Probability Theory and Related Fields (2024) 188:159–234
https://doi.org/10.1007/s00440-023-01222-8

From robust tests to Bayes-like posterior distributions

Yannick Baraud[1]

Received: 3 July 2022 / Revised: 25 April 2023 / Accepted: 16 June 2023 /
Published online: 13 July 2023

Abstract

In the Bayes paradigm, given a loss function and an n-sample, we present the construction of a new type of posterior distribution, that extends the classical Bayes one. The loss functions we have in mind are either those derived from the total variation and Hellinger distances or some $\mathbb{L}_j$-ones for $j > 1$. We prove that, with a probability close to one, this new posterior distribution concentrates its mass in a neighbourhood (for the chosen loss function) of the law of the data, provided that this law belongs to the support of the prior or, at least, lies close enough to it. We therefore establish that the new posterior distribution enjoys some robustness properties with respect to a possible misspecification of the prior, or more precisely, its support. We also show that the posterior distribution is stable with respect to the equidistribution assumption we started from. Besides, when the model is regular and well-specified and one uses the squared Hellinger loss, we show that our credible regions possess, at least for n sufficiently large, the same ellipsoidal shapes and approximately the same sizes as those we would derive from the classical Bayesian posterior distribution by using the Bernstein–von Mises theorem. Then we use our Bayesian-like approach to solve the following problems. We first consider the estimation of a location parameter or both the location and scale parameters of a density in a nonparametric framework. Then we tackle the problem of estimating a density, with the squared Hellinger loss, in a high-dimensional parametric model under some sparsity conditions on the parameter. Importantly, the results established in this paper are nonasymptotic and provide, as much as possible, bounds with explicit constants.

Keywords Bayes procedure · Gibbs estimator · Posterior distribution · Robustness · Hellinger distance · Total variation distance

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 811017.

✉ Yannick Baraud
yannick.baraud@uni.lu

[1] Department of Mathematics, University of Luxembourg, Maison du nombre, 6 Avenue de la Fonte, 4364 Esch-sur-Alzette, Grand Duchy of Luxembourg

Mathematics Subject Classification Primary 62G05 · 62G35 · 62F35 · 62F15

1 Introduction

Observe n i.i.d. random variables $X_1, \ldots, X_n$ with values in a measurable space $(E, \mathcal{E})$ and assume that their common distribution $P^\star$ belongs to a family $\mathscr{M}$ of candidate probabilities, or at least lies close enough to it in a suitable sense. We consider the problem of estimating $P^\star$ from the observation of $\boldsymbol{X} = (X_1, \ldots, X_n)$ and we evaluate the performance of an estimator with values in $\mathscr{M}$ by means of a given loss function $\ell : \mathscr{P} \times \mathscr{M} \to \mathbb{R}_+$, where $\mathscr{P}$ denotes a set of probabilities containing $P^\star$.

Our approach to solve this estimation problem has a Bayesian flavour. We endow $\mathscr{M}$ with a σ-algebra $\mathcal{A}$ and a probability measure π that plays the same role as the prior in the classical Bayes paradigm. Our aim is to design a posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$, solely based on $\boldsymbol{X}$ and the choice of ℓ, that concentrates its mass, with a probability close to one, on an ℓ-*ball,* namely a set of the form

$$\mathscr{B}(P^\star, r) = \left\{P \in \mathscr{M},\ \ell(P^\star, P) \leq r\right\} \quad \text{with} \quad r > 0. \tag{1}$$

This means that with a probability close to 1, a point $\widehat{P}$ which is randomly drawn according to our (random) distribution $\widehat{\pi}_{\boldsymbol{X}}$ is likely to estimate $P^\star$ with an accuracy (with respect to the chosen loss ℓ) not larger than r. Our objective is to design $\widehat{\pi}_{\boldsymbol{X}}$ in such a way that this concentration property holds for small values of r and under mild assumptions on $P^\star$ and $\mathscr{M}$.

In the literature, many authors have studied the concentration properties of the classical Bayes posterior distribution on Hellinger balls. We refer to the pioneering papers by van der Vaart and his co-authors—see for example Ghosal, Ghosh and van der Vaart [19]. They show that the concentration property around $P^\star$ holds, as n tends to infinity, provided that the prior π puts enough mass on sets of the form $\mathcal{K}(P^\star, \varepsilon) = \{P \in \mathscr{M},\ K(P^\star, P) < \varepsilon\}$ where ε is a positive number and $K(P^\star, P)$ the Kullback–Leibler divergence between $P^\star$ and P. This assumption may, however, be quite restrictive even in the favorable situation where $P^\star$ belongs to the model $\mathscr{M}$. Such sets may indeed be empty, and the condition therefore unsatisfied, when the probabilities in $\mathscr{M}$ are not equivalent. This is for example the case when $\mathscr{M}$ is the set of all uniform distributions P_θ on $[\theta - 1/2, \theta + 1/2]$, with $\theta \in \mathbb{R}$, although the problem of estimating $P^\star \in \mathscr{M}$ in this setting is quite easy, even in the Bayesian paradigm. The assumption appears even more restrictive when the probability $P^\star$ does not belong to $\mathscr{M}$, that is when the model is misspecified. For example, if the distributions in $\mathscr{M}$ are all equivalent and R is singular with respect to $\overline{P} \in \mathscr{M}$, $\mathcal{K}(P^\star, \varepsilon)$ is empty for $P^\star = (1 - 10^{-10})\overline{P} + 10^{-10}R$ although $P^\star$ and $\overline{P} \in \mathscr{M}$ are statistically indistinguishable from any n-sample of realistic size.

Unfortunately, it is in general impossible to get rid of the restrictive conditions we have mentioned above. It is well known that the Bayes posterior distribution can be unstable in case of a misspecification of the model. Examples that illustrate this weakness have been given in Jiang and Tanner [21] and Baraud and Birgé [6] for instance. This instability is due to the fact that the Bayes posterior distribution is

based on the log-likelihood function and similar issues are known for the maximum likelihood estimator.

In order to obtain the concentration and stability properties we look for, we replace the log-likelihood function by a more stable one. Substituting another function to the log-likelihood one is not new in the literature and leads to what is called *quasi-posterior distributions*. The resulting estimators, called *quasi-Bayesian estimators* or *Laplace type estimators*, have been studied by various statisticians among which Chernozhukov and Hong [18] and Bissiri et al. [16] (we also refer to the references therein). These papers, however, do not address the problem of misspecification. In contrast, it is addressed in Jiang and Tanner [21] for performing variable selection in the logistic model. The authors show that the classical Bayesian approach is no longer reliable when the model is slightly misspecified while their Gibbs posterior distribution performs well and offers thus a much safer alternative. The problem of estimating a high-dimensional parameter $\theta \in \mathbb{R}^d$ under a sparsity condition was considered in Atchadé [2]. His quasi-posterior distribution is obtained by replacing the joint density of the data by a more suitable one and by using some specific prior that forces sparsity. He proves that the so-defined posterior distribution contracts around the true parameter $\theta^\star$ at rate $\sqrt{(s^\star \log d)/n}$ (where $s^\star$ is the number of nonzero coordinates of $\theta^\star$) when both d and n tend to infinity. A common feature of the papers we have cited above lies in their asymptotic nature. This is not the case for Bhattacharya et al. [8] who replaced the likelihood function in the expression of the posterior distribution by the *fractional likelihood*, that is a suitable power of the likelihood function. The authors also consider the situation where the model is possibly misspecified but their result involves the α-divergence which, as the Kullback one, can be infinite even when the true distribution of the data is close to the model for the total variation distance or the Hellinger one.

Baraud and Birgé [6] propose a surrogate to the Bayes posterior distribution that is called the *ρ-posterior distribution* in reference to the theory of ρ-estimation that was developed in Baraud et al. [7] and Baraud and Birgé [5]. In the frequentist paradigm, this theory aimed at solving the various problems connected to the instability of the maximum likelihood method. The ρ-posterior distribution preserves some of the nice features of the classical Bayes one but also possesses the robustness property we are interested in. The authors show that their posterior distribution concentrates on a Hellinger ball around $P^\star$ as soon as the prior puts enough mass around a point which is close enough to $P^\star$. However their approach applies to specific dominated models $\mathscr{M} = \{P = p \cdot \mu,\ p \in \mathcal{M}\}$ only. They assume that the family $\mathcal{M}$ of densities that defines their model possesses some special combinatorial structure which is either met when $\mathcal{M}$ is finite or when it satisfies some VC-type condition (see their Section 5). As a consequence, the concentration radius they obtain not only depends on the choice of the prior but also on a complexity term that is linked to this structure. Unlike theirs, our approach makes no such assumptions on $\mathcal{M}$ and we are therefore able to get rid of this unpleasant complexity term while retaining a similar dependency with respect to the choice of the prior. Baraud and Birgé's posterior distribution has also the drawback to involve the supremum over the family $\mathcal{M}$ of an empirical process. Their posterior distribution is therefore difficult to calculate in practice, unless $\mathcal{M}$ is finite with a reasonable size. From a more theoretical point of view, it also raises some

unpleasant issues with regard to the measurability of this supremum in the situation where the family $\mathcal{M}$ is uncountable, which is the typical case. Finally, Baraud and Birgé's approach restricts to the squared Hellinger loss while ours applies to many others.

Closer to our approach are the aggregation methods and PAC-Bayesian techniques that have been popularized by Olivier Catoni in statistical learning (see Catoni [17]). This approach has mainly been applied for the purpose of empirical risk minimization and statistical learning (see for example Alquier [1]). Our aim is to extend these techniques toward a versatile tool that can solve our Bayes-like estimation problem for various loss functions simultaneously.

The problem of designing a good estimator of $P^\star$ for a given loss function ℓ was tackled in the frequentist paradigm in Baraud [4]. There, the author provides a general framework that enables one to deal with various loss functions of interest, among which the total variation, 1-Wasserstein, Hellinger, and $\mathbb{L}_j$-losses among others. His approach relies on the construction of a suitable family of robust tests and lies in the line of the former work of Le Cam [22], Birgé [9] and Birgé [11]. The aim of the present paper is to transpose this theory from the frequentist to the Bayesian paradigm. If ℓ is the Kullback–Leibler divergence, our construction recovers the classical Bayes posterior distribution even though this is not the choice we would recommend for the reasons we have explained before.

Quite surprisingly, the concentration properties that we establish here require almost no assumption on $\mathcal{M}$ and the distribution of the data (apart from independence). They mostly depend on the choices of the prior π and the loss function ℓ. For a suitable element P which belongs to the model $\mathcal{M}$ and lies close enough to $P^\star$, these concentration properties depend on the minimal value of the radius r over which the log-ratio $V(P,r) = \log[\pi(\mathcal{B}(P,2r))/\pi(\mathcal{B}(P,r))]$ (with $\mathcal{B}$ defined in (1)) increases at most linearly with r. This log-ratio was introduced in Birgé [12] for the purpose of analyzing the behaviour of the classical Bayes posterior distribution. In our Bayes-like paradigm, we show that the behaviour of the quantities $V(P,r)$ for $P \in \mathcal{M}$ and $r > 0$ completely encapsulates the complexity of the model $\mathcal{M}$. We prove that our posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ concentrates on an ℓ-ball centered at $P^\star$ and the radius $r = r(n)$ of which is usually of minimax order as n tends to infinity when the model is well-specified. From a nonasymptotic point of view, we show that $\widehat{\pi}_{\boldsymbol{X}}$ retains its nice concentration properties as long as $P^\star$ remains close enough to an element P in $\mathcal{M}$ around which the prior puts enough mass, that is, even in the situation where the model is slightly misspecified. Actually, we establish the stronger result that even when the data are only independent but not i.i.d., the above conclusion remains true for the average $\overline{P^\star}$ of their marginal distributions in place of $P^\star$. We therefore show that the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ enjoys some robustness properties with respect to the equidistribution assumption we started from. The main theorems involve as much as possible explicit numerical constants. We illustrate our results with examples which are deliberately chosen to be as general and simple as possible. Our aim is to give a flavour of the results that can be established with our Bayes-like posterior, avoiding as much as possible the technicalities that would result from the choice of *ad-hoc* priors introduced to solve specific problems. Instead, we wish to discuss the optimality and robustness properties of our construction for solving general parametric and nonparametric esti-

mation problems in the density framework under assumptions that we wish to be as weak as possible. These posterior distributions will therefore provide a benchmark for comparison with other methods. Their practical implementation will be the subject of future work.

Of special interest is the choice of ℓ given by the total variation distance or the Hellinger one. As we shall see, for such losses the stability of our posterior distribution automatically leads to estimators $\widehat{P} \sim \widehat{\pi}_{\boldsymbol{X}}$ that are naturally robust to the presence of outliers or contaminating data among the sample. These results contrast sharply with the instability of the classical Bayes posterior distribution we underlined earlier. Nevertheless, our posterior distribution also shares some similarities with the classical Bayes one. When the model is well-specified and one uses the squared Hellinger loss, we show that the credible regions of our posterior distribution asymptotically possess the same ellipsoidal shapes and approximately the same sizes as the ones we derive from the classical Bayes posterior by means of the Bernstein–von Mises theorem. Establishing an analogue of this theorem for our Bayes-like posterior distribution is, however, beyond the scope of the present paper.

Our paper is organized as follows. We present our statistical setting in Sect. 2. We consider there independent but not necessarily i.i.d. data in order to analyse later on the behaviour of our posterior distribution with respect to a possible departure from equidistribution. The construction of the posterior distribution is described in Sect. 3. In this section, we also show how more classical constructions based on the likelihood or the fractional likelihoods are particular cases of ours. We complete this section with some heuristics which, we hope, help understanding the main ideas of our approach. In particular, we bridge there the problem of designing robust posterior distributions to that of testing between two disjoint ℓ-balls. Section 4 is devoted to the main theorems. We describe there the concentration properties of our posterior distribution. The applications of these results to classical loss functions are presented in Sect. 5. We put a special emphasis on the cases of the total variation distance and the squared Hellinger loss. In the remaining part of the paper, we only focus on these two losses. In Sect. 6 we highlight some similarities and differences between the classical Bayes posterior and ours for the squared Hellinger loss. In Sect. 7 we explain how our posterior distribution can be used to solve the problem of estimating a density, or a parameter associated with it, in several statistical frameworks of interest. We discuss there how the concentration properties of our posterior distribution deteriorate in the case of a misspecification of the model by the prior. We also consider the problems of estimating a density in a location-scale family and a high-dimensional parameter in a parametric model under a sparsity constraint. We also show how our estimation strategy leads to unusual rates of convergence for estimating a translation parameter in a non-regular statistical model. In Sect. 8, we provide an evaluation of the concentration radius of our posterior distributions in the parametric framework. Finally, Sect. 9 is devoted to the proofs of the main theorems and Sect. 10 to the other proofs.

2 The statistical setting

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an n-tuple of independent random variables with values in a measurable space $(E, \mathcal{E})$ and joint distribution $\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star$. Even though this might not be true, we pretend that the X_i are i.i.d. and our aim is to estimate their (presumed) common distribution $P^\star$ from the observation of $\boldsymbol{X}$. To do so, we introduce a family $\mathscr{M}$ that consists of candidate probabilities (or merely finite signed measures in the case of the $\mathbb{L}_j$-loss). The reason for considering finite signed measures lies in the fact that statisticians sometimes estimate probability densities by integrable functions that are not necessarily densities but elements of a suitable linear space for instance (think of the case of projection estimators). We endow $\mathscr{M}$ with a σ-algebra $\mathcal{A}$ and a probability measure π, that we call *a prior* by analogy to the classical Bayesian framework, and we refer to the resulting pair $(\mathscr{M}, \pi)$ as our *model*. The model $(\mathscr{M}, \pi)$ plays here a similar role as in the classical Bayes paradigm. It encapsulates the *a priori* information that the statistician has on $P^\star$. Nevertheless, we do not assume that $P^\star$, if it ever exists, belongs to $\mathscr{M}$ nor that the true marginals $P_i^\star$ do. We rather assume that the model $(\mathscr{M}, \pi)$ is approximately correct in the sense that the average distribution

$$\overline{P}^\star = \frac{1}{n}\sum_{i=1}^n P_i^\star$$

is close enough to some point P in $\mathscr{M}$ around which the prior π puts enough mass. We assume that $\overline{P}^\star$ belongs to a given set $\mathscr{P}$ of probability measures on $(E, \mathcal{E})$ and we measure the estimation accuracy by means of a loss function $\ell : (\mathscr{M} \cup \mathscr{P}) \times \mathscr{M} \to \mathbb{R}_+$ which is not identical to 0 in order to avoid trivialities. Even though ℓ may not be a genuine distance in general, we assume that it shares some similar features and we interpret it as if it were. For this reason, we call *ℓ-ball* (or *ball* for short) centered at $P \in \mathscr{P} \cup \mathscr{M}$ with radius $r > 0$ the subset of $\mathscr{M}$

$$\mathscr{B}(P, r) = \{Q \in \mathscr{M},\ \ell(P, Q) \leq r\}.$$

Our aim is to built *a posterior distribution* (or posterior for short) $\widehat{\pi}_{\boldsymbol{X}}$ on $(\mathscr{M}, \mathcal{A})$, depending on our observation $\boldsymbol{X}$, which concentrates with a probability close to 1 on an ℓ-ball of the form $\mathscr{B}(\overline{P}^\star, r_n)$ where we wish the value of $r_n > 0$ to be small.

2.1 The special case of parametrized models

In many situations we consider statistical models $\mathscr{M} = \{P_\theta,\ \theta \in \Theta\}$ which are parametrized via a one-to-one mapping $\theta \mapsto P_\theta$. When $(\Theta, \mathfrak{B}, \nu)$ is a measurable space, we endow $\mathscr{M}$ with the σ-algebra $\mathcal{A} = \{A,\ \{\theta \in \Theta,\ P_\theta \in A\} \in \mathfrak{B}\}$. This choice possesses several advantages. First, the mapping $\theta \mapsto P_\theta$ is measurable from $(\Theta, \mathfrak{B})$ onto $(\mathscr{M}, \mathcal{A})$ and we may therefore define the prior π on $(\mathscr{M}, \mathcal{A})$ as the image of ν by this mapping. Besides, a function F is measurable on $(\mathscr{M}, \mathcal{A})$ if and only if the mapping $\theta \mapsto F \circ P_\theta$ is measurable on $(\Theta, \mathfrak{B})$. This property makes the measurability of F easier to check in general. In particular, the mapping $F : P_\theta \mapsto \theta$ is measurable on $(\mathscr{M}, \mathcal{A})$ because $\theta \mapsto F \circ P_\theta = \theta$ is measurable on $(\Theta, \mathfrak{B})$ and we may then define a posterior $\widehat{\nu}_{\boldsymbol{X}}$ on $(\Theta, \mathfrak{B})$ as the image by F of our posterior $\widehat{\pi}_{\boldsymbol{X}}$ on

$(\mathscr{M}, \mathcal{A})$. By definition of $\widehat{\nu}_{\boldsymbol{X}}$, for all $\theta \in \Theta$ and $r > 0$

$$\widehat{\pi}_{\boldsymbol{X}}\left(\mathscr{B}(P_\theta, r)\right) = \widehat{\nu}_{\boldsymbol{X}}\left(\left\{\theta' \in \Theta,\ \ell(\theta, \theta') \leq r\right\}\right) \tag{2}$$

where $\ell(\theta, \theta')$ denotes, slightly abusively, $\ell(P_\theta, P_{\theta'})$ for $\theta, \theta' \in \Theta$. The concentration of $\widehat{\pi}_{\boldsymbol{X}}$ on an ℓ-ball centered at P_θ with radius $r > 0$ is then equivalent to the concentration of $\widehat{\nu}_{\boldsymbol{X}}$ on the set $\{\theta' \in \Theta,\ \ell(\theta, \theta') \leq r\}$. Every time we consider a parametrized model, we assume that it is identifiable and implicitly use the construction that we presented above as well as its consequences.

2.2 Notation and conventions

Throughout this paper, we use the following notation and conventions. For $a, b \in \mathbb{R}$, $a \vee b$ and $a \wedge b$ denote $\min\{a, b\}$ and $\max\{a, b\}$ respectively. For $x \in \mathbb{R}$, $(x)_+ = x \vee 0$ while $(x)_- = (-x) \vee 0$. The Euclidean spaces $\mathbb{R}^k$ with $k \geq 1$ are equipped with their Borel σ-algebras. The cardinality of a set A is denoted $|A|$ and its complement ${}^{\complement}A$. In particular, for $P \in \mathscr{P} \cup \mathscr{M}$ and $r > 0$, ${}^{\complement}\mathscr{B}(P, r) = \{Q \in \mathscr{M},\ \ell(P, Q) > r\}$. The elements of $\mathbb{R}^k$ with $k > 1$ are denoted with bold letters, e.g. $\boldsymbol{x} = (x_1, \ldots, x_k)$ and $\mathbf{0} = (0, \ldots, 0)$. For $\boldsymbol{x} \in \mathbb{R}^k$, $|\boldsymbol{x}|_\infty = \max_{i \in \{1,\ldots,k\}} |x_i|$ while $|\boldsymbol{x}|$ denotes the Euclidean norm of $\boldsymbol{x}$. The inner product of $\mathbb{R}^k$ is denoted by $\langle \cdot, \cdot \rangle$ and the closed Euclidean ball centered at $\boldsymbol{x}$ with radius $r \geq 0$ by $\mathcal{B}(\boldsymbol{x}, r)$. By convention $\inf_\varnothing = +\infty$ unless otherwise specified. We write $f \equiv c$ when a function f is constant and equals c on its domain. For all suitable functions f on $(E^n, \mathcal{E}^{\otimes n})$, $\mathbb{E}\left[f(\boldsymbol{X})\right]$ means $\int_{E^n} f d\mathbf{P}^\star$ while for f on $(E, \mathcal{E})$, $\mathbb{E}_S\left[f(X)\right]$ denotes the integral $\int_E f dS$ with respect to the measure S on $(E, \mathcal{E})$. For $j \in [1, +\infty)$, we denote by $\mathscr{L}_j(E, \mathcal{E}, \mu)$, the set of measurable functions f on $(E, \mathcal{E})$ such that $\|f\|_{j,\mu} = [\int_E |f|^j d\mu]^{1/j} < +\infty$ while $\|f\|_\infty = \sup_{x \in E} |f(x)|$ is the supremum norm of a function f on E. If π' is a distribution on $(\mathscr{M}, \mathcal{A})$, $Q \sim \pi'$ means that Q is a random variable with distribution π'. Finally, all the measures that we consider are implicitly assumed to be σ-finite.

3 Construction of the posterior distribution

Throughout this section, the model $(\mathscr{M}, \pi)$ is assumed to be fixed.

3.1 The properties of our loss functions

The construction of the posterior not only depends on the prior π but also on the choice of the loss function. We first assume that ℓ satisfies some basic properties which are described below.

Assumption 1 For all $S \in \mathscr{P} \cup \mathscr{M}$, the mapping

$$\ell(S, \cdot) : \left| \begin{array}{rcl} (\mathscr{M}, \mathcal{A}) & \longrightarrow & \mathbb{R}_+ \\ P & \longmapsto & \ell(S, P) \end{array} \right.$$

is measurable.

Under such an assumption, ℓ-balls are measurable and the quantities $\pi(\mathscr{B}(P, r))$ for $P \in \mathscr{P} \cup \mathscr{M}$ and $r > 0$ are therefore well-defined.

Assumption 2 There exists a positive number τ such that, for all $S \in \mathscr{P}$ and $P, Q \in \mathscr{M}$,

$$\ell(S, Q) \leq \tau \left[\ell(S, P) + \ell(P, Q)\right] \tag{3}$$

$$\ell(S, Q) \geq \tau^{-1}\ell(P, Q) - \ell(S, P). \tag{4}$$

When ℓ is a genuine distance, inequalities (3) and (4) are satisfied with $\tau = 1$ since they correspond to the triangle inequality. When ℓ is the square of a distance, these inequalities are satisfied with $\tau = 2$.

Importantly, we assume that ℓ is associated with a family $\mathscr{T}(\ell, \mathscr{M}) = \left\{t_{(P,Q)}, (P, Q) \in \mathscr{M}^2\right\}$ of test statistics on $(E, \mathcal{E})$ which possesses the properties below. We shall see in Sect. 5 that many classical loss functions (among which the total variation distance, the squared Hellinger distance, etc.) can be associated with families $\mathscr{T}(\ell, \mathscr{M})$ satisfying the following assumptions.

Assumption 3 The elements $t_{(P,Q)}$ of $\mathscr{T}(\ell, \mathscr{M})$ satisfy:

(i) The mapping

$$t : \left| \begin{array}{l} (E \times \mathscr{M} \times \mathscr{M}, \mathcal{E} \otimes \mathcal{A} \otimes \mathcal{A}) \longrightarrow \mathbb{R} \\ \qquad\qquad (x, P, Q) \longmapsto t_{(P,Q)}(x) \end{array} \right.$$

is measurable.

(ii) For all $P, Q \in \mathscr{M}$, $t_{(P,Q)} = -t_{(Q,P)}$.

(iii) there exist positive numbers a_0, a_1 such that, for all $S \in \mathscr{P}$ and $P, Q \in \mathscr{M}$,

$$\mathbb{E}_S\left[t_{(P,Q)}(X)\right] \leq a_0\ell(S, P) - a_1\ell(S, Q). \tag{5}$$

(iv) For all $P, Q \in \mathscr{M}$,

$$\sup_{x \in E} t_{(P,Q)}(x) - \inf_{x \in E} t_{(P,Q)}(x) \leq 1.$$

Under assumption (ii), $t_{(P,P)} = 0$ and we deduce from (5) that $(a_0 - a_1)\ell(S, P) \geq 0$, hence that $a_0 \geq a_1$ since ℓ is not constantly equal to 0.

Some families $\mathscr{T}(\ell, \mathscr{M})$ may satisfy the stronger

Assumption 4 Additionally to Assumption 3, there exists $a_2 > 0$ such that

(iv) for all $S \in \mathscr{P}$ and $P, Q \in \mathscr{M}$,

$$\mathrm{Var}_S\left[t_{(P,Q)}(X)\right] \leq a_2\left[\ell(S, P) + \ell(S, Q)\right].$$

3.2 Construction of the posterior

Let $\mathscr{T}(\ell, \mathscr{M})$ be a family of test statistics that satisfies our Assumption 3 and let β and λ be two positive numbers such that

$$\lambda = (1+c)\beta \quad \text{with} \quad c > 0 \ \text{ satisfying } \ c_0 = (1+c) - c(a_0/a_1) > 0. \tag{6}$$

We set

$$\mathbf{T}(\boldsymbol{X}, P, Q) = \sum_{i=1}^{n} t_{(P,Q)}(X_i) \quad \text{for all } P, Q \in \mathscr{M}$$

and define $\widetilde{\pi}_{\boldsymbol{X}}(\cdot|P)$ as the probability on $(\mathscr{M}, \mathcal{A})$ with density

$$\frac{d\widetilde{\pi}_{\boldsymbol{X}}(\cdot|P)}{d\pi} : Q \mapsto \frac{\exp\left[\lambda \mathbf{T}(\boldsymbol{X}, P, Q)\right]}{\int_{\mathscr{M}} \exp\left[\lambda \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}.$$

Then, for $P \in \mathscr{M}$ we set

$$\begin{aligned}\mathbf{T}(\boldsymbol{X}, P) &= \int_{\mathscr{M}} \mathbf{T}(\boldsymbol{X}, P, Q) d\widetilde{\pi}_{\boldsymbol{X}}(Q|P) \\ &= \int_{\mathscr{M}} \mathbf{T}(\boldsymbol{X}, P, Q) \frac{\exp\left[\lambda \mathbf{T}(\boldsymbol{X}, P, Q)\right]}{\int_{\mathscr{M}} \exp\left[\lambda \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)} d\pi(Q).\end{aligned}$$

Finally, we define $\widehat{\pi}_{\boldsymbol{X}}$ as the posterior distribution on $(\mathscr{M}, \mathcal{A})$ with density

$$\frac{d\widehat{\pi}_{\boldsymbol{X}}}{d\pi} : P \mapsto \frac{\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]}{\int_{\mathscr{M}} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P)}. \tag{7}$$

Our Assumption 3–(i) ensures that $d\widetilde{\pi}_{\boldsymbol{X}}(\cdot|P)/d\pi$ is a measurable function of $(\boldsymbol{X}, P, Q)$ and $d\widehat{\pi}_{\boldsymbol{X}}/d\pi$ a measurable function of $(\boldsymbol{X}, P)$.

The posterior distribution depends on our choice of β and λ (or equivalently c) even though we drop this dependency with the notation $\widehat{\pi}_{\boldsymbol{X}}$.

3.3 Monte Carlo computation of functions of the posterior

Even though we focus on the concentration properties of the posterior $\widehat{\pi}_{\boldsymbol{X}}$, one may alternatively be interested in some estimators derived from it. For example, estimators of the form

$$I = \int_{\mathscr{M}} F(P) d\widehat{\pi}_{\boldsymbol{X}}(P)$$

where F is a real-valued π-integrable function on $(\mathscr{M}, \mathcal{A})$. For typical choices of F, I gives the expected mean, mode or median of the posterior whenever these quantities

make sense. One may also choose $F : P \mapsto \mathbb{1}_{P \in \mathscr{B}(P_0,\varepsilon)}$ with $P_0 \in \mathscr{M}$ and $\varepsilon > 0$ in order to compute the (posterior) probability that $\ell(P_0, \widehat{P})$ is not larger than ε when $\widehat{P} \sim \widehat{\pi}_{\boldsymbol{X}}$.

Interestingly, the integral I can be approximated by Monte Carlo as follows. Assume that the prior π admits a density of the form $C^{-1}\Pi$ with respect to a given probability measure $\mathfrak{m}$, where Π is a nonnegative $\mathfrak{m}$-integrable function on $(\mathscr{M}, \mathcal{A})$ and $C = \int_{\mathscr{M}} \Pi(P) d\mathfrak{m}(P) > 0$ a positive normalizing constant (that will not be involved in our calculation). Let $P_1, \ldots, P_N$ be an N-sample with distribution $\mathfrak{m}$ and for each $i \in \{1, \ldots, N\}$, $Q_i^{(1)}, \ldots, Q_i^{(N')}$ an independent N'-sample with the same distribution. We may approximate I by

$$\widehat{I}_{N,N'} = \sum_{i=1}^{N} F(P_i) \frac{\exp\left[-\beta W_{i,N'}(P_i)\right] \Pi(P_i)}{\sum_{i'=1}^{N} \exp\left[-\beta W_{i',N'}(P_{i'})\right] \Pi(P_{i'})}$$

where for all $i \in \{1, \ldots, N\}$,

$$W_{i,N'}(P_i) = \sum_{j=1}^{N'} T(\boldsymbol{X}, P_i, Q_i^{(j)}) \frac{\exp\left[\lambda T(\boldsymbol{X}, P_i, Q_i^{(j)})\right] \Pi(Q_i^{(j)})}{\sum_{j'=1}^{N'} \exp\left[\lambda T(\boldsymbol{X}, P_i, Q_i^{(j')})\right] \Pi(Q_i^{(j')})}.$$

It is then easy to check that, by the law of large numbers,

$$\lim_{N \to +\infty} \left[\lim_{N' \to +\infty} \widehat{I}_{N,N'} \right] = I.$$

3.4 Connection with the classical Bayes posterior distribution

The classical Bayes posterior turns out to be a particular case of the posterior-type ones introduced in Sect. 3.2. As we shall see now, they are associated with the Kullback–Leibler divergence loss. We recall that the Kullback–Leibler divergence $\ell(P, Q) = K(P, Q)$ between two probabilities P, Q on $(E, \mathcal{E})$ is defined by

$$K(P, Q) = \begin{cases} \displaystyle\int_E \log\left(\frac{dP}{dQ}\right) dP & \text{when } P \ll Q; \\ +\infty & \text{otherwise.} \end{cases}$$

Let us consider now a family $\mathscr{M}$ of probabilities that satisfy for some $a > 0$ and suitable versions of their densities dQ/dP the following inequalities:

$$e^{-a} \leq \frac{dP}{dQ}(x) \leq e^{a} \quad \text{for all } x \in E \text{ and } P, Q \in \mathscr{M}. \tag{8}$$

It follows from Baraud [4, Proposition 12] that the families of functions

$$\mathscr{T}(\ell, \mathscr{M}) = \left\{ t_{(P,Q)} = \frac{1}{2a} \log\left(\frac{dQ}{dP}\right),\ P, Q \in \mathscr{M} \right\} \tag{9}$$

satisfies our Assumptions 3 and 4 with $a_0 = a_1 = 1/(2a)$ and $a_2 = 2a/[\tanh(a/2)]$. Note that given $P, Q \in \mathscr{M}$, $P \neq Q$, the test based on the sign of $t_{(P,Q)}$ is the classical likelihood ratio test between P and Q.

