le 9$ defined in (15).

This result about the concentration of the ρ-posterior distribution is analogue to that one can obtain from a frequentist point of view by using a model selection method. Up to possible extra logarithmic terms, the ρ-posterior concentrates at a rate which achieves the best compromise between the approximation and complexity terms among the family of models.

8.3. *Model selection among exponential families.* In this section, we pretend that the observations $X_1, \dots, X_n$ are i.i.d. but keep in mind that the X_i might not be equidistributed so that their true joint density $\mathbf{s}$ might not be of the form $(s, \dots, s)$.

Hereafter, $\ell_2(\mathbb{N})$ denotes the Hilbert space of all square-summable sequences $\boldsymbol{\theta} = (\theta_j)_{j\ge0}$ of real numbers that we endow with the Hilbert norm $|\cdot|$ and the inner product $\langle\cdot,\cdot\rangle$. Let $\mathcal{M} = \mathbb{N}$, M be some positive number and for $m \in \mathcal{M}$, let $\boldsymbol{\Theta}'_m$ be the subset of $\ell_2(\mathbb{N})$ of these sequences $\boldsymbol{\theta} = (\theta_j)_{j\ge0}$ such that $\theta_j \in [-M, M]$ for $0 \le j \le m$ and $\theta_j = 0$ for all $j > m$.

For a sequence $\mathbf{T} = (T_j)_{j\ge0}$ of linearly independent measurable real-valued functions on $\mathscr{X}$ with $T_0 \equiv 1$ and $m \in \mathcal{M}$, we define the density model $\overline{S}_m$ as the exponential family

$$\overline{S}_m = \big\{t_{\boldsymbol{\theta}} = \exp\big[\langle\boldsymbol{\theta}, \mathbf{T}\rangle - A(\boldsymbol{\theta})\big], \boldsymbol{\theta} \in \boldsymbol{\Theta}'_m, \theta_m \ne 0\big\},$$

where A denotes the mapping from $\boldsymbol{\Theta} = \bigcup_{m\in\mathcal{M}} \boldsymbol{\Theta}'_m$ to $\mathbb{R}$ defined by

$$A(\boldsymbol{\theta}) = \log\int_{\mathscr{X}} \exp\big[\langle\boldsymbol{\theta}, \mathbf{T}(x)\rangle\big]\,d\mu(x),$$

and μ is a finite measure on $\mathscr{X}$. Note that, whatever $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $x \mapsto \langle \boldsymbol{\theta}, \mathbf{T}(x) \rangle$ is well-defined on $\mathscr{X}$ since only a finite number of coefficients of $\boldsymbol{\theta}$ are nonzero.

For all $m \in \mathcal{M}$, we endow $\overline{S}_m$ with the Borel σ-algebra $\mathscr{S}_m$ and the prior π_m which is the image of the uniform distribution on $\boldsymbol{\Theta}'_m$ (identified with $[-M, M]^{m+1}$) by the mapping $\boldsymbol{\theta} \mapsto t_{\boldsymbol{\theta}}$ on $\boldsymbol{\Theta}'_m$. Throughout this section, we consider the family of (disjunct) measured spaces $(\overline{S}_m, \mathscr{S}_m, \pi_m)$ with $m \in \mathcal{M}$ together with the choice $L_m = (m+1)\log 2$ for all $m \in \mathcal{M}$, so that $\sum_{m \in \mathcal{M}} e^{-L_m} = 1$. Then $\overline{S} = \bigcup_{m \in \mathcal{M}} \overline{S}_m$, π is given by (46) and for all $m \in \mathcal{M}$,

$$\operatorname{pen}(m) = c_5 n \overline{\varepsilon}_m^2 + \bigl(c_6 + \beta^{-1}\bigr) L_m \quad \text{with } \overline{\varepsilon}_m = \frac{11 c_0}{4} \sqrt{\frac{\overline{c}_n (m+3)}{n}} \log^{3/2}(en)$$

and c_0, c_5, c_6, $\overline{c}_n$ defined in (15). In such a situation, we derive the following result.

PROPOSITION 13. *Assume that, for all $m \in \mathcal{M}$, the restriction A_m of A to $\boldsymbol{\Theta}'_m$ is convex and twice differentiable on the interior of $\boldsymbol{\Theta}'_m$ with a Hessian whose eigenvalues lie in $(0, \sigma_m]$ for some $\sigma_m > 0$. Whatever the density $\mathbf{s}$ of $\boldsymbol{X}$, for all $\xi > 0$, with $\mathbb{P}_{\mathbf{s}}$-probability at least $1 - e^{-\xi}$,*

$$\overline{\pi}_{\boldsymbol{X}}\bigl(\mathscr{B}^{\overline{S}}(\mathbf{s}, r)\bigr) \geq 1 - e^{-\xi'} \quad \textit{for all } \xi' > 0 \textit{ and all } r \in [\overline{r}_n, 1]$$

with

$$\overline{r}_n^2 \leq C(\beta) \inf_{m \geq 1} \biggl[h^2(\mathbf{s}, \overline{S}_m) + \frac{m+1}{n} \bigl[\log^3(en) + \log\bigl(1 + n\sigma_m^2 M^2\bigr) \bigr] \biggr] + c_9 \frac{\xi + \xi' + 2.4}{n}$$

and some constant $C(\beta) > 0$ depending on β only.

