# Density estimation with multivariate histograms and best basis selection

Jussi Klemelä
Department of Economics, University of Mannheim
L 7 3-5 Verfügungsgebäude
68131 Mannheim, Germany
Email: klemela@rumms.uni-mannheim.de
Fax +49 621 1811931

November 17, 2006

**Abstract**

We consider estimation of multivariate densities with histograms which are based on data-dependent partitions. We find data-dependent partitions by minimizing a complexity penalized error criterion. The estimator may also be characterized as a series estimator whose basis is chosen empirically. We show that the estimator achieves minimax rates of convergence up to a logarithmic factor over a scale of smoothness classes containing functions with anisotropic and spatially varying smoothness. The method may also be viewed as based on the presmoothing of data. We show how the optimal amount of presmoothing depends on the spatial inhomogeneity of the density.

**Mathematics Subject Classifications (AMS 2000):** 62G07
**Key Words:** adaptive estimation, dyadic CART, multivariate density estimation, presmoothing, tree structured estimators.
**Short title:** Density estimation with histograms

# 1 Introduction

We consider density estimation based on i.i.d. multivariate random vectors taking values in $\mathbf{R}^d$. We estimate densities with histograms, which we define to be rectangularwise constant estimates, and the value of the estimate in

each rectangle is taken to be the empirical probability divided by the volume of the rectangle. The main problem is how to choose the partition defining the histogram in an optimal way.

Histograms with equispaced bins are not able to adapt to spatially varying smoothness. This problem appears already in the one dimensional case. Furthermore, in the multivariate case the density to be estimated may have anisotropic smoothness: the density function may vary more in one direction than in the other directions. We should choose bins to be thinner in the direction where the density varies more.

By choosing the partition in a flexible way we are not so vulnerable to the curse of dimensionality. Indeed, in high dimensional cases accurate estimation may be possible if the "effective dimension" of the density is small. Effective dimension could mean for example the number of variables with respect which the density has variability. Indeed, an extreme case of anisotropic smoothness occurs when the density is almost constant on its support with respect to some variables. These types of densities could be estimated well if we had a method of choosing the partition of the histogram economically: we should choose a partition which only delineates the support with respect to those directions where there is no variation.

We define the histogram estimator as a minimizer of a complexity penalized error criterion. As the error criterion we take the empirical risk with the $L_2$ contrast function, and the complexity of the histogram is defined to be the number of sets in the partition. The set of candidate partitions is fixed, and defined by the set of dyadic splitting sequences. Thus the estimator is similar to the dyadic regressograms considered in Donoho (1997).

An important property of the estimator is that we can define it in two ways: (1) as a histogram estimator and (2) as a series estimator associated to a basis of multivariate Haar functions. The characterization of the estimator as a series estimator makes it possible to analyze asymptotic properties of the estimator and the definition of the estimator as a histogram makes it possible to find a fast algorithm for evaluating the estimates. In the histogram characterization the partition is chosen empirically and in the series estimator characterization the basis is chosen empirically. Instead of thresholding the empirical coefficients in a fixed basis the method chooses empirically a basis where the thresholding is performed.

We show that the estimator achieves minimax rates up to a logarithmic factor over a scale of anisotropic smoothness classes, for the $L_2$ loss. We consider histograms with unequal binwidths in every direction and thus we nearly achieve the minimax rates over smoothness classes containing functions with considerable spatially varying smoothness. To apply the estimator we have to choose a bound for the maximal fineness of the partitions we consider. We

may increase the flexibility of the estimator by choosing the maximal allowed resolution to be fine. On the other hand this will increase the computational complexity of the estimator. We shall show how the bound for the maximal fineness depends on the spatial inhomogeneity of the density. We show also how the computational complexity depends on this bound for the maximal fineness. The method we propose may be seen as based on presmoothing the data since the estimator uses only the frequencies on the partition defined by the finest resolution level.

We give some references to the previous literature on histograms with irregular data-dependent partitions, and on other spatially flexible estimation methods.

1. *Multivariate regressograms.* Breiman, Friedman, Olshen and Stone (1984) introduced CART (Classification and Regression Trees) as a method for estimating classification and regression functions with piecewise constant estimates. They constructed data-dependent partitions by a two step procedure. First they found a set of candidate partitions by minimizing an empirical error criterion in a myopic fashion, and then they chose the final partition by minimizing an error-complexity criterion among the set of candidate partitions.

   Donoho (1997) considered 2-dimensional Gaussian regression on a fixed and regularly spaced design. He considers an estimator which is defined as a minimizer of an error-complexity criterion. Unlike in CART, where the set of candidate partitions is constructed empirically, he considered candidate partitions which are obtained by sequential dyadic splitting of the rectangle containing the support of the regression function.

2. *Multivariate histograms.* Density estimation with CART-type methods was considered by Shang (1994), Sutton (1994), Ooi (2002). Hüsemann and Terrell (1991) consider the problem of optimal fixed and variable cell dimensions in bivariate histograms. Lugosi and Nobel (1996) present $L_1$-consistency results on density estimators based on data dependent partitions. Barron, Birgé and Massart (1999) constructed a multivariate histogram which achieves asymptotic minimax rates over anisotropic Hölder classes for the $L_2$ loss. Their histograms had different number of bins in different directions but in a single direction bins were equispaced. A modified Akaike criterion for histogram estimation with irregular splits was studied in the multivariate case by Castellan (2000) who gives oracle inequalities for Kullback-Leibler and Hellinger loss.

3. *Other methods.* Multivariate density estimation based on wavelet expansions has been considered in Tribouley (1995). Neumann (2000) constructed an estimator based on wavelet expansions which achieves minimax rates up to a logarithmic factor over a large scale of anisotropic Besov classes in the Gaussian white noise model. Kerkyacharian, Lepski and Picard (2001) consider a kernel based adaptation scheme to cope with anisotropic smoothness.

In Section 2 we define the estimator in two ways as a histogram and as a series estimator. We present an algorithm for the computation of an estimate. In Section 3 we give the rates of convergence of the estimator. Some of the proofs are in the Appendices.

# 2 Estimators

## 2.1 Dyadic histogram

Let $X^1, \ldots, X^n \in \mathbf{R}^d$ be i.i.d. random vectors whose density function we want to estimate. A histogram with partition $\mathcal{P}$ is defined by

$$\hat{f}(x, \mathcal{P}) = \sum_{R \in \mathcal{P}} \frac{n_R}{n \operatorname{vol}(R)} I_R(x), \qquad x \in \mathbf{R}^d, \tag{1}$$

where $n_R = \#\{X^i \in R\}$ are the frequencies for the sets of the partition. We will define a set of partitions from which we search the optimal partition.

**Collection of partitions.** We define a collection of dyadic partition generating trees. The optimal partition will be searched from the collection of partitions generated by these partition generating trees.

**Definition 1** *A collection of dyadic partition generating trees $\mathbb{T}(R_0, J)$, associated with a rectangle $R_0 \subset \mathbf{R}^d$, and with a bound for split numbers $J = (J_1, \ldots, J_d)$, $J_l \in \{0, 1, \ldots\}$, consists of binary trees whose each node is annotated with a rectangle, and each non-leaf node is annotated with a splitting direction in $\{1, \ldots, d\}$.*

1. *The root node is annotated with $R_0$.*

2. *Let a non-leaf node be annotated with rectangle $R = \Pi_{m=1}^d [c_m, d_m]$ and direction $l \in \{1, \ldots, d\}$. The split point is*

$$s = (d_l - c_l)/2.$$

*Denote*

$$R_{l,s}^{(0)}(R) = \{x \in R : x_l < s\}, \qquad R_{l,s}^{(1)}(R) = \{x \in R : x_l \geq s\}.$$

*The left child of the node is annotated with* $R_{l,s}^{(0)}(R)$ *and the right child is annotated with* $R_{l,s}^{(1)}(R)$.

3. *In direction* $l$ *at most* $J_l$ *splits will be made,* $l = 1, \ldots, d$.

We make some remarks concerning the definition.

- In fact, a set of dyadic partition generating trees is completely determined by the initial rectangle and by the splitting directions; since the splits are always made at the midpoints of the sides of the rectangles, the annotation of the nodes with rectangles is redundant.
- The simplest dyadic partition generating tree is the tree which consists only of the root node, and this tree is the single member of $\mathbb{T}(R_0, 0)$.
- The bound $J$ for the split numbers implies a bound for the depth of the tree: the depth is at most $|J| = \sum_{i=1}^{d} J_i$. (We define the depth of a tree to be equal to the largest depth among the depths of its nodes and we stipulate that the depth of the root node is 0 and the depth of the children of the root is 1, and so on.)
- Note that a tree generating a dyadic partition may be an unbalanced tree: some terminal nodes may have depth equal to $|J|$ but the depth of some other terminal nodes may be less than $|J|$.

Each tree in the set $\mathbb{T}(R_0, J)$ generates a partition: the partition is the collection of the rectangles associated with the leaf nodes of the tree. This is the content of the definition below.

**Definition 2** (Collection of dyadic partitions.) *The dyadic partition associated to tree* $\mathcal{T} \in \mathbb{T}(R_0, J)$*, where* $\mathbb{T}(R_0, J)$ *is defined in Definition 1, is*

$$\mathcal{P}(\mathcal{T}) = \{R(t) : t \in \mathrm{Ter}(\mathcal{T})\}, \tag{2}$$

*where* $\mathrm{Ter}(\mathcal{T})$ *is the set of terminal nodes of* $\mathcal{T}$*, and* $R(t)$ *is the rectangle annotated to node* $t$. The collection of dyadic partitions $\mathbb{P} = \mathbb{P}(R_0, J)$*, with base rectangle* $R_0$ *and with depth bound* $J$ *is denoted with*

$$\mathbb{P}(R_0, J) = \{\mathcal{P}(\mathcal{T}) : \mathcal{T} \in \mathbb{T}(R_0, J)\}. \tag{3}$$

**Complexity penalized error criterion.** Define the empirical risk of a density estimator $\hat{f} : \mathbf{R}^d \to \mathbf{R}$ with

$$\gamma_n\left(\hat{f}\right) = \frac{1}{n}\sum_{i=1}^{n}\gamma\left(\hat{f}, X^i\right), \tag{4}$$

where $\gamma(g, x)$ is the $L_2$ contrast function,

$$\gamma(g, x) = -2g(x) + \|g\|_2^2, \qquad g : \mathbf{R}^d \to \mathbf{R}, \;\; x \in \mathbf{R}^d.$$

