Electronic Journal of Statistics
Vol. 12 (2018) 1256–1298
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1413

Fast adaptive estimation of log-additive exponential models in Kullback-Leibler divergence*

Cristina Butucea

LAMA(UMR 8050), UPEM, UPEC, CNRS, F-77454, Marne-la-Vallée, France
and CREST, ENSAE, Université Paris-Saclay, France.
e-mail: cristina.butucea@u-pem.fr

Jean-François Delmas

CERMICS, École des Ponts, UPE, Champs-sur-Marne, France.
e-mail: delmas@cermics.enpc.fr

Anne Dutfoy

EDF Research & Development, Industrial Risk Management Department,
Palaiseau, France
e-mail: anne.dutfoy@edf.fr

Richard Fischer

LAMA(UMR 8050), UPEM, UPEC, CNRS, F-77454, Marne-la-Vallée, France
CERMICS, École des Ponts, UPE, Champs-sur-Marne, France
EDF Research & Development, Industrial Risk Management Department,
Palaiseau, France
e-mail: fischerr@cermics.enpc.fr

Abstract: We study the problem of nonparametric estimation of probability density functions (pdf) with a product form on the domain $\triangle = \{(x_1, \ldots, x_d) \in \mathbb{R}^d, 0 \leq x_1 \leq \cdots \leq x_d \leq 1\}$. Such pdf's appear in the random truncation model as the joint pdf of the observations. They are also obtained as maximum entropy distributions of order statistics with given marginals. We propose an estimation method based on the approximation of the logarithm of the density by a carefully chosen family of basis functions. We show that the method achieves a fast convergence rate in probability with respect to the Kullback-Leibler divergence for pdf's whose logarithm belong to a Sobolev function class with known regularity. In the case when the regularity is unknown, we propose an estimation procedure using convex aggregation of the log-densities to obtain adaptability. The performance of this method is illustrated in a simulation study.

MSC 2010 subject classifications: 62G07, 62G05, 62G20.
Keywords and phrases: Adaptive density estimation, aggregation, exponential family, Kullback-Leibler divergence, product form, Sobolev classes, truncation model.

Received May 2017.

*This work is partially supported by the French “Agence Nationale de la Recherche”, CIFRE n° 1531/2012, and by EDF Research & Development, Industrial Risk Management Department

Contents

1. Introduction

In this paper, we estimate probability density functions (pdf's) with product form on the simplex $\triangle = \{(x_1, \ldots, x_d) \in \mathbb{R}^d, 0 \leq x_1 \leq \cdots \leq x_d \leq 1\}$ by a nonparametric approach given a sample of n independent observations $\mathbb{X}^n = (X^1, \ldots, X^n)$. We restrict our attention to pdf's which can be written in the form:

$$f^0(x) = \exp\left(\sum_{i=1}^{d} \ell_i^0(x_i) - \mathrm{a}_0\right) \mathbf{1}_\triangle(x), \text{ for } x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \tag{1.1}$$

with ℓ_i^0 bounded, centered, measurable functions on $I = [0, 1]$ for all $1 \leq i \leq d$, and normalizing constant a_0. There are two different approaches to arrive at probability densities of this form. First, given independent $[0, 1]$-valued random variables we take the order statistic and we obtain a vector of dependent random variables supported on the simplex Δ. Second, given an order statistic with fixed one-dimensional marginal probability densities the joint probability density with maximum entropy is the unique density with the previous product form on the simplex.

The first example is the random truncation model, which was first formulated in [32], and has various applications ranging from astronomy ([30]), economics ([21], [19]) to survival data analysis ([26], [22], [29]). For $d = 2$, let (Z_1, Z_2) be a pair of independent random variables on I such that Z_i has density function p_i

for $i \in \{1, 2\}$. Let us suppose that we can only observe realizations of (Z_1, Z_2) if $Z_1 \leq Z_2$. Let $(\bar{Z}_1, \bar{Z}_2)$ denote a pair of random variables distributed as (Z_1, Z_2) conditionally on $Z_1 \leq Z_2$. Then the joint density function f^0 of $(\bar{Z}_1, \bar{Z}_2)$ is given by, for $x = (x_1, x_2) \in I^2$:

$$f^0(x) = \frac{1}{\alpha}\, p_1(x_1) p_2(x_2) \mathbf{1}_{\triangle}(x), \tag{1.2}$$

with $\alpha = \int_{I^2} p_1(x_1) p_2(x_2) \mathbf{1}_{\triangle}(x)\, dx$. Notice that f^0 is of the form required in (1.1):

$$f^0(x) = \exp(\ell_1^0(x_1) + \ell_2^0(x_2) - \mathrm{a}_0) \mathbf{1}_{\triangle}(x),$$

with ℓ_i^0 defined as $\ell_i^0 = \log(p_i) - \int_I \log(p_i)$ for $i \in \{1, 2\}$. According to Corollary 5.7. of [11], f^0 is the density of the maximum entropy distribution of order statistics with marginals $\mathbf{f}_1$ and $\mathbf{f}_2$ given by:

$$\mathbf{f}_1(x_1) = \frac{1}{\alpha}\, p_1(x_1) \int_{x_1}^{1} p_2(s)\, ds \quad \text{and} \quad \mathbf{f}_2(x_2) = \frac{1}{\alpha}\, p_2(x_2) \int_0^{x_2} p_1(s)\, ds.$$

This brings us to our second motivating example, for general dimension $d \geq 2$. There is an important amount of literature on copula models for order statistics, see e.g. [3]. Following that line of research, [11] gives a necessary and sufficient condition for the existence of a distribution of order statistics with fixed marginal cumulative distribution functions $\mathbf{F}_i$, $1 \leq i \leq d$, which has maximum entropy. Moreover, its explicit expression is given as a function of the marginal distributions.

Let us suppose, for the sake of simplicity, that all $\mathbf{F}_i$ are absolutely continuous with density function $\mathbf{f}_i$ supported on $I = [0, 1]$, and that $\mathbf{F}_{i-1} > \mathbf{F}_i$ on $(0, 1)$ for $2 \leq i \leq d$. Then the maximum entropy density $f_{\mathbf{F}}$, when it exists, is given by, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = \mathbf{f}_1(x_1) \prod_{i=2}^{d} h_i(x_i) \exp\left(- \int_{x_{i-1}}^{x_i} h_i(s)\, ds \right) \mathbf{1}_{\triangle}(x),$$

with $h_i = \mathbf{f}_i / (\mathbf{F}_{i-1} - \mathbf{F}_i)$ for $2 \leq i \leq d$. The density $f_{\mathbf{F}}$ is of the same form as f^0 in (1.1) with ℓ_i^0 defined as:

$$\ell_1^0 = \log(\mathbf{f}_1) + K_2 \quad \text{and} \quad \ell_i^0 = \log(h_i) - K_i + K_{i+1} \quad \text{for} \quad 2 \leq i \leq d,$$

with K_i, $2 \leq i \leq d$ a primitive of h_i chosen such that ℓ_i^0 are centered, and $K_{d+1} = c$ a constant. In order to estimate the joint density $f^0 = f_{\mathbf{F}}$, we can estimate from the data the marginals $\mathbf{F}_i$ and $\mathbf{f}_i$ for i from 1 to d and plug them into the previous analytical formula. However, this procedure leads to untractable evaluations of both L^2 and Kullback-Leibler risks.

We present an additive exponential series model specifically designed to estimate such densities. This exponential model is a multivariate version of the exponential series estimator considered in [6] in the univariate setting. Essentially, we approximate the functions ℓ_i^0 by their projections on a family of Jacobi polynomials $(\varphi_{i,k}, k \in \mathbb{N})$, which are orthonormal for each $1 \leq i \leq d$ with respect to the i-th marginal of the Lebesgue measure on the support $\triangle$. The model takes the form, for $\theta = (\theta_{i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i)$ and $x = (x_1, \dots, x_d) \in \triangle$:

$$f_\theta(x) = \exp\left(\sum_{i=1}^{d} \sum_{k=1}^{m_i} \theta_{i,k}\varphi_{i,k}(x_i) - \psi(\theta)\right) \mathbf{1}_\triangle(x),$$

with $\psi(\theta) = \log\left(\int_\triangle \exp\left(\sum_{i=1}^d \sum_{k=1}^{m_i} \theta_{i,k}\varphi_{i,k}(x_i)\right) dx\right)$. Even though the Jacobi polynomials $(x \mapsto \varphi_{i,k}(x_i),\ k \in \mathbb{N})$ are orthonormal (with respect to the Lebesgue measure on $\triangle$) for each $1 \leq i \leq d$, if we take $i \neq j$, the families $(x \mapsto \varphi_{i,k}(x_i),\ k \in \mathbb{N})$ and $(x \mapsto \varphi_{j,k}(x_j),\ k \in \mathbb{N})$ are not orthogonal . However, we construct the Gram matrix of the family $(\varphi_{[i],k}, i \in 1, ..., d)$ on the simplex, where $\varphi_{[i],k}$ is $\varphi_{i,k}$ seen as a function of its i-th coordinate on the simplex. We calculate explicitly the largest and the smallest eigenvalues of this matrix in order to control the stochastic fluctuations of our estimator. The exact definition and further properties of these polynomials can be found in the Appendix. We estimate the parameters of the model by $\hat{\theta} = (\hat{\theta}_{i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i)$, obtained by solving the maximum likelihood equations:

$$\int_\triangle \varphi_{i,k}(x_i) f_{\hat{\theta}}(x)\, dx = \frac{1}{n}\sum_{j=1}^{n} \varphi_{i,k}(X_i^j) \qquad \text{for } 1 \leq i \leq d,\ 1 \leq k \leq m_i.$$

Approximation of log-densities by polynomials appears in [18] as an application of the maximum entropy principle, while [14] shows existence and consistency of the maximum likelihood estimation. We measure the quality of the estimator $f_{\hat{\theta}}$ of f^0 by the Kullback-Leibler divergence $D\left(f^0 \| f_{\hat{\theta}}\right)$ defined as:

$$D\left(f^0 \| f_{\hat{\theta}}\right) = \int_\triangle f^0 \log\left(f^0 / f_{\hat{\theta}}\right).$$

Convergence rates in Kullback-Leibler divergence of nonparametric density estimators have been given by [20] for kernel density estimators, [6] and [33] for the exponential series estimators, [5] for histogram-based estimators, and [25] for wavelet-based log-density estimators. Here, we give results for the convergence rate in probability when the functions ℓ_i^0 belong to a Sobolev space with regularity $r_i > d$ for all $1 \leq i \leq d$. We show that if we take $m = m(n) = (m_1(n), \dots, m_d(n))$ members of the families $(\varphi_{i,k}, k \in \mathbb{N})$, $1 \leq i \leq d$, and let m_i grow with n such that $(\sum_{i=1}^d m_i^{2d})(\sum_{i=1}^d m_i^{-2r_i})$ and $(\sum_{i=1}^d m_i)^{2d+1}/n$ tend to 0, then the maximum likelihood estimator $f_{\hat{\theta}_{m,n}}$ verifies:

$$D\left(f^0 \| f_{\hat{\theta}_{m,n}}\right) = O_{\mathbb{P}}\left(\sum_{i=1}^{d}\left(m_i^{-2r_i} + \frac{m_i}{n}\right)\right).$$

Notice that this is the sum of the same univariate convergence rates as in [6]. The fact that the underlying pdf is log-additive on its support explains why the global d-dimensional risk is reduced to the sum of d one-dimensional estimation risks. However, the support is a major constraint in our model and this adds technical difficulties. By choosing m_i proportional to $n^{1/(2r_i+1)}$, which gives the optimal convergence rate $O_{\mathbb{P}}(n^{-2r_i/(2r_i+1)})$ in the univariate case as shown in [35], our estimator achieves a convergence rate of $O_{\mathbb{P}}(n^{-2\min(r)/(2\min(r)+1)})$. Recall that in this paper the dimension d is fixed with n. Note that a global choice $m_i = n^{1/(2\min(r)+1)}$ also achieves the optimal convergence rate. Therefore by exploiting the special structure of the underlying density, and carefully choosing the basis functions, we managed to reduce the problem of estimating a d-dimensional pdf to that of estimating d one-dimensional pdf's. We highlight the fact that this constitutes a significant gain over convergence rates of general nonparametric multivariate density estimation methods.

In most cases the smoothness parameters r_i, $1 \leq i \leq d$, are not available, therefore a method which adapts to the unknown smoothness is required to estimate the density with the best possible convergence rate. Adaptive methods for function estimation based on a random sample include Lepski's method, model selection, wavelet thresholding and aggregation of estimators.

Lepski's method, originating from [28], consists of constructing a grid of regularities, and choosing among the minimax estimators associated to each regularity the best estimator by an iterative procedure based on the available sample. This method was extensively applied for Gaussian white noise model, regression, and density estimation, see [9] and references therein. Adaptation via model selection with a complexity penalization criterion was considered by [8] and [4] for a large variety of models including wavelet-based density estimation. Loss in the Kullback-Leibler distance for model selection was studied in [34] and [13] for mixing strategies, and in [36] for the information complexity minimization strategy. More recently, bandwidth selection for multivariate kernel density estimation was addressed in [17] for L^s risk, $1 \leq s < \infty$, and [27] for L^∞ risk. Wavelet based adaptive density estimation with thresholding was considered in [24] and [15], where an upper bound for the rate of convergence was given for a collection of Besov-spaces. Linear and convex aggregate estimators appear in the more recent work [31] with an application to adaptive density estimation in expected L^2 risk, with sample splitting.

Here we extend the convex aggregation scheme for the estimation of the logarithm of the density proposed in [12] to achieve adaptive optimality. We take the estimator $f_{\hat{\theta}_{m,n}}$ for different values of $m \in \mathcal{M}_n$, where $\mathcal{M}_n$ is a sequence of sets of parameter configurations with increasing cardinality. These estimators are not uniformly bounded as required in [12], but we show that they are uniformly bounded in probability and that it does not change the general result. The different values of m correspond to different values of the regularity parameters. The convex aggregate estimator f_λ takes the form:

$$f_\lambda(x) = \exp\left(\sum_{m\in\mathcal{M}_n} \lambda_m \left(\sum_{i=1}^{d}\sum_{k=1}^{m_i} \theta_{i,k}\varphi_{i,k}(x_i)\right) - \psi_\lambda\right) \mathbf{1}_\triangle(x),$$

with $\lambda \in \Lambda^+ = \{\lambda = (\lambda_m, m \in \mathcal{M}_n), \lambda_m \geq 0 \text{ and } \sum_{m\in\mathcal{M}_n} \lambda_m = 1\}$ and normalizing constant ψ_λ given by:

$$\psi_\lambda = \log\left(\int_\triangle \exp\left(\sum_{m\in\mathcal{M}_n} \lambda_m \left(\sum_{i=1}^{d}\sum_{k=1}^{m_i} \theta_{i,k}\varphi_{i,k}(x_i)\right)\right) dx\right).$$

To apply the aggregation method, we split our sample $\mathbb{X}^n$ into two parts $\mathbb{X}_1^n$ and $\mathbb{X}_2^n$, with size proportional to n. We use the first part to create the estimators $f_{\hat{\theta}_{m,n}}$, then we use the second part to determine the optimal choice of the aggregation parameter $\hat{\lambda}_n^*$. We select $\hat{\lambda}_n^*$ by maximizing a penalized version of the log-likelihood function. We show that this method gives a sequence of estimators $f_{\hat{\lambda}_n^*}$, free of the smoothness parameters $r_1, \ldots, r_d$, which verifies:

$$D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) = O_{\mathbb{P}}\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right).$$

In summary, we give an adaptive minimax estimator of the joint density of an order statistic with maximal entropy within the family having the same marginal distributions. In order to achieve this, we project the log-density on a family of Jacobi polynomials and estimate their coefficients by maximum likelihood for different smoothness values. The Jacobi polynomials are a natural choice on the simplex. As an alternative, wavelet bases on the simplex could be used, but the computational part is to the best of our knowledge much more involved. The main difficulty is to control the correlations induced by the fact that the family of Jacobi polynomials are not orthogonal with respect to the Lebesgue measure on the simplex $\triangle$. The algorithm was implemented in [10] on a set of real data issued from industrial applications. At the last step, we aggregate these estimators into an adaptive procedure, following the previous non asymptotic results in [12], and show here that there is no loss in the rate due to adaptation to the smoothness. Our estimator is a bona-fide probability density and having the support $\triangle$ of an order statistic.

We considered as a natural choice the Kullback-Leibler divergence, as a loss function in the context of maximum entropy distributions, hence the log-additive model for density estimation in this setup. One might consider a more classical L^2-risk and estimate the density under its coordinate-wise product form on the simplex $\triangle$. Then a projection on the family of Jacobi polynomials can be estimated by a least squares procedure for different smoothness values and analogous results can be established, using the tools developed in this paper. An aggregation procedure in L^2 would similarly provide an adaptive minimax procedure. However, the resulting estimator might take negative values and further transformations of such an estimator may change the smoothness or the dependence structure.

The rest of the paper is organized as follows. In Section 2 we introduce the notations used in the rest of the paper. In Section 3, we describe the additive exponential series model and the estimation procedure, then we show that the estimator converges to the true underlying density with a convergence rate that is the sum of the convergence rates for the same type of univariate model, see Theorem 3.3. We consider an adaptive method with convex aggregation of the logarithms of the previous estimators to adapt to the unknown smoothness of the underlying density in Section 4, see Theorem 4.1. We assess the performance of the adaptive estimator via a simulation study in Section 5. The definition of the basis functions and their properties used during the proofs are given in Section 6. The detailed proofs of the results in Section 3 and 4 are contained in Sections 7, 8 and 9.

2. Notations

Let $I = [0, 1]$, $d \geq 2$ and $\triangle = \{(x_1, \ldots, x_d) \in I^d, x_1 \leq x_2 \leq \ldots \leq x_d\}$ denote the simplex of I^d. For an arbitrary real-valued function h_i defined on I with $1 \leq i \leq d$, let $h_{[i]}$ be the function defined on $\triangle$ such that for $x = (x_1, \ldots, x_d) \in \triangle$:

$$h_{[i]}(x) = h_i(x_i)\mathbf{1}_\triangle(x). \tag{2.1}$$

Let q_i, $1 \leq i \leq d$ be the one-dimensional marginals of the Lebesgue measure on $\triangle$:

$$q_i(dt) = \frac{1}{(d-i)!(i-1)!}(1-t)^{d-i}t^{i-1}\,\mathbf{1}_I(t)\,dt. \tag{2.2}$$

If $h_i \in L^1(q_i)$, then we have: $\int_\triangle h_{[i]} = \int_I h_i q_i$.

