RHO-ESTIMATORS REVISITED: GENERAL THEORY AND APPLICATIONS

Yannick Baraud, Lucien Birgé

RHO-ESTIMATORS REVISITED: GENERAL THEORY AND APPLICATIONS

Y. BARAUD AND L. BIRGÉ

Abstract. Following Baraud, Birgé and Sart (2016), we pursue our attempt to design a universal and robust estimation method based on independent (but not necessarily i.i.d.) observations. Given such observations with an unknown joint distribution $\mathbf{P}$ and a dominated model $\mathscr{Q}$ for $\mathbf{P}$, we build an estimator $\widehat{\mathbf{P}}$ based on $\mathscr{Q}$ and measure its risk by an Hellinger-type distance. When $\mathbf{P}$ does belong to the model, this risk is bounded by some new notion of dimension which relies on the local complexity of the model in a vicinity of $\mathbf{P}$. In most situations this bound corresponds to the minimax risk over the model (up to a possible logarithmic factor). When $\mathbf{P}$ does not belong to the model, its risk involves an additional bias term proportional to the distance between $\mathbf{P}$ and $\mathscr{Q}$, whatever the true distribution $\mathbf{P}$. From this point of view, this new version of ρ-estimators improves upon the previous one described in Baraud, Birgé and Sart (2016) which required that $\mathbf{P}$ be absolutely continuous with respect to some known reference measure. Further additional improvements have been brought as compared to the former construction. In particular, it provides a very general treatment of the regression framework with random design as well as a computationally tractable procedure for aggregating estimators. Finally, we consider the situation where the Statistician has at disposal many different models and we build a penalized version of the ρ-estimator for model selection and adaptation purposes. In the regression setting, this penalized estimator not only allows to estimate the regression function but also the distribution of the errors.

1. Introduction

In a previous paper, namely Baraud, Birgé and Sart (2016), we introduced a new class of estimators that we called ρ-estimators for estimating the distribution $\mathbf{P}$ of a random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ with values in some measurable space $(\mathscr{X}, \mathscr{B})$ under the assumption that the X_i are independent but not necessarily i.i.d. These estimators are based on density models, a *density model* being a family of densities $\mathbf{t}$ with respect to some reference measure $\boldsymbol{\mu}$ on $\mathscr{X}$. We also assumed that $\mathbf{P}$ was absolutely continuous with respect to $\boldsymbol{\mu}$ with density $\mathbf{s}$ and we measured the performance of an estimator $\widehat{\mathbf{s}}$ in terms of $\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})$, where $\mathbf{h}$ is a Hellinger-type distance to be defined later. Originally, the motivations for this construction were to design an estimator $\widehat{\mathbf{s}}$ of $\mathbf{s}$ with the following properties.

— Given a density model $\mathbf{S}$, the estimator $\widehat{\mathbf{s}}$ should be nearly optimal over $\mathbf{S}$ from the minimax point of view, which means that it is possible to bound the risk of the estimator $\widehat{\mathbf{s}}$ over $\mathbf{S}$ from above by some quantity $CD(\mathbf{S})$ which is approximately of the order of the minimax risk over $\mathbf{S}$.

— Since in Statistics we typically have uncomplete information about the true distribution of the observation, when we assume that $\mathbf{s}$ belongs to $\mathbf{S}$ nothing ever warrants that

Date: June 15, 2016.

this is true. We may more reasonably expect that $\mathbf{s}$ is close to $\mathbf{S}$ which means that the model $\mathbf{S}$ is not exact but only approximate and that the quantity $\mathbf{h}(\mathbf{s},\mathbf{S}) = \inf_{\mathbf{t}\in\mathbf{S}} \mathbf{h}(\mathbf{s},\mathbf{t})$ might therefore be positive. In this case we would like that the risk of $\widehat{\mathbf{s}}$ be bounded by $C'\left[D(\mathbf{S}) + \mathbf{h}^2(\mathbf{s},\mathbf{S})\right]$ for some universal constant C'.

In the case of ρ-estimators, the previous bound can actually be slightly refined and expressed in the following way. It is possible to define on $\mathbf{S}$ a positive function R such that the risk of the ρ-estimator is not larger than $R(\mathbf{s})$, with $R(\mathbf{s}) \leq CD(\mathbf{S})$ if $\mathbf{s}$ belongs to the model $\mathbf{S}$ and not larger than $C' \inf_{\overline{\mathbf{s}}\in\mathbf{S}} \left[R(\overline{\mathbf{s}}) + \mathbf{h}(\mathbf{s},\overline{\mathbf{s}})\right]$ when $\mathbf{s}$ does not belong to $\mathbf{S}$.

The weak sensibility of this risk bound to small deviations with respect to the Hellinger-type distance $\mathbf{h}$ between $\mathbf{s}$ and an element $\overline{\mathbf{s}}$ of $\mathbf{S}$ covers some classical notions of robustness among which robustness to a possible contamination of the data and robustness to outliers. The first situation arises when the data $X_1,\ldots,X_n$ are assumed to be i.i.d. with density $\overline{s}$ belonging to some density model S while they are actually drawn from a density of the form $s = (1-\varepsilon)\overline{s} + \varepsilon u$ for some arbitrary density $u \neq \overline{s}$ and $\varepsilon > 0$. A portion ε of the original data are therefore contaminated by a sample of density u. In contrast, the presence of outliers can be modeled in the following way. The observation $\boldsymbol{X} = (X_1,\ldots,X_n)$ is assumed to have a product density of the form $\overline{s}^{\otimes n}$ with $\overline{s} \in S$ while the true density is actually of the form $\mathbf{s} = \bigotimes_{i=1}^n t_i$ with $t_i = \overline{s}$ when $i \notin J \subset \{1,\ldots,n\}$ and $t_i = \delta_i$ when $i \in J$, where δ_i is the density of a probability which is very concentrated around some arbitrary point x_i in $\mathscr{X}$ (weakly close to the Dirac measure at x_i). In this case, the random variables $X_1,\ldots,X_n$ are still independent but no longer i.i.d. We can prove that in each of these two cases the densities $\overline{\mathbf{s}} = \overline{s}^{\otimes n}$ and $\mathbf{s}$ are close with respect to $\mathbf{h}$, at least when ε and $|J|$ are small enough. The risk bound of the ρ-estimator at the true density $\mathbf{s}$ will therefore be close to that obtained under the presumed density $\overline{\mathbf{s}}$.

This is, of course, a simplified presentation of the main ideas underlying our work in Baraud, Birgé and Sart (2016). We proved there various non-asymptotic risk bounds for ρ-estimators. We also provided various applications to density and regression estimation. There are nevertheless some limitations to the properties of ρ-estimators as defined there.

i) The study of random design regression required that either the distribution of the design be known or that the errors have a symmetric distribution. We want to relax these assumptions and consider the random design regression framework with greater generality.

ii) We worked with some reference measure $\boldsymbol{\mu}$ and assumed that all the probabilities we considered, including the true distribution $\mathbf{P}$ of $\boldsymbol{X}$, were absolutely continuous with respect to $\boldsymbol{\mu}$. This is quite natural for the probabilities that belong to our models since the models are, by assumption, dominated and typically defined via a reference measure and a family of densities with respect to this measure. Nevertheless, the assumption that the true distribution $\mathbf{P}$ of the observations be also dominated by $\boldsymbol{\mu}$ is questionable. We therefore would like to get rid of it and let the true distribution be completely arbitrary, relaxing thus the assumption that the density $\mathbf{s}$ exists. Unexpectedly, such an extension leads to subtle complications as we shall see below and this generalization is actually far from being straightforward.

We also want here to design a method based on "probability models" rather than "density models", which means working with dominated models $\mathcal{P}$ consisting of probabilities rather than of densities as for $\mathbf{S}$. Of course, the choice of a dominating measure $\boldsymbol{\mu}$ and a specific set $\mathbf{S}$ of densities leads to a probability model $\mathcal{P}$. This is by the way what is actually done in Statistics, but the converse is definitely not true and there exist many ways of representing a dominated probability model by a reference measure and a set of densities. It turns out, as shown by our example in the next section, that the performance of a very familiar estimator, namely the MLE (Maximum Likelihood Estimator), can be strongly affected by the choice of a specific version of the densities. Our purpose here is to design an estimator the performance of which only depends on $\mathcal{P}$ and not on the choices of the reference measure and the densities that are used to represent it.

In order to get rid of the above-mentioned restrictions, we have to modify our original construction which leads to the new version that we present here. This new version retains all the nice properties that we proved in Baraud, Birgé and Sart (2016) and the numerous illustrations we considered there remain valid for the new version. It additionally provides a general treatment of conditional density estimation and regression, allowing to estimate both the regression function and the error distribution even when the distribution of the design is totally unknown and the errors admit no finite moments. From this point of view, our approach contrasts very much with that based on the classical least squares.

An alternative point of view on the particular problem of estimating a conditional density can be found in Sart (2015).

A thorough study of the performance of the least squares estimator (or truncated versions of it) can be found in Györfi *et al.* (2002) and we refer the reader to the references therein. A nice feature of these results lies in the fact that they hold without any assumption on the distribution of the design. While few moment conditions on the errors are necessary to bound the $\mathbb{L}_2$-integrated risk of their estimator, much stronger ones, typically boundedness of the errors, are necessary to obtain exponential deviation bounds. In contrast, in linear regression, Audibert and Catoni (2011) established exponential deviation bounds for the risk of some robust versions of the ordinary least squares estimator. Their idea is to replace the sum of squares by the sum of some truncated version of these in view of designing a new criterion which is less sensitive to possible outliers than the original least squares. Their way of modifying the least squares criterion possesses some similarity to our way of modifying the log-likelihood criterion, as we shall see below. However their results require some conditions on the distribution of the design as well as some (weak) moment condition on the errors while ours do not.

It is known, and we shall give an additional example below, that the MLE, which is often considered as a "universal" estimator, does not share, in general, the properties that we require and more specifically robustness. An illustration of the lack of robustness of the MLE with respect to Hellinger deviations is provided in Baraud and Birgé (2016). Some other weaknesses of the MLE have been described in Le Cam (1990) and Birgé (2006), among other authors, and various alternatives aimed at designing some sorts of universal estimators which would not suffer from the same weaknesses have been proposed in the past by Le Cam (1973) and (1975) followed by Birgé (1983) and (2006). The construction of ρ-estimators, as described in Baraud, Birgé and Sart (2016) was in this line. In that paper, we actually introduced ρ-estimators via a testing argument as was the case for Le Cam and Birgé for their methods. This argument remains valid for the generalized version we consider here, as we shall see in Lemma 2 of Section 5, but ρ-estimators can also be viewed as a generalization, and in fact a robustified version, of the MLE on discretized versions of the models.

To see this, let us assume that we observe $\boldsymbol{X} = (X_1, \ldots, X_n)$ where the X_i are i.i.d. with an unknown density q belonging to a set $\mathcal{Q}$ of densities with respect to some reference measure μ. We may write the log-likelihood of q as $\sum_{i=1}^n \log\left(q(X_i)\right)$ and the log-likelihood ratios as

$$\mathbf{L}(\boldsymbol{X}, q, q') = \sum_{i=1}^n \log\left(\frac{q'(X_i)}{q(X_i)}\right) = \sum_{i=1}^n \log\left(q'(X_i)\right) - \sum_{i=1}^n \log\left(q(X_i)\right),$$

so that maximizing the likelihood is equivalent to minimizing with respect to q

$$\mathbf{L}(\boldsymbol{X}, q) = \sup_{q' \in \mathcal{Q}} \sum_{i=1}^n \log\left(\frac{q'(X_i)}{q(X_i)}\right) = \sup_{q' \in \mathcal{Q}} \mathbf{L}(\boldsymbol{X}, q, q').$$

This happens simply because of the magic property of the logarithm which says that $\log(a/b) = \log a - \log b$. However, the use of the unbounded log function in the definition of $\mathbf{L}(\boldsymbol{X}, q)$ leads to various problems that are responsable for some weaknesses of the MLE. Replacing the log function by another function φ amounts to replace $\mathbf{L}(\boldsymbol{X}, q, q')$ by

$$\mathbf{T}(\boldsymbol{X}, q, q') = \sum_{i=1}^n \varphi\left(\frac{q'(X_i)}{q(X_i)}\right)$$

which is different from $\sum_{i=1}^n \varphi\left(q'(X_i)\right) - \sum_{i=1}^n \varphi\left(q(X_i)\right)$ since φ is not the log-function. We may nevertheless define the analogue of $\mathbf{L}(\boldsymbol{X}, q)$, namely

$$\boldsymbol{\Upsilon}(\boldsymbol{X}, q) = \sup_{q' \in \mathcal{Q}} \mathbf{T}(\boldsymbol{X}, q, q') = \sup_{q' \in \mathcal{Q}} \sum_{i=1}^n \varphi\left(\frac{q'(X_i)}{q(X_i)}\right) \tag{1}$$

and define our estimator $\widehat{q}(\boldsymbol{X})$ as a minimizer with respect to $q \in \mathcal{Q}$ of the quantity $\boldsymbol{\Upsilon}(\boldsymbol{X}, q)$. The resulting estimator is an alternative to the maximum likelihood estimator and we shall show that, for a suitable choice of a bounded function φ it enjoys various properties, among which robustness, that are not shared by the MLE.

To analyze the performance of this new estimator, we have to study the behaviour of the process $\mathbf{T}(\boldsymbol{X}, q, q')$ when q is fixed, $q \cdot \mu$ is close to the true distribution of the X_i and q' varies in $\mathcal{Q}$. Since the function φ is bounded, the process is similar to those considered in learning theory for the purpose of studying empirical risk minimization. As a consequence, the tools we shall use are also similar to those used in that case and described in great details in Koltchinskii (2006).

It is well-known that working with a single model for estimating an unknown distribution is not very efficient unless one has very precise pieces of information about the true distribution, which is rarely the case. Working with many models simultaneously and performing model selection improves the situation drastically. Refining the previous construction of ρ-estimators by adding suitable penalty terms to the statistic $\mathbf{T}(\boldsymbol{X}, q, q')$ allows to work with a finite or countable family of probability models $\{\mathcal{P}_m,\ m \in \mathcal{M}\}$ instead of a single one, each model $\mathcal{P}_m$ leading to a risk bound of the form $C'\left[D(\mathcal{P}_m) + \mathbf{h}^2\left(\mathcal{P}_m, \mathbf{P}\right)\right]$, and to choose from the observations a model with approximately the best possible bound which results in a final estimator $\widehat{\mathbf{P}}$ and a bound for $\mathbf{h}^2(\widehat{\mathbf{P}}, \mathbf{P})$ of the form $C'' \inf_{m \in \mathcal{M}}\left[D(\mathcal{P}_m) + \mathbf{h}^2\left(\mathcal{P}_m, \mathbf{P}\right) + \Delta_m\right]$ where the additional term Δ_m is connected to the complexity of the family of models we use.

The paper is organised as follows. In the next section we shall give some simplified overview of the paper. Section 3 illustrates by a very simple example involving the MLE on a translation family the sensibility of classical estimators to the choice of the version of the densities of the distributions that enter the statistical model. We shall then make our framework, which is based on families of probabilities rather than densities, precise. Section 4 will be devoted to the definition of models and of our generalized ρ-estimators, then to the assumptions that the function φ we use to define the statistic $\mathbf{T}$ should satisfy. Section 5 provides an analysis of these assumptions and their consequences as well as two examples of functions that satisfy such assumptions. In Section 6, we define the dimension function of a model, a quantity which measures the difficulty of estimation within the model and present the main result, namely the performance of these new ρ-estimators. Section 7provides various methods that allow to bound the dimension function on different types of models while Section 8 indicates how these bounds are to be used to bound the risk of ρ-estimators in typical situations with applications to the minimax risk over classical statistical models. We also provide a few examples. Many applications of such results have already been given in Baraud, Birgé and Sart (2016) and we provide here a new one: estimation of conditional distributions in Section 9. We apply this additional result in Section 10 to the special case of random design regression when the distribution of the design is unknown, a situation for which not many results are known. We provide here a complete treatment of this regression framework with simultaneous estimation of both the regression function and the density of the errors. Section 11 is devoted to estimator selection and aggregation: we show there how our procedure can be used either to select an element from a family of preliminary estimators or to aggregate them in a convex way. Section 12 contains the proof of the main results while our last section includes all other proofs.

2. A partial overview of the paper and the results therein

The aim of this section is to provide a reader friendly overview of the paper, removing thus some subtleties and only considering the most typical situations in order to keep our exposure as simple as possible. For simplicity, we shall assume for a moment that the observations $X_1, \dots, X_n$ are i.i.d. with distribution P and denote by $h(\cdot,\cdot)$ the usual Hellinger distance (as defined below by (8)) and by $\mathbb{P}$ the probability that gives $\boldsymbol{X} = (X_1, \dots, X_n)$ the distribution $P^{\otimes n}$.

2.1. Models.

Classical estimation procedures like the MLE or more generally M-estimators are based on statistical models which are sets of probabilities $\overline{\mathscr{Q}}$ to which the true distribution P of the X_i is assumed to belong. In order that these procedures work correctly, rather strong assumptions on these models are required. In order to avoid such restrictions, we shall base the construction of our estimators, called ρ-estimators, on simpler sets which serve as substitutes for these statistical models. Starting from such a model $\overline{\mathscr{Q}}$, we replace it by a countable, therefore dominated, approximation $\mathscr{Q}$. Choosing some dominating measure μ and versions of the densities $dQ/d\mu$ for $Q \in \mathscr{Q}$, we may represent $\mathscr{Q}$ as $\{q \cdot \mu,\ q \in \mathcal{Q}\}$ for some countable set $\mathcal{Q}$ of densities with respect to μ. In typical situations, $\mathscr{Q}$ is either a countable and dense subset of $\overline{\mathscr{Q}}$ or an η-net for $\overline{\mathscr{Q}}$ for a well-chosen value of η and with respect to the Hellinger metric. In the remainder of this Section 2, we shall assume that such a convenient approximation $\mathscr{Q}$ of $\overline{\mathscr{Q}}$ has been chosen, together with μ and $\mathcal{Q}$.

2.2. **Construction of a ρ-estimator.** A ρ-estimator of P based on the model $\overline{\mathscr{Q}}$ is any probability $\widehat{P} = \widehat{p} \cdot \mu \in \mathscr{Q}$ with $\widehat{p} \in \mathcal{Q}$ minimizing over $\mathcal{Q}$ the criterion

$$q \mapsto \Upsilon(\boldsymbol{X}, q) = \sup_{q' \in \mathcal{Q}} \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q'(X_i)}{q(X_i)}}\right) \quad \text{with} \quad \psi(x) = \frac{x-1}{x+1}.$$

Let us make here some remarks:

a) A comparison with (1) shows that $\varphi(\cdot) = \psi(\sqrt{\cdot})$.

b) Other functions ψ can be used (see Sections 4.4 and 5). Nevertheless, we recommend so far to prefer this one because it leads to better constants in the bounds for the risk of the estimator.

c) The computation of the ρ-estimator is in general difficult. The density $\widehat{p}$ is a saddle-point of the map defined on $\mathcal{Q} \times \mathcal{Q}$ by

$$(q, q') \mapsto T(\boldsymbol{X}, q, q') = \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q'(X_i)}{q(X_i)}}\right).$$

However, when $\mathcal{Q}$ is a convex set of functions and the maps $q \mapsto T(\boldsymbol{X}, q, q')$ and $q \mapsto T(\boldsymbol{X}, q', q)$ are respectively convex and concave over $\mathcal{Q}$ for all fixed $q' \in \mathcal{Q}$, the computation of the ρ-estimator can be done efficiently at least in the following cases:
— when $\mathcal{Q}$ is a subset of a parametric model with a small number of parameters;
— when we have at disposal a finite set of preliminary estimators and we want to select one of them or build a convex aggregate of them (see Section 11).

d) When there is no exact minimizer of $q \mapsto \Upsilon(\boldsymbol{X}, q)$ any approximate minimizer can be used as a ρ-estimator (see Section 4.2.)

e) Although the construction of a ρ-estimator built over $\mathscr{Q}$ depends on the choice of some dominating measure μ and some specific versions of the densities $dQ/d\mu$ for Q in $\mathscr{Q}$, the performance of our estimator remains independent of these choices and only depends on $\mathscr{Q}$ (see the discussion of Sections 4.1 and 4.2).

2.3. **What is the performance of a ρ-estimator?** In typical situations (not all of them, but at this stage we shall only consider such cases), one can choose the approximation $\mathscr{Q}$ of $\overline{\mathscr{Q}}$ in such a way that the performance of a ρ-estimator $\widehat{P}$ built on $\mathscr{Q}$ only depends on the model $\overline{\mathscr{Q}}$ and the number n of observations. Assuming that the approximation $\mathscr{Q}$ of $\overline{\mathscr{Q}}$ has been chosen in such a way, we may consider the ρ-estimator $\widehat{P}$ built on $\mathscr{Q}$ as associated to $\overline{\mathscr{Q}}$ which we shall most of the time do in the remainder of this Section 2.

The performance of $\widehat{P}$ is measured by its Hellinger distance $h(P, \widehat{P})$ to the true distribution P and, in typical cases, the result we shall establish takes the following form: whatever the true distribution P,

$$\mathbb{P}\left[Ch^2(P, \widehat{P}) \le \inf_{Q \in \overline{\mathscr{Q}}} h^2(P, Q) + n^{-1}\left[D_n(\overline{\mathscr{Q}}) + \xi\right]\right] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0, \tag{2}$$

where C is a universal positive constant and $D_n(\overline{\mathscr{Q}}) \ge 1$ some number depending on both the structure of the statistical model $\overline{\mathscr{Q}}$ and the number of observations n, this due to the fact that we build our estimator as if the true joint distribution of the observations actually belonged to the family $\{Q^{\otimes n},\ Q \in \overline{\mathscr{Q}}\}$. Bounds of the form (2) are established in Section 8 and allow to control the minimax quadratic risk on $\overline{\mathscr{Q}}$ since (2) implies that

$$\sup_{P \in \overline{\mathscr{Q}}} \mathbb{E}\left[h^2(P, \widehat{P})\right] \le (Cn)^{-1}\left[D_n(\overline{\mathscr{Q}}) + 1\right].$$

The quantity $D_n(\overline{\mathscr{Q}})$ allows to deal with many different situations and corresponds, for most statistical models $\overline{\mathscr{Q}}$ we studied, to the difficulty of uniform estimation on $\overline{\mathscr{Q}}$, up to possible extra logarithmic factors. A simple illustration is as follows. Let $\overline{\mathscr{Q}} = \{q \cdot \mu, q \in \overline{\mathcal{Q}}\}$ for some dominating measure μ and densities q of the form $\max\{f, 0\}$ or $\exp(f)$ or, more generally, $G(f)$ for some non-negative monotone function G, with f varying among a d-dimensional linear space, $d \le n$. Then one can use for $\mathscr{Q}$ a countable and dense subset of $\overline{\mathscr{Q}}$ (with respect to the Hellinger distance) and $D_n(\overline{\mathscr{Q}})$ is of order $d \log n$. More generally, this actually holds whenever $\overline{\mathcal{Q}}$ is a VC-subgraph class of functions with dimension $d \ge 1$ (see Sections 7.2 and 8.1).

It is also possible to define a ρ-estimator from huge classes of probabilities for which the corresponding classes $\overline{\mathcal{Q}}$ of densities are neither VC nor posess a finite entropy with respect to the Hellinger metric. For example, one may consider the class $\overline{\mathcal{Q}}$ of all densities q (with respect to the Lebesgue measure) that are monotone on some interval $I = I(q)$ (depending on the density) and vanish elsewhere, or even the larger set of those that are monotone on each element of a partition $m = m(q)$ (also depending on q) of $\mathbb{R}$ into at most $k \ge 3$ intervals. It also possible to replace the constraint of monotonicity by others such as convexity, concavity or log-concavity. This leads to a series of possible non-parametric statistical models for P for which the corresponding ρ-estimator does not degenerate and remains robust. The performance of ρ-estimators on such models has been studied in Baraud and Birgé (2016).

2.4. **Why is a ρ-estimator robust?** The bias term in (2), namely $\inf_{Q \in \overline{\mathscr{Q}}} h^2(P, Q)$, accounts for the robustness property of the ρ-estimator with respect to the Hellinger distance and measures the additional loss we get as compared to the case when P belongs to $\overline{\mathscr{Q}}$. If this quantity is small, the performance of the ρ-estimator will not deteriorate too much as compared to the ideal situation where P really belongs to $\overline{\mathscr{Q}}$. More precisely, if for some probability $\overline{P} \in \overline{\mathscr{Q}}$, $\inf_{Q \in \overline{\mathscr{Q}}} h^2(P, Q) = h^2(P, \overline{P})$ is small as compared to $D_n(\overline{\mathscr{Q}})/n$, everything is almost as if the ρ-estimator $\widehat{P}$ were built from an i.i.d. sample with distribution $\overline{P}$. The ρ-estimators under P and $\overline{P}$ would therefore look the same. This includes the following situations:

Misspecification. The true distribution P of the observations does not belong to $\overline{\mathscr{Q}}$ but is close to $\overline{\mathscr{Q}}$. For example, $\overline{\mathscr{Q}}$ is the set of all Gaussian distributions on $\mathbb{R}^k$ with identity covariance matrix and mean vector belonging to a given linear subspace $S \subset \mathbb{R}^k$ while the true distribution P has the same form except from the fact that its mean does not belong to S but its Euclidean distance to S is $\varepsilon > 0$. Then, it follows from classical formulas that

$$\inf_{Q \in \overline{\mathscr{Q}}} h^2(P, Q) = 1 - e^{-\varepsilon^2/8} \le \frac{\varepsilon^2}{8}.$$

Contamination. The true distribution P is of the form $(1-\varepsilon)\overline{P} + \varepsilon R$ with $\overline{P} \in \overline{\mathscr{Q}}$ and $R \ne \overline{P}$ but otherwise arbitrary. This situation arises when a proportion $\varepsilon \in (0, 1)$ of the sample $X_1, \ldots, X_n$ is contaminated by another sample. If follows from the convexity property of the Hellinger distance that

$$\inf_{Q \in \overline{\mathscr{Q}}} h^2(P, Q) \leq h^2(P, \overline{P}) \leq \varepsilon h^2(R, \overline{P}) \leq \varepsilon,$$

and this bound holds whatever the contaminating distribution R. From a more practical point of view, one can see the contaminated case as follows: for each i one decides between no contamination with a probability $1-\varepsilon$ and contamination with a probability ε and draws X_i accordingly with distribution $\overline{P}$ or R. This means that a proportion of the data will be contaminated. If it were possible to extract from the sample $X_1, \ldots, X_n$ those N data, with $N \sim \mathcal{B}(n, 1-\varepsilon)$, which are really distributed according to the distribution $\overline{P} \in \overline{\mathscr{Q}}$ we would build our ρ-estimator $\widetilde{P}$ on such data. The robustness property ensures that the ρ-estimator $\widehat{P}$ based on the whole data set remains close to $\widetilde{P}$. Everything works almost as if the ρ-estimator $\widehat{P}$ only considered the non-contamined subsample and ignored the other data, at least when ε is small enough.

More robustness. There is an additional aspect of robustness that is not apparent in (2). Our general result about the performance of ρ-estimators, as stated in (21) below, actually allows that our observations be independent but not necessarily i.i.d., in which case the joint distribution $\mathbf{P}$ of $(X_1, \ldots, X_n)$ is actually of the form $\bigotimes_{i=1}^n P_i$ but not necessarily of the form $P^{\otimes n}$. Of course we do not know whether $\mathbf{P}$ is of the first form or the other and, proceding as if $X_1, \ldots, X_n$ were i.i.d., we build a ρ-estimator $\widehat{P} \in \mathscr{Q}$ of the presumed common density P and make a mistake which is no longer $h^2(P, \widehat{P})$ but

$$\frac{1}{n}\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \quad \text{with} \quad \widehat{\mathbf{P}} = \widehat{P}^{\otimes n} \quad \text{and} \quad \mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) = \sum_{i=1}^n h^2(P_i, \widehat{P}),$$

which is consistent with the i.i.d. case $P_i = P$ for all i.

In this context, we actually prove the following analogue of (2):

$$\text{(3)} \quad \mathbb{P}\left[\frac{C}{n}\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \leq \inf_{Q \in \overline{\mathscr{Q}}}\left(\frac{1}{n}\sum_{i=1}^n h^2(P_i, Q)\right) + \frac{D_n(\overline{\mathscr{Q}}) + \xi}{n}\right] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0.$$

This allows much more possibilities of deviations between $\mathbf{P}$ and the statistical model $\{Q^{\otimes n},\ Q \in \overline{\mathscr{Q}}\}$. For instance, we may have $h(P_i, \overline{P}) \leq \varepsilon$ for some $\overline{P} \in \overline{\mathscr{Q}}$ and all i, $P_i \neq P_{i'}$ for all $i \neq i'$, and nevertheless

$$\inf_{Q \in \overline{\mathscr{Q}}}\left(\frac{1}{n}\sum_{i=1}^n h^2(P_i, Q)\right) \leq \varepsilon^2.$$

An alternative situation corresponds to a small number of "outliers", namely, $P_i = P$ except on a subset $J \subset \{1, \ldots, n\}$ of indices of small cardinality and, for $i \in J$, P_i is completely arbitrary, for instance a Dirac measure. In such a case, for any probability Q,

$$\left(1 - \frac{|J|}{n}\right) h^2(P, Q) \leq \frac{1}{n}\sum_{i=1}^n h^2(P_i, Q) \leq \left(1 - \frac{|J|}{n}\right) h^2(P, Q) + \frac{|J|}{n},$$

and we deduce from (3) that, on a set of probability at least $1 - e^{-\xi}$,

$$\frac{C(n-|J|)}{n} h^2(P, \widehat{P}) \leq \frac{C}{n}\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \leq \left[\left(\frac{n-|J|}{n}\right)\inf_{Q \in \overline{\mathscr{Q}}} h^2(P, Q) + \frac{|J|}{n}\right] + \frac{D_n(\overline{\mathscr{Q}}) + \xi}{n}.$$

Finally,

$$\mathbb{P}\left[Ch^2(P,\widehat{P}) \le \inf_{Q\in\overline{\mathscr{Q}}} h^2(P,Q) + \frac{|J| + D_n(\overline{\mathscr{Q}}) + \xi}{n-|J|}\right] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0.$$

When $|J|/n$ is small enough, this bound appears to be a slight modification of what we would get from (2) if $\mathbf{P}$ were of the form $P^{\otimes n}$. This means that the ρ-estimator $\widehat{P}$ is also robust with respect to a possible departure from the assumption that the X_i are i.i.d.

