Electronic Journal of Statistics
Vol. 6 (2012) 382–420
ISSN: 1935-7524
DOI: 10.1214/12-EJS677

Classification via local multi-resolution projections

Jean-Baptiste Monnier

Université Paris Diderot, Paris 7,
LPMA, office 5B01,
175 rue du Chevaleret,
75013, Paris, France
e-mail: monnier@math.jussieu.fr

Abstract: We focus on the supervised binary classification problem, which consists in guessing the label Y associated to a co-variate $X \in \mathbb{R}^d$, given a set of n independent and identically distributed co-variates and associated labels (X_i, Y_i). We assume that the law of the random vector (X, Y) is unknown and the marginal law of X admits a density supported on a set $\mathcal{A}$. In the particular case of plug-in classifiers, solving the classification problem boils down to the estimation of the regression function $\eta(X) = \mathbb{E}[Y|X]$. Assuming first $\mathcal{A}$ to be known, we show how it is possible to construct an estimator of η by localized projections onto a multi-resolution analysis (MRA). In a second step, we show how this estimation procedure generalizes to the case where $\mathcal{A}$ is unknown. Interestingly, this novel estimation procedure presents similar theoretical performances as the celebrated local-polynomial estimator (LPE). In addition, it benefits from the lattice structure of the underlying MRA and thus outperforms the LPE from a computational standpoint, which turns out to be a crucial feature in many practical applications. Finally, we prove that the associated plug-in classifier can reach super-fast rates under a margin assumption.

AMS 2000 subject classifications: Primary 62G05, 62G08; secondary 62H30, 62H12.
Keywords and phrases: Nonparametric regression, random design, multi-resolution analysis, supervised binary classification, margin assumption.

Received June 2011.

Contents

1. Introduction

1.1. Setting

The supervised binary classification problem is directly related to a wide range of applications such as spam detection or assisted medical diagnosis (see [25, chap. 1] for more details). It can be described as follows.

The supervised binary classification problem. Let $\mathcal{E}$ stand for a subset of $\mathbb{R}^d$ and write $\mathcal{Y} = \{0, 1\}$. Assume we observe n co-variates $X_i \in \mathcal{E}$ and associated labels $Y_i \in \mathcal{Y}$ such that the elements of $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$ are n independent realizations of the random vector $(X, Y) \in \mathcal{E} \times \mathcal{Y}$ of unknown law $\mathbb{P}_{X,Y}$. Given $\mathcal{D}_n$ and a new co-variate X_{n+1}, we want to predict the associated label Y_{n+1} so as to minimize the probability of making a mistake.

In other words, we want to build a **classifier** $h_n : \mathcal{E} \mapsto \mathcal{Y}$ upon the data $\mathcal{D}_n$, which minimizes $\mathbb{P}(h_n(X) \neq Y|\mathcal{D}_n)$. It is well known that the Bayes classifier $h^*(\tau) := \mathbb{1}_{\{\eta(\tau)\geq 1/2\}}$, where $\eta(\tau) := \mathbb{E}[Y|X = \tau] = \mathbb{P}(Y = 1|X = \tau)$ (unknown in practice), is optimal among all classifiers since, for any other classifier h_n, we have $\ell(h_n, h^*) := \mathbb{P}(h_n(X) \neq Y|\mathcal{D}_n) - \mathbb{P}(h^*(X) \neq Y) \geq 0$ (see [12]). As a consequence, we measure the classification risk $\mathscr{T}(h_n)$ associated to a classifier h_n as its average relative performance over all data sets $\mathcal{D}_n$, $\mathscr{T}(h_n) = \mathbb{E}^{\otimes n}\ell(h_n, h^*)$. As described in [12, Chap. 7], there is no classifier h_n such that $\mathscr{T}(h_n)$ goes to zero with n at a specified rate for all distributions $\mathbb{P}_{X,Y}$. We therefore make the assumption that $\mathbb{P}_{X,Y}$ belongs to a class of distributions $\mathcal{P}$ (as large as possible) and aim at constructing a classifier h_n such that

$$\inf_{\theta_n} \sup_{\mathbb{P}_{X,Y}\in\mathcal{P}} \mathscr{T}(\theta_n) \lesssim \sup_{\mathbb{P}_{X,Y}\in\mathcal{P}} \mathscr{T}(h_n) \lesssim (\log n)^{\delta} \inf_{\theta_n} \sup_{\mathbb{P}_{X,Y}\in\mathcal{P}} \mathscr{T}(\theta_n), \quad n \geq 1, \quad (1)$$

where the infinimum is taken over all measurable maps θ_n from $\mathcal{E}$ into $\mathcal{Y}$ and $\lesssim$ means lesser or equal up to a multiplicative constant factor independent of n. Any classifier h_n verifying eq. (1) will be said to be **(nearly) minimax optimal** when $\delta = 0$ ($\delta > 0$). $\mathcal{P}$ will stand for the set of all distributions such that the marginal law $\mathbb{P}_X$ of X admits a density μ on $\mathcal{E}$ and η belongs to a given smoothness class. Throughout the paper, we will denote by μ the density of $\mathbb{P}_X$.

Many classifiers have been suggested in the literature, such as k-nearest neighbors, neural networks, support vector machine (SVM) or decision trees (see [12, 25]). In this paper, we will exclusively focus on **plug-in classifiers** $h_n(\tau) := \mathbb{1}_{\{\eta_n(\tau)\geq 1/2\}}$, where η_n stands for an estimator of η. With such classifiers, it is shown in [48] that,

$$\mathscr{T}(h_n) \leq 2\mathbb{E}^{\otimes n}\mathbb{E}|\eta_n(X) - \eta(X)|, \quad (2)$$

where the term on the rhs is known as the regression loss (of the estimator η_n of η) in $\mathbb{L}_1(\mathcal{E},\mu)$-norm. Eq. (2) shows in particular that rates of convergence on the classification risk of a plug-in classifier h_n can be readily derived from rates of convergence on the regression loss of η_n. This prompts us to focus on the regression problem, which can be stated in full generality as follows.

The regression on a random design problem. Let $\mathcal{E},\mathcal{Y}$ stand for subsets of $\mathbb{R}^d$ and $\mathbb{R}$, respectively. Assume we dispose of n co-variates $X_i \in \mathcal{E}$ and associated observations $Y_i \in \mathcal{Y}$ such that the elements of $\mathcal{D}_n = \{(X_i, Y_i), i = 1,\dots,n\}$ are n independent realizations of the random vector $(X,Y) \in \mathcal{E}\times\mathcal{Y}$ of unknown law $\mathbb{P}_{X,Y}$. We define $\xi := Y - \eta(X)$, where $\eta(\tau) := \mathbb{E}[Y|X=\tau]$, so that by construction $\mathbb{E}[\xi|X]=0$. Given $\mathcal{D}_n$ and under the assumption that $\mathbb{P}_{X,Y}$ belongs to a large class of distributions $\mathcal{P}$, we want to come up with an estimator η_n of η, which is as accurate as possible for the wide range of losses $\mathscr{S}_p(\eta_n) = \mathbb{E}^{\otimes n}\mathbb{E}|\eta_n(X)-\eta(X)|^p$, $p \geq 1$.

As described previously, in the particular case where $\mathcal{Y}=\{0,1\}$, we fall back on the regression problem associated to the classification problem with plug-in classifiers. In this case, ξ is bounded such that $|\xi| \leq 1$. Notice however that the regression on a random design problem stated above permits for $\mathcal{Y}$ to be any subset of $\mathbb{R}$ (including $\mathbb{R}$ itself). To be more precise, and by analogy with eq. (1), our aim is to build an estimator η_n of η such that, for all $p \geq 1$,

$$\inf_{\theta_n}\sup_{\mathbb{P}_{X,Y}\in\mathcal{P}} \mathscr{S}_p(\theta_n) \lesssim \sup_{\mathbb{P}_{X,Y}\in\mathcal{P}} \mathscr{S}_p(\eta_n) \lesssim (\log n)^{\delta} \inf_{\theta_n}\sup_{\mathbb{P}_{X,Y}\in\mathcal{P}} \mathscr{S}_p(\theta_n), \quad n \geq 1, \quad (3)$$

where the infinimum is taken over all measurable maps θ_n from $\mathcal{E}$ into $\mathcal{Y}$. And η_n will be said to be (nearly) minimax optimal when $\delta = 0$ ($\delta > 0$).

1.2. Motivations

Many estimators η_n of η have been suggested in the literature to solve the regression on a random design problem. Among them, the celebrated local polynomial estimator (LPE) has been praised for its flexibility and strong theoretical performances (see [45, 46]). As is well known, the LPE is minimax optimal in any dimension $d \in \mathbb{N}$ and for any $\mathscr{S}_p$-loss, $p \in (0,\infty]$, over the set of laws $\mathcal{P}$ such that (i) μ is bounded from above and below on its support $\mathcal{A} := \mathrm{Supp}\mu = \{\tau : \mu(\tau) > 0\}$, (ii) η belongs to a Hölder ball $\mathscr{C}^s(\mathcal{E},M)$ of radius M and (iii) ξ has sub-Gaussian tails. As a drawback, the LPE is computationally expansive since it requires to perform a new regression at every single point $x \in \mathcal{A}$ where we want to estimate η.

Computational efficiency is however of primary importance in many practical applications. In this paper, we show that it is possible to construct a novel estimator η_n of η by localized projections onto multi-resolution analysis (MRA) of $\mathbb{L}_2(\mathbb{R}^d,\lambda)$ (where λ stands for the Lebesgue measure on $\mathcal{E}$), which presents similar theoretical performances and is computationally more efficient than the LPE.

1.3. The hypotheses

In this section, we summarize the assumptions on μ, $\mathcal{A}$, η and ξ that will be used throughout the paper.

Assumption on μ. Let us denote by $\mu_{\min}, \mu_{\max}$ two real numbers such that $0 < \mu_{\min} \leq \mu_{\max} < \infty$. As is standard in the regression on a random design setting, we assume that the density μ is bounded above and below on its support $\mathcal{A}$.
(**D1**) $\mu_{\min} \leq \mu(\tau) \leq \mu_{\max}$ for all $\tau \in \mathcal{A}$.
This guarantees that we have enough information at each point $x \in \mathcal{A}$ in order to estimate η with best accuracy. For a study with weaker assumptions on μ, the reader is referred to [17, 18], for example, and the references therein.

Assumption on $\mathcal{A}$. We first assume that,
(**S1**) $\mathcal{A} = \mathcal{E} = [0,1]^d$.
Therefore $\mathcal{A}$ is known under (**S1**). We will deal with the case where $\mathcal{A}$ is unknown in Section 9.

Assumption on η. Fix $r \in \mathbb{N}$. In the sequel, we will assume that,
($\mathbf{H}_{\mathbf{s}}^{\mathbf{r}}$) The regression function η belongs to the generalized Lipschitz ball $\mathscr{L}^s(\mathcal{E}, M)$ of radius M, for some $s \in (0, r)$.
Unless otherwise sated, s is **unknown** but belongs to the interval $(0, r)$, where r is **known**. For a detailed review of generalized Lipschitz classes, the reader is referred to the Appendix below.

Assumptions on the noise ξ. We will consider the two following assumptions,
(**N1**) Conditionally on X, the noise ξ is uniformly bounded, meaning that there exists an absolute constant $K > 0$ such that $|\xi| \leq K$.
(**N2**) The noise ξ is independent of X and normally distributed with mean zero and variance σ^2, which we will denote by $\xi \sim \Phi(0, \sigma^2)$.
Assumption (**N1**) is adapted to the supervised binary classification setting, where $\mathcal{Y} = \{0, 1\}$, while (**N2**) is more common in the regression on a random design setting, where $\mathcal{Y} = \mathbb{R}$.

Combination of assumptions. In the sequel, we will conveniently refer by (**CS1**) to the set of assumptions (**D1**), (**S1**), (**N1**) or (**N2**). As detailed below in Section 3, configuration (**CS1**) is comparable to what is customary in the regression on a random design setting.

2. Our results

Assuming at first $\mathcal{A}$ to be known, we introduce a novel nonparametric estimator $\eta^{@}$ of η built upon local regressions against a multi-resolution analysis (MRA) of $\mathbb{L}_2(\mathbb{R}^d, \lambda)$ and show that, under (**CS1**), it is adaptive nearly minimax optimal over a wide generalized Lipschitz scale and across the wide range of losses

$\mathbb{L}_p(\mathcal{E},\mu), p \in [1,\infty)$. We subsequently show that these results generalize to the case where $\mathcal{A}$ is unknown but belongs to a large class of (eventually disconnected) subsets of $\mathbb{R}^d$, provided we modify the estimator $\eta^{@}$ accordingly. We denote by $\eta^{\maltese}$ this latter estimator and prove that $\eta^{\maltese}$ can be used to build an adaptive nearly minimax optimal plug-in classifier, which can reach superfast rates under a margin assumption. The above results essentially hinge on an exponential upper-bound on the probability of deviation of $\eta^{@}$ from η at a point, as detailed in Theorem 7.1. These results either improve on the current literature or are interesting in their own right for the following reasons.

1) They show that it is possible to use MRAs to construct an adaptive nearly minimax optimal estimator $\eta^{@}$ of η under the sole set of assumptions (**CS1**). More precisely, our results (i) hold in any dimension d; (ii) over the wide range of $\mathbb{L}_p(\mathcal{E},\mu)$-losses, $p \in [1,\infty)$; (iii) and a large Lipschitz scale; (iv) and do not require any assumption on μ beyond (**D1**). It is noteworthy that, in contrary to most alternative MRA-based estimation methods, no smoothness assumption on μ is needed.
2) From a computational perspective, $\eta^{@}$ outperforms other estimators of η under (**D1**) since it takes full advantage of the lattice structure of the underlying MRA. In particular it requires at most as many regressions as there are data points to be computed everywhere on $\mathcal{E}$, while alternative kernel estimators must be recomputed at each single point of $\mathcal{E}$. We illustrate this latter feature through simulation.
3) Furthermore, and in contrary to alternative MRA-based estimators, the local nature of $\eta^{@}$ allows to relax the assumption that $\mathcal{A}$ is known. This latter configuration allows for μ to cancel on $\mathcal{E}$ as long as it remains bounded on its support $\mathcal{A}$, which is particularly appropriate to the supervised binary classification problem under a margin assumption.
4) In the regression on a random design setting, $\eta^{@}$ bridges in fact the gap between usual linear wavelet estimators and alternative kernel estimators, such as the LPE. On the one hand, $\eta^{@}$ inherits its computational efficiency from the lattice structure of the underlying MRA. On the other hand, it features similar theoretical performances as the LPE in the random design setting. In particular, it remains a (locally) linear estimator of the data (modulo a spectral thresholding of the local regression matrix), and cannot discriminate finer smoothness than the one described by (generalized) Lipschitz spaces.

Here is the paper layout. We start by a literature review in Section 3. We give a hand-waving introduction to the main ideas that underpin the local multiresolution estimation procedure in Section 4. We define notations that will be used throughout the paper and introduce MRAs in Section 5. Our actual estimation procedure is described in Section 6 and the results are detailed in Section 7. We show how these results can be fine-tuned under additional assumptions in Section 8. Assumption (**S1**) is relaxed and the properties of $\eta^{\maltese}$ are detailed in Section 9. We show how these latter results spread to the classification setting in Section 10. Results of a simulation study with $\eta^{@}$ under (**CS1**) are given in Section 11. Proofs of the regression results can be found in Section 12. The proofs of the classification results are simple modifications of the proofs given

in [4] and can be found in [39]. In addition, the Appendix contains a detailed review of generalized Lipschitz spaces and MRAs.

3. Literature review

Both the regression on a random design problem and the classification problem have a long-standing history in nonparametric statistics. We will therefore limit ourselves to a brief account of the corresponding literature that is relevant to the present paper.

3.1. Classification with plug-in classifiers

Let us start with a review of some of the classification literature dedicated to plug-in classifiers. The seminal work [37] showed that plug-in rules are asymptotically optimal. It has been subsequently pointed out in [36] that the classification problem is in fact only sensitive to the behavior of $\mathbb{P}_{X,Y}$ near the boundary line $\mathscr{M} := \{\tau \in \mathcal{E} : \eta(\tau) = 1/2\}$. So that assumptions on the behavior of $\mathbb{P}_{X,Y}$ away from this boundary are in fact unnecessary. Subsequent works such as [3] have shown that convex combinations of plug-in classifiers can reach fast rates (meaning faster than $n^{-1/2}$, and thus faster than nonparametric estimation rates). More recently, it has been shown in [4] that plug-in classifiers can reach super fast rates (that is faster than n^{-1}) under suitable conditions. All these results are derived under some sort of smoothness assumption on the regression function η (see [50]) and a margin assumption (**MA**) (see Section 10 for details). This latter assumption clarifies the behavior of $\mathbb{P}_{X,Y}$ in a neighborhood of $\mathscr{M}$ and kicks in naturally through the computation

$$\mathscr{T}(h_n) \leq \delta \mathbb{P}(0 < |2\eta(X) - 1| \leq \delta) + \mathbb{E}|\eta_n(X) - \eta(X)| \mathbb{1}_{\{|\eta_n(X) - \eta(X)| > \delta\}},$$

where δ is chosen such that it balances the two terms on the rhs. Finally, [4] exhibited optimal convergence rates under smoothness and margin assumptions and showed that they are attained with plug-in classifiers. Let us now turn to the regression on a random design problem.

3.2. Regression on a random design with wavelets

First results on multi-resolution analysis (MRA) and wavelet bases (see [34, 38]) emerged in the nonparametric statistics literature in the early 1990's (see [27, 14, 13, 15, 16]). It has been proved that, under (**CS1**) and in the particular case where μ is the uniform distribution on $\mathcal{E}$, thresholded wavelet estimators of η are nearly minimax optimal over a wide Besov scale and range of $\mathbb{L}_p(\mathcal{E}, \mu)$-losses (see [10]). In order to leverage on the power of MRAs and associated wavelet bases, several authors attempted to transpose these latter results to more general design densities μ. This, however, led to a considerable amount of difficulties.

