Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection

Nathalie Akakpo

LPMA, Université Pierre et Marie Curie

February 16, 2011

Abstract

This article is devoted to nonlinear approximation and estimation via piecewise polynomials built on partitions into dyadic rectangles. The approximation rate is studied over possibly inhomogeneous and anisotropic smoothness classes that contain Besov classes. Highlighting the interest of such a result in statistics, adaptation in the minimax sense to both inhomogeneity and anisotropy of a related multivariate density estimator is proved. Besides, that estimation procedure can be implemented with a computational complexity simply linear in the sample size.

Contents

1 Introduction

When estimating a multivariate function, it seems natural to consider that its smoothness is likely to vary either spatially, or with the direction, or both. We will refer to the first feature as (spatial) inhomogeneity. If the risk is measured in a $\mathbb{L}_q$-norm, measuring the smoothness in a $\mathbb{L}_p$-norm with $p < q$ allows to take into account such an inhomogeneity – all the greater as p is smaller – in the sense that functions with some localized singularities and otherwise flat parts may thus keep a high smoothness index. For the second feature, we will talk about anisotropy, which is usually described by different indices of smoothness according to the coordinate directions. Yet, statistical procedures that adapt both to possible inhomogeneity and anisotropy remain rather scarce. Indeed, the existing literature seems to amount to the following references. Neumann and Von Sachs [NvS97], for estimating the evolutionary spectrum of a locally stationary time series, and Neumann [Neu00], in the Gaussian white noise framework, study thresholding procedures in a tensor product wavelet basis. In a Gaussian regression framework, Donoho [Don97] proposes the dyadic CART procedure, a selection procedure among histograms built on partitions into dyadic rectangles, extended to the density estimation framework by Klemelä [Kle09]. Last, Kerkyacharian, Lepski and Picard [KLP01] introduce a kernel estimator with adaptive bandwidth in the Gaussian white noise model. These authors study the performance of their procedures for the $\mathbb{L}_2$-risk, apart from the latter who consider any $\mathbb{L}_q$-risk for $q \geq 1$. Neumann and Von Sachs [NvS97] measure the smoothness of the function to estimate in the Sobolev scale, whereas the others consider the finer Besov scale. Besides, the $\mathbb{L}_p$-norm in which the smoothness is measured is allowed to vary with the direction, except in [Don97], but always constrained to be greater than 1. Common to those few procedures is the ability to reach the minimax rate over a wide range of possibly inhomogeneous and anisotropic classes, up to a logarithmic factor, the unknown smoothness being as usually limited by the *a priori* fixed smoothness of the underlying wavelets, piecewise polynomials or kernel.

Adaptation results of the aforementioned type rely as much on Statistics as on Approximation Theory, oracle-type inequalities reflecting the interplay between both domains. Assume for instance that the function s to estimate lies in the set $\mathcal{F}([0,1]^d, \mathbb{R})$ of all real-valued functions defined over the unit cube $[0,1]^d$, let $(S_m)_{m \in \mathcal{M}}$ be a given family of linear subspaces of $\mathcal{F}([0,1]^d, \mathbb{R})$ and $\tilde{s}$ a statistical procedure somehow based on that family. An oracle-type inequality in the $\mathbb{L}_q$-norm roughly takes the form

$$\mathbb{E}_s \left[\|s - \tilde{s}\|_q^q \right] \leq C \inf_{m \in \mathcal{M}} \left\{ \inf_{t \in S_m} \|s - t\|_q^q + (\dim(S_m)/n)^{q/2} \right\}, \tag{1}$$

where C is some positive constant, indicating that $\tilde{s}$ is able to choose a model S_m in the family that approximately realizes the best compromise between the approximation error and the dimension of the model. Equivalently, it may be written as

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_q^q\right] \leq C \inf_{D\in\mathbb{N}^\star}\left\{\inf_{t\in\cup_{m\in\mathcal{M}_D}S_m}\|s-t\|_q^q + (D/n)^{q/2}\right\}, \tag{2}$$

where $\mathcal{M}_D = \{m \in \mathcal{M} \text{ s.t. } \dim(S_m) = D\}$. On the other hand, the collection $(S_m)_{m\in\mathcal{M}}$ should be chosen so as to have good approximation properties over various classes $\mathcal{S}(\alpha, p, R)$ of functions with smoothness α measured in a $\mathbb{L}_p$-norm and with semi-norm smaller than R. Otherwise said, each approximating space $\cup_{m\in\mathcal{M}_D}S_m$ – typically nonlinear to deal with inhomogeneous functions– should satisfy, for a wide range of values of α, p and R,

$$\sup_{s\in\mathcal{S}(\alpha,p,R)}\inf_{t\in\cup_{m\in\mathcal{M}_D}S_m}\|s-t\|_q \leq C(\alpha,p)RD^{-\alpha/d}, \tag{3}$$

for some positive real $C(\alpha, p)$ that only depends on α and p. Combining the oracle-type inequality (2) and the approximation result (3) then provides an estimator $\tilde{s}$ with rate at most of order $(Rn^{-\alpha/d})^{qd/(d+2\alpha)}$ over each class $\mathcal{S}(\alpha, p, R)$, which is usually the minimax rate. Having at one's disposal spaces $(S_m)_{m\in\mathcal{M}}$ that do no depend on any *a priori* knowledge about the smoothness of the function to estimate – other than the scale of spaces it belongs to – and reaching the approximation rate (3) is thus a real issue for statisticians. In order to deal with inhomogeneity only, in a multivariate framework, such results appear for instance in the following references. DeVore, Jawerth and Popov [DJP92], Birgé and Massart [BM00] or Cohen, Dahmen, Daubechies and DeVore [CDDD01] propose wavelet based approximation algorithms aimed in particular at Besov type smoothness. Applications of the approximation result of [BM00] to statistical estimation may be found in Birgé and Massart [BM97] or Massart [Mas07] for instance. DeVore and Yu [DeV98] are concerned with piecewise polynomials built on partitions into dyadic cubes, notably for functions with Besov type smoothness. But their result will wait until Birgé [Bir06] to be used in Statistics. More generally, such results are in fact hidden behind all adaptive procedures. Thus, for both inhomogeneous and anisotropic functions, we refer in particular to the articles cited in the first paragraph. Let us underline that the procedure studied by Donoho [Don97] and Klemelä [Kle09], though based on dyadic rectangles instead of cubes, does not rely on a nonlinear approximation result via piecewise polynomials such as [DY90]. Indeed, the adaptivity of that estimator follows from its characterization as a wavelet selection procedure among some tree-structured subfamily of the Haar basis. Other nonlinear wavelet based approximation results are proved in Hochmuth [Hoc02b] or [Lei03] for anisotropic Besov spaces. Last, piecewise constant approximation based on dyadic rectangles is studied in Cohen and Mirebeau [CM09] for nonstandard smoothness spaces under the constraint of continuous differentiability.

Our aim here is to provide an approximation result tailored for statisticians, whose interest is illustrated by a new statistical procedure. The first part of the article is devoted to piecewise polynomial approximation based on partitions into dyadic rectangles. Thanks to an approximation algorithm inspired from DeVore and Yu [DY90], we obtain approximation rates akin to (3) over possibly inhomogeneous and anisotropic smoothness classes that contain for instance the more traditional Besov classes. The approximation rate can be measured in any $\mathbb{L}_q$-norm, for $1 \leq q \leq \infty$, and we allow an arbitrarily high inhomogeneity in the sense that we measure the smoothness in a $\mathbb{L}_p$-norm with p allowed to be arbitrarily close to 0. Besides, we take into account a possible restriction on the minimal size of the dyadic rectangles, which may arise in statistical applications. For estimating a multivariate function, we then introduce a selection procedure that chooses from the data the best partition into dyadic rectangles and the best piecewise polynomial built on that partition thanks to a penalized least-squares type criterion. The degree of the polynomial may vary from one rectangle to another, and also according to the coordinate directions, so as to provide a good adaptation both to inhomogeneity and anisotropy. Thus, our procedure extends the dyadic histogram selection procedures of Donoho [Don97], Klemelä [Kle09] or Blanchard, Schäfer, Rozenholc and Müller [BSRM07], and the dyadic piecewise polynomial estimation procedure proposed in a univariate or isotropic framework by Willett and Nowak [WN07]. We study the theoretical performance of the procedure – with no need to resort to the "wavelet trick" used in [Don97, Kle09] – for the $\mathbb{L}_2$-risk in the density estimation framework, as [Kle09], but we propose a more refined form of penalty than [Kle09]. For such a penalty, we provide an oracle-type inequality and adaptivity results in the minimax sense over a wide range of possibly inhomogeneous and anisotropic smoothness classes that contain Besov type classes. We emphasize that, if the maximal degree of the polynomials does not depend on the sample size, we reach the minimax rate up to a constant factor only, contrary to *all* the previously mentioned estimators. This results not only from the good approximation properties of dyadic piecewise polynomials, but also from the moderate number of dyadic partitions of the same size. We can also allow the maximal degree of the polynomials to grow logarithmically with the sample size, in which case we reach the minimax rate on a growing range of smoothness classes, up to a logarithmic factor. Moreover, our procedure can be implemented with a computational complexity only linear in the sample size, possibly up to a logarithmic factor, depending on the way we choose the maximal degree.

The plan of the paper is as follows. Section 2 is devoted to piecewise polynomial approximation based on partitions into dyadic rectangles. In Section 3, we are concerned with density estimation based on a data-driven choice of a best dyadic piecewise polynomial. We study there the theoretical properties of the procedure and briefly describe the algorithm to implement it. Most proofs of Sections 2 and 3 are deferred respectively to Section 4 and to Sections 5 and 6.

2 Adaptive approximation by dyadic piecewise polynomials

In this section, we present an approximation algorithm by piecewise polynomials built on partitions into dyadic rectangles. We study its rate of approximation over some classes of functions that may present at the same time anisotropic and inhomogeneous smoothness.

2.1 Notation

Throughout the article, we fix $d \in \mathbb{N}^\star$, and throughout this section, we fix some d-uple of nonnegative integers $\boldsymbol{r} = (r_1, \ldots, r_d)$ that represent the maximal degree of polynomial approximation in each direction. We call dyadic rectangle of $[0,1]^d$ any set of the form $I_1 \times \ldots \times I_d$ where, for all $1 \leq l \leq d$,

$$I_l = [0, 2^{-j_l}] \quad \text{or} \quad I_l =]k_l 2^{-j_l}, (k_l+1)2^{-j_l}]$$

with $j_l \in \mathbb{N}$ and $k_l \in \{1, \ldots, 2^{j_l} - 1\}$. Otherwise said, a dyadic rectangle of $[0,1]^d$ is defined as a product of d dyadic intervals of $[0,1]$ that may have different lengths. For a partition m of $[0,1]^d$ into dyadic rectangles, we denote by $|m|$ the number of rectangles in m and by $S_{(m,\boldsymbol{r})}$ the space of all piecewise polynomial functions on $[0,1]^d$ which are polynomial with degree $\leq r_l$ in the l-th direction, $l = 1, \ldots, d$, over each rectangle of m. Besides, for $0 < p \leq \infty$, we denote by $\mathbb{L}_p([0,1]^d)$ the set of all real-valued and measurable functions s on $[0,1]^d$ such that the (quasi-)norm

$$\|s\|_p = \begin{cases} \left(\int_{[0,1]^d} |s(x)|^p \mathrm{d}\lambda_d(x) \right)^{1/p} & \text{if } 0 < p < \infty \\ \sup_{x \in [0,1]^d} |s(x)| & \text{if } p = \infty \end{cases}$$

is finite, where λ_d is the Lebesgue measure on $[0,1]^d$. Last, $C(\theta)$, $C_i(\theta)$ or $C'_i(\theta)$, $i \in \mathbb{N}^\star$ stand for a positive reals that only depend on the parameter θ. Their values may change from one line to another, unless otherwise said.

2.2 Approximation algorithm

Let us fix $1 \leq q \leq \infty$. In order to approximate a possibly anisotropic and inhomogeneous function s in the $\mathbb{L}_q$-norm, we propose an approximation algorithm inspired from [DY90]. We shall construct an adequate piecewise polynomial approximation on a partition into dyadic rectangles adapted to s, beginning with the trivial partition of the unit square $[0,1]^d$ and proceeding to successive refinements. For doing so, we consider the criterion

$$\mathcal{E}_{\boldsymbol{r}}(s, K)_q = \inf_{P \in \mathscr{P}_{\boldsymbol{r}}} \|(s - P)\mathbb{1}_K\|_q \tag{4}$$

measuring the error in approximating s on a rectangle $K \subset [0,1]^d$ by some element from the set $\mathscr{P}_{\boldsymbol{r}}$ of all polynomials on $[0,1]^d$ with degree $\leq r_l$ in the l-th direction. We also fix some threshold $\epsilon > 0$ – to be chosen later, according to the smoothness assumptions on s. But contrary to [DY90], we allow the degrees of smoothness of s to vary with the directions and describe them by a multi-index $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d) \in \prod_{l=1}^d (0, r_l + 1)$, in a sense that will be made precise in the next subsection. Thus, our algorithm is based on a special subcollection of dyadic rectangles adapted to an anisotropic smoothness measured by $\boldsymbol{\sigma}$. Indeed, for $j \in \mathbb{N}$, we define $\mathcal{D}_j^{\boldsymbol{\sigma}}$ as the set of all dyadic rectangles $I_1 \times \ldots \times I_d \subset [0,1]^d$ such that, for all $1 \leq l \leq d$,

$$I_l = \left[0, 2^{-\lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor}\right] \quad \text{or} \quad I_l = \left]k_l 2^{-\lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor}, (k_l+1) 2^{-\lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor}\right],$$

with $\underline{\boldsymbol{\sigma}} = \min_{1 \leq l \leq d} \sigma_l$ and $k_l \in \{1, \ldots, 2^{\lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor} - 1\}$, and we set $\mathcal{D}^{\boldsymbol{\sigma}} = \cup_{j \in \mathbb{N}} \mathcal{D}_j^{\boldsymbol{\sigma}}$. It should be noticed that, for all $j \in \mathbb{N}$, any $K \in \mathcal{D}_j^{\boldsymbol{\sigma}}$ can be partitioned into dyadic rectangles of $\mathcal{D}_{j+1}^{\boldsymbol{\sigma}}$, that we call children of K. For $d = 2$ and $\sigma_2 = 2\sigma_1$ for instance, a partition of $[0,1]^2$ into dyadic rectangles from $\mathcal{D}^{\boldsymbol{\sigma}}$ will thus be roughly twice as fine in the first direction, as illustrated by Figure 1.

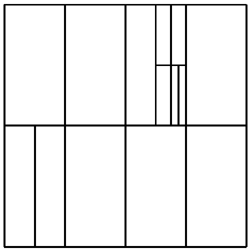

Figure 1: Example of partition of $[0,1]^2$ into dyadic rectangles from $\mathcal{D}^{\boldsymbol{\sigma}}$ for $\sigma_2 = 2\sigma_1$.

The algorithm begins with the set $\mathcal{I}^1(s,\epsilon)$ that only contains $[0,1]^d$. If $\mathcal{E}_{\boldsymbol{r}}(s,[0,1]^d)_q < \epsilon$, then the algorithm stops. Else, $[0,1]^d$ is replaced with his children in $\mathcal{I}^1(s,\epsilon)$, hence a new partition $\mathcal{I}^2(s,\epsilon)$. In the same way, the k-th step begins with a partition $\mathcal{I}^k(s,\epsilon)$ of $[0,1]^d$ into dyadic rectangles that belong to $\mathcal{D}^{\boldsymbol{\sigma}}$. If $\max_{K \in \mathcal{I}^k(s,\epsilon)} \mathcal{E}_{\boldsymbol{r}}(s,K)_q < \epsilon$, then the algorithm stops. Else, a dyadic rectangle $K \in \mathcal{I}^k(s,\epsilon)$ such that $\mathcal{E}_{\boldsymbol{r}}(s,K)_q \geq \epsilon$ is chosen and replaced with his children in $\mathcal{I}^k(s,\epsilon)$, hence a new partition $\mathcal{I}^{k+1}(s,\epsilon)$. Since $s \in \mathbb{L}_q([0,1]^d)$, $\mathcal{E}_{\boldsymbol{r}}(s,K)_q$ tends to 0 when the Lebesgue measure of K tends to 0, so the algorithm finally stops. The final partition $\mathcal{I}(s,\epsilon)$ only contains dyadic rectangles that belong to $\mathcal{D}^{\boldsymbol{\sigma}}$ and such that $\max_{K \in \mathcal{I}(s,\epsilon)} \mathcal{E}_{\boldsymbol{r}}(s,K)_q < \epsilon$. For all $K \in \mathcal{I}(s,\epsilon)$, we approximate s on K by $Q_K(s)$, a polynomial function with degree $\leq r_l$ in the l-th direction such that $\|(s - Q_K(s))\mathbb{1}_K\|_q = \mathcal{E}_{\boldsymbol{r}}(s,K)_q$. Otherwise said, we approximate s on the unit cube by

$$A(s,\epsilon) = \sum_{K \in \mathcal{I}(s,\epsilon)} Q_K(s),$$

thus committing the error

$$\|s - A(s,\epsilon)\|_q = \left(\sum_{K \in \mathcal{I}(s,\epsilon)} \|(s - Q_K(s))\mathbb{1}_K\|_q^q \right)^{1/q} < |\mathcal{I}(s,\epsilon)|^{1/q}\epsilon \tag{5}$$

if $1 \leq q < \infty$, and

$$\|s - A(s,\epsilon)\|_\infty = \max_{K \in \mathcal{I}(s,\epsilon)} \|(s - Q_K(s))\mathbb{1}_K\|_\infty < \epsilon \tag{6}$$

if $q = \infty$.