If we apply the construction described in Sect. 3.2 to the family $\mathscr{T}(\ell, \mathscr{M})$ we obtain that for all $P, Q, P_0 \in \mathscr{M}$,

$$\mathbf{T}(\boldsymbol{X}, P, Q) = \mathbf{T}(\boldsymbol{X}, P_0, Q) - \mathbf{T}(\boldsymbol{X}, P_0, P).$$

For all $\lambda > 0$, the density of $\widetilde{\pi}_{\boldsymbol{X}}(\cdot|P)$

$$Q \mapsto \frac{\exp[\lambda \mathbf{T}(\boldsymbol{X}, P, Q)]}{\int_{\mathscr{M}} \exp[\lambda \mathbf{T}(\boldsymbol{X}, P, Q)]\, d\pi(Q)} = \frac{\exp[\lambda \mathbf{T}(\boldsymbol{X}, P_0, Q)]}{\int_{\mathscr{M}} \exp[\lambda \mathbf{T}(\boldsymbol{X}, P_0, Q)]\, d\pi(Q)}$$

is independent of P and writing $\widetilde{\pi}_{\boldsymbol{X}}(\cdot)$ in place of $\widetilde{\pi}_{\boldsymbol{X}}(\cdot|P)$ we obtain that

$$\begin{aligned}
\mathbf{T}(\boldsymbol{X}, P) &= \int_{\mathscr{M}} \mathbf{T}(\boldsymbol{X}, P, Q) d\widetilde{\pi}_{\boldsymbol{X}}(Q) \\
&= \int_{\mathscr{M}} \mathbf{T}(\boldsymbol{X}, P_0, Q) d\widetilde{\pi}_{\boldsymbol{X}}(Q) - \mathbf{T}(\boldsymbol{X}, P_0, P) \\
&= C - \frac{1}{2a} \sum_{i=1}^{n} \log\left(\frac{dP}{dP_0}\right)(X_i) \text{ with } C = \int_{\mathscr{M}} \mathbf{T}(\boldsymbol{X}, P_0, Q) d\widetilde{\pi}_{\boldsymbol{X}}(Q).
\end{aligned}$$

Finally, the density of our posterior $\widehat{\pi}_{\boldsymbol{X}}$ at $P \in \mathscr{M}$ is given by

$$\frac{d\widehat{\pi}_{\boldsymbol{X}}}{d\pi}(P) = \frac{\exp[-\beta \mathbf{T}(\boldsymbol{X}, P)]}{\int_{\mathscr{M}} \exp[-\beta \mathbf{T}(\boldsymbol{X}, P)]\, d\pi(P)} = \frac{\left[\prod_{i=1}^{n} (dP/dP_0)(X_i)\right]^{\beta/(2a)}}{\int_{\mathscr{M}} \left[\prod_{i=1}^{n} (dP/dP_0)(X_i)\right]^{\beta/(2a)} d\pi(P)}.$$

This is the density of the classical Bayes posterior when $\beta = 2a$ while for other values of β it is that of fractional Bayes ones.

Nevertheless, in the present paper we restrict our study to loss functions that satisfy some triangle-type inequality – see Assumption 2. This excludes the Kullback–Leibler divergence unless one is ready to make strong assumptions on the unknown distribution of the data, which we do not want to do here.

3.5 Some heuristics

In this section, we present the basic ideas that underline our approach. In particular, we shall see how the estimation problem we want to solve is linked to the one of testing between two disjoint ℓ-balls $\mathscr{B}(P, r)$ and $\mathscr{B}(Q, r)$ with $P, Q \in \mathscr{M}$.

In order to avoid unnecessary details, we assume here that we observe i.i.d. data $X_1, \ldots, X_n$ with distribution $P^\star \in \mathscr{P}$ and that we have at disposal a family $\mathscr{T}(\ell, \mathscr{M})$ of functions that satisfies our Assumption 3. In particular it follows from Assumption 3-(iii) that

$$\mathbb{E}\left[\frac{\mathbf{T}(\boldsymbol{X}, P, Q)}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[t_{(P,Q)}(X_i)\right] \leq a_0\ell(P^\star, P) - a_1\ell(P^\star, Q).$$

The antisymmetric property required by Assumption 3-(ii) entails that

$$\mathbf{T}(\boldsymbol{X}, P, Q) = -\mathbf{T}(\boldsymbol{X}, Q, P)$$

and leads to the lower bound

$$\mathbb{E}\left[\frac{\mathbf{T}(\boldsymbol{X}, P, Q)}{n}\right] \geq a_1\ell(P^\star, P) - a_0\ell(P^\star, Q).$$

Assuming for the sake of simplicity that $a_0 = a_1 = 1$, these calculations show that $n^{-1}\mathbf{T}(\boldsymbol{X}, P, Q) = n^{-1}\sum_{i=1}^{n} t_{(P,Q)}(X_i)$ is an unbiased and consistent estimator of $\ell(P^\star, P) - \ell(P^\star, Q)$. In particular, if the two ℓ-balls $\mathscr{B}(P, r)$, $\mathscr{B}(Q, r)$ are disjoint and $P^\star$ belongs to one of them, the sign of $n^{-1}\mathbf{T}(\boldsymbol{X}, P, Q) = n^{-1}\sum_{i=1}^{n} t_{(P,Q)}(X_i)$ provides a consistent test for deciding which one contains $P^\star$. In fact, the test does not depend on the value of r and consequently chooses the element among $\{P, Q\}$ which is the closest to $P^\star$ (with respect to ℓ), at least when n is large enough. As compared to the classical likelihood ratio test between P and Q, this test has the advantage not to assume that $P^\star$ is either P or Q but only that it lies in a small enough ℓ-vicinity around one of these two probabilities. The test is said to be *robust* with respect to the model $\{P, Q\}$. Its nonasymptotic properties have been studied in Baraud [4].

Let us now explain how such families $\{\mathbf{T}(\boldsymbol{X}, P, Q), (P, Q) \in \mathscr{M}^2\}$ of test statistics can be used to build robust estimators and not only tests. In the frequentist paradigm, the construction of ℓ-estimators is based on the following heuristics. If, with a probability close to 1, $n^{-1}\mathbf{T}(\boldsymbol{X}, P, Q)$ is close to its expectation $\ell(P^\star, P) - \ell(P^\star, Q)$ uniformly with respect to $(P, Q) \in \mathscr{M}^2$ then $n^{-1}\mathbf{T}'(\boldsymbol{X}, P) = \sup_{Q\in\mathscr{M}}\left[n^{-1}\mathbf{T}(\boldsymbol{X}, P, Q)\right]$ is close to

$$\sup_{Q\in\mathscr{M}}\left[\ell(P^\star, P) - \ell(P^\star, Q)\right] = \ell(P^\star, P) - \inf_{Q\in\mathscr{M}}\ell(P^\star, Q).$$

We therefore expect that a minimizer over $\mathscr{M}$ of the function $P \in \mathscr{M} \mapsto n^{-1}\mathbf{T}'(\boldsymbol{X}, P)$ be close to a minimizer over $\mathscr{M}$ of the function $P \in \mathscr{M} \mapsto \ell(P^\star, P) - \inf_{Q\in\mathscr{M}}\ell(P^\star, Q)$, that is an element that minimizes the loss $\ell(P^\star, P)$ among the probabilities $P \in \mathscr{M}$.

In the Bayesian paradigm, we may argue in a similar way as follows. Replacing $n^{-1}\mathbf{T}(\boldsymbol{X}, P, Q)$ by its expectation $\ell(P^\star, P) - \ell(P^\star, Q)$, as we did before, amounts

to replacing $\mathbf{T}(\boldsymbol{X}, \mathbf{P})$ by

$$\begin{aligned}
&\overline{\mathbf{T}}(\boldsymbol{X}, \mathbf{P}) \\
&\quad = n \int_{\mathscr{M}} \big(\ell(P^\star, P) - \ell(P^\star, Q)\big) \frac{\exp\big[n\lambda\,(\ell(P^\star, P) - \ell(P^\star, Q))\big]\, d\pi(Q)}{\int_{\mathscr{M}} \exp\left[n\lambda\,(\ell(P^\star, P) - \ell(P^\star, Q))\right] d\pi(Q)} \\
&\quad = n\ell(P^\star, P) - n \int_{\mathscr{M}} \ell(P^\star, Q) \frac{\exp\big[-n\lambda\ell(P^\star, Q)\big]}{\int_{\mathscr{M}} \exp\left[-n\lambda\ell(P^\star, Q)\right] d\pi(Q)} d\pi(Q).
\end{aligned}$$

Note that the second term in the right-hand side does not depend on P. Consequently, replacing $\mathbf{T}(\boldsymbol{X}, \mathbf{P})$ by $\overline{\mathbf{T}}(\boldsymbol{X}, \mathbf{P})$ in the expression (7) of the density of $\widehat{\pi}_X$ leads to the density

$$P \mapsto \frac{\exp\big[-\beta\overline{\mathbf{T}}(\boldsymbol{X}, P)\big]}{\int_{\mathscr{M}} \exp\big[-\beta\overline{\mathbf{T}}(\boldsymbol{X}, P)\big]\, d\pi(P)} = \frac{\exp\big[-n\beta\ell(P^\star, P)\big]}{\int_{\mathscr{M}} \exp\left[-n\beta\ell(P^\star, P)\right] d\pi(P)}.$$

We recognize here the density of a Gibbs measure associated with the energy $\ell(P^\star, P)$ at point $P \in \mathscr{M}$ and inverse temperature $n\beta > 0$. We know that when the temperature goes to 0 (or equivalently $n\beta$ to infinity), Gibbs measures concentrate their masses in vicinities of low energy points in $\mathscr{M}$. In our case, these low energy points are those for which $\ell(P^\star, P)$ is minimal.

Similar ideas can be found in Catoni's work and more specifically in his construction of Gibbs estimators—see Catoni [17, Chapter 4]. There, Catoni shows how to aggregate a continuous family of estimators in order to minimize a risk. In the present paper, we do not aim at aggregating estimators but we use similar ideas and tools that are due to Catoni and his co-authors for the construction of our robust posterior distribution.

4 The main results

4.1 Linking the prior to the complexity of the model

For $P \in \mathscr{M}$ and $r > 0$, we recall that

$$V(P, r) = \log\left(\frac{\pi(\mathscr{B}(P, 2r))}{\pi(\mathscr{B}(P, r))}\right)$$

where we use the convention $a/0 = +\infty$ for all $a \geq 0$. We said in the Introduction that such quantities encapsulate in some sense the complexity of the model $(\mathscr{M}, \pi)$ and we shall now explain why. If $\mathscr{M} = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \mathbb{R}^k\}$ is a parametric model endowed with a loss ℓ such that $\ell(\boldsymbol{\theta}, \boldsymbol{\theta}') = \big|\boldsymbol{\theta} - \boldsymbol{\theta}'\big|$, so that $(\mathscr{M}, \ell)$ is isometric to $(\mathbb{R}^k, |\cdot|)$, and if the prior ν on $\Theta = \mathbb{R}^k$ is improper and given by the Lebesgue measure, we obtain that for all $P \in \mathscr{M}$ and $r > 0$

$$V(P, r) = \log\left(\frac{\pi(\mathscr{B}(P, 2r))}{\pi(\mathscr{B}(P, r))}\right) = \log\left(\frac{(2r)^k}{r^k}\right) = k \log 2. \tag{10}$$

We observe that $V(P, r)$ corresponds in this case to the usual dimension of $\mathbb{R}^k$ (up the factor $\log 2$). For more general models $(\mathscr{M}, \pi)$ and loss functions ℓ, we may interpret $V(P, r)$ as some notion of dimension (or complexity) associated with the element $P \in \mathscr{M}$ at the scale $r > 0$. As we do not consider improper priors but probability distributions, $\lim_{r\to+\infty} \pi(\mathscr{B}(P, r)) = 1$ and consequently $\lim_{r\to+\infty} V(P, r) = 0$. This means that the connection with the notion of "dimension" is only relevant for values of r which are not too large.

Given $\gamma \in (0, 1]$, the set

$$\mathcal{R}(\beta, P) = \left\{ r \geq \frac{1}{n\beta a_1}, \text{ such that } \sup_{r' \geq r} \frac{V(P, r')}{r'} \leq \gamma n \beta a_1 \right\}$$

is the subinterval of $\mathbb{R}_+$ on which the mapping $r \mapsto V(P, r)$ is not larger than $r \mapsto \gamma n \beta a_1 r$. We denote by

$$r_n(\beta, P) = \inf \mathcal{R}(\beta, P) \tag{11}$$

the left endpoint of $\mathcal{R}(\beta, P)$. Since $\mathcal{R}(\beta, P)$ is increasing with β with respect to set inclusion, $r_n(\beta, P)$ is a nonincreasing function of β. For example, in the ideal situation given in (10) where $V(P, r) \equiv k \log 2$ with $k \log 2 \geq 1$, $r_n(\beta, P) = (\gamma a_1)^{-1}[k \log 2/(n\beta)]$. When the model $\mathscr{M} = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\}$ is parametric and the parameter space Θ is an open subset of $\mathbb{R}^k$ endowed with a prior ν, we shall see in Sect. 8.2 that under suitable assumptions $r_n(\beta, P_{\boldsymbol{\theta}})$ is indeed of order $k/(n\beta)$, at least for n sufficiently large.

The Bayesian paradigm offers the possibility to favour some elements of $\mathscr{M}$ as compared to others. The order of magnitude of $r_n(\beta, P)$ allows one to quantify how much the prior π advantages or disadvantages $P \in \mathscr{M}$. It follows from the definition of $r_n(\beta, P)$ that

$$0 < \pi\left(\mathscr{B}(P, 2r)\right) \leq \exp\left(\gamma n \beta a_1 r\right) \pi\left(\mathscr{B}(P, r)\right) \quad \text{for all } r > r_n(\beta, P). \tag{12}$$

Letting r decrease to $r_n(\beta, P)$, we derive that (12) also holds for $r = r_n(\beta, P)$. In particular, $\pi\left(\mathscr{B}(P, r)\right) > 0$ for $r = r_n(\beta, P)$. If the prior puts no mass on the ℓ-ball $\mathscr{B}(P, r)$, which clearly corresponds to a situation where the prior disadvantages P, $r_n(\beta, P) > r$ and $r_n(\beta, P)$ is therefore large if r is large. In the opposite case, if the prior puts enough mass on $\mathscr{B}(P, r)$ in the sense that

$$\pi\left(\mathscr{B}(P, r)\right) \geq \exp\left(-\gamma n \beta a_1 r\right), \tag{13}$$

then for all $r' \geq r$,

$$\begin{aligned} \pi\left(\mathscr{B}(P, r')\right) &\geq \exp\left(-\gamma n \beta a_1 r\right) \geq \exp\left(-\gamma n \beta a_1 r'\right) \\ &\geq \exp\left(-\gamma n \beta a_1 r'\right) \pi\left(\mathscr{B}(P, 2r')\right) \end{aligned}$$

hence,

$$\frac{\pi\left(\mathscr{B}(P,2r')\right)}{\pi\left(\mathscr{B}(P,r')\right)} \leq \exp\left(\gamma n\beta a_1 r'\right) \quad \text{and } r_n(\beta, P) \leq r.$$

The quantity $r_n(\beta, P)$ is therefore small if r is small. Although (13) is not equivalent to (12) (it is actually stronger), the previous arguments provide a partial view on the relationship between π and r_n and conditions to decide whether P is favoured by π or not, according to the size of $r_n(\beta, P)$.

4.2 A general result on the concentration property of the posterior distribution

According to the discussion of Sect. 4.1, we see that, when the set

$$\mathscr{M}(\beta) = \left\{P \in \mathscr{M},\ r_n(\beta, P) \leq a_1^{-1}\beta\right\} \tag{14}$$

is nonempty, it contains the most favoured elements of the model $(\mathscr{M}, \pi)$ at level $a_1^{-1}\beta$. Since $r_n(\beta, P)$ is nonincreasing with β, the set $\mathscr{M}(\beta)$ is increasing with β with respect to set inclusion. If $a_1^{-1}\beta \geq (n\beta a_1)^{-1}$ or equivalently $\beta \geq 1/\sqrt{n}$, the set $\mathscr{M}(\beta)$ can alternatively be defined from $V(P, r)$ as follows:

$$\mathscr{M}(\beta) = \left\{P \in \mathscr{M},\ V(P,r) \leq \gamma n\beta a_1 r \text{ for all } r \geq a_1^{-1}\beta\right\}. \tag{15}$$

This set plays a crucial role in our first result.

Theorem 1 *Assume that the model $(\mathscr{M}, \pi)$ and the loss ℓ satisfy Assumptions 1 and 2 and the family $\mathscr{T}(\ell, \mathscr{M})$ Assumption 3. Let $\gamma < (c_0 \wedge c)/(2\tau)$ and $\beta \geq 1/\sqrt{n}$ be chosen in such a way that the set $\mathscr{M}(\beta)$ defined by (14) is not empty. Then, the posterior $\widehat{\pi}_{\boldsymbol{X}}$ defined by (7) possesses the following property. There exists $\kappa_0 > 0$ only depending on c, τ, γ and the ratio a_0/a_1 such that, for all $\xi > 0$ and any distribution $\mathbf{P}^\star$ with marginals in $\mathscr{P}$,*

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathrm{c}}\mathscr{B}(\overline{P}^\star, \kappa_0 r)\right)\right] \leq 2e^{-\xi} \tag{16}$$

with

$$r = \inf_{P \in \mathscr{M}(\beta)} \ell(\overline{P}^\star, P) + \frac{1}{a_1}\left(\beta + \frac{2\xi}{n\beta}\right). \tag{17}$$

In particular,

$$\mathbb{P}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathrm{c}}\mathscr{B}(\overline{P}^\star, \kappa_0 r)\right) \geq e^{-\xi/2}\right] \leq 2e^{-\xi/2}.$$

The value of κ_0 is given by (119) in the proof. It only depends on the choice of the family $\mathscr{T}(\ell, \mathscr{M})$ but not on the prior π. Hence, for a given family $\mathscr{T}(\ell, \mathscr{M})$, κ_0 is a numerical constant.

Let us now comment on Theorem 1. When $X_1, \ldots, X_n$ are truly i.i.d. with distribution $P^\star$ and the prior puts enough mass around $P^\star$, in the sense that $P^\star \in \mathscr{M}(\beta)$, then $r = a_1^{-1}[\beta + 2\xi/(n\beta)]$ in (17). When this ideal situation is not met, either because the data are not identically distributed or because $P^\star$ does not belong to $\mathscr{M}(\beta)$, r increases by at most an additive term of order $\inf_{P\in\mathscr{M}(\beta)} \ell(\overline{P^\star}, P)$. When this approximation term remains small as compared to $a_1^{-1}\beta$, the value of r does not deteriorate too much as compared to the previous situation.

The value of r given by (17) depends on the choice of the parameter β. Since the set $\mathscr{M}(\beta)$ is increasing (with respect to set inclusion) as β gets larger, the two terms $\inf_{P\in\mathscr{M}(\beta)} \ell(\overline{P^\star}, P)$ and $a_1^{-1}\beta$ vary in opposite directions as β increases. The set $\mathscr{M}(\beta)$ must be large enough to provide a suitable approximation of $\overline{P^\star}$ while β must not be too large in order to keep $a_1^{-1}\beta$ to a reasonable size. Practically, we recommend to choose $\beta = \beta(\alpha) \geq 1/\sqrt{n}$ such that

$$\pi\left(\mathscr{M}(\beta)\right) \geq 1 - \alpha \quad \text{for } \alpha \in (0, 1/10). \tag{18}$$

In Example 1 below and in Sect. 7.1, we give some examples of choices of β.

Example 1 Let $(\mathscr{M}, \pi)$ be a model where the prior π satisfies for some $k \geq 1$ and constants $0 < A \leq (2/e)B$,

$$(Ar)^k \wedge 1 \leq \pi\left(\mathscr{B}(P, r)\right) \leq (Br)^k \wedge 1 \quad \text{for all } P \in \mathscr{M} \text{ and } r > 0. \tag{19}$$

This means that the prior π behaves like the Lebesgue measure on an Euclidean space of dimension k for small enough values of r. Then,

$$V(P, r) = \log \frac{\pi\left(\mathscr{B}(P, 2r)\right)}{\pi\left(\mathscr{B}(P, r)\right)} \leq k \log\left(\frac{2B}{A}\right) \quad \text{for all } P \in \mathscr{M} \text{ and } r > 0 \tag{20}$$

which implies that for all $P \in \mathscr{M}$

$$r_n(P, \beta) \leq \frac{k}{\gamma a_1 n \beta} \log\left(\frac{2B}{A}\right). \tag{21}$$

The right-hand side is not larger than $a_1^{-1}\beta$ for

$$\beta = \sqrt{\frac{k \log(2B/A)}{\gamma n}} \tag{22}$$

which is larger than $1/\sqrt{n}$ since $(2B/A) \geq e$ and $\gamma \in (0, 1]$. For such a value of β, which does not depend on the distribution of the data, the element P belongs to $\mathscr{M}(\beta)$ given by (15), and since P is arbitrary we derive that $\mathscr{M}(\beta) = \mathscr{M}$. Applying

Theorem 1 we conclude that the distribution $\widehat{\pi}_{\boldsymbol{X}}$ concentrates on an ℓ-ball centered at $\overline{P}^{\star}$ with a radius r of order

$$\overline{r}_n = \inf_{P \in \mathscr{M}} \ell(\overline{P}^{\star}, P) + \frac{1}{a_1}\left(\sqrt{\frac{k}{n}} + \frac{2\xi}{\sqrt{nk}}\right). \tag{23}$$

4.3 A refined result under Assumption 4

Let us assume now that the family $\mathscr{T}(\ell, \mathscr{M})$ satisfies the stronger Assumption 4. We introduce the mapping

$$\phi : \begin{vmatrix} (0, +\infty) & \longrightarrow & \mathbb{R}_+ \\ z & \longmapsto & \phi(z) = \dfrac{2\,(e^z - 1 - z)}{z^2}. \end{vmatrix} \tag{24}$$

The function ϕ is increasing on $(0, +\infty)$ and tends to 1 when z tends to 0. Given $\beta > 0$ and a family $\mathscr{T}(\ell, \mathscr{M})$ that satisfies Assumption 4, we define

$$\overline{c}_1 = c_0 - \beta a_2 a_1^{-1} \tau^2 \phi\left[\beta(1 + 2c)\right](1 + 2c(1 + c)); \tag{25}$$

$$\overline{c}_2 = c - \beta a_2 a_1^{-1} \tau^2 \phi\left[\beta(1 + 2c)\right] c^2; \tag{26}$$

$$\overline{c}_3 = (2 + c) - \beta a_2 a_1^{-1} \tau^2 \phi\left[\beta(3 + 2c)\right](2 + c)^2. \tag{27}$$

Note that the value of $\overline{c}_1 \wedge \overline{c}_2 \wedge \overline{c}_3$ is positive for $\beta = 0$ and decreases continuously to $-\infty$ when β grows to infinity. Consequently, there exists some $\beta_0 > 0$ for which $\overline{c}_1 \wedge \overline{c}_2 \wedge \overline{c}_3 = 0$ and $\overline{c}_1 \wedge \overline{c}_2 \wedge \overline{c}_3$ is positive for all values $\beta \in (0, \beta_0)$.

Let us now present our second result on the concentration property of our posterior $\widehat{\pi}_{\boldsymbol{X}}$.

Theorem 2 *Assume that the model $(\mathscr{M}, \pi)$ and the loss ℓ satisfy Assumptions 1 and 2 and the family $\mathscr{T}(\ell, \mathscr{M})$ Assumption 4. For $\beta \in (0, \beta_0)$ and $\gamma < (\overline{c}_1 \wedge \overline{c}_2 \wedge \overline{c}_3)/(2\tau)$, the posterior $\widehat{\pi}_{\boldsymbol{X}}$ defined by (7) satisfies the following property. There exists $\kappa_0 > 0$ only depending on a_0/a_1, a_2/a_1, c, τ, β and γ such that, for all $\xi > 0$ and any distribution $\mathbf{P}^{\star}$ with marginals in $\mathscr{P}$,*

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{P}^{\star}, \kappa_0 r)\right)\right] \leq 2e^{-\xi} \tag{28}$$

with

$$r = \inf_{P \in \mathscr{M}} \left[\ell(\overline{P}^{\star}, P) + r_n(\beta, P)\right] + \frac{2\xi}{n\beta a_1}. \tag{29}$$

In particular,

$$\mathbb{P}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{P}^{\star}, \kappa_0 r)\right) \geq e^{-\xi/2}\right] \leq 2e^{-\xi/2}.$$

The value of κ_0 is given by (132) in the proof. Note that the constraints on β and γ, that are required in our Theorem 2, and that on c given in (6) only depend on a_0, a_1 and a_2, hence on the choice of the family $\mathscr{T}(\ell, \mathscr{M})$. When a_0, a_1 and a_2 do not depend on $\mathscr{M}$, the value of β can be chosen as a universal constant. In particular, it neither depends on the model $(\mathscr{M}, \pi)$ nor on the sample size n.

Example 2 (Example 1 continued) Let us go back to the framework of our Example 1 and assume that $\mathscr{T}(\ell, \mathscr{M})$ satisfies the requirements of Theorem 2, hence Assumption 4. Applying our construction with some numerical value of β which satisfies the constraint of our Theorem 2, we deduce from (21) that $\widehat{\pi}_X$ concentrates on an ℓ-ball with radius of order

$$\overline{r} = \inf_{P \in \mathscr{M}} \ell(\overline{P}^\star, P) + \frac{\log(2B/A)}{\gamma a_1 \beta} \frac{k}{n} + \frac{2}{a_1 \beta} \frac{\xi}{n}. \tag{30}$$

When the model is well-specified, $\inf_{P \in \mathscr{M}} \ell(\overline{P}^\star, P) = 0$ and the ball $\mathscr{B}(P^\star, \kappa_0 \overline{r})$ with radius $\overline{r} = \overline{r}(n)$ contracts at the rate $1/n$. Applying our Theorem 1 under Assumption 3, ignoring the fact that the family $\mathscr{T}(\ell, \mathscr{M})$ also satisfies Assumption 4, would lead to the weaker result that when the model is well-specified the posterior concentrates on an ℓ-ball with radius of order $\sqrt{k/n}$, hence at a rate $1/\sqrt{n}$, as shown by (23).

4.4 Concentrated priors

Theorem 1 and 2 show that starting from a prior π that puts enough mass around most of the elements of $\mathscr{M}$, the posterior $\widehat{\pi}_X$ concentrates on an ℓ-ball with radius of order $\inf_{P \in \mathscr{M}} \ell(\overline{P}^\star, P) + r_n$ where r_n is small, at least under suitable assumptions and for n sufficiently large. The situation we want to investigate now is what happens when the prior is very concentrated on a small ℓ-ball with radius $\varepsilon > 0$ around an element $\overline{Q} \in \mathscr{M}$ that might not be the true distribution of the data. More precisely, assume the following

Assumption 5 For $\overline{Q} \in \mathscr{M}$ and $\varepsilon > 0$,

$$\pi\left({}^{\mathrm{c}}\mathscr{B}(\overline{Q}, \varepsilon)\right) \leq e^{-(2\xi+1)} \pi\left(\mathscr{B}(\overline{Q}, \varepsilon)\right) \quad \text{with } \xi > 0.$$

In this case, we establish the following result.

Theorem 3 *Assume that the model $(\mathscr{M}, \pi)$ and the loss ℓ satisfy Assumptions 1 and 2 and the family $\mathscr{T}(\ell, \mathscr{M})$ Assumption 3. If Assumption 5 is satisfied, there exists $\kappa_0 > 0$ only depending on c, τ and the ratio a_0/a_1 such that for any distribution $\mathbf{P}^\star$ with marginals in $\mathscr{P}$,*

$$\mathbb{E}\left[\widehat{\pi}_X\left({}^{\mathrm{c}}\mathscr{B}(\overline{P}^\star, \kappa_0 r)\right)\right] \leq 2e^{-\xi} \quad \textit{with} \quad r = \ell(\overline{P}^\star, \overline{Q}) \vee \frac{\beta}{a_1} \vee \varepsilon. \tag{31}$$

In particular, for the choice $\beta = a_1 \varepsilon$, $r = \ell(\overline{P}^\star, \overline{Q}) \vee \varepsilon$.

If furthermore, Assumption 4 is satisfied and $\beta \in (0, \beta_0)$ *(where* β_0 *is defined in Sect. 4.3), there exists* $\kappa_0' > 0$ *only depending on* $\tau, \beta, a_0/a_1$ *and* a_2/a_1 *such that for any distribution* $\mathbf{P}^\star$ *with marginals in* $\mathscr{P}$*,*

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{P}^\star, \kappa_0' r)\right)\right] \leq 2e^{-\xi} \quad \textit{with} \quad r = \ell(\overline{P}^\star, \overline{Q}) \vee \varepsilon. \tag{32}$$

This result shows that for a suitable choice of β, the posterior $\widehat{\pi}_{\boldsymbol{X}}$ also concentrates on an ℓ-ball centred at $\overline{P}^\star$ with radius of order ε when the model is well-specified, that is, when the data are i.i.d. with distribution $\overline{P}^\star = \overline{Q}$. When the model is misspecified, the radius of the ball is of order $\ell(\overline{P}^\star, \overline{Q}) \vee \varepsilon$ and therefore does not inflate more than the distance of $\overline{P}^\star$ to the center $\overline{Q}$. This result illustrates the stability of the posterior $\widehat{\pi}_{\boldsymbol{X}}$ with respect to misspecification.

5 Applications to classical loss functions

The aim of this section is to show how our general construction can be applied to loss functions ℓ of interest. The propositions contained in this section about the corresponding families $\mathscr{T}(\ell, \mathscr{M})$ have been established in Baraud [4] except for the squared Hellinger loss for which we refer to Baraud and Birgé [5, Proposition 3]. The list of loss functions we present here is not exhaustive. Our results also apply to all loss functions that derive from a variational formula of the form

$$\ell(P, Q) = \sup_{f \in \mathscr{F}} \left[\int_E f dP - \int_E f dQ\right]$$

where $\mathscr{F}$ is a suitable class of bounded functions. For such losses, we refer the reader to Baraud [4].

In this section, we consider models $\mathscr{M} = \{P = p \cdot \mu,\ p \in \mathcal{M}\}$ which are dominated by a measure μ on $(E, \mathcal{E})$ and we denote by $\mathcal{M} \subset \mathscr{L}_1(E, \mathcal{E}, \mu)$ the corresponding families of densities with respect to μ. Elements $P, Q, \ldots$ in $\mathscr{M}$ are associated with their densities in $\mathcal{M}$ by using lower case letters $p, q, \ldots$. In all the cases we consider, $t_{(P,Q)}(x)$ is a measurable function of $(p(x), q(x))$ for $P, Q \in \mathscr{M}$ and $x \in E$. In order to satisfy our measurability Assumption 3-(i), it is therefore sufficient to assume that

$$\begin{aligned}(E \times \mathscr{M}, \mathcal{E} \otimes \mathcal{A}) &\longrightarrow \mathbb{R} \\ (x, P) &\longmapsto p(x)\end{aligned}$$

is measurable. In the case of a parametrized model $\mathscr{M} = \{P_\theta = p_\theta \cdot \mu, \theta \in \Theta\}$, as described in Sect. 2.1, this condition is satisfied as soon as the mapping

$$p : \left|\begin{aligned}(E \times \Theta, \mathcal{E} \otimes \mathfrak{B}) &\longrightarrow \mathbb{R}_+ \\ (x, \theta) &\longmapsto p_\theta(x)\end{aligned}\right.$$

is measurable. Throughout this section, we assume that such measurability assumptions are satisfied.