For a measurable function f, let $\| f \|_\infty$ be the usual sup norm of f on its domain of definition. For f defined on $\triangle$, let $\| f \|_{L^2} = \sqrt{\int_\triangle f^2}$. For f defined on I, let $\| f \|_{L^2(q_i)} = \sqrt{\int_I f^2 q_i}$.

For a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, let $\min(x)$ $(\max(x))$ denote the smallest (largest) component.

Let us denote the support of a probability density g by supp $(g) = \{x \in \mathbb{R}^d, g(x) > 0\}$. Let $\mathcal{P}(\triangle)$ denote the set of probability densities on $\triangle$. For $g, h \in \mathcal{P}(\triangle)$, the Kullback-Leibler distance $D\left(g\|h\right)$ is defined as:

$$D\left(g\|h\right) = \int_\triangle g \log\left(g/h\right).$$

Recall that $D\left(g\|h\right) \in [0, +\infty]$.

Definition 2.1. *We say that a probability density $f^0 \in \mathcal{P}(\triangle)$ has a product form if there exist $(\ell_i^0, 1 \leq i \leq d)$ bounded measurable functions defined on I such that $\int_I \ell_i^0 q_i = 0$ for $1 \leq i \leq d$ and a.e. on $\triangle$:*

$$f^0(x) = \exp\left(\ell^0(x) - \mathrm{a}_0\right)\mathbf{1}_\triangle(x), \tag{2.3}$$

with $\ell^0 = \sum_{i=1}^d \ell^0_{[i]}$ *and* $\mathrm{a}_0 = \log\left(\int_\triangle \exp\left(\ell^0\right)\right)$, *that is*

$$f^0(x) = \exp\left(\sum_{i=1}^d \ell_i^0(x_i) - \mathrm{a}_0\right)$$

for a.e. $x = (x_1, \ldots, x_d) \in \triangle$.

Definition 2.1 implies that supp $(f^0) = \triangle$ and f^0 is bounded. Let $\mathbb{X}^n = (X^1, \ldots, X^n)$ denote an i.i.d. sample of size n from the density f^0.

For $1 \leq i \leq d$, let $(\varphi_{i,k}, k \in \mathbb{N})$ be the family of orthonormal polynomials on I with respect to the measure q_i; see Section 6 for a precise definition of those polynomials and some of their properties. Recall $\varphi_{[i],k}(x) = \varphi_{i,k}(x_i)$ for $x = (x_1, \ldots, x_d) \in \triangle$. Notice that $(\varphi_{[i],k}, 1 \leq i \leq d, k \in \mathbb{N})$ is a family of normal polynomials with respect to the Lebesgue measure on $\triangle$, but not orthogonal.

Let $m = (m_1, \ldots, m_d) \in (\mathbb{N}^*)^d$ and set $|m| = \sum_{i=1}^d m_i$. We define the $\mathbb{R}^{|m|}$-valued function $\varphi_m = (\varphi_{[i],k}; 1 \leq k \leq m_i, 1 \leq i \leq d)$ and the $\mathbb{R}^{m_i}$-valued functions $\varphi_{i,m} = (\varphi_{i,k}; 1 \leq k \leq m_i)$ for $1 \leq i \leq d$. For $\theta = (\theta_{i,k}; 1 \leq k \leq m_i, 1 \leq i \leq d)$ and $\theta' = (\theta'_{i,k}; 1 \leq k \leq m_i, 1 \leq i \leq d)$ elements of $\mathbb{R}^{|m|}$, we denote the scalar product:

$$\theta \cdot \theta' = \sum_{i=1}^d \sum_{k=1}^{m_i} \theta_{i,k} \theta'_{i,k}$$

and the norm $\|\theta\| = \sqrt{\theta \cdot \theta}$. We define the function $\theta \cdot \varphi_m$ as follows, for $x \in \triangle$:

$$(\theta \cdot \varphi_m)(x) = \theta \cdot \varphi_m(x).$$

For a positive sequence $(a_n)_{n \in \mathbb{N}}$, the notation $O_{\mathbb{P}}(a_n)$ of stochastic boundedness for a sequence of random variables $(Y_n, n \in \mathbb{N})$ means that for every $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that:

$$\mathbb{P}\left(|Y_n / a_n| > C_\varepsilon\right) < \varepsilon \quad \text{for all } n \in \mathbb{N}.$$

3. Additive exponential series model

In this Section, we study the problem of estimation of an unknown density f^0 with a product form on the set $\triangle$, as described in (2.3), given the sample $\mathbb{X}^n$ drawn from f^0. Our goal is to give an estimation method based on a sequence of regular exponential models, which suits the special characteristics of the target density f^0. Estimating such a density with standard multidimensional nonparametric techniques naturally suffer from the curse of dimensionality, resulting in slow convergence rates for high-dimensional problems. We show that by taking into consideration that f^0 has a product form, we can recover the one-dimensional convergence rate for the density estimation, allowing for fast convergence of the estimator even if d is large. The quality of the estimators is measured by the Kullback-Leibler distance, as it has strong connections to the maximum entropy framework of [11].

We propose to estimate f^0 using the following additive exponential series model, for $m \in (\mathbb{N}^*)^d$:

$$f_\theta(x) = \exp\left(\theta \cdot \varphi_m(x) - \psi(\theta)\right) \mathbf{1}_\triangle(x), \tag{3.1}$$

with $\psi(\theta) = \log\left(\int_\triangle \exp\left(\theta \cdot \varphi_m\right)\right)$. This model is similar to the one introduced in [33], but there are two major differences. First, we have only kept the univariate terms in the multivariate exponential series estimator of [33] since the target probability density is the product of univariate functions. Second, we have restricted our model to $\triangle$ instead of the hyper-cube I^d, and we have chosen the basis functions $((\varphi_{i,k}, k \in \mathbb{N}), 1 \leq i \leq d)$ which are appropriate for this support.

Remark 3.1. In the general case, one has to be careful when considering a density f^0 with a product form and a support different from $\triangle$. Let f_i^0 denote the i-th marginal density function of f^0. If supp $(f_i^0) = A \subset \mathbb{R}$ for all $1 \leq i \leq d$, we can apply a strictly monotone mapping of A onto I to obtain a distribution with a product form supported on $\triangle$. When the supports of the marginals differ, there is no transformation that yields a random vector with a density as in Definition 2.1. A possible way to treat this case consists of constructing a family of basis functions which has similar properties with respect to supp (f^0) as the family $((\varphi_{i,k}, k \in \mathbb{N}), 1 \leq i \leq d)$ with respect to $\triangle$, which we discuss in detail in Section 6. Then we could define an exponential series model with this family of basis functions and support restricted to supp (f^0) to estimate f^0.

Let $m \in (\mathbb{N}^*)^d$. We define the following function on $\mathbb{R}^{|m|}$ taking values in $\mathbb{R}^{|m|}$ by:

$$A_m(\theta) = \int_\triangle \varphi_m f_\theta, \quad \theta \in \mathbb{R}^{|m|}. \tag{3.2}$$

According to Lemma 3 in [6], we have the following result on A_m.

Lemma 3.2. *The function A_m is one-to-one from $\mathbb{R}^{|m|}$ to $\Omega_m = A_m(\mathbb{R}^{|m|})$.*

We denote by $\Theta_m : \Omega_m \to \mathbb{R}^{|m|}$ the inverse of A_m. The empirical mean of the sample $\mathbb{X}^n$ of size n is:

$$\hat{\mu}_{m,n} = \frac{1}{n} \sum_{j=1}^n \varphi_m(X^j). \tag{3.3}$$

In Section 8.2 we show that $\hat{\mu}_{m,n} \in \Omega_m$ a.s. when $n \geq 2$.

For $n \geq 2$, we define a.s. the maximum likelihood estimator $\hat{f}_{m,n} = f_{\hat{\theta}_{m,n}}$ of f^0 by choosing:

$$\hat{\theta}_{m,n} = \Theta_m(\hat{\mu}_{m,n}). \tag{3.4}$$

The loss between the estimator $\hat{f}_{m,n}$ and the true underlying density f^0 is measured by the Kullback-Leibler divergence $D\left(f^0 \| \hat{f}_{m,n}\right)$.

For $r \in \mathbb{N}^*$, let $W_r^2(q_i)$ denote the Sobolev space of functions in $L^2(q_i)$, such that the $(r-1)$-th derivative is absolutely continuous and the L^2 norm of the

r-th derivative is finite:

$$W_r^2(q_i) = \left\{ h \in L^2(q_i); h^{(r-1)} \text{ is absolutely continuous and } h^{(r)} \in L^2(q_i) \right\}.$$

The main result is given by the following theorem whose proof is given in Section 8.3.

Theorem 3.3. *Let $f^0 \in \mathcal{P}(\triangle)$ be a probability density with a product form, see Definition 2.1. Assume the functions ℓ_i^0, defined in (2.3) belong to the Sobolev space $W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}$ with $r_i > d$ for all $1 \leq i \leq d$. Let $(X^n, n \in \mathbb{N}^*)$ be i.i.d. random variables with density distribution f^0. We consider a sequence $(m(n) = (m_1(n), \ldots, m_d(n)), n \in \mathbb{N}^*)$ such that $\lim_{n\to\infty} m_i(n) = +\infty$ for all $1 \leq i \leq d$, and which satisfies:*

$$\lim_{n\to\infty} |m|^{2d} \left(\sum_{i=1}^{d} m_i^{-2r_i} \right) = 0, \tag{3.5}$$

$$\lim_{n\to\infty} \frac{|m|^{2d+1}}{n} = 0. \tag{3.6}$$

The Kullback-Leibler distance $D\left(f^0 \| \hat{f}_{m,n}\right)$ of the maximum likelihood estimator $\hat{f}_{m,n}$ defined by (3.4) to f^0 converges in probability to 0 with the convergence rate:

$$D\left(f^0 \| \hat{f}_{m,n}\right) = O_{\mathbb{P}}\left(\sum_{i=1}^{d} m_i^{-2r_i} + \frac{|m|}{n} \right). \tag{3.7}$$

Remark 3.4. Let us take $(m^\circ(n) = (m_1^\circ(n), \ldots, m_d^\circ(n)), n \in \mathbb{N}^*)$ with $m_i^\circ(n) = \lfloor n^{1/(2r_i+1)} \rfloor$. This choice constitutes a balance between the bias and the variance term. Then the conditions (3.5) and (3.6) are satisfied, and we obtain that :

$$D\left(f^0 \| \hat{f}_{m^\circ,n}\right) = O_{\mathbb{P}}\left(\sum_{i=1}^{d} n^{-2r_i/(2r_i+1)} \right) = O_{\mathbb{P}}\left(n^{-2\min(r)/(2\min(r)+1)} \right).$$

Thus the convergence rate corresponds to the least smooth ℓ_i^0. This rate can also be obtained with a choice where all m_i are the same. Namely, with $(m^*(n) = (v^*(n), \ldots, v^*(n)), n \in \mathbb{N}^*)$ and $v^*(n) = \lfloor n^{1/(2\min(r)+1)} \rfloor$.

For $r = (r_1, \ldots, r_d) \in (\mathbb{N}^*)^d$, $r_i > d$ for $1 \leq i \leq d$, and a constant $\kappa > 0$, let :

$$\mathcal{K}_r(\kappa) = \left\{ f^0 = \exp\left(\sum_{i=1}^{d} \ell_{[i]}^0 - \mathrm{a}_0 \right) \in \mathcal{P}(\triangle); \| \ell_i^0 \|_\infty \leq \kappa, \| (\ell_i^0)^{(r_i)} \|_{L^2(q_i)} \leq \kappa \right\}. \tag{3.8}$$

The constants $\mathfrak{A}_1$ and $\mathfrak{A}_2$, appearing in the upper bounds during the proof of Theorem 3.3 (more precisely in Propositions 8.3 and 8.5), are uniformly bounded on $\mathcal{K}_r(\kappa)$, thanks to Corollary 6.13 and $\| \log(f^0) \|_\infty \leq 2d\kappa + |\log(d!)|$, which is due to (7.6). This yields the following corollary for the uniform convergence in probability on the set $\mathcal{K}_r(\kappa)$ of densities:

Corollary 3.5. *Under the assumptions of Theorem 3.3, we get the following result:*

$$\lim_{K\to\infty}\limsup_{n\to\infty}\sup_{f^0\in\mathcal{K}_r(\kappa)}\mathbb{P}\left(D\left(f^0\|\hat{f}_{m,n}\right)\geq\left(\sum_{i=1}^d m_i^{-2r_i}+\frac{|m|}{n}\right)K\right)=0.$$

Remark 3.6. Since we let r_i vary for each $1\leq i\leq d$, our class of densities $\mathcal{K}_r(\kappa)$ is anisotropic, i.e. the multivariate functions are not equally smooth in all directions. Estimation of anisotropic multivariate functions for L^s risk, $1\leq s\leq\infty$, was considered in multiple papers. For a Gaussian white noise model, [23] obtains minimax convergence rates on anisotropic Besov classes for L^s risk, $1\leq s<\infty$,while [7] gives the minimax rate of convergence on anisotropic Hölder classes for the L^∞ risk. For kernel density estimation, results on the minimax convergence rate for anisotropic Nikol'skii classes for L^s risk, $1\leq s<\infty$, can be found in [17]. These papers conclude in general, that if the considered class has smoothness parameters $\tilde{r}_i$ for the i-th coordinate, $1\leq i\leq d$, then the optimal convergence rate becomes $n^{-2\tilde{R}/(2\tilde{R}+1)}$ (multiplied with a logarithmic factor for L^∞ risk), with $\tilde{R}$ defined by the equation $1/\tilde{R}=\sum_{i=1}^d 1/\tilde{r}_i$. Since $\tilde{R}<\tilde{r}_i$ for all $1\leq i\leq d$, the convergence rate $n^{-2\min(r)/(2\min(r)+1)}$ is strictly better than the convergence rate for these anisotropic classes. In the isotropic case, when $r_i=r$ for all $1\leq i\leq d$, the minimax convergence rate specializes to $n^{-2r/(2r+d)}$ (which was obtained in [33] as an upper bound). This rate decreases exponentially when the dimension d increases. However, by exploiting the multiplicative structure of the model, we managed to obtain the univariate convergence rate $n^{-2r/(2r+1)}$, which is minimax optimal, see [35].

4. Adaptive estimation

Notice that the choice of the optimal series of estimators $\hat{f}_{m^*,n}$ with m^* defined in Remark 3.4 requires the knowledge of $\min(r)$ at least. When this knowledge is not available, we propose an adaptive method based on the proposed estimators in Section 3, which can mimic asymptotically the behaviour of the optimal choice. Let us introduce some notation first. We separate the sample $\mathbb{X}^n$ into two parts $\mathbb{X}_1^n$ and $\mathbb{X}_2^n$ of size $n_1=\lfloor C_e n\rfloor$ and $n_2=n-\lfloor C_e n\rfloor$ respectively, with some constant $C_e\in(0,1)$. The first part of the sample will be used to create our estimators, and the second half will be used in the aggregation procedure. Let $(N_n, n\in\mathbb{N}^*)$ be a sequence of non-decreasing positive integers depending on n such that $\lim_{n\to\infty}N_n=+\infty$. Let us denote:

$$\mathcal{N}_n=\left\{\lfloor n^{1/(2(d+j)+1)}\rfloor, 1\leq j\leq N_n\right\}\qquad\mathcal{M}_n=\left\{m=(v,\ldots,v)\in\mathbb{R}^d, v\in\mathcal{N}_n\right\}. \tag{4.1}$$

For $m\in\mathcal{M}_n$ let $\hat{f}_{m,n}$ be the additive exponential series estimator based on the first half of the sample, namely:

$$\hat{f}_{m,n}(x)=\exp\left(\hat{\theta}_{m,n}\cdot\varphi_m(x)-\psi(\hat{\theta}_{m,n})\right)\mathbf{1}_\triangle(x),$$

with $\hat{\theta}_{m,n}$ given by (3.4) using the sample $\mathbb{X}_1^n$ (replacing n with n_1 in the definition (3.3) of $\hat{\mu}_{m,n}$). Let :

$$\mathcal{F}_n = \{\hat{f}_{m,n}, m \in \mathcal{M}_n\}$$

denote the set of different estimators obtained by this procedure. Notice that Card $(\mathcal{F}_n) \leq$ Card $(\mathcal{M}_n) \leq N_n$. Recall that by Remark 3.4, we have that for $r = (r_1, \ldots, r_d)$ with $r_i > d$ and $n \geq \bar{n}$, where $\bar{n}$ is given by:

$$\bar{n} = \min\{n \in \mathbb{N}, N_n \geq \min(r) - d + 1\}, \tag{4.2}$$

the sequence of estimators $\hat{f}_{m^*,n}$, with $m^* = m^*(n) = (v^*, \ldots, v^*) \in \mathcal{M}_n$ given by $v^* = \lfloor n^{1/(2\min(r)+1)} \rfloor$, achieves the optimal convergence rate

$$O_{\mathbb{P}}(n^{-2\min(r)/(2\min(r)+1)}).$$

By letting N_n go to infinity, we ensure that for every combination of regularity parameters $r = (r_1, \ldots, r_d)$ with $r_i > d$, the sequence of optimal estimators $\hat{f}_{m^*,n}$ is included in the sets $\mathcal{F}_n$ for n large enough.

