FROM ROBUST TESTS TO BAYES-LIKE POSTERIOR DISTRIBUTIONS

YANNICK BARAUD

Abstract. In the Bayes paradigm and for a given loss function, we propose the construction of a new type of posterior distributions for estimating the law of an n-sample. The loss functions we have in mind are based on the total variation distance, the Hellinger distance as well as some $\mathbb{L}_j$-distances. We prove that, with a probability close to one, this new posterior distribution concentrates its mass in a neighbourhood of the law of the data, for the chosen loss function, 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. For the total variation and squared Hellinger losses, we also show that the posterior distribution keeps its concentration properties when the data are only independent, hence not necessarily i.i.d., provided that most of their marginals are close enough to some probability distribution around which the prior puts enough mass. The posterior distribution is therefore also stable with respect to the equidistribution assumption. We illustrate these results by several applications. We consider the problems of estimating a location parameter or both the location and the scale of a density in a nonparametric framework. Finally, we also tackle the problem of estimating a density, with the squared Hellinger loss, in a high-dimensional parametric model under some sparcity conditions. The results established in this paper are non-asymptotic and provide, as much as possible, explicit constants.

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. Our aim is to estimate $P^\star$ from the observation of $\boldsymbol{X} = (X_1, \ldots, X_n)$ and to evaluate the performance of an estimator with values

Date: July, 1st 2021.

1991 *Mathematics Subject Classification.* Primary 62G05, 62G35, 62F35, 62F15.

Key words and phrases. 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 N° 811017.

in $\mathscr{M}$, we introduce a loss function ℓ from $\mathscr{P} \times \mathscr{M}$ with values in $\mathbb{R}_+$, where $\mathscr{P}$ denotes a suitable set of probabilities containing $P^\star$.

The loss functions we have in mind are based on distances such as the Hellinger or the total variation one. Given two probabilities P, Q on $(E, \mathcal{E})$, we recall that the total variation $\|P - Q\|$ and the squared Hellinger distance $h^2(P, Q)$ between P and Q are respectively given by the formulas

$$\|P - Q\| = \frac{1}{2} \int_E \left| \frac{dP}{d\mu} - \frac{dQ}{d\mu} \right| d\mu \tag{1}$$

and

$$h^2(P, Q) = \frac{1}{2} \int_E \left(\sqrt{\frac{dP}{d\mu}} - \sqrt{\frac{dQ}{d\mu}} \right)^2 d\mu \tag{2}$$

where μ dominates both P and Q, the result being independent of the choice of the reference measure μ.

Our approach to solve the problem has a Bayesian flavour since we endow $\mathscr{M}$ with a σ-algebra $\mathcal{A}$ and a probability measure π on $(\mathscr{M}, \mathcal{A})$ that plays the same role as the prior in the 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,* i.e. a set of the form

$$\mathscr{B}(P^\star, r) = \{P \in \mathscr{M},\ \ell(P^\star, P) \leqslant r\} \quad \text{with} \quad r > 0.$$

This means that with a probability close to 1, a point 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 loss ℓ) not larger than r.

The classical Bayes posterior distribution has been studied at length by van der Vaart and his co-authors — see for example Ghosal, Ghosh and van der Vaart (2000). They show that it concentrates around $P^\star$ 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 is unfortunately quite restrictive since such sets may be empty, and the condition therefore unsatisfied, when the probabilities in $\mathscr{M}$ are not equivalent. The situation is even worse when the probability $P^\star$ does not belong to $\mathscr{M}$, even in the favourable situation where it lies close to it, since the quantity $K(P^\star, P)$ can be very large and possibly infinite even when the total variation distance between $P^\star$ and P is very small. It is actually well-known that Bayes estimators are not robust with respect to a misspecification of the model. This weakness is due to the fact that they are based on the log-likelihood function which can be unstable when an outlier belongs to the data set.

In order to overcome this issue and propose more stable posterior distributions, some authors have replaced the log-likelihood function by another

one, leading to *quasi-posterior distributions*. This is the strategy that is proposed in Baraud and Birgé (2020) (see also the references therein). The authors proposed a surrogate to the Bayes posterior distribution that is called the *ρ-posterior distribution* in reference to the theory of ρ-estimation that was developped in the series of papers Baraud *et al* (2017) and Baraud and Birgé (2018). The ρ-posterior distribution preserves some of the nice features of the classical Bayes one but also possesses the property of being robust with respect to the presence of outliers and contaminated data among the sample. They show that it concentrates on a Hellinger ball around $P^\star$ as soon as the prior puts enough mass around a point which is close enough to $P^\star$. This key property confers to the ρ-posterior distribution the robustness the authors were looking for.

However, the ρ-posterior distribution is difficult to compute and therefore mainly of theoretical interest. It provides a benchmark to compare against. These difficulties are linked to the calculations of some suprema of empirical processes that are involved in the definition of the density of the ρ-posterior distribution. More deceiving is the fact that the authors do not show any improvement of their Bayes-like approach as compared to the frequentist one based on ρ-estimation. For a suitable choice of the prior, an estimator based on the ρ-posterior distribution would satisfy similar risks bound as those established for ρ-estimators. As a consequence, ρ-Bayes estimators do not seem to benefit from any gain that would result from a good choice of a prior as compared to the frequentist approach that presumes nothing.

Closer to our approach are the aggregation methods and PAC-Bayesian technics that have been popularized by Olivier Catoni in statistical learning (see Catoni (2004)). This approach has mainly been applied for the purpose of empirical risk minimization and statistical learning (see for example Alquier (2008)) and our aim is to extend it toward a versatile tool that can solve our estimation problem for various loss functions simultaneously.

The problem of designing a good estimator of $P^\star$ for a given loss function ℓ was solved in a frequentist way in Baraud (2021). 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 (1973), Birgé (1983) and Birgé (2006). The aim of the present paper is to transpose this theory from the frequentist to the Bayesian paradigm.

For very general models $\mathscr{M}$, we prove that $\widehat{\pi}_{\boldsymbol{X}}$ concentrates around $P^\star$ (for the desired loss) at a rate $r = r(n)$ which usually corresponds to the minimax one when π puts enough mass around $P^\star$ or, at least, around an element $\overline{P}$ which is close enough to $P^\star$. We therefore show that the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ enjoys some optimality and robustness properties with respect to the choices of $\mathscr{M}$ and of the prior π. In fact, we also show

that the posterior distribution is robust to the equidistribution assumption we started from. In particular, when ℓ is the total variation or the squared Hellinger loss, the concentration properties of $\widehat{\pi}_{\boldsymbol{X}}$ remain valid as long as the data are independent and most of their marginals lie close enough to $\overline{P}$. This result contrasts sharply with the instability of the classical Bayesian posterior distribution that we mentioned earlier.

Quite surprisingly, the concentration properties that we establish here require almost no assumption on $P^\star$ and $\mathscr{M}$. They mostly depend on the prior π and the loss function ℓ that have been chosen. More precisely, for a suitable element P which belongs to the model $\mathscr{M}$ and lies close enough to $P^\star$, these properties depend on the minimal value of r over which the ratio $\pi(\mathscr{B}(P,2r))/\pi(\mathscr{B}(P,r))$ becomes large enough. This ratio was introduced earlier in Birgé (2015) for the purpose of analyzing the behaviour of the classical posterior distribution in the Bayes paradigm. In our Bayes-like paradigm, we show that the choice of the prior completely encapsulates the complexity of the model $\mathscr{M}$. In particular, no assumption on the VC nor on the metric dimension of the model $\mathscr{M}$ is required. From this point of view, the results we establish here are of different nature than those obtained in the frequentist and Bayesian paradigms in Baraud (2021) and Baraud and Birgé (2020) which do require such assumptions.

Another difference with Baraud and Birgé (2020) lies in the construction of the posterior distribution. In the present paper, it does not involve any suprema of empirical processes but only integrals. It is therefore easier to compute even though the calculations of these integrals may not be necessarily easy, especially in high dimension.

The present paper is organized as follows. We present our statistical setting in Section 2. Unlike what has been described in this introduction, we actually consider independent but not necessarily i.i.d. data in order to analyse the behaviour of the posterior distribution with respect to a possible departure from equidistribution. Our main assumptions on the loss function are given and commented on in Section 3. In the remaining part of the paper, we shall mainly focus on the total variation and squared Hellinger losses. The construction of the posterior and its properties are presented in Section 4. Applications can be found in Section 5. There, we consider the problems of estimating a density in a location-scale family as well as that of a high-dimensional parameter in a parametric model under a sparcity constraint. We also show how our estimation strategy may lead to unusual rates of convergence for estimating a translation parameter in a non-regular statistical model. Finally, Section 6 is devoted to the proofs of the main theorems and Section 7 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 $\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star$. The probabilities $P_i^\star$ are assumed to belong to a given set $\mathscr{P}$ of probability measures on $(E, \mathcal{E})$. 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. We endow $\mathscr{M}$ with a σ-algebra $\mathcal{A}$ and a probability measure π, that we call *a prior*, 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 most of the $P_i^\star$ are close enough to some point $\overline{P}$ in $\mathscr{M}$ around which the prior π puts enough mass. In order to be more specific, we introduce a loss function ℓ, which is a mapping from $(\mathscr{P} \cup \mathscr{M}) \times \mathscr{M}$ into $\mathbb{R}_+$, and write

$$\ell(\mathbf{P}^\star, Q) = \frac{1}{n} \sum_{i=1}^n \ell(P_i^\star, Q) \quad \text{for all } Q \in \mathscr{M}.$$

In order to avoid trivialities, we assume that ℓ is not constantly equal to 0 on $(\mathscr{M} \cup \mathscr{P}) \times \mathscr{M}$. 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 $P \in \mathscr{M}$ with radius $r > 0$ the subset of $\mathscr{M}$ defined and denoted by $\mathscr{B}(P, r) = \{Q \in \mathscr{M},\ \ell(P, Q) \leqslant r\}$. By extension, we set

$$\mathscr{B}(\mathbf{P}^\star, r) = \{Q \in \mathscr{M},\ \ell(\mathbf{P}^\star, Q) \leqslant r\} \quad \text{for all } r > 0.$$

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

Throughout this paper, we use the following notations. The subsets of $\mathbb{R}, \mathbb{R}^k$ will be equipped with their Borel σ-algebras. The cardinality of a set A is denoted $|A|$ and the elements of $\mathbb{R}^k$ with $k > 1$ are denoted with bold letters, e.g. $\mathbf{x} = (x_1, \ldots, x_k)$ and $\mathbf{0} = (0, \ldots, 0)$. For $\mathbf{x} \in \mathbb{R}^k$, $|\mathbf{x}|_\infty = \max_{i \in \{1,\ldots,k\}} |x_i|$ while $|\mathbf{x}|$ denotes the Euclidean norm of $\mathbf{x}$. For all suitable functions f on $(E^n, \mathcal{E}^{\otimes n})$, $\mathbb{E}[f(\boldsymbol{X})]$ means $\int_{E^n} f d\mathbf{P}^\star$ while for f on $(E, \mathcal{E})$, $\mathbb{E}_S[f(X)]$ denotes the integral $\int_E f dS$ with respect to some measure S on $(E, \mathcal{E})$. For $x \in \mathbb{R}$, $(x)_+ = \max(x, 0)$ and $(x)_- = \max\{-x, 0\}$. 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 $\|f\|_{j,\mu} = [\int_E |f|^j d\mu]^{1/j} < +\infty$. Finally, $\|f\|_\infty = \sup_{x\in E}|f(x)|$ is the supremum norm of a function f on E.

3. Our main assumptions on the loss function

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

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

is measurable and there exists a positive number τ such that for all $P \in \mathscr{P}$ and $\overline{P}, Q \in \mathscr{M}$

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

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

Under such an assumption, ℓ-balls are measurable, i.e. belong to $\mathcal{A}$. 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 $\tau = 2$.

The construction of the posterior distribution not only depends on the prior π but also on the loss function ℓ. For this reason, we introduce a family $\mathscr{T}(\ell,\mathscr{M}) = \left\{t_{(P,Q)},\ (P,Q)\in\mathscr{M}^2\right\}$ of functions on $(E,\mathcal{E})$ with the following properties.

Assumption 2. *The elements $t_{(P,Q)}$ of $\mathscr{T}(\ell,\mathscr{M})$ satisfy:*

(i) The mapping

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

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 $P,Q\in\mathscr{M}$,

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

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

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

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

Many classical loss functions (among which the total variation distance, the 1-Wasserstein distance, the squared Hellinger loss, etc.) can be associated to families $\mathscr{T}(\ell,\mathscr{M})$ satisfying our Assumption 2 — see Baraud (2021) —. Some of them possess the additional property given below.

Assumption 3. *Additionally to Assumption 2, there exists $a_2 > 0$ such that*

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

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

This assumption is typically satisfied when ℓ behaves as the square of a distance.

In the proposition below we provide families $\mathscr{T}(\ell, \mathscr{M})$ that do satisfy our requirements for some loss functions ℓ of interest. These results have been established in Baraud (2021) except for the squared Hellinger loss for which we refer to Baraud & Birgé (2018)[Proposition 3]. The list below is not exhaustive and other losses can also be considered, especially those that can be defined by 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. We refer to Baraud (2021) for more details on the way the families $\mathscr{T}(\ell, \mathscr{M})$ can be obtained from the loss functions ℓ.

Proposition 1. *The following holds:*

(1) Total variation. *Let $\overline{\mathscr{P}}$ be the set of all probability measures on $(E, \mathcal{E})$, $\mathscr{M}$ a subset of $\overline{\mathscr{P}}$ dominated by some reference measure μ and for $P, Q \in \overline{\mathscr{P}}$, $\ell(P,Q) = \|P - Q\|$ the total variation loss (TV-loss for short) between P and Q. The family $\mathscr{T}(\ell, \mathscr{M})$ of functions $t_{(P,Q)}$ defined for $P = p \cdot \mu, Q = q \cdot \mu \in \mathscr{M}$ by*

$$t_{(P,Q)} = \frac{1}{2}\left[\mathbb{1}_{q>p} - Q(q > p)\right] - \frac{1}{2}\left[\mathbb{1}_{p>q} - P(p > q)\right] \tag{6}$$

satisfies Assumption 2 with $a_0 = 3/2$ and $a_1 = 1/2$. If $\mathscr{T}(\ell, \mathscr{M})$ satisfies Assumption 2 whatever the model $\mathscr{M}$, it may also occasionally satisfy Assumption 3 for some specific $\mathscr{M}$. An example of such a model is given Section 5.2.