5.1 The case of the total variation distance

In this section, $\mathscr{P}$ is the set of all probability measures on $(E, \mathcal{E})$ and

$$\|P - Q\| = \frac{1}{2}\int_E |p - q|\, d\mu \tag{33}$$

denotes the total variation loss (TV-loss for short) between $P, Q \in \mathscr{P}$.

Proposition 1 *The family $\mathscr{T}(\ell, \mathscr{M})$ which consists of all the functions $t_{(P,Q)}$ defined for $P = p \cdot \mu$ and $Q = q \cdot \mu$ in $\mathscr{M}$ by*

$$t_{(P,Q)} = \frac{1}{2}\left[\mathbb{1}_{q>p} - Q(q > p)\right] - \frac{1}{2}\left[\mathbb{1}_{p>q} - P(p > q)\right] \tag{34}$$

satisfies Assumption 2 with $\tau = 1$ and Assumption 3 with $a_0 = 3/2$ and $a_1 = 1/2$.

It follows from Proposition 1 that we may apply our general construction to the so-defined family $\mathscr{T}(\ell, \mathscr{M})$ with the values $c = c_0 = 1/3$ (hence $\lambda = 4/3$). The reader can check that the value $\gamma = 1/100$ satisfies the requirement of our Theorem 1 and that (16) is satisfied with $\kappa_0 = 220$. Theorem 1 can therefore be rephrased as follows.

Corollary 1 *Let $\beta \geq 1/\sqrt{n}$, $c = 1/3$ and $\widehat{\pi}_X^{TV}$ be the posterior defined by (7) and associated with the family $\mathscr{T}(\ell, \mathscr{M})$ given in Proposition 1. For all $\xi > 0$ and any distribution $\mathbf{P}^\star$, with a probability at least $1 - 2e^{-\xi/2}$, the posterior $\widehat{\pi}_X^{TV}$ satisfies*

$$\widehat{\pi}_X^{TV}\left(\left\{P \in \mathscr{M},\ \ell(\overline{P}^\star, P) \leq 220\left[\inf_{P' \in \mathscr{M}(\beta)} \ell(\overline{P}^\star, P') + 2\left(\beta + \frac{2\xi}{n\beta}\right)\right]\right\}\right) \geq 1 - e^{-\xi/2} \tag{35}$$

where

$$\mathscr{M}(\beta) = \left\{P \in \mathscr{M},\ \sup_{r \geq 2\beta}\left[\frac{200}{nr}\log\left(\frac{\pi\left(\mathscr{B}(P, 2r)\right)}{\pi\left(\mathscr{B}(P, r)\right)}\right)\right] \leq \beta\right\}.$$

By convexity, we may write that

$$\inf_{P \in \mathscr{M}(\beta)}\left\|P - \overline{P}^\star\right\| \leq \inf_{P \in \mathscr{M}(\beta)}\left[\frac{1}{n}\sum_{i=1}^n \left\|P - P_i^\star\right\|\right]$$

and the left-hand side is therefore small when there exists $P \in \mathscr{M}(\beta)$ that approximates well enough most of the marginals of $\mathbf{P}^\star$. The concentration properties of $\widehat{\pi}_X^{\mathrm{TV}}$ remain thus stable with respect to a possible misspecification of the model and a departure from the equidistribution assumption.

In fact, as we shall see in our Example 3 below, the average distribution $\overline{P}^\star$ may belong to $\mathscr{M}(\beta)$ even when none of the marginals $P_i^\star$ does. This means that in good

situations, the posterior may concentrate around $\overline{P}^\star$, as it would do in the i.i.d. case when the distribution of the data does belong to $\mathscr{M}(\beta)$, even when the data are non-i.i.d. and their marginals do not belong to $\mathscr{M}(\beta)$.

Example 3 (Example 1 continued) Going back to Example 1 and taking for ℓ the TV-loss (then $a_1 = 1/2$), we deduce from (23) that

$$\overline{r}_n = \inf_{P \in \mathscr{M}} \left\| \overline{P}^\star - P \right\| + 2\left(\sqrt{\frac{k}{n}} + \frac{2\xi}{\sqrt{nk}}\right).$$

In particular, if for each $i \in \{1, \ldots, n\}$, $P_i^\star$ is the uniform distribution on $[(i-1)/n, i/n]$ and $\mathscr{M}$ contains the uniform distribution $\mathcal{U}([0,1])$ on $[0,1]$, $\mathscr{M}$ contains $\overline{P}^\star = \mathcal{U}([0,1])$, even if none of the marginals $P_i^\star$ belongs to $\mathscr{M}$. We then get that

$$\overline{r}_n = 2\left(\sqrt{\frac{k}{n}} + \frac{2\xi}{\sqrt{nk}}\right)$$

and the posterior concentrates around $\overline{P}^\star$ at a parametric rate.

5.2 Case of the $\mathbb{L}_j$-loss

Let $j \in (1, +\infty)$. We denote by $\mathscr{P}_j$ the set of all finite and signed measures on $(E, \mathcal{E}, \mu)$ which are of the form $P = p \cdot \mu$ with $p \in \mathscr{L}_j(E,\mu) \cap \mathscr{L}_1(E,\mu)$. Let ℓ_j be the loss defined by $\ell_j(P,Q) = \|p - q\|_{\mu,j}$ for all $P = p \cdot \mu$ and $Q = q \cdot \mu$ in $\mathscr{P}_j$. In this section, $\mathscr{P}$ is the subset that consists of all the probability measures in $\mathscr{P}_j$.

Proposition 2 *Let $\mathscr{M} = \{P = p \cdot \mu,\ p \in \mathcal{M}\}$ be a subset of $\mathscr{P}_j$ for which $\mathcal{M}$ satisfies for some $R > 0$*

$$\|p - q\|_\infty \le R\,\|p - q\|_{\mu,j} \quad \textit{for all } p, q \in \mathcal{M}. \tag{36}$$

Define for $P = p \cdot \mu$ and $Q = q \cdot \mu$ in $\mathscr{M}$,

$$f_{(P,Q)} = \frac{(p-q)_+^{j-1} - (p-q)_-^{j-1}}{\|p - q\|_{\mu,j}^{j-1}} \quad \textit{when } P \neq Q \quad \textit{and} \quad f_{(P,P)} = 0.$$

Then, the family $\mathscr{T}(\ell_j, \mathscr{M})$ which contains the functions $t_{(P,Q)}$ defined for $P, Q \in \mathscr{M}$ by

$$t_{(P,Q)} = \frac{1}{2R^{j-1}}\left[\int_E f_{(P,Q)} \frac{dP + dQ}{2} - f_{(P,Q)}\right] \tag{37}$$

satisfies Assumption 2 with $\tau = 1$ and Assumption 3 with $a_0 = 3/(4R^{j-1})$ and $a_1 = 1/(4R^{j-1})$.

When $j = 2$, (36) is typically satisfied when $\mathcal{M}$ is a subset of a linear space enjoying good connections between the $\mathbb{L}_2(\mu)$ and the supremum norms. Many finite dimensional linear spaces with good approximation properties do satisfy such connections (e.g. piecewise polynomials of a fixed degree on a regular partition of $[0, 1]$, trigonometric polynomials on $[0, 1)$ etc.). We refer the reader to Birgé and Massart [14, Section 3] for additional examples. The property may also hold for infinite dimensional linear spaces as proven in Baraud [4].

It follows from Proposition 2 that one may choose $c = c_0 = 1/3$ in (6) and $\gamma = 1/100$ in Theorem 1. Besides, Theorem 1 applies with $\kappa_0 = 220$.

Example 4 (Example 1 continued) Let us go back to our Example 1 with $\ell = \ell_j$ and $\mathscr{T}(\ell, \mathscr{M})$ given in Proposition 2. For the choice of β given in (22) and $\gamma = 1/100$, we deduce from (23) (with $a_1 = 1/(4R^{j-1})$) that the resulting posterior $\widehat{\pi}_X$ concentrates on an ℓ_j-ball around $\overline{P}^\star$ with a radius of order

$$\bar{r}_n = \inf_{p \in \mathcal{M}} \left\| \frac{1}{n} \sum_{i=1}^{n} p_i^\star - p \right\|_{\mu, j} + 4R^{j-1} \left(\sqrt{\frac{k}{n}} + \frac{2\xi}{\sqrt{nk}} \right).$$

5.3 The case of the squared Hellinger loss

Here, $\mathscr{P}$ is the set of all probability measures on $(E, \mathcal{E})$ and

$$\ell(P, Q) = h^2(P, Q) = \frac{1}{2} \int_E \left(\sqrt{p} - \sqrt{q} \right)^2 d\mu, \tag{38}$$

is the squared Hellinger distance between two probabilities $P, Q \in \mathscr{P}$.

Proposition 3 *Let ψ be the function defined by*

$$\psi : \left| \begin{array}{rcl} [0, +\infty] & \longrightarrow & [-1, 1] \\ x & \longmapsto & \begin{cases} \dfrac{x-1}{x+1} & \text{if } x \in [0, +\infty) \\ 1 & \text{if } x = +\infty. \end{cases} \end{array} \right.$$

The family $\mathscr{T}(\ell, \mathscr{M})$ containing the functions $t_{(P,Q)}$ defined for $P = p \cdot \mu$ and $Q = q \cdot \mu$ in $\mathscr{M}$ by

$$t_{(P,Q)} = \frac{1}{2} \psi \left(\sqrt{\frac{q}{p}} \right) \tag{39}$$

(with the conventions $0/0 = 1$ and $x/0 = +\infty$ for all $x > 0$) satisfies Assumption 2 with $\tau = 2$ and Assumption 4 with $a_0 = 2$, $a_1 = 3/16$, $a_2 = 3\sqrt{2}/4$.

With such a choice of family $\mathscr{T}(\ell, \mathscr{M})$, (6) is satisfied with $c = 1/125$, then $c_0 \in [0.922, 0.923]$, and the requirements of Theorem 2 are satisfied with $\beta = 2\gamma =$

$1/500$. Then the value $\kappa_0 = 1694$ suits. The definition (11) of $r_n(\beta, P)$ for $P \in \mathscr{M}$ becomes

$$r_n(\beta, P) = \inf\left\{ r \geq \frac{8000}{3n},\ \frac{\pi\left(\mathscr{B}(P, 2r')\right)}{\pi\left(\mathscr{B}(P, r')\right)} \leq \exp\left(\frac{3nr'}{8.10^6}\right) \text{ for all } r' \geq r \right\}, \tag{40}$$

with the convention $\sup \varnothing = 8000/(3n)$. Theorem 2 can then be rephrased as follows.

Corollary 2 *Let $\pi_{\boldsymbol{X}}^h$ be the posterior defined by (7) and associated with the family $\mathscr{T}(\ell, \mathscr{M})$ given in Proposition 3 and the choices $c = 1/125$ and $\beta = 1/500$. For all $\xi > 0$ and any distribution $\mathbf{P}^\star$, with a probability at least $1 - 2e^{-\xi/2}$,*

$$\widehat{\pi}_{\boldsymbol{X}}^h\left(\left\{P \in \mathscr{M},\ h^2\left(\overline{P}^\star, P\right) \leq 1694r\right\}\right) \geq 1 - e^{-\xi/2}$$

where

$$r = \inf_{P \in \mathscr{M}}\left[h^2\left(\overline{P}^\star, P\right) + r_n(\beta, P)\right] + \frac{5334\xi}{n}$$

and $r_n(\beta, P)$ is given by (40).

As for the total variation distance, we may write that

$$\inf_{P \in \mathscr{M}} h^2\left(\overline{P}^\star, P\right) \leq \inf_{P \in \mathscr{M}}\left[\frac{1}{n}\sum_{i=1}^n h^2\left(P_i^\star, P\right)\right].$$

The left-hand side is small when there exists an element $P \in \mathscr{M}$ that approximates well most of the marginal distribution $P_i^\star$. If for such a P, the quantity $r_n(\beta, P)$ is small enough, the posterior concentrates around $\overline{P}^\star$ just as it would do if the data were truly i.i.d. with distribution $P \in \mathscr{M}$.

Example 5 (Example 1 continued) Let us go back to Example 1, more precisely Example 2, with $\ell = h^2$ and $\mathscr{T}(\ell, \mathscr{M})$ given in Proposition 3. Inequality (21) is satisfied with $\beta = 2\gamma = 1/500$ and $a_1 = 3/16$. It follows from (30) that $\widehat{\pi}_{\boldsymbol{X}}^h$ concentrates on an h^2-ball around $\overline{P}^\star$ with a radius of order

$$\bar{r} = \inf_{P \in \mathscr{M}} h^2\left(\overline{P}^\star, P\right) + \frac{k + \xi}{n}.$$

6 Comparing the classical Bayesian approach to ours

In this section, our aim is to highlight some similarities and differences between the Bayesian posterior and ours. Throughout this section, we consider the squared

Hellinger loss $\ell = h^2$ and denote by $\widehat{\pi}_X^K$ the Bayes posterior associated with the model $(\mathscr{M}, \pi)$. The letter K in the notation $\widehat{\pi}_X^K$ refers to the fact that the Bayesian posterior can be obtained from our general construction by using the Kullback–Leibler divergence as explained in Sect. 3.4. When $\mathscr{M} = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\}$ is parametric with $\Theta \subset \mathbb{R}^k$, we denote by $\widehat{\nu}_X^K$ the Bayesian posterior on the parameter space Θ and $\widehat{\nu}_X^h$ that associated to $\widehat{\pi}_X^h$.

6.1 Some classical concentration results for the Bayes posterior distribution

Most of the results that have been established about the concentration properties of the Bayesian posterior are asymptotic in nature. It seems difficult to establish a general nonasymptotic version of those as we do for our posterior. One of the only exceptions we are aware of is Birgé [13].

When the data are i.i.d. with a distribution $P^\star \in \mathscr{M}$, a typical asymptotic form of these results is the following one (see Ghosal et al. [19] Theorems 2.1 and 2.4 for example). Let ε_n be a sequence of positive numbers that converges to zero when n goes to infinity. If $P^\star$ fulfils some suitable conditions, that we shall discuss later on and which depend on the prior π and ε_n, the following convergence in probability holds true

$$\widehat{\pi}_X^K(\{P \in \mathscr{M},\ h^2(P^\star, P) \geq M_n \varepsilon_n^2\}) \xrightarrow[n\to+\infty]{\mathrm{P}} 0 \quad \text{under } P^\star. \tag{41}$$

In (41), $M_n = M$ denotes some large enough positive constant if $n\varepsilon_n^2 \to +\infty$ as $n \to +\infty$ while M_n is increasing to infinity as $n \to +\infty$ if $\liminf n\varepsilon_n^2 > 0$ as $n \to +\infty$. The first condition on ε_n is typically satisfied when $\mathscr{M}$ is a nonparametric model while the second one generally applies to parametric ones.

In comparison, in this well-specified framework, our Corollary 2 leads to the following result. For all $P^\star \in \mathscr{M}$ and $\xi > 0$

$$\mathbb{P}\left[\widehat{\pi}_X^h\left(\left\{P \in \mathscr{M},\ h^2(P^\star, P) \geq \kappa_0'\left(r_n(\beta, P^\star) + \frac{\xi}{n}\right)\right\}\right) \geq e^{-\xi/2}\right] \leq 2e^{-\xi/2} \tag{42}$$

for some numerical constant $\kappa_0' > 0$. If $P^\star$ satisfies $r_n(\beta, P^\star) \leq \varepsilon_n^2$, we recover (41) by setting $\xi = \xi_n = (M_n/(\kappa_0') - 1)n\varepsilon_n^2$. However, our condition that $r_n(\beta, P^\star) \leq \varepsilon_n^2$ is not equivalent to that imposed on $P^\star$ by Ghosal, Ghosh and van der Vaart [19]. It is actually weaker. In their paper, this condition is fulfilled when the prior puts enough mass on Kullback–Leibler type balls around $P^\star$. Our approach allows one to consider Hellinger balls only, which are larger and make our assumption weaker. In fact, as already underlined in the Introduction, these Kullback–Leibler type balls could be empty, and the condition unsatisfied, while our theorem would still apply.

The result established by Birgé [13] provides an improvement as compared to the one presented above and established by Ghosal, Ghosh and van der Vaart. Birgé shows that it is essentially possible to get rid of the Kullback–Leibler divergence (see

his Theorem 2) but only when the model is parametric and well-specified. Apart for the nonparametric framework, this result leaves little place for improvement since we know that the Bayesian posterior may fail to concentrate around the true parameter when the model becomes slightly ill-specified.

Another consequence of our Corollary 2, as compared to (41), is that it allows one to control

$$\widehat{\pi}_{\boldsymbol{X}}^{h}\left(\left\{P\in\mathscr{M},\ h^2(P^\star,P)\geq\kappa_0'\left(\varepsilon_n^2+\frac{\xi}{n}\right)\right\}\right)$$

uniformly over the set $\{P^\star\in\mathscr{M}, r_n(\beta,P^\star)\leq\varepsilon_n^2\}$. For example, in the framework of Example 2, for the choice $\varepsilon_n^2=ck/n$ with $c=\log(2B/A)/(\gamma a_1\beta)$, we know that $r_n(\beta,P^\star)\leq\varepsilon_n^2$ for all $P^\star\in\mathscr{M}$ and we deduce from (42) that

$$\sup_{P^\star\in\mathscr{M}}\mathbb{P}\left[\widehat{\pi}_{\boldsymbol{X}}^{h}\left(\left\{P\in\mathscr{M},\ h^2(P^\star,P)\geq\kappa_0'\left(\varepsilon_n^2+\frac{\xi}{n}\right)\right\}\right)\geq e^{-\xi/2}\right]\leq 2e^{-\xi/2}.$$

The concentration properties of our posterior is therefore uniform over the statistical model $\mathscr{M}$.

6.2 About the shapes and sizes of the credible regions

A nice feature of the Bayesian approach lies in the fact that it allows one to build credible regions. In practice, they often play the same role as the confidence regions in the frequentist paradigm. When the data are i.i.d. with distribution $P^\star=P_{\boldsymbol{\theta}^\star}$ in a parametric model $\mathscr{M}=\{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta}\in\Theta\}$, $\Theta\subset\mathbb{R}^k$, a credible set for the parameter $\boldsymbol{\theta}^\star$ is a subset $\widehat{\Theta}_{n,\boldsymbol{X}}\subset\Theta$ (only depending on observable quantities) that satisfies $\widehat{\nu}_{\boldsymbol{X}}^{K}(\widehat{\Theta}_{n,\boldsymbol{X}})\geq 1-e^{-\xi}$ for some choice of $\xi>0$. When $\mathscr{M}$ is a regular parametric model with a nonsingular Fisher information matrix $\mathbf{J}$, and provided that it satisfies additional assumptions—see van der Vaart [24]—the Bernstein–von Mises theorem applies and tells us that

$$\left\|\widehat{\nu}_{\boldsymbol{X}}^{K}-\mathcal{N}\left(\widehat{\boldsymbol{\theta}}_n,(n\mathbf{J}(\boldsymbol{\theta}^\star))^{-1}\right)\right\|\underset{n\to+\infty}{\overset{\mathrm{P}}{\longrightarrow}}0\quad\text{under } P_{\boldsymbol{\theta}^\star}$$

where $\widehat{\boldsymbol{\theta}}_n$ denotes the Maximum Likelihood Estimator (MLE for short). Denoting by $\overline{\chi}_k^{-1}(\xi)$ the $(1-e^{-\xi})$-quantile of a chi-square random variable with k degrees of freedom and

$$\Theta_{n,\boldsymbol{X}}=\left\{\boldsymbol{\theta}\in\Theta,\ n\left|\mathbf{J}^{1/2}(\boldsymbol{\theta}^\star)\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right)\right|^2\leq\overline{\chi}_k^{-1}(\xi)\right\},\tag{43}$$

we deduce that

$$\left|\widehat{\nu}_{\boldsymbol{X}}^{K}\left(\Theta_{n,\boldsymbol{X}}\right)-(1-e^{-\xi})\right|\leq\left\|\widehat{\nu}_{\boldsymbol{X}}^{K}-\mathcal{N}\left(\widehat{\boldsymbol{\theta}}_n,(n\mathbf{J}(\boldsymbol{\theta}^\star))^{-1}\right)\right\|\underset{n\to+\infty}{\overset{\mathrm{P}}{\longrightarrow}}0$$

hence

$$\widehat{\nu}_{\boldsymbol{X}}^{K}\left(\Theta_{n,\boldsymbol{X}}\right) \underset{n\to+\infty}{\overset{\mathrm{P}}{\longrightarrow}} 1-e^{-\xi} \quad \text{under } P_{\boldsymbol{\theta}^\star}.$$

The asymptotic level of "credibility" of the set $\Theta_{n,\boldsymbol{X}}$ is therefore $1-e^{-\xi}$. This set is not, however, a genuine credible region since it depends on the unknown parameter $\boldsymbol{\theta}^\star$. We would obtain a genuine credible region by replacing $\boldsymbol{\theta}^\star$ by $\widehat{\boldsymbol{\theta}}_n$ in the expression of $\Theta_{n,\boldsymbol{X}}$. This substitution would change the level of credibility but not the shape of the region, which is an ellipsoid centred at $\widehat{\boldsymbol{\theta}}_n$ and the axes of which are given by the eigenvectors of the Fisher information matrix.

The aim of this section is to show that our posterior concentrates its mass on regions that have the same shape and approximately the same size. The size of $\Theta_{n,\boldsymbol{X}}$ is determined by the value of the quantile $\overline{\chi}_k^{-1}(\xi)$. The aim of the following lemma is to specify the order of magnitude of this quantile as a function of k and ξ. In fact, we consider below the more general case of the quantiles of a gamma distribution $\gamma(s,\sigma)$ with parameters $s,\sigma>0$, that is, the distribution with density $x\mapsto (x^{s-1}e^{-s/\sigma})/(\sigma^s\Gamma(s))$ with respect to the Lebesgue measure on $\mathbb{R}_+$. The proof is postponed to Sect. 10.1.

Lemma 1 *For $s,\sigma,\xi>0$, let $\overline{\gamma}_{s,\sigma}^{-1}(\xi)$ be the $(1-e^{-\xi})$-quantile of the gamma distribution $\gamma(s,\sigma)$ and $\overline{\Phi}^{-1}(\xi)$ that of a standard Gaussian random variable. Then,*

$$\overline{\gamma}_{s,\sigma}^{-1}(\xi)\le\sigma\left(\sqrt{s}+\sqrt{\xi}\right)^2 \tag{44}$$

and for all $s=t+1>1$ and $\xi\ge\log 2+1/(12t)$,

$$\overline{\gamma}_{s,\sigma}^{-1}(\xi)\ge\sigma\left[t+\left[\sqrt{t}\,\overline{\Phi}^{-1}\left(\xi-\frac{1}{12t}\right)\right]\vee\left[\xi+\log\left(\frac{e^{-1/(12t)}}{\sqrt{2\pi t}}\right)\right]\right]. \tag{45}$$

Since $\overline{\Phi}^{-1}(\xi)$ is equivalent to $\sqrt{2\xi}$ for large values of $\xi>0$, these two inequalities show that for s and ξ large enough, $\overline{\gamma}_{s,\sigma}^{-1}(\xi)$ is of order $\sigma\,[s+\xi]$. In particular, $\overline{\chi}_k^{-1}(\xi)=\overline{\gamma}_{k/2,2}^{-1}(\xi)$ is of order $k+\xi$ for k and ξ large enough.

To compare ourselves with the classical Bayesian paradigm, we prove in Sect. 10.2 the result below for our posterior. This result is based on the assumption that the statistical model $\mathscr{M}$ is regular in the sense that is defined in Ibragimov and Has'minskiĭ [20]. In order to avoid too many technicalities here, we refer the reader to our Sect. 8.3, more precisely Corollary 4, for a complete description of the assumptions on the statistical model $\mathscr{M}$.

Theorem 4 *Assume that the statistical model $\mathscr{M}$ satisfies the assumptions of Corollary 4. If $X_1,\ldots,X_n$ are i.i.d. with distribution $P_{\boldsymbol{\theta}^\star}\in\mathscr{M}$, for all $\xi>0$ and n large enough, with a probability $1-2e^{-\xi}$,*

$$\widehat{\nu}_{\boldsymbol{X}}^{h}\left(\left\{\boldsymbol{\theta}\in\Theta,\ n\left|\mathbf{J}^{1/2}(\boldsymbol{\theta}^\star)\left(\boldsymbol{\theta}-\boldsymbol{\theta}^\star\right)\right|^2\le\kappa^\star\,(k+\xi)\right\}\right)\ge 1-e^{-\xi} \tag{46}$$

where $\kappa^\star$ is a positive numerical constant.

The set

$$\left\{\boldsymbol{\theta} \in \Theta,\ n\left|\mathbf{J}^{1/2}(\boldsymbol{\theta}^\star)\left(\boldsymbol{\theta}-\boldsymbol{\theta}^\star\right)\right|^2 \leq \kappa^\star\,(k+\xi)\right\}$$

possesses the same shape and, by Lemma 1, approximately the same size as the set $\Theta_{n,\boldsymbol{X}}$ defined by (43). We deduce from Theorem 4 that the classical Bayes posterior and ours concentrate both on similar sets. If $\widehat{\boldsymbol{\theta}}_n$ is an asymptotically efficient estimator of $\boldsymbol{\theta}^\star$, it is therefore reasonable to look for a credible region of the form

$$\left\{\boldsymbol{\theta} \in \Theta,\ n\left|\mathbf{J}^{1/2}(\widehat{\boldsymbol{\theta}}_n)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right)\right|^2 \leq t\right\}, t>0$$

for $\widehat{\nu}^h_{\boldsymbol{X}}$ as we would do for the classical Bayes one.

6.3 Robustness

As already mentioned, our approach allows the statistician to design robust posteriors by choosing as a loss function the squared Hellinger loss or the total variation one. In this section, we illustrate this property on a concrete example. Consider the statistical model $\mathscr{M} = \{P_\theta = \mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$ and the prior π associated with the distribution $\nu = \mathcal{N}(0, 1)$ on $\Theta = \mathbb{R}$. Then, the Bayes posterior on Θ is $\widehat{\nu}^K_{\boldsymbol{X}} = \mathcal{N}(\widehat{m}_n, \sigma_n^2)$ with $\widehat{m}_n = (n+1)^{-1}\sum_{i=1}^n X_i$ and $\sigma_n^2 = 1/(n+1)$. It concentrates on intervals of the form $[\widehat{m}_n - c/\sqrt{n+1}, \widehat{m}_n + c/\sqrt{n+1}]$ for $c > 0$ large enough. If the distribution of the data is contaminated so that $X_1, \ldots, X_n$ are i.i.d. with distribution

$$P^\star = \left(1-\frac{1}{n}\right)P_0 + \frac{1}{n}\mathcal{N}\Big(10^4(n+1), 1/n\Big),$$

then with a probability at least $1-(1-1/n)^n \geq 1-1/e > 63\%$, the posterior concentrates around $\widehat{m}_n \approx 10^4$, hence far away from 0, even though $P^\star$ and P_0 are close: $\|P^\star - P_0\| \leq 1/n$.

In this specific framework, the model $\mathscr{M}$ is regular, the Fisher information is constant and positive, ν admits a positive density which is continuous at $\theta^\star = 0$ and for all $\theta, \theta' \in \Theta, h^2(\theta, \theta') = 1 - e^{-|\theta-\theta'|^2/8}$. We shall see in Sect. 8, more precisely in Corollary 4, that for such regular statistical models $r_n(\beta, P_0) \leq \kappa^\star/n$ for some numerical constant $\kappa^\star > 0$, at least for n large enough. Since $h^2(P^\star, P_0) \leq \|P^\star - P_0\| \leq 1/n$, we deduce from Corollary 2 that the posterior $\widehat{\nu}^h_{\boldsymbol{X}}$ concentrates on a set of the form

$$\left\{\theta \in \mathbb{R},\ h^2(\theta, 0) \leq \frac{c}{n}\right\} = \left\{\theta \in \mathbb{R},\ |\theta| \leq \sqrt{8\log\left(\frac{1}{1-c/n}\right)}\right\}$$

with $c > 0$. This set is an interval around 0 of approximate length $1/\sqrt{n}$, at least for n sufficiently large. Despite the contamination of the data, the concentration property of $\widehat{\nu}_X^h$ remains thus the same as in the well-specified case.

7 Applications

7.1 How to choose β in Theorem 1 for a translation model?

In this section, we consider the translation model $\mathscr{M} = \{P_\theta = p(\cdot - \theta) \cdot \mu,\ \theta \in \mathbb{R}\}$ where p is a density on $\mathbb{R}$ with respect to the Lebesgue measure μ. Our aim is to estimate the translation parameter θ by using a prior ν_σ on $\Theta = \mathbb{R}$ with a density (with respect to μ) of the form $q(\cdot/\sigma)/\sigma$ for some density q and positive number σ. We evaluate the estimation error by means of the total variation loss. In order to use our construction we need to tune the parameter β. In Sect. 4.2, we suggested to choose $\beta \geq 1/\sqrt{n}$ satisfying (18). In order to find such a value of $\beta = \beta(\alpha)$, we may proceed as follows. Consider a symmetric bounded interval $I = [-l/2, l/2] \subset \mathbb{R}$ of length $l > 0$ satisfying $\nu_\sigma(I) \geq 1 - \alpha$, hence concentrating most of the mass of the prior ν_σ. If the set $\mathscr{M}(\beta)$ is large enough to contain $\{P_\theta,\ \theta \in I\}$,

$$\pi\left(\mathscr{M}(\beta)\right) \geq \pi\left(\{P_\theta,\ \theta \in I\}\right) = \nu_\sigma(I) \geq 1 - \alpha \tag{47}$$

and β satisfies (18). We deduce from our Corollary 1 that the corresponding posterior $\widehat{\pi}_X^{\mathrm{TV}}$ concentrates with a probability at least $1 - 2e^{-\xi/2}$ on a TV-ball with a radius of order

$$\inf_{P' \in \mathscr{M}(\beta)} \ell(\overline{P}^\star, P') + 2\left(\beta + \frac{2\xi}{n\beta}\right) \leq \inf_{\theta \in I} \ell(\overline{P}^\star, P_\theta) + 2\beta + \frac{4\xi}{\sqrt{n}} = r(\beta). \tag{48}$$

The approximation term $\inf_{\theta \in I} \ell(\overline{P}^\star, P_\theta)$ is small as soon as $\overline{P}^\star$ is close enough to a distribution $P_{\theta^\star}$ whose parameter $\theta^\star$ belongs to I. If we want to prevent us from the situation where $\operatorname{argmin}_{\theta \in \Theta} \ell(\overline{P}^\star, P_\theta)$ is far from 0, we need to increase I (or equivalently diminish α). What would be the consequence on the value of $\beta = \beta(\alpha)$? What if we increase σ, to make the prior distribution flatter, or diminish σ to make it more picky? Finally, what is the influence of the choice of the density q on the size of β?

These are the questions we want to answer in this section. In order to simplify the presentation of our results and avoid technicalities, we make the change of variables $l = 2\sigma t$, or equivalently $t = l/(2\sigma) > 0$, and assume the following.

