Constr Approx (2010) 31: 195–229
DOI 10.1007/s00365-009-9062-2

CONSTRUCTIVE APPROXIMATION

Maxisets for Model Selection

F. Autin · E. Le Pennec · J.M. Loubes · V. Rivoirard

Received: 27 February 2008 / Revised: 22 December 2008 / Accepted: 6 February 2009 /
Published online: 7 July 2009

Abstract We address the statistical issue of determining the maximal spaces (maxisets) where model selection procedures attain a given rate of convergence. By considering first general dictionaries, then orthonormal bases, we characterize these maxisets in terms of approximation spaces. These results are illustrated by classical choices of wavelet model collections. For each of them, the maxisets are described in terms of functional spaces. We give special attention to the issue of calculability and measure the induced loss of performance in terms of maxisets.

Communicated by Gerard Kerkyacharian.

F. Autin
Centre de Mathématiques et d'Informatique, 39, rue F. Joliot Curie, 13453 Marseille Cedex 13, France
e-mail: autin@cmi.univ-mrs.fr

E. Le Pennec (✉)
Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599, Université Paris Diderot, 175 rue du Chevaleret, 75013 Paris, France
e-mail: lepennec@math.jussieu.fr

J.M. Loubes
Institut de Mathématiques de Toulouse, Equipe de Probabilités et de Statistique, Université de Toulouse Paul Sabatier, 118 Route de Narbonne, 31000 Toulouse, France
e-mail: Jean-Michel.Loubes@math.ups-tlse.fr

V. Rivoirard
Laboratoire de Mathématiques, UMR 8628, Université Paris-Sud., Bât 425, 91405, Orsay cedex, France
e-mail: Vincent.Rivoirard@math.u-psud.fr

V. Rivoirard
Département de Mathématiques et Applications, UMR 8553, Ecole Normale Supérieure, 45, rue d'Ulm, 75230, Paris Cedex 05, France

Keywords Approximation spaces · Approximation theory · Besov spaces · Estimation · Maxiset · Model selection · Rates of convergence

Mathematics Subject Classification (2000) 62G05 · 62G20 · 41A25 · 42C40

1 Introduction

The topic of this paper lies on the frontier between statistics and approximation theory. Our goal is to characterize the functions well estimated by a special class of estimation procedures: the model selection rules. Our purpose is not to build new model selection estimators but to determine thoroughly the functions for which well-known model selection procedures achieve good performances. Of course, approximation theory plays a crucial role in our setting, but surprisingly, its role is even more important than that of statistical tools. This statement will be emphasized by the use of the *maxiset approach*, which illustrates the well-known fact that "well estimating is well approximating."

More precisely, we consider the classical Gaussian white noise model

$$dY_{n,t} = s(t)\,dt + \frac{1}{\sqrt{n}}\,dW_t, \quad t \in \mathcal{D},$$

where $\mathcal{D} \subset \mathbb{R}$, s is the unknown function, W is the Brownian motion in $\mathbb{R}$, and $n \in \mathbb{N}^* = \{1, 2, \dots\}$. This model means that for any $u \in \mathbb{L}_2(\mathcal{D})$,

$$Y_n(u) = \int_{\mathcal{D}} u(t)\,dY_{n,t} = \int_{\mathcal{D}} u(t)s(t)\,dt + \frac{1}{\sqrt{n}} W_u$$

is observable where $W_u = \int_{\mathcal{D}} u(t)\,dW_t$ is a centered Gaussian process such that for all functions u and u',

$$\mathbb{E}[W_u W_{u'}] = \int_{\mathcal{D}} u(t)u'(t)\,dt.$$

We take a noise level of the form $1/\sqrt{n}$ to refer to the asymptotic equivalence between the Gaussian white noise model and the classical regression model with n equispaced observations (see [24]).

Two questions naturally arise: how to construct an estimator $\hat{s}$ of s based on the observation $dY_{n,t}$, and how to measure its performance. Many estimators have been proposed in this setting, including wavelet thresholding, kernel rules, and Bayesian procedures. In this paper, we only focus on model selection techniques described precisely in the next paragraph.

1.1 Model Selection Procedures

The model selection methodology consists in constructing an estimator by minimizing an empirical contrast γ_n over a given set, called a model. The pioneer work in model selection goes back to the 1970's with Mallows [18] and Akaike [1]. Birgé

and Massart develop the whole modern theory of model selection in [9–11] or [7], for instance. Estimation of a regression function with model selection estimators is considered by Baraud in [5, 6], while inverse problems are tackled by Loubes and Ludeña [16, 17]. Finally model selection techniques nowadays provide valuable tools in statistical learning (see Boucheron et al. [12]).

In nonparametric estimation, performances of estimators are usually measured by using the quadratic norm, which gives rise to the following empirical quadratic contrast:

$$\gamma_n(u) = -2Y_n(u) + \|u\|^2$$

for any function u, where $\|\cdot\|$ denotes the norm associated to $\mathbb{L}_2(\mathcal{D})$. We assume that we are given a dictionary of functions of $\mathbb{L}_2(\mathcal{D})$, denoted by $\Phi = (\varphi_i)_{i\in\mathcal{I}}$, where $\mathcal{I}$ is a countable set, and we consider $\mathcal{M}_n$, a collection of models spanned by some functions of Φ. For any $m \in \mathcal{M}_n$, we denote by $\mathcal{I}_m$ the subset of $\mathcal{I}$ such that

$$m = \operatorname{span}\{\varphi_i : i \in \mathcal{I}_m\}$$

and $D_m \le |\mathcal{I}_m|$ the dimension of m. Let $\hat{s}_m$ be the function that minimizes the quadratic empirical criterion $\gamma_n(u)$ with respect to $u \in m$. A straightforward computation shows that the estimator $\hat{s}_m$ is the projection of the data onto the space m. So, if $\{e_1^m, \dots, e_{D_m}^m\}$ is an orthonormal basis (not necessarily related to Φ) of m, and

$$\hat{\beta}_i^m = Y_n\big(e_i^m\big) = \int_{\mathcal{D}} e_i^m(t)\, dY_{n,t},$$

then

$$\hat{s}_m = \sum_{i\in\mathcal{I}_m} \hat{\beta}_i^m e_i^m, \quad \text{and} \quad \gamma_n(\hat{s}_m) = -\sum_{i\in\mathcal{I}_m} \big(\hat{\beta}_i^m\big)^2.$$