We use the second part of the sample $\mathbb{X}_2^n$ to create an aggregate estimator based on $\mathcal{F}_n$, which asymptotically mimics the performance of the optimal sequence $\hat{f}_{m^*,n}$. We will write $\hat{\ell}_{m,n} = \hat{\theta}_{m,n} \cdot \varphi_m$ to ease notation. We define the convex combination $\hat{\ell}_\lambda$ of the functions $\hat{\ell}_{m,n}$, $m \in \mathcal{M}_n$:

$$\hat{\ell}_\lambda = \sum_{m \in \mathcal{M}_n} \lambda_m \hat{\ell}_{m,n},$$

with aggregation weights $\lambda \in \Lambda^+ = \{\lambda = (\lambda_m, m \in \mathcal{M}_n) \in \mathbb{R}^{\mathcal{M}_n}, \lambda_m \geq 0$ and $\sum_{m \in \mathcal{M}_n} \lambda_m = 1\}$. For such a convex combination, we define the probability density function f_λ as:

$$f_\lambda = \exp(\hat{\ell}_\lambda - \psi_\lambda)\mathbf{1}_\triangle, \tag{4.3}$$

with $\psi_\lambda = \log\left(\int_\triangle \exp(\hat{\ell}_\lambda)\right)$. We apply the convex aggregation method for log-densities developed in [12] to get an aggregate estimator which achieves adaptability. Notice that the reference probability measure in this paper corresponds to $d!\mathbf{1}_\triangle(x)dx$. This implies that ψ_λ here differs from the ψ_λ of [12] by the constant $\log(d!)$, but this does not affect the calculations. The aggregation weights are chosen by maximizing the penalized maximum likelihood criterion H_n defined as:

$$H_n(\lambda) = \frac{1}{n_2} \sum_{X^j \in \mathbb{X}_2^n} \hat{\ell}_\lambda(X^j) - \psi_\lambda - \frac{1}{2} \text{pen}\,(\lambda), \tag{4.4}$$

with the penalizing function pen $(\lambda) = \sum_{m \in \mathcal{M}_n} \lambda_m \, D\left(f_\lambda \| \hat{f}_{m,n}\right)$. The convex aggregate estimator $f_{\hat{\lambda}_n^*}$ is obtained by setting:

$$\hat{\lambda}_n^* = \underset{\lambda \in \Lambda^+}{\text{argmax}}\ H_n(\lambda). \tag{4.5}$$

The main result of this section is given by the next theorem which asserts that if we choose $N_n = o(\log(n))$ such that $\lim_{n\to\infty} N_n = +\infty$, the series of convex aggregate estimators $f_{\hat{\lambda}^*_n}$ converge to f^0 with the optimal convergence rate, i.e. as if the smoothness was known.

Theorem 4.1. *Let $f^0 \in \mathcal{P}(\triangle)$ be a probability density with a product form given by (2.3). Assume the functions ℓ^0_i belongs to the Sobolev space $W^2_{r_i}(q_i)$, $r_i \in \mathbb{N}$ with $r_i > d$ for all $1 \le i \le d$. Let $(X^n, n \in \mathbb{N}^*)$ be i.i.d. random variables with density f^0. Let $N_n = o(\log(n))$ such that $\lim_{n\to\infty} N_n = +\infty$. The convex aggregate estimator $f_{\hat{\lambda}^*_n}$ defined by (4.3) with $\hat{\lambda}^*_n$ given by (4.5) converges to f^0 in probability with the convergence rate:*

$$D\left(f^0 \| f_{\hat{\lambda}^*_n}\right) = O_{\mathbb{P}}\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right). \tag{4.6}$$

The proof of this theorem is provided in Section 9. Similarly to Corollary 3.5, we have uniform convergence over sets of densities with increasing regularity. Recall the definition (3.8) of the set $\mathcal{K}_r(\kappa)$. Let $\mathcal{R}_n = \{j, d+1 \le j \le R_n\}$, where R_n satisfies the three inequalities:

$$R_n \le N_n + d, \tag{4.7}$$

$$R_n \le \left\lfloor n^{\frac{1}{2(d+N_n)+1}} \right\rfloor, \tag{4.8}$$

$$R_n \le \frac{\log(n)}{2\log(\log(N_n))} - \frac{1}{2}. \tag{4.9}$$

Corollary 4.2. *Under the assumptions of Theorem 4.1, we get the following result:*

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{r\in(\mathcal{R}_n)^d} \sup_{f^0\in\mathcal{K}_r(\kappa)} \mathbb{P}\left(D\left(f^0 \| f_{\hat{\lambda}^*_n}\right) \ge \left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) K\right) = 0.$$

Remark 4.3. For example when $N_n = \log(n)/(2\log(\log(n)))$, then (4.7), (4.8) and (4.9) are satisfied with $R_n = N_n$ for n large enough.

5. Simulation study: random truncation model

In this section we present the results of Monte Carlo simulation studies on the performance of the additive exponential series estimator. We take the example of the random truncation model introduced in Section 1 with $d = 2$, which is used in many applications. This model naturally satisfies our model assumptions.

Let $Z = (Z_1, Z_2)$ be a pair of independent random variable with density functions p_1, p_2 respectively such that $\triangle \subset \text{supp}\,(p)$, where $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ is the joint density function of Z. Suppose that we only observe pairs (Z_1, Z_2) if $0 \le Z_1 \le Z_2 \le 1$. Then the joint density function f of the observable pairs is given by, for $x = (x_1, x_2) \in \mathbb{R}^2$:

$$f(x) = \frac{p_1(x_1)p_2(x_2)}{\int_\triangle p(y)\,dy} \mathbf{1}_\triangle(x).$$

This corresponds to the form (1.2).

We will choose the densities p_1, p_2 from the following distributions:

- Normal(μ, σ^2) with $\mu \in \mathbb{R}, \sigma > 0$:

$$f_{\mu,\sigma^2}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \, \mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}},$$

- NormalMix$(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w)$ with $w \in (0,1)$:

$$f(t) = w f_{\mu_1,\sigma_1^2}(t) + (1-w) f_{\mu_2,\sigma_2^2}(t),$$

- Beta(α, β, a, b) with $0 < \alpha < \beta$, $a < 0$, $b > 1$:

$$f(t) = \frac{(t-a)^{\alpha-1}(b-t)^{\beta-\alpha-1}}{(b-a)^{\beta-1} B(\alpha, \beta-\alpha)} \mathbf{1}_{(a,b)}(t),$$

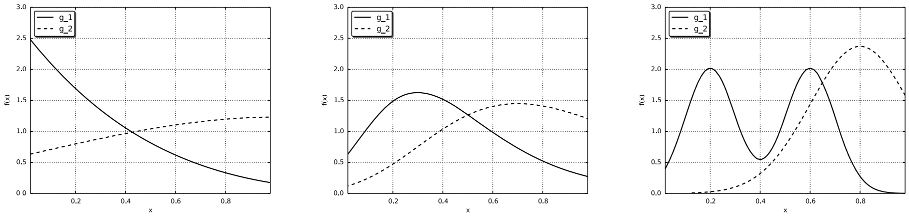
- Gumbel(α, β) with $\alpha > 0$, $\beta \in \mathbb{R}$:

$$f(t) = \alpha \, \mathrm{e}^{-\alpha(t-\beta) - \mathrm{e}^{-\alpha(t-\beta)}}.$$

The exact choices for densities p_1, p_2 are given in Table 1. Figure 1 shows the resulting density functions g_1 and g_2 for each case.

TABLE 1
Distributions for the left-truncated model used in the simulation study.

Model	p_1	p_2
Beta	Beta$(1, 6, -1, 2)$	Beta$(3, 5, -1, 2)$
Gumbel	Gumbel$(4, 0.3)$	Gumbel$(2.4, 0.7)$
Normal mix	NormalMix$(0.2, 0.1, 0.6, 0.1, 0.5)$	Normal$(0.8, 0.2)$



(a) Beta (B) (b) Gumbel (G) (c) Normal mix (NMix)

FIG 1. *Density functions g_1, g_2 of the left-truncated models used in the simulation study.*

To calculate the parameters $\hat{\theta}_{m,n}$, we recall that $\hat{\theta}_{m,n}$ is the solution of the equation (3.4), therefore can be also characterized as:

$$\hat{\theta}_{m,n} = \operatorname{argmax}_{\theta \in \mathbb{R}^{|m|}} \theta \cdot \hat{\mu}_{m,n} - \psi(\theta), \tag{5.1}$$

with $\hat{\mu}_{m,n}$ defined by (3.3), see Lemma 7.4 . We use a numerical optimisation method to solve (5.1) and obtain the parameters $\hat{\theta}_{m,n}$. We estimate our model with $m_1 = m_2 = \bar{m}$, and $\bar{m} = 1, 2, 3, 4$. We compute the final estimator based on the convex aggregation method proposed in Section 4. We ran 100 estimations with increasing sample sizes $n \in \{200, 500, 1000\}$, and we calculated the average Kullback-Leibler distance as well as the L^2 distance between f^0 and its estimator. We used 80% of the sample to calculate the initial estimators, and the remaining 20% to perform the aggregation. The distances were calculated by numerical integration. We compare the results with a truncated kernel density estimator with Gaussian kernel functions and bandwidth selection based on Scott's rule. The results are summarized in Table 2 and Table 3.

Table 2

Average Kullback-Leibler distances for the additive exponential series estimator (AESE) and the truncated kernel estimator (Kernel) based on 100 samples of size n. Variances provided in parenthesis.

KL	n=200		n=500		n=1000	
	AESE	Kernel	AESE	Kernel	AESE	Kernel
B	**0.0137**	0.0524	**0.0048**	0.0395	**0.0028**	0.0339
	(8.94E-05)	(1.73E-04)	(9.51E-06)	(4.61E-05)	(3.50E-06)	(2.14E-05)
G	**0.0204**	0.0249	**0.0089**	0.0180	**0.0050**	0.0154
	(1.48E-04)	(8.03E-05)	(2.88E-05)	(2.07E-05)	(6.70E-06)	(1.03E-05)
N	**0.0545**	0.0774	**0.0337**	0.0559	**0.0259**	0.0433
Mix	(4.51E-04)	(7.29E-05)	(1.88E-04)	(2.95E-05)	(2.50E-05)	(1.52E-05)

Table 3

Average L^2 distances for the additive exponential series estimator (AESE) and the truncated kernel estimator (Kernel) based on 100 samples of size n. Variances provided in parenthesis.

$\mathbb{L}^2$	n=200		n=500		n=1000	
	AESE	Kernel	AESE	Kernel	AESE	Kernel
B	**0.0536**	0.2107	**0.0200**	0.1660	**0.0120**	0.1429
	(1.42E-03)	(2.60E-03)	(2.27E-04)	(8.04E-04)	(7.45E-05)	(3.52E-04)
G	**0.0683**	0.0856	**0.0297**	0.0621	**0.0166**	0.0522
	(1.95E-03)	(9.94E-04)	(3.61E-04)	(2.49E-04)	(8.74E-05)	(1.19E-04)
N	**0.2314**	0.3534	**0.1489**	0.2545	**0.1112**	0.1952
Mix	(1.17E-02)	(1.43E-03)	(5.53E-03)	(6.95E-04)	(9.25E-04)	(3.83E-04)

We can conclude that the additive exponential series estimator outperforms the kernel density estimator both with respect to the Kullback-Leibler distance and the L^2 distance. As expected, the performance of both methods increases with the sample size. The boxplot of the 100 values of the Kullback-Leibler and L^2 distance for the different sample sizes can be found in Figures 2, 4 and 6. Figures 3, 5 and 7 illustrate the different estimators compared to the true joint density function for the three cases obtained with a sample size of 1000. We can observe that the additive exponential series method leads to a smooth estimator compared to the kernel method.

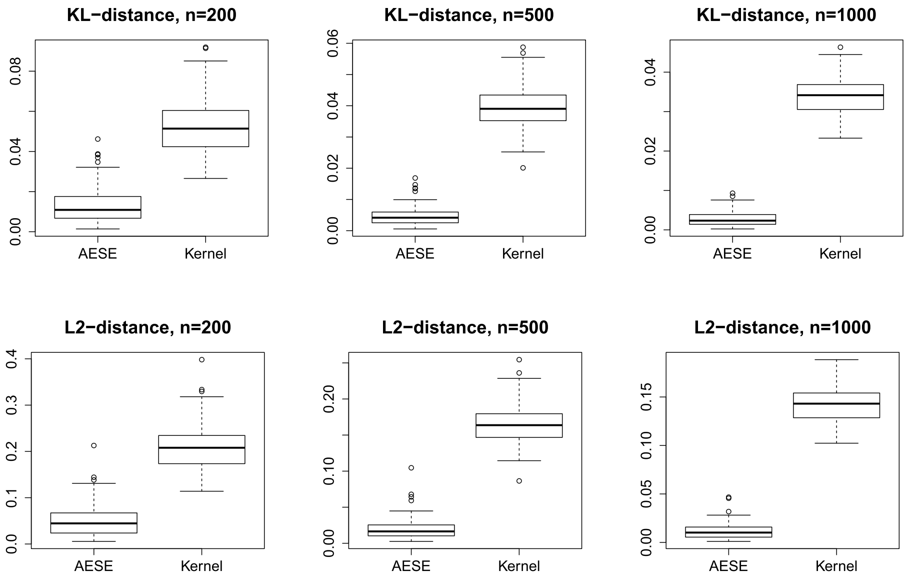
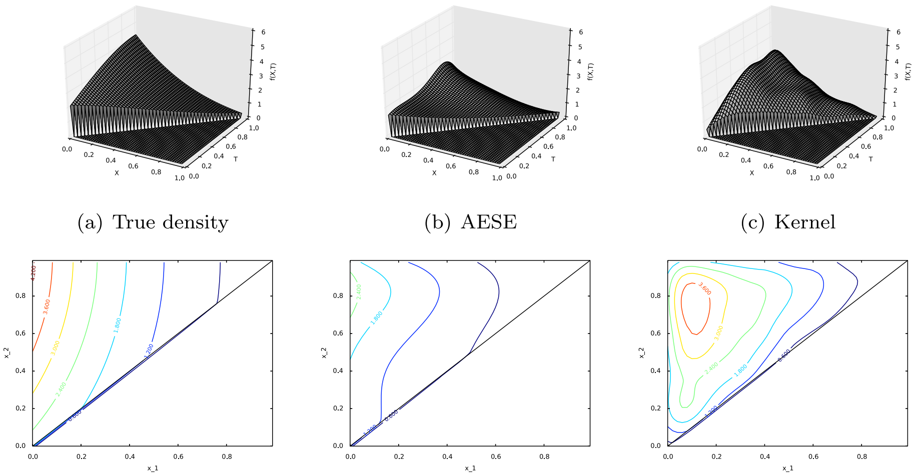


Fig 2. *Boxplot of the Kullback-Leibler and L^2 distances for the additive exponential series estimator (AESE) and the truncated kernel estimators with Beta marginals.*



(a) True density (b) AESE (c) Kernel

(d) True density (e) AESE (f) Kernel

Fig 3. *Joint density functions of the true density and its estimators with Beta marginals.*

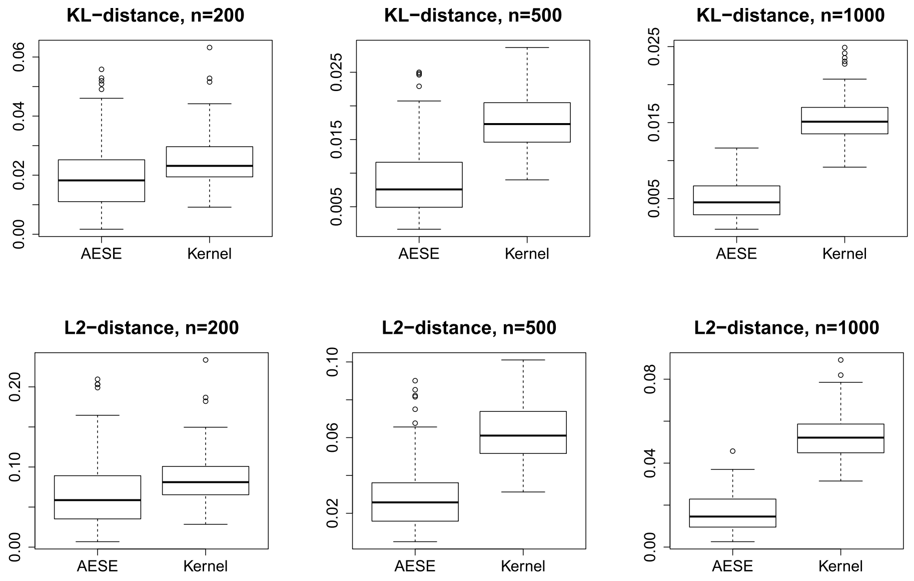
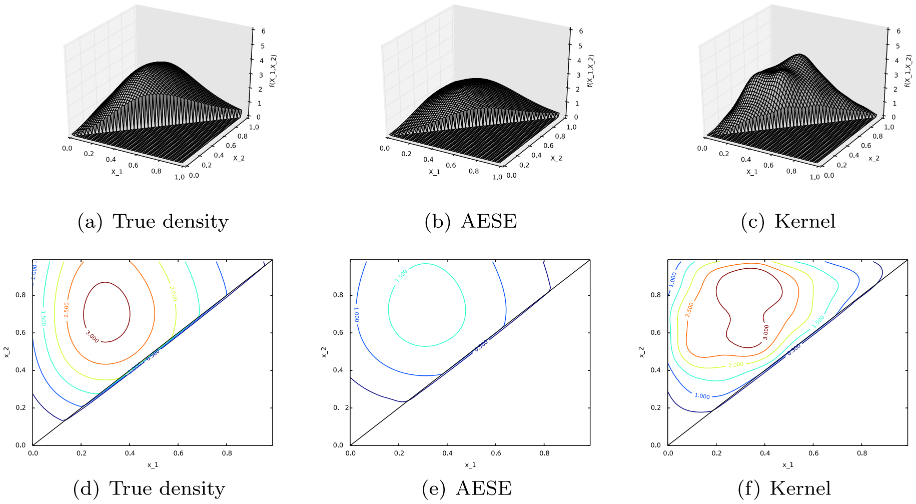


FIG 4. *Boxplot of the Kullback-Leibler and L^2 distances for the additive exponential series estimator (AESE) and the truncated kernel estimators with Gumbel marginals.*



(a) True density (b) AESE (c) Kernel

(d) True density (e) AESE (f) Kernel

FIG 5. *Joint density functions of the true density and its estimators with Gumbel marginals.*