Assumption 6 The density q is positive, symmetric and decreasing on $\mathbb{R}_+$. There exists some nonnegative and nondecreasing function $\varphi : [0, 1) \to \mathbb{R}_+$ such that

$$\|P_0 - P_\theta\| \leq r \iff |\theta| \leq \varphi(r) \quad \text{for all } r \in [0, 1).$$

When p is symmetric and nonincreasing on $\mathbb{R}_+$, the total variation distance between P_0 and P_θ is given by

$$\|P_0 - P_\theta\| = 2P_0([0, |\theta|/2]) \quad \text{for all } \theta \in \mathbb{R}.$$

Our Assumption 6 is then satisfied with $\varphi(r) = F_0^{-1}[(r+1)/2]$ for all $r \in [0, 1)$, where F_0^{-1} denotes the quantile function of the distribution P_0. We set

$$\overline{\Gamma} = \max\left\{\left[\sup_{0<r\le 1/4} \frac{\varphi(2r)}{\varphi(r)}\right] q(0), \frac{1}{2\varphi(1/4)}\right\} \tag{49}$$

and assume that this quantity is finite. Note that it only depends on $q(0)$ and p. For example, if p is the density $x \mapsto (1/2)e^{-|x|}$,

$$\|P_0 - P_\theta\| = 1 - \exp[-|\theta|/2] \quad \text{and} \quad \varphi : r \mapsto -2\log(1-r).$$

Since the mapping $r \mapsto [\varphi(2r)/\varphi(r)]$ is increasing, we obtain in this case

$$\overline{\Gamma} = \frac{1}{\log(4/3)} \max\left\{q(0)\log 2, \frac{1}{4}\right\}.$$

If now $p : x \mapsto (s/2)(1-|x|)^{s-1}\mathbb{1}_{|x|<1}$ with $s > 0$,

$$\|P_0 - P_\theta\| = 1 - (1-|\theta|/2)^s \quad \text{and} \quad \varphi : r \mapsto 2[1-(1-r)^{1/s}].$$

The mapping $r \mapsto \varphi(2r)/\varphi(r)$ has a continuous extension on $[0, 1/4]$ and is therefore bounded. Given $q(0)$, $\overline{\Gamma}$ is therefore a finite number.

The following result is proven in Sect. 10.3.

Proposition 4 *Assume that Assumption 6 is satisfied and $\overline{\Gamma}$ is finite. Let t be a $(1-\alpha/2)$-quantile of q with $\alpha \le 1/2$. The set $\mathscr{M}(\beta)$ contains the subset $\{P_\theta, \theta \in [-\sigma t, \sigma t]\}$ and therefore satisfies (47) if*

$$\beta \ge \overline{\beta} = \sqrt{\frac{1}{n\gamma} \max\left\{\log\left(\frac{\overline{\Gamma}(\sigma \vee 1)}{q(2t)}\right), \log 4\right\}}. \tag{50}$$

Let us now comment on this result. The quantity $\overline{\beta}$ may be written as $C/\sqrt{n}$ with

$$C = \sqrt{\frac{1}{\gamma} \max\left\{\log\left(\frac{\overline{\Gamma}(\sigma \vee 1)}{q(2t)}\right), \log 4\right\}}.$$

Increasing the value of σ or that of t enlarges the interval $I = [-\sigma t, \sigma t]$. It also makes the value of $C = C(\sigma, t)$ larger. Increasing σ makes the prior ν_σ flatter and for

a fixed value of $t > 0$, $C = C(\sigma)$ increases as $\sqrt{\log \sigma}$ when σ is larger than 1. In the other case, for a fixed value of σ, $C = C(t)$ increases as $\sqrt{\log(1/q(2t))}$. For example, when q is the density of a standard Gaussian random variable, $\sqrt{\log(1/q(2t))}$ is of order t, while for the Laplace and the Cauchy distributions it is of order $\sqrt{t}$ and $\sqrt{\log t}$ respectively. This result illustrates the fact that it is safer to use priors with heavy tails when the size of the location parameter is uncertain. In case of a light-tailed prior, it may be wise to introduce a scaling parameter $\sigma > 1$. By taking $\sigma = 10$, the concentration radius only increases by a factor less than 1.6, while the interval I is ten times longer.

7.2 Fast rates

We go back to the statistical framework described in Sect. 7.1 and consider the special case of the density $p : x \mapsto sx^{s-1}\mathbb{1}_{(0,1]}$ with $s \in (0, 1]$. As before, we choose the TV-loss. In this specific situation,

$$\|P_\theta - P_{\theta'}\| = \left|\theta - \theta'\right|^s \wedge 1 \quad \text{for all } \theta, \theta' \in \mathbb{R} \tag{51}$$

and consequently, $\varphi(r) = r^{1/s}$ for all $r \in [0, 1)$. Besides, the family $\mathscr{T}(\ell, \mathscr{M})$ given by (34) satisfies not only Assumption 3 but also Assumption 4 with $a_2 = 1$. These two facts are proven in Baraud [4, Examples 5 and 6]. As a consequence, Theorem 2 applies. The reader can check that the constants $c = \beta = 0.1$ and $\gamma = 0.01$ satisfy the requirements of Theorem 2 and that its conclusion holds true with $\kappa_0 = 144$.

In order to be more specific about the concentration radius of our posterior $\widehat{\pi}_{\boldsymbol{X}}^{\mathrm{TV}}$, the following proposition provides an upper bound for the quantity $r_n(\beta, P_\theta)$. The proof is postponed to Sect. 10.4.

Proposition 5 *Let t_0 be the third quartile of ν_1. If the density q is positive, symmetric and decreasing on $[0, +\infty)$, for all $\theta \in \mathbb{R}$ the quantity $r_n(\beta, P_\theta)$ is not larger than*

$$\overline{r}_n(\beta, P_\theta) = \frac{2000}{n} \max\left\{ \log\left(\frac{\overline{\Gamma}(\sigma \vee 1)}{q\left[2\left(\frac{|\theta|}{\sigma} \vee t_0\right)\right]} \right), \log 4 \right\}. \tag{52}$$

Then, our Theorem 2 tells us that for all $\xi > 0$, with a probability at least $1 - 2e^{-\xi/2}$, the posterior satisfies

$$\widehat{\pi}_{\boldsymbol{X}}^{\mathrm{TV}}\left(\mathscr{B}(\overline{P}^\star, 144r)\right) \geq 1 - e^{-\xi/2}$$

with

$$r \leq \inf_{\theta \in \mathbb{R}} \left[\left\| \overline{P}^\star - P_\theta \right\| + \overline{r}_n(\beta, P_\theta) \right] + \frac{40\xi}{n}. \tag{53}$$

When the data are i.i.d. with distribution $P_{\theta^\star}$, with probability close to 1, a randomized estimator $P_{\widehat{\theta}}$ with distribution $\widehat{\pi}_{\boldsymbol{X}}^{\mathrm{TV}}$ satisfies with high probability

$$\left|\theta^\star - \widehat{\theta}\right|^s \wedge 1 = \left\| P_{\theta^\star} - P_{\widehat{\theta}} \right\| \leq \frac{C(\xi, s, q, \theta^\star, \sigma)}{n}.$$

This inequality implies, at least for n large enough, that

$$\left|\theta^\star - \widehat{\theta}\right| \leq \frac{C^{1/s}(\xi, s, q, \theta^\star, \sigma)}{n^{1/s}},$$

which means that the parameter $\theta^\star$ is estimated at rate $n^{-1/s}$. This rate is much faster than the usual $(1/\sqrt{n})$-parametric one that is reached by an estimator based on a moment method for instance. For example, when $s = 1/3$ and $n = 100$, a moment estimator provides an accuracy of order 10^{-1} while that of $\widehat{\theta}$ is of order 10^{-6}. Since p is unbounded, note that the maximum likelihood estimator for $\theta^\star$ does not exist and is therefore useless.

It follows from the work of Le Cam that in a translation model $\mathscr{M}$ of the form $\{P_\theta = p(\cdot - \theta) \cdot \mu, \theta \in \mathbb{R}\}$, where p is a density with respect to the Lebesgue measure μ, it is impossible to estimate a distribution $P^\star \in \mathscr{M}$ from an n-sample at a rate faster than $1/n$ for the TV-loss. Because of (51), the rate we get is not only optimal for estimating the distribution $P_{\theta^\star}$ but also for estimating the parameter $\theta^\star$ with respect to the Euclidean distance.

An alternative rate-optimal estimator for estimating $\theta^\star$ is that given by the minimum of the observations. This estimator is unfortunately obviously non-robust to the presence of an outlier among the sample. Our construction provides an estimator which possesses the property of being both rate-optimal and robust.

It also interesting to see how the quantity $\overline{r}_n(\beta, P_\theta)$ given in (52) deteriorates under a misspecification of the prior ν_σ, that is, when the size of the parameter $\theta^\star$ is large compared to σ. When q is Gaussian, $\overline{r}_n(\beta, P_{\theta^\star})$ increases by a factor of order $(\theta^\star/\sigma)^2$ while for the Laplace and Cauchy distributions it is of order $|\theta^\star|/\sigma$ and $\log(|\theta^\star|/\sigma)$ respectively. From these results, we conclude as before that the Cauchy distribution possesses some advantages over the other two distributions when little information is available on the location of the parameter $\theta^\star$.

7.3 A general result under entropy

In this section, we equip $E = \mathbb{R}^k$ with the Lebesgue measure μ and the norm $|\cdot|_\infty$. We consider the TV-loss and the location-scale family

$$\mathscr{M} = \left\{ P_{(p,\mathbf{m},\sigma)} = \frac{1}{\sigma^k} p\left(\frac{\cdot - \mathbf{m}}{\sigma}\right) \cdot \mu, \ p \in \mathcal{M}_0, \ \mathbf{m} \in \mathbb{R}^k, \sigma > 0 \right\}, \tag{54}$$

where $\mathcal{M}_0$ is a set of densities on $\mathbb{R}^k$. Given independent observations $X_1, \ldots, X_n$ with presumed distribution $P^\star = P_{(p^\star, \mathbf{m}^\star, \sigma^\star)} \in \mathscr{M}$, our aim is to estimate the density

$p^\star \in \mathcal{M}_0$, the location parameter $\mathbf{m}^\star \in \mathbb{R}^k$ and the scale parameter $\sigma^\star > 0$, hence the parameter $\theta^\star = (p^\star, \mathbf{m}^\star, \sigma^\star) \in \Theta = \mathcal{M}_0 \times \mathbb{R}^k \times (0, +\infty)$. We assume that the set of densities $\mathcal{M}_0$ satisfies the following conditions.

Assumption 7 Let $\widetilde{D}$ be a continuous nonincreasing mapping from $(0, +\infty)$ to $[1, +\infty)$ such that $\lim_{\eta \to +\infty} \eta^{-2} \widetilde{D}(\eta) = 0$. For all $\eta > 0$, there exists a finite subset $\mathcal{M}_0[\eta] \subset \mathcal{M}_0$ satisfying

$$|\mathcal{M}_0[\eta]| \leq \exp\left[\widetilde{D}(\eta)\right] \tag{55}$$

such that for all $p \in \mathcal{M}_0$, there exists $\overline{p} \in \mathcal{M}_0[\eta]$ that satisfies

$$\left\| P_{(p,\mathbf{0},1)} - P_{(\overline{p},\mathbf{0},1)} \right\| = \frac{1}{2} \int_{\mathbb{R}^k} |p - \overline{p}| \, d\mu \leq \eta. \tag{56}$$

Besides, we assume that there exist $A, s > 0$ such that for all $p \in \mathcal{M}_0$, $\mathbf{m} \in \mathbb{R}^k$ and $\sigma \geq 1$,

$$\left\| P_{(p,\mathbf{0},1)} - P_{(p,\mathbf{m},\sigma)} \right\| \leq \left[A \left(\left(\left| \frac{\mathbf{m}}{\sigma} \right|_\infty \right)^s + \left(1 - \frac{1}{\sigma} \right)^s \right) \right] \bigwedge 1. \tag{57}$$

The first part of Assumption 7, which corresponds to inequalities (55) and (56), aims at measuring the size of the set $\mathcal{M}_0$ by means of its entropy. The entropy of a set controls its metric dimension and usually determines the minimax rate of convergence over it as shown in Birgé [9]. With the second part of Assumption 7, namely inequality (57), we require some regularity properties of the TV-loss with respect to the location and scale parameters. It will be commented on later. We shall see that this condition may be satisfied even when the densities in $\mathcal{M}_0$ are not smooth.

Let us now turn to the choice of our prior. We first consider a countable subset of the parameter space Θ that will be proven to possess good approximation properties. Namely, we define for $\eta, \delta > 0$

$$\Theta[\eta, \delta] = \left\{ \left(\overline{p}, (1+\delta)^{j_0} \delta \mathbf{j}, (1+\delta)^{j_0} \right), \ (\overline{p}, j_0, \mathbf{j}) \in \mathcal{M}_0[\eta] \times \mathbb{Z} \times \mathbb{Z}^k \right\}$$

and we associate a positive weight L_θ with any element $\theta = \theta(\overline{p}, j_0, \mathbf{j}) \in \Theta[\eta, \delta]$ as follows

$$L_\theta = (k+1)L + \log |\mathcal{M}_0[\eta]| + 2 \sum_{i=0}^{k} \log(1 + |j_i|) \tag{58}$$

with $L = \log\left[(\pi^2/3) - 1\right]$. It is not difficult to check that $\sum_{\theta \in \Theta[\eta,\delta]} e^{-L_\theta} = 1$, and we may therefore endow $\mathscr{M}$ with the (discrete) prior π defined as

$$\pi(\{P_\theta\}) = e^{-L_\theta} \quad \text{for all } \theta \in \Theta[\eta, \delta]. \tag{59}$$

With such a prior, our posterior $\widehat{\pi}_{\boldsymbol{X}}^{\mathrm{TV}}$ given in Corollary 1 possesses the following properties.

Corollary 3 *Let $\xi > 0$ $K > 1$. Assume that $\mathcal{M}_0$ satisfies Assumption 7 and define*

$$\eta = \eta_n = \inf \mathscr{D}_n \quad \textit{with} \quad \mathscr{D}_n = \left\{ \eta > 0,\ \widetilde{D}(\eta) \leq \frac{n\eta^2}{24} \right\} \tag{60}$$

$$\delta = \delta_n = \left(\frac{\eta_n}{2A} \right)^{1/s}, \tag{61}$$

$$\beta = \beta_n = \frac{1}{2} \left[K\eta_n + 2\sqrt{\frac{18.6(k+1)}{n}} \right] \tag{62}$$

and the subset $\mathscr{M}_n(K)$ of $\mathscr{M}$ that consists of the elements $P_{(p,\mathbf{m},\sigma)}$ for which

$$|\log \sigma| \vee \left| \frac{\mathbf{m}}{\sigma} \right|_\infty \leq \Lambda_n = \exp \left[\frac{(K^2-1)n\eta_n^2}{48(k+1)} + \log\log(1+\delta_n) \right]. \tag{63}$$

Then, the posterior $\widehat{\pi}_{\boldsymbol{X}}^{TV}$ satisfies the following property: there exists a numerical constant $\kappa_0' > 0$ such that for all $\xi > 0$,

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}} \left({}^{\mathsf{c}}\mathscr{B}(\overline{P}^\star, \kappa_0' r_n) \right) \right] \leq 2e^{-\xi} \tag{64}$$

with

$$r_n = \inf_{P \in \mathscr{M}_n(K)} \ell(\overline{P}^\star, P) + K\eta_n + \sqrt{\frac{k+1}{n}} + \frac{\xi}{\sqrt{n(k+1)}} \wedge \frac{\xi}{Kn\eta_n}. \tag{65}$$

Let us now comment on this result. The radius r_n is the sum of three main terms, omitting the dependency with respect to ξ. The first one, $\inf_{P \in \mathscr{M}_n(K)} \ell(\overline{P}^\star, P)$, corresponds to the approximation of $\overline{P}^\star$ by an element of $\mathscr{M}$ whose location and scale parameters satisfy the constraints given in (63). The quantity η_n, involved in the second term, usually corresponds to the minimax rate for solely estimating a density $p \in \mathcal{M}_0$ from an n-sample. Finally, the third term $\sqrt{(k+1)/n}$ corresponds to the rate we would get for solely estimating the location and translation parameters $(\mathbf{m}, \sigma) \in \mathbb{R}^{k+1}$ when the density p is known.

Let us now provide some examples for which our condition (57) is satisfied. We start with an example where the densities in $\mathcal{M}_0$ are smooth.

Lemma 2 *Assume that the set $\mathcal{M}_0$ consists of densities p that are supported on $[0,1]^k$, satisfy $\sup_{p \in \mathcal{M}_0} \|p\|_\infty \leq L_0$ and*

$$\sup_{p \in \mathcal{M}_0} \left| p(\boldsymbol{x}) - p(\boldsymbol{x}') \right| \leq L_1 \left| \boldsymbol{x} - \boldsymbol{x}' \right|^s \quad \textit{for all } \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^k, \tag{66}$$

with constants $L_0, L_1 > 0$ and $s \in (0,1]$. Then (57) is satisfied with $A = L_1 \vee [(1 + L_1 k^{s/2} + L_0)/2]$.

Nevertheless, condition (57) may also be satisfied for families $\mathcal{M}_0$ of densities which are not smooth, as shown in Lemma 3 below. It makes it possible to consider the following example.

Example 6 We consider here the situation where $k = 1$ and $\mathcal{M}_0$ is the set of all nonincreasing densities on $[0, 1]$ that are bounded by $B > 1$. Then, $\mathscr{M}$ consists of all the probabilities whose densities are supported on intervals I with positive lengths, nonincreasing on I and which are bounded by $B/\mu(I)$. Birman and Solomjak [15] proved that $\mathcal{M}_0$ satisfies Assumption 7 with $\widetilde{D}(\eta)$ of order $(1/\eta) \vee 1$ (up to some constant that depends on B). We deduce from (60) that η_n is therefore of order $n^{-1/3}$. Besides, it follows from Lemma 3 below that (57) is satisfied with $A = B$ and $s = 1$. We may therefore apply Corollary 3. For a value of K large enough compared to 1, Λ_n defined by (63) is larger than $\exp\left[CK^2n^{1/3}\right]$ for some constant $C > 0$ (depending on A). In particular, if $X_1, \ldots, X_n$ are i.i.d. with a density of the form

$$x \mapsto p^\star(x) = \frac{1}{\sigma^\star} p\left(\frac{x - m^\star}{\sigma^\star}\right)$$

where $p \in \mathcal{M}_0$, $|m^\star/\sigma^\star| \leq \exp\left[CK^2n^{1/3}\right]$ and

$$\exp\left[-\exp\left[CK^2n^{1/3}\right]\right] \leq \sigma^\star \leq \exp\left[\exp\left[CK^2n^{1/3}\right]\right],$$

(64) is satisfied with r_n of order $C'n^{-1/3}$ where the constant $C' > 0$ only depends on ξ, K, B but not on $m^\star$ and $\sigma^\star$. This means that the concentration properties of $\widehat{\pi}_{\boldsymbol{X}}$ hold true uniformly over a huge range of translation and scale parameters $\mathbf{m}$ and σ when n is large enough.

Lemma 3 *Let p be a nonincreasing density on $(0, +\infty)$. For all $\sigma \geq 1$*

$$\frac{1}{2}\int_{\mathbb{R}} \left|\frac{1}{\sigma} p\left(\frac{x}{\sigma}\right) - p(x)\right| dx \leq \left(1 - \frac{1}{\sigma}\right). \tag{67}$$

If, furthermore, p is bounded by $B \geq 1$, for all $m \in \mathbb{R}$,

$$\frac{1}{2}\int_{\mathbb{R}} |p(x) - p(x - m)|\, dx \leq (|m|B) \wedge 1. \tag{68}$$

In particular, for all $m \in \mathbb{R}$ and $\sigma \geq 1$,

$$\frac{1}{2}\int_{\mathbb{R}} \left|\frac{1}{\sigma} p\left(\frac{x - m}{\sigma}\right) - p(x)\right| dx \leq \left[B\left|\frac{m}{\sigma}\right| + \left(1 - \frac{1}{\sigma}\right)\right] \wedge 1. \tag{69}$$

7.4 Estimating a parameter under sparsity

Let us consider a parametric dominated model $\mathscr{M} = \left\{P_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}} \cdot \mu,\ \boldsymbol{\theta} \in \mathbb{R}^k\right\}$ where the dimension k of the parameters is large. We presume, even though this might not

be true, that the data are i.i.d. with distribution $P_{\boldsymbol{\theta}^\star} \in \mathscr{M}$ and that the coordinates of the true parameter $\boldsymbol{\theta}^\star = (\theta_1^\star, \ldots, \theta_k^\star)$ are all zero except for a small number of them. Our aim is to estimate $P_{\boldsymbol{\theta}^\star}$ from the observation of $X_1, \ldots, X_n$ by using the squared Hellinger loss.

To tackle this problem, we partition the model $\mathscr{M}$ into the sub-models $\{\mathscr{M}_m,\ m \subset \{1, \ldots, k\}\}$ where $\mathscr{M}_m$ consists of those distributions $P_{\boldsymbol{\theta}} \in \mathscr{M}$ for which the coordinates of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ are all zero except those with an index $i \in m$. We denote by Θ_m the set of such parameters, so that $\mathscr{M}_m = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_m\}$, and we use the conventions $\Theta_\varnothing = \{\mathbf{0}\}$ and $\mathscr{M}_\varnothing = \{P_{\mathbf{0}}\}$. Given some positive number $R > 0$, we equip each parameter space $\Theta_m, m \subset \{1, \ldots, k\}$, with the uniform distribution ν_m on $\Theta_m(R) = [-R, R]^k \cap \Theta_m$ when $m \neq \varnothing$ and the Dirac mass $\nu_\varnothing = \delta_{\mathbf{0}}$ at $\mathbf{0} \in \mathbb{R}^k$ when $m = \varnothing$. We may then define on $\mathbb{R}^k = \bigcup_{m \subset \{1,\ldots,k\}} \Theta_m$, the hierarchical prior

$$\nu = \sum_{m \subset \{1,\ldots,k\}} e^{-L_m} \nu_m \quad \text{with} \quad L_m = |m| \log k + k \log \left(1 + \frac{1}{k}\right). \tag{70}$$

We endow $\mathscr{M}$ with the σ-algebra and the prior π as described in Sect. 2.1. Besides, we assume that there exists $s \in (0, 1]$ and a positive number $B_k = B_k(R)$, possibly depending on k and R (although we drop the dependency with respect to R), such that

$$h^2\left(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}'}\right) \leq B_k \left|\boldsymbol{\theta} - \boldsymbol{\theta}'\right|_\infty^s \quad \text{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in [-R, R]^k. \tag{71}$$

The following result is proven in Sect. 10.8.

Proposition 6 *Assume that*

$$p : \left| \begin{array}{rcl} E \times \mathbb{R}^k & \longrightarrow & \mathbb{R}_+ \\ (x, \boldsymbol{\theta}) & \longmapsto & p_{\boldsymbol{\theta}}(x) \end{array} \right.$$

is measurable. If $R B_k^{1/s} \geq 1$ *there exists a numerical constant* $\kappa_0' > 0$ *such that for any distribution* $\mathbf{P}^\star$ *and* $\xi > 0$

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}^h \left({}^{\mathsf{c}}\mathscr{B}(\overline{P}^\star, \kappa_0' r)\right)\right] \leq 2e^{-\xi}$$

where

$$r = \inf_{m \subset \{1,\ldots,k\}} \left[\inf_{\boldsymbol{\theta} \in \Theta_m(R)} \ell(\overline{P}^\star, P_{\boldsymbol{\theta}}) + \frac{|m| \log\left(2kR(nB_k)^{1/s}\right) + \xi}{n} \right]. \tag{72}$$

Let us now comment on this result. First of all, the mapping

$$R \mapsto \sup \left\{ \frac{h^2\left(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}'}\right)}{\left|\boldsymbol{\theta} - \boldsymbol{\theta}'\right|_\infty^s},\ \boldsymbol{\theta} \neq \boldsymbol{\theta}',\ \boldsymbol{\theta}, \boldsymbol{\theta}' \in [-R, R]^k \right\}$$

being nondecreasing, our condition $RB_k^{1/s} = R[B_k(R)]^{1/s} \geq 1$ is always satisfied for a value of R sufficiently large.

When B_k does not increase faster than a power of k, the radius r given in (72) only depends logarithmically on the dimension k of the parameter space, as expected.

Let us now illustrate Proposition 6 by choosing some specific models $\mathscr{M} = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \mathbb{R}^k\}$. If $P_{\boldsymbol{\theta}}$ is the Gaussian distribution with mean $\boldsymbol{\theta} \in \mathbb{R}^k$ and covariance matrix $\sigma^2 I_k$, where I_k denotes the $k \times k$ identity matrix,

$$h^2(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}'}) = 1 - \exp\left[-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2}{8\sigma^2}\right] \leq \frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2}{8\sigma^2} \leq \frac{k\,|\boldsymbol{\theta} - \boldsymbol{\theta}'|_\infty^2}{8\sigma^2}.$$

Then, inequality (71) is satisfied with $B_k = k/(8\sigma^2)$ and $s = 2$. In particular, our condition $RB_k^{1/s} \geq 1$ is equivalent to $R \geq 2\sigma\sqrt{(2/k)}$. In this case, the value of r given by (72) is of order

$$\inf_{m \subset \{1,\ldots,k\}} \left[\inf_{\boldsymbol{\theta} \in \Theta_m(R)} \ell(\overline{P}^\star, P_{\boldsymbol{\theta}}) + \frac{|m| \log(knR/\sigma) + \xi}{n}\right].$$

More generally, if $\mathscr{M} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \mathbb{R}^k\}$ is a regular statistical model with a nonsingular Fisher information matrix $\mathbf{J}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbb{R}^k$, we know from the book of Ibragimov and Has'minskiĭ [20, Theorem 7.1, p. 81] that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^k$ such that $\boldsymbol{\theta}, \boldsymbol{\theta}' \in [-R, R]^k$

$$h^2(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}'}) \leq \frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2}{8} \sup_{\boldsymbol{\theta}'' \in \mathbb{R}^k, |\boldsymbol{\theta}''|_\infty \leq R} \operatorname{tr}\left(\mathbf{J}(\boldsymbol{\theta}'')\right).$$

Then, Assumption (71) holds with $s = 2$ and we may take

$$B_k = \frac{k^2}{8} \sup_{\boldsymbol{\theta}'' \in \mathbb{R}^k, |\boldsymbol{\theta}''|_\infty \leq R} \varrho\left(\mathbf{J}(\boldsymbol{\theta}'')\right)$$

where $\varrho\left(\mathbf{J}(\boldsymbol{\theta}'')\right)$ denotes the largest eigenvalue of the matrix $\mathbf{J}(\boldsymbol{\theta}'')$. This value is independent of $\boldsymbol{\theta}''$ when $\mathscr{M}$ is a translation model.

Finally note that the second term in (72) only increases logarithmically with respect to R, at least when $B_k = B_k(R)$ does not increase faster than a power of R. By taking larger values of R one may therefore considerably enlarge the sizes of the cubes $\Theta_m(R)$, and therefore diminish the approximation term in (72), while only slightly increasing the second term $[|m| \log(2kR(nB_k)^{1/s}) + \xi]/n$.

8 Some tools for evaluating $r_n(\beta, P)$

The aim of this section is to provide some mathematical results that allow one to bound the quantity $r_n(\beta, P)$ from above, or at least evaluate its order of magnitude, when n is sufficiently large. Throughout this section, we consider a parametric statistical model

$\mathscr{M} = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\}$ where the parameter space $\Theta \subset \mathbb{R}^k$ is endowed with a prior ν which admits a density q with respect to the Lebesgue measure on $\mathbb{R}^k$. In order to use the definition (11) of the quantity $r_n(\beta, P)$, we assume that we have at disposal a family $\mathscr{T}(\ell, \mathscr{M})$ that satisfies our Assumption 3, which provides us with a value of $a_1 > 0$, as well as a value γ that satisfy the requirements of our main theorems. Our aim is to bound $r_n(\beta, P)$ as a function of a_1, γ, β, k and n under suitable assumptions on the density q and the behaviour of the loss ℓ. Once ℓ and $\mathscr{T}(\ell, \mathscr{M})$ are given, a_1 and γ can be considered as fixed numerical constants. The value of β can also be considered as a numerical constant when Theorem 2 applies. Otherwise, it can be chosen of order $\sqrt{k/n}$ as in our Example 1.

8.1 Bounding $r_n(\beta, P_\theta)$ in parametric models

In what follows, $|\cdot|_*$ denotes some arbitrary norm on $\mathbb{R}^k$ and $\mathcal{B}_*(\boldsymbol{x}, z)$ the corresponding closed ball centered at $\boldsymbol{x} \in \mathbb{R}^k$ with radius $z \geq 0$.

Assumption 8 Let $\boldsymbol{\theta}^\star$ be an element of $\Theta \subset \mathbb{R}^k$.

(i) There exist positive numbers $\underline{a}, \overline{a}$ and s such that

$$\underline{a}\left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_*^s \leq \ell(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) \leq \overline{a}\left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_*^s \quad \text{for all } \boldsymbol{\theta} \in \Theta. \tag{73}$$

(ii) There exists a positive nonincreasing function $\upsilon_{\boldsymbol{\theta}}$ on $\mathbb{R}_+$ such that

$$\nu(\mathcal{B}_*(\boldsymbol{\theta}^\star, 2x)) \leq \upsilon_{\boldsymbol{\theta}^\star}(x)\nu(\mathcal{B}_*(\boldsymbol{\theta}^\star, x)) \quad \text{for all } x > 0. \tag{74}$$

Under Assumption 8-(i), the loss function behaves like a power of a norm between the parameters.

The following result is an extension of Proposition 10 in Baraud and Birgé [6]. It was established there for the special case of the squared Hellinger loss and we provide here an extension to an arbitrary one. Since the proof follows the same lines, we omit it.

Proposition 7 *Under Assumption 8,*

$$r_n(\beta, P_{\boldsymbol{\theta}^\star}) \leq \inf\left\{r \geq \frac{1}{n\beta a_1},\ r \geq \frac{\varrho_0 \log\left[\upsilon_{\boldsymbol{\theta}^\star}\left([r/\overline{a}]^{1/s}\right)\right]}{\gamma n \beta a_1}\right\} \tag{75}$$

with $\varrho_0 = 1 + \log(2\overline{a}/\underline{a})/[s \log 2]$. *If* $\upsilon_{\boldsymbol{\theta}^\star} \equiv \upsilon > 0$, *then*

$$r_n(\beta, P_{\boldsymbol{\theta}^\star}) \leq \frac{(\varrho_0 \log \upsilon) \vee 1}{a_1 n \gamma \beta}. \tag{76}$$

If Assumption 8-(i) is satisfied and if the parameter space Θ *is convex and* q *satisfies*

$$\underline{b} \leq q(\boldsymbol{\theta}) \leq \overline{b} \quad \textit{for all } \boldsymbol{\theta} \in \Theta \quad \textit{with } 0 < \underline{b} \leq \overline{b}, \tag{77}$$

then Assumption 8-(ii) holds with $\upsilon_{\boldsymbol{\theta}^\star} \equiv 2^k(\overline{b}/\underline{b})$. *Consequently,*

$$r_n(\beta, P_{\boldsymbol{\theta}^\star}) \leq \frac{\varrho_1}{a_1\gamma}\frac{k}{n\beta} \quad \textit{with} \quad \varrho_1 = \left[\varrho_0 \log\left(2\left[\overline{b}/\underline{b}\right]^{1/k}\right)\right] \vee 1. \tag{78}$$

When Assumption 8-(i) is satisfied and ν admits a density which is bounded away from 0 and infinity on a convex parameter space $\Theta \subset \mathbb{R}^k$, $r_n(\beta, P_{\boldsymbol{\theta}})$ is of order $k/(n\beta)$ for all $\boldsymbol{\theta} \in \Theta$. This result may also hold true when the density is not bounded away from infinity as shown in the following example. If $k = 1$, $\Theta = [-1, 1]$ and $q : \theta \mapsto (t/2)|\theta|^{t-1}\mathbb{1}_{[-1,1]}(\theta)$ with $t \in (0, 1)$, Assumption 8-(ii) holds with $\upsilon_\theta \equiv 2^{1+t}\left(2^t - 1\right)^{-1}$ for all $\theta \in [-1, 1]$—see Baraud and Birgé [6, Proposition 11]. Then (76) still applies. In the other direction, when the density q takes very small values in the neighbourhood of the parameter $\boldsymbol{\theta}$, the function $\upsilon_{\boldsymbol{\theta}}$ may take large values around 0. This is for example the case when q is proportional to $\theta \mapsto \exp\left[-1/\left(2|\theta|^t\right)\right]\mathbb{1}_{[-1,1]}(\theta)$, $t > 0$, and $\theta = 0$. It follows from Baraud and Birgé [6, Proposition 12] (and its proof) that Assumption 8-(ii) is satisfied with $\upsilon_\theta : x \mapsto \exp(c(t)/x^t)$ for some quantity $c(t) > 0$. Applying (75) leads to an upper bound on $r_n(\beta, P_\theta)$ of order $(n\beta)^{-s/(s+t)}$.