2.3 Approximation rate over anisotropic function classes

In order to study the approximation rate of the previous algorithm, we introduce function spaces that arise naturally from the way the algorithm proceeds. Let us fix $\boldsymbol{\sigma} \in \prod_{l=1}^d (0, r_l + 1)$ and $0 < p, p' \leq \infty$. For $s \in \mathbb{L}_p([0,1]^d)$ and $k \in \mathbb{N}$, we set

$$e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s) = \inf_{P \in \Pi_k^{\boldsymbol{r},\boldsymbol{\sigma}}} \|s - P\|_p \tag{7}$$

where $\Pi_k^{\boldsymbol{r},\boldsymbol{\sigma}}$ is the set of all piecewise polynomial functions on $[0,1]^d$ that are polynomial with degree $\leq r_l$ in the l-th direction over each rectangle in $\mathcal{D}_k^{\boldsymbol{\sigma}}$. Then, we define $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$ as the set of all functions $s \in \mathbb{L}_p([0,1]^d)$ such that the quantity

$$N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'}(s) = \begin{cases} \left(\sum_{k \in \mathbb{N}} \left(2^{k\underline{\boldsymbol{\sigma}}} e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s) \right)^{p'} \right)^{1/p'} & \text{if } 0 < p' < \infty \\ \sup_{k \in \mathbb{N}} \left(2^{k\underline{\boldsymbol{\sigma}}} e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s) \right) & \text{if } p' = \infty \end{cases}$$

is finite. One can easily verify that $N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'}$ is a (quasi-)semi-norm on $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$, and that $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$ gets larger as p' increases since

$$N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'_2}(s) \leq N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'_1}(s) \text{ for } 0 < p'_1 \leq p'_2 \leq \infty. \tag{8}$$

If $p \geq q$, then $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$ is obviously embedded in the space $\mathbb{L}_q([0,1]^d)$ in which we measure the quality of approximation. The same property still holds for p smaller than q, under adequate assumptions on the harmonic mean $H(\boldsymbol{\sigma})$ of $\sigma_1, \dots, \sigma_d$, *i.e.*

$$H(\boldsymbol{\sigma}) = \left(\frac{1}{d} \sum_{l=1}^d \frac{1}{\sigma_l} \right)^{-1}.$$

Indeed, denoting by $(x)_+ = \max\{x, 0\}$ for any real x, we prove in Section 4 the following continuous embedding.

Proposition 1 *Let $\boldsymbol{\sigma} \in \prod_{l=1}^{d}(0, r_l+1)$, $0 < p, p' \leq \infty$ and $1 \leq q \leq \infty$. If*

$$H(\boldsymbol{\sigma})/d > (1/p - 1/q)_+ ,$$

then $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d)) \subset \mathbb{L}_q([0,1]^d)$ and, for all $s \in \mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$,

$$\|s\|_q \leq C(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, p', q)\left(\|s\|_p + N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'}(s)\right).$$

The reader familiar with classical function spaces will have noted the similarity between the definition and the embedding properties of spaces $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$ and those of Besov spaces. Before going further, let us recall the definition of the latter according to [ST87], for instance. We denote by $(\mathbf{b}_1, \ldots, \mathbf{b}_d)$ the canonical basis of $\mathbb{R}^d$ and set $\mathcal{R} = [0,1]^{\mathrm{d}}$. For all $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d) \in (0, +\infty)^d$, $0 < p, p' \leq \infty$, $s \in \mathbb{L}_p([0,1]^d)$, $h > 0$ and $1 \leq l \leq d$, we define

$$\mathcal{R}(\sigma_l, h) = \{x \in [0,1]^d \text{ s.t. } x, x + h\mathbf{b}_l, \ldots, x + (\lfloor \sigma_l \rfloor + 1)h\mathbf{b}_l \in \mathcal{R}\},$$

$$\Delta_{h\mathbf{b}_l}^{\sigma_l} s(x) = \sum_{k=0}^{\lfloor \sigma_l \rfloor + 1} \binom{\lfloor \sigma_l \rfloor + 1}{k} (-1)^{\lfloor \sigma_l \rfloor + 1 - k} s(x + kh\mathbf{b}_l), \text{ for } x \in \mathcal{R}(\sigma_l, h),$$

$$\omega_{\sigma_l}^{(l)}(s, y, \mathcal{R})_p = \sup_{0 < h \leq y} \|\Delta_{h\mathbf{b}_l}^{\sigma_l} s \mathbb{1}_{\mathcal{R}(\sigma_l, h)}\|_p, \text{ for } y \geq 0,$$

$$|s|_{\boldsymbol{\sigma},p,p'} = \begin{cases} \sum_{l=1}^{d} \left(\int_0^\infty \left[y^{-\sigma_l} \omega_{\sigma_l}^{(l)}(s, y, \mathcal{R})_p \right]^{p'} \frac{\mathrm{d}y}{y} \right)^{1/p'} & \text{if } 0 < p' < \infty \\ \sum_{l=1}^{d} \left(\sup_{y > 0} y^{-\sigma_l} \omega_{\sigma_l}^{(l)}(s, y, \mathcal{R})_p \right) & \text{if } p' = \infty. \end{cases}$$

For $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d) \in (0, +\infty)^d$, $0 < p, p' \leq \infty$, we denote by $\mathscr{B}_{p'}^{\boldsymbol{\sigma}}\left(\mathbb{L}_p([0,1]^d)\right)$ the space of all measurable functions $s \in \mathbb{L}_p([0,1]^d)$ such that $|s|_{\boldsymbol{\sigma},p,p'}$ is finite. According to the proposition below, Besov spaces $\mathscr{B}_{p'}^{\boldsymbol{\sigma}}\left(\mathbb{L}_p([0,1]^d)\right)$ are embedded in spaces $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$.

Proposition 2 *Let $\boldsymbol{\sigma} \in \prod_{l=1}^{d}(0, r_l+1)$, $0 < p < \infty$ and $0 < p' \leq \infty$. For all $s \in \mathscr{B}_{p'}^{\boldsymbol{\sigma}}\left(\mathbb{L}_p([0,1]^d)\right)$,*

$$N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'}(s) \leq C(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, p')|s|_{\boldsymbol{\sigma},p,p'}.$$

We shall not give a proof of that proposition here, since it relies exactly on the same arguments as those used by [Hoc02a] in the proof of Theorem 4.1 (beginning of page 197) combined with Inequality (14) in the same reference. It should be noticed that the space $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$ is in general larger than $\mathscr{B}_{p'}^{\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$. Indeed, contrary to $\mathscr{B}_{p'}^{\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$, the space $\mathcal{N}_{p'}^{\boldsymbol{r},\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$ contains discontinuous functions (piecewise polynomials, for instance) even for $H(\boldsymbol{\sigma})/d > 1/p$.

We are now able to state approximation rates over anisotropic classes of the form

$$\mathcal{S}(\boldsymbol{r}, \boldsymbol{\sigma}, p, p', R) = \{s \in \mathbb{L}_p([0,1]^d) \text{ s.t. } N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'}(s) \leq R\},$$

where $\boldsymbol{\sigma} \in \prod_{l=1}^d (0, r_l+1)$, $0 < p, p' \leq \infty$ and $R > 0$, thus extending the result of DeVore and Yu [DY90] (Corollary 3.3), which is only devoted to functions with isotropic smoothness. The approximation rate is related to the harmonic mean $H(\boldsymbol{\sigma})$ of $\sigma_1, \dots, \sigma_d$, which in case of isotropic smoothness of order σ, *i.e.* if $\sigma_1 = \dots = \sigma_d = \sigma$, reduces to σ.

Theorem 1 *Let $R > 0$, $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in \prod_{l=1}^d (0, r_l + 1)$, $0 < p < \infty$ and $1 \leq q \leq \infty$ such that*

$$H(\boldsymbol{\sigma})/d > (1/p - 1/q)_+.$$

Assume that $s \in \mathcal{S}(\boldsymbol{r}, \boldsymbol{\sigma}, p, p', R)$, where $p' = \infty$ if $0 < p \leq 1$ or $p \geq q$, and $p' = p$ if $1 < p < q$. Then, for all $k \in \mathbb{N}$, there exists some partition m of $[0,1]^d$ into dyadic rectangles, that may depend on $s, d, \boldsymbol{r}, \boldsymbol{\sigma}, p$ and q, such that

$$|m| \leq C_1(d, \boldsymbol{\sigma}, p) 2^{kd}$$

and

$$\inf_{t \in S_{(m,\boldsymbol{r})}} \|s - t\|_q \leq C_2(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) R 2^{-kH(\boldsymbol{\sigma})}. \tag{9}$$

The same result still holds whatever $0 < p' \leq \infty$ if $0 < p \leq 1$ or $p \geq q$, and whatever $0 < p' \leq p$ if $1 < p < q$, as a straightforward consequence of Theorem 1 and Inequality (8). Denoting by $\mathcal{M}_D$, $D \in \mathbb{N}^\star$, the set of all the partitions of $[0,1]^d$ into D dyadic rectangles, we obtain uniform approximation rates simultaneously over a wide range of classes $\mathcal{S}(\boldsymbol{r}, \boldsymbol{\sigma}, p, p', R)$ by considering the nonlinear approximating space $\cup_{m \in \mathcal{M}_D} S_{(m,\boldsymbol{r})}$. That property is stated more precisely in Corollary 1 below, which can be immediately derived from Theorem 1.

Corollary 1 *Let $R > 0$, $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in \prod_{l=1}^d (0, r_l + 1)$, $0 < p < \infty$, $0 < p' \leq \infty$ and $1 \leq q \leq \infty$ satisfying the assumptions of Theorem 1. For all $D \geq C_1(d, \boldsymbol{\sigma}, p)$, where $C_1(d, \boldsymbol{\sigma}, p)$ is given by Theorem 1,*

$$\sup_{s \in \mathcal{S}(\boldsymbol{r}, \boldsymbol{\sigma}, p, p', R)} \inf_{t \in \cup_{m \in \mathcal{M}_D} S_{(m,\boldsymbol{r})}} \|s - t\|_q \leq C_2'(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) R D^{-H(\boldsymbol{\sigma})/d}.$$

We also propose of a more refined version of Theorem 1 that allows to take into account constraints on the minimal dimensions of the dyadic rectangles, which will prove most useful for estimation purpose in the next section. We recall that $\underline{\boldsymbol{\sigma}} = \min_{1 \leq l \leq d} \sigma_l$.

Theorem 2 *Let $J \in \mathbb{N}$, $R > 0$, $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in \prod_{l=1}^d (0, r_l+1)$, $0 < p < \infty$, $0 < p' \leq \infty$ and $1 \leq q \leq \infty$ such that*

$$H(\boldsymbol{\sigma})/d > (1/p - 1/q)_+ .$$

Assume that $s \in \mathcal{S}(\boldsymbol{r}, \boldsymbol{\sigma}, p, p', R)$, where $p' = \infty$ if $0 < p \leq 1$ or $p \geq q$, and $p' = p$ if $1 < p < q$. Then, for all $k \in \mathbb{N}$, there exists some partition m of $[0,1]^d$, that may depend *on* $s, d, \boldsymbol{r}, \boldsymbol{\sigma}, p$ *and* q*, only contains dyadic rectangles with sidelength at least* $2^{-J\underline{\boldsymbol{\sigma}}/\sigma_l}$ *in the* l*-th direction,* $l = 1, \dots, d$*, and satisfies both*

$$|m| \leq C_1(d, \boldsymbol{\sigma}, p) 2^{kd}$$

and

$$\inf_{t \in S_{(m,\boldsymbol{r})}} \|s - t\|_q \leq C_3(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) R \left(2^{-Jd(H(\boldsymbol{\sigma})/d - (1/p - 1/q)_+)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} + 2^{-kH(\boldsymbol{\sigma})} \right). \quad (10)$$

Remark: Given $J \in \mathbb{N}$, that theorem relies on applying the approximation algorithm of Section 2.2 to an approximation of s from $S_{(m_J,\boldsymbol{r})}$, where m_J is the partition of $[0,1]^d$ into the dyadic rectangles from $\mathcal{D}^{\boldsymbol{\sigma}}_J$. Thus, the term $2^{-Jd(H(\boldsymbol{\sigma})/d - (1/p - 1/q)_+)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}$ in (10), which is of order $(\dim(S_{(m_J,\boldsymbol{r})}))^{-(H(\boldsymbol{\sigma})/d - (1/p-1/q)_+)}$, corresponds with an upper-bound for the linear approximation error $\inf_{t \in S_{(m_J,\boldsymbol{r})}} ||s - t||_q$. The upper-bound (10) is of the same order as (9) – up to a real that only depends on $d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q$ – as long as

$$k \leq J \frac{\underline{\boldsymbol{\sigma}}}{H(\boldsymbol{\sigma})} \left(\frac{H(\boldsymbol{\sigma})}{d} - \left(\frac{1}{p} - \frac{1}{q} \right)_+ \right) \frac{d}{H(\boldsymbol{\sigma})}. \quad (11)$$

If $p \geq q$ and $\underline{\boldsymbol{\sigma}} = H(\boldsymbol{\sigma})$, *i.e.* if s has homogeneous and isotropic smoothness, then that condition simply amounts to $k \leq J$. Otherwise, Condition (11) is all the more stringent as p is small by comparison with q or as $\underline{\boldsymbol{\sigma}}$ is small by comparison with $H(\boldsymbol{\sigma})$, *i.e.* all the more stringent as inhomogeneity or anisotropy are pronounced.

Given $J \in \mathbb{N}$, let us denote by $\mathcal{M}^J_D$ the set of all the partitions into D dyadic rectangles with sidelengths $\geq 2^{-J}$, for $D \in \mathbb{N}^\star$. We can still obtain uniform approximation rates simultaneously over a wide range of classes $\mathcal{S}(\boldsymbol{r}, \boldsymbol{\sigma}, p, p', R)$ under the constraint that the piecewise polynomial approximations are built over dyadic rectangles with sidelengths $\geq 2^{-J}$, by introducing this time the nonlinear approximation space $\cup_{m \in \mathcal{M}^J_D} S_{(m,\boldsymbol{r})}$. Indeed, as for all $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in \prod_{l=1}^d (0, r_l + 1)$ and $l = 1, \dots, d$, $2^{-J\underline{\boldsymbol{\sigma}}/\sigma_l} \geq 2^{-J}$, a straightforward consequence of Theorem 2 is Corollary 2 below.

Corollary 2 *Let* $J \in \mathbb{N}$*,* $R > 0$*,* $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in \prod_{l=1}^d (0, r_l + 1)$*,* $0 < p < \infty$*,* $0 < p' \leq \infty$ *and* $1 \leq q \leq \infty$ *satisfying the assumptions of Theorem 2. For all* $D \geq C_1(d, \boldsymbol{\sigma}, p)$*, where* $C_1(d, \boldsymbol{\sigma}, p)$ *is given by Theorem 2,*

$$\begin{aligned} &\sup_{s \in \mathcal{S}(\boldsymbol{r}, \boldsymbol{\sigma}, p, p', R)} \inf_{t \in \cup_{m \in \mathcal{M}^J_D} S_{(m,\boldsymbol{r})}} \|s - t\|_q \\ &\leq C'_3(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) R \left(2^{-Jd(H(\boldsymbol{\sigma})/d - (1/p - 1/q)_+)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} + 2^{-kH(\boldsymbol{\sigma})} \right). \end{aligned}$$

3 Application to density estimation

This section aims at illustrating the interest of the previous approximation results in statistics. More precisely, placing ourselves in the density estimation framework, we show that combining estimation via dyadic piecewise polynomial selection and the aforementioned approximation results leads to a new density estimator which is able to adapt to the unknown smoothness of the function to estimate, even though it is both anisotropic and inhomogeneous. Besides, we explain how such a procedure can be implemented efficiently.

3.1 Framework and notation

Let $n \in \mathbb{N}$, $n \geq 4$, we observe independent and identically distributed random variables $Y_1, \dots, Y_n$ defined on the same measurable space $(\Omega, \mathcal{A})$ and taking values in $[0,1]^d$. We assume that $Y_1, \dots, Y_n$ admit the same density s with respect to the Lebesgue measure λ_d on $[0,1]^d$ and that $s \in \mathbb{L}_2([0,1]^d)$. We denote by P_s the joint distribution of $(Y_1, \dots, Y_n)$, that is the probability measure with density

$$\frac{dP_s}{d\lambda_d^{\otimes n}} : (y_1, \dots, y_n) \in [0,1]^d \times \dots \times [0,1]^d \longmapsto \prod_{i=1}^{n} s(y_i),$$

while $\mathbb{P}_s$ stands for the underlying probability measure on $(\Omega, \mathcal{A})$, so that for all product B of n rectangles of $[0,1]^d$

$$P_s(B) = \mathbb{P}_s(\{\omega \in \Omega \text{ s.t. } (Y_1(\omega), \dots, Y_n(\omega)) \in B\}).$$

The expectation and variance associated with $\mathbb{P}_s$ are denoted by $\mathbb{E}_s$ and Var_s.

