Probab. Theory Relat. Fields (2011) 151:353–401
DOI 10.1007/s00440-010-0302-y

Estimator selection with respect to Hellinger-type risks

Yannick Baraud

Received: 11 May 2009 / Revised: 28 April 2010 / Published online: 22 May 2010

Abstract We observe a random measure N and aim at estimating its intensity s. This statistical framework allows to deal simultaneously with the problems of estimating a density, the marginals of a multivariate distribution, the mean of a random vector with nonnegative components and the intensity of a Poisson process. Our estimation strategy is based on estimator selection. Given a family of estimators of s based on the observation of N, we propose a selection rule, based on N as well, in view of selecting among these. Little assumption is made on the collection of estimators and their dependency with respect to the observation N need not be known. The procedure offers the possibility to deal with various problems among which model selection, convex aggregation and construction of T-estimators as studied recently in Birgé (Ann Inst H Poincaré Probab Stat 42(3):273–325, 2006). For illustration, we shall consider the problems of estimation, complete variable selection and selection among linear estimators in possibly non-Gaussian regression settings.

Keywords Estimator selection · Model selection · Variable selection · T-estimator · Histogram · Estimator aggregation · Hellinger loss

Mathematics Subject Classification (2000) Primary 62G05; Secondary 62C12 · 62J05 · 62J12 · 62G07 · 62G35

Y. Baraud (✉)
Laboratoire J-A Dieudonné, Université de Nice Sophia-Antipolis, Parc Valrose,
06108 Nice Cedex 02, France
e-mail: baraud@unice.fr

1 Introduction

1.1 The statistical setting

Let $N_1, \ldots, N_k$ be k independent random measures. Each N_i is defined on an abstract probability space $(\Omega, \mathcal{T}, \mathbb{P})$ and takes its values in the class of positive measures on a measured space $(\mathcal{X}_i, \mathcal{A}_i, \mu_i)$. Besides, we assume that

$$\mathbb{E}[N_i(A)] = \int_A s_i d\mu_i < +\infty, \quad \text{for all } A \in \mathcal{A}_i \tag{1}$$

for some nonnegative and measurable function s_i on $\mathcal{X}_i$. We shall call s_i the intensity of N_i. Equality (1) implies that N_i is finite a.s. and that for all measurable and nonnegative functions f_i on $\mathcal{X}_i$,

$$\mathbb{E}\left[\int_{\mathcal{X}_i} f_i dN_i\right] = \int_{\mathcal{X}_i} f_i s_i d\mu_i. \tag{2}$$

Our aim is to estimate $s = (s_1, \ldots, s_k)$ from the observation of $N = (N_1, \ldots, N_k)$.

Throughout, we shall set $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_k)$, $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_k)$, $\mu = (\mu_1, \ldots, \mu_k)$ and denote by $\mathcal{L}$ the cone of nonnegative and measurable functions t of the form $(t_1, \ldots, t_k)$ where the t_i are nonnegative and integrable functions on $(\mathcal{X}_i, \mathcal{A}_i, \mu_i)$. Moreover, for $f = (f_1, \ldots, f_k) \in \mathcal{L}$ we shall use the notations

$$\int_{\mathcal{X}} f dN = \sum_{i=1}^k \int_{\mathcal{X}_i} f_i dN_i \quad \text{and} \quad \int_{\mathcal{X}} f d\mu = \sum_{i=1}^k \int_{\mathcal{X}_i} f_i d\mu_i.$$

Finally, $\mathcal{L}_0$ will denote a known subset of $\mathcal{L}$ containing the target function s.

This statistical framework allows to deal simultaneously with the more classical ones given below:

Example 1 (Density Estimation) Consider the problem of estimating a density s on $(\mathcal{X}, \mathcal{A}, \mu)$ from the observation of an n-sample $X_1, \ldots, X_n$ with distribution $P_s = s \cdot \mu$. To handle this problem, we shall take $k = 1$, $N = n^{-1} \sum_{i=1}^n \delta_{X_i}$ and $\mathcal{L}_0$ the set of densities on $(\mathcal{X}, \mathcal{A})$ with respect to μ.

Example 2 (Estimation of marginals) Let $X_1, \ldots, X_n$ be independent random variables with values in the measured spaces $(\mathcal{X}_1, \mathcal{A}_1, \mu_1), \ldots, (\mathcal{X}_n, \mathcal{A}_n, \mu_n)$ respectively. We assume that for all i, X_i admits a density s_i with respect to μ_i and our aim is to estimate $s = (s_1, \ldots, s_n)$ from the observation of $X = (X_1, \ldots, X_n)$. We shall deal with this problem by taking $k = n$ and $N_i = \delta_{X_i}$ for $i = 1, \ldots, n$. Note that this setting includes as a particular case that of the regression framework

$$X_i = f_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{3}$$

where the f_i are unknown real numbers and the $\varepsilon_i = X_i - f_i$ are i.i.d. random variables with known distribution q. In this case $s_i(x) = q(x - f_i)$ for all $i = 1, \ldots, n$ and the problem of estimating the densities of the X_i amounts to estimating the shift parameter $f = (f_1, \ldots, f_n)$.

Example 3 Let $X_1, \ldots, X_n$ be n independent, nonnegative and integrable random variables. Our aim is to estimate the function s given by $s(i) = \mathbb{E}(X_i) < +\infty$ for $i \in \mathcal{X} = \{1, \ldots, n\}$ on the basis of the observation $X = (X_1, \ldots, X_n)$. This statistical setting is a particular case of our general one by taking $k = 1$, $\mathcal{A} = \mathcal{P}(\mathcal{X})$, μ the counting measure on $(\mathcal{X}, \mathcal{A})$, $\mathcal{L}_0 = \mathcal{L}$ and N the measure defined for $A \subset \mathcal{X}$ by $N(A) = \sum_{i \in A} X_i$.

Among the marginal distributions of X we have in mind, we mention the Binomial or Gamma among others.

Example 4 (Estimating the intensity of a Poisson process) Consider the problem of estimating the intensity s of a possibly inhomogeneous Poisson process N on a measurable space $(\mathcal{X}, \mathcal{A})$. We shall assume that s is integrable. This statistical setting is a particular case of our general one by taking $k = 1$ and $\mathcal{L}_0 = \mathcal{L}$.

Hereafter, we shall deal with estimators with values in $\mathcal{L}_0$ and to measure their risks, endow $\mathcal{L}_0$ with the distance H defined for t, t' in $\mathcal{L}_0$ by

$$H^2(t, t') = \frac{1}{2} \int_{\mathcal{X}} \left(\sqrt{t} - \sqrt{t'} \right)^2 d\mu = \frac{1}{2} \sum_{i=1}^{k} \int_{\mathcal{X}_i} \left(\sqrt{t_i} - \sqrt{t_i'} \right)^2 d\mu_i .$$

When $k = 1$ and t, t' are densities with respect to μ, H is merely the Hellinger distance h between the corresponding probabilities. Given an estimator $\hat{s}$ of s, i.e. a measurable function of N with $\hat{s} \in \mathcal{L}_0$, we define its risk by $\mathbb{E}[H^2(s, \hat{s})]$.

1.2 An account of the results

We start with an arbitrary collection $\mathcal{E} = \left\{ \hat{s}_\lambda,\ \lambda \in \Lambda \right\}$ of estimators based on N together with a family $\mathbb{S}$ of subsets of $\mathcal{L}_0$. The family $\mathcal{E}$ need not be countable even though we shall assume so in order to avoid measurability problems. In fact, the reader can check that the cardinality of $\mathcal{E}$ will play no role in our results. In contrast, the family $\mathbb{S}$ should be countable (we shall use the word countable for finite or countable) and its complexity is measured by means of a mapping Δ from $\mathbb{S}$ into $[1, +\infty)$ satisfying

$$\Sigma = \sum_{S \in \mathbb{S}} e^{-\Delta(S)} < +\infty. \tag{4}$$

When $\Sigma = 1$, $e^{-\Delta}$ corresponds to a prior distribution on the family $\mathbb{S}$ and gives thus a Bayesian flavor to our statistical procedure. The fact that Δ is assumed to be not smaller than 1 (although one usually assumes $\Delta \geq 0$) is only here to simplify the

presentation of the results. Hereafter, the elements S of $\mathbb{S}$ will be called models and assumed to have finite metric dimensions $D(S)$ (in an appropriate sense).

In the present paper, the problem we consider is that of *estimator selection*. More precisely, our aim is to select some estimator $\hat{s}_{\hat{\lambda}}$ among the collection $\mathcal{E}$ from the same observation N in view of achieving the smallest risk bound over $\mathcal{E}$. For appropriate choices of $\mathcal{E}$ and $\mathbb{S}$, our approach allows us to deal simultaneously with the problems of model selection, (convex) aggregation and construction of T-estimators. As we shall see, very little assumptions on the estimators $\hat{s}_\lambda$ will be required and in fact the way they depend on N need not even be known. Nevertheless, there should be some connections between the families $\mathcal{E}$ and $\mathbb{S}$. Typically, each $\hat{s}_\lambda$ should belong (or at least should be close enough) to $\bigcup_{S\in\mathbb{S}} S$. More precisely, we associate to each estimator $\hat{s}_\lambda$ a (possibly random) subfamily $\mathbb{S}_\lambda \subset \mathbb{S}$ of approximation models and introduce the accuracy index of $\hat{s}_\lambda$ with respect to $\mathbb{S}_\lambda$ as the (random) quantity

$$A(\hat{s}_\lambda, \mathbb{S}_\lambda) = \inf_{S\in\mathbb{S}_\lambda} \inf_{t\in S} \left[H^2(\hat{s}_\lambda, t) + \mathrm{pen}_\lambda(t) \right], \tag{5}$$

where pen_λ is a penalty function from $\bigcup_{S\in\mathbb{S}_\lambda} S$ into $\mathbb{R}_+$. Typically, pen_λ is of the form

$$\mathrm{pen}_\lambda(t) = c_0\tau \inf_{S\in\mathbb{S}_\lambda(t)} (D(S) + \Delta(S)), \tag{6}$$

where c_0 is a universal constant in $(0, 1-1/\sqrt{2})$, $\mathbb{S}_\lambda(t) = \{S \in \mathbb{S}_\lambda,\ t \in S\}$ and τ is a scaling parameter depending on the statistical setting (τ is of order $1/n$ for Example 1 and of order a universal constant for Examples 2, 3 and 4). The quantity $A(\hat{s}_\lambda, \mathbb{S}_\lambda)$ measures in some sense the complexity of the estimator $\hat{s}_\lambda$ with respect to the collection $\mathbb{S}_\lambda$. For a choice of the penalty given by (6), the model S_λ which minimizes (5) over $\mathbb{S}_\lambda$ achieves the best trade-off between the approximation term $\inf_{t\in S} H^2(\hat{s}_\lambda, t)$ and the complexity term $\tau(D(S) \vee \Delta(S))$. The selection procedure we propose leads to an estimator $\tilde{s} = \hat{s}_{\hat{\lambda}}$ which satisfies the following inequality for some constant $C \in (0, 1)$ which neither depends on τ nor s

$$C\mathbb{E}\left[H^2(s, \tilde{s})\right] \le \inf_{\lambda\in\Lambda} \left\{ \mathbb{E}\left[H^2\left(s, \hat{s}_\lambda\right)\right] + \mathbb{E}\left[A(\hat{s}_\lambda, \mathbb{S}_\lambda)\right] \right\}. \tag{7}$$

Inequality (7) leads to an oracle inequality as soon as the quantity

$$\mathbb{E}\left[A(\hat{s}_\lambda, \mathbb{S}_\lambda)\right] = \mathbb{E}\left[\inf_{S\in\mathbb{S}_\lambda} \inf_{t\in S} \left(H^2(\hat{s}_\lambda, t) + \mathrm{pen}_\lambda(t) \right) \right]$$

is not larger that $\mathbb{E}\left[H^2\left(s, \hat{s}_\lambda\right)\right]$ up to a universal constant whatever s. Such a property depends on the choice of the subfamilies $\mathbb{S}_\lambda$. From a theoretical point of view, the choices $\mathbb{S}_\lambda = \mathbb{S}$ for all $\lambda \in \Lambda$ lead to the smallest values of $A(\hat{s}_\lambda, \mathbb{S}_\lambda)$. Nevertheless, for computational reasons it may be sometimes convenient to reduce the family $\mathbb{S}_\lambda$ to a smaller number of models.

Selecting among estimators is an old problem in statistics. In density or regression, most of the statisticians use resampling techniques (cross-validation, V-fold,...).

They seem to give satisfactory results in practice but little is known on the theoretical performances of the resulting choice. In the opposite, we provide a non-asymptotic risk bound for the estimator we select but more needs to be done to make our procedure practical. We shall point out the difficulties to be overcome in view of computing the final estimator $\tilde{s}$ and also describe some situations for which these computations are indeed feasible.

1.3 Connections with Birgé's T-estimators

The starting point of this paper originates from a series of papers by Birgé [16–18] providing a new perspective on estimation theory. His approach relies on ideas borrowed from old papers by Le Cam [37,38] and Birgé [13–15], showing how to derive good estimators from families of robust tests between simple hypotheses, and also from more recent ones about complexity and model selection such as Barron and Cover [11] and Barron, Birgé and Massart [10]. More precisely, given a model S with a finite metric dimension, the construction of Birgé's estimators (called T-estimators) is based on a good discretization of S and on the use of a robust test in view of selecting among the discretization points. T-estimators are naturally robust under misspecification and, from this point of view, may outperform the well-known maximum likelihood estimators which are not. If one considers discretization points as candidate estimators, T-estimators result from an estimator selection procedure which crucially relies on the ability of finding a robust test with respect to a given distance d and, to our knowledge, no general recipe for this is available. Our approach provides a general machinery to build such robust tests for Hellinger-type distances and allows us to build T-estimators in various statistical settings, recovering Birgé's results in the contexts of Examples 1 and 4 and establishing new ones in the cases of Examples 2 and 3. Unlike Birgé's approach which allows to select among deterministic points only, ours can deal with arbitrary collection of estimators.

1.4 Connections with model selection

Consider a collection of models $\mathbb{S} = \{S_m,\ m \in \mathcal{M}\}$ (say linear spaces) together with a mapping Δ from $\mathbb{S}$ into $[1, +\infty)$ satisfying (4) and associate to each $m \in \mathcal{M}$ an estimator $\hat{s}_m$ with values in S_m such that, for some distance d on $\mathcal{L}_0$ and some positive constant C,

$$C\mathbb{E}\left[d^2(s, \hat{s}_m)\right] \leq \inf_{t \in S_m} d^2(s, t) + \tau D(S_m).$$

In view of estimating s at best, an ideal choice of m is given by the index m^*, usually called the oracle, for which $\hat{s}_{m^*}$ achieves the best possible risk bound among the collection of estimators $\left\{\hat{s}_m,\ m \in \mathcal{M}\right\}$. In practice, this oracle is inaccessible since it depends on the unknown parameter s and the art of model selection is to design a rule solely based on the data in order to mimic $\hat{s}_{m^*}$. From this point of view, the model selection problem is a particular case of that of estimator selection. The

following oracle-type inequality is typical of what is usually proved in the literature: for all $s \in \mathcal{L}_0$

$$C'\mathbb{E}\left[d^2(s,\hat{s}_{\hat{m}})\right] \leq \inf_{m\in\mathcal{M}}\left\{\mathbb{E}\left[d^2(s,\hat{s}_m)\right] + \tau\,(D(S_m)\vee\Delta(S_m))\right\}. \tag{8}$$

There exist many different ways of designing a selection rule. Some are based on the minimization of a penalized criterion. For example, let us mention Castellan [24,25] and Massart [45, Chapter 7] for the problem of estimating a density, Reynaud-Bouret [48] for that of estimating the intensity of a Poisson process and in the regression setting Baraud [7], Birgé and Massart [20] and Yang [54] among other references. Another way, usually called Lepski's method, appears in a series of papers by Lepski [39–42] and was originally designed to perform model selection among collections of nested models. In a more abstract way, Birgé [16] proposed a way of selecting among T-estimators and closer to ours, Baraud and Birgé [8] suggested to compare pair by pair histogram-type estimators in the statistical frameworks described in Examples 1 and 4 (among others). Finally, we mention that other procedures based on resampling have interestingly emerged from the work of Arlot [3,4] and Célisse [27].

Our approach to estimator selection provides an alternative to solve the problem of model selection. By choosing $d = H$, $\Lambda = \mathcal{M}$, $\mathbb{S}_m = \{S_m\}$ and pen_m given by (6) for all $m \in \mathcal{M}$, we have

$$A(\hat{s}_m, \mathbb{S}_m) \leq 2c_0\tau\,(D(S_m)\vee\Delta(S_m)), \quad \forall m \in \mathcal{M}$$

and it is then straightforward to deduce from (7) an oracle-type inequality such as (8) for the estimator we select. Compared to the model selection procedures mentioned above, we shall see that ours possesses the advantage to apply in many statistical settings simultaneously and to require very few assumptions on the collection of models.

1.5 Organization of the paper

The paper is organized as follows. The basic ideas underlying our approach are described in Sect. 2. The main assumptions on the measure N and the family $\mathbb{S}$ are presented and discussed in Sect. 3 on the basis of Examples 1 to 4. The selection procedure and the main result can be found in Sect. 4. In Sect. 5 we consider models S which consist of piecewise constant functions on partitions of $\mathcal{X}$. For the problem of estimating a density, we give a practical way of choosing the number of cells of a regular histogram (regular in the sense that each cell contains the same number of data, with a possible exception for the right-most). In Sect. 6, we deal with models with bounded metric dimensions and, as an application, handle the problems of (convex) aggregation and construction of T-estimators. The problem of estimating the means of nonnegative random variables as presented in Example 3 is tackled in Sect. 7. We establish there uniform rates of estimation over various classes of means and provide a lower bound on the minimax estimation rate over classes for which $\sqrt{s} = (\sqrt{s(1)}, \ldots, \sqrt{s(n)})$ belong to a given linear space. In Sect. 8, we consider the regression framework presented in Example 3 and deal with the problems of model

selection, complete variable selection and that of selecting among linear estimators under weak integrability properties of the errors. Finally, Sect. 9 is devoted to the proofs.

Throughout, we use the following notations. The quantity $|E|$ denotes the cardinality of a finite set E. For $x \in \mathbb{R}_+$, $\lfloor x \rfloor = \sup\{n \in \mathbb{N}, n \leq x\}$. The Euclidean norm of $\mathbb{R}^n$ is denoted $\| \|$. We set $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$, $\mathbb{R}_+^* = \mathbb{R}_+\backslash\{0\}$ and for $t \in \mathbb{R}_+^{*n}$, we denote by $\sqrt{t}$ the vector $\left(\sqrt{t_1}, \ldots, \sqrt{t_n}\right)$. Given a closed convex subset A of an Hilbert space, Π_A denotes the projection operator onto A. For $t \in \mathcal{L}_0$ and $\mathcal{F} \subset \mathcal{L}_0$, we set $H(t, \mathcal{F}) = \inf_{f\in\mathcal{F}} H(t, f)$ and for $y > 0$, $\mathcal{B}(t, y) = \left\{t' \in \mathcal{L}_0,\ H(t, t') \leq y\right\}$. Throughout $C, C', C'', \ldots$ denote constants that may vary from line to line.

2 Basic formulas and basic ideas

The aim of this section is to present the basic formulas and ideas underlying our approach. For the sake of simplicity, we assume that $k = 1$ until further notice. For $t \in \mathcal{L}_0$, we define $\rho(s, t) = \int_{\mathcal{X}} \sqrt{st}\, d\mu$. This quantity corresponds to the Hellinger affinity whenever s and t are densities. The squared distance $H^2(s, t)$ is related to $\rho(s, t)$ by the formula $2H^2(s, t) = \int_{\mathcal{X}} s d\mu + \int_{\mathcal{X}} t d\mu - 2\rho(s, t)$. Throughout, t, t' denote two elements of $\mathcal{L}_0$ one should think of as estimators of s. One would prefer t' to t if $H^2(s, t')$ is smaller than $H^2(s, t)$ or equivalently if

$$\left[\rho(s, t') - \frac{1}{2}\int_{\mathcal{X}} t' d\mu\right] - \left[\rho(s, t) - \frac{1}{2}\int_{\mathcal{X}} t d\mu\right] \geq 0.$$

Since $\int_{\mathcal{X}} t d\mu$ and $\int_{\mathcal{X}} t' d\mu$ are both known, deciding whether t' is preferable to t amounts to estimating $\rho(s, t)$ and $\rho(s, t')$ in a suitable way. In the following sections, we present the material that enables us to estimate these quantities on the basis of the observation N.

2.1 An approximation of $\rho(.,.)$

For a measure ν on $(\mathcal{X}, \mathcal{A})$ and $t, r \in \mathcal{L}_0$, we set

$$\rho_r(\nu, t) = \frac{1}{2}\left[\rho(t, r) + \int_{\mathcal{X}} \sqrt{\frac{t}{r}} d\nu\right] \leq +\infty \tag{9}$$

(using the conventions $0/0 = 0$ and $a/0 = +\infty$ for all $a > 0$). We start with the following result showing that $\rho_r(s \cdot \mu, t)$ over-approximates $\rho(s, t)$.