8.2 Some asymptotic order of magnitude

In Sect. 8.2, we have given some general tools for controlling the quantity $r_n(\beta, P_{\boldsymbol{\theta}})$ for a given value of n. In this section, we present some sufficient conditions under which $r_n(\beta, P_{\boldsymbol{\theta}})$ is of order $k/(n\beta)$ at least when n is large enough. These conditions are not the weakest possible ones but they have the advantage to be relatively easy to check on many examples.

Assumption 9 The density q is continuous and positive at $\boldsymbol{\theta}^\star \in \Theta$. The loss function ℓ satisfies the following properties for some positive number $s > 0$ and a norm $|\cdot|_*$ on $\mathbb{R}^k$.

(i) For all $\varepsilon > 0$, there exists $z = z(\varepsilon) > 0$ such that

$$(1-\varepsilon)\left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_*^s \leq \ell(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) \leq (1+\varepsilon)\left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_*^s \quad \text{for all } \boldsymbol{\theta} \in \mathcal{B}_*(\boldsymbol{\theta}^\star, z).$$

(ii) There exists a subset $\mathcal{K} \subset \Theta$, the interior of which contains $\boldsymbol{\theta}^\star$, that satisfies for some positive numbers $\underline{a}_{\mathcal{K}}$ and η:

$$\underline{a}_{\mathcal{K}}\left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_*^s \leq \ell(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) \quad \text{for } \boldsymbol{\theta} \in \mathcal{K} \text{ and for } \boldsymbol{\theta} \notin \mathcal{K} \quad \ell(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) \geq \eta > 0. \tag{79}$$

Under these assumptions, we establish the following proposition, the proof of which is postponed to Sect. 10.10.

Proposition 8 *Under Assumption 9, at least for n sufficiently large,*

$$r_n\left(\beta, P_{\boldsymbol{\theta}^\star}\right) \leq \frac{(1+1/s)}{a_1\gamma}\frac{k}{n\beta}. \tag{80}$$

8.3 The case of the squared Hellinger loss on a regular statistical model

Of particular interest is the situation where the statistical model $\mathscr{M} = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset \mathbb{R}^k$, is regular. There exist several ways of defining a regular model in statistics and we adopt here the definition of Ibragimov and Has'minskiĭ [20].

Definition 1 Let μ be a measure on $(E, \mathcal{E})$ and Θ an open subset of $\mathbb{R}^k$. The statistical model $\mathscr{M} = \{P_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}} \cdot \mu,\ \boldsymbol{\theta} \in \Theta\}$ is said to be *regular* if the family of functions $\{\zeta_{\boldsymbol{\theta}} = \sqrt{p_{\boldsymbol{\theta}}},\ \theta \in \Theta\} \subset \mathscr{L}_2(E, \mathcal{E}, \mu)$ satisfies the following properties.

(i) For μ-almost all $x \in E$, $\boldsymbol{\theta} \mapsto \zeta_{\boldsymbol{\theta}}(x)$ is continuous.
(ii) For all $\boldsymbol{\theta} \in \Theta$, there exists $\dot{\boldsymbol{\zeta}}_{\boldsymbol{\theta}} = (\dot{\zeta}_{\boldsymbol{\theta},1}, \ldots, \dot{\zeta}_{\boldsymbol{\theta},k}) : E \to \mathbb{R}^k$ such that

$$\int_E \left|\dot{\boldsymbol{\zeta}}_{\boldsymbol{\theta}}(x)\right|^2 d\mu(x) < +\infty$$

and

$$\int_E \left|\zeta_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}(x) - \zeta_{\boldsymbol{\theta}}(x) - \langle \dot{\boldsymbol{\zeta}}_{\boldsymbol{\theta}}(x), \boldsymbol{\epsilon}\rangle\right|^2 d\mu(x) = o(|\boldsymbol{\epsilon}|^2) \quad \text{when } |\boldsymbol{\epsilon}| \to 0.$$

(iii) For all $i \in \{1, \ldots, k\}$, the mapping $\boldsymbol{\theta} \mapsto \dot{\zeta}_{\boldsymbol{\theta},i}$ is continuous in $\mathscr{L}_2(E, \mathcal{E}, \mu)$.

When the model is regular, the matrix

$$\mathbf{J}(\boldsymbol{\theta}) = \left(4 \int_E \dot{\zeta}_{\boldsymbol{\theta},i}(x)\dot{\zeta}_{\boldsymbol{\theta},j}(x) d\mu(x)\right)_{\substack{1 \le i \le k \\ 1 \le j \le k}},$$

is called the *Fisher information matrix*.

The matrix $\mathbf{J}(\boldsymbol{\theta})$ is symmetric and nonnegative and we may therefore consider its square root $\mathbf{J}^{1/2}(\boldsymbol{\theta})$, that is, the symmetric $(k \times k)$-nonnegative matrix that satisfies $\mathbf{J}^{1/2}(\boldsymbol{\theta})\mathbf{J}^{1/2}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta})$.

Regular statistical models enjoy nice metric properties that are described in Proposition 9 below. For a proof we refer the reader to Ibragimov and Has'minskiĭ [20]—Lemma 7.1 page 65, Theorem 7.6 page 81 and its proof.

Proposition 9 *Let Θ be an open subset of $\mathbb{R}^k$ and $\boldsymbol{\theta}^\star \in \Theta$. If $\mathscr{M} = \{P_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}} \cdot \mu,\ \boldsymbol{\theta} \in \Theta\}$ is regular and the Fisher information matrix $\mathbf{J}(\boldsymbol{\theta}^\star)$ nonsingular at $\boldsymbol{\theta}^\star \in \Theta$, Assumption 9-(i) is satisfied with $\ell = h^2$, $s = 2$ and for the norm $|\cdot|_*$ defined by*

$$|\boldsymbol{x}|_* = \frac{1}{\sqrt{8}}\left|\mathbf{J}^{1/2}(\boldsymbol{\theta}^\star)\boldsymbol{x}\right| \quad \textit{for all } \boldsymbol{x} \in \mathbb{R}^k. \tag{81}$$

Besides, for any compact subset $\mathcal{K} \subset \Theta$ there exist positive numbers $\overline{a}_{\mathcal{K}}, \underline{a}_{\mathcal{K}}$ such that

$$\underline{a}_{\mathcal{K}}\left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_*^2 \le h^2\left(\boldsymbol{\theta}, \boldsymbol{\theta}^\star\right) \le \overline{a}_{\mathcal{K}}\left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_*^2 \quad \textit{for all } \boldsymbol{\theta} \in \mathcal{K}. \tag{82}$$

Using Proposition 8, we immediately infer the following result.

Corollary 4 *Let Θ be an open subset of $\mathbb{R}^k$. Assume that $\mathscr{M} = \{P_\theta = p_\theta \cdot \mu,\ \theta \in \Theta\}$ is regular and the Fisher information matrix $\mathbf{J}(\theta^\star)$ nonsingular at $\theta^\star \in \Theta \subset \mathbb{R}^k$. Assume that there exists a compact set $\mathcal{K} \subset \Theta$, containing $\theta^\star$ in its interior, such that $h(\theta, \theta^\star) \geq \eta > 0$ for all $\theta \notin \mathcal{K}$. Assume furthermore that the density q is continuous and positive at $\theta^\star$. Then, $r_n(\beta, P_{\theta^\star}) \leq [3/(2a_1\gamma\beta)](k/n)$, at least for n sufficiently large.*

9 Proofs of Theorems 1, 2 and 3

Throughout this proof we fix some $\overline{Q} \in \mathscr{M}$, $r, \beta > 0$ and use the following notation: $c_1 = 1 + c$, $c_2 = 2 + c$,

$$\mathcal{V}(\pi, \overline{Q}) = \left\{r > 0, \pi\left(\mathscr{B}(\overline{Q}, r)\right) > 0\right\}$$

and for $r \in \mathcal{V}(\pi, \overline{Q})$, $\mathscr{B} = \mathscr{B}(\overline{Q}, r)$ and $\pi_{\mathscr{B}} = [\pi(\mathscr{B})]^{-1}\, \mathbb{1}_{\mathscr{B}} \cdot \pi$.

9.1 Main parts of the proofs of Theorems 1 and 2

Throughout the proofs of these two theorems we fix some positive number z, that will be chosen later on, $r \geq r_n(\beta, \overline{Q})$ and set

$$A = \left\{\int_{\mathscr{M}} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P) > z\right\}.$$

It follows from the definition (7) of $\widehat{\pi}_{\boldsymbol{X}}$ that for all $J \in \mathbb{N}$

$$\begin{aligned}
\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)\right)\right] &= \mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)\right) \mathbb{1}_{{}^{\mathsf{c}}A}\right] + \mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)\right) \mathbb{1}_{A}\right] \\
&\leq \mathbb{P}({}^{\mathsf{c}}A) + \frac{1}{z}\mathbb{E}\left[\int_{{}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P)\right] \\
&= \mathbb{P}({}^{\mathsf{c}}A) + \frac{1}{z}\int_{{}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P).
\end{aligned} \tag{83}$$

In a first step, we prove that for some well chosen values of β, z, r and for J large enough, each of the two terms in the right-hand side of (83) is not larger than $e^{-\xi}$. To achieve this goal, we bound the first term of the right-hand side of (83) by applying Markov's inequality

$$\begin{aligned}
\mathbb{P}({}^{\mathsf{c}}A) &= \mathbb{P}\left[\int_{\mathscr{M}} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P) \leq z\right] \\
&= \mathbb{P}\left[\left[\int_{\mathscr{M}} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P)\right]^{-1} \geq z^{-1}\right]
\end{aligned}$$

$$\leq z\mathbb{E}\left[\frac{1}{\int_{\mathscr{M}}\exp\left[-\beta\mathbf{T}(\boldsymbol{X},P)\right]d\pi(P)}\right]$$

and then by using Lemma 6, we obtain that

$$\mathbb{P}({}^{\mathsf{c}}A)\leq\frac{z}{\pi^2(\mathscr{B})}\left[\int_{\mathscr{B}^2}\exp\left[-\mathbf{L}(P,Q)\right]d\pi_{\mathscr{B}}(P)d\pi_{\mathscr{B}}(Q)\right]^{-1}. \tag{84}$$

We therefore have a control of $\mathbb{P}({}^{\mathsf{c}}A)$ by choosing z small enough. We bound the second term of (83) by using Lemma 5.

We then finish the proofs of Theorems 1 and 2 as follows. In the context of Theorem 1, we finally establish that for a suitable value of J and all $\overline{Q}\in\mathscr{M}(\beta)$,

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{Q},2^Jr)\right)\right]\leq 2e^{-\xi}\quad\text{with}\quad r=r(\overline{Q})=\ell(\overline{P}^{\star},\overline{Q})+a_1^{-1}\left(\beta+\frac{2\xi}{n\beta}\right).$$

By (3), $\mathscr{B}(\overline{Q},2^Jr)\subset\mathscr{B}(\overline{P}^{\star},\tau\ell(\overline{P}^{\star},\overline{Q})+\tau 2^Jr)$ for all $\overline{Q}\in\mathscr{M}(\beta)$, and consequently $\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{P}^{\star},\overline{r}\right)\right]\leq 2e^{-\xi}$ with

$$\overline{r}=\overline{r}(\overline{Q})=\tau\left[\ell(\overline{P}^{\star},\overline{Q})+2^Jr\right]=\tau\left[(1+2^J)\ell(\overline{P}^{\star},\overline{Q})+2^Ja_1^{-1}\left(\beta+\frac{2\xi}{n\beta}\right)\right].$$

We obtain (16) by monotone convergence, taking a sequence $(\overline{Q}_N)_{N\geq 0}\subset\mathscr{M}(\beta)$ such that $\ell(\overline{P}^{\star},\overline{Q}_N)$ is nonincreasing to $\inf_{P\in\mathscr{M}(\beta)}\ell(\overline{P}^{\star},P)$, so that

$$\begin{aligned}\lim_{N\to+\infty}\overline{r}(\overline{Q}_N)&=\tau\left[(1+2^J)\inf_{\overline{Q}\in\mathscr{M}(\beta)}\ell(\overline{P}^{\star},\overline{Q})+2^Ja_1^{-1}\left(\beta+\frac{2\xi}{n\beta}\right)\right]\\&\leq\tau(1+2^J)\left[\inf_{\overline{Q}\in\mathscr{M}(\beta)}\ell(\overline{P}^{\star},\overline{Q})+a_1^{-1}\left(\beta+\frac{2\xi}{n\beta}\right)\right]\end{aligned}$$

and (16) holds provided that $\kappa_0\geq\tau(2^J+1)$.

In the context of Theorem 2, we show that for some suitable value of J and all $\overline{Q}\in\mathscr{M}$,

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{Q},2^Jr)\right)\right]\leq 2e^{-\xi}\quad\text{with}\quad r=\ell(\overline{P}^{\star},\overline{Q})+r_n(\overline{Q},\beta)+\frac{2\xi}{n\beta a_1},$$

and we get (28) by arguing similarly.

9.2 Preliminary results

In the proofs of Theorems 1 and 2, we use the following consequence of our Assumption 3. We may write

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[t_{(P,Q)}(X_i)\right]=\mathbb{E}_S\left[t_{(P,Q)}(X)\right]\quad\text{with } S=\overline{P}^{\star}=\frac{1}{n}\sum_{i=1}^{n}P_i^{\star}\in\mathscr{P}$$

and we deduce from (5) that for all $P,\,Q\in\mathscr{M}$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[t_{(P,Q)}(X_i)\right]\le a_0\ell(\overline{P}^{\star},\,P)-a_1\ell(\overline{P}^{\star},\,Q). \tag{85}$$

Besides, using the antisymmetry property (ii) we also obtain that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[t_{(P,Q)}(X_i)\right]\ge a_1\ell(\overline{P}^{\star},\,P)-a_0\ell(\overline{P}^{\star},\,Q). \tag{86}$$

For the proof of Theorems 2, we additionnally use the following consequence of our Assumption 4. By taking $S=\overline{P}^{\star}$ and using the convexity of the mapping $u\mapsto u^2$, we deduce that for all $P,\,Q\in\mathscr{M}$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left[t_{(P,Q)}(X_i)\right]&=\mathbb{E}_S\left[t^2_{(P,Q)}(X)\right]-\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[t_{(P,Q)}(X_i)\right]\right)^2\\&\le\mathbb{E}_S\left[t^2_{(P,Q)}(X)\right]-\left(\mathbb{E}_S\left[t_{(P,Q)}(X)\right]\right)^2\\&=\mathrm{Var}_S\left[t_{(P,Q)}(X)\right]\end{aligned}$$

and it follows then from Assumption 4 (iv) that for all $P,\,Q\in\mathscr{M}$

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left[t_{(P,Q)}(X_i)\right]\le a_2\left[\ell(\overline{P}^{\star},\,P)+\ell(\overline{P}^{\star},\,Q)\right]. \tag{87}$$

The proofs of our main results rely on the following lemmas.

Lemma 4 *Let $(U,\,V)$ be a pair of random variables with values in a product space $(E\times F,\,\mathcal{E}\otimes\mathcal{F})$ and marginal distributions P_U and P_V respectively. For all measurable function h on $(E\times F,\,\mathcal{E}\otimes\mathcal{F})$,*

$$\mathbb{E}_U\left[\frac{1}{\mathbb{E}_V\left[\exp\left[-h(U,\,V)\right]\right]}\right]\le\left[\mathbb{E}_V\left[\frac{1}{\mathbb{E}_U\left[\exp\left[h(U,\,V)\right]\right]}\right]\right]^{-1}.$$

This lemma is proven in Audibert and Catoni [3, Lemma 4.2, p. 28].

Lemma 5 *For $P, Q \in \mathscr{M}$, we set*

$$\mathbf{M}(P, Q) = \log\left[\int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right] d\pi(Q')\right]\right].$$

For all $r \in \mathcal{V}(\pi, \overline{Q})$ and $P \in \mathscr{M}$,

$$\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P)\right]\right] \leq \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}} \exp\left[-\mathbf{M}(P, Q)\right] d\pi_{\mathscr{B}}(Q)\right]^{-1}. \tag{88}$$

Proof Let $r \in \mathcal{V}(\pi, \overline{Q})$. For $P, Q \in \mathscr{M}$, we set

$$I(\boldsymbol{X}, P, Q) = c_1\beta\mathbf{T}(\boldsymbol{X}, P, Q) - \log\int_{\mathscr{M}} \exp\left[c\beta\mathbf{T}(\boldsymbol{X}, P, Q')\right] d\pi(Q').$$

Then,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[-I(\boldsymbol{X}, P, Q)\right]\right] \\
&\quad= \mathbb{E}\left[\exp\left[-c_1\beta\mathbf{T}(\boldsymbol{X}, P, Q) + \log\int_{\mathscr{M}} \exp\left[c\beta\mathbf{T}(\boldsymbol{X}, P, Q')\right] d\pi(Q')\right]\right] \\
&\quad= \mathbb{E}\left[\int_{\mathscr{M}} \exp\left[c\beta\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\beta\mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q')\right] \\
&\quad= \exp\left[\mathbf{M}(P, Q)\right].
\end{aligned} \tag{89}$$

Since $\lambda = c_1\beta = (1 + c)\beta$, it follows from the convexity of the exponential that

$$\begin{aligned}
\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P)\right]\right] &= \mathbb{E}\left[\exp\left[\int_{\mathscr{M}} [-\beta\mathbf{T}(\boldsymbol{X}, P, Q)] d\widetilde{\pi}_{\boldsymbol{X}}(Q|P)\right]\right] \\
&\leq \mathbb{E}\left[\int_{\mathscr{M}} \exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P, Q)\right] d\widetilde{\pi}_{\boldsymbol{X}}(Q|P)\right] \\
&= \mathbb{E}\left[\frac{\int_{\mathscr{M}} \exp\left[c\beta\mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}{\int_{\mathscr{M}} \exp\left[c_1\beta\mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}\right] \\
&\leq \mathbb{E}\left[\frac{\int_{\mathscr{M}} \exp\left[c\beta\mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}{\int_{\mathscr{B}} \exp\left[c_1\beta\mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P)\right]\right] &\leq \mathbb{E}\left[\frac{1}{\int_{\mathscr{B}} \exp\left[I(\boldsymbol{X}, P, Q)\right] d\pi(Q)}\right] \\
&= \frac{1}{\pi(\mathscr{B})}\mathbb{E}\left[\frac{1}{\int_{\mathscr{B}} \exp\left[I(\boldsymbol{X}, P, Q)\right] d\pi_{\mathscr{B}}(Q)}\right].
\end{aligned}$$

Applying Lemma 4 with $U = \boldsymbol{X}$, $V = Q$ with distribution $\pi_{\mathscr{B}}$, and $h(U, V) = -I(\boldsymbol{X}, P, Q)$, we obtain that

$$\mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] \leq \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}} \frac{1}{\mathbb{E}\left[\exp\left[-I(\boldsymbol{X}, P, Q)\right]\right]} d\pi_{\mathscr{B}}(Q)\right]^{-1}$$

and (88) follows from (89). □

Lemma 6 *For $P, Q \in \mathscr{M}$, we set*

$$\mathbf{L}(P, Q) = \log \int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c_2 \mathbf{T}(\boldsymbol{X}, P, Q') - c_1 \mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] d\pi(Q').$$

For all $r \in \mathcal{V}(\pi, \overline{Q})$,

$$\mathbb{E}\left[\frac{1}{\int_{\mathscr{M}} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P)}\right] \leq \frac{1}{\pi^2(\mathscr{B})}\left[\int_{\mathscr{B}^2} \exp\left[-\mathbf{L}(P, Q)\right] d\pi_{\mathscr{B}}(P) d\pi_{\mathscr{B}}(Q)\right]^{-1}.$$

Proof For $P, Q \in \mathscr{M}$, we set

$$H(\boldsymbol{X}, P, Q) = \beta c_1 \mathbf{T}(\boldsymbol{X}, P, Q) - \log\left[\int_{\mathscr{M}} \exp\left[c_2 \beta \mathbf{T}(\boldsymbol{X}, P, Q')\right] d\pi(Q')\right].$$

Then,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[-H(\boldsymbol{X}, P, Q)\right]\right] \\
&= \mathbb{E}\left[\exp\left[-\beta c_1 \mathbf{T}(\boldsymbol{X}, P, Q)\right] \int_{\mathscr{M}} \exp\left[c_2 \beta \mathbf{T}(\boldsymbol{X}, P, Q')\right] d\pi(Q')\right] \\
&= \mathbb{E}\left[\int_{\mathscr{M}} \exp\left[\beta\left(c_2 \mathbf{T}(\boldsymbol{X}, P, Q') - c_1 \mathbf{T}(\boldsymbol{X}, P, Q)\right)\right] d\pi(Q')\right] \\
&= \exp\left[\mathbf{L}(P, Q)\right]. \qquad (90)
\end{aligned}$$

It follows from the convexity of the exponential and the fact that $\lambda = c_1 \beta$ that for all $P \in \mathscr{M}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left[\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] &= \mathbb{E}\left[\exp\left[\int_{\mathscr{M}} \left[\beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\widetilde{\pi}_{\boldsymbol{X}}(Q|P)\right]\right] \\
&\leq \mathbb{E}\left[\int_{\mathscr{M}} \exp\left[\beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\widetilde{\pi}_{\boldsymbol{X}}(Q|P)\right] \\
&= \mathbb{E}\left[\frac{\int_{\mathscr{M}} \exp\left[c_2 \beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}{\int_{\mathscr{M}} \exp\left[c_1 \beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}\right]
\end{aligned}$$

$$= \mathbb{E}\left[\frac{1}{\int_{\mathscr{M}} \exp\left[H(\boldsymbol{X}, P, Q)\right] d\pi(Q)}\right].$$

Applying Lemma 4 with $U = \boldsymbol{X}$, $V = Q$ with distribution π, and $h(U, V) = -H(\boldsymbol{X}, P, Q)$ we obtain that

$$\mathbb{E}\left[\exp\left[\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] \leq \left[\int_{\mathscr{M}} \frac{1}{\mathbb{E}\left[\exp\left[-H(\boldsymbol{X}, P, Q)\right]\right]} d\pi(Q)\right]^{-1}.$$

We deduce from (90) that for all $P \in \mathscr{M}$

$$\begin{aligned}\mathbb{E}\left[\exp\left[\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] &\leq \left[\int_{\mathscr{M}} \exp\left[-\mathbf{L}(P, Q)\right] d\pi(Q)\right]^{-1} \\ &\leq \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}} \exp\left[-\mathbf{L}(P, Q)\right] d\pi_{\mathscr{B}}(Q)\right]^{-1}. \end{aligned} \tag{91}$$

Applying Lemma 4 with $U = \boldsymbol{X}$, $V = P$ with distribution π and $h(U, V) = \beta \mathbf{T}(\boldsymbol{X}, P)$, gives

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\int_{\mathscr{M}} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P)}\right] &\leq \left[\int_{\mathscr{M}} \frac{1}{\mathbb{E}\left[\exp\left[\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right]} d\pi(P)\right]^{-1} \\ &\leq \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}} \frac{1}{\mathbb{E}\left[\exp\left[\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right]} d\pi_{\mathscr{B}}(P)\right]^{-1}\end{aligned}$$

which together with (91) leads to the result. □

The proofs of Theorems 1 and 2 rely on suitable bounds on the Laplace transforms of sums of independent random variables and on a summation lemma. These results are presented below.

Lemma 7 *For all $\beta \in \mathbb{R}$ and random variable U with values in an interval of length $l \in (0, +\infty)$,*

$$\log \mathbb{E}\left[\exp\left[\beta U\right]\right] \leq \beta \mathbb{E}\left[U\right] + \frac{\beta^2 l^2}{8}. \tag{92}$$

Lemma 8 *Let U be a squared integrable random variable not larger than $b > 0$. For all $\beta > 0$,*

$$\log \mathbb{E}\left[\exp\left[\beta U\right]\right] \leq \beta \mathbb{E}\left[U\right] + \beta^2 \mathbb{E}\left[U^2\right] \frac{\phi(\beta b)}{2}, \tag{93}$$

where ϕ is defined by (24).

The proofs of Lemmas 7 and 8 can be found on pages 21 and 23 in Massart [23] (where our function ϕ is defined as twice his).

Lemma 9 *Let $J \in \mathbb{N}$, $\gamma > 0$ and $\overline{Q} \in \mathscr{M}$. If r satisfies $n\beta a_1 r \geq 1$ and (12), for all $\gamma_0 > 2\gamma$*

$$\begin{aligned}\int_{\complement\mathscr{B}(\overline{Q},2^J r)} &\exp\left[-\gamma_0 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P) \\ &\leq \pi\left(\mathscr{B}\right) \exp\left[\Xi - (\gamma_0 - 2\gamma)\, n\beta a_1 2^J r\right] \end{aligned} \tag{94}$$

with

$$\Xi = -\gamma + \log\left[\frac{1}{1 - \exp\left[-(\gamma_0 - 2\gamma)\right]}\right]$$

Besides,

$$\int_{\mathscr{M}} \exp\left[-\gamma_0 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P) \leq \pi(\mathscr{B}) \exp\left[\Xi'\right] \tag{95}$$

with

$$\Xi' = \log\left[1 + \frac{\exp\left[-(\gamma_0 - \gamma)\right]}{1 - \exp\left[-(\gamma_0 - 2\gamma)\right]}\right].$$

Proof From (12), we deduce by induction that for all $j \geq 0$

$$\begin{aligned}\pi\left(\mathscr{B}(\overline{Q}, 2^{j+1} r)\right) &\leq \exp\left[\gamma n\beta a_1 r \sum_{k=0}^{j} 2^k\right] \pi\left(\mathscr{B}\right) \\ &= \exp\left[(2^{j+1} - 1)\gamma n\beta a_1 r\right] \pi\left(\mathscr{B}\right)\end{aligned}$$

Consequently,

$$\begin{aligned}&\int_{\complement\mathscr{B}(\overline{Q},2^J r)} \exp\left[-\gamma_0 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P) \\ &= \sum_{j \geq J} \int_{\mathscr{B}(\overline{Q},2^{j+1} r)\setminus\mathscr{B}(\overline{Q},2^j r)} \exp\left[-\gamma_0 \beta n a_1 \ell(\overline{Q}, P)\right] d\pi(P) \\ &\leq \pi\left(\mathscr{B}\right) \sum_{j \geq J} \frac{\pi\left(\mathscr{B}(\overline{Q}, 2^{j+1} r)\right)}{\pi\left(\mathscr{B}\right)} \exp\left[-\gamma_0 n\beta a_1 2^j r\right] \\ &\leq \pi\left(\mathscr{B}\right) \sum_{j \geq J} \exp\left[\gamma n\beta a_1 (2^{j+1} - 1) r - \gamma_0 n\beta a_1 2^j r\right]\end{aligned}$$

$$= \pi(\mathscr{B}) \exp\left[-\gamma n\beta a_1 r\right] \sum_{j\geq J} \exp\left[-(\gamma_0 - 2\gamma) n\beta a_1 2^j r\right]$$
$$= \pi(\mathscr{B}) \exp\left[-\gamma n\beta a_1 r\right] \sum_{j\geq 0} \exp\left[-(\gamma_0 - 2\gamma) n\beta a_1 2^j 2^J r\right].$$

Since $2^j \geq j+1$ for all $j \geq 0$ we obtain that

$$\begin{aligned}
&\int_{\mathscr{C}_{\mathscr{B}}(\overline{Q}, 2^J r)} \exp\left[-\gamma_0 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P) \\
&\leq \pi(\mathscr{B}) \exp\left[-\gamma n\beta a_1 r\right] \sum_{j\geq 0} \exp\left[-(\gamma_0 - 2\gamma) n\beta a_1 (j+1) 2^J r\right] \\
&\leq \pi(\mathscr{B}) \exp\left[-\gamma n\beta a_1 r - (\gamma_0 - 2\gamma) n\beta a_1 2^J r\right] \sum_{j\geq 0} \exp\left[-j(\gamma_0 - 2\gamma) n\beta a_1 2^J r\right] \\
&= \pi(\mathscr{B}) \frac{\exp\left[-\gamma n\beta a_1 r\right]}{1 - \exp\left[-(\gamma_0 - 2\gamma) n\beta a_1 2^J r\right]} \exp\left[-(\gamma_0 - 2\gamma) n\beta a_1 2^J r\right].
\end{aligned}$$

which leads to (94) since $n\beta a_1 2^J r \geq n\beta a_1 r \geq 1$. Finally, by applying this inequality with $J = 0$ we obtain that

$$\begin{aligned}
&\int_{\mathscr{M}} \exp\left[-\gamma_0 \beta n a_1 \ell(\overline{Q}, P)\right] d\pi(P) \\
&= \int_{\mathscr{B}} \exp\left[-\gamma_0 \beta n a_1 \ell(\overline{Q}, P)\right] d\pi(P) + \int_{\mathscr{C}_{\mathscr{B}}} \exp\left[-\gamma_0 \beta n a_1 \ell(\overline{Q}, P)\right] d\pi(P) \\
&\leq \pi(\mathscr{B}) \left[1 + \frac{\exp\left[-\gamma - (\gamma_0 - 2\gamma) n\beta a_1 r\right]}{1 - \exp\left[-(\gamma_0 - 2\gamma)\right]}\right] \\
&\leq \pi(\mathscr{B}) \left[1 + \frac{\exp\left[-(\gamma_0 - \gamma)\right]}{1 - \exp\left[-(\gamma_0 - 2\gamma)\right]}\right],
\end{aligned}$$

which is (95). □

9.3 Proof of Theorem 1

For all $i \in \{1, \ldots, n\}$ and $P, Q, Q' \in \mathscr{M}$, let us set

$$\begin{aligned}
U_i &= c\left(t_{(P,Q')}(X_i) - \mathbb{E}\left[t_{(P,Q')}(X_i)\right]\right) \\
&\quad - c_1\left(t_{(P,Q)}(X_i) - \mathbb{E}\left[t_{(P,Q)}(X_i)\right]\right) \qquad (96) \\
V_i &= c_2\left(t_{(P,Q')}(X_i) - \mathbb{E}\left[t_{(P,Q')}(X_i)\right]\right) \\
&\quad - c_1\left(t_{(P,Q)}(X_i) - \mathbb{E}\left[t_{(P,Q)}(X_i)\right]\right). \qquad (97)
\end{aligned}$$

The random variables U_i are independent and under Assumption 3-(iv), they takes their values in an interval of length $l_1 = c + c_1 = 1 + 2c$. The V_i are also independent

and they takes their values in an interval of length $l_2 = c_1 + c_2 = 3 + 2c$. Applying Lemma 7, we obtain that

$$\prod_{i=1}^{n} \mathbb{E}\left[\exp\left[\beta U_i\right]\right] \leq \exp\left[\frac{l_1^2 n\beta^2}{8}\right] \tag{98}$$

and

$$\prod_{i=1}^{n} \mathbb{E}\left[\exp\left[\beta V_i\right]\right] \leq \exp\left[\frac{l_2^2 n\beta^2}{8}\right]. \tag{99}$$