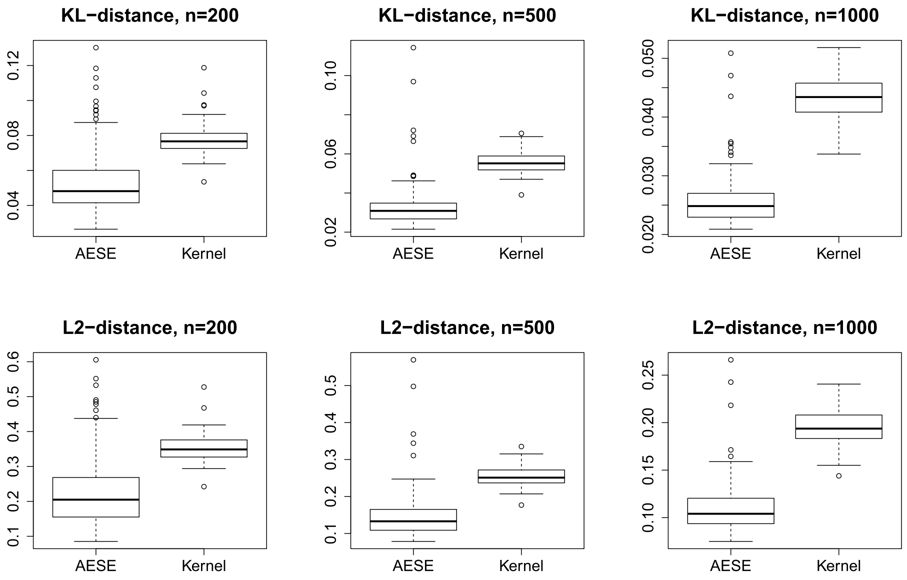
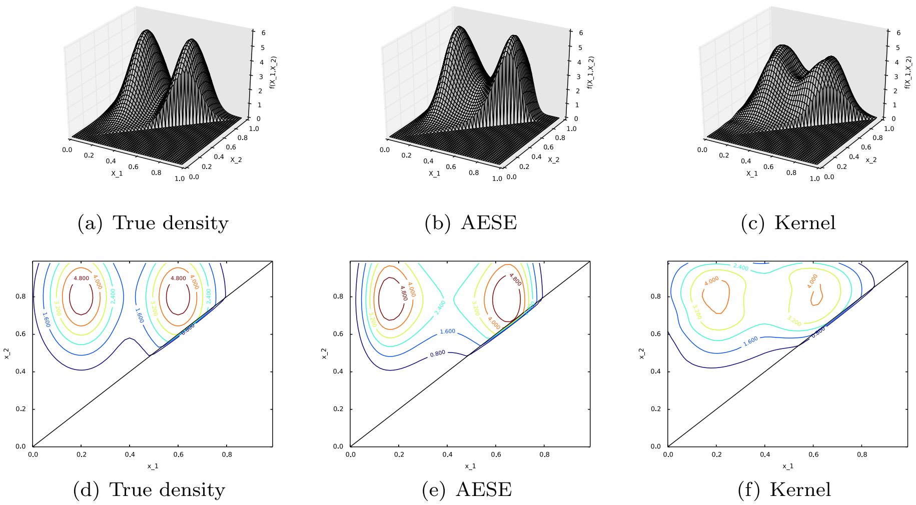


FIG 6. *Boxplot of the Kullback-Leibler and L^2 distances for the additive exponential series estimator (AESE) and the truncated kernel estimators with Normal mix marginals.*



(a) True density

(b) AESE

(c) Kernel

(d) True density

(e) AESE

(f) Kernel

FIG 7. *Joint density functions of the true density and its estimators with Normal mix marginals.*

Remark 5.1. The additive exponential series model encompasses a lot of popular choices for the marginals p_1, p_2. For example, the exponential distribution is included in the model for $m_i = 1$, and the normal distribution is included for $m_i = 2$. Thus we expect that if we choose exponential or normal distributions for p_1, p_2, we obtain even better results for the additive exponential series estimator, which was confirmed by the numerical experiments (not included here for brevity).

6. Appendix: Orthonormal series of polynomials

6.1. Jacobi polynomials

The following results can be found in [2] p. 774. The Jacobi polynomials $\left(P_k^{(\alpha,\beta)},\right.$ $k \in \mathbb{N})$ for $\alpha, \beta \in (-1, +\infty)$ are series of orthogonal polynomials with respect to the measure $w_{\alpha,\beta}(t)\mathbf{1}_{[-1,1]}(t)\,dt$, with $w_{\alpha,\beta}(t) = (1-t)^\alpha(1+t)^\beta$ for $t \in [-1,1]$. They are given by Rodrigues' formula, for $t \in [-1,1]$, $k \in \mathbb{N}$:

$$P_k^{(\alpha,\beta)}(t) = \frac{(-1)^k}{2^k k! w_{\alpha,\beta}(t)} \frac{d^k}{dt^k}\left[w_{\alpha,\beta}(t)(1-t^2)^k\right].$$

The normalizing constants are given by:

$$\begin{aligned} &\int_{-1}^{1} P_k^{(\alpha,\beta)}(t) P_\ell^{(\alpha,\beta)}(t) w_{\alpha,\beta}(t)\,dt \\ = \;& \mathbf{1}_{\{k=\ell\}} \frac{2^{\alpha+\beta+1}}{2k+\alpha+\beta+1} \frac{\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{\Gamma(k+\alpha+\beta+1)k!}. \end{aligned} \tag{6.1}$$

In what follows, we will be interested in Jacobi polynomials with $\alpha = d - i$ and $\beta = i - 1$, which are orthogonal to the weight function $w_{d-i,i-1}(t) = \mathbf{1}_{[-1,1]}(t)(1-t)^{d-i}(1+t)^{i-1}$. The leading coefficient of $P_k^{(d-i,i-1)}$ is:

$$\omega'_{i,k} = \frac{(2k+d-1)!}{2^k k!(k+d-1)!}. \tag{6.2}$$

Let $r \in \mathbb{N}^*$. Recall that $P_k^{(\alpha,\beta)}$ has degree k. The derivatives of the Jacobi polynomials $P_k^{(d-i,i-1)}$, $r \le k$, verify, for $t \in I$ (see Proposition 1.4.15 of [16]):

$$\frac{d^r}{dt^r} P_k^{(d-i,i-1)}(t) = \frac{(k+d-1+r)!}{2^r(k+d-1)!} P_{k-r}^{(d-i+r,i-1+r)}(t). \tag{6.3}$$

We also have:

$$\sup_{t\in[-1,1]} \left|P_k^{(d-i,i-1)}(t)\right| = \max\left(\frac{(k+d-i)!}{k!(d-i)!}, \frac{(k+i-1)!}{k!(i-1)!}\right). \tag{6.4}$$

6.2. Definition of the basis functions

Based on the Jacobi polynomials, we define a shifted version, normalized and adapted to the interval $I = [0, 1]$.

Definition 6.1. *For $1 \leq i \leq d$, $k \in \mathbb{N}$, we define for $t \in I$:*

$$\varphi_{i,k}(t) = \rho_{i,k} \sqrt{(d-i)!(i-1)!}\, P_k^{(d-i,i-1)}(2t-1),$$

with

$$\rho_{i,k} = \sqrt{\frac{(2k+d)k!(k+d-1)!}{((k+d-i)!(k+i-1)!)}}. \tag{6.5}$$

Recall the definition (2.2) of the marginals q_i of the Lebesgue measure on the simplex. According to the following Lemma, the polynomials $(\varphi_{i,k}, k \in \mathbb{N})$ form an orthonormal basis of $L^2(q_i)$ for all $1 \leq i \leq d$. Notice that $\varphi_{i,k}$ has degree k.

Lemma 6.2. *For $1 \leq i \leq d$, $k, \ell \in \mathbb{N}$, we have:*

$$\int_I \varphi_{i,k}\varphi_{i,\ell}\, q_i = \mathbf{1}_{\{k=\ell\}}.$$

Proof. We have, for $k, \ell \in \mathbb{N}$:

$$\begin{aligned}\int_I \varphi_{i,k}\varphi_{i,\ell}\, q_i &= \rho_{i,k}\rho_{i,\ell} \int_0^1 P_k^{(d-i,i-1)}(2t-1) P_\ell^{(d-i,i-1)}(2t-1)(1-t)^{d-i} t^{i-1}\, dt \\ &= \frac{\rho_{i,k}\rho_{i,\ell}}{2^d} \int_{-1}^1 P_k^{(d-i,i-1)}(s) P_\ell^{(d-i,i-1)}(s) w_{d-i,i-1}(s)\, ds \\ &= \mathbf{1}_{\{k=\ell\}},\end{aligned}$$

where we used (6.1) for the last equality. □

6.3. Mixed scalar products

Recall notation (2.1), so that $\varphi_{[i],k}(x) = \varphi_{i,k}(x_i)$ for $x = (x_1, \ldots, x_d) \in \triangle$. Notice that $(\varphi_{[i],k}, k \in \mathbb{N})$ is a family of orthonormal polynomials with respect to the Lebesgue measure on $\triangle$, for all $1 \leq i \leq d$.

We give the mixed scalar products of $(\varphi_{[i],k}, k \in \mathbb{N})$ and $(\varphi_{[j],\ell}, \ell \in \mathbb{N})$, $1 \leq i < j \leq d$ with respect to the Lebesgue measure on the simplex $\triangle$.

Lemma 6.3. *For $1 \leq i < j \leq d$ and $k, \ell \in \mathbb{N}$, we have:*

$$\int_\triangle \varphi_{[i],k}\, \varphi_{[j],\ell} = \mathbf{1}_{\{k=\ell\}} \sqrt{\frac{(j-1)!(d-i)!}{(i-1)!(d-j)!}} \sqrt{\frac{(k+d-j)!(k+i-1)!}{(k+d-i)!(k+j-1)!}}.$$

We also have $0 \leq \int_\triangle \varphi_{[i],k}\, \varphi_{[j],\ell} \leq 1$ for all $k, \ell \in \mathbb{N}$.

Proof. By integrating with respect to x_ℓ for $\ell \in \{1, \dots, d\} \backslash \{i, j\}$, we obtain:

$$\begin{aligned}
&\int_\triangle \varphi_{[i],k}\, \varphi_{[j],\ell} \\
&= \int_0^1 \left(\int_0^{x_j} \frac{x_i^{i-1}}{(i-1)!} \frac{(x_j - x_i)^{j-i-1}}{(j-i-1)!} \varphi_{i,k}(x_i)\, dx_i \right) \varphi_{j,\ell}(x_j) \frac{(1-x_j)^{d-j}}{(d-j)!}\, dx_j.
\end{aligned}$$

We deduce that:

$$\int_\triangle \varphi_{[i],k}\, \varphi_{[j],\ell} = \int_I r_k \varphi_{j,\ell}\, q_j,$$

with r_k a polynomial defined on I given by:

$$r_k(s) = (j-1)! \int_0^1 \frac{t^{i-1}}{(i-1)!} \frac{(1-t)^{j-i-1}}{(j-i-1)!} \varphi_{i,k}(st)\, dt.$$

Notice that r_k is a polynomial of degree at most k as $\varphi_{i,k}$ is a polynomial with degree k. Therefore if $k < \ell$, we have $\int_\triangle \varphi_{[i],k}\varphi_{[j],\ell} = 0$ since $\varphi_{j,\ell}$ is orthogonal (with respect to the measure q_j) to any polynomial of degree less than ℓ. Similar calculations show that if $k > \ell$, the integral is also 0.

Let us consider now the case $k = \ell$. We compute the coefficient ν_k of t^k in the polynomial r_k. We deduce from (6.2) that the leading coefficient $\omega_{i,k}$ of $\varphi_{i,k}$ is given by:

$$\omega_{i,k} = \rho_{i,k} \sqrt{(d-i)!(i-1)!} \cdot \omega'_{i,k} 2^k = \rho_{i,k} \sqrt{(d-i)!(i-1)!} \frac{(2k+d-1)!}{k!(k+d-1)!}.$$

Using this we obtain for ν_k :

$$\begin{aligned}
\nu_k &= (j-1)! \omega_{i,k} \int_0^1 \frac{t^{k+i-1}}{(i-1)!} \frac{(1-t)^{j-i-1}}{(j-i-1)!}\, dt \\
&= \omega_{i,k} \frac{(k+i-1)!(j-1)!}{(k+j-1)!(i-1)!},
\end{aligned}$$

and thus r_k has degree k. The orthonormality of $(\varphi_{j,k}, k \in \mathbb{N})$ ensures that $\int_I r_k \varphi_{j,k}\, q_j = \nu_k / \omega_{j,k}$. Therefore, we obtain:

$$\int_\triangle \varphi_{[i],k}\varphi_{[j],k} = \frac{\nu_k}{\omega_{j,k}} = \sqrt{\frac{(j-1)!(d-i)!}{(i-1)!(d-j)!}} \sqrt{\frac{(k+d-j)!(k+i-1)!}{(k+d-i)!(k+j-1)!}}.$$

Since $(j-1)!/(i-1)! \leq (k+j-1)!/(k+i-1)!$, and $(d-i)!/(d-j)! \leq (k+d-i)!/(k+d-j)!$, we can conclude that $0 \leq \int_\triangle \varphi_{[i],k}\varphi_{[j],k} \leq 1$. □

This shows that the family of functions $\varphi = (\varphi_{i,k}, 1 \leq i \leq d, k \in \mathbb{N})$ is not orthogonal with respect to the Lebesgue measure on $\triangle$. For $k \in \mathbb{N}^*$, let us consider the matrix $R_k \in \mathbb{R}^{d \times d}$ with elements:

$$R_k(i,j) = \int_\triangle \varphi_{[i],k}\varphi_{[j],k}. \tag{6.6}$$

If $Y = (Y_1, \ldots, Y_d)$ is uniformly distributed on $\triangle$, then R_k is the correlation matrix of the random variable $(\varphi_{1,k}(Y_1), \ldots, \varphi_{d,k}(Y_d))$. Therefore it is symmetric and positive semi-definite. Let $\zeta_{k,d} \leq \ldots \leq \zeta_{k,1}$ denote the eigenvalues of R_k. We aim to find a lower bound and an upper bound for these eigenvalues which is independent of k.

Lemma 6.4. *For $k \in \mathbb{N}^*$, the largest eigenvalue $\zeta_{k,1}$ and the smallest eigenvalue $\zeta_{k,d}$ of R_k are given by:*

$$\zeta_{k,1} = \frac{k+d}{k+1} \quad \text{and} \quad \zeta_{k,d} = \frac{k}{k+d-1},$$

and we have $1/d \leq \zeta_{k,d} \leq \zeta_{k,1} \leq (d+1)/2 \leq d$.

Proof. It is easy to check that the inverse R_k^{-1} of R_k exists and is symmetric tridiagonal with diagonal entries D_i, $1 \leq i \leq d$ and lower (and upper) diagonal elements Q_i, $1 \leq i \leq d-1$ given by:

$$D_i = \frac{(k+d-1)(k+1) + 2(i-1)(d-i)}{k(k+d)}$$

and

$$Q_i = -\frac{\sqrt{i(d-i)(k+i)(k+d-i)}}{k(k+d)}.$$

The matrix R_k^{-1} is positive definite, since all of its principal minors have a positive determinant. In particular, this ensures that the eigenvalues of R_k and R_k^{-1} are all positive.

It is easy to check that $\zeta_\circ = (k+1)/(k+d)$ is an eigenvalue of R_k^{-1} with corresponding eigenvector $w = (w_1, \ldots, w_d)$ given by, for $1 \leq i \leq d$:

$$w_i = \sqrt{\frac{(d-1)!}{(d-i)!} \frac{(k+d-i)!}{(k+d-1)!} \frac{(k+i-1)!}{k!} \frac{1}{(i-1)!}}.$$

This implies that w is an eigenvector of R_k with eigenvalue $\zeta_\circ^{-1}$. The matrix R_k has positive elements. We can apply the Perron-Frobenius theorem for positive matrices: the largest eigenvalue of R_k has multiplicity one and is the only eigenvalue with corresponding eigenvector x such that $x > 0$. Since $w > 0$, we deduce that $\zeta_\circ^{-1}$ is the largest eigenvalue of R_k.

Let $c_i(\zeta)$, $1 \leq i \leq d$ denote the i-th leading principal minor of the matrix $R_k^{-1} - \zeta I_d$, where I_d is the d-dimensional identity matrix. The eigenvalues of R_k^{-1} are exactly the roots of the characteristic polynomial $c_d(\zeta)$. Since R_k^{-1} is symmetric and tridiagonal, we have the following recurrence relation for $c_i(\zeta)$, $1 \leq i \leq d$:

$$c_i(\zeta) = (D_i - \zeta)c_{i-1}(\zeta) - Q_{i-1}^2 c_{i-2}(\zeta),$$

with initial values $c_0(\zeta) = 1$, $c_{-1}(\zeta) = 0$.

Let M_k be the symmetric tridiagonal matrix $d \times d$ with diagonal entries D_i, $1 \leq i \leq d$ and lower (and upper) diagonal elements $|Q_i|$, $1 \leq i \leq d-1$. Notice the characteristic polynomial of M_k is also $c_d(\zeta)$. So M_k and R_k^{-1} have the same eigenvalues.

It is easy to check that $\zeta^* = (k+d-1)/k$ is an eigenvalue of M_k with corresponding eigenvector $v = (v_1, \dots, v_d)$ given by, for $1 \leq i \leq d$:

$$v_i = \sqrt{\frac{(d-1)!}{(d-i)!} \frac{(k+d-1)!}{(k+d-i)!} \frac{k!}{(k+i-1)!} \frac{1}{(i-1)!}}.$$

(One can check that $v' = (v'_1, \dots, v'_d)$, with $v'_i = (-1)^{i-1} v_i$, is an eigenvector of R_k^{-1} with eigenvalue ζ^*.)

The matrix M_k has non-negative elements, with positive elements in the diagonal, sub- and superdiagonal. Therefore M_k is irreducible, and we can apply the Perron-Frobenius theorem for non-negative, irreducible matrices: the largest eigenvalue of M_k has multiplicity one and is the only eigenvalue with corresponding eigenvector x such that $x > 0$. Since $v > 0$, we deduce that ζ^* is the largest eigenvalue of M_k. It is also the largest eigenvalue of R_k^{-1}. Thus $1/\zeta^* = k/(k+d-1)$ is the lowest eigenvalue of R_k.