3.2 Dyadic piecewise polynomial estimators

Let m be some partition of $[0,1]^d$ into dyadic rectangles and $\boldsymbol{\rho} = (\boldsymbol{\rho}_K)_{K \in m}$ a sequence such that, for all $K \in m$, $\boldsymbol{\rho}_K = (\rho_K(1), \dots, \rho_K(d)) \in \mathbb{N}^d$. We denote by $S_{(m,\boldsymbol{\rho})}$ the space of all functions $t : [0,1]^d \to \mathbb{R}$ such that, for all $K \in m$, t is polynomial with degree $\leq \rho_K(l)$ in the l-th direction on the rectangle K. In particular, if $\boldsymbol{\rho}$ is constant and equal to $\boldsymbol{r}$, then $S_{(m,\boldsymbol{\rho})}$ coincides with the space $S_{(m,\boldsymbol{r})}$ introduced in Section 2. Let $\langle ., . \rangle$ be the usual scalar product on $\mathbb{L}_2([0,1]^d)$. We recall that s minimizes over $t \in \mathbb{L}_2([0,1]^d)$

$$\|s - t\|_2^2 - \|s\|_2^2 = \|t\|_2^2 - 2\langle t, s \rangle = \mathbb{E}_s[\gamma(t)],$$

where

$$\gamma(t) = \|t\|_2^2 - \frac{2}{n} \sum_{i=1}^{n} t(Y_i)$$

only depends on the observed variables. Thus, a natural estimator of s with values in $S_{(m,\boldsymbol{\rho})}$ is

$$\hat{s}_{(m,\boldsymbol{\rho})} = \underset{t \in S_{(m,\boldsymbol{\rho})}}{\operatorname{argmin}} \gamma(t),$$

that we will call a dyadic piecewise polynomial estimator. Such an estimator is just a projection estimator of s on $S_{(m,\boldsymbol{\rho})}$. Indeed, if for each dyadic rectangle K we set $\Lambda(\boldsymbol{\rho}_K) = \prod_{l=1}^{d} \{0, \ldots, \rho_K(l)\}$ and denote by $(\Phi_{K,\boldsymbol{k}})_{\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}_K)}$ an orthonormal basis of the space of polynomial functions over K with degree $\leq \rho_K(l)$ in the l-th direction, then simple computations lead to

$$\hat{s}_{(m,\boldsymbol{\rho})} = \sum_{K \in m} \sum_{\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}_K)} \left(\frac{1}{n} \sum_{i=1}^{n} \Phi_{K,\boldsymbol{k}}(Y_i) \right) \Phi_{K,\boldsymbol{k}}.$$

For theoretical reasons, we shall choose in the remaining of the article an orthonormal basis $(\Phi_{K,\boldsymbol{k}})_{\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}_K)}$ derived from the Legendre polynomials in the following way. Let $(Q_j)_{j \in \mathbb{N}}$ be the orthogonal family of the Legendre polynomials in $\mathbb{L}_2([-1,1])$. For $K = \prod_{l=1}^{d} [u_i, v_i]$ rectangle of $[0,1]^d$, $\boldsymbol{k} = (k(1), \ldots, k(d)) \in \mathbb{N}^d$ and $x = (x_1, \ldots, x_d) \in [0,1]^d$, we set

$$\pi(\boldsymbol{k}) = \prod_{l=1}^{d} (2k(l) + 1)$$

and

$$\Phi_{K,\boldsymbol{k}}(x) = \sqrt{\frac{\pi(\boldsymbol{k})}{\lambda_d(K)}} \prod_{l=1}^{d} Q_{k(l)} \left(\frac{2x_l - u_l - v_l}{v_l - u_l} \right) \mathbb{1}_K(x).$$

We recall that, for all $j \in \mathbb{N}$, Q_j satisfies

$$\|Q_j\|_\infty = 1 \quad \text{and} \quad \|Q_j\|_2^2 = \frac{2}{(2j+1)}.$$

Therefore, for K rectangle in $[0,1]^d$ and $\boldsymbol{\rho}_K \in \mathbb{N}^d$, $(\Phi_{K,\boldsymbol{k}})_{\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}_K)}$ is a basis of the space of piecewise polynomial functions with support K and degree $\leq \rho_K(l)$ in the l-th direction, which is orthonormal for the norm $\|.\|_2$ and satisfies

$$\|\Phi_{K,\boldsymbol{k}}\|_\infty^2 = \frac{\pi(\boldsymbol{k})}{\lambda_d(K)}. \tag{12}$$

For each partition m of $[0,1]^d$ into dyadic rectangles and each $\boldsymbol{\rho} = (\boldsymbol{\rho}_K)_{K \in m} \in (\mathbb{N}^d)^{|m|}$, we can evaluate the performance of $\hat{s}_{(m,\boldsymbol{\rho})}$ by giving an upper-bound for its quadratic risk. For that purpose, we introduce the orthogonal projection $s_{(m,\boldsymbol{\rho})}$ of s on $S_{(m,\boldsymbol{\rho})}$, the dimension $\dim(S_{(m,\boldsymbol{\rho})})$ of $S_{(m,\boldsymbol{\rho})}$, *i.e.*

$$\dim(S_{(m,\boldsymbol{\rho})}) = \sum_{K \in m} |\Lambda(\boldsymbol{\rho}_K)| = \sum_{K \in m} \prod_{l=1}^{d} (\rho_K(l) + 1),$$

and define $\boldsymbol{\rho_{\max}} = (\rho_{\max}(1), \dots, \rho_{\max}(d))$ by

$$\rho_{\max}(l) = \max_{K \in m} \rho_K(l), l = 1, \dots, d. \tag{13}$$

Proposition 3 *Let m be a partition of $[0,1]^d$ into dyadic rectangles and $\boldsymbol{\rho} = (\boldsymbol{\rho}_K)_{K \in m} \in (\mathbb{N}^d)^{|m|}$. If $s \in \mathbb{L}_2([0,1]^d)$, then*

$$\mathbb{E}_s \left[\|s - \hat{s}_{(m,\boldsymbol{\rho})}\|_2^2 \right] = \|s - s_{(m,\boldsymbol{\rho})}\|_2^2 + \frac{1}{n} \sum_{K \in m} \sum_{\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}_K)} \mathrm{Var}_s(\Phi_{K,\boldsymbol{k}}(Y_1)).$$

If $\|s\|_\infty$ is finite, then

$$\mathbb{E}_s \left[\|s - \hat{s}_{(m,\boldsymbol{\rho})}\|_2^2 \right] \leq \|s - s_{(m,\boldsymbol{\rho})}\|_2^2 + \pi(\boldsymbol{\rho_{\max}}) \|s\|_\infty \frac{\dim(S_{(m,\boldsymbol{\rho})})}{n}.$$

Proof: Pythagoras' Equality gives

$$\mathbb{E}_s \left[\|s - \hat{s}_{(m,\boldsymbol{\rho})}\|_2^2 \right] = \|s - s_{(m,\boldsymbol{\rho})}\|_2^2 + \mathbb{E}_s \left[\|s_{(m,\boldsymbol{\rho})} - \hat{s}_{(m,\boldsymbol{\rho})}\|_2^2 \right].$$

Then, we deduce the first equality in Proposition 3 from the expressions of $\hat{s}_{(m,\boldsymbol{\rho})}$ and $s_{(m,\boldsymbol{\rho})}$ in the orthonormal basis $(\Phi_{K,\boldsymbol{k}})_{K \in m, \boldsymbol{k} \in \Lambda(\boldsymbol{\rho}_K)}$ of $S_{(m,\boldsymbol{\rho})}$ and the fact that $Y_1, \dots, Y_n$ are independent and identically distributed.

If s is bounded, we deduce from (12) that, for all $K \in m$ and $\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}_K)$,

$$\mathbb{E}_s \left[\Phi_{K,\boldsymbol{k}}^2(Y_1) \right] \leq \langle s, \mathbb{1}_K \rangle \frac{\pi(\boldsymbol{k})}{\lambda_d(K)} \leq \|s\|_\infty \pi(\boldsymbol{\rho_{\max}}),$$

hence the upper-bound for $\mathbb{E}_s \left[\|s - \hat{s}_{(m,\boldsymbol{\rho})}\|_2^2 \right]$. ■

Thus, we recover that, for bounded densities at least, choosing a model $S_{(m,\boldsymbol{\rho})}$ that realizes a good compromise between the approximation error and the dimension of the model leads to an estimator $\hat{s}_{(m,\boldsymbol{\rho})}$ with small risk. Such a choice reveals in fact optimal for densities presenting the kind of smoothness described in Section 2.3. More precisely, for $\boldsymbol{\sigma} \in (0, +\infty)^d$, $0 < p, p' \leq \infty$, $R > 0$ and $L > 0$, we set $\lfloor \boldsymbol{\sigma} \rfloor = (\lfloor \sigma_1 \rfloor, \dots, \lfloor \sigma_d \rfloor)$ and consider the class $\mathcal{P}(\boldsymbol{\sigma}, p, p', R, L)$ of all the probability densities s with respect to λ_d such that $s \in \mathcal{S}(\lfloor \boldsymbol{\sigma} \rfloor + 1, \boldsymbol{\sigma}, p, p', R)$ and $\|s\|_\infty \leq L$. Thanks to the upper-bound of Proposition 3, we obtain in Proposition 4 below that any statistical procedure which is able to realize approximately $\inf_{m \in \mathcal{M}, \boldsymbol{\rho} \in \mathbb{N}^d} \mathbb{E}_s \left[\|s - \hat{s}_{(m,\boldsymbol{\rho})}\|_2^2 \right]$, where $\mathcal{M}$ is the collection of all the partitions of $[0,1]^d$ into dyadic rectangles, enjoys adaptivity properties: it also reaches approximately the minimax risk over a wide range of classes $\mathcal{P}(\boldsymbol{\sigma}, p, p', R, L)$.

Proposition 4 *For $0 < p < \infty$, let $p' = \infty$ when $0 < p \leq 1$ or $p \geq 2$, and $p' = p$ when $1 < p < 2$. For all $L > 0$ and $R \geq n^{-1/2}$, if $\boldsymbol{\sigma} \in (0, +\infty)^d$ and $0 < p < \infty$ satisfy $H(\boldsymbol{\sigma})/d > (1/p - 1/2)_+$, then*

$$\begin{aligned}
&\sup_{s \in \mathcal{P}(\boldsymbol{\sigma}, p, p', R, L)} \inf_{m \in \mathcal{M}, \boldsymbol{\rho} \in \mathbb{N}^d} \mathbb{E}_s \left[\| s - \hat{s}_{(m, \boldsymbol{\rho})} \|_2^2 \right] \\
&\leq C(d, \boldsymbol{\sigma}, p, L) \left(R n^{-H(\boldsymbol{\sigma})/d} \right)^{2d/(d + 2H(\boldsymbol{\sigma}))} \\
&\leq C(d, \boldsymbol{\sigma}, p, L) \inf_{\hat{s}} \sup_{s \in \mathcal{P}(\boldsymbol{\sigma}, p, p', R, L)} \mathbb{E}_s \left[\| s - \hat{s} \|_2^2 \right]
\end{aligned}$$

where the last infimum is taken over all the estimators $\hat{s}$ of s.

Proof: Let us fix $\boldsymbol{\sigma}, p, p', R, L$ satisfying the assumptions of Proposition 4 and choose $\boldsymbol{\rho} = \lfloor \boldsymbol{\sigma} \rfloor + 1$. For all $s \in \mathcal{P}(\boldsymbol{\sigma}, p, p', R, L)$, we deduce from Proposition 3 and Theorem 1 that

$$\inf_{m \in \mathcal{M}, \boldsymbol{\rho} \in \mathbb{N}^d} \mathbb{E}_s \left[\| s - \hat{s}_{(m, \boldsymbol{\rho})} \|_2^2 \right] \leq C(d, \boldsymbol{\sigma}, p, L) \inf_{k \in \mathbb{N}} \left\{ R^2 2^{-2kH(\boldsymbol{\sigma})} + \frac{2^{kd}}{n} \right\}.$$

We then choose $k_\star$ as the greatest integer $k \in \mathbb{N}$ such that $2^{kd}/n \leq R^2 2^{-2kH(\boldsymbol{\sigma})}$, *i.e.* such that $2^k \leq (nR^2)^{1/(d + 2H(\boldsymbol{\sigma}))}$ so as to bound the infimum on the right-hand side, which provides the first inequality in Proposition 4.

Let us define the Besov class $\mathcal{B}(\boldsymbol{\sigma}, p, p', R, L)$ of all the probability densities s with respect to λ_d such that $|s|_{\boldsymbol{\sigma}, p, p'} \leq R$ (where $|.|_{\boldsymbol{\sigma}, p, p'}$ is defined in Section 2.3) and $\|s\|_\infty \leq L$. We deduce from Proposition 2 that, for $0 < p \leq 1$ or $p \geq 2$, there exists some positive real $C(\boldsymbol{\sigma}, p)$ such that $\mathcal{P}(\boldsymbol{\sigma}, p, \infty, R, L)$ contains $\mathcal{B}(\boldsymbol{\sigma}, \infty, \infty, C(\boldsymbol{\sigma}, p)R, L)$, and, for $1 < p < 2$, there exists some positive real $C(\boldsymbol{\sigma}, p)$ such that $\mathcal{P}(\boldsymbol{\sigma}, p, p, R, L)$ contains $\mathcal{B}(\boldsymbol{\sigma}, p, p, C(\boldsymbol{\sigma}, p)R, L)$. Besides, according to Triebel [Tri11] (Proposition 10), for all $\epsilon > 0$, the Kolmogorov ϵ-entropy in $\mathbb{L}_2([0,1]^d)$ of the Besov space $\mathscr{B}_q^{\boldsymbol{\sigma}}(\mathbb{L}_q([0,1]^d))$ is $\epsilon^{-H(\boldsymbol{\sigma})/d}$ for $H(\boldsymbol{\sigma})/d > (1/q - 1/2)_+$. Thus, the second inequality in Proposition 4 follows from the lower-bounds for minimax risks proved in [YB99] (Proposition 1, *ii*)). ■

In the sequel, our problem will thus be to build a statistical procedure that requires no prior knowldege on s but whose risk behaves almost as $\inf_{m \in \mathcal{M}, \boldsymbol{\rho} \in \mathbb{N}^d} \mathbb{E}_s \left[\| s - \hat{s}_{(m, \boldsymbol{\rho})} \|_2^2 \right]$.

3.3 Dyadic piecewise polynomial selection

Let us fix $\boldsymbol{r}_\star \in \mathbb{N}^d$, $J_\star \in \mathbb{N}$, and denote by $\mathcal{M}_\star$ the set of all partitions of $[0,1]^d$ into dyadic rectangles with sidelengths at least $2^{-J_\star}$. We consider the family $\boldsymbol{\mathcal{M}_\star^{deg}}$ of all couples $(m, \boldsymbol{\rho})$ with $m \in \mathcal{M}_\star$ and $\boldsymbol{\rho} = (\boldsymbol{\rho}_K)_{K \in m}$ such that, for all $K \in m$, $\boldsymbol{\rho}_K \in \Lambda(\boldsymbol{r}_\star)$. Ideally, we would like to choose the couple $(m, \boldsymbol{\rho})$ that minimizes $\mathbb{E}_s \left[\| s - \hat{s}_{(m, \boldsymbol{\rho})} \|_2^2 \right]$ among the elements of $\boldsymbol{\mathcal{M}_\star^{deg}}$. This is hopeless without knowing s, but from Pythagora's Equality and

Proposition (3.2), we have, for all $(m, \boldsymbol{\rho}) \in \boldsymbol{\mathcal{M}}_{\star}^{\boldsymbol{deg}}$,

$$\mathbb{E}_s\left[\|s - \hat{s}_{(m,\boldsymbol{\rho})}\|_2^2\right] - \|s\|_2^2 = -\|s_{(m,\boldsymbol{\rho})}\|_2^2 + \frac{1}{n}\sum_{K\in m}\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}_K)} \mathrm{Var}_s(\Phi_{K,\boldsymbol{k}}(Y_1)).$$

Thus, we propose to select an adequate partition $\hat{m}$ and the associated sequence of maximal degrees $\hat{\boldsymbol{\rho}} = (\hat{\boldsymbol{\rho}}_K)_{K\in\hat{m}}$ from the data so that

$$\begin{aligned}(\hat{m}, \hat{\boldsymbol{\rho}}) &= \operatorname*{argmin}_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_{\star}^{\boldsymbol{deg}}} \{-\|\hat{s}_{(m,\boldsymbol{\rho})}\|_2^2 + \mathrm{pen}(m, \boldsymbol{\rho})\}\\ &= \operatorname*{argmin}_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_{\star}^{\boldsymbol{deg}}} \{\gamma(\hat{s}_{(m,\boldsymbol{\rho})}) + \mathrm{pen}(m, \boldsymbol{\rho})\}\end{aligned}$$

where pen : $\boldsymbol{\mathcal{M}}_{\star}^{\boldsymbol{deg}} \to \mathbb{R}^+$ is a so-called penalty function. We then estimate the density s by

$$\tilde{s} = \hat{s}_{(\hat{m},\hat{\boldsymbol{\rho}})}.$$

According to the proof of Proposition 4, in view of proving the adaptivity of the penalized estimator $\tilde{s}$, the penalty pen should be chosen so that $\tilde{s}$ satisfies an inequality akin to

$$\mathbb{E}_s[\|s - \tilde{s}\|_2^2] \le C \min_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_{\star}^{\boldsymbol{deg}}} \left\{\|s - s_{(m,\boldsymbol{\rho})}\|_2^2 + \frac{\dim(S_{(m,\boldsymbol{\rho})})}{n}\right\} \tag{14}$$

where C is a positive real that does not depend on n.

In order to define an adequate form of penalty, we introduce the set $\mathcal{D}_\star$ of all dyadic rectangles of $[0,1]^d$ with sidelengths $\ge 2^{-J_\star}$ and, for all $K \in \mathcal{D}_\star$ and $\boldsymbol{k} \in \Lambda(\boldsymbol{r}_\star)$, we set

$$\hat{\sigma}_{K,\boldsymbol{k}}^2 = \frac{1}{n(n-1)}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\Phi_{K,\boldsymbol{k}}(Y_i) - \Phi_{K,\boldsymbol{k}}(Y_j)\right)^2,$$

which is an unbiased estimator of $\mathrm{Var}_s(\Phi_{K,\boldsymbol{k}}(Y_1))$. We also set

$$\widehat{M}_{1,\star} = \frac{1}{n}\max_{K\in\mathcal{D}_\star}\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{r}_\star)}\sqrt{\frac{\pi(\boldsymbol{k})}{\lambda_d(K)}}\left|\sum_{i=1}^{n}\Phi_{K,\boldsymbol{k}}(Y_i)\right| \quad\text{and}\quad \widehat{M}_{2,\star} = \frac{1}{n}\max_{K\in\mathcal{D}_\star}\max_{\boldsymbol{k}\in\Lambda(\boldsymbol{r}_\star)}\sum_{i=1}^{n}\Phi_{K,\boldsymbol{k}}^2(Y_i),$$

that overestimate respectively

$$\max_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_{\star}^{\boldsymbol{deg}}}\|s_{(m,\boldsymbol{\rho})}\|_\infty \quad\text{and}\quad \max_{K\in\mathcal{D}_\star}\max_{\boldsymbol{k}\in\Lambda(\boldsymbol{r}_\star)}\mathbb{E}_s\left[\Phi_{K,\boldsymbol{k}}^2(Y_1)\right].$$

The following theorem suggests a form of penalty yielding an inequality close to (14).