Minimization of $\|\hat{f} - f\|_2^2$ over estimators $\hat{f}$ is equivalent to the minimization of $\|\hat{f} - f\|_2^2 - \|f\|_2^2$, and minimization of $\gamma_n(\hat{f})$ amounts to the minimization of $\|\hat{f} - f\|_2^2 - \|f\|_2^2$, up to the approximation $\int_{\mathbf{R}^d}\hat{f}f \approx n^{-1}\sum_{i=1}^{n}\hat{f}(X^i)$. Indeed,

$$\begin{aligned}
\|\hat{f} - f\|_2^2 - \|f\|_2^2 &= -2\int_{\mathbf{R}^d} f\hat{f} + \|\hat{f}\|_2^2 \\
&\approx -2n^{-1}\sum_{i=1}^{n}\hat{f}(X^i) + \|\hat{f}\|_2^2 \\
&= \gamma_n\left(\hat{f}\right).
\end{aligned} \tag{5}$$

A histogram $\hat{f}(\cdot, \mathcal{P})$ is uniquely defined through its partition $\mathcal{P}$ and we use the notation

$$ERR_n(\mathcal{P}) = \gamma_n\left(\hat{f}(\cdot, \mathcal{P})\right). \tag{6}$$

We have that

$$ERR_n(\mathcal{P}) = -\sum_{R\in\mathcal{P}}\frac{n_R^2}{n^2\mathrm{vol}(R)} = -\left\|\hat{f}^2(\cdot, \mathcal{P})\right\|_2^2. \tag{7}$$

The complexity of a histogram is taken to be the number of sets in the partition of the histogram. Let $0 \le \alpha < \infty$ and define the complexity penalized error criterion as

$$COPERR_n(\mathcal{P}, \alpha) = ERR_n(\mathcal{P}) + \alpha \cdot \#\mathcal{P}. \tag{8}$$

**Definition of the dyadic histogram.** The dyadic histogram is defined as a minimizer of the complexity penalized empirical risk, when we minimize the complexity penalized empirical risk over the set of dyadic partitions.

**Definition 3** (Dyadic histogram.) *Define the partition corresponding to parameter $\alpha$ as*

$$\hat{\mathcal{P}}_\alpha = \mathrm{argmin}_{\mathcal{P}\in\mathbb{P}(R_0,J)}COPERR_n(\mathcal{P}, \alpha), \tag{9}$$

*where* $\mathbb{P}(R_0, J)$ *is defined in (3). The* dyadic histogram *is defined as*

$$\hat{f}_{n,\alpha} = \hat{f}(\,\cdot\,, \hat{\mathcal{P}}_\alpha). \tag{10}$$

*where* $\hat{f}(\,\cdot\,, \mathcal{P})$ *is defined in (1).*

**Remark 1** The estimator depends, besides the smoothing parameter $\alpha$, also on the maximal directionwise split numbers $J$ and on the initial rectangle $R_0$. Theorem 1 gives conditions for the choice of these parameters. In particular, $\alpha$ and $J$ will depend on the sample size $n$. In Theorem 1 we take $R_0 = [0,1]^d$ but in practice one would estimate $R_0$. A reasonable choice is to define $R_0$ as the smallest rectangle containing the observations, whose sides are parallel to the coordinate axis.

## 2.2 Series estimator

We define a series estimator by using a basis of Haar-wavelets. We will prove that the series estimator is in fact identical with a dyadic histogram. A dyadic histogram is a useful representation of the estimator when we want to find algorithms for the calculation of the estimates. The representation of the estimator as a series estimator is useful when we want to find asymptotic properties of the estimator.

Denote

$$\tilde{f}(x, W, \Theta, \mathcal{B}) = I_{[0,1]^d}(x) + \sum_{\phi \in \mathcal{B}} w_\phi \theta_\phi \phi(x), \qquad x \in \mathbf{R}^d, \tag{11}$$

where $\mathcal{B}$ is an orthonormal system of functions in $L_2([0,1]^d)$, $W = (w_\phi)_{\phi\in\mathcal{B}} \in \{0,1\}^{\mathcal{B}}$, $\Theta = (\theta_\phi)_{\phi\in\mathcal{B}} \in \mathbf{R}^{\mathcal{B}}$. Vector $W$ chooses a subset of $\mathcal{B}$ and vector $\Theta$ gives the coefficients of the expansion. We will assume that $\int_{[0,1]^d} \phi = 0$ for all $\phi \in \mathcal{B}$ and since we estimate densities we may include the indicator $I_{[0,1]^d}$ to all expansions.

**Multivariate Haar wavelets.** The univariate Haar scaling function is $\eta^{(0)} = I_{[0,1]}$ and the univariate Haar wavelet is $\eta^{(1)} = I_{[1/2,1)} - I_{[0,1/2)}$. Denote

$$\eta^{(\iota)}_{j_m,k_m}(t) = 2^{j_m/2}\eta^{(\iota)}(2^{j_m}t - k_m), \qquad t \in [0,1],$$

with $\iota \in \{0,1\}$, $j_m \in \{0,1,\ldots\}$, and $k_m \in \{0,\ldots,2^{j_m}-1\}$. Let

$$\phi^{(l)}_{j,k}(x) = \eta^{(1)}_{j_l,k_l}(x_l)\Pi^d_{m=1,m\neq l}\eta^{(0)}_{j_m,k_m}(x_m), \qquad x = (x_1,\ldots,x_d) \in \mathbf{R}^d, \tag{12}$$

where $l \in \{1, \ldots, d\}$, $j = (j_1, \ldots, j_d) \in \{0, 1, \ldots\}^d$, $k = (k_1, \ldots, k_d) \in K_j$, and

$$K_j = \{k = (k_1, \ldots, k_d) : k_l = 0, \ldots, 2^{j_l} - 1, l = 1, \ldots, d\} \tag{13}$$

is the set of translation coefficients corresponding to resolution index $j$. Function $\Pi_{m=1}^d \eta_{j_m,k_m}^{(0)}(x_m)$ is (a constant times) the indicator of a rectangle but we have multiplied with Haar wavelet $\eta_{j_l,k_l}^{(1)}(x_l)$ in (12).

**Dyadic rectangles.** Define the rectangle corresponding to the pair of multi-indeces $(j, k) \in \{0, 1, \ldots\}^d \times K_j$ as

$$R_{jk} = \Pi_{l=1}^d \left[ \frac{k_l}{2^{j_l}}, \frac{k_l + 1}{2^{j_l}} \right), \tag{14}$$

where $K_j$ is defined in (13). We have defined in Definition 1 a collection of dyadic partition generating trees. When the root node is annotated with rectangle $[0, 1)^d$, then every node of the tree is annotated with a dyadic rectangle. We have a bijective correspondence between dyadic rectangles and pairs of multi-indeces, defined by (14). We denote with $\mathcal{I}(t)$ the pair of multi-indeces associated with a node, that is, when a node is annotated with rectangle $R_{jk}$, then $\mathcal{I}(t) = (j, k)$.

**Collection of pre-bases.** Definition 4 of a collection of pre-bases is a counterpart of Definition 2 of a collection of partitions. A difference is that now we take the initial rectangle $R_0 = [0, 1)^d$. We defined in Definition 2 a collection of partitions generated by a collection of partition generating trees. We define analogously a collection of pre-bases generated by a collection of partition generating trees.

In (2) we defined the partition associated with a partition generating tree. We define analogously a pre-basis $\mathcal{B}(\mathcal{T})$ annotated with a partition generating tree. Collection $\mathcal{B}(\mathcal{T})$ is a finite orthonormal system and $\int_{[0,1]^d} \phi = 0$ for each $\phi \in \mathcal{B}(\mathcal{T})$. We call these collections "pre-bases" since it is possible to extend these to be bases of $L_2([0, 1]^d)$.

**Definition 4** (Collection of pre-bases.) *When $\mathcal{T} \in \mathbb{T}([0, 1]^d, J)$ is a dyadic partition generating tree, where $\mathbb{T}([0, 1]^d, J)$ is defined in Definition 1, and $t$ is a node of $\mathcal{T}$, let $s(t) \in \{1, \ldots, d\}$ be the direction annotated with $t$ and let $\mathcal{I}(t)$ be the pair of multi-indeces annotated with $t$. Denote with $NT(\mathcal{T})$ the set of non-terminal nodes of $\mathcal{T}$. The* pre-basis associated to tree $\mathcal{T}$ *is*

$$\mathcal{B}(\mathcal{T}) = \left\{ \phi_{\mathcal{I}(t)}^{(s(t))} : t \in NT(\mathcal{T}) \right\}, \tag{15}$$

*where* $\phi_{j,k}^{(l)}$ *is defined in (12).* The collection of pre-bases $\mathcal{L}(\mathcal{J})$*, with depth bound* $J = (J_1, \ldots, J_d)$*, is*

$$\mathcal{L}(J) = \left\{ \mathcal{B}(\mathcal{T}) : \mathcal{T} \in \mathbb{T}([0,1]^d, J) \right\}. \tag{16}$$

**Collection of tree weights.** We define a series estimator whose terms are a subset of a pre-basis $\mathcal{B}(\mathcal{T})$. The series estimator is defined with the help of 0-1-weights which choose a subset of the pre-basis. In order the series estimator to be equivalent with a dyadic histogram we need to make a restriction to the weights of the series estimator. The pre-basis $\mathcal{B}(\mathcal{T})$ is associated with tree $\mathcal{T}$ and we require that the weights are such that they correspond to a pruning of the associated tree. The *collection of tree-weights* $\mathcal{W}_{tree,J} = \mathcal{W}_{tree,J}(\mathcal{B})$, associated with $\mathcal{B} \in \mathcal{L}(J)$, is the set of vectors $W = (w_\phi)_{\phi \in \mathcal{B}} \in \{0,1\}^{\mathcal{B}}$, which satisfy the condition that a weight can be zero only when all the "ancestor" weights are zero at the coarser resolution levels. Define

$$\mathcal{W}_{tree,J} = \left\{ (w_\phi)_{\phi \in \mathcal{B}} \in \{0,1\}^{\mathcal{B}} : \text{ if } w_\phi = 0 \text{ then } w_{\phi'} = 0 \text{ for all } \phi' \subset \phi \right\}. \tag{17}$$

Here $\phi' \subset \phi$ means for $\phi = \phi_{\mathcal{I}(t)}^{(s(t))}, \phi' = \phi_{\mathcal{I}(t')}^{(s(t'))} \in \mathcal{B}$, that $R_{\mathcal{I}(t')} \subset R_{\mathcal{I}(t)}$, where $R_{jk}$ is defined in (14).

When $\phi' \subset \phi$ then we say that $\phi'$ is a child of $\phi$. The tree condition says that if $w_\phi = 0$, then $w_{\phi'} = 0$ for all children $\phi'$ of $\phi$. Choosing a subset of $\mathcal{B}(\mathcal{T})$ with the help of weights $W \in \mathcal{W}_{tree,J}(\mathcal{B}(\mathcal{T}))$ is equivalent to the pruning of tree $\mathcal{T} \in \mathbb{T}([0,1]^d, J)$.