Since $\zeta_{k,d}$ is increasing in k, we have the uniform lower bound $1/d$. □

Remark 6.5. We conjecture that the eigenvalues $\zeta_{k,i}$ of R_k are given by, for $1 \leq i \leq d$:

$$\zeta_{k,i} = \frac{k(k+d)}{(k+i)(k+i-1)}.$$

6.4. Bounds between different norms

In this Section, we will give inequalities between different types of norms for functions defined on the simplex $\triangle$. These inequalities are used during the proof of Theorem 3.3. Let $m = (m_1, \dots, m_d) \in (\mathbb{N}^*)^d$. Recall the notation φ_m and $\theta \cdot \varphi_m$ with $\theta = (\theta_{i,k}; 1 \leq k \leq m_i, 1 \leq i \leq d) \in \mathbb{R}^{|m|}$ from Section 3.

For $1 \leq i \leq d$, we set $\theta_i = (\theta_{i,k}, 1 \leq k \leq m_i) \in \mathbb{R}^{m_i}$, $\varphi_{i,m} = (\varphi_{i,k}, 1 \leq k \leq m_i)$ and:

$$\theta_i \cdot \varphi_{i,m} = \sum_{k=1}^{m_i} \theta_{i,k} \varphi_{i,k} \quad \text{and} \quad \theta_i \cdot \varphi_{[i],m} = \sum_{k=1}^{m_i} \theta_{i,k} \varphi_{[i],k},$$

with $\varphi_{[i],m} = (\varphi_{[i],k}, 1 \leq k \leq m_i)$. In particular, we have $\varphi_m = \sum_{i=1}^d \varphi_{[i],m}$ and $\theta \cdot \varphi_m = \sum_{i=1}^d \theta_i \cdot \varphi_{[i],m}$. We first give lower and upper bounds on $\| \theta \cdot \varphi_m \|_{L^2}$.

Lemma 6.6. *For all $\theta \in \mathbb{R}^{|m|}$ we have:*

$$\frac{\| \theta \|}{\sqrt{d}} \leq \| \theta \cdot \varphi_m \|_{L^2} \leq \sqrt{d} \, \| \theta \| .$$

Proof. We have:

$$\|\,\theta\cdot\varphi_m\,\|_{L^2}^2 = \sum_{i=1}^{d}\sum_{k=1}^{m_i}\theta_{i,k}^2 + 2\sum_{i<j}\sum_{k=1}^{\min(m_i,m_j)}\theta_{i,k}\theta_{j,k}\int_{\triangle}\varphi_{[i],k}\varphi_{[j],k}, \tag{6.7}$$

where we used the normality of $\varphi_{[i],k}$ with respect to the Lebesgue measure on $\triangle$ and Lemma 6.3 for the cross products. We can rewrite this in a matrix form:

$$\|\,\theta\cdot\varphi_m\,\|_{L^2}^2 = \sum_{k=1}^{\max(m)}(\theta_k^*)^T R_k\theta_k^*,$$

where $R_k \in \mathbb{R}^{d\times d}$ is given by (6.6) and $\theta_k^* = (\theta_{1,k}^*,\dots,\theta_{d,k}^*) \in \mathbb{R}^d$ is defined, for $1 \le i \le d$, $1 \le k \le \max(m)$, as:

$$\theta_{i,k}^* = \theta_{i,k}\mathbf{1}_{\{k\le m_i\}}.$$

Since, according to Lemma 6.4, all the eigenvalues of R_k are uniformly larger than $1/d$ and smaller than d, this gives:

$$\frac{\|\,\theta\,\|^2}{d} = \frac{1}{d}\sum_{k=1}^{\max(m)}\|\,\theta_k^*\,\|^2 \le \|\,\theta\cdot\varphi_m\,\|_{L^2}^2 \le d\sum_{k=1}^{\max(m)}\|\,\theta_k^*\,\|^2 = d\,\|\,\theta\,\|^2\,.$$

This conclues the proof. □

We give an inequality between different norms for polynomials defined on I.

Lemma 6.7. *If h is a polynomial of degree less than or equal to n on I, then we have for all $1 \le i \le d$:*

$$\|\,h\,\|_\infty \le \sqrt{2(d-1)!}(n+d)^d\,\|\,h\,\|_{L^2(q_i)}$$

Proof. There exists $(\beta_k, 0 \le k \le n)$ such that $h = \sum_{k=0}^n \beta_k\varphi_{i,k}$. By the Cauchy-Schwarz inequality, we have:

$$|h| \le \left(\sum_{k=0}^{n}\beta_k^2\right)^{1/2}\left(\sum_{k=0}^{n}\varphi_{i,k}^2\right)^{1/2}. \tag{6.8}$$

We deduce from Definition 6.1 of $\varphi_{i,k}$ and (6.4) that:

$$\begin{aligned}\|\,\varphi_{i,k}\,\|_\infty &= \sqrt{\frac{(2k+d)(k+d-1)!}{k!}}\\ &\quad\cdot\max\left(\sqrt{\frac{(i-1)!(k+d-i)!}{(d-i)!(k+i-1)!}},\sqrt{\frac{(d-i)!(k+i-1)!}{(i-1)!(k+d-i)!}}\right).\end{aligned}$$

For all $1 \le i \le d$, we have the uniform upper bound:

$$\|\varphi_{i,k}\|_\infty \le \sqrt{(d-1)!}\sqrt{2k+d}\frac{(k+d-1)!}{k!}\cdot \tag{6.9}$$

This implies that for $t \in I$:

$$\begin{aligned}\sum_{k=0}^{n} \varphi_{i,k}^2(t) &\le \sum_{k=0}^{n} \|\varphi_{i,k}^2\|_\infty \\ &\le (d-1)!\sum_{k=0}^{n}(2k+d)\left(\frac{(k+d-1)!}{k!}\right)^2 \le 2(d-1)!(n+d)^{2d}.\end{aligned}$$

Bessel's inequality implies that $\sum_{k=0}^{n}\beta_k^2 \le \|h\|_{L^2(q_i)}^2$. We conclude the proof using (6.8). □

We recall the notation S_m of the linear space spanned by $(\varphi_{[i],k}; 1 \le k \le m_i, 1 \le i \le d)$, and the different norms introduced in Section 7.

Lemma 6.8. *Let $m \in (\mathbb{N}^*)^d$ and $\kappa_m = \sqrt{2d!}\sqrt{\sum_{i=1}^{d}(m_i+d)^{2d}}$. Then we have for every $g \in S_m$: $\|g\|_\infty \le \kappa_m \|g\|_{L^2}$.*

Proof. Let $g \in S_m$. We can write $g = \theta \cdot \varphi_m$ for a unique $\theta \in \mathbb{R}^{|m|}$. Let $g_i = \theta_i \cdot \varphi_{i,m}$ so that $g = \sum_{i=1}^{d} g_{[i]}$, where g_i is a polynomial defined on I of degree at most m_i for all $1 \le i \le d$. We have:

$$\begin{aligned}\|g\|_\infty &\le \sum_{i=1}^{d}\|g_i\|_\infty \\ &\le \sqrt{2(d-1)!}\sum_{i=1}^{d}(m_i+d)^d\|g_i\|_{L^2(q_i)} \\ &\le \frac{\kappa_m}{\sqrt{d}}\left(\sum_{i=1}^{d}\|g_i\|_{L^2(q_i)}^2\right)^{1/2} \\ &= \frac{\kappa_m}{\sqrt{d}}\|\theta\| \le \kappa_m\|\theta\cdot\varphi_m\|_{L^2} = \kappa_m\|g\|_{L^2},\end{aligned}$$

where we used Lemma 6.7 for the second inequality, Cauchy-Schwarz for the third inequality, and Lemma 6.6 for the fourth inequality. □

Remark 6.9. For d fixed, κ_m as a function of m verifies:

$$\kappa_m = O\left(\sqrt{\sum_{i=1}^{d} m_i^{2d}}\right) = O(|m|^d).$$

6.5. Bounds on approximations

Now we bound the L^2 and L^∞ norm of the approximation error of additive functions where each component belongs to a Sobolev space. Let $m = (m_1, \dots, m_d) \in (\mathbb{N}^*)^d$, $r = (r_1, \dots, r_d) \in (\mathbb{N}^*)^d$ such that $m_i + 1 \geq r_i$ for all $1 \leq i \leq d$. Let $\ell = \sum_{i=1}^d \ell_{[i]}$ with $\ell_i \in W^2_{r_i}(q_i)$ and $\int_I \ell_i q_i = 0$ for $1 \leq i \leq d$. Let ℓ_{i,m_i} be the orthogonal projection in $L^2(q_i)$ of ℓ_i on the span of $(\varphi_{i,k}, 0 \leq k \leq m_i)$ given by $\ell_{i,m_i} = \sum_{k=1}^{m_i} \left(\int_I \ell_i \varphi_{i,k} q_i \right) \varphi_{i,k}$. Then $\ell_m = \sum_{i=1}^d \ell_{[i],m_i}$ is the approximation of ℓ on S_m given by (7.10). We start by giving a bound on the $L^2(q_i)$ norm of the error when we approximate ℓ_i by ℓ_{i,m_i}.

Lemma 6.10. *For each $1 \leq i \leq d$, $m_i + 1 \geq r_i$ and $\ell_i \in W^2_{r_i}(q_i)$, we have:*

$$\| \ell_i - \ell_{i,m_i} \|^2_{L^2(q_i)} \leq \frac{2^{-2r_i}(m_i+1-r_i)!(m_i+d)!}{(m_i+1)!(m_i+d+r_i)!} \| \ell_i^{(r_i)} \|^2_{L^2(q_i)} . \tag{6.10}$$

Proof. Notice that (6.3) implies that the series $(\varphi^{(r_i)}_{i,k}, k \geq r_i)$ is orthogonal on I with respect to the weight function $v_i(t) = (1-t)^{d-i+r_i} t^{i-1+r_i}$, and the normalizing constants $\kappa_{i,k} \geq 0$ are given by:

$$\begin{aligned}
\kappa^2_{i,k} &= \int_0^1 \left(\varphi^{(r_i)}_{i,k}(t) \right)^2 v_i(t)\, dt \\
&= \rho^2_{i,k} (d-i)!(i-1)! \int_0^1 \left(\frac{d^{r_i}}{dt^{r_i}} P_k^{(d-i,i-1)}(2t-1) \right)^2 v_i(t)\, dt \\
&= \rho^2_{i,k} (d-i)!(i-1)! \frac{((k+d-1+r_i)!)^2}{2^{d+2r_i}((k+d-1)!)^2} \\
&\cdot \int_{-1}^1 \left(P_{k-r_i}^{(d-i+r_i,i-1+r_i)}(s) \right)^2 w_{d-i+r_i,i-1+r_i}(s)\, ds \\
&= (d-i)!(i-1)! \frac{k!(k+d-1+r_i)!}{(k-r_i)!(k+d-1)!},
\end{aligned} \tag{6.11}$$

where we used the definition of $\varphi_{i,k}$ for the second equality, (6.3) for the third equality and (6.1) for the fourth equality. Notice that $\kappa_{i,k}$ is non-decreasing as a function of k. Since $\ell_i - \ell_{i,m_i} = \sum_{k=m_i+1}^\infty \beta_{i,k}\varphi_{i,k}$, we have:

$$\begin{aligned}
\| \ell_i - \ell_{i,m_i} \|^2_{L^2(q_i)} &= \sum_{k=m_i+1}^\infty \beta^2_{i,k} \leq \frac{1}{\kappa^2_{i,m_i+1}} \sum_{k=m_i+1}^\infty \kappa^2_{i,k} \beta^2_{i,k} \\
&\leq \frac{1}{\kappa^2_{i,m_i+1}} \sum_{k=r_i}^\infty \kappa^2_{i,k} \beta^2_{i,k},
\end{aligned} \tag{6.12}$$

where the first inequality is due to the monotonicity of $\kappa_{i,k}$ as k increases. Thanks to (6.3) and the definition of $\kappa_{i,k}$, we get that $(\varphi_{i,k}^{(r_i)}/\kappa_{i,k}, k \geq r_i)$ is an orthonormal basis of $L^2(v_i)$. Therefore, we have

$$\sum_{k=r_i}^{\infty} \kappa_{i,k}^2 \beta_{i,k}^2 = \int_0^1 \left(\ell_i^{(r_i)}(t)\right)^2 v_i(t)\, dt \leq \frac{(d-i)!(i-1)!}{2^{2r_i}} \| \ell_i^{(r_i)} \|_{L^2(q_i)}^2, \tag{6.13}$$

since $\sup_{t\in I} q_i(t)/v_i(t) = (d-i)!(i-1)!/2^{2r_i}$. This and (6.12) implies (6.10). □

Lemma 6.10 yields a simple bound on the L^2 norm of the approximation error $\ell - \ell_m$.

Corollary 6.11. *For $m = (m_1, \ldots, m_d)$, $m_i + 1 \geq r_i$ and $\ell_i \in W_{r_i}^2(q_i)$ for all $1 \leq i \leq d$, we get:*

$$\| \ell - \ell_m \|_{L^2} = O\left(\sqrt{\sum_{i=1}^d m_i^{-2r_i}}\right).$$

Proof. We have:

$$\| \ell - \ell_m \|_{L^2} \leq \sum_{i=1}^d \| \ell_i - \ell_{i,m_i} \|_{L^2(q_i)} = O\left(\sum_{i=1}^d m_i^{-r_i}\right) = O\left(\sqrt{\sum_{i=1}^d m_i^{-2r_i}}\right),$$

where we used (6.10) for the first equality. □

Lastly, we bound the L^∞ norm of the approximation error.

Lemma 6.12. *For each $1 \leq i \leq d$, $m_i + 1 \geq r_i > d$ and $\ell_i \in W_{r_i}^2(q_i)$, we have:*

$$\| \ell_i - \ell_{i,m_i} \|_\infty \leq \frac{2^{-r_i}\sqrt{2(d-1)!}\, \mathrm{e}^{r_i}}{\sqrt{2r_i - 2d - 1}} \frac{1}{(m_i + r_i)^{r_i - d - \frac{1}{2}}} \| \ell_i^{(r_i)} \|_{L^2(q_i)}. \tag{6.14}$$

Proof. We first give a lower bound for the constants $\kappa_{i,k}$, $1 \leq i \leq d$, $d < r_i \leq m_i + 1 \leq k$ given by (6.11). We have for $k \geq m_i + 1$:

$$\begin{aligned}\frac{\kappa_{i,k}}{(d-i)!(i-1)!(k+r_i)^{2r_i}} &= \frac{k!(k+d-1+r_i)!}{(k-r_i)!(k+d-1)!(k+r_i)^{2r_i}} \\ &\geq \frac{(k+r_i)!}{(k-r_i)!(k+r_i)^{2r_i}} \geq \frac{(2r_i)!}{(2r_i)^{2r_i}}.\end{aligned}$$

Since $n! \geq n^n \mathrm{e}^{-n}$ for $n \in \mathbb{N}^*$, we deduce that

$$\kappa_{i,k}^2 \geq (d-i)!(i-1)!(k+r_i)^{2r_i} \mathrm{e}^{-2r_i}. \tag{6.15}$$

Since $\ell_i - \ell_{i,m_i} = \sum_{k=m_i+1}^{\infty} \beta_{i,k}\varphi_{i,k}$ we have:

$$\begin{aligned}
\| \ell_i - \ell_{i,m_i} \|_\infty &= \left\| \sum_{k=m_i+1}^{\infty} \beta_{i,k}\varphi_{i,k} \right\|_\infty \\
&\leq \sum_{k=m_i+1}^{\infty} |\beta_{i,k}| \, \|\varphi_{i,k}\|_\infty \\
&\leq \sqrt{\sum_{k=m_i+1}^{\infty} \frac{\|\varphi_{i,k}\|_\infty^2}{\kappa_{i,k}^2}} \sqrt{\sum_{k=m_i+1}^{\infty} \kappa_{i,k}^2 \beta_{i,k}^2} \\
&\leq \sqrt{\sum_{k=m_i+1}^{\infty} \frac{2(d-1)!(k+d)^{2d}}{\kappa_{i,k}^2}} \sqrt{\frac{(d-i)!(i-1)!}{2^{2r_i}}} \, \| \ell_i^{(r_i)} \|_{L^2(q_i)} \\
&\leq \sqrt{\sum_{k=m_i+1}^{\infty} \frac{2(d-1)!}{(d-i)!(i-1)!} \frac{\mathrm{e}^{2r_i}}{(k+r_i)^{2r_i-2d}}} \sqrt{\frac{(d-i)!(i-1)!}{2^{2r_i}}} \, \| \ell_i^{(r_i)} \|_{L^2(q_i)} \\
&\leq \frac{2^{-r_i}\sqrt{2(d-1)!}\,\mathrm{e}^{r_i}}{\sqrt{2r_i-2d-1}\sqrt{(m_i+r_i)^{2r_i-2d-1}}} \, \| \ell_i^{(r_i)} \|_{L^2(q_i)},
\end{aligned}$$

where we used Cauchy-Schwarz for the second inequality, (6.9) and (6.13) for the third inequality, (6.15) for the fourth inequality, and $\sum_{k=m_i+1}^{\infty}(k+r_i)^{-2r_i+2d} \leq (2r_i - 2d - 1)^{-1}(m_i + r_i)^{-2r_i+2d+1}$ for the fifth inequality. □

Corollary 6.13. *There exists a constant $\mathcal{C} > 0$ such that for all $\ell_i \in W_{r_i}^2(q_i)$ and $m_i + 1 \geq r_i > d$ for all $1 \leq i \leq d$, we have:*

$$\| \ell - \ell_m \|_\infty \leq \mathcal{C} \sum_{i=1}^{d} \| \ell_i^{(r_i)} \|_{L^2(q_i)} .$$

Proof. Notice that for $m_i + 1 \geq r_i > d$, we have:

$$\frac{2^{-r_i}\sqrt{2(d-1)!}\,\mathrm{e}^{r_i}}{\sqrt{2r_i-2d-1}} \frac{1}{(m_i+r_i)^{r_i-d-\frac{1}{2}}} \leq \frac{2^{-r_i}\sqrt{2(d-1)!}\,\mathrm{e}^{r_i}}{\sqrt{2r_i-2d-1}} \frac{1}{(2r_i-1)^{r_i-d-\frac{1}{2}}},$$

and that the right hand side is bounded by a constant $\mathcal{C} > 0$ for all $r_i \in \mathbb{N}^*$. Therefore:

$$\| \ell - \ell_m \|_\infty \leq \sum_{i=1}^{d} \| \ell_i - \ell_{i,m_i} \|_\infty \leq \mathcal{C} \sum_{i=1}^{d} \| \ell_i^{(r_i)} \|_{L^2(q_i)} . \qquad □$$

7. Preliminary elements for the proof of Theorem 3.3

We adapt the results from [6] to our setting, by following their lines of proof. Even though some results appear to be very similar, for the reader's convenience we repeat the statements and their proofs and carefully use the parameters from our context (dimension d, particular basis of functions, etc.) Let us recall Lemmas 1 and 2 of [6].