Theorem 3 *Let $\boldsymbol{r}_\star \in \mathbb{N}^d$ and $J_\star \in \mathbb{N}$ be such that $|\Lambda(\boldsymbol{r}_\star)| \leq \max\{\exp(n)/n, n^d\}$ and $2^{dJ_\star} \leq n/\log(n|\Lambda(\boldsymbol{r}_\star)|)$. Let $(L_{(m,\boldsymbol{\rho})})_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}}$ be a family of nonnegative real numbers, that may depend on n, satisfying*

$$\sum_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}} \exp(-L_{(m,\boldsymbol{\rho})}|m|) \leq 1. \tag{15}$$

If s is bounded and pen *is defined on $\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$ by*

$$\begin{aligned}\mathrm{pen}(m,\boldsymbol{\rho}) = \frac{1}{n}\sum_{K\in m}\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}_K)} &\left(\kappa_1\hat{\sigma}^2_{K,\boldsymbol{k}} + \kappa_2\pi(\boldsymbol{k})\right)\\ &+ \left(\left(\kappa_3\widehat{M}_{2,\star} + \kappa_4\pi(\boldsymbol{r}_\star)\right)|\Lambda(\boldsymbol{r}_\star)| + \kappa_5\widehat{M}_{1,\star}\right)\frac{L_{(m,\boldsymbol{\rho})}|m|}{n}\end{aligned}$$

where $\kappa_1,\ldots,\kappa_5$ are large enough positive constants, then

$$\begin{aligned}\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\right] \leq \min_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}} &\left\{\kappa'_1\|s-s_{(m,\boldsymbol{\rho})}\|_2^2 + \kappa'_2\frac{1}{n}\sum_{K\in m}\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}_K)}\mathrm{Var}_s(\Phi_{K,\boldsymbol{k}}(Y_1))\right.\\ &\left. + \kappa'_3\pi(\boldsymbol{r}_\star)\frac{\dim(S_{(m,\boldsymbol{\rho})})}{n} + \kappa'_4\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\|s\|_\infty\frac{L_{(m,\boldsymbol{\rho})}|m|}{n}\right\}\\ &+ \kappa'_5\|s\|_\infty^2\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\frac{1}{n}.\end{aligned}$$

where $\kappa'_1,\ldots,\kappa'_5$ are positive reals, $\kappa'_1,\ldots,\kappa'_4$ only depend on $\kappa_1,\ldots,\kappa_5$, and κ'_5 also depends on d.

Thus, the penalty associated to each $(m,\boldsymbol{\rho}) \in \boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$ is composed of two terms: an additive term that overestimates the variance over the model $S_{(m,\boldsymbol{\rho})}$, and a term linear in the size of the partition m, up to the weight $L_{(m,\boldsymbol{\rho})}$, that overestimates the upper-bound given in Proposition 3 for the variance over $S_{(m,\boldsymbol{\rho})}$. There remains to choose those weights under the constraint (15). According to Proposition 5 below, each model in $\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$ can be assigned the same weight that only depends on d and $\boldsymbol{r}_\star$.

Proposition 5 *If $\kappa_1,\ldots,\kappa_5$ are large enough positive constants, then the penalty defined on $\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$ by*

$$\begin{aligned}\mathrm{pen}(m,\boldsymbol{\rho}) = \frac{1}{n}\sum_{K\in m}\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}_K)} &\left(\kappa_1\hat{\sigma}^2_{K,\boldsymbol{k}} + \kappa_2\pi(\boldsymbol{k})\right)\\ &+ \left(\left(\kappa_3\widehat{M}_{2,\star} + \kappa_4\pi(\boldsymbol{r}_\star)\right)|\Lambda(\boldsymbol{r}_\star)| + \kappa_5\widehat{M}_{1,\star}\right)\frac{\log(8d|\Lambda(\boldsymbol{r}_\star)|)|m|}{n}\end{aligned} \tag{16}$$

satisfies the assumptions of Theorem 3. Moreover, if $|\Lambda(\boldsymbol{r}_\star)| \le \max\{\exp(n)/n, n^d\}$, $2^{dJ_\star} \le n/\log(n|\Lambda(\boldsymbol{r}_\star)|)$ *and* s *is bounded, then for such a penalty*

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\right] \le \kappa'' \min_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_\star^{deg}} \left\{\|s-s_{(m,\boldsymbol{\rho})}\|_2^2 + \pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\|s\|_\infty^2 \frac{\log(8ed|\Lambda(\boldsymbol{r}_\star)|)|m|}{n}\right\}.$$

where κ'' *is a positive real that only depends on* $\kappa_1, \ldots, \kappa_5$ *and* d.

Proof: First, for all $D \in \mathbb{N}^\star$, the number of partitions of $[0,1]^d$ into D dyadic rectangles satisfies

$$|\mathcal{M}_D| \le (4d)^D. \tag{17}$$

Indeed, as illustrated by Figure 2, each partition in $\mathcal{M}_D$ can be described by a complete dyadic tree with D leaves whose edges are labeled with a sequence of $D-1$ integers in $\{1, \ldots, d\}$ giving the cutting directions to obtain the partition from the unit square.

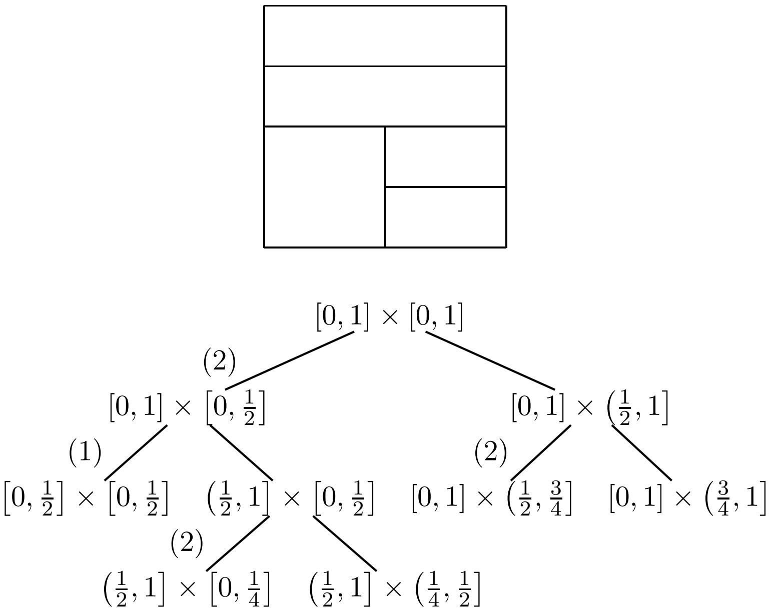
$[0,1]\times[0,1]$
(2)
$[0,1]\times[0,\frac{1}{2}]$
$[0,1]\times(\frac{1}{2},1]$
(1)
(2)
$[0,\frac{1}{2}]\times[0,\frac{1}{2}]$
$(\frac{1}{2},1]\times[0,\frac{1}{2}]$
$[0,1]\times(\frac{1}{2},\frac{3}{4}]$
$[0,1]\times(\frac{3}{4},1]$
(2)
$(\frac{1}{2},1]\times[0,\frac{1}{4}]$
$(\frac{1}{2},1]\times(\frac{1}{4},\frac{1}{2}]$

Figure 2: Top: Partition of $[0,1]^2$ into dyadic rectangles. Bottom: Binary tree labeled with the sequence of cutting directions $(2,1,2,2)$ corresponding with that partition.

The number of complete dyadic trees with D leaves is given by the Catalan number

$$\frac{1}{D}\binom{2(D-1)}{D-1} \le 4^D,$$

hence (17). We deduce from (17) that, for all positive real L,

$$\begin{aligned}\sum_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}} \exp(-L|m|) &\le \sum_{D\in\mathbb{N}^\star}\sum_{m\in\mathcal{M}_D}\sum_{\boldsymbol{\rho}\in\Lambda(\boldsymbol{r}_\star)^D}\exp(-L|m|)\\ &\le \sum_{D\in\mathbb{N}^\star}(4d|\Lambda(\boldsymbol{r}_\star)|)^D\exp(-LD)\\ &\le 1/\left(\exp\left(L-\log(4d|\Lambda(\boldsymbol{r}_\star)|)\right)-1\right)\end{aligned}$$

So, we can choose $L \ge \log(8d|\Lambda(\boldsymbol{r}_\star)|)$ for Condition (15) to be fulfilled.

Since $\|s\|_\infty \ge 1$, the upper-bound for $\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\right]$ is then a straightforward consequence of Theorem 3. ■

It is worth pointing out that penalty (16) is more refined than the penalties proposed by [Kle09] or [AD10] for density estimation via dyadic histogram selection based on a least-squares type criterion. Indeed, when $\boldsymbol{r}_\star$ is null, penalty (16) is not simply proportional to the dimension of the partition.

With a penalty chosen as above, we recover an inequality close to (14), that allows to prove the adaptivity of $\tilde{s}$ over a wide range of classes $\mathcal{P}(\boldsymbol{\sigma},p,p',R,L)$ as defined in Section 3.2. For that purpose, we introduce

$$q(d,\boldsymbol{\sigma},p)=\frac{\underline{\boldsymbol{\sigma}}}{H(\boldsymbol{\sigma})}\frac{d+2H(\boldsymbol{\sigma})}{H(\boldsymbol{\sigma})}\left(\frac{H(\boldsymbol{\sigma})}{d}-\left(\frac{1}{p}-\frac{1}{2}\right)_+\right)$$

and

$$w(\boldsymbol{r}_\star)=\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\log\left(8ed|\Lambda(\boldsymbol{r}_\star)|\right).$$

Theorem 4 *Let $\boldsymbol{r}_\star\in\mathbb{N}^d$ and $J_\star\in\mathbb{N}$ be such that $|\Lambda(\boldsymbol{r}_\star)|\le\max\{\exp(n)/n,n^d\}$ and $J_\star=\max\{J\in\mathbb{N}$ s.t. $2^{Jd}\le n/\log(n|\Lambda(\boldsymbol{r}_\star)|)\}$, and* pen *be the penalty given by Proposition 5. For all $p>0$, let $p'=\infty$ if $0<p\le 1$ or $p\ge 2$, and $p'=p$ if $1<p<2$. For all $L>0$, $\boldsymbol{\sigma}\in\prod_{l=1}^d(0,r_\star(l)+1)$, $p>0$ such that $H(\boldsymbol{\sigma})/d>(1/p-1/2)_+$ and $q(d,\boldsymbol{\sigma},p)>1$, for all R such that $w(\boldsymbol{r}_\star)/n\le R^2\le(n/\log(n|\Lambda(\boldsymbol{r}_\star)|))^{q(d,\boldsymbol{\sigma},p)-1}$,*

$$\sup_{s\in\mathcal{P}(\boldsymbol{\sigma},p,p',R,L)}\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\right]\le Cw(\boldsymbol{r}_\star)^{2H(\boldsymbol{\sigma})/(d+2H(\boldsymbol{\sigma}))}\inf_{\hat{s}}\sup_{s\in\mathcal{P}(\boldsymbol{\sigma},p,p',R,L)}\mathbb{E}_s\left[\|s-\hat{s}\|_2^2\right],$$

where C only depends on $d,\boldsymbol{\sigma},p,L$ and the penalty constants $\kappa_1,\ldots,\kappa_5$ and the above infimum is taken over all the estimators of s.

Thus, if $\boldsymbol{r}_\star$ is chosen as a constant with respect to n, then $\tilde{s}$ reaches the minimax risk, up to a constant factor, over a wide range of classes that contain functions with possibly anisotropic and inhomogeneous smoothness limited by the maximal degrees $\boldsymbol{r}_\star$. Another strategy consists in allowing the maximal degrees $\boldsymbol{r}_\star$ to increase with the sample size n, while $w(\boldsymbol{r}_\star)$ varies slowly with n. For instance, with $r_\star(l)=\log(n)$ for all $l=1,\ldots,d$, our estimator $\tilde{s}$ still approximately reaches the minimax risk over a range of classes all the wider as n increases. The price to pay is only a logarithmic factor, proportional to $(\log(\log(n))\log^{2d}(n))^{2H(\boldsymbol{\sigma})/(d+2H(\boldsymbol{\sigma}))}$ over classes with smoothness $H(\boldsymbol{\sigma})$. Thus, such a result may be seen as a nonasymptotic and multivariate counterpart of Theorem 1 in Willett and Nowak [WN07].

Remark: Contrary to [NvS97, Neu00, KLP01, Kle09], we have chosen here the smoothing parameter $J_\star$ independently of the smoothness of s, hence the restriction on $q(d,\boldsymbol{\sigma},p)$, that could disappear otherwise. Setting $\mu_{\boldsymbol{\sigma}} = H(\boldsymbol{\sigma})/\underline{\boldsymbol{\sigma}}$, the condition $q(d,\boldsymbol{\sigma},p) > 1$ is equivalent to $H(\boldsymbol{\sigma})/d > \nu(\boldsymbol{\sigma},p)$, where

$$\nu(\boldsymbol{\sigma},p) = \frac{1}{2}\left(\frac{1}{2}(\mu_{\boldsymbol{\sigma}}-1) + \left(\frac{1}{p}-\frac{1}{2}\right)_+ + \sqrt{\left(\frac{1}{2}(\mu_{\boldsymbol{\sigma}}-1) + \left(\frac{1}{p}-\frac{1}{2}\right)_+\right)^2 + 2\left(\frac{1}{p}-\frac{1}{2}\right)_+}\right).$$

In case of isotropic and homogeneous smoothness, *i.e.* when $\mu_{\boldsymbol{\sigma}} = 1$ and $p \geq 2$, $q(d,\boldsymbol{\sigma},p) > 1$ is simply equivalent to $H(\boldsymbol{\sigma})/d > 0$. In case of isotropic and inhomogeneous smoothness, *i.e.* when $\mu_{\boldsymbol{\sigma}} = 1$ and $p < 2$, $q(d,\boldsymbol{\sigma},p) > 1$ is equivalent to $H(\boldsymbol{\sigma})/d > \nu(\boldsymbol{\sigma},p)$ where $\nu(\boldsymbol{\sigma},p) \in (1/p-1/2, 1/p)$. This is slightly stronger than $H(\boldsymbol{\sigma})/d > 1/p-1/2$, but still better than the restriction $H(\boldsymbol{\sigma})/d > 1/p$ which is often encountered in the literature. Otherwise, $\nu(\boldsymbol{\sigma},p)$ increases with $\mu_{\boldsymbol{\sigma}}$ and $1/p$, *i.e.* with the anisotropy and the inhomogeneity.

3.4 Implementing the dyadic piecewise polynomial selection procedure

We end this article with a brief discussion about the implementation of our estimator $\tilde{s}$ for the penalty defined in Proposition 5. Let us fix the penalty constants $\kappa_1,\ldots,\kappa_5$ and set, for all dyadic rectangle $K \in \mathcal{D}_\star$ and all $\boldsymbol{r} \in \Lambda(\boldsymbol{r}_\star)$,

$$\begin{aligned}\widehat{W}(K,\boldsymbol{r}) = \sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{r})}\left(-\left(\frac{1}{n}\sum_{i=1}^n \Phi_{K,\boldsymbol{k}}(Y_i)\right)^2 + \kappa_1\frac{\hat{\sigma}^2_{K,\boldsymbol{k}}}{n} + \kappa_2\frac{\pi(\boldsymbol{k})}{n}\right)\\ + \frac{\log(8d|\Lambda(\boldsymbol{r}_\star)|)}{n}\left(\left(\kappa_3\widehat{M}_{2,\star} + \kappa_4\pi(\boldsymbol{r}_\star)\right)|\Lambda(\boldsymbol{r}_\star)| + \kappa_5\widehat{M}_{1,\star}\right)\end{aligned}$$

and

$$\hat{\boldsymbol{r}}_K = \underset{\boldsymbol{r}\in\Lambda(\boldsymbol{r}_\star)}{\operatorname{argmin}}\ \widehat{W}(K,\boldsymbol{r}).$$

Given the decomposition of $\hat{s}_{(m,\boldsymbol{\rho})}$ in the basis $(\Phi_{K,\boldsymbol{k}})_{K\in m,\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}_K)}$, the model $(\hat{m},\hat{\boldsymbol{\rho}})$ to select in $\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$ is characterized by

$$(\hat{m},\hat{\boldsymbol{\rho}}) = \underset{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}}{\operatorname{argmin}} \sum_{K\in m}\widehat{W}(K,\boldsymbol{\rho}_K),$$

so

$$\hat{m} = \underset{m\in\mathcal{M}_\star}{\operatorname{argmin}} \sum_{K\in m}\widehat{W}(K,\hat{\boldsymbol{r}}_K) \text{ and, for all } K\in\hat{m}, \hat{\boldsymbol{\rho}}_K = \hat{\boldsymbol{r}}_K.$$

Thus, the steps leading to $\tilde{s}$ are

1. Compute $\widehat{M}_{1,\star}$ and $\widehat{M}_{2,\star}$.

2. For all $K \in \mathcal{D}_\star$ and all $\boldsymbol{k} \in \Lambda(\boldsymbol{r}_\star)$, compute $\hat{\sigma}^2_{K,\boldsymbol{k}}$.

3. For all $K \in \mathcal{D}_\star$, compute $\hat{\boldsymbol{r}}_K$ and $\widehat{W}(K, \hat{\boldsymbol{r}}_K)$.

4. Determine the best partition $\hat{m} = \operatorname{argmin}_{m \in \mathcal{M}_\star} \sum_{K \in m} \widehat{W}(K, \hat{\boldsymbol{r}}_K)$.

5. Set, for all $K \in \hat{m}$, $\hat{\boldsymbol{\rho}}_K = \hat{\boldsymbol{r}}_K$.

6. Compute $\tilde{s} = \hat{s}_{(\hat{m}, \hat{\boldsymbol{\rho}})}$.

Since $\hat{m}$ is the partition in $\mathcal{M}_\star$ that minimizes a given additive criterion, it can be determined via the algorithm inspired from Donoho [Don97] and described in [BSR04] (beginning of Section 3), with a computational complexity at most of order $\mathcal{O}(|\mathcal{D}_\star|)$. Therefore, one easily verifies that the whole steps only require a computational complexity at most of order $\mathcal{O}(|\Lambda(\boldsymbol{r}_\star)||\mathcal{D}_\star|)$. Since $|\mathcal{D}_\star| = (2^{J_\star+1} - 1)^d$, if we choose $J_\star$ as prescribed by Theorem 3, then determining $\tilde{s}$ requires at most $\mathcal{O}(n)$ computations when $\boldsymbol{r}_\star$ is constant, and at most $\mathcal{O}(n \log^d(n))$ when $r_\star(l) = \log(n)$ for all $l = 1, \dots, d$. Last, regarding the choice of the penalty constants $\kappa_1, \dots, \kappa_5$, they can be calibrated via simulations over a wide collection of test densities. Such a method has already proved to yield good results in practice, even though several constants have to be chosen, as shown for instance in [CR04].