**Definition of the series estimator.** Analogously to (8) we define a complexity penalized error criterion

$$\mathcal{E}_n(W, \Theta, \mathcal{B}, \alpha) = \gamma_n \left( \tilde{f}(\cdot, W, \Theta, \mathcal{B}) \right) + \alpha \cdot D(W), \tag{18}$$

where $\gamma_n$ is defined in (4), and the complexity penalization is taken to be the number of terms in the expansion:

$$D(W) = \#\{w_\phi : w_\phi = 1\} + 1, \tag{19}$$

where $W = (w_\phi)_{\phi \in \mathcal{B}} \in \{0,1\}^{\mathcal{B}}$. We have added 1 in the definition of $D(W)$ since the function $I_{[0,1]^d}$ is also in the expansion (11). The series estimator $f_{n,\alpha}^*$ is a minimization estimator where we search a best pre-basis $\mathcal{B}_{n,\alpha}^*$ and a best sub-set of $\mathcal{B}_{n,\alpha}^*$ so that the tree condition is satisfied. The coefficients of the expansion are given by the empirical coefficients $\Theta_n(\mathcal{B})$:

$$\Theta_n(\mathcal{B}) = \left( \hat{\theta}_\phi \right)_{\phi \in \mathcal{B}}, \qquad \hat{\theta}_\phi = \frac{1}{n} \sum_{i=1}^{n} \phi(X^i). \tag{20}$$

**Definition 5** (Dyadic series estimator.) *The empirical choice for the basis* $\mathcal{B}$ *and for the coefficient vector* $W$ *is defined by*

$$\left(\mathcal{B}^*_{n,\alpha}, W^*_{n,\alpha}\right) = \text{argmin}_{\mathcal{B}\in\mathcal{L}(J), W\in\mathcal{W}_{tree,J}(\mathcal{B})}\mathcal{E}_n(W, \Theta_n(\mathcal{B}), \mathcal{B}, \alpha). \tag{21}$$

*The* dyadic series estimator *is defined by*

$$f^*_{n,\alpha}(x) = \tilde{f}\left(x, W^*_{n,\alpha}, \Theta_n(\mathcal{B}^*_{n,\alpha}), \mathcal{B}^*_{n,\alpha}\right), \qquad x \in \mathbf{R}^d, \tag{22}$$

*where* $\tilde{f}(\cdot, W, \Theta, \mathcal{B})$ *is defined in (11).*

## 2.3 Equivalence between estimators

We may prove that the a dyadic histogram is equivalent to a series estimator.

**Lemma 1** *We have that* $\hat{f}_{n,\alpha} = f^*_{n,\alpha}$*, where* $\hat{f}_{n,\alpha}$ *is defined in (10) and* $f^*_{n,\alpha}$ *is defined in (22), when the initial rectangle of the dyadic histogram is* $R_0 = [0,1)^d$*.*

A proof of Lemma 1 may be found in the technical report. See also Engel (1994).

## 2.4 Algorithms and computational complexity

Let us discuss algorithms for solving the minimization problem (9). The solution is the partition defining the estimator. One may solve the minimization problem by first building a large multitree which contains all paths leading to partitions, and then pruning the tree.

### 2.4.1 Growing the tree.

First we construct a multitree with a single root node and at most $2d$ children for every node. The root node will correspond to the initial rectangle $R_0$. We have $d$ ways of choosing the splitting direction and each binary split gives two bins. Thus $2d$ children will represent the rectangles resulting from the binary splits in $d$ directions. At most $J_l$ splits will be made in direction $l$, thus the depth of the tree will is $|J|_{max} = \max_{l=1,\ldots,d}|J_l|$. We shall record the number of observations $n_R$ in each bin R, and calculate $-n_R^2/(n^2\text{vol}(R))$, so that we are able to calculate (7) for all partitions. When some bin is empty of observations we shall not split it anymore. The resulting tree will have at most

$$\sum_{i=0}^{|J|_{max}} (2d)^i = O\left((2d)^{|J|_{nax}}\right) \tag{23}$$

nodes. For the choice $J = J_n$ as in (28), there is $O\left(n^{a \log_2(2d)}\right)$ nodes in the tree.

### 2.4.2 Pruning the tree.

To prune the tree we start from the next to the highest level, and travel to the root node one level at a time. For each node we find out whether the split in some of the $d$ directions helps (whether it results to a smaller complexity penalized error criterion). If the split does not help, we shall cut the tree below the node. This is a multivariate version of the Fast algorithm for Dyadic CART given in Donoho (1997). The number of flops required by the algorithm is bounded by the number of nodes of the tree given in (23).

We formulate a lemma which states that the minimization problem may be solved by this bottom-up algorithm.

**Lemma 2** *Let $T$ be the tree grown in Section 2.4.1. Let $t$ be some non-terminal node of $T$ and $t_{il}$, $i = 1, 2$, $l = 1, \ldots, d$, be the children of $t$. Denote with $R_t$ and $R_{il}$, respectively, the rectangles annotated with these nodes. Denote the partition minimizing the complexity penalized error criterion, when we localize to rectangle $R$ which is annotated to a node of $T$, by*

$$\widehat{\mathcal{P}}_{n,\alpha}(R) = \operatorname{argmin}_{\mathcal{P} \in \tilde{\mathbb{P}}(R)} COPERR_n\left(\mathcal{P}, \alpha\right),$$

*where $\tilde{\mathbb{P}}(R)$ is the set of partitions $\mathbb{P}(R, J')$, defined in Definition 2 and $J' = J - \operatorname{depth}(R)$, where $\operatorname{depth}(R)$ is the vector of the number of splits which has been made in each direction to reach $R$. Let*

$$\begin{aligned} \mathcal{M} \;=\; & \min\left\{COPERR_n\left(\{R_t\}, \alpha\right), \right. \\ & \left. COPERR_n\left(\widehat{\mathcal{P}}_{n,\alpha}(R_{1l}), \alpha\right) + COPERR_n\left(\widehat{\mathcal{P}}_{n,\alpha}(R_{2l}), \alpha\right), l = 1, \ldots d\right\}. \end{aligned}$$

*Then,*

$$\widehat{\mathcal{P}}_{n,\alpha}(R_t) \;=\; \begin{cases} \{R_t\}, \text{ when } \mathcal{M} = COPERR_n\left(\{R_t\}, \alpha\right), \\ \widehat{\mathcal{P}}_{\alpha}(R_{1l}) \cup \widehat{\mathcal{P}}_{\alpha}(R_{2l}), \text{ when} \\ \mathcal{M} = COPERR_n\left(\widehat{\mathcal{P}}_{n,\alpha}(R_{1l}), \alpha\right) + COPERR_n\left(\widehat{\mathcal{P}}_{n,\alpha}(R_{2l}), \alpha\right). \end{cases}$$

*Proof.* When $\mathcal{P}_{il} \in \tilde{\mathbb{P}}(R_{il})$, $i = 1, 2$, $l = 1, \ldots, d$, then

$$COPERR_n\left(\mathcal{P}_{1l} \cup \mathcal{P}_{2l}, \alpha\right) = COPERR_n\left(\mathcal{P}_{1l}, \alpha\right) + COPERR_n\left(\mathcal{P}_{2l}, \alpha\right). \tag{24}$$

Indeed, $\mathcal{P}_{1l}$ and $\mathcal{P}_{2l}$ are partitions of disjoint rectangles and thus (24) follows from (7) and the fact that $\#(\mathcal{P}_{1l} \cup \mathcal{P}_{2l}) = \#\mathcal{P}_{1l} + \#\mathcal{P}_{2l}$. On the other hand

$$\tilde{\mathbb{P}}(R_t) = \{\{R_t\}\} \cup \left\{\mathcal{P}_{1l} \cup \mathcal{P}_{2l} : \mathcal{P}_{il} \in \tilde{\mathbb{P}}(R_{il}), i = 1, 2, l = 1, \ldots, d\right\}.$$

We have proved the lemma. □

**Remark 2** In particular, when we choose $t$ in Lemma 2 to be the root of tree $T$, then $R_t = R_0$ and $\widehat{\mathcal{P}}_{n,\alpha}(R_0) = \widehat{\mathcal{P}}_{n,\alpha}$ is the global solution defined in (9).

# 3 Rates of convergence

We prove that the estimator achieves optimal rates of convergence up to a logarithmic factor over anisotropic Besov classes $B_{sp}(L)$. The parameter $p = (p_1, \ldots, p_d)$ of the Besov ball may be such that $p_l < 2$ for each $l = 1, \ldots, d$. In order to reach optimal rates of convergence over such function classes containing functions with high spatial variability, it is essential that the bin widths may have variable length in any single direction. We denote the intersection of the Besov ball with the set of bounded densities as

$$\mathcal{F} = \mathcal{F}_{sp}(L, B_\infty) = B_{sp}(L) \cap \left\{f : \int_{[0,1]^d} f = 1, 0 \le f \le B_\infty\right\}, \tag{25}$$

where $0 < B_\infty < \infty$, and we define the anisotropic Besov ball $B_{sp}(L)$, where $s = (s_1, \ldots, s_d) \in (0, 1]^d$, $p = (p_1, \ldots, p_d) \in [1, \infty]^d$, $0 < L < \infty$, to be the set of functions $f : [0, 1]^d \to \mathbf{R}$ satisfying for $l = 1, \ldots, d$,

$$\|D_h^l f\|_{L_{p_l}(A_h^l)} \le L h^{s_l},$$

where $0 < h < 1$, $D_h^l f(x) = f(x + he_l) - f(x)$, $e_l \in \mathbf{R}^d$ with the l:th coordinate one and he other coordinates zero, and

$$A_h^l = \{(x_1, \ldots, x_d) : 0 \le x_m \le 1, m \ne l, 0 \le x_l < 1 - h\}. \tag{26}$$

For more on anisotropic Besov spaces, see Nikol'skii (1975).

**The result.** The exponent $r$ of the optimal rate of convergence and the anisotropic smoothness index $\sigma$ are defined by

$$r = \frac{\sigma}{2\sigma + 1}, \qquad \sigma = \left(\sum_{l=1}^{d} s_l^{-1}\right)^{-1}. \tag{27}$$

Besides the smoothing parameter $\alpha$, the estimator depends on the vector of maximal directionwise split numbers $J$ and we take

$$J_n = (J_{n,1}, \ldots, J_{n,d}), \qquad J_{n,l} = \left\lceil \frac{\sigma}{s_l}\, a \log_2 n \right\rceil, \tag{28}$$

where $a \geq 0$ is the fineness parameter. The initial rectangle of the dyadic histogram is $R_0 = [0,1]^d$.