Lemma 7.1 (Lemma 1 of [6])**.** *Let $g, h \in \mathcal{P}(\triangle)$. If $\| \log(g/h) \|_\infty < +\infty$, then we have:*

$$D\left(g \| h\right) \geq \frac{1}{2}\, \mathrm{e}^{-\, \| \log(g/h) \|_\infty} \int_\triangle g \log^2\left(g/h\right), \tag{7.1}$$

and for any $\kappa \in \mathbb{R}$:

$$D\left(g \| h\right) \leq \frac{1}{2}\, \mathrm{e}^{\| \log(g/h) - \kappa \|_\infty} \int_\triangle g \left(\log\left(g/h\right) - \kappa\right)^2, \tag{7.2}$$

$$\int_\triangle \frac{(g-h)^2}{g} \leq \mathrm{e}^{2\left(\| \log(g/h) - \kappa \|_\infty - \kappa\right)} \int_\triangle g \left(\log\left(g/h\right) - \kappa\right)^2. \tag{7.3}$$

Lemma 7.1 readily implies the following Corollary.

Corollary 7.2. *Let $g, h \in \mathcal{P}(\triangle)$. If $\| \log(g/h) \|_\infty < +\infty$, then we have, for any constant $\kappa \in \mathbb{R}$:*

$$D\left(g \| h\right) \leq \frac{1}{2}\, \mathrm{e}^{\| \log(g/h) - \kappa \|_\infty} \| g \|_\infty \int_\triangle \left(\log\left(g/h\right) - \kappa\right)^2, \tag{7.4}$$

and:

$$\| g - h \|_{L^2} \leq \| g \|_\infty \, \mathrm{e}^{\left(\| \log(g/h) - \kappa \|_\infty - \kappa\right)} \| \log\left(g/h\right) - \kappa \|_{L^2}. \tag{7.5}$$

Recall Definition 2.1 for densities f^0 with a product form on $\triangle$. We give a few bounds between the L^∞ norms of $\log(f^0)$, ℓ^0 and the constant a_0.

Lemma 7.3. *Let $f^0 \in \mathcal{P}(\triangle)$ given by Definition 2.1. Then we have:*

$$|\mathrm{a}_0| \leq \| \ell^0 \|_\infty + |\log(d!)|, \quad \| \log(f^0) \|_\infty \leq 2 \| \ell^0 \|_\infty + |\log(d!)|, \tag{7.6}$$

$$|\mathrm{a}_0| \leq \| \log(f^0) \|_\infty, \qquad \| \ell^0 \|_\infty \leq 2 \| \log(f^0) \|_\infty. \tag{7.7}$$

Proof. The first part of (7.6) can be obtained by bounding ℓ^0 with $\| \ell^0 \|_\infty$ in the definition of a_0. The second part is a direct consequence of this. The first part of (7.7) can be deduced from the fact that $\int_\triangle \ell^0 = 0$. The second part is again a direct consequence of the first part. □

Let $m \in (\mathbb{N}^*)^d$. Recall the application A_m defined in (3.2) and set $\Omega_m = A_m(\mathbb{R}^{|m|})$. For $\alpha \in \mathbb{R}^{|m|}$, we define the function $\mathscr{F}_\alpha$ on $\mathbb{R}^{|m|}$ by:

$$\mathscr{F}_\alpha(\theta) = \theta \cdot \alpha - \psi(\theta). \tag{7.8}$$

Recall also the additive exponential series model f_θ given by (3.1).

Lemma 7.4 (Lemma 3 of [6])**.** *Let $m \in (\mathbb{N}^*)^d$. The application A_m is one-to-one from $\mathbb{R}^{|m|}$ onto Ω_m, with inverse say Θ_m. Let $f \in \mathcal{P}(\triangle)$ such that $\alpha = \int_\triangle \varphi_m f$ belongs to Ω_m. Then for all $\theta \in \mathbb{R}^{|m|}$, we have with $\theta^* = \Theta_m(\alpha)$:*

$$D\left(f \| f_\theta\right) = D\left(f \| f_{\theta^*}\right) + D\left(f_{\theta^*} \| f_\theta\right). \tag{7.9}$$

Furthermore, θ^* achieves $\max_{\theta \in \mathbb{R}^{|m|}} \mathscr{F}_\alpha(\theta)$ as well as $\min_{\theta \in \mathbb{R}^{|m|}} D\left(f \| f_\theta\right)$.

Definition 7.5. *Let $m \in (\mathbb{N}^*)^d$. For $f \in \mathcal{P}(\triangle)$ such that $\alpha = \int_\triangle \varphi_m f \in \Omega_m$, the probability density f_{θ^*}, with $\theta^* = \Theta_m(\alpha)$ (that is $\int_\triangle \varphi_m f = \int_\triangle \varphi_m f_{\theta^*}$), is called the information projection of f.*

The information projection of a density f is the closest density in the exponential family (3.1) with respect to the Kullback-Leibler distance to f.

We consider the linear space of real valued functions defined on $\triangle$ and generated by φ_m:

$$S_m = \{\theta \cdot \varphi_m ; \theta \in \mathbb{R}^{|m|}\}. \tag{7.10}$$

Let $\kappa_m = \sqrt{2d!}\sqrt{\sum_{i=1}^d (m_i + d)^{2d}}$. The following Lemma summarizes Lemmas 6.6 and 6.8.

Lemma 7.6. *Let $m \in (\mathbb{N}^*)^d$. We have for all $g \in S_m$:*

$$\| g \|_\infty \le \kappa_m \| g \|_{L^2}, \tag{7.11}$$

For all $\theta \in \mathbb{R}^{|m|}$, we have:

$$\frac{\| \theta \|}{\sqrt{d}} \le \| \theta \cdot \varphi_m \|_{L^2} \le \sqrt{d} \, \| \theta \| . \tag{7.12}$$

Now we give upper and lower bounds for the Kullback-Leibler distance between two members of the exponential family f_θ and $f_{\theta'}$ in terms of the Euclidean distance $\| \theta - \theta' \|$. Note that $\| \log(f_\theta) \|_\infty = \sup_{x \in \triangle} |\log(f_\theta(x))|$ is finite, for all $\theta \in \mathbb{R}^{|m|}$.

Lemma 7.7. *Let $m \in (\mathbb{N}^*)^d$. For $\theta, \theta' \in \mathbb{R}^{|m|}$, we have:*

$$\| \log(f_\theta / f_{\theta'}) \|_\infty \le 2\sqrt{d}\, \kappa_m \| \theta - \theta' \|, \tag{7.13}$$

$$D\left(f_\theta \| f_{\theta'}\right) \le \frac{d}{2} \, \mathrm{e}^{\| \log(f_\theta) \|_\infty + \sqrt{d}\, \kappa_m \, \|\theta - \theta'\|} \, \| \theta - \theta' \|^2, \tag{7.14}$$

$$D\left(f_\theta \| f_{\theta'}\right) \ge \frac{1}{2d} \, \mathrm{e}^{- \| \log(f_\theta) \|_\infty - 2\sqrt{d}\, \kappa_m \, \|\theta - \theta'\|} \, \| \theta - \theta' \|^2 . \tag{7.15}$$

Proof. Since $\psi(\theta') - \psi(\theta) = \log\left(\int_\triangle \mathrm{e}^{(\theta' - \theta)\cdot \varphi_m} f_\theta\right)$, we get $|\psi(\theta') - \psi(\theta)| \le \| (\theta' - \theta) \cdot \varphi_m \|_\infty$. This implies that:

$$\begin{aligned} \| \log(f_\theta / f_{\theta'}) \|_\infty &\le 2 \, \| (\theta - \theta') \cdot \varphi_m \|_\infty \\ &\le 2\kappa_m \, \| (\theta - \theta') \cdot \varphi_m \|_{L^2} \\ &\le 2\sqrt{d}\, \kappa_m \, \| \theta - \theta' \|, \end{aligned}$$

where we used (3.1) for the first inequality, (7.11) for the second and (7.12) for the third. The proof of (7.14) and (7.15) follows the proof of Lemma 4 in [6] and is not reproduced. □

Now we will show that the application Θ_m is locally Lipschitz.

Lemma 7.8. *Let $m \in (\mathbb{N}^*)^d$ and $\theta \in \mathbb{R}^{|m|}$. If $\alpha \in \mathbb{R}^{|m|}$ satisfies:*

$$\| A_m(\theta) - \alpha \| \leq \frac{\mathrm{e}^{-(1+\|\log(f_\theta)\|_\infty)}}{6d^{\frac{3}{2}}\kappa_m}, \tag{7.16}$$

Then α belongs to Ω_m and $\theta^* = \Theta_m(\alpha)$ exists. Let τ be such that:

$$6d^{\frac{3}{2}}\,\mathrm{e}^{1+\|\log(f_\theta)\|_\infty}\,\kappa_m\,\| A_m(\theta) - \alpha \| \leq \tau \leq 1.$$

Then θ^* satisfies:

$$\| \theta - \theta^* \| \leq 3d\,\mathrm{e}^{\tau+\|\log(f_\theta)\|_\infty}\,\| A_m(\theta) - \alpha \|, \tag{7.17}$$

$$\| \log(f_\theta/f_{\theta^*}) \|_\infty \leq 6d^{\frac{3}{2}}\,\mathrm{e}^{\tau+\|\log(f_\theta)\|_\infty}\,\kappa_m\,\| A_m(\theta) - \alpha \| \leq \tau, \tag{7.18}$$

$$D\left(f_\theta \| f_{\theta^*}\right) \leq 3d\,\mathrm{e}^{\tau+\|\log(f_\theta)\|_\infty}\,\| A_m(\theta) - \alpha \|^2 . \tag{7.19}$$

Proof. Suppose that $\alpha \neq A_m(\theta)$ (otherwise the results are trivial). Recall $\mathscr{F}_\alpha$ defined in (7.8). We have, for all $\theta' \in \mathbb{R}^{|m|}$:

$$\begin{aligned} \mathscr{F} := \mathscr{F}_\alpha(\theta) - \mathscr{F}_\alpha(\theta') &= (\theta - \theta') \cdot \alpha + \psi(\theta') - \psi(\theta) \\ &= D\left(f_\theta \| f_{\theta'}\right) - (\theta - \theta') \cdot (A_m(\theta) - \alpha). \end{aligned} \tag{7.20}$$

Using (7.15) and the Cauchy-Schwarz inequality, we obtain the strict inequality:

$$\mathscr{F} > \frac{1}{3d}\,\mathrm{e}^{-\,\|\log(f_\theta)\|_\infty\,-2\sqrt{d}\,\kappa_m\,\|\theta-\theta'\|}\,\| \theta - \theta' \|^2 - \| \theta - \theta' \|\,\| A_m(\theta) - \alpha \| .$$

We consider the ball centered at θ: $B_r = \{\theta' \in \mathbb{R}^{|m|}, \| \theta - \theta' \| \leq r\}$ with radius $r = 3d\,\mathrm{e}^{\tau+\|\log(f_\theta)\|_\infty}\,\| A_m(\theta) - \alpha \|$. For all $\theta' \in \partial B_r$, we have:

$$\mathscr{F} > \left(\mathrm{e}^{\tau-6d^{\frac{3}{2}}\kappa_m\,\|A_m(\theta)-\alpha\|\,\mathrm{e}^{\tau+\|\log(f_\theta)\|_\infty}} - 1\right) 3d\,\mathrm{e}^{\tau+\|\log(f_\theta)\|_\infty}\,\| A_m(\theta) - \alpha \|^2 .$$

The right hand side is non-negative as $6d^{\frac{3}{2}}\,\mathrm{e}^{1+\|\log(f_\theta)\|_\infty}\,\kappa_m\,\| A_m(\theta) - \alpha \| \leq \tau \leq 1$, see the condition on τ. Thus, the value of $\mathscr{F}_\alpha$ at θ, an interior point of B_r, is larger than the values of $\mathscr{F}_\alpha$ on ∂B_r. Therefore $\mathscr{F}_\alpha$ is maximal at a point, say θ^*, in the interior of B_r. Since the gradient of $\mathscr{F}_\alpha$ at θ^* equals 0, we have $\nabla \mathscr{F}_\alpha(\theta^*) = \alpha - \int_\triangle \varphi_m f_{\theta^*} = 0$, which means that $\alpha \in \Omega_m$ and $\theta^* = \Theta_m(\alpha)$. Since θ^* is inside B_r, we get (7.17). The upper bound (7.18) is due to (7.13) of Lemma 7.7. To prove (7.19), we use (7.20) and the fact that $\mathscr{F}_\alpha(\theta) - \mathscr{F}_\alpha(\theta^*) \leq 0$, which gives:

$$\begin{aligned} D\left(f_\theta \| f_{\theta^*}\right) &\leq (\theta - \theta^*) \cdot (A_m(\theta) - \alpha) \leq \| \theta - \theta^* \|\,\| A_m(\theta) - \alpha \| \\ &\leq 3d\,\mathrm{e}^{\tau+\|\log(f_\theta)\|_\infty}\,\| A_m(\theta) - \alpha \|^2 . \end{aligned}$$

□

8. Proof of Theorem 3.3

In this Section, we first show that the information projection f_{θ^*} of f^0 onto $\{f_\theta, \theta \in \mathbb{R}^{|m|}\}$ exists for all $m \in (\mathbb{N}^*)^d$. Moreover, the maximum likelihood estimator $\hat{\theta}_{m,n}$, defined in (3.4) based on an i.i.d sample $\mathbb{X}^n$, verifies almost surely $\hat{\theta}_{m,n} = \Theta_m(\hat{\mu}_{m,n})$ for $n \geq 2$ with $\hat{\mu}_{m,n}$ the empirical mean given by (3.3). Recall $\Omega_m = A_m(\mathbb{R}^{|m|})$ with A_m defined by (3.2).

Lemma 8.1. *The mean $\alpha = \int_\triangle \varphi_m f^0$ verifies $\alpha \in \Omega_m$ and the empirical mean $\hat{\mu}_{m,n}$ verifies $\hat{\mu}_{m,n} \in \Omega_m$ almost surely when $n \geq 2$.*

Remark 8.2. By Lemma 7.4, this also means that $\hat{\theta}_{m,n} = \text{argmax}_{\,\theta\in\mathbb{R}^{|m|}} \mathcal{F}_{\hat{\mu}_{m,n}}(\theta)$, and since $\mathcal{F}_{\hat{\mu}_{m,n}}(\theta) = (1/n)\sum_{j=1}^n \log(f_\theta(X^j))$, the estimator $\hat{f}_{m,n} = f_{\hat{\theta}_{m,n}}$ is the maximum likelihood estimator of f^0 in the model $\{f_\theta, \theta \in \mathbb{R}^m\}$ based on $\mathbb{X}^n$.

Proof. Notice that $\psi(\theta) = \log(\mathbb{E}[\exp(\theta \cdot \varphi_m(U))]) - \log(d!)$, where U is a random vector uniformly distributed on $\triangle$. The Hessian matrix $\nabla^2\psi(\theta)$ is equal to the covariance matrix of $\varphi_m(X)$, where X has density f_θ. Therefore $\nabla^2\psi(\theta)$ is positive semi-definite, and we show that it is positive definite too. Indeed, for $\lambda \in \mathbb{R}^{|m|}$, $\lambda^T\nabla^2\psi(\theta)\lambda = 0$ is equivalent to $\mathbb{E}[(\lambda \cdot \varphi_m(X))^2] = 0$, which implies that $\lambda \cdot \varphi_m(X) = 0$ a.e. on $\triangle$. Since $(\varphi_{[i],k}, 1 \leq i \leq d, 1 \leq k \leq m_i)$ are linearly independent, this means $\lambda = 0$. Thus $\nabla^2\psi(\theta)$ is positive definite, providing that $\theta \mapsto \psi(\theta)$ is a strictly convex function.

Let $\psi^* : \mathbb{R}^{|m|} \to \mathbb{R} \cup \{+\infty\}$ denote the Legendre-Fenchel transformation of the function $\theta \mapsto \psi(\theta)$, i.e. for $\alpha \in \mathbb{R}^{|m|}$:

$$\psi^*(\alpha) = \sup_{\theta\in\mathbb{R}^{|m|}} \alpha \cdot \theta - \psi(\theta) = \sup_{\theta\in\mathbb{R}^{|m|}} \mathcal{F}_\alpha(\theta).$$

Suppose that $\alpha \in \Omega_m$. Then according to Lemma 7.4, $\psi^*(\alpha) = \mathcal{F}_\alpha(\theta^*)$ with $\theta^* = \Theta_m(\alpha)$, thus $\psi^*(\alpha)$ is finite. Therefore $\Omega_m \subseteq \text{Dom }(\psi^*)$, where $\text{Dom }(\psi^*) = \{\alpha \in \mathbb{R}^{|m|} : \psi^*(\alpha) < +\infty\}$. By Lemma 7.8, we have that Ω_m is an open subset of $\mathbb{R}^{|m|}$. So, we get $\Omega_m \subseteq \text{int }(\text{Dom }(\psi^*))$, where $\text{int }(A)$ is the interior of a set $A \subseteq \mathbb{R}^{|m|}$. Inversely, let $\alpha \in \text{int }(\text{Dom }(\psi^*))$. This insures that $\theta^* = \text{argmax}_{\,\theta\in\mathbb{R}^{|m|}} \mathcal{F}_\alpha(\theta)$ exists uniquely and that $\nabla\mathcal{F}_\alpha(\theta^*) = 0$. Therefore, we get:

$$0 = \nabla\mathcal{F}_\alpha(\theta^*) = \alpha - \int_\triangle \varphi_m f_{\theta^*} = \alpha - A_m(\theta^*),$$

giving $\alpha \in \Omega_m$. Thus we obtain $\Omega_m = \text{Dom }(\psi^*)$. Set

$$\Upsilon = \text{int }(\text{cv }(\text{supp }(\varphi_m(U)))),$$

where $\text{cv }(A)$ is the convex hull of a set $A \subseteq \mathbb{R}^{|m|}$. Thanks to Lemma 4.1. of [1], we have $\Upsilon = \text{int }(\text{Dom }(\psi^*)) = \Omega_m$, and thus Υ is non-empty. The proof is complete as soon as we prove that $\alpha := \int_\triangle \varphi_m f^0 \in \Upsilon$ and $\hat{\mu}_{m,n} \in \Upsilon$ almost surely when $n \geq 2$. Notice that the probability measures of $\varphi_m(X^0)$, where X^0 has density f^0, and $\varphi_m(U)$ are equivalent, so that $\Upsilon = \text{int }(\text{cv }(\text{supp }(\varphi_m(X^0))))$.