4 Proofs of the approximation results

For $j \in \mathbb{N}$ and $K \in \mathcal{D}^{\boldsymbol{\sigma}}_j$, we recall that the children of K are all the dyadic rectangles of $\mathcal{D}^{\boldsymbol{\sigma}}_{j+1}$ that are included in K. We will also refer to K as the parent of its children and will often use the fact that the children of K form a partition of K into

$$\prod_{l=1}^{d} 2^{\lfloor (j+1)\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor - \lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor} \leq 2^d 2^{d\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \tag{18}$$

dyadic rectangles from $\mathcal{D}^{\boldsymbol{\sigma}}_{j+1}$.

In all the proofs, the notation $C(\theta)$ stands for a positive real that only depends on the parameter θ, and whose value is allowed to change from one occurrence to another.

4.1 Proof of Proposition 1

For $p \geq q$, Proposition 1 follows from the continuous embedding of $\mathbb{L}_p([0,1]^d)$ in $\mathbb{L}_q([0,1]^d)$. For $p < q$, it corresponds with the second point in the more general result below.

Proposition 6 *Let $R > 0$, $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in \prod_{l=1}^{d}(0, r_l + 1)$, $1 \leq q \leq \infty$ and $0 < p < q$ such that $H(\boldsymbol{\sigma})/d > 1/p - 1/q$. For $s \in \mathbb{L}_p([0,1]^d)$, $k \in \mathbb{N}$, and any dyadic rectangle $K \in \mathcal{D}_k^{\boldsymbol{\sigma}}$, we set*

$$e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s, K) = \inf_{P \in \Pi_k^{\boldsymbol{r},\boldsymbol{\sigma}}} \|(s - P)\mathbb{1}_K\|_p.$$

If $N_{\boldsymbol{r},\boldsymbol{\sigma},p,\infty}(s) \leq R$, then

i) for all $j \in \mathbb{N}$ and all $K \in \mathcal{D}_j^{\boldsymbol{\sigma}}$,

$$\mathscr{E}_{\boldsymbol{r}}(s, K)_q \leq C(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) \sum_{k \geq j} 2^{-kd(H(\boldsymbol{\sigma})/d + 1/q - 1/p)\underline{\boldsymbol{\sigma}}/H(\sigma)} 2^{k\underline{\boldsymbol{\sigma}}} e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s, K). \tag{19}$$

ii) $s \in \mathbb{L}_q([0,1]^d)$ and $\|s\|_q \leq C(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q)(\|s\|_p + R)$.

Proof: Let us fix $1 \leq q \leq \infty$, $0 < p < q$, $j \in \mathbb{N}$ and $K \in \mathcal{D}_j^{\boldsymbol{\sigma}}$. For all $k \geq j$, we denote by $\mathcal{C}_k(K)$ the set of all rectangles from $\mathcal{D}_k^{\boldsymbol{\sigma}}$ that are included in K. Thus, $\mathcal{C}_j(K)$ is reduced to $\{K\}$, $\mathcal{C}_{j+1}(K)$ is the set of all the children of K, *etc.*... For any rectangle $I \subset [0,1]^d$, we denote by $P_I(s)$ a polynomial function on I with degree $\leq r_l$ in the l-th direction such that

$$\|(s - P_I(s))\mathbb{1}_I\|_p = \mathscr{E}_{\boldsymbol{r}}(s, I)_p,$$

where $\mathscr{E}_{\boldsymbol{r}}(s, I)_p$ is defined as in (4). For all $k \geq j$, we set

$$\Sigma_k(s, K) = \sum_{I \in \mathcal{C}_k(K)} P_I(s)\mathbb{1}_I$$

and, in order to alleviate the notation, we simply write $e_k(s, K)$ instead of $e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s, K)$ in the whole proof. It should be noticed that $e_k(s, [0,1]^d) = e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s)$ as defined by (7), and that

$$e_k(s, K) = \|(s - \Sigma_k(s, K))\mathbb{1}_K\|_p = \left(\sum_{I \in \mathcal{C}_k(K)} \mathscr{E}_{\boldsymbol{r}}^p(s, I)_p \right)^{1/p}.$$

Therefore,

$$\|(s - \Sigma_k(s, K))\mathbb{1}_K\|_p \leq \left(\sum_{I \in \mathcal{D}_k^{\boldsymbol{\sigma}}} \mathscr{E}_{\boldsymbol{r}}^p(s, I)_p \right)^{1/p} = e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s) \leq 2^{-k\underline{\boldsymbol{\sigma}}} R$$

so that the sequence $(\Sigma_k(s, K))_{k \geq j}$ converges to $s\mathbb{1}_K$ in $\mathbb{L}_p([0,1]^d)$.

Let us prove that $(\Sigma_k(s, K))_{k \geq j}$ also converges to $s\mathbb{1}_K$ in $\mathbb{L}_q([0,1]^d)$. We now fix $k \geq j$. When $0 < p < q \leq \infty$ as assumed here, Markov Inequality for polynomials asserts that, for all rectangle I of $[0,1]^d$, and all polynomial function $P \in \mathscr{P}_{\boldsymbol{r}}$,

$$\|P\mathbb{1}_I\|_q \leq C(d, \boldsymbol{r}, p, q)(\lambda_d(I))^{(1/q - 1/p)} \|P\mathbb{1}_I\|_p. \tag{20}$$

We refer to Lemma 5.1 in [Hoc02a] for a proof (that still holds for $q = \infty$). Let us first assume that $0 < p < q < \infty$. We then deduce from (20) that

$$\begin{aligned}
&\|\Sigma_{k+1}(s,K) - \Sigma_k(s,K)\|_q^q \\
&= \sum_{I \in \mathcal{C}_{k+1}(K)} \|(\Sigma_{k+1}(s,K) - \Sigma_k(s,K))\mathbb{1}_I\|_q^q \\
&\leq C(d,\boldsymbol{r},p,q) \sum_{I \in \mathcal{C}_{k+1}(K)} (\lambda_d(I))^{q(1/q-1/p)} \|(\Sigma_{k+1}(s,K) - \Sigma_k(s,K))\mathbb{1}_I\|_p^q \\
&\leq C(d,\boldsymbol{r},p,q) 2^{q(k+1)d(1/p-1/q)\underline{\boldsymbol{\sigma}}/H(\sigma)} \sum_{I \in \mathcal{C}_{k+1}(K)} \|(\Sigma_{k+1}(s,K) - \Sigma_k(s,K))\mathbb{1}_I\|_p^q. \qquad (21)
\end{aligned}$$

Let us also fix $I \in \mathcal{C}_{k+1}(K)$. Then

$$(\Sigma_{k+1}(s,K) - \Sigma_k(s,K))\mathbb{1}_I = (P_I(s) - P_{\tilde{I}}(s))\mathbb{1}_I$$

where $\tilde{I} \in \mathcal{C}_k(K)$ is the parent of I. Let $\kappa(p) = 2^{1/p}$ if $p < 1$, and $\kappa(p) = 1$ otherwise. From the (quasi-)triangle inequality satisfied by $\|.\|_p$, we then get

$$\begin{aligned}
\|(\Sigma_{k+1}(s,K) - \Sigma_k(s,K))\mathbb{1}_I\|_p &\leq \kappa(p) \left(\|(s - P_I(s))\mathbb{1}_I\|_p + \|(s - P_{\tilde{I}}(s))\mathbb{1}_I\|_p\right) \\
&\leq \kappa(p) \left(\mathcal{E}_{\boldsymbol{r}}(s,I)_p + \mathcal{E}_{\boldsymbol{r}}(s,\tilde{I})_p\right),
\end{aligned}$$

hence, by convexity of $x \mapsto x^q$,

$$\|(\Sigma_{k+1}(s,K) - \Sigma_k(s,K))\mathbb{1}_I\|_p^q \leq 2^{q-1}\kappa^q(p) \left(\mathcal{E}_{\boldsymbol{r}}^q(s,I)_p + \mathcal{E}_{\boldsymbol{r}}^q(s,\tilde{I})_p\right).$$

By grouping all the rectangles $I \in \mathcal{C}_{k+1}(K)$ that have the same parent, we obtain

$$\begin{aligned}
&\textstyle\sum_{I \in \mathcal{C}_{k+1}(K)} \|(\Sigma_{k+1}(s,K) - \Sigma_k(s,K))\mathbb{1}_I\|_p^q \\
&\leq 2^{q-1}\kappa^q(p) \left(\textstyle\sum_{I \in \mathcal{C}_{k+1}(K)} \mathcal{E}_{\boldsymbol{r}}^q(s,I)_p + 2^{d(1+\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))} \sum_{\tilde{I} \in \mathcal{C}_k(K)} \mathcal{E}_{\boldsymbol{r}}^q(s,\tilde{I})_p\right).
\end{aligned}$$

The classical inequality between ℓ_p and ℓ_q-(quasi-)norms

$$\left(\sum_i |a_i|^q\right)^{1/q} \leq \left(\sum_i |a_i|^p\right)^{1/p}, \text{ for } 0 < p \leq q < \infty \qquad (22)$$

then provides

$$\sum_{I \in \mathcal{C}_{k+1}(K)} \|(\Sigma_{k+1}(s,K) - \Sigma_k(s,K))\mathbb{1}_I\|_p^q \leq 2^{q-1}\kappa^q(p) \left(e_{k+1}^q(s,K) + 2^{d(1+\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))} e_k^q(s,K)\right).$$

Since $(e_k(s,K))_k \in \mathbb{N}$ is a decreasing sequence, by setting $\tau = H(\boldsymbol{\sigma})/d + 1/q - 1/p$ and combining Inequality (21) with the above inequality, we obtain

$$\begin{aligned}\|\Sigma_{k+1}(s,K) - \Sigma_k(s,K)\|_q &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)\left(2^{k\underline{\boldsymbol{\sigma}}}e_k(s,K)\right)2^{-kd\tau\underline{\boldsymbol{\sigma}}/H(\sigma)} \\ &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)R2^{-kd\tau\underline{\boldsymbol{\sigma}}/H(\sigma)}.\end{aligned}$$

We can prove in the same way that such an upper-bound still holds for $q=\infty$. Since $\tau > 0$, for all $0 < p < q \leq \infty$, $(\Sigma_k(s,K))_{k\geq j}$ also converges in $\mathbb{L}_q([0,1]^d)$ to $s\mathbb{1}_K$. In particular, we have thus proved that $s \in \mathbb{L}_q([0,1]^d)$.

From the definition of $\mathcal{E}_{\boldsymbol{r}}(s,K)_q$ and the triangle inequality, it follows that

$$\begin{aligned}\mathcal{E}_{\boldsymbol{r}}(s,K)_q &\leq \|(s - P_K(s))\mathbb{1}_K\|_q \\ &\leq \sum_{k\geq j}\|\Sigma_{k+1}(s,K) - \Sigma_k(s,K)\|_q \\ &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)\sum_{k\geq j}2^{-kd\tau\underline{\boldsymbol{\sigma}}/H(\sigma)}2^{k\underline{\boldsymbol{\sigma}}}e_k(s,K). \qquad (23)\end{aligned}$$

We have thus proved (19), and the above inequality for $K=[0,1]^d$ combined with Markov Inequality (20) also provides

$$\begin{aligned}\|s\|_q &\leq \|P_{[0,1]^d}(s)\|_q + \|s - P_{[0,1]^d}(s)\|_q \\ &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)(\|P_{[0,1]^d}(s)\|_p + R) \\ &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)(\|s\|_p + \mathcal{E}_{\boldsymbol{r}}(s,[0,1]^d)_p + R) \\ &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)(\|s\|_p + R).\end{aligned}$$

■

4.2 Proofs of Theorems 1 and 2

A first approximation result for the algorithm decsribed in Section 2 can be stated as follows.

Proposition 7 *Let $k \in \mathbb{N}$, $R > 0$, $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_d) \in \prod_{l=1}^d (0, r_l+1)$, $0 < p < \infty$, $1 \leq q \leq \infty$ and $s \in \mathbb{L}_q([0,1]^d)$. Assume that*

$$H(\boldsymbol{\sigma})/d > (1/p - 1/q)_+$$

and that

$$\sup_{j\in\mathbb{N}} 2^{jd(\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))(H(\boldsymbol{\sigma})/d-(1/p-1/q)_+)}\left(\sum_{K\in\mathcal{D}_j^{\boldsymbol{\sigma}}}\mathcal{E}_{\boldsymbol{r}}^p(s,K)_q\right)^{1/p} \leq R. \qquad (24)$$

Then, there exists some partition m of $[0,1]^d$ that only contains dyadic rectangles from $\mathcal{D}^{\boldsymbol{\sigma}}$ and $s_{(m,\boldsymbol{r})} \in S_{(m,\boldsymbol{r})}$ such that

$$|m| \leq C_1(d, \boldsymbol{\sigma}, p) 2^{kd}$$

and

$$\|s - s_{(m,\boldsymbol{r})}\|_q \leq C_2(d, \boldsymbol{\sigma}, p, q) R 2^{-kH(\boldsymbol{\sigma})}.$$

Besides, if for some $J \in \mathbb{N}$, s is polynomial with coordinate degree $\leq r$ over each rectangle of $\mathcal{D}^{\boldsymbol{\sigma}}_J$, then m only contains dyadic rectangles from $\cup_{j=0}^{J} \mathcal{D}^{\boldsymbol{\sigma}}_j$.

Proof: For $k = 0$, we can just choose m as the trivial partition of $[0,1]^d$ and $s_{(m,\boldsymbol{r})}$ as the polynomial of best $\mathbb{L}_q$-approximation over $[0,1]^d$ in $\mathscr{P}_{\boldsymbol{r}}$. Indeed, we then have

$$\|s - s_{(m,\boldsymbol{r})}\|_q = \mathcal{E}_{\boldsymbol{r}}(s, [0,1]^d)_q \leq R,$$

where the last inequality follows from (24). Let us now fix $k \geq 1$, set

$$\tau = H(\boldsymbol{\sigma})/d - (1/p - 1/q)_+ \quad \text{and} \quad \lambda = 2^{(1+(1+\tau p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))d/p},$$

and choose

$$\epsilon = \lambda R 2^{-kd(\tau + 1/p)}.$$

If $\mathcal{I}(s, \epsilon)$ is trivial, then the upper-bound (5) provides

$$\|s - A(s, \epsilon)\|_q \leq \epsilon \leq \lambda R 2^{-kH(\boldsymbol{\sigma})}.$$

Let us now assume that $\mathcal{I}(s, \epsilon)$ is not trivial and fix $j \geq 1$ such that $\mathcal{I}(s, \epsilon) \cap \mathcal{D}^{\boldsymbol{\sigma}}_j$ is not empty. If $K \in \mathcal{I}(s, \epsilon) \cap \mathcal{D}^{\boldsymbol{\sigma}}_j$, then K is a child of a dyadic rectangle $\tilde{K} \in \mathcal{D}^{\boldsymbol{\sigma}}_{j-1}$ such that

$$\epsilon \leq \mathcal{E}_{\boldsymbol{r}}(s, \tilde{K})_q,$$

hence

$$\epsilon^p \leq 2^{-(j-1)dp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} 2^{(j-1)dp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \mathcal{E}^p_{\boldsymbol{r}}(s, \tilde{K})_q.$$

By grouping all the rectangles $K \in \mathcal{I}(s, \epsilon) \cap \mathcal{D}^{\boldsymbol{\sigma}}_j$ having the same parent in $\mathcal{D}^{\boldsymbol{\sigma}}_{j-1}$, and taking into account Remark (18), we obtain

$$|\mathcal{I}(s, \epsilon) \cap \mathcal{D}^{\boldsymbol{\sigma}}_j| \epsilon^p \leq 2^{d(1+(1+p\tau)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))} 2^{-jdp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} R^p.$$

Replacing ϵ by its value, we deduce that

$$|\mathcal{I}(s, \epsilon) \cap \mathcal{D}^{\boldsymbol{\sigma}}_j| \leq 2^{kd(1+p\tau)} 2^{-jdp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}. \tag{25}$$

Besides, for all $j \geq 1$,

$$|\mathcal{I}(s, \epsilon) \cap \mathcal{D}^{\boldsymbol{\sigma}}_j| \leq |\mathcal{D}^{\boldsymbol{\sigma}}_j| \leq 2^{jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}.$$

Let us denote by J the greatest integer $j \geq 1$ such that

$$2^{jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \leq 2^{kd(1+p\tau)} 2^{-jdp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})},$$

i.e. such that

$$2^{jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \leq 2^{kd}.$$

Since $\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}) \leq 1$, the last inequality is satisfied by $k \geq 1$ for instance, so that J is well-defined. Besides, J is characterized by

$$2^{Jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \leq 2^{kd} < 2^{(J+1)d\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}.$$

Therefore,

$$\begin{aligned}
|\mathcal{I}(s,\epsilon)| &= \sum_{j\geq 1} |\mathcal{I}(s,\epsilon) \cap \mathcal{D}_j^{\boldsymbol{\sigma}}| \\
&\leq \sum_{j=1}^{J} 2^{jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} + 2^{kd(1+p\tau)} \sum_{j\geq J+1} 2^{-jdp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \\
&\leq C_1(d,\boldsymbol{\sigma},p) 2^{kd}
\end{aligned}$$

where

$$C_1(d,\boldsymbol{\sigma},p) = \frac{2^{d\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}}{2^{d\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} - 1} + \frac{1}{1 - 2^{-dp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}}.$$

Moreover, we deduce from (5) that, if $1 \leq q < \infty$, then

$$\|s - A(s,\epsilon)\|_q \leq |\mathcal{I}(s,\epsilon)|^{1/q}\epsilon \leq C_1^{1/q}(d,\boldsymbol{\sigma},p) R 2^{-kH(\boldsymbol{\sigma})},$$

and we deduce from (6) that, if $q = \infty$, then

$$\|s - A(s,\epsilon)\|_\infty < \epsilon \leq \lambda R 2^{-kH(\boldsymbol{\sigma})}.$$

So Proposition 7 is satisfied for

$$C_2(d,\boldsymbol{\sigma},p,q) = \begin{cases} C_1^{1/q}(d,\boldsymbol{\sigma},p)\lambda & \text{if } 1 \leq q < \infty \\ \lambda & \text{if } q = \infty, \end{cases}$$

$m = \mathcal{I}(s,\epsilon)$ and $s_{(m,\boldsymbol{r})} = A(s,\epsilon)$. The last assertion in Proposition 7 is a straightforward consequence of the approximation algorithm. ■

The following lemma allows to link Assumption (24) with the (quasi-)semi-norm $N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'}$.