The literature relative to the study of wavelet estimators on an **unknown** random design breaks down into two main streams. (i) The first one aims at constructing new wavelet bases adapted to the (empirical) measure of the design (see [29, 30, 9, 47]). (ii) The second one aims at coming up with new algorithms to estimate the coefficients of the expansion of η on traditional wavelet bases (see [1, 23, 31, 41, 44]). The present paper belongs to this second line of research.

As described in [23], the success of the LPE on a random design results from the fact that it is built as a "ratio", which cancels out most of the influence of the design. In a wavelet context, a first suggestion has therefore been to use the ratio estimator of η (see [2, 42], for example), well known from the statistics literature on orthogonal series decomposition (see [20, 21] and [12, Chap. 17] and the references therein). Roughly speaking, the ratio estimator is the wavelet equivalent of the Nadaraya-Watson estimator (see [40, 49]). It is elaborated on the simple observation that $\eta(x) = \eta(x)\mu(x)/\mu(x)$ for all $x \in \mathcal{A}$, where both $g(.) = \eta(.)\mu(.)$ and $\mu(.)$ are easily estimated via traditional wavelet methods. The ratio estimator relies thus unfortunately on the estimation of μ itself and must therefore assume as much smoothness on μ as on η.

To address that issue, an other approach has been introduced in [6, 28]. They work with $d = 1$ and take $\mathcal{E}$ to be the unit interval $[0, 1]$. Their approach relies on the wavelet estimation of $\eta \circ G^{-1}$, where G stands for the cumulative distribution of the design and G^{-1} for its generalized inverse. Results are therefore stated in term of regularity of $f \circ G^{-1}$. Unfortunately, this method does not readily generalize to the the multi-dimensional case, where G admits no inverse.

Finally, [5] obtains adaptive near-minimax optimal wavelet estimators over a wide Besov scale under (**CS1**) by means of model selection techniques. His results are hence valid for the $\mathbb{L}_2(\mathcal{E}, \mu)$-loss only.

Other relevant references that proceed with hybrid estimators (LPE and kernel estimator or LPE and wavelet estimator) are [19] and [51]. They both work under (**CS1**), with $d = 1$ and assume that μ is at least continuous.

4. A primer on local multi-resolution estimation under (CS1)

In order to fix the ideas, let us now give a hand-waving introduction to the local multi-resolution estimation method. Throughout the paper, we will work with r-MRAs of $\mathbb{L}_2(\mathbb{R}^d, \lambda)$, for some $r \in \mathbb{N}$, consisting of nested approximation spaces $\mathcal{V}_j \subset \mathcal{V}_{j+1}$ built upon compactly supported scaling functions (see Section 5.2 and Appendix). Under the assumption that η belongs to the generalized Lipschitz ball $\mathscr{L}^s(\mathcal{E}, M)$ of radius M, the essential supremum of the remainder of the orthogonal projection $\mathscr{P}_j\eta$ of η onto $\mathcal{V}_j$ decreases like 2^{-js} (see Appendix). The regression function η can therefore be legitimately approximated by $\mathscr{P}_j\eta$. As an element of $\mathcal{V}_j$, $\mathscr{P}_j\eta$ may be written as an infinite linear combination of scaling functions at level j. In particular, there exists a partition $\mathcal{F}_j$ of $\mathcal{E}$ into hypercubes of edge-length 2^{-j} such that, for all $\mathcal{H} \in \mathcal{F}_j$ and all $x \in \mathcal{H}$, we can write $\mathscr{P}_j\eta(x) = \sum_{k \in \mathcal{S}_j(\mathcal{H})} \alpha_{j,k}\varphi_{j,k}(x)$, where $\mathcal{S}_j(\mathcal{H})$ stands for a finite subset of $\mathbb{Z}^d$ (see Figure 1). This leaves us in turn with the estimation of coefficients

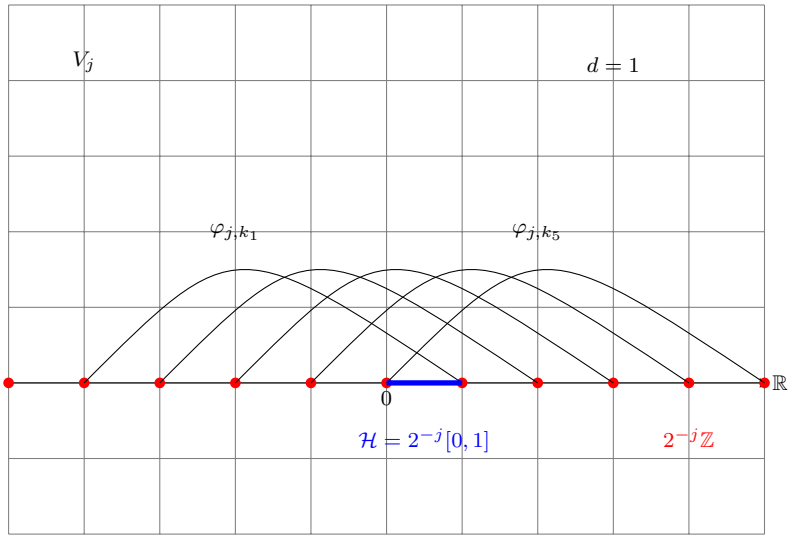


FIG 1. *Description of the localization cells $\mathcal{H}$ and their relations to the Supp$\varphi_{j,k}$.*

$(\alpha_{j,k})_{k\in\mathcal{S}_j(\mathcal{H})}$ for all $\mathcal{H} \in \mathcal{F}_j$, which is achieved by least-squares and provides us with the estimator $\eta_j^{@}$ of η on $\mathcal{H}$. It is noteworthy that the local estimator $\eta_j^{@}$ of η is exclusively built upon scaling functions and does not require the estimation of wavelet coefficients. In particular, it does not involve any sort of wavelet coefficient thresholding. To the best of the author knowledge, this is the first time that this local estimation procedure is proposed and studied from both a theoretical and computational perspective. In addition, we show that Lepski's method (see [32], for example) can be used to adaptively choose the resolution level j. Notice that Lepski's method has already been used in a MRA setting in [43]. In what follows, we detail the local multi-resolution estimation method and establish the near minimax optimality of $\eta^{@}$.

5. Notations

5.1. Preliminary notations

In the sequel, we will denote by $\mathcal{B}_p(z,\rho)$ the closed ℓ_p-ball of $\mathbb{R}^d$ of center z and radius ρ. More generally, we adopt the following notations: for any subset $\mathcal{S}$ of a topological space $\mathcal{E}$, Closure($\mathcal{S}$) will stand for its closure and $\mathcal{S}^c$ for its complement in $\mathcal{E}$. For any subset $\mathcal{S}$ of $\mathbb{R}^d$, $z \in \mathbb{R}^d$ and $\tau \in \mathbb{R}^+$, we will write $z + \mathcal{S}$ and $\tau\mathcal{S}$ to mean the sets $\{z + u : u \in \mathcal{S}\}$ and $\{\tau u : u \in \mathcal{S}\}$, respectively. Finally, given a set (of functions) $\mathcal{R}$, Span$\mathcal{R}$ will denote the set of finite linear combinations of elements of $\mathcal{R}$.

For any $p \in \mathbb{N}$, vectors v of $\mathbb{R}^p$ will be seen as elements of $\mathcal{M}_{p,1}$, that is matrix with p rows and one column. For any two $u, v \in \mathbb{R}^p$, $\langle u, v\rangle$ will denote

their Euclidean scalar product. In addition, for any $p, q \in \mathbb{N}$ and $M \in \mathcal{M}_{p,q}$, M^t will stand for the transpose of M. For any two matrices M, P, $M \cdot P$ will denote their matrix product when it makes sense. $[M]_{k,\ell}$ and $[M]_{k,\bullet}$ will respectively stand for the element of M located at line k, column ℓ and the k^{th} row of M. Finally, $\|M\|_S$ will denote the spectral norm of M (see [26, §5.6.6]).

We denote by $\lfloor z \rfloor$ the integer part of $z \in \mathbb{R}$ defined as $\max\{a \in \mathbb{Z} : a \leq z\}$. More generally, given $z \in \mathbb{R}^d$, we write $\lfloor z \rfloor$ the integer part of z, meant in a coordinate-wise sense. In the same way, we denote by $\lceil z \rceil$ the smallest integer greater than z (in a coordinate-wise sense). We write rhs (resp. lhs) to mean *right-* (resp. *left-*) *hand-side* and sometimes write $:=$ to mean *equal by definition*. Throughout the paper, we will refer to constants independent of n as *absolute constants* and c, C will stand for absolute constants whose value may vary from line to line. For any two sequences a_n, b_n of n, we will write $a_n \lesssim b_n$ to mean $a_n \leq C b_n$ for some absolute constant C and $a_n \approx b_n$ to mean that there exist two constants c, C independent of n such that $c b_n \leq a_n \leq C b_n$.

5.2. The polynomial reproduction property

In what follows, we will exclusively consider MRAs built upon Daubechies' scaling functions $\varphi_{j,k}$ (see Appendix and [8, 35, 7, 24]). Given a natural integer r, we will refer by r-MRA to a MRA whose nested approximation spaces $\mathcal{V}_j$ reproduce polynomials up to order $r-1$. Daubechies' scaling functions $\varphi_{j,k}$ are appealing in the estimation framework since they are compactly supported and have minimal volume supports among scaling functions that give rise to r-MRAs. Recall finally that a r-MRA can explain Lipschitz smoothness s for any $s \in (0, r)$.

5.3. General notations

Consider the Daubechies' r-MRA of $\mathbb{L}_2(\mathbb{R}^d, \lambda)$ built upon Daubechies' scaling function φ, as described in the Appendix. We will denote by $\mathrm{Supp}\varphi_{j,k} = \{\tau \in \mathbb{R}^d : \varphi_{j,k}(\tau) > 0\}$ the support of $\varphi_{j,k}$. Recall that $\mathrm{Supp}\varphi = (-(r-1), r)^d$. To alleviate notations, we will write φ_k in place of $\varphi_{0,k}$ and φ_j in place of $\varphi_{j,0}$. Notice that $\mathrm{Closure}(\mathrm{Supp}\varphi_{j,k})$ is in fact a closed hyper-cube of $\mathbb{R}^d$ whose corners lie on the lattice $2^{-j}\mathbb{Z}^d$. For any $x \in \mathcal{A}$, we write

$$\mathcal{S}_j(x) = \{\nu \in \mathbb{Z}^d : x \in \mathrm{Supp}\varphi_{j,\nu}\}.$$

Furthermore, we write $\mathcal{F}_j := 2^{-j}((0,1)^d + \mathbb{Z}^d) \cap \mathcal{E}$. It defines a partition of $\mathcal{E}$ into 2^{jd} hypercubes of edge length 2^{-j}, modulo a λ-null set. For the sake of concision, we write $R = 2r-1$ in the sequel. We have the following proposition, whose proof is straightforward and thus left to the reader.

Proposition 5.1. *$\mathcal{S}_j$ verifies the following properties,*

1. *$\mathcal{S}_j$ is constant on each element $\mathcal{H} \in \mathcal{F}_j$. We will denote by $\mathcal{S}_j(\mathcal{H})$ its value on $\mathcal{H}$.*

2. *Moreover, for any two* $\mathcal{H}_1, \mathcal{H}_2 \in \mathcal{F}_j$, $\mathcal{H}_1 \neq \mathcal{H}_2$, $\mathcal{S}_j(\mathcal{H}_1)$ *differs from* $\mathcal{S}_j(\mathcal{H}_2)$ *by at least one element.*
3. *Finally, for any* $\mathcal{H} \in \mathcal{F}_j$, $\#\mathcal{S}_j(\mathcal{H}) = R^d$

It is a direct consequence of Proposition 5.1 that in the case where $r = 1$, we have $\#\mathcal{S}_j(\mathcal{H}) = 1$ for all $\mathcal{H} \in \mathcal{F}_j$. We denote its single element by $\nu(\mathcal{H})$. It is in fact easy to show that $\nu(\mathcal{H}) = \lfloor 2^j x \rfloor$ for any $x \in \mathcal{H}$. For any $\mathcal{H} \in \mathcal{F}_j$, we write

$$\alpha_{\mathcal{H}} = (\alpha_{j,\nu})_{\nu \in \mathcal{S}_j(\mathcal{H})} \in \mathbb{R}^{R^d},$$
$$\varphi_{\mathcal{H}}(.) = (\varphi_{j,\nu}(.)\mathbb{1}_{\mathcal{H}}(.))_{\nu \in \mathcal{S}_j(\mathcal{H})} \in \mathbb{R}^{R^d}.$$

and denote by $Y_{\mathcal{H}} = (Y_i \mathbb{1}_{\mathcal{H}}(X_i))_{1 \leq i \leq n}$.

6. Construction of the local estimator $\eta^{@}$

Assume we are under (**CS1**) and work with the Daubechies' r-MRA of $\mathbb{L}_2(\mathbb{R}^d, \lambda)$. The estimation procedure is local, so that we start by selecting a point $x \in \mathcal{A}$. By construction, there exists $\mathcal{H} \in \mathcal{F}_j$ such that $x \in \mathcal{H}$. We want to estimate η at point x. As detailed in the Appendix, an estimator of η can be reduced to an estimator of the orthogonal projection $\mathscr{P}_j\eta$ of η onto $\mathcal{V}_j$, modulo an error $\mathscr{R}_j\eta$, such that $|\mathscr{R}_j\eta| \leq M2^{-js}$ when η belongs to the generalized Lipschitz ball $\mathscr{L}^s(\mathcal{E}, M)$ of radius M. Now, we can write

$$\mathscr{P}_j\eta(x) = \sum_{k \in \mathbb{Z}^d} \alpha_{j,k}\varphi_{j,k}(x) = \sum_{k \in \mathcal{S}_j(\mathcal{H})} \alpha_{j,k}\varphi_{j,k}(x) = \langle \alpha_{\mathcal{H}}, \varphi_{\mathcal{H}}(x) \rangle.$$

This leaves us with exactly R^d coefficients $\alpha_{j,\nu}, \nu \in \mathcal{S}_j(\mathcal{H})$ to estimate, which are valid for any $x \in \mathcal{H}$. We evaluate these coefficients by least-squares. Denote by $B_{\mathcal{H}} \in \mathcal{M}_{n,R^d}$ the matrix whose rows are the vectors $\varphi_{\mathcal{H}}(X_i)^t$ for $1 \leq i \leq n$. Let us denote by $k_1, \ldots, k_{R^d}$ the elements of $\mathcal{S}_j(\mathcal{H})$. Then we choose

$$\begin{aligned} \alpha_{\mathcal{H}}^{\diamond} &\in \arg\min_{a \in \mathbb{R}^{R^d}} \sum_{i=1}^{n} \left(Y_i - \sum_{t=1}^{R^d} a_t \varphi_{j,k_t}(X_i) \right)^2 \mathbb{1}_{\mathcal{H}}(X_i) \\ &= \arg\min_{a \in \mathbb{R}^{R^d}} \|Y_{\mathcal{H}} - B_{\mathcal{H}} \cdot a\|^2_{\ell_2(\mathbb{R}^n)}, \end{aligned} \tag{4}$$

where we set $\alpha_{\mathcal{H}}^{\diamond} = 0$ if the arg min above contains more than one element. Let us write $Q_{\mathcal{H}} = B_{\mathcal{H}}^t \cdot B_{\mathcal{H}}/n \in \mathcal{M}_{R^d,R^d}$. As is well known, when $Q_{\mathcal{H}}$ is invertible, the arg min on the rhs of eq. (4) admits one single element which writes as follows,

$$\alpha_{\mathcal{H}}^{\diamond} = Q_{\mathcal{H}}^{-1} \cdot \frac{1}{n} B_{\mathcal{H}}^t \cdot Y_{\mathcal{H}}. \tag{5}$$

Naturally, we will denote the corresponding estimator of $\mathscr{P}_j\eta$ at point x by $\eta_{\mathcal{H}}^{\diamond}(x) = \langle \alpha_{\mathcal{H}}^{\diamond}, \varphi_{\mathcal{H}}(x) \rangle$.