2.5. **What can be done in the regression setting?** Our density framework applies to regression with random design where we observe i.i.d. pairs $X_i = (W_i, Y_i)$ of random variables with values in $\mathscr{W} \times \mathbb{R}$ and related by the equation

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

where f is the so called regression function mapping $\mathscr{W}$ into $\mathbb{R}$ and the ε_i are i.i.d. unobserved random variables called "noise" or "errors" that we shall assume to be independent of the W_i and to have some density s with respect to the Lebesgue measure λ on $\mathbb{R}$. Typically, neither the distribution P_W of the design points W_i nor the density s of the errors are known. A model $\overline{\mathscr{Q}}$ for the distribution P of the X_i can be designed on the basis of a model $\overline{F}$ for the regression function f, say a linear space with dimension d or more generally a VC-subgraph class of functions, and a candidate (unimodal) density r for s. Taking $\mu = P_W \otimes \lambda$ leads to the family of densities

$$\overline{\mathcal{Q}} = \left\{q_{r,g} : (w,u) \mapsto r(u - g(w)),\ g \in \overline{F}\right\}$$

which does not depend on the unknown distribution P_W of the design. We may therefore use our procedure to get an estimator $\widehat{f}$ of the regression f as an element $\widehat{f} \in \overline{F}$ satisfying $\widehat{p} = q_{r,\widehat{f}} \in \mathcal{Q}$. This estimator satisfies

$$\text{(4)} \qquad \mathbb{P}\left[Cd_s^2(f,\widehat{f}) \le \inf_{g\in\overline{F}} d_s^2(f,g) + h^2(s,r) + \frac{d\log(en/d) + \xi}{n}\right] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0$$

where C is a universal constant and $d_s(\cdot,\cdot)$ some loss function depending on the properties of the density s. If f and all the elements of $\overline{F}$ are bounded by some constant b and s is the Gaussian, Cauchy or Laplace density, $d_s(\cdot,\cdot)$ is equivalent (up to constants depending on b) to the classical $\mathbb{L}_2(P_W)$-distance while, for s being the uniform or the exponential density, $d_s^2(\cdot,\cdot)$ is equivalent to the $\mathbb{L}_1(P_W)$-distance. Inequality (4) is free of any assumption on the distribution of the design which can therefore be arbitrary. Besides, it shows that the ρ-estimator of f is robust not only with respect to a possible misspecification of the model $\overline{F}$ for f but also of the distribution of the errors ε_i.

2.6. **Introducing several models.** Of course, dealing with a single model $\overline{F}$ for f and a single candidate density r for s in the regression setting is strongly restrictive. Hopefully ρ-estimation can also be combined with a model selection procedure which is detailed in Section 8 allowing to deal with a family $\mathbb{F}$ of candidate models $\overline{F}$ for f and a family $\mathcal{R}$ of candidate densities for s, as described in Section 10. Using all these models simultaneously, we obtain a pair of ρ-estimators $(\widehat{f},\widehat{s})$ with $\widehat{f}$ estimating the regression function f and $\widehat{s}$ the density s of the errors. We shall prove that both losses $d_s^2(f,\widehat{f})$ and $h^2(s,\widehat{s})$ are not larger, with probability at least $1-e^{-\xi}$, $\xi>0$, than

$$\inf_{\overline{F}\in\mathbb{F}} d_s^2(f,\overline{F})+\inf_{r\in\mathcal{R}} h^2(s,r)+\frac{D_n(\overline{F})+\Delta(\overline{F})+\Delta'(r)+\xi}{n}$$

up to a universal factor $C\geq 1$, the nonnegative maps $\Delta(\cdot)$ and $\Delta'(\cdot)$ accounting for the complexities of the collections $\mathbb{F}$ and $\mathcal{R}$ respectively and $D_n(\overline{F})$ corresponding to the "dimension" of the model $\overline{F}$ as in (2). This possibility of introducing many models in order to approximate more closely the true unknown distribution brings a lot of flexibility (at least from a theoretical point of view) to our method.

3. Our new framework

As already mentioned, our method is based on statistical models which are sets of probability distributions, in opposition with more classical models which are sets of densities with respect to a given dominating measure. In order to emphasize the originality of this approach, let us first consider the impact of this difference on the behaviour of the MLE.

3.1. **A counterexample.** In Baraud, Birgé and Sart (2016), all probabilities that we used, including the true joint distribution of the observations, were dominated by some given measure $\boldsymbol{\mu}$ with well-defined densities viewed as functions on the observation space (rather than elements of $\mathbb{L}_1(\boldsymbol{\mu})$). This is actually the common practice in Statistics. When we say that we work with the Gaussian model $\{P_\theta=\mathcal{N}(\theta,1),\ \theta\in\mathbb{R}\}$ we actually mean that we work with the family of density functions $\left(1/\sqrt{2\pi}\right)\exp\left[-(x-\theta)^2/2\right]$ with respect to the Lebesgue measure μ on $\mathbb{R}$, these densities being viewed as classical functions, not elements of $\mathbb{L}_1(\mu)$, although this is not explicitely mentioned. This is an important fact because choosing other versions of the densities $dP_\theta/d\mu$ would change the value of many classical estimators, in particular the most celebrated one, namely the MLE. This would not only change the value of the MLE but also its performance as shown by the following example.

Proposition 1. *Let us consider a sequence of random variables $(X_k)_{k\geq1}$ defined on a measurable space $(\Omega,\mathscr{A},\mathbb{P})$ with distribution $P_\theta=\mathcal{N}(\theta,1)$ for some $\theta\in\mathbb{R}$. We choose the family of densities*

$$x\mapsto p_\theta(x)=\exp\left[\theta x-\left(\theta^2/2\right)+(\theta^2/2)\exp\left(x^2\right)\mathbb{1}_\theta(x)\mathbb{1}_{(0,+\infty)}(\theta)\right]\quad \text{for } \theta\in\mathbb{R}, \tag{5}$$

with respect to the reference measure μ with density $\left(1/\sqrt{2\pi}\right)\exp\left[-x^2/2\right]$ with respect to the Lebesgue measure. Whatever the true parameter, on a set of probability tending to 1 when n goes to infinity, the MLE is given by $X_{(n)}=\max\{X_1,\ldots,X_n\}$ and is therefore inconsistent.

Proof. Let us denote by $X_{(1)},\ldots,X_{(n)}$ the order statistics, by $\overline{X}_n$ the mean of the observations and let us work on the set $\Omega_n\subset\Omega$ on which the following properties are satisfied:

$$\overline{X}_n\notin\left\{X_{(1)},\ldots,X_{(n)}\right\}\qquad\text{and}\qquad X_{(n)}\geq\sqrt{\log(4n)}>\left|\overline{X}_n\right|. \tag{6}$$

Since $\mathbb{P}(\Omega_n)\underset{n\to+\infty}{\longrightarrow}1$ whatever $\theta\in\mathbb{R}$, it is enough to show that the MLE is attained at $X_{(n)}$ when the event Ω_n holds which we now assume. It implies in particular that

$$n^{-1}\exp\left[X_{(n)}^2\right]-1\geq 3. \tag{7}$$

For all $\theta \in \mathbb{R}$, the log-likelihood writes $nL_n(\theta)$ with

$$L_n(\theta) = \theta\overline{X}_n - \frac{\theta^2}{2} + \frac{\theta^2}{2n}\left(\sum_{k=1}^{n} \exp\left[X_{(k)}^2\right] \mathbb{1}_{X_{(k)}}(\theta)\right) \mathbb{1}_{(0,+\infty)}(\theta).$$

If either $\theta < 0$ or $\theta > X_{(n)}$, by (6) and (7),

$$L_n(\theta) = \theta\overline{X}_n - \left(\theta^2/2\right) \leq \overline{X}_n^2/2 < X_{(n)}\overline{X}_n + \frac{X_{(n)}^2}{2}\left(n^{-1}\exp\left[X_{(n)}^2\right] - 1\right) = L_n\big(X_{(n)}\big)\,,$$

so that the MLE necessarily belongs to the interval $[0, X_{(n)}]$. Moreover,

$$L_n(\theta) \leq \overline{L}_n(\theta) = \theta\overline{X}_n + \left(\theta^2/2\right)\left(n^{-1}\exp\left[X_{(n)}^2\right] - 1\right) \quad \text{for all } \theta$$

and by (7), $\overline{L}_n$ is strictly convex. Since by (6) and (7),

$$\overline{L}_n(0) = 0 < X_{(n)}\overline{X}_n + X_{(n)}^2 \leq \overline{L}_n\big(X_{(n)}\big)\,,$$

by convexity the unique maximum of the function $\overline{L}_n$ on the interval $[0, X_{(n)}]$ is reached at the point $X_{(n)}$. The conclusion follows since $L_n\big(X_{(n)}\big) = \overline{L}_n\big(X_{(n)}\big)$ and $L_n(\theta) \leq \overline{L}_n(\theta)$ for all θ. □

Note that the usual choice of $p_\theta : x \mapsto e^{-x\theta+\theta^2/2}$ for $dP_\theta/d\mu$ is purely conventional. Mathematically speaking our choice (5) is perfectly correct but leads to an inconsistent MLE. Also note that the usual tools that are used to prove consistency of the MLE, like bracketing entropy (see for instance Theorem 7.4 of van de Geer (2000)) are not stable with respect to changes of versions of the densities in the family. The same is true for arguments based on VC-classes that we used in Baraud, Birgé and Sart (2016). Choosing a convenient set of densities to work with is well-grounded as long as the reference measure $\boldsymbol{\mu}$ not only dominates the model but also the true distribution $\mathbf{P}$. If not, sets of nul measure with respect to $\boldsymbol{\mu}$ might have a positive probability under $\mathbf{P}$ and it becomes unclear how the choice of this family influences the performance of the estimator. In order to avoid these problems we shall work with probabilities rather than densities which accounts for the framework below which contrasts with the one we considered in Baraud, Birgé and Sart (2016).

3.2. **A purely probabilistic framework.** We observe a random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ defined on some probability space $(\Omega, \Xi, \mathbb{P})$ with independent components X_i and values in the measurable product space $(\mathscr{X}, \mathscr{B}) = (\prod_{i=1}^n \mathscr{X}_i, \bigotimes_{i=1}^n \mathscr{B}_i)$. We denote by $\mathscr{P}$ the set of all product probabilities on $(\mathscr{X}, \mathscr{B})$ and by $\mathbf{P} = \bigotimes_{i=1}^n P_i \in \mathscr{P}$ the true distribution of $\boldsymbol{X}$. We identify an element $\mathbf{Q} = \bigotimes_{i=1}^n Q_i$ of $\mathscr{P}$ with the n-uplet $(Q_1, \ldots, Q_n)$ and extend this identification to the elements $\boldsymbol{\mu} = \bigotimes_{i=1}^n \mu_i$ of the set $\mathscr{M}$ of all product measures on $(\mathscr{X}, \mathscr{B})$.

Our aim is to estimate the unknown distribution $\mathbf{P} = (P_1, \ldots, P_n)$ from the observation of $\boldsymbol{X}$. In order to evaluate the performance of an estimator $\widehat{\mathbf{P}}(\boldsymbol{X}) \in \mathscr{P}$ of $\mathbf{P}$, we shall introduce, following Le Cam (1975), an Hellinger-type distance $\mathbf{h}$ on $\mathscr{P}$. We recall that given two probabilities Q and Q' on a measurable space $(\mathscr{X}, \mathscr{B})$, the Hellinger distance and the Hellinger affinity between Q and Q' are respectively given by

$$h^2(Q,Q') = \frac{1}{2}\int_{\mathscr{X}}\left(\sqrt{\frac{dQ}{d\mu}}-\sqrt{\frac{dQ'}{d\mu}}\right)^2 d\mu, \tag{8}$$

$$\rho(Q,Q') = \int_{\mathscr{X}}\sqrt{\frac{dQ}{d\mu}\frac{dQ'}{d\mu}}d\mu = 1-h^2(Q,Q'),$$

where μ denotes any measure dominating both Q and Q', the result being independent of the choice of μ. The Hellinger-type distance $\mathbf{h}(\mathbf{Q},\mathbf{Q}')$ and affinity $\boldsymbol{\rho}(\mathbf{Q},\mathbf{Q}')$ between two elements $\mathbf{Q}=(Q_1,\ldots,Q_n)$ and $\mathbf{Q}'=(Q'_1,\ldots,Q'_n)$ of $\mathscr{P}$ are then given by the formulas

$$\mathbf{h}^2(\mathbf{Q},\mathbf{Q}') = \sum_{i=1}^n h^2(Q_i,Q'_i) = \sum_{i=1}^n \left[1-\rho(Q_i,Q'_i)\right] = n-\boldsymbol{\rho}(\mathbf{Q},\mathbf{Q}').$$

4. Our estimation strategy

4.1. **Models and their representations.** Our estimation strategy is based on what we call *ρ-models*.

Definition 1. *A ρ-model is a* countable *(which in this paper always means either finite or infinite and countable) subset $\mathscr{Q}$ of $\mathscr{P}$.*

One should think of a ρ-model as a probability set for which we believe that the true distribution $\mathbf{P}$ is reasonnably close to it (with respect to the Hellinger-type distance $\mathbf{h}$) and our estimator is defined as a random element of $\mathscr{Q}$ or of its closure in the metric space $(\mathscr{P},\mathbf{h})$. The construction of our estimator is actually based on a specific representation of our ρ-model via a dominating measure and a family of densities. In order to be more precise, let us first introduce some additional notations.

Given a measure μ on a measurable space $(\mathscr{X},\mathscr{B})$ we call *density with respect to* μ any measurable function q from $\mathscr{X}$ to $\mathbb{R}_+$ such that $\int_{\mathscr{X}} q(x)\,d\mu(x)=1$ and denote by $\mathcal{L}(\mu)$ the set of all densities with respect to μ. We write $Q=q\cdot\mu$ for the probability on $\mathscr{X}$ with density q. When a probability $\mathbf{Q}=\bigotimes_{i=1}^n Q_i\in\mathscr{P}$ is such that $Q_i=q_i\cdot\mu_i$ for $i=1,\ldots,n$, we write $\mathbf{Q}=\mathbf{q}\cdot\boldsymbol{\mu}$ where $\mathbf{q}$ is the n-uplet $(q_1,\ldots,q_n)$ and we say that $\boldsymbol{\mu}$ dominates $\mathbf{Q}$ and $\mathbf{q}$ is a density for $\mathbf{Q}$ with respect to $\boldsymbol{\mu}$. Note that since a ρ-model $\mathscr{Q}$ is countable, it is necessarily dominated.

Definition 2. *Given a ρ-model $\mathscr{Q}\subset\mathscr{P}$ we call* representation of the ρ-model $\mathscr{Q}$ *a pair $\mathcal{R}(\mathscr{Q})=(\boldsymbol{\mu},\boldsymbol{\mathcal{Q}})$ where $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)$ is a measure which dominates $\mathscr{Q}$ and $\boldsymbol{\mathcal{Q}}$ is a family of densities $\mathbf{q}=(q_1,\ldots,q_n)$ with $q_i\in\mathcal{L}(\mu_i)$ for $i=1,\ldots,n$ such that $\mathscr{Q}=\{\mathbf{q}\cdot\boldsymbol{\mu},\ \mathbf{q}\in\boldsymbol{\mathcal{Q}}\}$.*

Clearly a given ρ-model $\mathscr{Q}$ has many different representations depending on the choice of the dominating measure $\boldsymbol{\mu}$ and the versions of the densities $q_i=dQ_i/d\mu_i$. Unfortunately, as we have seen in Section 3.1, such choices may have strong consequences not only on the expression of the resulting estimator but also on its performance. However, this is not the case for ρ-estimators, the performances of which can be bounded independently of the chosen representation.

4.2. **Construction of a ρ-estimator.** The construction of our estimator depends on three elements with specific properties to be made precise below: a function ψ from $[0,+\infty]$ to $[-1,1]$ which will serve as a substitute for the logarithm to derive an alternative to the MLE, a ρ-model $\boldsymbol{\mathscr{Q}}$ together with a given, although arbitrary, representation $\mathcal{R}(\boldsymbol{\mathscr{Q}})=(\boldsymbol{\mu},\boldsymbol{\mathcal{Q}})$ of it and a *penalty function* "pen" mapping $\boldsymbol{\mathscr{Q}}$ to $\mathbb{R}_+$, the role of which will be explained later in Section 6. We may, in a first reading, assume that this penalty is identically 0. It is essential to note that the reference measure $\boldsymbol{\mu}$ is chosen by the statistician and that there is no reason that the true distribution $\mathbf{P}$ of $\boldsymbol{X}$ satisfies $\mathbf{P}\ll\boldsymbol{\mu}$.

Given the function ψ and the representation $\mathcal{R}(\boldsymbol{\mathscr{Q}})$, we define the real-valued function $\mathbf{T}$ on $\mathscr{X}\times\boldsymbol{\mathcal{Q}}\times\boldsymbol{\mathcal{Q}}$ by

$$\mathbf{T}(\mathbf{x},\mathbf{q},\mathbf{q}')=\sum_{i=1}^{n}\psi\left(\sqrt{\frac{q_i'(x_i)}{q_i(x_i)}}\right)\quad\text{for } \mathbf{x}=(x_1,\ldots,x_n)\in\mathscr{X}\ \text{ and }\ \mathbf{q},\mathbf{q}'\in\boldsymbol{\mathcal{Q}},\tag{9}$$

with the conventions $0/0=1$ and $a/0=+\infty$ for all $a>0$. We then set (with $\mathbf{Q}=\mathbf{q}\cdot\boldsymbol{\mu}$ and $\mathbf{Q}'=\mathbf{q}'\cdot\boldsymbol{\mu}$)

$$\boldsymbol{\Upsilon}(\boldsymbol{X},\mathbf{q})=\sup_{\mathbf{q}'\in\boldsymbol{\mathcal{Q}}}\left[\mathbf{T}(\boldsymbol{X},\mathbf{q},\mathbf{q}')-\operatorname{pen}(\mathbf{Q}')\right]+\operatorname{pen}(\mathbf{Q})\quad\text{for all } \mathbf{q}\in\boldsymbol{\mathcal{Q}}.\tag{10}$$

Finally, given $\kappa>0$ to be made precise later by (22), we define our estimator $\widehat{\mathbf{P}}=\widehat{\mathbf{P}}(\boldsymbol{X})$ as any (measurable) element belonging to the closure in $\boldsymbol{\mathscr{P}}$ (with respect to the distance $\mathbf{h}$) of the random (and non-void) set

$$\mathscr{E}(\psi,\boldsymbol{X})=\left\{\mathbf{Q}=\mathbf{q}\cdot\boldsymbol{\mu},\ \mathbf{q}\in\boldsymbol{\mathcal{Q}}\ \text{ such that }\ \boldsymbol{\Upsilon}(\boldsymbol{X},\mathbf{q})\le\inf_{\mathbf{q}'\in\boldsymbol{\mathcal{Q}}}\boldsymbol{\Upsilon}(\boldsymbol{X},\mathbf{q}')+\frac{\kappa}{25}\right\}.\tag{11}$$

Since all the probabilities in $\mathscr{E}(\psi,\boldsymbol{X})$ are dominated by $\boldsymbol{\mu}$, so is $\widehat{\mathbf{P}}$ and there exists a random density $\widehat{\mathbf{p}}=(\widehat{p}_1,\ldots,\widehat{p}_n)$ on $(\mathscr{X},\boldsymbol{\mathscr{B}})$ with $\widehat{p}_i\in\mathcal{L}(\mu_i)$ for $i=1,\ldots,n$ such that $\widehat{\mathbf{P}}=\widehat{\mathbf{p}}\cdot\boldsymbol{\mu}$. We shall call such an estimator a *ρ-estimator*.

It is clear from this construction that the ρ-estimator $\widehat{\mathbf{P}}$ depends on the chosen representation $\mathcal{R}(\boldsymbol{\mathscr{Q}})$ of the ρ-model $\boldsymbol{\mathscr{Q}}$ and that there are different versions of the ρ-estimators associated to a given ρ-model $\boldsymbol{\mathscr{Q}}$. The important point, that we shall prove in Section 6, is that the risk bounds we shall derive only depend on the ρ-model $\boldsymbol{\mathscr{Q}}$ and the function "pen" but not on the chosen representation, which allows us to choose the more convenient one for the construction. By the way, most of the time, the ρ-model will be directly given by a specific representation, that is a family $\boldsymbol{\mathcal{Q}}$ of densities with respect to some reference measure $\boldsymbol{\mu}$.

4.3. **Notations and conventions.** Throughout this paper, given a representation $\mathcal{R}(\boldsymbol{\mathscr{Q}})=(\boldsymbol{\mu},\boldsymbol{\mathcal{Q}})$ of the ρ-model $\boldsymbol{\mathscr{Q}}$, we shall use lower case letters $\mathbf{q},\mathbf{q}',\ldots$ and $q_i,q_i',\ldots$ for denoting the chosen densities of $\mathbf{Q},\mathbf{Q}',\ldots$ and $Q_i,Q_i',\ldots$ with respect to the reference measures $\boldsymbol{\mu}$ and μ_i respectively for all $i=1,\ldots,n$.

Of special interest is the situation where the X_i are i.i.d. with values in a measurable set $(\mathscr{X},\mathscr{B})$ in which case $\mathbf{P}$ is of the form $P^{\otimes n}$ for some probability measure P on $(\mathscr{X},\mathscr{B})$. Since estimating $\mathbf{P}$ amounts to estimating the marginal P in such a situation, we model the probability P rather than $\mathbf{P}$ and use unbold letters in the following way: $\mathscr{P}$ and $\mathscr{M}$ denote respectively the set of all probability distributions and all positive measures on $(\mathscr{X},\mathscr{B})$. A ρ-model $\mathscr{Q}$ for P is a countable subset of $\mathscr{P}$ and the corresponding ρ-model $\boldsymbol{\mathscr{Q}}$ for $\mathbf{P}$ is simply $\{\mathbf{Q} = Q^{\otimes n},\ Q \in \mathscr{Q}\}$. Note that the Hellinger distance $h(\cdot,\cdot)$ on $\mathscr{P}$ is related to the Hellinger-type distance $\mathbf{h}(\cdot,\cdot)$ on $\boldsymbol{\mathscr{P}}$ in the following way:

$$\mathbf{h}^2(\mathbf{Q},\mathbf{Q}') = nh^2(Q,Q') \ \text{ for all } \ Q,Q' \in \mathscr{P}.$$

We set $\log_+(x) = \max\{\log x, 0\}$ for all $x > 0$; $|A|$ denotes the cardinality of the set A; $\mathcal{P}(A)$ is the class of all subsets of A; $\boldsymbol{\mathscr{B}}(\mathbf{P}, r) = \{\mathbf{Q} \in \boldsymbol{\mathscr{P}} \mid \mathbf{h}(\mathbf{P},\mathbf{Q}) \le r\}$ is the closed Hellinger-type ball in $\boldsymbol{\mathscr{P}}$ with center $\mathbf{Q}$ and radius r (with respect to the distance $\mathbf{h}$). Given a set E, a nonnegative function ℓ on $E \times E$, $x \in E$ and $A \subset E$, we set $\ell(x,A) = \inf_{y \in A} \ell(x,y)$. In particular, for $\boldsymbol{\mathscr{R}} \subset \boldsymbol{\mathscr{P}}$, $\mathbf{h}(\mathbf{P},\boldsymbol{\mathscr{R}}) = \inf_{\mathbf{R} \in \boldsymbol{\mathscr{R}}} \mathbf{h}(\mathbf{P},\mathbf{R})$. We set $x \vee y$ and $x \wedge y$ for $\max\{x,y\}$ and $\min\{x,y\}$ respectively. By convention $\sup_\varnothing = 0$, the ratio $u/0$ equals $+\infty$ for $u > 0$, $-\infty$ for $u < 0$ and 1 for $u = 0$.

4.4. **Our assumptions.** Given a ρ-model $\boldsymbol{\mathscr{Q}}$, let us now specify which properties of the function ψ are required in order to control the risk of the resulting ρ-estimators.

Assumption 1. *Let $\boldsymbol{\mathscr{Q}}$ be a ρ-model.*

i) The function ψ is monotone from $[0,+\infty]$ to $[-1,1]$ and satisfies

$$\psi(x) = -\psi(1/x) \ \text{ for all } \ x \in [0,+\infty). \tag{12}$$

ii) There exist three positive constants a_0, a_1, a_2 with $a_0 \ge 1 \ge a_1$ and $a_2^2 \ge 1 \vee (6a_1)$ such that, whatever the representation $\mathcal{R}(\boldsymbol{\mathscr{Q}}) = (\boldsymbol{\mu}, \boldsymbol{\mathcal{Q}})$ of $\boldsymbol{\mathscr{Q}}$, the densities $\mathbf{q},\mathbf{q}' \in \boldsymbol{\mathcal{Q}}$ and the probability $\mathbf{R} \in \boldsymbol{\mathscr{P}}$,

$$\int_{\mathscr{X}_i} \psi\left(\sqrt{\frac{q_i'}{q_i}}\right) dR_i \le a_0 h^2(R_i,Q_i) - a_1 h^2(R_i,Q_i') \quad \textit{for all } i \in \{1,\ldots,n\} \tag{13}$$

and

$$\int_{\mathscr{X}_i} \psi^2\left(\sqrt{\frac{q_i'}{q_i}}\right) dR_i \le a_2^2 \left[h^2(R_i,Q_i) + h^2(R_i,Q_i')\right] \quad \textit{for all } i \in \{1,\ldots,n\}. \tag{14}$$

In view of checking that a given function ψ satisfies Assumption 1, the next result to be proved in Section 13.1 is useful.

Proposition 2. *Let ψ satisfy (12). If for a particular representation $\mathcal{R}(\boldsymbol{\mathscr{Q}}) = (\boldsymbol{\mu}, \boldsymbol{\mathcal{Q}})$ of $\boldsymbol{\mathscr{Q}}$ and any probability $\mathbf{R} \in \boldsymbol{\mathscr{P}}$ which is absolutely continuous with respect to $\boldsymbol{\mu}$ the function ψ satisfies (13) and (14) for positive constants $a_0 > 2$, $a_1 \le [(a_0-2)/2] \wedge 1$ and $a_2^2 \ge 1 \vee (6a_1)$, then it satisfies Assumption 1 with the same constants a_0, a_1 and a_2.*

This proposition means that, up to a possible adjustment of the constants a_0 and a_1, it is actually enough to check that (13) and (14) hold true for a given representation $(\boldsymbol{\mu}, \boldsymbol{\mathcal{Q}})$ of $\boldsymbol{\mathscr{Q}}$ and all probabilities $\mathbf{R} \ll \boldsymbol{\mu}$.