(2) Hellinger distance. *Let $\overline{\mathscr{P}}$ be the set of all probability measures on $(E, \mathcal{E})$, $\mathscr{M}$ a subset of $\overline{\mathscr{P}}$ dominated by some reference measure μ and for $P, Q \in \overline{\mathscr{P}}$, $\ell(P,Q) = h^2(P,Q)$ the squared Hellinger distance between P and Q. Besides, 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 $\mathscr{T}(\ell, \mathscr{M})$ *of functions* $t_{(P,Q)}$ *defined for* $P = p \cdot \mu, Q = q \cdot \mu \in \mathscr{M}$ *by*

$$t_{(P,Q)} = \frac{1}{2}\psi\left(\sqrt{\frac{q}{p}}\right) \tag{7}$$

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

(3) $\mathbb{L}_j$-loss *with* $1 < j < +\infty$. *For* $j \in (1, +\infty)$, *let* $\overline{\mathscr{P}}_j$ *be the set of finite and signed measures on* $(E, \mathcal{E})$ *of the form* $P = p \cdot \mu$ *with* $p \in \mathscr{L}_j(E, \mu) \cap \mathscr{L}_1(E, \mu)$ *and* $\mathscr{M} = \{P = p \cdot \mu,\ p \in \mathcal{M}\}$ *be a subset of* $\overline{\mathscr{P}}_j$. *Assume that the family* $\mathcal{M}$ *of densities satisfies*

$$\|p - q\|_\infty \leqslant R \|p - q\|_{\mu,j} \quad \text{for all } p, q \in \mathcal{M} \text{ and some } R > 0. \tag{8}$$

For $P = p \cdot \mu$ *and* $Q = q \cdot \mu$ *in* $\mathscr{M}$, *define*

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

The family $\mathscr{T}(\ell, \mathscr{M})$ *of functions* $t_{(P,Q)}$ *defined for* $P, Q \in \mathscr{M}$ *by*

$$t_{(P,Q)} = \frac{1}{2R^{j-1}}\left[\int_E f_{(P,Q)} \frac{dP + dQ}{2} - f_{(P,Q)}\right] \tag{9}$$

satisfies Assumption 2 with $a_0 = 3/(4R^{j-1})$ *and* $a_1 = 1/(4R^{j-1})$ *for the loss* $\ell = \ell_j$ *with* $\ell_j(P, Q) = \|p - q\|_{\mu,j}$ *for all* $P = p \cdot \mu$ *and* $Q = q \cdot \mu$ *in* $\overline{\mathscr{P}}_j$.

When $j = 2$, (8) is typically satisfied when $\mathcal{M}$ is a subset of a linear space enjoying good connections between the $\mathbb{L}_2$ 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 (1998)[Section 3] for additional examples. The property may also hold for infinite dimensional linear spaces as proven in Baraud (2021).

4. Construction of the posterior distribution and main results

4.1. **Construction of the posterior distribution.** It relies on two positive numbers β and

$$\lambda = (1 + c)\beta \quad \text{with} \quad c > 0 \text{ such that } c_0 = (1 + c) - c(a_0/a_1) > 0. \tag{10}$$

Given the family $\mathscr{T}(\ell, \mathscr{M})$, 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 $(\mathscr{M}, \mathcal{A}, \pi)$ with density

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

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

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

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

$$\frac{\widehat{\pi}_{\boldsymbol{X}}}{d\pi} : P \mapsto \frac{\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]}{\int_{\mathscr{M}} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P)}. \tag{11}$$

Our Assumption 2-(i) ensures that $\widetilde{\pi}_{\boldsymbol{X}}(\cdot|P)/d\pi$ is a measurable function of $(\boldsymbol{X}, P, Q)$ and $\widehat{\pi}_{\boldsymbol{X}}/d\pi$ a measurable function of $(\boldsymbol{X}, P)$.

4.2. **The influence of the prior.** The Bayesian paradigm offers the possibility to favour some elements of $\mathscr{M}$ as compared to others. In order to evaluate how much the prior π advantages or disadvantages an element $\overline{P} \in \mathscr{M}$, we fix some number $\gamma > 0$, introduce the set

$$\mathcal{R}(\beta, \overline{P}) = \left\{ r \geqslant \frac{1}{n\beta a_1},\ \frac{\pi\left(\mathscr{B}(\overline{P}, 2r)\right)}{\pi\left(\mathscr{B}(\overline{P}, r)\right)} > \exp\left(\gamma n \beta a_1 r\right) \right\}$$

with the convention $a/0 = +\infty$ for all $a \geqslant 0$ and finally define

$$r_n(\beta, \overline{P}) = \sup \mathcal{R}(\beta, \overline{P}) \quad \text{with} \quad \sup \varnothing = 1/(n\beta a_1). \tag{12}$$

It follows from the definition of $r_n(\beta, \overline{P})$ that

$$0 < \pi\left(\mathscr{B}(\overline{P}, 2r)\right) \leqslant \exp\left(\gamma n \beta a_1 r\right) \pi\left(\mathscr{B}(\overline{P}, r)\right) \quad \text{for all } r > r_n(\beta, \overline{P}). \tag{13}$$

Letting r decrease to $r_n(\beta, \overline{P})$, we derive that (13) holds for $r = r_n(\beta, \overline{P})$. In particular, $\pi\left(\mathscr{B}(\overline{P}, r)\right) > 0$ for $r = r_n(\beta, \overline{P})$.

We shall see below that the performance of the posterior $\widehat{\pi}_{\boldsymbol{X}}$ depends on those $\overline{P} \in \mathscr{M}$ such that $r_n(\beta, \overline{P})$ is small enough. The connection between the behaviour of the prior π in the vicinity of an element $\overline{P} \in \mathscr{M}$ and the quantity $r_n(\beta, \overline{P})$ can be made as follows. Clearly, if the prior puts no mass on the ℓ-ball $\mathscr{B}(\overline{P}, r)$ then $r_n(\beta, \overline{P}) > r$ and $r_n(\beta, \overline{P})$ is therefore large if r is large. In the opposite case, if the prior puts enough mass on $\mathscr{B}(\overline{P}, r)$ in the sense that

$$\pi\left(\mathscr{B}(\overline{P}, r)\right) \geqslant \exp\left(-\gamma n \beta a_1 r\right), \tag{14}$$

then for all $r' \geqslant r$,

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

which implies that $r_n(\beta, \overline{P}) \leqslant r$ and $r_n(\beta, \overline{P})$ is therefore small if r is small. Although (14) is not equivalent to (13) (it is actually stronger), the previous arguments provide a partial view on the relationship between π and r_n and conditions to decide whether $\overline{P}$ is favoured by π or not, according to the size of $r_n(\beta, \overline{P})$.

4.3. **A first result on the concentration property of the posterior distribution.** Following the above discussion, when the set

$$\mathscr{M}(\beta) = \left\{ \overline{P} \in \mathscr{M},\ r_n(\beta, \overline{P}) \leqslant a_1^{-1}\beta \right\} \tag{15}$$

is non-empty, it gathers the most favoured elements of the model $(\mathscr{M}, \pi)$ at level $a_1^{-1}\beta$. The set $\mathscr{M}(\beta)$ can alternatively be defined as

$$\mathscr{M}(\beta) = \left\{ \overline{P} \in \mathscr{M},\ \sup_{r \geqslant a_1^{-1}\beta} \left[\frac{1}{\gamma n a_1 r} \log \left(\frac{\pi\left(\mathscr{B}(\overline{P}, 2r)\right)}{\pi\left(\mathscr{B}(\overline{P}, r)\right)} \right) \right] \leqslant \beta \right\}. \tag{16}$$

It is sometimes easier to use this latter form for the calculations. The set $\mathscr{M}(\beta)$ will play a crucial role in our first result below.

Theorem 1. *Let Assumptions 1 and 2 be satisfied, fix $\gamma < (c_0 \wedge c)/(2\tau)$ and let $\beta \geqslant 1/\sqrt{n}$ be chosen in such a way that the set $\mathscr{M}(\beta)$ defined by (15) is not empty. Then, the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ defined by (11) possesses the following property. Given $\xi > 0$, there exists a number $\kappa_0 \geqslant 2$ depending only on c, τ, γ, ξ and the ratio a_0/a_1 such that, for any distribution $\mathbf{P}^\star$,*

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\mathbf{P}^\star, \kappa_0 r)\right)\right] \leqslant 2e^{-\xi} \quad \textit{with} \quad r = \inf_{P \in \mathscr{M}(\beta)} \ell(\mathbf{P}^\star, P) + a_1^{-1}\beta. \tag{17}$$

In particular,

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

A suitable choice for κ_0 is given by (81) and it is of the form $A + B\xi$ where A and B only depends on c, τ, γ and a_0/a_1. It therefore only depends on ξ linearly and on the choice of our loss function ℓ but not on the prior π. Hence, given a loss function ℓ and a confidence level $1 - 2e^{-\xi}$, κ_0 is a numerical constant.

Our posterior distribution depends on the parameter $\lambda = (1 + c)\beta$ where $c > 0$ satisfies (10). By Proposition 1, this constraint is satisfied for any $c \in (0, 1/2)$ when ℓ is the TV or an ℓ_j loss. For the choice $c = 1/3$, $c_0 = c$ and our condition on γ becomes $\gamma \leqslant (6\tau)^{-1}$.

When the data are truly i.i.d. and the prior puts enough mass around their common distribution $P^\star$, in the sense that $P^\star \in \mathscr{M}(\beta)$, then $r = a_1^{-1}\beta$. 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 $\inf_{P \in \mathscr{M}(\beta)} \ell(\mathbf{P}^\star, P)$. When this quantity 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 concentration properties of the posterior distribution is therefore stable with respect to a possible misspecification of the model and a departure from the equidistribution assumption.

The value of r given by (17) depends on the choice of the parameter β. Since the set $\mathscr{M}(\beta)$ is increasing with β (for the inclusion), the two terms $\inf_{P\in\mathscr{M}(\beta)} \ell(\mathbf{P}^\star, P)$ and $a_1^{-1}\beta$ vary in opposite directions when β increases. The set $\mathscr{M}(\beta)$ must be large enough to provide a suitable approximation of $\mathbf{P}^\star$, and therefore include as many elements of $\mathscr{M}$ as possible since $\mathbf{P}^\star$ is unknown, but β must not be too large in order to keep $a_1^{-1}\beta$ to a reasonable size.

In order to illustrate and comment further on the choice of the parameter β, let us consider the following simple example.

Example 1. Assume that the data are truly i.i.d. with distribution $P^\star$ and that $\mathscr{M}$ is a parametric model on which π behaves like the uniform distribution on a suitable bounded subset of $\mathbb{R}^D$. More precisely, assume that for all $P \in \mathscr{M}$ and $r > 0$

$$(Ar)^D \wedge 1 \leqslant \pi\left(\mathscr{B}(P, r)\right) \leqslant (Br)^D \wedge 1$$

for some positive numbers $A \leqslant B$ and $D \geqslant 1$. Note that this assumption implies that $\pi\left(\mathscr{B}(P, A^{-1})\right) = 1$ for all $P \in \mathscr{M}$ so that the diameter of the support of π is bounded by $2\tau A^{-1}$. Then,

$$\frac{\pi\left(\mathscr{B}(P, 2r)\right)}{\pi\left(\mathscr{B}(P, r)\right)} \leqslant \left(\frac{2B}{A}\right)^D \quad \text{for all } P \in \mathscr{M} \quad \text{and} \quad r > 0 \tag{18}$$

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

$$\sup_{r \geqslant a^{-1}\beta} \left[\frac{1}{\gamma n a_1 r} \log\left(\frac{\pi\left(\mathscr{B}(P, 2r)\right)}{\pi\left(\mathscr{B}(P, r)\right)}\right)\right] \leqslant \frac{D}{\gamma n \beta} \log\left(\frac{2B}{A}\right)$$

and we note that the right-hand side is not larger than $\beta \geqslant 1/\sqrt{n}$ for

$$\beta = \sqrt{\frac{D \log(2B/A)}{\gamma n}} \vee \frac{1}{\sqrt{n}}. \tag{19}$$

This means that for such a value of β, $P \in \mathscr{M}(\beta)$, and since P is arbitrary we obtain that $\mathscr{M}(\beta) = \mathscr{M}$. We derive from Theorem 1 that the distribution $\widehat{\pi}_{\boldsymbol{X}}$ concentrates on an ℓ-ball centered at $\mathbf{P}^\star$ with a radius of order

$$r = \inf_{P\in\mathscr{M}} \ell(\mathbf{P}^\star, P) + a_1^{-1}\sqrt{\frac{D}{n}}.$$

In particular we derive, using Proposition 1, that for the TV-loss

$$r = \inf_{P\in\mathscr{M}} \left[\frac{1}{n}\sum_{i=1}^n \|P_i^\star - P\|\right] + \frac{1}{2}\sqrt{\frac{D}{n}}.$$

Provided that $\mathscr{M}$ is of the form $\mathscr{M} = \{P = p \cdot \mu,\ p \in \mathcal{M}\}$ with $\mathcal{M}$ satisfying (8), we also derive that for the ℓ_j-loss and any distribution $\mathbf{P}^\star = \bigotimes_{1=1}^n (p_i^\star \cdot \mu)$ with $p_1^\star, \ldots, p_n^\star \in \mathscr{L}_j(E, \mathcal{E}, \mu)$,

$$r_n = \inf_{p \in \mathcal{M}} \left[\frac{1}{n} \sum_{i=1}^n \|p_i^\star - p\|_{\mu,j} \right] + 4R^{j-1} \sqrt{\frac{D}{n}}.$$

This example shows that in the context of Theorem 1, the parameter β must be tuned as a suitable function of n, that depends on the properties of the prior distribution, in such a way that all (or most) of the elements of $\mathscr{M}$ belong to $\mathscr{M}(\beta)$.

Quite surprisingly, the situation changes drastically when Assumption 3 is met as we shall see in the next section.