Acknowledgements. The first author has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 811017.

The second author was supported by the grant ANR-17-CE40-0001-01 of the French National Research Agency ANR (project BASICS) and by Laboratoire J.A. Dieudonné (Nice).

SUPPLEMENTARY MATERIAL

Supplement to "Robust Bayes-like estimation: Rho-Bayes estimation" (DOI: 10.1214/20-AOS1948SUPP; .pdf). This Supplement provides the proofs of most results given in the paper.

REFERENCES

ATCHADÉ, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *Ann. Statist.* **45** 2248–2273. MR3718168 https://doi.org/10.1214/16-AOS1526

BARAUD, Y. (2016). Bounding the expectation of the supremum of an empirical process over a (weak) VC-major class. *Electron. J. Stat.* **10** 1709–1728. MR3522658 https://doi.org/10.1214/15-EJS1055

BARAUD, Y. and BIRGÉ, L. (2016). Rho-estimators for shape restricted density estimation. *Stochastic Process. Appl.* **126** 3888–3912. MR3565484 https://doi.org/10.1016/j.spa.2016.04.013

BARAUD, Y. and BIRGÉ, L. (2017). Robust bayes-like estimation: Rho-bayes estimation. Technical report, arXiv:1711.08328v1.

BARAUD, Y. and BIRGÉ, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.* **46** 3767–3804. MR3852668 https://doi.org/10.1214/17-AOS1675

BARAUD, Y. and BIRGÉ, L. (2020). Supplement to "Robust Bayes-like estimation: Rho-Bayes estimation." https://doi.org/10.1214/20-AOS1948SUPP

BARAUD, Y., BIRGÉ, L. and SART, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.* **207** 425–517. MR3595933 https://doi.org/10.1007/s00222-016-0673-5

BHATTACHARYA, A., PATI, D. and YANG, Y. (2019). Bayesian fractional posteriors. *Ann. Statist.* **47** 39–66. MR3909926 https://doi.org/10.1214/18-AOS1712

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129 https://doi.org/10.1007/BF00532480

BIRGÉ, L. (1984). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. Henri Poincaré Probab. Stat.* **20** 201–223. MR0762855

BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **71** 271–291. MR0816706 https://doi.org/10.1007/BF00332312

BIRGÉ, L. (2006a). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **42** 273–325. MR2219712 https://doi.org/10.1016/j.anihpb.2005.04.004

BIRGÉ, L. (2006b). Statistical estimation with model selection. *Indag. Math. (N.S.)* **17** 497–537. MR2320111 https://doi.org/10.1016/S0019-3577(07)00004-3

BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1103–1130. MR3557191 https://doi.org/10.1111/rssb.12158

CATONI, O. (2007). *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. MR2483528

CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346. MR1984779 https://doi.org/10.1016/S0304-4076(03)00100-3

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **303**. Springer, Berlin. MR1261635 https://doi.org/10.1007/978-3-662-02888-9

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation: Asymptotic Theory. Applications of Mathematics* **16**. Springer, New York–Berlin. Translated from the Russian by Samuel Kotz. MR0620321

JIANG, W. and TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36** 2207–2231. MR2458185 https://doi.org/10.1214/07-AOS547

KLEIJN, B. J. K. and VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395 https://doi.org/10.1214/009053606000000029

KLEIJN, B. J. K. and VAN DER VAART, A. W. (2012). The Bernstein–Von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381. MR2988412 https://doi.org/10.1214/12-EJS675

LE CAM, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic Processes and Related Topics* (*Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind.*, 1974, *Vol.* 1; *Dedicated to Jerzy Neyman*) 13–54. MR0395005

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics*. Springer, New York. MR0856411 https://doi.org/10.1007/978-1-4612-4946-7

LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879

PANOV, M. and SPOKOINY, V. (2015). Finite sample Bernstein–von Mises theorem for semiparametric problems. *Bayesian Anal.* **10** 665–710. MR3420819 https://doi.org/10.1214/14-BA926

VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory. Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge Univ. Press, Cambridge. MR1739079

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2