View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Open Repository and Bibliography - Luxembourg

TESTS AND ESTIMATION STRATEGIES ASSOCIATED TO SOME LOSS FUNCTIONS

YANNICK BARAUD

Abstract. We consider the problem of estimating the joint distribution of n independent random variables. Our approach is based on a family of candidate probabilities that we shall call *a model* and which is chosen to either contain the true distribution of the data or at least to provide a good approximation of it with respect to some loss function. The aim of the present paper is to describe a general estimation strategy that allows to adapt to both the specific features of the model and the choice of the loss function in view of designing an estimator with good estimation properties. The losses we have in mind are based on the total variation, Hellinger, Wasserstein and $\mathbb{L}_p$-distances to name a few. We show that the risk of the resulting estimator with respect to the loss function can be bounded by the sum of an approximation term accounting for the loss between the true distribution and the model and a complexity term that corresponds to the bound we would get if this distribution did belong to the model. Our results hold under mild assumptions on the true distribution of the data and are based on exponential deviation inequalities that are non-asymptotic and involve explicit constants. When the model reduces to two distinct probabilities, we show how our estimation strategy leads to a robust test whose errors of first and second kinds only depend on the losses between the true distribution and the two tested probabilities.

1. Introduction

Observe n independent random variables $X_1, \ldots, X_n$ with values in a measured space $(E, \mathcal{E}, \mu)$ and assume they are i.i.d. with common distribution $P^\star$. Consider now a loss function ℓ for evaluating the performance of an estimator of $P^\star$ together with a model $\mathscr{M}$, i.e. a family of candidate probabilities for $P^\star$, that may or may not contain $P^\star$ but which is believed to provide a suitable approximation of it. The purpose of the present paper is to design a generic method for estimating $P^\star$ that takes into account our

Date: March 27, 2020.

1991 *Mathematics Subject Classification.* Primary 62F35, 62G35, 62G07; Secondary 62C20, 62G10.

Key words and phrases. Density estimation, parametric estimation, robust estimation, Wasserstein loss, total variation loss, L_p-loss, minimax theory, robust testing.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017.

choices of the loss function ℓ and the model $\mathscr{M}$ in view of building estimators with good estimation properties.

To be more specific, we build an estimator $\widehat{P}$ with values in $\mathscr{M}$ and measure its accuracy by the quantity $\ell(P^\star, \widehat{P})$ where the loss ℓ is a nonnegative function defined on $\mathscr{P} \times \mathscr{M}$ for some suitable set $\mathscr{P}$ containing the true distribution $P^\star$. Even though ℓ may not be a distance (it may neither be symmetrical nor satisfy the triangle inequality), we shall interpret the loss $\ell(P^\star, \widehat{P})$ just as if it were: small values of $\ell(P^\star, \widehat{P})$ would mean $\widehat{P}$ is "close" to $P^\star$ while large value of $\ell(P^\star, \widehat{P})$ would in contrast be understood as $\widehat{P}$ is "far" from it. Our purpose is to design an estimation procedure which guarantees small values of $\ell(P^\star, \widehat{P})$ whenever $\mathscr{M}$ provides a good approximation of $P^\star$ (i.e. $\inf_{P\in\mathscr{M}} \ell(P^\star, P)$ is small) and the *dimension* of $\mathscr{M}$, defined in a suitable way, remains small compared to n.

This problem was solved for the Hellinger loss in Baraud *et al.* (2017) and Baraud and Birgé (2018) and to give an account of these results, let us first recall that the squared Hellinger distance $h^2(P, Q)$ between two probabilities P and Q on E is given by the formula

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

where ν denotes an arbitrary measure on $(E, \mathcal{E})$ that dominates both P and Q, the result being independent of the choice of ν. The estimator $\widehat{P}(\boldsymbol{X})$ which results from their procedure (named ρ-estimation) typically satisfies an inequality of the form

$$\mathbb{E}\left[h^2\left(P^\star, \widehat{P}(\boldsymbol{X})\right)\right] \leqslant C\left[\inf_{P\in\mathscr{M}} h^2(P^\star, P) + \frac{D_n(\mathscr{M})}{n}\right], \tag{2}$$

where C is a positive numerical constant and $D_n(\mathscr{M})$ a complexity term that may depend on the number n of observations and the dimension (in some sense) of the statistical model $\mathscr{M}$. This inequality essentially says that the loss between $P^\star$ and $\widehat{P}$ is not larger $CD_n(\mathscr{M})/n$ when $P^\star$ belongs to the model $\mathscr{M}$ and that this bound does not deteriorate too much as long as $\inf_{P\in\mathscr{M}} h^2(P^\star, P)$ remains sufficiently small, i.e. as long as $P^\star$ remains close enough to the model with respect to the Hellinger loss. An interesting feature of this result lies in the following facts: the inequality (2) is true under very weak assumptions on both the statistical model $\mathscr{M}$ and the underlying distribution $P^\star$ and, in all cases we know, the quantity $D_n(\mathscr{M})/n$ turns out to be the best possible bound that can be achieved uniformly over the model $\mathscr{M}$ (up to a possible logarithmic factor).

In the present paper, we would like to extend this result to other losses. One may think for instance to the total variation distance, the Wasserstein distance, the Kullback-Leibler divergence and the $\mathbb{L}_j$-distances with $j > 1$, among others. Unfortunately, for most of these losses, there is no hope to obtain a risk bound which is similar to (2) under weak assumptions on both the model $\mathscr{M}$ and the true distribution $P^\star$, as in the case of the Hellinger loss. If, for instance, $\mathscr{M}$ is the set of all uniform distributions on $[\theta, \theta+1]$ with $\theta \in \mathbb{R}$ and ℓ is the Kullback-Leibler divergence, $\sup_{P^\star \in \mathscr{M}} \ell(P^\star, \widehat{P}) = +\infty$ whatever the choice of the estimator $\widehat{P}$ and there is consequently no way of controlling the risk as in (2). The situation is not much better with the $\mathbb{L}_j$-loss since it requires that the densities associated to the probabilities in $\mathscr{M}$ belong to $\mathbb{L}_j(E, \mathcal{E}, \mu)$, that the unknown probability $P^\star$ be dominated by the reference measure μ and that its density also belong to $\mathbb{L}_j(E, \mathcal{E}, \mu)$. In view of these disappointing observations, we see that specific assumptions need to be made in order to deal with such loss functions. Our point of view in this paper is to make (possibly strong) assumptions on the model $\mathscr{M}$, since it is chosen by the statistician, but to assume as little as possible on $P^\star$ since it is unknown.

Despite some differences, our approach shares some similarities with that developed for the Hellinger loss in Baraud *et al.* (2017) and Baraud and Birgé (2018). In particular, it is also based on the existence of suitable tests between probability "balls" (with a suitable sense when ℓ is not a genuine distance). We shall give some general recipe about how to build such tests for the various loss functions we consider. When applied to some specific losses, our construction allows to recover some well-known tests while for other losses, these tests are to our knowledge new. For example, for the total variation distance, our testing procedure is the same as that proposed by Devroye and Lugosi (see Devroye and Lugosi (2001)[Chapter 6]) based on the seminal paper by Yatracos (1985). For the Kullback-Leibler divergence we obtain the classical likelihood ratio test while, for the $\mathbb{L}_2$-distance, we obtain the test based on the $\mathbb{L}_2$-contrast. For the Wasserstein distance and the $\mathbb{L}_j$-losses with $j \neq 2$, the tests we derive from our method seem to be new in the literature.

Our estimation procedure is based on the above-mentioned tests and results in a new class of estimators that we shall call *ℓ-estimators* and which generalize ρ-estimators. The study of these estimators can interestingly be made within a unified framework even though, in order to keep the present paper to a reasonable size, we shall mainly discuss the cases of the total variation and $\mathbb{L}_2$-losses.

Throughout the paper, we shall provide several applications to illustrate the performances of ℓ-estimators. Some of these will give us the opportunity to contextualize these estimators within the literature by recovering some known results. Other applications will produce new ones.

For the total variation loss, we show that ℓ-estimators can achieve a convergence rate which is faster than the usual $1/\sqrt{n}$ rate. Such results contrast with those obtained previously by Birgé (2006) with T-estimators (see his Corollary 6) and Devroye and Lugosi (2001) with skeleton estimators. Closer to our approach (for this loss) is that of Gao *et al.* (2018). In their paper, these authors proposed a robust estimation of the mean of a Gaussian vector based on the observation of an n-sample. The estimator proposed by Gao *et al.* in this specific framework shares some similarities with ours. It is also obtained as the minimizer of the supremum of a random functional over a suitable class of functions. However, our construction differs from theirs by the choices of the classes over which the supremum and infimum are computed.

We also address the problem of estimating a density on $\mathbb{R}^d$ with respect to the Lebesgue measure when the risk is defined through the $\mathbb{L}_2$-loss and without assuming that the true density is bounded in sup-norm. We are only aware of very few results in this direction. Birgé and Massart (1998) studied the performances of minimum $\mathbb{L}_2$-contrast estimators on linear spaces V on which the $\mathbb{L}_\infty$-norm is suitably controlled by the $\mathbb{L}_2$-one. Their results, however, suffer from two limitations: the functions in V are supported on a compact set, say $[0,1]^d$, and V is finite dimensional. Our theory allows us to relax these restrictions and generalize their results to an infinite dimensional linear space of functions with possibly unbounded support ($\mathbb{R}^d$ typically). For a suitable choice of the linear space V we shall derive a uniform risk bound over the class of all squared integrable densities that lie in a Besov space $B^\alpha_{s,\infty}(\mathbb{R}^d)$ with parameters $s \geqslant 2$ and $\alpha > 0$. This result is to our knowledge new and generalizes that obtained by Rivoirard *et al.* (2011) in the case of $d = 1$ and $s > 2$ (we also refer to Rivoirard and Reynaud-Bouret (2010) for a lower bound on the minimax risk).

Our paper is organized as follows. We first present the statistical framework as well as our main assumptions in Section 2. We shall actually consider a more general framework than the one described in this introduction since we assume the observations $X_1, \ldots, X_n$ to be independent but not necessarily i.i.d. We shall also allow our model $\mathscr{M}$ to contain finite and possibly signed measures, hence not only probabilities. Such models will be useful when dealing with $\mathbb{L}_j$-losses, $j > 1$. The heuristics underlying our approach is also described in Section 2 as well as our main assumptions on the loss functions we use. The estimation procedure and the general results on the performance of ℓ-estimators are presented in Section 3 and some consequences of these results in Section 4 where we deal with the cases of the Wasserstein and the $\mathbb{L}_2$-losses. Our results about the risk of ℓ-estimators for the $\mathbb{L}_2$-loss over Besov balls on $\mathbb{R}^d$ will also be found there. We then put a special emphasis on the total variation loss in Section 5. In particular, we provide an illustration to the problem of estimating a non-increasing density on a half line for the $\mathbb{L}_1$-distance. We shall also see how ℓ-estimators associated to the total variation loss can reach rates which are faster than $1/\sqrt{n}$. Hellinger and Kullback-Leibler-losses will be considered in Section 6 while Section 7 will be devoted to the comparison of the properties of ρ- and ℓ-estimators for the total variation loss. Our procedure is based on the existence of a family of robust tests between two probabilities whose performance will be described in Section 8 with an emphasis on the cases of the total variation and $\mathbb{L}_j$-losses. Finally Section 9 is devoted to the proofs of the main theorems and Section 10 to the other proofs.

2. The statistical framework and our main assumptions

The statistical framework of the present paper is slightly different from our previous description since we actually do not assume that the observations $X_1, \dots, X_n$ are i.i.d. but only independent with respective marginals $P_1^\star, \dots, P_n^\star$. In many cases, our statistical model will be based on the assumption that the data are i.i.d. although this might not be true and our aim will then be to analyze the behaviour of our estimator with respect to a possible departure from the assumption that the X_i are i.i.d.

Throughout this paper, $\mathscr{P}$ denotes a set of probabilities on $(E, \mathcal{E})$ that contains the marginal distributions $P_1^\star, \dots, P_n^\star$, $\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star$ is the distribution our observation $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{\mathscr{P}} = \{\mathbf{P} = \bigotimes_{i=1}^n P_i,\ P_i \in \mathscr{P}\}$ denotes the set of all product probabilities with marginals in $\mathscr{P}$. Hence, $\mathbf{P}^\star$ belongs to $\boldsymbol{\mathscr{P}}$.

2.1. Notations.

For convenience, we shall identify an element $\mathbf{P} = \bigotimes_{i=1}^n P_i$ of $\boldsymbol{\mathscr{P}}$ with the n-uplet $(P_1, \dots, P_n)$. Hence, depending on the context, we shall write $\mathbf{P}$ either as a product of probabilities or as an n-uplet. When we write $\mathbb{E}[g(\boldsymbol{X})]$, we assume that the distribution of $\boldsymbol{X}$ is $\mathbf{P}^\star$ while $\mathbb{E}_S[f(X)]$ means the expectation of $f(X)$ when the distribution of X is S. We use the same conventions for $\mathrm{Var}\left(g(\boldsymbol{X})\right)$ and $\mathrm{Var}_S\left(f(X)\right)$.

Beside these conventions, we shall use the following notations. For $x \in \mathbb{R}$ and $k > 0$, $x_+^k = \max\{0, x\}^k$ and $x_-^k = \max\{0, -x\}^k$; for $x \in \mathbb{R}^d$, $|x|$ denotes the Euclidean norm of x and $B(x, r)$ the closed Euclidean ball centered at x with radius $r \geqslant 0$. Given a σ-finite measure μ on $(E, \mathcal{E})$ and $j \geqslant 1$, we shall denote by $\mathscr{L}_j(E, \mu)$ the set of measurable functions f on $(E, \mathcal{E}, \mu)$ such that $\|f\|_j^j = \int_E |f|^j d\mu < +\infty$ and $\mathbb{L}_j(E, \mu)$ the associated set of equivalent classes on which two functions that coincide for μ-almost all $x \in E$ are indistinguishable. If d is a positive integer, we shall write $\mathscr{L}_j(\mathbb{R}^d)$ and $\mathbb{L}_j(\mathbb{R}^d)$ for $\mathscr{L}_j(E, \mu)$ and $\mathbb{L}_j(E, \mu)$ respectively when $E = \mathbb{R}^d$, $\mathcal{E}$ its Borel σ-algebra and $\mu = \lambda$ the Lebesgue measure on $\mathbb{R}^d$. Finally, we denote by $\|f\|_\infty$ the quantity $\sup_{x \in E} |f(x)| < +\infty$.

2.2. Models.

As already mentioned, our strategy for estimating $\mathbf{P}^\star$ is based on models. This means that we assume to have at disposal a family $\boldsymbol{\mathscr{M}}$ of elements of the form $(P_1, \dots, P_n)$ where the P_i are finite measures on $(E, \mathcal{E})$, possibly signed, which belong to some set $\mathscr{M}$. In most cases, the P_i will be probabilities but it will sometimes be more convenient to consider signed measures of the form $p \cdot \mu$ where p is not necessarily a density of probability but an element of $\mathscr{L}_j(E, \mu) \cap \mathscr{L}_1(E, \mu)$ for some $j > 1$.

In the density setting, i.e. when we believe that the observations $X_1 \ldots, X_n$ are i.i.d., although this might not be true, we shall consider a model $\boldsymbol{\mathscr{M}}$ which corresponds to this belief, choosing $\boldsymbol{\mathscr{M}}$ of the form $\{(P,\ldots,P),\ P \in \mathscr{M}\}$ and specifying $\mathscr{M}$ only. In other statistical frameworks such as the regression one, the marginals $P_1^\star,\ldots,P_n^\star$ of $\mathbf{P}^\star$ may have different features and it shall then be convenient to consider in $\boldsymbol{\mathscr{M}}$ elements of the form $(P_1,\ldots,P_n)$ with possibly different entries P_i.

Throughout this paper we shall assume that $\mathscr{M}$ (and therefore $\boldsymbol{\mathscr{M}} \subset \mathscr{M}^n$) is *at most countable* (which means finite or countable) in order to avoid measurability issues. Since the model $\boldsymbol{\mathscr{M}}$ is only assumed to provide an approximation of $\mathbf{P}^\star$ and may not contain it, this condition is not very restrictive: most of the models that statisticians use are separable and can therefore be well approximated by countable subsets.

Since $\mathscr{M}$ is countable, it is dominated and there exists a σ-finite measure μ on $(E,\mathcal{E})$ for which we may write any element $P \in \mathscr{M}$ as $P = p\cdot\mu$ with p an integrable function on $(E,\mathcal{E})$. Throughout the paper, we shall assume the measure μ associated to $\mathscr{M}$ to be fixed once and for all and that the statistician has chosen for each $P \in \mathscr{M}$ a convenient version $p \in \mathscr{L}_1(E,\mu)$ of $dP/d\mu$. We shall systematically use the corresponding lower case letter to denote this density. This construction results in a family of densities $\mathcal{M}$ associated to $\mathscr{M}$. We shall sometime rather start from a countable family $\mathcal{M}$ of densities in $\mathscr{L}_1(E,\mu)$ (which may not be probability densities) and then define $\mathscr{M}$ as the family of (possibly signed) finite measures $\{P = p\cdot\mu,\ p \in \mathcal{M}\}$.

Given the previous framework, the observation $\boldsymbol{X}$ and the model $\boldsymbol{\mathscr{M}}$, we want to design an estimator $\widehat{\mathbf{P}} = \widehat{\mathbf{P}}(\boldsymbol{X})$ of $\mathbf{P}^\star$ with values in $\boldsymbol{\mathscr{M}}$. To evaluate its performance, we introduce a loss function ℓ defined on $\mathscr{P}\times\mathscr{M}$ with values in $\mathbb{R}_+$. For $\mathbf{P} = (P_1,\ldots,P_n) \in \mathscr{P}^n$ and $\mathbf{Q} = (Q_1,\ldots,Q_n) \in \mathscr{M}^n$, we set

$$\boldsymbol{\ell}(\mathbf{P},\mathbf{Q}) = \sum_{i=1}^{n} \ell(P_i,Q_i) \tag{3}$$

and measure the quality of an estimator $\widehat{\mathbf{P}}$ by $\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}})$. The smaller this quantity, the better the estimator. Since, by construction, $\widehat{\mathbf{P}} \in \boldsymbol{\mathscr{M}}$, $\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}})$ is lower bounded by $\inf_{\mathbf{Q}\in\boldsymbol{\mathscr{M}}} \boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q}) = \boldsymbol{\ell}(\mathbf{P}^\star,\boldsymbol{\mathscr{M}})$ and the best we can expect is that $\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}})$ be close to $\boldsymbol{\ell}(\mathbf{P}^\star,\boldsymbol{\mathscr{M}})$.

2.3. **Some heuristics.** To simplify the presentation of our heuristics, let us assume that the X_i are actually i.i.d. with distribution $P^\star$ and that the elements of $\boldsymbol{\mathscr{M}}$ take the form $P^{\otimes n}$ with $P \in \mathscr{M}$ so that $\boldsymbol{\ell}(\mathbf{P},\mathbf{Q}) = n\ell(P,Q)$ by (3). If $P^\star$ were known, the loss function ℓ would provide an ordering between the elements of $\mathscr{M}$ by saying that P is better than Q if $\ell(P^\star,P) \leqslant \ell(P^\star,Q)$ and an ideal point in $\mathscr{M}$ for estimating $P^\star$ would be a point $\overline{P} \in \mathscr{M}$ such that $\ell(P^\star,\overline{P}) = \inf_{P\in\mathscr{M}} \ell(P^\star,P)$. Since $P^\star$ is unknown, one cannot find $\overline{P}$.

Assume nevertheless that we are able to approximate $\ell(P^\star, P) - \ell(P^\star, Q)$ by some statistic $T(\boldsymbol{X}, P, Q)$ with an error bounded by Δ. We can use $T(\boldsymbol{X}, P, Q)$ for testing between P and Q, deciding according to the sign of $T(\boldsymbol{X}, P, Q)$. This results in a robust test (because we do not assume that $P^\star$ is either P or Q and not even very close to any of them) between P and Q which decides correctly provided that $|\ell(P^\star, P) - \ell(P^\star, Q)| > \Delta$. When (P, Q) varies among all possible pairs of probabilities within our statistical model, we obtain a family $\{T(\boldsymbol{X}, P, Q), (P, Q) \in \overline{\mathscr{M}^2}\}$ of robust tests which we can use to build an estimator of $P^\star$, or rather of $\overline{P}$, as defined above.

Deriving an estimator from a family of robust tests is not a new problem and methods for that have been developed a long time ago by Le Cam (1973) and then Birgé (1983), more recently by Baraud (2011) and then Baraud *et al.* (2017) and it is actually this last recipe that we shall use here. In Baraud *et al.* (2017) it was used to handle the loss $\ell = h^2$ derived from the Hellinger distance h to build ρ-estimators. It worked because we could (approximately) express $h^2(P^\star, P) - h^2(P^\star, Q)$ as the expectation of $T(\boldsymbol{X}, P, Q)$ or, more precisely, view $T(\boldsymbol{X}, P, Q)$ as an empirical version of an approximation of $h^2(P^\star, P) - h^2(P^\star, Q)$, then use the properties of the corresponding empirical process indexed by (P, Q) to build a suitable estimator. To mimic this construction, we need that similar arguments could be applied to the loss ℓ and the choice of the various losses we shall consider below follows from these requirements. We shall explain more precisely in Sections 4.1 and 6.1 what properties of the loss imply the assumptions that are needed for our proofs.

As to the performances of the robust tests based on the sign of $T(\boldsymbol{X}, P, Q)$ that we mentioned previously, they are interesting by themselves and will be studied in details in Section 8.

2.4. **Loss functions.** Let us now provide the definitions of the various loss-functions we shall consider in this paper.

Total variation loss (TV-loss). The *total variation distance* $\|P - Q\|$ between two probabilities P, Q on $(E, \mathcal{E})$ is usually defined as follows:

$$\|P - Q\| = \sup_{A \in \mathcal{E}} [P(A) - Q(A)]. \tag{4}$$

The total variation loss is $\ell(P, Q) = \|P - Q\|$. We shall write TV for *total variation* hereafter.

Hellinger loss. The Hellinger loss is related to the *Hellinger distance* h, which is defined by (1), by $\ell(P, Q) = h^2(P, Q)$. We recall that the quantity $\rho(P, Q) = 1 - h^2(P, Q)$ is the Hellinger affinity between P and Q.

Kullback-Leibler loss (KL-loss). The *Kullback-Leibler divergence* $K(P, Q)$

between two probabilities $P = p \cdot d\mu$ and $Q = q \cdot \mu$ on $(E, \mathcal{E})$ is defined as

$$K(P,Q) = \begin{cases} \int_E \log(p/q)\, p\, d\mu & \text{when } P \ll Q \\ +\infty & \text{otherwise,} \end{cases} \tag{5}$$

with the following conventions:

$$\text{For } x \in E, \qquad \log\left(\frac{p}{q}\right)(x) = \begin{cases} 0 & \text{if } p(x) = q(x) = 0 \\ +\infty & \text{if } p(x) > 0 \text{ and } q(x) = 0 \\ -\infty & \text{if } p(x) = 0 \text{ and } q(x) > 0. \end{cases}$$

In particular, $\exp\left[\log(p(x)/q(x))\right] = p(x)/q(x)$ for all $x \in E$ with the conventions $0/0 = 1$ and $a/0 = +\infty$ for all $a > 0$. This results in the KL-loss, $\ell(P,Q) = K(P,Q)$.

Wasserstein loss. The *(first) Wasserstein distance* between two probabilities P and Q on $E = [0,1]$ (with $\mathcal{E}$ its Borel σ-algebra) associated to the Euclidean metric is

$$W(P,Q) = \inf_{X \sim P, Y \sim Q} \mathbb{E}\left[|X - Y|\right] = \sup_f \left[\mathbb{E}\left(f(X)\right) - \mathbb{E}\left(f(Y)\right)\right] \tag{6}$$

where the infimum runs among all pairs (X,Y) with marginal distributions P and Q and the supremum among all functions f on $[0,1]$ which are Lipschitz with Lipschitz constant not larger than 1. We refer to Villani (2009) [pages 77 and 78].

$\mathbb{L}_j$-loss. Given some positive measure μ on $(E, \mathcal{E})$ and $j \in (1, +\infty)$, we consider the set $\overline{\mathscr{P}}_j$ of finite and signed measures P on $(E, \mathcal{E})$ of the form $P = p \cdot \mu$ with $p \in \mathscr{L}_j(E,\mu) \cap \mathscr{L}_1(E,\mu)$. It is a normed linear space with $\mathbb{L}_j$-norm given by $\|P\|_j = \left[\int_E |p|^j \, d\mu\right]^{1/j}$. Given two elements $P = p \cdot \mu$ and $Q = q \cdot \mu$ in $\overline{\mathscr{P}}_j$, we define the $\mathbb{L}_j$*-loss* ℓ_j on $\overline{\mathscr{P}}_j$ by

$$\ell_j(P,Q) = \|P - Q\|_j = \left[\int_E |p - q|^j \, d\mu\right]^{1/j}. \tag{7}$$

Unlike the losses we have seen so far, the $\mathbb{L}_j$-loss between P and Q depends on the choice of the reference measure μ. Changing μ would automatically change the value of $\ell_j(P,Q)$.

2.5. **Assumptions.** As already mentioned in Section 2.3, the construction we use here only applies to some specific loss functions ℓ and countable models $\mathscr{M}$. They are characterized by the fact that one can find a family

$$\mathscr{T}(\ell, \mathscr{M}) = \left\{\phi_{(P,Q)}, \ (P,Q) \in \mathscr{M}^2\right\}$$

of measurable functions on $(E, \mathcal{E})$ with the following properties.

Assumption 1. *The elements $\phi_{(P,Q)}$ of $\mathscr{T}(\ell, \mathscr{M})$ satisfy:*

(i) for all $P, Q \in \mathscr{M}$, $\phi_{(P,Q)} = -\phi_{(Q,P)}$. In particular, $\phi_{(P,P)} = 0$ for all $P \in \mathscr{M}$;

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

$$\mathbb{E}_S\left[\phi_{(P,Q)}(X)\right] \leqslant a_0\ell(S,P) - a_1\ell(S,Q); \tag{8}$$

(iii) whatever P *and* Q *in* $\mathscr{M}$*,*

$$\sup_{x,x' \in E} |\phi_{(P,Q)}(x) - \phi_{(P,Q)}(x')| \leqslant 1.$$

Note that (i) and (8) together imply that $\mathbb{E}_S\left[\phi_{(Q,P)}(X)\right] \geqslant a_1\ell(S,Q) - a_0\ell(S,P)$. Exchanging the roles of P and Q we get

$$\mathbb{E}_S\left[\phi_{(P,Q)}(X)\right] \geqslant a_1\ell(S,P) - a_0\ell(S,Q),$$

which, together with (8), implies that whatever $S \in \mathscr{P}$ and $P, Q \in \mathscr{M}$

$$a_1\ell(S,P) - a_0\ell(S,Q) \leqslant a_0\ell(S,P) - a_1\ell(S,Q).$$

If $\mathscr{P} \cap \mathscr{M} \neq \varnothing$, setting $S = Q \in \mathscr{P} \cap \mathscr{M}$ in the previous inequality leads to $a_1 \leqslant a_0$.

We shall sometimes reinforce Assumption 1 in the following way in order to establish a more accurate risk bound for our estimator.

Assumption 2. *Additionally to (i), (ii) and (iii) of Assumption 1, the following inequality holds for some constant* $a_2 \geqslant a_1$*,*

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

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

It is clear that if a function $\phi_{(P,Q)}$ satisfies (i), (ii) and (iv) for some positive numbers a_0, a_1 and a_2, so does $C\phi_{(P,Q)}$ for $C > 0$ with the constants Ca_0, Ca_1 and C^2a_2 in place of a_0, a_1 and a_2 respectively. Condition (iii) may therefore be interpreted as a normalizing condition which can be applied to any bounded function $\phi_{(P,Q)}$ that satisfies (i), (ii) and possibly (iv).

In (iv), the condition $a_2 \geqslant a_1$ is only present for convenience in order to simplify some proofs; it can easily be satisfied by enlarging the constant a_2 whenever necessary.

We shall see in Section 4.1 that the loss functions we have introduced in the previous section can be associated to families $\mathscr{T}(\ell, \mathscr{M})$ that satisfy Assumption 1 (and sometimes Assumption 2).

3. The ℓ-estimator and its risk bound on a model

3.1. **The two-points model and robust tests.** Let us start by considering the case of a two-points model $\mathscr{M} = \{\mathbf{P}, \mathbf{Q}\}$ with $\mathbf{P} = (P_1, \dots, P_n)$, $\mathbf{Q} = (Q_1, \dots, Q_n)$, the marginals P_i and Q_i belonging to $\mathscr{M}$ for all i, while the true distribution of $\boldsymbol{X}$ is $\mathbf{P}^\star = (P_1^\star, \dots, P_n^\star)$ which typically does not belong to $\mathscr{M}$. Estimating $\mathbf{P}^\star$ using the model $\mathscr{M}$ amounts to testing between $\{\mathbf{P}\}$ and $\{\mathbf{Q}\}$ under misspecification. In this context the natural decision, if $\mathbf{P}^\star$ were known, would be to choose the point in $\mathscr{M}$ closest to $\mathbf{P}^\star$, closeness being defined in terms of the loss function ℓ. Using the functions $\phi_{(P,Q)} \in \mathscr{T}(\ell, \mathscr{M})$ which satisfy Assumption 1 we define

$$\mathbf{T}(\mathbf{x}, \mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{n} \phi_{(P_i,Q_i)}(x_i) \quad \text{for all } \mathbf{x} = (x_1, \dots, x_n) \in \mathbf{E}. \tag{9}$$

Proposition 1. *Under Assumption 1-(i) and (ii), whatever the true distribution* $\mathbf{P}^\star \in \mathscr{P}$*,*

$$\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})\right] \leqslant a_0 \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{P}) - a_1 \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q}) \tag{10}$$

and

$$\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})\right] \geqslant a_1 \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{P}) - a_0 \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q}). \tag{11}$$

Proof. Applying (ii) to $S = P_i^\star$, $P = P_i$, and $Q = Q_i$ for all $i \in \{1, \dots, n\}$ and summing the resulting inequalities leads to (10). Analogously, using (i) and exchanging the roles of $\mathbf{P}$ and $\mathbf{Q}$ in (10) leads to (11). □

In order to analyze the consequences of Proposition 1, let us pretend that $a_1 = a_0$ so that $a_0^{-1}\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})\right] = \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{P}) - \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q})$. As a consequence, the sign of $\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})\right]$ indicates which element among $\{\mathbf{P}, \mathbf{Q}\}$ minimizes the loss to $\mathbf{P}^\star$. Since $\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})\right]$ is unknown, it is natural to replace it by its empirical version $\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})$ and to introduce a test $\Phi_{(\mathbf{P},\mathbf{Q})}$ with values in $\{0, 1\}$ satisfying

$$\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = \begin{cases} 1 & \text{if} \quad \mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q}) > 0 \\ 0 & \text{if} \quad \mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q}) < 0. \end{cases} \tag{12}$$

This means that we decide that $\mathbf{P}$ is closer to $\mathbf{P}^\star$ when $\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 0$ and that $\mathbf{Q}$ is closer to $\mathbf{P}^\star$ when $\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 1$, the choice between $\mathbf{P}$ and $\mathbf{Q}$ being unimportant, as well as the value of $\Phi_{(\mathbf{P},\mathbf{Q})}$, when $\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q}) = 0$. The performance of the test $\Phi_{(\mathbf{P},\mathbf{Q})}$ under Assumption 1 or 2 will be discussed in more detail in Section 8.