Now the issue is the selection of the best model $\hat{m}$ from the data which gives rise to the *model selection estimator* $\hat{s}_{\hat{m}}$. For this purpose, a penalized rule is considered, which aims at selecting an estimator close enough to the data but still lying in a small space to avoid overfitting issues. Let $\operatorname{pen}_n(m)$ be a penalty function which increases when D_m increases. The model $\hat{m}$ is selected using the following penalized criterion:

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}_n} \big\{\gamma_n(\hat{s}_m) + \operatorname{pen}_n(m)\big\}. \tag{1.1}$$

The choice of the model collection and the associated penalty are then the key issues handled by model selection theory. We point out that the choices of both the model collection and the penalty function should depend on the noise level. This is emphasized by the subscript n for $\mathcal{M}_n$ and $\operatorname{pen}_n(m)$.

The asymptotic behavior of model selection estimators has been studied by many authors. We refer to Massart [19] for general references and recall hereafter the main oracle-type inequality. Such an oracle inequality provides a non-asymptotic control on the estimation error with respect to a bias term $\|s - s_m\|$, where s_m stands for the

best approximation (in the $\mathbb{L}_2$ sense) of the function s by a function of m. In other words, s_m is the orthogonal projection of s onto m, defined by

$$s_m = \sum_{i \in \mathcal{I}_m} \beta_i^m e_i^m, \quad \beta_i^m = \int_{\mathcal{D}} e_i^m(t) s(t)\, dt.$$

Theorem 1 (Theorem 4.2 of [19]) *Let $n \in \mathbb{N}^\star$ be fixed, and let $(x_m)_{m \in \mathcal{M}_n}$ be some family of positive numbers such that*

$$\sum_{m \in \mathcal{M}_n} \exp(-x_m) = \Sigma_n < \infty. \tag{1.2}$$

Let $\kappa > 1$, and assume that

$$\mathrm{pen}_n(m) \geq \frac{\kappa}{n}\left(\sqrt{D_m} + \sqrt{2x_m}\right)^2. \tag{1.3}$$

Then, almost surely, there exists some minimizer $\hat{m}$ of the penalized least-squares criterion

$$\gamma_n(\hat{s}_m) + \mathrm{pen}_n(m)$$

over $m \in \mathcal{M}_n$. Moreover, the corresponding penalized least-squares estimator $\hat{s}_{\hat{m}}$ is unique, and the following inequality is valid:

$$\mathbb{E}\left[\|\hat{s}_{\hat{m}} - s\|^2\right] \leq C\left[\inf_{m \in \mathcal{M}_n}\left\{\|s_m - s\|^2 + \mathrm{pen}_n(m)\right\} + \frac{1 + \Sigma_n}{n}\right], \tag{1.4}$$

where C depends only on κ.

Equation (1.4) is the key result to establish optimality of penalized estimators under oracle or minimax points of view. In this paper, we focus on an alternative to these approaches: the maxiset point of view.

1.2 The Maxiset Point of View

Before describing the maxiset approach, let us briefly recall that for a given procedure $s^* = (s_n^*)_n$, the minimax study of s^* consists in comparing the rate of convergence of s^* achieved on a given functional space $\mathcal{F}$ with the best possible rate achieved by any estimator. More precisely, let $\mathcal{F}(R)$ be the ball of radius R associated with $\mathcal{F}$; then the procedure $s^* = (s_n^*)_n$ achieves the rate $\rho^* = (\rho_n^*)_n$ on $\mathcal{F}(R)$ if

$$\sup_n \left\{(\rho_n^*)^{-2} \sup_{s \in \mathcal{F}(R)} \mathbb{E}\left[\|s_n^* - s\|^2\right]\right\} < \infty.$$

To check that a procedure is optimal from the minimax point of view (said to be minimax), it must be proved that its rate of convergence achieves the best rate among any procedure on each ball of the class. This minimax approach is extensively used, and many methods cited above are proved to be minimax in different statistical frameworks.

However, the choice of the function class is subjective and, in the minimax framework, statisticians have no idea whether there are other functions well-estimated at the rate ρ^* by their procedure. A different point of view is to consider the procedure s^* as given and search all the functions s that are well estimated at a given rate ρ^*: this is the *maxiset* approach, which has been proposed by Kerkyacharian and Picard [15]. The maximal space, or maxiset, of the procedure s^* for this rate ρ^* is defined as the set of all these functions. Obviously, the larger the maxiset, the better the procedure. We set the following definition:

Definition 1 Let $\rho^* = (\rho_n^*)_n$ be a decreasing sequence of positive real numbers, and let $s^* = (s_n^*)_n$ be an estimation procedure. The maxiset of s^* associated with the rate ρ^* is

$$MS(s^*, \rho^*) = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \sup_n \left\{ (\rho_n^*)^{-2} \mathbb{E}\left[\|s_n^* - s\|^2 \right] \right\} < \infty \right\},$$

and the ball of radius $R > 0$ of the maxiset is defined by

$$MS(s^*, \rho^*)(R) = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \sup_n \left\{ (\rho_n^*)^{-2} \mathbb{E}\left[\|s_n^* - s\|^2 \right] \right\} \leq R^2 \right\}.$$

Of course, there exist connections between maxiset and minimax points of view: s^* achieves the rate ρ^* on $\mathcal{F}$ if and only if

$$\mathcal{F} \subset MS(s^*, \rho^*).$$

In the white noise setting, the maxiset theory has been investigated for a wide range of estimation procedures, including kernel, thresholding and Lepski procedures, and Bayesian or linear rules. We refer to [3, 4, 8, 13, 15, 21], and [22] for general results. Maxisets have also been investigated for other statistical models, see [2] and [23].