We now introduce a thresholded version of $\eta^{\diamond}_{\mathcal{H}}$ based on the spectral thresholding of $Q_{\mathcal{H}}$. We denote by $\lambda_{\min}(Q_{\mathcal{H}})$ the smallest eigenvalue of $Q_{\mathcal{H}}$ in the case where $r \geq 2$, when $Q_{\mathcal{H}}$ is actually a matrix, and $Q_{\mathcal{H}}$ itself in the case where $r = 1$, when it is a real number. Furthermore, we define

$$\eta^{@}_{\mathcal{H}}(x) = \begin{cases} 0 & \text{if } \pi_n^{-1} > \lambda_{\min}(Q_{\mathcal{H}}) \\ \eta^{\diamond}_{\mathcal{H}}(x) & \text{otherwise} \end{cases}, \tag{6}$$

where π_n is a tuning parameter. In practice, and unless otherwise stated, we choose $\pi_n = \log n$. Moreover, we assume throughout the paper that n is large enough so that $\pi_n^{-1} \leq \min(\frac{g_{\min}}{2}, 1)$, where, for reasons that will clarified later, we have denoted,

$$g_{\min} := \mu_{\min} c_{\min}, \tag{7}$$

and $c_{\min}$ stands for the strictly positive constant defined in the proof of Proposition 12.4. Ultimately, the estimator $\eta^{@}_{j}$ of $\mathscr{P}_j \eta$ is defined as,

$$\eta^{@}_{j}(x) = \sum_{\mathcal{H} \in \mathcal{F}_j} \eta^{@}_{\mathcal{H}}(x) \mathbb{1}_{\mathcal{H}}(x), \qquad x \in \mathcal{E}. \tag{8}$$

7. The results

Let r be a natural integer, denote by $\mathcal{P}$ the set of all distributions on $\mathcal{E} \times \mathcal{Y}$ and write

$$\mathcal{P}(\mathbf{CS1}, \mathbf{H}^{\mathbf{r}}_{\mathbf{s}}) := \{\mathbb{P} \in \mathcal{P} : \ (\mathbf{CS1}) \text{ and } (\mathbf{H}^{\mathbf{r}}_{\mathbf{s}}) \text{ hold true}\}. \tag{9}$$

Furthermore, we define j_r, j_s, J and $t(n)$ such that,

$$2^{j_r} = \lfloor n^{\frac{1}{2r+d}} \rfloor, \qquad 2^{j_s} = \lfloor n^{\frac{1}{2s+d}} \rfloor,$$
$$2^{Jd} = \lfloor n t(n)^{-2} \rfloor, \qquad t(n)^2 = \kappa \pi_n^2 \log n,$$

where κ is a positive real number to be chosen later. In addition, we write $\mathcal{J}_n = \{j_r, j_r + 1, \ldots, J-1, J\}$. Notice that j_s strikes the balance between bias and variance in the sense that, for $\log n \geq (2s+d)\log 2$ and $s \in (0, r)$, one has got

$$n^{-\frac{1}{2}} 2^{j_s \frac{d}{2}} \leq 2^{-j_s s}, \tag{10a}$$
$$2^{-j_s s} \leq 2^{r + \frac{d}{2}} 2^{j_s \frac{d}{2}} n^{-\frac{1}{2}}, \tag{10b}$$
$$2^{-j_s s} \leq 2^r n^{-\frac{s}{2s+d}}. \tag{10c}$$

Throughout the sequel, we assume that n is large enough so that the latter inequalities hold true. Our first result gives an upper bound on the probability of deviation of $\eta^{@}_{j}$ form η at a point $x \in \mathcal{A}$.

Theorem 7.1. *Fix $r \in \mathbb{N}$ and assume we are under* (**CS1**) *and* ($\mathbf{H}_\mathbf{s}^\mathbf{r}$). *Recall that $\eta_j^{@}$ is defined in eq. (8). Then, for all $j \in \mathcal{J}_n$, all $\delta > 2M2^{-js}\max(1, 3\pi_n R^d \mu_{\max})$ and all $x \in \mathcal{A}$, we have got*

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathbf{CS1},\mathbf{H}_\mathbf{s}^\mathbf{r})} \mathbb{P}^{\otimes n}(|\eta(x) - \eta_j^{@}(x)| \geq \delta)$$
$$\leq 2R^{2d}\exp\left(-n2^{-jd}\frac{\pi_n^{-2}}{2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}\pi_n^{-1}}\right)\mathbb{1}_{\{\delta\leq M\}} + R^d\Lambda\left(\frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^d}\right), \tag{11}$$

where Λ is defined as follows,

$$\Lambda(\delta) = \begin{cases} 2\exp\left(-\dfrac{n\delta^2}{18K^2\mu_{\max} + 4K2^{j\frac{d}{2}}\delta}\right), & \text{under } (\mathbf{N1}) \\ 1 \wedge \left\{\dfrac{2\sigma(\mu_{\max} + 2^{j\frac{d}{2}}\delta)^{\frac{1}{2}}}{\delta\sqrt{2\pi n}}\exp\left(-\dfrac{n\delta^2\sigma^{-2}}{\mu_{\max} + 2^{j\frac{d}{2}}\delta}\right)\right\} & \\ \quad +2\exp\left(-\dfrac{n\delta^2}{2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta}\right), & \text{under } (\mathbf{N2}) \end{cases}$$

As a consequence of the above theorem, we can deduce the (near) minimax optimality of $\eta_{j_s}^{@}$ over generalized Lipschitz balls.

Corollary 7.1. Fix $r \in \mathbb{N}$ and assume we are under (**CS1**) and ($\mathbf{H}_\mathbf{s}^\mathbf{r}$). Then, for any $p \in [1, \infty)$ and $j \in \mathcal{J}_n$, one has got

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathbf{CS1},\mathbf{H}_\mathbf{s}^\mathbf{r})} \mathbb{E}^{\otimes n}\|\eta - \eta_j^{@}\|_{\mathbb{L}_p(\mathcal{E},\mu)}^p \leq C(p)\pi_n^p \max\left(2^{-js}, \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p, \tag{12}$$

where $\eta_j^{@}$ and $C(p)$ are defined in eq. (8) and Proposition 12.1 below, respectively. *A fortiori*, when s is **known**, we can choose $j = j_s$ and apply eq. (10a) and eq. (10c) above to obtain

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathbf{CS1},\mathbf{H}_\mathbf{s}^\mathbf{r})} \mathbb{E}^{\otimes n}\|\eta - \eta_{j_s}^{@}\|_{\mathbb{L}_p(\mathcal{E},\mu)}^p \leq C(p)2^{rp}\pi_n^p n^{-\frac{sp}{2s+d}}.$$

This, together with the lower-bound of Theorem 7.3, proves that $\eta_{j_s}^{@}$ is (nearly) minimax optimal over the generalized Lipschitz ball $\mathscr{L}^s(\mathcal{E}, M)$ of radius M.

The next Theorem shows that the approximation level j can be determined from the data $\mathcal{D}_n$ so that we obtain adaptation over a wide generalized Lipschitz scale.

Theorem 7.2. *Fix $r \in \mathbb{N}$ and assume we are under* (**CS1**) *and* ($\mathbf{H}_\mathbf{s}^\mathbf{r}$). *We define*

$$g(j,k) := \left(\frac{2^{j\frac{d}{2}}}{\sqrt{n}}t(n) + \frac{2^{k\frac{d}{2}}}{\sqrt{n}}t(n)\right),$$
$$j^{@}(x) := \inf\{j \in \mathcal{J}_n : |\eta_j^{@}(x) - \eta_k^{@}(x)| \leq g(j,k), \forall k \in \mathcal{J}_n, k > j\}, \quad x \in \mathcal{A},$$

where $\eta_j^{@}$ *is defined in eq. (8) and* $\inf\emptyset = \max(\mathcal{J}_n) = J$. *If* κ *is chosen large enough, meaning* $\kappa \geq \frac{p}{2}C_9^{-1}$, *where* C_9 *is defined in Proposition 12.2, then we obtain*

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathbf{CS1},\mathbf{H_s^r})} \mathbb{E}^{\otimes n}\|\eta_{j^{@}(.)}^{@}(.) - \eta(.)\|_{\mathbb{L}_p(\mathcal{E},\mu)}^p \leq 5^p 2^{rp} t(n)^p n^{-\frac{sp}{2s+d}}.$$

So that $\eta_{j^{@}(.)}^{@}(.)$ *is a nearly minimax adaptive estimator of* η *over the generalized Lipschitz scale* $\bigcup_{0<s<r}\mathscr{L}^s(\mathcal{E},M)$.

Finally, we prove that $\eta^{@}$ is indeed (nearly) minimax optimal by giving the corresponding lower-bound result.

Theorem 7.3. *Assume we are under* (**CS1**) *and* ($\mathbf{H_s^r}$). *We write* $\inf_{\theta_n}$ *the infinimum over all estimators* θ_n *of* η, *that is all measurable functions of the data* $\mathcal{D}_n$. *Then, for* $d \geq 1$, $s > 0$, *we have, for all* $1 \leq p < \infty$,

$$\inf_{\theta_n}\sup_{\mathbb{P}\in\mathcal{P}(\mathbf{CS1},\mathbf{H_s^r})} \mathbb{E}^{\otimes n}\|\theta_n - \eta\|_{\mathbb{L}_p(\mathcal{E},\mu)}^p \gtrsim n^{-\frac{sp}{2s+d}}.$$

The next section shows how these results can be improved in the case where we benefit from additional information on μ or η.

8. Refinement of the results

As can be seen from Corollary 7.1 and Theorem 7.2 above, π_n appears as a multiplicative factor in the upper-bounds and thus deteriorates them by a multiplicative $\log n$ term. However, this needs not be the case, and under appropriate additional assumptions, π_n can be chosen to be a constant. Consider indeed the following two assumptions.

(**O1**) We know $\mu_{\min}^* \in \mathbb{R}$, such that $0 < \mu_{\min}^* \leq \mu_{\min}$.

(**O2**) We know a finite positive real number M such that $\|\eta\|_{\mathbb{L}_\infty(\mathcal{E},\lambda)} \leq M$.

Under (**O1**), we know a lower bound $\mu_{\min}^*$ of $\mu_{\min}$, and therefore a lower bound $g_{\min}^*$ of $g_{\min}$ (see eq. (7)). Under (**O1**), we will thus choose $\pi_n^{-1} = \min(\frac{g_{\min}^*}{2}, 1)$. It is straightforward to show that Theorem 7.1 is still valid with this new value of π_n (see Remark 12.1 in the proof of Theorem 7.1), and thus all the subsequent results follow as well. Under (**O2**), we know an upper bound M of the essential supremum of η on $\mathcal{E}$. In that case, we redefine

$$\eta_{\mathcal{H}}^{@}(x) = T_M(\eta_{\mathcal{H}}^{\diamond}(x))\mathbb{1}_{\{\lambda_{\min}(Q_{\mathcal{H}})>0\}}, \tag{13}$$

where, for any $z \in \mathbb{R}$, we have written $T_M(z) = z\mathbb{1}_{\{|z|\leq M\}} + M\text{sign}(z)\mathbb{1}_{\{|z|>M\}}$. Once again, it is straightforward to show that Theorem 7.1 is now valid with $\pi_n^{-1} = \min(\frac{g_{\min}}{2}, 1)$ and $2M$ in place of M in the indicator function on the rhs of eq. (11) (see Remark 12.1 in the proof of Theorem 7.1), and thus all the subsequent results follow as well.

Notice that π_n is an absolute constant under (**O1**) and (**O2**), while it is an increasing sequence of n to be fine-tuned by the statistician otherwise. Hence π_n appears to be the price to pay for not knowing a lower bound of $\mu_{\min}$ or an upper bound of the essential supremum of η on $\mathcal{E}$.

9. Relaxation of assumption (S1)

9.1. The problem

Now, we would like to relax assumption (**S1**) and allow for $\mathcal{A}$ to be an unknown subset of $\mathcal{E}$, eventually disconnected. Under (**CS1**), the success of $\eta^{@}$ stems from the fact that it is constructed upon an approximation grid of the form $2^{-j}\mathbb{Z}^d \cap [0,1]^d$, whose edges coincide exactly with the boundary of $\mathcal{A}$. In the case where $\mathcal{A}$ is unknown, some cells of the lattice might straddle the boundary of $\mathcal{A}$ and thus require a new treatment.

In order to handle this new configuration, we will need to make a smoothness assumption on the boundary of $\mathcal{A}$ and allow for the estimation cells to move with the point at which we want to estimate η. Ultimately, we devise a new estimator $\eta^{\maltese}$ of η which is built upon a moving approximation grid. In fact, this new estimation method ensures that the point x at which we want to estimate η always belongs to a cell $\mathcal{H}$ of $\mathcal{F}_j$ at resolution level j, whose center belongs to $\mathcal{A}$. This will ensure that local regressions performed on cells that straddle the boundary of $\mathcal{A}$ are still meaningful.

The smoothness assumption we will make on $\mathcal{A}$ might be compared to the support assumption made in [4, eq. (2.1)] in the classification context. In substance, it is assumed in [4] that $\mathcal{A}$ is locally ball-shaped to be compatible with the ball-shaped support of the LPE kernel, which they use to estimate η. In our case, we perform estimation with multi-dimensional scaling functions whose supports are cube-shaped and will thus assume that $\mathcal{A}$ is locally cube-shaped.

9.2. Smoothness assumption on $\mathcal{A}$

Let us now make these informal arguments more precise. To that end we introduce assumption (**S2**) as an alternative to (**S1**) above. Fix an absolute constant $\mathfrak{m}_0 \in (0,1)$ and recall that $2^{j_s} = \lfloor n^{\frac{1}{2s+d}} \rfloor$. With these notations, (**S2**) goes as follows,

(**S2**) $\mathcal{E} = \mathbb{R}^d$ and $\mathcal{A}$ belongs to $\mathscr{A}_{j_s}$, where

$$\mathscr{A}_{j_s} := \{\mathcal{A} \subset \mathbb{R}^d : \exists \mathfrak{m} \geq \mathfrak{m}_0, \forall x \in \mathcal{A}, \\ \exists z_x \in \mathbb{R}^d, 0 \in \mathcal{B}_\infty(z_x, \mathfrak{m}) \subset 2^{j_s}(\mathcal{A} - x)\},$$

In words, (**S2**) means that if we zoom close enough to any $x \in \mathcal{A}$, we can find a hypercube $\mathcal{B}_\infty(z_x, \mathfrak{m})$ that contains x and is a subset of $\mathcal{A}$. Notice readily that for all $j_1 \geq j_2$, the component of $2^{j_2}(\mathcal{A} - x)$ that contains 0 is a subset of the component of $2^{j_1}(\mathcal{A} - x)$ that contains 0, so that $\mathscr{A}_{j_2} \subset \mathscr{A}_{j_1}$. Therefore $\mathscr{A}_{j_s}$ grows with n and shrinks with s. Of course, (**S1**) is a particular case of (**S2**). Setting (**S2**) allows $\mathcal{A}$ to be unknown and belong to a wide class of subsets of $\mathbb{R}^d$, eventually disconnected (see Figure 2).

In the sequel, we will conveniently refer by (**CS2**) to the set of assumptions (**D1**), (**S2**), (**N1**) or (**N2**).

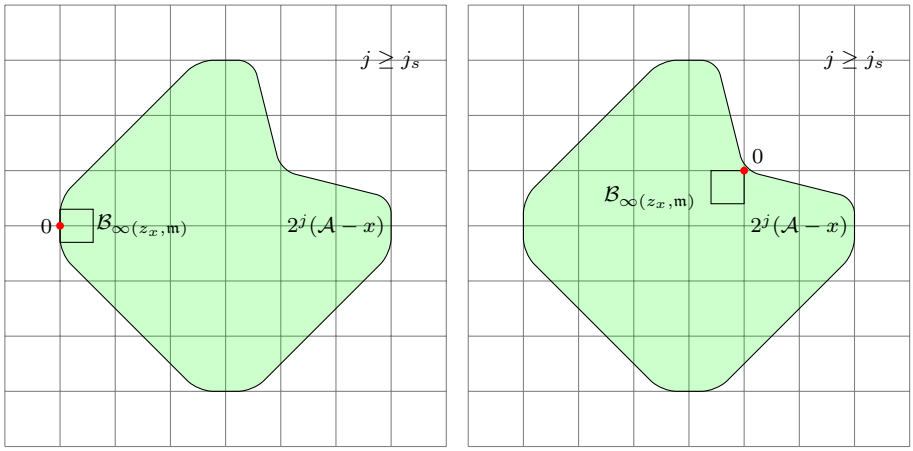


FIG 2. (**S2**) *allows for $\mathcal{A}$ to be non-convex and eventually disconnected.*

9.3. Moving local estimation under (CS2)

As detailed above, η^{Ψ} is obtained by local regression on a moving approximation grid. Let us describe the construction of η^{Ψ} more precisely.

First of all, we split the sample into two pieces. For simplicity, let us assume that we dispose of $2n$ data points. The first half of the sample points, which we denote by $\mathcal{D}'_n = \{(X'_i, Y'_i), i = 1, \dots, n\}$, will be used to identify the support $\mathcal{A}$ of μ, while the second half, which we denote by $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \dots, n\}$, will be used to estimate the scaling functions coefficients by local regressions.

Let us denote by $\mathcal{H}_0$ the cell $2^{-j}[0,1]^d$ of the lattice $2^{-j}\mathbb{Z}^d$ at resolution j. And denote by $\mathcal{H}_0(x)$ the same cell centered in x, that is $\mathcal{H}_0(x) = x - 2^{-j-1} + 2^{-j}[0,1]^d$. Then, the construction of $\eta_j^{\Psi}(x)$ at a point $x \in \mathbb{R}^d$ goes as follows. (i) If none of the design points (X'_i) of the sample $\mathcal{D}'_n$ lie in $\mathcal{H}_0(x)$, then take $\eta^{\Psi}(x) = 0$. (ii) If one or more design points of the sample $\mathcal{D}'_n$ lie in $\mathcal{H}_0(x)$, we select one of them and denote it by X'_{i_x} (the selection procedure is of no importance beyond computational considerations). By construction, x belongs to the cell $\mathcal{H}_0(X'_{i_x})$ centered in $X'_{i_x} \in \mathcal{A}$. Since X'_{i_x} belongs to $\mathcal{A}$, it makes sense to perform a local regression on $\mathcal{H}_0(X'_{i_x})$ with the sample points $\mathcal{D}_n$, which gives rise to an estimator η^{Ψ} of η valid at any point of $\mathcal{H}_0(X'_{i_x}) \cap \mathcal{A}$. It is noteworthy that this procedure uses the sample $\mathcal{D}'_n$ to identify the support $\mathcal{A}$ of μ.

Interestingly, the above estimation procedure requires at most as many regressions as there are data points in $\mathcal{D}'_n$ to return an estimator η^{Ψ} of η at every single point $x \in \mathcal{A}$. It is therefore computationally more efficient than any other kernel estimator, such as the LPE. The computational performance of η^{Ψ} can in fact be further improved in the sense that the local regression on the cell $\mathcal{H}_0(X'_i)$ can be omitted if the cell $\mathcal{H}_0(X'_i)$ is itself included in the union of cells centered at other design points of $\mathcal{D}'_n$. In particular, we can choose X'_{i_x} to be a design point X'_i of $\mathcal{D}'_n$ that belongs to $\mathcal{H}_0(x)$ and for which a local regression

has already been performed, if it exists, or any one of the X'_i that belong to $\mathcal{H}_0(x)$ otherwise.