3.2. **Our estimation procedure.** The basic idea underlying our estimation procedure is based on the following heuristics. On the basis of Proposition 1 (with $a_1 = a_0$ as before for the sake of simplicity), $a_0^{-1}\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})$ is an estimator of the difference $\boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{P}) - \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q})$ for $\mathbf{P}$ and $\mathbf{Q}$ in $\mathscr{M}$. If for a fixed $\mathbf{P} \in \mathscr{M}$ we believe that this estimation is uniformly good over all $\mathbf{Q} \in \mathscr{M}$, the quantity $a_0^{-1} \sup_{\mathbf{Q} \in \mathscr{M}} \mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})$ should be close to

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

Since this latter quantity is minimum for the best approximation point of $\mathbf{P}^\star$ in $\mathscr{M}$ (whenever it exists), it is natural to define our estimator as a minimizer over $\mathscr{M}$ of the map

$$\mathbf{P} \mapsto \mathbf{T}(\boldsymbol{X}, \mathbf{P}) = \sup_{\mathbf{Q} \in \mathscr{M}} \mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q}).$$

This minimizer may not exist but hopefully only an ϵ-*minimizer* is necessary. More precisely, given $\epsilon > 0$, we define an ℓ-*estimator* of $\mathbf{P}^\star$ in $\mathscr{M}$ as any element $\widehat{\mathbf{P}}$ of the set

$$\mathscr{E}(\boldsymbol{X}) = \left\{\mathbf{P} \in \mathscr{M},\ \mathbf{T}(\boldsymbol{X}, \mathbf{P}) \leqslant \inf_{\mathbf{P}' \in \mathscr{M}} \mathbf{T}(\boldsymbol{X}, \mathbf{P}') + \epsilon\right\}. \tag{13}$$

As we shall see below, it is preferable to choose ϵ small (not much larger than 1) in order to improve the risk bound of an ℓ-estimator. In particular, if there exists an element $\mathbf{P} \in \mathscr{M}$ (not necessarily unique) such that

$$\mathbf{T}(\boldsymbol{X}, \mathbf{P}) = \inf_{\mathbf{P}' \in \mathscr{M}} \mathbf{T}(\boldsymbol{X}, \mathbf{P}'),$$

we should choose it as our estimator $\widehat{\mathbf{P}}$.

It follows from (9) and (i) that $\mathbf{T}(\boldsymbol{X}, \mathbf{P}) \geqslant \mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{P}) = 0$ for all $\mathbf{P} \in \mathscr{M}$ and consequently any element $\widehat{\mathbf{P}}$ that satisfies $0 \leqslant \mathbf{T}(\boldsymbol{X}, \widehat{\mathbf{P}}) \leqslant \epsilon$ is an ℓ-estimator.

3.3. **Risk bounds of an ℓ-estimator on a model.** As suggested by the previous heuristics, the performance of our estimator will depend on the closeness of $\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})$ from its expectation which itself depends on the behaviour of the process $\overline{\mathbf{Z}}$ defined on $\mathscr{M}^2$ by:

$$\begin{aligned} (\overline{\mathbf{P}}, \mathbf{Q}) \mapsto \overline{\mathbf{Z}}(\boldsymbol{X}, \overline{\mathbf{P}}, \mathbf{Q}) &= \mathbf{T}(\boldsymbol{X}, \overline{\mathbf{P}}, \mathbf{Q}) - \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{P}}, \mathbf{Q})\right] \\ &= \sum_{i=1}^{n} \left[\phi_{(\overline{P}_i, Q_i)}(X_i) - \mathbb{E}\left[\phi_{(\overline{P}_i, Q_i)}(X_i)\right]\right]. \end{aligned} \tag{14}$$

To analyze this process, we introduce the following sets, to be called *balls* in the sequel, even if ℓ is not a distance

$$\mathscr{B}(\mathbf{P}^\star, y) = \{\boldsymbol{Q} \in \mathscr{M},\ \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q}) \leqslant y\} \quad \text{for } y \geqslant 0. \tag{15}$$

We then define the quantity

$$\mathbf{w}(\overline{\mathbf{P}}, y) = \mathbb{E}\left[\sup_{\mathbf{Q} \in \mathscr{B}(\mathbf{P}^\star, y)} \left|\overline{\mathbf{Z}}(\boldsymbol{X}, \overline{\mathbf{P}}, \mathbf{Q})\right|\right]. \tag{16}$$

The following theorem, to be proven in Section 9 below and which generalizes Theorem 1 of Baraud and Birgé (2018), shows that the performance of an ℓ-estimator only depends on the approximation quality of the model $\mathscr{M}$ and the properties of the process $\overline{\mathbf{Z}}$ respectively described by the functions $\overline{\mathbf{P}} \mapsto \boldsymbol{\ell}(\mathbf{P}^\star, \overline{\mathbf{P}})$ and $\mathbf{w}$.

Theorem 1. *Let Assumption 1 be satisfied, $\overline{\mathbf{P}}$ be an arbitrary element of $\mathscr{M}$, $\kappa \in (0,1)$, $c_0 = \kappa a_1/6$ and set*

$$v(\overline{\mathbf{P}}) = \left[\frac{1}{\sqrt{n}} \sup\left\{y > 0 \,\middle|\, \mathbf{w}(\overline{\mathbf{P}}, y) > c_0 y\right\}\right] \bigvee \frac{1}{c_0\sqrt{2}}. \tag{17}$$

Any ℓ-estimator $\widehat{\mathbf{P}}$, i.e. any element of the random set $\mathscr{E}(\boldsymbol{X})$ defined by (13), satisfies, with probability at least $1-0.37e^{-\xi}$,

$$\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}}) \leqslant \frac{2}{1-\kappa}\left[\frac{a_0}{a_1}\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}}) + \sqrt{n}\left(\frac{\kappa}{3}v(\overline{\mathbf{P}}) + \sqrt{\frac{\xi}{2a_1^2}}\right) + \frac{\epsilon}{2a_1}\right] - \boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M})$$

for all $\xi>0$ and all $\mathbf{P}^\star\in\mathscr{P}$.

The proof of this theorem is postponed to Section 9. Integrating with respect to $\xi>0$ and using the fact that $\overline{\mathbf{P}}$ is arbitrary in $\mathscr{M}$, we derive from Theorem 1 an upper bound for the integrated loss of the form

$$\mathbb{E}\left[\frac{1}{n}\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}})\right] \leqslant C\inf_{\overline{\mathbf{P}}\in\mathscr{M}}\left[\frac{1}{n}\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}}) + \frac{v(\overline{\mathbf{P}})}{\sqrt{n}}\right] \tag{18}$$

where C depends on the constants a_0, a_1 and the choices of κ and ϵ. The quantity $v(\overline{\mathbf{P}})$ is related to the complexity of the model $\mathscr{M}$ in the neighbourhood of $\overline{\mathbf{P}}$. This notion of complexity is related to that introduced by V. Koltchinskii (2006) in risk minimization in statistical learning. Note that the minimum in (18) might not be achieved for the best approximation point of $\mathbf{P}^\star$ in $\mathscr{M}$ but rather by some element $\overline{\mathbf{P}}\in\mathscr{M}$ that provides the best tradeoff between approximation and complexity at that point. However, in many situations, the quantity $v(\overline{\mathbf{P}})$ can be bounded uniformly over $\mathscr{M}$ by some quantity $v_n = v(\mathscr{M}, n)$ that only depends on n and the model. In this case (18) leads to

$$C^{-1}\mathbb{E}\left[\frac{1}{n}\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}})\right] \leqslant \inf_{\overline{\mathbf{P}}\in\mathscr{M}}\frac{1}{n}\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}}) + \frac{v_n}{\sqrt{n}} = \frac{1}{n}\boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M}) + \frac{v_n}{\sqrt{n}}. \tag{19}$$

The quantity $v_n/\sqrt{n}$ corresponds to the bound we would get if $\mathbf{P}^\star$ did belong to $\mathscr{M}$ while $\boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M})/n$ corresponds to an approximation term due to a possible misspecification of the model. When $\mathscr{M}$ is a product of models for each marginal, i.e. is of the form

$$\mathscr{M} = \{P_1\otimes\ldots\otimes P_n \text{ with } P_1\in\mathscr{M}_1,\ldots,P_n\in\mathscr{M}_n\} \tag{20}$$

with $\mathscr{M}_i\subset\mathscr{M}$ for $i\in\{1,\ldots,n\}$, then

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

In density estimation where $\mathscr{M}_i=\mathscr{M}$ for all i, (19) becomes

$$\mathbb{E}\left[\frac{1}{n}\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}})\right] \leqslant C\left[\frac{1}{n}\sum_{i=1}^n\ell(P_i^\star,\mathscr{M}) + \frac{v_n}{\sqrt{n}}\right]. \tag{21}$$

Note that the approximation term can be small even in a situation where none of the true marginals $P_i^\star$ belongs to $\mathscr{M}$.

We immediately notice that the bounds (19) and (21) become much simpler when $\boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M})=0$ which suggests the introduction of the following notations to be used throughout the paper and which generalize the case of $\mathbf{P}^\star \in \mathscr{M}$ ($\mathbf{P}^\star$ belongs to the model).

$$\overline{\mathscr{M}} = \{\mathbf{P} \in \mathscr{P} \mid \ell(\mathbf{P}, \mathscr{M})\} = 0 \quad \text{and} \quad \overline{\mathscr{M}} = \{P \in \mathscr{P} \mid \ell(P, \mathscr{M})\} = 0 \tag{22}$$

In particular, when the data are truly i.i.d. with distribution $P^\star \in \overline{\mathscr{M}}$, then

$$\frac{1}{n}\boldsymbol{\ell}(\mathbf{P}^\star, \widehat{\mathbf{P}}) = \ell(P^\star, \widehat{P}), \qquad \frac{1}{n}\sum_{i=1}^{n} \ell(P_i^\star, \mathscr{M}) = 0$$

and we deduce from (21) that

$$\sup_{P^\star \in \overline{\mathscr{M}}} \mathbb{E}\left[\ell(P^\star, \widehat{P})\right] \leqslant \frac{Cv_n}{\sqrt{n}}.$$

This means that when v_n is independent of n, the minimax rate over $\overline{\mathscr{M}}$ is at most of order $1/\sqrt{n}$. This bound can be improved in the following way under Assumption 2.

Theorem 2. *Let Assumption 2 be satisfied, $\overline{\mathbf{P}}$ be an arbitrary element of $\mathscr{M}$, $\kappa \in (0,1)$, $c_1 = 2\kappa^2 a_1^2/(225 a_2)$ and set*

$$D(\overline{\mathbf{P}}) = \sup\left\{y > 0 \,\middle|\, \mathbf{w}(\overline{\mathbf{P}}, y) > c_1 y\right\} \vee c_1^{-1}. \tag{23}$$

Any ℓ-estimator $\widehat{\mathbf{P}}$ satisfies, whatever $\mathbf{P}^\star \in \mathscr{P}^{\otimes n}$ and $\xi > 0$,

$$\begin{aligned}
&\boldsymbol{\ell}(\mathbf{P}^\star, \widehat{\mathbf{P}}) \\
&\quad \leqslant \frac{2}{1-\kappa}\left[\left(\frac{a_0}{a_1} + \frac{2\kappa}{15}\right)\boldsymbol{\ell}(\mathbf{P}^\star, \overline{\mathbf{P}}) + \frac{\kappa}{3}D(\overline{\mathbf{P}}) + 2\left(\frac{1}{a_1} + \frac{15 a_2}{2\kappa a_1^2}\right)\xi + \frac{\epsilon}{2a_1}\right] \\
&\qquad - \boldsymbol{\ell}(\mathbf{P}^\star, \mathscr{M}),
\end{aligned}$$

with probability at least $1 - 0.42 e^{-\xi}$.

The proof of this theorem is also postponed to Section 9. As does $v(\overline{\mathbf{P}})$, the quantity $D(\overline{\mathbf{P}})$ measures the complexity of the model $\mathscr{M}$ in the neighbourhood of $\overline{\mathbf{P}}$. In the common situation where $D(\overline{\mathbf{P}})$ can be bounded by some positive number D_n independently of $\overline{\mathbf{P}}$, we may derive from Theorem 2 an upper bound for the integrated risk of the form

$$\mathbb{E}\left[\frac{1}{n}\boldsymbol{\ell}(\mathbf{P}^\star, \widehat{\mathbf{P}})\right] \leqslant C\left[\frac{1}{n}\sum_{i=1}^{n}\ell(P_i^\star, \mathscr{M}_i) + \frac{D_n}{n}\right]$$

for some positive constant C depending on a_0, a_1, a_2 and the choices of κ and ϵ. In density estimation, if $\mathbf{P}^\star = (P^\star)^{\otimes n}$ with $P^\star \in \overline{\mathscr{M}}$ and D_n is independent of n, we derive that the minimax rate over $\overline{\mathscr{M}}$ with respect to the loss ℓ is not larger than $1/n$ (up to a numerical constant). This is an improvement over inequality (21) which is solely based on Assumption 1.

4. Examples of ℓ-estimators and their performances

4.1. **Building suitable families $\mathscr{T}(\ell,\mathscr{M})$.** In order to apply Theorems 1 or 2, we have to find families $\mathscr{T}(\ell,\mathscr{M})$ which satisfy Assumptions 1 or 2. Let us first explain how to build such families for three of our loss functions, i.e. Wasserstein, $\mathbb{L}_j$ and TV which share the property that they can be defined via a variational formula. Let us more precisely assume the following: the loss function ℓ is defined on $\overline{\mathscr{P}}\times\overline{\mathscr{P}}$ where $\overline{\mathscr{P}}$ denotes a convex subset of the space of finite and possibly signed measures on $(E,\mathcal{E})$ and ℓ takes the form

$$\ell(P,Q)=\sup_{f\in\mathcal{F}}\left[\int_E f\,dP-\int_E f\,dQ\right], \tag{24}$$

where $\mathcal{F}$ is a class of measurable functions on $(E,\mathcal{E})$. We moreover require that the following assumption be satisfied:

Assumption 3. *The class $\mathcal{F}$ is symmetric, i.e. if $f\in\mathcal{F}$ then $-f\in\mathcal{F}$, and it contains 0. For all P,Q in $\overline{\mathscr{P}}$ there exists a function $f^\star(P,Q)\in\mathcal{F}$ such that*

$$\sup_{f\in\mathcal{F}}\left[\int_E f\,dP-\int_E f\,dQ\right]=\int_E f^\star(P,Q)dP-\int_E f^\star(P,Q)dQ. \tag{25}$$

Finally, there exists $b<+\infty$ such that

$$\sup_{(P,Q)\in\overline{\mathscr{P}}^2}\|f^\star(P,Q)\|_\infty=\sup_{(P,Q)\in\overline{\mathscr{P}}^2,\,x\in E}|f^\star(P,Q)(x)|\leqslant b. \tag{26}$$

Note that one can always take $f^\star(Q,P)=-f^\star(P,Q)$ and $f^\star(P,P)=0$ since $\mathcal{F}$ is symmetric and contains 0. Under Assumption 3, it is easy to check that the loss function ℓ is nonnegative, symmetric and satisfies the triangle inequality. It therefore satisfies all the requirements for being a distance except from the fact that $\ell(P,Q)=0$ does not necessarily implies that $P=Q$.

Given a model $\mathscr{M}\subset\overline{\mathscr{P}}$ we consider the family a functions $\{f_{(P,Q)},(P,Q)\in\mathscr{M}^2\}\subset\mathcal{F}$ where $f_{(P,Q)}=f^\star(P,Q)$ is given by (25), $f_{(Q,P)}=-f_{(P,Q)}$ and $f_{(P,P)}=0$. We then set

$$\phi_{(P,Q)}=\frac{1}{2b}\left\{\int_E f_{(P,Q)}\frac{dP+dQ}{2}-f_{(P,Q)}\right\}, \tag{27}$$

with b provided by (26). We can now derive the following result to be proven in Section 10.2.

Proposition 2. *Let $\mathcal{F}$ be a class of measurable functions that satisfies Assumption 3 and ℓ the loss function defined by (24). Assume that our set of probabilities $\mathscr{P}$ and our model $\mathscr{M}$ are subsets of $\overline{\mathscr{P}}$. The family $\mathscr{T}(\ell,\mathscr{M})$ which consists of all functions $\phi_{(P,Q)}$ defined by (27) for $P,Q\in\mathscr{M}$ satisfies Assumption 1 with $a_0=3/(4b)$ and $a_1=1/(4b)$.*

We shall now be able to deal successively with the Wasserstein, $\mathbb{L}_j$ and TV-losses which do satisfy Assumption 3.

4.2. **The Wasserstein loss.** As already seen in (6), the 1-Wasserstein distance between P and Q satisfies the variational formula

$$W(P,Q) = \sup_{f\in\mathcal{F}} [\mathbb{E}_P(f) - \mathbb{E}_Q(f)] \tag{28}$$

where $\mathcal{F}$ is the class of 1-Lipschitz functions on $\mathbb{R}$. Actually we can restrict the class $\mathcal{F}$ to those functions which are 1-Lipschitz and bounded by 1, as shown by the following result to be proven in Section 10.7.

Proposition 3. *The supremum in (28) is reached for the function*

$$f_{(P,Q)}(x) = \int_0^x \left[\mathbb{1}_{F_Q(t)>F_P(t)} - \mathbb{1}_{F_P(t)>F_Q(t)}\right] dt \quad \text{for all } x \in \mathbb{R}, \tag{29}$$

which is 1-Lipschitz and satisfies $\|f_{(P,Q)}\|_\infty \leqslant 1$.

In this case a suitable family $\mathscr{T}(\ell,\mathscr{M})$ can be defined according to (27) and the following result is an immediate consequence of Proposition 2 with $b = 1$.

Corollary 1. *Let $\mathscr{P}$ be the set of all probabilities on $([0,1],\mathscr{B}([0,1]))$, $\mathscr{M}$ a countable subset of $\mathscr{P}$ and ℓ the loss defined for $P,Q \in \mathscr{P}$ by $\ell(P,Q) = W(P,Q)$. The family $\mathscr{T}(\ell,\mathscr{M})$ of functions $\phi(P,Q)$ given by*

$$\phi_{(P,Q)} = \frac{1}{2}\left\{\frac{1}{2}\left(\mathbb{E}_P\left[f_{(P,Q)}\right] + \mathbb{E}_Q\left[f_{(P,Q)}\right]\right) - f_{(P,Q)}\right\} \quad \text{for } P,Q \in \mathscr{M}, \tag{30}$$

with $f_{(P,Q)}$ defined by (29), satisfies Assumption 1 with $a_0 = 3/4$ and $a_1 = 1/4$.

The following result holds.

Proposition 4. *Let P and Q be two probabilities in $\mathscr{P}$ with distribution functions F_P and F_Q respectively. Then,*

$$\begin{aligned}
&\mathbf{T}(\boldsymbol{X}, P^{\otimes n}, Q^{\otimes n}) \\
&\quad = \frac{1}{2}\int_0^1 \left[\mathbb{1}_{F_Q(t)>F_P(t)} - \mathbb{1}_{F_P(t)>F_Q(t)}\right]\left[\widehat{F}_n(t) - \frac{F_P(t)+F_Q(t)}{2}\right] dt,
\end{aligned}$$

where $\widehat{F}_n$ denotes the empirical distribution function.

Proof. It follows from (29) that for all $i \in \{1,\ldots,n\}$,

$$\begin{aligned}
f_{(P,Q)}(X_i) &= \int_0^1 \left[\mathbb{1}_{F_Q(t)>F_P(t)} - \mathbb{1}_{F_P(t)>F_Q(t)}\right] \mathbb{1}_{t<X_i} dt \\
&= \int_0^1 \left[\mathbb{1}_{F_Q(t)>F_P(t)} - \mathbb{1}_{F_P(t)>F_Q(t)}\right] [1 - \mathbb{1}_{X_i\leqslant t}]\, dt
\end{aligned}$$

and for all probability $R \in \mathscr{P}$ with distribution function F_R,

$$\mathbb{E}_R\left[f_{(P,Q)}(X)\right] = \int_0^1 \left[\mathbb{1}_{F_Q(t)>F_P(t)} - \mathbb{1}_{F_P(t)>F_Q(t)}\right] [1 - F_R(t)]\, dt.$$

Hence, for all $i \in \{1, \ldots, n\}$

$$\begin{aligned}\phi_{(P,Q)}(X_i) &= \frac{1}{2}\left\{\frac{1}{2}\left(\mathbb{E}_P\left[f_{(P,Q)}\right] + \mathbb{E}_Q\left[f_{(P,Q)}\right]\right) - f_{(P,Q)}(X_i)\right\} \\ &= \frac{1}{2}\left[\int_0^1 \left[\mathbb{1}_{F_Q(t)>F_P(t)} - \mathbb{1}_{F_P(t)>F_Q(t)}\right]\left[\mathbb{1}_{X_i \leqslant t} - \frac{F_Q(t)+F_P(t)}{2}\right]\right]\end{aligned}$$

and since $\mathbf{T}(\boldsymbol{X}, P^{\otimes n}, Q^{\otimes n}) = \sum_{i=1}^n \phi_{(P,Q)}(X_i)$, the result follows by summing over $i \in \{1, \ldots, n\}$. □

Example 1. Let $\mathscr{P}$ be the set of all probability distributions on $[0, 1]$ and our observations $X_1, \ldots, X_n$ be independent with values in $[0, 1]$ but presumably i.i.d. with a common distribution close to a model $\mathscr{M} \subset \mathscr{P}$. Our aim is to estimate $\mathbf{P}^\star$ using the Wasserstein loss. One can then derive the following consequence of Theorem 1.

Corollary 2. *Whatever the model $\mathscr{M}$ and $\xi > 0$, any ℓ-estimator $\widehat{P} \in \mathscr{M}$ satisfies, with probability at least $1 - 0.37e^{-\xi}$,*

$$\frac{1}{n}\sum_{i=1}^n W(P_i^\star, \widehat{P}) \leqslant \inf_{P \in \mathscr{M}} \frac{5}{n}\sum_{i=1}^n W(P_i^\star, P) + \frac{8}{\sqrt{n}}\left[\sqrt{2} + \sqrt{\frac{\xi}{2}} + \frac{\epsilon}{2\sqrt{n}}\right].$$

If, in particular, the data are truly i.i.d. with distribution $P^\star \in \overline{\mathscr{M}}$, for all $\xi > 0$ and with probability at least $1 - 0.37e^{-\xi}$,

$$W(P^\star, \widehat{P}) \leqslant \frac{8}{\sqrt{n}}\left[\sqrt{2} + \sqrt{\frac{\xi}{2}} + \frac{\epsilon}{2\sqrt{n}}\right].$$

The proof of this corollary is postponed to Section 10.10. Note that the bound does not depend on the complexity of the model $\mathscr{M}$ which can therefore be as large as desired in $\mathscr{P}$ provided that it remains countable.

Let us notice that, whenever the empirical measure $\widehat{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ belongs to the model $\mathscr{M}$, it is an ℓ-estimator. It indeed follows from Proposition 4 that for all $Q \in \mathscr{M}$,

$$\begin{aligned}&\mathbf{T}(\boldsymbol{X}, \widehat{P}_n^{\otimes n}, Q^{\otimes n}) \\ &\quad = \frac{1}{2}\int_0^1 \left[\mathbb{1}_{F_Q(t)>\widehat{F}_n(t)} - \mathbb{1}_{\widehat{F}_n(t)>F_Q(t)}\right]\left[\widehat{F}_n(t) - \frac{1}{2}\left(\widehat{F}_n(t) + F_Q(t)\right)\right] dt \\ &\quad = \frac{1}{4}\int_0^1 \left[\mathbb{1}_{F_Q(t)>\widehat{F}_n(t)} - \mathbb{1}_{\widehat{F}_n(t)>F_Q(t)}\right]\left[\widehat{F}_n(t) - F_Q(t)\right] dt \\ &\quad = -\frac{1}{4}\int_0^1 \left|\widehat{F}_n(t) - F_Q(t)\right| dt.\end{aligned}$$

Hence $\sup_{Q \in \mathscr{M}} \mathbf{T}(\boldsymbol{X}, \widehat{P}_n^{\otimes n}, Q^{\otimes n}) = 0$ if $\widehat{P}_n \in \mathcal{M}$, which proves that $\widehat{P}_n$ is an ℓ-estimator.

4.3. **The $\mathbb{L}_j$-loss for $j \in (1, +\infty)$.** It is well-known that

$$\ell_j(P, Q) = \sup_{f \in \mathcal{F}} \int_E (p - q) f d\mu, \tag{31}$$

which is (24), where $\mathcal{F}$ is the class of functions $f \in \mathscr{L}_{j'}(E, \mu)$ satisfying $\|f\|_{j'} \leqslant 1$ where j' is the conjugate exponent $j/(j-1)$ of j. It follows from Hölder inequality (actually from the case of equality), that the supremum in (31) is reached for

$$f_{(P,Q)} = \frac{(p-q)_+^{j-1} - (p-q)_-^{j-1}}{\|p-q\|_j^{j-1}} \quad \text{when } P \neq Q \quad \text{and} \quad f_{(P,P)} = 0. \tag{32}$$

In particular, when $j = 2$, we get $f_{(P,Q)} = (p - q)/\|p - q\|_2$.

Corollary 3. *Let $j \in (1, +\infty)$. Assume that the set of probabilities $\mathscr{P}$ and the countable model $\mathscr{M}$ are two subsets of $\overline{\mathscr{P}}_j$ and that there exists a number $R > 0$ such that*

$$\|p - q\|_\infty \leqslant R \|p - q\|_j \quad \textit{for all } P, Q \in \mathscr{M}. \tag{33}$$

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

$$\phi_{(P,Q)} = \frac{1}{2R^{j-1}} \left\{ \frac{1}{2} \int_E f_{(P,Q)} (p + q) \, d\mu - f_{(P,Q)} \right\} \tag{34}$$

with $f_{(P,Q)}$ given by (32), satisfies Assumption 1 with $a_0 = 3/(4R^{j-1})$ and $a_1 = 1/(4R^{j-1})$ for the loss ℓ_j.

Proof. It follows from (32) and (33) that $\left\| f_{(P,Q)} \right\|_\infty = \|p - q\|_\infty^{j-1} \|p - q\|_j^{1-j} \leqslant R^{j-1}$ and we simply apply Proposition 2 with $b = R^{j-1}$. □

When $j = 2$, we get

$$\phi_{(P,Q)} = \frac{\left(2q - \|q\|_2^2\right) - \left(2p - \|p\|_2^2\right)}{4R \|p - q\|_2}, \tag{35}$$

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

$$\begin{aligned} \frac{4R\|p-q\|_2}{n} & \mathbf{T}(\boldsymbol{X}, P^{\otimes n}, Q^{\otimes n}) \\ &= \left[\frac{2}{n} \sum_{i=1}^n q(X_i) - \|q\|_2^2 \right] - \left[\frac{2}{n} \sum_{i=1}^n p(X_i) - \|p\|_2^2 \right]. \end{aligned}$$

4.3.1. *The quadratic loss and linear models of densities.* In this section, we assume that the marginal distributions $P_i^\star$ of the data $\boldsymbol{X} = (X_1, \ldots, X_n)$ admit densities $p_i^\star$ with respect to some positive dominating measure μ and that $p_1^\star, \ldots, p_n^\star$ belong to $\mathscr{L}_2(E, \mu)$. Our set $\mathscr{P}$ is the set of all probabilities $P = p \cdot \mu$ with $p \in \mathscr{L}_2(E, \mu)$. We shall consider the density framework, assuming that the data are i.i.d., even though this might not be true. The presumed common density of the data will be approximated by a model

$\mathcal{M} \subset \mathscr{L}_2(E,\mu) \cap \mathscr{L}_1(E,\mu)$ which may contain functions p that might not be probability densities. For $\mathbf{P} \in \mathscr{P}^n$ and $\mathbf{Q} \in \mathscr{M}^n$, according to (7), we consider the loss function

$$\boldsymbol{\ell}_2(\mathbf{P},\mathbf{Q}) = \sum_{i=1}^{n} \ell_2(p_i,q_i) = \sum_{i=1}^{n} \left\|p_i - q_i\right\|_2,$$

where p_i and q_i denote, as usual, versions of the densities of the components P_i and Q_i with respect to μ for all $i \in \{1,\ldots,n\}$. We shall moreover assume here that $\mathcal{M}$ is a subset of some linear subspace V of $\mathscr{L}_2(E,\mu)$ such that the restriction to V of $\|\cdot\|_2$ is a norm which turns V into a Hilbert space satisfying the following property:

Assumption 4. *There exists a positive number R such that*

$$\|t\|_\infty \leqslant R\,\|t\|_2 \quad \text{for all } t \in V. \tag{36}$$

When E is a compact space, typically $[0,1]$, this assumption is met for many finite dimensional spaces with good approximation properties as shown in Birgé and Massart (1998)[Section 3]. Nevertheless, our approach allows us to consider more general situations where the set E is not compact and V possibly infinite dimensional. Illustrations will be given in Section 4.3.2.

In this framework, we shall use the family $\mathscr{T}(\ell,\mathscr{M})$ of functions given by (35) to build our ℓ-estimator. Its performance is given by the following result to be proven in Section 10.15.