By using Assumption 2 and the fact that $c_0 = c_1 - ca_0/a_1 > 0$,

$$\begin{aligned}
&c\left(a_0\ell(\overline{P}^\star, P) - a_1\ell(\overline{P}^\star, Q')\right) - c_1\left(a_1\ell(\overline{P}^\star, P) - a_0\ell(\overline{P}^\star, Q)\right)\\
&\quad = -\left(c_1a_1 - ca_0\right)\ell(\overline{P}^\star, P) - ca_1\ell(\overline{P}^\star, Q') + c_1a_0\ell(\overline{P}^\star, Q)\\
&\quad \leq -c_0a_1\left[\tau^{-1}\ell(\overline{Q}, P) - \ell(\overline{P}^\star, \overline{Q})\right] - ca_1\left[\tau^{-1}\ell(\overline{Q}, Q') - \ell(\overline{P}^\star, \overline{Q})\right]\\
&\qquad + \tau c_1a_0\left[\ell(\overline{P}^\star, \overline{Q}) + \ell(\overline{Q}, Q)\right]\\
&\quad = e_0a_1\ell(\overline{P}^\star, \overline{Q}) - \tau^{-1}c_0a_1\ell(\overline{Q}, P) - \tau^{-1}ca_1\ell(\overline{Q}, Q') + \tau c_1a_0\ell(\overline{Q}, Q)
\end{aligned} \tag{100}$$

with

$$e_0 = c_0 + c + \frac{\tau c_1a_0}{a_1}. \tag{101}$$

It follows from (100) and Assumptions 3-(iii), more precisely its consequences (85) and (86), that

$$\begin{aligned}
&n^{-1}\left\{c\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, P, Q')\right] - c_1\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, P, Q)\right]\right\}\\
&\leq c\left[a_0\ell(\overline{P}^\star, P) - a_1\ell(\overline{P}^\star, Q')\right] - c_1\left[a_1\ell(\overline{P}^\star, P) - a_0\ell(\overline{P}^\star, Q)\right]\\
&\leq e_0a_1\ell(\overline{P}^\star, \overline{Q}) - \tau^{-1}c_0a_1\ell(\overline{Q}, P) - \tau^{-1}ca_1\ell(\overline{Q}, Q') + \tau c_1a_0\ell(\overline{Q}, Q).
\end{aligned} \tag{102}$$

Since $a_0 \geq a_1$ and $c_2 > c_1$, $c_0' = c_2(a_0/a_1) - c_1 > 0$ and by arguing as above, we obtain similarly that

$$\begin{aligned}
&n^{-1}\left\{c_2\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, P, Q')\right] - c_1\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, P, Q)\right]\right\}\\
&\quad \leq c_2\left(a_0\ell(\overline{P}^\star, P) - a_1\ell(\overline{P}^\star, Q')\right) - c_1\left(a_1\ell(\overline{P}^\star, P) - a_0\ell(\overline{P}^\star, Q)\right)\\
&\quad = c_0'a_1\ell(\overline{P}^\star, P) - c_2a_1\ell(\overline{P}^\star, Q') + c_1a_0\ell(\overline{P}^\star, Q)\\
&\quad \leq \tau c_0'a_1\left[\ell(\overline{P}^\star, \overline{Q}) + \ell(\overline{Q}, P)\right] - c_2a_1\left[\tau^{-1}\ell(\overline{Q}, Q') - \ell(\overline{P}^\star, \overline{Q})\right]\\
&\qquad + \tau c_1a_0\left[\ell(\overline{P}^\star, \overline{Q}) + \ell(\overline{Q}, Q)\right]
\end{aligned}$$

$$\leq (e_1 + c_2)\, a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_0' a_1 \ell(\overline{Q}, P) - \tau^{-1} c_2 a_1 \ell(\overline{Q}, Q') + \tau c_1 a_0 \ell(\overline{Q}, Q), \tag{103}$$

with

$$e_1 = \tau \left[c_0' + c_1 a_0/a_1\right] = \tau \left[c_2(a_0/a_1) + c_1 (a_0/a_1 - 1)\right]. \tag{104}$$

Using (98) and (102), we deduce that for all $P, Q, Q' \in \mathscr{M}$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[\beta\left(c\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left[\beta\left(c t_{(P,Q')}(X_i) - c_1 t_{(P,Q)}(X_i)\right)\right]\right] \\
&= \exp\left[\beta\left(c\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, P, Q')\right] - c_1\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, P, Q)\right]\right)\right] \prod_{i=1}^{n} \mathbb{E}\left[\exp\left[\beta U_i\right]\right] \\
&\leq \exp\left[n\beta\left[\Delta_1(P, Q) - \tau^{-1} c a_1 \ell(\overline{Q}, Q')\right]\right]
\end{aligned} \tag{105}$$

with

$$\Delta_1(P, Q) = e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 \ell(\overline{Q}, Q) + \frac{l_1^2 \beta}{8} - \tau^{-1} c_0 a_1 \ell(\overline{Q}, P). \tag{106}$$

Using (99) and (103), we obtain similarly that for all $P, Q, Q' \in \mathscr{M}$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[\beta\left(c_2\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] \\
&\quad \leq \exp\left[n\beta\left[\Delta_2(P, Q) - \tau^{-1} c_2 a_1 \ell(\overline{Q}, Q')\right]\right]
\end{aligned} \tag{107}$$

with

$$\begin{aligned}
\Delta_2(P, Q) = (e_1 + c_2)\, a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_0' a_1 \ell(\overline{Q}, P) + \tau c_1 a_0 \ell(\overline{Q}, Q) \\
+ \frac{l_2^2 \beta}{8}.
\end{aligned} \tag{108}$$

Since $2\gamma < \tau^{-1} c < \tau^{-1} c_2$, we may apply Lemma 9 with $\gamma_0 = \tau^{-1} c$ and $\gamma_0 = \tau^{-1} c_2$ successively which leads to

$$\int_{\mathscr{M}} \exp\left[-\tau^{-1} c n \beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q') \leq \pi\left(\mathscr{B}\right) \exp\left[\Xi_1\right] \tag{109}$$

and

$$\int_{\mathscr{M}} \exp\left[-\tau^{-1} c_2 n \beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q') \leq \pi\left(\mathscr{B}\right) \exp\left[\Xi_1\right] \tag{110}$$

with

$$\Xi_1 = \log\left[1 + \frac{\exp\left[-\left(\tau^{-1}c - \gamma\right)\right]}{1 - \exp\left[-\left(\tau^{-1}c - 2\gamma\right)\right]}\right] \geq \log\left[1 + \frac{\exp\left[-\left(\tau^{-1}c_2 - \gamma\right)\right]}{1 - \exp\left[-\left(\tau^{-1}c_2 - 2\gamma\right)\right]}\right]. \tag{111}$$

Putting (107) and (110) together leads to

$$\begin{aligned}\exp\left[\mathbf{L}(P, Q)\right] &= \int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c_2\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] d\pi(Q') \\ &\leq \exp\left[n\beta\Delta_2(P, Q)\right] \int_{\mathscr{M}} \exp\left[-\tau^{-1}c_2 n\beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q') \\ &\leq \pi\left(\mathscr{B}\right) \exp\left[\Xi_1 + n\beta\Delta_2(P, Q)\right],\end{aligned}$$

and since, for all $(P, Q) \in \mathscr{B}^2$, by definition (108) of $\Delta_2(P, Q)$,

$$\begin{aligned}\Delta_2(P, Q) &\leq (e_1 + c_2)\, a_1 \ell(\overline{P}^\star, \overline{Q}) + \left[\tau c_0' a_1 + \tau c_1 a_0\right] r + \frac{l_2^2\beta}{8} \\ &= (e_1 + c_2)\, a_1 \ell(\overline{P}^\star, \overline{Q}) + e_1 a_1 r + \frac{l_2^2\beta}{8} = \Delta_2\end{aligned} \tag{112}$$

we derive that

$$\left[\int_{\mathscr{B}^2} \exp\left[-\mathbf{L}(P, Q)\right] d\pi_{\mathscr{B}}(P) d\pi_{\mathscr{B}}(Q)\right]^{-1} \leq \pi\left(\mathscr{B}\right) \exp\left[\Xi_1 + n\beta\Delta_2\right].$$

We deduce from (84) that

$$\mathbb{P}(^{\mathsf{c}}A) \leq \frac{z}{\pi\left(\mathscr{B}\right)} \exp\left[\Xi_1 + n\beta\Delta_2\right].$$

In particular, $\mathbb{P}(^{\mathsf{c}}A) \leq e^{-\xi}$ for z satisfying

$$\log\left(\frac{1}{z}\right) = \xi + \log\frac{1}{\pi(\mathscr{B})} + \Xi_1 + n\beta\Delta_2. \tag{113}$$

Putting (105) and (109) together, we obtain that

$$\begin{aligned}&\exp\left[\mathbf{M}(P, Q)\right] \\ &= \int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] d\pi(Q') \\ &\leq \exp\left[n\beta\Delta_1(P, Q)\right] \int_{\mathscr{M}} \exp\left[-\tau^{-1}cn\beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q')\end{aligned}$$

$$\leq \pi\left(\mathscr{B}\right) \exp\left[\Xi_1 + n\beta\Delta_1(P, Q)\right].$$

It follows from the definition (106) of $\Delta_1(P, Q)$ that for all $P \in \mathscr{M}$ and for all $Q \in \mathscr{B}$,

$$\Delta_1(P, Q) \leq e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - \tau^{-1} c_0 a_1 \ell(\overline{Q}, P),$$

and consequently, for all $P \in \mathscr{M}$ and $Q \in \mathscr{B}$

$$\exp\left[\mathbf{M}(P, Q)\right] \leq \pi\left(\mathscr{B}\right) \exp\left[\Xi_1 + n\beta\left(e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - \tau^{-1} c_0 a_1 \ell(\overline{Q}, P)\right)\right].$$

We derive from Lemma 5 that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P)\right]\right] \\
&\quad\leq \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}} \exp\left[-\mathbf{M}(P, Q)\right] d\pi_{\mathscr{B}}(Q)\right]^{-1} \\
&\quad\leq \exp\left[\Xi_1 + n\beta\left(e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - \tau^{-1} c_0 a_1 \ell(\overline{Q}, P)\right)\right],
\end{aligned}$$

hence,

$$\begin{aligned}
&\int_{\mathrm{c}_{\mathscr{B}}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P) \\
&\qquad\leq \exp\left[\Xi_1 + n\beta\left(e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8}\right)\right] \\
&\qquad\quad\times \int_{\mathrm{c}_{\mathscr{B}}(\overline{Q}, 2^J r)} \exp\left[-\tau^{-1} c_0 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P). \qquad (114)
\end{aligned}$$

Applying Lemma 9 with $\gamma_0 = \tau^{-1} c_0 > 2\gamma$ and setting $e_2 = \tau^{-1} c_0 - 2\gamma$, we get

$$\int_{\mathrm{c}_{\mathscr{B}}(\overline{Q}, 2^J r)} \exp\left[-\tau^{-1} c_0 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P) \leq \pi\left(\mathscr{B}\right) \exp\left[\Xi_2 - e_2 n\beta a_1 2^J r\right]$$

with

$$\Xi_2 = -\gamma + \log\left[\frac{1}{1 - \exp\left[-e_2\right]}\right], \qquad (115)$$

which together with (114) leads to

$$\log \int_{\mathrm{c}_{\mathscr{B}}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)$$

$$\leq \log\left[\pi\left(\mathscr{B}\right)\right] + \Xi_1 + \Xi_2 + n\beta\left[e_0 a_1 \ell(\overline{P}^\star, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - e_2 a_1 2^J r\right]. \tag{116}$$

Using the definitions (113) of z and (112) of Δ_2 we deduce from (116) that

$$\begin{aligned}
&\log\left[\frac{1}{z}\int_{\mathscr{C}_{\mathscr{B}}(\overline{Q},2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)\right] \\
&\leq \log\left(\frac{1}{z}\right) + \log\left[\pi\left(\mathscr{B}\right)\right] + \Xi_1 + \Xi_2 \\
&\quad + n\beta\left[e_0 a_1 \ell(\overline{P}^\star, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - e_2 a_1 2^J r\right] \\
&= \xi + \log\frac{1}{\pi(\mathscr{B})} + \Xi_1 + n\beta\Delta_2 + \log\left[\pi\left(\mathscr{B}\right)\right] + \Xi_1 + \Xi_2 \\
&\quad + n\beta\left[e_0 a_1 \ell(\overline{P}^\star, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - e_2 a_1 2^J r\right] \\
&= n\beta\left[\left(e_1 + c_2 + e_0\right) a_1 \ell(\overline{P}^\star, \overline{Q}) + e_1 a_1 r + \frac{l_2^2\beta}{8} + + \tau c_1 a_0 r + \frac{l_1^2\beta}{8}\right] \\
&\quad + \xi + 2\Xi_1 + \Xi_2 - e_2 n\beta a_1 2^J r \\
&= n\beta\left[\left(e_0 + e_1 + c_2\right) a_1 \ell(\overline{P}^\star, \overline{Q}) + \left[e_1 + \frac{\tau c_1 a_0}{a_1}\right] a_1 r + \frac{(l_1^2 + l_2^2)\beta}{8}\right] \\
&\quad + \xi + 2\Xi_1 + \Xi_2 - e_2 n\beta a_1 2^J r.
\end{aligned} \tag{117}$$

Setting,

$$C_1 = e_0 + e_1 + c_2 \quad \text{and} \quad C_2 = e_1 + \frac{\tau c_1 a_0}{a_1},$$

we see that the right-hand side of (117) is not larger than $-\xi$, provided that

$$e_2 n\beta a_1 2^J r \geq 2\xi + 2\Xi_1 + \Xi_2 + n\beta\left[C_1 a_1 \ell(\overline{P}^\star, \overline{Q}) + C_2 a_1 r + \frac{(l_1^2 + l_2^2)\beta}{8}\right]$$

or equivalently if

$$2^J \geq \frac{1}{e_2}\left[\frac{2\xi + 2\Xi_1 + \Xi_2}{\beta n a_1 r} + \frac{C_1 \ell(\overline{P}^\star, \overline{Q})}{r} + C_2 + \frac{\left[l_1^2 + l_2^2\right]\beta}{8 a_1 r}\right]. \tag{118}$$

Choosing $\overline{Q}$ in $\mathscr{M}(\beta)$ and using the inequalities $a_1^{-1}\beta \geq r_n(\beta, \overline{Q}) \geq 1/(\beta n a_1)$, for

$$r = \ell(\overline{P}^\star, \overline{Q}) + \frac{1}{a_1}\left(\beta + \frac{2\xi}{n\beta}\right) \geq \frac{1}{\beta n a_1}$$

we obtain that the right-hand side of (118) satisfies

$$\begin{aligned}
&\frac{1}{e_2}\left[\frac{2\xi + 2\Xi_1 + \Xi_2}{\beta n a_1 r} + \frac{C_1\ell(\overline{P}^\star, \overline{Q}) + C_2 r}{r} + \frac{\left[l_1^2 + l_2^2\right]\beta}{8a_1 r}\right] \\
&\quad \leq \frac{1}{e_2}\left[C_2 + 2\Xi_1 + \Xi_2 + \frac{C_3}{r}\left(\ell(\overline{P}^\star, \overline{Q}) + \frac{1}{a_1}\left(\beta + \frac{2\xi}{n\beta}\right)\right)\right] \\
&\quad = \frac{1}{e_2}\left[C_2 + 2\Xi_1 + \Xi_2 + C_3\right]
\end{aligned}$$

with $C_3 = \max\{1, C_1, \left[l_1^2 + l_2^2\right]/8\}$. Inequality (118) is therefore satisfied for $J \in \mathbb{N}$ such that

$$2^J \geq \frac{C_2 + 2\Xi_1 + \Xi_2 + C_3}{e_2} \vee 1 > 2^{J-1},$$

and we may take

$$\kappa_0 = \tau\left[\frac{2\left(C_2 + 2\Xi_1 + \Xi_2 + C_3\right)}{e_2} \vee 1 + 1\right] \geq \tau\left(2^J + 1\right). \tag{119}$$

We recall below, the list of constants depending on a_0, a_1, c, τ and γ and we have used along the proof.

$$\begin{aligned}
&c_0 = 1 + c - \frac{ca_0}{a_1}, && c_1 = 1 + c, && c_2 = 2 + c, \\
&c_0' = \frac{c_2 a_0}{a_1} - c_1, && l_1 = 1 + 2c, && l_2 = 3 + 2c, \\
&e_0 = c_0 + c + \frac{\tau c_1 a_0}{a_1}, && e_1 = \tau\left[c_0' + c_1\frac{a_0}{a_1}\right], && e_2 = \tau^{-1}c_0 - 2\gamma, \\
&C_1 = e_0 + e_1 + c_2, && C_2 = e_1 + \frac{\tau c_1 a_0}{a_1}, && C_3 = \max\left\{1, C_1, \frac{l_1^2 + l_2^2}{8}\right\},
\end{aligned}$$

and

$$\Xi_1 = \log\left[1 + \frac{\exp\left[-\left(\tau^{-1}c - \gamma\right)\right]}{1 - \exp\left[-\left(\tau^{-1}c - 2\gamma\right)\right]}\right], \quad \Xi_2 = -\gamma + \log\left[\frac{1}{1 - \exp\left[-e_2\right]}\right].$$

9.4 Proof of Theorem 2

The proof follows the same lines as that of Theorem 1. Under Assumption 3-(iv), the random variables U_i and V_i defined by (96) and (97) are not larger than with $b = c + c_1 = l_1$ and $b = c_2 + c_1 = l_2$ respectively. Since under Assumption 4, more precisely its consequence (87), that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[U_i^2\right] &\le 2\left[\frac{c^2}{n}\sum_{i=1}^{n}\mathrm{Var}\left[t_{(P,Q')}(X_i)\right]+\frac{c_1^2}{n}\sum_{i=1}^{n}\mathrm{Var}\left[t_{(P,Q)}(X_i)\right]\right]\\ &\le 2a_2\left[(c^2+c_1^2)\ell(\overline{P}^\star,P)+c^2\ell(\overline{P}^\star,Q')+c_1^2\ell(\overline{P}^\star,Q)\right]\end{aligned}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[V_i^2\right] \le 2a_2\left[(c_2^2+c_1^2)\ell(\overline{P}^\star,P)+c_2^2\ell(\overline{P}^\star,Q')+c_1^2\ell(\overline{P}^\star,Q)\right]$$

we may apply Lemma 8 and using the notation $\Lambda_1 = \tau\phi(\beta l_1)$, $\Lambda_2 = \tau\phi(\beta l_2)$ and Assumption 1, we get

$$\begin{aligned}&\frac{1}{n\beta}\log\left[\prod_{i=1}^{n}\mathbb{E}\left[\exp\left[\beta U_i\right]\right]\right]\\ &\quad\le \phi(\beta l_1)\beta a_2\left[(c^2+c_1^2)\ell(\overline{P}^\star,P)+c^2\ell(\overline{P}^\star,Q')+c_1^2\ell(\overline{P}^\star,Q)\right]\\ &\quad\le 2\Lambda_1\beta a_2\left[c^2+c_1^2\right]\ell(\overline{P}^\star,\overline{Q})\\ &\qquad+\Lambda_1\beta a_2\left[(c^2+c_1^2)\ell(\overline{Q},P)+c^2\ell(\overline{Q},Q')+c_1^2\ell(\overline{Q},Q)\right]\end{aligned}\tag{120}$$

and similarly

$$\begin{aligned}&\frac{1}{n\beta}\log\left[\prod_{i=1}^{n}\mathbb{E}\left[\exp\left[\beta V_i\right]\right]\right]\\ &\quad\le 2\Lambda_2\beta a_2\left[c_2^2+c_1^2\right]\ell(\overline{P}^\star,\overline{Q})\\ &\qquad+\Lambda_2\beta a_2\left[(c_2^2+c_1^2)\ell(\overline{Q},P)+c_2^2\ell(\overline{Q},Q')+c_1^2\ell(\overline{Q},Q)\right].\end{aligned}\tag{121}$$

It follows from (102) that

$$\begin{aligned}E_1 = n^{-1}&\left\{c\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q')\right]-c_1\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q)\right]\right\}\\ &+2\Lambda_1\beta a_2\left[c^2+c_1^2\right]\ell(\overline{P}^\star,\overline{Q})\\ &+\Lambda_1\beta a_2\left[(c^2+c_1^2)\ell(\overline{Q},P)+c^2\ell(\overline{Q},Q')+c_1^2\ell(\overline{Q},Q)\right]\end{aligned}$$

$$
\begin{aligned}
&\leq \left[e_0 a_1 + 2\Lambda_1 \beta a_2 \left(c^2 + c_1^2\right)\right] \ell(\overline{P}^\star, \overline{Q}) \\
&\quad - \left[\tau^{-1} c_0 a_1 - \Lambda_1 \beta a_2 (c^2 + c_1^2)\right] \ell(\overline{Q}, P) \\
&\quad - \left[\tau^{-1} c a_1 - \Lambda_1 \beta a_2 c^2\right] \ell(\overline{Q}, Q') \\
&\quad + \left[\tau c_1 a_0 + \Lambda_1 \beta a_2 c_1^2\right] \ell(\overline{Q}, Q).
\end{aligned}
$$

Using the definitions (25) of $\overline{c}_1$ and (26) of $\overline{c}_2$,, that is,

$$
\overline{c}_1 = c_0 - \tau \Lambda_1 \beta a_2 a_1^{-1} (c^2 + c_1^2) \quad \text{and} \quad \overline{c}_2 = c - \tau \Lambda_1 \beta a_2 a_1^{-1} c^2
$$

and setting

$$
\begin{aligned}
e_3 &= e_0 + 2\Lambda_1 \beta \frac{a_2 \left(c^2 + c_1^2\right)}{a_1} \\
e_4 &= \frac{1}{a_1} \left[\tau c_1 a_0 + \Lambda_1 \beta a_2 c_1^2\right]
\end{aligned}
$$

and arguing as in the proof of inequality (105), we deduce from (120) that

$$
\begin{aligned}
&\log \mathbb{E}\left[\exp\left[\beta \left(c\mathbf{T}(\boldsymbol{X}, P, Q') - c_1 \mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] \\
&\quad \leq n\beta E_1 \\
&\quad \leq n\beta a_1 \left[e_3 \ell(\overline{P}^\star, \overline{Q}) - \tau^{-1} \left[\overline{c}_1 \ell(\overline{Q}, P) + \overline{c}_2 \ell(\overline{Q}, Q')\right] + e_4 \ell(\overline{Q}, Q)\right]. \qquad (122)
\end{aligned}
$$

It follows from (103) that

$$
\begin{aligned}
E_2 &= n^{-1} \left\{c_2 \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, P, Q')\right] - c_1 \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, P, Q)\right]\right\} \\
&\quad + 2\Lambda_2 \beta a_2 \left[c_2^2 + c_1^2\right] \ell(\overline{P}^\star, \overline{Q}) \\
&\quad + \Lambda_2 \beta a_2 \left[(c_2^2 + c_1^2) \ell(\overline{Q}, P) + c_2^2 \ell(\overline{Q}, Q') + c_1^2 \ell(\overline{Q}, Q)\right] \\
&\leq (e_1 + c_2)\, a_1 \ell(\overline{P}^\star, \overline{Q}) + \tau c_0' a_1 \ell(\overline{Q}, P) - \tau^{-1} c_2 a_1 \ell(\overline{Q}, Q') \\
&\quad + \tau c_1 a_0 \ell(\overline{Q}, Q) + 2\Lambda_2 \beta a_2 \left[c_2^2 + c_1^2\right] \ell(\overline{P}^\star, \overline{Q}) \\
&\quad + \Lambda_2 \beta a_2 \left[(c_2^2 + c_1^2) \ell(\overline{Q}, P) + c_2^2 \ell(\overline{Q}, Q') + c_1^2 \ell(\overline{Q}, Q)\right] \\
&= \left[(e_1 + c_2)\, a_1 + 2\Lambda_2 \beta a_2 \left(c_2^2 + c_1^2\right)\right] \ell(\overline{P}^\star, \overline{Q}) \\
&\quad + \left[\tau c_0' a_1 + \Lambda_2 \beta a_2 (c_2^2 + c_1^2)\right] \ell(\overline{Q}, P) \\
&\quad - \left[\tau^{-1} c_2 a_1 - \Lambda_2 \beta a_2 c_2^2\right] \ell(\overline{Q}, Q') \\
&\quad + \left[\tau c_1 a_0 + \Lambda_2 \beta a_2 c_1^2\right] \ell(\overline{Q}, Q).
\end{aligned}
$$

Using the definition (27) of $\overline{c}_3$,, that is,

$$\overline{c}_3 = c_2 - \tau\Lambda_2\beta a_2 a_1^{-1} c_2^2,$$

and setting

$$e_5 = e_1 + c_2 + 2\Lambda_2\beta\frac{a_2\left(c_2^2 + c_1^2\right)}{a_1}, \qquad e_6 = \tau c_0' + \Lambda_2\beta\frac{a_2(c_2^2 + c_1^2)}{a_1}$$
$$e_7 = \frac{1}{a_1}\left[\tau c_1 a_0 + \Lambda_2\beta a_2 c_1^2\right],$$

and arguing as in the proof of (107), we deduce from (121) that

$$\log\mathbb{E}\left[\exp\left[\beta\left(c_2\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] \le n\beta E_2$$
$$= n\beta a_1\left(e_5\ell(\overline{P}^\star, \overline{Q}) + e_6\ell(\overline{Q}, P) - \tau^{-1}\overline{c}_3\ell(\overline{Q}, Q') + e_7\ell(\overline{Q}, Q)\right). \quad (123)$$

Under our assumption on β, we know that the quantities $\overline{c}_2$ and $\overline{c}_3$ are positive and that $2\gamma < \tau^{-1}(\overline{c}_2 \wedge \overline{c}_3)$. We may therefore apply Lemma 9 with $\gamma_0 = \tau^{-1}\overline{c}_2$ and $\gamma_0 = \tau^{-1}\overline{c}_3$ successively and get

$$\int_{\mathscr{M}}\exp\left[-\tau^{-1}\overline{c}_2 n\beta a_1\ell(\overline{Q}, Q')\right]d\pi(Q') \le \pi(\mathscr{B})\exp\left[\overline{\Xi}_1\right] \quad (124)$$

and

$$\int_{\mathscr{M}}\exp\left[-\tau^{-1}\overline{c}_3 n\beta a_1\ell(\overline{Q}, Q')\right]d\pi(Q') \le \pi(\mathscr{B})\exp\left[\overline{\Xi}_1\right] \quad (125)$$

with

$$\overline{\Xi}_1 = \log\left[1 + \frac{\exp\left[-\left(\tau^{-1}(\overline{c}_2 \wedge \overline{c}_3) - \gamma\right)\right]}{1 - \exp\left[-\left(\tau^{-1}(\overline{c}_2 \wedge \overline{c}_3) - 2\gamma\right)\right]}\right]. \quad (126)$$

Putting (123) and (125) together, we obtain that for all $(P, Q) \in \mathscr{B}^2$

$$\exp\left[\mathbf{L}(P, Q)\right] = \int_{\mathscr{M}}\mathbb{E}\left[\exp\left[\beta\left(c_2\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right]d\pi(Q')$$
$$\le \exp\left[n\beta a_1\left(e_5\ell(\overline{P}^\star, \overline{Q}) + e_6\ell(\overline{Q}, P) + e_7\ell(\overline{Q}, Q)\right)\right]$$
$$\times\int_{\mathscr{M}}\exp\left[-\tau^{-1}\overline{c}_3 n\beta a_1\ell(\overline{Q}, Q')\right]d\pi(Q')$$
$$\le \pi(\mathscr{B})\exp\left[\overline{\Xi}_1 + n\beta a_1\left(e_5\ell(\overline{P}^\star, \overline{Q}) + e_6\ell(\overline{Q}, P) + e_7\ell(\overline{Q}, Q)\right)\right]$$
$$\le \pi(\mathscr{B})\exp\left[\overline{\Xi}_1 + n\beta a_1\left(e_5\ell(\overline{P}^\star, \overline{Q}) + (e_6 + e_7)r\right)\right].$$

Consequently,

$$\left[\int_{\mathscr{B}^2} \exp\left[-\mathbf{L}(P,\,Q)\right] d\pi_{\mathscr{B}}(P) d\pi_{\mathscr{B}}(Q)\right]^{-1} \leq \pi\left(\mathscr{B}\right) \exp\left[\overline{\Xi}_1 + n\beta a_1 \left(e_5 \ell(\overline{P}^{\star}, \overline{Q}) + (e_6 + e_7) r\right)\right].$$

We deduce from (84) that

$$\mathbb{P}({}^{\mathsf{c}}A) \leq \frac{z}{\pi\left(\mathscr{B}\right)} \exp\left[\overline{\Xi}_1 + n\beta a_1 \left(e_5 \ell(\overline{P}^{\star}, \overline{Q}) + (e_6 + e_7) r\right)\right].$$

In particular, $\mathbb{P}({}^{\mathsf{c}}A) \leq e^{-\xi}$ for z satisfying

$$\log\left(\frac{1}{z}\right) = \xi + \log\frac{1}{\pi(\mathscr{B})} + \overline{\Xi}_1 + n\beta a_1 \left[e_5 \ell(\overline{P}^{\star}, \overline{Q}) + (e_6 + e_7) r\right]. \tag{127}$$

Putting (122) and (124) together, we obtain that for all $Q \in \mathscr{B}$

$$\begin{aligned}
&\exp\left[\mathbf{M}(P,\,Q)\right] \\
&= \int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c\mathbf{T}(\boldsymbol{X},\,P,\,Q') - c_1\mathbf{T}(\boldsymbol{X},\,P,\,Q)\right)\right]\right] d\pi(Q') \\
&\leq \exp\left[n\beta a_1 \left(e_3 \ell(\overline{P}^{\star}, \overline{Q}) - \tau^{-1}\overline{c}_1 \ell(\overline{Q}, P) + e_4 \ell(\overline{Q}, Q)\right)\right] \\
&\quad \times \int_{\mathscr{M}} \exp\left[-\tau^{-1}\overline{c}_2 n\beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q') \\
&\leq \pi(\mathscr{B}) \exp\left[\overline{\Xi}_1 + n\beta a_1 \left(e_3 \ell(\overline{P}^{\star}, \overline{Q}) + e_4 r - \tau^{-1}\overline{c}_1 \ell(\overline{Q}, P)\right)\right].
\end{aligned}$$

We derive from Lemma 5 that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X},\,P)\right]\right] \\
&\leq \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}} \exp\left[-\mathbf{M}(P,\,Q)\right] d\pi_{\mathscr{B}}(Q)\right]^{-1} \\
&\leq \exp\left[\overline{\Xi}_1 + n\beta a_1 \left(e_3 \ell(\overline{P}^{\star}, \overline{Q}) + e_4 r - \tau^{-1}\overline{c}_1 \ell(\overline{Q}, P)\right)\right],
\end{aligned}$$

and consequently,

$$\begin{aligned}
&\int_{{}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X},\,P)\right]\right] d\pi(P) \\
&\leq \exp\left[\overline{\Xi}_1 + n\beta a_1 \left(e_3 \ell(\overline{P}^{\star}, \overline{Q}) + e_4 r\right)\right] \\
&\quad \times \int_{{}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)} \exp\left[-\tau^{-1}\overline{c}_1 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P).
\end{aligned} \tag{128}$$

Since under our assumptions, $\overline{c}_1 > 0$ and $2\gamma < \tau^{-1}\overline{c}_1$ we may apply Lemma 9 with $\gamma_0 = \tau^{-1}\overline{c}_1$, and setting $e_8 = \tau^{-1}\overline{c}_1 - 2\gamma$ which leads to

$$\int_{\mathsf{c}_{\mathscr{B}}(\overline{Q},2^J r)} \exp\left[-\tau^{-1}\overline{c}_1 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P) \le \pi(\mathscr{B}) \exp\left[\overline{\Xi}_2 - e_8 n\beta a_1 2^J r\right].$$

with

$$\overline{\Xi}_2 = -\gamma + \log\left[\frac{1}{1-\exp[-e_8]}\right], \tag{129}$$

which together with (128) leads to

$$\begin{aligned}&\int_{\mathsf{c}_{\mathscr{B}}(\overline{Q},2^J r)} \mathbb{E}\left[\exp[-\beta \mathbf{T}(\boldsymbol{X}, P)]\right] d\pi(P)\\ &\quad\le \pi(\mathscr{B}) \exp\left[\overline{\Xi}_1 + \overline{\Xi}_2 + n\beta a_1 \left(e_3 \ell(\overline{P}^\star, \overline{Q}) + e_4 r - e_8 2^J r\right)\right].\end{aligned} \tag{130}$$