**Theorem 1** *Let $X^1, \ldots, X^n$ be i.i.d. observations from the distribution of density $f \in \mathcal{F}$. When $s_l$, $p_l$, and the fineness parameter $a$ in (28) are such that*

$$\sigma - (1/p_l - 1/2)_+ > 0, \qquad l = 1, \ldots, d \tag{29}$$

*and*

$$\frac{\sigma}{2\sigma + 1} \frac{1}{\sigma - (1/p_l - 1/2)_+} < a < 1, \qquad l = 1, \ldots, d, \tag{30}$$

*then*

$$\limsup_{n\to\infty} \left( \frac{n}{\log_e n} \right)^{2r} \sup_{f\in\mathcal{F}} E_f \int_{[0,1]^d} \left( f - \hat{f}_{n,\alpha_n} \right)^2 < \infty,$$

*where $\hat{f}_{n,\alpha}$ is defined in (10),*

$$\alpha_n = C B_\infty \frac{\log_e n}{n}, \tag{31}$$

*and $C > 0$ is a sufficiently large constant.*

A proof of Theorem 1 is given in Section 3.1.

**Remark 3** (Adaptiveness of the estimator.) The choice of penalization parameter $\alpha$ in Theorem 1 does not depend on the smoothness parameters $s_1, \ldots, s_d$, nor on $p_1, \ldots, p_d$, or $L$. Vector $J$ depends on $s_1, \ldots, s_d$ and on fineness parameter $a$. The lower bound for $a$ depends on the parameters $s_l$ and $p_l$, but we may take parameter $a$ arbitrarily close to 1.

**Remark 4** (The fineness parameter and restrictions on the smoothness.) Because $s_i \leq 1$ we have $\sigma \leq 1/d$. By (29), $\sigma > (1/p_l - 1/2)_+$. Thus, Theorem 1 holds only for $\sigma$ satisfying

$$\max_{l=1,\ldots,d} \left( \frac{1}{p_l} - \frac{1}{2} \right)_+ < \sigma \leq \frac{1}{d}.$$

Thus, for large values of $d$, parameters $p_l$ cannot be much smaller than 2. When $\min_{l=1,\ldots,d} p_l \geq 2$, Theorem 1 holds for $0 < \sigma \leq 1/d$. Figure 1 shows the

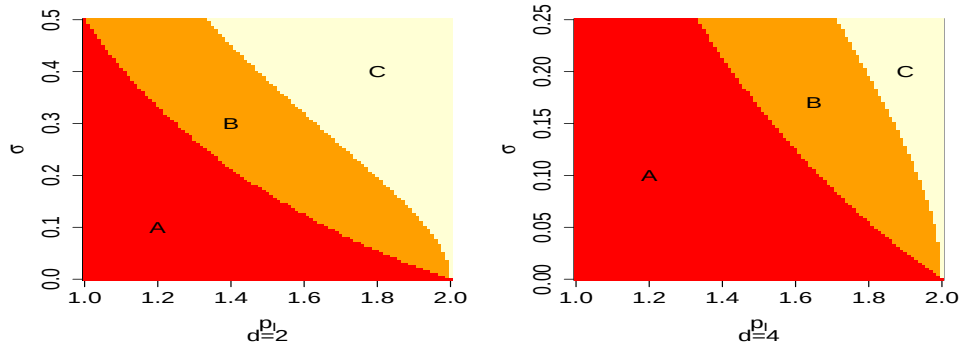


Figure 1: The admissible range $C$ of $\sigma$ and $p_l$ for a) $d = 2$ and b) $d = 4$.

possible values of $(p_l, \sigma) \in [1, 2] \times [0, d^{-1}]$ for a) $d = 2$ and b) $d = 4$. Region $A$ is the region where condition (29) is violated, so that $\sigma \leq (1/p_l - 1/2)_+$. Region $B$ is the region where condition (30) may not be satisfied because $c' = \frac{\sigma}{2\sigma+1} \frac{1}{\sigma-(1/p_l-1/2)_+} > 1$. Region $C$ is the region where condition (29) is satisfied and $c' \leq 1$.

## 3.1 Proof of Theorem 1

Since dyadic histograms are equivalent to dyadic series estimators, as proved in Lemma 1, it is enough prove the theorem for the series estimator. We go through the steps of the proof and after that we give details for the proofs of step 2 and step 4. Details for step 1 are given in Appendix A and step 3 is proved in the technical report.

**Step 1.** *(Application of an oracle inequality.)* The first step is to bound the MISE of the estimator by a minimal complexity penalized approximation error. We have for series estimator $f^*_{n,\alpha_n}$, defined in (22), when $\alpha_n$ is defined in (31) and $J_n$ is defined in (28) with fineness parameter $0 < a < 1$, for continuous densities $f : [0, 1]^d \to \mathbf{R}$, that

$$E_f \left\| f^*_{n,\alpha_n} - f \right\|_2^2 \leq C_1 \min_{(W,\Theta,\mathcal{B})\in\mathbb{K}_0} K(f, W, \Theta, \mathcal{B}, \alpha_n) + C_2 n^{-1}, \tag{32}$$

where $C_1$ and $C_2$ are positive constants,

$$K(f, W, \Theta, \mathcal{B}, \alpha) = \left\| \tilde{f}(\cdot, W, \Theta, \mathcal{B}) - f \right\|_2^2 + \alpha \cdot D(W), \tag{33}$$

$\tilde{f}$ is defined in (11), $\alpha \geq 0$, $D(W)$ is defined in (19) ,

$$\mathbb{K}_0 = \left\{ (W, \Theta, \mathcal{B}) \in \mathcal{W}(\mathcal{B}) \times \mathbf{R}^{\mathcal{B}} \times \mathcal{L} : \|\tilde{f}(\cdot, W, \Theta, \mathcal{B})\|_\infty \leq 2B_\infty \right\}, \quad (34)$$

where $\mathcal{W}(\mathcal{B}) = \mathcal{W}_{tree,J_n}(\mathcal{B})$ as defined in (17), $\mathcal{L} = \mathcal{L}(J_n)$ is defined in (16), and $B_\infty > \|f\|_\infty$ is a positive constant. Eq. (32) is proved in Appendix A.

**Step 2.** *(Choosing a basis.)* We bound the approximation error by finding a pre-basis $\mathcal{B}^*_{\alpha_n} \in \mathcal{L}(J_n)$, which is in a sense the best pre-basis for $f \in B_{sp}(L)$. After fixing the pre-basis to be $\mathcal{B}^*_{\alpha_n}$, we choose the vector of coefficients to be the coefficients of $f$ in the pre-basis $\mathcal{B}^*_{\alpha_n}$; $\Theta = \Theta_f(\mathcal{B}^*_{\alpha_n})$, where

$$\Theta_f(\mathcal{B}) = \left( \int_{\mathbf{R}^d} f\phi \right)_{\phi \in \mathcal{B}}. \quad (35)$$

We have the upper bound

$$\min_{(W,\Theta,\mathcal{B}) \in \mathbb{K}_0} K(f, W, \Theta, \mathcal{B}, \alpha_n) \leq \min_{W \in \mathcal{W}(\mathcal{B}^*_{\alpha_n})} K(f, W, \Theta_f(\mathcal{B}^*_{\alpha_n}), \mathcal{B}^*_{\alpha_n}, \alpha_n),$$

where $\mathcal{W}(\mathcal{B}^*_{\alpha_n}) = \mathcal{W}_{tree,J_n}(\mathcal{B}^*_{\alpha_n})$ is defined in (17).

**Step 3.** The minimization is restricted to the tree weights. One may show that this restriction does not increase much the complexity penalized approximation error: we do not get much better approximation by minimizing the weights over a larger collection of weights. When (29) holds,

$$\begin{aligned} &\sup_{f \in B_{sp}(L)} \min_{W \in \mathcal{W}_{tree,J_n}(\mathcal{B}^*_{\alpha_n})} K(f, W, \Theta_f(\mathcal{B}^*_{\alpha_n}), \mathcal{B}^*_{\alpha_n}, \alpha_n) \\ &\quad \leq C \sup_{f \in B_{sp}(L)} \min_{W \in \{0,1\}^{\mathcal{B}^*_{\alpha_n}}} K(f, W, \Theta_f(\mathcal{B}^*_{\alpha_n}), \mathcal{B}^*_{\alpha_n}, \alpha_n), \end{aligned} \quad (36)$$

for a positive constant $C$, depending on $s, p, L, d$. A proof of (36) may be found in the technical report. See also Donoho (1997).

**Step 4.** The last step is to bound the complexity penalized approximation error in (36). We have that

$$\sup_{f \in B_{sp}(L)} \min_{W \in \{0,1\}^{\mathcal{B}^*_{\alpha_n}}} K(f, W, \Theta_f(\mathcal{B}^*_{\alpha_n}), \mathcal{B}^*_{\alpha_n}, \alpha_n) \leq C\alpha_n^{2r}. \quad (37)$$

This proves the theorem by the choice of $\alpha_n$ in (31).

### 3.1.1 Step 2

We define a best basis for the approximation of functions in $B_{sp}(L)$. Let $h : \{1, 2, \ldots\} \to \{1, \ldots, d\}$. We apply function $h$ as a *direction selection rule.* We define trees and multi-indeces with the help of a direction selection rule:

- Every tree in $\mathbb{T}([0,1]^d, J)$, defined in Definition 1, is uniquely determined by the splitting directions. Let $\mathcal{T}_{h,M}$ be the partition generating tree determined by the following rules.

  1. Split the root node in direction $h(1)$.
  2. The $2^m$ nodes at depth $m$ are splitted in direction $m+1$, for $m \in \{0, \ldots, M-1\}$.

- For any direction selection rule $h$ we define the corresponding sequence of multi-indeces $\mathcal{J} = \mathcal{J}_h : \{0, 1, \ldots\} \to \{0, 1, \ldots\}^d$, by $\mathcal{J} = (j_1, \ldots, j_d)$, where $j_l(m)$ is the number of times direction $l$ was chosen by $h$ up to step $m$: $j_l(0) = 0$ and

$$j_l(m) = \#\{m' \leq m : h(m') = l\}, \qquad m = 1, 2, \ldots, \tag{38}$$

  $l = 1, \ldots, d$.