Lemma 1 *Let $R > 0$, $0 < p < \infty$, $1 \leq q \leq \infty$ and $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in \prod_{l=1}^{d}(0, r_l + 1)$ such that $H(\boldsymbol{\sigma})/d > (1/p - 1/q)_+$. Assume that $s \in \mathcal{S}(\boldsymbol{r}, \boldsymbol{\sigma}, p, p', R)$, where $p' = \infty$ if $0 < p \leq 1$ or $p \geq q$ and $p' = p$ if $1 < p < q$, then*

$$\sup_{j \in \mathbb{N}} 2^{jd(\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))(H(\boldsymbol{\sigma})/d - (1/p - 1/q)_+)} \left(\sum_{K \in \mathcal{D}_j^{\boldsymbol{\sigma}}} \mathcal{E}_{\boldsymbol{r}}^{p}(s, K)_q \right)^{1/p} \leq C(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) R. \tag{26}$$

Proof: If $p \geq q$, then the left-hand side of Inequality (26) is upper-bounded by

$$\sup_{j \in \mathbb{N}} 2^{j\underline{\boldsymbol{\sigma}}} \left(\sum_{K \in \mathcal{D}_j^{\boldsymbol{\sigma}}} \mathcal{E}_{\boldsymbol{r}}^{p}(s, K)_p \right)^{1/p} = \sup_{j \in \mathbb{N}} 2^{j\underline{\boldsymbol{\sigma}}} e_{\boldsymbol{r}, \boldsymbol{\sigma}, p, j}(s) \leq R.$$

Let us now assume that $p < q$ and set $\tau = H(\boldsymbol{\sigma})/d + 1/q - 1/p$. From Inequality (19) in Proposition 6, we deduce that

$$\begin{aligned} & 2^{jd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \left(\sum_{K \in \mathcal{D}_j^{\boldsymbol{\sigma}}} \mathcal{E}_{\boldsymbol{r}}^{p}(s, K)_q \right)^{1/p} \\ & \qquad \leq C(d, r, \boldsymbol{\sigma}, p, q) 2^{jd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \left(\sum_{K \in \mathcal{D}_j^{\boldsymbol{\sigma}}} \left(\sum_{k \geq j} 2^{-kd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} 2^{k\underline{\boldsymbol{\sigma}}} e_k(s, K) \right)^{p} \right)^{1/p}. \end{aligned} \tag{27}$$

If $0 < p \leq 1$, then the classical inequality between ℓ_p and ℓ_1-(quasi-)norms recalled in (22) leads to

$$\begin{aligned} & 2^{jd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \left(\sum_{K \in \mathcal{D}_j^{\boldsymbol{\sigma}}} \mathcal{E}_{\boldsymbol{r}}^{p}(s, K)_q \right)^{1/p} \\ & \leq C(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) 2^{jd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \left(\sum_{K \in \mathcal{D}_j^{\boldsymbol{\sigma}}} \sum_{k \geq j} 2^{-kpd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} 2^{kp\underline{\boldsymbol{\sigma}}} e_k^{p}(s, K) \right)^{1/p} \\ & \leq C(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) 2^{jd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \left(\sum_{k \geq j} 2^{-kpd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} 2^{kp\underline{\boldsymbol{\sigma}}} e_k^{p}(s, [0,1]^d) \right)^{1/p} \\ & \leq C(d, \boldsymbol{r}, \boldsymbol{\sigma}, p, q) \sup_{k \geq j} \left(2^{k\underline{\boldsymbol{\sigma}}} e_k(s, [0,1]^d) \right) 2^{jd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \left(\sum_{k \geq j} 2^{-kpd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \right)^{1/p} \end{aligned}$$

hence Inequality (26). If $1 < p < \infty$, then there exists $1 < p^\star < \infty$ such that $1/p+1/p^\star = 1$, so we obtain by applying Hölder inequality to (27) that

$$\begin{aligned}
&2^{jd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}\left(\sum_{K\in\mathcal{D}_j^{\boldsymbol{\sigma}}}\mathcal{E}_{\boldsymbol{r}}^p(s,K)_q\right)^{1/p}\\
&\le C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)2^{jd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}\left(\sum_{K\in\mathcal{D}_j^{\boldsymbol{\sigma}}}\left(\sum_{k\ge j}2^{-p^\star kd\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}\right)^{p/p^\star}\left(\sum_{k\ge j}2^{kp\underline{\boldsymbol{\sigma}}}e_k^p(s,K)\right)\right)^{1/p}\\
&\le C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)\left(\sum_{k\ge j}2^{kp\underline{\boldsymbol{\sigma}}}\sum_{K\in\mathcal{D}_j^{\boldsymbol{\sigma}}}e_k^p(s,K)\right)^{1/p}\\
&\le C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)\left(\sum_{k\ge j}2^{kp\underline{\boldsymbol{\sigma}}}e_k^p(s,[0,1]^d)\right)^{1/p}
\end{aligned}$$

hence Inequality (26). ■

Last, Lemma 2 provides an upper-bound for the linear approximation error of $\mathcal{S}(\boldsymbol{r},\boldsymbol{\sigma},p,p',R)$ by $\Pi_J^{\boldsymbol{r},\boldsymbol{\sigma}}$ in the $\mathbb{L}_q$-norm.

Lemma 2 *Let $R>0$, $0<p<\infty$, $1\le q\le\infty$, $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_d)\in\prod_{l=1}^d(0,r_l+1)$ such that $H(\boldsymbol{\sigma})/d>(1/p-1/q)_+$, and $\kappa(p)=2^{1/p}$ if $0<p\le 1$, and 1 otherwise. Assume that $s\in\mathcal{S}(\boldsymbol{r},\boldsymbol{\sigma},p,p',R)$ where $p'=\infty$ if $0<p\le 1$ or $p\ge q$, and $p'=p$ if $1<p<q$. Then, for all $J\in\mathbb{N}$, there exists a function $s_J\in\Pi_J^{\boldsymbol{r},\boldsymbol{\sigma}}$ such that $s_J\in\mathcal{S}(\boldsymbol{r},\boldsymbol{\sigma},p,p',2\kappa(p)R)$ and*

$$\|s-s_J\|_q\le C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)2^{-Jd(H(\boldsymbol{\sigma})/d-(1/p-1/q)_+)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}R. \tag{28}$$

Proof: For all $K\in\mathcal{D}_J^{\boldsymbol{\sigma}}$, we denote by $P_K(s)$ a polynomial function on K with degree $\le r_l$ in the l-th direction such that

$$\|(s-P_K(s))\mathbb{1}_I\|_p=\mathcal{E}_{\boldsymbol{r}}(s,K)_p,$$

and we set

$$s_J=\sum_{K\in\mathcal{D}_J^{\boldsymbol{\sigma}}}P_K(s)\mathbb{1}_K.$$

In order to alleviate the notation, we simply write $e_k(s)$ instead of $e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s)$, and $e_k(s,K)$ instead of $e_{\boldsymbol{r},\boldsymbol{\sigma},p,k}(s,K)$, as in the proof of Proposition 6.

Since $s_J \in \Pi_J^{\boldsymbol{r},\boldsymbol{\sigma}}$, $e_k(s_J) = 0$ for $k \geq J$. If $k < J$, then the (quasi-)triangle inequality, the definition of s_J and the inclusion $\Pi_k^{\boldsymbol{r},\boldsymbol{\sigma}} \subset \Pi_J^{\boldsymbol{r},\boldsymbol{\sigma}}$ provide successively

$$\begin{aligned} e_k(s_J) &\leq \kappa(p)\left(\|s - s_J\|_p + e_k(s)\right) \\ &\leq \kappa(p)\left(e_J(s) + e_k(s)\right) \\ &\leq 2\kappa(p) e_k(s). \end{aligned}$$

Therefore, $N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'}(s_J) \leq 2\kappa(p) N_{\boldsymbol{r},\boldsymbol{\sigma},p,p'}(s)$, so that $s_J \in \mathcal{S}(\boldsymbol{r},\boldsymbol{\sigma},p,p',2\kappa(p)R)$.

If $p \geq q$, then

$$\|s - s_J\|_q \leq \|s - s_J\|_p = \left(\sum_{K \in \mathcal{D}_J^{\boldsymbol{\sigma}}} \mathcal{E}_{\boldsymbol{r}}^p(s,K)_p\right)^{1/p} = e_{\boldsymbol{r},\boldsymbol{\sigma},p,J}(s) \leq 2^{-J\underline{\boldsymbol{\sigma}}} R.$$

If $p < q < \infty$, then we deduce from Inequality (23) in the proof of Proposition 6 and from Inequality (22) between ℓ_p and ℓ_q-(quasi-)norms that

$$\begin{aligned} \|s - s_J\|_q &= \left(\sum_{K \in \mathcal{D}_J^{\boldsymbol{\sigma}}} \|(s - P_K(s))\mathbb{1}_K\|_q^q\right)^{1/q} \\ &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q)\left(\sum_{K \in \mathcal{D}_J^{\boldsymbol{\sigma}}}\left(\sum_{k \geq J} 2^{-kd(H(\sigma)/d+1/q-1/p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} 2^{k\underline{\boldsymbol{\sigma}}} e_k(s,K)\right)^p\right)^{1/p}. \end{aligned}$$

We then obtain Inequality (28) either thanks to the inequality between ℓ_1 and ℓ_p-(quasi-)norms in case $0 \leq p \leq 1$, or thanks to Hölder Inequality otherwise. Last, if $q = \infty$, then we still deduce from Inequality (23) that

$$\begin{aligned} \|s - s_J\|_\infty &= \max_{K \in \mathcal{D}_J^{\boldsymbol{\sigma}}} \|(s - P_K(s))\mathbb{1}_K\|_\infty \\ &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q) \max_{K \in \mathcal{D}_J^{\boldsymbol{\sigma}}}\left(\sum_{k \geq J} 2^{-kd(H(\sigma)/d+1/q-1/p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} 2^{k\underline{\boldsymbol{\sigma}}} e_k(s,K)\right) \\ &\leq C(d,\boldsymbol{r},\boldsymbol{\sigma},p,q) 2^{-Jd(H(\sigma)/d+1/q-1/p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} R. \end{aligned}$$

■

Theorem 1 is then a straightforward consequence of Proposition 7 and Lemma 1. To prove Theorem 2, for each $J \in \mathbb{N}$, we just have to apply Proposition 7 and Lemma 1 to the function s_J given by Lemma 2 and use the triangle inequality

$$\inf_{t \in S_{(m,\boldsymbol{r})}} \|s - t\|_q \leq \|s - s_J\|_q + \inf_{t \in S_{(m,\boldsymbol{r})}} \|s_J - t\|_q$$

where m can be any partition of $[0,1]^d$ into dyadic rectangles.

5 Proof of Theorem 3

In the following proof, we denote by $(w_{(m,\boldsymbol{\rho})})_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}}$ a family of nonnegative reals and set $\Sigma = \sum_{(m,\boldsymbol{\rho})\in\boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}} \exp(-w_{(m,\boldsymbol{\rho})})$. We fix $(m,\boldsymbol{\rho}) \in \boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$ as well as some positive reals ζ, $\theta_1,\ldots,\theta_8$ such that $2\theta_1(1+\theta_2) < 1$ and $\theta_8 < 1$.

From the definition of $\tilde{s} = \hat{s}_{(\hat{m},\hat{\boldsymbol{\rho}})}$, it follows that

$$\gamma(\tilde{s}) + \mathrm{pen}(\hat{m},\hat{\boldsymbol{\rho}}) \leq \gamma(\hat{s}_{(m,\boldsymbol{\rho})}) + \mathrm{pen}(m,\boldsymbol{\rho}). \tag{29}$$

For all $t,u \in \mathbb{L}_2([0,1]^d)$,

$$\gamma(t) - \gamma(u) = \|s-t\|_2^2 - \|s-u\|_2^2 - 2\nu(t-u), \tag{30}$$

where

$$\nu(t) = \frac{1}{n}\sum_{i=1}^{n} \left(t(Y_i) - \langle t, s\rangle\right).$$

Besides, for all $(m',\boldsymbol{\rho}') \in \boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$, setting

$$\chi(m',\boldsymbol{\rho}') = \|s_{(m',\boldsymbol{\rho}')} - \hat{s}_{(m',\boldsymbol{\rho}')}\|_2,$$

we obtain by developing $s_{(m',\boldsymbol{\rho}')}$ and $\hat{s}_{(m',\boldsymbol{\rho}')}$ in the orthonormal basis $(\Phi_{K,\boldsymbol{k}})_{K\in m',\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}'_K)}$ and using the linearity of ν

$$\chi^2(m',\boldsymbol{\rho}') = \sum_{K\in m'}\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}'_K)} \nu^2(\Phi_{K,\boldsymbol{k}}) = \nu\left(\hat{s}_{(m',\boldsymbol{\rho}')} - s_{(m',\boldsymbol{\rho}')}\right). \tag{31}$$

From Equalities (30), (31), Pythagoras' Equality and the linearity of ν, we deduce

$$\gamma(\tilde{s}) - \gamma(\hat{s}_{(m,\boldsymbol{\rho})}) = \|s-\tilde{s}\|_2^2 - \|s - s_{(m,\boldsymbol{\rho})}\|_2^2 + \chi^2(m,\boldsymbol{\rho}) - 2\chi^2(\hat{m},\hat{\boldsymbol{\rho}}) - 2\nu\left(s_{(\hat{m},\hat{\boldsymbol{\rho}})} - s_{(m,\boldsymbol{\rho})}\right),$$

which, combined with Inequality (29), leads to

$$\begin{aligned}\|s-\tilde{s}\|_2^2 \leq \|s - s_{(m,\boldsymbol{\rho})}\|_2^2 &+ \mathrm{pen}(m,\boldsymbol{\rho}) - \chi^2(m,\boldsymbol{\rho}) \\ &+ 2\chi^2(\hat{m},\hat{\boldsymbol{\rho}}) + 2\nu\left(s_{(\hat{m},\hat{\boldsymbol{\rho}})} - s_{(m,\boldsymbol{\rho})}\right) - \mathrm{pen}(\hat{m},\hat{\boldsymbol{\rho}}).\end{aligned} \tag{32}$$

We shall now provide an upper-bound for the term $2\nu\left(s_{(\hat{m},\hat{\boldsymbol{\rho}})} - s_{(m,\boldsymbol{\rho})}\right)$ on an event with great probability. From Bernstein's Inequality, as stated for instance in [Mas07] (Section 2.2.3), for all bounded function $t : [0,1]^d \to \mathbb{R}$ and all $x > 0$,

$$\mathbb{P}_s\left(\nu(t) \geq \sqrt{2\mathbb{E}_s[t^2(Y_1)]\frac{x}{n}} + \frac{\|t\|_\infty}{3}\frac{x}{n}\right) \leq \exp(-x). \tag{33}$$

Let us fix $(m', \boldsymbol{\rho}') \in \boldsymbol{\mathcal{M}}_{\star}^{\boldsymbol{deg}}$ and apply Bernstein's Inequality to $t = s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})}$. Since $\Phi_{K,\boldsymbol{k}}$ has support K,

$$\|s_{(m',\boldsymbol{\rho}')}\|_\infty = \max_{K \in m'} \left\| \sum_{\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}'_K)} \langle s, \Phi_{K,\boldsymbol{k}} \rangle \Phi_{K,\boldsymbol{k}} \right\|_\infty \leq M_{1,\star}$$

where

$$M_{1,\star} = \max_{K \in \mathcal{D}_\star} \sum_{\boldsymbol{k} \in \Lambda(\boldsymbol{r}_\star)} \sqrt{\frac{\pi(\boldsymbol{k})}{\lambda_d(K)}} |\langle s, \Phi_{K,\boldsymbol{k}} \rangle|,$$

so

$$\|s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})}\|_\infty \leq 2M_{1,\star}.$$

Since $S_{(m,\boldsymbol{\rho})}$ and $S_{(m',\boldsymbol{\rho}')}$ are both subspaces of $S_{(m_\star,\boldsymbol{r}_\star)}$,

$$\begin{aligned} \mathbb{E}_s \left[\left(s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})} \right)^2 (Y_1) \right] &= \int_{[0,1]^d} s_{(m_\star,\boldsymbol{r}_\star)} \left(s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})} \right)^2 \mathrm{d}\lambda_d \\ &\leq M_{1,\star} \|s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})}\|_2^2. \end{aligned}$$

From (33), there exists a set $\Omega(m, \boldsymbol{\rho}, m', \boldsymbol{\rho}', \zeta)$ such that $\mathbb{P}_s \left(\Omega(m, \boldsymbol{\rho}, m', \boldsymbol{\rho}', \zeta) \right) \geq 1 - \exp(-(w_{(m',\boldsymbol{\rho}')} + \zeta))$ and over which

$$\nu \left(s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})} \right) \leq \sqrt{2M_{1,\star} \|s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})}\|_2^2 \frac{w_{(m',\boldsymbol{\rho}')} + \zeta}{n}} + \frac{2}{3} M_{1,\star} \frac{w_{(m',\boldsymbol{\rho}')} + \zeta}{n}.$$

We recall that, for all $a, b \geq 0$ and $\theta > 0$,

$$2ab \leq \theta a^2 + \theta^{-1} b^2. \tag{34}$$

Thus, on $\Omega(m, \boldsymbol{\rho}, m', \boldsymbol{\rho}', \zeta)$, we have

$$\nu \left(s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})} \right) \leq \theta_1 \|s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})}\|_2^2 + \left(2/3 + \theta_1^{-1} \right) M_{1,\star} \frac{w_{(m',\boldsymbol{\rho}')} + \zeta}{n}.$$

Besides, using the triangle inequality, (34), and Pythagoras' Equality, we obtain

$$\begin{aligned} \|s_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})}\|_2^2 &\leq \left(\|s - s_{(m',\boldsymbol{\rho}')}\|_2 + \|s - s_{(m,\boldsymbol{\rho})}\|_2 \right)^2 \\ &\leq (1 + \theta_2) \|s - s_{(m',\boldsymbol{\rho}')}\|_2^2 + \left(1 + \theta_2^{-1} \right) \|s - s_{(m,\boldsymbol{\rho})}\|_2^2 \\ &\leq (1 + \theta_2) \|s - \hat{s}_{(m',\boldsymbol{\rho}')}\|_2^2 - (1 + \theta_2) \chi^2(m', \boldsymbol{\rho}') + \left(1 + \theta_2^{-1} \right) \|s - s_{(m,\boldsymbol{\rho})}\|_2^2. \end{aligned}$$

Therefore, the set $\Omega_{(m,\boldsymbol{\rho})}(\zeta) = \cap_{(m',\boldsymbol{\rho}') \in \boldsymbol{\mathcal{M}}_{\star}^{\boldsymbol{deg}}} \Omega(m, \boldsymbol{\rho}, m', \boldsymbol{\rho}', \zeta)$ is an event with probability

$$\mathbb{P}_s \left(\Omega_{(m,\boldsymbol{\rho})}(\zeta) \right) \geq 1 - \exp(-\zeta) \Sigma \tag{35}$$

over which

$$2\nu\left(s_{(\hat{m},\hat{\boldsymbol{\rho}})} - s_{(m,\boldsymbol{\rho})}\right) \leq 2\theta_1(1+\theta_2)\|s-\tilde{s}\|_2^2 + 2\theta_1\left(1+\theta_2^{-1}\right)\|s - s_{(m,\boldsymbol{\rho})}\|_2^2 \\ - 2\theta_1(1+\theta_2)\chi^2(\hat{m},\hat{\boldsymbol{\rho}}) + 2\left(2/3+\theta_1^{-1}\right)M_{1,\star}\frac{w_{(\hat{m},\hat{\boldsymbol{\rho}})}+\zeta}{n}. \quad (36)$$

Let us now provide a concentration inequality for $\chi^2(\hat{m},\hat{\boldsymbol{\rho}})$. For that purpose, we first prove the following result.