Corollary 4. *Assume that $\mathcal{M}$ is a subset of a Hilbert space $V \subset \mathscr{L}_2(E,\mu) \cap \mathscr{L}_1(E,\mu)$ which satisfies Assumption 4. Any ℓ-estimator $\widehat{P} = \widehat{p}\cdot\mu$ for the $\mathbb{L}_2$-loss based on $\mathcal{M}$ satisfies for all $\xi > 0$*

$$\frac{1}{n}\sum_{i=1}^{n}\|p_i^\star - \widehat{p}\|_2 \leqslant \inf_{p\in\mathcal{M}}\left[\frac{5}{n}\sum_{i=1}^{n}\|p_i^\star - p\|_2\right] + \left[1 + \frac{\sqrt{\xi}}{2} + \frac{\epsilon}{2\sqrt{2n}}\right]\frac{8R\sqrt{2}}{\sqrt{n}}, \tag{37}$$

with probability at least $1 - 0.37e^{-\xi}$. In the case of truly i.i.d. observations X_i with density $p^\star$, we get

$$\mathbb{E}\left[\|p^\star - \widehat{p}(\boldsymbol{X})\|_2^2\right] \leqslant C\left[\inf_{p\in\mathcal{M}}\|p^\star - p\|_2^2 + \left[1 + \sqrt{\xi} + \frac{\epsilon}{\sqrt{n}}\right]^2 \frac{R^2}{n}\right], \tag{38}$$

for some universal constant $C > 0$.

It is important to notice that the bound we get does not depend on the dimension of the linear space V (which can therefore be infinite) but rather on the constant R that controls the ratio between the sup-norm and the $\mathbb{L}_2$-norm on V.

In density estimation when $p_i^\star = p^\star \in \mathscr{L}_2(\mu)$ for all i, there exists a large amount of literature on the problem of estimating $p^\star$ with respect to the $\mathbb{L}_2$-norm. A nice feature of (37) lies in the fact that it does not involve the sup-norm of the density $p^\star$ which may therefore be unbounded. Birgé and Massart (1998)[Theorem 2 p. 343] studied the property of the projection estimator on finite dimensional linear spaces V satisfying (36), typically linear spaces of functions on $[0,1]^d$. Our result generalizes theirs.

4.3.2. *Risk bounds for the quadratic loss over Besov spaces.* In this section we consider the problem of estimating a density $p^\star$ with respect to the Lebesgue measure on $E = \mathbb{R}^d$, under the assumption that $p^\star$ is close to a given Besov space $B^\alpha_{s,\infty}(\mathbb{R}^d)$ with $\alpha > 0$ and $s \in [2, +\infty)$. We refer to Meyer (1992) for a definition of these classes of functions and to Giné and Nickl (2016) Section 4.3.6 for their characterisation in terms of coefficients in a suitable wavelet basis. We choose the $\mathbb{L}_2$-norm as our loss function.

Proposition 5. *Let $s \geqslant 2$, $d \geqslant 1$ and $\alpha > 0$. There exist two constants K, K' depending on d, α and s with the following properties. For all $J \geqslant 0$, there exists a linear subspace V_J of $\mathscr{L}_1(\mathbb{R}^d) \cap \mathscr{L}_2(\mathbb{R}^d)$ such that $(V_J, \|\cdot\|_2)$ is a Hilbert space satisfying Assumption 4 with $R = K2^{Jd/2}$ and, for all $f \in B^\alpha_{s,\infty}(\mathbb{R}^d) \cap \mathscr{L}_1(\mathbb{R}^d) \cap \mathscr{L}_2(\mathbb{R}^d)$,*

$$\inf_{t \in V_J} \|f - t\|_2^2 \leqslant K' \, |f|_{\alpha,s,\infty}^{s/(s-1)} \, \|f\|_1^{(s-2)/(s-1)} \, 2^{-Js\alpha/(s-1)} \tag{39}$$

where $|f|_{\alpha,s,\infty}$ is the Besov semi-norm of f in $B^\alpha_{s,\infty}(\mathbb{R}^d)$.

The proof of this approximation result is postponed to Section 10.8. In the right-hand side of (39), we use the convention $0^0 = 0$ when $s = 2$ and $\|f\|_1 = 0$. Note that this approximation bound neither depends on the $\mathbb{L}_2$-norm nor on the sup-norm of f which may therefore be arbitrarily large.

Corollary 5. *Let $s \geqslant 2$, $\alpha > 0$, $r > 0$, $d \geqslant 1$ and $\mathcal{F}^d_{\alpha,s,\infty}(r)$ be the class of all probability densities p on $\mathbb{R}^d$ that belong to $B^\alpha_{s,\infty}(\mathbb{R}^d) \cap \mathscr{L}_2(\mathbb{R}^d)$ and such that their Besov semi-norms are bounded by $r > 0$. There exists an ℓ-estimator (ℓ being the $\mathbb{L}_2$-loss) $\widehat{p}$ that satisfies, whatever the density $p^\star$ of the X_i,*

$$\mathbb{E}\left[\|p^\star - \widehat{p}(\boldsymbol{X})\|_2^2\right] \leqslant C \left[\inf_{p \in \mathcal{F}^d_{\alpha,s,\infty}(r)} \|p^\star - p\|_2^2 + \frac{r^{ds/[d(s-1)+s\alpha]}}{n^{\alpha s/[d(s-1)+s\alpha]}} + \frac{1}{n}\right],$$

where C is a positive number that depends on s, d, α and ϵ only.

An interesting feature of this result lies in the fact that the class $\mathcal{F}^d_{\alpha,s,\infty}(r)$ contains densities that are neither compactly supported nor bounded in supremum norm when $\alpha < 1/s$. We are not aware of many results in this direction. When $d = 1$ and for $r, r' > 0$, the bound we get is known to be optimal (up to a constant that depends on r', α and s) over the smaller set of densities $p^\star$ which satisfy $\|p^\star\|_2 \vee \|p^\star\|_\infty \leqslant r'$ and belong to $B^\alpha_{s,\infty}(\mathbb{R})$ with Besov norms bounded by r. We refer the reader to Rivoirard *et al.* (2011)[Theorem 4] and the references therein. Besides, the authors obtain there (see their Theorem 3) an upper bound which is similar to ours. However there exist a few differences between their bound and ours: our result does not require that the densities $p^\star$ be uniformly bounded in $\mathbb{L}_2(\mathbb{R})$ and it includes the case where $s = 2$ while theirs is only true for $s > 2$. Their estimator is adaptive with respect to the parameters of the Besov space while ours is not. This could explain the extra-logarithmic factor that appears in their risk bound. Nevertheless, we believe that this extra-logarithmic factor is actually not necessary for adaptation.

Proof. Throughout this proof, we fix some probability density $\overline{p}$ in $\mathcal{F}^d_{\alpha,s,\infty}(r)$. Let J be the nonnegative integer which satisfies

$$2^J \leqslant 1 \vee \left(nr^{s/(s-1)}\right)^{(s-1)/[d(s-1)+s\alpha]} < 2^{J+1}$$

and V_J be the Hilbert space provided by Proposition 5 for this value of J. We consider the model of (signed) densities $\mathcal{M} = V_J$ (or more precisely a countable dense subset of it with respect to the $\mathbb{L}_2$-norm). Since by Proposition 5 the space V_J satisfies Assumption 4 with $R = K2^{Jd/2}$, Corollary 4 applies and (38) implies that an ℓ-estimator $\widehat{p}$ of $p^\star$ based on $\mathcal{M}$ satisfies

$$\mathbb{E}\left[\|p^\star - \widehat{p}(\boldsymbol{X})\|_2^2\right] \leqslant C_0 \left[\|p^\star - \overline{p}\|_2^2 + \inf_{p\in\mathcal{M}} \|p - \overline{p}\|_2^2 + \frac{2^{Jd}}{n}\right], \tag{40}$$

where C_0 is a positive constant that only depends on d, s, α and ϵ. Since $\overline{p}$ belongs to $B^\alpha_{s,\infty}(\mathbb{R}^d) \cap \mathscr{L}_2(\mathbb{R}^d)$ and satisfies $|\overline{p}|_{\alpha,s,\infty} \leqslant r$, it follows from (39) that we may choose $p \in \mathcal{M}$ such that

$$\|p - \overline{p}\|_2^2 \leqslant K' r^{s/(s-1)} 2^{-Js\alpha/(s-1)}$$

with a possibly enlarged value of K'. Our choice of J then implies that

$$\|p - \overline{p}\|_2^2 \leqslant K' \left(r^d n^{-\alpha}\right)^{s/[d(s-1)+s\alpha]} \quad \text{and} \quad \frac{2^{Jd}}{n} \leqslant \left(r^d n^{-\alpha}\right)^{s/[d(s-1)+s\alpha]} + \frac{1}{n}.$$

The final bound on $\mathbb{E}\left[\|p^\star - \widehat{p}(\boldsymbol{X})\|_2^2\right]$ follows from (40) and a minimization with respect to $\overline{p} \in \mathcal{F}^d_{\alpha,s,\infty}(r)$. □

4.3.3. *The $\mathbb{L}_j$-loss for models of piecewise constant functions.* Let us consider the $\mathbb{L}_j$-loss with $j \in (1, +\infty)$ and the problem of evaluating the performance of an ℓ-estimator relative to the set $\overline{\mathcal{M}}_D$ of all densities (with respect to some probability μ on E) which are piecewise constant on a fixed partition $\mathcal{I}$ of E into $D \geqslant 2$ pieces satisfying $\mu(I) = 1/D$ for all $I \in \mathcal{I}$. The following result to be proven in Section 10.11 holds.

Corollary 6. *Let $D \in \{1, \ldots, n\}$, $j \in (1, +\infty)$ and $\mathcal{M}$ be a countable and dense subset of $\overline{\mathcal{M}}_D$ (for the $\mathbb{L}_j$-norm). Assume that the data are i.i.d. with distribution $P^\star = p^\star \cdot \mu$ with $p^\star \in \mathbb{L}_j(\mu)$ and $j \in (1, +\infty)$ and set*

$$\|\overline{p}_D\|_{j/2} = \left[\int_E |\overline{p}_D|^{j/2}\, d\mu\right]^{2/j} \quad \textit{with} \quad \overline{p}_D = \sum_{I\in\mathcal{I}} \left[D \int_I p^\star(x)\, d\mu(x)\right] \mathbb{1}_I.$$

The ℓ-estimator $\widehat{p}$ of $p^\star$ on $\mathcal{M}$ for the $\mathbb{L}_j$-loss satisfies for some constant $C > 0$ depending on j only and all $\xi > 0$,

$$\|p^\star - \widehat{p}\|_j \leqslant C \left[\inf_{p \in \overline{\mathcal{M}}_D} \|p^\star - p\|_j + \left(1 + \sqrt{\xi} + \frac{\epsilon}{\sqrt{n}}\right) \sqrt{\frac{D}{n} \|\overline{p}_D\|_{j/2}} \right], \tag{41}$$

with probability at least $1 - 0.37e^{-\xi}$.

Up to a constant depending on ϵ, the quantity

$$B_{j,n,D}(p^\star) = \sqrt{\frac{D}{n} \|\overline{p}_D\|_{j/2}}$$

is the risk bound we would get if $p^\star$ did belong to $\overline{\mathcal{M}}_D$. Let us further analyze this quantity. The function $\overline{p}_D$ is a density which belongs to $\overline{\mathcal{M}}_D$. If we set $a_I = D \int_I p^\star(x)\, d\mu(x) \in [0, D]$ for all $I \in \mathcal{I}$, it satisfies

$$\|\overline{p}_D\|_{j/2}^{j/2} = \frac{1}{D} \sum_{I \in \mathcal{I}} a_I^{j/2} \qquad \text{and} \qquad \|\overline{p}_D\|_1 = \frac{1}{D} \sum_{I \in \mathcal{I}} a_I = 1.$$

In particular, $B_{2,n,D}(p^\star) = \sqrt{D/n}$. It follows from convexity arguments that for $j > 2$,

$$1 = \left(\frac{1}{D} \sum_{I \in \mathcal{I}} a_I \right)^{j/2} \leqslant \|\overline{p}_D\|_{j/2}^{j/2} \leqslant \max_{I \in \mathcal{I}} a_I^{j/2-1} \leqslant D^{j/2-1}$$

while for $j \in (1, 2)$,

$$D^{j/2-1} \leqslant \min_{I \in \mathcal{I}} a_I^{j/2-1} \leqslant \|\overline{p}_D\|_{j/2}^{j/2} \leqslant \left(\frac{1}{D} \sum_{I \in \mathcal{I}} a_I \right)^{j/2} = 1.$$

As a consequence,

$$\sqrt{\frac{D}{n}} \leqslant B_{j,n,D}(p^\star) \leqslant \frac{D^{1-1/j}}{\sqrt{n}} \quad \text{for } j > 2\ , \tag{42}$$

while

$$\frac{D^{1-1/j}}{\sqrt{n}} \leqslant B_{j,n,D}(p^\star) \leqslant \sqrt{\frac{D}{n}} \quad \text{for } j \in (1, 2). \tag{43}$$

The lower bound in (42) corresponds to the situation where $\overline{P}_D = \overline{p}_D \cdot \mu = \mu$ (uniform distribution on E with respect to μ) while the upper bound is achieved in the least favorable situation where $\overline{p}_D = D\mathbb{1}_I$ for some $I \in \mathcal{I}$ (uniform distribution on I). The situation is exactely the opposite when $j \in (1, 2)$. Note that when j approches 1 and the distribution $\overline{P}_D$ is uniform over one of the intervals $I \in \mathcal{I}$, the bound we get is almost of order $1/\sqrt{n}$.

5. The case of the TV-loss

5.1. **Building suitable families $\mathscr{T}(\ell,\mathscr{M})$.** It is well-known that the TV-distance $\|P-Q\|$ between two probabilities P,Q on $(E,\mathcal{E})$ given by (4) can equivalently be written as $\sup_{f\in\mathcal{F}}[\mathbb{E}_P(f)-\mathbb{E}_Q(f)]$ where $\mathcal{F}$ is the set of all measurable functions f with values in $[0,1]$. This class does not satisfy Assumption 3 by lack of symmetry but we can equivalently write

$$\|P-Q\|=\sup_{f\in\mathcal{F}}[\mathbb{E}_P(f)-\mathbb{E}_Q(f)], \tag{44}$$

where $\mathcal{F}$ is now the set of all measurable functions with values in $[-1/2,1/2]$, which is symmetric and satisfies $\sup_{f\in\mathcal{F}}\|f\|_\infty\leqslant 1/2$. The supremum in (44) is then reached for $f_{(P,Q)}=\mathbb{1}_{p>q}-1/2$ where p and q denote versions of the respective densities of P and Q with respect to some common dominating measure. It follows that (26) holds with $b=1/2$ and a straightforward application of Proposition 2 leads to

Corollary 7. *Let $\mathscr{P}$ be the set of all probabilities on $(E,\mathcal{E})$, $\mathscr{M}$ a subset of $\mathscr{P}$ and ℓ be the TV-loss defined by $\ell(P,Q)=\|P-Q\|$ for all $P,Q\in\mathscr{P}$. The family $\mathscr{T}(\ell,\mathscr{M})$ of functions $\phi_{(P,Q)}$ defined for all $P,Q\in\mathscr{M}$ by*

$$\phi_{(P,Q)}=\frac{1}{2}\left[P\left(p>q\right)+Q\left(p>q\right)\right]-\mathbb{1}_{p>q} \tag{45}$$

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

The ℓ-estimator based on $\mathcal{M}$ is therefore an ϵ-minimizer of the function defined for $p\in\mathcal{M}$ by

$$\sup_{q\in\mathcal{M}}\left[\frac{1}{2}\left[P\left(p>q\right)+Q\left(p>q\right)\right]-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{p>q}(X_i)\right]. \tag{46}$$

5.2. **Risk bounds based on VC-dimensions.** In this section, $\mathscr{P}$ is the set of all probabilities on $(E,\mathcal{E})$ and we pretend that our observations $X_1,\ldots,X_n$ are i.i.d. with a distribution $\overline{P}$ belonging to a statistical model $\mathscr{M}\subset\mathscr{P}$. Given some dominating measure μ, we recall that we may associate to $\mathscr{M}$ a family $\mathcal{M}$ of densities on E such that $\mathscr{M}=\{P=p\cdot\mu,\ p\in\mathcal{M}\}$. Given a density $\overline{p}\in\mathcal{M}$, we consider the following assumption relative to it.

Assumption 5. *The class of subsets of E: $\{\{x\,|\,\overline{p}(x)>q(x)\},\ q\in\mathcal{M}\}$ is VC with dimension not larger than $V(\overline{p})\geqslant 1$.*

We refer the reader to Dudley (1984) for the definition of the VC-dimension of a class of sets. The family of sets of the form $\{p>q\}$ with $p,q\in\mathcal{M}$ are known as the Yatracos class associated to $\mathcal{M}$. Assumption 5 is weaker than the usual assumption that the Yatracos class $\{\{x\,|\,p(x)>q(x)\},\ p,q\in\mathcal{M}\}$ is VC (see Devroye and Lugosi (2001) for example). In particular, we shall see how to take advantage of this weaker form in our Example 3.

Corollary 8. *Let $\overline{p} \in \mathcal{M}$ satisfy Assumption 5. For any ℓ–estimator $\widehat{P} \in \mathscr{M}$ and all $\xi > 0$, with a probability at least $1 - 0.37e^{-\xi}$,*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|P_i^\star - \widehat{P}\right\| \leqslant \frac{6}{n}\sum_{i=1}^{n}\left\|P_i^\star - \overline{P}\right\| - \inf_{P\in\mathscr{M}}\frac{1}{n}\sum_{i=1}^{n}\|P_i^\star - P\| \\ + 179\sqrt{\frac{V(\overline{p})}{n}} + \sqrt{\frac{8\xi}{n}} + \frac{2\epsilon}{n}. \tag{47}$$

In particular, if Assumption 5 is satisfied for all $\overline{p} \in \mathcal{M}$ and $\sup_{\overline{p}\in\mathcal{M}} V(\overline{p}) = V < +\infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|P_i^\star - \widehat{P}\right\| \leqslant \inf_{P\in\mathscr{M}}\frac{5}{n}\sum_{i=1}^{n}\|P_i^\star - P\| + 179\sqrt{\frac{V}{n}} + \sqrt{\frac{8\xi}{n}} + \frac{2\epsilon}{n}. \tag{48}$$

The proof of this corollary is postponed to Section 10.12.

When the X_i are truly i.i.d. with distribution $P^\star$, (48) becomes

$$\left\|P^\star - \widehat{P}\right\| \leqslant 5\inf_{P\in\mathscr{M}}\|P^\star - P\| + 179\sqrt{\frac{V}{n}} + \sqrt{\frac{8\xi}{n}} + \frac{2\epsilon}{n}. \tag{49}$$

Whenever $P^\star = p^\star \cdot \mu$ is absolutely continuous with respect to μ, the above result immediately translates into an upper bound on the $\mathbb{L}_1$-loss between the densities of $P^\star$ and $\widehat{P}$ via the well-known formula

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

Integrating (49) with respect to ξ, we deduce a risk bound of the form

$$\mathbb{E}\left[\int_E |p^\star - \widehat{p}|\, d\mu\right] \leqslant C\left[\inf_{p\in\mathcal{M}}\int_E |p^\star - p|\, d\mu + \sqrt{\frac{V}{n}}\right],$$

for some positive number $C > 0$. Up to the numerical constant $C > 0$, this bound is similar to that obtained for the minimum distance estimator in Devroye and Lugosi (2001).

When $P^\star$ belongs to $\overline{\mathscr{M}}$, i.e. the closure of $\mathscr{M}$ with respect to the TV-distance, (49) tells us that our ℓ-estimator provides an estimation accuracy of order $\sqrt{V/n}$ with probability close to 1. The approximation term $\inf_{P\in\mathscr{M}}\|P^\star - P\|$ shows that this accuracy is still of the same order when the distance between $P^\star$ and $\mathscr{M}$ remains small enough compared to $\sqrt{V/n}$, which means that the estimator possesses some robustness property with respect to some possible misspecification of the model. Inequality (48) shows that it is also robust with respect to the assumption that the data are truly i.i.d.: if the true distribution $\mathbf{P}^\star$ of $\boldsymbol{X}$ is close to some distribution $\overline{\mathbf{P}} = \overline{P}^{\otimes n}$

with $\overline{P}\in\mathscr{M}$, we deduce from (48) that, with probability at least $1-0.37e^{-\xi}$,

$$\begin{aligned}\left\|\overline{P}-\widehat{P}\right\| &\leqslant \frac{1}{n}\sum_{i=1}^{n}\left\|P_i^\star-\overline{P}\right\|+\frac{1}{n}\sum_{i=1}^{n}\left\|P_i^\star-\widehat{P}\right\|\\ &\leqslant \frac{6}{n}\sum_{i=1}^{n}\left\|P_i^\star-\overline{P}\right\|+179\sqrt{\frac{V}{n}}+\sqrt{\frac{8\xi}{n}}+\frac{2\epsilon}{n}. \qquad (50)\end{aligned}$$

The performance of $\widehat{P}$, viewed as an estimator of $\overline{P}\in\mathscr{M}$, is only slightly modified when $\mathbf{P}^\star\neq\overline{P}^{\otimes n}$ provided that $\sum_{i=1}^{n}\left\|P_i^\star-\overline{P}\right\|/n$ is small enough compared to $\sqrt{V/n}$.

Example 2. To illustrate the robustness property of the ℓ-estimator for the TV-loss, let us focus on the following problem. The observations are presumed to be i.i.d. and have a common Gaussian distribution $P_{m^\star}=\mathcal{N}(m^\star,I_d)$ in $\mathbb{R}^d$ with mean vector $m^\star$ and identity covariance matrix, but they are actually contaminated so that, for $1\leqslant i\leqslant n$, the true distribution of X_i is actually $P_i^\star=(1-\alpha_i)P_{m^\star}+\alpha_iR_i$ for some arbitrary probabilities R_i and small numbers $\alpha_i\in[0,1]$. We choose for our model the family $\mathscr{M}$ of Gaussian distributions P_m with mean $m\in\mathbb{Q}^d$ and identity covariance matrix. Denoting by p_m the density of P_m, we see that for all $m,\overline{m}\in\mathbb{Q}^d$, the set $\{x\in\mathbb{R}^d,\ p_m(x)\geqslant p_{\overline{m}}(x)\}$ is either $\mathbb{R}^d$ (when $\overline{m}=m$) or closed half-spaces of $\mathbb{R}^d$. The VC-dimension of this class is not larger than $V=d+1$ (see Devroye and Lugosi (2001), Corollary 4.2 page 33). The class of the complementary sets possess the same VC-dimension, hence Assumption 5 is satisfied with $V(\overline{p})=V=d+1$ for all $\overline{p}\in\mathcal{M}$. Besides, the following lemma to be proven in Section 10.13, allows to relate the TV-distance between P_m and $P_{m'}$ to the Euclidean distance between the parameters m and m'.

Lemma 1. *For all $m,m'\in\mathbb{R}^d$,*

$$\|P_m-P_{m'}\|=\mathbb{P}\left[|Z|\leqslant\left|m-m'\right|/2\right] \qquad (51)$$

where Z is a standard real-valued Gaussian random variable. Consequently,

$$0.78\min\left\{1,\frac{|m-m'|}{\sqrt{2\pi}}\right\}\leqslant\|P_m-P_{m'}\|\leqslant\min\left\{1,\frac{|m-m'|}{\sqrt{2\pi}}\right\}. \qquad (52)$$

This means that when m' is close to m the quantity $\|P_m-P_{m'}\|$ is of order $|m-m'|/\sqrt{2\pi}$ while it is of order 1 when m' is far away from m.

As a consequence of (50) with $\overline{P}=P_{m^\star}$ and (52) we deduce that, whatever $\xi>0$, with probability at least $1-0.37e^{-\xi}$, the ℓ-estimator $\widehat{P}=P_{\widehat{m}}$ satisfies

$$\begin{aligned}0.78\min\left\{1,\frac{|\overline{m}-\widehat{m}|}{\sqrt{2\pi}}\right\} &\leqslant \|P_{\overline{m}}-P_{\widehat{m}}\|\\ &\leqslant \frac{6}{n}\sum_{i=1}^{n}\|P_i^\star-P_{\overline{m}}\|+179\sqrt{\frac{d+1}{n}}+\sqrt{\frac{8\xi}{n}}+\frac{2\epsilon}{n}. \qquad (53)\end{aligned}$$

Since $\overline{m}$ can be arbitrarily close to $m^\star$ and the map $m \mapsto \|P_i^\star - P_m\|$ is continuous with respect to the Euclidean norm on $\mathbb{R}^d$ for all $i \in \{1, \ldots, n\}$, we obtain that, for $\xi > 0$, with probability at least $1 - 0.37e^{-\xi}$, the ℓ-estimator $\widehat{P} = P_{\widehat{m}}$ satisfies

$$\min\left\{1, \frac{|m^\star - \widehat{m}|}{\sqrt{2\pi}}\right\} \leqslant \frac{7.7}{n}\sum_{i=1}^{n}\|P_i^\star - P_{m^\star}\| + 230\sqrt{\frac{d+1}{n}} + \sqrt{\frac{13.2\xi}{n}} + \frac{2.6\epsilon}{n}. \tag{54}$$

In particular, since $P_i^\star - P_{m^\star} = \alpha_i(R_i - P_{m^\star})$ for each $i \in \{1, \ldots, n\}$

$$\min\left\{1, \frac{|m^\star - \widehat{m}|}{\sqrt{2\pi}}\right\} \leqslant \frac{7.7}{n}\sum_{i=1}^{n}\alpha_i + 230\sqrt{\frac{d+1}{n}} + \sqrt{\frac{13.2\xi}{n}} + \frac{2.6\epsilon}{n}, \tag{55}$$

which establishes some robustness property of the ℓ-estimator $\widehat{m}$ with respect to a possible contamination of the data.

When $\alpha_i = \alpha$ for all i, (55) is similar to the bound obtained in Gao *et al.* (2018)[Theorem 3.1] for TV-Gan in this setting.

An interesting feature of Corollary 8 and more precisely (47) lies in the fact that the upper bound involves the quantity $V(\overline{p})$ which may depend on the choice of $\overline{p}$. This means that the best choice of $\overline{p}$ in view of minimizing the right-hand side of (47) might not be the density of the best approximation point of $\mathbf{P}^\star$ in $\mathscr{M}$. From this point of view, (47) contrasts with (48) which requires that for all $\overline{p} \in \mathcal{M}$ this quantity be bounded independently of $\overline{p}$. This subtle difference allows us to deal with statistical models for which the quantity $V(\overline{p})$ can be very different from a density $\overline{p}$ to another and possibly even infinite for some $\overline{p}$. The following example provides a good illustration of this fact.

Example 3. Let us consider the problem of estimating a density which is presumably belonging to the set $\overline{\mathcal{M}}$ of all non-increasing densities supported by some unknown half line, i.e. densities (with respect to the Lebesgue measure λ) which are non-increasing on an interval of the form $(a, +\infty)$ with $a \in \mathbb{R}$ and vanish elsewhere. For $d \geqslant 1$, let $\overline{\mathcal{M}}_d$ be the subset of $\overline{\mathcal{M}}$ of those densities of the form $\overline{p} = \sum_{I \in \mathcal{I}} a_I \mathbb{1}_I$ where $\mathcal{I}$ is a set of at most d disjoint intervals with positive lengths and $a_I > 0$ for all $I \in \mathcal{I}$. In other words, $\overline{\mathcal{M}}_d$ is the set of all non-increasing piecewise constant densities the supports of which are the unions of at most d (non-trivial) intervals. We shall denote by $\overline{\mathscr{M}}_d = \{\overline{p} \cdot \lambda,\, \overline{p} \in \overline{\mathcal{M}}_d\}$ the corresponding set of probabilities and by $\mathcal{M}_d$ and $\mathcal{M}$ respectively some countable and dense subsets of $\overline{\mathcal{M}}_d$ and $\overline{\mathcal{M}}$ for the $\mathbb{L}_1(\lambda)$-distance. We shall assume with no loss of generality that $\mathcal{M}_d \subset \mathcal{M}$ for all $d \geqslant 1$.

Given $q \in \mathcal{M}$ and $\overline{p} \in \mathcal{M}_d$, the set $\{\overline{p} > q\}$ is the union of at most d intervals and by applying Lemma 1 in Baraud and Birgé (2016) we deduce that Assumption 5 is satisfied with $V(\overline{p}) = 2d$. We may then apply Corollary 8 with an arbitrary choice of $d \geqslant 1$ and $\overline{p} \in \mathcal{M}_d$ (with $\overline{P} = \overline{p} \cdot \lambda$). Since $\mathcal{M}_d$ is dense in $\overline{\mathcal{M}}_d$ for all $d \geqslant 1$, we get

Proposition 6. *For all $\xi > 0$, with a probability at least $1 - 0.37e^{-\xi}$, the ℓ–estimator $\widehat{P} = \widehat{p} \cdot \lambda$ based on $\mathcal{M}$ satisfies*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|P_i^\star - \widehat{P}\right\| \leqslant \inf_{d\geqslant 1}\left[\inf_{\overline{P}\in\overline{\mathscr{M}}_d}\frac{6}{n}\sum_{i=1}^{n}\left\|P_i^\star - \overline{P}\right\| + 179\sqrt{\frac{2d}{n}}\right] + \sqrt{\frac{8\xi}{n}} + \frac{2\epsilon}{n}. \tag{56}$$

In particular, if the data are i.i.d. with density $p^\star$,

$$\|p^\star - \widehat{p}\|_1 \leqslant \inf_{d\geqslant 1}\left[6\inf_{\overline{p}\in\overline{\mathcal{M}}_d}\|p^\star - \overline{p}\|_1 + 179\sqrt{\frac{2d}{n}}\right] + \sqrt{\frac{8\xi}{n}} + \frac{2\epsilon}{n} \tag{57}$$

with probability at least $1 - 0.37e^{-\xi}$, for all $\xi > 0$.

Let $\overline{\mathcal{M}}(H,L)$ be the subset of $\overline{\mathcal{M}}$ consisting of those densities p such that $I = \{x \,|\, p(x) > 0\}$ is an interval of length not larger than $L > 0$ and the variation of p on I, i.e. the quantity $\sup_{x\in I} p(x) - \inf_{x\in I} p(x)$, is not larger than $H \geqslant 0$. The following approximation result which is due to Birgé (1987)[see Section 2 pages 1014-1015] allows us to derive uniform risk bounds over $\overline{\mathcal{M}}(H,L)$.