Intuitively, the computational efficiency of $\eta^{\maltese}$ stems from the fact that the design points (X'_i) provide some valuable information on the unknown support $\mathcal{A}$ of μ, which can be exploited under (**CS2**). In particular, and as we will see below, (**D1**) guarantees that the design points of $\mathcal{D}'_n$ populate $\mathcal{A}$ densely enough so that, as long as $j \leq J$, the cells $\mathcal{H}_0(X'_i)$, $1 \leq i \leq n$, form a cover of $\mathcal{A}$, modulo a set whose μ-measure decreases almost exponentially fast toward zero with n.

9.4. Construction of the local estimator $\eta^{\maltese}$

Assume we are under (**S2**) and work with the Daubechies' r-MRA of $\mathbb{L}_2(\mathbb{R}^d, \lambda)$. Obviously, shifting the approximation grid is equivalent to shifting the data points (X_i) of $\mathcal{D}_n$ and keeping the lattice fixed. For ease of notations and clarity, we adopt this second point of view. In order to compute $\eta^{\maltese}$ at a point $x \in \mathcal{H}_0(X'_{i_x}) \cap \mathcal{A}$, we want to shift the design points in such a way that X'_{i_x} falls right in the middle of $\mathcal{H}_0$. In other words, we want X'_{i_x} to be shifted at point $2^{-j-1} \in \mathbb{R}^d$ (whose coordinates are worth $2^{-j-1} \in \mathbb{R}$). This corresponds to the change of variable $\widetilde{X}_i = X_i - (X'_{i_x} - 2^{-j-1})$, where we have denoted by X_i and $\widetilde{X}_i$ the representations of a same data point in the canonical and shifted coordinate systems of $\mathbb{R}^d$, respectively. In order to compute $\eta^{\maltese}$ at point $x \in \mathcal{H}_0(X'_{i_x}) \cap \mathcal{A}$, it is therefore enough to perform a local regression on $\mathcal{H}_0$ against the shifted data points,

$$\widetilde{\mathcal{D}}_x = \{(\widetilde{X}_i, Y_i), i = 1, \ldots, n\}.$$

For the sake of concision, we will denote by $\widetilde{u} = u - (X'_{i_x} - 2^{-j-1})$ the coordinate representation of a point u in the shifted coordinate system of $\mathbb{R}^d$. Let us denote by $k_1, \ldots, k_{R^d}$ the elements of $\mathcal{S}_j(\mathcal{H}_0)$. With these notations, eq. (4) must be corrected and written as

$$\alpha^{\diamond}_{\mathcal{H}_0} \in \arg\min_{a \in \mathbb{R}^{R^d}} \sum_{i=1}^{n} \left(Y_i - \sum_{t=1}^{R^d} a_t \varphi_{j,k_t}(\widetilde{X}_i) \right)^2 \mathbb{1}_{\mathcal{H}}(\widetilde{X}_i), \tag{14}$$

where we set $\alpha^{\diamond}_{\mathcal{H}_0} = 0$ if the arg min above contains more than one element. The notations introduced in Section 5.3 can be updated to this new setting as follows. $B_{\mathcal{H}_0}$ stands now for the random matrix of $\mathcal{M}_{n,R^d}$ whose rows are the $\varphi_{\mathcal{H}_0}(\widetilde{X}_i)^t$, $i = 1, \ldots, n$. In addition, we recall that we have defined $Q_{\mathcal{H}_0} = B^t_{\mathcal{H}_0} \cdot B_{\mathcal{H}_0}/n \in \mathcal{M}_{R^d,R^d}$. Its coefficients write thus as

$$[Q_{\mathcal{H}_0}]_{\nu,\nu'} = \frac{1}{n} \sum_{i=1}^{n} \varphi_{j,\nu}(\widetilde{X}_i) \varphi_{j,\nu'}(\widetilde{X}_i) \mathbb{1}_{\mathcal{H}_0}(\widetilde{X}_i), \qquad \nu, \nu' \in \mathcal{S}_j(\mathcal{H}_0).$$

Notice here that $\mathcal{S}_j(\mathcal{H}_0) = \{\nu \in \mathbb{Z}^d : 2^{-1} \in \mathrm{Supp}\varphi_\nu\}$, which neither depends on j nor x. Therefore, and for later reference, we denote

$$\mathfrak{S} := \{\nu \in \mathbb{Z}^d : 2^{-1} \in \mathrm{Supp}\varphi_\nu\}, \tag{15}$$

In addition, if we write $Y_{\mathcal{H}_0} = (Y_i \mathbb{1}_{\mathcal{H}_0}(\widetilde{X}_i))_{1\leq i\leq n}$, then eq. (5) still holds true when the solution to eq. (14) is unique. So that, for all $x \in \mathcal{H}_0(X_{i_x}) \cap \mathcal{A}$, we can write $\eta^{\diamond}_{\mathcal{H}_0}(\widetilde{x}) = \langle \alpha^{\diamond}_{\mathcal{H}_0}, \varphi_{\mathcal{H}_0}(\widetilde{x})\rangle$. Finally eq. (6) remains valid with X_i replaced by $\widetilde{X}_i$ and $\mathcal{H}$ by $\mathcal{H}_0$, $\eta^{@}_{\mathcal{H}_0}$ redefined as $\eta^{\maltese}_{\mathcal{H}_0}$ and $g_{\min}$ redefined as

$$g_{\min} = \mu_{\min} c_{\min}, \tag{16}$$

where $c_{\min}$ is the strictly positive constant defined in Lemma 12.1 below. So that ultimately, the estimator $\eta^{\maltese}_j$ of $\mathscr{P}_j\eta$ at a point $x \in \mathbb{R}^d$ writes as

$$\eta^{\maltese}_j(x) = \eta^{\maltese}_{\mathcal{H}_0}(\widetilde{x}), \qquad\qquad x \in \mathcal{E}. \tag{17}$$

Notice that by contrast with eq. (8) above, the sum over the hypercubes of $\mathcal{F}_j$ has disappeared. This is due to the fact that the approximation grid moves with x so that we end up virtually always performing estimation on the same hypercube $\mathcal{H}_0$.

9.5. The results

Interestingly, $\eta^{\maltese}$ still verifies similar results as the ones described in Section 7. To be more precise, recall that we work with a sample of size $2n$ broken up into two pieces $\mathcal{D}_n$ and $\mathcal{D}'_n$ of size n. Let us redefine $\mathcal{J}_n$ so that $\mathcal{J}_n = \{j_s, j_s + 1, \ldots, J-1, J\}$ where $2^{j_s} = \lfloor n^{\frac{1}{2s+d}} \rfloor$. Then, we obtain the following result in place of Theorem 7.1.

Theorem 9.1. *Fix $r \in \mathbb{N}$ and assume we are under* (**CS2**) *and* ($\mathbf{H}^{\mathbf{r}}_{\mathbf{s}}$). *Recall that $\eta^{\maltese}_j$ is defined in eq. (17). Then, for all $j \in \mathcal{J}_n$, all $\delta > 2M2^{-j_s}\max(1, 3\pi_n R^d \mu_{\max})$ and all $x \in \mathcal{A}$, we have got*

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}(\mathbf{CS2},\mathbf{H}^{\mathbf{r}}_{\mathbf{s}})} \mathbb{P}^{\otimes n}(|\eta(x) - \eta^{\maltese}_j(x)| \geq \delta) \\
&\leq 3R^{2d} \exp\left(-n2^{-jd} \frac{\pi_n^{-2}}{2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}\pi_n^{-1}}\right) \mathbb{1}_{\{\delta\leq M\}} \\
&+ R^d \Lambda\left(\frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^d}\right),
\end{aligned}$$

where Λ has been defined in Theorem 7.1.

Left aside the fact that $\eta^{\maltese}$ is constructed upon a sample of size $2n$, the sole difference with the result of Theorem 7.1 is that the leading constant in front of the exponential on the second line has changed from $2R^d$ to $3R^d$. Furthermore, it is straightforward to deduce from Theorem 9.1 results similar to Corollary 7.1, Theorem 7.2 and Theorem 7.3, and a fortiori the refined results obtained in Section 8, for $\eta^{\maltese}$ under (**CS2**). The proofs of these results for $\eta^{\maltese}$ under the set of assumptions (**CS2**) follow, for the most part, exactly the same lines as the proofs given for $\eta^{@}$ under (**CS1**). Details can be found in Section 12.2.

10. Classification via local multi-resolution projections

Recall from [4] that the margin assumption can be written as,
(**MA**) There exist constants $C_* > 0$ and $\vartheta \geq 0$ such that

$$\mathbb{P}(0 < |2\eta(X) - 1| \leq t) \leq C_* t^{\vartheta}, \quad \forall t > 0.$$

The binary classification setting corresponds to (**CS2**), under assumptions (**N1**) and (**O2**). Notice besides that we have $K = 1$ in (**N1**) and $M = 1$ in ($\mathbf{H}_{\mathbf{s}}^{\mathbf{r}}$). Since we are under (**O2**), it follows from Section 8 that $\pi_n = \pi_0 = \min(1, \frac{g_{\min}}{2})$ is independent of n and $\eta^{\maltese}$ is capped at $M = 1$ as in eq. (13). For the sake of coherence, we denote by $j^{\maltese}$ the adaptive resolution level built upon $\eta^{\maltese}$, as described in Theorem 7.2, and define $\mathcal{P}(\mathbf{CS2}, \mathbf{H}_{\mathbf{s}}^{\mathbf{r}})$ by analogy with eq. (9) above. Finally, we recall that $\eta^{\maltese}$ is built upon a sample of size $2n$ split into two sub-samples $\mathcal{D}_n$ and $\mathcal{D}'_n$ of size n.

As a consequence of Theorem 9.1, we can use the plug-in classifier built upon $\eta^{\maltese}$ to obtain similar results as the ones given in [4, Lemma 3.1] for LPE based plug-in classifiers.

Corollary 10.1. Fix $r \in \mathbb{N}$ and assume we are in the binary classification setting. Assume moreover that ($\mathbf{H}_{\mathbf{s}}^{\mathbf{r}}$) and (**MA**) hold true. Consider the plug-in classifiers $h_{j_s}^{\maltese}(.) = \mathbb{1}_{\{\eta_{j_s}^{\maltese}(.) \geq \frac{1}{2}\}}$ and $h_{j^{\maltese}}^{\maltese}(.) = \mathbb{1}_{\{\eta_{j^{\maltese}(.)}^{\maltese}(.) \geq \frac{1}{2}\}}$. Then, as soon as $\kappa > C_0(1 + \vartheta)$, we have

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_{\mathbf{s}}^{\mathbf{r}}, \mathbf{MA})} \mathscr{T}(h_{j_s}^{\maltese}) \leq C_1 n^{-\frac{s}{2s+d}(1+\vartheta)}, \tag{18}$$

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_{\mathbf{s}}^{\mathbf{r}}, \mathbf{MA})} \mathscr{T}(h_{j^{\maltese}}^{\maltese}) \leq C_2 (\log n)^{\frac{1+\vartheta}{2}} n^{-\frac{s}{2s+d}(1+\vartheta)}, \tag{19}$$

where the classification risk $\mathscr{T}(.)$ has been defined in Section 1 and the constants C_0, C_1, C_2 are made explicit in [39] and only depend on $\mu_{\max}, \mu_{min}, r, d$ and ϑ.

In fact, it can be shown that the classifiers $h^{\maltese}$ defined in Corollary 10.1 are (nearly) minimax optimal. Proofs of Corollary 10.1 and the associated lower-bound can be found in [39].

11. Simulation study

In order to illustrate the performance of $\eta_{j^{@}}^{@}$, we have carried out a simulation study in the regression setting in the one-dimensional case, that is with $d = 1$. As detailed earlier, the sole purpose of this simulation is to show that (1) $\eta^{@}$ can be easily implemented and is computationally efficient, (2) $\eta^{@}$ works well in practice in the case where the density of the design μ is discontinuous, (3) and to give an intuitive visual feel for $\eta^{@}$, which is built upon the juxtaposition of local regressions against a set of scaling functions. In particular, we run our simulation against benchmark signals, which allows to compare them with the ones detailed in the literature for alternative kernel estimators (see simulation

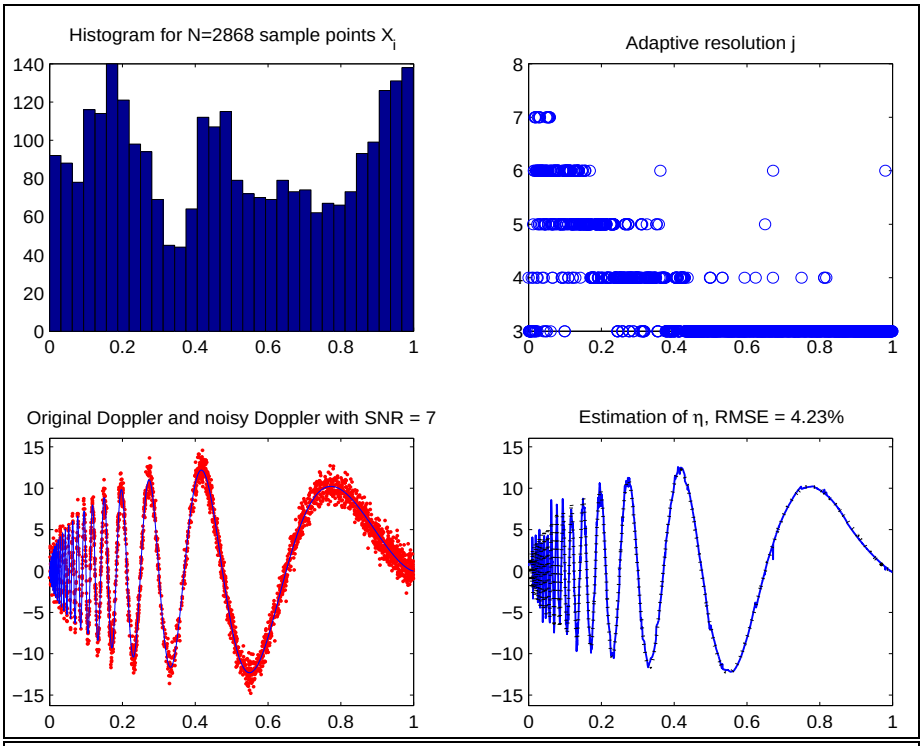
study in [32], for example). We have run them under (**CS1**), which corresponds to the case where $\eta_j^{@}$ can be completely computed with exactly 2^j regressions. We have in particular $\mathcal{E} = [0,1] = \mathcal{A}$. We focus on the functions η introduced in [14] and used as a benchmark in numerous subsequent simulation studies. They are made available through the Wavelab850 library freely available at `http://www-stat.stanford.edu/~wavelab/`. In addition we have chosen the noise ξ to be standard normal, that is we are working under (**N2**) with $\sigma = 1$. In all cases, we have chosen the signal-to-noise ratio (SNR) to be equal to 7. To be more specific, we are working on a dyadic grid G of $[0,1]$ of resolution 2^{-15}. We compute the root-mean-squared-error (RMSE) of both the signal and the noise on that grid and rescale the signal so that its RMSE be seven times bigger than the one of the noise.

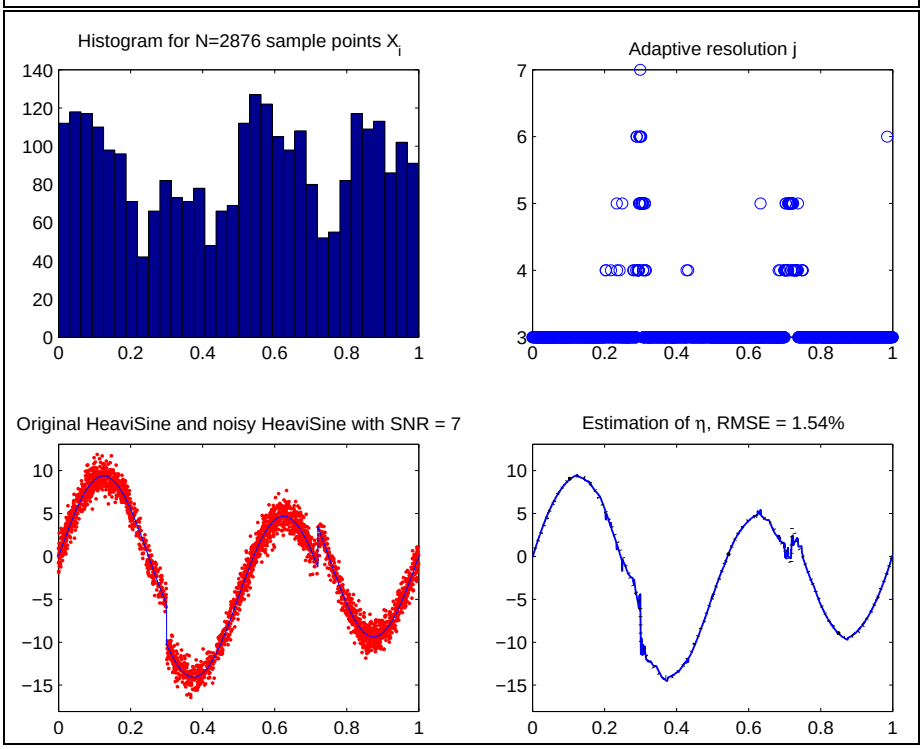
Let us now give details about the simulation of the sample points and the computation of the estimator. We divide the unit-interval into ten sub-segments $A_k := 10^{-1}[k, k+1]$ for $k = 0, \dots, 9$. We define the density of X as follows.