This directly implies that $\alpha = \mathbb{E}[\varphi_m(X^0)] \in \Upsilon$. To show that $\hat{\mu}_{m,n} \in \Upsilon$, since the probability measures of $\varphi_m(X^0)$ and $\varphi_m(U)$ are equivalent, it is sufficient to prove that, when $n \geq 2$, $(1/n)\sum_{j=1}^n \varphi_m(U^j) \in \Upsilon$, with $(U^1, \ldots, U^n)$ independent random vectors distributed as U. Let cl (A) denote the closure of a set $A \subset \mathbb{R}^{|m|}$. Let $1 \leq i \leq d$. Recall $\varphi_{i,m} = (\varphi_{i,k}, 1 \leq k \leq m_i)$. The linear independence of $(\varphi_{i,k}, 1 \leq k \leq m_i)$ implies that all hyper-plane H_i tangent to $\Upsilon_i :=$ cv $(\{\varphi_{i,m}(x);\, x \in [0,1]\})$ is such that $\{x \in [0,1];\, \varphi_{i,m}(x) \in H\}$ has zero Lebesgue measure. This readily implies that the probability for $\varphi_m(U_\ell)$ and $\varphi_m(U_j)$, with $\ell \neq j$, to belongs to the same tangent hyper-plane H of Υ is zero. We get that a.s. $(1/n)\sum_{j=1}^n \varphi_m(U^j) \in \Upsilon$ for all $n \geq 2$. Thus, the proof is complete. □

We divide the proof of Theorem 3.3 into two parts: first we bound the error due to the bias of the proposed exponential model, then we bound the error due to the variance of the sample estimation. We formulate the results in two general Propositions, which can be later specified to get Theorem 3.3.

8.1. Bias of the estimator

The bias error comes from the information projection of the true underlying density f^0 onto the family of the exponential series model $\{f_\theta, \theta \in \mathbb{R}^{|m|}\}$. We recall the linear space S_m spanned by $(\varphi_{[i],k}, 1 \leq k \leq m_i, 1 \leq i \leq d)$ where $\varphi_{i,k}$ is a polynomial of degree k, and the form of the probability density f^0 given in (2.3). For $1 \leq i \leq d$, let $\ell^0_{i,m}$ be the orthogonal projection in $L^2(q_i)$ of ℓ^0_i on the vector space spanned by $(\varphi_{i,k}, 0 \leq k \leq m_i)$ or equivalently on the vector space spanned by $(\varphi_{i,k}, 1 \leq k \leq m_i)$, as we assumed that $\int_I \ell^0_i q_i = 0$. We set $\ell^0_m = \sum_{i=1}^d \ell^0_{[i],m}$ the approximation of ℓ^0 on S_m. In particular we have $\ell^0_m = \theta^0 \cdot \varphi_m$ for some $\theta^0 \in \mathbb{R}^{|m|}$. Let:

$$\Delta_m = \|\, \ell^0 - \ell^0_m \,\|_{L^2} \quad \text{and} \quad \gamma_m = \|\, \ell^0 - \ell^0_m \,\|_\infty$$

denote the L^2 and L^∞ errors of the approximation of ℓ^0 by ℓ^0_m on the simplex $\triangle$.

Proposition 8.3. *Let $f^0 \in \mathcal{P}(\triangle)$ have a product form given by Definition 2.1. Let $m \in (\mathbb{N}^*)^d$. The information projection f_{θ^*} of f^0 exists (with $\theta^* \in \mathbb{R}^{|m|}$ and $\int_\triangle \varphi_m f_{\theta^*} = \int_\triangle \varphi_m f^0$) and verifies, with $\mathfrak{A}_1 = \frac{1}{2}\,\mathrm{e}^{\gamma_m + \|\log(f^0)\|_\infty}$:*

$$D\left(f^0 \| f_{\theta^*}\right) \leq \mathfrak{A}_1 \Delta_m^2. \tag{8.1}$$

Proof. The existence of θ^* is due to Lemma 8.1. Thanks to Lemma 7.4 and (7.4) with $\kappa = \psi(\theta^0) - \mathrm{a}_0$, we can deduce that:

$$\begin{aligned} D\left(f^0 \| f_{\theta^*}\right) &\leq D\left(f^0 \| f_{\theta^0_m}\right) \leq \frac{1}{2}\,\mathrm{e}^{\|\ell^0 - \ell^0_m\|_\infty}\, \|\, f^0 \,\|_\infty \, \|\, \ell^0 - \ell^0_m \,\|^2_{L^2} \\ &\leq \frac{1}{2}\,\mathrm{e}^{\gamma_m + \|\log(f^0)\|_\infty}\, \Delta_m^2. \end{aligned}$$

□

Set:

$$\varepsilon_m = 6d^{\frac{5}{2}}\kappa_m \Delta_m \,\mathrm{e}^{(4\gamma_m + 2\,\|\log(f^0)\|_\infty + 1)}\,. \tag{8.2}$$

We need the following lemma to control $\|\log(f^0/f_{\theta^*})\|_\infty$.

Lemma 8.4. *If $\varepsilon_m \le 1$, we also have:*

$$\|\log(f^0/f_{\theta^*})\|_\infty \le 2\gamma_m + \varepsilon_m \le 2\gamma_m + 1. \tag{8.3}$$

Proof. To show (8.3), let $f^0_m = f_{\theta^0}$ denote the density function in the exponential family corresponding to θ^0, and $\alpha^0 = \int_\triangle \varphi_m f^0$. For each $1 \le i \le d$, the functions $\varphi_{i,m} = (\varphi_{[i],k},\ 1 \le k \le m_i)$ form an orthonormal set with respect to the Lebesgue measure on $\triangle$. We set $\alpha^0_{i,m} = \int_\triangle \varphi_{i,m} f^0$ and $A_{i,m}(\theta^0) = \int_\triangle \varphi_{i,m} f_{\theta^0}$. By Bessel's inequality, we have for $1 \le i \le d$:

$$\|\alpha^0_{i,m} - A_{i,m}(\theta^0)\| \le \|f^0 - f^0_m\|_{L^2}\,.$$

Summing up these inequalities for $1 \le i \le d$, we get:

$$\begin{aligned}
\|\alpha^0 - A_m(\theta^0)\| &\le \sum_{i=1}^{d} \|\alpha^0_{i,m} - A_{i,m}(\theta^0)\| \\
&\le d\,\|f^0 - f^0_m\|_{L^2} \\
&\le d\,\|f^0\|_\infty\, \mathrm{e}^{\left(\|\ell^0 - \ell^0_m\|_\infty - (\psi(\theta^0) - \mathrm{a}_0)\right)}\,\|\ell^0 - \ell^0_m\|_{L^2} \\
&\le d\,\mathrm{e}^{\|\log(f^0)\|_\infty + 2\gamma_m}\,\Delta_m,
\end{aligned}$$

where we used (7.5) with $\kappa = \psi(\theta^0) - \mathrm{a}_0$ for the third inequality and the inequality $\left|\psi(\theta^0) - \mathrm{a}_0\right| \le \gamma_m$ (due to $\psi(\theta^0) - \mathrm{a}_0 = \log(\int \exp(\ell^0_m - \ell^0) f^0)$) for the fourth inequality. The latter argument also ensures that $\|\log(f^0/f^0_m)\|_\infty \le 2\gamma_m$. In order to apply Lemma 7.8 with $\theta = \theta^0$, $\alpha = \alpha^0$, we check condition (7.16), which is implied by:

$$d\,\mathrm{e}^{\|\log(f^0)\|_\infty + 2\gamma_m}\,\Delta_m \le \frac{\mathrm{e}^{-(1 + \|\log(f^0_m)\|_\infty)}}{6d^{\frac{3}{2}}\kappa_m}.$$

Since $\|\log(f^0_m)\|_\infty \le \|\log(f^0)\|_\infty + \|\log(f^0/f^0_m)\|_\infty \le \|\log(f^0)\|_\infty + 2\gamma_m$, this condition is ensured whenever $\varepsilon_m \le 1$. In this case we deduce, thanks to (7.18) with $\tau = 1$, that $\|\log(f^0_m/f_{\theta^*})\|_\infty \le \varepsilon_m$. By the triangle inequality, we obtain $\|\log(f^0/f_{\theta^*})\|_\infty \le 2\gamma_m + \varepsilon_m$. This completes the proof. □

8.2. Variance of the estimator

We control the variance error due to the parameter estimation by the size of the sample. We keep the notations used in Section 8.1. In particular ε_m is defined by (8.2) and $\kappa_m = \sqrt{2d!}\sqrt{\sum_{i=1}^{d}(m_i + d)^{2d}}$. The results are summarized in the following proposition.

Proposition 8.5. *Let $f^0 \in \mathcal{P}(\triangle)$ have a product form given by Definition 2.1. Let $m \in (\mathbb{N}^*)^d$ and suppose that $\varepsilon_m \leq 1$. Set:*

$$\delta_{m,n} = 6d^{\frac{3}{2}}\kappa_m \sqrt{\frac{|m|}{n}}\, \mathrm{e}^{2\gamma_m + \|\log(f^0)\|_\infty + 2}.$$

If $\delta_{m,n} \leq 1$, then for every $0 < K \leq \delta_{m,n}^{-2}$, we have:

$$\mathbb{P}\left(D\left(f_{\theta^*} \| \hat{f}_{m,n} \right) \geq \mathfrak{A}_2 \frac{|m|}{n} K \right) \leq \exp(\| \log(f^0) \|_\infty)/K. \tag{8.4}$$

where $\mathfrak{A}_2 = 3d\, \mathrm{e}^{2\gamma_m + \varepsilon_m + \|\log(f^0)\|_\infty + \tau}$, and $\tau = \delta_{m,n}\sqrt{K} \leq 1$.

Proof. Let θ^* be defined in Proposition 8.3. Let $X = (X_1, \ldots, X_d)$ denote a random variable with density f^0. Let θ in Lemma 7.8 be equal to θ^*, which gives $A_m(\theta^*) = \alpha^0 = \mathbb{E}[\varphi_m(X)]$, and for α, we take the empirical mean $\hat{\mu}_{m,n}$. With this setting, we have:

$$\| \alpha - \alpha^0 \|^2 = \sum_{i=1}^{d} \sum_{k=1}^{m_i} \left(\hat{\mu}_{m,n,i,k} - \mathbb{E}[\varphi_{i,k}(X_i)] \right)^2.$$

By Chebyshev's inequality $\| \alpha - \alpha^0 \|^2 \leq |m|\, K/n$ except on a set whose probability verifies:

$$\mathbb{P}\left(\| \alpha - \alpha^0 \|^2 > \frac{|m|}{n} K \right) \leq \frac{1}{|m|\, K} \sum_{i=1}^{d} \sum_{k=1}^{m_i} \sigma_{i,k}^2.$$

with $\sigma_{i,k}^2 = \mathrm{Var}\,[\varphi_{i,k}(X_i)]$. We have the upper bound $\sigma_{i,k}^2 \leq \| f^0 \|_\infty \int_\triangle \varphi_{[i],k}^2 \leq \mathrm{e}^{\|\log(f^0)\|_\infty}$ by the normality of $\varphi_{i,k}$. Therefore we obtain:

$$\mathbb{P}\left(\| \alpha - \alpha^0 \|^2 > \frac{|m|}{n} K \right) \leq \frac{\mathrm{e}^{\|\log(f^0)\|_\infty}}{K}.$$

We can apply Lemma 7.8 on the event $\{\| \alpha - \alpha^0 \| \leq \sqrt{|m|\, K/n}\}$ if:

$$\sqrt{\frac{|m|}{n} K} \leq \frac{\mathrm{e}^{-(1 + \|\log(f_{\theta^*})\|_\infty)}}{6d^{\frac{3}{2}}\kappa_m}. \tag{8.5}$$

Thanks to (8.3) we have:

$$\| \log(f_{\theta^*}) \|_\infty \leq \| \log(f^0/f_{\theta^*}) \|_\infty + \| \log(f^0) \|_\infty \leq 2\gamma_m + \varepsilon_m + \| \log(f^0) \|_\infty. \tag{8.6}$$

Since $\varepsilon_m \leq 1$, (8.5) holds if $\delta_{m,n}^2 \leq 1/K$. Then except on a set of probability less than $\mathrm{e}^{\|\log(f^0)\|_\infty}/K$, the maximum likelihood estimator $\hat{\theta}_{m,n}$ satisfies, thanks to (7.19) with $\tau = \delta_{m,n}\sqrt{K}$:

$$D\left(f_{\theta^*} \| f_{\hat{\theta}_{m,n}} \right) \leq 3d\, \mathrm{e}^{\|\log(f_{\theta^*})\|_\infty + \tau} \frac{|m|}{n} K \leq 3d\, \mathrm{e}^{2\gamma_m + \varepsilon_m + \|\log(f^0)\|_\infty + \tau} \frac{|m|}{n} K. \tag{8.7}$$

□

8.3. Proof of Theorem 3.3

Recall that $r = (r_1, \dots, r_d) \in \mathbb{N}^d$ is fixed. We assume $\ell_i^0 \in W_{r_i}^2(q_i)$ for all $1 \leq i \leq d$. Corollary 6.11 ensures $\Delta_m = O(\sqrt{\sum_{i=1}^d m_i^{-2r_i}})$ and the boundedness of γ_m when $m_i > r_i$ for all $1 \leq i \leq d$ is due to Corollary 6.13. By Remark 6.9, we have that $\kappa_m = O(|m|^d)$. If (3.5) holds, then $\kappa_m \Delta_m$ converges to 0. Therefore for m large enough, we have that ε_m defined in (8.2) is less than 1. By Proposition 8.3, the information projection f_{θ^*} of f^0 exists. For such m, by Lemma 7.4, we have that for all $\theta \in \mathbb{R}^{|m|}$:

$$D\left(f^0 \| f_\theta\right) = D\left(f^0 \| f_{\theta^*}\right) + D\left(f_{\theta^*} \| f_\theta\right).$$

Proposition 8.3 and $\Delta_m = O(\sqrt{\sum_{i=1}^d m_i^{-2r_i}})$ ensures that the $D\left(f^0 \| f_{\theta^*}\right) = O(\sum_{i=1}^d m_i^{-2r_i})$. The condition $\delta_{m,n} \leq 1$ in Proposition 8.5 is verified for n large enough since γ_m is bounded and (3.6) holds, giving $\lim_{n\to\infty} \delta_{m,n} = 0$. Proposition 8.5 then ensures that $D\left(f_{\theta^*} \| \hat{f}_{m,n}\right) = O_{\mathbb{P}}(|m| / n)$. Therefore the proof is complete.

9. Proof of Theorem 4.1

In this section we provide the elements of the proof of Theorem 4.1. We assume the hypotheses of Theorem 4.1. Recall the notation of Section 4. We shall stress out when we use the inequalities (4.7), (4.8) and (4.9) to achieve uniformity in r in Corollary 4.2.

First recall that ℓ^0 from (2.3) admits the following representation: $\ell^0 = \sum_{i=1}^d \sum_{k=1}^\infty \theta_{i,k}^0 \varphi_{[i],k}$. For $m = (m_1, \dots, m_d) \in (\mathbb{N}^*)^d$, let

$$\ell_m^0 = \sum_{i=1}^d \sum_{k=1}^{m_i} \theta_{i,k}^0 \varphi_{[i],k} \text{ and } f_m^0 = \exp(\ell_m^0 - \psi(\theta_m^0)).$$

Using Corollary 6.13 and $\left|\psi(\theta_m^0) - \mathrm{a}_0\right| \leq \| \ell_m^0 - \ell^0 \|_\infty$, we get $\| \log(f_m^0/f^0) \|_\infty$ is bounded for all $m \in (\mathbb{N}^*)^d$ such that $m_i \geq r_i$:

$$\| \log(f_m^0/f^0) \|_\infty \leq 2\gamma_m \leq 2\gamma, \tag{9.1}$$

with $\gamma_m = \| \ell_m^0 - \ell^0 \|_\infty$, and $\gamma = \mathcal{C} \sum_{i=1}^d \| (\ell_i^0)^{(r_i)} \|_{L^2(q_i)}$ with $\mathcal{C}$ defined in Corollary 6.13 which does not depend on r or m. For $m = (v, \dots, v) \in \mathcal{M}_n$, we have that $a_n \leq v \leq b_n$, with a_n, b_n given by:

$$a_n = \left\lfloor n^{1/(2(d+N_n)+1)} \right\rfloor \quad \text{and} \quad b_n = \left\lfloor n^{1/(2(d+1)+1)} \right\rfloor. \tag{9.2}$$

The upper bound (9.1) is uniform over $m \in \mathcal{M}_n$ and $r \in (\mathcal{R}_n)^d$ when (4.8) holds. Since $N_n = o(\log(n))$, we have $\lim_{n\to+\infty} a_n = +\infty$. Hence, for n large enough, say $n \geq n^*$, we have $\varepsilon_m \leq 1$ for all $m = (v, \dots, v) \in \mathcal{M}_n$ with ε_m given by (8.2), since $\kappa_m \Delta_m = O(a_n^{d-\min(r)})$. According to Lemma 8.4 and its proof, this means that the information projection $f_{\theta_m^*}$ of f^0 onto the set of functions $(\varphi_{[i],k}, 1 \leq i \leq d, 1 \leq k \leq v)$ verify, by (7.18) with $\tau = 1$, for all $m \in \mathcal{M}_n$:

$$\| \log(f_{\theta_m^*}/f_m^0) \|_\infty \leq 1. \tag{9.3}$$

Recall the notation $A_m^0 = \int_\triangle \varphi_m f^0$ for the expected value of $\varphi_m(X^1)$, $\hat{\mu}_{m,n}$ the corresponding empirical mean based on the sample $\mathbb{X}_1^n$ of size $n_1 = \lfloor C_e n \rfloor$, and $\hat{\ell}_{m,n} = \hat{\theta}_{m,n} \cdot \varphi_m$ where $\hat{\theta}_{m,n}$ is the maximum likelihood estimate given by (3.4). Let $T_n > 0$ be defined as:

$$T_n = \frac{n_1 \, \mathrm{e}^{-4\gamma-4-2\,\|\log(f^0)\|_\infty}}{72 d^5 d! b_n (b_n + d)^{2d} \log(b_n)}, \tag{9.4}$$

with b_n given by (9.2) and γ as in (9.1). We define the sets:

$$\mathcal{B}_{m,n} = \{\| A_m^0 - \hat{\mu}_{m,n} \|^2 > |m| \, T_n \log(b_n)/n_1\} \quad \text{and} \quad \mathcal{A}_n = \left(\bigcup_{m \in \mathcal{M}_n} \mathcal{B}_{m,n} \right)^c.$$

We first show that with probability converging to 1, the estimators are uniformly bounded.