Proposition 7. *Let $p \in \overline{\mathcal{M}}(H,L)$ with $H \geqslant 0$ and $L > 0$. For each $d \geqslant 1$, there is an histogram $\overline{p}_d \in \overline{\mathcal{M}}_d$ such that*

$$\|p - \overline{p}_d\|_1 \leqslant \exp\left[\frac{\log(HL+1)}{d}\right] - 1. \tag{58}$$

A remarkable feature of this result lies in the fact that, for large enough values of d, the approximation bound is of order $\log(1+HL)/d$ and therefore only depends logarithmically on HL. From this point of view, it significantly improves the usual approximation bound HL/d which can easily be obtained by approximating p with an histogram based on a regular partition of the support of p into d pieces.

Using Proposition 7 together with (57) and optimizing with respect to d leads to the following risk bound.

Proposition 8. *Let $\widehat{p}$ be the ℓ-estimator of Proposition 6. There exists a universal constant $C > 0$ such that whatever $H \geqslant 0$, $L > 0$, $p^\star \in \overline{\mathcal{M}}(H,L)$ and $\xi > 0$,*

$$\|p^\star - \widehat{p}\|_1 \leqslant C\left[\left[\frac{\log(1+HL)}{n}\right]^{1/3} + \left[\frac{\log(1+HL)+1+\xi}{n}\right]^{1/2} + \frac{\epsilon}{n}\right] \tag{59}$$

with probability at least $1 - e^{-\xi}$.

5.3. **Robust regression with unimodal errors.** In this section $\mathscr{P}$ is the set of all probabilities on $E = \mathbb{R}$, q a given density on $\mathbb{R}$ (with respect to the Lebesgue measure) and P_θ the distribution with density $q_\theta = q(\cdot - \theta)$ for $\theta \in \mathbb{R}$. For $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$, we denote by $\mathbf{P}_{\boldsymbol{\theta}}$ the distribution $P_{\theta_1} \otimes \ldots \otimes P_{\theta_n}$, i.e. the distribution of a random vector of the form $\boldsymbol{X}' = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$ where the components $\varepsilon_1, \ldots, \varepsilon_n$ of $\boldsymbol{\varepsilon}$ are i.i.d. with density q. The vector $\boldsymbol{\theta}$ will be called the location parameter of the distribution $\mathbf{P}_{\boldsymbol{\theta}}$. We actually assume that the true distribution $\mathbf{P}^\star = P_1^\star \otimes \ldots \otimes P_n^\star$ of our observation $\boldsymbol{X}$ is close to a probability of the form $\mathbf{P}_{\overline{\boldsymbol{\theta}}}$ and, in view of estimating the location parameter $\overline{\boldsymbol{\theta}}$, we make the assumption that it belongs to some (countable) subset Θ of $\mathbb{R}^n$. Our model for the distribution $\mathbf{P}^\star$ is therefore $\mathscr{M} = \{\mathbf{P}_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\}$ and we shall use it to estimate $\mathbf{P}^\star$ with the TV-distance as our loss.

Assumption 6. *The density q is unimodal on $\mathbb{R}$ and Θ is a subset of a linear subspace of $\mathbb{R}^n$ with dimension $d \geqslant 1$.*

Under this assumption, we shall prove in Section 10.16 the following risk bound when ℓ is the TV-loss and $\mathscr{T}(\ell, \mathscr{M})$ has been chosen accordingly by (45).

Corollary 9. *If Assumption 6 is satisfied, any ℓ-estimator $\mathbf{P}_{\widehat{\boldsymbol{\theta}}} = \bigotimes_{i=1}^n \mathbf{P}_{\widehat{\theta}_i} \in \mathscr{M}$ satisfies, for all $\xi > 0$, with a probability at least $1 - 0.37e^{-\xi}$,*

$$\frac{1}{n}\sum_{i=1}^n \left\| P_i^\star - P_{\widehat{\theta}_i} \right\| \leqslant 5 \inf_{\boldsymbol{\theta} \in \Theta} \frac{1}{n}\sum_{i=1}^n \| P_i^\star - P_{\theta_i} \| + 552\sqrt{\frac{d+1}{n}} + \sqrt{\frac{8\xi}{n}} + \frac{2\epsilon}{n}. \tag{60}$$

To illustrate this result, let us consider the following example.

Example 4. Let $X_1, \ldots, X_n$ be independent random variables with respective distributions $P_1^\star, \ldots, P_n^\star$. We pretend, even though this might not be true, that the observations are of the form

$$X_i = \theta_i^\star + \varepsilon_i \quad \text{for i=1}, \ldots, n, \tag{61}$$

where $\boldsymbol{\theta}^\star = (\theta_1^\star, \ldots, \theta_n^\star)$ belongs to $\mathbb{R}^n$ and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with density $q : x \mapsto [\pi(1 + x^2)]^{-1}$ (Cauchy distribution) and our purpose is to estimate $\boldsymbol{\theta}^\star$. This framework can be viewed as a regression where the errors are Cauchy distributed and the θ_i correspond to the values of a regression function at some fixed point. The proof of the following lemma is postponed to Section 10.14.

Lemma 2. *For all $\theta, \theta' \in \mathbb{R}$,*

$$\| P_\theta - P_{\theta'} \| = \frac{2}{\pi} \arctan \frac{|\theta - \theta'|}{2}, \tag{62}$$

in particular

$$0.639 \left[\frac{|\theta - \theta'|}{\pi} \wedge 1 \right] \leqslant \| P_\theta - P_{\theta'} \| \leqslant \frac{|\theta - \theta'|}{\pi} \wedge 1. \tag{63}$$

We can now deduce the following result from Corollary 9.

Corollary 10. *Let V be a linear subspace of $\mathbb{R}^n$ with dimension $d \geqslant 1$ and consider the (possibly approximate) statistical model given by (61) for the data $X_1, \ldots, X_n$. There exists an ℓ-estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$ with values in V such that for all $\xi > 0$, with probability at least $1 - e^{-\xi}$,*

$$\frac{0.639}{n}\sum_{i=1}^{n}\left[\frac{\left|\theta_i^\star - \widehat{\theta}_i\right|}{\pi} \wedge 1\right] \leqslant \frac{6}{n}\sum_{i=1}^{n}\left\|P_i^\star - P_{\theta_i^\star}\right\| + \frac{5}{n}\inf_{\boldsymbol{\theta}\in V}\sum_{i=1}^{n}\left[\frac{|\theta_i^\star - \theta_i|}{\pi} \wedge 1\right] + 552\sqrt{\frac{d+1}{n}} + \frac{8\xi}{n} + \frac{2\epsilon}{n}.$$

Let us now analyze this bound. The first term $\sum_{i=1}^{n}\left\|P_i^\star - P_{\theta_i^\star}\right\|/n$ measures how well our model (61) describes the data. In particular, if $P_i^\star = P_{\theta_i^\star}$ for all i except for those in a subset $I \subset \{1, \ldots, n\}$,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|P_i^\star - P_{\theta_i^\star}\right\| \leqslant \frac{|I|}{n}$$

and the risk bound we get does not deteriorate much as long as $|I|$ remains small enough compared to $\sqrt{n(d+1)}$. When $P_i^\star = P_{\theta_i^\star}$ for all i, the bound we get corresponds, up to a remainder term, to the classical decomposition of the risk into an approximation term

$$\frac{1}{n}\inf_{\boldsymbol{\theta}\in V}\sum_{i=1}^{n}\left[\frac{|\theta_i^\star - \theta_i|}{\pi} \wedge 1\right] \leqslant \frac{1}{\pi n}\inf_{\boldsymbol{\theta}\in V}\sum_{i=1}^{n}|\theta_i^\star - \theta_i|,$$

which measures how well the linear space V approximates the parameter $\boldsymbol{\theta}^\star$ and a complexity term $\sqrt{(d+1)/n}$ that depends on the dimension of the model.

Proof. Let Θ be a countable and dense subset of V with respect to the Euclidean norm. Since Θ is a subset of a linear space with dimension $d \geqslant 1$ and q is unimodal, we deduce from Corollary 9 that there exists an ℓ-estimator $\widehat{\boldsymbol{\theta}}$ with values in $\Theta \subset V$ that satisfies (60). It follows from Lemma 2 that $\theta \mapsto \sum_{i=1}\|P_i^\star - P_{\theta_i}\|$ is continuous on $\mathbb{R}^n$, hence

$$\inf_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}\|P_i^\star - P_{\theta_i}\| = \inf_{\boldsymbol{\theta}\in V}\frac{1}{n}\sum_{i=1}\|P_i^\star - P_{\theta_i}\|, \tag{64}$$

and that

$$\frac{0.639}{n}\sum_{i=1}^{n}\left[\frac{\left|\theta_i^\star - \widehat{\theta}_i\right|}{\pi} \wedge 1\right] \leqslant \frac{1}{n}\sum_{i=1}^{n}\left\|P_{\theta_i^\star} - P_{\widehat{\theta}_i}\right\| \leqslant \frac{1}{n}\sum_{i=1}^{n}\left\|P_{\theta_i^\star} - P_i^\star\right\| + \frac{1}{n}\sum_{i=1}^{n}\left\|P_i^\star - P_{\widehat{\theta}_i}\right\|.$$

The result follows by combining this last inequality with (60), (63) and (64). □

5.4. **Faster rates under Assumption 2.** Although they share some similar features, the ℓ-estimator based on the approximate minimization with respect to $p \in \mathcal{M}$ of (46) looks different from that proposed by Devroye and Lugosi (2001)[Chapter 7] and Gao *et al.* (2018). Unlike their results, we shall prove that the ℓ-estimator may converge at a rate which can be faster than $1/\sqrt{n}$ provided that the model $\mathscr{M}$ satisfies our Assumption 2. To check whether this assumption is fulfilled on $\mathcal{M}$, one may use the following result and the fact that $\mathrm{Var}_S\left[\phi_{(P,Q)}(X)\right] = S(p>q)S(p\leqslant q)$ for all probabilities S.

Proposition 9. *If there exists a constant $a_2' \geqslant 0$ such that*

$$P(p \leqslant q) \wedge Q(p > q) \leqslant a_2' \|P - Q\| \tag{65}$$

for all probabilities P, Q in $\mathscr{M}$, then for all probabilities $S \in \mathscr{P}$

$$S(p>q)S(p\leqslant q) \leqslant S(p>q) \wedge S(p \leqslant q) \leqslant a_2\left[\|S-P\| + \|S-Q\|\right] \tag{66}$$

with $a_2 = 1 + a_2'$. Hence the family $\mathscr{T}(\ell, \mathscr{M})$ defined in Corollary 7 satisfies Assumption 2.

The proof is postponed to Section 10.3. Let us comment on our Condition (65). The testing affinity between two probabilities P and Q (see Le Cam (1973; 1986)) is defined as

$$\pi(P,Q) = \int_E (p \wedge q)\, d\mu = P(p \leqslant q) + Q(p > q) = P(p < q) + Q(p \geqslant q)$$

and it satisfies $\pi(P,Q) = 1 - \|P-Q\|$. It corresponds to the sum of the errors of first and second kinds of the (optimal) test function $\mathbb{1}_{q \geqslant p}$ when testing P versus Q on the basis of a single observation. In many situations, when P and Q are close with respect to the TV-distance, both errors are close to 1/2 but this is not the case when (65) holds: one of the testing errors is close to 0 and the other close to 1. To illustrate this fact, let us present two examples in the translation model, i.e. when $\mathcal{M} = \{p_\theta = p(\cdot - \theta),\ \theta \in \mathbb{Q}\}$ for some density p with respect to the Lebesgue measure μ on $\mathbb{R}$. We shall denote by P_θ the probability associated to the density p_θ.

Example 5. The density $p = \mathbb{1}_{[-1/2,1/2]}$ is that of the uniform distribution on $[-1/2, 1/2]$. It is easy to see that for all $\theta, \theta' \in \mathbb{R}$, $P_{\theta'}(p_\theta > p_{\theta'}) = 0$. Hence

$$P_\theta(p_\theta \leqslant p_{\theta'}) \wedge P_{\theta'}(p_\theta > p_{\theta'}) = P_{\theta'}(p_\theta > p_{\theta'}) = 0 \leqslant a_2' \left\|P_\theta - P_{\theta'}\right\|$$

and Condition (65) is therefore satisfied with $a_2' = 0$.

Example 6. We take for p the unbounded density $x \mapsto \alpha x^{\alpha-1}\mathbb{1}_{(0,1]}$ for some $\alpha \in (0,1)$. Note that for $\theta > \theta'$, $P_\theta(p_\theta \leqslant p_{\theta'}) = P_\theta(p_\theta < p_{\theta'}) = 0$, hence

$$P_\theta(p_\theta \leqslant p_{\theta'}) \wedge P_{\theta'}(p_\theta > p_{\theta'}) \leqslant P_\theta(p_\theta \leqslant p_{\theta'}) = 0$$

and for $\theta < \theta'$,

$$P_\theta(p_\theta \leqslant p_{\theta'}) \wedge P_{\theta'}(p_\theta > p_{\theta'}) \leqslant P_{\theta'}(p_\theta > p_{\theta'}) = 0.$$

Condition (65) is therefore satisfied with $a_2' = 0$.

Let us now go back to the framework of Section 5.2 assuming moreover that the observations $X_1, \ldots, X_n$ are (truly) i.i.d. with distribution $P^\star$ and that the family $\mathcal{M}$ of densities associated to our statistical model $\mathscr{M}$ satisfies Assumption 5.

When the true distribution $P^\star$ has a density $p^\star$ with respect to μ which belongs to $\overline{\mathcal{M}} = \{p_\theta = p(\cdot - \theta),\ \theta \in \mathbb{R}\}$, (49) shows that $\widehat{P}$ estimates $P^\star$ with an accuracy at least of order $1/\sqrt{n}$ with respect to the TV-distance. This order of magnitude (with respect to the number n of observations, omitting here the dependency with respect to V) is optimal in many statistical models including the Gaussian one that we considered at the end of Section 5.2. However, for some other models it is well-known to be suboptimal. This is for example the case for the family of uniform distributions on $[\theta - 1/2, \theta + 1/2]$ with $\theta \in \mathbb{R}$ for which one can estimate both the true distribution (with respect to the TV-distance) and the parameter θ (with respect to the Euclidean distance) at the rate $1/n$. We have seen in Example 5 that this statistical model actually fulfills our Assumption 2. Under this assumption, the result of Corollary 8 can actually be improved as shown by the following result to be proven in Section 10.17.

Corollary 11. *Assume that $X_1, \ldots X_n$ are i.i.d. with distribution $P^\star$ and that Assumption 5 and Condition (65) are both satisfied. Then any ℓ-estimator $\widehat{P} \in \mathscr{M}$ based on the family $\mathscr{T}(\ell, \mathscr{M})$ provided by Corollary 7 satisfies, for all $\xi > 0$, with a probability at least $1 - 0.42e^{-\xi}$,*

$$\begin{aligned}\left\|P^\star - \widehat{P}\right\| \leqslant\ & 6 \inf_{P \in \mathscr{M}} \|P^\star - P\| \\ & + c a_2^3 \frac{V}{n} \log\left(\frac{2en}{V \wedge n}\right) + 9.16\,(1 + 120 a_2)\,\frac{\xi}{n} + 2.3 \frac{\epsilon}{n},\end{aligned} \tag{67}$$

where c is a positive numerical constant ($c = 3 \times 10^{11}$ suits).

In view of illustrating this result, let us go back to our Example 6 for which we know that (65) holds with $a_2' = 0$, hence one may take $a_2 = 1$. For all $\theta, \theta' \in \mathbb{R}$, the set $\{x \in \mathbb{R},\ p_\theta(x) > p_{\theta'}(x)\}$ is an interval and such a class of subsets of $\mathbb{R}$ cannot shatter more than 2 points. Consequently, our Assumption 5 is satisfied with $V = 2$. It then follows from Corollary 11 that, whatever the true distribution $P^\star$ of our observations, with a probability at least $1 - 0.42e^{-\xi}$,

$$\left\|P^\star - \widehat{P}\right\| \leqslant C\left[\inf_{P \in \mathscr{M}} \|P^\star - P\| + \frac{\log n + 1 + \xi}{n}\right], \tag{68}$$

for some constant $C > 0$ that only depends on the choice of ϵ.

For this particular translation model, the TV-distance between two probabilities P_θ and $P_{\theta'}$ in $\mathscr{M}$ can be related to the Euclidean distance between their parameters by arguing as follows. First of all, it is not difficult to check that the testing affinity between P_θ and $P_{\theta'}$ (with $\theta < \theta'$) writes as

$$\begin{aligned}\pi(P_\theta, P_{\theta'}) &= \int_{\mathbb{R}} (p_\theta \wedge p_{\theta'})\, d\lambda = \int_\theta^{\theta'} 0 d\lambda + \int_{\theta'}^{1+\theta} p_\theta d\lambda + \int_{1+\theta}^{1+\theta'} 0 d\lambda \\ &= [(x-\theta)^\alpha]_{\theta'}^{1+\theta} = 1 - \left|\theta' - \theta\right|^\alpha \quad \text{when } \theta' \leqslant \theta + 1\end{aligned}$$

and $\pi(P_\theta, P_{\theta'}) = 0$ for $\theta' > \theta + 1$ since the supports of P_θ and $P_{\theta'}$ are then disjoint. Consequently, for all $\theta, \theta' \in \mathbb{R}$

$$\|P_\theta - P_{\theta'}\| = 1 - \pi(P_\theta, P_{\theta'}) = \left|\theta - \theta'\right|^\alpha \wedge 1,$$

which means, using the triangle inequality, that if $P^\star$ is close to some distribution $P_{\overline{\theta}} \in \mathscr{M}$, by (68), the estimator $\widehat{P} = P_{\widehat{\theta}}$ of $P^\star$ satisfies, with a probability at least $1 - 0.42e^{-\xi}$,

$$\left[\left|\overline{\theta} - \widehat{\theta}\right|^\alpha \wedge 1\right] = \left\|P_{\overline{\theta}} - P_{\widehat{\theta}}\right\| \leqslant C\left[2\left\|P^\star - P_{\overline{\theta}}\right\| + \frac{\log n + 1 + \xi}{n}\right].$$

In particular, if $P^\star$ belongs to $\overline{\mathscr{M}}$, i.e. $P^\star = P_{\overline{\theta}}$ for some $\overline{\theta} \in \mathbb{R}$, and if n is large enough, the estimator $\widehat{\theta}$ estimates $\overline{\theta}$ with an accuracy of order $(\log n/n)^{1/\alpha}$. This rate is much faster than $1/\sqrt{n}$ whatever $\alpha \in (0,1)$ and is optimal up to the logarithmic factor.

It is not difficult to check that the above calculations extend to the case $\alpha = 1$, i.e. when the statistical model is the translation of the uniform density $p = \mathbb{1}_{[-1/2,1/2]}$ as in Example 5. The ℓ-estimator then converges at rate (at least) $\log n/n$ in this case. In particular, it does not coincide with the empirical median which is known to converge at rate $1/\sqrt{n}$. Note that this result is not contradictory to our Proposition 14 (presented in our Section 7.1 below) since the density p is not a decreasing function of $|x|$. This proves, in passing, that our assumption that f is decreasing is necessary in our Assumption 7.

6. Hellinger and KL-losses

6.1. Building suitable families $\mathscr{T}(\ell, \mathscr{M})$. The Hellinger and KL-losses cannot be defined by variational formulas like (24) and (25) but, as we shall see, satisfy the following alternative expressions:

$$\ell(P,Q) = \sup_{f \in \mathcal{F}}\left[\int_E f dP - \Lambda(Q,f)\right] = \int_E f_{(P,Q)} dP - \Lambda(Q, f_{(P,Q)}), \tag{69}$$

for some class of functions $\mathcal{F}$ and a fixed function Λ on $\overline{\mathscr{P}} \times \mathcal{F}$ (where $\overline{\mathscr{P}}$ denotes a convex set of probabilities. Observe that (24) and (25) are actually a special case of (69) when $\Lambda(Q,f) = \int_E f dQ$.

A common feature of losses of the forms (69) and (24) lies in the fact that we know where the supremum is reached, i.e. we have identified a function

$f_{(P,Q)}$ such that $\ell(P,Q) = \int_E f_{(P,Q)} dP - \Lambda(Q, f_{(P,Q)})$. When $\Lambda(Q,f) = \int_E f dQ$, Proposition 2 tells us that a recipe to build up candidate functions $\phi_{(P,Q)}$ satisfying our Assumption 1 is given by (27). We shall actually use the same recipe for other functions Λ, setting, for $P,Q \in \overline{\mathscr{P}}$, $R = (P+Q)/2 \in \overline{\mathscr{P}}$ (since $\overline{\mathscr{P}}$ is convex) and

$$\phi_{(P,Q)} = C\left[\left(f_{(R,P)} - \Lambda(P, f_{(R,P)})\right) - \left(f_{(R,Q)} - \Lambda(Q, f_{(R,Q)})\right)\right], \tag{70}$$

where C denotes a positive normalizing constant that is chosen for $\phi_{(P,Q)}$ to satisfy our Assumption 1-(iii). This expression of $\phi_{(P,Q)}$ is motivated by the equality

$$\mathbb{E}_R\left[\phi_{(P,Q)}(X)\right] = C\left[\ell(R,P) - \ell(R,Q)\right]$$

which means that the sign of $\mathbb{E}_R\left[\phi_{(P,Q)}(X)\right]$ is the same as that of $\ell(R,P) - \ell(R,Q)$.

Note that when $\Lambda(Q,f) = \int_E f dQ$ and $\mathcal{F}$ is symmetric

$$\ell(R,P) = \sup_{f\in\mathcal{F}}\left[\int_E f dR - \int_E f dP\right] = \frac{1}{2}\sup_{f\in\mathcal{F}}\left[\int_E f dQ - \int_E f dP\right] = \frac{1}{2}\ell(P,Q)$$

and we may therefore choose $f_{(R,P)} = f_{(Q,P)} = -f_{(P,Q)} = -f_{(R,Q)}$ which together with (70) gives

$$\begin{aligned}\phi_{(P,Q)} &= C\left[\left(f_{(R,P)} - \int_E f_{(R,P)} dP\right) - \left(f_{(R,Q)} - \int_E f_{(R,Q)} dQ\right)\right]\\ &= 2C\left[\int_E f_{(P,Q)} \frac{dP + dQ}{2} - f_{(P,Q)}\right].\end{aligned}$$

Up to the normalizing constant, we recover (27).

6.2. **The Hellinger distance.**

An alternative way of defining the Hellinger distance given by (1) is provided by the following proposition (with the conventions $0/0 = 1$ and $a/0 = +\infty$ for all $a > 0$). This result will be proven in Section 10.4.

Proposition 10. *Let $\mathcal{G}$ be the class of all measurable functions g on $(E,\mathcal{E})$ with values in $[0,+\infty]$. For all probabilities P,Q on $(E,\mathcal{E})$,*

$$h^2(P,Q) = \frac{1}{2}\sup_{g\in\mathcal{G}}\left[\mathbb{E}_P\left(1-g\right) + \mathbb{E}_Q\left(1 - 1/g\right)\right], \tag{71}$$

and the supremum is reached for $g = g_{(P,Q)} = \sqrt{q/p}$. In particular, the Hellinger affinity between P and Q satisfies

$$\rho(P,Q) = \frac{1}{2}\inf_{g\in\mathcal{G}}\left[\mathbb{E}_P\left(g\right) + \mathbb{E}_Q\left(1/g\right)\right]. \tag{72}$$

With the change of functions $f = 1 - g$, (69) is satisfied for the class $\mathcal{F}$ of functions with values in $[-\infty, 1]$, $\Lambda(Q,f) = \int_E [f/(1-f)] dQ$ and $f_{(P,Q)} =$

$1-\sqrt{q/p}$. Applying (70), we obtain that

$$\begin{aligned}\phi_{(P,Q)} &= C\left[\left(f_{(R,P)}-\Lambda(P,f_{(R,P)})\right)-\left(f_{(R,Q)}-\Lambda(Q,f_{(R,Q)})\right)\right]\\ &= C\left[\frac{\sqrt{q}-\sqrt{p}}{\sqrt{r}}+\int_E\sqrt{r}(\sqrt{q}-\sqrt{p})d\mu\right]\end{aligned}$$

and $\phi_{(P,Q)}$ corresponds thus to the test proposed in Baraud (2011). In particular, we obtain the following result the proof of which is postponed to Section 10.5.

Proposition 11. *Let $\mathscr{P}$ be the set of probabilities on $(E,\mathcal{E})$ dominated by μ, $\mathscr{M}$ a countable subset of $\mathscr{P}$ and consider the loss ℓ defined by $\ell(P,Q)=h^2(P,Q)$ for all $P,Q\in\mathscr{P}$. The family $\mathscr{T}(\ell,\mathscr{M})$ of functions $\phi_{(P,Q)}$ defined for $P,Q\in\mathscr{M}$ by*

$$\phi_{(P,Q)}=\frac{1}{2\sqrt{2}}\left[\rho(R,Q)-\rho(R,P)+\frac{\sqrt{q}-\sqrt{p}}{\sqrt{r}}\right]\quad \text{with } R=\frac{P+Q}{2}, \tag{73}$$

satisfies Assumption 2 with $a_0=(\sqrt{2}+1)/2$, $a_1=(\sqrt{2}-1)/2$, $a_2=3/2$.

It is possible to design other families $\mathscr{T}(\ell,\mathscr{M})$ that satisfy Assumption 2 on the larger set of all probabilities on $(E,\mathcal{E})$, i.e. that are not necessarily dominated by μ, but this requires more technicalities. We prefer to avoid them here and rather refer the reader to Baraud and Birgé (2018).

6.3. **The Kullback-Leibler divergence.** We mention the Kullback-Leibler divergence as an example of loss function that fits our assumptions. Nevertheless, we would probably not recommend it in general as a loss function. As seen in the introduction, an estimator $\widehat{\theta}$ of a parameter θ can be very good in the sense that the associated probabilities $P_{\widehat{\theta}}$ and P_θ would be difficult to distinguish (say from a sample of size 10^6) while $K(P_{\widehat{\theta}},P_\theta)=+\infty$.

The KL-divergence given by (5) can alternatively be defined via the following variational formula:

$$K(P,Q)=\sup_{f\in\mathcal{F}}\left[\mathbb{E}_P\left[f\right]-\log\mathbb{E}_Q\left(e^f\right)\right], \tag{74}$$

which corresponds to (69) with $\Lambda(Q,f)=\log\mathbb{E}_Q(e^f)$ and $\mathcal{F}$ is the class of all measurable functions on E with values in $[-\infty,+\infty]$ such that $\Lambda(Q,f)<+\infty$. Equality holds in (74) for $f=f_{(P,Q)}=\log(p/q)$. Using that $\Lambda(P,f_{(R,P)})=\Lambda(Q,f_{(R,Q)})=0$ and applying (70) we get that $\phi_{(P,Q)}$ is proportional to

$$\begin{aligned}&\left[f_{(R,P)}-\Lambda(P,f_{(R,P)})\right]-\left[f_{(R,Q)}-\Lambda(Q,f_{(R,Q)})\right]\\ &\qquad\qquad=\log\frac{r}{p}-\log\frac{r}{q}=\log q-\log p\end{aligned}$$

and therefore corresponds to the well-known likelihood ratio test. The following holds.

Proposition 12. *Let $\mathscr{P}$ be the set of all probabilities S on $(E,\mathcal{E})$ which are dominated by μ and whose densities s satisfy $\mathbb{E}_S[|\log s|] < +\infty$. Assume that $\mathscr{M} = \{p\cdot\mu, p\in\mathcal{M}\}$ is a countable subset of $\mathscr{P}$ and that $\mathcal{M}$ satisfies for some constant $a>0$,*

$$e^{-a} \leqslant \frac{p}{q}(x) \leqslant e^{a} \quad \textit{for all } p,q\in\mathcal{M} \textit{ and } x\in E. \tag{75}$$

The family $\mathscr{T}(\ell,\mathscr{M})$ of functions $\phi_{(P,Q)}$ given by

$$\phi_{(P,Q)} = \frac{1}{2a}\log\left(\frac{q}{p}\right) \quad \textit{for all } P,Q\in\mathscr{M} \tag{76}$$

satisfies Assumption 2 with $a_0 = a_1 = 1/(2a)$ and $a_2 = 1/[a(2\wedge a)]$ for the KL-loss $\ell(P,Q) = K(P,Q)$, $P,Q\in\mathscr{P}$.

The proof is postponed to Section 10.6.

Under (75), the squared Hellinger distance and the Kullback-Leibler divergence turn out to be equivalent on $\mathscr{M}$. More precisely, it follows from Lemma 7.23 in Massart (2007) that if $\mathcal{M}$ satisfies (75)

$$2h^2(P,Q) \leqslant K(P,Q) \leqslant 2(2+a)h^2(P,Q) \quad \text{for all } P,Q\in\mathscr{M}.$$

If the data are i.i.d. with distribution $P^\star$, we also have $2h^2(P^\star,P)\leqslant K(P^\star,P)$ for all $P\in\mathscr{M}$ but (75) says nothing on how much larger $K(P^\star,P)$ is compared to $h^2(P^\star,P)$. This means that the result of Theorem 2 for the Kullback-Leibler divergence cannot be deduced from that established for the squared Hellinger distance.

7. ℓ-estimators based on the TV-distance versus ρ-estimators

As explained in Section 5.2, a nice feature of ℓ-estimators based on the TV-distance lies in their robustness properties with respect to a possible misspecification of the model. As described in details in Baraud *et al* (2017) and Baraud and Birgé (2018), ρ-estimators also possess robustness properties except from the fact that these properties are expressed in terms of the Hellinger distance between probabilities and not the TV one. Since these two distances are not equivalent in general, it is worth analyzing further the main differences between ℓ-estimators based on the TV-distance and ρ-estimators.

7.1. **Robustness and optimality.**

Let us go back to Example 2 in the simple situation where $d=1$. We assume that the data are i.i.d. with distribution $P^\star = (1-\alpha)P_{m^\star} + \alpha R$ for some probability R on $\mathbb{R}$ and $\alpha\in[0,1/10]$. Then, $7.7\|P^\star - P_{m^\star}\| = 7.7\alpha\|P_{m^\star} - R\| \leqslant 0.77$ and choosing $\epsilon = 10^{-10}$ and fixing the value of $\xi>0$, we deduce from (55) that when n is sufficiently large compared to ξ (so that the right-hand side of (55) is smaller than 1), with probability at least $1-e^{-\xi}$ the ℓ-estimator $\widehat{m}$ of $m^\star$ satisfies

$$|m^\star - \widehat{m}| \leqslant C\left[\|P^\star - P_{m^\star}\| + \sqrt{\frac{1+\xi}{n}}\right] \tag{77}$$

for some universal constant $C>0$.