Proposition 8 *Let $(m',\boldsymbol{\rho}') \in \boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}, x > 0,$*

$$V_{(m',\boldsymbol{\rho}')} = \frac{1}{n}\sum_{K\in m'}\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}'_K)} \mathrm{Var}_s\left(\Phi_{K,\boldsymbol{k}}^2(Y_1)\right) = \mathbb{E}_s\left[\|\hat{s}_{(m',\boldsymbol{\rho}')} - s_{(m,\boldsymbol{\rho})}\|_2^2\right]$$

and

$$M_{2,\star} = \max_{K\in\mathcal{D}_\star,\boldsymbol{k}\in\Lambda(\boldsymbol{r}_\star)} \mathbb{E}_s[\Phi_{K,\boldsymbol{k}}^2(Y_1)].$$

There exist an event $\Omega_\star$ that does not depend on $(m',\boldsymbol{\rho}')$ and an event $\Omega_{(m',\boldsymbol{\rho}')}(x)$ such that

$$\mathbb{P}_s(\Omega_\star^c) \leq 2^{d+1}/(n^2\log(n)), \quad (37)$$

$$\mathbb{P}_s(\Omega_{(m',\boldsymbol{\rho}')}^c(x)) \leq \exp(-x),$$

and, on $\Omega_{(m',\boldsymbol{\rho}')}(x)$,

$$\chi^2(m',\boldsymbol{\rho}')\mathbb{1}_{\Omega_\star} \leq (1+\theta_3)(1+\theta_4)V_{(m',\boldsymbol{\rho}')} \\ + 4\left(1+\theta_4^{-1}\right)|\Lambda(\boldsymbol{r}_\star)|\left(\left(4/3+\theta_3^{-1}\right)M_{2,\star} + (5/3)\left(1+3\theta_3^{-1}\right)\pi(\boldsymbol{r}_\star)\right)\frac{x}{n}.$$

Proof: Let us fix $x > 0$, and set, for all $K \in \mathcal{D}_\star$ and $\boldsymbol{k} \in \Lambda(\boldsymbol{r}_\star)$,

$$\sigma_{K,k}^2 = \mathrm{Var}_s\left(\Phi_{K,\boldsymbol{k}}(Y_1)\right), \quad \varepsilon_{K,\boldsymbol{k}} = \sqrt{6\sigma_{K,k}^2} + 2\sqrt{\pi(\boldsymbol{k})},$$

$$\Omega_\star = \bigcap_{K\in\mathcal{D}_\star}\bigcap_{\boldsymbol{k}\in\Lambda(\boldsymbol{r}_\star)}\left\{|\nu(\Phi_{K,\boldsymbol{k}})| < \varepsilon_{K,\boldsymbol{k}}\sqrt{\lambda_d(K)}\right\}.$$

From Bernstein's Inequality (see for instance [Mas07], Section 2.2.3), for all $K \in \mathcal{D}_\star$, $\boldsymbol{k} \in \Lambda(\boldsymbol{r}_\star)$ and $x > 0$,

$$\mathbb{P}_s\left(|\nu(\Phi_{K,\boldsymbol{k}})| \geq \sqrt{2\sigma_{K,\boldsymbol{k}}^2\frac{x}{n}} + \frac{2\sqrt{\pi(\boldsymbol{k})}}{3\sqrt{\lambda_d(K)}}\frac{x}{n}\right) \leq 2\exp(-x),$$

so

$$\mathbb{P}_s\left(|\nu(\Phi_{K,\boldsymbol{k}})| \geq \varepsilon_{K,\boldsymbol{k}}\sqrt{\lambda_d(K)}\right) \leq 2\exp(-3n\lambda_d(K)) \leq 2\exp(-3n2^{-dJ_\star}).$$

Besides, there are $2^{J_\star+1}-1$ dyadic intervals of $[0,1]$ with length $\geq 2^{-J_\star}$, so $|\mathcal{D}_\star| \leq 2^{d(1+J_\star)}$. And we assume that $2^{dJ_\star} \leq n/\log(n|\Lambda(\boldsymbol{r}_\star)|)$, hence the upper-bound for $\mathbb{P}_s(\Omega_\star^c)$.

Let us also fix $(m', \boldsymbol{\rho}') \in \boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$, set

$$v_{(m',\boldsymbol{\rho}')} = \max_{K\in m'} \sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}'_K)} \varepsilon_{K,\boldsymbol{k}}\sqrt{\pi(\boldsymbol{k})} \quad \text{and} \quad b_{(m',\boldsymbol{\rho}')}(x) = \sqrt{\frac{nv_{(m',\boldsymbol{\rho}')}}{2(1/3+\theta_3^{-1})x}},$$

choose $\mathscr{T}_{(m',\boldsymbol{\rho}')}$ a countable and dense subset of $\mathcal{T}_{(m',\boldsymbol{\rho}')} = \left\{t \in S_{(m',\boldsymbol{\rho}')}/\|t\|_2 = 1, \|t\|_\infty \leq b_{(m',\boldsymbol{\rho}')}(x)\right\}$, and define

$$Z(m',\boldsymbol{\rho}') = \sup_{t\in\mathcal{T}_{(m',\boldsymbol{\rho}')}} \nu(t) = \sup_{t\in\mathscr{T}_{(m',\boldsymbol{\rho}')}} \nu(t).$$

Since $\Phi_{K,\boldsymbol{k}}$ has support K, for all $t \in S_{(m',\boldsymbol{\rho}')}$,

$$\begin{aligned}
\mathbb{E}_s\left[t^2(Y_1)\right] &= \mathbb{E}_s\left[\sum_{K\in m'}\left(\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho'}_K)} \langle t, \Phi_{K,k}\rangle \Phi_{K,k}\right)^2\right] \\
&\leq \sum_{K\in m'} |\Lambda(\boldsymbol{\rho}'_K)| \sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}'_K)} \langle t, \Phi_{K,k}\rangle^2 \mathbb{E}_s\left[\Phi_{K,k}^2(Y_1)\right] \\
&\leq |\Lambda(\boldsymbol{r}_\star)| M_{2,\star}\|t\|_2^2.
\end{aligned}$$

So Talagrand's Inequality, as stated for instance in [Mas07] (Chapter 5, Inequality (5.50)), ensures that there exists an event $\Omega_{(m',\boldsymbol{\rho}')}(x)$ such that $\mathbb{P}_s(\Omega_{(m',\boldsymbol{\rho}')}(x)) \geq 1-\exp(-x)$ and over which

$$Z(m',\boldsymbol{\rho}') \leq (1+\theta_3)\mathbb{E}_s\left[Z(m',\boldsymbol{\rho}')\right] + \sqrt{2|\Lambda(\boldsymbol{r}_\star)|M_{2,\star}\frac{x}{n}} + \sqrt{2(1/3+\theta_3^{-1})v_{(m',\boldsymbol{\rho}')}\frac{x}{n}}.$$

Since ν is linear, we deduce from Cauchy-Scwharz Inequality and its equality case that

$$\chi(m',\boldsymbol{\rho}') = \sup_{t\in S_{(m',\boldsymbol{\rho}')}, \|t\|_2=1} \nu(t) = \nu(t^\bullet_{(m',\boldsymbol{\rho}')})$$

where

$$t^\bullet_{(m',\boldsymbol{\rho}')} = \sum_{K\in m'}\sum_{\boldsymbol{k}\in\Lambda(\boldsymbol{\rho}'_K)} \frac{\nu(\Phi_{K,\boldsymbol{k}})}{\chi(m',\boldsymbol{\rho}')}\Phi_{K,\boldsymbol{k}}.$$

Therefore,

$$\mathbb{E}_s\left[Z(m',\boldsymbol{\rho}')\right] \leq \mathbb{E}_s\left[\chi(m',\boldsymbol{\rho}')\right] \leq \sqrt{\mathbb{E}_s\left[\chi^2(m',\boldsymbol{\rho}')\right]} = \sqrt{V_{(m',\boldsymbol{\rho}')}}.$$

Moreover, on the set $\Omega_{(m',\boldsymbol{\rho}')}(x) \cap \Omega_\star$, either $\chi(m',\boldsymbol{\rho}') \geq \sqrt{2(1/3+\theta_3^{-1})v_{(m',\boldsymbol{\rho}')}x/n}$, in which case $t^\bullet_{(m',\boldsymbol{\rho}')} \in \mathcal{T}_{(m',\boldsymbol{\rho}')}$, so that

$$\begin{aligned}\chi(m',\boldsymbol{\rho}') &= Z(m',\boldsymbol{\rho}') \\ &\qquad \leq (1+\theta_3)\sqrt{V_{(m',\boldsymbol{\rho}')}} + \sqrt{2|\Lambda(\boldsymbol{r}_\star)|M_{2,\star}\frac{x}{n}} + \sqrt{2(1/3+\theta_3^{-1})v_{(m',\boldsymbol{\rho}')}\frac{x}{n}},\end{aligned}$$

or $\chi(m',\boldsymbol{\rho}') < \sqrt{2(1/3+\theta_3^{-1})v_{(m',\boldsymbol{\rho}')}x/n}$, and the above inequality is still satisfied. Applying Inequality (34) with $\theta = 1$, we get

$$v_{(m',\boldsymbol{\rho}')} \leq \max_{K \in m'} \sum_{\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}'_K)} \left(\sigma^2_{K,\boldsymbol{k}} + 5\pi(\boldsymbol{k})\right) \leq |\Lambda(\boldsymbol{r}_\star)| \left(M_{2,\star} + 5\pi(\boldsymbol{r}_\star)\right).$$

Consequently, on $\Omega_{(m',\boldsymbol{\rho}')}(x)$,

$$\begin{aligned}\chi(m',\boldsymbol{\rho}')\mathbb{1}_{\Omega_\star} &\leq (1+\theta_3)\sqrt{V_{(m',\boldsymbol{\rho}')}} \\ &\qquad + \left(\sqrt{M_{2,\star}} + \sqrt{(1/3+\theta_3^{-1})(M_{2,\star}+5\pi(\boldsymbol{r}_\star))}\right)\sqrt{2|\Lambda(\boldsymbol{r}_\star)|\frac{x}{n}}.\end{aligned}$$

Thus, applying twice Inequality (34), with $\theta = \theta_4$ and $\theta = 1$, we get the concentration inequality for $\chi(m',\boldsymbol{\rho}')$ stated in Proposition 8. ■

From Proposition 8, we deduce that $\Omega_\chi(\zeta) = \cap_{(m',\boldsymbol{\rho}') \in \boldsymbol{\mathcal{M}}^{\boldsymbol{deg}}_\star} \Omega_{(m',\boldsymbol{\rho}')}(w_{(m',\boldsymbol{\rho}')} + \zeta)$ is an event with probability

$$\mathbb{P}_s\left(\Omega_\chi(\zeta)\right) \geq 1 - \exp(-\zeta)\Sigma$$

over which

$$\begin{aligned}\chi^2(\hat{m},\hat{\boldsymbol{\rho}})\mathbb{1}_{\Omega_\star} &\leq (1+\theta_3)(1+\theta_4)V_{(\hat{m},\hat{\boldsymbol{\rho}})} \\ &\quad + 4\left(1+\theta_4^{-1}\right)|\Lambda(\boldsymbol{r}_\star)|\left(\left(4/3+\theta_3^{-1}\right)M_{2,\star} + (5/3)\left(1+3\theta_3^{-1}\right)\pi(\boldsymbol{r}_\star)\right)\frac{w_{(\hat{m},\hat{\boldsymbol{\rho}})}+\zeta}{n}. \qquad (38)\end{aligned}$$

Our main task is then to estimate the unknown variance terms $V_{(m',\boldsymbol{\rho}')}, M_{1,\star}, M_{2,\star}$. Lemma 1 in [RBRTM10] remains valid with the same constants even though the Y_i's take values in $\mathbb{R}^d$ with $d \geq 1$. Let us set $\gamma = 3 + \log|\Lambda(\boldsymbol{r}_\star)|/\log(n)$. Since $|\Lambda(\boldsymbol{r}_\star)| \leq n^d$, γ is bounded independently of n ($3 \leq \gamma \leq 3+d$)). So, from the proof of Lemma 1 in [RBRTM10], for all $K \in \mathcal{D}_\star$ and $\boldsymbol{k} \in \Lambda(\boldsymbol{r}_\star)$, there exists an event $\Omega_{K,\boldsymbol{k}}$ such that $\mathbb{P}_s(\Omega^c_{K,\boldsymbol{k}}) \leq C(\theta_5,d)/(n^3|\Lambda(\boldsymbol{r}_\star)|)$ and over which

$$\begin{aligned}\mathrm{Var}_s\left(\Phi_{K,\boldsymbol{k}}(Y_1)\right) &\leq (1+\theta_5)\left(\hat{\sigma}^2_{K,\boldsymbol{k}} + 2\|\Phi_{K,\boldsymbol{k}}\|_\infty\sqrt{2\gamma\hat{\sigma}^2_{K,\boldsymbol{k}}\frac{\log(n)}{n}} + 8\gamma\|\Phi_{K,\boldsymbol{k}}\|^2_\infty\frac{\log(n)}{n}\right) \\ &\leq (1+\theta_5)\left(\hat{\sigma}^2_{K,\boldsymbol{k}} + 2\sqrt{8\hat{\sigma}^2_{K,\boldsymbol{k}}\pi(\boldsymbol{k})} + 32\pi(\boldsymbol{k})\right).\end{aligned}$$

Applying Inequality (34) with $a = \hat{\sigma}_{K,\boldsymbol{k}}$, $b = \sqrt{8\pi(\boldsymbol{k})}$ and $\theta = \theta_6$, we get, on $\Omega_{K,\boldsymbol{k}}$,

$$\sigma^2_{K,\boldsymbol{k}} \leq (1+\theta_5)\left((1+\theta_6)\hat{\sigma}^2_{K,\boldsymbol{k}} + 8(4+\theta_6^{-1})\pi(\boldsymbol{k})\right).$$

For all $(m', \boldsymbol{\rho}') \in \boldsymbol{\mathcal{M}}^{\boldsymbol{deg}}_{\star}$, let us introduce

$$\widehat{V}_{(m',\boldsymbol{\rho}')}(\theta_6) = \frac{1}{n}\sum_{K \in m'}\sum_{\boldsymbol{k} \in \Lambda(\boldsymbol{\rho}'_K)} \left((1+\theta_6)\hat{\sigma}^2_{K,\boldsymbol{k}} + 8(4+\theta_6^{-1})\pi(\boldsymbol{k})\right).$$

We have just proved that the set $\Omega_\sigma = \cap_{K\in\mathcal{D}_\star}\cap_{\boldsymbol{k}\in\Lambda(\boldsymbol{r}_\star)}\Omega_{K,\boldsymbol{k}}$ is an event with probability

$$\mathbb{P}_s\left(\Omega_\sigma\right) \geq 1 - 2^d C(\theta_5, d)/(n^2\log(n)) \tag{39}$$

over which

$$V_{(\hat{m},\hat{\boldsymbol{\rho}})} \leq (1+\theta_5)\widehat{V}_{(\hat{m},\hat{\boldsymbol{\rho}})}(\theta_6). \tag{40}$$