1.3 Overview of the Paper

The goal of this paper is to investigate maxisets of model selection procedures. Following the classical model selection literature, we only use penalties proportional to the dimension D_m of m:

$$\mathrm{pen}_n(m) = \frac{\lambda_n}{n} D_m, \tag{1.5}$$

with λ_n to be specified. Our main result characterizes these maxisets in terms of approximation spaces. More precisely, we establish an equivalence between the statistical performance of $\hat{s}_{\hat{m}}$ and the approximation properties of the model collections $\mathcal{M}_n$. With

$$\rho_{n,\alpha} = \left(\frac{\lambda_n}{n} \right)^{\frac{\alpha}{1+2\alpha}} \tag{1.6}$$

for any $\alpha > 0$, Theorem 2, combined with Theorem 1, proves that, for a given function s, the quadratic risk $\mathbb{E}[\|s - \hat{s}_{\hat{m}}\|^2]$ decays at the rate $\rho_{n,\alpha}^2$ if and only if the

deterministic quantity

$$Q(s,n) = \inf_{m\in\mathcal{M}_n}\left\{\|s_m - s\|^2 + \frac{\lambda_n}{n}D_m\right\} \tag{1.7}$$

decays at the rate $\rho^2_{n,\alpha}$ as well. This result holds with mild assumptions on λ_n and under an embedding assumption on the model collections ($\mathcal{M}_n \subset \mathcal{M}_{n+1}$). Once we impose additional structure on the model collections, the deterministic condition can be rephrased as a linear approximation property and a non-linear one as stated in Theorem 3. We illustrate these results for three different model collections based on wavelet bases. The first one deals with sieves in which all the models are embedded, the second one with the collection of all subspaces spanned by vectors of a given basis. For these examples, we handle the issue of calculability and give explicit characterizations of the maxisets. In the third example, we provide an intermediate choice of model collections and use the fact that the embedding condition on the model collections can be relaxed. Finally, performances of these estimators are compared and discussed.

The paper is organized as follows. Section 2 describes the main general results established in this paper. More precisely, we specify results valid for general dictionaries in Sect. 2.1. In Sect. 2.2, we focus on the case where Φ is an orthonormal family. Section 3 is devoted to the illustrations of these results for some model selection estimators associated with wavelet methods. In particular, a comparison of maxiset performances are provided and discussed. Section 4 gives the proofs of our results.

2 Main Results

As explained in the introduction, our goal is to investigate maxisets associated with model selection estimators $\hat{s}_{\hat{m}}$ where the penalty function is defined in (1.5), and with the rate $\rho_\alpha = (\rho_{n,\alpha})_n$ where $\rho_{n,\alpha}$ is specified in (1.6). Observe that $\rho_{n,\alpha}$ depends on the choice of λ_n. It can, for instance, be polynomial, or can take the classical form

$$\rho_{n,\alpha} = \left(\frac{\log n}{n}\right)^{\frac{\alpha}{1+2\alpha}}.$$

So we wish to determine

$$MS(\hat{s}_{\hat{m}},\rho_\alpha) = \left\{s\in\mathbb{L}_2(\mathcal{D}) : \sup_n\left\{\rho^{-2}_{n,\alpha}\mathbb{E}\left[\|\hat{s}_{\hat{m}} - s\|^2\right]\right\} < \infty\right\}.$$

In the sequel, we use the following notation: if $\mathcal{F}$ is a given space,

$$MS(\hat{s}_{\hat{m}},\rho_\alpha) :=: \mathcal{F}$$

means that for any $R > 0$, there exists $R' > 0$ such that

$$MS(\hat{s}_{\hat{m}},\rho_\alpha)(R) \subset \mathcal{F}(R') \tag{2.1}$$

and for any $R' > 0$, there exists $R > 0$ such that

$$\mathcal{F}(R') \subset MS(\hat{s}_{\hat{m}}, \rho_\alpha)(R). \tag{2.2}$$

2.1 The Case of General Dictionaries

In this section, we make no assumption on Φ. Theorem 1 is a non-asymptotic result, while maxiset results deal with rates of convergence (with asymptotics in n). Therefore, obtaining maxiset results for model selection estimators requires a structure on the sequence of model collections. We first focus on the case of nested model collections ($\mathcal{M}_n \subset \mathcal{M}_{n+1}$). Note that this does not imply a strong structure on the model collection for a given n. In particular, this does not imply that the models are nested. Identifying the maxiset $MS(\hat{s}_{\hat{m}}, \rho_\alpha)$ is a two-step procedure. We need to establish inclusion (2.1) and inclusion (2.2). Recall that we have previously introduced:

$$Q(s, n) = \inf_{m \in \mathcal{M}_n} \left\{ \|s_m - s\|^2 + \frac{\lambda_n}{n} D_m \right\}.$$

Roughly speaking, Theorem 1 established by Massart proves that any function s satisfying

$$\sup_n \left\{ \rho_{n,\alpha}^{-2} Q(s, n) \right\} \leq (R')^2$$

belongs to the maxiset $MS(\hat{s}_{\hat{m}}, \rho_\alpha)$ and thus provides inclusion (2.2). The following theorem establishes inclusion (2.1) and highlights the fact that $Q(s, n)$ plays a capital role:

Theorem 2 *Let $0 < \alpha_0 < \infty$ be fixed. Let us assume that for any n, the sequence of model collections satisfies*

$$\mathcal{M}_n \subset \mathcal{M}_{n+1}, \tag{2.3}$$

and that the sequence of positive numbers $(\lambda_n)_n$ is non-decreasing and satisfies

$$\lim_{n \to +\infty} n^{-1} \lambda_n = 0, \tag{2.4}$$

and there exist $n_0 \in \mathbb{N}^*$ and two constants $0 < \delta \leq \frac{1}{2}$ and $0 < p < 1$ such that for $n \geq n_0$,

$$\lambda_{2n} \leq 2(1 - \delta)\lambda_n, \tag{2.5}$$

$$\sum_{m \in \mathcal{M}_n} e^{-\frac{(\sqrt{\lambda_n} - 1)^2 D_m}{2}} \leq \sqrt{1 - p}, \tag{2.6}$$

and

$$\lambda_{n_0} \geq \Upsilon(\delta, p, \alpha_0), \tag{2.7}$$

where $\Upsilon(\delta, p, \alpha_0)$ *is a positive constant depending only on* α_0, p *and* δ *defined in* (4.3) *of Sect.* 4. *Then, the penalized rule* $\hat{s}_{\hat{m}}$ *is such that for any* $\alpha \in (0, \alpha_0]$, *for any* $R > 0$, *there exists* $R' > 0$ *such that for* $s \in \mathbb{L}_2(\mathcal{D})$,

$$\sup_n \left\{ \rho_{n,\alpha}^{-2} \mathbb{E}\left[\|\hat{s}_{\hat{m}} - s\|^2 \right] \right\} \leq R^2 \Rightarrow \sup_n \left\{ \rho_{n,\alpha}^{-2} Q(s,n) \right\} \leq (R')^2.$$