*(Definition of $h^*$.)* We define a direction selection rule $h^*$, which depends on the vector of smoothness indeces $s = (s_1, \ldots, s_d)$ of the anisotropic Besov space $B_{sp}(L)$. Define the sequence $\mathcal{Z}(m) = \mathcal{Z}_s(m) = (z_1(m), \ldots, z_d(m)) \in [0, \infty)^d$, $m = 0, 1, \ldots$, satisfying $\mathcal{Z}(0) = (0, \ldots, 0)$, and

$$\begin{cases} z_1(m)s_1 = \cdots = z_d(m)s_d \\ z_1(m) + \cdots + z_d(m) = m. \end{cases} \tag{39}$$

That is, $\mathcal{Z}(1) \in \{x \in [0, \infty)^d : x_1 s_1 = \cdots = x_d s_d\}$ is such that $\sum_{l=1}^d z_l(1) = 1$, and $\mathcal{Z}(m) = m\mathcal{Z}(1)$ for $m \geq 1$ integer. We define $h^*$ so that $\mathcal{J}^* = \mathcal{J}_{h^*}$ is an approximation to $\mathcal{Z}_s$, taking values on a grid. The direction selection rule $h^*$ is defined by the following rules.

1. Choose $h^*(1) = \text{argmax}_{l \in \{1,\ldots,d\}} z_l(1)$, that is, $h^*(1) = \text{argmin}_{l \in \{1,\ldots,d\}} s_l$.
2. Write $\mathcal{J}^*(m) = (j_1^*(m), \ldots, j_d^*(m))$. Define for $m = 1, 2, \ldots$,

$$h^*(m+1) = \text{argmax}_{l \in \{1,\ldots,d\}} z_l(m) - j_l^*(m).$$

   That is, we choose the direction where $\mathcal{J}^*(m)$ is furthest below from $\mathcal{Z}(m)$.

By the definition, $z_l(m)s_l = m\sigma$, for $l = 1, \ldots, d$. Thus,

$$j_l^*(m)s_l \sim m\sigma \tag{40}$$

as $m \to \infty$ where $\sigma$ is defined in (27). This means that the proportion in which direction $l$ was chosen, $j_l^*(m)/m$, is approximately equal to $\sigma/s_l$.

*(Definition of the pre-basis.)* We choose

$$M_{\alpha_n}^* = \left[a \log \alpha_n^{-1}\right].$$

We have, see Eq. (40),

$$j_l^*(M_{\alpha_n}^*) \le J_{n,l}, \qquad l = 1, \ldots, d, \tag{41}$$

where $J_n = (J_{n,1}, \ldots, J_{n,d})$ is defined in (28). Thus

$$\mathcal{T}_{h^*, M_{\alpha_n}^*} \in \mathbb{T}([0,1]^d, J_n). \tag{42}$$

Define the best pre-basis

$$\mathcal{B}_{\alpha_n}^* = \mathcal{B}(\mathcal{T}_{h^*, M_{\alpha_n}^*}), \tag{43}$$

where $\mathcal{B}(\mathcal{T})$ is defined in (15). Eq. (42) implies that

$$\mathcal{B}_{\alpha_n}^* \in \mathcal{L}(J_n)$$

where $\mathcal{L}(J)$ is defined in (16). Define the basis

$$\mathcal{B}^* = \left\{I_{[0,1]^d}\right\} \cup \mathcal{B}\left(\mathcal{T}_{h^*,\infty}\right). \tag{44}$$

This is a spatially homogeneous anisotropic basis. A proof that $\mathcal{B}^*$ is a basis of $L_2([0,1]^d)$ may be found in the technical report.

### 3.1.2 Step 4

**Largeness of the wavelet coefficients in basis $\mathcal{B}^*$.** We need a bound to the coefficients $\int_{[0,1]^d} f\phi$, $\phi \in \mathcal{B}^*$. To give the bound it is convenient to write the basis in terms of Haar-basis functions. We have that

$$\mathcal{B}^* = \bigcup_{m=0}^{\infty} \left\{\phi_{\mathcal{J}^*(m),k}^{(h^*(m+1))} : k \in K_{\mathcal{J}^*(m)}\right\},$$

where $\phi_{j,k}^{(l)}$ are defined in (12) and $K_j$ is defined in (13). We denote the coefficients of $f$ by

$$\tau_{mk} = \int_{[0,1]^d} f\phi_{\mathcal{J}^*(m),k}^{(h^*(m+1))}, \tag{45}$$

where $m = 0, 1, \ldots$ and $k \in K_{\mathcal{J}^*(m)}$.

**Lemma 3** *Let $f \in B_{sp}(L)$, $s = (s_1, \ldots, s_d) \in (0,1]^d$, $p = (p_1, \ldots, p_d) \in [1,\infty]^d$. Then we have for $m = 0, 1, \ldots$, that*

$$\left( \sum_{k \in K_{\mathcal{J}^*(m)}} |\tau_{mk}|^{\tilde{p}_{l^*_m}} \right)^{1/\tilde{p}_{l^*_m}} \leq 2^{d/2} L 2^{-m(\sigma+1/2-1/\tilde{p}_{l^*_m})}, \tag{46}$$

*where we use the notation $l^*_m = h^*(m+1)$ and*

$$\tilde{p}_l = \min\{p_l, 2\}, \qquad l = 1, \ldots, d, \tag{47}$$

*and $\sigma$ is defined in (27).*

A proof of Lemma 3 may be found in the technical report.

**Final lemma.** Eq. (37) follows from Lemma 4 below.

**Lemma 4** *Let (29) be satisfied, let $\tilde{a}$ satisfy*

$$\tilde{a} > \frac{\sigma}{2\sigma+1} \frac{1}{\sigma - (1/p_l - 1/2)_+}, \qquad l = 1, \ldots, d, \tag{48}$$

*where $\sigma$ is defined in (27). Let $M = M_\alpha$ be an integer satisfying*

$$M_\alpha \geq \tilde{a} \log_2 \alpha^{-1} \tag{49}$$

*and let $\mathcal{B}^*_\alpha = \mathcal{B}(\mathcal{T}_{h^*, M_\alpha})$. Then*

$$\sup_{f \in B_{sp}(L)} \min_{W \in \{0,1\}^{\mathcal{B}^*_\alpha}} K\left(f, W, \Theta_f(\mathcal{B}^*_\alpha), \mathcal{B}^*_\alpha, \alpha\right) \leq C \alpha^{2r}, \tag{50}$$

*for a positive constant $C$, depending on $s, p, L, d$, when $0 < \alpha < 1$ is sufficiently small, where $r = \sigma/(2\sigma+1)$ is defined in (27).*

*Proof.* As before, we denote $\mathcal{J}^* = \mathcal{J}_{h^*}$. We prove that

$$\sum_{m=0}^{M} \sum_{k \in K_{\mathcal{J}^*(m)}} \min\left\{\tau^2_{mk}, \alpha\right\} \leq \left(2 + \frac{2C_{p,L,d}}{2^{\beta_*} - 1}\right) \cdot \alpha^{2\sigma/(2\sigma+1)} \tag{51}$$

and

$$\sum_{m=M+1}^{\infty} \sum_{k \in K_{\mathcal{J}^*(m)}} \tau^2_{mk} \leq \frac{2^d L^2}{2^{2\beta_*} - 1} \cdot \alpha^{2\sigma/(2\sigma+1)}, \tag{52}$$

when $0 < \alpha < 1$ is sufficiently small, where $\tau_{mk}$ is defined in (45), $C_{p,L,d} = \max_{l=1,\ldots,d}(2^{d/2} L)^{\tilde{p}_l}$, and $\beta_* = \min_{l=1,\ldots,d}(\sigma + 1/2 - 1/\tilde{p}_l) = \sigma + 1/2 - 1/p_*$, $p_* = \min_{l=1,\ldots,d} \tilde{p}_l$. This implies the lemma, when we apply Lemma 5 with $\mathcal{B} = \mathcal{B}^*_\alpha$ and with the basis $\mathcal{B}_\infty = \mathcal{B}^*$.

**Proof of (51).** Let $m^* \geq 1$ be defined by

$$m^* = \left\lceil \frac{1}{2\sigma+1} \log_2 \alpha^{-1} \right\rceil. \tag{53}$$

Note that $m^* < M$ since $\tilde{a} > 1/(2\sigma+1)$ by the lower bound in (48). Write

$$\sum_{m=0}^{M} \sum_{k \in K_{\mathcal{J}^*(m)}} \min\left\{\tau_{mk}^2, \alpha\right\} \leq A + B, \tag{54}$$

where

$$\begin{aligned} A &\stackrel{def}{=} \sum_{m=0}^{m^*} \sum_{k \in K_{\mathcal{J}^*(m)}} \alpha = \alpha \sum_{m=0}^{m^*} 2^m = \alpha(2^{m^*+1} - 2) \\ &\leq 2\alpha\alpha^{-1/(2\sigma+1)} = 2\alpha^{2\sigma/(2\sigma+1)} \end{aligned} \tag{55}$$

by the definition of $m^*$ in (53), and

$$\begin{aligned} B &\stackrel{def}{=} \sum_{m=m^*+1}^{M} \sum_{k \in K_{\mathcal{J}^*(m)}} \min\left\{\tau_{mk}^2, \alpha\right\} \\ &\leq \sum_{m=m^*+1}^{M} \alpha^{1-\tilde{p}_{l_m^*}/2} \sum_{k \in K_{\mathcal{J}^*(m)}} |\tau_{mk}|^{\tilde{p}_{l_m^*}}, \end{aligned} \tag{56}$$

where $\tilde{p}_l$ is defined in (47), and we use the notation $l_m^* = h^*(m+1)$. Above we used the fact $\min\{\tau_{mk}^2, \alpha\} \leq \alpha^{1-\tilde{p}_l/2}|\tau_{mk}|^{\tilde{p}_l}$, for $l = 1, \ldots, d$. Indeed, we have that when $\tau_{mk}^2 \leq \alpha$, then $\alpha^{1-\tilde{p}_l/2}|\tau_{mk}|^{\tilde{p}_l} \geq \tau_{mk}^2$ and when $\tau_{mk}^2 > \alpha$, then $\alpha^{1-\tilde{p}_l/2}|\tau_{mk}|^{\tilde{p}_l} \geq \alpha$. We have from Lemma 3 that

$$\sum_{k \in K_{\mathcal{J}^*(m)}} |\tau_{mk}|^{\tilde{p}_{l_m^*}} \leq \left(2^{d/2} L 2^{-m\beta_m}\right)^{\tilde{p}_{l_m^*}}, \qquad \beta_m = \sigma + 1/2 - 1/\tilde{p}_{l_m^*}. \tag{57}$$

Continuing from (56),

$$\begin{aligned} B &\leq C_{p,L,d} \sum_{m=m^*+1}^{M} \alpha^{1-\tilde{p}_{l_m^*}/2} 2^{-\tilde{p}_{l_m^*} m\beta_m} && (58) \\ &= C_{p,L,d} \sum_{m=m^*+1}^{M} \alpha^{1-\tilde{p}_{l_m^*}/2} 2^{-\tilde{p}_{l_m^*} m^*\beta_m} 2^{\tilde{p}_{l_m^*}(m^*-m)\beta_m} \\ &\leq \frac{C_{p,L,d}}{2^{\beta_*} - 1} \cdot \alpha 2^{m^*} && (59) \\ &\leq \frac{2C_{p,L,d}}{2^{\beta_*} - 1} \cdot \alpha^{2\sigma/(2\sigma+1)}, && (60) \end{aligned}$$

where in (58) we applied (57). In (59) we applied

$$\alpha^{-\tilde{p}_{l_m^*}/2} 2^{-p_{l_m^*} m^*(\sigma+1/2)} \leq 1$$

which holds due to the choice of $m^*$ in (53), and we applied in (59) also the fact

$$\sum_{m=m^*+1}^{\infty} 2^{\tilde{p}_{l_m^*}(m^*-m)\beta_m} \leq \sum_{m=1}^{\infty} (2^{-\beta_*})^m = \frac{1}{2^{\beta_*}-1}$$

which holds because $\tilde{p}_{l_m^*} \geq 1$, because $\sum_{m=1}^{\infty} r^m = r/(1-r)$ for $0 < r < 1$, and because $\beta_* > 0$ which is assumed in (29). The claim (51) follows from (54), (55), and (60).