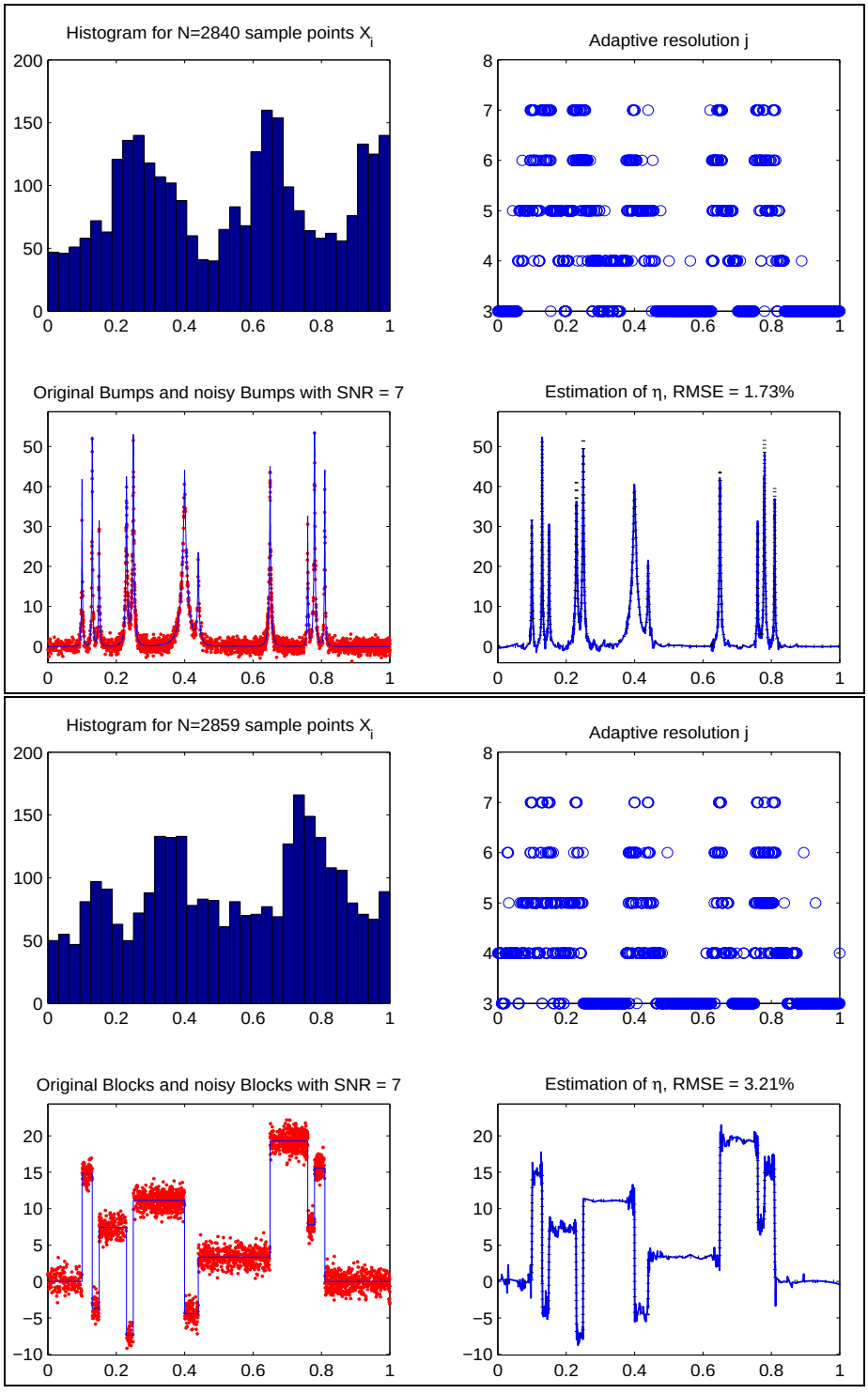
$$\mu(x) = \sum_{k=0}^{9} p_k \lambda(A_k)^{-1} \mathbb{1}_{A_k}(x).$$

We choose the p_k's at random. To that end, we denote by $(u_k)_{0 \le k \le 9}$ ten realizations of the uniform random variable on $[.25, 1]$, write $v = u_0 + \cdots + u_9$ and set $p_k = u_k v^{-1}$. Notice that this guarantees that $\mu \ge \min_{0 \le k \le 9} 10 p_k \ge \mu_{\min} = 0.25$ on $[0,1]$. We then simulate 3000 sample points X_i according to μ. Finally, we bring the points back on the grid G by assimilating them to their nearest grid node. Since the X_i's are supposed to be drawn from a law that is absolutely continuous with respect to the Lebesgue measure on $[0,1]$, we must keep only one data point per grid node. This reduces the number of data points from 3000 to the number that is reported on top of each of the histograms.

In order to compute the adaptive estimator at sample points X_i, we use the boundary-corrected scaling functions coded into Wavelab850 for $r = 3$ and for which we must have $j \ge 3$. We set $J = \lceil \log(n/\log n)/\log 2 \rceil$. The elimination of redundant sample points on the grid removes on average 150 points so that we obtain $J = 10$. We therefore have $\mathcal{J}_n = \{3, 4, \dots, 10\}$. Notice interestingly that the computation of $\eta_3^{@}$ requires only 8 regressions and $\eta_{10}^{@}$ requires $1,024$ of them. This is much smaller than for the LPE whose computation necessitates as many regressions as there are sample points at each resolution level. In practice, we compute the minimum eigenvalues of all regression matrices across partitions and resolution levels and choose π_n^{-1} to be the first decile of this set of values. When proving theoretical results, we have chosen $\eta_j^{@}$ to be zero on the small probability event where the minimum eigenvalue of the regression matrix is smaller than π_n^{-1}. In practice we can choose it to be an average value of the nearby cells in order to get an estimator that is overall more appealing to the eye. In our simulation, we in fact do not use that modification. Instead, we modify $j^{@}$ to be the highest $j \in \{3, \dots, j^{@}\}$ such that $\eta_{j^{@}}^{@}$ has been computed from a valid regression matrix, meaning a regression matrix whose smallest eigenvalue is greater than the threshold π_n^{-1}.


Histogram for N=2868 sample points X_i
Adaptive resolution j
Original Doppler and noisy Doppler with SNR = 7
Estimation of η, RMSE = 4.23%


Histogram for N=2876 sample points X_i
Adaptive resolution j
Original HeaviSine and noisy HeaviSine with SNR = 7
Estimation of η, RMSE = 1.54%


Histogram for N=2840 sample points X_i
Adaptive resolution j
Original Bumps and noisy Bumps with SNR = 7
Estimation of η, RMSE = 1.73%
Histogram for N=2859 sample points X_i
Adaptive resolution j
Original Blocks and noisy Blocks with SNR = 7
Estimation of η, RMSE = 3.21%

In practice, for a given signal, we generate μ at random and compute $\eta_{j^@}^{@}$ for 100 samples drawn from μ. We quantify the performance $\eta_{j^@}^{@}$ by its relative RMSE, meaning its RMSE computed at sample points X_i divided by the amplitude of the true signal, that is its maximal absolute value on the underlying dyadic grid. We display results for "Doppler", "HeaviSine", "Bumps" and "Blocks" corresponding to the median performance among the 100 trials. Each figure displays four graphs. Clockwise from the top left corner, they display in turn, an histogram of sample points X_i; the adaptive level $j^@$ at sample points X_i; the true signal (black dots) and the estimator $\eta_{j^@}^{@}$ at sample points X_i (solid blue line) and its corresponding relative RMSE in the title; and finally the original signal (solid blue line) with its noisy version at sample points X_i (red dots).

12. Proofs

12.1. Proof of the upper-bound results under (CS1)

12.1.1. Proof of Corollary 7.1

Consider the term

$$I = \int_{\mathcal{A}} \mathbb{E}[|\eta(x) - \eta_j^{@}(x)|^p]\mu(x)dx.$$

Now, apply Proposition 12.1 and notice that $\int_{\mathcal{A}} \mu(x)dx = 1$ to show that I is upper-bounded by the term that appears on the rhs of eq. (12) stated in Corollary 7.1. In particular, for all $1 \leq p < \infty$, we obtain $I \leq C(p)\pi_n^p t(n)^{-p} \leq C(p) < \infty$. This in turn proves that we can apply the Fubini-Tonelli theorem to get

$$I = \mathbb{E}[\|\eta - \eta_j^{@}\|^p_{\mathbb{L}_p(\mathcal{E},\mu)}],$$

and concludes the proof. □

12.1.2. Proof of Theorem 7.1

Let $x \in \mathcal{A}$ and $j \in \mathcal{J}_n$. There exists $\mathcal{H} \in \mathcal{F}_j$ such that $x \in \mathcal{H}$. Let us work on the set $\{\lambda_{\min}(Q_{\mathcal{H}}) \geq \pi_n^{-1}\}$ on which $Q_{\mathcal{H}}$ is invertible. On that set, we can write

$$\begin{aligned}
|\mathscr{P}_j\eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)| &= |\langle \alpha_{\mathcal{H}} - \alpha_{\mathcal{H}}^{\diamond}, \varphi_{\mathcal{H}}(x)\rangle| \\
&= \left|\left\langle Q_{\mathcal{H}}^{-1} \cdot \left(\frac{B_{\mathcal{H}}^t}{n} \cdot (B_{\mathcal{H}} \cdot \alpha_{\mathcal{H}} - Y_{\mathcal{H}})\right), \varphi_{\mathcal{H}}(x)\right\rangle\right| \\
&\leq \|Q_{\mathcal{H}}^{-1}\|_S \left\|\frac{B_{\mathcal{H}}^t}{n} \cdot (B_{\mathcal{H}} \cdot \alpha_{\mathcal{H}} - Y_{\mathcal{H}})\right\|_{\ell^2(\mathbb{R}^{R^d})} \|\varphi_{\mathcal{H}}(x)\|_{\ell^2(\mathbb{R}^{R^d})} \\
&\leq R^{\frac{d}{2}} 2^{j\frac{d}{2}} \lambda_{\min}(Q_{\mathcal{H}})^{-1} \left\|\frac{B_{\mathcal{H}}^t}{n} \cdot (B_{\mathcal{H}} \cdot \alpha_{\mathcal{H}} - Y_{\mathcal{H}})\right\|_{\ell^2(\mathbb{R}^{R^d})}.
\end{aligned}$$

Now, notice that for all $X_i \in \mathcal{H}$, we have $Y_i = \langle \alpha_{\mathcal{H}}, \varphi_{\mathcal{H}}(X_i)\rangle + \mathscr{R}_j\eta(X_i) + \xi_i$. Write $\mathscr{R}_{\mathcal{H}} = (\mathscr{R}_j\eta(X_i)\mathbb{1}_{\mathcal{H}}(X_i))_{1\le i\le n}$ and $\xi_{\mathcal{H}} = (\xi_i\mathbb{1}_{\mathcal{H}}(X_i))_{1\le i\le n}$. Then, we have,

$$W_{\mathcal{H}} = \left|\frac{B_{\mathcal{H}}^t}{n}\cdot(B_{\mathcal{H}}\cdot\alpha_{\mathcal{H}} - Y_{\mathcal{H}})\right| = \left|\frac{B_{\mathcal{H}}^t}{n}\cdot(\xi_{\mathcal{H}} + \mathscr{R}_{\mathcal{H}})\right| \in \mathbb{R}^{R^d}.$$

Thus, a direct application of Proposition 12.5 allows to write, for $\delta > 2M2^{-js}\max(1, 3\pi_n R^d\mu_{\max})$,

$$\begin{aligned}
&\mathbb{P}(|\eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)| \ge \delta, \lambda_{\min}(Q_{\mathcal{H}}) \ge \pi_n^{-1}) \\
&\qquad \le \mathbb{P}\left(\|W_{\mathcal{H}}\|_{\ell_2(\mathbb{R}^{R^d})} \ge \frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^{\frac{d}{2}}}\right) \\
&\qquad \le R^d \sup_{k\in\mathcal{S}_j(\mathcal{H})} \mathbb{P}\left([W_{\mathcal{H}}]_k \ge \frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^d}\right) \\
&\qquad \le R^d\Lambda\left(\frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^d}\right).
\end{aligned}$$

By definition, we have $\eta_j^{@}(x) = \eta_{\mathcal{H}}^{@}(x)$, so that we have

$$\begin{aligned}
\mathbb{P}(|\eta(x) - \eta_j^{@}(x)| \ge \delta) &= \mathbb{P}(|\eta(x) - \eta_{\mathcal{H}}^{@}(x)| \ge \delta, \lambda_{\min}(Q_{\mathcal{H}}) \ge \pi_n^{-1}) \\
&\quad + \mathbb{P}(|\eta(x) - \eta_{\mathcal{H}}^{@}(x)| \ge \delta, \lambda_{\min}(Q_{\mathcal{H}}) < \pi_n^{-1}).
\end{aligned} \tag{20}$$

By construction, $\eta_{\mathcal{H}}^{@}(x) = \eta_{\mathcal{H}}^{\diamond}(x)$ on the event $\{\lambda_{\min}(Q_{\mathcal{H}}) \ge \pi_n^{-1}\}$ and $\eta_{\mathcal{H}}^{@}(x) = 0$ on its complement. So that we obtain $|\eta(x) - \eta_{\mathcal{H}}^{@}(x)| = |\eta(x)| \le M$ on the rhs of eq. (20). Notice in addition that $M2^{-js} \ge |\mathscr{R}_j\eta(x)|$ under $(\mathbf{H_s^r})$ (see Appendix). Finally, we obtain, for $\frac{\delta}{2} > M2^{-js} \ge |\mathscr{R}_j\eta(x)|$,

$$\begin{aligned}
&\mathbb{P}(|\eta(x) - \eta_j^{@}(x)| \ge \delta) \\
&\quad \le \mathbb{P}(|\mathscr{P}_j\eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)| \ge \frac{\delta}{2}, \lambda_{\min}(Q_{\mathcal{H}}) \ge \pi_n^{-1}) + \mathbb{P}(\lambda_{\min}(Q_{\mathcal{H}}) < \pi_n^{-1})\mathbb{1}_{\{\bar{M}\ge\delta\}},
\end{aligned}$$

where we have written $\bar{M} = M$. The term on the lhs has been dealt with above. The term on the rhs is tackled using Proposition 12.3. This concludes the proof.

Remark 12.1. Under (**O2**), we have $|\eta_{\mathcal{H}}^{@}(x)| \le M$, and since $\eta \in \mathscr{L}^s(\mathcal{E}, M)$, we obtain $|\eta(x) - \eta_{\mathcal{H}}^{@}(x)| \le 2M$ on the rhs of eq. (20). While on the lhs, it is straightforward that (see [22, Chap. 10])

$$|\eta(x) - \eta_{\mathcal{H}}^{@}(x)| = |\eta(x) - T_M(\eta_{\mathcal{H}}^{\diamond}(x))| \le |\eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)|.$$

Under (**O1**), the proof remains unchanged. So that the proof still holds with

$$\bar{M} = \begin{cases} 2M, & \text{under } (\mathbf{O2}), \\ M, & \text{otherwise.} \end{cases}$$

□

12.1.3. Proof of Theorem 7.2

This result is obtained after a slight modification of [32, Proposition 3.4]. In the same way as in the proof of Theorem 7.1, we are brought back to controlling $\mathbb{E}|\eta^{@}_{j^{@}(x)}(x) - \eta(x)|^p$ for all $x \in \mathcal{A}$. To that end, we split this term as follows

$$\begin{aligned}\mathbb{E}|\eta^{@}_{j^{@}(x)}(x) - \eta(x)|^p &= \mathbb{E}|\eta^{@}_{j^{@}(x)}(x) - \eta(x)|^p(\mathbb{1}_{\{j^{@}(x)\leq j_s\}} + \mathbb{1}_{\{j^{@}(x)>j_s\}}) \\ &= I + II.\end{aligned}$$

Let us first deal with I. Notice that

$$2^{1-p}|\eta^{@}_{j^{@}(x)}(x) - \eta(x)|^p \leq |\eta^{@}_{j^{@}(x)}(x) - \eta^{@}_{j_s}(x)|^p + |\eta^{@}_{j_s}(x) - \eta(x)|^p.$$

The last term is of the good order since

$$\begin{aligned}\mathbb{E}|\eta^{@}_{j_s}(x) - \eta(x)|^p &\leq C(p)\pi_n^p \max\left(2^{-j_s s}, \frac{2^{j_s\frac{d}{2}}}{\sqrt{n}}\right)^p \\ &= \frac{C(p)}{(\kappa\log n)^{\frac{p}{2}}}\left(t(n)2^r n^{-\frac{s}{2s+d}}\right)^p,\end{aligned}$$

according to Proposition 12.1, eq. (10a) and eq. (10c). Regarding the first term, notice that on the event $\{j^{@}(x) \leq j_s\}$, one has got

$$\begin{aligned}|\eta^{@}_{j^{@}(x)}(x) - \eta^{@}_{j_s}(x)| \leq g(j^{@}(x), j_s) &\leq \sup_{j_\varpi \leq k \leq j_s} g(k, j_s) \\ &\leq g(j_s, j_s) = 2t(n)\frac{2^{j_s\frac{d}{2}}}{\sqrt{n}} \leq 2t(n)2^r n^{-\frac{s}{2s+d}},\end{aligned}$$

where we have used eq. (10a) and eq. (10c) and which is of the good order too. Let us now turn to II. For any two $j < k$, we write

$$\mathcal{G}(x, j, k) = \{|\eta^{@}_j(x) - \eta^{@}_k(x)| > g(j, k)\}.$$

Write $\mathcal{J}_n(j) = \{k \in \mathcal{J}_n : k > j\}$. Notice first that we have the following inclusions

$$\begin{aligned}\{j^{@}(x) = j\} &\subseteq \bigcup_{k\in\mathcal{J}_n(j-1)} \mathcal{G}(x, j-1, k), \\ \{j^{@}(x) > j_s\} = \bigcup_{j\in\mathcal{J}_n(j_s)} \{j^{@}(x) = j\} &\subseteq \bigcup_{j\in\mathcal{J}_n(j_s)} \bigcup_{k\in\mathcal{J}_n(j-1)} \mathcal{G}(x, j-1, k).\end{aligned}$$