5. About the function ψ

5.1. **Some comments on Assumption 1.**

a) It follows from (12) that necessarily $\psi(1) = 0$. Also observe that, with the convention $1/0 = +\infty$, $\psi(+\infty) = -\psi(0)$.

b) We deduce from (12) that

$$\mathbf{T}(\boldsymbol{X},\mathbf{q},\mathbf{q}') = -\mathbf{T}(\boldsymbol{X},\mathbf{q}',\mathbf{q}) \quad \text{for all } \mathbf{q},\mathbf{q}' \in \boldsymbol{\mathscr{Q}}. \tag{15}$$

c) In view of (13) and (14), the conditions $a_0 \geq 1$, $a_1 \leq 1$ and $a_2^2 \geq 1 \vee (6a_1)$ can always be satisfied by enlarging a_0 and a_2 and diminishing a_1 if necessary. The conditions $a_0 \geq 1 \geq a_1$ and $a_2 \geq 1$ turn out to be necessary when $\psi(+\infty) = 1$ and there exist two probabilities $\mathbf{Q},\mathbf{Q}' \in \boldsymbol{\mathscr{Q}}$ such that $h(Q_i, Q_i') = 1$ for some $i \in \{1,\dots,n\}$. In this case, for $R_i = Q_i'$ and any reference measure μ_i, $\psi\left(\sqrt{(q_i'/q_i)(x)}\right) = \psi(+\infty) = 1$ for R_i-almost all $x \in \mathscr{X}_i$ so that the left-hand side of (13) equals 1 while the right-hand side equals $a_0 h^2(R_i, Q_i) = a_0$ leading to the inequality $a_0 \geq 1$. The same argument applies to (14) and leads to the inequality $a_2 \geq 1$. Taking now $R_i = Q_i$, $\psi\left(\sqrt{q_i'/q_i}(x)\right) = \psi(0) = -1$ for R_i-almost all $x \in \mathscr{X}_i$ so that the left-hand side of (13) equals -1 while the right hand-side equals $-a_1 h^2(R_i, Q_i') = -a_1$ which leads to the inequality $a_1 \leq 1$.
d) As we have just seen, a_0, a_1 and a_2 are not uniquely defined but, in the sequel, when we shall say that Assumption 1 holds, this will mean that the function ψ satisfies (13) and (14) with given values of these three constants which will therefore be considered as fixed once ψ has been chosen.
e) Note that the left-hand sides of (13) and (14) depend on the choices of the reference measures μ_i and versions of the densities $q_i = dQ_i/d\mu_i$ and $q_i' = dQ_i'/d\mu_i$ while the corresponding right-hand sides do not.
f) Inequality (14) is satisfied as soon as ψ is Lipschitz on $[0,+\infty)$. It actually follows from Lemma 1 below (to be proved in Section 13.2) that (14) holds as soon as there exists a constant $L > 0$ such that

$$|\psi(x)| = |\psi(x) - \psi(1)| \leq L\,|x-1| \quad \text{for all} \quad x \in \mathbb{R}_+. \tag{16}$$

Lemma 1. *If ψ satisfies Assumption 1-i) and (16) for some constant $L > 0$, inequality (14) is satisfied with $a_2 = 2L + 1$.*

5.2. **Some consequences of Assumption 1.** It follows from (12) and (13) that, for all $\mathbf{P} \in \boldsymbol{\mathscr{P}}$, $\mathbf{Q},\mathbf{Q}' \in \boldsymbol{\mathscr{Q}}$ and $i \in \{1,\dots,n\}$,

$$\int_{\mathscr{X}_i} \psi\left(\sqrt{\frac{q_i}{q_i'}}\right) dP_i \geq a_1 h^2(P_i, Q_i') - a_0 h^2(P_i, Q_i).$$

By exchanging the roles of q_i and q_i' in the above inequality, we obtain the following inequalities which hold for all $i = 1,\dots,n$:

$$a_1 h^2(P_i, Q_i) - a_0 h^2(P_i, Q_i') \leq \int_{\mathscr{X}_i} \psi\left(\sqrt{\frac{q_i'}{q_i}}\right) dP_i \leq a_0 h^2(P_i, Q_i) - a_1 h^2(P_i, Q_i').$$

Summing these inequalities with respect to i and using (9) leads to

$$a_1 \mathbf{h}^2(\mathbf{P},\mathbf{Q}) - a_0 \mathbf{h}^2(\mathbf{P},\mathbf{Q}') \leq \mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\mathbf{q},\mathbf{q}')\right] \leq a_0 \mathbf{h}^2(\mathbf{P},\mathbf{Q}) - a_1 \mathbf{h}^2(\mathbf{P},\mathbf{Q}'). \tag{17}$$

The right-hand side of the inequality shows that if $\mathbf{h}^2(\mathbf{P},\mathbf{Q}) < a_1 a_0^{-1} \mathbf{h}^2(\mathbf{P},\mathbf{Q}')$, then $\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\mathbf{q},\mathbf{q}')\right]$ is negative while the left-hand side shows that it is positive when $\mathbf{h}^2(\mathbf{P},\mathbf{Q}') < a_1 a_0^{-1} \mathbf{h}^2(\mathbf{P},\mathbf{Q})$. In particular if $\mathbf{T}(\boldsymbol{X},\mathbf{q},\mathbf{q}')$ is close enough to its expectation, its sign may be used as a test statistic to decide which of the two probabilities $\mathbf{Q}$ and $\mathbf{Q}'$ is closer to $\mathbf{P}$. Bounds for the probabilities of errors of this test are provided by the following lemma.

Lemma 2. *Let ψ satisfy Assumption 1, $\mathbf{P} \in \mathscr{P}$, $\mathbf{Q}, \mathbf{Q}' \in \mathscr{Q}$ and $x \geq 0$. Whatever the representation $\mathcal{R}(\mathscr{Q})$ of the ρ-model $\mathscr{Q}$, if $\mathbf{h}^2(\mathbf{P}, \mathbf{Q}) < a_1 a_0^{-1} \mathbf{h}^2(\mathbf{P}, \mathbf{Q}')$, then*

$$\mathbb{P}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}') \geq x\right] \leq \exp\left[\frac{-\left(a_1 \mathbf{h}^2(\mathbf{P}, \mathbf{Q}') - a_0 \mathbf{h}^2(\mathbf{P}, \mathbf{Q}) + x\right)^2}{2\left[(a_2^2 + a_1/3)\mathbf{h}^2(\mathbf{P}, \mathbf{Q}') + (a_2^2 - a_0/3)\mathbf{h}^2(\mathbf{P}, \mathbf{Q}) + x/3\right]}\right]$$

while if $\mathbf{h}^2(\mathbf{P}, \mathbf{Q}') < a_1 a_0^{-1} \mathbf{h}^2(\mathbf{P}, \mathbf{Q})$, then

$$\mathbb{P}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}') \leq -x\right] \leq \exp\left[\frac{-\left(a_1 \mathbf{h}^2(\mathbf{P}, \mathbf{Q}) - a_0 \mathbf{h}^2(\mathbf{P}, \mathbf{Q}') + x\right)^2}{2\left[(a_2^2 + a_1/3)\mathbf{h}^2(\mathbf{P}, \mathbf{Q}) + (a_2^2 - a_0/3)\mathbf{h}^2(\mathbf{P}, \mathbf{Q}') + x/3\right]}\right].$$

Proof. Using the right-hand side of (17), we may write

$$\begin{aligned}
&\mathbb{P}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}') \geq x\right] \\
&\quad = \mathbb{P}\Big[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}') - \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}')\right] \geq x - \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}')\right]\Big] \\
&\quad \leq \mathbb{P}\Big[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}') - \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}')\right] \geq a_1 \mathbf{h}^2(\mathbf{P}, \mathbf{Q}') - a_0 \mathbf{h}^2(\mathbf{P}, \mathbf{Q}) + x\Big].
\end{aligned}$$

We obtain the first inequality by applying the Bernstein deviation inequality (see Massart (2007), inequality (2.16)) to $\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}') - \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}')\right]$. Indeed, $\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}')$ is a sum of independent random variables with absolute value not larger than 1 (since $|\psi|$ is bounded by 1) and by (14),

$$\sum_{i=1}^{n} \int_{\mathscr{X}_i} \psi^2\left(\sqrt{\frac{q_i'}{q_i}}\right) dP_i \leq a_2^2\left[\mathbf{h}^2(\mathbf{P}, \mathbf{Q}) + \mathbf{h}^2(\mathbf{P}, \mathbf{Q}')\right]. \tag{18}$$

To complete the proof, note that the second inequality is a consequence of the first one by exchanging the roles of $\mathbf{Q}$ and $\mathbf{Q}'$ and using (15). ☐

5.3. **Examples of functions ψ.** Let us now introduce two functions ψ which satisfy Assumption 1.

Proposition 3. *Let ψ_1 and ψ_2 be the functions taking the value 1 at $+\infty$ and defined for $x \in \mathbb{R}_+$ by*

$$\psi_1(x) = \frac{x-1}{\sqrt{x^2+1}} \qquad \text{and} \qquad \psi_2(x) = \frac{x-1}{x+1}.$$

These two functions are continuously increasing from $[0, +\infty]$ to $[-1, 1]$ and satisfy Assumption 1 for all ρ-models $\mathscr{Q}$ with $a_0 = 4.97$, $a_1 = 0.083$, $a_2^2 = 3 + 2\sqrt{2}$ for ψ_1 and $a_0 = 4$, $a_1 = 3/8$, $a_2^2 = 3\sqrt{2}$ for ψ_2.

This proposition will be proved in Section 13.3.

6. The performance of ρ-estimators on ρ-models

The deviation $\mathbf{h}(\mathbf{P}, \widehat{\mathbf{P}})$ between the true distribution $\mathbf{P}$ and a ρ-estimator $\widehat{\mathbf{P}}$ built on some ρ-model $\mathscr{Q}$ is controlled by two terms which are the analogue of the classical bias and variance terms and we shall first introduce a function that replaces here the variance. For $y > 0$, $\mathbf{P} \in \mathscr{P}$ and $\overline{\mathbf{P}} \in \mathscr{Q}$ we define

$$\mathscr{B}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) = \left\{\mathbf{Q} \in \mathscr{Q} \;\middle|\; \mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \mathbf{h}^2(\mathbf{P}, \mathbf{Q}) \leq y^2\right\}$$

and

$$\mathbf{w}^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}},y)=\inf_{\mathcal{R}(\mathscr{Q})}\mathbb{E}\left[\sup_{\mathbf{Q}\in\mathscr{B}^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}},y)}\left|\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]\right|\right], \tag{19}$$

where the infimum runs among all possible representations $\mathcal{R}(\mathscr{Q})$ of the ρ-model $\mathscr{Q}$. We recall that we use the convention $\sup_{\varnothing}=0$. Since $\mathscr{B}^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}},y)$ is a subset of $\mathscr{Q}$, it is also countable, the supremum of $\left|\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\cdot)-\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\cdot)\right]\right|$ is therefore measurable and $\mathbf{w}^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}},y)$ is well-defined.

Definition 3. *Given ψ satisfying Assumption 1 with constants a_0,a_1 and a_2, we define the* dimension function $D^{\mathscr{Q}}$ *of the ρ-model $\mathscr{Q}$ as the function from $\mathscr{P}\times\mathscr{Q}$ to $[1,+\infty)$ given by*

$$D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}})=\left[\beta^2\sup\left\{z>0\;\middle|\;\mathbf{w}^{\mathscr{Q}}\left(\mathbf{P},\overline{\mathbf{P}},\sqrt{z}\right)>\frac{a_1z}{8}\right\}\right]\bigvee 1\quad\textit{with}\quad\beta=\frac{a_1}{4a_2}. \tag{20}$$

This function, which depends on the local fluctuation of the empirical process $\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})$ indexed by $\mathbf{q}$, is quite similar to the local Rademacher complexity introduced in Koltchinskii (2006) for the purpose of studying empirical risk minimization.

Our first theorem, to be proven at the end of this section, deals with the simplest situation of a nul penalty function.

Theorem 1. *Let $\mathbf{P}$ be an arbitrary distribution in $\mathscr{P}$, $\mathscr{Q}$ a ρ-model for $\mathbf{P}$, $\overline{\mathbf{P}}$ an arbitrary element of $\mathscr{Q}$, ψ a function satisfying Assumption 1, $D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}})$ the quantity given by (20) and* $\mathrm{pen}(\mathbf{Q})=0$ *for all $\mathbf{Q}\in\mathscr{Q}$. Then any ρ-estimator $\widehat{\mathbf{P}}$, that is any random element belonging to the closure in $(\mathscr{P},\mathbf{h})$ of the set $\mathscr{E}(\psi,\boldsymbol{X})$ defined by (11), satisfies for all $\xi>0$,*

$$\mathbb{P}\left[\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\le\gamma\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})-\mathbf{h}^2(\mathbf{P},\mathscr{Q})+\frac{4\kappa}{a_1}\left(\frac{D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}})}{4.7}+1.49+\xi\right)\right]\ge 1-e^{-\xi}, \tag{21}$$

with

$$\gamma=\frac{4(a_0+16)}{a_1}+2+\frac{168}{a_2^2};\qquad\kappa=\frac{35a_2^2}{a_1}+74,\quad\textit{hence}\quad\frac{\kappa}{25}\ge 11.36. \tag{22}$$

Introducing a non-trivial penalty function allows to favour some densities as compared to others in $\mathscr{Q}$ and gives thus a Bayesian flavour to our estimation procedure. We shall mainly use it when we have at disposal not only a single ρ-model for $\mathbf{P}$ but rather a countable collection $\{\mathscr{Q}_m,\ m\in\mathcal{M}\}$ of candidate ones in which case we set $\mathscr{Q}=\bigcup_{m\in\mathcal{M}}\mathscr{Q}_m$. Note that $\mathscr{Q}$ is still a ρ-model to which the construction of Section 4.2 applies. The penalty function may not only be used for estimating $\mathbf{P}$ but also for performing model selection among the family $\{\mathscr{Q}_m,\ m\in\mathcal{M}\}$ by deciding that the procedure selects the ρ-model $\mathscr{Q}_{\widehat{m}}$ if the resulting estimator $\widehat{\mathbf{P}}$ belongs to $\mathscr{Q}_{\widehat{m}}$. Since $\widehat{\mathbf{P}}$ may belong to several ρ-models, this selection procedure may result in a (random) set of possible ρ-models for $\mathbf{P}$ and a common way of selecting one is to choose that with the smallest *complexity* in a suitable sense. In the present paper, the complexity of a ρ-model will be measured by means of a weight function $\Delta(\cdot)$ mapping $\{\mathscr{Q}_m,\ m\in\mathcal{M}\}$ into $\mathbb{R}_+$ and which satisfies

$$\sum_{m\in\mathcal{M}}e^{-\Delta(\mathscr{Q}_m)}\le 1. \tag{23}$$

The number 1 is chosen for convenience. Note that when equality holds in (23) $e^{-\Delta(\cdot)}$ can be viewed as a prior distribution on the family of ρ-models $\{\mathscr{Q}_m,\ m\in\mathcal{M}\}$.

In such a context, we shall describe how our penalty term should depend on this weight function Δ in view of selecting a suitable ρ-model for $\mathbf{P}$. The next theorem is proved in Section 12.1.

Theorem 2. *Let $\mathbf{P}$ be an arbitrary distribution in $\mathscr{P}$, $\{\boldsymbol{\mathscr{Q}}_m,\ m\in\mathcal{M}\}$ be a countable collection of ρ-models for $\mathbf{P}$, Δ a weight function satisfying (23), $\overline{\mathbf{P}}$ an arbitrary element of $\boldsymbol{\mathscr{Q}}=\bigcup_{m\in\mathcal{M}}\boldsymbol{\mathscr{Q}}_m$, ψ a function satisfying (1) and κ a number satisfing (22). For all $m\in\mathcal{M}$, $D^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P},\overline{\mathbf{P}})$ is the dimension function of $\boldsymbol{\mathscr{Q}}_m$ defined by (20). Assume that the penalty function* pen *from $\boldsymbol{\mathscr{Q}}$ to $\mathbb{R}_+$ satisfies*

$$G(\mathbf{P},\overline{\mathbf{P}})+\mathrm{pen}(\mathbf{Q})\geq\kappa\inf_{\{m\in\mathcal{M}\,|\,\boldsymbol{\mathscr{Q}}_m\ni\mathbf{Q}\}}\left[\frac{D^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P},\overline{\mathbf{P}})}{4.7}+\Delta(\boldsymbol{\mathscr{Q}}_m)\right]\quad\textit{for all}\ \ \mathbf{Q}\in\boldsymbol{\mathscr{Q}},\tag{24}$$

for some real-valued function G on $\mathscr{P}\times\boldsymbol{\mathscr{Q}}$. Then, for all $\xi>0$, any ρ-estimator $\widehat{\mathbf{P}}$ satisfies

$$\mathbb{P}\left[\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\leq\gamma\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})-\mathbf{h}^2(\mathbf{P},\boldsymbol{\mathscr{Q}})+\frac{4}{a_1}\left[G(\mathbf{P},\overline{\mathbf{P}})+\mathrm{pen}(\overline{\mathbf{P}})+\kappa(1.49+\xi)\right]\right]\geq1-e^{-\xi},$$

with γ given by (22).

Proof of Theorem 1. Theorem 1 is obtained by applying Theorem 2 with $\{\boldsymbol{\mathscr{Q}}_m,\ m\in\mathcal{M}\}=\{\boldsymbol{\mathscr{Q}}\}$ (so that $\mathcal{M}$ reduces to a singleton), $\Delta(\boldsymbol{\mathscr{Q}})=0$, which clearly satisfies (23), and $G(\mathbf{P},\overline{\mathbf{P}})=\kappa D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}})/4.7$. Since $\mathrm{pen}(\mathbf{Q})=0$ for all $\mathbf{Q}\in\boldsymbol{\mathscr{Q}}$, equality holds in (24). The lower bound for $\kappa/25$ in (22) follows from our assumption that $a_2^2\geq6a_1$.

7. Bounding the dimension function of a ρ-model

Throughout this section we fix the function ψ satisfying Assumption 1 and the corresponding values of a_1 and a_2 and when we shall say that some quantity depends on ψ, this means that it actually depends on a_1 and a_2.

In view of the bounds that we have established in Section 6, of special interest is the situation where the dimension function $D^{\boldsymbol{\mathscr{Q}}}$ of $\boldsymbol{\mathscr{Q}}$ is uniformly bounded from above.

Corollary 1. *If*

$$D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}})\leq\overline{D}(\boldsymbol{\mathscr{Q}})\quad\textit{for all}\ \mathbf{P}\in\mathscr{P}\ \textit{and}\ \overline{\mathbf{P}}\in\boldsymbol{\mathscr{Q}},\tag{25}$$

then

$$\mathbb{P}\left[C\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\leq\mathbf{h}^2(\mathbf{P},\boldsymbol{\mathscr{Q}})+\overline{D}(\boldsymbol{\mathscr{Q}})+\xi\right]\geq1-e^{-\xi}\quad\textit{for all}\ \ \mathbf{P}\in\mathscr{P}\ \ \textit{and}\ \ \xi>0,$$

where C denotes a constant that only depends on ψ.

Proof. Choosing $\overline{\mathbf{P}}\in\boldsymbol{\mathscr{Q}}$ such that $\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})\leq\mathbf{h}^2(\mathbf{P},\boldsymbol{\mathscr{Q}})+\kappa/(25\gamma a_1)$ leads by (21) to a control of the loss $\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})$ of the form

$$\mathbb{P}\left[\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\leq(\gamma-1)\mathbf{h}^2(\mathbf{P},\boldsymbol{\mathscr{Q}})+\frac{4\kappa}{a_1}\left(\frac{\overline{D}(\boldsymbol{\mathscr{Q}})}{4.7}+1.5+\xi\right)\right]\geq1-e^{-\xi}\quad\text{for all}\ \xi>0.$$

The conclusion follows since $D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}})$, hence $\overline{D}(\boldsymbol{\mathscr{Q}})$, is bounded from below by 1. $\square$

Let us begin by some elementary considerations. We immediately derives from the definition of $D^{\mathscr{Q}}$ that

$$(26)\qquad D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}}) \le D \vee 1 \quad \text{if} \quad \mathbf{w}^{\mathscr{Q}}\left(\mathbf{P},\overline{\mathbf{P}},y\right) \le a_1 y^2/8 \quad \text{for all } y \ge \beta^{-1}\sqrt{D}.$$

In particular, since $|\psi| \le 1$, the expectation in (19) is never larger than $2n$ so that $\mathbf{w}^{\mathscr{Q}}\left(\mathbf{P},\overline{\mathbf{P}},y\right) \le a_1 y^2/8$ for $y \ge 4\sqrt{(n/a_1)}$ and (26) holds with

$$\sqrt{D} = 4\beta\sqrt{(n/a_1)} = \sqrt{na_1}/a_2 \qquad \text{or equivalently} \qquad D = na_1/a_2^2 \le n/6.$$

Therefore,

$$(27)\qquad D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}}) \le n/6 \quad \text{for all } \mathbf{P} \in \mathscr{P} \text{ and } \overline{\mathbf{P}} \in \mathscr{Q},$$

whatever the choices of ψ, a_1 and a_2.

More precise bounds will now be given that depend on some specific features of $\mathscr{Q}$.

7.1. **The finite case.** Given a finite set $\mathscr{Q} \subset \mathscr{P}$, let us set

$$(28)\qquad \mathscr{H}(\mathscr{Q},y) = \sup_{\mathbf{P}\in\mathscr{P}} \log_+\left(2\left|\mathscr{Q} \cap \mathscr{B}(\mathbf{P},y)\right|\right) \quad \text{for all } y > 0$$

and

$$(29)\qquad \overline{\eta} = \sup\left\{z > 0 \,\middle|\, \sqrt{\mathscr{H}(\mathscr{Q},z/\beta)} > z/x_0\right\} \quad \text{with} \quad x_0 = \sqrt{2}\left[\sqrt{1+(\beta/a_2)}+1\right].$$

Since $\mathscr{Q}$ is finite, the function $y \mapsto \mathscr{H}(\mathscr{Q},y)$ is bounded by $\log\left(2|\mathscr{Q}|\right)$ and since $\beta/a_2 = a_1/(4a_2^2) \le 1/24$,

$$(30)\qquad \overline{\eta} \le x_0\sqrt{\log\left(2|\mathscr{Q}|\right)} < 3\sqrt{\log\left(2|\mathscr{Q}|\right)}.$$

Proposition 4. *If the ρ-model $\mathscr{Q}$ is finite and $\overline{\eta}$ is defined by (29), then*

$$D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}}) \le \overline{D}(\mathscr{Q}) = \overline{\eta}^2 \vee 1 < 9\log\left(2|\mathscr{Q}|\right) \quad \textit{for all } \mathbf{P} \in \mathscr{P} \textit{ and } \overline{\mathbf{P}} \in \mathscr{Q}.$$

The proof of this result is given in Section 13.4. The first upper bound $\overline{\eta}^2 \vee 1$ for $D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}})$ neither depends on $\mathbf{P}$ nor on $\overline{\mathbf{P}}$ but might depend on β. The second bound only depends on the cardinality of $\mathscr{Q}$ and holds whatever ψ, a_1 and a_2.

7.2. **Bounds based on the VC-index.** In this section we investigate the case where $\mathscr{Q} = \{\mathbf{q}\cdot\boldsymbol{\mu},\ \mathbf{q} \in \boldsymbol{\mathcal{Q}}\}$ is possibly infinite. A density $\mathbf{q} = (q_1,\ldots,q_n) \in \boldsymbol{\mathcal{Q}}$ is viewed as a function on $\overline{\mathscr{X}} = \bigcup_{i=1}^n \{i\} \times \mathscr{X}_i$ defined, for $\overline{x} = (i,x)$ with $x \in \mathscr{X}_i$, by $\mathbf{q}(i,x) = q_i(x)$ so that $\boldsymbol{\mathcal{Q}}$ is now viewed as a class of functions on $\overline{\mathscr{X}}$. A common notion of dimension for the class $\boldsymbol{\mathcal{Q}}$ is the following one.

Definition 4. *A class $\mathscr{F}$ of functions from a set $\mathscr{X}$ with values in $(-\infty,+\infty]$ is VC-subgraph with index $\overline{V}$ (or equivalently with dimension $\overline{V}-1 \ge 0$) if the class of subgraphs $\{(x,u) \in \mathscr{X}\times\mathbb{R} \,|\, f(x) > u\}$ as f varies in $\mathscr{F}$ is a VC-class of sets in $\mathscr{X}\times\mathbb{R}$ with index $\overline{V}$ (or dimension $\overline{V}-1$).*

We recall that, by definition, the index $\overline{V}$ of a VC-class is a positive integer, hence its dimension $\overline{V}-1 \in \mathbb{N}$. For additional details about VC-classes and related notions, we refer to van der Vaart and Wellner (1996) and Baraud, Birgé and Sart (2016, Section 8).

The dimension function $D^{\mathscr{Q}}$ relates to the VC-index of $\boldsymbol{\mathcal{Q}}$ as follows.

Proposition 5. *Let ψ satisfy Assumption 1. If $\boldsymbol{\mathscr{Q}}$ is VC-subgraph on $\overline{\mathscr{X}}$ with index not larger than $\overline{V} \le n$, then*

$$D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P}, \overline{\mathbf{P}}) \le \overline{D}(\boldsymbol{\mathscr{Q}}) = C_1 \overline{V} \left[1 + \log_+ \left(n/\overline{V}\right)\right] \quad \textit{for all } \mathbf{P} \in \mathscr{P} \textit{ and } \overline{\mathbf{P}} \in \boldsymbol{\mathscr{Q}},$$

where C_1 is a universal constant.

The proof is given in Section 13.5. A nice feature of this bound lies in the fact that it neither depends on the choices of ψ, a_1 and a_2 nor on the cardinality of $\boldsymbol{\mathscr{Q}}$ which can therefore be arbitrarily large. In the particular case of density estimation the following result is useful in view of applying Proposition 5.

Proposition 6. *If $\mathcal{Q}$ is a density ρ-model for P which is VC-subgraph on $\mathscr{X}$ with index $\overline{V}$, then $\boldsymbol{\mathcal{Q}} = \{\mathbf{q} = (q, \ldots, q),\ q \in \mathcal{Q}\}$ is a density ρ-model for $\mathbf{P}$ which is VC-subgraph on $\mathscr{X}$ with index not larger than $\overline{V}$.*

Proof. If the class of subgraphs $\left\{(\overline{x}, u) \in \overline{\mathscr{X}} \times \mathbb{R} \,\middle|\, \mathbf{q}(\overline{x}) > u\right\}$, with $\mathbf{q}$ running in $\boldsymbol{\mathcal{Q}}$, shatters the subset $\{(\overline{x}_1, u_1), \ldots, (\overline{x}_k, u_k)\}$ of $\overline{\mathscr{X}} \times \mathbb{R}$, then, whatever $J \subset \{1, \ldots, k\}$, there exists $\mathbf{q} \in \boldsymbol{\mathcal{Q}}$ such that

$$j \in J \quad \Longleftrightarrow \quad \mathbf{q}(\overline{x}_j) = q(x_j) > u_j.$$

Hence, the class of subgraphs $\{(x, u) \in \mathscr{X} \times \mathbb{R} \,|\, q(x) > u\}$ with q running in $\mathcal{Q}$ shatters the subset $\{(x_1, u_1), \ldots, (x_k, u_k)\}$ of $\mathscr{X} \times \mathbb{R}$ and therefore $k + 1 \le \overline{V}$. □

Following Baraud (2016), we introduce the following definition.

Definition 5. *A class of functions $\mathscr{F}$ on a set $\mathscr{X}$ with values in $[-\infty, +\infty]$ is said to be weak VC-major with dimension not larger than $k \in \mathbb{N}$ if for all $u \in \mathbb{R}$, the class of subsets*

$$\mathscr{C}_u(\mathscr{F}) = \left\{\{x \in \mathscr{X} \,|\, f(x) > u\},\ f \in \mathscr{F}\right\}$$

is a VC-class with dimension not larger than k (index not larger than $k + 1$). The weak dimension of $\mathscr{F}$ is the smallest of such integers k.

Definition 6. *Let $\mathscr{F}$ be a class of real-valued functions on $\mathscr{X}$. We shall say that an element $\overline{f} \in \mathscr{F}$ is extremal in $\mathscr{F}$ with degree $d \in \mathbb{N}$ if the class of functions*

$$\left(\mathscr{F}/\overline{f}\right) = \left\{f/\overline{f},\ f \in \mathscr{F}\right\}$$

is weak VC-major with dimension d.

When $\overline{\mathbf{p}}$ is extremal in $\boldsymbol{\mathcal{Q}}$ the bound we get on the quantity $D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P}, \overline{\mathbf{P}})$ depends on $\overline{\mathbf{P}} = \overline{\mathbf{p}} \cdot \boldsymbol{\mu}$.

Proposition 7. *Let ψ satisfy Assumption 1 and assume that $\overline{\mathbf{p}}$ is extremal in $\boldsymbol{\mathcal{Q}}$ with degree not larger than $d \le n$. Then*

$$D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P}, \overline{\mathbf{P}}) \le 33d \left[\log\left(e^2 n/d\right)\right]^3 \quad \textit{for all } \mathbf{P} \in \mathscr{P}.$$

The proof is given in Section 13.6. This upper bound, though depending on the specific features of $\overline{\mathbf{P}}$, is free from the choices of ψ, a_1 and a_2.

In the particular case of density estimation, the following result turns to be useful in view of applying Proposition 7.

Proposition 8. *Let $\mathcal{Q}$ be a density ρ-model for P. If $\overline{p}$ is extremal in $\mathcal{Q}$ with degree d, $\overline{\mathbf{p}} = (\overline{p}, \ldots, \overline{p})$ is extremal in $\boldsymbol{\mathcal{Q}} = \{\mathbf{q} = (q, \ldots, q),\ q \in \mathcal{Q}\}$ with degree not larger than d.*

Proof. Let $u \in \mathbb{R}$. If $\mathscr{C}_u\big((\boldsymbol{\mathcal{Q}}/\overline{\mathbf{p}})\big)$ shatters $\{\overline{x}_1, \ldots, \overline{x}_k\} \subset \overline{\mathscr{X}}$, for all $J \subset \{1, \ldots, k\}$ there exists $\mathbf{q} \in \boldsymbol{\mathcal{Q}}$ such that

$$j \in J \quad \Longleftrightarrow \quad \frac{\mathbf{q}}{\overline{\mathbf{p}}}(\overline{x}_j) = \frac{q}{\overline{p}}(x_j) > u.$$

Hence $\mathscr{C}_u\big((\mathcal{Q}/\overline{p})\big)$ shatters $\{x_1, \ldots, x_k\} \subset \mathscr{X}$ which is only possible for $k \le d$. $\square$

7.3. **Some illustrations.** In this section, we consider the situation where $X_1, \ldots, X_n$ are i.i.d. with unknown distribution $P \in \mathscr{P}$ on $(\mathscr{X}, \mathscr{B})$ and $\mathscr{Q}$ is a ρ-model for P. Our aim is to give some examples of density ρ-models $\mathcal{Q}$ to which Propositions 5 or 7 apply.