**Proof of (52).** We have that

$$\begin{aligned}
\sum_{m=M+1}^{\infty} \sum_{k \in K_{\mathcal{J}^*(m)}} \tau_{mk}^2 &\leq \sum_{m=M+1}^{\infty} \left( \sum_{k \in K_{\mathcal{J}^*(m)}} |\tau_{mk}|^{\tilde{p}_{l_m^*}} \right)^{2/\tilde{p}_{l_m^*}} && (61)\\
&\leq 2^d L^2 \sum_{m=M+1}^{\infty} 2^{-2m\beta_m} && (62)\\
&\leq \frac{2^d L^2}{2^{2\beta_*}-1} \cdot 2^{-2\beta_* M} && (63)\\
&\leq \frac{2^d L^2}{2^{2\beta_*}-1} \cdot \alpha^{2\beta_* \tilde{a}} && (64)\\
&\leq \frac{2^d L^2}{2^{2\beta_*}-1} \cdot \alpha^{2\sigma/(2\sigma+1)}, && (65)
\end{aligned}$$

where in (61) we applied the subadditivity of the function $x \mapsto x^{\tilde{p}_l/2}$, in (62) we applied (57), in (63) we applied that for $0 < r < 1$, $\sum_{m=M+1}^{\infty} r^m = r^{M+1}/(1-r)$ and the fact that $\beta_* > 0$, in (64) we applied the choice of $M$ in (49), and in (65) we applied the lower bound for $\tilde{a}$ in (48). We have proved Lemma 4. □

# Acknowledgments

Writing of this article was financed by Deutsche Forschungsgemeinschaft under project MA1026/8-2.

# References


Barron, A., Birgé, L. and Massart, P. (1999), 'Risk bounds for model selection via penalization', *Probab. Theory Relat. Fields* **113**, 301–413.

Bousquet, O. (2002), 'A Bennett concentration inequality and its application to suprema of empirical processes', *C. R. Acad. Sci. Paris, Ser. I* **334**, 495–500.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. J. (1984), *Classification and Regression Trees*, Chapman & Hall, New York.

Castellan, G. (2000), 'Sélection d'histogrammes à l'aide d'un critère de type Akaike', *C. R. Acad. Sci., Paris, Sér. I, Math.* **330**(8), 729–732.

Donoho, D. L. (1997), 'Cart and best-ortho-basis: A connection.', *Ann. Statist.* **25**, 1870–1911.

Engel, J. (1994), 'A simple wavelet approach to nonparametric regression from recursive partitioning schemes', *J. Multivariate Anal.* **49**, 242–254.

Hüsemann, J. A. and Terrell, G. R. (1991), 'Optimal parameter choice for error minimization in bivariate histograms', *J. Multivariate Anal.* **37**, 85–103.

Kerkyacharian, G., Lepski, O. and Picard, D. (2001), 'Nonlinear estimation in anisotropic multi-index denoising', *Probab. Theory Relat. Fields* **121**, 137–170.

Lugosi, G. and Nobel, A. (1996), 'Consistency of data-driven histogram methods for density estimation and classification', *Ann. Statist.* **24**, 687–706.

Neumann, M. H. (2000), 'Multivariate wavelet thresholding in anisotropic function spaces', *Stat. Sin.* **10**, 399–431.

Nikol'skii, S. M. (1975), *Approximation of Functions of Several Variables and Imbedding Theorems.*, Springer-Verlag, Berlin.

Ooi, H. (2002), 'Density visualization and mode hunting using trees', *J. Comput. Graph. Statist.* **11**, 328–347.

Shang, N. (1994), Tree-structured density estimation and dimensionality reduction, *in* 'Proceedings of the 26rd Symposium on the Interface', pp. 172–176.

Sutton, C. D. (1994), 'Tree structured density estimation', *Computing Science and Statistics* **26**, 167–171.

Tribouley, K. (1995), 'Practical estimation of multivariate densities using wavelet methods', *Statist. Neerlandica* **49**, 41–62.

# A Oracle inequality

## A.1 General setting

We will state an oracle inequality in a general setting, in order to simplify the notation and exposition. We denote

$$\hat{f}_{n,\alpha}(x) = \tilde{f}\left(x, \hat{\Lambda}_{n,\alpha}\right), \qquad x \in \mathbf{R}^d, \tag{66}$$

where

$$\tilde{f}(x, \Lambda) = \sum_{\phi \in \mathcal{D}} \lambda_\phi \phi(x), \qquad x \in \mathbf{R}^d, \tag{67}$$

where $\mathcal{D} \subset L_2([0,1]^d)$ is a collection of functions, we will assume that $\mathcal{D}$ has finite cardinality, and $\Lambda = (\lambda_\phi)_{\phi \in \mathcal{D}} \in \mathbf{R}^{\mathcal{D}}$ gives the coefficients of the expansion,

$$\hat{\Lambda}_{n,\alpha} = \mathrm{argmin}_{\Lambda \in \mathbb{K}} \mathcal{E}_n(\Lambda, \alpha), \tag{68}$$

$$\mathcal{E}_n(\Lambda, \alpha) = \gamma_n\left(\tilde{f}(\cdot, \Lambda)\right) + \alpha \cdot D(\Lambda),$$

where $\gamma_n$ is defined in (4), $\alpha \geq 0$,

$$D(\Lambda) = \#\{\lambda_\phi : \lambda_\phi \neq 0\}, \tag{69}$$

and $\mathbb{K} \subset \mathbf{R}^{\mathcal{D}}$.

**Theorem 2** *We have for the estimator $\hat{f}_{n,\alpha}$ defined in (66), based on i.i.d. observations $X^1, \ldots, X^n$ from the distribution of a continuous density $f : [0,1]^d \to \mathbf{R}$, that*

$$E_f \left\| \hat{f}_{n,\alpha_n} - f \right\|_2^2 1_{\tilde{\Omega}} \leq C_1 \inf_{\Lambda \in \mathbb{K}_0} K\left(f, \Lambda, \alpha_n\right) + C_2 n^{-1}, \tag{70}$$

*where*

$$K(f, \Lambda, \alpha) = \left\| \tilde{f}(\,\cdot\,, \Lambda) - f \right\|_2^2 + \alpha \cdot D(\Lambda), \tag{71}$$

$$\mathbb{K}_0 = \left\{ \Lambda \in \mathbb{K} : \|\tilde{f}(\cdot, \Lambda)\|_\infty \leq 2B_\infty \right\}, \tag{72}$$

*where* $B_\infty > \|f\|_\infty$ *is a positive constant.*

$$\alpha_n = C_L \, B_\infty \, \frac{\log_e(\#\mathcal{D})}{n}\,, \tag{73}$$

*where* $C_L$, $C_1$, $C_2$ *are positive constants, and* $1_{\tilde{\Omega}}$ *is the indicator of the event*

$$\tilde{\Omega} = \left( \|\hat{f}_{n,\alpha_n}\|_\infty \leq 2B_\infty \right). \tag{74}$$

## A.2 Proof of Theorem 2

Denote $\hat{f} = \hat{f}_{n,\alpha_n}$ and $\hat{\Lambda} = \hat{\Lambda}_{n,\alpha_n}$. We condition on the set $\tilde{\Omega}$ so that $\hat{\Lambda} \in \mathbb{K}_0$. Let $f$ be the true density and let $\Lambda^0 \in \mathbb{K}_0$. Denote

$$\zeta = C_1 K(f, \Lambda^0, \alpha_n),$$

where $C_1$ is a positive constant to be chosen later. We have that

$$\begin{aligned}
& E\|\hat{f} - f\|_2^2 \\
&= \int_0^\infty P\left( \|\hat{f} - f\|_2^2 > t \right) dt \\
&\leq \zeta + \int_\zeta^\infty P\left( \|\hat{f} - f\|_2^2 > t \right) dt \\
&= \zeta + C_2 n^{-1} \int_0^\infty P\left( \|\hat{f} - f\|_2^2 > C_2 n^{-1} t + \zeta \right) dt,
\end{aligned} \tag{75}$$

where $C_2$ is a positive constant to be chosen later. Let $a > 0$. Now

$$\begin{aligned}
A &\stackrel{def}{=} \left( \|\hat{f} - f\|_2^2 > C_2 n^{-1} t + \zeta \right) \\
&= \Big( (a+1)\|\hat{f} - f\|_2^2 \\
&\qquad > a\|\hat{f} - f\|_2^2 + C_1 K(f, \Lambda^0, \alpha_n) + C_2 n^{-1} t \Big).
\end{aligned} \tag{76}$$

Lemma 6 implies that the theoretical error-complexity of the minimization estimator may be bounded by the theoretical error-complexity of $f^0 = \tilde{f}(\cdot, \Lambda^0)$, with the additional empirical term:

$$\begin{aligned}
& K(f, \hat{\Lambda}, \alpha_n) \leq K(f, \Lambda^0, \alpha_n) + 2\nu_n \left( \hat{f} - f^0 \right) \\
& \Leftrightarrow \quad \|\hat{f} - f\|_2^2 \leq K(f, \Lambda^0, \alpha_n) - \alpha_n D(\hat{\Lambda}) + 2\nu_n \left( \hat{f} - f^0 \right).
\end{aligned}$$