Alternatively, in this statistical setting, we may use a ρ-estimator $\widetilde{m}$ for estimating $m^\star$. More precisely, we may apply the following result that can be derived by combining Corollary 3 of Baraud and Birgé (2018) with Proposition 42 of Baraud *et al* (2017).

Proposition 13. *Let $\mathcal{M}=\{p_\theta = p(\cdot-\theta),\ \theta\in\mathbb{Q}\}$ be a translation model for the data $X_1,\ldots,X_n$ where p is a unimodal density. One can build a ρ-estimator $\widetilde{\theta}=\widetilde{\theta}(\boldsymbol{X})$ such that for all $\theta^\star\in\mathbb{Q}$ and $\xi>0$, with probability at least $1-e^{-\xi}$,*

$$h^2(P_{\theta^\star}, P_{\widetilde{\theta}}) \leqslant C\left[\frac{1}{n}\sum_{i=1}^n h^2(P_i^\star, P_{\theta^\star}) + \frac{\log n+\xi}{n}\right] \tag{78}$$

where C denotes some positive universal constant.

Applying this result to our statistical model and using the fact that for all $m,m'\in\mathbb{R}$

$$(1-e^{-1})\left[\frac{(m-m')^2}{4}\wedge 1\right] \leqslant h^2(P_m,P_{m'}) = 1-e^{-\frac{(m-m')^2}{4}} \leqslant \frac{(m-m')^2}{4}\wedge 1$$

we deduce that, provided that n is large enough compared to ξ, with probability at least $1-e^{-\xi}$,

$$|m^\star - \widetilde{m}| \leqslant C'\left[h(P^\star, P_{m^\star}) + \sqrt{\frac{\log n+\xi}{n}}\right], \tag{79}$$

for some universal $C'>0$.

If we forget about the logarithmic factor and the universal constants C, C', the main difference between inequalities (77) and (79) lies in the expression of the approximation terms $\|P^\star, P_{m^\star}\|$ and $h(P^\star, P_{m^\star})$. Since, for all probabilities P,Q, $\|P-Q\|\leqslant\sqrt{2}h(P,Q)$, the accuracy of $\widehat{m}$ cannot be much worse than that of $\widetilde{m}$ but it can indeed be much better: when the probabilities R and $P_{m^\star}$ are singular,

$$\|P^\star - P_{m^\star}\| = \alpha\,\|P_{m^\star}-R\| = \alpha$$

while

$$h^2(P^\star, P_{m^\star}) = h^2((1-\alpha)P_{m^\star}+\alpha R, P_{m^\star}) = 1-\sqrt{1-\alpha}$$

and we deduce that

$$\begin{aligned}|m^\star - \widehat{m}| &\leqslant C\left[\alpha + \sqrt{\frac{1+\xi}{n}}\right]\\ |m^\star - \widetilde{m}| &\leqslant C'\left[(1-\sqrt{1-\alpha})^{1/2} + \sqrt{\frac{\log n + \xi}{n}}\right].\end{aligned}$$

For small values of α, $(1-\sqrt{1-\alpha})^{1/2}$ is of order $\sqrt{\alpha/2}$ and is therefore much larger than α. This means that the bound we get on $\widehat{m}$ does not deteriorate too much compared to the ideal situation where the model is exact, i.e. $P^\star = P_{m^\star}$, as long as α remains small compared to $1/\sqrt{n}$. The same conclusion would be true for $\widetilde{m}$ only if α remains small enough compared to $(\log n)/n$. From this point of view, the estimator $\widetilde{m}$ seems less robust than $\widehat{m}$. This disappointing result (for ρ-estimators) is actually not restricted to this Gaussian model and can actually be generalized to many other situations for which the TV-distance and the Hellinger one are equivalent on the model $\mathscr{M}$.

The apparent superiority of ℓ-estimator over ρ-estimators must nevertheless be nuanced in the light of the following example. Assume that the data $X_1, \ldots, X_n$ are truly i.i.d. from a translation model $\mathcal{M} = \{p_\theta = p(\cdot - \theta),\ \theta \in \mathbb{Q}\}$ where the density p satisfies the following.

Assumption 7. *There exists a positive decreasing function f on $(0, +\infty)$ such that $p(x) = f(|x|)$ for all $x \in \mathbb{R} \setminus \{0\}$.*

In this situation, an ℓ-estimator for the location parameter θ is not difficult to compute. Putting aside the fact that our statistical model is parametrized by $\mathbb{Q}$ and not $\mathbb{R}$ in order to make it countable, the empirical median turns out to be an ℓ-estimator. More precisely, let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the order statistics associated to the n-sample $X_1, \ldots, X_n$ with $n \geqslant 2$ and define the empirical median as $X_{(\lceil n/2 \rceil)}$ where

$$\lceil x \rceil = \min\{k \in \mathbb{N},\ k \geqslant x\} \quad \text{for all } x > 0,$$

that is

$$\sum_{i=1}^n 1\!\!1_{X_i < X_{(\lceil n/2\rceil)}} < \frac{n}{2} \leqslant \sum_{i=1}^n 1\!\!1_{X_i \leqslant X_{(\lceil n/2\rceil)}}. \tag{80}$$

The proof of the following result is postponed to Section 10.1.

Proposition 14. *Let Assumption 7 be satisfied and consider the TV-loss. Any element $\widehat{\theta} \in \mathbb{Q}$ that satisfies $X_{(\lceil n/2\rceil)} < \widehat{\theta} < X_{(\lceil n/2\rceil+1)}$ is an ℓ-estimator of θ for the choice $\epsilon = 1/2$.*

It is not difficult to find an example of a translation model satisfying Assumption 7 for which the empirical median would lead to a sub-optimal estimator. The choice

$$p : x \mapsto \frac{\alpha}{2(1+\alpha)}\left[\frac{1}{|x|^{1-\alpha}} \wedge \frac{1}{x^2}\right] 1\!\!1_{|x|>0} \quad \text{with } \alpha \in (0,1)$$

actually suits. For this density, it is possible to check that the empirical median converges at rate $n^{-1/(2\alpha)}$ (with respect to the Euclidean loss) while the minimax rate is actually of order $n^{-1/\alpha}$. Since for θ, θ' close enough $|\theta - \theta'|^{\alpha}$ is of order $h^2(P_\theta, P_{\theta'})$, one can easily derive from Proposition 13 that a ρ-estimator would converge at a rate which is at least $(\log n/n)^{1/\alpha}$ and would therefore be optimal (up to a logarithmic factor).

In conclusion, our construction of ℓ-estimators for the TV-loss warrants robustness but not optimality.

7.2. **Logarithmic factors.** The above discussion puts aside the logarithmic factor that appears in the right-hand side of (78) compared to (50). In fact, the results obtained for the Hellinger loss in Baraud and Birgé (2018; 2016) often involve such logarithmic factors. Compared to the minimax risk over the model $\mathscr{M}$, these factors turn out to be sometimes necessary.

This is for example the case when $\mathcal{M}$ is a countable and dense subset (with respect the Hellinger distance) of the set $\overline{\mathcal{M}}$ of all histograms on $\mathbb{R}$ with at most $d \geqslant 1$ pieces, i.e. $p \in \overline{\mathcal{M}}$ is of the form

$$\sum_{i=1}^{d} a_i \mathbb{1}_{(b_i, b_{i+1}]} \quad \text{with } -\infty < b_1 < \ldots < b_{d+1} < +\infty$$

and $a_1, \ldots, a_d \in \mathbb{R}_+$ satisfying $\sum_{i=1}^{d+1} a_i(b_{i+1} - b_i) = 1$. It is proven in Baraud and Birgé (2016) that if the data $X_1, \ldots, X_n$ are i.i.d. with a density $p^\star \in \overline{\mathcal{M}}$, the ρ-estimator $\widetilde{p}$ on $\mathcal{M}$ satisfies, for some universal constant $C > 0$,

$$\sup_{p^\star \in \overline{\mathcal{M}}} \mathbb{E}\left[h^2(p^\star, \widetilde{p})\right] \leqslant C \frac{d \max\{\log^{3/2}(n/d), 1\}}{n}. \tag{81}$$

It is shown in Birgé and Massart (1998)[Proposition 2] that the minimax rate is at least $(d/n)\max\{\log(n/d), 1\}$ (up to some universal constant). The logarithmic factor appearing in the right-hand side of (81) is therefore necessary (with a possibly smaller power though). A look at the proof of Proposition 2 in Birgé and Massart shows that this logarithmic factor is due to some combinatoric arguments based on the fact that $\mathcal{M}$ contains histograms based on possibly irregular partitions of $[0, 1]$.

Surprisingly, this logarithmic factor disappears for the TV-loss. It is easy to see that for $\overline{p}, q \in \mathcal{M}$, the sets $\{\overline{p} > q\}$ are the union of at most $d+1$ intervals and Assumption 5 is therefore satisfied with $V(\overline{p}) = 2(d+1)$ for all $\overline{p} \in \mathcal{M}$. It follows then from Proposition 6, more precisely (57), that the ℓ-estimator of $\widehat{p}$ would satisfy, for some constant $C' > 0$ depending on ϵ only,

$$\sup_{p^\star \in \overline{\mathcal{M}}_d} \mathbb{E}\left[\|p^\star - \widehat{p}\|_1^2\right] \leqslant C' \frac{d}{n}. \tag{82}$$

This result shows that, on this example at least, the minimax rates with respect to the squared TV and Hellinger losses may differ by at least a logarithmic factor (in fact, it can be proven that the the minimax rate with respect to squared TV loss is indeed of order d/n for $d \geqslant 2$).

8. Application to robust testing

The heuristics we have developed in Section 2.3 and which lead to the construction of ℓ-estimators are based on the fact that one can roughly estimate $\ell(P^\star, P) - \ell(P^\star, Q)$ by a statistic $\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})/n$. This statistics can also be used to build robust tests between P and Q. More precisely, Proposition 1 implies that

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

so that $\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})\right]$ roughly (exactly indeed if $a_0 = a_1$) tells us which of $\mathbf{P}$ or $\mathbf{Q}$ is closer to $\mathbf{P}^\star$. Replacing $\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})\right]$ by its empirical version $\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})$ and basing our decision on its sign, i.e. the test statistic $\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X})$ given by (12), then leads to a test between $\mathbf{P}$ and $\mathbf{Q}$ the errors of which can be controlled provided that the ratio $\boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{P})/\boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q})$ is far away enough from 1.

The performance of $\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X})$ as a test between $\mathbf{P}$ and $\mathbf{Q}$ is provided by the following result to be proven in Section 10.9.

Proposition 15. *Let Assumption 1 hold and $\mathbf{P}^\star \in \mathscr{P}$ be such that $\gamma = a_0 \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{P})/[a_1 \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q})] < 1$. Then*

$$\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 1\right] \leqslant \exp\left[-\frac{\boldsymbol{\ell}^2(\mathbf{P}^\star, \mathbf{Q})}{2n}\left[a_1(1-\gamma)\right]^2\right]. \tag{83}$$

If, moreover, Assumption 2-(iv) is satisfied,

$$\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 1\right] \leqslant \exp\left[-\frac{\boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q})}{2}\frac{a_1(1-\gamma)^2}{[(1-\gamma)/3] + [(1+\gamma)a_2/a_1]}\right]. \tag{84}$$

Inequalities (83) and (84) both say that if $\mathbf{P}^\star$ is close enough to $\mathbf{P}$ and far enough from $\mathbf{Q}$ with respect to the loss $\boldsymbol{\ell}$, the test $\Phi_{(\mathbf{P},\mathbf{Q})}$ decides $\mathbf{P}$ with probability close to 1. In view of the symmetry of the assumptions with respect to $\mathbf{P}$ and $\mathbf{Q}$, to bound $\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 0\right]$ it suffices to exchange their roles, now assuming that $\gamma = a_0 \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q})/[a_1 \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{P})] < 1$.

Recalling from Section 2.5 that $a_1 \leqslant a_0$, note that if

$$a_1/a_0 \leqslant \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{P})/\boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q}) \leqslant a_0/a_1,$$

one cannot say anything about the performance of the test.

In order to comment these results further, let us consider the density framework with $\mathbf{P} = \mathbf{P}^\star = (P^\star)^{\otimes n}$ and $\mathbf{Q} = Q^{\otimes n}$ for some probability Q on $(E, \mathcal{E})$. When Assumption 1 is satisfied, it is interesting to notice that the test accepts the hypothesis $P^\star = P$ with probability close to one as soon as $\ell(P^\star, Q) = \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q})/n$ is large enough compared to $1/\sqrt{n}$. The situation is even better under Assumption 2 since it is actually enough that

$\ell(P^\star, Q)$ be large enough compared to $1/n$. It is well-known, mainly from the work of Le Cam (1973), that it is impossible to distinguish between two probabilities P and Q from an n-sample when the Hellinger distance $h(P,Q)$ is small enough compared to $1/\sqrt{n}$. As a consequence, the test $\Phi_{(\mathbf{P},\mathbf{Q})}$ is optimal under Assumption 1 when the loss ℓ is of the order of the Hellinger distance and optimal under Assumption 2 when it is of order the square of the Hellinger distance.

As we have seen earlier, most loss functions of interest are actually powers of some distance on $\mathscr{P}$. As an illustration let us focus on the case of $\ell = h^2$ and let Assumption 2 hold, in which case (84) becomes, according to Proposition 11,

$$\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 1\right] \leqslant \exp\left[-\frac{3\left(\sqrt{2}-1\right)^2(1-\gamma)^2 n h^2(P^\star, Q)}{4\left[(1-\gamma)\left(\sqrt{2}-1\right)+9(1+\gamma)\right]}\right], \tag{85}$$

provided that

$$\gamma = \left(3+2\sqrt{2}\right)\frac{h^2(P^\star,P)}{h^2(P^\star,Q)} < 1. \tag{86}$$

An interesting feature of this result is that it holds even if both $h(P^\star, P)$ and $h(P^\star, Q)$ are larger than $h(P,Q)/2$ provided that (86) is satisfied. This would not be the case of tests between Hellinger balls centered at P and Q respectively. More generally, if $\ell = \mathrm{d}^j$ for some distance d and $j \geqslant 1$, the test will perform nicely if $\mathrm{d}(P^\star, Q)/\mathrm{d}(P^\star, P)$ is large enough, even if $\mathrm{d}(P^\star, Q)$ is much larger than $\mathrm{d}(P,Q)$.

8.1. **Case of the TV-distance.** Let us assume here that $\mathbf{P} = P^{\otimes n}$ and $\mathbf{Q} = Q^{\otimes n}$ where P and Q have densities p, q with respect to some dominating measure μ. It follows from (45) that the test statistic $\mathbf{T}(\boldsymbol{X}, \mathbf{P}, \mathbf{Q})$ for testing between $\mathbf{P}$ and $\mathbf{Q}$ satisfies

$$\frac{\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})}{n} = \frac{P(p>q)+Q(p>q)}{2} - \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{p>q}(X_i).$$

We derive from Proposition 15 and Corollary 7 the following result for the test $\Phi_{(\mathbf{P},\mathbf{Q})}$.

Proposition 16. *Let P and Q be probabilities on $(E, \mathcal{E})$. If $\boldsymbol{X} = (X_1, \ldots, X_n)$ has distribution $\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star$ and*

$$\gamma = \frac{3\sum_{i=1}^n \|P_i^\star - P\|}{\sum_{i=1}^n \|P_i^\star - Q\|} < 1.$$

The test $\Phi_{(\mathbf{P},\mathbf{Q})}$ defined by (12) satisfies

$$\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 1\right] \leqslant \exp\left[-\frac{(1-\gamma)^2 n}{8}\left(\frac{1}{n}\sum_{i=1}^n \|P_i^\star - Q\|\right)^2\right].$$

In particular, if $X_1,\ldots,X_n$ are i.i.d. with distribution $P^\star$,

$$\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X})=1\right]\leqslant\exp\left[-\frac{(1-\gamma)^2n}{8}\left\|P^\star-Q\right\|^2\right]$$

provided that $\gamma=3\left\|P^\star-P\right\|/\left\|P^\star-Q\right\|<1$.

Given two real numbers $a<b$ with mean $m=(a+b)/2$, the intervals $(-\infty,m)=\{x\in\mathbb{R},\ m-x>0\}$ and $(m,+\infty)=\{x\in\mathbb{R},\ m-x<0\}$ also correspond to the sets $\{x\in\mathbb{R},\ |x-a|<|b-x|\}$ and $\{x\in\mathbb{R},\ |x-a|>|b-x|\}$ of those x which are closer to a and b respectively. Using the fact that

$$Q(p>q)=\int_E 1\!\!1_{p>q}q\,d\mu\leqslant\int_E 1\!\!1_{p>q}p\,d\mu=P(p>q),$$

the corresponding test $\Phi_{(\mathbf{P},\mathbf{Q})}$ can be reformulated equivalently as $\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X})=0$ when

$$\left|\frac{1}{n}\sum_{i=1}^n 1\!\!1_{p>q}(X_i)-P(p>q)\right|<\left|\frac{1}{n}\sum_{i=1}^n 1\!\!1_{p>q}(X_i)-Q(p>q)\right| \tag{87}$$

and $\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X})=1$ when

$$\left|\frac{1}{n}\sum_{i=1}^n 1\!\!1_{p>q}(X_i)-Q(p>q)\right|<\left|\frac{1}{n}\sum_{i=1}^n 1\!\!1_{p>q}(X_i)-P(p>q)\right|.$$

Under this form and using the convention that the test takes the value 1 when equality holds in (87), we recover the test proposed by Devroye and Lugosi (2001) [Chapter 6].

8.2. **Case of the $\mathbb{L}_j$-loss.** We assume here that $\mathbf{P}=P^{\otimes n}$ and $\mathbf{Q}=Q^{\otimes n}$ where P,Q are not necessarily probabilities but possibly signed measures with densities p and q with respect to some dominating measure μ. We consider the $\mathbb{L}_j$-loss for $j\in(1,+\infty)$ and assume that p and q belong to $\mathscr{L}_j(E,\mu)\cap\mathscr{L}_1(E,\mu)$. Clearly, (33) is satisfied as soon as $R=\left\|p-q\right\|_\infty/\left\|p-q\right\|_j<+\infty$ (assuming $P\neq Q$) and it follows from Corollary 3 that

$$\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})=\frac{1}{2}\left[\sum_{i=1}^n\left(\frac{\sigma|p-q|^{j-1}}{\|p-q\|_\infty^{j-1}}\right)(X_i)-\int_E\frac{\sigma|p-q|^{j-1}}{\|p-q\|_\infty^{j-1}}\frac{p+q}{2}\,d\mu\right]$$

where $\sigma(x)=1\!\!1_{q>p}(x)-1\!\!1_{p>q}(x)$ for all $x\in E$. Note that for $j=2$,

$$4\left\|p-q\right\|_\infty\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})=\left[2\sum_{i=1}^n q(X_i)-\|q\|_2^2\right]-\left[2\sum_{i=1}^n p(X_i)-\|p\|_2^2\right]$$

and the test $\Phi_{(\mathbf{P},\mathbf{Q})}$ between P and Q is that associated to the classical $\mathbb{L}_2$-contrast function.

We deduce from Proposition 15 and Corollary 3 the following result.

Proposition 17. *Let $j \in (1,+\infty)$, $P = p\cdot\mu$, $Q = q\cdot\mu$ be two distinct and possibly signed measures on $(E,\mathcal{E})$ with $p,q \in \mathscr{L}_j(E,\mu)\cap\mathscr{L}_1(E,\mu)$. Assume that $X_1,\ldots,X_n$ are independent with respective densities $p_1^\star,\ldots,p_n^\star \in \mathscr{L}_j(E,\mu)$. If*

$$\gamma = \frac{3\sum_{i=1}^n \|p_i^\star - p\|_j}{\sum_{i=1}^n \|p_i^\star - q\|_j} < 1 \quad \textit{and} \quad R = \frac{\|p-q\|_\infty}{\|p-q\|_j} < +\infty$$

the test $\Phi_{(\mathbf{P},\mathbf{Q})}$ defined by (12) satisfies

$$\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 1\right] \leqslant \exp\left[-\frac{(1-\gamma)^2 n}{32R^{2(j-1)}}\left(\frac{1}{n}\sum_{i=1}^n \|p_i^\star - q\|_j\right)^2\right].$$

In particular, if $X_1,\ldots,X_n$ are i.i.d. with density $p^\star \in \mathscr{L}_j(E,\mu)$,

$$\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X}) = 1\right] \leqslant \exp\left[-\frac{(1-\gamma)^2 n}{32R^{2(j-1)}}\|p^\star - q\|_j^2\right]$$

provided that $\gamma = 3\|p^\star - p\|_j / \|p^\star - q\|_j < 1$.

9. Proofs of Theorems 1 and 2

Let $\overline{\mathbf{P}}$ be an arbitrary point in $\mathscr{M}$ and $\zeta > 0$. We recall that $\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q}) = \sum_{i=1}^n \phi_{(\overline{P}_i,Q_i)}(X_i)$ for all $\mathbf{Q} = (Q_1,\ldots,Q_n) \in \mathscr{M}$ and we shall repeatedly use the equalities and inequalities below that immediately derive from Assumption 1-(i) and (ii) and a summation over $i \in \{1,\ldots,n\}$:

$$\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q}) = -\mathbf{T}(\boldsymbol{X},\mathbf{Q},\overline{\mathbf{P}}) \tag{88}$$

$$\Delta(\overline{\mathbf{P}},\mathbf{Q}) = \mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q})\right] = -\Delta(\mathbf{Q},\overline{\mathbf{P}}) \tag{89}$$

$$\Delta(\overline{\mathbf{P}},\mathbf{Q}) \leqslant a_0\boldsymbol{\ell}\left(\mathbf{P}^\star,\overline{\mathbf{P}}\right) - a_1\boldsymbol{\ell}\left(\mathbf{P}^\star,\mathbf{Q}\right) \tag{90}$$

$$\geqslant a_1\boldsymbol{\ell}\left(\mathbf{P}^\star,\overline{\mathbf{P}}\right) - a_0\boldsymbol{\ell}\left(\mathbf{P}^\star,\mathbf{Q}\right). \tag{91}$$

Furthermore, since the X_i are independent, when Assumption 2 is satisfied

$$\operatorname{Var}\left(\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q})\right) \leqslant a_2\left[\boldsymbol{\ell}\left(\mathbf{P}^\star,\overline{\mathbf{P}}\right) + \boldsymbol{\ell}\left(\mathbf{P}^\star,\mathbf{Q}\right)\right]. \tag{92}$$

It follows from (88) and the definition (13) of $\widehat{\mathbf{P}}$ that

$$-\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\widehat{\mathbf{P}}) = \mathbf{T}(\boldsymbol{X},\widehat{\mathbf{P}},\overline{\mathbf{P}}) \leqslant \mathbf{T}(\boldsymbol{X},\widehat{\mathbf{P}}) \leqslant \mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}}) + \epsilon. \tag{93}$$

Setting

$$\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}}) = \sup_{\mathbf{Q}\in\mathscr{M}}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q}) - (1-\kappa)\Delta(\overline{\mathbf{P}},\mathbf{Q}) - \zeta\right] \tag{94}$$

we deduce from (90) and the fact that $\kappa \in (0,1)$ that

$$\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}}) = \sup_{\mathbf{Q}\in\mathscr{M}}\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q}) \leqslant \mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}}) + (1-\kappa)\sup_{\mathbf{Q}\in\mathscr{M}}\Delta(\overline{\mathbf{P}},\mathbf{Q}) + \zeta$$

$$\leqslant \mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}}) + (1-\kappa)\left[a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}}) - a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M})\right] + \zeta, \tag{95}$$

which together with (93) leads to

$$\begin{aligned} & -\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\widehat{\mathbf{P}}) \\ & \leqslant \mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}})+(1-\kappa)\left[a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}})-a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M})\right]+\zeta+\epsilon. \end{aligned} \tag{96}$$

Using (90) again with $\mathbf{Q}=\widehat{\mathbf{P}}$, we obtain that

$$\begin{aligned} (1-\kappa)a_1\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}}) & \\ & \leqslant (1-\kappa)a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}})-(1-\kappa)\Delta(\overline{\mathbf{P}},\widehat{\mathbf{P}}) \\ & = (1-\kappa)a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}})+\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\widehat{\mathbf{P}})-(1-\kappa)\Delta(\overline{\mathbf{P}},\widehat{\mathbf{P}})-\zeta\right] \\ & \quad -\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\widehat{\mathbf{P}})+\zeta. \end{aligned}$$

which with (94) and (96) gives

$$\begin{aligned} (1-\kappa)a_1\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}}) & \\ & \leqslant (1-\kappa)a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}})+\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}}) \\ & \quad +\left[\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}})+(1-\kappa)\left(a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}})-a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M})\right)+\zeta+\epsilon\right]+\zeta \\ & \leqslant (1-\kappa)\left[2a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}})-a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M})\right]+2\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}})+2\zeta+\epsilon. \end{aligned} \tag{97}$$

Let us now control $\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}})$. To do so introduce $\delta=2$, y_0 a positive numbers to be chosen later on and for all $j\in\mathbb{N}$,

$$\mathscr{M}_j=\left\{\mathbf{Q}\in\mathscr{M},\ (\delta^j-1)y_0\leqslant\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})<(\delta^{j+1}-1)y_0\right\}, \tag{98}$$

so that $\mathscr{M}=\bigcup_{j\geqslant 0}\mathscr{M}_j$. Besides, for $j\in\mathbb{N}$ and $\mathbf{x}=(x_1,\ldots,x_n)\in\mathbf{E}$ we set

$$\overline{\mathbf{Z}}_j(\mathbf{x},\overline{\mathbf{P}})=\sup_{\mathbf{Q}\in\mathscr{M}_j}\left[\mathbf{T}(\mathbf{x},\overline{\mathbf{P}},\mathbf{Q})-\Delta(\overline{\mathbf{P}},\mathbf{Q})\right] \tag{99}$$

$$=\sup_{\mathbf{Q}\in\mathscr{M}_j}\left[\sum_{i=1}^n\left(\phi_{(\overline{P}_i,Q_i)}(x_i)-\mathbb{E}\left[\phi_{(\overline{P}_i,Q_i)}(X_i)\right]\right)\right]. \tag{100}$$

For all $j\in\mathbb{N}$ and $\mathbf{Q}\in\mathscr{M}_j$, we deduce from (89) and (91) that

$$\begin{aligned} -\Delta(\overline{\mathbf{P}},\mathbf{Q})=\Delta(\mathbf{Q},\overline{\mathbf{P}}) & \geqslant a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})-a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}}) \\ & \geqslant a_1(\delta^j-1)y_0-a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}}) \end{aligned}$$

hence,

$$\begin{aligned} \mathbf{Z}(\boldsymbol{X},\overline{\mathbf{P}}) & =\sup_{j\in\mathbb{N}}\sup_{\mathbf{Q}\in\mathscr{M}_j}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q})-(1-\kappa)\Delta(\overline{\mathbf{P}},\mathbf{Q})-\zeta\right] \\ & \leqslant\sup_{j\in\mathbb{N}}\left[\overline{\mathbf{Z}}_j(\boldsymbol{X},\overline{\mathbf{P}})-\kappa\inf_{\mathbf{Q}\in\mathscr{M}_j}\Delta(\mathbf{Q},\overline{\mathbf{P}})-\zeta\right] \\ & =\kappa a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}})+\sup_{j\in\mathbb{N}}\Xi_j \end{aligned} \tag{101}$$

with

$$\Xi_j=\overline{\mathbf{Z}}_j(\boldsymbol{X},\overline{\mathbf{P}})-\kappa a_1(\delta^j-1)y_0-\zeta. \tag{102}$$

Putting (97) and (101) together leads to the inequality

$$(1-\kappa)a_1\boldsymbol{\ell}(\mathbf{P}^\star,\widehat{\mathbf{P}}) \leqslant 2a_0\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}}) - (1-\kappa)a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathscr{M}) + 2\zeta + \epsilon + 2\max_{j\geqslant 0}\Xi_j. \tag{103}$$

It remains to control the random variables Ξ_j for $j\in\mathbb{N}$. This is the purpose of the two following lemmas.