Therefore, we can write

$$\begin{aligned}II &\leq \sum_{j\in\mathcal{J}_n(j_s)} \mathbb{E}|\eta^{@}_{j^{@}(x)}(x) - \eta(x)|^p \mathbb{1}_{\{j^{@}(x)=j\}} \\ &\leq \sum_{j\in\mathcal{J}_n(j_s)} \sum_{k\in\mathcal{J}_n(j-1)} \mathbb{E}|\eta^{@}_j(x) - \eta(x)|^p \mathbb{1}_{\mathcal{G}(x,j-1,k)}.\end{aligned}$$

Now, we notice that

$$|\eta_j^{@}(x) - \eta_k^{@}(x)| \leq |\eta_j^{@}(x) - \eta(x)| + |\eta(x) - \eta_k^{@}(x)|.$$

So that

$$\begin{aligned}\mathcal{G}(x,j,k) &= \{|\eta_j^{@}(x) - \eta_k^{@}(x)| > g(j,k)\} \\ &\subset \left\{|\eta_j^{@}(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}} t(n)\right\} \bigcup \left\{|\eta_k^{@}(x) - \eta(x)| > \frac{2^{k\frac{d}{2}}}{\sqrt{n}} t(n)\right\}, \\ \mathbb{P}(\mathcal{G}(x,j,k)) &\leq \mathbb{P}\left(|\eta_j^{@}(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}} t(n)\right) + \mathbb{P}\left(|\eta_k^{@}(x) - \eta(x)| > \frac{2^{k\frac{d}{2}}}{\sqrt{n}} t(n)\right).\end{aligned}$$

So that a direct application of the Cauchy-Schwarz inequality leads to

$$\mathbb{E}|\eta_j^{@}(x) - \eta(x)|^p \mathbb{1}_{\mathcal{G}(x,j-1,k)} \leq (\mathbb{E}|\eta_j^{@}(x) - \eta(x)|^{2p})^{\frac{1}{2}} \mathbb{P}(\mathcal{G}(x,j-1,k))^{\frac{1}{2}}.$$

Now, a direct application of Proposition 12.1 for $j_s \leq j \leq J$ gets us

$$(\mathbb{E}|\eta_j^{@}(x) - \eta(x)|^{2p})^{\frac{1}{2}} \leq \sqrt{C(2p)} \pi_n^p \max\left(2^{-js}, \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p \leq \sqrt{C(2p)} (\kappa \log n)^{-\frac{p}{2}}.$$

Besides, notice that for $j_s \leq j < k \leq J$, we can apply Proposition 12.2 with $\kappa \geq \frac{p}{2} C_9^{-1}$ to obtain

$$\mathbb{P}\left(|\eta_j^{@}(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}} t(n)\right) \vee \mathbb{P}\left(|\eta_k^{@}(x) - \eta(x)| > \frac{2^{k\frac{d}{2}}}{\sqrt{n}} t(n)\right) \leq 5R^{2d} n^{-\frac{p}{2}}.$$

To conclude the proof, it remains to notice that $\#\mathcal{J}_n \leq \log n$ and remark that the multiplicative constant in the upper-bound of Theorem 7.2 is indeed smaller than, say, 5 for n large enough. □

12.1.4. A few useful Propositions and Lemmas

Proposition 12.1. *Fix $r \in \mathbb{N}$ and assume we are under* (**CS1**) *and* ($\mathbf{H}_\mathbf{s}^\mathbf{r}$). *Then, For any $x \in \mathcal{A}$ and $j \in \mathcal{J}_n$, one has got*

$$\mathbb{E}[|\eta(x) - \eta_j^{@}(x)|^p] \leq C(p) \pi_n^p \max\left(2^{-js}, \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p,$$

where

$$C(p) = 3^p M^p \max(1, R^{2d} \mu_{\max})^p + C_5(r,d,p,\mu_{\max}; K, \sigma) + 2M^p R^{2d},$$

and C_5 is made explicit in the proof at eq. (21).

Proof. For any $x \in \mathcal{A}$, take $\delta = 3M2^{-js}\max(1, 3\pi_n R^d \mu_{\max})$. Notice first that $\max(1, 3\pi_n R^d \mu_{\max}) \leq \pi_n \max(1, 3R^d \mu_{\max})$ since, by construction, $\pi_n^{-1} \leq 1$ in any case. Now, write

$$\begin{aligned}\mathbb{E}[|\eta(x) - \eta_j^{@}(x)|^p] &= \int_{\mathbb{R}^+} pt^{p-1}\mathbb{P}(|\eta(x) - \eta_j^{@}(x)| \geq t)dt \\ &\leq \delta^p + \int_{\delta}^{+\infty} pt^{p-1}\mathbb{P}(|\eta(x) - \eta_j^{@}(x)| \geq t)dt.\end{aligned}$$

As δ has been fixed, we only need to tackle the rhs above, which we will denote by II. Using Theorem 7.1, we can write

$$\begin{aligned}II &\leq 2R^{2d}\exp\left(-n2^{-jd}\frac{\pi_n^{-2}}{2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}\pi_n^{-1}}\right)\int_0^M pt^{p-1}dt \\ &+ R^d\int_0^{\infty} pt^{p-1}\Lambda\left(\frac{t2^{-j\frac{d}{2}}}{2\pi_n R^d}\right)dt.\end{aligned}$$

Denote by II_1 and II_2 the lhs and rhs terms above, respectively. Now, recall that $j \leq J$, where $2^{Jd} \leq nt(n)^{-2}$ and $t(n)^2 = \kappa\pi_n^2 \log n$. Therefore, as soon as

$$\kappa \geq \frac{p}{2}\left(2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}\pi_n^{-1}\right),$$

we have $II_1 \leq 2M^p R^{2d} n^{-\frac{p}{2}}$. Let us now turn to II_2. Assume first that we are working under the bounded noise assumption, (**N1**). In that case, we have

$$\begin{aligned}II_2 &\leq 2R^d\int_0^{\infty} pt^{p-1}\exp\left(-\frac{n2^{-jd}t^2\pi_n^{-2}}{64K^2R^{2d}\mu_{\max} + 8KR^d\pi_n^{-1}t}\right)dt \\ &\leq C_2(r, d, p, \mu_{\max}, K)\left(\pi_n\frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p.\end{aligned}$$

where the last inequality results from the change of variable $u = \sqrt{n}2^{-j\frac{d}{2}}\pi_n^{-1}t$ together with the fact that $2^{jd} \leq n$ and we have written

$$C_2 := 2R^d\int_0^{\infty} pt^{p-1}\exp\left(-\frac{t^2}{64K^2R^{2d}\mu_{\max} + 8KR^d t}\right)dt.$$

Assume now that we are working under the Gaussian noise assumption (**N2**). In that case, we have

$$\begin{aligned}II_2 \leq R^d\int_0^{\infty} pt^{p-1}&\left(1 \wedge \left\{\frac{2\sigma R^{\frac{d}{2}}(4R^d\mu_{\max} + 2t\pi_n^{-1})^{\frac{1}{2}}}{t\pi_n^{-1}2^{-j\frac{d}{2}}\sqrt{2\pi n}}\right.\right. \\ &\left.\left.\exp\left(-\frac{n2^{-jd}\pi_n^{-2}t^2\sigma^{-2}}{4R^{2d}\mu_{\max} + 2R^d\pi_n^{-1}t}\right)\right\}\right)dt \\ &+ 2R^d\int_0^{\infty} pt^{p-1}\exp\left(-\frac{n2^{-jd}\pi_n^{-2}t^2}{8R^{2d}\mu_{\max} + \frac{8}{3}R^d\pi_n^{-1}t}\right)dt.\end{aligned}$$

Denote by II_3 and II_4 the first and second term, respectively. They can both be handled in the exact same way as II_2, which leads to

$$II_4 \leq C_4(r,d,p,\mu_{\max})\left(\pi_n \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p,$$

where we have written

$$C_4 := 2R^d \int_0^\infty pt^{p-1} \exp\left(-\frac{t^2}{8R^{2d}\mu_{\max} + \frac{8}{3}R^d t}\right) dt,$$

and

$$II_3 \leq C_3(r,d,p,\mu_{\max},\sigma)\left(\pi_n \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p,$$

where we have written

$$C_3 := R^d \int_0^\infty pt^{p-1}\left(1 \wedge \left\{\frac{2\sigma R^{\frac{d}{2}}(4R^d\mu_{\max}+2t)^{\frac{1}{2}}}{t\sqrt{2\pi}} \exp\left(-\frac{t^2\sigma^{-2}}{4R^{2d}\mu_{\max}+2R^d t}\right)\right\}\right)dt.$$

To conclude, let us write

$$C_5(r,d,p,\mu_{\max};K,\sigma) = \begin{cases} C_2(r,d,p,\mu_{\max},K) & \text{under } (\mathbf{N1}) \\ C_3(r,d,p,\mu_{\max},\sigma) + C_4(r,d,p,\mu_{\max}) & \text{under } (\mathbf{N2}) \end{cases} \tag{21}$$

Therefore, we ultimately obtain

$$\mathbb{E}[|\eta(x) - \eta_j^{@}(x)|^p] \leq \left(3^p M^p \max(1, 3R^d\mu_{\max})^p + C_5 + 2M^p R^{2d}\right) \pi_n^p \max\left(2^{-js}, \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p,$$

which concludes the proof. □

Proposition 12.2. *Fix r in $\mathbb{N}$ and assume we are under* (**CS1**) *and* ($\mathbf{H_s^r}$). *This means in particular that $s \in (0,r)$. Let j be such that $j_s \leq j \leq J$. Let $t(n)^2 = \kappa\pi_n^2 \log n$, and define*

$$C_9(r,d,\mu_{\max},\pi_n;K,\sigma) := \begin{cases} C_6(r,d,\mu_{\max},K,\pi_n), & \text{under } (\mathbf{N1}) \\ C_6(r,d,\mu_{\max},\sigma,\pi_n), & \text{under } (\mathbf{N2}) \end{cases},$$

where C_6 is defined in eq. (22) below. Then we have, for n large enough,

$$\mathbb{P}\left(|\eta_j^{@}(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}} t(n)\right) \leq 5R^{2d} n^{-\kappa C_9}.$$

Proof. The proof relies on a direct application of Theorem 7.1. Write $C_0 = 2M\max(1, 3\pi_n R^d \mu_{\max})$ and notice indeed that the theorem applies since for $j \geq j_s$, we get $2^{j\frac{d}{2}} n^{-\frac{1}{2}} \geq 2^{-(r+\frac{d}{2})} 2^{-j_s s}$ (see eq. (10b)) and, as soon as n is large enough, we have $t(n) \geq 2^{r+\frac{d}{2}} C_0$. This leads us to

$$\mathbb{P}\left(|\eta_j^{@}(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}} t(n)\right) \leq 2R^{2d} \exp\left(-n2^{-jd} \frac{\pi_n^{-2}}{2\mu_{\max} R^{4d} + \frac{4}{3} R^{2d} \pi_n^{-1}}\right) + R^d \Lambda\left(\frac{t(n)}{2\pi_n R^d \sqrt{n}}\right).$$

Let us denote the first term by I and the second one by II. I is easily tackled noticing that for $j \leq J$, $n2^{-jd} \geq n2^{-Jd} \geq t(n)^2 = \kappa \pi_n^2 \log n$. So that, we obtain $I \leq 2R^{2d} n^{-\kappa C_6}$, where we have written

$$C_6(r, d, \mu_{\max}, K, \pi_n) := \frac{\min(1, K^{-2})}{64\mu_{\max} R^{2d} + 8R^d \pi_n^{-1}}. \tag{22}$$

Let us now turn to II. Assume first we work under (**N1**). Then we can write

$$II \leq 2R^d \exp\left(-\frac{t(n)^2 \pi_n^{-2}}{64R^{2d} K^2 \mu_{\max} + 8R^d K \pi_n^{-1} \frac{2^{j\frac{d}{2}} t(n)}{\sqrt{n}}}\right).$$

Notice first that $2^{j\frac{d}{2}} t(n) \leq \sqrt{n}$. Therefore, we obtain $II \leq 2R^d n^{-\kappa C_6}$. Assume now that we work under (**N2**). In that case, we obtain

$$II \leq R^d \left(1 \wedge \left\{ \frac{2R^{\frac{d}{2}} \sigma (4R^d \mu_{\max} + 2\pi_n^{-1} \frac{2^{j\frac{d}{2}} t(n)}{\sqrt{n}})^{\frac{1}{2}}}{t(n) \pi_n^{-1} \sqrt{2\pi}} \right.\right.$$
$$\left.\left. \exp\left(-\frac{t(n)^2 \pi_n^{-2} \sigma^{-2}}{4R^{2d} \mu_{\max} + 2R^d \pi_n^{-1} \frac{2^{j\frac{d}{2}} t(n)}{\sqrt{n}}}\right) \right\}\right)$$
$$+ 2R^d \exp\left(-\frac{t(n)^2 \pi_n^{-2}}{8R^{2d} \mu_{\max} + \frac{8}{3} R^d \pi_n^{-1} \frac{2^{j\frac{d}{2}} t(n)}{\sqrt{n}}}\right).$$

We proceed exactly as under (**N1**). So that we obtain $II \leq C_7 n^{-\kappa C_8}$, where

$$C_8(r, d, \mu_{\max}, \sigma, \pi_n) := \frac{\min(1, \sigma^{-2})}{4R^{2d} \mu_{\max} + 2R^d \pi_n^{-1}},$$
$$C_7(r, d, \mu_{\max}, \sigma, \pi_n, t(n)) = R^d \frac{2R^{\frac{d}{2}} \sigma (4R^d \mu_{\max} + 2\pi_n^{-1})^{\frac{1}{2}}}{t(n) \pi_n^{-1} \sqrt{2\pi}} + 2R^d.$$

So that $C_7 \leq 3R^d$ for n large enough. Notice finally that $C_8(r, d, \mu_{\max}, t, \pi_n) \geq C_6(r, d, \mu_{\max}, t, \pi_n)$. This concludes the proof. □

Proposition 12.3. *Fix an integer $r \geq 1$ and assume we are under* (**CS1**). *Let $x \in \mathcal{A}$ and $j \in \mathcal{J}_n$. By construction, there exists $\mathcal{H} \in \mathcal{F}_j$ such that $x \in \mathcal{H}$. Recall besides that $\#\mathcal{S}_j(\mathcal{H}) = R^d$, where $R = 2r-1$ is obviously independent of both x and j. Write $\|.\| = \|.\|_{\ell_2(\mathbb{R}^{R^d})}$ and assume there exists a strictly positive constant $g_{\min}$ independent of x and j such that*

$$\lambda_{\min}(\mathbb{E}Q_{\mathcal{H}}) = \min_{u \in \mathbb{R}^{R^d} : \|u\|=1} \langle u, \mathbb{E}Q_{\mathcal{H}} u \rangle \geq g_{\min}. \tag{23}$$

Then, for any real number t such that $0 < t \leq \frac{g_{\min}}{2}$, we have

$$\mathbb{P}(\lambda_{\min}(Q_{\mathcal{H}}) \leq t) \leq 2R^{2d} \exp\left(-n2^{-jd} \frac{t^2}{2\mu_{\max} R^{4d} + \frac{4}{3} R^{2d} t} \right).$$

Proof. Under the assumption described in eq. (23), we get

$$\begin{aligned}
\lambda_{\min}(Q_{\mathcal{H}}) &\geq \min_{u \in \mathbb{R}^{R^d} : \|u\|=1} \langle u, \mathbb{E}Q_{\mathcal{H}} u \rangle + \min_{u \in \mathbb{R}^{R^d} : \|u\|=1} \langle u, (Q_{\mathcal{H}} - \mathbb{E}Q_{\mathcal{H}}) u \rangle \\
&\geq 2t - \sum_{\nu, \nu' \in \mathcal{S}_j(\mathcal{H})} |[Q_{\mathcal{H}}]_{\nu,\nu'} - [\mathbb{E}Q_{\mathcal{H}}]_{\nu,\nu'}|.
\end{aligned}$$

Write $T_i = \varphi_{j,\nu}(X_i)\varphi_{j,\nu'}(X_i)\mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E}\varphi_{j,\nu}(X)\varphi_{j,\nu'}(X)\mathbb{1}_{\mathcal{H}}(X)$, so that $\mathbb{E}T_i = 0$, $\mathbb{V}\mathrm{ar}T_i \leq \mu_{\max} 2^{jd}$ and $|T_i| \leq 2^{jd+1}$. A direct application of Bernstein inequality for any $\delta > 0$ leads to

$$\begin{aligned}
&\mathbb{P}(|[Q_{\mathcal{H}}]_{\nu,\nu'} - [\mathbb{E}Q_{\mathcal{H}}]_{\nu,\nu'}| \geq \delta) \\
&= \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \varphi_{j,\nu}(X_i)\varphi_{j,\nu'}(X_i)\mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E}\varphi_{j,\nu}(X)\varphi_{j,\nu'}(X)\mathbb{1}_{\mathcal{H}}(X) \right| \geq \delta \right) \\
&\leq 2\exp\left(-\frac{n2^{-jd}\delta^2}{2\mu_{\max} + \frac{4}{3}\delta} \right).
\end{aligned}$$

To conclude, we write

$$\begin{aligned}
\mathbb{P}(\lambda_{\min}(Q_{\mathcal{H}}) \leq t) &\leq \mathbb{P}\left(\sum_{\nu, \nu' \in \mathcal{S}_j(\mathcal{H})} |[Q_{\mathcal{H}}]_{\nu,\nu'} - [\mathbb{E}Q_{\mathcal{H}}]_{\nu,\nu'}| \geq t \right) \\
&\leq 2R^{2d} \exp\left(-n2^{-jd} \frac{t^2}{2\mu_{\max} R^{4d} + \frac{4}{3} R^{2d} t} \right).
\end{aligned}$$