Piecewise constant functions. Let k be a positive integer and $\mathscr{X}$ an arbitrary interval of $\mathbb{R}$ (possibly $\mathscr{X} = \mathbb{R}$). We define $\mathscr{F}_k$ as the class of functions f on $\mathscr{X}$ such that there exists a partition $\mathscr{I}(f)$ of $\mathscr{X}$ into at most k intervals and f is constant on each of these intervals. Note that $\mathscr{I}(f)$ depends on f. The following result is to be proved in Section 13.7.

Proposition 9. *The set $\mathscr{F}_k$ is VC-subgraph with dimension bounded by $2k$.*

Let us apply this to histogram estimation on $\mathscr{X} = \mathbb{R}$. For a positive integer D we denote by $\overline{\mathcal{Q}}_D$ the subset of $\mathscr{F}_{D+2}$ of densities with respect to the Lebesgue measure μ, that is the set of piecewise constant densities on $\mathbb{R}$ with at most D pieces. Let the true distribution P have a density s which belongs to $\overline{\mathcal{Q}}_D$ and let $\widehat{P}$ be the ρ-estimator of P built on a countable and dense subset $\mathscr{Q}$ of $\overline{\mathscr{Q}} = \{q \cdot \mu,\ q \in \overline{\mathcal{Q}}_D\}$. We derive from Propositions 5 and 9 that for some universal constant $C > 0$,

$$D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P}, \overline{\mathbf{P}}) \le CD\left[1 + \log_+(n/D)\right] \quad \text{for all} \quad \overline{\mathbf{P}} \in \boldsymbol{\mathscr{Q}}.$$

The logarithmic factor in this bounds turns out be necessary. The argument is as follows. Since $\mathbf{P} \in \overline{\boldsymbol{\mathscr{Q}}}$, Corollary 2 implies that

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}})\right] \le C'D\left[1 + \log_+(n/D)\right] \tag{31}$$

for some universal constant $C' > 0$. This inequality appears to be optimal (up to the numerical constant C') in view of the lower bound established in Proposition 2 of Birgé and Massart (1998). It also shows that the logarithmic factor involved in the bound of the dimension function established in Proposition 5 is necessary, at least for some VC-subgraph classes.

Piecewise exponential families. We start with the following definition.

Definition 7. *Let $g_1, \ldots, g_J$ be $J \ge 1$ real-valued functions on a set $\mathscr{X}$. We shall say that a class $\mathscr{F}$ of positive functions on $\mathscr{X}$ is an exponential family based on $g_1, \ldots, g_J$ if the elements f of $\mathscr{F}$ are of the form*

$$f = \exp\left[\sum_{j=1}^{J} \beta_j g_j\right] \quad \textit{for} \quad \beta_1, \ldots, \beta_J \in \mathbb{R}. \tag{32}$$

If $\mathscr{X}$ is a nontrivial interval of $\mathbb{R}$ and k a positive integer, we shall say that $\mathscr{F}$ is a k-piecewise exponential family based on $g_1, \ldots, g_J$ if for all $f \in \mathscr{F}$ there exists a partition $\mathscr{I}(f)$ of $\mathscr{X}$ into at most k intervals such that for all $I \in \mathscr{I}(f)$, the restriction f_I of f to I is of the form (32).

The properties of exponential and piecewise exponential families are described by the following proposition to be proven in Section 13.8.

Proposition 10. *Let $\mathcal{Q}$ be a class of functions on $\mathscr{X}$.*

i) If $\mathcal{Q}$ is an exponential family based on $J \geq 1$ functions, $\mathcal{Q}$ is VC-subgraph with index not larger than $J+2$.
ii) Let $\mathscr{I}$ be a partition of $\mathscr{X}$ with cardinality not larger than $k \geq 1$. If for all $I \in \mathscr{I}$ the family $\mathcal{Q}_I$ consisting of the restrictions of the functions q in $\mathcal{Q}$ to the set I is an exponential family on I based on $J \geq 1$ functions, $\mathcal{Q}$ is VC-subgraph with index not larger than $k(J+2)$.
iii) If $\mathscr{X}$ is a non-trivial interval of $\mathbb{R}$ and $\mathcal{Q}$ is a k-piecewise exponential family based on J functions, all densities $\overline{p} \in \mathcal{Q}$ are extremal in $\mathcal{Q}$ with degree d not larger than $\lceil 9.4k(J+2) \rceil$, which is the smallest integer j not smaller than $9.4k(J+2)$.

8. The performance of ρ-estimators on usual statistical models

As already mentioned in Section 2.2, classical statistical models are sets $\overline{\mathscr{Q}}$ of probabilities which are typically separable but uncountable and cannot consequently be used as ρ-models. It is therefore important, given some statistical model $\overline{\mathscr{Q}}$ to find how to choose a suitable ρ-model to replace it and to understand how the performance of the resulting ρ-estimator is connected to the properties of $\overline{\mathscr{Q}}$. As we do for ρ-models, we associate to the separable, hence dominated set of probabilities $\overline{\mathscr{Q}}$ a representation by some dominating measure $\boldsymbol{\mu}$ and a family of densities $\overline{\boldsymbol{\mathcal{Q}}}$ so that $\overline{\mathscr{Q}} = \left\{\mathbf{q}\cdot\boldsymbol{\mu},\, \mathbf{q} \in \overline{\boldsymbol{\mathcal{Q}}}\right\}$.

8.1. Statistical models which are VC-subgraph classes.

When the family of densities $\overline{\boldsymbol{\mathcal{Q}}}$ is VC-subgraph with index not larger than $\overline{V}$, one can choose for $\mathscr{Q}$ an arbitrary countable and dense subset of $\overline{\mathscr{Q}}$ and for $\boldsymbol{\mathcal{Q}}$ the corresponding subset of $\overline{\boldsymbol{\mathcal{Q}}}$. In this case $\mathbf{h}^2(\mathbf{P},\mathscr{Q}) = \mathbf{h}^2\big(\mathbf{P},\overline{\mathscr{Q}}\big)$ since $\mathscr{Q}$ is dense in $\overline{\mathscr{Q}}$ and $\boldsymbol{\mathcal{Q}}$ is also VC-subgraph with index not larger than $\overline{V}$. We therefore deduce from Proposition 5 and Corollary 1 the following result.

Corollary 2. *Under the assumptions of Theorem 1, if $\overline{\boldsymbol{\mathcal{Q}}}$ is VC-subgraph with index $\overline{V}$, any ρ-estimator $\widehat{\mathbf{P}}$ based on a countable and dense subset $\mathscr{Q}$ of $\overline{\mathscr{Q}}$ (with respect to $\mathbf{h}$) satisfies for all $\mathbf{P} \in \mathscr{P}$ and $\xi > 0$,*

$$\mathbb{P}\left[C\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}}) \leq \mathbf{h}^2\left(\mathbf{P},\overline{\mathscr{Q}}\right) + \overline{V}\left[1+\log_+\big(n/\overline{V}\big)\right] + \xi\right] \geq 1 - e^{-\xi}, \tag{33}$$

where the constant C only depends on the choice of ψ.

Instead of a countable and dense subset of $\overline{\mathscr{Q}}$, we could take for $\mathscr{Q}$ some subset of $\overline{\mathscr{Q}}$ which is an η-net with $\eta \leq \overline{V}$. We would then still get (33) with a different constant C.

8.2. Totally bounded statistical models.

If the statistical model $\overline{\mathscr{Q}}$ is a totally bounded subset of the metric space $(\mathscr{P},\mathbf{h})$ and $\eta > 0$, one can cover $\overline{\mathscr{Q}}$ by a finite number of closed balls of radius η and the set $\mathscr{Q}[\eta]$ of their centers is an η-net for $\overline{\mathscr{Q}}$, which means that $\mathbf{h}(\mathbf{Q},\mathscr{Q}[\eta]) \leq \eta$ for all $\mathbf{Q} \in \overline{\mathscr{Q}}$. The function $y \mapsto \mathscr{H}(\mathscr{Q}[\eta], y)$ measures in a sense the massiveness of $\mathscr{Q}[\eta]$ and turns out to be a useful tool to measure that of $\overline{\mathscr{Q}}$. We shall in particular use the following classical notions of dimension based on the metric structure of $\overline{\mathscr{Q}}$.

Definition 8. *A set* $\overline{\mathscr{Q}} \subset \mathscr{P}$ *is said to have a metric dimension bounded by* $\widetilde{D}$*, where* $\widetilde{D}$ *is a right-continuous function from* $(0,+\infty)$ *to* $[1/2,+\infty]$*, if, for all* $\eta > 0$*, there exists an* η*-net* $\mathscr{Q}[\eta]$ *for* $\overline{\mathscr{Q}}$ *which satisfies*

$$\mathscr{H}(\mathscr{Q}[\eta], y) \le (y/\eta)^2 \widetilde{D}(\eta) \quad \text{for all } y \ge 2\eta. \tag{34}$$

We shall say that $\overline{\mathscr{Q}}$ *has an entropy dimension bounded by* $V \ge 0$ *if, for all* $\eta > 0$*, there exists an* η*-net* $\mathscr{Q}[\eta]$ *of* $\overline{\mathscr{Q}}$ *such that*

$$\mathscr{H}(\mathscr{Q}[\eta], y) \le V \log(y/\eta) \quad \text{for all} \quad y \ge 2\eta. \tag{35}$$

For the sake of convenience, we have slightly modified the original definition of the metric dimension due to Birgé (2006) (Definition 6 p. 293) which is actually obtained by replacing the left-hand side of (34) by $\mathscr{H}(\mathscr{Q}[\eta], y) - \log 2$. Since in both definitions the metric dimension is not smaller than $1/2$, it is easy to check that if $\overline{\mathscr{Q}}$ has a metric dimension bounded by D_M in Birgé's sense it has a metric dimension bounded by $\widetilde{D} = (1 + (\log 2)/2)D_M$ in our sense and conversely if $\overline{\mathscr{Q}}$ has a metric dimension bounded by $\widetilde{D}$ in our sense it also has a metric dimension bounded by $\widetilde{D}$ in Birgé's sense. Hence, changing $\widetilde{D}$ into D_M only changes the numerical constants.

The logarithm being a slowly varying function, it is not difficult to see that the notion of metric dimension is more general than the entropy one in the sense that if $\overline{\mathscr{Q}}$ has an entropy dimension bounded by some V, then it also has a metric dimension bounded by $\widetilde{D}(\cdot)$ with

$$\widetilde{D}(\eta) \le (1/2) \vee [V(\log 2)/4] \quad \text{for all } \eta > 0. \tag{36}$$

If $\overline{\mathscr{Q}}$ has a metric dimension bounded by $\widetilde{D}$ and if η is a positive number satisfying

$$\widetilde{D}(\eta) \le (\beta\eta/x_0)^2, \tag{37}$$

with x_0 given by (29), we deduce from (34) there exists an η-net $\mathscr{Q} = \mathscr{Q}[\eta]$ for $\overline{\mathscr{Q}}$ for which

$$\sqrt{\mathscr{H}(\mathscr{Q}[\eta], z/\beta)} \le z/x_0 \quad \text{for all } z \ge 2\beta\eta.$$

It then follows that $\overline{\eta}$, as defined in (29), satisfies $\overline{\eta} \le 2\beta\eta$ and we deduce from Proposition 4 that the dimension function $D^{\mathscr{Q}}$ satisfies

$$D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}) \le \overline{D}(\mathscr{Q}) = (2\beta\eta)^2 \vee 1 \quad \text{for all } \mathbf{P} \in \mathscr{P} \text{ and } \overline{\mathbf{P}} \in \mathscr{Q}. \tag{38}$$

If, in particular, $\overline{\mathscr{Q}}$ has an entropy dimension bounded by $V \ge 0$ we deduce from (36) that (37) holds for

$$\eta^2 = \frac{x_0^2}{2\beta^2}\left(1 \bigvee \frac{V\log 2}{2}\right) < \frac{9}{2\beta^2}\left(1 \bigvee \frac{V\log 2}{2}\right) \tag{39}$$

and we derive from (38) that for $\mathscr{Q} = \mathscr{Q}[\eta]$

$$D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}) \le \overline{D}(\mathscr{Q}) = 18\left(1 \bigvee \frac{V\log 2}{2}\right) \quad \text{for all } \mathbf{P} \in \mathscr{P} \text{ and } \overline{\mathbf{P}} \in \mathscr{Q}. \tag{40}$$

Since in both cases $\mathbf{h}(\mathbf{P}, \mathscr{Q}) \le \mathbf{h}(\mathbf{P}, \overline{\mathscr{Q}}) + \eta$ for all $\mathbf{P} \in \mathscr{P}$ since $\mathscr{Q}$ is an η-net for $\overline{\mathscr{Q}}$, it follows from (38), (40) and Corollary 1 that

Corollary 3. *Let the assumptions of Theorem 1 hold.*

i) If $\overline{\mathscr{Q}}$ has a metric dimension bounded by $\widetilde{D}$ and η satisfies (37), any ρ-estimator $\widehat{\mathbf{P}}$ based on a suitable η-net $\mathscr{Q}$ for $\overline{\mathscr{Q}}$ satisfies

$$\mathbb{P}\left[C\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}}) \le \mathbf{h}^2(\mathbf{P},\overline{\mathscr{Q}}) + \left(\eta^2 \vee 1\right) + \xi\right] \ge 1 - e^{-\xi} \quad \textit{for all } \mathbf{P} \in \mathscr{P} \textit{ and } \xi > 0. \tag{41}$$

ii) If $\overline{\mathscr{Q}}$ has an entropy dimension bounded by V and η satisfies (39), any ρ-estimator $\widehat{\mathbf{P}}$ based on a suitable η-net $\mathscr{Q}$ for $\overline{\mathscr{Q}}$ satisfies

$$\mathbb{P}\left[C\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}}) \le \mathbf{h}^2\left(\mathbf{P},\overline{\mathscr{Q}}\right) + (V \vee 1) + \xi\right] \ge 1 - e^{-\xi} \quad \textit{for all } \mathbf{P} \in \mathscr{P} \textit{ and } \xi > 0. \tag{42}$$

In both cases, C is a constant depending only on the choice of ψ.

8.3. **Bounding the minimax risk over a statistical model $\overline{\mathscr{Q}}$.** As we have seen in the previous sections — see (33), (41) or (42) —, there are various situations for which, starting from a statistical model $\overline{\mathscr{Q}}$, we can find a ρ-model $\mathscr{Q}$ such that (25) holds with $\overline{D}(\mathscr{Q}) = \overline{D}\left(\overline{\mathscr{Q}}\right)$ where the quantity $\overline{D}\left(\overline{\mathscr{Q}}\right)$ only depends on the properties of $\overline{\mathscr{Q}}$ and the number of observations n. In such a case, Corollary 1 implies that any ρ-estimator $\widehat{\mathbf{P}}$ built on $\mathscr{Q}$ satisfies

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\right] \le C\left[\mathbf{h}^2(\mathbf{P},\overline{\mathscr{Q}}) + \overline{D}(\overline{\mathscr{Q}})\right] \quad \text{for all } \mathbf{P} \in \mathscr{P},$$

where the constant C only depends on ψ. In particular, the minimax risk (with respect to the squared Hellinger loss) over the statistical model $\overline{\mathscr{Q}}$ is bounded by $C\overline{D}\left(\overline{\mathscr{Q}}\right)$ since

$$\inf_{\widetilde{\mathbf{P}}} \sup_{\mathbf{P}\in\overline{\mathscr{Q}}} \mathbb{E}\left[\mathbf{h}^2(\mathbf{P},\widetilde{\mathbf{P}})\right] \le \sup_{\mathbf{P}\in\overline{\mathscr{Q}}} \mathbb{E}\left[\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\right] \le C\overline{D}\left(\overline{\mathscr{Q}}\right),$$

where the infimum runs over all possible estimators $\widetilde{\mathbf{P}}$.

Unfortunately, there are also situations for which one is unable to bound the dimension function $D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}})$ for some well-chosen $\mathscr{Q}$, uniformly with respect to $\mathbf{P}$ and $\overline{\mathbf{P}}$. This is the case when the statistical model $\overline{\mathscr{Q}}$ has not a finite metric dimension and, viewed as a set of densities, is not VC-subgraph. One such example is the set of probabilities on the real line with densities that are unimodal on some arbitrary compact interval. This set can nevertheless been analyzed via Proposition 7 as shown in Baraud and Birgé (2016). Another illustration is provided in Baraud, Birgé and Sart (2016, Section 6.5). In both cases, we were nevertheless able to bound satisfactorily the risk of ρ-estimators without bounding $D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}})$ uniformly.

Indeed, if $\overline{\mathscr{Q}}$ is a statistical model, $\mathscr{Q}$ an arbitrary ρ-model and $\widehat{\mathbf{P}}(\mathscr{Q})$ is a ρ-estimator built on $\mathscr{Q}$ it follows from (21) that the maximal risk of $\widehat{\mathbf{P}}(\mathscr{Q})$ over $\overline{\mathscr{Q}}$ is bounded by

$$\sup_{\mathbf{P}\in\overline{\mathscr{Q}}} \mathbb{E}\left[\mathbf{h}^2\Big(\mathbf{P},\widehat{\mathbf{P}}(\mathscr{Q})\Big)\right] \le C\mathscr{R}(\mathscr{Q}) \quad \text{with} \quad \mathscr{R}(\mathscr{Q}) = \sup_{\mathbf{P}\in\overline{\mathscr{Q}}} \inf_{\overline{\mathbf{P}}\in\mathscr{Q}} \left[\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}}) + D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}})\right].$$

We see from this bound that the quality of the ρ-model $\mathscr{Q}$ for estimating an arbitrary $\mathbf{P} \in \overline{\mathscr{Q}}$ using a ρ-estimator over $\mathscr{Q}$ can be characterized by the quantity $\mathscr{R}(\mathscr{Q})$. The smaller $\mathscr{R}(\mathscr{Q})$, the better the ρ-model $\mathscr{Q}$. If we assume that $\mathbf{P} \in \overline{\mathscr{Q}}$, we should use, for the construction of a ρ-estimator, some ρ-model which minimizes, at least approximately, the function $\mathscr{R}$ with respect to all possible ρ-models. Indeed, if

$$\mathscr{D}\left(\overline{\boldsymbol{\mathscr{Q}}}\right)=\inf_{\boldsymbol{\mathscr{Q}}}\mathscr{R}(\boldsymbol{\mathscr{Q}})=\inf_{\boldsymbol{\mathscr{Q}}}\sup_{\mathbf{P}\in\overline{\boldsymbol{\mathscr{Q}}}}\inf_{\overline{\mathbf{P}}\in\boldsymbol{\mathscr{Q}}}\left[\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})+D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}})\right], \tag{43}$$

where the infimum $\inf_{\boldsymbol{\mathscr{Q}}}$ is over all possible ρ-models $\boldsymbol{\mathscr{Q}}$, and if the ρ-model $\boldsymbol{\mathscr{Q}}'$ satisfies $\mathscr{R}(\boldsymbol{\mathscr{Q}}')\le c\mathscr{D}\left(\overline{\boldsymbol{\mathscr{Q}}}\right)$, then

$$\sup_{\mathbf{P}\in\overline{\boldsymbol{\mathscr{Q}}}}\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{P},\widehat{\mathbf{P}}(\boldsymbol{\mathscr{Q}}')\right)\right]\le Cc\mathscr{D}\left(\overline{\boldsymbol{\mathscr{Q}}}\right). \tag{44}$$

The quantity $\mathscr{D}\left(\overline{\boldsymbol{\mathscr{Q}}}\right)$ given by (43), which we shall call the *ρ-dimension of the statistical model $\overline{\boldsymbol{\mathscr{Q}}}$*, is a characteristic of this set that reflects the difficulty of ρ-estimation on $\overline{\boldsymbol{\mathscr{Q}}}$ as shown by (44) and provides an upper bound (up to some multiplicative constant depending on our choice of ψ) for the minimax risk over $\overline{\boldsymbol{\mathscr{Q}}}$ with respect to the quadratic Hellinger loss. If, in particular, (25) holds with $\overline{D}(\boldsymbol{\mathscr{Q}})=\overline{D}\left(\overline{\boldsymbol{\mathscr{Q}}}\right)$, then $\mathscr{D}\left(\overline{\boldsymbol{\mathscr{Q}}}\right)\le\overline{D}\left(\overline{\boldsymbol{\mathscr{Q}}}\right)$. As shown by Corollaries 2 and 3, this property holds when $\overline{\boldsymbol{\mathscr{Q}}}$ is VC-subgraph or is totally bounded. In all situations we know, $\mathscr{D}\left(\overline{\boldsymbol{\mathscr{Q}}}\right)$ corresponds, up to possible logarithmic and constant factors, to the minimax risk over $\overline{\boldsymbol{\mathscr{Q}}}$.

The choice of a suitable ρ-model $\boldsymbol{\mathscr{Q}}$ for $\overline{\boldsymbol{\mathscr{Q}}}$ is not always straightforward and even in the simple situation where $\overline{\boldsymbol{\mathscr{Q}}}$ is already countable, nothing warrants that the choice $\boldsymbol{\mathscr{Q}}=\overline{\boldsymbol{\mathscr{Q}}}$ be optimal. Actually, it may happen that a ρ-estimator built on a suitably discretized version $\boldsymbol{\mathscr{Q}}$ of $\overline{\boldsymbol{\mathscr{Q}}}$ possesses a much smaller risk that that built on the initial ρ-model $\overline{\boldsymbol{\mathscr{Q}}}$.

8.4. **The case of several models.** When we have at disposal of collection $\{\overline{\boldsymbol{\mathscr{Q}}}_m,\, m\in\mathcal{M}\}$ of classical statistical models endowed with a weight function $\overline{\Delta}\ge 0$ satisfying

$$\sum_{m\in\mathcal{M}}e^{-\overline{\Delta}(\overline{\boldsymbol{\mathscr{Q}}}_m)}\le 1,$$

one may associate to each $\overline{\boldsymbol{\mathscr{Q}}}_m$ a suitable ρ-model $\boldsymbol{\mathscr{Q}}_m$ as we did before and endow the resulting collection of ρ-models $\{\boldsymbol{\mathscr{Q}}_m,\, m\in\mathcal{M}\}$ with a weight function Δ defined by $\Delta(\boldsymbol{\mathscr{Q}}_m)=\overline{\Delta}(\overline{\boldsymbol{\mathscr{Q}}}_m)$ for all $m\in\mathcal{M}$, which therefore satisfies (23). If for all $m\in\mathcal{M}$ the dimension function $D^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P},\overline{\mathbf{P}})$ can be bounded from above by some number $\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m)$ which is independent of $(\mathbf{P},\overline{\mathbf{P}})\in\mathscr{P}\times\boldsymbol{\mathscr{Q}}$ with $\boldsymbol{\mathscr{Q}}=\bigcup_{m\in\mathcal{M}}\boldsymbol{\mathscr{Q}}_m$ and only depends on n and the specific features of $\overline{\boldsymbol{\mathscr{Q}}}_m$, we may apply Theorem 2 with $G(\mathbf{P},\overline{\mathbf{P}})=0$ for all $(\mathbf{P},\overline{\mathbf{P}})\in\mathscr{P}\times\boldsymbol{\mathscr{Q}}$,

$$\operatorname{pen}(\mathbf{Q})=\kappa\inf_{\{m\in\mathcal{M}\,|\,\boldsymbol{\mathscr{Q}}_m\ni\mathbf{Q}\}}\left[\frac{\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m)}{4.7}+\overline{\Delta}(\overline{\boldsymbol{\mathscr{Q}}}_m)\right]\quad\text{for all }\mathbf{Q}\in\boldsymbol{\mathscr{Q}}, \tag{45}$$

and $\overline{\mathbf{P}}\in\boldsymbol{\mathscr{Q}}$ such that

$$\gamma\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})+\frac{4}{a_1}\operatorname{pen}(\overline{\mathbf{P}})\le\inf_{m\in\mathcal{M}}\left[\gamma\mathbf{h}^2(\mathbf{P},\boldsymbol{\mathscr{Q}}_m)+\frac{4\kappa}{a_1}\left(\frac{\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m)}{4.7}+\overline{\Delta}(\overline{\boldsymbol{\mathscr{Q}}}_m)\right)\right]+\frac{\kappa}{25a_1}.$$

We derive that $\widehat{\mathbf{P}}$ satisfies, for all $\mathbf{P}\in\mathscr{P}$ and all $\xi>0$,

$$\mathbb{P}\left[\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\le\inf_{m\in\mathcal{M}}\left(\gamma\mathbf{h}^2(\mathbf{P},\boldsymbol{\mathscr{Q}}_m)+\frac{4\kappa}{a_1}\left[\frac{\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m)}{4.7}+\overline{\Delta}(\overline{\boldsymbol{\mathscr{Q}}}_m)+1.5+\xi\right]\right)\right]\ge 1-e^{-\xi}. \tag{46}$$

Since $\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m)\ge 1$ for all m, it follows that

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\right]\le C\inf_{m\in\mathcal{M}}\left[\mathbf{h}^2(\mathbf{P},\boldsymbol{\mathscr{Q}}_m)+\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m)+\overline{\Delta}(\overline{\boldsymbol{\mathscr{Q}}}_m)\right].$$

If, in particular, for some universal constant $c > 0$ and all $m \in \mathcal{M}$ the ρ-model $\boldsymbol{\mathscr{Q}}_m$ is a $c[\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m)]^{1/2}$-net for $\overline{\boldsymbol{\mathscr{Q}}}_m$, we deduce that

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{P},\widehat{\mathbf{P}})\right] \le C \inf_{m\in\mathcal{M}} \left[2\mathbf{h}^2(\mathbf{P},\overline{\boldsymbol{\mathscr{Q}}}_m) + (1+2c^2)\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m) + \overline{\Delta}(\overline{\boldsymbol{\mathscr{Q}}}_m)\right].$$

This shows that the penalized ρ-estimator $\widehat{\mathbf{P}}$ achieves, up to a positive multiplicative factor, the best trade-off between the approximation term $\mathbf{h}^2(\mathbf{P},\overline{\boldsymbol{\mathscr{Q}}}_m)$ and the complexity term $\overline{D}(\overline{\boldsymbol{\mathscr{Q}}}_m) + \overline{\Delta}(\overline{\boldsymbol{\mathscr{Q}}}_m)$ over the collection $\{\overline{\boldsymbol{\mathscr{Q}}}_m,\, m \in \mathcal{M}\}$.

9. Estimating a conditional distribution

9.1. **Description of the framework.** Let us now apply our result to the estimation of a conditional distribution. We consider i.i.d. pairs $X_i = (W_i, Y_i)$, $i = 1,\dots,n$ of random variables with values in the product space $(\mathscr{W}\times\mathscr{Y}, \mathscr{B}(\mathscr{W})\otimes\mathscr{B}(\mathscr{Y}))$ and common distribution P. We denote by P_W the marginal distribution of W and assume the existence of a conditional distribution P_w of Y when $W = w$ which means that for all bounded measurable functions f on $\mathscr{Y}$,

$$\mathbb{E}[f(Y)\,|\,W=w] = \int f(y)\,dP_w(y) \quad P_W\text{-a.s.}$$

and for all bounded measurable functions g on $\mathscr{W}\times\mathscr{Y}$,

$$\mathbb{E}[g(W,Y)] = \int g(w,y)\,dP_w(y)\,dP_W(w).$$

Our purpose is to estimate the conditional distribution P_w without the knowledge of P_W which may therefore be completely arbitrary.

In order to do this we consider a reference measure λ on $(\mathscr{Y},\mathscr{B}(\mathscr{Y}))$ and the set $\mathcal{L}_c(\mathscr{W},\lambda)$ of conditional densities with respect to λ, that is the set of measurable functions t from $(\mathscr{W}\times\mathscr{Y}, \mathscr{B}(\mathscr{W})\otimes\mathscr{B}(\mathscr{Y}))$ to $\mathbb{R}_+$ such that for all $w \in \mathscr{W}$, the function $t_w : y \mapsto t(w,y) \in \mathcal{L}(\lambda)$. Then, to each element $t \in \mathcal{L}_c(\mathscr{W},\lambda)$ is associated a conditional distribution $t_w\cdot\lambda$ for Y. In order to build our estimators we first introduce a countable family $\mathbb{S} = \{S_m,\ m \in \mathcal{M}\}$ of ρ-models of conditional densities, that is of countable subsets of $\mathcal{L}_c(\mathscr{W},\lambda)$, together with a weight function $\Delta : \mathbb{S} \mapsto \mathbb{R}_+$ satisfying

$$\sum_{m\in\mathcal{M}} e^{-\Delta(S_m)} \le 1. \tag{47}$$

To each S_m, we associate a probability ρ-model $\mathscr{Q}_m = \{Q_t,\, t \in S_m\}$ for P, where the probability Q_t on $\mathscr{W}\times\mathscr{Y}$ is given by

$$Q_t(A\times B) = \int_A \left[\int_B t_w(y)\,d\lambda(y)\right] dP_W(w), \quad \text{or equivalently} \quad \frac{dQ_t}{dP_W\otimes d\lambda}(w,y) = t(w,y).$$

This means that Q_t has a marginal distribution P_W on $\mathscr{W}$ and a conditional distribution given $W = w$ with density t_w with respect to λ. Note that the ρ-models $\mathscr{Q}_m$ depend on the unknown distribution P_W but the densities with respect to the dominating measure $P_W\otimes\lambda$ do not. If we set $\Delta(\mathscr{Q}_m) = \Delta(S_m)$ for all $m \in \mathcal{M}$ and introduce a suitable penalty pen on $\mathscr{Q} = \cup_{m\in\mathcal{M}}\mathscr{Q}_m$, we may build a ρ-estimator of P from our sample $X_1,\dots,X_n$ according to the recipe of Section 4.2 since its values only depend on the family of densities in $\mathcal{Q} = \cup_{m\in\mathcal{M}} S_m$. As a consequence, our estimation strategy neither needs to know P_W nor to estimate it. Such a ρ-estimator will be of the form $Q_{\widehat{s}} = \widehat{s}\cdot(P_W\otimes\lambda)$ and provides an estimator $\widehat{s}_w\cdot\lambda$ of the conditional probability P_w.