Lemma 3. *Under the assumptions of Theorem 1 and for the choices* $y_0 = \sqrt{n}v(\overline{\mathbf{P}})$ *and* $\zeta = 2c_0y_0 + \sqrt{n\xi/2}$

$$\mathbb{P}\left[\Xi_j > 0\right] \leqslant \exp\left[-\xi - (\delta^{j+1}-1)^2\right] \quad \textit{for all } j\in\mathbb{N}.$$

Lemma 4. *Under the assumptions of Theorem 2 and for the choices*

$$y_0 = D(\overline{\mathbf{P}}) \quad \textit{and} \quad \zeta = \frac{2\kappa a_1}{15}\boldsymbol{\ell}(\mathbf{P}^\star,\overline{\mathbf{P}}) + \frac{\kappa a_1}{3}y_0 + 2\left[1+\frac{15a_2}{2\kappa a_1}\right]\xi, \tag{104}$$

we obtain that

$$\mathbb{P}\left[\Xi_j > 0\right] \leqslant \exp\left[-\xi - 2^{j+1}+1\right] \quad \textit{for all } j\geqslant 0.$$

The proofs of Lemma 3 and 4 are given in Section 9.1 and 9.2 respectively. We conclude the proofs of the main theorems as follows. Theorem 1 follows from (103) and Lemma 3 together with the fact that since $\delta = 2$

$$\sum_{j\geqslant 0}\mathbb{P}\left[\Xi_j > 0\right] \leqslant \exp\left[-\xi - (2^{j+1}-1)^2\right] \leqslant e^{-\xi}\sum_{j\geqslant 0} e^{-(2^{j+1}-1)^2} \leqslant 0.37e^{-\xi}.$$

Similarly, we obtain Theorem 2 from Lemma 4 and the fact that

$$\sum_{j\geqslant 0}\mathbb{P}\left[\Xi_j > 0\right] \leqslant \exp\left[-\xi - 2^{j+1}+1\right] \leqslant 0.42e^{-\xi}.$$

9.1. **Proof of Lemma 3.** Let us fix $j\in\mathbb{N}$. It follows from the definition (17) of $v(\overline{\mathbf{P}}) = y_0/\sqrt{n}$ and the fact that $y\mapsto \mathbf{w}(\overline{\mathbf{P}},y)$ is nondecreasing

$$\mathbf{w}(\overline{\mathbf{P}},y_0) \leqslant \mathbf{w}(\overline{\mathbf{P}},y) \leqslant c_0y \quad \text{for all } y > y_0. \tag{105}$$

In particular by letting y decrease to y_0, we also obtain that $\mathbf{w}(\overline{\mathbf{P}},y_0)\leqslant c_0y_0$. Since $\mathscr{M}_j$ is a subset of $\mathscr{B}(\mathbf{P}^\star,r_j)$ with $r_j = (\delta^{j+1}-1)y_0 \geqslant y_0$,

$$\overline{\mathbf{Z}}_j(\boldsymbol{X},\overline{\mathbf{P}}) \leqslant \sup_{\mathbf{Q}\in\mathscr{B}(\mathbf{P}^\star,r_j)}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q}) - \Delta(\overline{\mathbf{P}},\mathbf{Q})\right]$$

and because of (105) (and the facts that the inequality is also true for $y=y_0$ and $r_0 = (\delta-1)y_0 = y_0$),

$$\mathbb{E}\left[\overline{\mathbf{Z}}_j(\boldsymbol{X},\overline{\mathbf{P}})\right] \leqslant \mathbf{w}(\overline{\mathbf{P}},r_j) \leqslant c_0r_j. \tag{106}$$

Under Assumption 1-(iii), for all $i\in\{1,\ldots,n\}$, $\mathbf{Q}\in\mathscr{M}_j$ and $x,x'\in E$ the quantity $|\phi_{(\overline{P}_i,Q_i)}(x) - \phi_{(\overline{P}_i,Q_i)}(x')|$ is bounded by 1 and it follows thus from the expression (100) of $\overline{\mathbf{Z}}_j$ that for all $\mathbf{x}\in\mathbf{E}$, and $x_i'\in E$

$$\left|\overline{\mathbf{Z}}_j((x_1,\ldots,x_i,\ldots,x_n),\overline{\mathbf{P}}) - \overline{\mathbf{Z}}_j((x_1,\ldots,x_i',\ldots,x_n),\overline{\mathbf{P}})\right| \leqslant 1.$$

The random variables $X_1, \ldots, X_n$ being independent, Theorem 5.1 of Massart (2007) applies to $\mathbf{x} \to \overline{\mathbf{Z}}_j(\mathbf{x}, \overline{\mathbf{P}})$ and we obtain that whatever $z > 0$, with a probability at least $1 - e^{-z}$,

$$\overline{\mathbf{Z}}_j(\boldsymbol{X}, \overline{\mathbf{P}}) \leqslant \mathbb{E}\left[\overline{\mathbf{Z}}_j(\boldsymbol{X}, \overline{\mathbf{P}})\right] + \sqrt{nz/2} \leqslant c_0 r_j + \sqrt{nz/2}, \tag{107}$$

by (106). We recall that $r_j = (\delta^{j+1} - 1)y_0$, $y_0 = \sqrt{n}v(\overline{\mathbf{P}})$ and that $\zeta = 2c_0 y_0 + \sqrt{n\xi/2}$. We choose $z = z_j = 2c_0^2(\delta^{j+1} - 1)^2 y_0^2/n + \xi$ which satisfies

$$z_j = \left[2c_0^2 v^2(\overline{\mathbf{P}})\right](\delta^{j+1} - 1)^2 + \xi = (\delta^{j+1} - 1)^2 + \xi,$$

since by (17) $2c_0^2 v^2(\overline{\mathbf{P}}) \geqslant 1$. Using the subadditive property of the square root, we infer from (107) that Ξ_j defined by (102) satisfies with a probability at least $1 - e^{-z_j} \geqslant 1 - e^{-(\delta^{j+1}-1)^2 - \xi}$,

$$\begin{aligned}
\Xi_j &\leqslant c_0\left(\delta^{j+1} - 1\right) y_0 + \sqrt{c_0^2(\delta^{j+1} - 1)^2 y_0^2 + n\xi/2} - \kappa a_1(\delta^j - 1)y_0 - \zeta \\
&\leqslant \left(\delta^{j+1} - 1\right) y_0 \left[2c_0 - \kappa a_1 \frac{\delta^j - 1}{\delta^{j+1} - 1}\right] - 2c_0 y_0.
\end{aligned}$$

Finally note that the the right-hand side is always non-positive: when $j = 0$ it vanishes and for $j \geqslant 1$ it is negative since

$$2c_0 - \kappa a_1 \frac{\delta^j - 1}{\delta^{j+1} - 1} = \kappa a_1 \left[\frac{1}{3} - \frac{2^j - 1}{2^{j+1} - 1}\right] \leqslant 0.$$

9.2. **Proof of Lemma 4.** Arguing as in the proof of Lemma 4, we deduce similarly from the definition (23) of $D(\overline{\mathbf{P}})$ that for all $y \geqslant y_0 = D(\overline{\mathbf{P}})$, $\mathbf{w}(\overline{\mathbf{P}}, y_0) \leqslant \mathbf{w}(\overline{\mathbf{P}}, y) \leqslant c_1 y$ and consequently that

$$\mathbb{E}\left[\left|\overline{\mathbf{Z}}_j(\boldsymbol{X}, \overline{\mathbf{P}})\right|\right] \leqslant c_1\left(\delta^{j+1} - 1\right) y_0 \quad \text{for all } j \geqslant 0. \tag{108}$$

Let us recall the following version of Talagrand's inequality that can be found in Baraud, Birgé and Sart (2017)

Proposition 18. *Let T be some finite set, $U_1, \ldots, U_n$ be independent centered random vectors with values in $\mathbb{R}^T$ and $Z = \sup_{t \in T} \left|\sum_{i=1}^n U_{i,t}\right|$. If for some positive numbers b and v,*

$$\max_{i=1,\ldots,n} |U_{i,t}| \leqslant b \qquad \text{and} \qquad \sum_{i=1}^n \mathbb{E}\left[U_{i,t}^2\right] \leqslant v^2 \quad \text{for all } t \in T,$$

then, for all positive c and x,

$$\mathbb{P}\left[Z \leqslant (1 + c)\mathbb{E}(Z) + (8b)^{-1} c v^2 + 2\left(1 + 8c^{-1}\right) bx\right] \geqslant 1 - e^{-x}. \tag{109}$$

The above result extends to countable sets T (by monotone convergence) and we may therefore take $T = \mathscr{M}_j$, $U_{i,\mathbf{Q}} = \phi_{(\overline{P}_i, Q_i)}(X_i) - \mathbb{E}\left[\phi_{(\overline{P}_i, Q_i)}(X_i)\right]$

for all $i \in \{1, \ldots, n\}$, so that $Z = \left|\overline{\mathbf{Z}}_j(\boldsymbol{X}, \overline{\mathbf{P}})\right|$, and $b = 1$ because of Assumption 1-(iii). Furthermore, using Assumption 2, the definition of $\mathscr{M}_j$ and (92) we obtain that

$$\begin{aligned} v^2 &= a_2\left[\boldsymbol{\ell}(\mathbf{P}^\star, \overline{\mathbf{P}}) + (\delta^{j+1} - 1)y_0\right] \\ &\geqslant a_2 \sup_{\mathbf{Q}\in\mathscr{M}_j}\left[\boldsymbol{\ell}(\mathbf{P}^\star, \overline{\mathbf{P}}) + \boldsymbol{\ell}(\mathbf{P}^\star, \mathbf{Q})\right] \geqslant \sup_{\mathbf{Q}\in\mathscr{M}_j} \mathrm{Var}[\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{P}}, \mathbf{Q})]. \end{aligned}$$

Applying Proposition 18 with $x = z_j = (\delta^{j+1} - 1) + \xi$ and using (108) with the fact that $y_0 \geqslant c_1^{-1}$, we obtain that with a probability at least $1 - e^{-z_j}$,

$$\begin{aligned} \overline{\mathbf{Z}}_j(\boldsymbol{X}, \overline{\mathbf{P}}) &\leqslant (1+c)\mathbb{E}\left[Z\right] + (a_2 c/8)\left[\boldsymbol{\ell}(\mathbf{P}^\star, \overline{\mathbf{P}}) + (\delta^{j+1} - 1)y_0\right] \\ &\quad + 2[1 + (8/c)]z_j \\ &\leqslant (1+c)c_1\left(\delta^{j+1} - 1\right)y_0 + (a_2 c/8)\left[\boldsymbol{\ell}(\mathbf{P}^\star, \overline{\mathbf{P}}) + (\delta^{j+1} - 1)y_0\right] \\ &\qquad + 2(1 + (8/c))\left[(\delta^{j+1} - 1) + \xi\right] \\ &\leqslant \left[(1+c)c_1 + a_2 c/8 + 2c_1[1 + (8/c)]\right](\delta^{j+1} - 1)y_0 \\ &\qquad + (a_2 c/8)\boldsymbol{\ell}(\mathbf{P}^\star, \overline{\mathbf{P}}) + 2[1 + (8/c)]\xi. \end{aligned} \tag{110}$$

Let us now choose $c = 16\kappa l a_1/[a_2(1+\delta)]$ with $l = 0.2$. Note that $c_1 = 2l^2\kappa^2 a_1^2/[a_2(1+\delta)^2]$ so that $(c/8)^2 = 2c_1/a_2$ and $c \leqslant 16l/(1+\delta)$ since $(a_1/a_2) \vee \kappa \leqslant 1$. With such choices (and the fact that $\delta = 2$),

$$\begin{aligned} \frac{c_1}{a_2}\left(3 + c + \frac{16}{c}\right) + \frac{c}{8} &= \frac{1}{2}\left(\frac{c}{8}\right)^2\left(3 + c + \frac{16}{c}\right) + \frac{c}{8} \\ &= \frac{c}{4}\left(1 + \frac{c(3+c)}{32}\right) \leqslant \frac{c}{4}\left(1 + \frac{l\left[3(1+\delta) + 16l\right]}{2(1+\delta)^2}\right) \\ &= \frac{\kappa a_1}{a_2(1+\delta)} \times 4l\left(1 + \frac{l\left[3(1+\delta) + 16l\right]}{2(1+\delta)^2}\right) \\ &< \frac{\kappa a_1}{a_2(1+\delta)} = \frac{\kappa a_1}{3a_2} \end{aligned} \tag{111}$$

and

$$2\left(1 + \frac{8}{c}\right) = 2\left(1 + \frac{(1+\delta)a_2}{2\kappa l a_1}\right) = 2\left(1 + \frac{15a_2}{2\kappa a_1}\right). \tag{112}$$

Using (110), (111) and (112) with the fact that ζ defined by (104) satisfies $\zeta = (a_2 c/8)\boldsymbol{\ell}(\mathbf{P}^\star, \overline{\mathbf{P}}) + \kappa a_1 y_0/3 + 2(1 + 8/c)\xi$, we obtain that Ξ_j defined by

(102) satisfies with a probability at least $1-e^{-z_j}$,

$$\begin{aligned}\Xi_j &\leqslant \left[\left[(1+c)c_1+\frac{a_2c}{8}+2c_1\left(1+\frac{8}{c}\right)\right]-\kappa a_1\frac{\delta^j-1}{\delta^{j+1}-1}\right]\left(\delta^{j+1}-1\right)y_0-\frac{\kappa a_1}{3}y_0\\ &=\left[\frac{c_1}{a_2}\left(3+c+\frac{16}{c}\right)+\frac{c}{8}-\kappa\frac{a_1}{a_2}\frac{\delta^j-1}{\delta^{j+1}-1}\right]a_2\left(\delta^{j+1}-1\right)y_0-\frac{\kappa a_1}{3}y_0\\ &\leqslant\left[\frac{\kappa a_1}{a_2(1+\delta)}-\kappa\frac{a_1}{a_2}\frac{\delta^j-1}{\delta^{j+1}-1}\right]a_2\left(\delta^{j+1}-1\right)y_0-\frac{\kappa a_1}{3}y_0\\ &=\left[\frac{1}{1+\delta}-\frac{\delta^j-1}{\delta^{j+1}-1}\right]\kappa a_1\left(\delta^{j+1}-1\right)y_0-\frac{\kappa a_1}{3}y_0\end{aligned}$$

Since $\delta=2$, when $j=0$ the right-hand side vanishes and for $j\geqslant 1$,

$$\frac{1}{1+\delta}-\frac{\delta^j-1}{\delta^{j+1}-1}=\frac{1}{3}-\frac{2^j-1}{2^{j+1}-1}\leqslant 0.$$

We conclude that $\Xi_j\leqslant 0$ with probability at least $1-e^{-z_j}=1-e^{-\xi-(2^{j+1}-1)}$ for all $j\in\mathbb{N}$.

10. Other proofs

10.1. **Proof of Proposition 14.** Let $\theta,\theta'\in\mathbb{Q}$, $\theta\neq\theta'$. Since f is decreasing on $(0,+\infty)$, for all $x\in\mathbb{R}\setminus\{\theta,\theta'\}$,

$$\begin{aligned}p_{\theta'}(x)>p_\theta(x)&\iff f(|x-\theta'|)>f(|x-\theta|)\iff|x-\theta'|<|x-\theta|\\ &\iff\begin{cases}x>(\theta+\theta')/2 & \text{if}\quad \theta'>\theta\\ x<(\theta+\theta')/2 & \text{if}\quad\theta'<\theta\end{cases}\end{aligned}\tag{113}$$

and

$$\begin{aligned}p_{\theta'}(x)=p_\theta(x)&\iff f(|x-\theta'|)=f(|x-\theta|)\iff|x-\theta'|=|x-\theta|\\ &\iff x=\frac{\theta+\theta'}{2}\in\mathbb{Q}.\end{aligned}\tag{114}$$

Using that p is symmetric and doing the change of variables $u=\theta+\theta'-x$, i.e. $x=\theta+\theta'-u$, we obtain that

$$\begin{aligned}P_\theta\left[p_{\theta'}>p_\theta\right]&=\int_{\mathbb{R}}1\!\!1_{p(x-\theta')>p(x-\theta)}p(x-\theta)dx=\int_{\mathbb{R}}1\!\!1_{p(\theta-u)>p(\theta'-u)}p(\theta'-u)du\\ &=\int_{\mathbb{R}}1\!\!1_{p(u-\theta)>p(u-\theta')}p(u-\theta')du=P_{\theta'}\left[p_\theta>p_{\theta'}\right]\end{aligned}$$

and by (114),

$$P_\theta\left[p_{\theta'}=p_\theta\right]=P_{\theta'}\left[p_{\theta'}=p_\theta\right]=0.$$

It follows from Corollary 7 and the fact that with probability 1 $\{X_1,\ldots,X_n\}\cap\mathbb{Q}=\varnothing$, that with probability 1, for all $i\in\{1,\ldots,n\}$

$$\begin{aligned}\phi_{(P_\theta,P_{\theta'})}(X_i)&=\frac{1}{2}\left[P_\theta\left(p_\theta>p_{\theta'}\right)+P_{\theta'}\left(p_\theta>p_{\theta'}\right)\right]-\mathbb{1}_{p_\theta>p_{\theta'}(X_i)}\\&=\frac{1}{2}\left[1-P_\theta\left(p_\theta=p_{\theta'}\right)+P_{\theta'}\left(p_\theta>p_{\theta'}\right)-P_\theta\left(p_{\theta'}>p_\theta\right)\right]\\&\quad-\mathbb{1}_{p_\theta>p_{\theta'}(X_i)}\\&=\frac{1}{2}-\mathbb{1}_{p_\theta>p_{\theta'}(X_i)}\end{aligned}$$

and by (113) we obtain that

$$\mathbf{T}(\boldsymbol{X},\mathbf{P}_\theta,\mathbf{P}_{\theta'})=\frac{n}{2}-\sum_{i=1}^n\mathbb{1}_{p_\theta>p_{\theta'}}(X_i)=\frac{n}{2}-\begin{cases}\sum_{i=1}^n\mathbb{1}_{X_i>(\theta+\theta')/2}&\text{if}\quad\theta>\theta',\\\sum_{i=1}^n\mathbb{1}_{X_i<(\theta+\theta')/2}&\text{if}\quad\theta<\theta'.\end{cases}$$

Let us now take $\theta=\widehat{\theta}\in(X_{(\lceil n/2\rceil)},X_{(\lceil n/2\rceil+1)})$. It follows from (80) that if $\theta'>\widehat{\theta}$,

$$\sum_{i=1}^n\mathbb{1}_{X_i<(\widehat{\theta}+\theta')/2}\geqslant\sum_{i=1}^n\mathbb{1}_{X_i\leqslant\widehat{\theta}}\geqslant\sum_{i=1}^n\mathbb{1}_{X_i\leqslant X_{(\lceil n/2\rceil)}}\geqslant\frac{n}{2}$$

and consequently, $\mathbf{T}(\boldsymbol{X},\mathbf{P}_{\widehat{\theta}},\mathbf{P}_{\theta'})\leqslant 0$.

If now $\theta'<\widehat{\theta}$ we may distinguish between two cases. Since $(\theta'+\widehat{\theta})/2\in\mathbb{Q}$,

$$\text{either}\quad\frac{\theta'+\widehat{\theta}}{2}<X_{(\lceil n/2\rceil)}<\widehat{\theta}\quad\text{or}\quad X_{(\lceil n/2\rceil)}<\frac{\theta'+\widehat{\theta}}{2}<\widehat{\theta}<X_{(\lceil n/2\rceil+1)}.$$

In the first case, using (80) again we obtain that

$$\sum_{i=1}^n\mathbb{1}_{X_i>(\theta'+\widehat{\theta})/2}=n-\sum_{i=1}^n\mathbb{1}_{X_i\leqslant(\theta'+\widehat{\theta})/2}\geqslant n-\sum_{i=1}^n\mathbb{1}_{X_i<X_{(\lceil n/2\rceil)}}>\frac{n}{2}$$

hence, $\mathbf{T}(\boldsymbol{X},\mathbf{P}_{\widehat{\theta}},\mathbf{P}_{\theta'})<0$. In the second case,

$$\sum_{i=1}^n\mathbb{1}_{X_i>(\widehat{\theta}+\theta')/2}=n-\left\lceil\frac{n}{2}\right\rceil\geqslant\frac{n-1}{2}$$

which implies that $\mathbf{T}(\boldsymbol{X},\mathbf{P}_{\widehat{\theta}},\mathbf{P}_{\theta'})\leqslant 1/2$.

Putting all these bounds together, we finally obtain that

$$\mathbf{T}(\boldsymbol{X},\mathbf{P}_{\widehat{\theta}})=\sup_{\theta'\in\mathbb{Q}}\mathbf{T}(\boldsymbol{X},\mathbf{P}_{\widehat{\theta}},\mathbf{P}_{\theta'})\leqslant\frac{1}{2}\leqslant\inf_{\theta\in\mathbb{Q}}\sup_{\theta'\in\mathbb{Q}}\mathbf{T}(\boldsymbol{X},\mathbf{P}_\theta,\mathbf{P}_{\theta'})+\frac{1}{2}$$

since $\sup_{\theta'\in\mathbb{Q}}\mathbf{T}(\boldsymbol{X},\mathbf{P}_\theta,\mathbf{P}_{\theta'})\geqslant\mathbf{T}(\boldsymbol{X},\mathbf{P}_\theta,\mathbf{P}_\theta)=0$ for all $\theta\in\mathbb{Q}$. Hence $\widehat{\theta}$ is an ℓ–estimator for the choice $\epsilon=1/2$.

10.2. **Proof of Proposition 2.** Let $P,Q\in\overline{\mathscr{P}}$. Since $f_{(P,Q)}=-f_{(Q,P)}$, $\phi_{(P,Q)}=-\phi_{(Q,P)}$ and since $\|f_{(P,Q)}\|_\infty\leqslant b$, for all $x,x'\in E$,

$$\phi_{(P,Q)}(x)-\phi_{(P,Q)}(x')\leqslant\frac{1}{2b}\left(f_{(Q,P)}(x')-f_{(Q,P)}(x)\right)\leqslant 1.$$

Using the definitions of ℓ and $f_{(P,Q)}\in\mathcal{F}$ we derive that for all $S\in\mathscr{P}\subset\overline{\mathscr{P}}$,

$$\begin{aligned}2b\mathbb{E}_S\left[\phi_{(P,Q)}\right]&=\int_E f_{(P,Q)}\frac{dP+dQ}{2}-\int_E f_{(P,Q)}dS\\&=\int_E f_{(P,Q)}dP-\int_E f_{(P,Q)}dS\\&\quad+\frac{1}{2}\left[\int_E f_{(P,Q)}dP+\int_E f_{(P,Q)}dQ\right]-\int_E f_{(P,Q)}dP\\&=\int_E f_{(Q,P)}dS-\int_E f_{(Q,P)}dP-\frac{1}{2}\left[\int_E f_{(P,Q)}dP-\int_E f_{(P,Q)}dQ\right]\\&\leqslant\ell(S,P)-\frac{1}{2}\ell(P,Q).\end{aligned}$$

Finally, since by the triangle inequality $\ell(P,Q)\geqslant\ell(S,Q)-\ell(S,P)$ the conclusion follows.

10.3. **Proof of Proposition 9.** Let $S\in\mathscr{P}$. Let us first prove (66). Using the definition (44) of the TV-distance,

$$S(p>q)\leqslant S(p>q)-Q(p>q)+Q(p>q)\leqslant\|S-Q\|+Q(p>q)\tag{115}$$

and similarly, $S(p\leqslant q)\leqslant\|S-P\|+P(p\leqslant q)$. Inequality (65) with the triangle inequality leads to

$$\begin{aligned}S(p>q)\wedge S(p\leqslant q)&\leqslant\|S-P\|\vee\|S-Q\|+P(p\leqslant q)\wedge Q(p>q)\\&\leqslant\|S-P\|+\|S-Q\|+a_2'\|P-Q\|\\&\leqslant(1+a_2')\left[\|S-P\|+\|S-Q\|\right]\end{aligned}$$

which is (66).

Let us now establish Condition (iv). Using the definition (45) of $\phi_{(P,Q)}$ with (66), we obtain that

$$\begin{aligned}\mathrm{Var}_S\left[\phi_{(P,Q)}\right]&=S(p>q)S(p\leqslant q)\leqslant S(p>q)\wedge S(p\leqslant q)\\&\leqslant a_2\left[\|S-P\|+\|S-Q\|\right]\end{aligned}$$

and the result follows from the fact that $a_2=1+a_2'\geqslant 1/2=a_1$.

10.4. **Proof of Proposition 10.** For all $x\in E$ and $g\in\mathcal{G}$

$$\sqrt{p(x)q(x)}\leqslant\frac{1}{2}\left[g(x)p(x)+(1/g(x))q(x)\right]$$

with the conventions $(+\infty)\times 0=0$ and $(+\infty)\times a=(+\infty)$ for all $a>0$. Note that equality holds for $g=g_{(P,Q)}=\sqrt{q/p}$ with our conventions. Integrating with respect to μ gives

$$\int_E \sqrt{pq}d\mu = 1 - h^2(P,Q) \leqslant \frac{1}{2}\left[\mathbb{E}_P(g) + \mathbb{E}_Q(1/g)\right] \in [0,+\infty]$$

and consequently for all $g \in \mathcal{G}$

$$h^2(P,Q) \geqslant \frac{1}{2}\left[\mathbb{E}_P(1-g) + \mathbb{E}_Q(1-1/g)\right] \in [-\infty, 1]$$

and equality holds for $g = g_{(P,Q)}$ which leads to the result.

10.5. **Proof of Proposition 11.** Let us set $\overline{\phi}_{(P,Q)} = \phi_{(P,Q)}\sqrt{2}$ and denote by $r = (p+q)/2$ the density of R with respect to μ. Using that $(p \vee q)/r \leqslant 2$, for all $x, x' \in E$

$$\overline{\phi}_{(P,Q)}(x) - \overline{\phi}_{(P,Q)}(x') \leqslant \frac{1}{2}\left[\sqrt{\frac{q}{r}}(x) + \sqrt{\frac{p}{r}}(x')\right] \leqslant \sqrt{2}$$

hence, $\phi_{(P,Q)} = \overline{\phi}_{(P,Q)}/\sqrt{2}$ takes its values in $[-1,1]$. For $T = t \cdot \mu \in \{P,Q\}$, we set

$$\rho_r(S,t) = \frac{1}{2}\left[\rho(R,T) + \mathbb{E}_S\left(\sqrt{\frac{t}{r}}(X)\right)\right],$$

so that

$$\begin{aligned}\mathbb{E}_S\left[\overline{\phi}_{(P,Q)}(X)\right] &= \rho_r(S,q) - \rho_r(S,p) \\ &= \rho_r(S,q) - \rho(S,Q) + \rho(S,Q) - \rho(S,P) + \rho(S,P) - \rho_r(S,p).\end{aligned}$$

By Proposition 1 of Baraud (2011) (which requires that $S \ll \mu$), $0 \leqslant \rho_r(S,t) - \rho(S,T) \leqslant [h^2(S,P) + h^2(S,Q)]/\sqrt{2}$ for all $T \in \{P,Q\}$ and since $\rho(S,Q) - \rho(S,P) = h^2(S,P) - h^2(S,Q)$, we deduce that

$$\begin{aligned}\mathbb{E}_S\left[\overline{\phi}_{(P,Q)}(X)\right] &\leqslant \frac{1}{\sqrt{2}}\left[h^2(S,P) + h^2(S,Q)\right] + h^2(S,P) - h^2(S,Q) \\ &\leqslant \left(1 + \frac{1}{\sqrt{2}}\right)h^2(S,P) - \left(1 - \frac{1}{\sqrt{2}}\right)h^2(S,Q).\end{aligned}$$

Hence (ii) is satisfied with $a_0 = (\sqrt{2}+1)/2$ and $a_1 = (\sqrt{2}-1)/2$. Using that

$$4\,\mathrm{Var}_S\left[\overline{\phi}_{(P,Q)}(X)\right] = \mathrm{Var}_S\left[\frac{\sqrt{p}-\sqrt{q}}{\sqrt{r}}(X)\right] \leqslant \mathbb{E}_S\left[\frac{\left(\sqrt{p(X)} - \sqrt{q(X)}\right)^2}{r(X)}\right]$$

condition (iv) with $a_2 = 3/2$ follows from the proof of Proposition 3 of Baraud (2011).

10.6. **Proof of Proposition 12.** It is clear from the definition (76) that $\phi_{(P,Q)} = -\phi_{(Q,P)}$ and that under (75) $\phi_{(P,Q)}(x) - \phi_{(P,Q)}(x') \leqslant 1$ for all $x, x' \in E$. Using the definition of the Kullback-Liebler divergence and the fact that $\int_E s\,|\log s|\,d\mu$ is finite, we obtain that

$$\begin{aligned}\mathbb{E}_S\left[\phi_{(P,Q)}\right] &= \frac{1}{2a}\mathbb{E}_S\left[\log\left(\frac{q}{p}\right)\right] = \frac{1}{2a}\mathbb{E}_S\left[\log\left(\frac{s}{p}\right) - \log\left(\frac{s}{q}\right)\right]\\ &= \frac{1}{2a}\left[K(S,P) - K(S,Q)\right].\end{aligned}$$

Assumption 1 is therefore satisfied with $a_0 = a_1 = 1/(2a)$. For all $u > -1$, $\log^2(1+u) \leqslant 2u - 2\log(1+u)$ and applying this inequality to the function $-1 + t/s > -1$ when $t \in \{p, q\}$, we obtain that

$$\int_E s\log^2\left(\frac{t}{s}\right) d\mu \leqslant 2\int_E s\left[\frac{t}{s} - 1 - \log\left(\frac{t}{s}\right)\right] d\mu \leqslant 2K(S, t\cdot\mu).$$

Consequently,

$$\begin{aligned}\mathrm{Var}_S\left[\phi_{(P,Q)}(X)\right] \leqslant \mathbb{E}_S\left[\phi^2_{(P,Q)}(X)\right] &= \frac{1}{4a^2}\int_E \log^2\left(\frac{q}{p}\right) s d\mu\\ &\leqslant \frac{1}{2a^2}\left[\int_E \log^2\left(\frac{p}{s}\right) s d\mu + \int_E \log^2\left(\frac{q}{s}\right) s d\mu\right]\\ &\leqslant \frac{1}{a^2}\left[K(S,P) + K(S,Q)\right],\end{aligned}$$

and Assumption 2-(iv) is therefore satisfied with $a_2 = \max\{1/a^2, a_1\} = 1/[a(2\wedge a)]$.

10.7. **Proof of Proposition 3.** Let us denote by $\mathrm{sgn} = \mathrm{sgn}(P, Q, \cdot) = \mathbb{1}_{F_Q > F_P} - \mathbb{1}_{F_P > F_Q}$ the function corresponding to the sign of $F_Q - F_P$ on the set $\{F_Q \neq F_P\}$ and which vanishes elsewhere. We write $f = f_{(P,Q)}$ for short. The function sgn takes its values in $[-1, 1]$ and since P and Q are supported on $[0, 1]$, f vanishes outside the interval $[0, 1]$. It is therefore 1-Lipschitz and bounded by 1. Besides, by Fubini's theorem

$$\begin{aligned}&\mathbb{E}_P\left[f(X)\right] - \mathbb{E}_Q\left[f(X)\right]\\ &= \int_0^1\left[\int_0^1 \mathrm{sgn}(t)\mathbb{1}_{0\leqslant t<x}dt\right] dP(x) - \int_0^1\left[\int_0^1 \mathrm{sgn}(t)\mathbb{1}_{0\leqslant t<x}dt\right] dQ(x)\\ &= \int_0^1 \mathrm{sgn}(t)(1 - F_P(t))dt - \int_0^1 \mathrm{sgn}(t)(1 - F_Q(t))dt\\ &= \int_0^1 \mathrm{sgn}(t)\left(F_Q(t) - F_P(t)\right) dt = \int_0^1 \left|F_P(t) - F_Q(t)\right| dt = W(P,Q).\end{aligned}$$

For the last equality, we refer to Shorack and Wellner (1986)[Page 64].