4.4. **The concentration property of the posterior distribution under Assumption 3.** Let us define the mapping

$$\phi : \left| \begin{array}{rcl} (0, +\infty) & \longrightarrow & \mathbb{R}_+ \\ z & \longmapsto & \phi(z) = \dfrac{2\,(e^z - 1 - z)}{z^2}. \end{array} \right. \tag{20}$$

The function ϕ is increasing on $(0, +\infty)$ and tends to 1 when z tends to 0.

Theorem 2. *Assume that Assumptions 1 and 3 hold and define*

$$c_1 = c_0 - \beta a_2 a_1^{-1} \tau^2 \phi\left[\beta(1+2c)\right](1 + 2c(1+c)); \tag{21}$$

$$c_2 = c - \beta a_2 a_1^{-1} \tau^2 \phi\left[\beta(1+2c)\right] c^2; \tag{22}$$

$$c_3 = (2+c) - \beta a_2 a_1^{-1} \tau^2 \phi\left[\beta(3+2c)\right](2+c)^2. \tag{23}$$

Let $\gamma < (c_1 \wedge c_2 \wedge c_3)/(2\tau)$ and β_0 be the value of β for which $c_1 \wedge c_2 \wedge c_3 = 0$. Then, for $\beta \in (0, \beta_0)$ and $n \geqslant 1/(\beta a_1)$, the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ defined by (11) satisfies the following property. Given $\xi > 0$, there exists a number κ_0 depending only on $a_0, a_1, a_2, c, \tau, \beta, \gamma$ and ξ such that, for any distribution $\mathbf{P}^\star$,

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\mathbf{P}^\star, \kappa_0 r)\right)\right] \leqslant 2e^{-\xi} \text{ with } r = \inf_{P \in \mathscr{M}} \left[\ell(\mathbf{P}^\star, P) + r_n(\beta, P)\right]. \tag{24}$$

In particular,

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

It follows from the proof that one may take κ_0 given by (93), which is of the form $(A' + B'\xi)r$ with A' and B' depending only on $a_0, a_1, a_2, c, \tau, \beta$ and γ. Note that the constraints on c, β and γ that are required in our Theorem 2 only depend on a_0, a_1 and a_2, hence on the choice of the loss function ℓ, but not on the model $(\mathscr{M}, \pi)$. In particular, unlike Theorem 1, the value of β can be chosen as a universal constant for a given loss function. For example,

when $\ell = h^2$, we know from Proposition 1 that $a_0 = 2$, $a_1 = 3/16$ and $a_2 = 3\sqrt{2}/4$, and we may take $c = 0.05$, $\beta = 0.01$, $\gamma = 0.01$ and $\tau = 2$ in (3).

In order to illustrate this new result and compare it with that of Theorem 1, let us go back to the framework of Example 1.

Example 2 (Example 1 continued)**.** Assume that $\ell = h^2$ is the squared Hellinger loss and that the quantities c, β and γ have been chosen as above.

We see that that the right-hand side of (18) is not larger than $\exp(\gamma n a_1 \beta r)$ provided that

$$r \geqslant \frac{D \log(2B/A)}{\gamma n a_1 \beta}.$$

For $\gamma \leqslant D \log(2B/A)$, the right-hand side of this latter inequality is not smaller than $1/(n\beta a_1)$ and we derive from the definition (12) of $r_n(\beta, P)$ that

$$r_n(\beta, P) \leqslant \frac{D \log(2B/A)}{\gamma n a_1 \beta} \quad \text{for all } P \in \mathscr{M}.$$

By applying Theorem 2 we obtain that the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ concentrates on an ℓ-ball with radius of order $\inf_{P \in \mathscr{M}} \ell(\mathbf{P}^\star, P) + D/n$, hence at rate $1/n$ when the data are i.i.d. with distribution $P^\star \in \mathscr{M}$.

Applying our Theorem 1 under the only Assumption 2 and ignoring the fact that the loss $\ell = h^2$ additionally satisfies Assumption 3, would lead, by arguing as for the TV-loss, to the weaker result that the posterior distribution concentrates on an ℓ-ball with radius of order $\inf_{P \in \mathscr{M}} \ell(\mathbf{P}^\star, P) + \sqrt{D/n}$.

We conclude that Theorem 2 leads to a stronger result on the concentration properties of $\widehat{\pi}_{\boldsymbol{X}}$ as compared to Theorem 1 when the loss function ℓ satisfies Assumption 3 on the model $\mathscr{M}$.

5. Applications

5.1. **How big is the set $\mathscr{M}(\beta)$ in a translation model?** In this section, we consider the translation model $\mathscr{M} = \{P_\theta = p(\cdot - \theta) \cdot \mu, \theta \in \mathbb{R}\}$ associated to a density p on $\mathbb{R}$ with respect to the Lebesgue measure μ. Given a density q on the real line and a scale parameter σ, we estimate the location parameter θ by choosing the prior ν_σ with density $q_\sigma : \theta \mapsto \sigma^{-1} q(\theta/\sigma)$ with respect to μ. The prior π on $\mathscr{M}$ is the image of ν_σ by the mapping $\theta \mapsto P_\theta$ and we use the total variation distance to measure the quality of an estimator of P_θ.

By Theorem 1, we know that the concentration properties of the posterior $\widehat{\pi}_{\boldsymbol{X}}$ depend on the size of the set $\mathscr{M}(\beta)$ given by (16). Given a compact interval $I \subset \mathbb{R}$, our aim is to find a value of $\beta \geqslant 1/\sqrt{n}$ for which $\mathscr{M}(\beta)$ contains the subset $\{P_\theta,\ \theta \in I\} \subset \mathscr{M}$. We assume the following.

Assumption 4. *The density q is positive, symmetric and decreasing on $\mathbb{R}_+$. Besides, there exists $L \in (0, +\infty]$ such that the mapping H defined by*

$$H : \left| \begin{array}{rcl} [0, L) & \longrightarrow & [0, 1) \\ t & \longmapsto & \|P_t - P_0\| \end{array} \right. \tag{25}$$

is bijective.

Under Assumption 4, H is necessarily increasing on $[0, L)$ and we may define its inverse $G : [0, 1) \to [0, L)$. We set

$$\overline{\Gamma} = \max \left\{ \left[\sup_{0<r\leqslant 1/4} \frac{G(2r)}{G(r)} \right] q(0), \frac{1}{2G(1/4)} \right\} \tag{26}$$

and assume that this quantity is finite. Note that $\overline{\Gamma}$ only depends on the choices of p and q. For example, when $p(x) = (1/2)e^{-|x|}$, $L = +\infty$, $H : t \mapsto 1-\exp[-t/2]$, $G : r \mapsto -2\log(1-r)$ and since the mapping $r \mapsto [G(2r)/G(r)]$ is increasing,

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

If $p : x \mapsto (\alpha/2)(1-|x|)^{-1+\alpha} \mathbb{1}_{|x|<1}$ with $\alpha > 0$, $L = 2$, $H : t \mapsto 1-(1-t/2)^\alpha$, $G : r \mapsto 2[1-(1-r)^{1/\alpha}]$. Since $G(r) \sim 2r/\alpha$ in a neighbourhood of 0, the mapping $r \mapsto G(2r)/G(r)$ is continuous on $[0, 1/4]$ and therefore bounded. Given $q(0)$, $\overline{\Gamma}$ is therefore a finite number.

The following result is proven in Section (7.1).

Proposition 2. *Let Assumption 4 hold, $\overline{\Gamma}$ the quantity defined by (26), $\gamma \leqslant \log 4$ and t a positive number such that $\nu_1([t, +\infty)) \leqslant 1/4$. The set $\mathscr{M}(\beta)$ contains the subset $\{P_\theta, \theta \in [-\sigma t, \sigma t]\}$ if*

$$\beta \geqslant \sqrt{\frac{1}{n\gamma} \max \left\{ \log \left(\frac{\overline{\Gamma}\,(\sigma \vee 1)}{q(2t)} \right), \log 4 \right\}}. \tag{27}$$

Note that the interval $I = [-\sigma t, \sigma t]$ can be enlarged by increasing the value of σ or that of t. In the first case, increasing σ makes the prior ν_σ flatter and for a fixed value of $t > 0$, the right-hand side of (27) increases as $\sqrt{\log \sigma}$ when σ becomes larger than 1. In the other case, for a fixed value of σ, the right-hand side of (27) increases as $\sqrt{\log(1/q(2t))}$. 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.

5.2. **Fast rates.** We go back to the statistical framework described in Section 5.1 in the special case where p is the density $x \mapsto \alpha x^{\alpha-1} \mathbb{1}_{(0,1]}$ with $\alpha \in (0, 1]$. As before, we choose the TV-loss. Since ℓ is a distance we may take $\tau = 1$ in Assumption 1. Besides, we have seen in Proposition 1 that the family $\mathscr{T}(\mathscr{M}, \ell)$ given by (6) satisfies our Assumption 2 with $a_0 = 3/2$

and $a_1 = 1/2$. In fact, it turns out that in the specific situation we consider here, the TV-loss also satisfies Assumption 3 with $a_2 = 1$. This means that from a more statistical point of view, the TV-loss rather behaves here as the square of a distance. In addition, some simple calculations show that

$$\|P_\theta - P_{\theta'}\| = \left|\theta - \theta'\right|^\alpha \wedge 1 \quad \text{for all } \theta, \theta' \in \mathbb{R}. \tag{28}$$

These facts are proven in Baraud (2021) [Examples 5 and 6].

Since Assumption 3 holds true, we may apply Theorem 2 and the reader can check that the constants $c = 0.3$, $\beta = 0.1$ and $\gamma = 0.01$ satisfy the requirements of this theorem.

To estimate the location parameter θ, we choose a prior $\nu_\sigma = \sigma^{-1} q(\cdot/\sigma)\cdot\mu$ associated to a density q that satisfies the requirements of Assumption 4 so that, by (28), this assumption holds true with $L = 1$, $G : r \to r^{1/\alpha}$ and

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

We prove in Section 7.2 the following result.

Proposition 3. *Let t_0 be the third quartile of ν_1. If the density q is positive, symmetric and decreasing on $[0, +\infty)$, for all $\theta \in \mathbb{R}$,*

$$r_n(\beta, P_\theta) \leqslant \frac{2000}{n} \max\left\{\log\left(\overline{\Gamma}\,(\sigma \vee 1)\right) - \log\left(q\left[2\left(\frac{|\theta|}{\sigma} \vee t_0\right)\right]\right), \log 4\right\}.$$

By applying Theorem 2, we conclude that for all $\xi > 0$, with a probability at least $1 - 2e^{-\xi/2}$, the posterior distribution satisfies

$$\widehat{\pi}_{\boldsymbol{X}}\left(\mathscr{B}(\mathbf{P}^\star, \kappa_0' r)\right) \geqslant 1 - e^{-\xi/2}$$

for some some $\kappa_0' \geqslant 2$ only depending on $\xi, \alpha, q(0)$ and for

$$r = \inf_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^n \|P_i^\star - P_\theta\| + \frac{1}{n}\left[1 - \log\left(q\left[2\left(\frac{|\theta|}{\sigma} \vee t_0\right)\right]\right) + \log(\sigma \vee 1)\right]\right].$$

In particular, when the data are i.i.d. with distribution $P_{\theta^\star}$, with probability close to 1, an element $P_{\widehat{\theta}}$ drawn randomly according to the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ satisfies with a probability close to 1,

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

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

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

The parameter is therefore estimated at rate $n^{-1/\alpha}$ which is much faster than the usual $1/\sqrt{n}$-parametric one that is reached by an estimator based on a moment method for instance. Since the densities are unbounded, note that the maximum likelihood estimator does not exist and is therefore useless. It is well-known, mainly from the work of Le Cam, that is impossible to

estimate a distribution in a translation model at a rate faster than $1/n$ for the TV-loss. Because of (28), the rate we get is not only optimal for estimating $P_{\theta^\star}$ but also for estimating $\theta^\star$ with respect to the Euclidean distance. An alternative 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 optimal and robust.

It is also interesting to see how the critical radius r behaves under a misspecification of the prior ν_σ, i.e. when the size of the parameter $\theta^\star$ is large compared to σ. When q is Gaussian, r 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.

5.3. **A general result under entropy.** In this section, $(E,\mathcal{E}) = (\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k))$ that we equip with the Lebesgue measure μ and the norm $|\cdot|_\infty$. We consider the TV-loss ℓ and the location-scale family

$$\mathscr{M} = \left\{ P_{(p,\mathbf{m},\sigma)} = \frac{1}{\sigma^k} p\left(\frac{\cdot - \mathbf{m}}{\sigma}\right) \cdot \mu,\ p \in \mathcal{M}_0,\ \mathbf{m} \in \mathbb{R}^k, \sigma > 0 \right\}, \tag{29}$$

where $\mathcal{M}_0$ is a set of densities on $(\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k), \mu)$ that satisfies the following entropy condition:

Assumption 5. *Let $\widetilde{D}$ be a continuous non-increasing 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*

$$|\mathcal{M}_0[\eta]| \leqslant \exp\left[\widetilde{D}(\eta)\right] \tag{30}$$

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

$$\ell\left(P_{(p,\mathbf{0},1)}, P_{(\overline{p},\mathbf{0},1)}\right) = \frac{1}{2}\int_{\mathbb{R}^k} |p - \overline{p}|\, d\mu \leqslant \eta. \tag{31}$$

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

$$\ell\left(P_{(p,\mathbf{0},1)}, P_{(p,\mathbf{m},\sigma)}\right) \leqslant \left[A\left(\left(\left|\frac{\mathbf{m}}{\sigma}\right|_\infty\right)^\alpha + \left(1 - \frac{1}{\sigma}\right)^\alpha\right)\right] \bigwedge 1. \tag{32}$$

For $\eta, \delta > 0$, we define

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

and for $\theta = \theta(\overline{p}, j_0, \mathbf{j}) \in \Theta[\eta,\delta]$, set

$$L_\theta = (k+1)L + \log|\mathcal{M}_0[\eta]| + 2\sum_{i=0}^{k} \log(1 + |j_i|) \tag{33}$$

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 prior π defined as

$$\pi(\{P_\theta\}) = e^{-L_\theta} \quad \text{for all } \theta \in \Theta[\eta,\delta]. \tag{34}$$