□

Proposition 12.4. *Fix an integer $r \geq 1$ and assume we are under* (**CS1**). *For any $x \in \mathcal{A}$ and $j \in \mathcal{J}_n$, we denote by $\mathcal{H}$ the unique hypercube of $\mathcal{F}_j$ such that $x \in \mathcal{H}$. Then, there exists a strictly positive absolute constant $g_{\min}$ independent of both x and j such that, for all $j \in \mathcal{J}_n$ and all $x \in \mathcal{A}$, we have $\lambda_{\min}(\mathbb{E}Q_{\mathcal{H}}) \geq g_{\min} > 0$.*

Proof. For any $u \in \mathbb{R}^{R^d}$ such that $\|u\|_{\ell_2(\mathbb{R}^{R^d})} = 1$, we can write

$$\langle u, \mathbb{E}Q_{\mathcal{H}} \cdot u \rangle = \int_{\mathcal{A}} \left(\sum_{\nu \in \mathcal{S}_j(\mathcal{H})} u_\nu \varphi_{j,\nu}(w) \mathbb{1}_{\mathcal{H}}(w) \right)^2 \mu(w) dw$$

$$\geq \mu_{\min} \int_{\mathcal{H}} \left(\sum_{\nu \in \mathcal{S}_j(\mathcal{H})} u_\nu \varphi_{j,\nu}(w) \right)^2 dw, \tag{24}$$

$$= \mu_{\min} \int_{[0,1]^d} \left(\sum_{\nu \in \mathfrak{S}} u_\nu \varphi_\nu(w) \right)^2 dw, \tag{25}$$

where $\mathfrak{S}$ has been defined in eq. (15) and the last equality results from the fact that the value of the integral on the rhs of eq. (24) is invariant with $\mathcal{H}$. Let us denote by $\mathbb{S}^{R^d-1}$ the unit-sphere of $\mathbb{R}^{R^d}$. As detailed in [39], the map

$$u \in \mathbb{S}^{R^d-1} \mapsto \int_{[0,1]^d} \left(\sum_{\nu \in \mathfrak{S}} u_\nu \varphi_\nu(w) \right)^2 dw,$$

is absolutely continuous with respect to u on the compact subset $\mathbb{S}^{R^d-1}$ of $\mathbb{R}^{R^d}$. It therefore reaches its minimum at some point $u^* \in \mathbb{S}^{R^d-1}$. It is a direct consequence of the **local linear independence property** of the scaling functions (φ_k) (see Proposition 12.7) that

$$\int_{[0,1]^d} \left(\sum_{\nu \in \mathfrak{S}} u^*_\nu \varphi_\nu(w) \right)^2 dw = c_{\min} > 0,$$

where $c_{\min}$ is a constant that is both independent from x and j. This concludes the proof with $g_{\min} = \mu_{\min} c_{\min}$. □

Proposition 12.5. *Let $(X_i)_{i=1,\dots,n}$ and $(\xi_i)_{i=1,\dots,n}$ be sequences of independent random variables such that $\mathbb{E}(\xi|X) = 0$. Take any $j \geq j_r$. Moreover, assume we are given a function $\mathscr{R}_j(.)$ such that $\|\mathscr{R}_j(.)\|_{\mathbb{L}_\infty(\mathcal{E},\lambda)} \leq M 2^{-js}$, a subset $\mathcal{H}$ of $\mathcal{E}$ and a scaling function $\varphi_{j,k}$. Write*

$$W_{j,k} = \frac{1}{n} \sum_{i=1}^{n} \varphi_{j,k}(X_i) \mathbb{1}_{\mathcal{H}}(X_i)(\mathscr{R}_j(X_i) + \xi_i),$$

and define

$$\Lambda(\delta) = \begin{cases} 2 \exp\left(-\dfrac{n\delta^2}{18K^2\mu_{\max} + 4K 2^{j\frac{d}{2}}\delta} \right), & \text{under } (\mathbf{N1}) \\ 1 \wedge \left\{ \dfrac{2\sigma(\mu_{\max} + 2^{j\frac{d}{2}}\delta)^{\frac{1}{2}}}{\delta\sqrt{2\pi n}} \exp\left(-\dfrac{n\delta^2\sigma^{-2}}{\mu_{\max} + 2^{j\frac{d}{2}}\delta} \right) \right\} & \\ \quad + 2 \exp\left(-\dfrac{n\delta^2}{2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta} \right), & \text{under } (\mathbf{N2}) \end{cases}$$

Then, for all $\delta > 3\mu_{\max} M 2^{-j(s+\frac{d}{2})}$, we have

$$\mathbb{P}(|W_{j,k}| \geq \delta) \leq \Lambda(\delta).$$

Proof. Notice indeed that

$$\begin{aligned} W_{j,k} &\leq \left| \frac{1}{n} \sum_{i=1}^{n} \varphi_{j,k}(X_i)\xi_i \mathbb{1}_{\mathcal{H}}(X_i) \right| \\ &+ \left| \frac{1}{n} \sum_{i=1}^{n} \varphi_{j,k}(X_i)\mathscr{R}_j(X_i)\mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E}\varphi_{j,k}(X)\mathscr{R}_j(X)\mathbb{1}_{\mathcal{H}}(X) \right| \\ &+ |\mathbb{E}\varphi_{j,k}(X)\mathscr{R}_j(X)\mathbb{1}_{\mathcal{H}}(X)| \\ &= I + II + III. \end{aligned}$$

So that we can write

$$\mathbb{P}(|W_{j,k}| \geq \delta) \leq \mathbb{P}(I \geq \delta/3) + \mathbb{P}(II \geq \delta/3) + \mathbb{P}(III \geq \delta/3).$$

Now it is enough to notice that

$$\begin{aligned} III &\leq \int |\varphi_{j,k}(w)\mathscr{R}_j(w)|\mathbb{1}_{\mathcal{H}}(w)\mu(w)dw \\ &\leq \mu_{\max} \int_{\mathcal{E}} |\varphi_{j,k}(w)\mathscr{R}_j(w)|dw \\ &\leq \mu_{\max} \|\varphi_{j,k}\|_{\mathbb{L}_1(\mathcal{E},\lambda)} \|\mathscr{R}_j\|_{\mathbb{L}_\infty(\mathcal{E},\lambda)} \\ &\leq \mu_{\max} M 2^{-j(s+\frac{d}{2})}. \end{aligned}$$

So that $\mathbb{P}(III \geq \delta/3) = 0$ as soon as $\delta > 3\mu_{\max} M 2^{-j(s+\frac{d}{2})}$.

Now, turn to II and write $II = |\sum T_i/n|$ with $T_i = \varphi_{j,k}(X_i)\mathscr{R}_j(X_i)\mathbb{1}_{\mathcal{H}}(X_i)$ $-\mathbb{E}\varphi_{j,k}(X)\mathscr{R}_j(X)\mathbb{1}_{\mathcal{H}}(X)$. Obviously $\mathbb{E}T_i = 0, \mathbb{V}\mathrm{ar}T_i \leq \mathbb{E}(\varphi_{j,\nu}(X)\mathscr{R}_j(X)\mathbb{1}_{\mathcal{H}}(X))^2$ $\leq \mu_{\max} M^2 2^{-2js}$ and $|T_i| \leq M 2^{-js} 2^{j\frac{d}{2}+1}$. So that we can apply Bernstein inequality to get

$$\mathbb{P}(II \geq \delta/3) \leq 2\exp\left(-\frac{n2^{2js}\delta^2}{18\mu_{\max}M^2 + 4M2^{j\frac{d}{2}}2^{js}\delta}\right).$$

And finally, turn to III. Assume first that the noise ξ is bounded by K. We have obviously $\mathbb{E}\varphi_{j,k}(X_i)\xi_i\mathbb{1}_{\mathcal{H}}(X_i) = 0$, $\mathbb{V}\mathrm{ar}(\varphi_{j,k}(X_i)\xi_i\mathbb{1}_{\mathcal{H}}(X_i)) \leq K^2\mu_{\max}$ and $|\varphi_{j,k}(X_i)\xi_i\mathbb{1}_{\mathcal{H}}(X_i)| \leq K2^{j\frac{d}{2}+1}$, so that

$$\mathbb{P}(I \geq \delta/3) \leq 2\exp\left(-\frac{n\delta^2}{18K^2\mu_{\max} + 4K2^{j\frac{d}{2}}\delta}\right).$$

Now, it is enough to notice that for all $s > 0$ and j such that $j \geq \frac{1}{s}\log_2\frac{M}{K}$ (which becomes a constraint for $K < M$ only),

$$\frac{n2^{2js}\delta^2}{18\mu_{\max}M^2 + 4M2^{j\frac{d}{2}}2^{js}\delta} \geq \frac{n\delta^2}{18K^2\mu_{\max} + 4K2^{j\frac{d}{2}}\delta},$$

which concludes the proof under (**N1**). When $j \geq \frac{1}{s}\log_2 3M$, the conclusion under (**N2**) is a direct consequence of Proposition 12.6. □

Proposition 12.6. *Let $\varphi_{j,k}$ be a scaling function and $\mathcal{H}$ a subset of $\mathcal{E}$. Define*

$$I = \frac{1}{n}\sum_{i=1}^{n} \varphi_{j,k}(X_i)\xi_i \mathbb{1}_{\mathcal{H}}(X_i).$$

Assume now that the noise ξ is conditionally Gaussian, that is we are under (**N2**). *Then, we notice that, conditionally on $X_1,\dots,X_n$, $I \sim \Phi(0, \sigma\rho_{j,k}/\sqrt{n})$, where $\rho_{j,k}^2 = n^{-1}\sum_{i=1}^n \varphi_{j,k}(X_i)^2\mathbb{1}_{\mathcal{H}}(X_i)$. Then, for all $\delta > 0$, one can write*

$$\mathbb{P}(|I| \geq \delta) \leq 1 \wedge \left\{ \frac{2\sigma(\mu_{\max} + 2^{j\frac{d}{2}}\delta)^{\frac{1}{2}}}{\delta\sqrt{2\pi n}} \exp\left(-\frac{n\delta^2\sigma^{-2}}{\mu_{\max} + 2^{j\frac{d}{2}}\delta}\right)\right\} + 2\exp\left(-\frac{n\delta^2}{2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta}\right).$$

Proof. For any $\delta > 0$, we write

$$\mathcal{C}_{j,k}(\delta) = \left\{ |\frac{1}{n}\sum_{i=1}^{n} \varphi_{j,k}(X_i)^2\mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E}\varphi_{j,k}(X)^2\mathbb{1}_{\mathcal{H}}(X)| \leq \delta \right\}.$$

Notice first that

$$\mathcal{C}_{j,k}(2^{j\frac{d}{2}}\delta) \subset \{\rho_{j,k}^2 \leq \mu_{\max} + 2^{j\frac{d}{2}}\delta\}.$$

So that

$$\begin{aligned}\mathbb{1}_{\{|I|\geq\delta\}} &\leq \mathbb{1}_{\{|I|\geq\delta\}}\mathbb{1}_{\{\rho_{j,k}^2 \leq \mu_{\max}+2^{j\frac{d}{2}}\delta\}} + \mathbb{1}_{\{|I|\geq\delta\}}\mathbb{1}_{\mathcal{C}_{j,k}^c(2^{j\frac{d}{2}}\delta)} \\ &\leq \mathbb{1}_{\{|I|\geq\delta\}}\mathbb{1}_{\{\rho_{j,k}^2 \leq \mu_{\max}+2^{j\frac{d}{2}}\delta\}} + \mathbb{1}_{\mathcal{C}_{j,k}^c(2^{j\frac{d}{2}}\delta)}.\end{aligned}$$

The first term is handled thanks to a regular Gaussian tail inequality. Notice indeed that

$$\begin{aligned}&\mathbb{P}(|I| \geq \delta | X_1,\dots,X_n)\mathbb{1}_{\{\rho_{j,k}^2 \leq \mu_{\max}+2^{j\frac{d}{2}}\delta\}} \\ &\quad\leq 1 \wedge \left\{ \frac{2\rho_{j,k}\sigma}{\delta\sqrt{2\pi n}} \exp\left(-\frac{n\delta^2}{\rho_{j,k}^2\sigma^2}\right)\right\}\mathbb{1}_{\{\rho_{j,k}^2 \leq \mu_{\max}+2^{j\frac{d}{2}}\delta\}} \\ &\quad\leq 1 \wedge \left\{ \frac{2\sigma(\mu_{\max} + 2^{j\frac{d}{2}}\delta)^{\frac{1}{2}}}{\delta\sqrt{2\pi n}} \exp\left(-\frac{n\delta^2\sigma^{-2}}{\mu_{\max} + 2^{j\frac{d}{2}}\delta}\right)\right\}.\end{aligned}$$

In addition, notice that $\mathbb{E}\varphi_{j,k}(X)^4\mathbb{1}_{\mathcal{H}}(X) \leq \mu_{\max}2^{jd}$ and $|\varphi_{j,k}(X)^2\mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E}\varphi_{j,k}(X)^2\mathbb{1}_{\mathcal{H}}(X)| \leq 2^{jd+1}$, so that a direct application of Bernstein inequality leads to

$$\mathbb{P}(\mathcal{C}_{j,k}(2^{j\frac{d}{2}}\delta)^c) \leq 2\exp\left(-\frac{n2^{jd}\delta^2}{2^{jd}(2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta)}\right) = 2\exp\left(-\frac{n\delta^2}{2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta}\right),$$

which concludes the proof. □

Proposition 12.7. *Let $\mathfrak{m}$ be a constant such that $\mathfrak{m} > 0$ and fix $z \in \mathbb{R}^d$ such that $z \in \mathcal{B}_\infty(2^{-1}, \mathfrak{m})$. Write $\mathfrak{S} := \{k \in \mathbb{Z}^d : 2^{-1} \in Supp\varphi_k\}$, the set of indexes corresponding to the scaling functions whose support $Supp\varphi_k$ contains the point $2^{-1} \in \mathbb{R}^d$. The scaling functions (φ_k) verify the* ***local linear independence property*** *in the sense that $\sum_{k\in\mathfrak{S}} \alpha_k\varphi_k = 0$ on the domain $\mathcal{B}_\infty(z, \mathfrak{m})$ if and only if $\alpha_k = 0$ for all $k \in \mathfrak{S}$.*

Proof. This result is derived from [33] and its proof can be found in [39]. □

12.2. Proof of the upper-bound results under (CS2)

Recall that under (**CS2**), we work with a sample of size $2n$ split into two subsamples denoted by $\mathcal{D}_n$ and $\mathcal{D}'_n$. As detailed previously, similar results as the ones described in Section 7, Section 8 and Section 12.1.4 are still valid with $\eta^{\maltese}$ under (**CS2**). They in fact all stem from Theorem 9.1. The proofs remain for the most part unchanged, with $\mathcal{J}_n$ redefined as $\mathcal{J}_n = \{j_s, j_s+1, \dots, J-1, J\}$ where $2^{j_s} = \lfloor n^{\frac{1}{2s+d}} \rfloor$, $\eta^{\maltese}$ in place of $\eta^{@}$, $\widetilde{X}_i$ in place of X_i (where we have written $\widetilde{u} = u - X'_{i_x} + 2^{-j-1}$), and $\mathcal{H}_0$ in place of $\mathcal{H}$. The sole differences appear in the proofs of Theorem 9.1 and Proposition 12.4. Let us start with the proof of Theorem 9.1.