Proposition 1 *Let $s, t, r \in \mathcal{L}_0$. We have,*

$$\rho_r(s \cdot \mu, t) - \rho(s, t) = \frac{1}{2}\int_{\mathcal{X}} \sqrt{\frac{t}{r}}\left(\sqrt{s} - \sqrt{r}\right)^2 d\mu \geq 0.$$

Besides, if $r = (t + t')/2$ *with* $t' \in \mathcal{L}_0$ *then*

$$0 \leq \rho_r(s \cdot \mu, t) - \rho(s, t) \leq \frac{1}{\sqrt{2}} \left[H^2(s, t) + H^2(s, t') \right]. \tag{10}$$

Proof It follows from the definition of ρ_r that

$$\begin{aligned} 2\left[\rho_r(s \cdot \mu, t) - \rho(s, t)\right] &= \int_{\mathcal{X}} \sqrt{tr}\, d\mu + \int_{\mathcal{X}} \sqrt{\frac{t}{r}}\, s d\mu - 2 \int_{\mathcal{X}} \sqrt{st}\, d\mu \\ &= \int_{\mathcal{X}} \sqrt{\frac{t}{r}} \left(\sqrt{s} - \sqrt{r}\right)^2 d\mu. \end{aligned}$$

For the second part, note that $(t/r)(x) \leq 2$ for all $x \in \mathcal{X}$ and therefore $\rho_r(s \cdot \mu, t) - \rho(s, t) \leq \sqrt{2}\, H^2(s, r)$. It remains to bound $H^2(s, r)$ from above. The concavity of the map $t \mapsto \sqrt{t}$ implies that $\rho(s, r) \geq \left[\rho(s, t) + \rho(s, t')\right]/2$ and therefore $2H^2(s, r) \leq H^2(s, t) + H^2(s, t')$, which leads to the result. □

The important point about Proposition 1 (more precisely inequality (10)) lies in the fact that the constant $1/\sqrt{2}$ is smaller than 1. This makes it possible to use the (sign of the) difference

$$T(s \cdot \mu, t, t') = \left[\rho_r(s \cdot \mu, t') - \frac{1}{2} \int_{\mathcal{X}} t' d\mu \right] - \left[\rho_r(s \cdot \mu, t) - \frac{1}{2} \int_{\mathcal{X}} t d\mu \right]$$

with $r = (t + t')/2$ as an alternative benchmark to find which between t and t' is the closest element to s (up to a multiplicative constant). More precisely, we can deduce from Proposition 1 the following corollary.

Corollary 1 *If* $T(s \cdot \mu, t, t') \geq 0$*, then*

$$H^2(s, t') \leq \frac{\sqrt{2} + 1}{\sqrt{2} - 1} H^2(s, t).$$

Proof Using inequality (10) and the assumption, we have

$$\begin{aligned} H^2\left(s, t'\right) - H^2\left(s, t\right) &= \left[\rho(s, t) - \frac{1}{2} \int_{\mathcal{X}} t d\mu \right] - \left[\rho(s, t') - \frac{1}{2} \int_{\mathcal{X}} t' d\mu \right] \\ &= \left[\rho_r(s \cdot \mu, t) - \frac{1}{2} \int_{\mathcal{X}} t d\mu \right] - \left[\rho_r(s \cdot \mu, t') - \frac{1}{2} \int_{\mathcal{X}} t' d\mu \right] \\ &\quad + \rho\left(s, t\right) - \rho_r\left(s \cdot \mu, t\right) + \rho_r\left(s \cdot \mu, t'\right) - \rho\left(s, t'\right) \\ &\leq \frac{1}{\sqrt{2}} \left[H^2\left(s, t\right) + H^2\left(s, t'\right) \right] \end{aligned}$$

which leads to the result. □

2.2 An estimator of $\rho_r(.,.)$

Throughout, given $t, t' \in \mathcal{L}_0$, we set

$$r = \frac{t+t'}{2} \in \mathcal{L}_0.$$

The superiority of the quantity $\rho_r(s \cdot \mu, t)$ over $\rho(s,t)$ lies in the fact that the former can easily be estimated by its empirical counterpart, namely

$$\rho_r(N,t) = \frac{1}{2}\left[\rho(t,r) + \int \sqrt{\frac{t}{r}} dN\right]. \tag{11}$$

Note that $\rho_r(N,t)$ is an unbiased estimator of $\rho_r(s \cdot \mu, t)$ because of (2). Consequently, a natural way of deciding which between t and t' is the closest to s is to consider the test statistic

$$T(N,t,t') = \left[\rho_r(N,t') - \frac{1}{2}\int_{\mathcal{X}} t' d\mu\right] - \left[\rho_r(N,t) - \frac{1}{2}\int_{\mathcal{X}} t d\mu\right].$$

Replacing the "ideal" test statistic $T(s \cdot \mu, t, t')$ by its empirical counterpart leads to an estimation error given by the process $Z(N,\cdot,\cdot)$ defined on $\mathcal{L}_0^2$ by

$$\begin{aligned} Z(N,t,t') &= T(N,t,t') - T(s \cdot \mu, t, t') \\ &= \left[\rho_r\left(N,t'\right) - \rho_r\left(s \cdot \mu, t'\right)\right] - \left[\rho_r(N,t) - \rho_r(s \cdot \mu, t)\right] \\ &= \int_{\mathcal{X}} \psi(t,t',x) dN - \int_{\mathcal{X}} \psi(t,t',x) s d\mu \end{aligned}$$

where $\psi(t,t',x)$ is the function on $\mathcal{L}_0^2 \times \mathcal{X}$ with values in $[-1/\sqrt{2}, 1/\sqrt{2}]$ given by

$$\psi(t,t',x) = \frac{1}{\sqrt{2}}\left[\sqrt{\frac{1}{1+t(x)/t'(x)}} - \sqrt{\frac{1}{1+t'(x)/t(x)}}\right]. \tag{12}$$

The study of the empirical process $Z(N,\cdot,\cdot)$ over the product space $S \times S'$ is at the core of our techniques.

2.3 The multidimensional case $k > 1$

In the multidimensional case, the same results can be obtained by reasoning component by component. More precisely, the formulas of the above sections extend by using the convention that for all k-uplets $\nu = (\nu_1, \ldots, \nu_k)$ of measures on

$(\mathcal{X}_1, \mathcal{A}_1), \ldots, (\mathcal{X}_k, \mathcal{A}_k)$ respectively,

$$\int_{\mathcal{X}} \phi(s,t,t',r)d\nu = \sum_{i=1}^{k} \int_{\mathcal{X}_i} \phi(s_i,t_i,t_i',r_i)d\nu_i,$$

whatever the functions $s, t, t', r \in \mathcal{L}_0$ and mappings ϕ from $\mathbb{R}_+^4$ into $\mathbb{R}$.

3 Assumptions on N and $\mathbb{S}$

Let τ, γ be positive numbers. For $(t,t') \in \mathcal{L}_0^2$ and $y > 0$, let us set

$$w^2(t,t',y) = \left[H^2(s,t) + H^2\left(s,t'\right)\right] \vee y^2.$$

We assume that the family $\mathbb{S}$ and the measure N satisfy the following.

Assumption 1 *Let τ and γ be fixed positive numbers and $c_0 \in (0, 1-1/\sqrt{2})$. For all pairs $(S, S') \in \mathbb{S}^2$, there exist positive numbers $D(S)$, $D(S')$ such that for all $\xi > 0$ and $y^2 \geq \tau\left(D(S) \vee D(S') + \xi\right)$,*

$$\mathbb{P}\left[\sup_{(t,t') \in S \times S'} \frac{Z(N,t,t')}{w^2(t,t',y)} > c_0\right] \leq \gamma e^{-\xi}.$$

This assumption means that for ξ large enough the random process $Z(N,t,t')$ is uniformly controlled by $w^2(t,t',y)$ over $S \times S'$ with probability close to 1. As we shall see on examples, such a property can be derived from concentration inequalities. Under suitable assumptions, the quantities $D(S)$ measure (in some sense) the massiveness of the parameter sets S. Assumption 1 is met in the following typical examples.

3.1 Discrete models

When the collection $\mathbb{S}$ consists of discrete models S, Assumption 1 holds under mild conditions on N. The proof of the following proposition is postponed to Sect. 9.1.

Proposition 2 *Let a, b, c and M be nonnegative numbers and $c_0 \in (0, 1-1/\sqrt{2})$. Assume that N satisfies for all $y, \xi > 0$*

$$\sup_{t,t' \in \mathcal{B}(s,y)} \mathbb{P}\left[Z(N,t,t') > \xi\right] \leq b \exp\left[-\frac{a\xi^2}{y^2 + c\xi}\right]. \tag{13}$$

Besides, assume that for all $S \in \mathbb{S}$ there exists $\eta(S) \geq 1/2$ such that for all $R \geq 2\eta(S)$

$$\left|S \cap \mathcal{B}(s, R\sqrt{\tau})\right| \leq M \exp\left(\frac{R^2}{2}\right) \quad \text{with } \tau = \frac{4(2+cc_0)}{ac_0^2}. \tag{14}$$

Then, Assumption 1 holds with $\gamma = bM^2$ *and* $D(S) = 4\eta^2(S)$ *for all* $S \in \mathbb{S}$.

Inequality (14) imposes that the number of points of S within balls of radii $R \geq 2\eta(S)$ be not larger that $M \exp\left(R^2/2\right)$. One needs to choose the parameter $\eta(S)$ large enough if the set S is too massive, that is, if it contains a large number of points within small balls. As to inequality (13), it typically derives from Bernstein's and is met in all the examples mentioned in the introduction.

Proposition 3 *Inequality* (13) *holds with* $a = n^2/12$, $b = 1$ *and* $c = n\sqrt{2}/6$ *for Example* 1, *with* $a = 1/12$, $b = 1$ *and* $c = \sqrt{2}/6$ *for Example* 2 *and with* $a = 1/12$, $b = 1$ *and* $c = \sqrt{2}/36$ *for Example* 4. *As to Example* 3, *if there exist nonnegative numbers* σ *and* β *such that for all* $i \in \{1, \ldots, n\}$, X_i *satisfies*

$$\mathbb{E}\left[e^{u(X_i - s(i))}\right] \leq \exp\left[\frac{u^2\sigma s(i)}{2(1 - |u|\beta)}\right] \quad \forall u \in (-1/\beta, 1/\beta), \tag{15}$$

then (13) *holds with* $a = 1/(12\sigma)$, $b = 1$ *and* $c = \beta\sqrt{2}/(12\sigma)$.

The value of τ given in Proposition 2 is of order $1/n$ in the density case and is of order a constant in the other cases. The proof of Proposition 3 is postponed to Sect. 9.2.

3.2 Piecewise constant parameter sets

Assume that $k = 1$ and define for any finite partition m of $\mathcal{X}$ the set S_m gathering the elements of $\mathcal{L}_0$ which are piecewise constant on each element of the partition m, that is

$$S_m = \left\{ \sum_{I \in m} a_I \mathbb{1}_I \middle| \, (a_I)_{I \in m} \in \mathbb{R}^{|m|} \right\} \bigcap \mathcal{L}_0.$$

Let $\mathcal{M}$ be a countable set consisting of such partitions. The family $\mathbb{S} = \{S_m,\ m \in \mathcal{M}\}$ and the measure N satisfy Assumption 1 provided that the following holds.

Proposition 4 *Let* a *and* δ *be positive numbers. For any finite partition* m *of* $\mathcal{X}$, *set*

$$\mathcal{X}^2(m) = \sum_{I \in m} \left(\sqrt{N(I)} - \sqrt{\mathbb{E}(N(I))}\right)^2$$

and assume that N *satisfies for all* $\xi > 0$

$$\mathbb{P}\left[\mathcal{X}^2(m) \geq a\left(|m| + \xi\right)\right] \leq e^{-\xi}. \tag{16}$$

Besides, assume that for all $m, m' \in \mathcal{M}$,

$$\left|m \vee m'\right| \leq \delta\left(|m| \vee \left|m'\right|\right) \tag{17}$$

where $m \vee m' = \{I \cap I',\ (I, I') \in m \times m'\}$. *Then Assumption* 1 *holds with* $\gamma = 1$, $\tau = 20ac_0^{-2}$ *and* $D(S_m) = \delta|m|$ *for all* $m \in \mathcal{M}$.

In this case, the parameter $D(S_m)$ is proportional to the dimension of the linear space generated by S_m. The assumptions given by (16) and (17) also appeared in Baraud and Birgé [8] as Assumptions H and H' in their Theorem 6. Inequality (16) can be obtained from concentration inequalities of suprema of empirical processes (based on N) over classes of uniformly bounded functions. In particular, the following result is proved in Baraud and Birgé [8].

Proposition 5 *Inequality* (16) *holds with* $a = 200/n$ *in the case of Example* 1, *with* $a = 6$ *in the case of Example* 4 *and, in the case of Example* 3, *with*

$$a = 3\kappa\left(1/\sqrt{2} + \sqrt{\left(\frac{\beta}{\kappa} - \frac{1}{2}\right)_+}\right)$$

provided that for some $\beta \geq 0$ *and* $\kappa > 0$, *the* X_i *satisfy for* $i = 1, \ldots, n$

$$\mathbb{E}\left[e^{u(X_i - s(i))}\right] \leq \exp\left[\kappa \frac{u^2 s(i)}{2(1 - u\beta)}\right] \quad \textit{for all } u \in \left[0, \frac{1}{\beta}\right[,$$

with the (convention $1/\beta = +\infty$ *if* $\beta = 0$*), and*

$$\mathbb{E}\left[e^{-u(X_i - s(i))}\right] \leq \exp\left[\kappa \frac{u^2 s(i)}{2}\right] \quad \textit{for all } u \geq 0.$$

One can check that the value of τ given in Proposition 4 is then of order $1/n$ in the density case and otherwise is of order a constant.

4 The selection procedure and the main result

Let $\{\hat{s}_\lambda,\ \lambda \in \Lambda\}$ be a collection of estimators of s with values in $\mathcal{L}_0$ based on the observation N and let $\mathbb{S}$ be a countable family of subsets S of $\mathcal{L}_0$. We recall that together with $\mathbb{S}$, we consider a nonnegative map Δ on $\mathbb{S}$ satisfying (4) and in order to simplify the presentation of our results, we assume that $\Delta(S) \geq 1$ for all $S \in \mathbb{S}$. Throughout, c_0 is an arbitrary number in $(0, 1 - 1/\sqrt{2})$.

4.1 The estimation procedure

We associate to each $\lambda \in \Lambda$, both a family of (possibly random) subsets $\mathbb{S}_\lambda$ of $\mathbb{S}$ and a penalty function pen_λ from $\bigcup_{S \in \mathbb{S}_\lambda} S$ into $\mathbb{R}_+$. The procedure includes three steps.
Step 1: *Construction of intermediate estimators*.
Let $\tau > 0$. For each λ, define $\tilde{s}_\lambda$ as any element of $\bigcup_{S \in \mathbb{S}_\lambda} S$ such that

$$H^2(\hat{s}_\lambda, \tilde{s}_\lambda) + \text{pen}_\lambda(\tilde{s}_\lambda) \leq A(\hat{s}_\lambda, \mathbb{S}_\lambda) + c_0\tau \tag{18}$$

where $A(\hat{s}_\lambda, \mathbb{S}_\lambda)$ is given by (5). When $\mathbb{S}_\lambda$ reduces to a single model S_λ containing $\hat{s}_\lambda$ with probability one and when pen_λ is constant over S_λ, one may choose $\tilde{s}_\lambda = \hat{s}_\lambda$, what we shall do for simplicity throughout this paper.
Step 2: *Pairwise comparison of the estimators* $\tilde{s}_\lambda$.
Given a pair $(\tilde{s}_\lambda, \tilde{s}_{\lambda'})$ such that $\tilde{s}_\lambda \neq \tilde{s}_{\lambda'}$, we consider the test statistic

$$\mathbf{T}(N, \tilde{s}_\lambda, \tilde{s}_{\lambda'}) = \left[\rho_r(N, \tilde{s}_{\lambda'}) - \frac{1}{2} \int_{\mathcal{X}} \tilde{s}_{\lambda'} d\mu - \text{pen}_{\lambda'}(\tilde{s}_{\lambda'}) \right] - \left[\rho_r(N, \tilde{s}_\lambda) - \frac{1}{2} \int_{\mathcal{X}} \tilde{s}_\lambda d\mu - \text{pen}_\lambda(\tilde{s}_\lambda) \right] \tag{19}$$

where $r = (\tilde{s}_\lambda + \tilde{s}_{\lambda'})/2$ and $\rho_r(N, .)$ is given by (11). We set

$$\mathcal{E}(\tilde{s}_\lambda) = \{\tilde{s}_{\lambda'},\ \mathbf{T}(N, \tilde{s}_\lambda, \tilde{s}_{\lambda'}) \geq 0\}$$

and note that either $\tilde{s}_\lambda \in \mathcal{E}(\tilde{s}_{\lambda'})$ or $\tilde{s}_{\lambda'} \in \mathcal{E}(\tilde{s}_\lambda)$ since $\mathbf{T}(N, \tilde{s}_\lambda, \tilde{s}_{\lambda'}) = -\mathbf{T}(N, \tilde{s}_{\lambda'}, \tilde{s}_\lambda)$. Then, we define

$$\mathcal{D}(\tilde{s}_\lambda) = \sup \left\{ H^2(\tilde{s}_\lambda, \tilde{s}_{\lambda'}) \,\middle|\, \tilde{s}_{\lambda'} \in \mathcal{E}(\tilde{s}_\lambda) \right\} \quad \text{if} \quad \mathcal{E}(\tilde{s}_\lambda) \neq \varnothing$$

and $\mathcal{D}(\tilde{s}_\lambda) = 0$ otherwise.
Step 3: *The final selection*.
Select $\tilde{\lambda}$ among Λ as any element satisfying

$$\mathcal{D}(\tilde{s}_{\tilde{\lambda}}) \leq \mathcal{D}(\tilde{s}_\lambda) + c_0 \tau, \quad \forall \lambda \in \Lambda$$

and $\hat{\lambda}$ as any element of Λ such that

$$H^2(\hat{s}_{\hat{\lambda}}, \tilde{s}_{\tilde{\lambda}}) \leq \inf_{\lambda \in \Lambda} H^2(\hat{s}_\lambda, \tilde{s}_{\tilde{\lambda}}) + c_0 \tau.$$

Our final estimator is $\tilde{s} = \hat{s}_{\hat{\lambda}}$.

4.2 Discussion about the procedure

The choice of the value c_0 is arbitrary in $(0, 1 - 1/\sqrt{2})$ and can be fixed to $(\sqrt{2} - 1)/(2\sqrt{2})$. It seemed interesting to show how the constants we get in the risk bounds were depending upon the choice of c_0, at least in the proofs. Even though an optimization with respect to c_0 looks theoretically untractable, the computations show that choices of c_0 too close to 0 or to $1 - 1/\sqrt{2}$ lead to large constants and should therefore be avoided.

Let us now discuss the implementation issues. The computation of the estimator $\tilde{s}$ requires the comparison pair by pair of the estimators $\tilde{s}_\lambda$ defined in Step 1. The whole

procedure may therefore be performed in about $|\Lambda|^2$ steps once the $\tilde{s}_\lambda$ are available. When the cardinality of Λ is not too large, the main difficulty lies in the computations of the $\tilde{s}_\lambda$. In the most favorable situations, one may take $\mathbb{S}_\lambda$ as a (possibly random) singleton S_λ for each $\lambda \in \Lambda$ and the penalty function pen_λ to be constant over S_λ. In this case, $\tilde{s}_\lambda$ may be chosen as the best approximation of $\hat{s}_\lambda$ in S_λ. Whenever the model S_λ is simple enough, this step can therefore be performed in a reasonable amount of time. Nevertheless, we sometimes use models S which result from an abstract discretization of manifolds and, in this least favorable case , the $\tilde{s}_\lambda$ are abstract as well.

4.3 The main result

We recall that for all $t \in \mathcal{L}_0$ and $\lambda \in \Lambda$,

$$\mathbb{S}_\lambda(t) = \{S \in \mathbb{S}_\lambda,\ t \in S\}.$$

We obtain the following result the proof of which postponed to Sect. 9.4.