10.8. **Proof of Proposition 5.** Let $I = \{0,1\}^d \setminus \{(0,\ldots,0)\}$ and consider a multivariate tensor product wavelet basis

$$\{\mathbf{\Phi_k}, \mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}},\ \mathbf{k} \in \mathbb{Z}^d, j \geqslant 0, \mathbf{i} \in I\}$$

of $\mathbb{L}_2(\mathbb{R}^d)$ based on the father and mother wavelets ϕ and ψ defined on $\mathbb{R}$, with compact support, regularity $r > \alpha$ and $\mathbb{L}_2$-norms equal to 1. This means that, for all $\mathbf{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d$, $\mathbf{k} = (k_1,\ldots,k_d) \in \mathbb{Z}^d$, $j \geqslant 0$ and $\mathbf{i} = (i_1,\ldots,i_d) \in I$,

$$\mathbf{\Phi_k}(\mathbf{x}) = \prod_{l=1}^{d} \phi(x_l - k_l) \quad \text{and} \quad \mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}} = 2^{jd/2} \prod_{l=1}^{d} \varphi^{(i_l)}\left(2^j x_l - k_l\right),$$

with $\varphi^{(1)} = \psi$ and $\varphi^{(0)} = \phi$. If a function $f \in \mathscr{L}_2(\mathbb{R}^d)$ writes as

$$f = \sum_{\mathbf{k}\in\mathbb{Z}^d} \left[\langle f, \mathbf{\Phi_k}\rangle \mathbf{\Phi_k} + \sum_{j\geqslant 0}\sum_{\mathbf{i}\in I} \langle f, \mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}\rangle \mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}\right] \quad \text{a.e.} \tag{116}$$

and if it belongs to the Besov space $B^{\alpha}_{s,\infty}(\mathbb{R}^d)$, then

$$|f|'_{\alpha,s,\infty} = \sup_{j\geqslant 0} 2^{j(\alpha+d/2-d/s)} \left(\sum_{\mathbf{k}\in\mathbb{Z}^d,\mathbf{i}\in I} \left|\langle f, \mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}\rangle\right|^s\right)^{1/s} < +\infty \tag{117}$$

and the quantity $|f|'_{\alpha,s,\infty}$ is equivalent to the Besov semi-norm associated to $B^{\alpha}_{s,\infty}(\mathbb{R}^d)$ (up to constants that depend on α, s, d, ϕ, ψ). Therefore, replacing $|f|'_{\alpha,s,\infty}$ by $|f|_{\alpha,s,\infty}$ will only change the values of the constants in what follows. We refer the reader to Section 4.3 of the book by Nickl and Giné (2016) for more details on Besov spaces on $\mathbb{R}^d$ and their connections with multivariate tensor product wavelet bases with regularity r. Since the father and mother wavelets φ, ψ have compact support on $\mathbb{R}$, the functions $\mathbf{\Phi_k}$ and $\mathbf{\Psi}^{\mathbf{i}}_{\mathbf{k}}$ also have compact support on $\mathbb{R}^d$ for all $\mathbf{k} \in \mathbb{Z}^d$, $j \geqslant 0$ and $\mathbf{i} \in I$. In fact, there exists a number $K_0 > 0$, depending on d, φ and ψ only such that for all $\mathbf{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d$, $j \geqslant 0$ the sets

$$\Lambda(\mathbf{x}) = \left\{\mathbf{k} \in \mathbb{Z}^d, |\mathbf{\Phi_k}(\mathbf{x})| > 0\right\} \quad \text{and} \quad \Lambda_j(\mathbf{x}) = \left\{\mathbf{k} \in \mathbb{Z}^d, \sum_{\mathbf{i}\in I} \left|\mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}(\mathbf{x})\right| > 0\right\}$$

have cardinalities not larger than K_0. In particular, for $J \geqslant 0$, the functions t of the form

$$t(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \left[\beta_{\mathbf{k},\mathbf{0}} \mathbf{\Phi_k}(\mathbf{x}) + \sum_{j=0}^{J}\sum_{\mathbf{i}\in I} \beta_{j,\mathbf{k},\mathbf{i}} \mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}(\mathbf{x})\right] \quad \text{for all } \mathbf{x} \in \mathbb{R}^d \tag{118}$$

with

$$\sum_{\mathbf{k}\in\mathbb{Z}^d} \left[\beta^2_{\mathbf{k},\mathbf{0}} + \sum_{j=0}^{J}\sum_{\mathbf{i}\in I} \beta^2_{j,\mathbf{k},\mathbf{i}}\right] < +\infty$$

are well-defined since the series in (118) only involves a finite number of non-zero terms. We define V_J as the linear space of these functions t given by (118) and, for all $j \geqslant 0$, the linear space U_j as the space of functions u of the form

$$u = \sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\beta_{j,\mathbf{k},\mathbf{i}}\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}} \quad \text{with} \quad \sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\beta_{j,\mathbf{k},\mathbf{i}}^2 < +\infty.$$

Since the functions $\boldsymbol{\Phi}_{\mathbf{k}}$ and $\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}$ form an orthonormal system in $\mathbb{L}_2(\mathbb{R}^d)$ for $\mathbf{k} \in \mathbb{Z}^d$, $j \geqslant 0$ and $\mathbf{i} \in I$, the linear spaces $(V_J, \|\cdot\|_2)$ and $(U_j, \|\cdot\|_2)$ with $j \geqslant 0$ are Hilbert spaces. Moreover, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}
&\sum_{\mathbf{k}\in\mathbb{Z}^d}\left[\boldsymbol{\Phi}_{\mathbf{k}}^2(\mathbf{x}) + \sum_{j=0}^{J}\sum_{\mathbf{i}\in I}\left(\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\right)^2(\mathbf{x})\right] \\
&\qquad = \sum_{\mathbf{k}\in\Lambda(\mathbf{x})}\boldsymbol{\Phi}_{\mathbf{k}}^2(\mathbf{x}) + \sum_{j=0}^{J}\left[\sum_{\mathbf{k}\in\Lambda_j(\mathbf{x})}\sum_{\mathbf{i}\in I}\left(\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\right)^2(\mathbf{x})\right] \\
&\qquad \leqslant K_0\left[\|\phi\|_\infty^{2d} + 2^d\max_{\mathbf{i}\in I}\left\|\boldsymbol{\Psi}_{0,\mathbf{0}}^{\mathbf{i}}\right\|_\infty^2\sum_{j=0}^{J}2^{jd}\right] \leqslant K_1^2 2^{Jd},
\end{aligned}$$

where K_1 only depends on d, ϕ and ψ. It follows from (118) and Cauchy-Schwarz inequality that, for all $\mathbf{x} \in \mathbb{R}^d$ and $t \in V_J$,

$$\begin{aligned}
|t(\mathbf{x})|^2 &= \left|\sum_{\mathbf{k}\in\mathbb{Z}^d}\left[\beta_{\mathbf{k},\mathbf{0}}\boldsymbol{\Phi}_{\mathbf{k}}(\mathbf{x}) + \sum_{j=0}^{J}\sum_{\mathbf{i}\in I}\beta_{j,\mathbf{k},\mathbf{i}}\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}(\mathbf{x})\right]\right|^2 \\
&\leqslant \left[\sum_{\mathbf{k}\in\mathbb{Z}^d}\left(\beta_{\mathbf{k},\mathbf{0}}^2 + \sum_{j=0}^{J}\sum_{\mathbf{i}\in I}\beta_{j,\mathbf{k},\mathbf{i}}^2\right)\right]\left[\sum_{\mathbf{k}\in\mathbb{Z}^d}\left(\boldsymbol{\Phi}_{\mathbf{k}}^2(\mathbf{x}) + \sum_{j=0}^{J}\sum_{\mathbf{i}\in I}\left(\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\right)^2(\mathbf{x})\right)\right] \\
&\leqslant \|t\|_2^2 \times K_1^2 2^{Jd}
\end{aligned}$$

which implies that V_J satisfies Assumption 4 with $R = K_1 2^{Jd/2}$.

For all $\mathbf{x} \in \mathbb{R}^d$ and $t \in U_j$ with $j \geqslant 0$

$$\begin{aligned}
|t(\mathbf{x})|^s &= \left|\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\langle t, \boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\rangle\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}(\mathbf{x})\right|^s = \left|\sum_{\mathbf{k}\in\Lambda_j(\mathbf{x})}\sum_{\mathbf{i}\in I}\langle t, \boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\rangle\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}(\mathbf{x})\right|^s \\
&\leqslant (|\Lambda_j(\mathbf{x})|\,|I|)^{s-1}\sum_{\mathbf{k}\in\Lambda_j(\mathbf{x})}\sum_{\mathbf{i}\in I}\left|\langle t, \boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\rangle\right|^s\left|\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}(\mathbf{x})\right|^s
\end{aligned}$$

$$\leqslant (K_0 2^d)^{s-1}\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\left|\langle t, \boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\rangle\right|^s\left|\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}(\mathbf{x})\right|^s. \tag{119}$$

Since, for all $\mathbf{i} \in I$ and $\mathbf{k} \in \mathbb{Z}^d$, $\|\boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\|_s = 2^{jd(1/2-1/s)}\|\boldsymbol{\Psi}_{0,\mathbf{0}}^{\mathbf{i}}\|_s$, integrating (119) with respect to $\mathbf{x} \in \mathbb{R}^d$ leads to the bound,

$$\|t\|_s \leqslant K_2 2^{jd(1/2-1/s)}\left[\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\left|\langle t, \boldsymbol{\Psi}_{j,\mathbf{k}}^{\mathbf{i}}\rangle\right|^s\right]^{1/s} \quad \text{for all } t \in U_j, \tag{120}$$

where K_2 depends on d, ϕ, ψ and s.

Let us now consider a function f in $B^{\alpha}_{s,\infty}\cap\mathscr{L}_1(\mathbb{R}^d)\cap\mathscr{L}_2(\mathbb{R}^d)$. It follows from (116) that f can be expanded in the wavelet basis as $\overline{f}_J+\sum_{j>J}f_j$ a.e. with $\overline{f}_J\in V_J$ and

$$f_j=\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\langle f,\mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}\rangle\mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}\in U_j\quad\text{for all } j>J.$$

Since f belongs to $\mathscr{L}_1(\mathbb{R}^d)$, for all $j\geqslant 0$

$$\begin{aligned}\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\left|\langle f,\mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}\rangle\right| &\leqslant\int_{\mathbb{R}^d}|f(\mathbf{x})|\left[\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\left|\mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}(\mathbf{x})\right|\right]d\mathbf{x}\\ &\leqslant K_0 2^{jd/2}\max_{\mathbf{i}\in I}\left\|\mathbf{\Psi}^{\mathbf{i}}_{0,\mathbf{0}}\right\|_\infty\|f\|_1\end{aligned}$$

and similarly,

$$\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}|\langle f,\mathbf{\Phi}_{\mathbf{k}}\rangle|\leqslant K_0\|\phi\|^d_\infty\|f\|_1.$$

As a consequence, $\overline{f}_J$ and f_j for $j>J$ belong to $\mathscr{L}_1(\mathbb{R}^d)$ and

$$\begin{aligned}\|f_j\|_1=\int_{\mathbb{R}^d}|f_j(\mathbf{x})|\,d\mathbf{x}&\leqslant\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\left|\langle f,\mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}\rangle\right|\int_{\mathbb{R}^d}\left|\mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}(\mathbf{x})\right|d\mathbf{x}\\ &=2^{-jd/2}\max_{\mathbf{i}\in I}\left\|\mathbf{\Psi}^{\mathbf{i}}_{0,\mathbf{0}}\right\|_1\sum_{\mathbf{k}\in\mathbb{Z}^d}\sum_{\mathbf{i}\in I}\left|\langle f,\mathbf{\Psi}^{\mathbf{i}}_{j,\mathbf{k}}\rangle\right|\leqslant K_3\|f\|_1\end{aligned}\tag{121}$$

where K_3 depends on d,ϕ and ψ only. Besides, since f belongs to $B^{\alpha}_{s,\infty}$ we deduce from (117) and (120) that

$$\|f_j\|_s\leqslant K_4|f|_{\alpha,s,\infty}2^{-j\alpha}\quad\text{for all } j>J,\tag{122}$$

where K_4 depends on d,ϕ,ψ,s and α. Combining (121) and (122) and using the fact that $s\geqslant 2$, we derive that for all $j>J$ and $z_j>0$

$$\begin{aligned}\|f_j\|_2^2&=\int_{\mathbb{R}^d}f_j^2(\mathbf{x})\mathbb{1}_{|f_j|\leqslant z_j}d\mathbf{x}+\int_{\mathbb{R}^d}f_j^2(\mathbf{x})\mathbb{1}_{|f_j|>z_j}d\mathbf{x}\\ &\leqslant z_j\|f_j\|_1+\frac{\|f_j\|_s^s}{z_j^{s-2}}\leqslant z_jK_3\|f\|_1+z_j^{-(s-2)}K_4^s|f|^s_{\alpha,s,\infty}2^{-js\alpha}.\end{aligned}$$

Setting $z_j=[K_4^s|f|^s_{\alpha,s,\infty}/(K_3\|f\|_1)]^{1/(s-1)}2^{-js\alpha/(s-1)}$ when $\|f\|_1\neq 0$ and letting z_j tend to $+\infty$ otherwise, we derive that for all $j>J$

$$\|f_j\|_2^2\leqslant K_5|f|^{s/(s-1)}_{\alpha,s,\infty}\|f\|_1^{(s-2)/(s-1)}2^{-js\alpha/(s-1)},\tag{123}$$

where K_5 only depends on d,ϕ,ψ,α and s (with the convention $0^0=0$). Since the spaces U_j are mutually orthogonal, we derive from (123) that

$$\begin{aligned}\left\|f-\overline{f}_J\right\|_2^2=\sum_{j>J}\|f_j\|_2^2&\leqslant K_5^2|f|^{s/(s-1)}_{\alpha,s,\infty}\|f\|_1^{(s-2)/(s-1)}\sum_{j>J}2^{-js\alpha/(s-1)}\\ &\leqslant K_6|f|^{s/(s-1)}_{\alpha,s,\infty}\|f\|_1^{(s-2)/(s-1)}2^{-Js\alpha/(s-1)}\end{aligned}$$

where K_6 depends on d,ϕ,ψ,s and α, which concludes the proof.

10.9. **Proof of Proposition 15.** Let $z = a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q}) - a_0\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{P})$. Since it follows from Proposition 1 that $z \leqslant -\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})\right]$, we derive from (12) that

$$\begin{aligned}\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X})=1\right] &\leqslant \mathbb{P}\left[\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})\geqslant 0\right]\\ &\leqslant \mathbb{P}\left[\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})-\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})\right]\geqslant z\right].\end{aligned}$$

Note that

$$\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})-\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\mathbf{P},\mathbf{Q})\right]=\sum_{i=1}^{n}\left(\phi_{(P_i,Q_i)}(X_i)-\mathbb{E}\left[\phi_{(P_i,Q_i)}(X_i)\right]\right)$$

is a sum of n independent centered random variables with values in $[-1,1]$ to which one can apply Hoeffding's inequality, which gives $\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X})=1\right] \leqslant \exp\left[-z_+^2/(2n)\right]$ when $z = z_+ > 0$. This inequality remains valid when $z_+ = 0$ which proves (83).

When Assumption 2-(iv) is satisfied we proceed in the same way, replacing Hoeffding's inequality by Bernstein's (see inequality (2.16) in Massart (2007)) with $v = a_2\left[\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})+\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{P})\right]$ and $b = 1$ (with Massart's notations). When $a_0\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{P}) = \gamma a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})$ with $0 \leqslant \gamma < 1$ so that $z > 0$, we derive that

$$\begin{aligned}&\mathbb{P}\left[\Phi_{(\mathbf{P},\mathbf{Q})}(\boldsymbol{X})=1\right]\\ &\qquad\leqslant \exp\left[-\frac{1}{2}\frac{z^2}{v+(bz/3)}\right]\\ &\qquad= \exp\left[-\frac{1}{2}\frac{\left[a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})-a_0\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{P})\right]^2}{\left[a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})-a_0\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{P})\right]/3+a_2\left[\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})+\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{P})\right]}\right] \qquad (124)\\ &\qquad= \exp\left[-\frac{1}{2}\frac{(1-\gamma)^2\left[a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})\right]^2}{\left[(1-\gamma)a_1\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})/3\right]+\left[(1+\gamma)a_2\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})\right]}\right]\\ &\qquad= \exp\left[-\frac{\boldsymbol{\ell}(\mathbf{P}^\star,\mathbf{Q})}{2}\frac{a_1(1-\gamma)^2}{\left[(1-\gamma)3\right]+\left[(1+\gamma)a_2/a_1\right]}\right].\end{aligned}$$

which is (84).

10.10. **Proof of Corollary 2.** Let f be two functions on $[0,1]$ that satisfy the following property: there exists a function f' on $[0,1]$ such that $\|f'\|_\infty \leqslant 1$ and

$$f(x)=\int_0^x f'(u)du=\int_0^1 f'(u)\mathbb{1}_{x\geqslant u}du \quad \text{for all } x\in[0,1].$$

Using Fubini's theorem, for such a function f we obtain that

$$\begin{aligned}\left|\sum_{i=1}^{n} f(X_i)-\mathbb{E}\left[f(X_i)\right]\right| &= \left|\sum_{i=1}^{n}\left[\int_0^1 f'(u)\left(\mathbb{1}_{X_i\geqslant u}-\mathbb{P}\left[X_i\geqslant u\right]\right)du\right]\right|\\ &= \left|\int_0^1 f'(u)\sum_{i=1}^{n}\left(\mathbb{1}_{X_i\geqslant u}-\mathbb{P}\left[X_i\geqslant u\right]\right)du\right|\\ &\leqslant \int_0^1\left|\sum_{i=1}^{n}\left(\mathbb{1}_{X_i\geqslant u}-\mathbb{P}\left[X_i\geqslant u\right]\right)\right|du\end{aligned}$$

It follows from Proposition 3 that the functions $f_{(P,Q)}$ defined by (29) satisfy this property for all probabilities $P,Q\in\mathscr{P}$. Let $\overline{P}\in\mathscr{M}$ and $y>0$. Using the definitions of $\phi_{(\overline{P},Q)}$ and $\mathbf{W}(\overline{\mathbf{P}},y)$ (with $\overline{\mathbf{P}}=\overline{P}^{\otimes n}$) given respectively by (30) and (16), we deduce that

$$\begin{aligned}\mathbf{w}(\overline{\mathbf{P}},y) &= \mathbb{E}\left[\sup_{\mathbf{Q}\in\mathscr{B}(\overline{\mathbf{P}},y)}\left|\sum_{i=1}^{n}\phi_{(\overline{P},Q)}(X_i)-\mathbb{E}\left[\phi_{(\overline{P},Q)}(X_i)\right]\right|\right]\\ &= \frac{1}{2}\mathbb{E}\left[\sup_{\mathbf{Q}\in\mathscr{B}(\overline{\mathbf{P}},y)}\left|\sum_{i=1}^{n} f_{(Q,\overline{P})}(X_i)-\mathbb{E}\left[f_{(Q,\overline{P}}(X_i)\right]\right|\right]\\ &\leqslant \frac{1}{2}\int_0^1\mathbb{E}\left[\left|\sum_{i=1}^{n}\left(\mathbb{1}_{X_i\geqslant u}-\mathbb{P}\left[X_i\geqslant u\right]\right)\right|\right]du\\ &\leqslant \frac{1}{2}\int_0^1\sqrt{\sum_{i=1}^{n}\mathrm{Var}(\mathbb{1}_{X_i\geqslant u})}du\leqslant\frac{\sqrt{n}}{4}.\end{aligned}$$

Consequently, $\mathbf{w}(\overline{\mathbf{P}},y)\leqslant c_0y$ for all $y\geqslant\sqrt{n}/(4c_0)$, hence $v(\overline{\mathbf{P}})\leqslant 1/(c_0\sqrt{2})$. Applying Theorem 1 with the values of a_0 and a_1 provided by Proposition 1 and using the fact that $\overline{P}$ is arbitrary in $\mathscr{M}$, we obtain that for all $\xi>0$ with a probability at least $1-0.37e^{-\xi}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}W(P_i^\star,\widehat{P}) &\leqslant \frac{5+\kappa}{1-\kappa}\inf_{P\in\mathscr{M}}\left[\frac{1}{n}\sum_{i=1}^{n}W(P_i^\star,P)\right]\\ &\quad+\frac{8}{(1-\kappa)\sqrt{n}}\left[\sqrt{2}+\sqrt{\frac{\xi}{2}}+\frac{\epsilon}{2\sqrt{n}}\right].\end{aligned}$$

We conclude by letting κ tend to 0.

10.11. **Proof of Corollary 6.** Let V be the linear space spanned by the D indicator functions $\mathbb{1}_I$ for $I\in\mathcal{I}$. Since for all $t=\sum_{I\in\mathcal{I}}t_I\mathbb{1}_I\in V$,

$$\|t\|_j^j=\sum_{I\in\mathcal{I}}|t_I|^jD^{-1}\geqslant D^{-1}\max_{I\in\mathcal{I}}|t_I|^j=D^{-1}\|t\|_\infty^j,$$

inequality (33) is satisfied with $R=D^{1/j}$. Moreover, given $\overline{p},q\in\mathcal{M}$ with $\overline{p}\neq q$, $(q-\overline{p})/\left\|q-\overline{p}\right\|_j$ writes as $\sum_{I\in\mathcal{I}}b_I 1\!\!1_I$ with

$$1=\left\|\sum_{I\in\mathcal{I}}b_I 1\!\!1_I\right\|_j=|b|_j\,D^{-1/j}\quad\text{and}\quad|b|_j=\left(\sum_{I\in\mathcal{I}}|b_I|^j\right)^{1/j}=R. \tag{125}$$

Hence,

$$\left(\frac{q-\overline{p}}{\left\|q-\overline{p}\right\|_j}\right)_+^{j-1}=\sum_{I\in\mathcal{I}}(b_I)_+^{j-1}1\!\!1_I\quad\text{and}\quad\left(\frac{q-\overline{p}}{\left\|q-\overline{p}\right\|_j}\right)_-^{j-1}=\sum_{I\in\mathcal{I}}(b_I)_-^{j-1}1\!\!1_I,$$

so that, by the definition (32) of $f_{(Q,P)}$

$$f_{(Q,P)}-\mathbb{E}\left(f_{(Q,P)}\right)=2\sum_{I\in\mathcal{I}}\left((b_I)_+^{j-1}-(b_I)_-^{j-1}\right)\left(1\!\!1_I-P^\star(I)\right).$$

Since by Corollary 3, $\phi_{(P,Q)}-\mathbb{E}\left[\phi_{(P,Q)}\right]=\left[f_{(Q,P)}-\mathbb{E}\left(f_{(Q,P)}\right)\right]/(2R^{j-1})$,

$$\begin{aligned}\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q})&=\sum_{i=1}^n\left[\phi_{(\overline{P},Q)}-\mathbb{E}\left(\phi_{(\overline{P},Q)}\right)\right]\\&=\frac{1}{R^{j-1}}\sum_{I\in\mathcal{I}}\left((b_I)_+^{j-1}-(b_I)_-^{j-1}\right)\sum_{i=1}^n\left[1\!\!1_I(X_i)-P^\star(I)\right]\\&\leqslant\frac{1}{R^{j-1}}\sum_{I\in\mathcal{I}}|b_I|^{j-1}\left|\sum_{i=1}^n\left[1\!\!1_I(X_i)-P^\star(I)\right]\right|.\end{aligned}\tag{126}$$

Using (125) and Hölder inequality with the conjugate exponents $j/(j-1)$ and j we obtain that

$$\begin{aligned}\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q})&\leqslant\left(\frac{|b|_j}{R}\right)^{j-1}\left[\sum_{I\in\mathcal{I}}\left|\sum_{i=1}^n\left[1\!\!1_I(X_i)-P^\star(I)\right]\right|^j\right]^{1/j}\\&=\left[\sum_{I\in\mathcal{I}}\left|\sum_{i=1}^n\left[1\!\!1_I(X_i)-P^\star(I)\right]\right|^j\right]^{1/j}.\end{aligned}$$

It follows from Jensen's inequality that for all $y\geqslant 0$

$$\mathbf{w}(\overline{\mathbf{P}},y)\leqslant\left[\sum_{I\in\mathcal{I}}\mathbb{E}\left|\sum_{i=1}^n\left[1\!\!1_I(X_i)-P^\star(I)\right]\right|^j\right]^{1/j}.$$

When $j>2$, we may use Theorem 15.10 [Page 442] in Boucheron *et al* (2013) with $Z=\sum_{i=1}^n 1\!\!1_I(X_i)$ and the fact that $1\!\!1_I(X_1),\ldots,1\!\!1_I(X_n)$ are independent nonnegative random variables bounded by 1. We obtain that for some constant $c_j>1$,

$$c_j^{-1}\mathbb{E}\left|\sum_{i=1}^n\left[1\!\!1_I(X_i)-P^\star(I)\right]\right|^j\leqslant 1+\left[nP^\star(I)\right]^{j/2}.$$

Using the inequality below that holds for all $j' \geqslant 1$

$$\|\overline{p}_D\|_{j'/2}^{j'/2} = \frac{1}{D}\sum_{i\in\mathcal{I}}\left(D\int_I p^\star d\mu\right)^{j'/2} = D^{j'/2-1}\sum_{i\in\mathcal{I}}[P^\star(I)]^{j'/2}$$

and the fact that $u \mapsto u^{1/j}$ is sub-additive, by summing (126) over $I \in \mathcal{I}$, we deduce that for all $y \geqslant 0$

$$\begin{aligned}\mathbf{w}(\overline{\mathbf{P}}, y) &\leqslant c_j^{1/j}\left[D + n^{j/2}\sum_{I\in\mathcal{I}}[P^\star(I)]^{j/2}\right]^{1/j} = c_j^{1/j}\left[D + n^{j/2}D^{1-j/2}\|\overline{p}_D\|_{j/2}^{j/2}\right]^{1/j} \\ &\leqslant c_j^{1/j}\left[D^{1/j} + D^{1/j-1/2}\sqrt{n\|\overline{p}_D\|_{j/2}}\right].\end{aligned}$$

Since $\overline{p}^\star$ is a density, μ a probability and $j > 2$, $1 = \|\overline{p}^\star\|_1 \leqslant \|\overline{p}_D\|_{j/2}$ and consequently for $D \leqslant n$

$$\mathbf{w}(\overline{\mathbf{P}}, y) \leqslant 2c_j^{1/j}D^{1/j-1/2}\sqrt{n\|\overline{p}_D\|_{j/2}}.$$

We may Theorem 1 with $a_0 = 3/(4R^{j-1})$, $a_1 = 1/(4R^{j-1})$, $R = D^{1/j}$, $c_0^{-1} = 6/(\kappa a_1) = 24D^{1-1/j}/\kappa$, and the quantity $v(\overline{\mathbf{P}})$ defined by (17) then satisfies

$$v(\overline{\mathbf{P}}) \leqslant \frac{2(c_j)^{1/j}}{c_0}D^{1/j-1/2}\|\overline{p}_D\|_{j/2}^{1/2} = \frac{48(c_j)^{1/j}}{\kappa}\sqrt{D\|\overline{p}_D\|_{j/2}}.$$

We obtain (41) for $j > 2$ by using the fact that $\overline{p}$ is an arbitrary element of a dense subset of $\overline{\mathcal{M}}_D$.

When $j \in (1, 2]$ we argue as follows. By Jensen's inequality,

$$\mathbb{E}\left|\sum_{i=1}^n[\mathbb{1}_I(X_i) - P^\star(I)]\right|^{2(j/2)} \leqslant [nP^\star(I)]^{j/2}$$

and arguing as before, we obtain that for all $y \geqslant 0$,

$$\mathbf{w}(\overline{\mathbf{P}}, y) \leqslant \sqrt{n}\left[\sum_{I\in\mathcal{I}}[P^\star(I)]^{j/2}\right]^{1/j} = D^{1/j-1/2}\sqrt{n\|\overline{p}_D\|_{j/2}}.$$

The remainder of the proof is then similar to that of the case $j > 2$ and leads to (41) for $j \in (1, 2]$.

10.12. **Proof of Corollary 8.** Let us fix an arbitrary density $\overline{p}$ in $\mathcal{M}$ and denote by $\overline{P}$ the corresponding probability in $\mathscr{M}$. It follows from Assumption 5 that the classes of subsets

$$\mathcal{C} = \{C_p = \{p < \overline{p}\},\ p \in \mathcal{M}\}$$

are VC with dimension not larger than $V = V(\overline{p})$. Applying Proposition 3.1 of Baraud (2016) (with $\sigma = 1$) we obtain that

$$\mathbb{E}\left[\sup_{C\in\mathcal{C}}\left|\sum_{i=1}^n(\mathbb{1}_C(X_i) - P_i^\star(C))\right|\right] \leqslant 10\sqrt{5nV}. \tag{127}$$

Using (45), for $P \in \mathscr{M}$ with density $p \in \mathcal{M}$, the quantity $\overline{\mathbf{Z}}\left(\boldsymbol{X}, \overline{\mathbf{P}}, \mathbf{P}\right)$ defined by (14) satisfies

$$\left|\overline{\mathbf{Z}}\left(\boldsymbol{X}, \overline{\mathbf{P}}, \mathbf{P}\right)\right| \leqslant \sup_{C \in \mathcal{C}} \left|\sum_{i=1}^{n} \left(\mathbb{1}_C(X_i) - P_i^{\star}(C)\right)\right|$$

and it follows from (16) and (127) that for all $y \geqslant 0$

$$\mathbf{w}(\overline{\mathbf{P}}, y) \leqslant \mathbb{E}\left[\sup_{\mathbf{P} \in \mathscr{M}} \left|\overline{\mathbf{Z}}\left(\boldsymbol{X}, \overline{\mathbf{P}}, \mathbf{P}\right)\right|\right] \leqslant 10\sqrt{5nV}.$$

Hence $v(\overline{\mathbf{P}})$ defined by (17) satisfies $v(\mathbf{P}_{(\overline{\boldsymbol{\theta}},q)}) = 10c_0^{-1}\sqrt{5V}$. We know from Corollary 7 that the family $\mathscr{T}(\ell, \mathscr{M})$ satisfies Assumption 1 with $a_0 = 3/2$ and $a_1 = 1/2$ and we may therefore apply our Theorem 1. Whatever $\kappa \in (0,1)$ and $\xi > 0$, any ℓ–estimator $\widehat{P} \in \mathscr{M}$ with a probability at least $1 - 0.37e^{-\xi}$,

$$\begin{aligned}\sum_{i=1}^{n} \left\|P_i^{\star} - \widehat{P}\right\| &< \frac{6}{1-\kappa} \sum_{i=1}^{n} \left\|P_i^{\star} - \overline{P}\right\| - \inf_{P \in \mathscr{M}} \sum_{i=1}^{n} \left\|P_i^{\star} - P\right\| \\ &\quad + \frac{4\sqrt{n}}{1-\kappa}\left[2 \times 10\sqrt{5V} + \sqrt{\frac{\xi}{2}} + \frac{\epsilon}{2\sqrt{n}}\right].\end{aligned}$$

The first part of the result follows by letting κ tend to 0 and by normalising by n.