Corollary 1. *Let $\xi > 0$, $K > 1$, ℓ be the TV-loss and $\mathscr{M}$ the family of probabilities given by (29) where $\mathcal{M}_0$ satisfies Assumption 5. Consider the parameters*

$$\eta_n = \inf \mathscr{D}_n \quad \textit{with} \quad \mathscr{D}_n = \left\{\eta > 0,\ \widetilde{D}(\eta) \leqslant \frac{n\eta^2}{24}\right\} \tag{35}$$

$$\delta = \delta_n = \left(\frac{\eta_n}{2A}\right)^{1/\alpha}, \tag{36}$$

$$\beta = \beta_n = \frac{1}{2}\left[K\eta_n + 2\sqrt{\frac{18.6(k+1)}{n}}\right] \tag{37}$$

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

$$|\log\sigma| \vee \left|\frac{\mathbf{m}}{\sigma}\right|_\infty \leqslant \Lambda_n = \exp\left[\frac{(K^2-1)n\eta_n^2}{48(k+1)} + \log\log(1+\delta_n)\right]. \tag{38}$$

Then, the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ associated to the value $\lambda = 4\beta/3$, the prior π given by (34) and the family $\mathscr{T}(\ell,\mathscr{M})$ given by (6) possesses the following property: there exists $\kappa_0 \geqslant 2$ only depending on ξ such that

$$\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\mathbf{P}^\star, \kappa_0 r)\right)\right] \leqslant 2e^{-\xi} \tag{39}$$

with

$$r = \inf_{P\in\mathscr{M}_n(K)} \ell(\mathbf{P}^\star, P) + K\eta_n + \sqrt{\frac{k+1}{n}}. \tag{40}$$

To comment on Condition (32), let us assume that the set $\mathcal{M}_0$ consists of densities p that are supported on $[0,1]^k$, satisfies $\sup_{p\in\mathcal{M}_0}\|p\|_\infty \leqslant L_0$ and

$$\sup_{p\in\mathcal{M}_0} \left|p(\mathbf{x}) - p(\mathbf{x}')\right| \leqslant L_1\left|\mathbf{x}-\mathbf{x}'\right|^\alpha \quad \text{for all } \mathbf{x},\mathbf{x}' \in \mathbb{R}^k \tag{41}$$

for some constants $L_0, L_1 > 0$ and $\alpha \in (0,1]$. For all $p \in \mathcal{M}_0$, $\sigma \geqslant 1$ and $\mathbf{m} \in \mathbb{R}^k$, the supports of the functions $\mathbf{x} \mapsto p(\mathbf{x}/\sigma)$ and $\mathbf{x} \mapsto p((\mathbf{x}-\mathbf{m})/\sigma)$ are included in the set $\mathcal{K} = [0,\sigma]^k \cup \{\mathbf{m}+\mathbf{x},\ \mathbf{x}\in[0,\sigma]^k\}$ the Lebesgue measure of which is not larger than $2\sigma^k$. Consequently, using (41), we deduce that

for all $p \in \mathcal{M}_0$, $\sigma \geqslant 1$ and $\mathbf{m} \in \mathbb{R}^k$,

$$
\begin{aligned}
&\ell\left(P_{(p,\mathbf{0},1)}, P_{(p,\mathbf{m},\sigma)}\right) \\
&\quad \leqslant \ell\left(P_{(p,\mathbf{0},1)}, P_{(p,\mathbf{0},\sigma)}\right) + \ell\left(P_{(p,\mathbf{0},\sigma)}, P_{(p,\mathbf{m},\sigma)}\right) \\
&\quad = \frac{1}{2}\int_{\mathbb{R}^k}\left|p(\mathbf{x}) - \frac{1}{\sigma^k}p\left(\frac{\mathbf{x}}{\sigma}\right)\right| d\mathbf{x} + \frac{1}{2\sigma^k}\int_{\mathbb{R}^k}\left|p\left(\frac{\mathbf{x}}{\sigma}\right) - p\left(\frac{\mathbf{x}-\mathbf{m}}{\sigma}\right)\right| d\mathbf{x} \\
&\quad \leqslant \frac{1}{2}\int_{\mathbb{R}^k}\left|p(\mathbf{x}) - \frac{1}{\sigma^k}p\left(\mathbf{x}\right)\right| d\mathbf{x} + \frac{1}{2\sigma^k}\int_{\mathbb{R}^k}\left|p(\mathbf{x}) - p\left(\frac{\mathbf{x}}{\sigma}\right)\right| d\mathbf{x} \\
&\qquad + \frac{1}{2\sigma^k}\int_{\mathbb{R}^k}\left|p\left(\frac{\mathbf{x}}{\sigma}\right) - p\left(\frac{\mathbf{x}-\mathbf{m}}{\sigma}\right)\right| d\mathbf{x} \\
&\quad \leqslant \frac{1}{2}\int_{\mathbb{R}^k}\left|p(\mathbf{x}) - \frac{1}{\sigma^k}p\left(\mathbf{x}\right)\right| d\mathbf{x} + \frac{1}{2\sigma^k}\int_{[0,1]^k}\left|p(\mathbf{x}) - p\left(\frac{\mathbf{x}}{\sigma}\right)\right| d\mathbf{x} \\
&\qquad + \frac{1}{2\sigma^k}\int_{[0,\sigma]^k\setminus[0,1]^k}\left|p\left(\frac{\mathbf{x}}{\sigma}\right)\right| d\mathbf{x} + \frac{1}{2\sigma^k}\int_{\mathcal{K}}\left|p\left(\frac{\mathbf{x}}{\sigma}\right) - p\left(\frac{\mathbf{x}-\mathbf{m}}{\sigma}\right)\right| d\mathbf{x} \\
&\quad \leqslant \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)^{\alpha}|\mathbf{x}|^{\alpha}\, d\mathbf{x} \\
&\qquad + \frac{1}{2}\int_{[0,1]^k\setminus[0,1/\sigma]^k}|p(\mathbf{x})|\, d\mathbf{x} + \frac{L_1}{2\sigma^k}\int_{\mathcal{K}}\left|\frac{\mathbf{m}}{\sigma}\right|^{\alpha} d\mathbf{x} \\
&\quad \leqslant \frac{1}{2}\left(1 - \frac{1}{\sigma^k}\right) + \frac{L_1 k^{\alpha/2}}{2\sigma^k}\left(1-\frac{1}{\sigma}\right)^{\alpha} + \frac{L_0}{2}\left(1-\frac{1}{\sigma^k}\right) + L_1\left|\frac{\mathbf{m}}{\sigma}\right|^{\alpha} \\
&\quad \leqslant \frac{1}{2}\left[1 + L_1 k^{\alpha/2} + L_0\right]\left(1-\frac{1}{\sigma}\right)^{\alpha} + L_1\left|\frac{\mathbf{m}}{\sigma}\right|^{\alpha}
\end{aligned}
$$

and (32) is therefore satisfied with $A = L_1 \vee [(1 + L_1 k^{\alpha/2} + L_0)/2]$.

In Lemma 1 below, we show that (32) may also be satisfied in a situation where the densities in $\mathcal{M}_0$ are irregular and possibly discontinuous. It makes it possible to consider the following example.

Example 3. We consider here the situation where $k = 1$ and $\mathcal{M}_0$ is the set of all non-increasing densities on $[0, 1]$ that are bounded by $B > 1$. Then, $\mathscr{M}$ consists of all the probabilities whose densities are non-increasing and supported on intervals I with positive lengths and which are bounded by $B/\mu(I)$. Birman and Solomjak (1967) proved that $\mathcal{M}_0$ satisfies Assumption 5 with $\widetilde{D}(\eta)$ of order $(1/\eta) \vee 1$ (up to some constant that depends on B) and consequently η_n is of order $n^{-1/3}$. Besides, it follows from Lemma 1 below that (32) is satisfied with $A = B$ and $\alpha = 1$. We may therefore apply Corollary 1. For a value of K large enough compared to 1, Λ_n defined by (38) is larger than $\exp\left[CK^2 n^{1/3}\right]$ 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 $p \in \mathcal{M}_0$, $|m^\star/\sigma^\star| \leqslant \exp\left[CK^2 n^{1/3}\right]$ and

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

the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ satisfies for all $\xi > 0$, with a probability at least $1 - 2e^{-\xi/2}$, $\widehat{\pi}_{\boldsymbol{X}}\left[\mathscr{B}(P^\star, C' n^{-1/3})\right] \geqslant 1 - e^{-\xi/2}$ 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 over a huge range of translation and scale parameters when n is large enough.

Lemma 1. *Let p be a non-increasing density on $(0, +\infty)$. For all $\sigma \geqslant 1$*

$$\frac{1}{2}\int_{\mathbb{R}} \left|\frac{1}{\sigma} p\left(\frac{x}{\sigma}\right) - p(x)\right| dx \leqslant \left(1 - \frac{1}{\sigma}\right). \tag{42}$$

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

$$\frac{1}{2}\int_{\mathbb{R}} |p(x) - p(x-m)|\, dx \leqslant (|m| B) \wedge 1. \tag{43}$$

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

$$\frac{1}{2}\int_{\mathbb{R}} \left|\frac{1}{\sigma} p\left(\frac{x-m}{\sigma}\right) - p(x)\right| dx \leqslant \left[B\left|\frac{m}{\sigma}\right| + \left(1 - \frac{1}{\sigma}\right)\right] \wedge 1. \tag{44}$$

5.4. **Estimating a parameter under sparcity.** Let us consider a family of distributions $\mathscr{M} = \left\{P_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}} \cdot \mu,\ \boldsymbol{\theta} \in \mathbb{R}^k\right\}$ that are parametrized by $\mathbb{R}^k$ where the dimension k is large. We presume, even though this might not be true, that the data are i.i.d. with a distribution $P_{\boldsymbol{\theta}^\star} \in \mathscr{M}$ associated to a parameter $\boldsymbol{\theta}^\star$ the coordinates of which are all zero except, maybe, a small number of these. For $m \subset \{1, \ldots, k\}$, $m \neq \varnothing$, we introduce the sub-family $\mathscr{M}_m$ that gathers the 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 $\mathscr{M}_m = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_m\}$ for all $m \subset \{1, \ldots, k\}$, $m \neq \varnothing$.

Throughout this section we consider the squared Hellinger loss and, given some $R > 0$, we assume that there exist $\alpha \in (0, 1]$ and a positive number $B_k = B_k(R)$ possibly depending on k and R, although we do not specify the dependency with respect to R, such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^k$ with $|\boldsymbol{\theta}|_\infty \vee |\boldsymbol{\theta}'|_\infty \leqslant R$

$$h\left(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}'}\right) \leqslant \sqrt{B_k}\, |\boldsymbol{\theta} - \boldsymbol{\theta}'|_\infty^\alpha. \tag{45}$$

As a consequence, the mapping $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$ is continuous. We endow $\mathscr{M}$ with the Borel σ-algebra $\mathcal{A}$ associated to the Hellinger distance so that, for all probabilities P on $(E, \mathcal{E})$, the mapping $Q \mapsto \ell(P, Q) = h^2(P, Q)$ is continuous, hence measurable, on $(\mathscr{M}, \mathcal{A})$, and Assumption 1 is satisfied with $\tau = 2$.

Given a nonempty subset m of $\{1, \ldots, p\}$, we endow Θ_m with the uniform distribution ν_m on the cube $\Theta_m(R) = \{\boldsymbol{\theta} \in \Theta_m,\ |\boldsymbol{\theta}|_\infty \leqslant R\}$. This leads

to a prior $\pi_m = \pi_m(R)$ on $\mathscr{M}_m$ defined as the image of ν_m by the mapping $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$. For $m = \varnothing$, we set $\Theta_\varnothing = \Theta_\varnothing(R) = \{\mathbf{0}\}$ and we endow it with the Dirac mass at $\mathbf{0}$ so that $\mathscr{M}_\varnothing = \{P_{\mathbf{0}}\}$ and $\pi_\varnothing$ is the Dirac mass at $P_{\mathbf{0}}$. We finally define our prior $\pi = \pi(R)$ on $(\mathscr{M}, \mathcal{A})$ as

$$\pi = \sum_{m \subset \{1,\dots,k\}} e^{-L_m} \pi_m \quad \text{with} \quad L_m = |m| \log k + k \log\left(1 + \frac{1}{k}\right). \tag{46}$$

It is not difficult to check that $\sum_{m \subset \{1,\dots,k\}} e^{-L_m} = 1$, hence that π is a genuine probability on $(\mathscr{M}, \mathcal{A})$.

The following result holds.

Corollary 2. *Let $\ell = h^2$ be the squared Hellinger distance, $\mathscr{M} = \{P_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}} \cdot \mu,\ \boldsymbol{\theta} \in \mathbb{R}^k\}$ a dominated statistical model satisfying (45) and the assumption that the mapping*

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

is measurable. We assume that $RB_k^{1/(2\alpha)} \geqslant 1$ and endow $\mathscr{M}$ with the prior $\pi = \pi(R)$ defined by (46) and define the posterior distribution $\widehat{\pi}_{\boldsymbol{X}}$ by (11) with $\beta = 0.01$, $\lambda = 1.05\beta$ and the family $\mathscr{T}(\ell, \mathscr{M})$ given by (7). Then there exists $\kappa_0 \geqslant 2$, only depending on the value of $\xi > 0$, such that

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

where

$$r = \inf_{m \subset \{1,\dots,k\}} \left[\inf_{\boldsymbol{\theta} \in \Theta_m(R)} \ell(\mathbf{P}^\star, P_{\boldsymbol{\theta}}) + \frac{|m| \log\left(2kR(nB_k)^{1/(2\alpha)}\right) + 1}{n} \right]. \tag{47}$$

Let us now comment on and illustrate this result.

When B_k does not increase faster than a power of k, r only depends logarithmically on the dimension k, as expected.

Since the mapping

$$R \mapsto \sup\left\{ \frac{h\left(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}'}\right)}{\left|\boldsymbol{\theta} - \boldsymbol{\theta}'\right|_\infty^\alpha} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^k, \boldsymbol{\theta} \neq \boldsymbol{\theta}',\ |\boldsymbol{\theta}|_\infty \vee \left|\boldsymbol{\theta}'\right|_\infty \leqslant R \right\}$$

is non-decreasing, our condition $RB_k^{1/(2\alpha)} \geqslant 1$ holds for R large enough.