Technical Assumptions (2.4), (2.5), (2.6), and (2.7) are very mild and could be partly relaxed while preserving the results. Assumption (2.4) is necessary to deal with rates converging to 0. Note that the classical cases $\lambda_n = \lambda_0$ or $\lambda_n = \lambda_0 \log(n)$ satisfy (2.4) and (2.5). Furthermore, Assumption (2.7) is always satisfied when $\lambda_n = \lambda_0 \log(n)$ or when $\lambda_n = \lambda_0$ with λ_0 large enough. Assumption (2.6) is very close to Assumptions (1.2)–(1.3). In particular, if there exist two constants $\kappa > 1$ and $0 < p < 1$ such that for any n,

$$\sum_{m \in \mathcal{M}_n} e^{-\frac{(\sqrt{\kappa^{-1}\lambda_n}-1)^2 D_m}{2}} \leq \sqrt{1-p}, \tag{2.8}$$

then, since

$$\mathrm{pen}_n(m) = \frac{\lambda_n}{n} D_m,$$

Conditions (1.2), (1.3) and (2.6) are all satisfied. The assumption $\alpha \in (0, \alpha_0]$ can be relaxed for particular model collections, which will be highlighted in Proposition 2 of Sect. 3.1. Finally, Assumption (2.3) can be removed for some special choice of model collection $\mathcal{M}_n$ at the price of a slight over-penalization, as shall be shown in Proposition 1 and Sect. 3.3.

Combining Theorems 1 and 2 gives a first characterization of the maxiset of the model selection procedure $\hat{s}_{\hat{m}}$:

Corollary 1 *Let* $\alpha_0 < \infty$ *be fixed. Assume that Assumptions* (2.3), (2.4), (2.5), (2.7), *and* (2.8) *are satisfied. Then for any* $\alpha \in (0, \alpha_0]$,

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) := \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \sup_n \left\{ \rho_{n,\alpha}^{-2} Q(s,n) \right\} < \infty \right\}.$$

The maxiset of $\hat{s}_{\hat{m}}$ is characterized by a deterministic approximation property of s with respect to the models $\mathcal{M}_n$. It can be related to some classical approximation properties of s in terms of approximation rates if the functions of Φ are orthonormal.

2.2 The Case of Orthonormal Bases

From now on, $\Phi = \{\varphi_i\}_{i \in \mathcal{I}}$ is assumed to be an orthonormal basis (for the $\mathbb{L}_2$ scalar product). We also assume that the model collections $\mathcal{M}_n$ are constructed through restrictions of a single model collection $\mathcal{M}$. Namely, given a collection of models $\mathcal{M}$ we introduce a sequence $\mathcal{J}_n$ of increasing subsets of the indices set $\mathcal{I}$, and we define the intermediate collection $\mathcal{M}'_n$ as

$$\mathcal{M}'_n = \left\{ m' = \mathrm{span}\{\varphi_i : i \in \mathcal{I}_m \cap \mathcal{J}_n\} : m \in \mathcal{M} \right\}. \tag{2.9}$$

The model collections $\mathcal{M}'_n$ do not necessarily satisfy the embedding condition (2.3). Thus, we define

$$\mathcal{M}_n = \bigcup_{k \le n} \mathcal{M}'_k$$

so $\mathcal{M}_n \subset \mathcal{M}_{n+1}$. The assumptions on Φ and on the model collections allow to give an explicit characterization of the maxisets. We set $\widetilde{\mathcal{M}} = \bigcup_n \mathcal{M}_n = \bigcup_n \mathcal{M}'_n$. Note that without any further assumption, $\widetilde{\mathcal{M}}$ can be a larger model collection than $\mathcal{M}$. Now let us denote by $V = (V_n)_n$ the sequence of approximation spaces defined by

$$V_n = \operatorname{span}\{\varphi_i : i \in \mathcal{J}_n\}$$

and consider the corresponding approximation space

$$\mathcal{L}^\alpha_V = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \sup_n \left\{ \rho_{n,\alpha}^{-1} \| P_{V_n} s - s \| \right\} < \infty \right\},$$

where $P_{V_n} s$ is the projection of s onto V_n. Define also another kind of approximation set:

$$\mathcal{A}^\alpha_{\widetilde{\mathcal{M}}} = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \sup_{M>0} \left\{ M^\alpha \inf_{\{m \in \widetilde{\mathcal{M}} :\, D_m \le M\}} \| s_m - s \| \right\} < \infty \right\}.$$

The corresponding balls of radius $R > 0$ are defined, as usual, by replacing ∞ by R in the previous definitions. We have the following result:

Theorem 3 *Let $\alpha_0 < \infty$ be fixed. Assume that* (2.4), (2.5), (2.7), *and* (2.8) *are satisfied. Then, the penalized rule $\hat{s}_{\hat{m}}$ satisfies the following result: for any $\alpha \in (0, \alpha_0]$,*

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) :=: \mathcal{A}^\alpha_{\widetilde{\mathcal{M}}} \cap \mathcal{L}^\alpha_V.$$

The result pointed out in Theorem 3 links the performance of the estimator to an approximation property for the estimated function. This approximation property is decomposed into a linear approximation measured by $\mathcal{L}^\alpha_V$ and a non-linear approximation measured by $\mathcal{A}^\alpha_{\widetilde{\mathcal{M}}}$. The linear condition is due to the use of the reduced model collection $\mathcal{M}_n$ instead of $\mathcal{M}$, which is often necessary to ensure either the calculability of the estimator or Condition (2.8). It plays the role of a minimum regularity property that is easily satisfied.