Thus we may continue (76) with

$$\begin{aligned} A &\subset \left(2\nu_n\left(\hat{f}-f^0\right) > \frac{a}{a+1}\,\|\hat{f}-f\|_2^2 + \alpha_n D(\hat{\Lambda})\right. \\ &\quad \left. + \left(\frac{C_1}{a+1} - 1\right) K(f,\Lambda^0,\alpha_n) + \frac{C_2}{a+1}\, n^{-1}t\right) \\ &\subset \left(\nu_n\left(\hat{f}-f^0\right) > w(\hat{\Lambda})\xi\right) \\ &\subset \left(\sup_{\Lambda\in\mathbb{K}_0} \frac{\nu_n(\tilde{f}(\cdot,\Lambda)) - \nu_n(f^0)}{w(\Lambda)} > \xi\right) \\ &\stackrel{def}{=} B, \end{aligned}$$

where we used the fact that on $\tilde{\Omega}$, $\hat{\Lambda} \in \mathbb{K}_0$, and we denote

$$\xi = \frac{1}{2}\,\min\left\{\frac{C_1}{a+1} - 1,\, \frac{a}{a+1}\right\},$$

$$\begin{aligned} w(\Lambda) &= \left\|\tilde{f}(\cdot,\Lambda) - f\right\|_2^2 + \left\|f - f^0\right\|_2^2 + \frac{\tau(\Lambda)}{2n}\,, \\ \tau(\Lambda) &= C_\tau\left[n\alpha_n\left(C_{\tau,1}D(\Lambda^0) + C_{\tau,2}D(\Lambda)\right) + t\right], \qquad (77) \\ C_\tau &= \frac{1}{\xi}\,\frac{C_2}{a+1}\,,\ \ C_{\tau,1} = \frac{C_1 - a - 1}{C_2}\,,\ \ C_{\tau,2} = \frac{a+1}{C_2}\,. \end{aligned}$$

We need to choose $C_1$, $C_2$, and $a$ so that $2C_\tau^{-1} \le \xi^2$. This inequality will be needed in (86). This choice is possible: we take $2(a+1)/C_2 \le \xi$. We need also $C_1/(a+1) - 1 > 0$ to guarantee $\xi > 0$. We have

$$P(A) \le P(B). \qquad (78)$$

We prove that

$$P(B) \le C\exp\{-t(C_L B_\infty)^{-1}\}, \qquad (79)$$

where $C$ and $C_L$ are positive constants. This proves the theorem, when we combine (75) and (78).

**Proof of (79).** For $\Phi \subset \mathcal{D}$, let $\mathbb{K}_{0,\Phi}$ be the set of coefficients in $\mathbb{K}_0$ which are non-zero exactly at the positions given by set $\Phi$:

$$\mathbb{K}_{0,\Phi} = \{\Lambda \in \mathbb{K}_0 : \lambda_\phi \neq 0 \text{ if and only if } \phi \in \Phi\}\,,$$

where we use again the notation $\Lambda = (\lambda_\phi)_{\phi\in\mathcal{D}}$. Let for $l = 1, 2, \ldots,$

$$\mathbb{D}_l = \{\Phi \subset \mathcal{D} : \#\Phi = l\}$$

be the set of subsets of $\mathcal{D}$ of cardinality $l$. We may write

$$\mathbb{K}_0 = \bigcup_{l=1}^{\infty} \bigcup_{\Phi \in \mathbb{D}_l} \mathbb{K}_{0,\Phi}.$$

That is, we make a countable partition of $\mathbb{K}_0$ and each member of the partition is the set of vectors $\Lambda$ which have exactly $l$ non-zero elements. We have that

$$B \subset \bigcup_{l=1}^{\infty} \bigcup_{\Phi \in \mathbb{D}_l} B_\Phi,$$

where

$$B_\Phi = \left( \sup_{\Lambda \in \mathbb{K}_{0,\Phi}} \frac{\nu_n(\tilde{f}(\cdot,\Lambda)) - \nu_n(f^0)}{w(\Lambda)} > \xi \right).$$

For $\Phi \in \mathbb{D}_l$,

$$P\left(B_\Phi\right) \leq 2\exp\{-(C_L B_\infty)^{-1}(t + lL)\}, \tag{80}$$

where $C_L$ is a positive constant defined in (90) and

$$L = C_L B_\infty \log_e(\#\mathcal{D}).$$

We prove (80) below. We have that

$$\#\,\mathbb{D}_l = \left( \begin{array}{c} \#\mathcal{D} \\ l \end{array} \right) \leq \left( \frac{e\#\mathcal{D}}{l} \right)^l. \tag{81}$$

Thus,

$$\begin{aligned} P\left(B\right) &\leq 2\sum_{l=1}^{\infty} \sum_{\Phi \in \mathbb{D}_l} \exp\{-(C_L B_\infty)^{-1}(t + lL)\} \\ &\leq 2\sum_{l=1}^{\infty} \left( \frac{e\#\mathcal{D}}{l} \right)^l \exp\{-(C_L B_\infty)^{-1}(t + lL)\} \\ &\leq C\exp\{-(C_L B_\infty)^{-1} t\}, \end{aligned} \tag{82}$$

by the choice of $L$. We have proved (79) up to proving (80).

**Proof of (80).** Denote

$$Z = \sup_{g \in \mathcal{G}} \nu_n(g),$$

where

$$\mathcal{G} = \mathcal{G}_\Phi = \left\{ \frac{\tilde{f}(\cdot,\Lambda) - f^0}{w(\Lambda)} : \Lambda \in \mathbb{K}_{0,\Phi} \right\}, \qquad \Phi \in \mathbb{D}_l.$$

Denote

$$v_0 = \sup_{g \in \mathcal{G}} \|g\|_2^2 = \sup_{\Lambda \in \mathbb{K}_{0,\Phi}} \frac{\|\tilde{f}(\cdot, \Lambda) - f^0\|_2^2}{w^2(\Lambda)}$$

and

$$\tau = C_\tau \left[ n\alpha_n \left( C_{\tau,1} D(\Lambda^0) + C_{\tau,2}\, l \right) + t \right].$$

We have for $\Phi \in \mathbb{D}_l$, $\Lambda \in \mathbb{K}_{0,\Phi}$, that $\tau(\Lambda) = \tau$, where $\tau(\Lambda)$ is defined in (77). Thus, for $\Phi \in \mathbb{D}_l$ and $\Lambda \in \mathbb{K}_{0,\Phi}$,

$$w(\Lambda) \geq \frac{1}{2} \left( \left\| \tilde{f}(\cdot, \Lambda) - f^0 \right\|_2^2 + \frac{\tau}{n} \right) \geq \left\| \tilde{f}(\cdot, \Lambda) - f^0 \right\|_2 \left( \frac{\tau}{n} \right)^{1/2}. \tag{83}$$

Thus

$$v_0 \leq \frac{n}{\tau}. \tag{84}$$

We have that

$$(EZ)^2 \leq \|f\|_\infty \left( l + D(\Lambda^0) \right) \tau^{-1}. \tag{85}$$

We prove equation (85) below in page 27. Denote

$$\eta^2 = \tau^{-1}(t + C_{\tau,2} L l).$$

Then we have

$$\begin{aligned} (EZ + \eta)^2 &\leq 2 \left[ (EZ)^2 + \eta^2 \right] \\ &\leq 2\tau^{-1} \left[ B_\infty \left( l + D(\Lambda^0) \right) + t + C_{\tau,2} L l \right] \\ &\leq 2C_\tau^{-1} \\ &\leq \xi^2 \end{aligned} \tag{86}$$

since

$$C_\tau^{-1} \tau \geq L \left( C_{\tau,1} D(\Lambda^0) + C_{\tau,2}\, l \right) + t \geq B_\infty D(\Lambda^0) + (B_\infty + C_{\tau,2} L)\, l + t,$$

where we used $n\alpha_n = L$ and $C_{\tau,1} C_L \log_e(\#\mathcal{D}) \geq 1$. Eq. (86) implies that

$$P(B_\Phi) = P\left( Z \geq \xi \right) \leq P\left( Z \geq EZ + \eta \right). \tag{87}$$

Denote

$$b = \sup\{\|g\|_\infty : g \in \mathcal{G}\},$$

and

$$v = \sup\{\mathrm{Var}_f(g(X^1)) : g \in \mathcal{G}\}.$$

By Talagrand's theorem, as given in Bousquet (2002), Theorem 2.3, by applying the inequality $h(t) \geq t^2/(2+2t/3)$, $t > 0$, for $h(t) = (1+t)\log(1+t)-t$, we get

$$P\left(Z \geq EZ + \eta\right) \leq \exp\left\{\frac{-n\eta^2}{2[v + 2bEZ + \eta b/3]}\right\}. \tag{88}$$

We have

$$v \leq \|f\|_\infty v_0 \leq \|f\|_\infty \frac{n}{\tau}.$$

Also, for for $\Phi \in \mathbb{D}_l$ and $\Lambda \in \mathbb{K}_{0,\Phi}$,

$$w(\Lambda) \geq \frac{\tau}{2n} \tag{89}$$

and thus

$$b \leq \frac{2n}{\tau} \sup_{\Lambda \in \mathbb{K}_{0,\Phi}} \|\tilde{f}(\cdot,\Lambda) - f^0\|_\infty \leq \frac{4n}{\tau} B, \qquad B = 2\|f\|_\infty,$$

by the definition of $\mathbb{K}_0$. Thus, applying inequalities $EZ \leq \xi$, $\eta \leq \xi$,

$$\begin{aligned} v + 2bEZ + \eta b/3 &\leq \frac{n}{\tau}\|f\|_\infty (1 + 8 \cdot \xi(2 + 1/3)) \\ &\stackrel{def}{=} \frac{n}{\tau}\|f\|_\infty C_L/2. \end{aligned} \tag{90}$$

Thus

$$P\left(Z \geq EZ + \eta\right) \leq \exp\{-(t + lL)/(B_\infty C_L)\}. \tag{91}$$

Eq. (80) follows from (87) and (91).