Using the definition (127) of z, we deduce that

$$\begin{aligned}&\log\left[\frac{1}{z}\int_{\mathsf{c}_{\mathscr{B}}(\overline{Q},2^J r)} \mathbb{E}\left[\exp[-\beta \mathbf{T}(\boldsymbol{X}, P)]\right] d\pi(P)\right]\\ &\le \log\left(\frac{1}{z}\right) + \log \pi(\mathscr{B}) + \overline{\Xi}_1 + \overline{\Xi}_2 + n\beta a_1 \left(e_3 \ell(\overline{P}^\star, \overline{Q}) + e_4 r - e_8 2^J r\right)\\ &= \xi + \log\frac{1}{\pi(\mathscr{B})} + \overline{\Xi}_1 + n\beta a_1 \left[e_5 \ell(\overline{P}^\star, \overline{Q}) + (e_6 + e_7) r\right]\\ &\quad + \log \pi(\mathscr{B}) + \overline{\Xi}_1 + \overline{\Xi}_2 + n\beta a_1 \left(e_3 \ell(\overline{P}^\star, \overline{Q}) + e_4 r - e_8 2^J r\right)\\ &= \xi + 2\overline{\Xi}_1 + \overline{\Xi}_2 + n\beta a_1 \left[(e_3 + e_5)\, \ell(\overline{P}^\star, \overline{Q}) + (e_4 + e_6 + e_7) r\right]\\ &\quad - e_8 n\beta a_1 2^J r.\end{aligned}$$

The right-hand side is not larger than $-\xi$ provided that

$$2^J \ge \frac{1}{e_8}\left[\frac{2\xi + 2\overline{\Xi}_1 + \overline{\Xi}_2}{n\beta a_1 r} + \left[(e_3 + e_5)\frac{\ell(\overline{P}^\star, \overline{Q})}{r} + e_4 + e_6 + e_7\right]\right]. \tag{131}$$

Using the fact that $r_n(\beta, \overline{Q}) \ge 1/(n\beta a_1)$, with the choice

$$r = \ell(\overline{P}^\star, \overline{Q}) + r_n(\beta, \overline{Q}) + \frac{2\xi}{n\beta a_1} \ge \ell(\overline{P}^\star, \overline{Q}) + \frac{1 + 2\xi}{n\beta a_1} \ge \frac{1}{n\beta a_1},$$

the right-hand side of (131) satisfies

$$\frac{1}{e_8}\left[\frac{2\xi + 2\overline{\Xi}_1 + \overline{\Xi}_2}{n\beta a_1 r} + \left[(e_3 + e_5)\frac{\ell(\overline{P}^\star, \overline{Q})}{r} + e_4 + e_6 + e_7\right]\right]$$

$$\leq \frac{1}{e_8}\left[2\overline{\Xi}_1 + \overline{\Xi}_2 + e_4 + e_6 + e_7 + \frac{(e_3+e_5)\vee 1}{r}\left(\ell(\overline{P}^\star, \overline{Q}) + \frac{2\xi}{n\beta a_1}\right)\right]$$
$$\leq \frac{2\overline{\Xi}_1 + \overline{\Xi}_2 + e_4 + e_6 + e_7 + (e_3+e_5)\vee 1}{e_8}.$$

Inequality (131) holds for $J \in \mathbb{N}$ such that

$$2^J \geq \frac{2\overline{\Xi}_1 + \overline{\Xi}_2 + e_4 + e_6 + e_7 + (e_3+e_5)\vee 1}{e_8} \vee 1 > 2^{J-1},$$

and we may take

$$\kappa_0 = \tau\left[\frac{2\left[2\overline{\Xi}_1 + \overline{\Xi}_2 + e_4 + e_6 + e_7 + (e_3+e_5)\vee 1\right]}{e_8} \vee 1 + 1\right]$$
$$\geq \tau\left(2^J + 1\right). \tag{132}$$

In complements to constants listed at the end of the proof of Theorem 1, we recall that

$$\Lambda_1 = \tau\phi(\beta l_1), \quad \Lambda_2 = \tau\phi(\beta l_2)$$

$$\overline{c}_1 = c_0 - \tau\Lambda_1\beta\frac{a_2(c^2+c_1^2)}{a_1}, \quad \overline{c}_2 = c - \tau\Lambda_1\beta\frac{a_2c^2}{a_1}, \quad \overline{c}_3 = c_2 - \tau\Lambda_2\beta\frac{a_2c_2^2}{a_1},$$

$$e_3 = e_0 + 2\Lambda_1\beta\frac{a_2\left(c^2+c_1^2\right)}{a_1}, \qquad e_4 = \frac{1}{a_1}\left[\tau c_1 a_0 + \Lambda_1\beta a_2 c_1^2\right],$$
$$e_5 = e_1 + c_2 + 2\Lambda_2\beta\frac{a_2\left(c_2^2+c_1^2\right)}{a_1}, \qquad e_6 = \tau c_0' + \Lambda_2\beta\frac{a_2(c_2^2+c_1^2)}{a_1},$$
$$e_7 = \frac{1}{a_1}\left[\tau c_1 a_0 + \Lambda_2\beta a_2 c_1^2\right], \qquad e_8 = \tau^{-1}\overline{c}_1 - 2\gamma,$$

and

$$\overline{\Xi}_1 = \log\left[1 + \frac{\exp\left[-\left(\tau^{-1}(\overline{c}_2\wedge\overline{c}_3) - \gamma\right)\right]}{1-\exp\left[-\left(\tau^{-1}(\overline{c}_2\wedge\overline{c}_3) - 2\gamma\right)\right]}\right],$$
$$\overline{\Xi}_2 = -\gamma + \log\left[\frac{1}{1-\exp\left[-e_8\right]}\right].$$

9.5 Proof of Theorem 3

Let us take $r \geq \varepsilon$ and set $\varpi = 2\xi + 1$ so that

$$\pi\left({}^{\mathsf{c}}\mathscr{B}\right) \leq \pi\left({}^{\mathsf{c}}\mathscr{B}(\overline{Q}, \varepsilon)\right) \leq e^{-\varpi}\pi\left(\mathscr{B}(\overline{Q}, \varepsilon)\right) \leq e^{-\varpi}\pi\left(\mathscr{B}\right).$$

In order to prove the first part, let us go back to the proof of Theorem 1. Clearly,

$$\int_{\mathscr{M}} \exp\left[-\tau^{-1} cn\beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q') \leq 1 = \pi\left(\mathscr{B}\right) + \pi\left({}^{\mathsf{c}}\mathscr{B}\right) \leq \pi\left(\mathscr{B}\right)\left(1 + e^{-\varpi}\right)$$

and similarly,

$$\int_{\mathscr{M}} \exp\left[-\tau^{-1} c_2 n\beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q') \leq \pi\left(\mathscr{B}\right)\left(1 + e^{-\varpi}\right).$$

Inequalities (109) and (110) are therefore satisfied with $\Xi_1 = \log(1+e^{-1})$. Moreover,

$$\begin{aligned}&\int_{{}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)} \exp\left[-\tau^{-1} c_0 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P)\\ &\quad\leq \exp\left[-\tau^{-1} c_0 n\beta a_1 2^J r\right] \pi\left({}^{\mathsf{c}}\mathscr{B}\right) \leq \pi\left(\mathscr{B}\right) \exp\left[-\varpi - \tau^{-1} c_0 n\beta a_1 2^J r\right].\end{aligned}$$

We deduce from (114) that

$$\begin{aligned}&\int_{{}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)\\ &\quad\leq \exp\left[\Xi_1 + n\beta\left(e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2 \beta}{8}\right)\right]\\ &\qquad\times \pi\left(\mathscr{B}\right) \exp\left[-\varpi - \tau^{-1} c_0 n\beta a_1 2^J r\right],\end{aligned}$$

and consequently,

$$\begin{aligned}&\log \int_{{}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)\\ &\quad\leq \log \pi\left(\mathscr{B}\right) + \Xi_1 - \varpi\\ &\qquad + n\beta\left[e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2 \beta}{8} - \tau^{-1} c_0 a_1 2^J r\right].\end{aligned}$$

Using the definitions (113) of z and (112) of Δ_2, we deduce that

$$\log\left[\frac{1}{z}\int_{{}^{\mathsf{c}}\mathscr{B}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)\right]$$

$$
\begin{aligned}
&\leq \xi + \log\frac{1}{\pi(\mathscr{B})} + \Xi_1 + n\beta\Delta_2 + \log\pi\left(\mathscr{B}\right) + \Xi_1 - \varpi \\
&\quad + n\beta\left[e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - \tau^{-1} c_0 a_1 2^J r\right] \\
&= \xi + 2\Xi_1 + n\beta\left[\left(e_1 + c_2\right) a_1 \ell(\overline{P}^{\star}, \overline{Q}) + e_1 a_1 r + \frac{l_2^2\beta}{8}\right] - \varpi \\
&\quad + n\beta\left[e_0 a_1 \ell(\overline{P}^{\star}, \overline{Q}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - \tau^{-1} c_0 a_1 2^J r\right] \\
&= \xi + 2\Xi_1 - \varpi \\
&\quad + n\beta a_1\left[C_1 \ell(\overline{P}^{\star}, \overline{Q}) + C_2 r + \frac{(l_1^2 + l_2^2)\beta}{8a_1} - \tau^{-1} c_0 2^J r\right],
\end{aligned}
$$

where the constants C_1 and C_2 are the same as those defined in the proof of Theorem 1. If we choose $r = \ell(\overline{P}^{\star}, \overline{Q}) \vee (\beta/a_1) \vee \varepsilon$ and J such that $\tau^{-1} c_0 2^J \geq C_1 + C_2 + (l_1^2 + l_2^2)/8$, we obtain that

$$
\log\left[\frac{1}{z}\int_{\mathscr{C}_{\mathscr{B}}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)\right] \leq \xi + 2\Xi_1 - \varpi \leq -\xi
$$

since $\varpi = 2\xi + 1 \geq 2(\xi + \Xi_1)$. We conclude as in the proof of Theorem 1.

In order to prove the second part of Theorem 3, we go back to the proof of Theorem 2. The arguments are similar. As before,

$$
\int_{\mathscr{M}} \exp\left[-\tau^{-1}\overline{c}_2 n\beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q') \leq \pi\left(\mathscr{B}\right)\left(1 + e^{-\varpi}\right)
$$

and

$$
\int_{\mathscr{M}} \exp\left[-\tau^{-1}\overline{c}_3 n\beta a_1 \ell(\overline{Q}, Q')\right] d\pi(Q') \leq \pi\left(\mathscr{B}\right)\left(1 + e^{-\varpi}\right).
$$

Inequalities (124) and (125) are therefore both satisfied with $\overline{\Xi}_1 = \log(1 + e^{-1})$. Moreover

$$
\begin{aligned}
&\int_{\mathscr{C}_{\mathscr{B}}(\overline{Q}, 2^J r)} \exp\left[-\tau^{-1}\overline{c}_1 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P) \\
&\quad \leq \pi\left(\mathscr{B}\right) \exp\left[-\varpi - \tau^{-1}\overline{c}_1 n\beta a_1 2^J r\right],
\end{aligned}
$$

and we deduce from (128) that

$$
\int_{\mathscr{C}_{\mathscr{B}}(\overline{Q}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)
$$

$$
\begin{aligned}
&\leq \exp\left[\overline{\Xi}_1 + n\beta a_1\left(e_3\ell(\overline{P}^\star, \overline{Q}) + e_4 r\right)\right] \\
&\quad \times \int_{\mathsf{c}_{\mathscr{B}(\overline{Q}, 2^J r)}} \exp\left[-\tau^{-1}\overline{c}_1 n\beta a_1 \ell(\overline{Q}, P)\right] d\pi(P) \\
&\leq \pi\left(\mathscr{B}\right) \exp\left[\overline{\Xi}_1 + n\beta a_1\left[e_3\ell(\overline{P}^\star, \overline{Q}) + e_4 r - \tau^{-1}\overline{c}_1 2^J r\right] - \varpi\right].
\end{aligned}
$$

Using the definition (127) of z, we deduce that

$$
\begin{aligned}
&\log\left[\frac{1}{z}\int_{\mathsf{c}_{\mathscr{B}(\overline{Q}, 2^J r)}} \mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)\right] \\
&\leq \xi + \log\frac{1}{\pi(\mathscr{B})} + \overline{\Xi}_1 + n\beta a_1\left[e_5\ell(\overline{P}^\star, \overline{Q}) + (e_6 + e_7) r\right] \\
&\quad + \log\pi\left(\mathscr{B}\right) + \overline{\Xi}_1 + n\beta a_1\left[e_3\ell(\overline{P}^\star, \overline{Q}) + e_4 r - \tau^{-1}\overline{c}_1 2^J r\right] - \varpi \\
&= \xi + 2\overline{\Xi}_1 - \varpi \\
&\quad + n\beta a_1\left[(e_3 + e_5)\ell(\overline{P}^\star, \overline{Q}) + (e_4 + e_6 + e_7) r - \tau^{-1}\overline{c}_1 2^J r\right].
\end{aligned}
$$

Taking $r = \ell(\overline{P}^\star, \overline{Q}) \vee \varepsilon \geq \varepsilon$ and $J \geq 0$ such that

$$
\tau^{-1}\overline{c}_1 2^J \geq e_3 + e_5 + e_4 + e_6 + e_7
$$

we obtain that

$$
\log\left[\frac{1}{z}\int_{\mathsf{c}_{\mathscr{B}(\overline{Q}, 2^J r)}} \mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P)\right] \leq -\xi
$$

and we conclude as before.

10 Other proofs

10.1 Proof of Lemma 1

Let Y be a random variable with gamma distribution $\gamma(s, 1)$. Since $\sigma Y \sim \gamma(s, \sigma)$, it is sufficient to prove the result for $\sigma = 1$. Using the inequality $\log(1-x) \geq -x/(1-x)$ which holds for all $x \in [0, 1)$, we obtain that

$$
\log\mathbb{E}\left[e^{\beta(Y-s)}\right] = -s\left[\log(1-\beta) + \beta\right] \leq \frac{s\beta^2}{1-\beta} \quad \text{for all } \beta \in [0, 1).
$$

Applying Lemma 8.2 in Birgé [10] with $a = \sqrt{s}$ and $b = 1$, we obtain that

$$
\mathbb{P}\left[Y \geq s + 2\sqrt{s\xi} + \xi\right] \leq e^{-\xi} \quad \text{for all } \xi \geq 0
$$

which proves (44). Let us now turn to the lower bound. For $x \geq 0$, let us set

$$g(x) = x - \log(1+x) \leq \left(\frac{x^2}{2}\right) \wedge x.$$

For all $t, u \geq 0$,

$$\begin{aligned}\int_{t+u}^{+\infty} x^t e^{-x} dx &= \int_u^{+\infty} (t+y)^t e^{-t-y} dy = t^t e^{-t} \int_u^{+\infty} e^{-tg(y/t)} dy \\ &\geq t^t e^{-t} \left(\int_u^{+\infty} e^{-y^2/(2t)} dy \vee \int_u^{+\infty} e^{-y} dy \right) \\ &= t^t e^{-t} \left[\left(\sqrt{2\pi t}\, \overline{F}\left(\frac{u}{\sqrt{t}}\right)\right) \vee e^{-u} \right],\end{aligned}$$

where $\overline{F}(z) = \mathbb{P}[\mathcal{N}(0,1) \geq z]$ for all $z \in \mathbb{R}$. Using the the following inequalities

$$t^{t-1/2} e^{-t} \sqrt{2\pi} \leq \Gamma(t) \leq t^{t-1/2} e^{-t} \sqrt{2\pi} \exp[1/(12t)], \tag{133}$$

that can be found in Whittaker and Watson [25, p. 253], with $t = s - 1 > 0$, we deduce that

$$\begin{aligned}\mathbb{P}[Y \geq t+u] &= \frac{1}{\Gamma(t+1)} \int_{t+u}^{+\infty} x^t e^{-x} dx = \frac{1}{t\Gamma(t)} \int_{t+u}^{+\infty} x^t e^{-x} dx \\ &\geq \left[\overline{F}\left(\frac{u}{\sqrt{t}}\right) e^{-1/(12t)} \right] \vee \left[\frac{e^{-u-1/(12t)}}{\sqrt{2\pi t}} \right].\end{aligned}$$

Using the fact that $\overline{F}(\overline{\Phi}^{-1}(z)) = e^{-z}$ for all $z \geq 0$, we obtain that for the choice

$$u = \left[\sqrt{t}\, \overline{\Phi}^{-1}\left(\xi - \frac{1}{12t}\right)\right] \vee \log\left(\frac{e^{\xi - 1/(12t)}}{\sqrt{2\pi t}}\right),$$

which is nonnegative for $\xi \geq \log 2 + 1/(12t)$, the quantity $\mathbb{P}[Y \geq t+u]$ is at least $e^{-\xi}$, which proves (45).

10.2 Proof of Theorem 4

Throughout this proof, $a_0 = 2$, $a_1 = 3/16$, $\beta = 2\gamma = 1/500$ and κ denotes a positive numerical constant that may vary from line to line. It follows from Corollary 4 that for n large enough, $r_n(\beta, P_{\theta^\star}) \leq r_n^\star = \kappa k/n$. Applying our Corollary 2 with $\ell = h^2$ (and 2ξ in place of ξ), we obtain that for n large enough, with a probability at least $1 - 2e^{-\xi}$,

$$1 - e^{-\xi} \leq \widehat{\nu}_{\boldsymbol{X}}^h \left(\left\{ \boldsymbol{\theta} \in \Theta,\ h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) \leq r_n(\xi) \right\} \right) \text{ with } r_n(\xi) = \frac{\kappa(k+\xi)}{n}.$$

We know by Proposition 9 that under the assumptions of Corollary 4, Assumption 9-(i) is satisfied with $s = 2$, $|\cdot|_*$ given by (81) and $\varepsilon = 1/2$. This implies that for n large

$$\left\{\boldsymbol{\theta} \in \Theta,\ h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) \le r_n(\xi)\right\} \subset \left\{\boldsymbol{\theta} \in \Theta,\ \left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_*^2 \le 2r_n(\xi)\right\},$$

which leads to (46).

10.3 Proof of Proposition 4

Let us denote by F_σ the distribution function of ν_σ. Throughout this proof, we fix some $\theta^\star \in [-\sigma t, \sigma t]$. Our aim is to prove that $P_{\theta^\star}$ belongs to $\mathscr{M}(\overline{\beta})$.

Since the total variation distance is translation invariant, $\|P_\theta - P_{\theta^\star}\| = \|P_{\theta-\theta^\star} - P_0\| = \|P_{\theta^\star-\theta} - P_0\|$ and consequently, for all $r \in [0, 1)$,

$$\{\theta \in \Theta,\ \|P_\theta - P_{\theta^\star}\| \le r\} = \left\{\theta \in \Theta,\ \left|\theta^\star - \theta\right| \le \varphi(r)\right\} \quad \text{for all } r \in [0, 1)$$

while for $r \ge 1$, $\{\theta \in \Theta,\ \|P_\theta - P_{\theta^\star}\| \le r\} = \Theta = \mathbb{R}$.

We set $r_0 = \sup\{r > 0,\ \varphi(r) \le \sigma t\}$ and distinguish between two cases.

Case 1 Assume $r_0 \le 1/4$. For all $r < r_0, \varphi(r) < \sigma t, 2r < 1$, and since q is symmetric, positive and decreasing on $\mathbb{R}_+$,

$$\begin{aligned}
\frac{\pi(\mathscr{B}(P_{\theta^\star}, 2r))}{\pi(\mathscr{B}(P_{\theta^\star}, r))} &= \frac{\nu_\sigma\left(\{\theta \in \mathbb{R},\ \|P_\theta - P_{\theta^\star}\| \le 2r\}\right)}{\nu_\sigma\left(\{\theta \in \mathbb{R},\ \|P_\theta - P_{\theta^\star}\| \le r\}\right)} \\
&= \frac{\nu_\sigma\left(\{\theta \in \mathbb{R},\ |\theta - \theta^\star| \le \varphi(2r)\}\right)}{\nu_\sigma\left(\{\theta \in \mathbb{R},\ |\theta - \theta^\star| \le \varphi(r)\}\right)} \le \frac{2q_\sigma(0)\varphi(2r)}{2q_\sigma(|\theta^\star| + \varphi(r))\varphi(r)} \\
&\le \frac{q_\sigma(0)\varphi(2r)}{q_\sigma(|\theta^\star| + \sigma t)\varphi(r)} \le \frac{q_\sigma(0)\varphi(2r)}{q_\sigma(2\sigma t)\varphi(r)} \\
&= \frac{q(0)\varphi(2r)}{q(2t)\varphi(r)} \le \frac{\overline{\Gamma}}{q(2t)}.
\end{aligned}$$

For all $r_0 < r < 1$, $|\theta^\star| \le \sigma t < \varphi(r)$, hence $F_\sigma(|\theta^\star| - \varphi(r)) \le F_\sigma(0) = 1/2$ and $F_\sigma(|\theta^\star| + \varphi(r)) \ge F_\sigma(\varphi(r)) \ge F_\sigma(\sigma t) = F_1(t) \ge 3/4$ under our assumption on t. Consequently,

$$\begin{aligned}
\frac{\pi(\mathscr{B}(P_{\theta^\star}, 2r))}{\pi(\mathscr{B}(P_{\theta^\star}, r))} &\le \frac{1}{\nu_\sigma\left(\{\theta \in \mathbb{R},\ |\theta - \theta^\star| \le \varphi(r)\}\right)} \\
&= \frac{1}{F_\sigma\left(|\theta^\star| + \varphi(r)\right) - F(|\theta^\star| - \varphi(r))} \\
&\le \frac{1}{3/4 - 1/2} = 4.
\end{aligned}$$

Note that the result also holds for $r = r_0$ by letting r decrease to r_0.

Case 2 Assume that $r_0 > 1/4$. Then $\varphi(1/4) \leq \sigma t$ and arguing as before, we obtain that for all $r \leq 1/4 < r_0$,

$$\frac{\pi(\mathscr{B}(P_{\theta^\star}, 2r))}{\pi(\mathscr{B}(P_{\theta^\star}, r))} \leq \frac{2q_\sigma(0)\varphi(2r)}{2q_\sigma(|\theta^\star| + \varphi(r))\varphi(r)} = \frac{q_\sigma(0)\varphi(2r)}{q_\sigma(|\theta^\star| + \varphi(1/4))\varphi(r)}$$
$$\leq \frac{q_\sigma(0)\varphi(2r)}{q_\sigma(2\sigma t)\varphi(r)} \leq \frac{\overline{\Gamma}}{q(2t)}.$$

For all $r \in (1/4, 1)$, $\varphi(r) \geq \varphi(1/4)$ and

$$\frac{\pi(\mathscr{B}(P_{\theta^\star}, 2r))}{\pi(\mathscr{B}(P_{\theta^\star}, r))} \leq \frac{1}{\nu_\sigma(\{\theta \in \mathbb{R},\ |\theta - \theta^\star| \leq \varphi(r)\})}$$
$$\leq \frac{1}{\nu_\sigma(\{\theta \in \mathbb{R},\ |\theta - \theta^\star| \leq \varphi(1/4)\})}$$
$$\leq \frac{1}{2q_\sigma(|\theta^\star| + \varphi(1/4))\varphi(1/4)}$$
$$\leq \frac{1}{2q_\sigma(2\sigma t)\varphi(1/4)} \leq \frac{\overline{\Gamma}\sigma}{q(2t)}.$$

We obtain that in any case, for all $r \in (0, 1)$ and $\theta^\star \in [-\sigma t, \sigma t]$,

$$\log\left(\frac{\pi(\mathscr{B}(P_{\theta^\star}, 2r))}{\pi(\mathscr{B}(P_{\theta^\star}, r))}\right) \leq \max\left\{\log\left(\frac{\overline{\Gamma}(\sigma \vee 1)}{q(2t)}\right), \log 4\right\}. \tag{134}$$

The inequality is also clearly true for $r \geq 1$ since then $\pi(\mathscr{B}(P_{\theta^\star}, 2r)) = \pi(\mathscr{B}(P_{\theta^\star}, r)) = 1$. Hence, for all $r \geq a_1^{-1}\beta$

$$\frac{1}{n\gamma a_1 r}\log\left(\frac{\pi(\mathscr{B}(P_{\theta^\star}, 2r))}{\pi(\mathscr{B}(P_{\theta^\star}, r))}\right) \leq \frac{1}{n\gamma\beta}\sup_{r>0}\log\left(\frac{\pi(\mathscr{B}(P_{\theta^\star}, 2r))}{\pi(\mathscr{B}(P_{\theta^\star}, r))}\right)$$
$$\leq \frac{1}{n\gamma\beta}\max\left\{\log\left(\frac{\overline{\Gamma}(\sigma \vee 1)}{q(2t)}\right), \log 4\right\}.$$

The right-hand side is not larger than β provided that it satisfies (50) and this lower bound is not smaller than $1/\sqrt{n}$ since $\gamma \leq 1$. We conclude by using (15).

10.4 Proof of Proposition 5

Under our assumption on q, Assumption 6 is satisfied and

$$\overline{\Gamma} = 2^{1/s}\max\left\{q(0), 2^{(1/s)-1}\right\}.$$

Let $t = (|\theta|/\sigma) \vee t_0$. Then, $\theta \in [-\sigma t, \sigma t]$, $\nu_1([t, +\infty)) \leq 1/4$ and inequality (134) holds true. We deduce from (11) that

$$r_n(\beta, P_\theta) \leq \frac{1}{\gamma n a_1 \beta} \max\left\{\log\left(\frac{\overline{\Gamma}(\sigma \vee 1)}{q(2t)}\right), \log 4\right\}$$

and the result follows from our specific choices of a_1, γ and β.

10.5 Proof of Corollary 3

We set for short $\Theta = \Theta[\eta, \delta]$ with the parameters η and δ defined by (60) and (61) respectively and also define

$$J_n = \exp\left[\frac{(K^2-1)\gamma\tau^4 a_1^2 n \eta_n^2}{2(k+1)}\right] \tag{135}$$

so that $\mathscr{M}_n(K)$ contains the elements $P = P_{(p,\mathbf{m},\sigma)}$ of $\mathscr{M}$ such that

$$|\log\sigma| \vee \left|\frac{\mathbf{m}}{\sigma}\right|_\infty \leq \log(1+\delta) J_n.$$

Hereafter we fix $P = P_{(p,\mathbf{m},\sigma)} \in \mathscr{M}_n(K)$. There exist $\theta = \theta(P) = (\overline{Q}, \overline{\mathbf{m}}, \overline{\sigma}) \in \Theta$ with $\overline{\sigma} = (1+\delta)^{j_0}$, $\overline{\mathbf{m}} = \overline{\sigma}\delta\mathbf{j}$, $(j_0, \mathbf{j}) \in \mathbb{Z} \times \mathbb{Z}^k$ such that

$$\frac{\overline{\sigma}}{(1+\delta)} \leq \sigma < \overline{\sigma} \quad \text{and} \quad \overline{m}_i = j_i \overline{\sigma}\delta \leq m_i < \overline{m}_i + \overline{\sigma}\delta, \tag{136}$$

for all $i \in \{1, \ldots, k\}$. Consequently,

$$0 \leq \left(1 - \frac{\sigma}{\overline{\sigma}}\right) \leq \frac{\delta}{1+\delta} < \delta \quad \text{and} \quad \left|\frac{\mathbf{m} - \overline{\mathbf{m}}}{\overline{\sigma}}\right|_\infty \leq \delta, \tag{137}$$

and we infer from (56) and (57) and the fact that the total variation loss is translation and scale invariant that P_θ satisfies

$$\begin{aligned}
\ell\left(P_{(p,\mathbf{m},\sigma)}, P_\theta\right) &\leq \ell\left(P_{(p,\mathbf{m},\sigma)}, P_{(\overline{Q},\mathbf{m},\sigma)}\right) + \ell\left(P_{(\overline{Q},\mathbf{m},\sigma)}, P_{(\overline{Q},\overline{\mathbf{m}},\overline{\sigma})}\right) \\
&\leq \ell\left(P_{(p,\mathbf{0},1)}, P_{(\overline{Q},\mathbf{0},1)}\right) + \ell\left(P_{(\overline{Q},\mathbf{0},1)}, P_{(\overline{Q},\frac{\overline{\mathbf{m}}-\mathbf{m}}{\sigma},\frac{\overline{\sigma}}{\sigma})}\right) \\
&\leq \eta + \left[A\left(\left|\frac{\mathbf{m}-\overline{\mathbf{m}}}{\overline{\sigma}}\right|_\infty^s + \left(1 - \frac{\sigma}{\overline{\sigma}}\right)^s\right)\right] \wedge 1 \\
&\leq \eta + 2A\delta^s = 2\eta.
\end{aligned}$$

Besides, the parameters $(j_0, \mathbf{j}) \in \mathbb{Z} \times \mathbb{Z}^k$ can be controlled in the following way. Using that $\sigma \leq \overline{\sigma}$, the inequality $\log(1+\delta) \leq \delta$ and (137), we obtain that for all

$i \in \{1, \ldots, k\}$,

$$|j_i| = \left|\frac{\overline{m}_i}{\overline{\sigma}\delta}\right| = \frac{1}{\overline{\sigma}\delta}\,|\overline{m}_i - m_i + m_i| \le \frac{1}{\overline{\sigma}\delta}\left[\overline{\sigma}\delta + \sigma\left|\frac{m_i}{\sigma}\right|\right] \le 1 + \frac{1}{\log(1+\delta)}\left|\frac{m_i}{\sigma}\right|.$$

Besides,

$$\begin{aligned} j_0 &= \frac{\log\overline{\sigma}}{\log(1+\delta)} = \frac{1}{\log(1+\delta)}\left[-\log\left(1 + \frac{\sigma}{\overline{\sigma}} - 1\right) + \log\sigma\right] \\ &\le \frac{1}{\log(1+\delta)}\left[-\log\left(1 - \frac{\delta}{1+\delta}\right) + |\log\sigma|\right] \\ &= \frac{1}{\log(1+\delta)}\left[\log(1+\delta) + |\log\sigma|\right] \le 1 + \frac{|\log\sigma|}{\log(1+\delta)} \end{aligned}$$

and using the inequality $\log(1+2x) \le 2\log(1+x)$, which holds for all $x \ge 0$, we obtain that

$$j_0 \ge \frac{\log\sigma}{\log(1+\delta)} \ge -\frac{|\log\sigma|}{\log(1+\delta)} \ge -\left[1 + \frac{|\log\sigma|}{\log(1+\delta)}\right].$$

Putting these inequalities together and using the fact that $P \in \mathscr{M}_n(K)$, we get

$$|(j_0, \mathbf{j})|_\infty \le 1 + \frac{1}{\log(1+\delta)}\left[|\log\sigma| \vee \left|\frac{\mathbf{m}}{\sigma}\right|_\infty\right] \le 1 + J_n. \tag{138}$$

For all $r > 0$, $e^{-L_\theta} \le \pi\left(\mathscr{B}(P_\theta, r)\right) \le 1$ and these two inequalities together with the definition (60) of η and Assumption 7 imply that for all $r > 0$

$$\begin{aligned} \frac{\pi\left(\mathscr{B}(P_\theta, 2r)\right)}{\pi\left(\mathscr{B}(P_\theta, r)\right)} &\le \exp\left[L_\theta\right] \le \exp\left[\widetilde{D}(\eta) + 2\sum_{i=0}^{k}\left[\frac{L}{2} + \log(1 + |j_i|)\right]\right] \\ &\le \exp\left[\gamma\tau^4 a_1^2 n\eta^2 + (k+1)\left[L + 2\log(1 + |(j_0, \mathbf{j})|_\infty)\right]\right]. \end{aligned}$$

Using (138), the definition (135) of J_n and the fact that $\log(2+x) \le \log 3 + \log x$ for all $x \ge 1$, we derive that

$$\begin{aligned} \frac{\pi\left(\mathscr{B}(P_\theta, 2r)\right)}{\pi\left(\mathscr{B}(P_\theta, r)\right)} &\le \exp\left[\gamma\tau^4 a_1^2 n\eta^2 + (k+1)L + 2(k+1)\log(2 + J_n)\right], \\ &\le \exp\left[K^2\gamma\tau^4 a_1^2 n\eta^2 + (k+1)\left(L + \log 9\right)\right] \end{aligned}$$

and since $\gamma = 1/6 \le L' = L + \log 9 < 3.1$,

$$\frac{1}{n\beta a_1} \le r_n(\beta, P_\theta) \le \frac{1}{\gamma n\beta a_1}\left[K^2\gamma\tau^4 a_1^2 n\eta^2 + (k+1)L'\right]$$

$$= \frac{1}{a_1\beta}\left[K^2\tau^4 a_1^2\eta^2 + \frac{(k+1)L'}{\gamma n}\right].$$

For the choice of $\beta = \beta_n$ given by (62),

$$\beta \geq \sqrt{K^2\tau^4 a_1^2\eta^2 + \frac{(k+1)L'}{\gamma n}} \geq \sqrt{\frac{k+1}{n}} \vee \frac{K\eta}{2}$$

hence, $r_n(\beta, P_\theta) \leq a_1^{-1}\beta$ and $P_\theta \in \mathscr{M}(\beta)$. This implies that

$$\begin{aligned}\inf_{P' \in \mathscr{M}(\beta)} \ell(\overline{P}^\star, P') + a_1^{-1}\beta &\leq \ell(\overline{P}^\star, P_\theta) + a_1^{-1}\beta \\ &\leq \ell(\overline{P}^\star, P) + \ell(P, P_\theta) + a_1^{-1}\beta \\ &\leq \ell(\overline{P}^\star, P) + 2\eta + \left[K\tau^2\eta + \frac{1}{a_1}\sqrt{\frac{(k+1)L'}{\gamma n}}\right],\end{aligned}$$

and the result follows by applying Corollary 1 and by using the fact that P is arbitrary in $\mathscr{M}_n(K)$.