Observe that if we have one model collection, that is, for any k and k', $\mathcal{M}_k = \mathcal{M}_{k'} = \mathcal{M}$, $\mathcal{J}_n = \mathcal{I}$ for any n and thus $\widetilde{\mathcal{M}} = \mathcal{M}$. Then

$$\mathcal{L}^\alpha_V = \operatorname{span}\{\varphi_i : i \in \mathcal{I}\},$$

and Theorem 3 gives

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) :=: \mathcal{A}^\alpha_{\mathcal{M}}.$$

The spaces $\mathcal{A}^\alpha_{\widetilde{\mathcal{M}}}$ and $\mathcal{L}^\alpha_V$ depend highly on the models and the approximation space. At first glance, the best choice seems to be $V_n = \mathbb{L}_2(\mathcal{D})$ and

$$\mathcal{M} = \{m : \mathcal{I}_m \subset \mathcal{I}\},$$

since the infimum in the definition of $\mathcal{A}^{\alpha}_{\widetilde{\mathcal{M}}}$ becomes smaller when the collection is enriched. There is, however, a price to pay when enlarging the model collection: the penalty has to be larger to satisfy (2.8), which deteriorates the convergence rate. A second issue comes from the tractability of the minimization (1.1) itself which will further limit the size of the model collection.

To avoid considering the union of $\mathcal{M}'_k$, which can dramatically increase the number of models considered for a fixed n, leading to large penalties, we can relax the assumption that the penalty is proportional to the dimension. Namely, for any n and for any $m \in \mathcal{M}'_n$, there exists $\tilde{m} \in \mathcal{M}$ such that

$$m = \operatorname{span}\{\varphi_i : i \in \mathcal{I}_{\tilde{m}} \cap \mathcal{J}_n\}.$$

Then for any model $m \in \mathcal{M}'_n$, we replace the dimension D_m by the larger dimension $D_{\tilde{m}}$, and we set

$$\widetilde{\operatorname{pen}}_n(m) = \frac{\lambda_n}{n} D_{\tilde{m}}.$$

The minimization of the corresponding penalized criterion over all model in $\mathcal{M}'_n$ leads to a result similar to Theorem 3. Mimicking its proof, we can state the following proposition that will be used in Sect. 3.3:

Proposition 1 *Let $\alpha_0 < \infty$ be fixed. Assume* (2.4), (2.5), (2.7), *and* (2.8) *are satisfied. Then the penalized estimator $\hat{s}_{\tilde{m}}$, where*

$$\tilde{m} = \operatorname*{argmin}_{m \in \mathcal{M}'_n} \big\{ \gamma_n(\hat{s}_m) + \widetilde{\operatorname{pen}}_n(m) \big\},$$

satisfies the following result: for any $\alpha \in (0, \alpha_0]$,

$$MS(\tilde{s}_{\tilde{m}}, \rho_\alpha) :=: \mathcal{A}^{\alpha}_{\mathcal{M}} \cap \mathcal{L}^{\alpha}_{V}.$$

Note that $\mathcal{M}_n$, $\mathcal{L}^{\alpha}_{V}$, and $\mathcal{A}^{\alpha}_{\widetilde{\mathcal{M}}}$ can be defined in a similar fashion for any arbitrary dictionary Φ. However, one can only obtain the inclusion $MS(\hat{s}_{\hat{m}}, \rho_\alpha) \subset \mathcal{A}^{\alpha}_{\widetilde{\mathcal{M}}} \cap \mathcal{L}^{\alpha}_{V}$ in the general case.

3 Comparisons of Model Selection Estimators

The aim of this section is twofold. First, we propose to illustrate our previous maxiset results to different model selection estimators built with wavelet methods by identifying precisely the spaces $\mathcal{A}^{\alpha}_{\widetilde{\mathcal{M}}}$ and $\mathcal{L}^{\alpha}_{V}$. Second, comparisons between the performances of these estimators are provided and discussed.

We briefly recall the construction of periodic wavelet bases of the interval $[0, 1]$. Let ϕ and ψ be two compactly supported functions of $\mathbb{L}_2(\mathbb{R})$, and denote for all $j \in \mathbb{N}$, all $k \in \mathbb{Z}$, and all $x \in \mathbb{R}$, $\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. Those functions can be periodized in such a way that

$$\Psi = \big\{ \phi_{00}, \psi_{jk} : j \geq 0,\ k \in \big\{0, \ldots, 2^j - 1\big\} \big\}$$

constitutes an orthonormal basis of $\mathbb{L}_2([0,1])$. Some popular examples of such bases are given in [14]. The function ϕ is called the scaling function and ψ the corresponding wavelet. Any periodic function $s \in \mathbb{L}_2([0,1])$ can be represented as:

$$s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk},$$

where

$$\alpha_{00} = \int_{[0,1]} s(t)\phi_{00}(t)\,dt,$$

and for any $j \in \mathbb{N}$ and for any $k \in \{0,\dots,2^j-1\}$,

$$\beta_{jk} = \int_{[0,1]} s(t)\psi_{jk}(t)\,dt.$$

Finally, we recall the characterization of Besov spaces using wavelets. Such spaces will play an important role in the following. In this section we assume that the multiresolution analysis associated with the basis Ψ is r-regular with $r \geq 1$ as defined in [20]. In this case, for any $0 < \alpha < r$ and any $1 \leq p, q \leq \infty$, the periodic function s belongs to the Besov space $\mathcal{B}^{\alpha}_{p,q}$ if and only if $|\alpha_{00}| < \infty$ and

$$\sum_{j=0}^{\infty} 2^{jq(\alpha+\frac{1}{2}-\frac{1}{p})}\|\beta_{j.}\|^q_{\ell_p} < \infty \quad \text{if } q < \infty,$$

$$\sup_{j\in\mathbb{N}} 2^{j(\alpha+\frac{1}{2}-\frac{1}{p})}\|\beta_{j.}\|_{\ell_p} < \infty \quad \text{if } q = \infty,$$

where $(\beta_{j.}) = (\beta_{jk})_k$. This characterization allows us to recall the following embeddings:

$$\mathcal{B}^{\alpha}_{p,q} \subsetneq \mathcal{B}^{\alpha'}_{p',q'} \quad \text{as soon as } \alpha - \frac{1}{p} \geq \alpha' - \frac{1}{p'},\ p < p' \text{ and } q \leq q'$$

and

$$\mathcal{B}^{\alpha}_{p,\infty} \subsetneq \mathcal{B}^{\alpha}_{2,\infty} \quad \text{as soon as } p > 2.$$