Theorem 1 *Let $c_0 \in (0, 1 - 1/\sqrt{2})$. Assume that N and $\mathbb{S}$ satisfy Assumption 1 for some positive constants τ and γ and let Δ be some mapping from $\mathbb{S}$ into $[1, +\infty)$ satisfying (4). By applying the selection procedure of Sect. 4.1 with penalties pen_λ satisfying for all $\lambda \in \Lambda$ and $t \in \bigcup_{S\in\mathbb{S}_\lambda} S$,*

$$\mathrm{pen}_\lambda(t) \geq c_0\tau \inf\{D(S) + \Delta(S),\ S \in \mathbb{S}_\lambda(t)\}, \tag{20}$$

the estimator $\tilde{s} = \hat{s}_{\hat{\lambda}}$ satisfies for all $\xi > 0$,

$$\mathbb{P}\left[CH^2(s,\tilde{s}) \geq \inf_{\lambda\in\Lambda}\left[H^2\left(s,\hat{s}_\lambda\right) + A(\hat{s}_\lambda, \mathbb{S}_\lambda)\right] + \tau\xi\right] \leq \left(\gamma\Sigma^2 e^{-\xi}\right) \wedge 1,$$

where C is a positive constant depending on c_0 only and

$$A(\hat{s}_\lambda, \mathbb{S}_\lambda) = \inf_{S\in\mathbb{S}_\lambda}\inf_{t\in S}\left[H^2(\hat{s}_\lambda, t) + \mathrm{pen}_\lambda(t)\right], \quad \forall\lambda \in \Lambda.$$

In particular, by integration with respect to ξ, for some C' depending on c_0, γ and Σ only

$$\begin{aligned} C'\mathbb{E}\left[H^2(s,\tilde{s})\right] &\leq \mathbb{E}\left[\inf_{\lambda\in\Lambda}\left\{H^2\left(s,\hat{s}_\lambda\right) + A(\hat{s}_\lambda, \mathbb{S}_\lambda)\right\}\right] \\ &\leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2\left(s,\hat{s}_\lambda\right)\right] + \mathbb{E}\left[A(\hat{s}_\lambda, \mathbb{S}_\lambda)\right]\right\}. \end{aligned} \tag{21}$$

We deduce from Theorem 1 the following corollary.

Corollary 2 *Under the assumptions of Theorem 1, if for all* $\lambda \in \Lambda$, $\hat{s}_\lambda \in \bigcup_{S\in\mathbb{S}_\lambda} S$ *with probability 1 and if equality holds in* (20)*, then*

$$C'\mathbb{E}\left[H^2\left(s,\tilde{s}\right)\right] \leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2\left(s,\hat{s}_\lambda\right)\right] + 2\mathbb{E}\left[v^2(\hat{s}_\lambda)\right]\right\} \tag{22}$$

where C' *depends on* c_0, γ *and* Σ *only and*

$$v^2(\hat{s}_\lambda) = \tau\left[\inf_{S\in\mathbb{S}_\lambda(\hat{s}_\lambda)} D(S)\vee\Delta(S)\right] \quad \text{for all } \lambda\in\Lambda. \tag{23}$$

Inequality (22) compares the risk of the resulting estimator $\tilde{s}$ to those of the $\hat{s}_\lambda$ plus an additional term $\mathbb{E}\left[v^2(\hat{s}_\lambda)\right]$. If for some deterministic $S \in \mathbb{S}_\lambda$, the estimator $\hat{s}_\lambda$ belongs to S with probability 1, we obtain that

$$v^2(\hat{s}_\lambda) \leq \tau\left[D(S)\vee\Delta(S)\right] \tag{24}$$

and hence $\mathbb{E}\left[v^2(\hat{s}_\lambda)\right]$ is small compared to the risk of $\hat{s}_\lambda$ as soon as for some universal constant $C'' > 0$,

$$C''\mathbb{E}\left[H^2\left(s,\hat{s}_\lambda\right)\right] \geq \tau D(S) \quad \text{for all } s\in\mathcal{L}_0.$$

We emphasize that (24) does not depend on the cardinality of the collection of estimators $\left\{\hat{s}_\lambda,\ \lambda\in\Lambda\right\}$. In particular, if with probability one all the estimators $\hat{s}_\lambda$ belong to a same deterministic model $S \in \bigcap_{\lambda\in\Lambda}\mathbb{S}_\lambda$, by setting $\Delta(S) = 1$, the resulting estimator $\tilde{s}$ satisfies

$$C'''\mathbb{E}\left[H^2(s,\tilde{s})\right] \leq \inf_{\lambda\in\Lambda}\mathbb{E}\left[H^2\left(s,\hat{s}_\lambda\right)\right] + \tau\left(D(S)\vee 1\right)$$

no matter how large this collection is. For a choice of $\left\{\hat{s}_\lambda,\ \lambda\in\Lambda\right\}$ which is countable and dense in S, we therefore get

$$\sup_{s\in S}\mathbb{E}\left[H^2(s,\tilde{s})\right] \leq C'\tau\left(D(S)\vee 1\right),$$

showing thus that the quantity $\tau\left(D(S)\vee 1\right)$ is an upper bound for the minimax rate over S.

5 Selecting among histogram-type estimators

Throughout this section, $k = 1$ and we consider a countable family $\mathcal{M}$ of finite partitions of $\mathcal{X}$ satisfying (17). We associate to $\mathcal{M}$ the family of models $\mathbb{S} = \{S_m, m\in\mathcal{M}\}$ described in Sect. 3.2. We assume that the measure N satisfies (16) and set $\tau = 20ac_0^{-2}$. As already seen in Proposition 4, inequalities (17) together with (16) imply that Assumption 1 holds and we may therefore apply our main theorem (Theorem 1).

The statistical settings we have in mind include Examples 1, 3 and 4 for which we already know from Proposition 5 that (16) holds true for a suitable value of a. We restrict ourselves to families of estimators $\hat{s}_\lambda$ with values in $\bigcup_{m\in\mathcal{M}} S_m$ which allows us to associate to each λ some (possibly random) partition $\hat{m}(\lambda) \in \mathcal{M}$ such that $\hat{s}_\lambda \in S_{\hat{m}(\lambda)}$ with probability 1. For all $\lambda \in \Lambda$, we choose $\mathbb{S}_\lambda = \left\{S_{\hat{m}(\lambda)}\right\}$ and pen_λ constant over $S_{\hat{m}(\lambda)}$ for the sake of simplicity. We may therefore take $\tilde{s}_\lambda = \hat{s}_\lambda$ for all $\lambda \in \Lambda$ in our selection procedure which only depends now on the choice of $\text{pen}_\lambda(\hat{s}_\lambda)$. We deduce from Theorem 1 the following result.

Theorem 2 *Assume that N and $\mathbb{S}$ satisfy (16) and (17) respectively. Let $\left\{\hat{s}_\lambda,\ \lambda \in \Lambda\right\}$ be a collection of estimators of s with values in $\bigcup_{m\in\mathcal{M}} S_m$. If for all $\lambda \in \Lambda$,*

$$\text{pen}_\lambda(\hat{s}_\lambda) \geq c_0\tau\left[\delta|\hat{m}(\lambda)| + \Delta(S_{\hat{m}(\lambda)})\right], \tag{25}$$

the estimator $\tilde{s}$ satisfies for some positive constant C depending on c_0 and Σ only,

$$\begin{aligned} C\mathbb{E}\left[H^2\left(s,\tilde{s}\right)\right] &\leq \mathbb{E}\left[\inf_{\lambda\in\Lambda}\left[H^2\left(s,\hat{s}_\lambda\right) + \text{pen}_\lambda\left(\hat{s}_\lambda\right)\right]\right] \\ &\leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2\left(s,\hat{s}_\lambda\right)\right] + \mathbb{E}\left[\text{pen}_\lambda\left(\hat{s}_\lambda\right)\right]\right\}. \end{aligned}$$

In particular, if equality holds in (25)

$$C'\mathbb{E}\left[H^2\left(s,\tilde{s}\right)\right] \leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2\left(s,\hat{s}_\lambda\right)\right] + \tau\mathbb{E}\left[|\hat{m}(\lambda)| \vee \Delta(S_{\hat{m}(\lambda)})\right]\right\},$$

for some C' depending on c_0, Σ and δ only.

Let us now turn to examples.

5.1 Model selection

Theorem 2 holds for any choices of estimators $\left\{\hat{s}_\lambda,\ \lambda \in \Lambda\right\}$ with values in $\bigcup_{m\in\mathcal{M}} S_m$. However, the estimators $\hat{s}_m$ defined by

$$\hat{s}_m = \sum_{I\in m} \frac{N(I)}{\mu(I)}\mathbb{1}_I \quad \text{for } m \in \mathcal{M} \tag{26}$$

are of special interest. Note that when $\mu(I) = 0$, $\mathbb{E}(N(I)) = \int_I s d\mu = 0$ and $N(I) = 0$ a.s., the estimator $\hat{s}_m$ is well-defined with the convention $0/0 = 0$ and $c/\infty = 0$ for $c > 0$. Besides, in the context of Example 1, $\hat{s}_m$ belongs to $\mathcal{L}_0$ (that is, $\hat{s}_m$ is a density on $\mathcal{X}$) as soon as μ is finite on $\mathcal{X}$. For these estimators, one can prove (we refer to Baraud and Birgé [8]) that for all $m \in \mathcal{M}$,

$$\mathbb{E}\left[H^2(s,\hat{s}_m)\right] \leq 4\left(H^2(s,S_m) + \tau|m|\right).$$

By applying Theorem 2 with $\Lambda = \mathcal{M}$ and for all $m \in \mathcal{M}$, $\mathbb{S}_m = \{S_m\}$ and

$$\text{pen}_m(t) = c_0\tau\,(\delta|m| + \Delta(S_m)), \quad \forall t \in S_m$$

we obtain the following result.

Corollary 3 *Assume N and $\mathbb{S}$ satisfy (16) and (17) respectively. The estimator $\tilde{s}$ satisfies for some constant C depending on c_0, δ and Σ only*

$$C\mathbb{E}\left[H^2\,(s,\tilde{s})\right] \leq \inf_{m\in\mathcal{M}}\left\{\mathbb{E}\left[H^2(s,\hat{s}_m)\right] + \tau\,(|m| \vee \Delta_m)\right\}$$
$$\leq \inf_{m\in\mathcal{M}}\left[H^2\,(s,S_m) + \tau(|m| \vee \Delta_m)\right].$$

This corollary recovers the results of Theorem 6 in Baraud and Birgé [8] even though the selection procedure is different. The choice of a suitable family $\mathcal{M}$ of partitions is of course a crucial point in view of deducing nice statistical properties for $\tilde{s}$. This point has been discussed in Baraud and Birgé [8] (see their Section 3). However, the families $\mathcal{M}$ proposed there have large cardinalities and the computation of $\tilde{s}$ becomes unfortunately NP-hard.

5.2 Selecting among histograms based on random partitions

In this section, we focus on the problem of estimating a density or the intensity of a Poisson process by an histogram(-type) estimator. More precisely, we consider the frameworks described in Examples 1 and 4 with $\mathcal{X} = [0, 1)$ and μ the Lebesgue measure. As already mentioned, the model selection approach developed in the previous section has the advantage to design an estimator possessing nice theoretical properties for suitable choices of families $\mathcal{M}$ but also has the drawback to be practically intractable for such choices. In view of designing a practical procedure for the problem of estimating a density, one may prefer families containing fewer but possibly data-dependent partitions. For illustration, we consider a family of random partitions the elements of which contain a same number of data (with a possible exception for the rightmost). Our selection procedure gives a practical way of choosing the number of data that should be put in each bin of an histogram and provide s thus an alternative to the cross-validation techniques which are quite popular in density for selecting such tuning parameters. This family of partitions can also be used for the problem of estimating the intensity of a Poisson process, the only difference being that the number $n = N([0, 1))$ of observations becomes now randomly drawn from a Poisson distribution with parameter $\bar{n} = \int_0^1 s(x)dx$ (that we assume to be positive).

The procedure

For $n \geq 1$, let $\widehat{M}$ be the random variable defined as the smallest integer for which each interval $[(i-1)/M, i/M[$ with $i = 1, \ldots, M$ contains at most one datum. For $\lambda \in \{1, \ldots, n\}$, define $\hat{m}(\lambda)$ as any partition of $[0, 1)$ based on the grid

$\{i/\widehat{M},\ i=1,\ldots,\widehat{M}\}$ for which $|I\cap\{X_1,\ldots,X_n\}|=\lambda$ for all intervals I in $\hat{m}(\lambda)$ (with a possible exception for the rightmost). Note that $|\hat{m}(\lambda)|\leq n\lambda^{-1}+1$. Let us now define our collection of estimators. It would be natural to take $\Lambda=\{1,\ldots,n\}$ in order to index the family of estimators $\hat{s}_{\hat{m}(\lambda)}$ defined by (26). However, this choice of Λ has the drawback to be random in the Poisson case and we rather take $\Lambda=\mathbb{N}^*$ and define $\hat{s}_\lambda$ as $\hat{s}_{\hat{m}(\lambda)}$ when $\lambda\leq n$ and as $n\mathbb{1}_{[0,1[}$ when $\lambda>n$. Note that both collections $\left\{\hat{s}_\lambda,\ \lambda\in\Lambda\right\}$ and $\left\{\hat{s}_{\hat{m}(\lambda)},\ \lambda\in\{1,\ldots,n\}\right\}$ coincide on the event $\{n\geq 1\}$. As to the family $\mathcal{M}$, we introduce for each positive integer M the family of partitions $\mathcal{M}_M$ consisting of intervals with end points belonging to $\{i/M,\ i=0,\ldots,M\}$ and set $\mathcal{M}=\bigcup_{M\geq 1}\mathcal{M}_M$. Finally, we define Δ as follows. First, for $M=1$ $\mathcal{M}_1$ reduces to $\{[0,1)\}$ and we set $\Delta(S_{\{[0,1)\}})=1$. Then we define recursively for $M\geq 2$ and $m\in\mathcal{M}_M\setminus\bigcup_{k=1}^{M-1}\mathcal{M}_k$, $\Delta(S_m)=(|m|+2)\log(M)$. Our choice of Δ satisfies (4) since

$$\begin{aligned}\sum_{m\in\mathcal{M}}e^{-\Delta(S_m)}&\leq\sum_{M\geq 1}\sum_{D=1}^{M}\sum_{m\in\mathcal{M}_M,|m|=D}e^{-(D+2)\log(M)}\\&\leq\sum_{M\geq 1}M^{-2}\left(1+M^{-1}\right)^M\leq\frac{e\pi^2}{6}<+\infty.\end{aligned}$$

The computation of $\tilde{s}$ requires $n(n-1)/2$ steps to compare the estimators $\hat{s}_\lambda$ pair by pair plus n additional steps to minimize $\lambda\mapsto\mathcal{D}(\hat{s}_\lambda)$ for $\lambda\in\{1,\ldots,n\}$. The whole selection procedure can therefore by implemented in about n^2 steps.

In view of facilitating the comparison of the risk bounds between the density and Poisson frameworks, we shall use the notation h for the Hellinger distance in the density case and the normalized Hellinger-type distance $H/\sqrt{\overline{n}}$ in the Poisson case.

Corollary 4 *Assume that* $\|s\|_{\mathbb{L}_q}=\left(\int_0^1 s^q\right)^{1/q}<+\infty$ *for some* $q>1$. *If for all* $\lambda\in\Lambda$,

$$\mathrm{pen}_\lambda(\hat{s}_\lambda)=2c_0\tau\left[(n\lambda^{-1}+1)\log(e+\widehat{M})\mathbb{1}_{\lambda\leq n}+\mathbb{1}_{\lambda>n}\right],$$

the estimator $\tilde{s}=\hat{s}_{\hat{\lambda}}$ *satisfies for some* C *depending on* q *only*

$$C\mathbb{E}\left[h^2(s,\tilde{s})\right]\leq\inf_{1\leq\lambda\leq n}\left[\mathbb{E}\left[h^2\left(s,\hat{s}_\lambda\right)\right]+\frac{\log\left(e+n^2\|s\|_{\mathbb{L}_q}\right)}{\lambda}\right]$$

in the density case, and in the Poisson case

$$C\mathbb{E}\left[\frac{H^2(s,\tilde{s})}{\overline{n}}\right]\leq\inf_{\lambda\geq 1}\left[\mathbb{E}\left[\frac{H^2\left(s,\hat{s}_\lambda\right)}{\overline{n}}\right]+\frac{\log\left(e+\overline{n}^2\|s/\overline{n}\|_{\mathbb{L}_q}\right)}{\lambda\wedge\overline{n}}\right].$$

Even though the estimators $\hat{s}_\lambda$ are widely used in practice, at least for the purpose of estimating a density, little is known about their risks. In density estimation, the only result we are aware of is due to Lugosi and Nobel [44] showing that if $\lambda = \lambda(n)$ satisfies both $\lambda(n) \to +\infty$ and $\lambda(n)/n \to 0$ as n tends to infinity, the $\mathbb{L}_1$-norm between s and $\hat{s}_{\lambda(n)}$ tends to 0 a.s. (and therefore so does the Hellinger distance). The assumption that for some $q > 1$, s^q is integrable is technical and ensures that the cardinality $\widehat{M}$ of the random grid $\{i/\widehat{M},\ i = 1, \ldots, \widehat{M}\}$ keeps to a reasonable size as n increases.

6 Collections of models with bounded metric dimensions

Throughout this section, we consider a family $\overline{\mathbb{S}}$ of subsets of $\mathcal{L}_0$ with metric dimensions bounded by $\overline{D}(.,.)$ (in the sense of Definition 6 in Birgé [16]). More precisely, we assume that for some universal constant $M > 0$, all $\overline{S} \in \overline{\mathbb{S}}$ and $\eta > 0$ there exist a number $\overline{D}(\overline{S}, \eta) \in [1/2, +\infty)$ and a discrete subset $\overline{S}[\eta] \subset \mathcal{L}_0$ such that

$$H(t, \overline{S}[\eta]) \leq \eta\sqrt{\tau}, \quad \text{for all } t \in \overline{S} \tag{27}$$

and for all $s \in \mathcal{L}_0$ and $R \geq 2\eta$,

$$\left|\overline{S}[\eta] \cap \mathcal{B}(s, R\sqrt{\tau})\right| \leq M \exp\left[\overline{D}(\overline{S}, \eta)\left(\frac{R}{\eta}\right)^2\right]. \tag{28}$$

Furthermore, it is assumed that the mapping $\eta \mapsto \overline{D}(\overline{S}, \eta)$ is right-continuous and, with no loss of generality, that is also non-increasing on $(0, +\infty)$. It follows from (27) and (28) that sets $\overline{S}$ with bounded metric dimensions may be approximated at any scale by discrete sets with controlled massiveness. Any compact set $\overline{S}$ has a bounded metric dimension: by definition for all $\eta > 0$ one can find a finite subset $\overline{S}[\eta]$ of $\overline{S}$ satisfying (27) and hence (28) holds with $M = 1$ and $\overline{D}(\overline{S}, \eta) = 4^{-1} \max\left\{\log\left|\overline{S}[\eta]\right|, 2\right\} \geq 1/2$.

As to the measure N, we assume that it satisfies (13) and we take $\tau = 4(2 + cc_0)/(ac_0^2)$ all along. As shown by Proposition 3, this assumption on N is met in all the examples of Sect. 1.1 and τ is then of order a constant (except in the density case where it is of order $1/n$). To our knowledge, the results of this section are new for the problems presented in Examples 2 and 3 and we believe that they can also be solved by using the robust tests described in Birgé [15] and Birgé [17] in the contexts of Examples 1 and 4 respectively.

6.1 The selection procedure and the main result

We start with a collection of estimators $\left\{\hat{s}_\lambda,\ \lambda \in \Lambda\right\}$, a family $\overline{\mathbb{S}}$ of models $\overline{S}$ with bounded metric dimensions and a mapping $\overline{\Delta}$ on $\overline{\mathbb{S}}$ with values in $[1, +\infty)$ satisfying (4). We associate to each model $\overline{S} \in \overline{\mathbb{S}}$ its discrete version

$$S = \overline{S}[\eta] \quad \text{with} \quad \eta = \eta(\overline{S}) = \inf\left\{\eta > 0,\ \eta^2 \geq 2\overline{D}(\overline{S}, \eta)\right\} \geq 1 \tag{29}$$

and consider the family $\mathbb{S}$ of those S when $\overline{S}$ runs among $\overline{\mathbb{S}}$. Since the mappings $\overline{D}(\overline{S}, .)$ are right-continuous and non-increasing, the sets $\left\{\eta > 0,\ \eta^2 \geq 2\overline{D}(\overline{S}, \eta)\right\}$ are non-void and contain their smallest elements $\eta(\overline{S})$ for all $\overline{S} \in \overline{\mathbb{S}}$. In order to select among the family $\left\{\hat{s}_\lambda,\ \lambda \in \Lambda\right\}$, we use the selection procedure described in Sect. 4.1 with the choices $\mathbb{S}_\lambda = \mathbb{S}$ for all $\lambda \in \Lambda$, $\Delta(S) = \overline{\Delta}(\overline{S})$ for all $\overline{S} \in \overline{\mathbb{S}}$ and

$$\mathrm{pen}_\lambda(t) = \mathrm{pen}(t) = c_0\tau \inf_{S \in \mathbb{S}(t)} \left[4\eta^2(\overline{S}) + \overline{\Delta}(\overline{S})\right]$$

for all $t \in \bigcup_{S\in\mathbb{S}} S$ and $\lambda \in \Lambda$.

For the choice of $\eta(\overline{S})$ given by (29), inequality (14) holds for all $S \in \mathbb{S}$ and therefore by Proposition 2, N and $\mathbb{S}$ satisfy Assumption 1 with $D(S) = 4\eta^2(\overline{S})$ and $\gamma = bM^2$. In particular, pen satisfies (20). Finally, note that under (27) for all $\lambda \in \Lambda$ and $\overline{S} \in \overline{\mathbb{S}}$,

$$\begin{aligned} A\left(\hat{s}_\lambda, \mathbb{S}\right) &\leq \inf_{t\in S}\left[H^2\left(\hat{s}_\lambda, t\right) + \mathrm{pen}(t)\right] \\ &\leq 2H^2\left(\hat{s}_\lambda, \overline{S}\right) + 2\tau\eta^2\left(\overline{S}\right) + c_0\tau\left[4\eta^2(\overline{S}) + \overline{\Delta}(\overline{S})\right] \\ &\leq 2H^2\left(\hat{s}_\lambda, \overline{S}\right) + \tau\left(6\eta^2(\overline{S}) + \overline{\Delta}(\overline{S})\right). \end{aligned}$$

By applying Theorem 1 we therefore deduce the following result.