For the second part, we argue as follows. Let $(\overline{P}_k)_{k\geqslant 1}$ be a sequence of probabilities in $\mathscr{M}$ such that $b_k = \sum_{i=1}^{n} \left\|P_i^{\star} - \overline{P}_k\right\|/n$ is non-increasing toward $\inf_{P \in \mathscr{M}} \sum_{i=1}^{n} \left\|P_i^{\star} - P\right\|/n$ and for $k \geqslant 1$, Ω_k the set

$$\frac{1}{n}\sum_{i=1}^{n} \left\|P_i^{\star} - \widehat{P}\right\| \leqslant 6b_k - \inf_{P \in \mathscr{M}} \frac{1}{n}\sum_{i=1}^{n} \left\|P_i^{\star} - P\right\| + 179\sqrt{\frac{V}{n}} + \sqrt{\frac{8\xi}{n}} + \frac{2\epsilon}{n}.$$

For all $k \geqslant 1$, $\mathbb{P}(\Omega_k) \geqslant 1 - 0.37e^{-\xi}$ and since $(b_k)_{k\geqslant 1}$ is non-increasing, $\Omega_{k+1} \subset \Omega_k$. Besides, $\Omega = \bigcap_{k\geqslant 1} \Omega_k$ is the set on which (48) is satisfied. We conclude using the fact that

$$\mathbb{P}(\Omega) = \lim_{k \to +\infty} \mathbb{P}(\Omega_k) \geqslant 1 - 0.37e^{-\xi}.$$

10.13. **Proof of Lemma 1.** Let us denote by P_m the Gaussian distribution with mean $m \in \mathbb{R}^d$ and identity covariance matrix and by p_m its density with respect to the Lebesgue measure. Since the Lebesgue measure is translation invariant, $\|P_m - P_{m'}\| = \|P_{m-m'} - P_0\|$ for all $m, m' \in \mathbb{R}^d$ and it suffices to prove the result for $m' = 0$. Let $m \in \mathbb{R}^d$. Since the results clearly hold for $m = 0$, with no loss of generality we may assume that with $m \neq 0$ what we shall do hereafter. As we have seen in Section 5.1,

$$\begin{aligned}
&\|P_m - P_0\| \\
&= P_m\left[\{x\in\mathbb{R}^d,\ p_m(x) > p_0(x)\}\right] - P_0\left[\{x\in\mathbb{R}^d,\ p_m(x) > p_0(x)\}\right] \\
&= P_0\left[\{x\in\mathbb{R}^d,\ p_m(x+m) > p_0(x+m)\}\right] - P_0\left[\{x\in\mathbb{R}^d,\ p_m(x) > p_0(x)\}\right] \\
&= P_0\left[\{x\in\mathbb{R}^d,\ |x|^2 < |x+m|^2\}\right] - P_0\left[\{x\in\mathbb{R}^d,\ |x-m|^2 > |x|^2\}\right] \\
&= P_0\left[\left\{x\in\mathbb{R}^d,\ \left\langle x, \frac{m}{|m|}\right\rangle > -\frac{|m|}{2}\right\}\right] - P_0\left[\left\{x\in\mathbb{R}^d,\ \left\langle x, \frac{m}{|m|}\right\rangle < \frac{|m|}{2}\right\}\right].
\end{aligned}$$

The first conclusion follows from the fact that the image of the probability P_0 by the mapping $x \mapsto \langle x, m\rangle / |m|$ is a standard real-valued Gaussian random variable.

For the second part, we argue as follows. Clearly, $\|P_m - P_0\| \leqslant 1$ and for all $m \in \mathbb{R}^d$, p_m is bounded by $1/\sqrt{2\pi}$. Consequently

$$\|P_m - P_0\| = 2\mathbb{P}\left[0 \leqslant Z \leqslant \frac{|m|}{2}\right] \leqslant 2\int_0^{|m|/2} p_m(x)dx \leqslant \frac{|m|}{\sqrt{2\pi}}$$

which leads to the right-hand side of (52). For the left-hand side, we check that the minimum of the mapping $m \to \mathbb{P}\left[|Z| \leqslant |m|/2\right] / \min\{1, |m|/\sqrt{2\pi}\}$ is reached for $|m| = \sqrt{2\pi}$ and is not smaller than 0.78.

10.14. **Proof of Lemma 2.** Since the TV-distance is translation invariant and q is symmetric, i.e.

$$\|P_\theta - P_{\theta'}\| = \left\|P_0 - P_{|\theta'-\theta|}\right\|$$

and it suffices thus to prove the result for $\theta' = 0$ and $\theta > 0$.

Note that for all $x \in \mathbb{R}$,

$$q(x) \wedge q_\theta(x) = \begin{cases} q_\theta(x) & \text{if } x \leqslant \theta/2 \\ q(x) & \text{if } x > \theta/2. \end{cases}$$

Consequently, using again that q is symmetric,

$$\begin{aligned}
\int_{\mathbb{R}} q(x) \wedge q_\theta(x)dx &= \int_{-\infty}^{\theta/2} q(x-\theta)dx + \int_{\theta/2}^{+\infty} q(x)dx \\
&= 2\int_{\theta/2}^{+\infty} q(x)dx = 1 - \frac{2}{\pi}\arctan(\theta/2)
\end{aligned}$$

and

$$\|P_0 - P_\theta\| = 1 - \int_{\mathbb{R}} q(x) \wedge q_\theta(x)dx = \frac{2}{\pi}\arctan(\theta/2).$$

This equality leads to (63). We obtain the right-hand side of 63 by using the fact that $\arctan(u) \leqslant u \wedge (\pi/2)$ for all $u \geqslant 0$. Since $u \mapsto \arctan(u)$ is concave on $[0, \pi/2]$ and increasing on $[\pi/2, +\infty)$,

$$\frac{\arctan(u)}{u} \geqslant \frac{\arctan(\pi/2)}{\pi/2} \quad \text{for all } u \in (0, \pi/2]$$

and

$$\arctan(u) \geqslant \arctan(\pi/2) \quad \text{for all } u \in [\pi/2, +\infty].$$

These inequalities lead to

$$\frac{2}{\pi}\arctan(u) \geqslant \left(\frac{2u}{\pi} \wedge 1\right)\frac{\arctan(\pi/2)}{\pi/2} > 0.639\left(\frac{2u}{\pi} \wedge 1\right) \quad \text{for all } u \geqslant 0$$

which implies (63).

10.15. **Proof of Corollary 4.** As a subset of $\mathscr{L}_2(E,\mu)$, V is also separable and admits an (at most countable) Hilbert basis $(\varphi_I)_{I\in\mathcal{I}}$. As a consequence of (36), for any non-decreasing sequence (for the inclusion) of subsets $(\mathcal{I}_k)_{k\geqslant 1}$ of $\mathcal{I}$ satisfying $\bigcup_{k\geqslant 1}\mathcal{I}_k = \mathcal{I}$,

$$\left\|\sum_{I\in\mathcal{I}_k}\varphi_I^2\right\|_\infty = \sup_{x\in E}\sup_{\sum_{I\in\mathcal{I}_k}c_I^2=1}\left|\sum_{I\in\mathcal{I}_k}c_I\varphi_I(x)\right|^2 \leqslant \sup_{t\in V,\ \|t\|_2=1}\|t\|_\infty^2 \leqslant R^2.$$

By letting $\mathcal{I}_k$ grow toward $\mathcal{I}$ we deduce that the above inequality remains true for $\mathcal{I}$ in place of $\mathcal{I}_k$ and for all $t \in V$, the sequence

$$t_k = \sum_{\lambda\in\mathcal{I}_k}\langle t,\varphi_I\rangle\varphi_I \quad \text{with} \quad \langle t,\varphi_I\rangle = \int_E t\varphi_I d\mu$$

converges both uniformly on E and in $\mathbb{L}_2(E,\mu)$ toward $t \in V$. The equality $t = \sum_{I\in\mathcal{I}}\langle t,\varphi_I\rangle\phi_I$ therefore holds pointwise and in $\mathbb{L}_2(E,\mu)$ and we may write that

$$\left\|\sum_{I\in\mathcal{I}}\varphi_I^2\right\|_\infty = \sup_{x\in E}\sup_{\sum_I c_I^2=1}\left|\sum_{I\in\mathcal{I}}c_I\varphi_I(x)\right|^2 = \sup_{t\in V,\ \|t\|_2=1}\|t\|_\infty^2 \leqslant R^2. \tag{128}$$

It also follows from (36) that for all $p,q \in \mathcal{M}$,

$$\|p-q\|_\infty \leqslant R\,\|p-q\|_2 \tag{129}$$

and, as seen in Section 4.3 (Proposition 3), the test between $P = p\cdot\mu$ and $Q = q\cdot\mu$, $P \neq Q$, given by

$$\phi_{(P,Q)} = \frac{(2q-\|q\|_2^2)-(2p-\|p\|_2^2)}{4R\,\|p-q\|} \tag{130}$$

satisfies then our Assumption 1 for the loss ℓ_2 with $a_0 = 3/(4R)$ and $a_1 = 1/(4R)$. Given two distinct elements $Q = q\cdot\mu$ and $\overline{P} = \overline{p}\cdot\mu$ in $\mathscr{M}$, we may write

$$q-\overline{p} = \sum_{I\in\mathcal{I}}c_I\varphi_I \quad \text{with} \quad \sum_{I\in\mathcal{I}}c_I^2 = \|\overline{p}-q\|_2^2 > 0$$

and since by Cauchy-Schwarz inequality and (128)

$$\mathbb{E}\left[\sum_{I\in\mathcal{I}}|c_I|\,|\varphi_I(X_i)|\right] \leqslant R\,\|\overline{p}-q\|_2 < +\infty$$

it follows from Fubini's Theorem and Cauchy-Schwarz inequality that

$$\begin{aligned}\left|\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q})\right| &= \left|\sum_{i=1}^{n}\left[\phi_{(\overline{P},Q)}(X_i)-\mathbb{E}\left(\phi_{(\overline{P},Q)}(X_i)\right)\right]\right|\\ &= \frac{1}{2R\left\|\overline{p}-q\right\|_2}\left|\sum_{I\in\mathcal{I}}c_I\left[\sum_{i=1}^{n}\left(\varphi_I(X_i)-\mathbb{E}\left[\varphi_I(X_i)\right]\right)\right]\right|\\ &\leqslant \frac{1}{2R}\left(\sum_{I\in\mathcal{I}}\left[\sum_{i=1}^{n}\left(\varphi_I(X_i)-\mathbb{E}\left[\varphi_I(X_i)\right]\right)\right]^2\right)^{1/2}. \qquad (131)\end{aligned}$$

Since $\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\overline{\mathbf{P}})=0$, note that this inequality is also true for $Q=\overline{P}$. Taking the supremum over all $Q\in\mathscr{M}$, the expectation on both sides of (131) and using Jensen inequality together with (128) again, we obtain that $\mathbf{w}(\overline{\mathbf{P}},\cdot)$ defined by (16) satisfies for all $y\geqslant 0$

$$\begin{aligned}\mathbf{w}(\overline{\mathbf{P}},y) &= \mathbb{E}\left[\sup_{\mathbf{Q}\in\mathscr{B}(\mathbf{P}^\star,y)}\left|\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{Q})\right|\right]\\ &\leqslant \frac{1}{2R}\mathbb{E}\left[\left(\sum_{I\in\mathcal{I}}\left[\sum_{i=1}^{n}\left(\varphi_I(X_i)-\mathbb{E}\left[\varphi_I(X_i)\right]\right)\right]^2\right)^{1/2}\right]\\ &\leqslant \frac{1}{2R}\left[\sum_{I\in\mathcal{I}}\mathbb{E}\left[\sum_{i=1}^{n}\left(\varphi_I(X_i)-\mathbb{E}\left[\varphi_I(X_i)\right]\right)^2\right]\right]^{1/2}\\ &= \frac{1}{2R}\left[\sum_{I\in\mathcal{I}}\sum_{i=1}^{n}\mathrm{Var}[\varphi_I(X_i)]\right]^{1/2} \leqslant \frac{1}{2R}\left(\sum_{I\in\mathcal{I}}\sum_{i=1}^{n}\int_E\varphi_I^2p_i^\star d\mu\right)^{1/2}\\ &\leqslant \frac{1}{2R}\left(\sum_{i=1}^{n}\left\|\sum_{I\in\mathcal{I}}\varphi_I^2\right\|_\infty\right)^{1/2} \leqslant \frac{\sqrt{n}}{2}.\end{aligned}$$

This means that $v(\overline{\mathbf{P}})$ defined by (17) satisfies

$$v(\overline{\mathbf{P}})=\max\left\{\frac{1}{2c_0},\frac{1}{c_0\sqrt{2}}\right\}=\frac{1}{c_0\sqrt{2}}$$

and by applying our Theorem 1 with a density $\overline{p}\in\mathcal{M}$ we obtain that for all $\kappa\in(0,1)$, any ℓ–estimator $\widehat{P}=\widehat{p}\cdot\mu$ satisfies with probability at least $1-0.37e^{-\xi}$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|p_i^\star-\widehat{p}\|_2 &\leqslant \frac{6}{(1-\kappa)n}\sum_{i=1}^{n}\|p_i^\star-\overline{p}\|_2-\inf_{p\in\mathcal{M}}\frac{1}{n}\sum_{i=1}^{n}\|p_i^\star-p\|_2\\ &\quad+\left(\sqrt{2}+\sqrt{\xi/2}\right)\frac{8R}{(1-\kappa)\sqrt{n}}+\frac{4R\epsilon}{(1-\kappa)n}.\end{aligned}$$

The result follows by letting κ tend to 0 and using the fact that $\overline{p}$ is arbitrary in $\mathcal{M}$.

10.16. **Proof of Corollary 9.** Throughout this section we shall identify a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ with the function on $\mathscr{X} = \{1, \ldots, n\} \times E$ defined by $(k, x) \mapsto \theta_k$ and for conveniency we shall denote by $\boldsymbol{\theta}$ both the vector and the corresponding function. We consider the class $\mathcal{F}$ of functions on $\mathscr{X}$ which are of the form $q_{\boldsymbol{\theta}} : (k, x) \mapsto q(x - \boldsymbol{\theta}(k, x)) = q(x - \theta_k)$. The linear space Θ (viewed as a space of functions on $\mathscr{X}$) is VC-subgraph with dimension not larger than $d+1$, so is the class of functions of the form $(k, x) \mapsto x - \boldsymbol{\theta}(x, k)$ by applying Proposition 42-(i) of Baraud *et al.* (2017) with $g : (k, x) \mapsto x$. Since q is unimodal it follows from Proposition 42-(vi) of Baraud *et al.* (2017) that $\mathcal{F}$ is VC-subgraph with dimension not larger than $9.41(d+1)$. Let us fix $\overline{\boldsymbol{\theta}} \in \Theta$. Using Proposition 42-(i) again, we obtain that the class $\left\{q_{\boldsymbol{\theta}} - q_{\overline{\boldsymbol{\theta}}},\ \boldsymbol{\theta} \in \Theta\right\}$ is VC-subgraph with dimension not larger than $9.41(d + 1)$ and the VC-dimension of the class (of subsets of $\mathscr{X}$)

$$\mathcal{C} = \left\{C_{\boldsymbol{\theta}} = \{q_{\boldsymbol{\theta}} - q_{\overline{\boldsymbol{\theta}}} < 0\},\ \boldsymbol{\theta} \in \Theta\right\}$$

as well. Applying Proposition 3.1 of Baraud (2016) (with $\sigma = 1$ and V in place of d) we obtain that whatever the independent random variables $Y_1, \ldots, Y_n$ with values in $\mathscr{X}$ and distributions $\widetilde{P}_1, \ldots, \widetilde{P}_n$ respectively,

$$\mathbb{E}\left[\sup_{C \in \mathcal{C}} \left|\sum_{i=1}^{n} \left(\mathbb{1}_C(Y_i) - \widetilde{P}_i(C)\right)\right|\right] \leqslant 10\sqrt{5nV} < 69\sqrt{n(d+1)}. \tag{132}$$

Using (45) and for $i \in \{1, \ldots, n\}$ denoting $\widetilde{P}_i^\star$ the probability on the random variable $Y_i = (i, X_i)$, i.e. $\widetilde{P}_i^\star = \delta_i \otimes P_i^\star$ (where δ_k denotes the Dirac mass at $k \in \{1, \ldots, n\}$), the quantity $\overline{\mathbf{Z}}\left(\boldsymbol{X}, \mathbf{P}_{(\overline{\boldsymbol{\theta}},q)}, \mathbf{P}_{(\boldsymbol{\theta},q)}\right)$ defined by (14) satisfies for all $\boldsymbol{\theta} \in \Theta$

$$\left|\overline{\mathbf{Z}}\left(\boldsymbol{X}, \mathbf{P}_{(\overline{\boldsymbol{\theta}},q)}, \mathbf{P}_{(\boldsymbol{\theta},q)}\right)\right| \leqslant \sup_{C \in \mathcal{C}} \left|\sum_{i=1}^{n} \left(\mathbb{1}_C(Y_i) - \widetilde{P}_i^\star(C)\right)\right|$$

and it follows from (16) and (132) that for all $y \geqslant 0$

$$\mathbf{w}(\mathbf{P}_{(\overline{\boldsymbol{\theta}},q)}, y) \leqslant \mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \Theta} \left|\overline{\mathbf{Z}}\left(\boldsymbol{X}, \mathbf{P}_{(\overline{\boldsymbol{\theta}},q)}, \mathbf{P}_{(\boldsymbol{\theta},q)}\right)\right|\right] < 69\sqrt{n(d+1)}.$$

Hence $v(\mathbf{P}_{(\overline{\boldsymbol{\theta}},q)})$ defined by (17) satisfies $v(\mathbf{P}_{(\overline{\boldsymbol{\theta}},q)}) = 69c_0^{-1}\sqrt{d+1}$. We know from Corollary 7 that the family $\mathscr{T}(\ell, \mathscr{M})$ satisfies Assumption 1 with $a_0 = 3/2$ and $a_1 = 1/2$ and we may therefore apply our Theorem 1. Whatever $\kappa \in (0, 1)$ and $\xi > 0$, any ℓ–estimator $\mathbf{P}_{(\widehat{\boldsymbol{\theta}},q)} \in \mathscr{M}$ with a probability at least $1 - 0.37e^{-\xi}$,

$$\begin{aligned}\sum_{i=1}^{n} \left\|P_i^\star - P_{(\widehat{\theta}_i,q)}\right\| &< \frac{6}{1-\kappa} \sum_{i=1}^{n} \left\|P_i^\star - P_{(\overline{\theta}_i,q)}\right\| - \inf_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} \left\|P_i^\star - P_{(\theta_i,q)}\right\| \\ &\quad + \frac{4\sqrt{n}}{1-\kappa}\left[2 \times 69\sqrt{d+1} + \sqrt{\frac{\xi}{2}} + \frac{\epsilon}{2\sqrt{n}}\right].\end{aligned}$$

Using the fact that $\overline{\boldsymbol{\theta}}$ is arbitrary in Θ, the result follows by letting κ tend to 0 and by normalising by n.

10.17. **Proof of Corollary 11.** We know from Corollary 7 that the family $\mathscr{T}(\ell,\mathscr{M})$ satisfies Assumption 1 with $a_0=3/2$ and $a_1=1/2$. Besides, since Condition (65) holds true, we may apply Proposition 9 and deduce that Assumption 2 is satisfied with $a_2=1+a_2'$. As a consequence Theorem 2 applies and it remains to bound from above the function D uniformly over $\boldsymbol{\mathscr{M}}=\{P^{\otimes n},\ P\in\mathscr{M}\}$.

Let $y\geqslant 0$ and denote by $\mathscr{M}(y)$ the subset of $\mathscr{M}$ gathering those probability P that satisfy $\|P-P^{\star}\|\leqslant y/n$, or equivalently for which $\mathbf{P}=P^{\otimes n}$ belongs to the set $\boldsymbol{\mathscr{B}}(\mathbf{P}^{\star},y)$ defined by (15) (here $\boldsymbol{\ell}(\mathbf{P}^{\star},\mathbf{P})=n\,\|P^{\star}-P\|$ since the data are assumed to be i.i.d. with distribution $P^{\star}$). We fix some probability $\overline{P}\in\mathscr{M}$ with density $\overline{p}$ and define

$$B(V,y)=\sup_{\mathcal{C}}\mathbb{E}\left[\left|\sum_{i=1}^{n}\left(\mathbb{1}_C(X_i)-P^{\star}(C)\right)\right|\right]$$

where the supremum runs among all the classes $\mathcal{C}$ of subsets of E which are VC with dimension not larger than V and satisfy

$$\sup_{C\in\mathcal{C}}P^{\star}(C)\leqslant\left[a_2\left(\|P^{\star}-\overline{P}\|+\frac{y}{n}\right)\right]\wedge 1=\sigma^2.$$

The set $\mathscr{M}(y)$ is the union of

$$\mathscr{M}_0(y)=\{P\in\mathscr{M}(y),\ P^{\star}(\overline{p}>p)\leqslant 1/2\},$$
$$\mathscr{M}_1(y)=\{P\in\mathscr{M}(y),\ P^{\star}(p\geqslant\overline{p})\leqslant 1/2\}.$$

Besides, by Assumption 5, the classes of sets $\{\overline{p}>p\},\ p\in\mathcal{M}\}$ and the class of their complementaries $\{\{p\leqslant\overline{p}\},\ p\in\mathcal{M}\}$ are both VC with dimension not larger than V.

For $P\in\mathscr{M}_0(y)$, using (45) we may write $\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{P})$ defined by (14) as

$$\begin{aligned}\left|\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{P})\right|&=\left|\sum_{i=1}^{n}\left[\phi_{(\overline{P},P)}(X_i)-\mathbb{E}\left(\phi_{(\overline{P},P)}\right)\right]\right|\\&=\left|\sum_{i=1}^{n}\left[\mathbb{1}_{\overline{p}>p}(X_i)-P^{\star}(\overline{p}>p)\right]\right|\end{aligned}$$

Since Condition (65) is satisfied we deduce from Proposition 9, more precisely (66), and the definition of $\mathscr{M}_0(y)$ and $\mathscr{M}(y)$ that for all $P\in\mathscr{M}_0(y)$

$$P^{\star}(\overline{p}>p)=P^{\star}(\overline{p}>p)\wedge P^{\star}(\overline{p}\leqslant p)\leqslant a_2\left[\left\|P^{\star}-\overline{P}\right\|+y/n\right]$$

and consequently,

$$\mathbb{E}\left[\sup_{P\in\mathscr{M}_0(y)}\left|\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{P})\right|\right]\leqslant B(V,y).$$

Arguing similarly for $P\in\mathscr{M}_1(y)$, we obtain that

$$\left|\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{P})\right| = \left|\sum_{i=1}^{n}\left[1\!\!1_{\overline{p}\leqslant p}(X_i)-P^\star(\overline{p}\leqslant p)\right]\right|$$

and

$$P^\star(\overline{p}\leqslant p)=P^\star(\overline{p}\leqslant p)\wedge P^\star(\overline{p}>p)\leqslant a_2\left[\left\|P^\star-\overline{P}\right\|+y/n\right],$$

leading to

$$\mathbb{E}\left[\sup_{P\in\mathscr{M}_1(y)}\left|\overline{\mathbf{Z}}(\boldsymbol{X},\overline{\mathbf{P}},\mathbf{P})\right|\right]\leqslant B(V,y).$$

The quantity $\mathbf{w}(\overline{\mathbf{P}},y)$ defined by (16) therefore satisfies $\mathbf{w}(\overline{\mathbf{P}},y)\leqslant 2B(V,y)$. Let us know bound from above $B(V,y)$ and to do so introduce

$$\overline{\Gamma}_n(V)=\log\left[2\sum_{j=0}^{V\wedge n}\binom{n}{j}\right]\leqslant (V\wedge n)L\quad\text{with}\quad L=\log\left(\frac{2en}{V\wedge n}\right) \tag{133}$$

$a=\left(32\sqrt{\overline{\Gamma}_n(V)/n}\right)\wedge 1\leqslant\left(32\sqrt{(V\wedge n)L/n}\right)\wedge 1$ and

$$\overline{H}(x)=x\sqrt{V\left(5+\log\left(\frac{1}{x}\right)\right)}\quad\text{for all } x\in(0,1].$$

For a fixed $\kappa\in(0,1)$, we consider a value of y that satisfies

$$\frac{y}{n}\geqslant\|P^\star-\overline{P}\|+\frac{2788a_2}{c_1^2}\frac{VL}{n}\quad\text{with}\quad c_1=\frac{2\kappa^2a_1^2}{225a_2}=\frac{\kappa^2}{450a_2}. \tag{134}$$

By using (133), the definition of a and the facts that $L>1$, $a_2\sqrt{2788}/c_1\geqslant 32>1/\sqrt{2e}$ we deduce from (134) that

$$\begin{aligned}\sigma=\sqrt{a_2\left(\|P^\star-\overline{P}\|+\frac{y}{n}\right)}\wedge 1&\geqslant\left[\frac{a_2\sqrt{2788}}{c_1}\sqrt{\frac{(V\wedge n)L}{n}}\right]\wedge 1\\ &\geqslant\sqrt{\frac{V\wedge n}{2en}}\vee a.\end{aligned} \tag{135}$$

Besides, it follows from (134) again that $\sigma\leqslant\sqrt{2a_2y/n}$. Applying Proposition 4 in Baraud (2016) (with V in place of d) and using (135) with the fact that $L\geqslant 1+\log 2$, we obtain that

$$\begin{aligned}\mathbf{w}(\overline{\mathbf{P}},y)\leqslant 2B(V,y)\leqslant 20\sqrt{n}\,\overline{H}(\sigma\vee a)&=20n\sigma\sqrt{\frac{V}{n}\left(5+\log\left(\frac{1}{\sigma}\right)\right)}\\ &\leqslant 20n\sigma\sqrt{\frac{V}{n}\left(5+\frac{L}{2}\right)}\leqslant 20n\sigma\sqrt{\left(\frac{5}{1+\log 2}+\frac{1}{2}\right)\frac{VL}{n}}\\ &\leqslant 20n\sqrt{\frac{2a_2y}{n}}\times\sqrt{\frac{3.46c_1^2y}{2788a_2n}}\leqslant c_1y.\end{aligned}$$

By definition (23) of $D(\overline{\mathbf{P}})$ we deduce that

$$\frac{D(\overline{\mathbf{P}})}{n} \leqslant \|P^\star - \overline{P}\| + \frac{2788a_2}{c_1^2}\frac{VL}{n} = \|P^\star - \overline{P}\| + \frac{c_3 a_2^3}{\kappa^4}\frac{VL}{n}$$

with $c_3 = 2788 \times (450)^2$ and by applying Theorem 2 we get that for all $\xi > 0$, with a probability at least $1 - 0.42e^{-\xi}$, the ℓ–estimator $\widehat{\mathbf{P}} = \widehat{P}^{\otimes n}$ satisfies

$$\begin{aligned}\|P^\star - \widehat{P}\| &\leqslant \frac{4}{1-\kappa}\left[\left(\frac{3}{2} + \frac{\kappa}{15}\right)\|P^\star - \overline{P}\| + \frac{\kappa D(\overline{\mathbf{P}})}{6n} + 2\left(1 + \frac{15a_2}{\kappa}\right)\frac{\xi}{n} + \frac{\epsilon}{2n}\right] \\ &\quad - \inf_{P \in \mathscr{M}} \|P^\star - P\| \\ &\leqslant \frac{4}{1-\kappa}\left[\left(\frac{3}{2} + \frac{7\kappa}{30}\right)\|P^\star - \overline{P}\| + \frac{c_3 a_2^3}{6\kappa^3}\frac{VL}{n} + 2\left(1 + \frac{15a_2}{\kappa}\right)\frac{\xi}{n} + \frac{\epsilon}{2n}\right] \\ &\quad - \inf_{P \in \mathscr{M}} \|P^\star - P\|.\end{aligned}$$

For $\kappa = 0.126$, the right-hand side is not larger than

$$7\|P^\star - \overline{P}\| - \inf_{P \in \mathscr{M}} \|P^\star - P\| + ca_2^3\frac{VL}{n} + 9.16\,(1 + 120a_2)\,\frac{\xi}{n} + 2.3\frac{\varepsilon}{n}$$

with $c = 3 \times 10^{11} > 4c_3/[6(1-\kappa)\kappa^3]$ and the result follows by using the fact that $\overline{P}$ is arbitrary in $\mathscr{M}$.

References

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401.

Baraud, Y. (2016). Bounding the expectation of the supremum of an empirical process over a (weak) vc-major class. *Electron. J. Statist.*, 10(2):1709–1728.

Baraud, Y. and Birgé, L. (2016). Rho-estimators for shape restricted density estimation. *Stochastic Process. Appl.*, 126(12):3888–3912.

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

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. (1987). On the risk of histograms for estimating decreasing densities. *Ann. Statist.*, 15(3):1013–1022.

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

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford.

Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York.

Dudley, R. M. (1984). A course on empirical processes. In *École d'été de Probabilités de Saint-Flour, XII—1982*, volume 1097 of *Lecture Notes in Math.*, pages 1–142. Springer, Berlin.

Gao, C., Liu, J., Yao, Y., and Zhu, W. (2018). Robust estimation and generative adversarial nets. Technical report, arXiv:1810.02030.

Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656.

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

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag, New York.

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.

Meyer, Y. (1992). *Wavelets and operators*, volume 37 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. Translated from the 1990 French original by D. H. Salinger.

Reynaud-Bouret, P. and Rivoirard, V. (2010). Near optimal thresholding estimation of a Poisson intensity on the real line. *Electron. J. Stat.*, 4:172–238.

Reynaud-Bouret, P., Rivoirard, V., and Tuleau-Malot, C. (2011). Adaptive density estimation: a curse of support? *J. Statist. Plann. Inference*, 141(1):115–139.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Villani, C. (2009). *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Old and new.

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.*, 13(2):768–774.

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`