3.1 Collection of Sieves

We first consider a single model collection corresponding to a class of nested models

$$\mathcal{M}^{(s)} = \left\{m = \operatorname{span}\left\{\phi_{00}, \psi_{jk} : j < N_m, 0 \leq k < 2^j\right\} : N_m \in \mathbb{N}\right\}.$$

For such a model collection, Theorem 3 could be applied with $V_n = \mathbb{L}_2$. One can even remove Assumption (2.7), which imposes a minimum value on λ_{n_0} that depends on the rate ρ_α:

Proposition 2 *Let $0 < \alpha < r$, and let $\hat{s}^{(s)}_{\hat{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(s)}$. Then, under Assumptions (2.4), (2.5), and (2.8),*

$$MS\big(\hat{s}^{(s)}_{\hat{m}}, \rho_\alpha\big) =: \mathcal{B}^\alpha_{2,\infty}.$$

Note that it suffices to choose $\lambda_n \geq \lambda_0$ with λ_0, independent of α, large enough to ensure Condition (2.8).

It is important to notice that the estimator $\hat{s}^{(s)}_{\hat{m}}$ cannot be computed in practice because to determine the best model $\hat{m}$ one needs to consider an infinite number of models, which cannot be done without computing an infinite number of wavelet coefficients. To overcome this issue, we specify a maximum resolution level $j_0(n)$ for estimation where $n \mapsto j_0(n)$ is non-decreasing. This modification is also in the scope of Theorem 3: it corresponds to

$$V_n = \operatorname{span}\big\{\phi_{00}, \psi_{jk} : 0 \leq j < j_0(n),\ 0 \leq k < 2^j\big\}$$

and the model collection $\mathcal{M}^{(s)}_n$ defined as follows:

$$\mathcal{M}^{(s)}_n = \mathcal{M}'^{(s)}_n = \big\{m \in \mathcal{M}^{(s)} :\ N_m < j_0(n)\big\}.$$

For the specific choice

$$2^{j_0(n)} \leq n\lambda_n^{-1} < 2^{j_0(n)+1}, \tag{3.1}$$

we obtain:

$$\begin{aligned}
\mathcal{L}^\alpha_V &= \left\{ s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 :\ \sup_{n\in\mathbb{N}^*} 2^{\frac{2j_0(n)\alpha}{1+2\alpha}} \|s - P_{V_n} s\|^2 < \infty \right\} \\
&= \left\{ s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 :\ \sup_{n\in\mathbb{N}^*} 2^{\frac{2j_0(n)\alpha}{1+2\alpha}} \sum_{j\geq j_0(n)}\sum_{k} \beta^2_{jk} < \infty \right\} \\
&= \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty}.
\end{aligned}$$

Since $\mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{B}^\alpha_{2,\infty}$ reduces to $\mathcal{B}^\alpha_{2,\infty}$, arguments of the proofs of Theorem 3 and Proposition 2 give:

Proposition 3 *Let $0 < \alpha < r$, and let $\hat{s}^{(st)}_{\hat{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(s)}_n$. Then, under Assumptions (2.4), (2.5), and (2.8),*

$$MS\big(\hat{s}^{(st)}_{\hat{m}}, \rho_\alpha\big) =: \mathcal{B}^\alpha_{2,\infty}.$$

This tractable procedure is thus as efficient as the original one. We obtain the maxiset behavior of the non-adaptive linear wavelet procedure pointed out in [21], but here the procedure is completely data-driven.

3.2 The Largest Model Collections

In this paragraph we enlarge the model collections in order to obtain much larger maxisets. We start with the following model collection:

$$\mathcal{M}^{(l)} = \big\{m = \operatorname{span}\big\{\phi_{00}, \psi_{jk} : (j,k) \in \mathcal{I}_m\big\} \colon \mathcal{I}_m \in \mathcal{P}(\mathcal{I})\big\},$$

where

$$\mathcal{I} = \bigcup_{j\geq 0}\big\{(j,k) : k \in \big\{0,1,\ldots,2^j-1\big\}\big\}$$

and $\mathcal{P}(\mathcal{I})$ is the set of all subsets of $\mathcal{I}$. This model collection is so rich that whatever the sequence $(\lambda_n)_n$, Condition (2.8) (or even Condition (1.2)) is not satisfied. To reduce the cardinality of the collection, we restrict the maximum resolution level to the resolution level $j_0(n)$ defined in (3.1) and consider the collections $\mathcal{M}_n^{(l)}$ defined from $\mathcal{M}^{(l)}$ by

$$\mathcal{M}_n^{(l)} = \mathcal{M}_n'^{(l)} = \big\{m \in \mathcal{M}^{(l)} : \mathcal{I}_m \in \mathcal{P}\big(\mathcal{I}^{j_0}\big)\big\},$$

where

$$\mathcal{I}^{j_0} = \bigcup_{0\leq j<j_0(n)} \big\{(j,k) : k \in \big\{0,1,\ldots,2^j-1\big\}\big\}.$$

Note that this corresponds to the same choice of V_n as in the previous paragraph and that the corresponding estimator fits perfectly within the framework of Theorem 3.