For illustration, assume that $P_{\boldsymbol{\theta}}$ 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. Then,

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

and we obtain that (45) is satisfied with $B_k = k/(8\sigma^2)$ and $\alpha = 1$. In particular, our condition $RB_k^{1/(2\alpha)} \geqslant 1$ is equivalent to $R \geqslant 2\sigma\sqrt{(2/k)}$. In this case, we obtain that the value of r given by (47) is of order

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

More generally, when $\mathscr{M} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \mathbb{R}^k\}$ is a regular statistical model with Fisher information $\mathbf{I}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$, we know from the book of Ibragimov and Has'minskiĭ (1981)[Theorem 7.1 p.81] that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^k$ such that $|\boldsymbol{\theta}|_\infty \vee |\boldsymbol{\theta}'|_\infty \leqslant R$

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

Then, Assumption (45) holds with $\alpha = 1$ and

$$B_k = \frac{k}{2\sqrt{2}}\sup_{\boldsymbol{\theta}''\in\mathbb{R}^k,|\boldsymbol{\theta}''|\leqslant R}\sqrt{\varrho\left(\mathbf{I}(\boldsymbol{\theta}'')\right)}$$

where $\varrho\left(\mathbf{I}(\boldsymbol{\theta}'')\right)$ denotes the largest eigenvalue of the matrix $\mathbf{I}(\boldsymbol{\theta}'')$. It is well known that this value is independent of $\boldsymbol{\theta}''$ when $\mathscr{M}$ is a translation model.

6. Proofs of Theorems 1 and 2

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

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

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

6.1. Preliminary results.

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

Lemma 2. *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 $(E\times F, \mathcal{E}\otimes\mathcal{F})$,*

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

This lemma is proven in Audibert and Catoni (2011) [Lemma 4.2, P. 28].

Lemma 3. *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 $r \in \mathcal{V}(\pi, \overline{P})$ *and* $P \in \mathscr{M}$*,*

$$\mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] \leqslant \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}} \exp\left[-\mathbf{M}(P,Q)\right] d\pi_{\mathscr{B}}(Q)\right]^{-1}. \tag{48}$$

Proof. Let $r \in \mathcal{V}(\pi, \overline{P})$. 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,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[-I(\boldsymbol{X}, P, Q)\right]\right] \\
&= \mathbb{E}\left[\exp\left[-c_1\beta \mathbf{T}(\boldsymbol{X}, P, Q) + \log \int_{\mathscr{M}} \exp\left[c\beta \mathbf{T}(\boldsymbol{X}, P, Q')\right] d\pi(Q')\right]\right] \\
&= \mathbb{E}\left[\int_{\mathscr{M}} \exp\left[c\beta \mathbf{T}(\boldsymbol{X}, P, Q') - c_1\beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q')\right] \\
&= \exp\left[\mathbf{M}(P,Q)\right]. \qquad (49)
\end{aligned}$$

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

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

Hence,

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

Applying Lemma 2 with $U = \boldsymbol{X}$, $V = Q$ with distribution $\pi_{\mathscr{B}}$, and $h(U,V) = -I(\boldsymbol{X}, P, Q)$, we obtain that

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

and (48) follows from (49). □

Lemma 4. *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 $r \in \mathcal{V}(\pi, \overline{P})$,

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

Proof. Let $r \in \mathcal{V}(\pi, \overline{P})$. 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,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[-H(\boldsymbol{X}, P, Q)\right]\right] \\
&= \mathbb{E}\left[\exp\left[-\beta c_1 \mathbf{T}(\boldsymbol{X}, P, Q)\right] \int_{\mathscr{M}} \exp\left[c_2 \beta \mathbf{T}(\boldsymbol{X}, P, Q')\right] d\pi(Q')\right] \\
&= \mathbb{E}\left[\int_{\mathscr{M}} \exp\left[\beta\left(c_2 \mathbf{T}(\boldsymbol{X}, P, Q') - c_1 \mathbf{T}(\boldsymbol{X}, P, Q)\right)\right] d\pi(Q')\right] \\
&= \exp\left[\mathbf{L}(P, Q)\right]. \qquad (50)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[\exp\left[\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] &= \mathbb{E}\left[\exp\left[\int_{\mathscr{M}} \left[\beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\widetilde{\pi}_{\boldsymbol{X}}(Q)\right]\right] \\
&\leqslant \mathbb{E}\left[\int_{\mathscr{M}} \exp\left[\beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\widetilde{\pi}_{\boldsymbol{X}}(Q)\right] \\
&= \mathbb{E}\left[\frac{\int_{\mathscr{M}} \exp\left[c_2 \beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}{\int_{\mathscr{M}} \exp\left[c_1 \beta \mathbf{T}(\boldsymbol{X}, P, Q)\right] d\pi(Q)}\right] \\
&= \mathbb{E}\left[\frac{1}{\int_{\mathscr{M}} \exp\left[H(\boldsymbol{X}, P, Q)\right] d\pi(Q)}\right].
\end{aligned}$$

Applying Lemma 2 with $U = \boldsymbol{X}$ and $V = Q$ with distribution π we obtain that

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

We deduce from (50) that for all $P \in \mathscr{M}$

$$\begin{aligned}
\mathbb{E}\left[\exp\left[\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] &\leqslant \left[\int_{\mathscr{M}} \exp\left[-\mathbf{L}(P, Q)\right] d\pi(Q)\right]^{-1} \\
&\leqslant \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}} \exp\left[-\mathbf{L}(P, Q)\right] d\pi_{\mathscr{B}}(Q)\right]^{-1}. \qquad (51)
\end{aligned}$$

Applying Lemma 2 with $U = \boldsymbol{X}$, $V = P$ with distribution π and $h(U,V) = \beta \mathbf{T}(\boldsymbol{X}, P)$, gives

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

which together with (51) 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 5. *For all $\beta \in \mathbb{R}$ and random variable U with values in an interval of length $l \in (0, +\infty)$,*

$$\log \mathbb{E}\left[\exp\left[\beta U\right]\right] \leqslant \beta \mathbb{E}\left[U\right] + \frac{\beta^2 l^2}{8}. \tag{52}$$

Lemma 6. *Let U be a squared integrable random variable not larger than $b > 0$. For all $\beta > 0$,*

$$\log \mathbb{E}\left[\exp\left[\beta U\right]\right] \leqslant \beta \mathbb{E}\left[U\right] + \beta^2 \mathbb{E}\left[U^2\right] \phi(\beta b), \tag{53}$$

where ϕ is defined by (20).

The proofs of Lemmas 5 and 6 can be found on pages 21 and 23 in Massart (2007).

Lemma 7. *Let $J \in \mathbb{N}$, $\gamma > 0$ and $\overline{P} \in \mathscr{M}$. If r satisfies $n\beta a_1 r \geqslant 1$ and (13), for all $\gamma_0 > 2\gamma$*

$$\int_{\complement \mathscr{B}(\overline{P}, 2^J r)} \exp\left[-\gamma_0 n\beta a_1 \ell(\overline{P}, P)\right] d\pi(P)$$
$$\leqslant \pi\left(\mathscr{B}\right) \exp\left[\Xi - \left(\gamma_0 - 2\gamma\right) n\beta a_1 2^J r\right] \tag{54}$$

with

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

Besides,

$$\int_{\mathscr{M}} \exp\left[-\gamma_0 n\beta a_1 \ell(\overline{P}, P)\right] d\pi(P) \leqslant \pi(\mathscr{B}) \exp\left[\Xi'\right] \tag{55}$$

with

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

Proof. From (13), we deduce by induction that for all $j \geqslant 0$

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

Consequently,

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

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

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

which leads to (54) since $n\beta a_1 2^J r \geqslant n\beta a_1 r \geqslant 1$. Finally, by applying this inequality with $J = 0$ we obtain that

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

which is (55). □

6.2. **Main parts of the proofs of Theorems 1 and 2.** Throughout the proofs of these two theorems, we fix some arbitrary element $\overline{P} \in \mathscr{M}$ and $r \geqslant r_n(\beta, \overline{P})$. It follows from the definition of $r_n(\beta, \overline{P})$ that r satisfies both $n\beta a_1 r \geqslant 1$ and inequality (13). For a positive number z, that will be chosen later as well, we 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 (11) of $\widehat{\pi}_{\boldsymbol{X}}$ that, given $J \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{P}, 2^J r)\right)\right] &= \mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{P}, 2^J r)\right) \mathbb{1}_{{}^{\mathsf{c}}A}\right] + \mathbb{E}\left[\widehat{\pi}_{\boldsymbol{X}}\left({}^{\mathsf{c}}\mathscr{B}(\overline{P}, 2^J r)\right) \mathbb{1}_{A}\right] \\
&\leqslant \mathbb{P}({}^{\mathsf{c}}A) + \frac{1}{z}\mathbb{E}\left[\int_{{}^{\mathsf{c}}\mathscr{B}(\overline{P}, 2^J r)} \exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right] d\pi(P)\right] \\
&= \mathbb{P}({}^{\mathsf{c}}A) + \frac{1}{z}\int_{{}^{\mathsf{c}}\mathscr{B}(\overline{P}, 2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X}, P)\right]\right] d\pi(P). \qquad (56)
\end{aligned}$$

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 (56) is not larger than $e^{-\xi}$. To achieve this goal, we bound the first term of the right-hand side of (56) by first applying Markov's inequality

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

and then by using Lemma 4. This leads to

$$\mathbb{P}({}^{\mathsf{c}}A) \leqslant \frac{z}{\pi^2(\mathscr{B})}\left[\int_{\mathscr{B}^2} \exp\left[-\mathbf{L}(P, Q)\right] d\pi_{\mathscr{B}}(P) d\pi_{\mathscr{B}}(Q)\right]^{-1}. \qquad (57)$$

To show that the first term in the right hand-side of (56) is not larger than $e^{-\xi}$ we therefore prove that this is the case of the right-hand side of (57) for z small enough. We bound the second term of (56) by using Lemma 3.

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 $\overline{P} \in \mathscr{M}(\beta)$,

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

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

$$\overline{r} = \overline{r}(\overline{P}) = \tau\left[\ell(\mathbf{P}^\star, \overline{P}) + 2^J r\right] = \tau\left[(1 + 2^J)\ell(\mathbf{P}^\star, \overline{P}) + 2^J a_1^{-1}\beta\right].$$

We obtain (17) by monotone convergence, taking a sequence $(\overline{P}_N)_{N\geqslant 0} \subset \mathscr{M}(\beta)$ such that $\ell(\mathbf{P}^\star, \overline{P}_N)$ is non-increasing to $\inf_{P\in\mathscr{M}(\beta)} \ell(\mathbf{P}^\star, P)$, so that

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

and (17) holds provided that $\kappa_0 \geqslant \tau(2^J+1)$.

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

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

and we get (24) by arguing similarly.

6.3. **Proof of Theorem 1.** For all $i\in\{1,\dots,n\}$ and $P,Q,Q'\in\mathscr{M}$, let us set

$$\begin{aligned}U_i = c&\left(t_{(P,Q')}(X_i) - \mathbb{E}\left[t_{(P,Q')}(X_i)\right]\right)\\ &- c_1\left(t_{(P,Q)}(X_i) - \mathbb{E}\left[t_{(P,Q)}(X_i)\right]\right) \qquad (58)\\ V_i = c_2&\left(t_{(P,Q')}(X_i) - \mathbb{E}\left[t_{(P,Q')}(X_i)\right]\right)\\ &- c_1\left(t_{(P,Q)}(X_i) - \mathbb{E}\left[t_{(P,Q)}(X_i)\right]\right). \qquad (59)\end{aligned}$$

The random variables U_i are independent and under Assumption 2-(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 5, we obtain that

$$\prod_{i=1}^n \mathbb{E}\left[\exp\left[\beta U_i\right]\right] \leqslant \exp\left[\frac{l_1^2 n\beta^2}{8}\right] \qquad (60)$$

and

$$\prod_{i=1}^n \mathbb{E}\left[\exp\left[\beta V_i\right]\right] \leqslant \exp\left[\frac{l_2^2 n\beta^2}{8}\right]. \qquad (61)$$

By using Assumption 1 and the fact that $c_0 = c_1 - ca_0/a_1 > 0$,

$$\begin{aligned}&c\left(a_0\ell(\mathbf{P}^\star,P) - a_1\ell(\mathbf{P}^\star,Q')\right) - c_1\left(a_1\ell(\mathbf{P}^\star,P) - a_0\ell(\mathbf{P}^\star,Q)\right)\\ &= -\left(c_1a_1 - ca_0\right)\ell(\mathbf{P}^\star,P) - ca_1\ell(\mathbf{P}^\star,Q') + c_1a_0\ell(\mathbf{P}^\star,Q)\\ &\leqslant -c_0a_1\left[\tau^{-1}\ell(\overline{P},P) - \ell(\mathbf{P}^\star,\overline{P})\right] - ca_1\left[\tau^{-1}\ell(\overline{P},Q') - \ell(\mathbf{P}^\star,\overline{P})\right]\\ &\quad + \tau c_1a_0\left[\ell(\mathbf{P}^\star,\overline{P}) + \ell(\overline{P},Q)\right]\\ &= e_0a_1\ell(\mathbf{P}^\star,\overline{P}) - \tau^{-1}c_0a_1\ell(\overline{P},P) - \tau^{-1}ca_1\ell(\overline{P},Q') + \tau c_1a_0\ell(\overline{P},Q) \qquad (62)\end{aligned}$$

with

$$e_0 = c_0 + c + \frac{\tau c_1 a_0}{a_1}. \tag{63}$$

It follows from Assumptions 2-(iii) and (62) that

$$\begin{aligned}
&n^{-1}\left\{c\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q')\right] - c_1\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q)\right]\right\}\\
&\leqslant c\left[a_0\ell(\mathbf{P}^\star,P) - a_1\ell(\mathbf{P}^\star,Q')\right] - c_1\left[a_1\ell(\mathbf{P}^\star,P) - a_0\ell(\mathbf{P}^\star,Q)\right]\\
&\leqslant e_0a_1\ell(\mathbf{P}^\star,\overline{P}) - \tau^{-1}c_0a_1\ell(\overline{P},P) - \tau^{-1}ca_1\ell(\overline{P},Q') + \tau c_1a_0\ell(\overline{P},Q). \qquad (64)
\end{aligned}$$