10.6 Proof of Lemma 2

For all $p \in \mathcal{M}_0$, $\sigma \geq 1$ and $\mathbf{m} \in \mathbb{R}^k$, the supports of the functions $\boldsymbol{x} \mapsto p(\boldsymbol{x}/\sigma)$ and $\boldsymbol{x} \mapsto p((\boldsymbol{x} - \mathbf{m})/\sigma)$ are included in the set $\mathcal{K} = [0, \sigma]^k \cup \{\mathbf{m} + \boldsymbol{x},\ \boldsymbol{x} \in [0, \sigma]^k\}$ the Lebesgue measure of which is not larger than $2\sigma^k$. Consequently, using (66), we deduce that for all $p \in \mathcal{M}_0$, $\sigma \geq 1$ and $\mathbf{m} \in \mathbb{R}^k$,

$$\begin{aligned}&\left\|P_{(p,\mathbf{0},1)} - P_{(p,\mathbf{m},\sigma)}\right\| \\ &\quad\leq \left\|P_{(p,\mathbf{0},1)} - P_{(p,\mathbf{0},\sigma)}\right\| + \left\|P_{(p,\mathbf{0},\sigma)} - P_{(p,\mathbf{m},\sigma)}\right\| \\ &\quad= \frac{1}{2}\int_{\mathbb{R}^k}\left|p(\boldsymbol{x}) - \frac{1}{\sigma^k}p\left(\frac{\boldsymbol{x}}{\sigma}\right)\right|d\boldsymbol{x} + \frac{1}{2\sigma^k}\int_{\mathbb{R}^k}\left|p\left(\frac{\boldsymbol{x}}{\sigma}\right) - p\left(\frac{\boldsymbol{x}-\mathbf{m}}{\sigma}\right)\right|d\boldsymbol{x} \\ &\quad\leq \frac{1}{2}\int_{\mathbb{R}^k}\left|p(\boldsymbol{x}) - \frac{1}{\sigma^k}p(\boldsymbol{x})\right|d\boldsymbol{x} + \frac{1}{2\sigma^k}\int_{\mathbb{R}^k}\left|p(\boldsymbol{x}) - p\left(\frac{\boldsymbol{x}}{\sigma}\right)\right|d\boldsymbol{x} \\ &\qquad+ \frac{1}{2\sigma^k}\int_{\mathbb{R}^k}\left|p\left(\frac{\boldsymbol{x}}{\sigma}\right) - p\left(\frac{\boldsymbol{x}-\mathbf{m}}{\sigma}\right)\right|d\boldsymbol{x} \\ &\quad\leq \frac{1}{2}\int_{\mathbb{R}^k}\left|p(\boldsymbol{x}) - \frac{1}{\sigma^k}p(\boldsymbol{x})\right|d\boldsymbol{x} + \frac{1}{2\sigma^k}\int_{[0,1]^k}\left|p(\boldsymbol{x}) - p\left(\frac{\boldsymbol{x}}{\sigma}\right)\right|d\boldsymbol{x} \\ &\qquad+ \frac{1}{2\sigma^k}\int_{[0,\sigma]^k\setminus[0,1]^k}\left|p\left(\frac{\boldsymbol{x}}{\sigma}\right)\right|d\boldsymbol{x} + \frac{1}{2\sigma^k}\int_{\mathcal{K}}\left|p\left(\frac{\boldsymbol{x}}{\sigma}\right) - p\left(\frac{\boldsymbol{x}-\mathbf{m}}{\sigma}\right)\right|d\boldsymbol{x} \\ &\quad\leq \frac{1}{2}\left(1 - \frac{1}{\sigma^k}\right) + \frac{1}{2\sigma^k}\int_{[0,1]^k} L_1\left(1 - \frac{1}{\sigma}\right)^s |\boldsymbol{x}|^s\, d\boldsymbol{x}\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\int_{[0,1]^k\setminus[0,1/\sigma]^k}|p(\boldsymbol{x})|\,d\boldsymbol{x}+\frac{L_1}{2\sigma^k}\int_{\mathcal{K}}\left|\frac{\mathbf{m}}{\sigma}\right|^s d\boldsymbol{x}\\
&\le\frac{1}{2}\left(1-\frac{1}{\sigma^k}\right)+\frac{L_1k^{s/2}}{2\sigma^k}\left(1-\frac{1}{\sigma}\right)^s+\frac{L_0}{2}\left(1-\frac{1}{\sigma^k}\right)+L_1\left|\frac{\mathbf{m}}{\sigma}\right|^s\\
&\le\frac{1}{2}\left[1+L_1k^{s/2}+L_0\right]\left(1-\frac{1}{\sigma}\right)^s+L_1\left|\frac{\mathbf{m}}{\sigma}\right|^s
\end{aligned}
$$

and (57) is therefore satisfied with $A=L_1\vee[(1+L_1k^{s/2}+L_0)/2]$.

10.7 Proof of Lemma 3

By doing the change of variables $u=x-m$ in (68) if ever necessary, we may assume with no loss of generality that $m>0$. Then, since p is nonincreasing in $(0,+\infty)$ and vanishes elsewhere $p(x-m)\ge p(x)$ for all $x\ge m$ and $p(x)\ge p(x-m)=0$ for all $x\in(0,m)$. Consequently,

$$
\begin{aligned}
\int_{\mathbb{R}}|p(x)-p(x-m)|\,dx&=\int_0^m p(x)dx+\int_m^{+\infty}[p(x-m)-p(x)]\,dx\\
&=2\int_0^m p(x)dx+\int_m^{+\infty}p(x-m)dx-\int_0^{+\infty}p(x)dx\\
&\le 2mB+1-1,
\end{aligned}
$$

and we obtain (68).

Since $\sigma\ge1$, $p(x/\sigma)\ge p(x)$ and $p(x)/\sigma\le p(x)$ for all $x>0$. Hence,

$$
\begin{aligned}
&\int_{\mathbb{R}}\left|\frac{1}{\sigma}p\left(\frac{x}{\sigma}\right)-p(x)\right|dx\\
&\quad\le\int_{\mathbb{R}}\left|\frac{1}{\sigma}p\left(\frac{x}{\sigma}\right)-\frac{1}{\sigma}p(x)\right|dx+\int_{\mathbb{R}}\left|\frac{1}{\sigma}p\,(x)-p(x)\right|dx\\
&\quad=\frac{1}{\sigma}\int_{\mathbb{R}}\left(p\left(\frac{x}{\sigma}\right)-p(x)\right)dx+\int_{\mathbb{R}}\left(p(x)-\frac{1}{\sigma}p\,(x)\right)dx\\
&\quad=2\left(1-\frac{1}{\sigma}\right),
\end{aligned}
$$

which leads to (67).

Finally, by combining (68) and (67) we deduce that for all $m\in\mathbb{R}$ and $\sigma\ge1$

$$
\begin{aligned}
&\frac{1}{2}\int_{\mathbb{R}}\left|\frac{1}{\sigma}p\left(\frac{x-m}{\sigma}\right)-p(x)\right|dx\\
&\quad=\frac{1}{2}\int_{\mathbb{R}}\left|\frac{1}{\sigma}p\left(\frac{x-m}{\sigma}\right)-\frac{1}{\sigma}p\left(\frac{x}{\sigma}\right)\right|dx+\frac{1}{2}\int_{\mathbb{R}}\left|\frac{1}{\sigma}p\left(\frac{x}{\sigma}\right)-p(x)\right|dx\\
&\quad=\frac{1}{2}\int_{\mathbb{R}}\left|p\left(u-\frac{m}{\sigma}\right)-p(u)\right|du+\frac{1}{2}\int_{\mathbb{R}}\left|\frac{1}{\sigma}p\left(\frac{x}{\sigma}\right)-p(x)\right|dx
\end{aligned}
$$

$$\leq B\left|\frac{m}{\sigma}\right| + \left(1 - \frac{1}{\sigma}\right)$$

which yields to (69).

10.8 Proof of Proposition 6

This proposition is a consequence of Corollary 2. Let us first check that the assumptions of this corollary are satisfied. For all $S \in \mathscr{P}$, the mapping $\boldsymbol{\theta} \mapsto h(S, P_\theta)$ is continuous because of (71). It is therefore measurable and it follows from the definition of the algebra $\mathcal{A}$ that Assumption 1 is satisfied. Since the mapping $(x, \boldsymbol{\theta}) \mapsto p(x, \boldsymbol{\theta})$ is measurable, so are the mappings

$$\mathbf{p} : \left| \begin{array}{l} (\mathbb{R}^k \times E \times E \longrightarrow \mathbb{R}_+ \\ \quad (x, \boldsymbol{\theta}, \boldsymbol{\theta}') \longmapsto (p_{\boldsymbol{\theta}}(x),\, p_{\boldsymbol{\theta}'}(x)). \end{array} \right.$$

and $(x, \boldsymbol{\theta}, \boldsymbol{\theta}') \mapsto \psi\left(\sqrt{p_{\boldsymbol{\theta}'}(x)/p_{\boldsymbol{\theta}}(x)}\right)$, since ψ is measurable. We deduce that $(x, P, P') \mapsto t_{(P,P')}(x)$ is measurable on $(E \times \mathscr{M} \times \mathscr{M}, \mathcal{E} \otimes \mathcal{A} \otimes \mathcal{A})$ which proves that Assumption 3-(i) holds true. The requirements of Corollary 2 are therefore satisfied and we may apply it. In order to evaluate the quantity $r_n(\beta, P_{\boldsymbol{\theta}})$ for $\boldsymbol{\theta} \in \mathbb{R}^k$, we use the following lemma the proof of which is postponed to Sect. 10.9.

Lemma 10 *Let $\boldsymbol{\theta} \in [-R, R]^k$. For all $m \subset \{1, \ldots, k\}$ and $r > 0$*

$$\nu_m\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\ \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leq r\right\}\right)$$
$$= \begin{cases} \dfrac{1}{2^{|m|}} \displaystyle\prod_{i \in m} \left[\left(1 - \frac{|\theta_i|}{R}\right) \wedge \frac{r}{R} + \left(1 + \frac{|\theta_i|}{R}\right) \wedge \frac{r}{R}\right] & \text{if } |\theta_i| \leq r \text{ for all } i \notin m \\ 0 & \text{otherwise,} \end{cases}$$

with the convention $\prod_\varnothing = 1$. In particular, if $\boldsymbol{\theta} \in \Theta_m(R)$ and

$$\nu_m\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\ \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leq r\right\}\right) \geq \frac{1}{2^{|m|}}\left(\frac{r}{R} \wedge 1\right)^{|m|} \tag{139}$$

and for all $K > 1$

$$\frac{\nu_m\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\ \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leq Kr\right\}\right)}{\nu_m\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\ \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leq r\right\}\right)} \leq K^{|m|}. \tag{140}$$

Let us set $B = B_k$ for short and define $m^\star$ as the subset of $\{1, \ldots, k\}$ that minimizes over those $m \subset \{1, \ldots, k\}$ the mapping

$$m \mapsto \inf_{\boldsymbol{\theta} \in \Theta_m(R)} \ell(\overline{P}^\star, P_{\boldsymbol{\theta}}) + \frac{|m| \log\left(2kR(nB)^{1/s}\right) + 1}{\gamma n \beta a_1}.$$

Finally, let $\boldsymbol{\theta}^\star$ for some arbitrary element of $\Theta_{m^\star}(R)$. It follows from (71) and (139) that for all $r > 0$,

$$\begin{aligned}
1 &\geq \pi_m\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r)\right) \\
&= \nu_m\left(\left\{\boldsymbol{\theta} \in \mathbb{R}^k,\ h^2(P_{\boldsymbol{\theta}^\star}, P_{\boldsymbol{\theta}}) \leq r\right\}\right) \\
&\geq \nu_m\left(\left\{\boldsymbol{\theta} \in \mathbb{R}^k,\ \left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_\infty \leq (r/B)^{1/s}\right\}\right) \\
&\geq \frac{1}{2^{|m|}}\left(\frac{(r/B)^{1/s}}{R} \wedge 1\right)^{|m|} \geq \frac{1}{2^{|m|}}\left(\frac{(r \wedge 1)^{1/s}}{RB^{1/s}}\right)^{|m|}, \qquad (141)
\end{aligned}$$

where the last inequality holds true under the assumption that $RB^{1/s} \geq 1$.

We deduce from (141) that for all $r > 0$

$$\begin{aligned}
\frac{\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, 2r)\right)}{\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r)\right)} &\leq \frac{1}{\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r)\right)} \\
&\leq \frac{1}{\sum_{m \subset \{1,\ldots,k\}} e^{-L_m} \nu_m\left(\left\{\boldsymbol{\theta} \in \mathbb{R}^k,\ \left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_\infty \leq (r/B)^{1/s}\right\}\right)} \\
&\leq \frac{e^{L_{m^\star}}}{\nu_{m^\star}\left(\left\{\boldsymbol{\theta} \in \Theta_{m^\star},\ \left|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right|_\infty \leq (r/B)^{1/s}\right\}\right)} \\
&\leq \exp\left[L_{m^\star} + |m^\star| \log\left(\frac{2RB^{1/s}}{(r \wedge 1)^{1/s}}\right)\right] \\
&= \exp\left[|m^\star| \log\left(2kRB^{1/s}\right) + k \log\left(1 + \frac{1}{k}\right) + \frac{|m^\star|}{s} \log\left(\frac{1}{r} \vee 1\right)\right]. \qquad (142)
\end{aligned}$$

Provided that

$$r \geq \frac{|m^\star| \log\left(2kR(nB)^{1/s}\right) + 1}{\gamma n \beta a_1} \geq \frac{1}{n},$$

we obtain

$$\begin{aligned}
&|m^\star| \log\left(2kRB^{1/s}\right) + k \log\left(1 + \frac{1}{k}\right) + \frac{|m^\star|}{s} \log\left(\frac{1}{r} \vee 1\right) \\
&\quad \leq |m^\star| \log\left(2kRB^{1/s}\right) + k \log\left(1 + \frac{1}{k}\right) + |m^\star| \log\left(n^{1/s}\right) \\
&\quad \leq |m^\star| \log\left(2kR(nB)^{1/s}\right) + 1 \leq \gamma n \beta a_1 r
\end{aligned}$$

and deduce from (142) that $r_n(\beta, P_{\boldsymbol{\theta}^\star})$ defined by (11) satisfies

$$\frac{1}{n\beta a_1} \leq r_n(\beta, P_{\boldsymbol{\theta}^\star}) \leq \frac{|m^\star| \log\left(2kR(nB)^{1/s}\right) + 1}{\gamma n \beta a_1}.$$

Applying Corollary 2, we obtain that for some numerical constant $\kappa_0' > 0$,

$$\mathbb{E}\left[\widehat{\pi}_X\left({}^{\mathsf{c}}\mathscr{B}(\overline{P}^\star, \kappa_0' r(m^\star, \boldsymbol{\theta}^\star))\right)\right] \leq 2e^{-\xi}$$

with

$$r(m^\star, \boldsymbol{\theta}^\star) = \ell(\overline{P}^\star, P_{\boldsymbol{\theta}^\star}) + \frac{|m^\star| \log\left(2kR(nB)^{1/s}\right) + \xi}{\gamma n \beta a_1}.$$

Finally, the conclusion follows from the definition of $m^\star$ and the fact that $\boldsymbol{\theta}^\star$ is arbitrary in $\Theta_{m^\star}(R)$.

10.9 Proof of Lemma 10

Let $\theta \in \mathbb{R}$ and ν be the uniform distribution on $[-R, R]$. For all $\theta \in [-R, R]$ and $r > 0$,

$$\begin{aligned}
\nu\left([\theta - r, \theta + r]\right) &= \frac{1}{2R}\left[(\theta + r) \wedge R - (\theta - r) \vee (-R)\right]_+ \\
&= \frac{1}{2R}\left[(r + \theta) \wedge R + (r - \theta) \wedge R\right]_+ \\
&= \frac{1}{2R}\left[(r + |\theta|) \wedge R + (r - |\theta|) \wedge R\right]_+ \\
&= \frac{1}{2}\left[\left(1 - \frac{|\theta|}{R}\right) \wedge \frac{r}{R} + \left(1 + \frac{|\theta|}{R}\right) \wedge \frac{r}{R}\right].
\end{aligned}$$

Let now $\boldsymbol{\theta} \in \mathbb{R}^k$ such that $|\boldsymbol{\theta}|_\infty \leq R$. For all $m \subset \{1, \ldots, k\}$, $m \neq \varnothing$,

$$\nu_m\left(\left\{\boldsymbol{\theta}' \in \Theta_m,\ \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leq r\right\}\right) = 0$$

if there exists $i \notin m$ such that $|\theta_i| > r$. Otherwise

$$\begin{aligned}
\nu_m\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\ \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leq r\right\}\right) &= \nu_m\left(\left\{\boldsymbol{\theta}' \in \Theta_m,\ \max_{i \in m} |\theta_i' - \theta_i| \leq r\right\}\right) \\
&= \prod_{i \in m} \nu\left([\theta_i - r, \theta_i + r]\right) \\
&= \frac{1}{2^{|m|}} \prod_{i \in m}\left[\left(1 - \frac{|\theta_i|}{R}\right) \wedge \frac{r}{R} + \left(1 + \frac{|\theta_i|}{R}\right) \wedge \frac{r}{R}\right].
\end{aligned}$$

If $m = \varnothing$,

$$\nu_\varnothing\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\ \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leq r\right\}\right) = \mathbb{1}_{|\boldsymbol{\theta}|_\infty \leq r}.$$

Let us now turn to the proof of (140). Since $\boldsymbol{\theta} \in \Theta_m(R)$, for all $K' \in \{1, K\}$

$$\begin{aligned}
&\nu_m\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\ \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \le K'r\right\}\right) \\
&\quad = \nu_m\left(\left\{\boldsymbol{\theta}' \in \Theta_m,\ \max_{i\in m}\left|\theta_i' - \theta_i\right| \le K'r\right\}\right) \\
&\quad = \prod_{i\in m} \nu\left([\theta_i - K'r, \theta_i + K'r]\right),
\end{aligned}$$

It is therefore enough to show that for all $r > 0$ and $\theta \in [0, R]$

$$\Delta(r) = \frac{\nu\left([\theta - Kr, \theta + Kr]\right)}{\nu\left([\theta - r, \theta + r]\right)} \le K.$$

This is what we do now by distinguishing between several cases.

When $\theta + Kr \le R, \theta - Kr \ge 2\theta - R \ge -R$ and consequently, $\Delta(r) = K$. When $\theta + Kr > R$ and $-R \le \theta - Kr$,

$$\Delta(r) = \frac{R - (\theta - Kr)}{(\theta + r) \wedge R - (\theta - r)} = \begin{cases} \dfrac{R - \theta + Kr}{R - \theta + r} & \text{when } \theta + r > R \\ \dfrac{R - \theta + Kr}{2r} & \text{when } \theta + r \le R, \end{cases}$$

and the conclusion follows from the facts that $0 \le R - \theta \le Kr$. When $\theta + Kr > R$ and $\theta - Kr < -R, r \ge (\theta + R)/K \ge R/K$, hence $R + r - \theta \ge 2R/K$ and $R \le Kr$. Consequently,

$$\begin{aligned}
\Delta(r) &= \frac{2R}{(\theta + r) \wedge R - (\theta - r) \vee (-R)} \\
&= \begin{cases} \dfrac{2R}{2R} = 1 & \text{when } \theta + r > R \text{ and } \theta - r < -R \\ \dfrac{2R}{R + r - \theta} \le K & \text{when } \theta + r > R \text{ and } \theta - r \ge -R \\ \dfrac{2R}{2r} \le K & \text{when } \theta + r \le R, \end{cases}
\end{aligned}$$

which concludes the proof.

10.10 Proof of Proposition 8

Let ε be a small enough positive number. Since q is continuous and positive at $\boldsymbol{\theta}^\star$ and since $\mathcal{K}$ has a nonempty interior, there exists $z^\star > 0$ such that $\Theta^\star = \mathcal{B}_*(\boldsymbol{\theta}^\star, z^\star) \subset \mathcal{K}$,

$$0 < \underline{b}^\star \le q(\boldsymbol{\theta}) \le \overline{b}^\star \quad \text{with } \overline{b}^\star / \underline{b}^\star \le 1 + \varepsilon, \tag{143}$$

for all $\boldsymbol{\theta} \in \Theta^\star$ and

$$(1-\varepsilon)|\boldsymbol{\theta}-\boldsymbol{\theta}^\star|_*^s \le \ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le (1+\varepsilon)|\boldsymbol{\theta}-\boldsymbol{\theta}^\star|_*^s. \tag{144}$$

In particular, $\nu(\Theta^\star) > 0$ and we may define the distribution $\nu^\star = \nu(\cdot \cap \Theta^\star)/\nu(\Theta^\star)$ on $\Theta^\star$ with density $q^\star = q\mathbb{1}_{\Theta^\star}/\nu(\Theta^\star)$. Let $\mathscr{M}^\star = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta^\star\}$ and $\pi^\star$ be the prior on $\mathscr{M}^\star$ associated with $\nu^\star$. The parameter space $\Theta^\star$ is convex and it follows fom (144) that $(\Theta^\star, \boldsymbol{\theta}^\star, \ell, \nu^\star)$ satisfy Assumption 8-(i) with $\overline{a} = 1+\varepsilon$, $\underline{a} = 1-\varepsilon$. Besides, it follows from (143) that the density $q^\star$ satisfies condition (77) on $\Theta^\star$. We may apply Proposition 7 and deduce that for the model $(\mathscr{M}^\star, \pi^\star)$, $r_n^\star = r_n^\star(\beta, P_{\boldsymbol{\theta}^\star})$ is not larger than $\kappa_0^\star k/(\beta n)$ with

$$\kappa_0^\star = \frac{1}{a_1\gamma}\left\{\left[1+\frac{\log\left[2(1+\varepsilon)/(1-\varepsilon)\right]}{s\log 2}\right]\log\left(2(1+\varepsilon)\right)\right\} \vee 1 < \frac{(1+s^{-1})}{a_1\gamma}$$

for ε small enough. Consequently, by definition of $r_n^\star$, for all $r \ge r_n^\star$

$$\begin{aligned}
\pi^\star\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, 2r)\right) &= \frac{1}{\nu(\Theta^\star)}\nu\left(\{\boldsymbol{\theta}\in\Theta,\ \ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le 2r\} \cap \Theta^\star\right) \\
&\le \frac{\exp\left(\gamma n\beta a_1 r\right)}{\nu(\Theta^\star)}\nu\left(\{\boldsymbol{\theta}\in\Theta,\ \ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le r\} \cap \Theta^\star\right) \\
&\le \frac{\exp\left(\gamma n\beta a_1 r\right)}{\nu(\Theta^\star)}\nu\left(\{\boldsymbol{\theta}\in\Theta,\ \ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le r\}\right).
\end{aligned} \tag{145}$$

Let $r_1 = [(z^\star)^s \underline{a}_{\mathcal{K}}) \wedge \eta]/2$. If $r \in (0, r_1)$ and the parameter $\boldsymbol{\theta} \in \Theta$ satisfies $\ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le 2r$, then $\ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) < \eta$ and $\boldsymbol{\theta}$ necessarily belongs to $\mathcal{K}$ under Assumption 9-(ii). Applying (79) we deduce that for such a parameter $\boldsymbol{\theta} \in \Theta$

$$\underline{a}_{\mathcal{K}}\left|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\right|_*^s \le \ell\left(\boldsymbol{\theta},\boldsymbol{\theta}^\star\right) \le 2r < 2r_1 \le \underline{a}_{\mathcal{K}}(z^\star)^s,$$

which implies that $\boldsymbol{\theta} \in \Theta^\star$. For n large enough, $r_n^\star = \kappa_0^\star k/n < r_1$ and for $r \in (r_n^\star, r_1)$ we may therefore write, using (145),

$$\begin{aligned}
\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, 2r)\right) &= \nu\left(\{\boldsymbol{\theta}\in\Theta,\ \ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le 2r\}\right) \\
&= \nu\left(\{\boldsymbol{\theta}\in\Theta,\ \ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le 2r\} \cap \Theta^\star\right) \\
&\le \exp\left(\gamma n\beta a_1 r\right)\nu\left(\{\boldsymbol{\theta}\in\Theta,\ \ell(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le r\}\right) \\
&= \exp\left(\gamma n\beta a_1 r\right)\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r)\right).
\end{aligned} \tag{146}$$

Since q is bounded away from 0 in a neighbourhood of $\boldsymbol{\theta}^\star$, $\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r_1)\right) > 0$ and we may also write that for $r \ge r_1$ and n large enough

$$\begin{aligned}
\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r)\right) &\ge \pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r_1)\right) \\
&= \exp\left[\log\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r_1)\right) + \gamma n\beta a_1 r_1 - \gamma n\beta a_1 r_1\right] \\
&\ge \exp\left[-\gamma n\beta a_1 r_1\right] \ge \exp\left[-\gamma n\beta a_1 r\right]
\end{aligned}$$

$$\geq \exp\left[-\gamma n\beta a_1 r\right] \pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, 2r)\right). \tag{147}$$

Putting (146) and (147) together we obtain that for n large enough

$$\pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, 2r)\right) \leq \exp\left(\gamma n\beta a_1 r\right) \pi\left(\mathscr{B}(P_{\boldsymbol{\theta}^\star}, r)\right) \quad \text{for all } r \geq r_n^\star$$

and consequently that $r_n(\beta, P_{\boldsymbol{\theta}^\star}) \leq r_n^\star = \kappa_0^\star k/n$.

Acknowledgements The author is grateful to Lucien Birgé and the two anonymous referees for their support and suggestions, which contributed to improving a previous version of the present paper.

References

1. Alquier, P.: PAC-Bayesian bounds for randomized empirical risk minimizers. Math. Methods Stat. **17**(4), 279–304 (2008)
2. Atchadé, Y.A.: On the contraction properties of some high-dimensional quasi-posterior distributions. Ann. Stat. **45**(5), 2248–2273 (2017)
3. Audibert, J.-Y., Catoni, O.: Linear regression through PAC-Bayesian truncation. arXiv:1010.0072 (2011)
4. Baraud, Y.: Tests and estimation strategies associated to some loss functions. Probab. Theory Relat. Fields **180**(3), 799–846 (2021)
5. Baraud, Y., Birgé, L.: Rho-estimators revisited: general theory and applications. Ann. Stat. **46**(6B), 3767–3804 (2018)
6. Baraud, Y., Birgé, L.: Robust Bayes-like estimation: Rho-Bayes estimation. Ann. Stat. **48**(6), 3699–3720 (2020)
7. Baraud, Y., Birgé, L., Sart, M.: A new method for estimation and model selection: ρ-estimation. Invent. Math. **207**(2), 425–517 (2017)
8. Bhattacharya, A., Pati, D., Yang, Y.: Bayesian fractional posteriors. Ann. Stat. **47**(1), 39–66 (2019)
9. Birgé, L.: Approximation dans les espaces métriques et théorie de l'estimation. Z. Wahrsch. Verw. Gebiete **65**(2), 181–237 (1983)
10. Birgé, L.: An alternative point of view on Lepski's method. In: State of the Art in Probability and Statistics (Leiden, 1999), Volume 36 of IMS Lecture Notes Monograph Series, pp. 113–133. Institute of Mathematical Statistics, Beachwood (2001)
11. Birgé, L.: Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Ann. Inst. H. Poincaré Probab. Stat. **42**(3), 273–325 (2006)
12. Birgé, L.: About the non-asymptotic behaviour of Bayes estimators. J. Stat. Plan. Inference **166**, 67–77 (2015)
13. Birgé, L.: About the non-asymptotic behaviour of Bayes estimators. J. Stat. Plan. Inference **166**, 67–77 (2015)
14. Birgé, L., Massart, P.: Minimum contrast estimators on sieves: exponential bounds and rates of convergence. Bernoulli **4**(3), 329–375 (1998)
15. Birman, Mv., Solomjak, M.Z.: Piecewise polynomial approximations of functions of classes $W_p{}^\alpha$. Mat. Sb. (N.S.) **73**(115), 331–355 (1967)
16. Bissiri, P.G., Holmes, C.C., Walker, S.G.: A general framework for updating belief distributions. J. R. Stat. Soc. Ser. B. Stat. Methodol. **78**(5), 1103–1130 (2016)
17. Catoni, O.: Statistical learning theory and stochastic optimization. In: Lecture Notes from the 31st Summer School on Probability Theory Held in Saint-Flour, July 8–25, 2001. Springer, Berlin (2004)
18. Chernozhukov, V., Hong, H.: An MCMC approach to classical estimation. J. Econom. **115**(2), 293–346 (2003)
19. Ghosal, S., Ghosh, J.K., van der Vaart, A.W.: Convergence rates of posterior distributions. Ann. Stat. **28**(2), 500–531 (2000)
20. Ibragimov, I.A., Has'minskiĭ, R.Z.: Statistical Estimation. Asymptotic Theory, vol. 16. Springer, New York (1981)

21. Jiang, W., Tanner, M.A.: Gibbs posterior for variable selection in high-dimensional classification and data mining. Ann. Stat. **36**(5), 2207–2231 (2008)
22. Le Cam, L.: Convergence of estimates under dimensionality restrictions. Ann. Stat. **1**, 38–53 (1973)
23. Massart, P.: Concentration Inequalities and Model Selection, Volume 1896 of Lecture Notes in Mathematics. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003 (2007)
24. van der Vaart, A.W.: Asymptotic Statistics, Volume 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (1998)
25. Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis. Cambridge Mathematical Library. Cambridge University Press, Cambridge. An introduction to the general theory of infinite processes and of analytic functions; with an account of the principal transcendental functions, Reprint of the fourth (1927) edition (1996)