The classical logarithmic penalty

$$\operatorname{pen}_n(m) = \frac{\lambda_0 \log(n) D_m}{n},$$

which corresponds to $\lambda_n = \lambda_0 \log(n)$, is sufficient to ensure Condition (2.8) as soon as λ_0 is a constant large enough (the choice $\lambda_n = \lambda_0$ is not sufficient). The identification of the corresponding maxiset focuses on the characterization of the space $\mathcal{A}^\alpha_{\mathcal{M}^{(l)}}$ since, as previously, $\mathcal{L}^\alpha_V = \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty}$. We rely on sparsity properties of $\mathcal{A}^\alpha_{\mathcal{M}^{(l)}}$. In our context, sparsity means that there is a *small* proportion of *large* coefficients of a signal. We introduce, for $n \in \mathbb{N}^*$, the notation

$$|\beta|_{(n)} = \inf\big\{u : \operatorname{card}\big\{(j,k) \in \mathbb{N} \times \big\{0,1,\ldots,2^j-1\big\} : |\beta_{jk}| > u\big\} < n\big\}$$

to represent the non-increasing rearrangement of the wavelet coefficient of a periodic signal s:

$$|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(n)} \geq \cdots.$$

As the best model $m \in \mathcal{M}^{(l)}$ of prescribed dimension M is obtained by choosing the subset of index corresponding to the M largest wavelet coefficients, a simple identification of the space $\mathcal{A}^\alpha_{\mathcal{M}^{(l)}}$ is

$$\mathcal{A}^\alpha_{\mathcal{M}^{(l)}} = \left\{s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \sup_{M\in\mathbb{N}^*} M^{2\alpha} \sum_{i=M+1}^{\infty} |\beta|^2_{(i)} < \infty\right\}.$$

Theorem 2.1 of [15] provides a characterization of this space as a weak Besov space:

$$\mathcal{A}^{\alpha}_{\mathcal{M}^{(l)}} = \mathcal{W}_{\frac{2}{1+2\alpha}}$$

with, for any $q \in]0,2[$,

$$\mathcal{W}_q = \left\{ s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \sup_{n\in\mathbb{N}^*} n^{1/q}|\beta|_{(n)} < \infty \right\}.$$

Following their definitions, the larger α, the smaller $q = 2/(1+2\alpha)$ and the sparser the sequence $(\beta_{jk})_{j,k}$. Lemma 2.2 of [15] shows that the spaces $\mathcal{W}_q$ $(0 < q < 2)$ have other characterizations in terms of wavelet coefficients:

$$\begin{aligned}\mathcal{W}_q &= \left\{ s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \sup_{u>0} u^{q-2} \sum_j \sum_k \beta_{jk}^2 1_{|\beta_{jk}|\le u} < \infty \right\} \\ &= \left\{ s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \sup_{u>0} u^{q} \sum_j \sum_k 1_{|\beta_{jk}|> u} < \infty \right\}.\end{aligned}$$

We thus obtain the following proposition:

Proposition 4 *Let $\alpha_0 < r$ be fixed, let $0 < \alpha \le \alpha_0$, and let $\hat{s}^{(l)}_{\hat{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(s)}_n$. Then under Assumptions (2.4), (2.5), (2.7), and (2.8):*

$$MS\big(\hat{s}^{(l)}_{\hat{m}}, \rho_\alpha\big) =: \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{W}_{\frac{2}{1+2\alpha}}.$$

Observe that the estimator $\hat{s}^{(l)}_{\hat{m}}$ is easily tractable from a computational point of view, as the minimization can be rewritten coefficientwise:

$$\begin{aligned}\hat{m}(n) &= \operatorname*{argmin}_{m\in\mathcal{M}^{(l)}_n}\left\{ \gamma_n(\hat{s}_m) + \frac{\lambda_n}{n} D_m \right\} \\ &= \operatorname*{argmin}_{m\in\mathcal{M}^{(l)}_n}\left\{ \sum_{j=0}^{j_0(n)-1}\sum_{k=0}^{2^j-1} \left(\hat{\beta}^2_{jk}\mathbf{1}_{(j,k)\notin\mathcal{I}_m} + \frac{\lambda_n}{n}\mathbf{1}_{(j,k)\in\mathcal{I}_m} \right) \right\}.\end{aligned}$$

The best subset $\mathcal{I}_{\hat{m}}$ is thus the set $\{(j,k) \in \mathcal{I}^{j_0} : |\hat{\beta}_{jk}| > \sqrt{\lambda_n/n}\}$, and $\hat{s}^{(l)}_{\hat{m}}$ corresponds to the well-known hard thresholding estimator

$$\hat{s}^{(l)}_{\hat{m}} = \hat{\alpha}_{00}\phi_{00} + \sum_{j=0}^{j_0(n)-1}\sum_{k=0}^{2^j-1} \hat{\beta}_{jk}\mathbf{1}_{|\hat{\beta}_{jk}|>\sqrt{\frac{\lambda_n}{n}}}\psi_{jk}.$$

Proposition 4 thus corresponds to the maxiset result established by Kerkyacharian and Picard [15].

3.3 A Special Strategy for Besov Spaces

We now consider the model collection proposed by Massart [19]. This collection can be viewed as a hybrid collection between the collections of Sects. 3.1 and 3.2. This strategy turns out to be minimax for all Besov spaces $\mathcal{B}^{\alpha}_{p,\infty}$ when $\alpha > \max(1/p - 1/2, 0)$ and $1 \le p \le \infty$.

More precisely, for a chosen $\theta > 2$, define the model collection by

$$\mathcal{M}^{(h)} = \big\{m = \operatorname{span}\{\phi_{00}, \psi_{jk} : (j,k) \in \mathcal{I}_m\} : J \in \mathbb{N},\ \mathcal{I}_m \in \mathcal{P}_J(\mathcal{I})\big\},$$

where for any $J \in \mathbb{N}$, $\mathcal{P}_J(\mathcal{I})$ is the set of all subsets $\mathcal{I}_m$ of $\mathcal{I}$ that can be written

$$\mathcal{I}_m = \big\{(j,k) : 0 \le j < J, 0 \le k < 2^j\big\} \cup \bigcup_{j \ge J} \big\{(j,k) : k \in A_j, |A_j| = \big\lfloor 2^J (j - J + 1)^{-\theta} \big\rfloor\big\}$$

with $\lfloor x \rfloor := \max\{n \in \mathbb{N} :\ n \le x\}$.