Since $a_0 \geqslant a_1$ 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

$$\begin{aligned}
&n^{-1}\left\{c_2\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q')\right] - c_1\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q)\right]\right\}\\
&\leqslant c_2\left(a_0\ell(\mathbf{P}^\star,P) - a_1\ell(\mathbf{P}^\star,Q')\right) - c_1\left(a_1\ell(\mathbf{P}^\star,P) - a_0\ell(\mathbf{P}^\star,Q)\right)\\
&= c_0'a_1\ell(\mathbf{P}^\star,P) - c_2a_1\ell(\mathbf{P}^\star,Q') + c_1a_0\ell(\mathbf{P}^\star,Q)\\
&\leqslant \tau c_0'a_1\left[\ell(\mathbf{P}^\star,\overline{P}) + \ell(\overline{P},P)\right] - c_2a_1\left[\tau^{-1}\ell(\overline{P},Q') - \ell(\mathbf{P}^\star,\overline{P})\right]\\
&\quad + \tau c_1a_0\left[\ell(\mathbf{P}^\star,\overline{P}) + \ell(\overline{P},Q)\right]\\
&\leqslant (e_1+c_2)\,a_1\ell(\mathbf{P}^\star,\overline{P}) + \tau c_0'a_1\ell(\overline{P},P)\\
&\quad - \tau^{-1}c_2a_1\ell(\overline{P},Q') + \tau c_1a_0\ell(\overline{P},Q), \qquad (65)
\end{aligned}$$

with

$$e_1 = \tau\left[c_0' + c_1a_0/a_1\right] = \tau\left[c_2(a_0/a_1) + \ c_1\left(a_0/a_1 - 1\right)\right]. \tag{66}$$

Using (60) and (64), we deduce that for all $P, Q, Q' \in \mathscr{M}$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[\beta\left(c\mathbf{T}(\boldsymbol{X},P,Q') - c_1\mathbf{T}(\boldsymbol{X},P,Q)\right)\right]\right]\\
&= \prod_{i=1}^n \mathbb{E}\left[\exp\left[\beta\left(ct_{(P,Q')}(X_i) - c_1t_{(P,Q)}(X_i)\right)\right]\right]\\
&= \exp\left[\beta\left(c\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q')\right] - c_1\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q)\right]\right)\right]\prod_{i=1}^n \mathbb{E}\left[\exp\left[\beta U_i\right]\right]\\
&\leqslant \exp\left[n\beta\left[\Delta_1(P,Q) - \tau^{-1}ca_1\ell(\overline{P},Q')\right]\right] \qquad (67)
\end{aligned}$$

with

$$\Delta_1(P,Q) = e_0a_1\ell(\mathbf{P}^\star,\overline{P}) + \tau c_1a_0\ell(\overline{P},Q) + \frac{l_1^2\beta}{8} - \tau^{-1}c_0a_1\ell(\overline{P},P). \tag{68}$$

Using (61) and (65), we obtain similarly that for all $P, Q, Q' \in \mathscr{M}$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[\beta\left(c_2\mathbf{T}(\boldsymbol{X},P,Q') - c_1\mathbf{T}(\boldsymbol{X},P,Q)\right)\right]\right]\\
&\leqslant \exp\left[n\beta\left[\Delta_2(P,Q) - \tau^{-1}c_2a_1\ell(\overline{P},Q')\right]\right] \qquad (69)
\end{aligned}$$

with

$$\Delta_2(P,Q) = (e_1+c_2)\,a_1\ell(\mathbf{P}^\star,\overline{P}) + \tau c_0' a_1 \ell(\overline{P},P) + \tau c_1 a_0 \ell(\overline{P},Q) + \frac{l_2^2\beta}{8}. \tag{70}$$

Since $2\gamma < \tau^{-1}c < \tau^{-1}c_2$, we may apply Lemma 7 with $\gamma_0 = \tau^{-1}c$ and $\gamma_0 = \tau^{-1}c_2$ successively which leads to

$$\int_{\mathscr{M}} \exp\left[-\tau^{-1}cn\beta a_1\ell(\overline{P},Q')\right] d\pi(Q') \leqslant \pi\left(\mathscr{B}\right)\exp\left[\Xi_1\right] \tag{71}$$

and

$$\int_{\mathscr{M}} \exp\left[-\tau^{-1}c_2n\beta a_1\ell(\overline{P},Q')\right] d\pi(Q') \leqslant \pi\left(\mathscr{B}\right)\exp\left[\Xi_1\right] \tag{72}$$

with

$$\begin{aligned}\Xi_1 &= \log\left[1+\frac{\exp\left[-\left(\tau^{-1}c-\gamma\right)\right]}{1-\exp\left[-\left(\tau^{-1}c-2\gamma\right)\right]}\right] \\ &\geqslant \log\left[1+\frac{\exp\left[-\left(\tau^{-1}c_2-\gamma\right)\right]}{1-\exp\left[-\left(\tau^{-1}c_2-2\gamma\right)\right]}\right].\end{aligned} \tag{73}$$

Putting (69) and (72) together leads to

$$\begin{aligned}\exp\left[\mathbf{L}(P,Q)\right] &= \int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c_2\mathbf{T}(\boldsymbol{X},P,Q') - c_1\mathbf{T}(\boldsymbol{X},P,Q)\right)\right]\right] d\pi(Q') \\ &\leqslant \exp\left[n\beta\Delta_2(P,Q)\right]\int_{\mathscr{M}}\exp\left[-\tau^{-1}c_2n\beta a_1\ell(\overline{P},Q')\right]d\pi(Q') \\ &\leqslant \pi\left(\mathscr{B}\right)\exp\left[\Xi_1+n\beta\Delta_2(P,Q)\right],\end{aligned}$$

and since, for all $(P,Q)\in\mathscr{B}^2$, by definition (70) of $\Delta_2(P,Q)$,

$$\begin{aligned}\Delta_2(P,Q) &\leqslant (e_1+c_2)\,a_1\ell(\mathbf{P}^\star,\overline{P}) + \left[\tau c_0' a_1 + \tau c_1 a_0\right] r + \frac{l_2^2\beta}{8} \\ &= (e_1+c_2)\,a_1\ell(\mathbf{P}^\star,\overline{P}) + e_1a_1r + \frac{l_2^2\beta}{8} = \Delta_2\end{aligned} \tag{74}$$

we derive that

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

We deduce from (57) that

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

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

$$\log\left(\frac{1}{z}\right) = \xi + \log\frac{1}{\pi(\mathscr{B})} + \Xi_1 + n\beta\Delta_2. \tag{75}$$

Putting (67) and (71) together, we obtain that

$$\begin{aligned}
&\exp\left[\mathbf{M}(P,Q)\right] \\
&= \int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c\mathbf{T}(\boldsymbol{X},P,Q') - c_1\mathbf{T}(\boldsymbol{X},P,Q)\right)\right]\right] d\pi(Q') \\
&\leqslant \exp\left[n\beta\Delta_1(P,Q)\right] \int_{\mathscr{M}} \exp\left[-\tau^{-1}cn\beta a_1 \ell(\overline{P},Q')\right] d\pi(Q') \\
&\leqslant \pi\left(\mathscr{B}\right) \exp\left[\Xi_1 + n\beta\Delta_1(P,Q)\right].
\end{aligned}$$

It follows from the definition (68) of $\Delta_1(P,Q)$ that for all $P \in \mathscr{M}$ and for all $Q \in \mathscr{B}$,

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

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

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

We derive from Lemma 3 that

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

hence,

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

Applying Lemma 7 with $\gamma_0 = \tau^{-1}c_0 > 2\gamma$ and setting $e_2 = \tau^{-1}c_0 - 2\gamma$, we get

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

with

$$(77) \qquad \Xi_2 = -\gamma + \log\left[\frac{1}{1 - \exp\left[-e_2\right]}\right],$$

which together with (76) leads to

$$\begin{aligned}
&\log \int_{\mathscr{C}_{\mathscr{B}}(\overline{P},2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X},P)\right]\right] d\pi(P) \\
&\leqslant \log\left[\pi\left(\mathscr{B}\right)\right] + \Xi_1 + \Xi_2 \\
&\qquad + n\beta\left[e_0 a_1 \ell(\mathbf{P}^\star,\overline{P}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - e_2 a_1 2^J r\right]. \qquad (78)
\end{aligned}$$

Using the definitions (75) of z and (74) of Δ_2 we deduce from (78) that

$$\begin{aligned}
&\log\left[\frac{1}{z}\int_{\mathscr{C}_{\mathscr{B}}(\overline{P},2^J r)} \mathbb{E}\left[\exp\left[-\beta \mathbf{T}(\boldsymbol{X},P)\right]\right] d\pi(P)\right] \\
&\leqslant \log\left(\frac{1}{z}\right) + \log\left[\pi\left(\mathscr{B}\right)\right] + \Xi_1 + \Xi_2 \\
&\quad + n\beta\left[e_0 a_1 \ell(\mathbf{P}^\star,\overline{P}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - e_2 a_1 2^J r\right] \\
&= \xi + \log\frac{1}{\pi(\mathscr{B})} + \Xi_1 + n\beta\Delta_2 + \log\left[\pi\left(\mathscr{B}\right)\right] + \Xi_1 + \Xi_2 \\
&\quad + n\beta\left[e_0 a_1 \ell(\mathbf{P}^\star,\overline{P}) + \tau c_1 a_0 r + \frac{l_1^2\beta}{8} - e_2 a_1 2^J r\right] \\
&= n\beta\left[\left(e_1 + c_2 + e_0\right) a_1 \ell(\mathbf{P}^\star,\overline{P}) + e_1 a_1 r + \frac{l_2^2\beta}{8} + + \tau c_1 a_0 r + \frac{l_1^2\beta}{8}\right] \\
&\quad + \xi + 2\Xi_1 + \Xi_2 - e_2 n\beta a_1 2^J r \\
&= n\beta\left[\left(e_0 + e_1 + c_2\right) a_1 \ell(\mathbf{P}^\star,\overline{P}) + \left[e_1 + \frac{\tau c_1 a_0}{a_1}\right] a_1 r + \frac{(l_1^2 + l_2^2)\beta}{8}\right] \\
&\quad + \xi + 2\Xi_1 + \Xi_2 - e_2 n\beta a_1 2^J r. \qquad (79)
\end{aligned}$$

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 (79) is not larger than $-\xi$, provided that

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

or equivalently if

$$2^J \geqslant \frac{1}{e_2}\left[\frac{2\xi + 2\Xi_1 + \Xi_2}{\beta n a_1 r} + \frac{C_1\ell(\mathbf{P}^\star,\overline{P}) + C_2 r}{r} + \frac{\left[l_1^2 + l_2^2\right]\beta}{8 a_1 r}\right]. \qquad (80)$$

For $\overline{P} \in \mathscr{M}(\beta)$ and the choice $r = \ell(\mathbf{P}^\star,\overline{P}) + a_1^{-1}\beta \geqslant r_n(\beta,\overline{P}) \geqslant 1/(n a_1 \beta)$, (80) is satisfied if

$$2^J \geqslant \frac{1}{e_2}\left(2\xi + 2\Xi_1 + \Xi_2 + C_1 + C_2 + \frac{\left[l_1^2 + l_2^2\right]}{8}\right)$$

and the requirement $\kappa_0 \geqslant \tau(2^J+1)$ if

$$\kappa_0 = \tau\left[1+\frac{2}{e_2}\left(2\xi+2\Xi_1+\Xi_2+C_1+C_2+\frac{\left[l_1^2+l_2^2\right]}{8}\right)\right]. \tag{81}$$

6.4. **Proof of Theorem 2.** The proof follows the same lines as that of Theorem 1. Under Assumption 2-(iv), the random variables U_i and V_i defined by (58) and (59) are not larger than with $b = c+c_1 = l_1$ and $b = c_2+c_1 = l_2$ respectively. Since under Assumption 3, for all $i \in \{1,\dots,n\}$

$$\begin{aligned}\mathbb{E}\left[U_i^2\right] &\leqslant 2\left[c^2\operatorname{Var}\left[t_{(P,Q')}(X_i)\right]+c_1^2\operatorname{Var}\left[t_{(P,Q)}(X_i)\right]\right]\\ &\leqslant 2a_2\left[(c^2+c_1^2)\ell(P_i^\star,P)+c^2\ell(P_i^\star,Q')+c_1^2\ell(P_i^\star,Q)\right]\end{aligned}$$

and

$$\mathbb{E}\left[V_i^2\right] \leqslant 2a_2\left[(c_2^2+c_1^2)\ell(P_i^\star,P)+c_2^2\ell(P_i^\star,Q')+c_1^2\ell(P_i^\star,Q)\right]$$

we may apply Lemma 6 and using the notations $\Lambda_1 = \tau\phi(\beta l_1)$, $\Lambda_2 = \tau\phi(\beta l_2)$ as well as Assumption 1, we get

$$\begin{aligned}&\frac{1}{n\beta}\log\left[\prod_{i=1}^n\mathbb{E}\left[\exp\left[\beta U_i\right]\right]\right]\\ &\leqslant 2\phi(\beta l_1)\beta a_2\left[(c^2+c_1^2)\ell(\mathbf{P}^\star,P)+c^2\ell(\mathbf{P}^\star,Q')+c_1^2\ell(\mathbf{P}^\star,Q)\right]\\ &\leqslant 2\Lambda_1\beta a_2\left[c^2+c_1^2\right]\ell(\mathbf{P}^\star,\overline{P})\\ &\quad+\Lambda_1\beta a_2\left[(c^2+c_1^2)\ell(\overline{P},P)+c^2\ell(\overline{P},Q')+c_1^2\ell(\overline{P},Q)\right]\end{aligned} \tag{82}$$

and

$$\begin{aligned}&\frac{1}{n\beta}\log\left[\prod_{i=1}^n\mathbb{E}\left[\exp\left[\beta V_i\right]\right]\right]\\ &\leqslant 2\Lambda_2\beta a_2\left[c_2^2+c_1^2\right]\ell(\mathbf{P}^\star,\overline{P})\\ &\quad+\Lambda_2\beta a_2\left[(c_2^2+c_1^2)\ell(\overline{P},P)+c_2^2\ell(\overline{P},Q')+c_1^2\ell(\overline{P},Q)\right].\end{aligned} \tag{83}$$