Theorem 3 *Let $\left\{\hat{s}_\lambda,\ \lambda \in \Lambda\right\}$ be a family of estimators based on a measure N satisfying (13). Let us assume that the family $\overline{\mathbb{S}}$ consists of subsets $\overline{S}$ of $\mathcal{L}_0$ with bounded metric dimensions $\overline{D}(\overline{S}, \eta)$ and that $\overline{\Delta}$ is a mapping from $\overline{\mathbb{S}}$ into $[1, +\infty)$ satisfying (4). By applying the selection procedure described above, the resulting estimator $\tilde{s}$ satisfies*

$$\begin{aligned} &C\mathbb{E}\left[H^2\left(s, \tilde{s}\right)\right] \\ &\leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2\left(s, \hat{s}_\lambda\right)\right] + \mathbb{E}\left[\inf_{\overline{S}\in\overline{\mathbb{S}}}\left[H^2(\hat{s}_\lambda, \overline{S}) + \tau\left(\eta^2\left(\overline{S}\right) \vee \overline{\Delta}\left(\overline{S}\right)\right)\right]\right]\right\} \end{aligned}$$

where $\eta(\overline{S})$ is given by (29) and C is a positive number depending on c_0, b, M and Σ only. In particular, if for all $\lambda \in \Lambda$ there exists some (possibly random) model $\overline{S}_\lambda \in \overline{\mathbb{S}}$ such that $\hat{s}_\lambda \in \overline{S}_\lambda$ with probability 1,

$$C\mathbb{E}\left[H^2\left(s, \tilde{s}\right)\right] \leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2\left(s, \hat{s}_\lambda\right)\right] + \tau\mathbb{E}\left[\eta^2\left(\overline{S}_\lambda\right) \vee \overline{\Delta}\left(\overline{S}_\lambda\right)\right]\right\}. \tag{30}$$

Let us now turn to examples.

6.2 Aggregation of arbitrary points

We assume here that the estimators $\hat{s}_\lambda$ are deterministic and to emphasize the fact that they do not depend on N, denote them s_λ hereafter. Typically, one should think of the s_λ as estimators of s based on an independent copy N' of N in which case the result below should be understood as conditional on N'. In view of selecting among these points, we consider the family of models $\overline{\mathbb{S}}$ given by $\overline{\mathbb{S}} = \{\{s_\lambda\},\ \lambda \in \Lambda\}$. Since each element $\overline{S}$ of $\overline{\mathbb{S}}$ reduces to a single point, its metric dimension can be chosen as $\overline{D}(\overline{S}, \eta) = 1/2$ for all $\eta > 0$ and hence $\eta(\overline{S}) = 1$. We deduce from Theorem 3 the following result.

Corollary 5 *Let N be some random measure satisfying (13), $\{s_\lambda, \lambda \in \Lambda\}$ a countable collection of points in $\mathcal{L}_0$ and $\overline{\Delta}$ a mapping from $\overline{\mathbb{S}}$ into $[1, +\infty)$ satisfying (4). By applying the selection procedure described in Sect. 6.1 the estimator $\tilde{s}$ satisfies*

$$C\mathbb{E}\left[H^2(s, \tilde{s})\right] \leq \inf_{\lambda\in\Lambda}\left\{H^2(s, s_\lambda) + \tau\overline{\Delta}(s_\lambda)\right\}$$

for some positive number C depending on c_0, b and Σ only. In particular, if Λ is finite, by choosing $\overline{\Delta}(s_\lambda) = 1 + \log|\Lambda|$ we obtain

$$\mathbb{E}\left[H^2(s, \tilde{s})\right] \leq C \inf_{\lambda\in\Lambda}\left[H^2(s, s_\lambda) + \tau\log|\Lambda|\right].$$

6.3 Construction of T-estimators

Let us start with a family of models $\overline{\mathbb{S}}$ with finite metric dimensions and consider the family of estimators (in fact points) obtained by gathering the elements of the sets S defined in Sect. 6.1 as they run among $\mathbb{S}$. That is $\left\{\hat{s}_\lambda,\ \lambda \in \Lambda\right\} = \bigcup_{S\in\mathbb{S}} S$. Because of (27), this collection of estimators satisfies $\inf_{\lambda\in\Lambda} H\left(s, \hat{s}_\lambda\right) \leq H(s, \overline{S}) + \eta(\overline{S})\sqrt{\tau}$ for all $s \in \mathcal{L}_0$ and $\overline{S} \in \overline{\mathbb{S}}$. By applying the selection procedure of Sect. 6.1 the resulting estimator $\tilde{s}$ turns to be a T-estimator (according to Definition 2 in Birgé [16]). We deduce from Theorem 3 the following result for $\tilde{s}$.

Corollary 6 *Let N be some measure satisfying (13), $\overline{\mathbb{S}}$ a countable collection of subsets $\overline{S}$ of $\mathcal{L}_0$ with bounded metric dimensions $\overline{D}(\overline{S}, \eta)$ and $\overline{\Delta}$ a mapping on $\overline{\mathbb{S}}$ satisfying (4). By applying the selection procedure described in Sect. 6.1 to the family of points $\bigcup_{\overline{S}\in\overline{\mathbb{S}}} S$ with S defined by (29), the resulting estimator $\tilde{s}$ is a T-estimator which satisfies*

$$\mathbb{E}\left[H^2(s, \tilde{s})\right] \leq C \inf_{\overline{S}\in\overline{\mathbb{S}}}\left[H^2(s, \overline{S}) + \tau\left(\eta^2(\overline{S}) \vee \overline{\Delta}(\overline{S})\right)\right]$$

for some C depending on c_0, b, M and Σ only.

This result recovers Corollary 4 in Birgé [16] in the density case and Theorem 3 in Birgé [17] in the Poisson case. Example 2 was also considered in Birgé [16, Proposition 6] but for a different loss function. The case of Example 3 is to our knowledge new.

6.4 Estimators with values in a simplex and convex aggregation

In this section, we assume that $\mathcal{L}_0$ is a convex subset of $\mathcal{L}$ and consider a family $\{t_1, \ldots, t_M\}$ of $M \geq 2$ distinct points of $\mathcal{L}_0$. We denote by $\mathcal{M}$ the class of nonempty subsets m of $\{1, \ldots, M\}$ and define $\overline{S}_m$ as the convex hull of the t_i for $i \in m$, namely

$$\overline{S}_m = \left\{ \sum_{i \in m} q_i t_i \Big| \ (q_i)_{i \in m} \in \mathbb{R}_+^{|m|}, \ \sum_{i \in m} q_i = 1 \right\} \subset \mathcal{L}_0.$$

Along the section, we assume that the estimators $\hat{s}_\lambda$ take their values in the convex hull of $\{t_1, \ldots, t_M\}$, that is, in

$$\overline{\mathcal{C}} = \bigcup_{m \in \mathcal{M}} \overline{S}_m = \left\{ \sum_{i=1}^M q_i t_i \Big| \, q_1, \ldots, q_M \in \mathbb{R}_+, \ \sum_{i=1}^M q_i = 1 \right\} \subset \mathcal{L}_0. \tag{31}$$

As usual, our aim is to select some estimator $\tilde{s}$ among the collection $\left\{\hat{s}_\lambda, \lambda \in \Lambda\right\}$ at best.

6.4.1 The example of convex aggregation

When the set $\{t_1, \ldots, t_M\}$ corresponds to M preliminary estimators of s (say obtained from an independent copy N' of N), the problem of looking for some best convex combination of these is usually called *convex aggregation*. Given some integer $D \in \{1, \ldots, M\}$, we tackle this problem by considering the (countable) collection of estimators (in fact points) given by

$$\left\{ s_\lambda = \sum_{i=1}^M \lambda_i t_i, \ \lambda \in \Lambda \right\} \tag{32}$$

with

$$\Lambda = \Lambda_{D,M} = \left\{ \lambda \in \mathbb{Q}_+^M, \ \sum_{i=1}^M \lambda_i = 1, \ |\{i, \lambda_i \neq 0\}| \leq D \right\}.$$

The choices $D = 1$ and $D = M$ correspond to the problems of *estimator selection* and *convex aggregation* respectively. These problems are particular cases of that of *aggregation* which aims at designing a suitable combination of given estimators in order to outperform each of these separately (and even the best combination of these) up to a remaining term. Aggregation techniques can be found in Juditsky and Nemirovski [35], Nemirovski [46], Yang [55–57], Tsybakov [51], Wegkamp [53], Birgé [16], Rigollet and Tsybakov [49], Bunea et al. [22], Goldenshluger [32] for $\mathbb{L}_p$-losses, and Catoni [26] (we refer to his course of Saint Flour which takes back some mixing techniques he introduced earlier). Most of the aggregation procedures

are based on a sample splitting and therefore usually requires that the data be i.i.d.. In a non-i.i.d. case, some nice results of aggregation can be found in Leung and Barron [43] for the problem of mixing least-squares estimators of a mean of a Gaussian vector Y. In their paper, they assume that the components of Y are independent with a known common variance. Giraud [31] extended their results to the case where it is unknown.

6.4.2 The selection procedure

In order to select among the estimators $\{\hat{s}_\lambda, \lambda \in \Lambda\}$, we apply the selection procedure described in Sect. 6.1 with the family of models $\overline{\mathbb{S}} = \{\overline{S}_m, m \in \mathcal{M}\}$. To do so, we need to find a mapping $\overline{\Delta}$ on $\overline{\mathbb{S}}$ with values in $[1, +\infty)$ satisfying (4) and build an $\eta\sqrt{\tau}$-net $\overline{S}_m[\eta]$ of $\overline{S}_m$ for all $\eta > 0$ and $m \in \mathcal{M}$. Concerning $\overline{\Delta}$, we choose $\overline{\Delta}\left(\overline{S}_m\right) = |m|(1 + \log(eM/|m|))$ for all $m \in \mathcal{M}$. It satisfies (4) since for all $D \in \{1, \ldots, M\}$, $\binom{M}{D} \leq \log(eM/D)$ and hence

$$\sum_{m\in\mathcal{M}} e^{-\overline{\Delta}(\overline{S}_m)} \leq \sum_{D=1}^{M} \binom{M}{D} e^{-D(1+\log(eM/D))} \leq \sum_{D\geq 1} e^{-D} < 1.$$

Let us now fix some $\eta > 0$ and $m \in \mathcal{M}$ and discretize $\overline{S}_m$. If for some $i \in \{1, \ldots, M\}$ $m = \{i\}$, we can merely take $\overline{S}_{\{i\}}[\eta] = \overline{S}_i$. If $|m| \geq 2$, write m as $\{i_1, \ldots, i_{|m|}\}$ with $1 \leq i_1 < \cdots < i_{|m|} \leq M$ and for

$$\varepsilon = \min\left\{\frac{\eta^2\tau}{(|m|-1)\,\|t\|_1}; 1\right\} \quad \text{with} \quad \|t\|_1 = \max_{i=1,\ldots,M} \int_{\mathcal{X}} t_i \, d\mu \tag{33}$$

define $\overline{S}_m[\eta] \subset \overline{S}_m$ as the set gathering the elements of the form

$$\sum_{j=1}^{|m|-1} q_{i_j} t_{i_j} + \left(1 - \sum_{j=1}^{|m|-1} q_{i_j}\right) t_{i_{|m|}}$$

where the q_{i_j} vary among $\{\ell\varepsilon, \ell = 0, \ldots, \lfloor \varepsilon^{-1} \rfloor\}$ and satisfy $\sum_{j=1}^{|m|-1} q_{i_j} \leq 1$. The following result holds.

Proposition 6 *For all $\eta > 0$ and all non-void subsets m of $\{1, \ldots, M\}$, the subset $S_m[\eta]$ defined above is an $\eta\sqrt{\tau}$-net for $\overline{S}_m$ which satisfies*

$$\log|S_m[\eta]| \leq |m| \log\left(1 + \left\lfloor \varepsilon^{-1} \right\rfloor\right)$$

where ε is given by (33) with the convention $\varepsilon^{-1} = 0$ if $|m| = 1$. In particular, the metric dimension $\overline{D}(\overline{S}_m, \eta)$ of $\overline{S}_m$ may be chosen as

$$\overline{D}(\overline{S}_m, \eta) = |m| \log\left(1 + \frac{(|m|-1)\,\|t\|_1}{\eta^2\tau} \vee 1\right) \geq \frac{1}{2}.$$

Some simple calculations show that for all $m \in \mathcal{M}$, the quantity $\eta(\overline{S}_m)$ given by (29) satisfies

$$\eta^2(\overline{S}_m) \leq 2|m| \log\left(1 + \frac{(|m|-1)\,\|t\|_1}{|m|\tau} \vee 1\right) \quad \text{for all } m \in \mathcal{M}.$$

6.4.3 The main result

Hereafter, $\hat{m}(\lambda)$ denotes any (possibly random) element of $\mathcal{M}$ for which $\hat{s}_\lambda \in S_{\hat{m}(\lambda)}$. We deduce from Theorem 3 the following result.

Corollary 7 *Let N be some measure satisfying (13) and $\{\hat{s}_\lambda,\ \lambda \in \Lambda\}$ a collection of arbitrary estimators with values in the simplex $\overline{\mathcal{C}}$ given by (31). Let L be the mapping defined on $\mathbb{N}^*$ by*

$$L(k) = \begin{cases} \log(M) & \text{if } k = 1 \\ 1 + \max\left\{\log\left(Mk^{-1}\right); \log\left(\|t\|_1\,\tau^{-1}\right)\right\} & \text{otherwise.} \end{cases}$$

By applying the selection procedure described above, the resulting estimator $\tilde{s}$ satisfies

$$\begin{aligned} &C\mathbb{E}\left[H^2(s,\tilde{s})\right] \\ &\leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2(s,\hat{s}_\lambda)\right] + \mathbb{E}\left[\inf_{m\in\mathcal{M}}\left[H^2(\hat{s}_\lambda,\overline{S}_m) + \tau|m|L(|m|)\right]\right]\right\} \\ &\leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2(s,\hat{s}_\lambda)\right] + \tau\mathbb{E}\left[|\hat{m}(\lambda)|L(|\hat{m}(\lambda)|)\right]\right\} \end{aligned}$$

where C depends on c_0 and b only.

For the problem of convex aggregation, that is, by considering the collection of points $\{s_\lambda,\ \lambda \in \Lambda\}$ defined by (32) with $D = M$, the estimator $\tilde{s}$ satisfies

$$C\mathbb{E}\left[H^2(s,\tilde{s})\right] \leq \inf_{m\in\mathcal{M}}\left[H^2(s,\overline{S}_m) + \tau|m|L(|m|)\right]. \tag{34}$$

Let us comment on this result in the density framework. In this case, $\|t\|_1 = 1$ and τ is of order $1/n$. If one considers the collection of points $\{s_\lambda,\ \lambda \in \Lambda\}$ given by (32) with $D = M$, the right-hand side of (34) is of the form

$$\inf_{m\in\mathcal{M}}\left\{H^2(s,\overline{S}_m) + \frac{|m|L(|m|)}{n}\right\} = \inf_{D=1,\ldots,M}\left[\inf_{\lambda\in\Lambda_{\mathrm{D,M}}} H^2(s,s_\lambda) + \frac{DL(D)}{n}\right]$$

where $L(1) = \log(M)$ and $L(D)$ is of order $\max\{\log(eM/D); \log(n)\}$ otherwise. In the density case, the problem of aggregating M densities has also been considered in Birgé [16] and Rigollet and Tsybakov [49]. In this latter paper, it is shown that the optimal rate of aggregation associated to the simplex $\Lambda_{\mathrm{M,M}}$ is of order M/n for the $\mathbb{L}_2$-norm and leads thus to risk bounds which are similar to ours (up to constants and $\log(n)$ factors). The result we get is similar to Birgé [16] for the $\mathbb{L}_1$-norm (up to the

logarithmic factor). We do not know whether this extra logarithmic factor is due to our discretization procedure or to the loss function we use.

7 Estimating the means of nonnegative random variables

In this section, we consider the statistical setting described in Example 3 and assume that (15) holds. We recall that it is satisfied for a large class of distributions including any random variables with values in $[0, \beta]$ (then $\sigma = \beta$), the Binomial distribution (then $\sigma = 1 = \beta$), the Poisson distribution (for the same choice of parameters), or the Gamma distribution $\gamma(p, q)$ (with mean p/q and $\beta = 1/q = \sigma$). By expanding (15) in a vicinity of 0, it is easy to see that (15) implies that $\mathrm{Var}(X_i) \leq \sigma \mathbb{E}(X_i)$ for all $i = 1, \ldots, n$. Throughout, we identify with the same notation the functions t on $\mathcal{X} = \{1, \ldots, n\}$ with the vectors $(t_1, \ldots, t_n) = (t(1), \ldots, t(n))$. The distance $\sqrt{2}H$ between two elements $t, t' \in \mathcal{L}_0$ corresponds to the Euclidean distance between the vectors $\sqrt{t}$ and $\sqrt{t'}$ and it seems natural to approximate the parameter $\sqrt{s}$ with respect to the Euclidean norm $\| \ \|$. For this purpose, we introduce a family $\mathbb{V}$ of linear subspaces $\overline{V}$ of $\mathbb{R}^n$ with respective (linear) dimensions denoted $\overline{D}(\overline{V})$. Together with this family, we associate, as usual, a mapping $\overline{\Delta}$ on $\mathbb{V}$ satisfying (4). The aim of this section is to design an estimator $\tilde{s}$ satisfying the following risk bound

$$C' \mathbb{E}\left[H^2(s, \tilde{s})\right] \leq \inf_{\overline{V} \in \mathbb{V}} \left[\inf_{v \in \overline{V}} \left\| \sqrt{s} - v \right\|^2 + \overline{D}(\overline{V}) \vee \overline{\Delta}(\overline{V}) \right] \quad \text{for all } \ s \in \mathbb{R}^n_+. \tag{35}$$

To do so, we associate to each linear space $\overline{V}$, some discrete set $V \subset \mathbb{R}^n_+$ obtained by the discretization device described in Birgé [16]. More precisely, the following result holds.

Proposition 7 *Let $\tau > 0$ and $\overline{\mathcal{C}}$ be some closed convex subset of $\mathbb{R}^n$ and $\overline{V} \subset \mathbb{R}^n$ a linear subspace with dimension $\overline{D}(\overline{V})$. For all $\eta > 0$, there exists a discrete subset $\overline{V}(\eta) \subset \overline{\mathcal{C}}$ such that whatever $f \in \overline{\mathcal{C}}$,*

$$\inf_{v \in \overline{V}(\eta)} \|f - v\| \leq 4 \left[\inf_{v \in \overline{V}} \|f - v\| + \eta \sqrt{\tau} \right] \tag{36}$$

and for all $R \geq 2\eta$,

$$\left| \left\{ v \in \overline{V}(\eta), \ \|f - v\| \leq R\sqrt{\tau} \right\} \right| \leq \exp\left[5 \overline{D}(\overline{V}) \left(\frac{R}{\eta} \right)^2 \right]. \tag{37}$$

We apply the proposition with $\overline{\mathcal{C}} = \mathbb{R}^n_+$ and for $\overline{V} \in \mathbb{V}$, denote by $V = \overline{V}(\eta)$ the discrete set resulting from the choice $\eta = \eta(\overline{V}) = \sqrt{20 \overline{D}(\overline{V})}$ and set

$$S(\overline{V}) = \left\{ (v_1^2, \ldots, v_n^2), \ v \in V \right\} \quad \text{for all } \ \overline{V} \in \mathbb{V}.$$

We shall now apply the selection procedure described in Sect. 4.1 with the collection of the estimators (in fact points) $\hat{s}_\lambda = \lambda$ with $\lambda \in \Lambda = \bigcup_{\overline{V}\in\mathbb{V}} S(\overline{V})$ together with the choices $\Delta(S(\overline{V})) = \overline{\Delta}(\overline{V})$ and $\mathbb{S}_\lambda = \mathbb{S}$ for all $\overline{V} \in \mathbb{V}$ and $\lambda \in \Lambda$. We know from Proposition 3 that N satisfies (13) with $a = (12\sigma)^{-1}$, $b = 1$ and $c = \beta\sqrt{2}/(12\sigma)$ and since the mapping $t \mapsto \sqrt{t}$ from $(\mathbb{R}^n_+, \sqrt{2}H)$ into $(\mathbb{R}^n_+, \|\ \|)$ is an isometry, it follows from (37) and our choice of $\eta(\overline{V})$ that all the models $S(\overline{V})$ of $\mathbb{S}$ satisfy (14) with $M = 1$ and τ depending on σ and β only. By Proposition 2, N and $\mathbb{S}$ satisfy thus Assumption 1 with $D(S(\overline{V})) = 80\overline{D}(\overline{V})$ and τ given by (14). Finally, since (36) implies $4^{-1}\sqrt{2}H(s, S(\overline{V})) \leq \inf_{v\in\overline{V}} \|\sqrt{s} - v\| + \eta\sqrt{\tau}$ for all $\overline{V} \in \mathbb{V}$, we deduce from Theorem 1 that the selected estimator $\tilde{s}$ satisfies the following:

Theorem 4 *Let $\mathbb{V}$ be a countable family of linear spaces $\overline{V}$ with respective dimensions $\overline{D}(\overline{V})$ and $\overline{\Delta}$ a mapping from $\mathbb{V}$ into $[1, +\infty[$ satisfying (4). By applying the estimation procedure described above, one designs an estimator $\tilde{s}$ satisfying (35) for some constant C' depending on c_0, σ, β and Σ only.*

7.1 Uniform convergence rates

In this section, we assume that $\sqrt{s}$ is of the form

$$\sqrt{s} = \sqrt{s_F} = (F(x_1), \ldots, F(x_n))$$

for some unknown nonnegative function F on $[0, 1]$ and deterministic points $0 \leq x_1 < \cdots < x_n \leq 1$. For a suitable choice of $\mathbb{V}$, our aim is to deduce from Theorem 4 uniform rates of convergence over classes of means $\mathcal{S}^\alpha_{p,\infty}(R)$ of the form $\mathcal{S}^\alpha_{p,\infty}(R) = \left\{s_F,\ F \in \mathcal{B}^\alpha_{p,\infty}(R)\right\} \subset \mathbb{R}^n_+$ where $\mathcal{B}^\alpha_{p,\infty}(R)$ is a Besov ball with radius $R > 0$ and parameters $\alpha > 0$ and $p \in [1, +\infty]$. For a precise definition of Besov spaces, we refer to DeVore and Lorentz [29]. The following result derives from Theorem 1 and Proposition 1 in Birgé and Massart [19].