Proof of Theorem 9.1. Assume we are under (**CS2**) and want to control the probability of deviation of $\eta_j^{\maltese}(x)$ from $\eta(x)$ at a point $x \in \mathcal{A}$, for some $j \in \mathcal{J}_n$. Recall that $\mathcal{H}_0(x)$ stands for the cell $\mathcal{H}_0 = 2^{-j}[0,1]^d$ centered in x at level j, that is $\mathcal{H}_0(x) = x - 2^{-j-1} + 2^{-j}[0,1]^d$ and denote by $\mathcal{O}_x$ the event

$$\mathcal{O}_x = \{\#\{i : X'_i \in \mathcal{H}_0(x)\} \geq 1\}.$$

We can write

$$\begin{aligned}\mathbb{P}(|\eta(x) - \eta_j^{\maltese}(x)| \geq \delta) &= \mathbb{P}(|\eta(x) - \eta_j^{\maltese}(x)| \geq \delta, \mathcal{O}_x) \\ &\quad + \mathbb{P}(|\eta(x) - \eta_j^{\maltese}(x)| \geq \delta, \mathcal{O}_x^c).\end{aligned}$$

Focus first on what happens on the event $\mathcal{O}_x^c$. The last term can be controlled easily since the probability that no single design point X'_i of $\mathcal{D}'_n$ belongs to $\mathcal{H}_0(x)$ decreases exponentially fast with n. Notice indeed that, under (**CS2**),

$$\begin{aligned}\mathbb{P}(\mathcal{O}_x^c) &= (\mathbb{P}(X'_1 \notin \mathcal{H}_0(x)))^n \\ &= (1 - \mathbb{P}(X'_1 \in \mathcal{H}_0(x)))^n \\ &= \left(1 - \int_{\mathcal{A}\cap\mathcal{H}_0(x)} \mu(w)dw\right)^n \\ &\leq \left(1 - \mu_{\min} 2^{-jd} \lambda\left(2^j(\mathcal{A} - x) \cap [-2^{-1}, 2^{-1}]^d\right)\right)^n \\ &\leq (1 - \mu_{\min} 2^{-jd} \min(2\mathfrak{m}_0, 2^{-1})^d)^n \\ &\leq \exp(-\mu_{\min} \min(2\mathfrak{m}_0, 2^{-1})^d n 2^{-jd}),\end{aligned}$$

where the before last inequality is a direct consequence of (**S2**) and the last one comes from the fact that for any $x \in [0,1)$, $\ln(1-x) \leq -x$. Now, recall that $\eta_j^{\maltese}(x) = 0$ on $\mathcal{O}_x^c$ and $|\eta(x)| \leq M$ since $\eta \in \mathscr{L}^s(\mathbb{R}^d, M)$. So that we obtain

$$\mathbb{P}(|\eta(x) - \eta_j^{\maltese}(x)| \geq \delta, \mathcal{O}_x^c) \leq \exp(-\mu_{\min} \min(2\mathfrak{m}_0, 2^{-1})^d n 2^{-jd}) \mathbb{1}_{\{\delta \leq M\}},$$

which is smaller than the first term in the upper-bound of Theorem 9.1. Now focus on what happens on the event $\mathcal{O}_x$. We can write

$$\begin{aligned}\mathbb{P}(|\eta(x) - \eta_j^{\maltese}(x)| \geq \delta, \mathcal{O}_x) &= \mathbb{P}(\mathcal{O}_x)\mathbb{E}[\mathbb{P}(|\eta(x) - \eta_j^{\maltese}(x)| \geq \delta | X'_{i_x}) | \mathcal{O}_x] \\ &\leq \mathbb{E}[\mathbb{P}(|\eta(x) - \eta_j^{\maltese}(x)| \geq \delta | X'_{i_x}) | \mathcal{O}_x].\end{aligned}$$

Therefore, it is enough to control the probability of deviation of $\eta_j^{\maltese}(x)$ from $\eta(x)$ on $\mathcal{O}_x$, conditionally on X'_{i_x}. It is controlled in exactly the same way as the probability of deviation of $\eta_j^{@}(x)$ from $\eta(x)$ under (**CS1**), except that we now work with conditional probabilities and expectations with respect to X'_{i_x}. Interestingly, the random variable X'_{i_x} is independent of the points of $\mathcal{D}_n$ since it is built upon the design points (X'_i) of $\mathcal{D}'_n$ which are themselves independent of the points of $\mathcal{D}_n$. This is a key feature that makes theoretical computations tractable under (**CS2**) and allows to handle $\eta^{\maltese}$ in a similar way as $\eta^{@}$ under (**CS1**). As announced above, Proposition 12.4 is the sole result that is not obviously true under (**CS2**). However it can be extended to setting (**CS2**) without much trouble (see below). Ultimately, this proves that, on the event $\mathcal{O}_x$ and conditionally on X'_{i_x}, the probability of deviation of $\eta_j^{\maltese}(x)$ from $\eta(x)$ verifies Theorem 7.1. So that finally, it remains to put everything together to obtain the results announced in Theorem 9.1, which concludes the proof. □

As detailed in [39], the proof of Proposition 12.4 can be extended to setting (**CS2**), thanks to the **local linear independence property** of the scaling functions (see Proposition 12.7) and a compactness argument. In particular, we obtain the following result, which is proved in [39].

Lemma 12.1. *Let $r \in \mathbb{N}$. Let φ be the Daubechies' scaling function of regularity r and $\mathfrak{S} = \{\nu \in \mathbb{Z}^d : 2^{-1} \in Supp\varphi_\nu\}$. Then, there exists a strictly positive absolute constant $c_{\min}$ such that*

$$\inf_{u \in \mathbb{S}^{R^d - 1}} \inf_{\mathfrak{m} \geq \mathfrak{m}_0} \inf_{z \in \mathcal{B}_\infty(2^{-1}, \mathfrak{m})} \int_{\mathcal{B}_\infty(z, \mathfrak{m}) \cap [0,1]^d} \left(\sum_{\nu \in \mathfrak{S}} u_\nu \varphi_\nu(w) \right)^2 dw \geq c_{\min}, \tag{26}$$

$$\inf_{\mathfrak{m} \geq \mathfrak{m}_0} \inf_{z \in \mathcal{B}_\infty(2^{-1}, \mathfrak{m})} \int_{\mathcal{B}_\infty(z, \mathfrak{m})} \sum_{\nu \in \mathfrak{S}} \varphi_\nu(w)^2 dw \geq c_{\min}. \tag{27}$$

Appendix

Generalized Lipschitz spaces

Here, we sum up relevant facts about Lipschitz and Besov spaces on $\mathbb{R}^d$ as stated in [7, Chap. 3] for any $d \in \mathbb{N}$ and [11, Chap. 2, §9] for $d = 1$. Let us denote by

$\mathscr{C}(\mathbb{R}^d)$ and $\widetilde{\mathscr{C}}(\mathbb{R}^d)$ the spaces of continuous and absolutely continuous functions on $\mathbb{R}^d$, respectively. Let us denote by $\|.\|$ the Euclidean norm of $\mathbb{R}^d$, f a function defined on $\mathbb{R}^d$ and write $\Delta_h^1(f,x) = |f(x+h) - f(x)|$ for any $x \in \mathbb{R}^d$. For any $r \in \mathbb{N}$ and all $x \in \mathbb{R}^d$, we further define the r^{th}-finite difference by induction as follows,

$$\Delta_h^r(f,x) = \Delta_h^1(\Delta_h^{r-1}(f,x)),$$

and the r^{th}-modulus of smoothness of $f \in \mathscr{C}(\mathbb{R}^d)$ as follows

$$\omega_r(f,t)_\infty = \sup_{0 \le \|h\| \le t} \|\Delta_h^r(f,.)\|_{\mathbb{L}_\infty(\mathbb{R}^d,\lambda)}.$$

Write $s > 0$ and $r = \lfloor s \rfloor + 1$. The Besov space $B^s_{\infty,\infty}$ on $\mathbb{R}^d$, also known as the generalized Lipschitz space $\mathscr{L}^s(\mathbb{R}^d)$, is the collection of all functions $f \in \widetilde{\mathscr{C}}(\mathbb{R}^d) \cap \mathbb{L}_\infty(\mathbb{R}^d,\lambda)$ such that the semi-norm

$$|f|_{\mathscr{L}^s(\mathbb{R}^d)} := \sup_{t>0} \left(t^{-s}\omega_r(f,t)_\infty\right),$$

is finite. The norm for $\mathscr{L}^s(\mathbb{R}^d)$ is subsequently defined as

$$\|f\|_{\mathscr{L}^s(\mathbb{R}^d)} := \|f\|_{\mathbb{L}_\infty(\mathbb{R}^d,\lambda)} + |f|_{\mathscr{L}^s(\mathbb{R}^d)}.$$

Fix a real number $M > 0$. Throughout the paper, $\mathscr{L}^s(\mathbb{R}^d, M)$ refers to the ball of $\mathscr{L}^s(\mathbb{R}^d)$ of radius M. Obviously, the elements of $\mathscr{L}^s(\mathbb{R}^d, M)$ are λ-a.e. uniformly bounded by M on $\mathbb{R}^d$.

As described in [11, 7], there exists an alternative definition of Lipschitz spaces $\mathscr{C}^s(\mathbb{R}^d)$, also known as Hölder spaces, which goes as follows. For any integer d, multi-index $q = (q_1, \ldots, q_d) \in \mathbb{N}^d$ and $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we define the differential operator ∂^q as usual by $\partial^q := \frac{\partial^{q_1+\cdots+q_d}}{\partial^{q_1}x_1 \ldots \partial^{q_d}x_d}$. For any positive integer s, $\mathscr{C}^s(\mathbb{R}^d)$ consists of the functions f on $\mathbb{R}^d$ such that $\partial^q f$ is bounded and absolutely continuous on $\mathbb{R}^d$, for all $q \in \mathbb{N}^d$ such that $|q|_1 := q_1 + \cdots + q_d \le s$. This definition is extended to non-integer s as follows,

$$\mathscr{C}^s(\mathbb{R}^d) := \{f \in \widetilde{\mathscr{C}}(\mathbb{R}^d) \cap \mathbb{L}_\infty(\mathbb{R}^d,\lambda) : \sup_{x \in \mathbb{R}^d} \Delta_h^1(f,x) \le C|h|^s\}, \quad 0 < s < 1,$$

$$\mathscr{C}^s(\mathbb{R}^d) := \{f \in \widetilde{\mathscr{C}}(\mathbb{R}^d) \cap \mathbb{L}_\infty(\mathbb{R}^d,\lambda) : \partial^q f \in \mathscr{C}^{s-m}(\mathbb{R}^d), |q|_1 = m\}, \quad m < s < m+1,\ m \in \mathbb{N}.$$

It can be shown that, for all non-integer $s > 0$, $\mathscr{C}^s(\mathbb{R}^d) = \mathscr{L}^s(\mathbb{R}^d)$, while $\mathscr{C}^s(\mathbb{R}^d)$ is a strict subset of $\mathscr{L}^s(\mathbb{R}^d)$ when $s \in \mathbb{N}$ (see [11, p. 52] for examples of functions that belong to $\mathscr{L}^1([0,1])$ but not to $\mathscr{C}^1([0,1])$ in the particular case where $d = 1$).

Furthermore, we define these function spaces on the subset $\mathcal{E}$ of $\mathbb{R}^d$ as the restriction of their elements to $\mathcal{E}$. As explained in [7, Remark 3.2.4], function spaces on $\mathcal{E}$ can be defined by restriction or, alternatively, in an intrinsic way, and both definitions coincide for fairly general domains $\mathcal{E}$ of $\mathbb{R}^d$.

Looking at function spaces on $\mathcal{E}$ as function spaces on $\mathbb{R}^d$ restricted to $\mathcal{E}$ justifies the use of MRAs of $\mathbb{L}_2(\mathbb{R}^d,\lambda)$ in our local analysis.

MRAs and smoothness analysis

Multivariate MRAs will always be assumed to be obtained from a tensorial product of one-dimensional MRAs, as described in [7, §1.4, eq. (1.4.10)]. We will denote by $\varphi_{j,k}(.) = 2^{jd/2}\varphi(2^j. - k)$ the translated and dilated version of φ with $k \in \mathbb{Z}^d$. As usual, we write $\mathcal{V}_j$ to mean Closure(Span$\{\varphi_{j,k}, k \in \mathbb{Z}^d\}$), so that Closure$(\cup_{j\geq 0}\mathcal{V}_j) = \mathbb{L}_2(\mathbb{R}^d, \lambda)$ (where the closures are taken with respect to the $\mathbb{L}_2(\mathbb{R}^d, \lambda)$-metric).

The r-MRAs defined in Section 5.2 are intimately connected with generalized Lipschitz spaces. Assume we are given a r-MRA with $r \in \mathbb{N}$ and $\eta \in \mathscr{L}^s(\mathbb{R}^d, M)$, where $s \in (0, r)$ and $M > 0$. Denote by $\mathscr{P}_j\eta$ the orthogonal projection of η onto $\mathcal{V}_j$ and by $\mathscr{R}_j\eta = \eta - \mathscr{P}_j\eta$ the corresponding remainder. Then, we have for all $x \in \mathbb{R}^d$, $\eta(x) = \mathscr{P}_j\eta(x) + \mathscr{R}_j\eta(x)$ where $\|\mathscr{R}_j\eta\|_{\mathbb{L}_\infty(\mathbb{R}^d,\lambda)} \leq M2^{-js}$, as detailed in [7, Corollary 3.3.1]. It is noteworthy that the above approximation results remain valid in the particular case where we work on the subset $\mathcal{E}$ of $\mathbb{R}^d$ and consider η to be the restriction to $\mathcal{E}$ of an element of $\mathscr{L}^s(\mathbb{R}^d)$.

Acknowledgement

The author would like to thank Dominique Picard for many fruitful discussions and suggestions. He would also like to acknowledge interesting conversations with Gérard Kerkyacharian. Finally, he would like to thank two anonymous referees and an associate editor whose constructive comments led to a full refactoring of the paper and largely contributed to improve it.

References

[1] Antoniadis, A., Grégoire, G. and Vial, P. (1997). Random design wavelet curve smoothing. *Statistics & Probability Letters* **35** 225-232. MR1484959

[2] Antoniadis, A. and Pham, D. T. (1998). Wavelet regression for random or irregular design. *Comput. Stat. Data An.* **28** 353-369. MR1659207

[3] Audibert, J.-Y. (2004). Classification under polynomial entropy and margin assumptions and randomized estimators. *Preprint, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6 and 7.* MR2096215

[4] Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Stat.* **35** 608-633. MR2336861

[5] Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.* **6** 127-146. MR1918295

[6] Cai, T. T. and Brown, L. D. (1998). Wavelet shrinkage for nonequispaced samples. *Ann. Stat.* **26** 1783-1799. MR1673278

[7] Cohen, A. (2003). *Numerical analysis of wavelet methods. Studies in mathematics and its applications* **32**. North-Holland. MR1990555

[8] Daubechies, I. (1992). *Ten lectures on wavelets. CBMS-NSF regional conference series in applied mathematics.* Society for Industrial and Applied Mathematics. MR1162107

[9] Delouille, V., Franke, J. and von Sachs, R. (2001). Nonparametric stochastic regression with design-adapted wavelets. *The Indian Journal of Statistics* **63** 328-366. MR1897046
[10] Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215-228.
[11] DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation. Grundlehren Der Mathematischen Wissenschaften*. Springer-Verlag. MR1261635
[12] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer-Verlag. MR1383093
[13] Donoho, D. L. (1995). De-Noising by Soft-Thresholding. *IEEE Trans. Inf. Theory* **41(3)** 613-627. MR1331258
[14] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425-455. MR1311089
[15] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet Shrinkage: Asymptotia? *J. Roy. Stat. Soc. B* **57(2)** 301-369. MR1323344
[16] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Stat.* **24(2)** 508-539. MR1394974
[17] Gaïffas, S. (2005). Convergence rates for pointwise curve estimation with a degenerate design. *Math. Methods Statist.* **1** 1-27. MR2158069
[18] Gaïffas, S. (2007). On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Statist.* **11** 344-364. MR2339297
[19] Gaïffas, S. (2007). Sharp estimation in sup norm with random design. *Statistics & Probability Letters* **77** 782-794. MR2369683
[20] Greblicki, W. and Pawlak, M. (1982). A classification procedure using the multiple Fourier series. *Inf. Sci.* **26** 115-126. MR0658750
[21] Greblicki, W. and Pawlak, M. (1985). Fourier and Hermite series estimates of regression functions. *Ann. Inst. Statist. Math.* **37** 443-454. MR0818041
[22] Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2001). *A distribution-free theory of nonparametric regression. Springer Series in Statistics*. Springer.
[23] Hall, P. and Turlach, B. A. (1997). Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *Ann. Stat.* **25** 1912-1925. MR1474074
[24] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. B. (1997). *Wavelets, approximation and statistical applications*. Springer Verlag, Berlin. MR1618204
[25] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, Berlin. MR1851606
[26] Horn, R. A. and Johnson, C. R. (1990). *Matrix analysis*. Cambridge University Press. MR1084815

[27] Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statistics & Probability Letters* **13** 15-24.
[28] Kerkyacharian, G. and Picard, D. (2004). Regression in random design and warped wavelets. *Bernoulli* **10** 1053-1105. MR2108043
[29] Kohler, M. (2003). Nonlinear orthogonal series estimates for random design regression. *J. Stat. Plan. Infer.* **115** 491-520. MR1985881
[30] Kohler, M. (2008). Multivariate orthogonal series estimates for random design regression. *J. Stat. Plan. Infer.* **138** 3217-3237. MR2442230
[31] Kovac, A. and Silverman, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statistical Assoc.* **95** 172-183.
[32] Lepski, O. V., Mammen, E. and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Stat.* **25** 392-947. MR1447734
[33] Malgouyres, G. and Lemarié-Rieusset, P.-G. (1991). On the support of the scaling function in a multi-resolution analysis. *Comptes rendus de l'Académie des sciences* **313** 377-380. MR1126417
[34] Mallat, S. (1989). Multiresolution approximations and wavelet orthonormal bases of L^2. *Trans. Amer. Math. Soc.* **315** 69-87. MR1008470
[35] Mallat, S. (2008). *A wavelet tour of signal processing: the sparse way.* Academic Press. MR2479996
[36] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Stat.* **27** 1808-1829. MR1765618
[37] Marron, J. S. (1983). Optimal rates of convergence to the Bayes risk in nonparametric discrimination. *Ann. Stat.* **11** 1142-1155. MR0720260
[38] Meyer, Y. (1992). *Wavelets and operators. Cambridge Studies in Advanced Mathematics* **37**. Cambridge University Press. MR1228209
[39] Monnier, J.-B. (2011). Classification via local multi-resolution projections Technical Report, LPMA, Université Paris Diderot - Paris 7.
[40] Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141-142.
[41] Neumann, M. H. and Spokoiny, V. G. (1995). On the efficiency of wavelet estimators under arbitrary error distributions. *Math. Methods Statist.* **4** 137-166. MR1335152
[42] Pensky, M. and Vidakovic, B. (2001). On non-equally spaced wavelet regression. *Ann. Inst. Statist. Math.* **53** 681-690. MR1879604
[43] Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Stat.* **28** 298-335. MR1762913
[44] Sardy, S., Percival, D. B., Bruce, A. G., Gao, H.-Y. and Stuetzle, W. (1999). Wavelet shrinkage for unequally spaced data. *Satistics and Computing* **9** 65-75.
[45] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Stat.* **8** 1348-1360. MR0594650
[46] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* **10** 1040-1053. MR0673642

[47] Sweldens, W. (1996). The lifting scheme: a custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal.* **3** 186-200. MR1385051

[48] Vapnik, V. N. (1998). *Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control*. John Wiley & Sons. MR1641250

[49] Watson, G. S. (1964). Smooth regression analysis. *The Indian Journal of Statistics* **26** 359-372. MR0185765

[50] Yang, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. Inf. Theory* **45** 2271-2284. MR1725115

[51] Zhang, S., Wong, M.-Y. and Zheng, Z. (2002). Wavelet threshold estimation of a regression function with random design. *J. Multivariate Anal.* **80** 256-284. MR1889776