**Proof of (85).** Let $\Lambda \in \mathbb{K}_{0,\Phi}$ where $\Phi \in \mathbb{D}_l$. Denote $\mathcal{D}(\Lambda, \Lambda^0) = \{\phi \in \mathcal{D} : \lambda_\phi \neq 0 \text{ or } \lambda^0_\phi \neq 0\}$. Let $\{\psi_1, \ldots, \psi_k\}$ be a basis of the span of $\mathcal{D}(\Lambda, \Lambda^0)$. We have $k \leq \#\mathcal{D}(\Lambda, \Lambda^0) \leq l + D(\Lambda^0)$ and, applying (83),

$$\sup_{\Lambda \in \mathbb{K}_{0,\Phi}} \frac{\|\tilde{f}(\cdot,\Lambda) - f^0\|_2^2}{w^2(\Lambda)} \leq \frac{n}{\tau}.$$

Thus we may apply Lemma 7 with $B_2^2 = n/\tau$ to get (85). □

## A.3 Proof of (32)

The series estimator $f^*_{n,\alpha}$ defined in (22) is equal to the estimator $\hat{f}_{n,\alpha}$ defined in (66), under suitable identifications. Indeed, we note that we can write

$$f^*_{n,\alpha}(x) = \tilde{f}\left(x, \hat{W}_{n,\alpha}, \hat{\Theta}_{n,\alpha}, \hat{\mathcal{B}}_{n,\alpha}\right), \qquad x \in \mathbf{R}^d, \tag{92}$$

where

$$\left(\hat{\mathcal{B}}_{n,\alpha}, \hat{\Theta}_{n,\alpha}, \hat{W}_{n,\alpha}\right) = \operatorname{argmin}_{\mathcal{B}\in\mathcal{L},\Theta\in\mathbf{R}^{\mathcal{B}},W\in\mathcal{W}(\mathcal{B})}\mathcal{E}_n\left(W,\Theta,\mathcal{B},\alpha\right). \tag{93}$$

We have that

$$f^*_{n,\alpha}(x) = \hat{f}_{n,\alpha}(x), \qquad x \in \mathbf{R}^d, \tag{94}$$

when we choose

$$\mathcal{D} = \left\{I_{[0,1]^d}\right\} \cup \bigcup_{\mathcal{B}\in\mathcal{L}} \mathcal{B},$$

and

$$\mathbb{K} \subset \left\{\Lambda \in \mathbf{R}^{\mathcal{D}} : \Lambda \in \mathcal{W}(\mathcal{B}) \text{ for some } \mathcal{B} \in \mathcal{L}\right\}. \tag{95}$$

Definition (95) is explained by the fact that the vectors $\Lambda = (\lambda_\phi)_{\phi\in\mathcal{D}}$ have to be such that components are non-zero only for a single basis $\mathcal{B}$: $\{\phi : \lambda_\phi \neq 0\} \subset \mathcal{B}$ for some $\mathcal{B} \in \mathcal{L}$. Eq. (94) holds since we may use the identification $\lambda_\phi = w_\phi\theta_\phi$.

**Cardinality of the library.** To apply Theorem 2 we have to check that the smoothing parameter in (31) has the same order as the smoothing parameter in (73). This follows becaus the cardinality of the dictionary $\mathcal{D} = \{I_{[0,1]^d}\} \cup \cup_{\mathcal{B}\in\mathcal{L}(J)}\mathcal{B}$ satisfies

$$\#\,\mathcal{D} \le C \cdot n^{a\log_2(2d)} \tag{96}$$

for a positive constant $C$. The calculation of the cardinality of the library is basically the same as the calculation in (23). Indeed, every function in $\mathcal{B}(\mathcal{T})$, defined in (15), may be obtained as a node of a large multitree which has 1 root node, where the number of children of each node is equal to $2d$, and the depth of the tree is equal to $|J_n|_{max} \le \lceil a\log_2 n\rceil$. The number of the nodes of the tree in question is $\sum_{i=0}^{|J_n|_{max}-1}(2d)^i \le (2d)^{|J_n|_{max}}$.

**Final detail.** Theorem 2 involves set $\tilde{\Omega}$ defined in (74). We need to show that the restriction to this set is not essential. We have the bound

$$E_f\left\|f^*_{n,\alpha_n} - f\right\|_2^2 1_{\tilde{\Omega}^c} \le \left(\left\|f^*_{n,\alpha_n}\right\|_\infty + \|f\|_\infty\right)^2 P\left(\tilde{\Omega}^c\right), \tag{97}$$

where we applied the fact that $\|g\|_2^2 \le \|g\|_\infty^2$, when the support of $g$ is contained in $[0,1]^d$. First, for all samples,

$$\|f^*_{n,\alpha_n}\|_\infty \le n^\kappa \tag{98}$$

for some $\kappa > 0$. We may prove (98) by noting that by Lemma 1, $\|f^*_{n,\alpha}\|_\infty = \|\hat{f}_{n,\alpha}\|_\infty$ and $\|\hat{f}_{n,\alpha}\|_\infty \le 2^{|J|}$, since $2^{|J|}$ is the minimal volume of the rectangles

in the partition of histogram $\hat{f}_{n,\alpha}$. Second, we need that for sufficiently large $n$,

$$P\left(\tilde{\Omega}^c\right) \leq \delta_n^* \stackrel{def}{=} n^{\kappa'} \exp\left\{-n^{1-a}\,\frac{3\|f\|_\infty}{8}\right\}, \tag{99}$$

for some $\kappa' > 0$, where $0 < a < 1$ is the fineness parameter in (28). Equation (99) follows from Bernstein's inequality. (Note that also in the proof of (99) we apply Lemma 1.) □

# B Auxiliary lemmas

## B.1 Complexity penalized approximation error

**Lemma 5** *Let $\mathcal{B}_\infty$ be a basis of $L_2([0,1]^d)$ such that $\mathcal{B} \subset \mathcal{B}_\infty$, where $\mathcal{B}$ is an orthonormal system. Then,*

$$\begin{aligned}\min_{W\in\{0,1\}^{\mathcal{B}}} & K\left(f, W, \Theta_f(\mathcal{B}), \mathcal{B}, \alpha\right) \\ = \;& \alpha + \sum_{\phi\in\mathcal{B}} \min\left\{\theta_{f,\phi}^2, \alpha\right\} + \sum_{\phi\in\mathcal{B}_\infty\setminus\mathcal{B}} \theta_{f,\phi}^2,\end{aligned}$$

*where $\theta_{f,\phi} = \int_{\mathbf{R}^d} f\phi$, $\Theta_f(\mathcal{B}) = (\theta_{f,\phi})_{\phi\in\mathcal{B}}$.*

## B.2 Pre-oracle inequality

Let

$$\mathcal{C} = \{g_\kappa : \kappa \in \mathbb{K}\}, \tag{100}$$

where $g_\kappa : \mathbf{R}^d \to \mathbf{R}$ and $\mathbb{K}$ is a set of parameters. Let $D : \mathbb{K} \to [0,\infty)$ be a penalization term and define the complexity penalized empirical risk as

$$\mathcal{E}_n(\kappa, \alpha) = \gamma_n\left(g_\kappa\right) + \alpha\, D(\kappa), \tag{101}$$

where $\gamma_n(g)$ is defined in (4) and $\alpha \geq 0$ is the smoothing parameter controlling the amount of penalization. We assume that $D(\kappa)$ takes larger values for more complex $g_\kappa$. Let $\hat{\kappa}$ be such that

$$\mathcal{E}_n\left(\hat{\kappa}, \alpha\right) \leq \inf_{\kappa\in\mathbb{K}} \mathcal{E}_n(\kappa, \alpha) + \epsilon, \tag{102}$$

where $\epsilon > 0$ and define the minimization estimator by

$$\hat{f} = g_{\hat{\kappa}}. \tag{103}$$

**Lemma 6** *Let $\mathcal{C} \subset L_2(\mathbf{R}^d)$ be parameterized by (100) and let $\hat{f} = g_{\hat{\kappa}} \in \mathcal{C}$ where $\hat{\kappa}$ satisfies (102). Then for each $f^0 = g_{\kappa^0} \in \mathcal{C}$,*

$$K(f, \hat{\kappa}, \alpha) \leq K(f, \kappa^0, \alpha) + \varepsilon + 2\nu_n\left(\hat{f} - f^0\right),$$

*where $f$ is the true density,*

$$K(f, \kappa, \alpha) = \|g_\kappa - f\|_2^2 + \alpha \cdot D(\kappa),$$

*and $\nu_n(g)$ is the centered empirical operator defined by*

$$\nu_n(g) = n^{-1}\sum_{i=1}^n g(X^i) - \int_{\mathbf{R}^d} gf,$$

*Proof.* We have for $g = \hat{f}$, $g = f^0$,

$$\|g - f\|_2^2 - \gamma_n(g) = \|f\|_2^2 - 2\int_{\mathbf{R}^d} fg + 2n^{-1}\sum_{i=1}^n g(X^i).$$

Thus,

$$\left\|\hat{f} - f\right\|_2^2 - \gamma_n\left(\hat{f}\right) + \gamma_n\left(f^0\right) - \left\|f^0 - f\right\|_2^2 = 2\nu_n\left(\hat{f} - f^0\right). \tag{104}$$

We have

$$\begin{aligned}
&K(f, \hat{\kappa}, \alpha) - K(f, \kappa^0, \alpha) \\
&= K(f, \hat{\kappa}, \alpha) - \mathcal{E}_n(\hat{\kappa}, \alpha) + \mathcal{E}_n(\hat{\kappa}, \alpha) - K(f, \kappa^0, \alpha) \\
&\leq K(f, \hat{\kappa}, \alpha) - \mathcal{E}_n(\hat{\kappa}, \alpha) + \mathcal{E}_n(\kappa^0, \alpha) + \varepsilon - K(f, \kappa^0, \alpha) \qquad (105) \\
&= \left\|\hat{f} - f\right\|_2^2 - \gamma_n(\hat{f}) + \gamma_n(f^0) + \varepsilon - \left\|f^0 - f\right\|_2^2 \\
&= 2\nu_n\left(\hat{f} - f^0\right) + \varepsilon. \qquad (106)
\end{aligned}$$

In (105) we applied (102) and in (106) we applied (104). □

## B.3 Expectation of the supremum

Let $\mathcal{G}$ be a set of linear combinations of an orthonormal system:

$$\mathcal{G} = \left\{\sum_{j=1}^k \theta_j\phi_j : \sum_{j=1}^k \theta_j^2 \leq B_2^2\right\}, \tag{107}$$

where $\{\phi_1, \ldots, \phi_k\}$ is an orthonormal system and $0 < B_2 < \infty$. We have a bound for $E\sup_{g\in\mathcal{G}}\nu_n(g)$ which depends essentially from $\sqrt{k/n}$.

**Lemma 7** *Let $\mathcal{G}$ be defined in (107). We have that*

$$E \sup_{g \in \mathcal{G}} \nu_n(g) \leq B_2 \|f\|_\infty^{1/2} (k/n)^{1/2} .$$

*Proof.* By the Cauchy-Schwartz inequality, for $g = \sum_{j=1}^k \theta_j \phi_j \in \mathcal{G}$,

$$\nu_n(g) = \sum_{j=1}^k \theta_j \nu_n(\phi_j) \leq \left( \sum_{j=1}^k \theta_j^2 \sum_{j=1}^k \nu_n(\phi_j)^2 \right)^{1/2} .$$

We have $E|X|^{1/2} \leq (E|X|)^{1/2}$. Thus,

$$E \sup_{g \in \mathcal{G}} \nu_n(g) \leq B_2 \left( \sum_{j=1}^k E\nu_n(\phi_j)^2 \right)^{1/2} .$$

We have

$$E\nu_n(\phi_j)^2 \leq \|f\|_\infty n^{-1},$$

which implies the lemma. □