Proposition 8 *For all $r \in \mathbb{N}^*$ and $J \in \mathbb{N}$, there exist positive numbers $C(r)$, $C'(r)$, $C''(r)$ and a family $\mathbf{V}_{r,J}$ of finite dimensional linear subspaces $\mathcal{V}$ of real-valued functions on $[0, 1]$ with the following properties:* $\dim(\mathcal{V}) \leq C(r)2^J$ *for all $\mathcal{V} \in \mathbf{V}_{r,J}$,* $\log|\mathbf{V}_{r,J}| \leq C'(r)2^J$ *and for all $\alpha \in (1/p, r)$ and all $F \in \mathcal{B}^\alpha_{p,\infty}(R)$ there exists $G \in \bigcup_{\mathcal{V}\in\mathbf{V}_{r,J}} \mathcal{V}$ such that*

$$\sup_{x\in[0,1]} |F(x) - G(x)| \leq C''(r)R2^{-J\alpha}.$$

For $\mathcal{V} \in \mathbf{V}_{r,J}$, we define $\overline{V} = \overline{V}(\mathcal{V}) = \{(G(x_1), \ldots, G(x_n)),\ G \in \mathcal{V}\} \subset \mathbb{R}^n$ and $\mathbb{V}_{r,J}$ as the collection of those linear subspaces $\overline{V}$ as $\mathcal{V}$ runs among $\mathbf{V}_{r,J}$. By applying Theorem 4 with $\mathbb{V} = \bigcup_{r\geq1, J\geq0} \mathbb{V}_{r,J}$ and $\overline{\Delta}(\overline{V}) = (C'(r)+1)2^J + r$ for all $\overline{V} \in \mathbb{V}_{r,J}$, $r \geq 1$ and $J \geq 0$ we deduce the following result.

Corollary 8 *By using the family of linear spaces $\mathbb{V}$ defined above, the estimator $\tilde{s}$ satisfies for all $p \in [1, +\infty]$, $\alpha > 1/p$ and $R > 1/\sqrt{n}$,*

$$\sup_{s \in \mathcal{S}^{\alpha}_{p,\infty}(R)} n^{-1} H^2(s, \tilde{s}) \leq C R^{2/(1+2\alpha)} n^{-2\alpha/(1+2\alpha)},$$

where C depends on c_0, τ and r.

To our knowledge, Example 3 has received little attention in the literature, especially from a non-asymptotic point of view. The only exceptions we are aware of are Antoniadis, Besbeas and Sapatinas [1] (see also Antoniadis and Sapatinas [2]) and Kolaczyk and Nowak [36]. In Antoniadis, Besbeas and Sapatinas [1], the authors estimate F^2 by a wavelet shrinkage procedure and show that the resulting estimator achieves the usual estimation rate of convergence over Sobolev classes with smoothness larger than $1/2$. Kolaczyk and Nowak [36] study the risk properties of some thresholding and partitioning estimators. Their approach assumes that s is bounded from above and below by positive numbers on $\mathcal{X} = \{1, \ldots, n\}$. Finally, Baraud and Birgé [8] tackled this problem but they restricted themselves to histogram-type estimators and smoothness $\alpha \leq 1$ only.

7.2 Lower bounds

Let $\overline{V}$ be a linear subspace of $\mathbb{R}^n$ such that $\overline{V} \cap \mathbb{R}^n_+ \neq \{0\}$. By applying Theorem 4 with $\overline{\Delta}(\overline{V}) = 1$, we have

$$\sup_{s \in \mathcal{S}} \mathbb{E}\left[H^2(s, \tilde{s})\right] \leq C(\tau) \overline{D}\left(\overline{V}\right) \quad \text{where } \mathcal{S} = \left\{s,\ \sqrt{s} \in \overline{V}\right\}.$$

The aim of this section is bound the minimax risk over $\mathcal{S}$ from below. We assume the following.

Assumption 2 *The distribution of the random vector $X = (X_1, \ldots, X_n)$ belongs to an exponential family of the form*

$$dP_\theta(x_1, \ldots, x_n) = \exp\left[\sum_{i=1}^{n} (\theta_i T(x_i) - A(\theta_i))\right] \bigotimes_{i=1}^{n} d\nu(x_i) \text{ with } \theta \in \Theta^n \tag{38}$$

where ν denotes some measure on $\mathbb{R}_+$, T is a map from $\mathbb{R}_+$ to $\mathbb{R}$, θ_i are parameters belonging to an open interval Θ such that

$$\Theta \subset \left\{a \in \mathbb{R},\ \int \exp\left[aT(x)\right] d\nu(x) < +\infty\right\}$$

and A denotes a smooth function from Θ into $\mathbb{R}$ satisfying $A''(a) \neq 0$ for all $a \in \Theta$.

Assumption 2 holds for the Poisson, Binomial and Gamma distributions (among others). For $\theta \in \Theta^n$, $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ will denote the probability and the expectation over P_θ. It is well-known that the function A is infinitely differentiable on Θ and that for all $i = 1, \ldots, n$

$$s(i) = \mathbb{E}_\theta [X_i] = A'(\theta_i) \quad \text{and} \quad \mathrm{Var}_\theta(X_i) = A''(\theta_i),$$

these quantities being all positive because we assume $X_i \geq 0$ for $i \in \{1, \ldots, n\}$.

We set

$$\mathcal{S} = \left\{ s \in A'(\Theta)^n, \ \sqrt{s} \in \overline{V} \right\}.$$

Let us fix some compact interval $I \subset \Theta$ and set $K = A'(I)$. Since A' and A'' are continuous and positive on I, there exists $\kappa > 0$ such that for all $\theta \in I^n$

$$0 < \mathbb{E}_\theta(X_i) \leq \kappa \mathrm{Var}_\theta (X_i) \quad \forall i = 1, \ldots, n. \tag{39}$$

The following result holds.

Theorem 5 *Let $R \in (0, (2\sqrt{\kappa})^{-1})$ with κ given by (39). Assume that Assumption 2 holds and that the linear space $\overline{V}$ is such that for some $u_0 \in \overline{V}$*

$$\left\{ (u_1^2, \ldots, u_n^2) \middle| \ u \in \overline{V}, \ \|u - u_0\| \leq R \right\} \subset K^n. \tag{40}$$

Then, whatever the estimator $\hat{s}$ based on $X_1, \ldots, X_n$,

$$\sup_{s \in \mathcal{S}} \mathbb{E}\left[H^2(s, \hat{s}) \right] \geq \frac{R^2}{30} \overline{D}(\overline{V}).$$

8 Estimation and variable selection in non-Gaussian regression

In this section, we use the notations of Example 2 and consider the regression setting

$$X_i = f_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{41}$$

where $f = (f_1, \ldots, f_n)$ is an unknown vector belonging to the cube $\overline{\mathcal{C}} = [-R, R]^n$ (for some $R > 0$) and the ε_i are i.i.d. random variables with density q on $\mathbb{R}$. Both q and R are assumed to be known. Our aim is to estimate f from the observation of $X = (X_1, \ldots, X_n)$ and to do so, we introduce a collection $\{\hat{f}_\lambda, \ \lambda \in \Lambda\}$ of estimators of f based on X and a family $\mathbb{V}$ of linear subspaces $\overline{V} \subset \mathbb{R}^n$ with respective dimensions $\overline{D}(\overline{V})$. As usual we consider a mapping $\overline{\Delta}$ from $\mathbb{V}$ into $[1, +\infty)$ satisfying (4).

8.1 The main assumption

For all $d \geq 1$ and $t \in \mathbb{R}^d$, we set

$$q_t(x) = (q_{t_1}(x_1), \ldots, q_{t_d}(x_d)) = (q(x_1 - t_1), \ldots, q(x_d - t_d)) \quad \forall x \in \mathbb{R}^d$$

and omit to specify the dependency of q_t with respect to the dimension of t. We assume that the density q satisfies the following.

Assumption 3 *For all real numbers $t, t' \in [-R, R]$,*

$$\underline{R} \left| t - t' \right| \leq h\left(q_t, q_{t'}\right) \leq \overline{R} \left| t - t' \right| \tag{42}$$

where h is the Hellinger distance between the densities q_t and $q_{t'}$.

Assumption 3 holds whenever $\sqrt{q}$ is regular enough (see Theorem 3A page 183 in Borovkov [21]) and is therefore satisfied for the Cauchy distribution. Note that such a distribution admits no finite moments.

8.2 The procedure and the results

For each $\lambda \in \Lambda$, let $\mathbb{V}_\lambda$ be a subset of $\mathbb{V}$ (possibly random depending on X). Associate to each linear space $\overline{V} \in \mathbb{V}$ the discrete subset V of $\overline{C}$ obtained by applying Proposition 7 with $\eta = \eta(\overline{V}) = \underline{R}^{-1} \left(10\overline{D}(\overline{V})\right)^{1/2}$. Then, set $S(\overline{V}) = \{q_t,\ t \in V\}$ and define $\mathbb{S}$ (respectively $\mathbb{S}_\lambda$) as the collection of those $S(\overline{V})$ as $\overline{V}$ runs among $\mathbb{V}$ (respectively $\mathbb{V}_\lambda$). Take $\Delta(S(\overline{V})) = \overline{\Delta}(\overline{V})$ for all $\overline{V} \in \mathbb{V}$ and select the estimator $\tilde{s} = \hat{s}_{\hat{\lambda}}$ among the family

$$\left\{\hat{s}_\lambda = q_{\tilde{f}_\lambda},\ \lambda \in \Lambda\right\} \quad \text{with } \tilde{f}_\lambda = \Pi_{\overline{C}} \hat{f}_\lambda$$

by applying the selection procedure described in Sect. 4.1 with τ given by (14), $a = 1/12$, $c = \sqrt{2}/6$ (hence τ only depends on c_0) and for all $\lambda \in \Lambda$,

$$\mathrm{pen}_\lambda(s') = c_0 \tau \inf\left\{40\overline{D}(\overline{V}) + \overline{\Delta}(\overline{V}) \middle| \ \overline{V} \in \mathbb{V}_\lambda\right\}.$$

Our final estimator $\tilde{f} = \hat{f}_{\hat{\lambda}}$ satisfies the following.

Theorem 6 *Consider the regression setting given by (41) where the mean f is known to belong to the cube $\overline{C} = [-R, R]^n$ for some $R > 0$ and assume that the density q of the ε_i is known and satisfies Assumption 3. Let $\mathbb{V}$ be a family of linear subspaces $\overline{V}$ of $\mathbb{R}^n$ with dimension $\overline{D}(\overline{V})$, $\overline{\Delta}$ a mapping from $\mathbb{V}$ into $[1, +\infty)$ satisfying (4) and $\{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ a collection of estimators of f based on X. By applying the selection procedure described above, the estimator $\tilde{f}$ satisfies*

$$C\mathbb{E}\left[\left\|f-\Pi_{\overline{\mathcal{C}}}\tilde{f}\right\|^2\right]$$
$$\leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\left\|f-\hat{f}_\lambda\right\|^2\right]+\mathbb{E}\left[\inf_{\overline{V}\in\mathbb{V}_\lambda}\left(\left\|\hat{f}_\lambda-\Pi_{\overline{V}}\hat{f}_\lambda\right\|^2+\overline{D}(\overline{V})\vee\overline{\Delta}(\overline{V})\right)\right]\right\}$$

where C depends on c_0, $\underline{R}$, $\overline{R}$ and Σ only. In particular if for all $\lambda\in\Lambda$ $\hat{f}_\lambda$ belongs to some $\overline{V}_\lambda\in\mathbb{S}_\lambda$ with probability 1,

$$C\mathbb{E}\left[\left\|f-\Pi_{\overline{\mathcal{C}}}\tilde{f}\right\|^2\right]\leq\inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\left\|f-\hat{f}_\lambda\right\|^2\right]+\mathbb{E}\left[\overline{D}(\overline{V}_\lambda)\vee\overline{\Delta}(\overline{V}_\lambda)\right]\right\}.$$

8.3 Model selection

In this section, we consider the problem of model selection among a collection $\mathbb{V}$ of linear spaces $\overline{V}$. Among the examples we have in mind is that of variable selection.

Problem 1 (Variable selection) *We assume that f is of the form*

$$f=\sum_{j=1}^{p}\beta_j v^{(j)}$$

where $\beta=(\beta_1,\ldots,\beta_p)$ is an unknown vector of $\mathbb{R}^p$ and $v^{(1)},\ldots,v^{(p)}$ are $p\geq 2$ known vectors in $\mathbb{R}^n$ that we call predictors. Since, the number p of those may be large and possibly larger than the number n of data, we assume that the vector β is sparse which means that $\left|\left\{j,\ \beta_j\neq 0\right\}\right|\leq p_{\max}$ for some integer $p_{\max}\leq n$. Our aim is to estimate f and the set $\left\{j,\ \beta_j\neq 0\right\}$. To do so, we consider the index set $\mathcal{M}$ consisting of all the subsets of $\{1,\ldots,p\}$ with cardinality not larger than $p_{\max}$ and $\mathbb{V}$ the family gathering the linear spaces $\overline{V}_m$ spanned by the $v^{(j)}$ for $j\in m$ when m varies among $\mathcal{M}$. By convention, $V_\varnothing=\{0\}$.

One way to address the problem of model selection is to associate to each $\overline{V}\in\mathbb{V}$ a family of points $\Lambda(\overline{V})$ which is countable and dense in $\overline{V}$ and then to apply the procedure described in Sect. 8.2 to the family of estimators (in fact points) given by $\hat{f}_\lambda=\lambda$ for $\lambda\in\bigcup_{\overline{V}\in\mathbb{V}}\Lambda(\overline{V})=\Lambda$. By applying Theorem 6 with $\mathbb{V}_\lambda=\mathbb{V}$ for all $\lambda\in\Lambda$, we deduce the following result without assuming any finite moments on the distribution of the ε_i.

Corollary 9 *Under the assumptions of Theorem 6, one can build an estimator $\hat{f}$ based on X such that*

$$C\mathbb{E}\left[\left\|f-\hat{f}\right\|^2\right]\leq\inf_{\overline{V}\in\mathbb{V}}\left\{\left\|f-\Pi_{\overline{V}}f\right\|^2+\overline{D}(\overline{V})\vee\overline{\Delta}(\overline{V})\right\}. \tag{43}$$

To our knowledge, such a result without any assumption on the integrability of the ε_i is new. In the context of variable selection with non-Gaussian errors, Theorems 5 and 6 by Dalalyan and Tsybakov [28] are probably the closest even though the risk bounds they get is slightly different and depend on the ℓ_1-norm of the β_j. These bounds are derived from sharp PAC-Bayesian ones and are difficult to compare to ours. Let us just say that Dalalyan and Tsybakov achieve better constants and do not assume that the distribution of the errors is known but require stronger assumptions both on the integrability of the errors and on the predictors $v^{(j)}$ to control the ℓ_1-norm of the β_j.

If the ε_i are centered and admit a finite variance σ^2, the family of candidate estimators used in Corollary 9 can be reduced to that of the least-squares $\{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ defined with the choice $\Lambda = \mathbb{V}$ by $\hat{f}_{\overline{V}} = \Pi_{\overline{V}} X$ for all $\overline{V} \in \mathbb{V}$. Since the risk of $\hat{f}_{\overline{V}}$ satisfies

$$\mathbb{E}\left[\left\| f - \hat{f}_{\overline{V}} \right\|^2\right] = \left\| f - \Pi_{\overline{V}} f \right\|^2 + \overline{D}(\overline{V})\sigma^2, \tag{44}$$

by applying Theorem 6 with $\mathbb{V}_{\overline{V}} = \left\{\overline{V}\right\}$ for all $\overline{V} \in \mathbb{V}$, we deduce the following.

Corollary 10 *Assume that the assumptions of Theorem 6 hold and that the ε_i are centered and admit a finite variance σ^2. By applying the selection procedure of Sect. 8.2 to the family of least-squares estimators $\left\{\Pi_{\overline{V}} X,\ \overline{V} \in \mathbb{V}\right\}$, one can select from the data X some linear space $\widehat{V}$ among $\mathbb{V}$ such that*

$$C\mathbb{E}\left[\left\| f - \Pi_{\overline{C}} \Pi_{\widehat{V}} X \right\|^2\right] \leq \inf_{\overline{V} \in \mathbb{V}} \left\{\left\| f - \Pi_{\overline{V}} f \right\|^2 + \overline{D}(\overline{V}) \vee \overline{\Delta}(\overline{V})\right\} \tag{45}$$

where C depends on c_0, σ, $\underline{R}$, $\overline{R}$ and Σ only.

Under the assumption that the ε_i are Gaussian, results of the same flavor (without any boundedness assumption on the vector f and therefore for $\overline{C} = \mathbb{R}$) were previously obtained by Birgé and Massart [20] when the variance is known and in Baraud, Giraud and Huet [9] when it is not. Nevertheless, these approaches as well as ours suffer from the same drawback: in the context of variable selection with $p_{\max}$ large enough, the collection $\{\hat{f}_{\overline{V}},\ \overline{V} \in \mathbb{V}\}$, though finite, is very large and the selection procedure becomes NP-hard and hence practically useless. In the recent years, many efforts have been done to design (practical) selection rules for the purpose of performing variable selection. Among the most popular ones, we mention the Lasso and the Dantzig selectors described respectively in Tibshirani [50] and Candès and Tao [23]. A theoretical analysis of these two procedures (separately and comparatively) has been done in Bickel, Ritov and Tsybakov [12]. Others based on random forest (see Genuer et al. [30]) or PLS regression (see Höskuldsson [34]) are also used in practice even though less is known on their theoretical performances. In any case, it seems that none of such procedures outperforms the others and it may therefore be reasonable to consider them all and let the data decide which is the most appropriate to estimate the truth.

In what follows we consider an arbitrary collection Λ of model selection procedures among $\mathbb{V}$, denote $\overline{V}_\lambda$ the model selected by the procedure $\lambda \in \Lambda$ and $\hat{f}_\lambda$ the

least-squares estimator on $\overline{V}_\lambda$. By applying the selection procedure of Sect. 8.2 with $\mathbb{V}_\lambda = \{\overline{V}_\lambda\}$ for all $\lambda \in \Lambda$, we deduce the following result.

Corollary 11 *Under the assumptions of Theorem 6, the estimator $\Pi_{\overline{\mathcal{C}}}\tilde{f}$ satisfies*

$$C\mathbb{E}\left[\left\|f - \Pi_{\overline{\mathcal{C}}}\tilde{f}\right\|^2\right] \leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right] + \mathbb{E}\left[D\left(\overline{V}_\lambda\right) \vee \overline{\Delta}\left(\overline{V}_\lambda\right)\right]\right\}$$

for some C depending on c_0, $\overline{R}$, $\underline{R}$ and Σ only.

The selection procedure we propose is unfortunately not practical in the general context of model selection. Not because the computations are NP-hard, at least as long as the family Λ of candidate procedures keeps to a reasonable size, but rather because our selection rule relies on a discretization device that is not practical yet. However, we mention that in the specific context of variable selection the procedure can be made feasible indeed by using the alternative family of models $\mathbb{V} = \{\overline{V}_m \cap \overline{\mathcal{C}}, m \in \mathcal{M}\}$. By assumption, the unknown parameter s belongs to one of these models and as subsets of linear spaces, their discretization can be easily done. Furthermore, only the models $\overline{V}_\lambda \cap \overline{\mathcal{C}}$ with $\lambda \in \Lambda$ actually need to be discretized.