Let us now fix $K \in \mathcal{D}_\star$ and $\boldsymbol{k} \in \Lambda(\boldsymbol{r}_\star)$. According to Bernstein's Inequality and Inequality (34), there exist events $\Omega^1_{K,\boldsymbol{k}}$ and $\Omega^2_{K,\boldsymbol{k}}$, each with $\mathbb{P}_s$-measure $\geq 1 - 2\exp(-3n\lambda_d(K))$, such that on $\Omega^1_{K,\boldsymbol{k}}$

$$\begin{aligned}\sqrt{\frac{\pi(\boldsymbol{k})}{\lambda_d(K)}}\left|\frac{1}{n}\sum_{i=1}^n \Phi_{K,\boldsymbol{k}}(Y_i) - \mathbb{E}_s\left[\Phi_{K,\boldsymbol{k}}(Y_1)\right]\right| &\leq \sqrt{6\mathbb{E}_s\left[\Phi^2_{K,\boldsymbol{k}}(Y_1)\right]\pi(\boldsymbol{k})} + \pi(\boldsymbol{k})\\ &\leq \theta_7\mathbb{E}_s\left[\Phi^2_{K,\boldsymbol{k}}(Y_1)\right] + (1+3\theta_7^{-1})\pi(\boldsymbol{k}),\end{aligned}$$

and on $\Omega^2_{K,\boldsymbol{k}}$

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \Phi^2_{K,\boldsymbol{k}}(Y_i) - \mathbb{E}_s\left[\Phi^2_{K,\boldsymbol{k}}(Y_1)\right]\right| &\leq \sqrt{6\|\Phi_{K,\boldsymbol{k}}\|^2_\infty\mathbb{E}_s\left[\Phi^2_{K,\boldsymbol{k}}(Y_1)\right]\lambda_d(\boldsymbol{k})} + \|\Phi_{K,\boldsymbol{k}}\|^2_\infty\lambda_d(\boldsymbol{k})\\ &\leq \theta_8\mathbb{E}_s\left[\Phi^2_{K,\boldsymbol{k}}(Y_1)\right] + (1+3\theta_8^{-1})\pi(\boldsymbol{k}).\end{aligned}$$

We thus obtain that $\Omega_M = \cap_{K\in\mathcal{D}_\star}\cap_{\boldsymbol{k}\in\Lambda(\boldsymbol{r}_\star)}(\Omega^1_{K,\boldsymbol{k}} \cap \Omega^2_{K,\boldsymbol{k}})$ is an event with probablity

$$\mathbb{P}_s(\Omega_M) \geq 1 - 4 \times 2^d/(n^2\log(n)) \tag{41}$$

over which

$$\begin{aligned}&M_{1,\star} \leq \widehat{M_{1,\star}} + \theta_7(1-\theta_8)^{-1}|\Lambda(\boldsymbol{r}_\star)|\widehat{M}_{2,\star} + \left(\theta_7(1-\theta_8)^{-1}(1+3\theta_8^{-1}) + (1+3\theta_7^{-1})\right)\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\\ &\widehat{M_{1,\star}} \leq M_{1,\star} + \theta_7|\Lambda(\boldsymbol{r}_\star)|M_{2,\star} + (1+3\theta_7^{-1})\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\\ &M_{2,\star} \leq (1-\theta_8)^{-1}\widehat{M}_{2,\star} + (1+3\theta_8^{-1})(1-\theta_8)^{-1}\pi(\boldsymbol{r}_\star)\\ &\widehat{M}_{2,\star} \leq (1+\theta_8)M_{2,\star} + (1+3\theta_8^{-1})\pi(\boldsymbol{r}_\star).\end{aligned} \tag{42}$$

Let us set $\Omega_\bullet = \Omega_\star \cap \Omega_\sigma \cap \Omega_M$ and

$$\begin{aligned}
C_0 &= 1 - 2\theta_1(1+\theta_2)\\
C_1 &= 1 + 2\theta_1(1+\theta_2^{-1})\\
C_2 &= (1+C_0)(1+\theta_3)(1+\theta_4)(1+\theta_5)\\
C_3 &= 4(1+C_0)(4/3+\theta_3^{-1})(1+\theta_4^{-1})(1-\theta_8)^{-1} + C_7\theta_7(1-\theta_8)^{-1}\\
C_4 &= 3C_3 + (20/3)(1+C_0)(1+3\theta_3^{-1})(1+\theta_4^{-1}) + C_7\left(1+3\theta_7^{-1}+\theta_7(1+3\theta_8^{-1})(1-\theta_8)^{-1}\right)\\
C_5 &= 2(2/3+\theta_1^{-1})\\
C_6 &= C_7 + 4(1+C_0)(4/3+\theta_3^{-1})(1+\theta_4^{-1})\\
C_7 &= (20/3)(1+C_0)(1+3\theta_3^{-1})(1+\theta_4^{-1}).
\end{aligned}$$

We choose pen such that, on $\Omega_\bullet$ and for all $(m',\boldsymbol{\rho}') \in \boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$,

$$\mathrm{pen}(m',\boldsymbol{\rho}') = C_2\widehat{V}_{(m',\boldsymbol{\rho}')}(\theta_6) + \left(\left(C_3\widehat{M}_{2,\star} + C_4\pi(\boldsymbol{r}_\star)\right)|\Lambda(\boldsymbol{r}_\star)| + C_5\widehat{M}_{1,\star}\right)\frac{w_{(m',\boldsymbol{\rho}')}}{n}.$$

Thus, combining Inequalities (32), (36), (38), (40), (42) with the upper-bounds

$$M_{1,\star} \le \pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\|s\|_\infty \quad \text{and} \quad M_{2,\star} \le \pi(\boldsymbol{r}_\star)\|s\|_\infty,$$

we obtain, on $\Omega_m(\zeta) \cap \Omega_\chi(\zeta) \cap \Omega_\bullet$,

$$C_0\|s-\tilde{s}\|_2^2 \le C_1\|s - s_{(m,\boldsymbol{\rho})}\|_2^2 + \mathrm{pen}(m,\boldsymbol{\rho}) + (C_6\|s\|_\infty + C_7)\,\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\frac{\zeta}{n}.$$

Setting

$$\begin{aligned}
C_3' &= C_3(1+\theta_8) + C_5(1+\theta_7)\\
C_4' &= C_3(1+3\theta_8^{-1}) + C_5(1+3\theta_7^{-1})
\end{aligned}$$

we deduce from (42) that, on $\Omega_\bullet$,

$$\mathrm{pen}(m,\boldsymbol{\rho}) \le C_2\widehat{V}_{(m,\boldsymbol{\rho})}(\theta_6) + \left(C_3'\|s\|_\infty + C_4'\right)\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\frac{w_{(m,\boldsymbol{\rho})}}{n},$$

so that, on $\Omega_m(\zeta) \cap \Omega_\chi(\zeta)$,

$$\begin{aligned}
C_0\|s-\tilde{s}\|_2^2 \mathbb{1}_{\Omega_\bullet} \le{}& C_1\|s - s_{(m,\boldsymbol{\rho})}\|_2^2 + C_2\widehat{V}_{(m,\boldsymbol{\rho})}(\theta_6)\\
&+ \left(C_3'\|s\|_\infty + C_4'\right)\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\frac{w_{(m,\boldsymbol{\rho})}}{n}\\
&+ (C_6\|s\|_\infty + C_7)\,\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\frac{\zeta}{n}.
\end{aligned}$$

Last, we recall that Fubini's Theorem yields, for all random variable U,

$$\mathbb{E}[U] \leq \mathbb{E}[U_+] = \int_0^\infty \mathbb{P}(U_+ > \zeta)\mathrm{d}\zeta = \int_0^\infty \mathbb{P}(U > \zeta)\mathrm{d}\zeta,$$

and we underline that

$$\mathbb{E}_s\left[\widehat{V}_{(m,\boldsymbol{\rho})}(\theta_6)\right] \leq (1+\theta_6)\mathbb{E}_s\left[\|\hat{s}_{(m,\boldsymbol{\rho})} - s_{(m,\boldsymbol{\rho})}\|_2^2\right] + 8(4+\theta_6^{-1})\pi(\boldsymbol{r}_\star)\frac{\dim(S_{(m,\boldsymbol{\rho})})}{n}.$$

Therefore,

$$\begin{aligned} C_0\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2 \mathbb{1}_{\Omega_\bullet}\right] \leq\ & C_1\|s - s_{(m,\boldsymbol{\rho})}\|_2^2 + (1+\theta_6)C_2\mathbb{E}_s\left[\|\hat{s}_{(m,\boldsymbol{\rho})} - s_{(m,\boldsymbol{\rho})}\|_2^2\right] \\ & + 8(4+\theta_6^{-1})C_2\pi(\boldsymbol{r}_\star)\frac{\dim(S_{(m,\boldsymbol{\rho})})}{n} \\ & + \left(C_3'\|s\|_\infty + C_4'\right)\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\frac{w_{(m,\boldsymbol{\rho})}}{n} \\ & + 2\left(C_6\|s\|_\infty + C_7\right)\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\frac{\Sigma}{n}. \end{aligned} \tag{43}$$

There remains to bound the risk of $\tilde{s}$ on $\Omega_\bullet^c$. According to (37), (39) and (41),

$$p_\bullet := \mathbb{P}_s(\Omega_\bullet^c) \leq \mathbb{P}_s(\Omega_\star^c) + \mathbb{P}_s(\Omega_\sigma^c) + \mathbb{P}_s(\Omega_M^c) \leq C(\theta_5, d)/(n^2\log(n)).$$

From Pythagoras' Equality and the inclusion of $S_{(\hat{m},\hat{\boldsymbol{\rho}})}$ into $S_{(m_\star,\boldsymbol{r}_\star)}$, we deduce

$$\|s-\tilde{s}\|_2^2 = \|s - \hat{s}_{(\hat{m},\hat{\boldsymbol{\rho}})}\|_2^2 + \chi^2(\hat{m},\hat{\boldsymbol{\rho}}) \leq \|s\|_2^2 + \chi^2(m_\star,\boldsymbol{r}_\star).$$

Therefore, it follows from Cauchy-Scwharz Inequality that

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\mathbb{1}_{\Omega_\bullet}\right] \leq p_\bullet\|s\|_2^2 + \sqrt{p_\bullet\mathbb{E}_s\left[\chi^4(m_\star,\boldsymbol{r}_\star)\right]}.$$

Let $\mathscr{S}_\star$ be some countable and dense subset of $\{t \in S_{(m_\star,\boldsymbol{r}_\star)} \text{ s.t. } \|t\|_2 = 1\}$. Since $\chi(m_\star,\boldsymbol{r}_\star) = \sup_{t\in\mathscr{S}_\star}|\nu(t)|$, we deduce from Theorem 12 in [BBLM05] that

$$\sqrt{\mathbb{E}_s\left[\chi^4(m_\star,\boldsymbol{r}_\star)\right]} \leq C\left(\mathbb{E}_s\left[\chi^2(m_\star,\boldsymbol{r}_\star)\right] + \sigma^2/n + M/n^2\right),$$

where M is any upper-bound for $\sup_{t\in\mathscr{S}_\star}\max_{1\leq i\leq n}|t(Y_i) - \langle t, s\rangle|$ and σ^2, any upper-bound for $n\sup_{t\in\mathscr{S}_\star}\mathrm{Var}_s(t(Y_1))$. Therefore, we obtain

$$\sqrt{\mathbb{E}_s\left[\chi^4(m_\star,\boldsymbol{r}_\star)\right]} \leq C\left(\frac{\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\|s\|_\infty}{\log(n)} + \frac{\|s\|_\infty}{n} + \frac{\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|}{n\log(n)}\right) \leq C\frac{\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\|s\|_\infty}{\log(n)},$$

hence

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\mathbb{1}_{\Omega_\bullet}\right] \leq C(\theta_5,d)\frac{\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\|s\|_\infty^2}{n\log^{3/2}(n)}. \tag{44}$$

Since $\|s\|_\infty \geq 1$, we conclude thanks to (43) and (44)

$$\begin{aligned}\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\right] \leq C_1''\|s-s_{(m,\boldsymbol{\rho})}\|_2^2 + C_2''\mathbb{E}_s\left[\|\hat{s}_{(m,\boldsymbol{\rho})}-s_{(m,\boldsymbol{\rho})}\|_2^2\right] + C_3''\pi(\boldsymbol{r}_\star)\frac{D_{(m,\boldsymbol{\rho})}}{n} \\ + \|s\|_\infty\pi(\boldsymbol{r}_\star)|\Lambda(\boldsymbol{r}_\star)|\left(C_4''\frac{w_{(m,\boldsymbol{\rho})}}{n} + C_5''\frac{\Sigma}{n} + C_6''\frac{\|s\|_\infty}{n\log^{3/2}(n)}\right)\end{aligned} \tag{45}$$

where

$$\begin{aligned}&C_1'' = C_1/C_0, \quad C_2'' = (1+\theta_6)C_2/C_0, \quad C_3'' = 8(4+\theta_6^{-1})C_2/C_0,\\ &C_4'' = (C_3'+C_4')/C_0, \quad C_5'' = 2(C_6+C_7)/C_0, \quad C_6'' = C(\theta_5, d).\end{aligned}$$

Choosing, for all $(m,\boldsymbol{\rho}) \in \boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$, $w_{(m,\boldsymbol{\rho})} = L_{(m,\boldsymbol{\rho})}|m|$, and taking in (45) the minimum over $(m,\boldsymbol{\rho}) \in \boldsymbol{\mathcal{M}}_\star^{\boldsymbol{deg}}$ allows to complete the proof.

6 Proof of Theorem 4

Let us fix $\boldsymbol{\sigma}, p, p', R, L$ satisfying the assumptions of the theorem and $s \in \mathcal{P}(\boldsymbol{\sigma}, p, p', R, L)$. For $J = J_\star$, all the partitions given by Theorem 2 belong to $\mathcal{M}_\star$, so according to Proposition 5 and Theorem 2 applied with $\boldsymbol{r} = \lfloor\boldsymbol{\sigma}\rfloor + 1$,

$$\begin{aligned}&\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\right]\\ &\leq C(\kappa'', d, \boldsymbol{\sigma}, p, L)\left(\inf_{k\in\mathbb{N}}\left\{R^2 2^{-2kH(\boldsymbol{\sigma})} + w(\boldsymbol{r}_\star)\frac{2^{kd}}{n}\right\} + R^2 2^{-2J_\star d(H(\boldsymbol{\sigma})/d-(1/p-1/2)_+)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}\right).\end{aligned}$$

In order to minimize approximately the above infimum, we choose

$$k_\star = \max\{k \in \mathbb{N} \text{ s.t. } w(\boldsymbol{r}_\star)2^{kd}/n \leq R^2 2^{-2kH(\boldsymbol{\sigma})}\}$$

which is well defined since $R^2 n/w(\boldsymbol{r}_\star) \leq 1$, and thus obtain

$$\begin{aligned}&\mathbb{E}_s\left[\|s-\tilde{s}\|_2^2\right]\\ &\leq C(\kappa'', d, \boldsymbol{\sigma}, p, L)\left(\left(R\left(n/w(\boldsymbol{r}_\star)\right)^{-H(\boldsymbol{\sigma})/d}\right)^{2d/(d+2H(\boldsymbol{\sigma}))} + R^2 2^{-2J_\star d(H(\boldsymbol{\sigma})/d-(1/p-1/2)_+)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}\right).\end{aligned}$$

Given the assumptions on $J_\star$ and R, the leading term in the right-hand sand is the first one. We then conclude thanks to Propostion 4.

References

[AD10] N. Akakpo and C. Durot. Histogram selection for possibly censored data. *Mathematical Methods of Statistics*, 19(3):189–218, 2010.

[BBLM05] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Annals of Probability*, 33(2):514–560, 2005.

[Bir06] L. Birgé. Statistical estimation with model selection. *Indag. Math. (N.S.)*, 17(4):497–537, 2006.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM00] L. Birgé and P. Massart. An adaptive compression algorithm in Besov spaces. *Constr. Approx.*, 16(1):1–36, 2000.

[BSR04] G. Blanchard, C. Schäfer, and Y. Rozenholc. Oracle bounds and exact algorithm for dyadic classification trees. *Learning theory*, pages 378–392, 2004.

[BSRM07] G. Blanchard, C. Schäfer, Y. Rozenholc, and K.R. Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2):209–241, 2007.

[CDDD01] Albert Cohen, Wolfgang Dahmen, Ingrid Daubechies, and Ronald DeVore. Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.*, 11(2):192–226, 2001.

[CM09] A. Cohen and J.M. Mirebeau. Adaptive and anisotropic piecewise polynomial approximation. *Multiscale, Nonlinear and Adaptive Approximation*, pages 75–135, 2009.

[CR04] F. Comte and Y. Rozenholc. A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.*, 56(3):449–473, 2004.

[DeV98] R. A. DeVore. Nonlinear approximation. In *Acta numerica, 1998*, volume 7 of *Acta Numer.*, pages 51–150. Cambridge Univ. Press, Cambridge, 1998.

[DJP92] Ronald A. DeVore, Björn Jawerth, and Vasil Popov. Compression of wavelet decompositions. *Amer. J. Math.*, 114(4):737–785, 1992.

[Don97] D. L. Donoho. CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911, 1997.

[DY90] R. A. DeVore and X. M. Yu. Degree of adaptive approximation. *Math. Comp.*, 55(192):625–635, 1990.

[Hoc02a] R. Hochmuth. Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208, 2002.

[Hoc02b] Reinhard Hochmuth. n-term approximation in anisotropic function spaces. *Math. Nachr.*, 244:131–149, 2002.

[Kle09] J. Klemelä. Multivariate histograms with data-dependent partitions. *Statist. Sinica*, 19(1):159–176, 2009.

[KLP01] G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields*, 121(2):137–170, 2001.

[Lei03] C. Leisner. Nonlinear wavelet approximation in anisotropic Besov spaces. *Indiana Univ. Math. J.*, 52(2):437–455, 2003.

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Neu00] M. H. Neumann. Multivariate wavelet thresholding in anisotropic function spaces. *Statist. Sinica*, 10(2):399–431, 2000.

[NvS97] M. H. Neumann and R. von Sachs. Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.*, 25(1):38–76, 1997.

[RBRTM10] P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Adaptive density estimation: a curse of support? *Journal of Statistical Planning and Inference*, 2010.

[ST87] H.-J. Schmeisser and H. Triebel. *Topics in Fourier analysis and function spaces*, volume 42 of *Mathematik und ihre Anwendungen in Physik und Technik [Mathematics and its Applications in Physics and Technology]*. Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig, 1987.

[Tri11] H. Triebel. Entropy numbers in function spaces with mixed integrability. *Revista Matemática Complutense*, 24(1):169–188, 2011.

[WN07] R. M. Willett and Robert D. Nowak. Multiscale Poisson intensity and density estimation. *IEEE Trans. Inform. Theory*, 53(9):3171–3187, 2007.

[YB99] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.