Within this framework, the Hellinger distance between the probabilities at hand writes, for any measure ν that dominates both P and $P_W \otimes \lambda$,

$$\begin{aligned} h^2(Q_t, P) &= \frac{1}{2}\int_{\mathscr{W}\times\mathscr{Y}} \left(\sqrt{t(w,y)\frac{dP_W(w)d\lambda(y)}{d\nu(w,y)}} - \sqrt{\frac{dP_W(w)dP_w(y)}{d\nu(w,y)}}\right)^2 d\nu(w,y) \\ &= \int_{\mathscr{W}} h^2(t_w \cdot \lambda, P_w)\, dP_W(w). \end{aligned}$$

Therefore

$$h^2(P, \mathscr{Q}_m) = \inf_{t \in S_m} \int_{\mathscr{W}} h^2(t_w \cdot \lambda, P_w)\, dP_W(w). \tag{48}$$

Note that $h^2(Q_t, P)$ can actually be viewed as a loss function for the conditional distributions, of the form $\ell(t_w \cdot \lambda, P_w)$ since it actually only depends on t_w and P_w.

9.2. **Assumptions and results.** Let us assume the following:

Assumption 2. *For all $m \in \mathcal{M}$, S_m is VC-subgraph with index not larger $\overline{V}_m$.*

We may then deduce from Theorem 2 the following result.

Corollary 4. *Let $\mathbb{S} = \{S_m,\ m \in \mathcal{M}\}$ be a family of ρ-models satisfying Assumption 2, Δ a weight function on $\mathbb{S}$ which satisfies (47), ψ a function satisfying Assumption 1, $\mathscr{Q} = \cup_{m\in\mathcal{M}}\{Q_t,\ t \in S_m\}$ and* pen $: \mathscr{Q} \to \mathbb{R}_+$ *given by*

$$\operatorname{pen}(Q) = \kappa \inf_{\{m \in \mathcal{M}\,|\, Q = Q_t \text{ with } t \in S_m\}} \left[\frac{C_1}{4.7}\overline{V}_m\left[1 + \log_+(n/\overline{V}_m)\right] + \Delta(S_m)\right] \quad \text{for all } Q \in \mathscr{Q},$$

where κ is given by (22) and C_1 is the constant appearing in Proposition 5. Then any ρ-estimator $Q_{\widehat{s}}$ based on $\mathscr{Q}$ with penalty pen *satisfies, for all $\xi > 0$ and some constant $C > 0$ depending on the choice of ψ only,*

$$\mathbb{P}\left[Ch^2(P, Q_{\widehat{s}}) \le \inf_{m \in \mathcal{M}} \left(h^2(P, \mathscr{Q}_m) + \frac{\overline{V}_m}{n}\left[1 + \log_+\left(\frac{n}{\overline{V}_m}\right)\right] + \Delta(S_m)\right) + \frac{\xi}{n}\right] \ge 1 - e^{-\xi}.$$

Note that this result does not require any information or assumption on the distribution of W. If, in particular, the conditional probability P_w is absolutely continuous with respect to λ for almost all w with density $dP_w/d\lambda = s_w$, P_W-a.s., one can write

$$h^2(P, Q_{\widehat{s}}) = \int_{\mathscr{W}} h^2(s_w \cdot \lambda, \widehat{s}_w \cdot \lambda)\, dP_W(w),$$

hence

$$h^2(P, \mathscr{Q}_m) = \inf_{t \in S_m} \int_{\mathscr{W}} h^2(P_w, t_w \cdot \lambda)\, dP_W(w).$$

Proof. Applying Propositions 5 and 6 to each ρ-model S_m with $m \in \mathcal{M}$, we obtain under Assumption 2 the existence of a universal constant $C_1 > 0$ such that

$$D^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P}, \overline{\mathbf{P}}) \le \overline{D}(\boldsymbol{\mathscr{Q}}_m) = C_1 \overline{V}_m\left[1 + \log_+(n/\overline{V}_m)\right] \quad \text{for all } \mathbf{P} \in \mathscr{P} \text{ and } \overline{\mathbf{P}} \in \boldsymbol{\mathscr{Q}}_m.$$

The penalty function therefore satisfies (45) with $\Delta(\boldsymbol{\mathscr{Q}}_m) = \Delta(S_m)$ for all $m \in \mathcal{M}$ and the result follows from (46). ∎

10. Regression with a random design

In this section we assume that the observations $X_i = (W_i, Y_i)$, $1 \le i \le n$ are i.i.d. copies of a random pair

$$X = (W, Y) \quad \text{with} \quad Y = f(W) + \epsilon, \tag{49}$$

where W is a random variable with distribution P_W on a measurable space $(\mathscr{W}, \mathscr{B}(\mathscr{W}))$, f is an unknown regression function mapping $\mathscr{W}$ into $\mathbb{R}$ and ϵ is a real-valued random variable with values in $\mathbb{R}$ and distribution P_ϵ, which is independent of W. Both distributions P_W and P_ϵ are assumed to be unknown. We shall use the specific notations introduced in Section 4.3 when the data are i.i.d. and denote by μ the product measure $P_W \otimes \lambda$ where λ is the Lebesgue mesure on $\mathbb{R}$. Note that μ is unknown since it depends on the distribution P_W of the design W.

If ϵ had a density s with respect to λ, the distribution P of $X = (W, Y)$ would be absolutely continuous with respect to μ with density p given by

$$p(w, y) = s(y - f(w)) \quad \text{for } (w, y) \in \mathscr{X}, \tag{50}$$

depending thus on two parameters: the density s of the errors and the regression function f.

Denoting by $\mathscr{D}$ the set of all densities on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$ and $\mathscr{F}$ the set of all measurable functions mapping $\mathscr{W}$ into $\mathbb{R}$, our aim is to estimate P assuming that it is close to some distribution of the form $p \cdot \mu$ with p given by (50) for some $s \in \mathscr{D}$ and $f \in \mathscr{F}$. Besides, when P is exactly of this form we shall also derive estimators for both s and f.

10.1. The main result.

For $r \in \mathscr{D}$ and $g \in \mathscr{F}$, we set

$$Q_{r,g} = q_{r,g} \cdot \mu \quad \text{with} \quad q_{r,g}(w, y) = r(y - g(w)),$$

which means that $Q_{r,g}$ is the distribution of X in (49) when $f = g$ and ϵ is distributed according to $R = r \cdot \lambda$. Given a density $r \in \mathscr{D}$ and a countable subset F of $\mathscr{F}$, we define the ρ-model

$$\mathscr{Q}_m = \{Q_{r,g},\ g \in F\} \quad \text{for } m = (r, F).$$

Given a countable subset $\mathcal{D}$ of $\mathscr{D}$ and a countable family $\mathbb{F}$ of countable subsets F of $\mathscr{F}$, we estimate P on the basis of the collection of ρ-models $\{\mathscr{Q}_m,\ m \in \mathcal{M}\}$ with $\mathcal{M} \subset \mathcal{D} \times \mathbb{F}$. We endow this family with a weight function Δ satisfying (23) and assume the following.

Assumption 3.

i) The densities $r \in \mathcal{D}$ are unimodal.
ii) Each F in $\mathbb{F}$ is VC-subgraph with index $\overline{V}(F)$.
iii) The function ψ satisfies Assumption 1 with $\mathscr{Q} = \cup_{m \in \mathcal{M}} \mathscr{Q}_m$.

Under Assumptions 3-*i*) and *ii*), the family of densities $\mathcal{Q}_m = \{q_{r,g},\ g \in F\}$ is VC-subgraph on $\mathscr{X}$ with index not larger than $\overline{V}_m = 9.41\overline{V}(F)$ for all $m = (r, F) \in \mathcal{M}$. This result derives from Baraud, Birgé and Sart (2016, Proposition 42). Besides, under Assumption 3-*iii*) Proposition 5 applies and we obtain that for some universal constant $C_1 > 0$, all $m \in \mathcal{M}$, $\mathbf{P} \in \mathscr{P}$ and $\overline{\mathbf{P}} \in \mathscr{Q}$, $D^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P}, \overline{\mathbf{P}}) \le \overline{D}(\boldsymbol{\mathscr{Q}}_m)$ with

$$\overline{D}(\boldsymbol{\mathscr{Q}}_m) = C_1 \overline{V}_m \left[1 + \log_+\left(n/\overline{V}_m\right)\right] \quad \text{for all } m \in \mathcal{M}. \tag{51}$$

Setting

$$\operatorname{pen}(\mathbf{Q}) = \kappa \inf_{\{m \in \mathcal{M} \,|\, \mathscr{Q}_m \ni Q\}} \left\{ \frac{\overline{D}(\boldsymbol{\mathscr{Q}}_m)}{4.7} + \Delta(\boldsymbol{\mathscr{Q}}_m) \right\},$$

so that (24) holds with $G(\mathbf{P}, \overline{\mathbf{P}}) = 0$, we may apply Theorem 2 which leads, in this particular case, to the following analogue of (46).

Theorem 3. *Assume that Assumption 3 holds. For any distribution $P \in \mathscr{P}$, any ρ-estimator $\widehat{\mathbf{P}} = (\widehat{P}, \ldots, \widehat{P})$ satisfies, for all $\xi > 0$,*

$$\mathbb{P}\left[h^2(P, \widehat{P}) \le \inf_{m \in \mathcal{M}} \left(\gamma h^2(P, \mathscr{Q}_m) + \frac{4\kappa}{na_1} \left[\overline{D}(\boldsymbol{\mathscr{Q}}_m) + \Delta(\boldsymbol{\mathscr{Q}}_m) + 1.5 + \xi \right] \right) \right] \ge 1 - e^{-\xi}, \tag{52}$$

where γ and κ are given by (22) and $\overline{D}(\boldsymbol{\mathscr{Q}}_m)$ by (51).

Some comments are in order.

a) This result holds without any assumption on the distribution P_W of the design.
b) The result is true even if the regression framework (49) is not exact as long as the X_i are i.i.d. In particular, the distribution P needs not have a density with respect to $\mu = P_W \otimes \lambda$.
c) If r admits k modes with $k > 1$ and F is VC-subgraph with index not larger than $\overline{V}$, $\mathscr{Q}_{(r,F)}$ remains VC-subgraph and its index is still bounded by $C\overline{V}$ for some constant C that now depends on k. Consequently the above result generalizes to families $\mathcal{D}$ of densities admitting more than a single mode in which case $\overline{V}(F)$ should be replaced by $c(r)\overline{V}(F)$ where $c(r)$ is a positive number depending on the number of modes of the density r.
d) With Theorem 3 at hand we could obtain in the present random design context an analogue of Corollary 39 in Baraud, Birgé and Sart (2016) which was established when the W_i were deterministic (fixed design regression).

10.2. **Estimation of s and f.** Let us now consider the situation where the regression framework (49) is exact and ϵ has an unknown density s with respect to the Lebesgue measure λ. Then $P = Q_{s,f}$ admits a density $q_{s,f}$ given by (50) with s belonging to $\mathscr{D}$ and f to $\mathscr{F}$ but not necessarily to our ρ-models $\mathcal{D}$ and $\mathcal{F} = \bigcup_{F \in \mathbb{F}} F$ respectively. Since we may choose our ρ-estimator of the form

$$\widehat{P} = q_{\widehat{s}, \widehat{f}} \cdot \mu \quad \text{with} \quad (\widehat{s}, \widehat{f}) \in \mathcal{D} \times \mathcal{F},$$

our procedure results in estimators $\widehat{s}$ and $\widehat{f}$ for s and f respectively and our aim in this section is to establish risk bounds for these two estimators. Since the map $(r, g) \mapsto Q_{r,g}$ is not necessarily one to one from $\mathscr{D} \times \mathscr{F}$ to $\mathscr{P}$, identifiability conditions are required on our ρ-model $\mathscr{Q}$ so that the equality $Q_{r,g} = Q_{r',g'}$ with $r, r' \in \mathcal{D}$ and $g, g' \in \mathcal{F}$ implies that $r = r'$ λ-a.e. and $g = g'$ P_W-a.s. For this purpose, we first introduce the following notation. For $a \in \mathbb{R}$, we shall denote by R_a the probability on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$ with density $r_a(\cdot) = r(\cdot - a)$. When ϵ has density r and $a = g(w)$ for some $w \in \mathscr{W}$, R_a can be viewed as the conditional distribution of $Y = g(W) + \epsilon$ given $W = w$. Given $r, r' \in \mathscr{D}$, $g, g' \in \mathscr{F}$ and $w \in \mathscr{W}$, the Hellinger distance between the probabilities $R_{g(w)}$, and $R'_{g'(w)}$ is given by

$$h^2\left(R_{g(w)}, R'_{g'(w)}\right) = \frac{1}{2} \int_{\mathbb{R}} \left[\sqrt{r\,(y - g(w))} - \sqrt{r'\,(y - g'(w))} \right]^2 d\lambda(y)$$

and the Hellinger distance between the corresponding probabilities $Q_{r,g}$ and $Q_{r',g'}$ on $(\mathscr{X},\mathscr{B})$ writes

$$h^2\left(Q_{r,g},Q_{r',g'}\right)=\int_{\mathscr{W}}h^2\left(R_{g(w)},R'_{g'(w)}\right)dP_W(w). \tag{53}$$

We recall that the Hellinger distance is translation invariant which means that for all densities $r,r'\in\mathscr{D}$, $a,a'\in\mathbb{R}$,

$$h^2(R_a,R'_{a'})=h^2(R_{a-a'},R'). \tag{54}$$

In particular, taking $a=g(w)$ and $a'=g'(w)$ for $g,g'\in\mathscr{F}$ and $w\in\mathscr{W}$ and integrating (54) with respect to P_W we obtain

$$h^2\left(Q_{r,g},Q_{r',g'}\right)=h^2\left(Q_{r,g-g'},Q_{r',0}\right)\quad\text{for all }(g,g')\in\mathscr{F}^2\text{ and }(r,r')\in\mathscr{D}^2. \tag{55}$$

In order to warrant identifiability, we assume the following.

Assumption 4. *There exists a positive constant A such that, for all $r,r'\in\mathcal{D}$,*

$$h(R,R')\le A\inf_{a\in\mathbb{R}}h(R_a,R')\quad\textit{with}\quad R=r\cdot\lambda\ \textit{ and }\ R'=r'\cdot\lambda.$$

When $Q_{r,g}=Q_{r',g'}$, (53) asserts that

$$h\left(R_{g(w)},R'_{g'(w)}\right)=0\quad\text{for }P_W\text{-almost all }w\in\mathscr{W}$$

and (54) implies that $h\left(R_{g(w)-g'(w)},R'\right)=0$ for such $w\in\mathscr{W}$. Applying Assumption 4 with $a=g(w)-g'(w)$ leads to $R=R'$ and $g(w)=g'(w)$ which solves our identifiability problem.

In order to evaluate the risk of our estimator $\widehat{f}$ of f, we endow $\mathscr{F}$ with the loss function d_s defined on $\mathscr{F}\times\mathscr{F}$ by

$$d_s^2(g,g')=\frac{1}{2}\int_{\mathscr{W}\times\mathbb{R}}\left(\sqrt{s_{g(w)}(y)}-\sqrt{s_{g'(w)}(y)}\right)^2dP_W(w)\,dy\quad\text{for}\quad g,g'\in\mathscr{F}.$$

This loss function depends on the true density s of the errors ϵ and on the distribution P_W of the design, hence on P. We have seen in Section 6.3 of Baraud, Birgé and Sart (2016) that if the density s is of order α with $\alpha\in(-1,1]$ (see Definition 26 of that paper for the order of a function) the restriction of d_s to the $\mathbb{L}_\infty(P_W)$-ball $\mathscr{B}_\infty(b)$ centred at 0 with radius b is equivalent (up to factors depending on b and s) to

$$\|g-g'\|_{1+\alpha,P_W}^{(1+\alpha)/2}\qquad\text{with}\qquad\|g-g'\|_{1+\alpha,P_W}=\left[\int_{\mathscr{W}}|g-g'|^{1+\alpha}dP_W\right]^{1/(1+\alpha)}.$$

In particular, if $\mathcal{F}\subset\mathscr{B}_\infty(b)$ and the true regression function f also belongs to $\mathscr{B}_\infty(b)$,

$$c(s,b)\,\|f-g\|_{1+\alpha,P_W}^{(1+\alpha)/2}\le d_s(f,g)\le C(s,b)\,\|f-g\|_{1+\alpha,P_W}^{(1+\alpha)/2}\quad\text{for all}\quad g\in\mathcal{F}$$

and suitable positive numbers $c(s,b)$ and $C(s,b)$. Of special interest is the case of $\alpha=1$ for which $d_s(f,g)$ is of the order of the $\mathbb{L}_2(P_W)$-distance between f and g for all $g\in\mathcal{F}$. This situation is met when the translation ρ-model associated to s is regular which is the case when s is Cauchy, Gaussian, Laplace, etc. When s is uniform or exponential, $d_s^2(\cdot,\cdot)$ is then equivalent to the $\mathbb{L}_1(P_W)$-norm.

In view of evaluating the risk of our estimator $\widehat{s}$ of the density s we shall consider the loss between two densities $r, r' \in \mathscr{D}$ induced by the Hellinger distance between the two corresponding measures $r \cdot \lambda$ and $r' \cdot \lambda$ and we shall write this loss $h(r, r')$ so that

$$h(r, r') = h(r \cdot \lambda, r' \cdot \lambda) = h\big(Q_{r,0}, Q_{r',0}\big) \quad \text{for all } r, r' \in \mathscr{D}.$$

We deduce from Theorem 3 the following result.

Corollary 5. *Assume that the X_i are i.i.d. with density p given by (50) and that Assumptions 3 and 4 are satisfied. For all $\xi > 0$ and all ρ-estimators $Q_{\widehat{s},\widehat{f}}$, with $\widehat{s} \in \mathcal{D}$ and $\widehat{f} \in \mathcal{F}$, based on the ρ-models $\mathscr{Q}_m$ defined in Section 10.1, with probability at least $1 - e^{-\xi}$,*

$$\max\left\{d_s^2(f, \widehat{f}), h^2(s, \widehat{s})\right\} \le C \inf_{(r,F)\in\mathcal{M}} \left[d_s^2(f, F) + h^2(s, r) + \frac{\overline{D}(\boldsymbol{\mathscr{Q}}_{(r,F)}) + \Delta(\boldsymbol{\mathscr{Q}}_{(r,F)}) + \xi}{n}\right],$$

where C is a positive constant depending on A and the choice of ψ.

The risk bound is the same for the two estimators and depends on the approximation properties of $\mathcal{F}$ and $\mathcal{D}$ with respect to f and s respectively. The proof of this corollary is given in Section 12.2.

11. Estimator selection and aggregation

In the case of density estimation, ρ-estimators can also be used to perform selection or aggregation of preliminary estimators. In this case, we assume that we have at hand a set $\boldsymbol{X}_1 = (X_1, \dots, X_n)$ of n independent random variables with an unknown joint distribution $\mathbf{P}$ to be estimated. We also have at hand a finite family $\boldsymbol{\mathscr{Q}} = \{\mathbf{P}_j,\ j \in \mathcal{J}\}$ of probabilities that can be considered as candidate estimators for $\mathbf{P}$. These are completely arbitrary but, in a typical situation, it is assumed (although this may not be true) that the observations X_i are i.i.d. and the $\mathbf{P}_j$ are preliminary estimators of the form $\mathbf{P}_j = \widehat{P}_j^{\otimes n}(\boldsymbol{X}_2)$, where $\boldsymbol{X}_2$ is a second sample independent from $\boldsymbol{X}_1$, and the $\widehat{P}_j$ are estimators that derive from various procedures applied to the sample $\boldsymbol{X}_2$.

11.1. Estimator selection.

Taking $\mathcal{M} = \mathcal{J}$, we view each probability $\mathbf{P}_j$ as a ρ-model $\boldsymbol{\mathscr{Q}}_j = \{\mathbf{P}_j\}$ with a single element. As a consequence, it follows from Proposition 4 that $D^{\boldsymbol{\mathscr{Q}}_j}(\mathbf{P}, \mathbf{P}_j) \le 9 \log 2 < 6.3$. Then we choose the weights $\Delta(\boldsymbol{\mathscr{Q}}_j) = \Delta(\mathbf{P}_j)$ satisfying (23). We may choose $\Delta(\boldsymbol{\mathscr{Q}}_j) = \log|\mathcal{J}|$ for all j but other more Bayesian choices are possible, or choices based on the confidence we have in the various procedures used to build the preliminary estimators. To compute the penalized ρ-estimator $\widehat{\mathbf{P}}$ of $\mathbf{P}$, we may use the penalty function $\operatorname{pen}(\mathbf{P}_j) = \kappa\Delta(\mathbf{P}_j)$ for all $j \in \mathcal{J}$, which, in view of (24), leads to $G(\mathbf{P}, \overline{\mathbf{P}}) = \kappa(6.3/4.7)$. Finally (46) shows that, for a suitable constant C depending on ψ only,

$$\mathbb{P}\left[\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le C \inf_{j\in\mathcal{J}} \big(\mathbf{h}^2(\mathbf{P}, \mathbf{P}_j) + \Delta(\mathbf{P}_j) + 1 + \xi\big)\right] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0 \text{ and } \mathbf{P} \in \mathscr{P}.$$

11.2. **Convex estimator aggregation.** In this case we set $\mathscr{J} = \{1, \ldots, N\}$, $N \geq 2$, we introduce some arbitrary countable and dense subset S of the N-dimensional simplex

$$\mathscr{C} = \left\{ (\alpha_1, \ldots, \alpha_N) \text{ such that } \alpha_j \geq 0 \text{ for } 1 \leq j \leq N \text{ and } \sum_{j=1}^{N} \alpha_j = 1 \right\}$$

and consider a single ρ-model

$$\boldsymbol{\mathscr{Q}} = \left\{ \sum_{j=1}^{N} \alpha_j \mathbf{P}_j \quad \text{for} \quad (\alpha_1, \ldots, \alpha_N) \in S \right\}.$$

To get a representation $\mathcal{R}(\boldsymbol{\mathscr{Q}})$, we select a dominating measure $\boldsymbol{\mu}$, densities $\mathbf{p}_j = d\mathbf{P}_j/d\boldsymbol{\mu}$ and set

$$\boldsymbol{\mathcal{Q}} = \left\{ \sum_{j=1}^{N} \alpha_j \mathbf{p}_j \quad \text{for} \quad (\alpha_1, \ldots, \alpha_N) \in S \right\}. \tag{56}$$

Note that this ρ-model is a subset of an N-dimensional linear space. It is therefore VC-subgraph with index $\overline{V}$ not larger than $N + 2$ and it follows from Proposition 5 that,

$$D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P}, \overline{\mathbf{P}}) \leq CN \log n \quad \text{for all } \mathbf{P} \in \mathscr{P} \text{ and } \overline{\mathbf{P}} \in \boldsymbol{\mathscr{Q}}.$$

Theorem 1 implies that, if $\widehat{\mathbf{P}}$ is the corresponding ρ-estimator of $\mathbf{P}$,

$$\mathbb{P}\left[\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \leq C \left(\mathbf{h}^2(\mathbf{P}, \boldsymbol{\mathscr{Q}}) + N \log n + \xi \right) \right] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0 \text{ and } \mathbf{P} \in \mathscr{P}.$$

It should be noted that there is no $\mathbb{L}_2$-type argument here, the densities $\mathbf{p}_j$ can be absolutely anything and the true distribution $\mathbf{P}$ should be a product measure but not necessarily of the form $P^{\otimes n}$.

11.3. **Practical implementation of the ρ-estimator for convex aggregation.** Let us consider the convex aggregation problem presented above. In the sequel, the function ψ will be either ψ_1 or ψ_2 as defined in Proposition 3. Note that for both choices, the function $u \mapsto \psi(\sqrt{u})$ is strictly concave. Let us now introduce the function

$$\mathbf{t} : (\boldsymbol{\alpha}, \boldsymbol{\beta}) \longmapsto \sum_{i=1}^{n} \psi\left(\sqrt{\frac{\sum_{j=1}^{N} \boldsymbol{\beta}_j \mathbf{p}_j(X_i)}{\sum_{j=1}^{N} \boldsymbol{\alpha}_j \mathbf{p}_j(X_i)}} \right) \quad \text{from } \mathscr{C} \times \mathscr{C} \text{ to } \mathbb{R}_+.$$

Theorem 4. *If the vectors*

$$\overrightarrow{\mathbf{p}}_j = \left(\mathbf{p}_j(X_1), \ldots, \mathbf{p}_j(X_n) \right), \quad j \in \mathscr{J},$$

are linearly independent with positive coordinates, the following facts hold.

i) For all $\boldsymbol{\alpha} \in \mathscr{C}$, the map $\boldsymbol{\beta} \mapsto \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is strictly concave on $\mathscr{C}$ and for all $\boldsymbol{\beta} \in \mathscr{C}$, the map $\boldsymbol{\alpha} \mapsto \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is strictly convex on $\mathscr{C}$.
ii) The map $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ possesses a unique saddle point on $\mathscr{C} \times \mathscr{C}$. This saddle point is of the form $(\boldsymbol{\alpha}^\star, \boldsymbol{\alpha}^\star)$ which means that

$$\mathbf{t}(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}) \leq \mathbf{t}(\boldsymbol{\alpha}^\star, \boldsymbol{\alpha}^\star) = 0 \leq \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^\star) \quad \textit{for all } (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathscr{C} \times \mathscr{C}.$$

iii) The probability

$$\widehat{\mathbf{P}} = \widehat{\mathbf{p}} \cdot \boldsymbol{\mu} \qquad \textit{with} \qquad \widehat{\mathbf{p}} = \sum_{j=1}^{N} \boldsymbol{\alpha}_j^\star \mathbf{p}_j$$

is a ρ-estimator for $\mathbf{P}$ over $\mathscr{Q}$ and it satisfies

$$\boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{\mathbf{p}}) = \inf_{\mathbf{q} \in \mathscr{Q}} \boldsymbol{\Upsilon}(\boldsymbol{X}, \mathbf{q}) = 0.$$

On the basis of this result and following Cherruault and Loridan (1973) for the search for a saddle point, we suggest the following algorithm to compute the ρ-estimator $\widehat{\mathbf{P}}$ of $\mathbf{P}$.

(1) `Initialization: choose` $\boldsymbol{\alpha} \in \mathscr{C}$ `and` $\varepsilon \in (0,1]$;

(2) `Compute`

$$\boldsymbol{\beta} = \arg\max_{\boldsymbol{\beta}' \in \mathscr{C}} \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta}');$$

(3) `if` $\mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \varepsilon$, `go to (2) with` $\boldsymbol{\alpha} = \boldsymbol{\beta}$;

(4) `else return` $\widehat{\mathbf{p}} = \left(\sum_{j=1}^{N} \boldsymbol{\alpha}_j \mathbf{p}_j\right)$.

Proof of Theorem 4. Under our assumption, $\mathbf{q}_{\boldsymbol{\alpha}} = \sum_{j=1}^{N} \boldsymbol{\alpha}_j \mathbf{p}_j$ only takes positive values at $X_1, \dots, X_n$ for all $\boldsymbol{\alpha} \in \mathscr{C}$. Since the function $u \mapsto \psi(\sqrt{u})$ is concave, for a given $\boldsymbol{\alpha} \in \mathscr{C}$, the map $\boldsymbol{\beta} \mapsto \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is concave on the convex set $\mathscr{C}$ as a sum of concave functions. Moreover, if for some $\lambda \in (0,1)$ and $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathscr{C}$

$$\mathbf{t}(\boldsymbol{\alpha}, \lambda\boldsymbol{\beta} + (1-\lambda)\boldsymbol{\beta}') = \lambda \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + (1-\lambda)\mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta}'),$$

then necessarily, for all $i \in \{1, \dots, n\}$,

$$\psi\left(\sqrt{\frac{\lambda \mathbf{q}_{\boldsymbol{\beta}}(X_i) + (1-\lambda)\mathbf{q}_{\boldsymbol{\beta}'}(X_i)}{\mathbf{q}_{\boldsymbol{\alpha}}(X_i)}}\right) = \lambda\psi\left(\sqrt{\frac{\mathbf{q}_{\boldsymbol{\beta}}(X_i)}{\mathbf{q}_{\boldsymbol{\alpha}}(X_i)}}\right) + (1-\lambda)\psi\left(\sqrt{\frac{\mathbf{q}_{\boldsymbol{\beta}'}(X_i)}{\mathbf{q}_{\boldsymbol{\alpha}}(X_i)}}\right).$$

Since the function $u \mapsto \psi(\sqrt{u})$ is strictly concave and $\mathbf{q}_{\boldsymbol{\alpha}}(X_i) > 0$, we obtain that

$$\sum_{j=1}^{N} \boldsymbol{\beta}_j \mathbf{p}_j(X_i) = \sum_{j=1}^{N} \boldsymbol{\beta}'_j \mathbf{p}_j(X_i) \quad \text{for all } i \in \{1, \dots, n\}$$

or equivalently, $\sum_{j=1}^{N} \left(\boldsymbol{\beta}_j - \boldsymbol{\beta'}_j\right) \overrightarrow{\mathbf{p}}_j = 0$. This implies that $\boldsymbol{\beta} = \boldsymbol{\beta}'$ since the vectors $\overrightarrow{\mathbf{p}}_j$, $j = 1, \dots, N$, are linearly independent. Moreover, $\mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\mathbf{t}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ since $\psi(1/x) = -\psi(x)$, therefore the map $\boldsymbol{\alpha} \mapsto \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is strictly convex for all $\boldsymbol{\beta} \in \mathscr{C}$ which concludes the proof of $i)$.