8.4 Selecting among linear estimators

In this section, we assume that the collection $\{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ consists of linear estimators of f. More precisely, we shall assume that Λ is an arbitrary collection of (deterministic) symmetric matrices and that $\hat{f}_\lambda = \lambda X$ for all $\lambda \in \Lambda$. Among the examples we have in mind is the following one.

Example 5 (Kernel estimation) Assume that f is of the form

$$f = (F(1/n), \ldots, F(n/n))$$

for some real-valued function F on $[0, 1]$. Given a symmetric kernel K, that is, a function from $\mathbb{R}^2$ into $\mathbb{R}$, such that the matrix $(K(i/n, j/n))_{1\leq i,j\leq n}$ is symmetric, we can associate the kernel estimator of f defined by

$$\hat{f}_{\lambda(K)} = \lambda(K)X \quad \text{with } \lambda_{i,j}(K) = \frac{K(i/n, j/n)}{n} \quad \text{for all } i, j = 1, \ldots, n.$$

This estimator corresponds to the Priestley and Chao [47] estimator evaluated at points k/n for $k = 1, \ldots, n$.

Other examples of linear estimators with symmetric matrices λ can be found in Arlot and Bach [5] (kernel ridge regression, spline smoothing, multiple kernel learning...). As also mentioned there, the classical Nadaraya-Watson kernel estimator is beyond the scope of this study because it corresponds to a non-symmetric matrix λ.

In view of selecting among the family $\{\hat{f}_\lambda,\ \lambda \in \Lambda\}$, we consider the family of models $\mathbb{V} = \{\overline{V}_\lambda,\ \lambda \in \Lambda\}$ defined as follows. For $\lambda \in \Lambda$, let $\lambda_{(1)} \geq \cdots \geq \lambda_{(n)}$

be the the eigenvalues of λ sorted by non-increasing order, $v_{(1)}, \ldots, v_{(n)}$ the corresponding eigenvectors and $D_\lambda = \max\left\{k, \lambda_{(k)} \geq 1/2\right\}$ with the convention $D_\lambda = 0$ if $\left\{k, \lambda_{(k)} \geq 1/2\right\} = \varnothing$. The linear space $\overline{V}_\lambda$ corresponds to the linear space generated by the $v_{(k)}$ for $k \leq D_\lambda$ with the convention $\overline{V}_\lambda = \{0\}$ if $D_\lambda = 0$. The threshold $1/2$ involved in the definition of D_λ is not magical and has been chosen for convenience. Any other choice of a constant in $(0, 1)$ would lead to a result which is similar to the one below (with a possibly different constant C).

Corollary 12 *Assume that the assumptions of Theorem 6 hold and that the ε_i are centered and admit a finite variance σ^2. Consider a collection $\{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ of linear estimators associated to symmetric matrices λ. By using the selection procedure described in Sect. 8.2 with the family of linear spaces $\mathbb{V}$ defined above and any mapping $\overline{\Delta}$ from $\mathbb{V}$ into $[1, +\infty)$ satisfying (4) the selected estimator satisfies*

$$C\mathbb{E}\left[\left\|f - \Pi_{\overline{\mathcal{C}}}\tilde{f}\right\|^2\right] \leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right] \vee \overline{\Delta}(\overline{V}_\lambda)\right\}$$

for some constant C depending on c_0, $\underline{R}$, $\overline{R}$, σ and Σ only.

The selection procedure allows to minimize the risk among the family of linear estimators $\{\lambda X,\ \lambda \in \Lambda\}$. In particular, it can be used to select a window or a kernel among a collection of those. A result of the same flavour as that of Corollary 12 can be found in Arlot and Bach [5] but under more restrictive assumptions on the matrices λ, the cardinality of Λ and the distribution of the ε_i. Nevertheless, their point of view is more practical than ours and leads to a concrete algorithm. In the Gaussian white noise model, Goldenshluger and Lepski [33] addressed the problem of structural adaptation by means of a suitable selection procedure among linear estimators based on kernels. Their assumptions on the family of kernels are slightly different from ours: they consider some convolution-type assumption between the kernels of the collection while we assume that the kernels are symmetric.

The proof of Corollary 12 is postponed to Sect. 9.11.

9 Proofs

9.1 Proof of Proposition 2

For $S, S' \in \mathbb{S}$ and $\xi > 0$,

$$y^2 \geq \tau\left[4\left(\eta^2(S) \vee \eta^2(S')\right) + \xi\right] \geq 4\tau\left(\eta^2(S) \vee \eta^2(S')\right).$$

We set $\mathcal{C}_0 = (S \cap \mathcal{B}(s, y)) \times (S' \cap \mathcal{B}(s, y))$ and for $j \geq 1$,

$$\mathcal{C}_j = \left\{(t, t') \in S \times S',\ \ 2^{j-1}y^2 < H^2(s, t) + H^2\left(s, t'\right) \leq 2^j y^2\right\}.$$

Note that for all $j \geq 0$, $\mathcal{C}_j \subset \big(S \cap \mathcal{B}(s, 2^{j/2}y)\big) \times \big(S' \cap \mathcal{B}(s, 2^{j/2}y)\big)$ and that for $(t, t') \in \mathcal{C}_j$, $w^2(t, t', y) = \big(H^2(s, t) + H^2(s, t')\big) \vee y^2 \geq (2^{j-1} \vee 1)y^2$. Using (13) and (14), we have

$$
\begin{aligned}
&\mathbb{P}\left[\sup_{(t,t')\in S\times S'} \frac{Z(N,t,t')}{w^2(t,t',y)} > c_0\right] \\
&\leq \sum_{(t,t')\in\mathcal{C}_0} \mathbb{P}\Big[Z(N,t,t') \geq c_0 y^2\Big] + \sum_{j\geq 1} \sum_{(t,t')\in\mathcal{C}_j} \mathbb{P}\Big[Z(N,t,t') \geq c_0 2^{j-1} y^2\Big] \\
&\leq b\,|S \cap \mathcal{B}(s,y)|\,\big|S' \cap \mathcal{B}(s,y)\big| \exp\left[-\frac{ac_0^2 y^4}{y^2 + cc_0 y^2}\right] \\
&\quad + b \sum_{j\geq 1} |S \cap \mathcal{B}(s, 2^{j/2}y)||S' \cap \mathcal{B}(s, 2^{j/2}y)| \exp\left[-\frac{ac_0^2 2^{2(j-1)} y^4}{2^j y^2 + cc_0 2^{j-1} y^2}\right] \\
&\leq bM^2 \exp\left[\left(\frac{1}{\tau} - \frac{ac_0^2}{1+cc_0}\right) y^2\right] + bM^2 \sum_{j\geq 1} \exp\left[\left(\frac{1}{\tau} - \frac{ac_0^2}{2(2+cc_0)}\right) 2^j y^2\right] \\
&\leq bM^2 \sum_{j\geq 0} \exp\left[-\frac{2^j y^2}{\tau}\right],
\end{aligned}
$$

recalling that $\tau = 4(2 + cc_0)/(ac_0^2)$. By using that $\tau^{-1}y^2 \geq 4(\eta^2(S) \vee \eta^2(S')) + \xi \geq 1 + \xi$ and the inequality $2^j \geq j + 1$ which holds for all $j \geq 0$, we finally obtain

$$
\mathbb{P}\left[\sup_{(t,t')\in S\times S'} \frac{Z(N,t,t')}{w^2(t,t',y)} > c_0\right] \leq bM^2 \sum_{j\geq 0} \exp\left[-(j+1)(1+\xi)\right] \leq bM^2 e^{-\xi}.
$$

9.2 Proof of Proposition 3

Cases of Examples 1 and 2 It suffices to prove the result in the case of Example 2, the result for Example 1 being obtained similarly by changing $Z(N, t, t')$ into $Z(N, t, t')/n$. Note that for all $t, t' \in \mathcal{L}_0$,

$$
Z(N,t,t') = \sum_{i=1}^{n} \Big(\psi(t_i, t_i', X_i) - \mathbb{E}\big[\psi(t_i, t_i', X_i)\big]\Big)
$$

is a sum of independent and centered random variables bounded by $\sqrt{2}$. Besides, by setting $r_i = (t_i + t_i')/2$ for $i = 1, \ldots, n$ and using that for all $x_i \in \mathcal{X}_i$, $(t(x_i) \vee$

$t_i'(x_i))/r_i(x_i) \le 2$ we have

$$\begin{aligned}
4\mathbb{E}\left[Z^2(N,t,t')\right] &\le \sum_{i=1}^n \int_{\mathcal{X}_i} \left(\sqrt{t_i} - \sqrt{t_i'}\right)^2 \frac{s_i}{r_i} d\mu_i \\
&= \sum_{i=1}^n \int_{\mathcal{X}_i} \left(\sqrt{t_i} - \sqrt{t_i'}\right)^2 \left(\sqrt{\frac{s_i}{r_i}} - 1 + 1\right)^2 d\mu_i \\
&\le 2\sum_{i=1}^n \int_{\mathcal{X}_i} \left(\sqrt{t_i} - \sqrt{t_i'}\right)^2 \left(\sqrt{\frac{s_i}{r_i}} - 1\right)^2 d\mu_i \\
&\quad + 2\sum_{i=1}^n \int_{\mathcal{X}_i} \left(\sqrt{t_i} - \sqrt{t_i'}\right)^2 d\mu_i \\
&= 2\sum_{i=1}^n \int_{\mathcal{X}_i} \frac{\left(\sqrt{t_i} - \sqrt{t_i'}\right)^2}{r_i} \left(\sqrt{s_i} - \sqrt{r_i}\right)^2 d\mu_i + 4H^2\left(t, t'\right) \\
&\le 8\left(H^2(s,r) + H^2(s,t) + H^2(s,t')\right).
\end{aligned}$$

Since the concavity of $u \mapsto \sqrt{u}$ implies that $2H^2(s,r) \le H^2(s,t) + H^2\left(s,t'\right)$, we obtain that for $t, t' \in \mathcal{B}(s,y)$

$$\mathrm{Var}\left(Z(N,t,t')\right) = \mathbb{E}\left[Z^2(N,t,t')\right] \le 3\left[H^2(s,t) + H^2\left(s,t'\right)\right] \le 6y^2.$$

Applying Bernstein's inequality, we obtain that (13) is fulfilled with $b = 1, a = 1/12$ and $c = \sqrt{2}/6$.

Case of Example 3 Under (15), for all $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ such that $\sum_{i=1}^n u_i^2 s(i) \le v^2$ and $\max_{i=1}^n |u_i| \le \gamma$, and all $\lambda \in (0, 1/(\beta\gamma))$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda \sum_{i=1}^n u_i (X_i - s(i))}\right] &= \prod_{i=1}^n \mathbb{E}\left[e^{\lambda u_i (X_i - s(i))}\right] \le \prod_{i=1}^n \exp\left[\frac{\lambda^2 \sigma u_i^2 s(i)}{2(1-\lambda\gamma\beta)}\right] \\
&\le \exp\left[\frac{\lambda^2 \sigma v^2}{2(1-\lambda\gamma\beta)}\right]. \qquad (46)
\end{aligned}$$

Under (46), we derive from Bernstein's inequality (see Massart [45, Corollary 2.10]),

$$\mathbb{P}\left[\sum_{i=1}^n u_i \left(X_i - s(i)\right) \ge \xi\right] \le \exp\left[-\frac{\xi^2}{2(\sigma v^2 + \gamma\beta\xi)}\right]. \qquad (47)$$

For $t, t' \in \mathcal{B}(s, y)$, let us now take $u = (\psi(t, t', 1), \ldots, \psi(t, t', n))$ (where ψ is defined by (12) on $\mathcal{X} = \{1, \ldots, n\}$) and note that

$$\sum_{i=1}^{n} \psi(t, t', i)\,(X_i - s(i)) = Z(N, t, t')$$

$$\max_{i=1,\ldots,n} \left|\psi(t, t', i)\right| \leq \frac{1}{\sqrt{2}} = \gamma.$$

Besides, arguing as for the case of Example 2, we get

$$\sum_{i=1}^{n} \psi^2(t, t', i)s(i) = \frac{1}{4}\sum_{i=1}^{n} \frac{\left(\sqrt{t(i)} - \sqrt{t'(i)}\right)^2 s(i)}{(t(i) + t'(i))/2}$$
$$\leq 3H^2(s, t) + 3H^2(s, t') \leq 6y^2 = v^2.$$

Consequently, we deduce from (47) that (13) is satisfied with $a = 1/(12\sigma)$, $b = 1$ and $c = \beta\sqrt{2}/(12\sigma)$ (then $\tau \leq 96c_0^{-2}(\sigma + \beta)$).

Case of Example 4 In this case,

$$Z(N, t, t') = \int_{\mathcal{X}} \psi\left(t, t', x\right)(dN(x) - s(x)d\mu)$$

where ψ is bounded with values in $[-1/\sqrt{2}, 1/\sqrt{2}]$ and, arguing as for Example 2, we see that it satisfies

$$\int_{\mathcal{X}} \psi^2\left(t, t', x\right) s(x)d\mu \leq 3\left(H^2(s, t) + H^2(s, t')\right) \leq 6y^2$$

for all $t, t' \in \mathcal{B}(s, y)$. By applying Proposition 7 in Reynaud-Bouret [48] we obtain that $Z(N, t, t')$ satisfies (13) with $a = 1/12$, $b = 1$ and $c = \sqrt{2}/36$.

9.3 Proof of Proposition 4

Let us fix $m, m' \in \mathcal{M}$, $\xi > 0$ and y such that

$$y^2 \geq \tau\left(D(S_m) \vee D(S_{m'}) + \xi\right).$$

All $t \in S_m$ and $t' \in S_{m'}$ are constant on the cells $I \in m \vee m'$ with value t_I, t'_I respectively and therefore so is $\psi(t, t', .)$. Namely, for all $x \in I$

$$\psi(t, t', x) = \psi(t_I, t'_I) = \frac{1}{\sqrt{2}}\left[\sqrt{\frac{1}{1 + t_I/t'_I}} - \sqrt{\frac{1}{1 + t'_I/t_I}}\right].$$

Using that $\left|\psi(t_I, t_I')\right| \leq 1/\sqrt{2}$ and Cauchy–Schwarz inequality, we get

$$\begin{aligned}
Z(N,t,t') &= \sum_{I\in m\vee m'} \psi(t_I,t_I')\left(N(I)-\mathbb{E}\left[N(I)\right]\right)\\
&= \sum_{I\in m\vee m'} \psi(t_I,t_I')\left(\sqrt{N(I)}-\sqrt{\mathbb{E}\left[N(I)\right]}\right)\left(\sqrt{N(I)}+\sqrt{\mathbb{E}\left[N(I)\right]}\right)\\
&= \sum_{I\in m\vee m'} \psi(t_I,t_I')\left(\sqrt{N(I)}-\sqrt{\mathbb{E}\left[N(I)\right]}\right)^2\\
&\quad +2\sum_{I\in m\vee m'} \psi(t_I,t_I')\sqrt{\mathbb{E}(N(I))}\left(\sqrt{N(I)}-\sqrt{\mathbb{E}\left[N(I)\right]}\right)\\
&\leq \frac{\mathcal{X}^2(m\vee m')}{\sqrt{2}} + 2\left[\sum_{I\in m\vee m'}\psi^2(t_I,t_I')\mathbb{E}(N(I))\right]^{1/2}\mathcal{X}(m\vee m')\\
&= \frac{\mathcal{X}^2(m\vee m')}{\sqrt{2}} + 2\left[\int \psi^2(t,t',x)sd\mu\right]^{1/2}\mathcal{X}(m\vee m').
\end{aligned}$$

Besides, arguing as in Sect. 9.2 (Example 2), we have

$$\int_{\mathcal{X}} \psi^2(t,t',x)sd\mu \leq 3\left(H^2(s,t)+H^2(s,t')\right)$$

and thus, using that $w^2(t,t',y)\geq y^2$ and $w^2(t,t',y)\geq \left(H^2(s,t)+H^2(s,t')\right)^{1/2}y$, we derive

$$\begin{aligned}
\sup_{(t,t')\in S_m\times S_{m'}}\frac{Z(N,t,t')}{w^2(t,t',y)} &\leq \frac{\mathcal{X}^2(m\vee m')}{\sqrt{2}y^2}+2\sqrt{3}\frac{\mathcal{X}(m\vee m')}{y}\\
&\leq \frac{2\sqrt{6}+1}{\sqrt{2}}\left(\frac{\mathcal{X}^2(m\vee m')}{y^2}\vee\frac{\mathcal{X}(m\vee m')}{y}\right).
\end{aligned}$$

Since $c_0\in(0,1)$,

$$\begin{aligned}
\left\{\sup_{(t,t')\in S_m\times S_{m'}}\frac{Z(N,t,t')}{w^2(t,t',y)}\geq c_0\right\} &\subset \left\{\frac{\mathcal{X}^2(m\vee m')}{y^2}\vee\frac{\mathcal{X}(m\vee m')}{y}\geq\frac{c_0\sqrt{2}}{2\sqrt{6}+1}\right\}\\
&\subset \left\{\frac{\mathcal{X}^2(m\vee m')}{y^2}\geq\frac{2c_0^2}{\left(2\sqrt{6}+1\right)^2}\right\}
\end{aligned}$$

and therefore

$$\mathbb{P}\left[\sup_{(t,t')\in S_m\times S_{m'}}\frac{Z(N,t,t')}{w^2(t,t',y)}\geq c_0\right]\leq\mathbb{P}\left[\mathcal{X}^2(m\vee m')\geq\frac{2c_0^2y^2}{\left(2\sqrt{6}+1\right)^2}\right].$$

We conclude by using (16) together with the fact that under (17),

$$y^2\geq\tau\left(D(S_m)\vee D(S_{m'})+\xi\right)\geq\frac{\left(2\sqrt{6}+1\right)^2}{2c_0^2}\times a\left(|m\vee m'|+\xi\right).$$

9.4 Proof of Theorem 1

Throughout we set $\kappa=c_0+1/\sqrt{2}\in(1/\sqrt{2},1)$ and fix some estimator $\hat{s}_\lambda$. We start with a preliminary result.

Preliminary result: For $\xi>0$ and $S,S'\in\mathbb{S}$, let us set

$$y^2(S,S',\xi)=\tau\left(D(S)\vee D(S')+\Delta(S)+\Delta(S')+\xi\right),$$

and

$$\Omega_\xi=\bigcap_{(S,S')\in\mathbb{S}^2}\left\{\sup_{(t,t')\in S\times S'}\frac{Z(N,t,t')}{w^2(t,t',y(S,S',\xi))}\leq c_0\right\}.$$

Note that under Assumption 1, $\mathbb{P}\left(\Omega_\xi\right)\geq 1-\gamma\Sigma^2e^{-\xi}$. Let us prove that on the set Ω_ξ,

$$\mathcal{D}(\tilde{s}_\lambda)\leq\frac{12}{1-\kappa}\left(H^2(s,\hat{s}_\lambda)+A(\hat{s}_\lambda,\mathbb{S}_\lambda)+\frac{1}{6}c_0\tau\xi\right).\tag{48}$$

Proof Since $\mathcal{D}(\tilde{s}_\lambda)=0$ whenever $\mathcal{E}(\tilde{s}_\lambda)=\varnothing$, we shall assume from now on that $\mathcal{E}(\tilde{s}_\lambda)\neq\varnothing$. Hence, there exists $\tilde{s}_{\lambda'}\in\mathcal{E}(\tilde{s}_\lambda)$ with $\tilde{s}_{\lambda'}\neq\tilde{s}_\lambda$. Using Proposition 1 with $r=(\tilde{s}_\lambda+\tilde{s}_{\lambda'})/2$ and the fact that $\mathbf{T}(N,\tilde{s}_\lambda,\tilde{s}_{\lambda'})\geq 0$, we get

$$\begin{aligned}H^2(s,\tilde{s}_{\lambda'})-H^2(s,\tilde{s}_\lambda)&=\left[\rho\left(s,\tilde{s}_\lambda\right)-\frac{1}{2}\int_{\mathcal{X}}\tilde{s}_\lambda d\mu\right]-\left[\rho\left(s,\tilde{s}_{\lambda'}\right)-\frac{1}{2}\int_{\mathcal{X}}\tilde{s}_{\lambda'}d\mu\right]\\&=-\mathbf{T}(N,\tilde{s}_\lambda,\tilde{s}_{\lambda'})+\mathrm{pen}(\tilde{s}_\lambda)-\mathrm{pen}(\tilde{s}_{\lambda'})\\&\quad+\left[\rho\left(s,\tilde{s}_\lambda\right)-\rho_r\left(s\cdot\mu,\tilde{s}_\lambda\right)\right]-\left[\rho\left(s,\tilde{s}_{\lambda'}\right)-\rho_r\left(s\cdot\mu,\tilde{s}_{\lambda'}\right)\right]\\&\quad+\left[\rho_r\left(s\cdot\mu,\tilde{s}_\lambda\right)-\rho_r\left(N,\tilde{s}_\lambda\right)\right]\\&\quad-\left[\rho_r\left(s\cdot\mu,\tilde{s}_{\lambda'}\right)-\rho_r\left(N,\tilde{s}_{\lambda'}\right)\right]\end{aligned}$$

$$\leq \frac{1}{\sqrt{2}}\left[H^2\left(s,\tilde{s}_\lambda\right)+H^2\left(s,\tilde{s}_{\lambda'}\right)\right] + Z(N,\tilde{s}_\lambda,\tilde{s}_{\lambda'})+\mathrm{pen}(\tilde{s}_\lambda)-\mathrm{pen}(\tilde{s}_{\lambda'})$$

and therefore,

$$\left(1-\frac{1}{\sqrt{2}}\right)H^2\left(s,\tilde{s}_{\lambda'}\right) \leq \left(1+\frac{1}{\sqrt{2}}\right)H^2\left(s,\tilde{s}_\lambda\right) + Z(N,\tilde{s}_\lambda,\tilde{s}_{\lambda'})+\mathrm{pen}(\tilde{s}_\lambda)-\mathrm{pen}(\tilde{s}_{\lambda'}).$$