Lemma 9.1. *Let $n \in \mathbb{N}^*$, $n \geq n^*$ and $\mathcal{M}_n$ as in (4.1). Then we have:*

$$\mathbb{P}(\mathcal{A}_n) \geq 1 - N_n 2dn^{C_{T_n}},$$

with C_{T_n} defined as:

$$C_{T_n} = \frac{1}{2d+3} \left(1 - \frac{T_n}{2 \| f^0 \|_\infty + C\sqrt{T_n}} \right),$$

with a finite constant C given by (9.9). Moreover, on the event $\mathcal{A}_n$, we have the following uniform upper bound for $\| \hat{\ell}_{m,n} \|_\infty$, $m \in \mathcal{M}_n$:

$$\| \hat{\ell}_{m,n} \|_\infty \leq 4 + 4\gamma + 2 \| \log(f^0) \|_\infty. \tag{9.5}$$

Remark 9.2. Notice that by the definition of b_n, $\lim_{n\to\infty} T_n = +\infty$. For n large enough, we have $C_{T_n} < -\varepsilon < 0$ for some positive ε, so that:

$$\lim_{n\to\infty} N_n 2dn^{C_{T_n}} = 0. \tag{9.6}$$

This ensures that $\lim_{n\to\infty} \mathbb{P}(\mathcal{A}_n) = 1$, that is $(\hat{\ell}_{m,n}, m \in \mathcal{M}_n)$ are uniformly bounded with probability converging to 1.

Proof. For $m = (v, \dots, v) \in \mathcal{M}_n$ fixed, in order to bound the distance between the vectors $\hat{\mu}_{m,n} = (\hat{\mu}_{m,n,i,k}, 1 \leq i \leq d, 1 \leq k \leq v)$ and $A_m^0 = \mathbb{E}[\hat{\mu}_{m,n}] =$

$(\alpha^0_{i,k}, 1 \leq i \leq d, 1 \leq k \leq v)$, we first consider a single term $\left|\alpha^0_{i,k} - \hat{\mu}_{m,n,i,k}\right|$. By Bernstein's inequality, we have for all $t > 0$:

$$\begin{aligned}
&\mathbb{P}\left(\left|\alpha^0_{i,k} - \hat{\mu}_{m,n,i,k}\right| > t\right) \leq 2\exp\left(-\frac{(n_1 t)^2/2}{n_1 \mathrm{Var}\ \varphi_{[i],k}(X^1) + 2n_1 t \parallel \varphi_{i,k} \parallel_\infty /3}\right) \\
&\leq 2\exp\left(-\frac{(n_1 t)^2/2}{n_1 \mathbb{E}\left[\varphi^2_{[i],k}(X^1)\right] + 2n_1 t\sqrt{2(d-1)!}(b_n + d)^{d-\frac{1}{2}}/3}\right) \\
&\leq 2\exp\left(-\frac{n_1 t^2/2}{\parallel f^0 \parallel_\infty + 2t\sqrt{2(d-1)!}(b_n + d)^{d-\frac{1}{2}}/3}\right),
\end{aligned}$$

where we used, thanks to (6.9):

$$\parallel \varphi_{i,k} \parallel_\infty \leq \sqrt{(d-1)!}\sqrt{2k+d}\frac{(k+d-1)!}{k!} \leq \sqrt{2(d-1)!}(b_n + d)^{d-\frac{1}{2}}$$

for the second inequality, and the orthonormality of $\varphi_{[i],k}$ for the third inequality. Let us choose $t = \sqrt{T_n \log(b_n)/n_1}$. This gives:

$$\begin{aligned}
&\mathbb{P}\left(\left|\alpha^0_{i,k} - \hat{\mu}_{m,n,i,k}\right| > \sqrt{\frac{T_n \log(b_n)}{n_1}}\right) \\
&\leq 2\exp\left(-\frac{T_n \log(b_n)/2}{\parallel f^0 \parallel_\infty + 2\sqrt{\frac{2T_n \log(b_n)(d-1)!(b_n+d)^{2d-1}}{9n_1}}}\right) \qquad (9.7) \\
&\leq 2b_n^{-\frac{T_n}{2\,\|f^0\|_\infty + C\sqrt{T_n}}}, \qquad (9.8)
\end{aligned}$$

with C given by:

$$C = \sup_{n \in \mathbb{N}^*} 4\sqrt{\frac{2\log(b_n)(d-1)!(b_n + d)^{2d-1}}{9n_1}}. \qquad (9.9)$$

Notice $C < +\infty$ since the sequence $\sqrt{\log(b_n)(b_n + d)^{2d-1}/9n_1}$ is $o(1)$. For the probability of $\mathcal{B}_{n,m}$ we have:

$$\begin{aligned}
\mathbb{P}(\mathcal{B}_{n,m}) &\leq \sum_{i=1}^{d}\sum_{k=1}^{v}\mathbb{P}\left(\left|\alpha^0_{i,k} - \hat{\mu}_{m,n,i,k}\right|^2 > \frac{T_n \log(b_n)}{n_1}\right) \\
&\leq \sum_{i=1}^{d}\sum_{k=1}^{v} 2b_n^{-\frac{T_n}{2\,\|f^0\|_\infty + C\sqrt{T_n}}} \\
&\leq 2dn^{C_{T_n}}.
\end{aligned}$$

This implies the following lower bound on $\mathbb{P}(\mathcal{A}_n)$:

$$\mathbb{P}(\mathcal{A}_n) = 1 - \mathbb{P}\left(\bigcup_{m \in \mathcal{M}_n} \mathcal{B}_{n,m}\right) \geq 1 - \sum_{m \in \mathcal{M}_n} \mathbb{P}(\mathcal{B}_{n,m}) \geq 1 - N_n 2dn^{C_{T_n}}.$$

On $\mathcal{A}_n$, by the definition of T_n, we have for all $m \in \mathcal{M}_n$:

$$\begin{aligned}\| A_m^0 - \hat{\mu}_{m,n} \| \, 6d^2\sqrt{2d!}(v+d)^d \, \mathrm{e}^{2\gamma_m+2} &\leq \sqrt{b_n \frac{T_n \log(b_n)}{n_1}} 6d^{\frac{5}{2}} \sqrt{2d!}(b_n+d)^d \, \mathrm{e}^{2\gamma+2} \\ &\leq \mathrm{e}^{-\, \|\log(f^0)\|_\infty} . \end{aligned} \tag{9.10}$$

Notice that whenever (9.10) holds, condition (7.16) of Lemma 7.8 is satisfied with $\theta = \theta_m^*$ and $\alpha = \hat{\mu}_{m,n}$, thanks to $\kappa_m \leq \sqrt{d2d!}(v+d)^d$ and:

$$\begin{aligned}\| \log(f_{\theta_m^*}) \|_\infty &\leq \| \log(f_{\theta_m^*}/f_m^0) \|_\infty + \| \log(f_m^0/f^0) \|_\infty + \| \log(f^0) \|_\infty \\ &\leq 1 + 2\gamma + \| \log(f^0) \|_\infty .\end{aligned}$$

According to Equation (7.18) with $\tau = 1$, we can deduce that on $\mathcal{A}_n$, we have:

$$\| \log(\hat{f}_{m,n}/f_{\theta_m^*}) \|_\infty \leq 1 \quad \text{for all } m \in \mathcal{M}_n, n \geq n^*.$$

This, along with (9.1) and (9.3), provide the following uniform upper bound for $(\| \hat{\ell}_{m,n} \|_\infty, m \in \mathcal{M}_n)$ on $\mathcal{A}_n$:

$$\begin{aligned}&\frac{1}{2} \| \hat{\ell}_{m,n} \|_\infty \leq \| \log(\hat{f}_{m,n}) \|_\infty \\ &\leq \| \log(\hat{f}_{m,n}/f_{\theta_m^*}) \|_\infty + \| \log(f_{\theta_m^*}/f_m^0) \|_\infty + \| \log(f_m^0/f^0) \|_\infty + \| \log(f^0) \|_\infty \\ &\leq 2 + 2\gamma + \| \log(f^0) \|_\infty,\end{aligned}$$

where we used (7.7) for the first inequality. □

We also give a sharp oracle inequality for the convex aggregate estimator $f_{\hat{\lambda}_n^*}$ conditionally on $\mathcal{A}_n$ with n fixed. The following lemma is a direct application of Theorem 3.1. of [12] and (9.5).

Lemma 9.3. *Let $n \in \mathbb{N}^*$ be fixed. Conditionally on $\mathcal{A}_n$, let $f_{\hat{\lambda}_n^*}$ be given by (4.3) with $\hat{\lambda}_n^*$ defined as in (4.5). Then for any $x > 0$ we have with probability greater than $1 - \exp(-x)$:*

$$D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) - \min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \leq \frac{\beta(\log(N_n) + x)}{n_2}, \tag{9.11}$$

with $\beta = 2\exp(6K + 2L) + 4K/3$, and $L, K \in \mathbb{R}$ given by :

$$L = \| \ell^0 \|_\infty, \qquad K = 4 + 4\gamma + 2 \| \log(f^0) \|_\infty,$$

with γ as in (9.1).

Now we prove Theorem 4.1. For $n \in \mathbb{N}^*$ and $C > 0$, we define the event $\mathcal{D}_n(C)$ as:

$$\mathcal{D}_n(C) = \left\{ D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) \geq C\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) \right\}.$$

Let $\varepsilon > 0$. To prove (4.6), we need to find $C_\varepsilon > 0$ such that for all n large enough:

$$\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon)\right) \leq \varepsilon. \tag{9.12}$$

We decompose the left hand side of (9.12) according to $\mathcal{A}_n$:

$$\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon)\right) \leq \mathbb{P}\left(\mathcal{D}_n(C_\varepsilon) \,|\, \mathcal{A}_n\right) \mathbb{P}(\mathcal{A}_n) + \mathbb{P}(\mathcal{A}_n^c). \tag{9.13}$$

The product $\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon) \,|\, \mathcal{A}_n\right) \mathbb{P}(\mathcal{A}_n)$ is bounded by:

$$\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon) \,|\, \mathcal{A}_n\right) \mathbb{P}(\mathcal{A}_n) \leq A_n(C_\varepsilon) + B_n(C_\varepsilon),$$

with $A_n(C_\varepsilon)$ and $B_n(C_\varepsilon)$ defined by:

$$A_n(C_\varepsilon) = \mathbb{P}\left(D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) - \min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \geq \frac{C_\varepsilon}{2}\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) \middle| \mathcal{A}_n \right),$$

$$B_n(C_\varepsilon) = \mathbb{P}\left(\min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \geq \frac{C_\varepsilon}{2}\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) \right).$$

To bound $A_n(C_\varepsilon)$ we apply Lemma 9.3 with $x = x_\varepsilon = -\log(\varepsilon/4)$:

$$\mathbb{P}\left(D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) - \min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \geq \frac{\beta(\log(N_n) + x_\varepsilon)}{n_2} \middle| \mathcal{A}_n \right) \leq \frac{\varepsilon}{4}.$$

Let us define $C_{\varepsilon,1}$ as:

$$C_{\varepsilon,1} = \sup_{n \in \mathbb{N}^*} \left(\frac{\beta(\log(N_n) + x_\varepsilon)}{n_2 n^{-\frac{2\min(r)}{2\min(r)+1}}} \right). \tag{9.14}$$

Since $N_n = o(\log(n))$, we have $C_{\varepsilon,1} < +\infty$ as the sequence on the right hand side of (9.14) is bounded. This bound is uniform over regularities in $(\mathcal{R}_n)^d$ thanks to (4.9) Therefore for all $C_\varepsilon \geq C_{\varepsilon,1}$, we have $A_n(C_\varepsilon) \leq \varepsilon/4$.

For $B_n(C_\varepsilon)$, note that if $n \geq \bar{n}$ with $\bar{n}$ given by (4.2), then $m^* = (v^*, \ldots, v^*) \in \mathcal{M}_n$ with $v^* = \lfloor n^{1/(2\min(r)+1)} \rfloor$. This holds for all $r \in (\mathcal{R}_n)^d$ due to (4.7). By Remark 3.4, we have that $D\left(f^0 \| \hat{f}_{m^*,n}\right) = O_{\mathbb{P}}(n^{-2\min(r)/(2\min(r)+1)})$. This ensure that there exists $C_{\varepsilon,2}$ such that for all $C_\varepsilon \geq C_{\varepsilon,2}$, $n \geq \bar{n}$:

$$B_n(C_\varepsilon) \leq \mathbb{P}\left(D\left(f^0 \| \hat{f}_{m^*,n}\right) \geq \frac{C_{\varepsilon,2}}{2}\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) \right) \leq \frac{\varepsilon}{4}.$$

We also have by (9.6) that there exists $\tilde{n} \in \mathbb{N}^*$ such that $\mathbb{P}(\mathcal{A}_n^c) \leq \varepsilon/2$ for all $n \geq \tilde{n}$. Therefore by setting $C_\varepsilon = \max(C_{\varepsilon,1}, C_{\varepsilon,2})$ in (9.13), we have for all $n \geq \max(n^*, \bar{n}, \tilde{n})$:

$$\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon)\right) \leq A_n(C_\varepsilon) + B_n(C_\varepsilon) + \mathbb{P}(\mathcal{A}_n^c) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which gives (9.12) and thus concludes the proof.

Acknowledgement

The authors would like to thank the referees for their valuable comments which helped to improve the presentation of the results.

References

[1] R. Abraham, J.-F. Delmas, and H. Guo. Critical multi-type Galton-Watson trees conditioned to be large. *arXiv preprint arXiv:1511.01721*, 2015. MR3368966

[2] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables.* Courier Dover Publications, 1970. MR1225604

[3] J. Avérous, C. Genest, and S. C. Kochar. On the dependence structure of order statistics. *Journal of multivariate analysis*, 94(1):159–171, 2005. MR2161215

[4] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999. MR1679028

[5] A. R. Barron, L. Gyorfi, and E. C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *Information Theory, IEEE Transactions on*, 38(5):1437–1454, 1992. MR1178189

[6] A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991. MR1126328

[7] K. Bertin. Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder classes. *Bernoulli*, 10(5):873–888, 2004. MR2093615

[8] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, pages 55–87. Springer New York, 1997. MR1462939

[9] C. Butucea. Exact adaptive pointwise estimation on Sobolev classes of densities. *ESAIM: Probability and Statistics*, 5:1–31, 2001. MR1845320

[10] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Nonparametric estimation of distributions of order statistics with application to nuclear engineering. In L. Podofillini, B. Sudret, B. Stojadinovic, E. Zio, and W. Kröger, editors, *Safety and Reliability of Complex Engineered Systems: ESREL 2015*, pages 2657–2665. CRC Press, 2015.

[11] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Maximum entropy distribution of order statistics with given marginals. *Bernoulli, to appear*, 2017. MR3706752

[12] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Optimal exponential bounds for aggregation of estimators for the kullback-leibler loss. *Electron. J. Statist.*, 11(1):2258–2294, 2017. MR3654825

[13] O. Catoni. The mixture approach to universal model selection. Technical report, École Normale Supérieure, 1997.

[14] B. R. Crain. An information theoretic approach to approximating a probability distribution. *SIAM Journal on Applied Mathematics*, 32(2):339–346, 1977. MR0436440

[15] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996. MR1394974

[16] C. F. Dunkl and Y. Xu. *Orthogonal polynomials of several variables*, volume 81. Cambridge University Press, 2001. MR1827871

[17] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011. MR2850214

[18] I. J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, pages 911–934, 1963. MR0150880

[19] E. Guerre, I. Perrigne, and Q. Vuong. Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68(3):525–574, 2000. MR1769378

[20] P. Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, pages 1491–1519, 1987. MR0913570

[21] T. Herbst. An application of randomly truncated data models in reserving IBNR claims. *Insurance: Mathematics and Economics*, 25(2):123–131, 1999.

[22] P. Joly, D. Commenges, and L. Letenneur. A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics*, pages 185–194, 1998.

[23] G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability Theory and Related Fields*, 121(2):137–170, 2001. MR1863916

[24] G. Kerkyacharian, D. Picard, and K. Tribouley. L_p adaptive density estimation. *Bernoulli*, pages 229–247, 1996. MR1416864

[25] J.-Y. Koo and W.-C. Kim. Wavelet density estimation by approximation of log-densities. *Statistics and Probability Letters*, 26(3):271–278, 1996. MR1394903

[26] S. W. Lagakos, L. Barraj, and V. De Gruttola. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75(3):515–523, 1988. MR0967591

[27] O. Lepski. Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *The Annals of Statistics*, 41(2):1005–1034, 2013. MR3099129

[28] O. V. Lepski. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, 35(3):454–466, 1991. MR1091202

[29] X. Luo and W.-Y. Tsai. Nonparametric estimation of bivariate distribution under right truncation with application to panic disorder. *Journal of Statistical Planning and Inference*, 139(4):1559–1568, 2009. MR2485148

[30] D. Lynden-Bell. A method of allowing for known observational selection in small samples applied to 3cr quasars. *Monthly Notices of the Royal*

Astronomical Society, 155(1):95–118, 1971.

[31] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007. MR2356821

[32] B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B*, 38(3):290–295, 1976. MR0652727

[33] X. Wu. Exponential series estimator of multivariate densities. *Journal of Econometrics*, 156(2):354–366, 2010. MR2609938

[34] Y. Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87, 2000. MR1762904

[35] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. MR1742500

[36] T. Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006. MR2291497