The set $\mathscr{C}$ being compact and convex, the strictly concave-convex function $\mathbf{t}$ admits a unique saddle point $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) \in \mathscr{C} \times \mathscr{C}$ such that

$$\mathbf{t}(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = \inf_{\boldsymbol{\alpha} \in \mathscr{C}} \sup_{\boldsymbol{\beta} \in \mathscr{C}} \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sup_{\boldsymbol{\beta} \in \mathscr{C}} \inf_{\boldsymbol{\alpha} \in \mathscr{C}} \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

by the min-max theorem and

$$\mathbf{t}(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = \inf_{\boldsymbol{\alpha} \in \mathscr{C}} \sup_{\boldsymbol{\beta} \in \mathscr{C}} \mathbf{t}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\boldsymbol{\alpha} \in \mathscr{C}} \sup_{\boldsymbol{\beta} \in \mathscr{C}} [-\mathbf{t}(\boldsymbol{\beta}, \boldsymbol{\alpha})] = -\sup_{\boldsymbol{\alpha} \in \mathscr{C}} \inf_{\boldsymbol{\beta} \in \mathscr{C}} \mathbf{t}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\mathbf{t}(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star),$$

hence $\mathbf{t}(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)=0$. Besides,

$$\mathbf{t}(\boldsymbol{\alpha}^\star,\boldsymbol{\alpha}^\star)=0=\mathbf{t}(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)=\sup_{\boldsymbol{\beta}\in\mathscr{C}}\mathbf{t}(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}).$$

Since the function $\boldsymbol{\beta}\mapsto\mathbf{t}(\boldsymbol{\alpha}^\star,\boldsymbol{\beta})$ is strictly concave, it has a unique maximum, hence $\boldsymbol{\alpha}^\star=\boldsymbol{\beta}^\star$ which proves ii). The conclusion of iii) follows from the definition of a ρ-estimator. □

12. Proof of the main results

12.1. **Proof of Theorem 2.** The proof is based on two auxiliary results to be proved in Sections 13.9 and 13.10 respectively. The first one allows us to bound

$$w(\mathcal{R}_1,\mathbf{P},\overline{\mathbf{P}},y)=\mathbb{E}\left[\sup_{\mathbf{Q}\in\mathscr{B}^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}},y)}\left|\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]\right|\right] \tag{57}$$

when $\overline{\mathbf{p}}$ and $\mathbf{q}$ are defined with a given representation $\mathcal{R}_1$ of $\mathscr{Q}$.

Proposition 11. *Whatever the representation $\mathcal{R}_1=\mathcal{R}_1(\mathscr{Q})$*

$$w(\mathcal{R}_1,\mathbf{P},\overline{\mathbf{P}},y)\le\mathbf{w}^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}},y)+16\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right)\quad\text{for all }\mathbf{P}\in\mathscr{P}\quad\text{and}\quad\overline{\mathbf{P}}\in\mathscr{Q}.$$

Let us now focus on our second auxiliary result.

Proposition 12. *Let $\mathbf{P}\in\mathscr{P}$, $\{\mathscr{Q}_m,\ m\in\mathcal{M}\}\subset\mathscr{P}$ be a countable family of ρ-models and Δ an associated weight function satisfying (23), $\overline{\mathbf{P}}\in\mathscr{Q}=\bigcup_{m\in\mathcal{M}}\mathscr{Q}_m$, $c_0>0$ and $B=1+a_2^2/(16c_0)$. Let $\mathcal{R}(\mathscr{Q})$ be an arbitrary representation to be used for building the statistic $\mathbf{T}$ and, for all $m\in\mathcal{M}$, let $\widetilde{D}_m(\mathbf{P},\overline{\mathbf{P}})$ be such that*

$$\mathbf{w}^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}},y)\le c_0y^2\quad\text{for all }y>\sqrt{\widetilde{D}_m(\mathbf{P},\overline{\mathbf{P}})}. \tag{58}$$

Whatever the numbers $\alpha>\delta>1$ and $\vartheta>1$ such that

$$2\exp[-\vartheta]+\sum_{j\ge1}\exp\left[-\vartheta\delta^j\right]\le1, \tag{59}$$

the following holds: for any $\xi>0$, with probability at least $1-e^{-\xi}$ and for all $m\in\mathcal{M}$ simultaneously,

$$\begin{aligned}&\sup_{\mathbf{Q}\in\mathscr{Q}_m}\left[\left|\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]\right|-c_0\alpha\left[\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})+\mathbf{h}^2(\mathbf{P},\mathbf{Q})\right]\right]\\ &\qquad\le\ K\left[c_0\widetilde{D}_m(\mathbf{P},\overline{\mathbf{P}})\bigvee\tau\left[\Delta(\mathscr{Q}_m)+\vartheta+\xi\right]\right]+16\left(1+\frac{4\sqrt{2}}{\sqrt{B\tau}}\right)\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}}),\end{aligned} \tag{60}$$

with

$$\tau\ge\tau_0=\frac{1}{2\delta B}\left[\sqrt{1+\frac{\alpha-\delta}{8\delta B}}-1\right]^{-2}\qquad\text{and}\qquad K=1+8\sqrt{\frac{2B}{\tau}}+\frac{4}{\tau}. \tag{61}$$

Let us now turn to the proof of Theorem 2. Let us fix $\xi>0$, $\overline{\mathbf{P}}\in\mathscr{Q}$ and a representation $\mathcal{R}(\mathscr{Q})$ of the ρ-model $\mathscr{Q}$. It follows from the definition of $D^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}})$ that (58) is satisfied with $c_0=a_1/8$ and $\widetilde{D}_m(\mathbf{P},\overline{\mathbf{P}})=\beta^{-2}D^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}})$ and we may therefore apply Proposition 12 with $c_0 = a_1/8$ and this choice of $\widetilde{D}_m(\mathbf{P}, \overline{\mathbf{P}})$ for all $m \in \mathcal{M}$, $\alpha = 4$, $\tau = \tau_0$, $\delta = 1.175$ and $\vartheta = 1.47$ which implies that (59) is satisfied. Let us then set

$$c_1 = 1 + 8\sqrt{\frac{2B}{\tau_0}} + \frac{4}{\tau_0}, \qquad c_2 = 16\left(1 + \frac{4\sqrt{2}}{\sqrt{B\tau_0}}\right) \qquad \text{and} \qquad \xi' = \kappa(\xi + \vartheta).$$

We deduce from (60) and the chosen values of the various constants involved that, on a set Ω_ξ the probability of which is at least $1 - e^{-\xi}$, for all $\mathbf{Q} \in \mathscr{Q}$ and all ρ-models $\mathscr{Q}_m$ containing $\mathbf{Q}$,

$$\begin{aligned} \mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q}) &\leq \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q})\right] + (a_1/2)\left[\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \mathbf{h}^2(\mathbf{P}, \mathbf{Q})\right] \\ &\quad + c_1\left[c_0\beta^{-2}D^{\mathscr{Q}_m}(\mathbf{P}, \overline{\mathbf{P}}) + \tau_0\left(\Delta(\mathscr{Q}_m) + \kappa^{-1}\xi'\right)\right] + c_2\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) \\ &= \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q})\right] + (a_1/2)\left[\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \mathbf{h}^2(\mathbf{P}, \mathbf{Q})\right] \\ &\quad + \tau_0 c_1\left[\frac{c_0}{\tau_0\beta^2}D^{\mathscr{Q}_m}(\mathbf{P}, \overline{\mathbf{P}}) + \Delta(\mathscr{Q}_m) + \kappa^{-1}\xi'\right] + c_2\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}). \end{aligned} \tag{62}$$

Now observe that $B = 1 + a_2^2/(2a_1)$, hence $B \geq 4$ since $a_2^2 \geq 6a_1$ and $(\alpha - \delta)/(8\delta B) < 0.07514$. Since

$$0.4909\,x < \sqrt{1+x} - 1 < 0.5\,x \quad \text{for } 0 < x < 0.07514,$$

it follows that

$$\frac{0.1475}{B} < \sqrt{1 + \frac{\alpha - \delta}{8\delta B}} - 1 < \frac{\alpha - \delta}{16\delta B} < \frac{0.1503}{B},$$

hence

$$18.837B < \tau_0 = \frac{1}{2\delta B}\left[\sqrt{1 + \frac{\alpha - \delta}{8\delta B}} - 1\right]^{-2} < 19.56B \tag{63}$$

and

$$\tau_0 c_1 = \tau_0 + 8\sqrt{2B\tau_0} + 4 < 69.6B + 4 \leq \frac{34.8a_2^2}{a_1} + 73.6 < \kappa.$$

Moreover, (63) and the inequality $B > a_2^2/(2a_1)$ imply that

$$\frac{c_0}{\tau_0\beta^2} = \frac{a_1}{8}\frac{16a_2^2}{a_1^2\tau_0} = \frac{2a_2^2}{a_1\tau_0} < \frac{2a_2^2}{18.837Ba_1} < \frac{2a_2^2}{18.837\left[a_2^2/(2a_1)\right]a_1} < \frac{1}{4.7}$$

and

$$\frac{c_2}{16} - 1 < \frac{4\sqrt{2}}{B\sqrt{18.837}} < \frac{8\sqrt{2}a_1}{\sqrt{18.837}a_2^2} < 2.6068\frac{a_1}{a_2^2}. \tag{64}$$

Then (62) becomes

$$\begin{aligned} \mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q}) &\leq \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q})\right] + (a_1/2)\left[\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \mathbf{h}^2(\mathbf{P}, \mathbf{Q})\right] \\ &\quad + \kappa\left[\frac{D^{\mathscr{Q}_m}(\mathbf{P}, \overline{\mathbf{P}})}{4.7} + \Delta(\mathscr{Q}_m)\right] + \xi' + c_2\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}). \end{aligned}$$

Since the last inequality is true for all $\mathscr{Q}_m$ containing $\mathbf{Q}$, we derive from (24) that

$$\begin{aligned} \mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q}) &\leq \mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q})\right] + (a_1/2)\left[\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \mathbf{h}^2(\mathbf{P}, \mathbf{Q})\right] \\ &\quad + G(\mathbf{P}, \overline{\mathbf{P}}) + \mathrm{pen}(\mathbf{Q}) + \xi' + c_2\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}), \end{aligned}$$

which, together with (17), leads to the following inequality which holds on Ω_ξ, for all $\mathbf{Q} \in \mathscr{Q}$ and with $A = a_0 + (a_1/2) + c_2$:

$$\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q}) \le A\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) - (a_1/2)\mathbf{h}^2(\mathbf{P}, \mathbf{Q}) + G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\mathbf{Q}) + \xi'.$$

We deduce from this inequality that, on Ω_ξ, for any (random) element $\widehat{\mathbf{P}} \in \mathscr{Q}$,

$$\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \widehat{\mathbf{p}}) \le A\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) - (a_1/2)\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) + G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\widehat{\mathbf{P}}) + \xi', \tag{65}$$

and that

$$\begin{aligned} \boldsymbol{\Upsilon}(\boldsymbol{X}, \overline{\mathbf{P}}) &= \sup_{\mathbf{Q} \in \mathscr{Q}} [\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q}) - \operatorname{pen}(\mathbf{Q})] + \operatorname{pen}(\overline{\mathbf{P}}) \\ &\le A\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) - (a_1/2)\mathbf{h}^2(\mathbf{P}, \mathscr{Q}) + G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\overline{\mathbf{P}}) + \xi'. \end{aligned} \tag{66}$$

Since $\mathbf{T}(\boldsymbol{X}, \widehat{\mathbf{p}}, \overline{\mathbf{p}}) = -\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \widehat{\mathbf{p}},)$, (65) leads to

$$\begin{aligned} (a_1/2)\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) &\le A\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) - \mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \widehat{\mathbf{p}}) + G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\widehat{\mathbf{P}}) + \xi' \\ &= A\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \left[\mathbf{T}(\boldsymbol{X}, \widehat{\mathbf{p}}, \overline{\mathbf{p}}) - \operatorname{pen}(\overline{\mathbf{P}})\right] + \operatorname{pen}(\widehat{\mathbf{P}}) \\ &\quad + G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\overline{\mathbf{P}}) + \xi' \\ &\le A\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{\mathbf{P}}) + G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\overline{\mathbf{P}}) + \xi'. \end{aligned} \tag{67}$$

If $\widehat{\mathbf{P}}$ belongs to the set $\mathscr{E}(\psi, \boldsymbol{X})$, $\boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{\mathbf{P}}) \le \boldsymbol{\Upsilon}(\boldsymbol{X}, \overline{\mathbf{P}}) + \kappa/25$ and by (66)

$$\boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{\mathbf{P}}) \le A\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) - (a_1/2)\mathbf{h}^2(\mathbf{P}, \mathscr{Q}) + G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\overline{\mathbf{P}}) + \xi' + (\kappa/25),$$

which, together with (67), shows that on the set Ω_ξ and for all $\widehat{\mathbf{P}} \in \mathscr{E}(\psi, \boldsymbol{X})$,

$$\frac{a_1}{2}\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le 2A\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) - \frac{a_1}{2}\mathbf{h}^2(\mathbf{P}, \mathscr{Q}) + 2\left[G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\overline{\mathbf{P}}) + \kappa(\xi + \vartheta + 0.02)\right].$$

This inequality, which extends to any element $\widehat{\mathbf{P}}$ belonging to the closure of $\mathscr{E}(\psi, \boldsymbol{X})$, writes

$$\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \frac{4A}{a_1}\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) - \mathbf{h}^2(\mathbf{P}, \mathscr{Q}) + \frac{4}{a_1}\left[G(\mathbf{P}, \overline{\mathbf{P}}) + \operatorname{pen}(\overline{\mathbf{P}}) + \kappa(\xi + \vartheta + 0.02)\right].$$

and the conclusion follows since, by (64), $A < a_0 + (a_1/2) + 16 + 42(a_1/a_2^2)$.

12.2. **Proof of Corollary 5.** Let us fix some arbitrary $m = (r, F) \in \mathcal{M}$ and $g \in F$. Under Assumption 4 and (54), for all $r' \in \mathcal{D}$, $g' \in \mathscr{F}$ and $w \in \mathscr{W}$,

$$h^2(r, r') = h^2(R, R') \le A^2 h^2\left(R_{g(w)-g'(w)}, R'\right) = A^2 h^2\left(R_{g(w)}, R'_{g'(w)}\right).$$

Integrating this inequality with respect to P_W gives

$$h(r, r') \le A h(Q_{r,g}, Q_{r',g'}) \quad \text{for all } g' \in \mathscr{F}. \tag{68}$$

For all $r' \in \mathcal{D}$, we deduce from (68), (54) and (55)

$$\begin{aligned} d_s(f, g') &= h(Q_{s,f}, Q_{s,g'}) \le h(Q_{s,f}, Q_{r',g'}) + h(Q_{r',g'}, Q_{r,g'}) + h(Q_{r,g'}, Q_{s,g'}) \\ &\le h(Q_{s,f}, Q_{r',g'}) + h(Q_{r',0}, Q_{r,0}) + h(Q_{r,0}, Q_{s,0}) \\ &= h(Q_{s,f}, Q_{r',g'}) + h(r, r') + h(s, r) \\ &\le h(Q_{s,f}, Q_{r',g'}) + A h(Q_{r',g'}, Q_{r,f}) + h(s, r) \\ &\le h(Q_{s,f}, Q_{r',g'}) + A\left[h(Q_{r',g'}, Q_{s,f}) + h(Q_{s,f}, Q_{r,f})\right] + h(s, r) \\ &= h(Q_{s,f}, Q_{r',g'}) + A\left[h(Q_{r',g'}, Q_{s,f}) + h(s, r)\right] + h(s, r) \\ &\le (1 + A)\left[h(Q_{s,f}, Q_{r',g'}) + h(s, r)\right] = (1 + A)\left[h(P, Q_{r',g'}) + h(s, r)\right]. \end{aligned}$$

Taking $r' = \widehat{s}$ and $g' = \widehat{f}$, we get

$$d_s(f, \widehat{f}) \le (1+A)\left[h(P, \widehat{P}) + h(s,r)\right] \quad \text{for all } r \in \mathcal{D}. \tag{69}$$

Besides,

$$h(P, \mathscr{Q}_m) \le h(Q_{s,f}, Q_{r,g}) \le h(Q_{s,f}, Q_{s,g}) + h(Q_{s,g}, Q_{r,g}) = d_s(f,g) + h(s,r),$$

hence in (52), since $\overline{D}(\boldsymbol{\mathscr{Q}}_m) \ge 1$,

$$\begin{aligned}
&\gamma h^2(P, \mathscr{Q}_m) + \frac{4\kappa}{na_1}\left[\overline{D}(\boldsymbol{\mathscr{Q}}_m) + \Delta(\boldsymbol{\mathscr{Q}}_m) + 1.5 + \xi\right] \\
&\qquad \le \quad C\left[d_s^2(f,g) + h^2(s,r) + n^{-1}\left(\overline{D}(\boldsymbol{\mathscr{Q}}_m) + \Delta(\boldsymbol{\mathscr{Q}}_m) + \xi\right)\right]
\end{aligned} \tag{70}$$

for some $C > 0$ depending on γ, a_1 and κ. Using (69) and Theorem 3, we obtain that, on a set Ω_ξ of probability at least $1 - e^{-\xi}$,

$$\begin{aligned}
d_s^2(f, \widehat{f}) &\le 2(1+A)^2\left[h^2(P, \widehat{P}) + h^2(s,r)\right] \\
&\le 2C(1+A)^2\left[d_s^2(f,g) + h^2(s,r) + n^{-1}\left(\overline{D}(\boldsymbol{\mathscr{Q}}_m) + \Delta(\boldsymbol{\mathscr{Q}}_m) + \xi\right)\right] \\
&\quad + 2(1+A)^2 h^2(s,r) \\
&\le C'\left[d_s^2(f,g) + h^2(s,r) + n^{-1}\left(\overline{D}(\boldsymbol{\mathscr{Q}}_m) + \Delta(\boldsymbol{\mathscr{Q}}_m) + \xi\right)\right],
\end{aligned} \tag{71}$$

where C' depends on A, γ, a_1 and κ. Using now (68) with $r' = \widehat{s}$ and $g' = \widehat{f}$, we deduce that

$$\begin{aligned}
h(s, \widehat{s}) &\le h(s,r) + h(r, \widehat{s}) \le h(s,r) + Ah(Q_{r,f}, Q_{\widehat{s},\widehat{f}}) \\
&\le h(s,r) + A\left[h(Q_{r,f}, Q_{s,f}) + h(Q_{s,f}, Q_{\widehat{s},\widehat{f}})\right] \\
&\le (1+A)\left[h(s,r) + h(P, \widehat{P})\right].
\end{aligned}$$

This bound is the same as that established in (69) for $d_s(f, \widehat{f})$ and arguing as before we obtain that on the same event Ω_ξ, $h^2(s, \widehat{s})$ is also not larger than the right-hand side of (71). The conclusion follows since $m = (r, F)$ is arbitrary in $\mathcal{M}$.

13. Other proofs

13.1. Proof of Proposition 2.

We start with the following result.

Lemma 3. *If (13) and (14) hold with $a_0 \ge 2(1+a_1)$ for a representation $(\boldsymbol{\mu}, \boldsymbol{\mathscr{Q}})$ and all $\mathbf{R} \ll \boldsymbol{\mu}$, they hold for this representation and all $\mathbf{R} \in \boldsymbol{\mathscr{P}}$.*

Proof. Let $\mathbf{q}, \mathbf{q}' \in \boldsymbol{\mathscr{Q}}$ and $\mathbf{R}$ be some probability on $(\mathscr{X}, \mathscr{B})$ which is not necessarily absolutely continuous with respect to $\boldsymbol{\mu}$. We stick here to the notations introduced in Section 4.3 and drop the index i. We may write $R = \delta^2 R' + (1-\delta^2)R''$ where R' is a probability which is absolutely continuous with respect to μ, R'' is a probability which is orthogonal to μ and $\delta \in [0,1]$. Taking $\overline{\mu} = R + Q$ which dominates both R and Q and using the fact that $(dR''/d\overline{\mu})(dQ/d\overline{\mu}) = 0$, $\overline{\mu}$-a.e. since $Q = q \cdot \mu$ is orthogonal to R'', we get

$$\rho(R,Q) = \int_{\mathscr{X}} \sqrt{\left[\delta^2 \frac{dR'}{d\overline{\mu}} + (1-\delta^2)\frac{dR''}{d\overline{\mu}}\right]\frac{dQ}{d\overline{\mu}}}\, d\overline{\mu} = \int_{\mathscr{X}} \delta \sqrt{\frac{dR'}{d\overline{\mu}}\frac{dQ}{d\overline{\mu}}}\, d\overline{\mu} = \delta\rho(R', Q) \le \delta.$$

Hence

$$\delta h^2(R',Q) = \delta - \delta\rho(R',Q) = \delta - \left[1 - h^2(R,Q)\right] \geq 0 \tag{72}$$

with similar results for $Q' = q' \cdot \mu$. Then, applying (14) to $R' \ll \mu$ and using that $a_2 \geq 1 \geq |\psi|$, we derive that

$$\begin{aligned}
\int_{\mathscr{X}} \psi^2\left(\sqrt{\frac{q'}{q}}\right) dR &\leq \delta^2 \int_{\mathscr{X}} \psi^2\left(\sqrt{\frac{q'}{q}}\right) dR' + 1 - \delta^2 \\
&\leq \delta^2 a_2^2\left[h^2(R',Q) + h^2(R',Q')\right] + a_2^2(1-\delta^2) \\
&= a_2^2\left[2\delta^2 - 2\delta + \delta\left[h^2(R,Q) + h^2(R,Q')\right] + 1 - \delta^2\right] \\
&= a_2^2\left[(1-\delta)^2 + \delta A\right],
\end{aligned}$$

with $A = h^2(R,Q) + h^2(R,Q') \geq 2(1-\delta)$, hence $(1-\delta)^2 + \delta A \leq (1-\delta)(A/2) + \delta A \leq A$ which leads to (14).

Let us now focus on (13). The same computations, using (72) and $h^2(R,Q) \geq 1-\delta$, lead to

$$\begin{aligned}
\int_{\mathscr{X}} \psi\left(\sqrt{\frac{q'}{q}}\right) dR &\leq \delta^2\left[a_0 h^2(R',Q) - a_1 h^2(R',Q')\right] + 1 - \delta^2 \\
&= \delta(\delta-1)(a_0 - a_1) + \delta\left[a_0 h^2(R,Q) - a_1 h^2(R,Q')\right] + 1 - \delta^2 \\
&= a_0 h^2(R,Q) - a_1 h^2(R,Q') \\
&\quad - (1-\delta)\left[\delta(a_0 - a_1) + a_0 h^2(R,Q) - a_1 h^2(R,Q') - 1 - \delta\right] \\
&\leq a_0 h^2(R,Q) - a_1 h^2(R,Q') \\
&\quad - (1-\delta)\left[\delta(a_0 - a_1) + a_0(1-\delta) - a_1 - 1 - \delta\right] \\
&= a_0 h^2(R,Q) - a_1 h^2(R,Q') - (1-\delta)\left[a_0 - (a_1+1)(1+\delta)\right].
\end{aligned}$$

Our conclusion follows since $a_0 - (a_1+1)(1+\delta) \geq a_0 - 2(a_1+1) \geq 0$ in view of our assumption on a_0 and a_1. □

Let us now proceed with the proof of Proposition 2. For all $i \in \{1,\ldots,n\}$, let ν_i be a privileged probability measure for $\mathscr{Q}_i = \{Q_i,\ \mathbf{Q} \in \boldsymbol{\mathscr{Q}}\}$, which means that, for all $A \in \mathscr{B}_i$,

$$\nu_i(A) = 0 \quad \Longleftrightarrow \quad Q_i(A) = 0 \ \text{ for all } Q_i \in \mathscr{Q}_i. \tag{73}$$

Such a dominating probability measure exists as soon as $\mathscr{Q}_i$ is separable with respect to the Hellinger distance which is the case here since $\mathscr{Q}_i$ is countable. We may therefore consider a representation $(\boldsymbol{\nu}, \boldsymbol{\mathcal{T}})$ of $\boldsymbol{\mathscr{Q}}$ based on $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_n)$. Let $(\boldsymbol{\mu}, \boldsymbol{\mathcal{Q}})$ be a an alternative one. For each i, $\nu_i \ll \mu_i$ and μ_i decomposes (in a unique way) as $\mu_i = \mu_i' + \mu_i''$ where $\mu_i' \ll \nu_i$, hence $\mu_i' = m_i \cdot \nu_i$ for some nonnegative function m_i on $\mathscr{X}_i$, and μ_i'' is orthogonal to ν_i. Consequently, for all $i \in \{1,\ldots,n\}$ and $Q_i \in \mathscr{Q}_i$

$$Q_i = q_i \cdot \mu_i = q_i m_i \cdot \nu_i + q_i \cdot \mu_i'' = t_i \cdot \nu_i \ \text{ with } \ q_i \in \mathcal{Q}_i \text{ and } t_i \in \mathcal{T}_i.$$

Since $Q_i \ll \nu_i$, $q_i = 0$ μ_i''-a.e. and $t_i = q_i m_i$ ν_i-a.s. Moreover $Q_i(\{m_i = 0\}) = 0$ for all $Q_i \in \mathscr{Q}_i$, hence $m_i > 0$ ν_i-a.s. If $Q_i' = t_i' \cdot \nu_i = q_i' \cdot \mu_i \in \mathscr{Q}_i$ then $t_i' = q_i' m_i$ ν_i-a.s. and

$$\psi\left(\sqrt{\frac{q_i'}{q_i}}\right) = \psi\left(\sqrt{\frac{q_i' m_i}{q_i m_i}}\right) = \psi\left(\sqrt{\frac{t_i'}{t_i}}\right) \quad \nu_i\text{-a.s.} \tag{74}$$

Since $q_i = q_i' = 0$ μ_i''-a.e., the first equality in (74) is also true μ_i''-a.e. for all $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$ and $i \in \{1, \dots, n\}$. Consequently, for any probability measure $\mathbf{R} \in \mathscr{P}$ such that $\mathbf{R} \ll \boldsymbol{\mu}$

$$\psi\left(\sqrt{\frac{q_i'}{q_i}}\right) = \psi\left(\sqrt{\frac{q_i' m_i}{q_i m_i}}\right) \quad R_i\text{-a.s.} \quad \text{for all } \mathbf{q}, \mathbf{q}' \in \mathcal{Q} \text{ and } i \in \{1, \dots, n\}. \tag{75}$$

If for the representation $(\boldsymbol{\mu}, \mathcal{Q})$ of $\mathscr{Q}$ the inequalities (13) and (14) hold for all $\mathbf{R} \ll \boldsymbol{\mu}$, they hold for all $\mathbf{R} \ll \boldsymbol{\nu} \ll \boldsymbol{\mu}$ and it follows from (74) that they also hold for the representation $(\boldsymbol{\nu}, \mathcal{T})$ and all $\mathbf{R} \ll \boldsymbol{\nu}$. Conversely, if the inequalities (13) and (14) are satisfied for the representation $(\boldsymbol{\nu}, \mathcal{T})$ and all $\mathbf{R} \ll \boldsymbol{\nu}$, they also hold for the representation

$$(\boldsymbol{\nu}, \mathcal{T}'), \quad \mathcal{T}' = \{(q_1 m_1, \dots, q_n m_n),\ \mathbf{q} \in \mathcal{Q}\},$$

and all $\mathbf{R} \ll \boldsymbol{\nu}$ because of (74). Lemma 3 then shows that they also hold for the representation $(\boldsymbol{\nu}, \mathcal{T}')$ and all $\mathbf{R} \in \mathscr{P}$, therefore also for the representation $(\boldsymbol{\mu}, \mathcal{Q})$ and all $\mathbf{R} \ll \boldsymbol{\mu}$ by (75). This completes the proof.