On Ω_ξ,

$$\begin{aligned} Z(N,\tilde{s}_\lambda,\tilde{s}_{\lambda'}) &\leq c_0 H^2(s,\tilde{s}_\lambda)+c_0H^2(s,\tilde{s}_{\lambda'}) \\ &\quad +c_0\inf\left\{y^2(S,S',\xi),\ (S,S')\in\mathbb{S}_\lambda(\tilde{s}_\lambda)\times\mathbb{S}_{\lambda'}(\tilde{s}_{\lambda'})\right\} \\ &\leq c_0H^2(s,\tilde{s}_\lambda)+c_0H^2(s,\tilde{s}_{\lambda'}) \\ &\quad +c_0\tau\inf_{(S,S')\in\mathbb{S}_\lambda(\tilde{s}_\lambda)\times\mathbb{S}_{\lambda'}(\tilde{s}_{\lambda'})}\left(D(S)+D(S')+\Delta(S)+\Delta(S')+\xi\right) \end{aligned}$$

and since for all $\lambda\in\Lambda$,

$$\mathrm{pen}_\lambda(\tilde{s}_\lambda)\geq c_0\tau\inf\left\{D(S)+\Delta(S),\ S\in\mathbb{S}_\lambda(\tilde{s}_\lambda)\right\},$$

we have

$$(1-\kappa)\,H^2\left(s,\tilde{s}_{\lambda'}\right)\leq(1+\kappa)\,H^2\left(s,\tilde{s}_\lambda\right)+2\,\mathrm{pen}_\lambda(\tilde{s}_\lambda)+c_0\tau\xi.$$

Since $\tilde{s}_{\lambda'}$ is arbitrary in $\mathcal{E}(\tilde{s}_\lambda)$, we deduce that on Ω_ξ,

$$\begin{aligned} \mathcal{D}(\tilde{s}_\lambda) &= \sup_{\tilde{s}_{\lambda'}\in\mathcal{E}(\tilde{s}_\lambda)} H^2\left(\tilde{s}_\lambda,\tilde{s}_{\lambda'}\right) \\ &\leq 2H^2\left(s,\tilde{s}_\lambda\right)+2\sup_{s_{\lambda'}\in\mathcal{E}(\tilde{s}_\lambda)}H^2\left(s,\tilde{s}_{\lambda'}\right) \\ &\leq 2\left(1+\frac{1+\kappa}{1-\kappa}\right)H^2\left(s,\tilde{s}_\lambda\right)+\frac{4}{1-\kappa}\,\mathrm{pen}_\lambda(\tilde{s}_\lambda)+\frac{2}{1-\kappa}c_0\tau\xi \\ &\leq \frac{4}{1-\kappa}\left(3H^2(s,\hat{s}_\lambda)+\frac{3}{2}H^2(\hat{s}_\lambda,\tilde{s}_\lambda)+\mathrm{pen}_\lambda(\tilde{s}_\lambda)+\frac{1}{2}c_0\tau\xi\right) \end{aligned}$$

and we conclude by using that $H^2(\hat{s}_\lambda,\tilde{s}_\lambda)+\mathrm{pen}_\lambda(\tilde{s}_\lambda)\leq A(\hat{s}_\lambda,\mathbb{S}_\lambda)+c_0\tau\leq 2A(\hat{s}_\lambda,\mathbb{S}_\lambda)$ because $\Delta(.)\geq 1$ on $\mathbb{S}$. □

End of the proof of Theorem 1 Using the triangular inequality and the fact that

$$H\left(\tilde{s}_{\tilde{\lambda}}, \tilde{s}\right) \leq H\left(\tilde{s}_{\tilde{\lambda}}, \hat{s}_{\lambda}\right) + \sqrt{c_0\tau} \leq H\left(\tilde{s}_{\tilde{\lambda}}, \tilde{s}_{\lambda}\right) + H\left(\tilde{s}_{\lambda}, \hat{s}_{\lambda}\right) + \sqrt{c_0\tau},$$

we have

$$H\left(s, \tilde{s}\right) \leq H(s, \hat{s}_{\lambda}) + 2H(\hat{s}_{\lambda}, \tilde{s}_{\lambda}) + 2H(\tilde{s}_{\lambda}, \tilde{s}_{\tilde{\lambda}}) + \sqrt{c_0\tau},$$

which with $c_0\tau \leq A(\hat{s}_{\lambda}, \mathbb{S}_{\lambda})$ and $H^2\left(\hat{s}_{\lambda}, \tilde{s}_{\lambda}\right) \leq 2A(\hat{s}_{\lambda}, \mathbb{S}_{\lambda})$ gives

$$3^{-1}H^2\left(s, \tilde{s}\right) \leq H^2(s, \hat{s}_{\lambda}) + (2\sqrt{2}+1)^2 A(\hat{s}_{\lambda}, \mathbb{S}_{\lambda}) + 4H^2(\tilde{s}_{\lambda}, \tilde{s}_{\tilde{\lambda}}).$$

On Ω_{ξ}, we deduce from (48)

$$\begin{aligned} H^2\left(\tilde{s}_{\lambda}, \tilde{s}_{\tilde{\lambda}}\right) &\leq \mathcal{D}(\tilde{s}_{\lambda}) \vee \mathcal{D}(\tilde{s}_{\tilde{\lambda}}) \leq \mathcal{D}(\tilde{s}_{\lambda}) + c_0\tau \\ &\leq \frac{12}{1-\kappa}\left[H^2(s, \hat{s}_{\lambda}) + A(\hat{s}_{\lambda}, \mathbb{S}_{\lambda}) + \frac{1}{6}c_0\tau\xi\right] + A(\hat{s}_{\lambda}, \mathbb{S}_{\lambda}), \end{aligned}$$

and get

$$\begin{aligned} 3^{-1}H^2\left(s, \tilde{s}\right) \leq &\left(1 + \frac{4\times 12}{1-\kappa}\right) H^2(s, \hat{s}_{\lambda}) + \frac{8}{1-\kappa}c_0\tau\xi \\ &+ \left((2\sqrt{2}+1)^2 + 4\frac{13-\kappa}{1-\kappa}\right) A(\hat{s}_{\lambda}, \mathbb{S}_{\lambda}). \end{aligned}$$

Finally, we conclude the first part by using that $\mathbb{P}\left(\Omega_{\xi}\right) \geq 1 - \gamma\Sigma^2 e^{-\xi}$ and the fact that $\hat{s}_{\lambda}$ is arbitrary. For the second part, it suffices to integrate with respect to ξ and to note that under the assumption that $\Delta \geq 1$ on $\mathbb{S}$,

$$A(\hat{s}_{\lambda}, \mathbb{S}_{\lambda}) \geq \inf_{S\in\mathbb{S}} \inf_{t\in S} \mathrm{pen}_{\lambda}(t) \geq c_0\tau, \quad \forall \lambda \in \Lambda.$$

□

9.5 Proof of Corollary 4

For Examples 1 and 4, we know from Proposition 5 that inequality (16) hold with $a = 200/n$ and $a = 6$ respectively. Besides, inequality (17) holds with $\delta = 2$ and since (25) is satisfied for all $\lambda \geq 1$, we may apply Theorem 2. To get the result, it remains to bound $\mathbb{E}\left[\mathrm{pen}_{\lambda}(\hat{s}_{\lambda})\right]$ from above. Let us first consider the case of density estimation. Note that if $n \geq 2$

$$\widehat{M} = \min\left\{M, \min_{i\neq j}\left|X_i - X_j\right| \geq \frac{1}{M}\right\} \leq \max_{i\neq j}\frac{1}{\left|X_i - X_j\right|} + 1$$

hence, by Hölder inequality with $q > 1$, $\overline{q} = q/(q-1) > 1$ and for $p = 2\overline{q} > 1$

$$\begin{aligned}\mathbb{E}\left[\widehat{M}^{1/p}\right] &\leq 1 + \mathbb{E}\left[\max_{i\neq j}\frac{1}{\left|X_i - X_j\right|^{1/p}}\right] \leq 1 + \frac{n(n-1)}{2}\mathbb{E}\left[\frac{1}{|X_1 - X_2|^{1/p}}\right] \\ &\leq 1 + \frac{n(n-1)}{2}\int_{[0,1)}\left[\int_{[0,1)}\frac{1}{|x-y|^{1/p}}s(y)dy\right]s(x)dx \\ &\leq 1 + 2n(n-1)\,\|s\|_{\mathbb{L}_q}. \end{aligned} \tag{49}$$

Since for $n = 1$, $\widehat{M} = 1$ a.s., this inequality remains true in this case. By using the concavity of the logarithm and of the map $t \mapsto t^{1/p}$, for all $\lambda \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\right] &= 2c_0\tau(n\lambda^{-1}+1)\mathbb{E}\left[\log(e+\widehat{M})\right] \\ &\leq 2c_0\tau(n\lambda^{-1}+1)\,p\mathbb{E}\left[\log(e+\widehat{M}^{1/p})\right] \\ &\leq 2c_0\tau(n\lambda^{-1}+1)\,p\log\left[e+\mathbb{E}\left(\widehat{M}^{1/p}\right)\right] \\ &\leq 2c_0\tau(n\lambda^{-1}+1)\,p\log\left[e+\left(1+2n(n-1)\,\|s\|_{\mathbb{L}_q}\right)\right] \\ &\leq 4\tau\left(n\lambda^{-1}+1\right)p\log\left(e+n(n-1)\,\|s\|_{\mathbb{L}_q}\right), \end{aligned} \tag{50}$$

and we conclude by using that in the density case τ equals $1/n$ up to a universal constant.

Let us now turn to the Poisson case. We decompose $\mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\right]$ as follows

$$\begin{aligned}\mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\right] &= \mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\mathbb{1}_{n=0}\right] + \mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\mathbb{1}_{n\geq 1}\right] \\ &= \mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\mathbb{1}_{n=0}\right] + \mathbb{E}\left[\mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\mathbb{1}_{n\geq 1}|n\right]\right]. \end{aligned}$$

On the event $\{n = 0\}$, $\mathrm{pen}_\lambda(\hat{s}_\lambda) = 2c_0\tau$ for all $\lambda \geq 1$ and therefore,

$$\mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\mathbb{1}_{n=0}\right] \leq 2c_0\tau. \tag{51}$$

Since $\bar{n} > 0$, $\mathbb{P}(n = k) > 0$ for all $k \geq 1$ and conditionally on the event $\{n = k\}$, $X_1, \ldots, X_k$ are i.i.d. with density $s/\bar{n}$. Hence, using (50), we deduce that

$$\mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\mathbb{1}_{n\geq 1}|n\right] \leq 4\tau\left(n\lambda^{-1}+1\right)p\log\left(e+\frac{n(n-1)\,\|s\|_{\mathbb{L}_q}}{\bar{n}}\right)\mathbb{1}_{n\geq 1}.$$

We now use the following inequality $\mathbb{E}(U\log V) \leq \mathbb{E}^{1/2}(U^2)\log(\mathbb{E}(V))$ which holds for all random variables U, V such that $U \geq 0$ and $V \geq e$. This inequality derives from Cauchy–Schwarz inequality together with the fact that the map $v \mapsto \log^2 v$ is concave on $[e, +\infty)$. We obtain

$$\mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\mathbb{1}_{n\geq 1}\right] \leq 4p\tau\mathbb{E}^{1/2}\left[\left(n\lambda^{-1}+1\right)^2\right]\log\left(e+\frac{\mathbb{E}\left(n(n-1)\right)\|s\|_{\mathbb{L}_q}}{\bar{n}}\right)$$

and using the fact that n is distributed as a Poisson random variable, we get

$$\mathbb{E}\left[\mathrm{pen}_\lambda(\hat{s}_\lambda)\mathbb{1}_{n\geq 1}\right] \leq 4p\tau\left(\bar{n}\lambda^{-1}+1\right)\log\left(e+\bar{n}\,\|s\|_{\mathbb{L}_q}\right). \tag{52}$$

We conclude by putting (51) and (52) together and by using that τ is a universal constant in the Poisson case.

9.6 Proof of Proposition 6

Clearly, the result is true for $|m| = 1$. Let us now assume $D = |m| \geq 2$. For $t = \sum_{j=1}^{D} q_{i_j} t_{i_j} \in \overline{S}_m$, define

$$\pi_m t = \sum_{j=1}^{D-1} \overline{q}_{i_j} t_j + \left(1 - \sum_{j=1}^{D-1} \bar{q}_{i_j}\right) t_{i_D} \quad \text{with } \bar{q}_{i_j} = \left\lfloor q_{i_j}\varepsilon^{-1}\right\rfloor \varepsilon$$

Note that for all $j \in \{1,\ldots,D\}$, $\bar{q}_{i_j} \geq 0$, $\left\lfloor q_{i_j}\varepsilon^{-1}\right\rfloor \in \left\{0,\ldots,\left\lfloor\varepsilon^{-1}\right\rfloor\right\}$, $\sum_{j=1}^{D-1}\bar{q}_{i_j} \leq \sum_{j=1}^{D-1} q_{i_j} \leq 1$ and therefore $\pi_m t \in \overline{S}_m[\eta]$. Besides, for all $j \in \{1,\ldots,D\}$ $\left|q_j - \bar{q}_j\right| \leq \varepsilon$ and hence, by using that for all $u, v \in \mathcal{L}_0$, $2H^2(u,v) \leq \int_{\mathcal{X}} |u-v|\,d\mu$ we get

$$\begin{aligned}
2H^2(t,\overline{S}_m[\eta]) \leq 2H^2(t,\pi_m t) &\leq \int_{\mathcal{X}} \left|\sum_{j=1}^{D}(q_j-\bar{q}_j)t_j\right| d\mu \\
&\leq \varepsilon\sum_{j=1}^{D-1}\int_{\mathcal{X}} t_j d\mu + \left|\sum_{j=1}^{D-1}\left(q_{i_j}-\bar{q}_{i_j}\right)\right|\int_X t_{i_D} d\mu \\
&\leq 2\varepsilon(D-1)\,\|t\|_1 \leq 2\eta^2\tau.
\end{aligned}$$

9.7 Proof of Proposition 7

We set $\overline{D} = \overline{D}(\overline{V})$ and consider an orthonormal basis $\left\{u_j,\ j=1,\ldots,\overline{D}\right\}$ of $\overline{V}$. It follows from Proposition 9 of Birgé [16] that the set

$$\mathcal{T} = \left\{\frac{2\eta\sqrt{\tau}}{\sqrt{\overline{D}}}\sum_{j=1}^{d} k_j u_j,\ (k_j)_{j=1,\ldots,\overline{D}} \in \mathbb{Z}^{\overline{D}}\right\}$$

is a $\eta\sqrt{\tau}$-net for $\overline{V}$ and for all $R \geq 2\eta$ and $h \in \mathbb{R}^n$

$$\left|\left\{t \in \mathcal{T},\ \|h - t\| \leq R\sqrt{\tau}\right\}\right| \leq \exp\left[0.458\overline{D}\left(\frac{R}{\eta}\right)^2\right].$$

The result follows by applying Proposition 12 of Birgé [16] (with $\overline{\pi} = \Pi_{\overline{\mathcal{C}}}$, $(M', d) = (\mathbb{R}^n, \|\ \|)$, $\mathcal{M}_0 = \overline{\mathcal{C}}$, $T = \mathcal{T}$ and $\lambda = 1 = \varepsilon$).

9.8 Proof of Corollary 8

Let us denote by $\|\ \|_\infty$ the supremum norm on $[0, 1]$. First note that (4) holds since

$$\sum_{\overline{V}\in\mathbb{V}} e^{-\Delta(\overline{V})} \leq \sum_{r\geq 1}\sum_{J\geq 0}\left|\mathbf{V}_{r,J}\right| e^{-(C'(r)+1)2^J - r} \leq \sum_{r\geq 1} e^{-r}\sum_{J\geq 0} e^{-2^J} < +\infty.$$

Let now $p \in [1, +\infty]$, $\alpha > 1/p$ and $R > 0$. There exists some $r \in \mathbb{N}^*$ such that $\alpha \in (1/p, r)$ and it follows from Proposition 8 that for all $J \geq 0$ there exits $\overline{V} \in \mathbb{V}_{r,J}$ such that $\overline{D}(\overline{V}) \leq C(r)2^J$ and for all $s_F \in \mathcal{S}^\alpha_{p,\infty}(R)$

$$\inf_{v\in\overline{V}}\left\|\sqrt{s_F} - v\right\| \leq n \inf_{\mathcal{V}\in\mathbf{V}_{r,J}}\inf_{G\in\mathcal{V}}\|F - G\|_\infty \leq C''(r)nR2^{-J\alpha}.$$

Hence, we deduce from (35) that for some constant C (depending on c_0, τ and r),

$$C\mathbb{E}\left[n^{-1}H^2(s, \tilde{s})\right] \leq \inf_{J\geq 0}\left(R^2 2^{-2J\alpha} + \frac{2^J}{n}\right)$$

and the result follows by choosing 2^J of order $(nR^2)^{1/(1+\alpha)} \geq 1$.

9.9 Proof of Theorem 5

Hereafter, $\rho(P, Q)$ and $h(P, Q)$ denote the Hellinger affinity and the Hellinger distance between the probabilities P, Q. For $\theta \in \Theta^n$, $A'(\theta)$ corresponds to the vector $t = (A'(\theta_1), \ldots, A'(\theta_n))$. Since the mapping from Θ^n into $\mathbb{R}^n_+$ defined by $\theta \mapsto A'(\theta)$ is one to one, for $s \in \mathcal{S}$ we denote (abusively) $\mathbb{P}_s$ and $\mathbb{E}_s$ the probability and expectation with respect to the probability P_θ where θ is the unique element of Θ^n satisfying $s = A'(\theta)$. We start with the following lemma.