It follows from (64) that

$$\begin{aligned}E_1 &= n^{-1}\left\{c\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q')\right]-c_1\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},P,Q)\right]\right\}\\ &\quad+2\Lambda_1\beta a_2\left[c^2+c_1^2\right]\ell(\mathbf{P}^\star,\overline{P})\\ &\quad+\Lambda_1\beta a_2\left[(c^2+c_1^2)\ell(\overline{P},P)+c^2\ell(\overline{P},Q')+c_1^2\ell(\overline{P},Q)\right]\\ &= \left[e_0a_1+2\Lambda_1\beta a_2\left(c^2+c_1^2\right)\right]\ell(\mathbf{P}^\star,\overline{P})\\ &\quad-\left[\tau^{-1}c_0a_1-\Lambda_1\beta a_2(c^2+c_1^2)\right]\ell(\overline{P},P)\\ &\quad-\left[\tau^{-1}ca_1-\Lambda_1\beta a_2c^2\right]\ell(\overline{P},Q')\\ &\quad+\left[\tau c_1a_0+\Lambda_1\beta a_2c_1^2\right]\ell(\overline{P},Q).\end{aligned}$$

Using the definitions (21) of $\tau^{-1}c_1$ and (22) of $\tau^{-1}c_2$ and setting

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

and arguing as in the proof of inequality (67), we deduce from (82) that

$$\begin{aligned}&\mathbb{E}\left[\exp\left[\beta\left(c\mathbf{T}(\boldsymbol{X},P,Q') - c_1\mathbf{T}(\boldsymbol{X},P,Q)\right)\right]\right]\\ &\leqslant \exp\left[n\beta E_1\right]\\ &= \exp\left[n\beta a_1\left(e_3\ell(\mathbf{P}^\star,\overline{P}) - \tau^{-1}\left[c_1\ell(\overline{P},P) + c_2\ell(\overline{P},Q')\right] + e_4\ell(\overline{P},Q)\right)\right]\end{aligned} \tag{84}$$

It follows from (65) that

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

Using the definition (23) of $\tau^{-1}c_3$, setting

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

and arguing as in the proof of (69), we deduce from (83) that

$$\begin{aligned}&\mathbb{E}\left[\exp\left[\beta\left(c_2\mathbf{T}(\boldsymbol{X},P,Q') - c_1\mathbf{T}(\boldsymbol{X},P,Q)\right)\right]\right]\\ &\leqslant \exp\left[n\beta E_2\right]\\ &= \exp\left[n\beta a_1\left(e_5\ell(\mathbf{P}^\star,\overline{P}) + e_6\ell(\overline{P},P) - \tau^{-1}c_3\ell(\overline{P},Q') + e_7\ell(\overline{P},Q)\right)\right].\end{aligned} \tag{85}$$

Under our assumption on β, we know that the quantities $\tau^{-1}c_2$ and $\tau^{-1}c_3$ are positive and that $2\gamma < \tau^{-1}\left(c_2 \wedge c_3\right)$. We may therefore apply Lemma 7

with $\gamma_0 = \tau^{-1}c_2$ and $\gamma_0 = \tau^{-1}c_3$ successively and get

$$\int_{\mathscr{M}} \exp\left[-\tau^{-1}c_2 n\beta a_1 \ell(\overline{P}, Q')\right] d\pi(Q') \leqslant \pi\left(\mathscr{B}\right) \exp\left[\overline{\Xi}_1\right] \tag{86}$$

and

$$\int_{\mathscr{M}} \exp\left[-\tau^{-1}c_3 n\beta a_1 \ell(\overline{P}, Q')\right] d\pi(Q') \leqslant \pi\left(\mathscr{B}\right) \exp\left[\overline{\Xi}_1\right] \tag{87}$$

with

$$\overline{\Xi}_1 = \log\left[1 + \frac{\exp\left[-\left(\tau^{-1}(c_2 \wedge c_3) - \gamma\right)\right]}{1 - \exp\left[-\left(\tau^{-1}(c_2 \wedge c_3) - 2\gamma\right)\right]}\right]. \tag{88}$$

Putting (85) and (87) together, we obtain that for all $(P, Q) \in \mathscr{B}^2$

$$\begin{aligned}
&\exp\left[\mathbf{L}(P,Q)\right] \\
&= \int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c_2\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] d\pi(Q') \\
&\leqslant \exp\left[n\beta a_1\left(e_5\ell(\mathbf{P}^\star, \overline{P}) + e_6\ell(\overline{P}, P) + e_7\ell(\overline{P}, Q)\right)\right] \\
&\quad \times \int_{\mathscr{M}} \exp\left[-\tau^{-1}c_3 n\beta a_1 \ell(\overline{P}, Q')\right] d\pi(Q') \\
&\leqslant \pi\left(\mathscr{B}\right) \exp\left[\overline{\Xi}_1 + n\beta a_1\left(e_5\ell(\mathbf{P}^\star, \overline{P}) + e_6\ell(\overline{P}, P) + e_7\ell(\overline{P}, Q)\right)\right] \\
&\leqslant \pi\left(\mathscr{B}\right) \exp\left[\overline{\Xi}_1 + n\beta a_1\left(e_5\ell(\mathbf{P}^\star, \overline{P}) + (e_6 + e_7)r\right)\right].
\end{aligned}$$

Consequently,

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

We deduce from (57) that

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

In particular, $\mathbb{P}(^{\complement}A) \leqslant e^{-\xi}$ for z satisfying

$$\log\left(\frac{1}{z}\right) = \xi + \log\frac{1}{\pi(\mathscr{B})} + \overline{\Xi}_1 + n\beta a_1\left[e_5\ell(\mathbf{P}^\star, \overline{P}) + (e_6 + e_7)r\right]. \tag{89}$$

Putting (84) and (86) together, we obtain that for all $Q \in \mathscr{B}$

$$\begin{aligned}
&\exp\left[\mathbf{M}(P,Q)\right] \\
&= \int_{\mathscr{M}} \mathbb{E}\left[\exp\left[\beta\left(c\mathbf{T}(\boldsymbol{X}, P, Q') - c_1\mathbf{T}(\boldsymbol{X}, P, Q)\right)\right]\right] d\pi(Q') \\
&\leqslant \exp\left[n\beta a_1\left(e_3\ell(\mathbf{P}^\star, \overline{P}) - \tau^{-1}c_1\ell(\overline{P}, P) + e_4\ell(\overline{P}, Q)\right)\right] \\
&\quad \times \int_{\mathscr{M}} \exp\left[-\tau^{-1}c_2 n\beta a_1 \ell(\overline{P}, Q')\right] d\pi(Q') \\
&\leqslant \pi(\mathscr{B}) \exp\left[\overline{\Xi}_1 + n\beta a_1\left(e_3\ell(\mathbf{P}^\star, \overline{P}) + e_4 r - \tau^{-1}c_1\ell(\overline{P}, P)\right)\right].
\end{aligned}$$

We derive from Lemma 3 that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X},P)\right]\right]\\
&\leqslant \frac{1}{\pi(\mathscr{B})}\left[\int_{\mathscr{B}}\exp\left[-\mathbf{M}(P,Q)\right]d\pi_{\mathscr{B}}(Q)\right]^{-1}\\
&\leqslant \exp\left[\overline{\Xi}_1+n\beta a_1\left(e_3\ell(\mathbf{P}^\star,\overline{P})+e_4r-\tau^{-1}c_1\ell(\overline{P},P)\right)\right],
\end{aligned}$$

and consequently,

$$\begin{aligned}
&\int_{\mathrm{c}_{\mathscr{B}}(\overline{P},2^Jr)}\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X},P)\right]\right]d\pi(P)\\
&\leqslant \exp\left[\overline{\Xi}_1+n\beta a_1\left(e_3\ell(\mathbf{P}^\star,\overline{P})+e_4r\right)\right]\\
&\qquad\times\int_{\mathrm{c}_{\mathscr{B}}(\overline{P},2^Jr)}\exp\left[-\tau^{-1}c_1n\beta a_1\ell(\overline{P},P)\right]d\pi(P). \qquad (90)
\end{aligned}$$

Since under our assumptions, $\tau^{-1}c_1>0$ and $2\gamma<\tau^{-1}c_1$ we may apply Lemma 7 with $\gamma_0=\tau^{-1}c_1$, which leads to

$$\int_{\mathrm{c}_{\mathscr{B}}(\overline{P},2^Jr)}\exp\left[-\tau^{-1}c_1n\beta a_1\ell(\overline{P},P)\right]d\pi(P)\leqslant\pi\left(\mathscr{B}\right)\exp\left[\overline{\Xi}_2-\left(\tau^{-1}c_1-2\gamma\right)n\beta a_1 2^Jr\right].$$

with

$$\overline{\Xi}_2=-\gamma+\log\left[\frac{1}{1-\exp\left[-\left(\tau^{-1}c_1-2\gamma\right)\right]}\right], \qquad (91)$$

which together with (90) leads to

$$\begin{aligned}
&\int_{\mathrm{c}_{\mathscr{B}}(\overline{P},2^Jr)}\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X},P)\right]\right]d\pi(P)\\
&\leqslant\pi\left(\mathscr{B}\right)\exp\left[\overline{\Xi}_1+\overline{\Xi}_2+n\beta a_1\left(e_3\ell(\mathbf{P}^\star,\overline{P})+e_4r-\left(\tau^{-1}c_1-2\gamma\right)2^Jr\right)\right]. \qquad (92)
\end{aligned}$$

Using the definition (89) of z, we deduce that

$$\begin{aligned}
&\log\left[\frac{1}{z}\int_{\mathrm{c}_{\mathscr{B}}(\overline{P},2^Jr)}\mathbb{E}\left[\exp\left[-\beta\mathbf{T}(\boldsymbol{X},P)\right]\right]d\pi(P)\right]\\
&\leqslant\log\left(\frac{1}{z}\right)+\log\pi(\mathscr{B})+\overline{\Xi}_1+\overline{\Xi}_2+n\beta a_1\left(e_3\ell(\mathbf{P}^\star,\overline{P})+e_4r-\left(\tau^{-1}c_1-2\gamma\right)2^Jr\right)\\
&=\xi+\log\frac{1}{\pi(\mathscr{B})}+\overline{\Xi}_1+n\beta a_1\left[e_5\ell(\mathbf{P}^\star,\overline{P})+(e_6+e_7)r\right]\\
&\quad+\log\pi(\mathscr{B})+\overline{\Xi}_1+\overline{\Xi}_2+n\beta a_1\left(e_3\ell(\mathbf{P}^\star,\overline{P})+e_4r-\left(\tau^{-1}c_1-2\gamma\right)2^Jr\right)\\
&=\xi+2\overline{\Xi}_1+\overline{\Xi}_2+n\beta a_1\left[\left(e_3+e_5\right)\ell(\mathbf{P}^\star,\overline{P})+(e_4+e_6+e_7)r\right]\\
&\quad-\left(\tau^{-1}c_1-2\gamma\right)n\beta a_1 2^Jr.
\end{aligned}$$

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

$$2^J\geqslant\frac{1}{\tau^{-1}c_1-2\gamma}\left[\frac{2\xi+2\overline{\Xi}_1+\overline{\Xi}_2}{n\beta a_1r}+\left[\left(e_3+e_5\right)\frac{\ell(\mathbf{P}^\star,\overline{P})}{r}+e_4+e_6+e_7\right]\right].$$

Choosing $r = \ell(\mathbf{P}^\star, \overline{P}) + r_n(\beta, \overline{P}) \geqslant 1/(n\beta a_1)$, this inequality is satisfied as soon as

$$2^J \geqslant \frac{1}{\tau^{-1}c_1 - 2\gamma}\left[2\xi + 2\overline{\Xi}_1 + \overline{\Xi}_2 + e_3 + e_5 + e_4 + e_6 + e_7\right]$$

and the requirement $\kappa_0 \geqslant \tau(2^J + 1)$ if

$$\kappa_0 = \tau\left[1 + \frac{2}{\tau^{-1}c_1 - 2\gamma}\left(2\xi + 2\overline{\Xi}_1 + \overline{\Xi}_2 + e_3 + e_5 + e_4 + e_6 + e_7\right)\right]. \tag{93}$$

7. Other proofs

7.1. **Proof of Proposition 2.** Let $|\overline{\theta}| \leqslant \sigma t$ and F_σ be the distribution function of ν_σ. For conveniency, when L is finite, we define $H(x) = 1$ for all $x \geqslant L$. Since the total variation distance is translation invariant,

$$\|P_\theta - P_{\overline{\theta}}\| = \left\|P_{\theta-\overline{\theta}} - P_0\right\| = \left\|P_{\overline{\theta}-\theta} - P_0\right\| = H\left(|\overline{\theta} - \theta|\right)$$

for all $\theta, \overline{\theta} \in \mathbb{R}$. We distinguish between two cases

Case 1: $r_0 = H(\sigma t) \leqslant 1/4$. Since q is symmetric, positive and decreasing on $\mathbb{R}_+$, for all $r \leqslant r_0$, $G(r) \leqslant \sigma t$ and

$$\begin{aligned}
\frac{\pi(\mathscr{B}(P_{\overline{\theta}}, 2r))}{\pi(\mathscr{B}(P_{\overline{\theta}}, r))} &= \frac{\nu_\sigma\left(\left\{\theta \in \mathbb{R},\ \|P_\theta - P_{\overline{\theta}}\| \leqslant 2r\right\}\right)}{\nu_\sigma\left(\left\{\theta \in \mathbb{R},\ \|P_\theta - P_{\overline{\theta}}\| \leqslant r\right\}\right)} = \frac{\nu_\sigma\left(\left\{\theta \in \mathbb{R},\ H\left(\left|\theta - \overline{\theta}\right|\right) \leqslant 2r\right\}\right)}{\nu_\sigma\left(\left\{\theta \in \mathbb{R},\ H\left(\left|\theta - \overline{\theta}\right|\right) \leqslant r\right\}\right)} \\
&= \frac{\nu_\sigma\left(\left\{\theta \in \mathbb{R},\ \left|\theta - \overline{\theta}\right| \leqslant G(2r)\right\}\right)}{\nu_\sigma\left(\left\{\theta \in \mathbb{R},\ \left|\theta - \overline{\theta}\right| \leqslant G(r)\right\}\right)} \leqslant \frac{2q_\sigma(0)G(2r)}{2q_\sigma(|\theta| + G(r))G(r)} \\
&\leqslant \frac{q_\sigma(0)G(2r)}{q_\sigma(|\theta| + G(r_0))G(r)} \leqslant \frac{q_\sigma(0)G(2r)}{q_\sigma(|\theta| + \sigma t)G(r)} \leqslant \frac{q_\sigma(0)G(2r)}{q_\sigma(2\sigma t)G(r)} \\
&= \frac{q(0)G(2r)}{q(2t)G(r)} \leqslant \frac{\overline{\Gamma}}{q(2t)}.
\end{aligned}$$