13.2. **Proof of Lemma 1.** Let $(\boldsymbol{\mu}, \mathcal{Q})$ be some representation of $\mathscr{Q}$, $\mathbf{q}, \mathbf{q}'$ two densities belonging to $\mathcal{Q}$, $\mathbf{R}$ an arbitrary element of $\mathscr{P}$ and i some index in $\{1, \dots, n\}$. Decompose R_i in the form $r_i \cdot \mu_i + R_i''$ with R_i'' orthogonal to μ_i. For simplicity we drop the index i and write $\mathscr{X}, \mu, R, R'', Q, Q', r, q, q'$ for $\mathscr{X}_i, \mu_i, R_i, R_i'', Q_i, Q_i', r_i, q_i, q_i'$. Using the inequality, valid for all $\alpha > 0$,

$$r = \left(\sqrt{r} - \sqrt{q} + \sqrt{q}\right)^2 \le (1+\alpha)\left(\sqrt{r} - \sqrt{q}\right)^2 + \left(1 + \alpha^{-1}\right) q,$$

the fact that ψ^2 is bounded by 1 and (16), we obtain that

$$\begin{aligned}
&\int_{\mathscr{X}} \psi^2\left(\sqrt{q'/q}\right) r \, d\mu \\
&\quad \le (1+\alpha) \int_{\mathscr{X}} \psi^2\left(\sqrt{q'/q}\right)\left(\sqrt{r} - \sqrt{q}\right)^2 d\mu + \left(1+\alpha^{-1}\right) \int_{\mathscr{X}} \psi^2\left(\sqrt{q'/q}\right) q \, d\mu \\
&\quad = (1+\alpha) \int_{\mathscr{X}} \psi^2\left(\sqrt{q'/q}\right)\left(\sqrt{r} - \sqrt{q}\right)^2 d\mu + \left(1+\alpha^{-1}\right) \int_{\mathscr{X} \cap \{q>0\}} \psi^2\left(\sqrt{q'/q}\right) q \, d\mu \\
&\quad \le (1+\alpha) \int_{\mathscr{X}} \left(\sqrt{r} - \sqrt{q}\right)^2 d\mu + \left(1+\alpha^{-1}\right) L^2 \int_{\mathscr{X} \cap \{q>0\}} \left(\sqrt{q'/q} - 1\right)^2 q \, d\mu \\
&\quad \le (1+\alpha) \int_{\mathscr{X}} \left(\sqrt{r} - \sqrt{q}\right)^2 d\mu + 2\left(1+\alpha^{-1}\right) L^2 h^2(Q, Q').
\end{aligned}$$

Since

$$2h^2(R, Q) = \int_{\mathscr{X}} \left(\sqrt{r} - \sqrt{q}\right)^2 d\mu + \int_{\mathscr{X}} dR'',$$

we get

$$\begin{aligned}
\int_{\mathscr{X}} \psi^2\left(\sqrt{q'/q}\right) dR &\le \int_{\mathscr{X}} \psi^2\left(\sqrt{q'/q}\right) r \, d\mu + \int_{\mathscr{X}} dR'' \\
&\le 2(1+\alpha) h^2(R, Q) + 2\left(1+\alpha^{-1}\right) L^2 h^2(Q, Q').
\end{aligned}$$

A similar bound holds for $\int_{\mathscr{X}} \psi^2\left(\sqrt{q/q'}\right) dR$. Using (12) and averaging the two bounds, then using $h^2(Q,Q') \le 2\left(h^2(R,Q)+h^2(R,Q')\right)$, we get

$$\begin{aligned}
\int_{\mathscr{X}} \psi^2\left(\sqrt{q'/q}\right) dR &= \int_{\mathscr{X}} \psi^2\left(\sqrt{q/q'}\right) dR \\
&\le (1+\alpha)\left[h^2(R,Q)+h^2(R,Q')\right] + 2\left(1+\alpha^{-1}\right) L^2 h^2(Q,Q') \\
&\le \left[h^2(R,Q)+h^2(R,Q')\right]\left[1+\alpha+4L^2(1+\alpha^{-1})\right].
\end{aligned}$$

The conclusion follows by choosing $\alpha = 2L$.

13.3. **Proof of Proposition 3.** It is clear that both functions are monotone and satisfy (12). Let $\boldsymbol{\mathscr{Q}}$ be an arbitrary ρ-model and $(\boldsymbol{\mu}, \boldsymbol{\mathscr{Q}})$ a representation of it. In view of Proposition 2, it is enough to prove (13) and (14) when $\mathbf{R}=\mathbf{S} \ll \boldsymbol{\mu}$, which we shall assume in the sequel, denoting by $\mathbf{s}$ the corresponding density. As in the proof of Lemma 1, we fix some $i \in \{1\ldots,n\}$, $\mathbf{q},\mathbf{q}' \in \boldsymbol{\mathscr{Q}}$ and then drop the index i in the notations to establish (13) and (14). Given two densities t,t' on $(\mathscr{X},\mathscr{B},\mu)$ we shall write $h(t,t')$ and $\rho(t,t')$ for the Hellinger distance and the Hellinger affinity between the probabilities $t\cdot\mu$ and $t'\cdot\mu$. The proof will repetedly use that $(a+b)^2 \le (1+\alpha)a^2 + \left(1+\alpha^{-1}\right)b^2$ for all $\alpha>0$.

Case of the function ψ_1. Let $r=(q+q')/2$. Our conventions $0/0=1$ and $a/0=+\infty$ for all $a>0$ imply that the equalities

$$\psi_1\left(\sqrt{\frac{q'}{q}}\right) = \frac{\sqrt{q'}-\sqrt{q}}{\sqrt{q+q'}}1\!\!1_{r>0} = \frac{\sqrt{q'}-\sqrt{q}}{\sqrt{2r}}1\!\!1_{r>0} \tag{76}$$

hold for all densities q,q'. Moreover the concavity of the square root implies that

$$\begin{aligned}
h^2(s,r) &= 1-\rho(s,r) = 1-\int\sqrt{\frac{sq+sq'}{2}}\,d\mu \\
&\le 1-\frac{1}{2}\left[\int\sqrt{sq}\,d\mu+\int\sqrt{sq'}\,d\mu\right] = \frac{1}{2}\left[h^2(s,q)+h^2(s,q')\right].
\end{aligned} \tag{77}$$

Squaring (76), integrating with respect to $S=s\cdot\mu$, using the bound $\left(\sqrt{q'}-\sqrt{q}\right)^2 \le 2r$ and then (77), we get,

$$\begin{aligned}
\int_{\mathscr{X}} \psi_1^2\left(\sqrt{\frac{q'}{q}}\right) s\,d\mu &= \int_{r>0} \frac{\left(\sqrt{q'}-\sqrt{q}\right)^2}{2r}\left(\sqrt{s}-\sqrt{r}+\sqrt{r}\right)^2 d\mu \\
&\le (1+\alpha)\int_{r>0}\frac{\left(\sqrt{q'}-\sqrt{q}\right)^2}{2r}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu \\
&\quad + \left(1+\alpha^{-1}\right)\int_{r>0}\frac{\left(\sqrt{q'}-\sqrt{q}\right)^2}{2r} r\,d\mu \\
&\le 2(1+\alpha)h^2(s,r) + \left(1+\alpha^{-1}\right)h^2(q,q') \\
&\le (1+\alpha)\left[h^2(s,q)+h^2(s,q')\right] + 2\left(1+\alpha^{-1}\right)\left[h^2(s,q)+h^2(s,q')\right].
\end{aligned}$$

Setting $\alpha=\sqrt{2}$ leads to $a_2^2 = 3+2\sqrt{2}$ which proves (14).

The proof of (13) is based on (77) and

$$0 \leq \sqrt{\frac{a+b}{2}} - \frac{\sqrt{a}+\sqrt{b}}{2} \leq \frac{\sqrt{2}-1}{2}\left|\sqrt{a}-\sqrt{b}\right| \quad \text{for all } a,b \geq 0. \tag{78}$$

The concavity of the square root leads to the left-hand side of (78). For the right-hand side, note that $z \mapsto \sqrt{(1+z^2)/2}$ is convex and its graph being under any of its chords, for all $z \in [0,1]$,

$$\sqrt{\frac{1+z^2}{2}} \leq \frac{1}{\sqrt{2}} + z\left(1-\frac{1}{\sqrt{2}}\right) = \frac{1+z}{2} + \frac{\sqrt{2}-1}{2}(1-z).$$

The result follows by applying this inequality to $z = \sqrt{(a \wedge b)/(a \vee b)}$ when $a \vee b \neq 0$, the case $a \vee b = 0$ being trivial.

Let us now turn to the proof of (13). We derive from (76) that

$$\begin{aligned}
&\int_{\mathscr{X}} \psi_1\left(\sqrt{\frac{q'}{q}}\right) s\, d\mu - \int_{r>0} \frac{\sqrt{q'}-\sqrt{q}}{\sqrt{2r}}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu \\
&\quad= \int_{r>0} \frac{\sqrt{q'}-\sqrt{q}}{\sqrt{2r}}\left(\sqrt{s}-\sqrt{r}+\sqrt{r}\right)^2 d\mu - \int_{r>0} \frac{\sqrt{q'}-\sqrt{q}}{\sqrt{2r}}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu \\
&\quad= \int_{r>0} \left(\sqrt{q'}-\sqrt{q}\right)\left(\sqrt{2s}-\sqrt{r/2}\right) d\mu \\
&\quad= \int \left(\sqrt{q'}-\sqrt{q}\right)\left(\sqrt{2s}-\sqrt{r/2}\right) d\mu \\
&\quad= \sqrt{2}\left[\rho(s,q')-\rho(s,q)\right] + \int \frac{\sqrt{q}-\sqrt{q'}}{\sqrt{2}}\sqrt{r}\, d\mu - \int \frac{q-q'}{2\sqrt{2}}\, d\mu \\
&\quad= \sqrt{2}\left[h^2(s,q)-h^2(s,q')\right] + \int \frac{\sqrt{q}-\sqrt{q'}}{\sqrt{2}}\left[\sqrt{\frac{q+q'}{2}} - \frac{\sqrt{q}+\sqrt{q'}}{2}\right] d\mu.
\end{aligned}$$

The inequality $|\sqrt{q}-\sqrt{q'}| \leq \sqrt{2r}$ and (77) imply that

$$\int_{r>0} \frac{\sqrt{q'}-\sqrt{q}}{\sqrt{2r}}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu \leq 2h^2(s,r) \leq h^2(s,q)+h^2(s,q')$$

and (78) yields

$$\int \frac{\sqrt{q}-\sqrt{q'}}{\sqrt{2}}\left[\sqrt{\frac{q+q'}{2}} - \frac{\sqrt{q}+\sqrt{q'}}{2}\right] d\mu \leq \frac{\sqrt{2}-1}{\sqrt{2}} h^2(q,q').$$

Therefore

$$\int_{\mathscr{X}} \psi_1\left(\sqrt{\frac{q'}{q}}\right) s\, d\mu \leq \left(1+\sqrt{2}\right) h^2(s,q) - \left(\sqrt{2}-1\right) h^2(s,q') + \frac{\sqrt{2}-1}{\sqrt{2}} h^2(q,q'),$$

hence

$$\begin{aligned}
\sqrt{2}\int_{\mathscr{X}} \psi_1\left(\sqrt{\frac{q'}{q}}\right) s\, d\mu &\leq \left[\left(2+\sqrt{2}\right) + \left(\sqrt{2}-1\right)(1+\alpha)\right] h^2(s,q) \\
&\quad - \left(\sqrt{2}-1\right)\left[\sqrt{2} - \left(1+\alpha^{-1}\right)\right] h^2(s,q').
\end{aligned}$$

The choice $\alpha = 7.7$ implies (13).

Case of the function ψ_2. Let us set

$$r = \left(\frac{\sqrt{q}+\sqrt{q'}}{\delta}\right)^2 \quad \text{with} \quad \delta^2 = \int_{\mathscr{X}} \left(\sqrt{q}+\sqrt{q'}\right)^2 d\mu = 2\left[1+\rho(q,q')\right] = 4\left[1-\frac{h^2(q,q')}{2}\right],$$

so that r is the density of a probability on $(\mathscr{X}, \mathscr{B})$,

$$\sqrt{2} \le \delta \le 2 \quad \text{and} \quad \frac{2}{\delta} = \frac{1}{\sqrt{1-(1/2)h^2(q,q')}} \ge 1 + \frac{h^2(q,q')}{4}, \tag{79}$$

by the convexity of the map $u \mapsto 1/\sqrt{1-u}$. Consequently,

$$h^2(s,r) = 1 - \int_{\mathscr{X}} \sqrt{sr}\, d\mu = 1 - \frac{1}{\delta}\left[\rho(s,q') + \rho(s,q)\right] \tag{80}$$

$$= 1 - \frac{2}{\delta} + \frac{1}{\delta}\left[h^2(s,q) + h^2(s,q')\right] \le \frac{h^2(s,q)+h^2(s,q')}{\delta} - \frac{h^2(q,q')}{4}. \tag{81}$$

The previous computations with this new value of r lead to

$$\begin{aligned}
\int_{\mathscr{X}} \psi_2^2\left(\sqrt{\frac{q'}{q}}\right) s\, d\mu &= \int_{r>0} \left(\frac{\sqrt{q'}-\sqrt{q}}{\sqrt{q'}+\sqrt{q}}\right)^2 \left(\sqrt{s}-\sqrt{r}+\sqrt{r}\right)^2 d\mu \\
&\le (1+\alpha)\int_{r>0} \left(\frac{\sqrt{q'}-\sqrt{q}}{\sqrt{q'}+\sqrt{q}}\right)^2 \left(\sqrt{s}-\sqrt{r}\right)^2 d\mu \\
&\quad + \left(1+\alpha^{-1}\right)\int_{r>0} \left(\frac{\sqrt{q'}-\sqrt{q}}{\sqrt{q'}+\sqrt{q}}\right)^2 \left(\frac{\sqrt{q}+\sqrt{q'}}{\delta}\right)^2 d\mu \\
&\le 2(1+\alpha)h^2(s,r) + 2\left(1+\alpha^{-1}\right)\delta^{-2}h^2(q,q')
\end{aligned}$$

and by (81),

$$\begin{aligned}
\int_{\mathscr{X}} \psi_2^2\left(\sqrt{\frac{q'}{q}}\right) s\, d\mu &\le 2(1+\alpha)\left[\frac{h^2(s,q)+h^2(s,q')}{\delta} - \frac{h^2(q,q')}{4}\right] + 2\left(1+\alpha^{-1}\right)\frac{h^2(q,q')}{\delta^2} \\
&= 2(1+\alpha)\delta^{-1}\left[h^2(s,q)+h^2(s,q')\right]
\end{aligned}$$

provided that $(1+\alpha)/4 = \delta^{-2}\left(1+\alpha^{-1}\right)$. Solving this equation with respect to α leads to $\alpha = 4\delta^{-2}$ hence $2(1+\alpha)\delta^{-1} = 2\delta^{-3}\left(\delta^2+4\right)$ which is a decreasing function of δ. We then conclude from (79) that $2(1+\alpha)\delta^{-1} \le 3\sqrt{2}$ which gives $a_2^2 = 3\sqrt{2}$.

Let us now turn to the proof of (13), setting

$$\rho_r(S,q) = \frac{1}{2}\left[\int_{r>0} \sqrt{qr}\, d\mu + \int_{r>0} \sqrt{\frac{q}{r}}\, dS\right]. \tag{82}$$

Then

$$\begin{aligned}
\int_{r>0} \sqrt{\frac{q}{r}}\, dS &= \int_{r>0} \sqrt{\frac{q}{r}} s\, d\mu = \int_{r>0} \sqrt{\frac{q}{r}}\left(\sqrt{s}-\sqrt{r}+\sqrt{r}\right)^2 d\mu \\
&= \int_{r>0} \sqrt{\frac{q}{r}}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu + \int_{r>0} \sqrt{qr}\, d\mu + 2\int_{r>0} \sqrt{q}\left(\sqrt{s}-\sqrt{r}\right) d\mu \\
&= \int_{r>0} \sqrt{\frac{q}{r}}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu - \int_{r>0} \sqrt{qr}\, d\mu + 2\int_{r>0} \sqrt{qs}\, d\mu
\end{aligned}$$

so that, since $r=0$ implies $q=0$,

$$\rho_r(S,q)=\rho(s,q)+\frac{1}{2}\int_{r>0}\sqrt{\frac{q}{r}}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu=\rho(s,q)+\frac{\delta}{2}\int_{r>0}\frac{\sqrt{q}}{\sqrt{q'}+\sqrt{q}}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu.$$

Moreover, since q and q' are densities, (82) leads to

$$\begin{aligned}\rho_r(S,q')-\rho_r(S,q) &= \frac{1}{2\delta}\int_{r>0}\left(\sqrt{q'}-\sqrt{q}\right)\left(\sqrt{q'}+\sqrt{q}\right)d\mu+\frac{\delta}{2}\int_{r>0}\frac{\sqrt{q'}-\sqrt{q}}{\sqrt{q'}+\sqrt{q}}\,dS\\ &= \frac{1}{2\delta}\int_{r>0}\left(q'-q\right)d\mu+\frac{\delta}{2}\int_{r>0}\psi_2\left(\sqrt{\frac{q'}{q}}\right)dS\\ &= \frac{\delta}{2}\int_{\mathscr{X}}\psi_2\left(\sqrt{\frac{q'}{q}}\right)dS,\end{aligned}$$

since $q=q'=0$ when $r=0$ and, by convention, $\psi_2(0/0)=\psi_2(1)=0$. Putting everything together, we derive that

$$\begin{aligned}\frac{\delta}{2}\int_{\mathscr{X}}\psi_2\left(\sqrt{\frac{q'}{q}}\right)dS &= \rho(s,q')-\rho(s,q)+\frac{\delta}{2}\int_{r>0}\frac{\sqrt{q'}-\sqrt{q}}{\sqrt{q'}+\sqrt{q}}\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu\\ &= \rho(s,q')-\rho(s,q)+\frac{\delta}{2}\int_{r>0}\psi_2\left(\sqrt{\frac{q'}{q}}\right)\left(\sqrt{s}-\sqrt{r}\right)^2 d\mu.\end{aligned}$$

Since $|\psi_2|$ is bounded by 1 we derive from (80) that

$$\frac{\delta}{2}\int_{\mathscr{X}}\psi_2\left(\sqrt{\frac{q'}{q}}\right)dS\le\rho(s,q')-\rho(s,q)+\delta h^2(s,r)=\delta-2\rho(s,q),$$

hence, by (79),

$$\begin{aligned}\int_{\mathscr{X}}\psi_2\left(\sqrt{\frac{q'}{q}}\right)dS &\le 2\left[1-\frac{2\rho(s,q)}{\delta}\right] \le 2\left[1-\rho(s,q)\left(1+\frac{h^2(q,q')}{4}\right)\right]\\ &= 2\left[h^2(s,q)\left(1+\frac{h^2(q,q')}{4}\right)-\frac{h^2(q,q')}{4}\right]\\ &\le \frac{1}{2}\left[5h^2(s,q)-h^2(q,q')\right].\end{aligned}\tag{83}$$

Since $h(q,q')\ge|h(s,q)-h(s,q')|$, we deduce from (83) with $\alpha=4$ that

$$\begin{aligned}\int_{\mathscr{X}}\psi_2\left(\sqrt{\frac{q'}{q}}\right)dS &\le \frac{1}{2}\left[5h^2(s,q)-\left(h^2(s,q)+h^2(s,q')-2\alpha^{1/2}h(s,q)\alpha^{-1/2}h(s,q')\right)\right]\\ &\le \frac{1}{2}\left[(4+\alpha)h^2(s,q)-(1-\alpha^{-1})h^2(s,q')\right] = 4h^2(s,q)-\frac{3}{8}h^2(s,q')\end{aligned}$$

as claimed.

13.4. **Proof of Proposition 4.** Applying Proposition 50 of Baraud, Birgé and Sart (2016) with $T = \mathscr{B}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y) \subset \boldsymbol{\mathscr{Q}} \cap \mathscr{B}(\mathbf{P},y)$, so that $\log_+(2|T|) \le \mathscr{H}(\boldsymbol{\mathscr{Q}},y)$ and $U_{i,t} = \psi\left(\sqrt{(q_i/\overline{p}_i)(X_i)}\right) \in [-1,1]$, for which one may take $b=1$ and $v^2 = a_2^2y^2$ by (14), we obtain that for all $y>0$ and $\mathbf{P},\overline{\mathbf{P}} \in \mathscr{P}$

$$\mathbf{w}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y) \le \mathscr{H}(\boldsymbol{\mathscr{Q}},y) + a_2y\sqrt{2\mathscr{H}(\boldsymbol{\mathscr{Q}},y)} = y^2\left[\frac{\mathscr{H}(\boldsymbol{\mathscr{Q}},y)}{y^2} + a_2\sqrt{\frac{2\mathscr{H}(\boldsymbol{\mathscr{Q}},y)}{y^2}}\right].$$

The inequality $x^2 + \sqrt{2}a_2x \le a_1/8$ is satisfied for

$$0 \le x \le \left(a_2/\sqrt{2}\right)\left[\sqrt{1+(\beta/a_2)} - 1\right] = \beta x_0^{-1}.$$

It follows from the definition of $\overline{\eta}$ that, if $\beta y > \overline{\eta}$,

$$\sqrt{\mathscr{H}(\boldsymbol{\mathscr{Q}},y)} \le \beta x_0^{-1}y \qquad \text{hence} \qquad \mathbf{w}^{\boldsymbol{\mathscr{Q}}}\left(\mathbf{P},\overline{\mathbf{P}},y\right) \le (a_1/8)y^2.$$

Consequently $D^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}}) \le \overline{\eta}^2 \vee 1$ by (26). The second bound derives from (30).

13.5. **Proof of Proposition 5.** By (27), we may restrict to the case of $\overline{V} \le n/6$. The proof being similar to that of Theorem 12 in Baraud, Birgé and Sart (2016), we only provide here a sketch of proof of the result. Since ψ is monotone and $\boldsymbol{\mathscr{Q}}$ is VC-subgraph on $\overline{\mathscr{X}}$ with index not larger than $\overline{V}$ so are the set $\left\{\psi\left(\sqrt{\mathbf{q}/\overline{\mathbf{p}}}\right),\ \mathbf{q} \in \boldsymbol{\mathscr{Q}}\right\}$ and its subset

$$\mathscr{F}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y) = \left\{\psi\left(\sqrt{\mathbf{q}/\overline{\mathbf{p}}}\right)\,\middle|\, \mathbf{Q} \in \mathscr{B}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y)\right\}. \tag{84}$$

Since the elements of $\mathscr{F}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y)$ are bounded by 1, it follows from Theorem 2.6.7 in van der Vaart and Wellner (1996) that, for some numerical constant K and any probability measure R on $\overline{\mathscr{X}}$, the $\mathbb{L}_2(R)$-entropy $\mathscr{H}(\mathscr{F}^{\boldsymbol{\mathscr{Q}}},R,\cdot)$ of $\mathscr{F}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y)$ satisfies for all $\varepsilon \in (0,1)$

$$\mathscr{H}(\mathscr{F}^{\boldsymbol{\mathscr{Q}}},R,\varepsilon) \le \log\left(K\overline{V}(16e)^{\overline{V}}\right) + 2(\overline{V}-1)\log(1/\varepsilon) \le 2\overline{V}\log(A/\varepsilon)$$

for some numerical constant $A \ge 2e$. Applying Lemma 49 in Baraud, Birgé and Sart (2016) to the random variables $(i,X_i) \in \overline{\mathscr{X}}$, $\mathscr{F} = \mathscr{F}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y)$, $v^2 = a_2^2y^2$ by (18) and $\overline{\mathscr{H}}(z) = 2\overline{V}\log_+(Az)$ with $z \ge 1/2$ (so that $L \le 3/2$, according to the proof of Theorem 12 in Baraud, Birgé and Sart (2016)), we get, for some numerical constant $C_0 > 0$,

$$\mathbf{w}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y) \le C_0\left[a_2y\sqrt{H} + H\right] \quad \text{with} \quad H = \overline{\mathscr{H}}\left(\frac{\sqrt{n}}{2a_2y}\bigvee\frac{1}{2}\right).$$

Let $D \ge \overline{V}$ to be chosen later on and $y \ge \beta^{-1}\sqrt{D}$. Since $\beta \le 1$, $a_2 \ge 1$, $\overline{V} \le n$ and $A \ge 2e$ we deduce that $y \ge \sqrt{\overline{V}}$,

$$H \le \overline{H} = \overline{\mathscr{H}}\left(\frac{\sqrt{n}}{2\sqrt{\overline{V}}}\right) = 2\overline{V}\log\left(\frac{A\sqrt{n}}{2\sqrt{\overline{V}}}\right) \quad \text{and} \quad \overline{H} \ge 2\overline{V}.$$

For all $y \ge \beta^{-1}\sqrt{D}$,

$$\begin{aligned}\mathbf{w}^{\boldsymbol{\mathscr{Q}}}(\mathbf{P},\overline{\mathbf{P}},y) &\le C_0\left[a_2y\sqrt{\overline{H}} + \overline{H}\right] = \frac{a_1}{8}y^2C_0\left[\frac{8a_2}{a_1}\frac{\sqrt{\overline{H}}}{y} + \frac{8}{a_1}\frac{\overline{H}}{y^2}\right]\\ &\le \frac{a_1}{8}y^2C_0\left[\frac{8a_2\beta}{a_1}\sqrt{\frac{\overline{H}}{D}} + \frac{8\beta^2}{a_1}\frac{\overline{H}}{D}\right].\end{aligned}$$

Since $\beta = a_1/(4a_2)$ and $a_2^2 \geq 6a_1$, we derive that

$$\mathbf{w}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) \leq \frac{a_1}{8} y^2 C_0 \left[2\sqrt{\frac{\overline{H}}{D}} + \frac{a_1}{2a_2^2} \frac{\overline{H}}{D} \right] \leq \frac{a_1}{8} y^2 C_0 \left[2\sqrt{\frac{\overline{H}}{D}} + \frac{\overline{H}}{12D} \right].$$

The inequality $2u + u^2/12 \leq C_0^{-1}$ is statisfied for $u \in [0, \overline{u}]$ with

$$\overline{u}^{-1} = C_0 \left(\sqrt{1 + \frac{1}{12C_0}} + 1 \right),$$

we deduce that for

$$D = \left(\overline{u}^{-2} \vee \frac{1}{2} \right) \overline{H} \geq \max\{\overline{u}^{-2}\overline{H}, \overline{V}\}$$

and all $y \geq \beta^{-1}\sqrt{D}$, $\mathbf{w}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) \leq a_1 y^2/8$. We conclude by (26).

13.6. **Proof of Proposition 7.** The function ψ being monotone, it follows from Proposition 3 of Baraud (2016) that the class of functions $\mathscr{F}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y)$ defined by (84) which satisfies

$$\mathscr{F}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) \subset \{\psi(g),\ g \in (\mathscr{Q}/\overline{\mathbf{p}})\}$$

is weak VC-major on $\overline{\mathscr{X}}$ with dimension not larger than $d \geq 1$. Besides, since ψ takes its values in $[-1, 1]$, $\mathscr{F}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y)$ is uniformly bounded by 1. Applying Corollary 1 of Baraud (2016) to $\mathscr{F}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y)$ with $b = 1$ and $\sigma^2 = (a_2^2 y^2/n) \wedge 1$ (because of (18)) and setting

$$\overline{\Gamma}(d) = \log \left(2 \sum_{j=0}^{d} \binom{n}{j} \right) \leq \log 2 + d \log \left(\frac{en}{d} \right) \leq dL \quad \text{with} \quad L = \log \left(\frac{e^2 n}{d} \right) \geq 2.$$

we obtain

$$\begin{aligned}
\mathbf{w}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) &\leq 4\sqrt{2n\overline{\Gamma}(d)} \times \sigma \log \left(\frac{e}{\sigma} \right) + 16\overline{\Gamma}(d) \leq 4\sqrt{2ndL} \times \sigma \log \left(\frac{e}{\sigma} \right) + 16dL \\
&\leq 4a_2 y \sqrt{2dL} \log \left(e\sqrt{\frac{n}{a_2^2 y^2} \vee 1} \right) + 16dL.
\end{aligned}$$

Let $D \geq d$ to be chosen later on. Since $a_2 \geq 1$, $\beta = a_1/(4a_2) \leq 1$ and $d \leq n$, for all $y \geq \beta^{-1}\sqrt{D} \geq \sqrt{d}$,

$$\log \left(e\sqrt{\frac{n}{a_2^2 y^2} \vee 1} \right) \leq \log \left(e\sqrt{\frac{n}{d}} \right) = \frac{L}{2}$$

hence, since $L \geq 2$,

$$\begin{aligned}
\mathbf{w}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) &\leq 2a_2 y \sqrt{2dL^3} + 16dL \leq 2a_2 y \sqrt{2dL^3} + 4dL^3 \\
&= \frac{a_1}{8} y^2 \left[\frac{16a_2}{a_1} \sqrt{\frac{dL^3}{y^2}} + \frac{32}{a_1} \frac{2dL^3}{y^2} \right] \leq \frac{a_1}{8} y^2 \left[\frac{16a_2\beta}{a_1} \sqrt{\frac{2dL^3}{D}} + \frac{32\beta^2}{a_1} \frac{dL^3}{D} \right] \\
&= \frac{a_1}{8} y^2 \left[4\sqrt{\frac{2dL^3}{D}} + \frac{2a_1}{a_2^2} \frac{dL^3}{D} \right] \leq \frac{a_1}{8} y^2 \left[4\sqrt{\frac{2dL^3}{D}} + \frac{dL^3}{3D} \right]
\end{aligned}$$

since $a_2^2 \geq 6a_1$. The inequality $4\sqrt{2}u + u^2/3 \leq 1$ is satisfied for $u \in [0, \overline{u}]$ with $\overline{u} = 1/\sqrt{33}$ and consequently, for $D = 33dL^3 \geq d$ and all $y \geq \beta^{-1}\sqrt{D}$, $\mathbf{w}^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) \leq a_1 y^2/8$. The conclusion then follows from (26).