Lemma 1 *Under Assumption 2, for all $\theta, \theta' \in I^n$, $t = A'(\theta)$ and $t' = A'(\theta')$,*

$$h^2(P_\theta, P_{\theta'}) \leq -\sum_{i=1}^n \log\rho\left(P_{\theta_i}, P_{\theta'_i}\right) \leq 4\kappa\sum_{i=1}^n H^2(t_i, t'_i) = 4\kappa H^2(t, t')$$

Proof Since

$$h^2(P_\theta, P_{\theta'}) = 1 - \rho(P_\theta, P_{\theta'}) = 1 - \exp\left[\sum_{i=1}^n \log\rho\left(P_{\theta_i}, P_{\theta'_i}\right)\right]$$
$$\leq -\sum_{i=1}^n \log\rho\left(P_{\theta_i}, P_{\theta'_i}\right),$$

it suffices to show that $-\sum_{i=1}^n \log\rho\left(P_{\theta_i}, P_{\theta'_i}\right) \leq 4\kappa \sum_{i=1}^n H^2(t_i, t'_i)$. Summing up over i, it is enough to show the inequality for $n = 1$, what we shall do. Let θ, θ' in I such that $t = A'(\theta)$ and $t' = A'(\theta')$. With no loss of generality, we may assume that $\theta' < \theta$ and set $\delta = (\theta - \theta')/2$. The Hellinger affinity between P_θ and $P_{\theta'}$ is given by

$$\rho(P_\theta, P_{\theta'}) = \exp\left[-\left(\frac{A(\theta) + A(\theta')}{2} - A\left(\frac{\theta + \theta'}{2}\right)\right)\right]$$

and therefore

$$\begin{aligned}
-\log\rho(P_\theta, P_{\theta'}) &= \frac{A(\theta) + A(\theta')}{2} - A\left(\frac{\theta + \theta'}{2}\right) \\
&= \frac{1}{2}\left[A(\theta) + A(\theta - 2\delta) - 2A(\theta - \delta)\right] \\
&= \frac{1}{2}\int_{\theta-\delta}^{\theta} \left(A'(u) - A'(u - \delta)\right) du \\
&= \frac{1}{2}\int_{\theta-\delta}^{\theta} \left[\int_{u-\delta}^{u} A''(v)dv\right] du.
\end{aligned}$$

Since $t, t' \in \mathbb{R}_+\backslash\{0\}$ and since A'' do not vanish on $[\theta', \theta]$ and A' is nondecreasing, for all $u \in [\theta - \delta, \theta]$ and $v \in [u - \delta, u]$

$$\begin{aligned}
A''(v) &= \frac{A''(v)}{2\sqrt{A'(v)}} \frac{A''(u)}{2\sqrt{A'(u)}} \frac{4\sqrt{A'(v)A'(u)}}{A''(u)} \\
&\leq \frac{A''(v)}{2\sqrt{A'(v)}} \frac{A''(u)}{2\sqrt{A'(u)}} \frac{4A'(u)}{A''(u)} \leq 4\kappa \frac{A''(v)}{2\sqrt{A'(v)}} \frac{A''(u)}{2\sqrt{A'(u)}}.
\end{aligned}$$

giving thus,

$$-\log\rho(P_\theta, P_{\theta'}) \leq 2\kappa \int_{\theta-\delta}^{\theta} \left[\int_{u-\delta}^{u} \frac{A''(v)}{2\sqrt{A'(v)}} \frac{A''(u)}{2\sqrt{A'(u)}} dv\right] du$$

$$\leq 2\kappa \int_{\theta'}^{\theta} \left[\int_{\theta'}^{\theta} \frac{A''(v)}{2\sqrt{A'(v)}} \frac{A''(u)}{2\sqrt{A'(u)}} dv \right] du$$

$$= 2\kappa \left(\int_{\theta'}^{\theta} \frac{A''(v)}{2\sqrt{A'(v)}} dv \right)^2 = 2\kappa \left(\sqrt{A'(\theta)} - \sqrt{A'(\theta')} \right)^2$$

$$= 2\kappa \left(\sqrt{t} - \sqrt{t'} \right)^2$$

□

The proof of Theorem 5 is based on Assouad's Lemma (see Assouad [6]), more precisely on the version given by Theorem 2.10 in Tsybakov [52]. Hereafter, $\{u_1, \ldots, u_{\overline{D}}\}$ denotes an orthonormal basis of $\overline{V}$ (we set $\overline{D} = \overline{D}(\overline{V})$) and $d(\varepsilon, \varepsilon')$ the Hamming distance between two elements ε and ε' of $\{0,1\}^{\overline{D}}$, that is $d(\varepsilon, \varepsilon') = \sum_{j=1}^{\overline{D}} \mathbb{1}_{\varepsilon_j \neq \varepsilon'_j}$. Besides, we set

$$\mathcal{S}_K = \left\{ s \in K^n,\ \sqrt{s} \in \overline{V} \right\} \subset \mathcal{S}.$$

Let $t_0 \in \mathcal{S}$ be such that $u_0 = \sqrt{t_0}$. Under (40), there exists $\beta_1, \ldots, \beta_{\overline{D}}$ such that $\sqrt{t^0} = \sum_{j=1}^{\overline{D}} \beta_j u_j$ and that for all $\varepsilon \in \{0,1\}^{\overline{D}}$ one can find $t^\varepsilon \in \mathcal{S}_K$ such that $\sqrt{t^\varepsilon} = \sum_{j=1}^{\overline{D}} \left(\beta_j + R\varepsilon_j \right) u_j$. Note that the for all $\varepsilon, \varepsilon' \in \{0,1\}^{\overline{D}}$,

$$2H^2(t^\varepsilon, t^{\varepsilon'}) = \left\| \sqrt{t^\varepsilon} - \sqrt{t^{\varepsilon'}} \right\|^2 = R^2 d(\varepsilon, \varepsilon').$$

Besides, whatever the estimator $\hat{s}$ and $s \in \mathcal{S}$

$$\sup_{s \in \mathcal{S}} \mathbb{E}_s \left[H^2 \left(s, \hat{s} \right) \right] \geq \sup_{s \in \mathcal{S}_K} \mathbb{E}_s \left[H^2 \left(s, \hat{s} \right) \right] \geq \sup_{\varepsilon \in \{0,1\}^{\overline{D}}} \mathbb{E}_{t^\varepsilon} \left[H^2 \left(t^\varepsilon, \hat{s} \right) \right]$$

$$\geq \inf_{\hat{\varepsilon}} \sup_{\varepsilon \in \{0,1\}^{\overline{D}}} \mathbb{E}_{t^\varepsilon} \left[H^2 \left(t^\varepsilon, t^{\hat{\varepsilon}} \right) \right] = \frac{R^2}{2} \inf_{\hat{\varepsilon}} \sup_{\varepsilon \in \{0,1\}^{\overline{D}}} \mathbb{E}_{t^\varepsilon} \left[d \left(\varepsilon, \hat{\varepsilon} \right) \right],$$

where the two last infima run among all estimators $\hat{\varepsilon}$ based on the observations $(X_1, \ldots, X_n)$ with values in $\{0,1\}^{\overline{D}}$. Theorem 2.10 in Tsybakov [52] asserts that

$$\inf_{\hat{\varepsilon}} \sup_{\varepsilon \in \{0,1\}^{\overline{D}}} \mathbb{E}_{t^\varepsilon} \left[d \left(\varepsilon, \hat{\varepsilon} \right) \right] \geq \frac{\overline{D}}{2} \left(1 - \sqrt{\alpha(2-\alpha)} \right)$$

provided that for all $\varepsilon, \varepsilon'$ such that $d(\varepsilon, \varepsilon') = 1$, $h^2 \left(P_{\theta^\varepsilon}, P_{\theta^{\varepsilon'}} \right) \leq \alpha < 1$ where the parameters θ^ε and $\theta^{\varepsilon'}$ are such that $t^\varepsilon = A'(\theta^\varepsilon)$ and $t^{\varepsilon'} = A'(\theta^{\varepsilon'})$. Taking $\alpha = 1/2$

and using Lemma 1, we get for all $\varepsilon, \varepsilon'$ such that $d(\varepsilon, \varepsilon') = 1$,

$$h^2\left(P_{\theta^\varepsilon}, P_{\theta^{\varepsilon'}}\right) \le 4\kappa H^2(t^\varepsilon, t^{\varepsilon'}) \le 2\kappa R^2 \le \frac{1}{2} = \alpha,$$

and hence,

$$\inf_{\hat{s}} \sup_{s\in S} \mathbb{E}_s\left[H^2\left(s, \hat{s}\right)\right] \ge \frac{1-\sqrt{3}/2}{4}\,\overline{D}R^2,$$

which concludes the proof.

9.10 Proof of Theorem 6

Let $\mathcal{L}_0 = \left\{q_t,\ t \in \overline{\mathcal{C}}\right\}$. Under Assumption 3, the mapping $t \mapsto q_t$ is a quasi isometry between $(\overline{\mathcal{C}}, \|\ \|)$ and $(\mathcal{L}_0, H)$. In particular, by using Proposition 7, for all $\overline{V} \in \mathbb{V}$, $q_t \in \mathcal{L}_0$ and number $r \ge 2(\underline{R}\eta(\overline{V}))$,

$$\begin{aligned}&\left|\left\{q_{t'} \in S(\overline{V}),\ H(q_t, q_{t'}) \le r\sqrt{\tau}\right\}\right| \\ &\quad\le \left|\left\{t' \in V,\ \left\|t - t'\right\| \le \underline{R}^{-1} r\sqrt{\tau}\right\}\right| \le \exp\left[5\overline{D}(\overline{V})\left(\frac{\underline{R}^{-1}r}{\eta(\overline{V})}\right)^2\right] \le \exp\left(\frac{r^2}{2}\right).\end{aligned}$$

Consequently, we deduce from Propositions 2 and 3 that $\mathbb{S}$ satisfies Assumption 1 with $\gamma = 1$ and $D(S(\overline{V})) = 40\overline{D}(\overline{V})$ for all $\overline{V} \in \mathbb{V}$. We may therefore apply Theorem 1 and get (recalling that $\tilde{f}_\lambda = \Pi_{\overline{\mathcal{C}}}\hat{f}_\lambda$ and $\hat{s}_\lambda = q_{\tilde{f}_\lambda}$)

$$C\mathbb{E}\left[H^2(q_f, \hat{s}_{\hat{\lambda}})\right] \le \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[H^2(q_f, q_{\tilde{f}_\lambda})\right] + \mathbb{E}\left[A(q_{\tilde{f}_\lambda}, \mathbb{S}_\lambda)\right]\right\},$$

and deduce that for some constant C' depending on c_0, $\overline{R}$, $\underline{R}$ and Σ only,

$$C'\mathbb{E}\left[\left\|f - \Pi_{\overline{\mathcal{C}}}\tilde{f}\right\|^2\right] \le \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\left\|f - \Pi_{\overline{\mathcal{C}}}\hat{f}_\lambda\right\|^2 + \mathbb{E}\left[A(q_{\tilde{f}_\lambda}, \mathbb{S}_\lambda)\right]\right]\right\}.$$

It remains to bound $A(q_{\tilde{f}_\lambda}, \mathbb{S}_\lambda)$ from above for all $\lambda \in \Lambda$. By using Proposition 7, for all $\lambda \in \Lambda$ and $\overline{V} \in \mathbb{V}_\lambda$,

$$\begin{aligned}A(\hat{s}_\lambda, \mathbb{S}_\lambda) &\le \inf_{t'\in V}\left[H^2(q_{\tilde{f}_\lambda}, q_{t'}) + \mathrm{pen}(q_{t'})\right] \\ &\le \overline{R}^2 \inf_{t'\in V}\left\|\Pi_{\overline{\mathcal{C}}}\hat{f}_\lambda - t'\right\|^2 + c_0\tau\left(40\overline{D}(\overline{V}) + \overline{\Delta}(\overline{V})\right) \\ &\le 8\overline{R}^2\left[\inf_{t'\in\overline{V}}\left\|\Pi_{\overline{\mathcal{C}}}\hat{f}_\lambda - t'\right\|^2 + \eta^2(\overline{V})\tau\right] + c_0\tau\left(40\overline{D}(\overline{V}) + \overline{\Delta}(\overline{V})\right)\end{aligned}$$

$$
\begin{aligned}
&\leq 8\overline{R}^2\left[\inf_{t'\in\overline{V}}\left\|\Pi_{\overline{\mathcal{C}}}\hat{f}_\lambda - f + f - \hat{f}_\lambda + \hat{f}_\lambda - t'\right\|^2 + \eta^2(\overline{V})\tau\right]\\
&\quad + c_0\tau\left(40\overline{D}(\overline{V}) + \overline{\Delta}(\overline{V})\right)\\
&\leq 24\overline{R}^2\left[\left\|f - \Pi_{\overline{\mathcal{C}}}\hat{f}_\lambda\right\|^2 + \left\|f - \hat{f}_\lambda\right\|^2 + \inf_{t'\in\overline{V}}\left\|\hat{f}_\lambda - t'\right\|^2 + \eta^2(\overline{V})\tau\right]\\
&\quad + c_0\tau\left(40\overline{D}(\overline{V}) + \overline{\Delta}(\overline{V})\right)\\
&\leq C''\left[\left\|f - \hat{f}_\lambda\right\|^2 + \left\|\hat{f}_\lambda - \Pi_{\overline{V}}\hat{f}_\lambda\right\|^2 + \overline{D}(\overline{V}) + \overline{\Delta}(\overline{V})\right]
\end{aligned}
$$

for some C'' depending on $\overline{R}$, $\underline{R}$ and c_0 only which concludes the proof.

9.11 Proof of Corollary 12

For each $\lambda \in \Lambda$, let $\overline{\lambda}$ be the linear map which coincides with λ on $\overline{V}_\lambda$ and takes the value 0 on its orthogonal. On the one hand,

$$
\begin{aligned}
\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right] &= \|f - \lambda f\|^2 + \operatorname{tr}\left(\lambda^2\right)\sigma^2\\
&= \sum_{k=1}^{n}(1-\lambda_{(k)})^2\langle f, v_{(k)}\rangle^2 + \sigma^2\sum_{k=1}^{n}\lambda_{(k)}^2.
\end{aligned}
$$

On the other hand, the definition of $D = D_\lambda$ entails that for all $k > D$, $\lambda_{(k)} \leq 1/2$ (with the convention $\lambda_{(n+1)} = 0$) and therefore

$$
\begin{aligned}
&\mathbb{E}\left[\inf_{t\in\overline{V}_\lambda}\left\|\hat{f}_\lambda - t\right\|^2\right]\\
&\leq \mathbb{E}\left[\left\|\lambda X - \overline{\lambda}X\right\|^2\right] = \mathbb{E}\left[\left\|(\lambda-\overline{\lambda})f + (\lambda-\overline{\lambda})\varepsilon\right\|^2\right]\\
&= \left\|(\lambda-\overline{\lambda})f\right\|^2 + \operatorname{tr}\left((\lambda-\overline{\lambda})^2\right)\sigma^2 = \sum_{k=1+D}^{n}\lambda_{(k)}^2\langle f, v_{(k)}\rangle^2 + \sigma^2\sum_{k=1+D}^{n}\lambda_{(k)}^2\\
&\leq \sum_{k=1}^{n}(1-\lambda_{(k)})^2\langle f, v_{(k)}\rangle^2 + \sigma^2\sum_{k=1}^{n}\lambda_{(k)}^2 = \mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right]
\end{aligned}
$$

and, since $D/4 \leq \sum_{k=1}^{D}\lambda_{(k)}^2 \leq \sigma^{-2}\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right]$, we conclude the proof by using Theorem 6.

Acknowledgements The author is thankful to Lucien Birgé for his careful reading of the paper and his thoughtful comments. We also thank the two referees for their comments and questions that have led to an improved version of the present paper.

References

1. Antoniadis, A., Besbeas, P., Sapatinas, T.: Wavelet shrinkage for natural exponential families with cubic variance functions. Sankhyā Ser. A, **63**(3), 309–327. Special issue on wavelets (2001)
2. Antoniadis, A., Sapatinas, T.: Wavelet shrinkage for natural exponential families with quadratic variance functions. Biometrika **88**(3), 805–820 (2001)
3. Arlot, S.: Rééchantillonnage et Sélection de modèles. PhD thesis, University Paris XI (2007)
4. Arlot, S.: Model selection by resampling penalization. Electron. J. Stat. **3**, 557–624 (2009)
5. Arlot, S., Bach, F.: Data-driven calibration of linear estimators with minimal penalties. Technical report, HAL : hal-00414774, version 1 (2009)
6. Assouad, P.: Deux remarques sur l'estimation. C. R. Acad. Sci. Paris Sér. I Math. **296**(23), 1021–1024 (1983)
7. Baraud, Y.: Model selection for regression on a fixed design. Probab. Theory Relat. Fields **117**(4), 467–493 (2000)
8. Baraud, Y., Birgé, L.: Estimating the intensity of a random measure by histogram type estimators. Probab. Theory Relat. Fields **143**(1–2), 239–284 (2009)
9. Baraud, Y., Giraud, C., Huet, S.: Gaussian model selection with an unknown variance. Ann. Stat. **37**(2), 630–672 (2009)
10. Barron, A., Birgé, L., Massart, P.: Risk bounds for model selection via penalization. Probab. Theory Relat. Fields **113**(3), 301–413 (1999)
11. Barron, A.R., Cover, T.M.: Minimum complexity density estimation. IEEE Trans. Inform. Theory **37**(4), 1034–1054 (1991)
12. Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of lasso and Dantzig selector. Ann. Stat. **37**(4), 1705–1732 (2009)
13. Birgé, L.: Approximation dans les espaces métriques et théorie de l'estimation. Z. Wahrsch. Verw. Gebiete **65**(2), 181–237 (1983)
14. Birgé, L.: Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. Ann. Inst. H. Poincaré Probab. Stat. **20**(3), 201–223 (1984a)
15. Birgé, L.: Sur un théorème de minimax et son application aux tests. Probab. Math. Stat. **3**(2), 259–282 (1984b)
16. Birgé, L.: Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Ann. Inst. H. Poincaré Probab. Stat. **42**(3), 273–325 (2006)
17. Birgé, L.: Model selection for Poisson processes. In: Cator, E., Jongbloed, G., Kraaikamp, C., Lopuhaä, R., Wellner, J. (eds.) Asymptotics: particles, processes and inverse problems, Festschrift for Piet Groeneboom, vol. 55, pp. 32–64. IMS Lecture Notes—Monograph Series (2007)
18. Birgé, L.: Model selection for density estimation with L2-loss. Technical report, arXiv:0808.1416 (2008)
19. Birgé, L., Massart, P.: An adaptive compression algorithm in Besov spaces. Constr. Approx. **16**(1), 1–36 (2000)
20. Birgé, L., Massart, P.: Gaussian model selection. J. Eur. Math. Soc. (JEMS) **3**(3), 203–268 (2001)
21. Borovkov, A.A.: Mathematical Statistics. Gordon and Breach, Amsterdam. Translated from the Russian by A. Moullagaliev and revised by the author (1998)
22. Bunea, F., Tsybakov, A.B., Wegkamp, M.H.: Aggregation for Gaussian regression. Ann. Stat. **35**(4), 1674–1697 (2007)
23. Candès, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**(6), 2313–2351 (2007)
24. Castellan, G.: Density estimation via exponential model selection. Technical report, 00.25 Université Paris XI, Orsay (2000)
25. Castellan, G.: Sélection d'histogrammes à l'aide d'un critère de type akaike. C.R.A.S. **330**, 729–732 (2000)
26. Catoni, O.: Statistical learning theory and stochastic optimization. In: Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. Springer, Berlin (2004)
27. Celisse, A.: Model selection via cross-validation in density estimation, regression, and change-points detection. PhD thesis, University Paris XI (2008)
28. Dalalyan, A.S., Tsybakov, A.B.: Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. Mach. Learning **72**(1–2), 39–61 (2008)
29. DeVore, R., Lorentz, G.: Constructive Approximation. Springer, Berlin (1993)

30. Genuer, R., Poggi, J.-M., Tuleau-Malot, C.: Variable selection using random forests. Pattern Recognit. Lett. (2010, to appear)
31. Giraud, C.: Mixing least-squares estimators when the variance is unknown. Bernoulli **14**(4), 1089–1107 (2008)
32. Goldenshluger, A.: A universal procedure for aggregating estimators. Ann. Stat. **37**(1), 542–568 (2009)
33. Goldenshluger, A., Lepski, O.: Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. Probab. Theory Relat. Fields **143**(1–2), 41–71 (2009)
34. Höskuldsson, A.: Variable and subset selection in PLS regression. Chemom. Intell. Lab. Syst. **55**, 23–38 (2001)
35. Juditsky, A., Nemirovski, A.: Functional aggregation for nonparametric regression. Ann. Stat. **28**(3), 681–712 (2000)
36. Kolaczyk, E.D., Nowak, R.D.: Multiscale likelihood analysis and complexity penalized estimation. Ann. Stat. **32**(2), 500–527 (2004)
37. Le Cam, L.: Convergence of estimates under dimensionality restrictions. Ann. Stat. **1**, 38–53 (1973)
38. Le Cam, L.: On local and global properties in the theory of asymptotic normality of experiments. In: Stochastic processes and related topics (Proc. Summer Res. Inst. Stat. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, vol. 1; dedicated to Jerzy Neyman), pp. 13–54. Academic Press, New York (1975)
39. Lepskiĭ, O.V.: A problem of adaptive estimation in Gaussian white noise. Teor. Veroyatnost. i Primenen. **35**(3), 459–470 (1990)
40. Lepskiĭ, O.V.: Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. Teor. Veroyatnost. i Primenen. **36**(4), 645–659 (1991)
41. Lepskiĭ, O.V.: Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. Teor. Veroyatnost. i Primenen. **37**(3), 468–481 (1992)
42. Lepskiĭ, O.V.: On problems of adaptive estimation in white Gaussian noise. In: Topics in Nonparametric Estimation. Adv. Soviet Math., vol. 12, pp. 87–106. American Mathematical Society, Providence (1992)
43. Leung, G., Barron, A.R.: Information theory and mixing least-squares regressions. IEEE Trans. Inform. Theory **52**(8), 3396–3410 (2006)
44. Lugosi, G., Nobel, A.: Consistency of data-driven histogram methods for density estimation and classification. Ann. Stat. **24**(2), 687–706 (1996)
45. Massart, P.: Concentration inequalities and model selection. Lecture Notes in Mathematics, vol. 1896. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. With a foreword by Jean Picard (2007)
46. Nemirovski, A.: Topics in non-parametric statistics. In: Lectures on Probability Theory and Statistics (Saint-Flour, 1998). Lecture Notes in Mathematics, vol. 1738, pp. 85–277. Springer, Berlin (2000)
47. Priestley, M.B., Chao, M.T.: Non-parametric function fitting. J. R. Stat. Soc. Ser. B **34**, 385–392 (1972)
48. Reynaud-Bouret, P.: Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. Probab. Theory Relat. Fields **126**(1), 103–153 (2003)
49. Rigollet, P., Tsybakov, A.B.: Linear and convex aggregation of density estimators. Math. Methods Stat. **16**(3), 260–280 (2007)
50. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
51. Tsybakov, A.B.: Optimal rates of aggregation. In: Proceedings of the 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines, pp. 303–313. Lecture Notes in Artificial Intelligence, vol. 2777. Springer, Berlin (2003)
52. Tsybakov, A.B.: Introduction à l'estimation non-paramétrique. Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 41. Springer, Berlin (2004)
53. Wegkamp, M.: Model selection in nonparametric regression. Ann. Stat. **31**, 252–273 (2003)
54. Yang, Y.: Model selection for nonparametric regression. Stat. Sinica **9**, 475–499 (1999)
55. Yang, Y.: Combining different procedures for adaptive regression. J. Multivar. Anal. **74**(1), 135–161 (2000)
56. Yang, Y.: Mixing strategies for density estimation. Ann. Stat. **28**(1), 75–87 (2000)
57. Yang, Y.: Adaptive regression by mixing. J. Am. Stat. Assoc. **96**(454), 574–588 (2001)