For all $r_0 < r < 1$, $|\overline{\theta}| \leqslant \sigma t = G(r_0) \leqslant G(r)$, hence $F_\sigma(|\overline{\theta}| - G(r)) \leqslant F_\sigma(0) = 1/2$ and $F_\sigma(|\overline{\theta}| + G(r)) \geqslant F_\sigma(G(r)) \geqslant F_\sigma(\sigma t) = F_1(t) \geqslant 3/4$. Consequently,

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

Case 2: $r_0 > 1/4$. Arguing as before, we obtain that for all $r \leqslant 1/4 < r_0$,

$$\begin{aligned}
\frac{\pi(\mathscr{B}(P_{\overline{\theta}}, 2r))}{\pi(\mathscr{B}(P_{\overline{\theta}}, r))} &\leqslant \frac{2q_\sigma(0)G(2r)}{2q_\sigma(|\theta| + G(r))G(r)} = \frac{q_\sigma(0)G(2r)}{q_\sigma(|\theta| + G(r_0))G(r)} \\
&\leqslant \frac{q_\sigma(0)G(2r)}{q_\sigma(2\sigma t)G(r)} \leqslant \frac{\overline{\Gamma}}{q(2t)}.
\end{aligned}$$

For all $r \in (1/4, 1)$, $G(r) \geqslant G(1/4)$ and $G(1/4) \leqslant \sigma t$

$$\begin{aligned}\frac{\pi(\mathscr{B}(P_{\overline{\theta}}, 2r))}{\pi(\mathscr{B}(P_{\overline{\theta}}, r))} &\leqslant \frac{1}{\nu_\sigma\left(\left\{\theta \in \mathbb{R},\ \left|\theta - \overline{\theta}\right| \leqslant G(r)\right\}\right)} \\ &\leqslant \frac{1}{\nu_\sigma\left(\left\{\theta \in \mathbb{R},\ \left|\theta - \overline{\theta}\right| \leqslant G(1/4)\right\}\right)} \\ &\leqslant \frac{1}{2q_\sigma(|\overline{\theta}| + G(1/4))G(1/4)} \\ &\leqslant \frac{1}{2q_\sigma(2\sigma t)G(1/4)} \leqslant \frac{\overline{\Gamma}\sigma}{q(2t)}.\end{aligned}$$

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

$$\log\left(\frac{\pi(\mathscr{B}(P_{\overline{\theta}}, 2r))}{\pi(\mathscr{B}(P_{\overline{\theta}}, r))}\right) \leqslant \max\left\{\log\left(\frac{\overline{\Gamma}\,(\sigma \vee 1)}{q(2t)}\right), \log 4\right\}, \tag{94}$$

hence, for all $r \geqslant a_1^{-1}\beta$

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

The right-hand side is not larger than β provided that it satisfies (27) and this lower bound is not smaller than $1/\sqrt{n}$ under the assumption $\eta \leqslant \log 4$. We conclude by using (16).

7.2. **Proof of Proposition 3.** Let us take $t = (|\overline{\theta}|/\sigma) \vee t_0$. For such a value of t, $\overline{\theta} \in [-\sigma t, \sigma t]$ and $\nu_1([t, +\infty)) \leqslant 1/4$. Since Assumption 4 is satisfied, (94) holds true and we deduce from (12) that

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

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

7.3. **Proof of Corollary 1.** Throughout this proof, $a_0 = 3/2$, $a_1 = 1/2$, $\tau = 1$, $\mathcal{A}$ the σ-algebra generated by all the subsets of $\mathscr{M}$ so that, Assumptions 1 and 2 hold. The constants $c = 1/3$, hence $\lambda = 4\beta/3$, and $\gamma = 1/6$ satisfy the constraints of Theorem 1. We may therefore apply it with these choices. Besides, we set $\delta = \delta_n$, $\eta = \eta_n$, $\beta = \beta_n$ and $\Theta = \Theta[\eta, \delta]$ for short and

$$J_n = \exp\left[\frac{(K^2 - 1)\gamma\tau^4 a_1^2 n\eta_n^2}{2(k+1)}\right] \tag{95}$$

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

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

Hereafter we fix $P = P_{(p,\mathbf{m},\sigma)} \in \mathscr{M}_n(K)$. There exist $\theta = \theta(P) = (\overline{p}, \overline{\mathbf{m}}, \overline{\sigma}) \in \Theta$ 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

$$\frac{\overline{\sigma}}{(1+\delta)} \leqslant \sigma < \overline{\sigma} \quad \text{and} \quad \overline{m}_i = j_i \overline{\sigma}\delta \leqslant m_i < \overline{m}_i + \overline{\sigma}\delta, \tag{96}$$

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

$$0 \leqslant \left(1 - \frac{\sigma}{\overline{\sigma}}\right) \leqslant \frac{\delta}{1+\delta} < \delta \quad \text{and} \quad \left|\frac{\mathbf{m} - \overline{\mathbf{m}}}{\overline{\sigma}}\right|_\infty \leqslant \delta, \tag{97}$$

and we infer from (31) and (32) and the fact that the total variation loss is translation and scale invariant that P_θ satisfies

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

Besides, the parameters $(j_0, \mathbf{j}) \in \mathbb{Z} \times \mathbb{Z}^k$ can be controlled in the following way. Using that $\sigma \leqslant \overline{\sigma}$, the inequality $\log(1+\delta) \leqslant \delta$ and (97), we obtain that for all $i \in \{1, \ldots, k\}$,

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

Besides,

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

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

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

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

$$|(j_0, \mathbf{j})|_\infty \leqslant 1 + \frac{1}{\log(1+\delta)}\left[|\log\sigma| \vee \left|\frac{\mathbf{m}}{\sigma}\right|_\infty\right] \leqslant 1 + J_n. \tag{98}$$

For all $r>0$, $e^{-L_\theta} \leqslant \pi\left(\mathscr{B}(P_\theta, r)\right) \leqslant 1$ and these two inequalities together with the definition (35) of η and Assumption 5 imply that for all $r>0$

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

Using (98), the definition (95) of J_n and the fact that $\log(2+x) \leqslant \log 3 + \log x$ for all $x \geqslant 1$, we derive that

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

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

$$\begin{aligned}\frac{1}{n\beta a_1} \leqslant r_n(\beta, P_\theta) &\leqslant \frac{1}{\gamma n\beta a_1}\left[K^2\gamma\tau^4 a_1^2 n\eta^2 + (k+1)L'\right] \\ &= \frac{1}{a_1\beta}\left[K^2\tau^4 a_1^2\eta^2 + \frac{(k+1)L'}{\gamma n}\right].\end{aligned}$$

For the choice of $\beta = \beta_n$ given by (37),

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

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

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

and the result follows from Theorem 1 and the fact that P is arbitrary in $\mathscr{M}_n(K)$.

7.4. **Proof of Lemma 1.** By doing the change of variables $u = x - m$ in (43) if ever necessary, we may assume with no loss of generality that $m > 0$. Then, since p is non-increasing in $(0, +\infty)$ and vanishes elsewhere $p(x-m) \geqslant p(x)$ for all $x \geqslant m$ and $p(x) \geqslant p(x-m) = 0$ for all $x \in (0, m)$. Consequently,

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

and we obtain (43).

Since $\sigma \geqslant 1$, $p(x/\sigma) \geqslant p(x)$ and $p(x)/\sigma \leqslant p(x)$ for all $x > 0$. Hence,

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

which leads to (42).

Finally, by combining (43) and (42) we deduce that for all $m \in \mathbb{R}$ and $\sigma \geqslant 1$

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

which yields to (44).

7.5. **Proof of Corollary 2.** It follows from Proposition 1 and our condition on p that the family $\mathscr{T}(\ell, \mathscr{M})$ satisfies Assumption 2 with $a_0 = 2$, $a_1 = 3/16$ and $a_2 = 3\sqrt{2}/4$ for the loss $\ell = h^2$. Besides, Assumption 1 holds true with $\tau = 2$ and the constants $\gamma = 0.01$, $\beta = 0.01$, $\lambda = (1+c)\beta$ with $c = 0.05$ satisfy the constraints of Theorem 2.

We also use the following lemma the proof of which is postponed to Section 7.6.

Lemma 8. *Let $\boldsymbol{\theta} \in \mathbb{R}^k$ be such that $|\boldsymbol{\theta}|_\infty \leqslant R$. For all $m \subset \{1, \ldots, k\}$ and $r > 0$*

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

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

$$\nu_m\left(\left\{ \boldsymbol{\theta}' \in \mathbb{R}^k,\, \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leqslant r \right\} \right) \geqslant \frac{1}{2^{|m|}} \left(\frac{r}{R} \wedge 1 \right)^{|m|} \tag{99}$$

and for all $K > 1$

$$\frac{\nu_m\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\, \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leqslant Kr\right\}\right)}{\nu_m\left(\left\{\boldsymbol{\theta}' \in \mathbb{R}^k,\, \left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|_\infty \leqslant r\right\}\right)} \leqslant K^{|m|}. \tag{100}$$

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

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

Finally, let $\overline{\boldsymbol{\theta}}$ for some arbitrary element of $\Theta_{\overline{m}}(R)$.

It follows from (45) and (99) that for all $r > 0$,

$$\begin{aligned}
1 &\geqslant \pi_m\left(\mathscr{B}(P_{\overline{\boldsymbol{\theta}}}, r)\right) \\
&= \nu_m\left(\left\{\boldsymbol{\theta} \in \mathbb{R}^k,\ h^2(P_{\overline{\boldsymbol{\theta}}}, P_{\boldsymbol{\theta}}) \leqslant r\right\}\right) \\
&\geqslant \nu_m\left(\left\{\boldsymbol{\theta} \in \mathbb{R}^k,\ \left|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right|_\infty \leqslant (r/B)^{1/(2\alpha)}\right\}\right) \\
&\geqslant \frac{1}{2^{|m|}}\left(\frac{(r/B)^{1/(2\alpha)}}{R} \wedge 1\right)^{|m|} \geqslant \frac{1}{2^{|m|}}\left(\frac{(r \wedge 1)^{1/(2\alpha)}}{RB^{1/(2\alpha)}}\right)^{|m|},
\end{aligned} \tag{101}$$

where the last inequality holds true under the assumption that $RB^{1/(2\alpha)} \geqslant 1$.

We deduce from (101) that for all $r > 0$

$$\begin{aligned}
&\frac{\pi\left(\mathscr{B}(P_{\overline{\boldsymbol{\theta}}}, 2r)\right)}{\pi\left(\mathscr{B}(P_{\overline{\boldsymbol{\theta}}}, r)\right)} \leqslant \frac{1}{\pi\left(\mathscr{B}(P_{\overline{\boldsymbol{\theta}}}, r)\right)} \\
&\leqslant \frac{1}{\sum_{m \subset \{1,\ldots,k\}} e^{-L_m} \nu_m\left(\left\{\boldsymbol{\theta} \in \mathbb{R}^k,\ \left|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right|_\infty \leqslant (r/B)^{1/(2\alpha)}\right\}\right)} \\
&\leqslant \frac{e^{L_{\overline{m}}}}{\nu_{\overline{m}}\left(\left\{\boldsymbol{\theta} \in \Theta_{\overline{m}},\ \left|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right|_\infty \leqslant (r/B)^{1/(2\alpha)}\right\}\right)} \\
&\leqslant \exp\left[L_{\overline{m}} + |\overline{m}| \log\left(\frac{2RB^{1/(2\alpha)}}{(r \wedge 1)^{1/(2\alpha)}}\right)\right] \\
&= \exp\left[|\overline{m}| \log\left(2kRB^{1/(2\alpha)}\right) + k \log\left(1 + \frac{1}{k}\right) + \frac{|\overline{m}|}{2\alpha} \log\left(\frac{1}{r} \vee 1\right)\right].
\end{aligned} \tag{102}$$

Provided that

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

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

and we deduce from (102) that $r_n(\beta, P_{\overline{\boldsymbol{\theta}}})$ defined by (12) satisfies

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

By applying Theorem 2, we conclude that (24) holds with

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

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

7.6. **Proof of Lemma 8.** 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 $|\boldsymbol{\theta}|_\infty \leqslant R$. For all $m \subset \{1, \ldots, k\}$, $m \neq \varnothing$,

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

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

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

If $m = \varnothing$,

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

Let us now turn to the proof of (100). Since $\boldsymbol{\theta} \in \Theta_m(R)$, for all $K' \in \{1, K\}$

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

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

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

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

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

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

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

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

which concludes the proof.

References

Alquier, P. (2008). PAC-Bayesian bounds for randomized empirical risk minimizers. *Math. Methods Statist.*, 17(4):279–304.

Audibert, J.-Y. and Catoni, O. (2011). Linear regression through PAC-Bayesian truncation. *arXiv:1010.0072*.

Baraud, Y. (2021). Tests and estimation strategies associated to some loss functions. *Probab. Theory Relat. Fields*, 180(3):799–846.

Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.*, 46(6B):3767–3804.

Baraud, Y. and Birgé, L. (2020). Robust bayes-like estimation: Rho-bayes estimation. *Ann. Statist.*, 48(6):3699–3720.

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325.

Birgé, L. (2015). About the non-asymptotic behaviour of Bayes estimators. *J. Statist. Plann. Inference*, 166:67–77.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Birman, M. v. and Solomjak, M. Z. (1967). Piecewise polynomial approximations of functions of classes $W_p{}^{\alpha}$. *Mat. Sb. (N.S.)*, 73 (115):331–355.

Catoni, O. (2004). 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-Verlag, Berlin.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*, volume 16. Springer-Verlag, New York.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53.

Massart, P. (2007). *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.

Department of Mathematics,
University of Luxembourg
Maison du nombre
6 avenue de la Fonte
L-4364 Esch-sur-Alzette
Grand Duchy of Luxembourg
Email address: yannick.baraud@uni.lu