13.7. **Proof of Proposition 9.** Let $(x_1, u_1), \ldots, (x_n, u_n)$ be $n = 2k+1$ points of $\mathscr{X} \times \mathbb{R}$ such that $x_1 \le x_2 \le \ldots \le x_n$. For all $j \in \{1, \ldots, k\}$, let i_j be an index in $\{2j-1, 2j\}$ such that $u_{i_j} = \max\{u_{2j-1}, u_{2j}\}$ and $\mathcal{K} = \{i_j, j \in \{1, \ldots, k\}\} \cup \{2k+1\}$. Let us prove that the subset $\{(x_i, u_i),\ i \in \mathcal{K}\}$ cannot be picked up by the subgraphs of the functions $f \in \mathscr{F}_k$. For any $f \in \mathscr{F}_k$, there exists a partition $\mathscr{I} = \mathscr{I}(f)$ of $\mathbb{R}$ into at most k intervals on which f is based. Since $n > 2k$, there exists at least one interval $I \in \mathscr{I}$ such that I contains three consecutive points x_i and among these three points, there exist two points $x_i, x_{i'}$ with indices $i \in \mathcal{K}$ and $i' \notin \mathcal{K}$ such that either (i, i') or (i', i) is of the form $(2j-1, 2j)$ for some $j \in \{1, \ldots, k\}$. Since f is piecewise constant on the elements on $\mathscr{I}$ and $u_{i'} \le u_i$ whenever the subgraph of f picks up (x_i, u_i) then it also picks the point $(x_{i'}, u_{i'})$. Hence, no subgraph of $f \in \mathscr{F}_k$ picks up the subset $\{(x_i, u_i),\ i \in \mathcal{K}\}$.

13.8. **Proof of Proposition 10.** Let us prove *i*). The linear span V of $\{g_1, \ldots, g_J\}$ is VC-subgraph with VC-index not larger than $J+2$ by Lemma 2.6.15 of van der Vaart and Wellner (1996). The function $u \mapsto e^u$ being increasing, it follows from Proposition 42, *ii*) of Baraud, Birgé and Sart (2016) that the class $\mathcal{F} = \{e^v, v \in V\}$ is also VC-subgraph with index not larger than $J+2$ and $\mathscr{Q} \subset \mathcal{F}$ as well, which concludes the proof of *i*).

By *i*), the families $\mathscr{Q}_I$ with $I \in \mathscr{I}$ are VC-subgraph on I with indices not larger than $J+2$ and since $\mathscr{I}$ is a partition of $\mathscr{X}$ with cardinality not larger than k, we deduce from Baraud and Birgé (2016)[Lemma 5.3)] that $\mathscr{Q}$ is VC-subgraph with index not larger than $k(J+2)$ which proves *ii*).

To prove *iii*) we fix some $\overline{p} \in \mathscr{Q}$. By assumption, the element $q \in \mathscr{Q}$ consists of functions which are of the form (32) on a partition $\mathscr{I}(q)$ of $\mathscr{X}$ into at most k intervals. Since

$$\mathscr{I}(q) \vee \mathscr{I}(\overline{p}) = \{I \cap I',\ I \in \mathscr{I}(q),\ I' \in \mathscr{I}'(\overline{p})\}$$

is a partition of $\mathscr{X}$ into at most $2k$ intervals and since on each element of such partition $q/\overline{p}$ is of the form (32), the class $(\mathscr{Q}/\overline{p})$ is a $(2k)$-piecewise exponential family based on J functions. Consequently, to prove *iii*) it suffices to show that a K-piecewise exponential family based on J functions is weak-VC major with dimension not larger than $\lceil 4.7K(J+2) \rceil$ and to apply the result with $K = 2k$. This is precisely the aim of the following proposition.

Proposition 13. *If $\mathscr{F}$ is a K-piecewise exponential family based on J functions on a non-trivial interval $\mathscr{X} \subset \mathbb{R}$, it is weak VC-major with dimension not larger than $d = \lceil 4.7K(J+2) \rceil$.*

Proof. Let $u \in \mathbb{R}$. If $u \le 0$, $\mathscr{C}_u(\mathscr{F})$ is reduced to the singleton $\{\mathscr{X}\}$ and is therefore VC on $\mathscr{X}$ with dimension $0 < d$. We may therefore assume from now on that $u > 0$. Let $x_1, \ldots, x_k$ be $k > d$ arbitrary points in $\mathscr{X}$. With no loss of generality we may assume that $x_1 < \ldots < x_k$. We need to prove that $\mathscr{C}_u(\mathscr{F})$ cannot shatter $\{x_1, \ldots, x_k\}$. To do so, it suffices to prove that

$$|\{\{i \in \mathcal{I}, f(x_i) > u\},\ f \in \mathscr{F}\}| < 2^k \quad \text{with} \quad \mathcal{I} = \{1, \ldots, k\}.$$

A partition $\mathscr{I}$ of $\mathscr{X}$ with cardinality not larger than $K \ge 1$ provides a partition of $\mathcal{I}$ into $L \le K$ non-void subsets $\{\mathcal{I}_\ell,\ \ell = 1, \ldots, L\}$ consisting of consecutive integers and each such partition is determined by a sequence $i_1 = 1 < i_2 < \ldots < i_L$ where i_ℓ denotes the first element of $\mathcal{I}_\ell$. For a given $L \in \{1, \ldots, K\}$, the number N_L of possible partitions of $\mathcal{I}$ into

L subsets is therefore the number of choices for $\{i_2, \ldots, i_L\}$, that is,

$$N_L = \binom{k-1}{L-1}. \tag{85}$$

We have seen in i) that the class

$$\mathscr{G} = \left\{ \exp\left(\sum_{j=1}^{J} \beta_j g_j\right),\ \beta_1, \ldots, \beta_J \in \mathbb{R} \right\}$$

is VC-subgraph with dimension (VC-dimension = VC-index−1) not larger than $J+1$, it is therefore weak VC-major with dimension not larger than $J+1$ by Proposition 1 of Baraud (2016). Hence the class of subsets $\mathscr{C}_u(\mathscr{G})$ is VC with dimension not larger than $J+1$. Given a partition $\{\mathcal{I}_\ell,\ \ell = 1, \ldots, L\}$ of $\{1, \ldots, k\}$, it follows from Sauer's Lemma (see Sauer (1972)) that

$$|\{\{i \in \mathcal{I}_\ell,\ g(x_i) > u\},\ g \in \mathscr{G}\}| \leq \left(\frac{e|\mathcal{I}_\ell|}{J+1}\right)^{J+1} \quad \text{for all} \quad \ell = 1, \ldots, L.$$

Hence,

$$|\{\{i \in \mathcal{I},\ f(x_i) > u\},\ f \in \mathscr{F}\}| \leq \sum_{L=1}^{K} \sum_{\mathcal{I}_1, \ldots, \mathcal{I}_L} \prod_{\ell=1}^{L} \left(\frac{e|\mathcal{I}_\ell|}{J+1}\right)^{J+1} \tag{86}$$

where the second sum varies among all possible partitions $\{\mathcal{I}_1, \ldots, \mathcal{I}_L\}$ of size L of $\mathcal{I}$ into consecutive integers. Using the concavity of the logarithm, the fact that $\sum_{\ell=1}^{L} |\mathcal{I}_\ell| = k$ and (85) we get,

$$\begin{aligned}
\sum_{L=1}^{K} \sum_{\mathcal{I}_1, \ldots, \mathcal{I}_L} \prod_{\ell=1}^{L} \left(\frac{e|\mathcal{I}_\ell|}{J+1}\right)^{J+1} &= \sum_{L=1}^{K} \sum_{\mathcal{I}_1, \ldots, \mathcal{I}_L} \left(\frac{e}{J+1}\right)^{L(J+1)} \exp\left[(J+1)L \times \frac{1}{L} \sum_{\ell=1}^{L} \log |\mathcal{I}_\ell|\right] \\
&\leq \sum_{L=1}^{K} \sum_{\mathcal{I}_1, \ldots, \mathcal{I}_L} \left(\frac{e}{J+1}\right)^{L(J+1)} \exp\left[(J+1)L \log\left(\frac{k}{L}\right)\right] \\
&= \sum_{L=1}^{K} N_L \left(\frac{ek}{L(J+1)}\right)^{L(J+1)}.
\end{aligned}$$

Since the function $x \mapsto (ek/x)^x$ is increasing on the interval $(0, k]$ and $k > d > K(J+1)$, then

$$\left(\frac{ek}{L(J+1)}\right)^{L(J+1)} \leq \left(\frac{ek}{K(J+1)}\right)^{K(J+1)} \quad \text{for all } L \in \{1, \ldots, K\},$$

so that

$$\sum_{L=1}^{K} \sum_{\mathcal{I}_1, \ldots, \mathcal{I}_L} \prod_{\ell=1}^{L} \left(\frac{e|\mathcal{I}_\ell|}{J+1}\right)^{J+1} \leq \left(\sum_{L=1}^{K} N_L\right) \left(\frac{ek}{K(J+1)}\right)^{K(J+1)}. \tag{87}$$

Since $\sum_{j=0}^{m} \binom{p}{j} \leq (ep/m)^m$ for $0 \leq m \leq p$, it follows from (85) that

$$\sum_{L=1}^{K} N_L = \sum_{L=1}^{K} \binom{k-1}{L-1} < \sum_{j=0}^{K} \binom{k}{j} \leq \left(\frac{ek}{K}\right)^K,$$

and (87) becomes, using again the concavity of the logarithm,

$$\begin{aligned}\frac{1}{K(J+2)}\log\left(\sum_{L=1}^{K}\sum_{\mathcal{I}_1,\ldots,\mathcal{I}_L}\prod_{\ell=1}^{L}\left[\frac{e|\mathcal{I}_\ell|}{J+1}\right]^{J+1}\right) &\le \frac{1}{(J+2)}\log\left(\frac{ek}{K}\right)+\frac{J+1}{J+2}\log\left(\frac{ek}{K(J+1)}\right)\\ &\le \log\left(\frac{2ek}{K(J+2)}\right).\end{aligned}$$

Finally (86) leads to

$$\frac{1}{k}\log\Big(\left|\{\{i\in\mathcal{I}, f(x_i)>u\},\ f\in\mathscr{F}\}\right|\Big)\le\frac{K(J+2)}{k}\log\left(\frac{2ek}{K(J+2)}\right).$$

One can easily check that the function $x\mapsto x^{-1}\log(2ex)$ is decreasing for $x>1/2$ and smaller than $\log 2$ for $x>4.7$ which implies that

$$\left|\{\{i\in\mathcal{I}, f(x_i)>u\},\ f\in\mathscr{F}\}\right|<2^k\quad\text{for } k>4.7K(J+2).$$

The conclusion follows. □

13.9. **Proof of Proposition 11.** Let $\boldsymbol{\nu}$ be a privileged probability for $\boldsymbol{\mathscr{Q}}$, that is a probability in $\mathscr{P}$ satisfying (73) for all $i\in\{1,\ldots,n\}$ leading to a representation $\mathcal{R}_1=(\boldsymbol{\nu},\boldsymbol{\mathcal{T}})$ of $\boldsymbol{\mathscr{Q}}$, and let $\mathcal{R}_2=(\boldsymbol{\mu},\boldsymbol{\mathcal{Q}})$ be an arbitrary representation of $\boldsymbol{\mathscr{Q}}$. It suffices to prove that

$$\sup_{y>0}\left|w(\mathcal{R}_1,\mathbf{P},\overline{\mathbf{P}},y)-w(\mathcal{R}_2,\mathbf{P},\overline{\mathbf{P}},y)\right|\le 8\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})\quad\text{for all } \mathbf{P}\in\mathscr{P}\text{ and }\overline{\mathbf{P}}\in\boldsymbol{\mathscr{Q}}. \tag{88}$$

Let us fix some $\mathbf{P}\in\mathscr{P}$ and $\overline{\mathbf{P}}\in\boldsymbol{\mathscr{Q}}$. For all $i\in\{1,\ldots,n\}$, μ_i and P_i can be decomposed (in a unique way) into a part which is absolutely continuous with respect to ν_i, denoted $\mu_i'=m_i\cdot\nu_i$ and $P_i'=p_i\cdot\nu_i$ respectively and a part which is orthogonal to ν_i, denoted μ_i'' and P_i'' respectively. For all $\mathbf{Q}=\mathbf{q}\cdot\boldsymbol{\mu}=\mathbf{t}\cdot\boldsymbol{\nu}\in\boldsymbol{\mathscr{Q}}$ and $i\in\{1,\ldots,n\}$,

$$Q_i=q_i\cdot\mu_i=q_im_i\cdot\nu_i+q_i\cdot\mu_i''=t_i\cdot\nu_i.$$

Let

$$A_i(\mathbf{Q})=\{t_i=q_im_i\text{ and }m_i>0\}\subset\mathscr{X}_i\qquad\text{and}\qquad A_i=\bigcup_{\mathbf{Q}\in\boldsymbol{\mathscr{Q}}}A_i(\mathbf{Q}).$$

Arguing as in the proof of Proposition 2, we derive that $\nu_i\left(A_i(\mathbf{Q})\right)=1$ for all $\mathbf{Q}\in\boldsymbol{\mathscr{Q}}$ and since $\boldsymbol{\mathscr{Q}}$ is countable, $\nu_i(A_i^c)=0$ for $i\in\{1,\ldots,n\}$. In particular, if $\overline{P}_i=\overline{p}_i\cdot\mu_i=\overline{s}_i\cdot\nu_i$, then $\overline{s}_i=\overline{p}_im_i$ on A_i. It then follows from the definition of A_i that everywhere

$$\psi\left(\sqrt{\frac{q_i}{\overline{p}_i}}\right)\mathbb{1}_{A_i}=\psi\left(\sqrt{\frac{t_i}{\overline{s}_i}}\right)\mathbb{1}_{A_i}\quad\text{for all } Q_i\in\mathscr{Q}_i. \tag{89}$$

For $\mathbf{Q}=\mathbf{q}\cdot\boldsymbol{\mu}=\mathbf{t}\cdot\boldsymbol{\nu}\in\boldsymbol{\mathscr{Q}}$, let

$$\mathbf{T}_{\mathbf{A}}(\mathbf{x},\overline{\mathbf{p}},\mathbf{q})=\sum_{i=1}^{n}\psi\left(\sqrt{\frac{q_i}{\overline{p}_i}}(x_i)\right)\mathbb{1}_{A_i}(x_i),\quad \mathbf{T}_{\mathbf{A}^c}(\mathbf{x},\overline{\mathbf{p}},\mathbf{q})=\sum_{i=1}^{n}\psi\left(\sqrt{\frac{q_i}{\overline{p}_i}}(x_i)\right)\mathbb{1}_{A_i^c}(x_i)$$

and define $\mathbf{T}_{\mathbf{A}}(\mathbf{x},\overline{\mathbf{s}},\mathbf{t})$ and $\mathbf{T}_{\mathbf{A}^c}(\mathbf{x},\overline{\mathbf{s}},\mathbf{t})$ in the same way. Then

$$\begin{aligned}\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right] &= \mathbf{T}_{\mathbf{A}}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}_{\mathbf{A}}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]\\ &\quad+\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right],\end{aligned}$$

with a similar decomposition for $\mathbf{T}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})\right]$. Since $|\psi|\le 1$,

$$\left|\mathbf{T}_{\mathbf{A}^c}(\mathbf{x},\overline{\mathbf{p}},\mathbf{q})\right|\le\sum_{i=1}^{n}1\!\!1_{A_i^c}(x_i)\qquad\text{and}\qquad\mathbb{E}_{\mathbf{P}}\left[\sum_{i=1}^{n}1\!\!1_{A_i^c}(X_i)\right]=\sum_{i=1}^{n}P_i(A_i^c).$$

Besides, since $\overline{P}_i\ll\nu_i$, $\overline{P}_i(A_i^c)=0$, hence, if the measure λ_i dominates both P_i and $\overline{P}_i$,

$$2h^2(\overline{P}_i,P_i)\ge\int_{A_i^c}\left(\sqrt{\frac{d\overline{P}_i}{d\lambda_i}}-\sqrt{\frac{dP_i}{d\lambda_i}}\right)^2d\lambda_i=P_i(A_i^c).$$

This implies that

$$\left|\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]\right|\le\sum_{i=1}^{n}1\!\!1_{A_i^c}(X_i)+2\mathbf{h}^2(\overline{\mathbf{P}},\mathbf{P}),$$

with the same bound for $\left|\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})\right]\right|$. For all $\mathbf{Q}\in\boldsymbol{\mathscr{Q}}$, $\mathbf{T}_{\mathbf{A}}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})=\mathbf{T}_{\mathbf{A}}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})$ by (89). Hence

$$\begin{aligned}
&\Big|\left|\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]\right|-\left|\mathbf{T}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})\right]\right|\Big|\\
&\quad\le\Big|\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]-\big(\mathbf{T}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})\right]\big)\Big|\\
&\quad=\left|\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]-\big(\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})\right]\big)\right|\\
&\quad\le\left|\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right]\right|+\left|\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})-\mathbb{E}_{\mathbf{P}}\left[\mathbf{T}_{\mathbf{A}^c}(\boldsymbol{X},\overline{\mathbf{s}},\mathbf{t})\right]\right|\\
&\quad\le 2\sum_{i=1}^{n}1\!\!1_{A_i^c}(X_i)+4\mathbf{h}^2(\overline{\mathbf{P}},\mathbf{P}),
\end{aligned}$$

which implies that (88) holds since

$$\sup_{y>0}\left|w(\mathcal{R}_1,\mathbf{P},\overline{\mathbf{P}},y)-w(\mathcal{R}_2,\mathbf{P},\overline{\mathbf{P}},y)\right|\le 2\sum_{i=1}^{n}P_i(A_i^c)+4\mathbf{h}^2(\overline{\mathbf{P}},\mathbf{P})\le 8\mathbf{h}^2(\overline{\mathbf{P}},\mathbf{P})$$

and concludes our proof.

13.10. **Proof of Proposition 12.** It relies on the following result — see Proposition 45 of Baraud, Birgé and Sart (2016) — which presents an extension of a version of Talagrand's Theorem on the suprema of empirical processes that is proved in Massart (2007).

Proposition 14. *Let T be some finite or countable 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,*

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

then, for all positive c and x,

$$\mathbb{P}\left[Z\le(1+c)\mathbb{E}(Z)+(8b)^{-1}cv^2+2\left(1+8c^{-1}\right)bx\right]\ge 1-e^{-x}.\tag{90}$$

Let $\xi>0$, $\alpha>\delta>1$, $\overline{\mathbf{P}}\in\boldsymbol{\mathscr{Q}}$ and m in $\mathcal{M}$ be fixed as well as a representation $\mathcal{R}(\boldsymbol{\mathscr{Q}})$ for the construction of $\mathbf{T}$. Setting

$$\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})=\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\right],$$

we recall from Proposition 11 that, whatever $m \in \mathcal{M}$,

$$\mathbb{E}_{\mathbf{P}}\left[\sup_{\mathbf{Q}\in\mathscr{B}^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}},y)} |\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})|\right] \leq \mathbf{w}^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}},y) + 16\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right) \quad \text{for all } y > 0. \tag{91}$$

For $j \in \mathbb{N}$ we define

$$x_0(m) = \left(\Delta(\mathscr{Q}_m) + \xi + \vartheta\right) \bigvee \left(\tau^{-1}c_0\widetilde{D}_m(\mathbf{P},\overline{\mathbf{P}})\right), \quad x_j(m) = \delta^j x_0(m), \quad y_j^2 = c_0{}^{-1}\tau x_j(m),$$

$$B_j^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}}) = \left\{\mathbf{Q} \in \mathscr{Q}_m \text{ such that } y_j^2 < \mathbf{h}^2(\mathbf{P},\mathbf{Q}) + \mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}}) \leq y_{j+1}^2\right\}$$

and

$$Z_j^{\mathscr{Q}_m}(\boldsymbol{X},\overline{\mathbf{p}}) = \sup_{\mathbf{Q}\in B_j^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}})} |\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})|.$$

Let us drop, for a while, the dependency of the quantities $x_j(m)$ with respect to m for $j \geq 0$. Since $B_j^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}}) \subset \mathscr{B}^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}},y_{j+1})$, it follows from (91) that

$$\mathbb{E}\left[Z_j^{\mathscr{Q}_m}(\boldsymbol{X},\overline{\mathbf{p}})\right] \leq \mathbf{w}^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}},y_{j+1}) + 16\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right). \tag{92}$$

For each $j \geq 0$, we may apply Proposition 14 to the supremum $Z_j^{\mathscr{Q}_m}(\boldsymbol{X},\overline{\mathbf{p}})$ by taking $T = B_j^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}})$ (which is countable as a subset of $\mathscr{Q}_m$) and

$$U_{i,\mathbf{q}} = \psi\left(\sqrt{\frac{q_i}{\overline{p}_i}}(X_i)\right) - \mathbb{E}\left[\psi\left(\sqrt{\frac{q_i}{\overline{p}_i}}(X_i)\right)\right] \quad \text{for all } i = 1,\ldots,n. \tag{93}$$

For such a choice, the assumptions of Proposition 14 are met with $b = 2$ (since ψ is bounded by 1) and $v^2 = a_2^2 y_{j+1}^2$ (by (14) and the definition of $B_j^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}})$). It therefore follows from (90) that, for all $c > 0$, with probability at least $1 - e^{-x_j}$ and for all $\mathbf{Q} \in B_j^{\mathscr{Q}_m}(\mathbf{P},\overline{\mathbf{P}})$,

$$\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q}) \leq Z_j^{\mathscr{Q}_m}(\boldsymbol{X},\overline{\mathbf{p}}) \leq (1+c)\mathbb{E}\left[Z_j^{\mathscr{Q}_m}(\boldsymbol{X},\overline{\mathbf{p}})\right] + (ca_2^2y_{j+1}^2/16) + 4\left(1+8c^{-1}\right)x_j. \tag{94}$$

Since $y_{j+1} > \sqrt{\widetilde{D}_m(\mathbf{P},\overline{\mathbf{P}})}$, we derive from (92) and (58) that

$$\mathbb{E}\left[Z_j^{\mathscr{Q}_m}(\boldsymbol{X},\overline{\mathbf{p}})\right] \leq c_0 y_{j+1}^2 + 16\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right).$$

Therefore (94) becomes, since $y_{j+1}^2 = \delta y_j^2$ and $x_j = c_0 y_j^2/\tau$,

$$\begin{aligned}
\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q}) &\leq y_{j+1}^2\left[c_0(1+c) + \frac{ca_2^2}{16}\right] + 4\left(1+8c^{-1}\right)x_j + 16(1+c)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right) \\
&= c_0 y_j^2\left[\delta c B + \frac{4\left(1+8c^{-1}\right)}{\tau} + \delta\right] + 16(1+c)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right).
\end{aligned}$$

Minimizing the bracketed term with respect to c leads to $c = 4\sqrt{2}[\delta B\tau]^{-1/2}$ and

$$\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q}) \leq c_0 y_j^2\left[\frac{4}{\tau} + 8\sqrt{\frac{2\delta B}{\tau}} + \delta\right] + 16\left(1 + \frac{4\sqrt{2}}{\sqrt{\delta B\tau}}\right)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right).$$

Since $\mathbf{h}^2(\mathbf{P},\mathbf{Q})+\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})>y_j^2$ on $B_j^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P},\overline{\mathbf{P}})$, we derive that

$$\begin{aligned}\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-c_0\alpha\left[\mathbf{h}^2(\mathbf{P},\mathbf{Q})+\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})\right]-16\left(1+\frac{4\sqrt{2}}{\sqrt{\delta B\tau}}\right)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right)&\\ <\ c_0y_j^2\left[\frac{4}{\tau}+8\sqrt{\frac{2\delta B}{\tau}}-(\alpha-\delta)\right]&\end{aligned}$$

and the bracketed factor is nonpositive provided that

$$\frac{1}{\tau}\leq 2\delta B\left[\sqrt{1+\frac{\alpha-\delta}{8\delta B}}-1\right]^2,$$

which is our condition (61). Finally, since $\delta>1$, with probability at least $1-e^{-x_j}$ and for all $\mathbf{Q}\in B_j^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P},\overline{\mathbf{P}})$,

$$\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-c_0\alpha\left[\mathbf{h}^2(\mathbf{P},\mathbf{Q})+\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})\right]-16\left(1+\frac{4\sqrt{2}}{\sqrt{B\tau}}\right)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right)<0.$$

Let us now define

$$Z^{\boldsymbol{\mathscr{Q}}_m}(\boldsymbol{X},\overline{\mathbf{p}})=\sup_{\mathbf{Q}\in\mathscr{B}^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P},\overline{\mathbf{P}},y_0)}|\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})|$$

and apply Proposition 14 in a similar way to $Z^{\boldsymbol{\mathscr{Q}}_m}(\boldsymbol{X},\overline{\mathbf{p}})$ with $x=x_0$. We then deduce analogously that, for all $c>0$, with probability at least $1-e^{-x_0}$ and for all $\mathbf{Q}\in\mathscr{B}^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P},\overline{\mathbf{P}},y_0)$,

$$\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})\leq Z^{\boldsymbol{\mathscr{Q}}_m}(\boldsymbol{X},\overline{\mathbf{p}})\leq c_0y_0^2\left[cB+\frac{4\left(1+8c^{-1}\right)}{\tau}+1\right]+16(1+c)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right).$$

With $c=4\sqrt{2/(B\tau)}$, we get, for all $\mathbf{Q}\in\mathscr{B}^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P},\overline{\mathbf{P}},y_0)$ and with probability at least $1-e^{-x_0}$,

$$\begin{aligned}\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})&\leq c_0y_0^2\left(\frac{4}{\tau}+8\sqrt{\frac{2B}{\tau}}+1\right)+16\left(1+\frac{4\sqrt{2}}{\sqrt{B\tau}}\right)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right)\\&=\left(\frac{4}{\tau}+8\sqrt{\frac{2B}{\tau}}+1\right)\tau x_0+16\left(1+\frac{4\sqrt{2}}{\sqrt{B\tau}}\right)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right).\end{aligned}$$

Combining all these bounds and reintroducing the dependency of the x_j with respect to m, we derive that, for all $\mathbf{Q}\in\boldsymbol{\mathscr{Q}}_m$ simultaneously,

$$\begin{aligned}&\mathbf{Z}(\boldsymbol{X},\overline{\mathbf{p}},\mathbf{q})-c_0\alpha\left[\mathbf{h}^2(\mathbf{P},\mathbf{Q})+\mathbf{h}^2(\mathbf{P},\overline{\mathbf{P}})\right]\\&\qquad\leq\left(\frac{4}{\tau}+8\sqrt{\frac{2B}{\tau}}+1\right)\tau x_0(m)+16\left(1+\frac{4\sqrt{2}}{\sqrt{B\tau}}\right)\mathbf{h}^2\left(\mathbf{P},\overline{\mathbf{P}}\right),\end{aligned}$$

with probability at least

$$1-\eta_m\quad\text{with}\quad\eta_m=2\exp\left[-x_0(m)\right]+\sum_{j\geq1}\exp\left[-x_j(m)\right].$$

In order to bound η_m, we observe that $x_j(m) \geq \Delta(\mathscr{Q}_m) + \xi + \vartheta\delta^j$ for all $j \in \mathbb{N}$, hence by (59),

$$\eta_m \leq \exp\left[-\xi - \Delta(\mathscr{Q}_m)\right]\left(2\exp[-\vartheta] + \sum_{j\geq 1}\exp\left[-\vartheta\delta^j\right]\right) \leq \exp\left[-\xi - \Delta(\mathscr{Q}_m)\right].$$

The result finally extends to all $\mathbf{Q} \in \mathscr{Q}$ by summing these bounds over $m \in \mathcal{M}$ and using (23).

References

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794.

Baraud, Y. (2016). Bounding the expectation of the supremum of an empirical process over a (weak) VC-major class. *To appear in Electron. J. Stat.*

Baraud, Y. and Birgé, L. (2016). Rho-estimators for shape restricted density estimation. *To appear in Stochastic Process. Appl.*

Baraud, Y., Birgé, L., and Sart, M. (2016). A new method for estimation and model selection: ρ-estimation. *To appear in Invent. math.*

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

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

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

Cherruault, Y. and Loridan, P. (1973). Méthodes pour la recherche de points de selle. *J. Math. Anal. Appl.*, 42:522–535.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, 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. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic processes and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 1; dedicated to Jerzy Neyman)*, pages 13–54. Academic Press, New York.

Le Cam, L. (1990). Maximum likelihood: An introduction. *Inter. Statist. Review*, 58(2):153–171.

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.

Sart, M. (2015). Estimating the conditional density by histogram type estimators and model selection. http://arxiv.org/abs/1512.07052.

Sauer, N. (1972). On the density of families of sets. *J. Combinatorial Theory Ser. A*, 13:145–147.

van de Geer, S. A. (2000). *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York.

Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351, 06100 Nice, France.

E-mail address: baraud@unice.fr

Sorbonne Universités, UPMC Univ. Paris 06, CNRS UMR 7599, LPMA - Case courrier 188, 75252 Paris Cedex 05, France.

E-mail address: lucien.birge@upmc.fr