As remarked in [19], for any $J \in \mathbb{N}$ and any $\mathcal{I}_m \in \mathcal{P}_J(\mathcal{I})$, the dimension D_m of the corresponding model m depends only on J and is such that

$$2^J \le D_m \le 2^J \left(1 + \sum_{n \ge 1} n^{-\theta}\right).$$

We denote by D_J this common dimension. Note that the model collection $\mathcal{M}^{(h)}$ does not vary with n. Using Theorem 3 with $V_n = \mathbb{L}_2$, we have the following proposition:

Proposition 5 *Let $\alpha_0 < r$ be fixed, let $0 < \alpha \le \alpha_0$, and let $\hat{s}^{(h)}_{\hat{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(h)}$. Then under Assumptions (2.4), (2.5), (2.7), and (2.8):*

$$MS\big(\hat{s}^{(h)}_{\hat{m}}, \rho_\alpha\big) =: \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}},$$

with

$$\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} = \Bigg\{ s = \alpha_{00}\phi_{00} + \sum_{j \ge 0} \sum_{k=0}^{2^j - 1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \sup_{J \ge 0} 2^{2J\alpha} \sum_{j \ge J} \sum_{k \ge \lfloor 2^J (j-J+1)^{-\theta} \rfloor} |\beta_j|^2_{(k)} < \infty \Bigg\},$$

where $(|\beta_j|_{(k)})_k$ is the reordered sequence of coefficients $(\beta_{jk})_k$:

$$|\beta_j|_{(1)} \ge |\beta_j|_{(2)} \cdots |\beta_j|_{(k)} \ge \cdots \ge |\beta_j|_{(2^j)}.$$

Note that, as in Sect. 3.1, as soon as $\lambda_n \ge \lambda_0$ with λ_0 large enough, Condition (2.8) holds.

This large set cannot be characterized in terms of classical spaces. Nevertheless it is undoubtedly a large functional space, since as proved in Sect. 4.4, for every $\alpha > 0$ and every $p \geq 1$ satisfying $p > 2/(2\alpha + 1)$, we get

$$\mathcal{B}^{\alpha}_{p,\infty} \subsetneq \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}. \tag{3.2}$$

This new procedure is not computable, since one needs an infinite number of wavelet coefficients to perform it. The problem of calculability can be solved by introducing, as previously, a maximum scale $j_0(n)$ as defined in (3.1). We consider the class of collection models $(\mathcal{M}^{(h)}_n)_n$ defined as follows:

$$\mathcal{M}^{(h)}_n = \big\{m = \operatorname{span}\big\{\phi_{00}, \psi_{jk} : (j,k) \in \mathcal{I}_m,\, j < j_0(n)\big\} \colon J \in \mathbb{N},\, \mathcal{I}_m \in \mathcal{P}_J(\mathcal{I})\big\}.$$

This model collection does not satisfy the embedding condition $\mathcal{M}^{(h)}_n \subset \mathcal{M}^{(h)}_{n+1}$. Nevertheless, we can use Proposition 1 with

$$\widetilde{\operatorname{pen}}_n(m) = \frac{\lambda_n}{n} D_J$$

if m is obtained from an index subset $\mathcal{I}_m$ in $\mathcal{P}_J(\mathcal{I})$. This slight over-penalization leads to the following result:

Proposition 6 *Let $\alpha_0 < r$ be fixed, let $0 < \alpha \leq \alpha_0$, and let $\hat{s}^{(ht)}_{\tilde{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(h)}_n$. Then under Assumptions* (2.4), (2.5), (2.7), *and* (2.8):

$$MS\big(\hat{s}^{(ht)}_{\tilde{m}}, \rho_\alpha\big) =: \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}.$$

Modifying Massart's strategy in order to obtain a practical estimator changes the maxiset performance. The previous set $\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}$ is intersected with the strong Besov space $\mathcal{B}^{\alpha/(1+2\alpha)}_{2,\infty}$. Nevertheless, as will be proved in Sect. 4.4, the maxiset $MS(\hat{s}^{(ht)}_{\tilde{m}}, \rho_\alpha)$ is still a large functional space. Indeed, for every $\alpha > 0$ and every p satisfying $p \geq \max(1, 2(\frac{1}{1+2\alpha} + 2\alpha)^{-1})$,

$$\mathcal{B}^{\alpha}_{p,\infty} \subseteq \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}. \tag{3.3}$$

3.4 Comparisons of Model Selection Estimators

In this paragraph, we compare the maxiset performances of the different model selection procedures described previously. For a chosen rate of convergence, let us recall that the larger the maxiset, the better the estimator. To begin, we propose to focus on the model selection estimators which are tractable from the computational point of view. Gathering Propositions 3, 4, and 6, we obtain the following comparison:

Proposition 7 *Let* $0 < \alpha < r$.

– *If for every* n, $\lambda_n = \lambda_0 \log(n)$ *with* λ_0 *large enough, then*

$$MS\big(\hat{s}^{(st)}_{\hat{m}}, \rho_\alpha\big) \subsetneq MS\big(\hat{s}^{(ht)}_{\tilde{m}}, \rho_\alpha\big) \subsetneq MS\big(\hat{s}^{(l)}_{\hat{m}}, \rho_\alpha\big). \tag{3.4}$$

– *If for every* n, $\lambda_n = \lambda_0$ *with* λ_0 *large enough, then*

$$MS\big(\hat{s}^{(st)}_{\hat{m}}, \rho_\alpha\big) \subsetneq MS\big(\hat{s}^{(ht)}_{\tilde{m}}, \rho_\alpha\big). \tag{3.5}$$

This means the following:

– If for every n, $\lambda_n = \lambda_0 \log(n)$ with λ_0 large enough, then, according to the maxiset point of view, the estimator $\hat{s}^{(l)}_{\hat{m}}$ strictly outperforms the estimator $\hat{s}^{(ht)}_{\tilde{m}}$, which strictly outperforms the estimator $\hat{s}^{(st)}_{\hat{m}}$.
– If for every n, $\lambda_n = \lambda_0$ or $\lambda_n = \lambda_0 \log(n)$ with λ_0 large enough, then, according to the maxiset point of view, the estimator $\hat{s}^{(ht)}_{\tilde{m}}$ strictly outperforms the estimator $\hat{s}^{(st)}_{\hat{m}}$.

The corresponding embeddings of functional spaces are proved in Sect. 4.4. The hard thresholding estimator $\hat{s}^{(l)}_{\hat{m}}$ appears as the best estimator when λ_n grows logarithmically, while estimator $\hat{s}^{(ht)}_{\tilde{m}}$ is the best estimator when λ_n is constant. In both cases, those estimators perform very well, since their maxiset contains all the Besov spaces $\mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{p,\infty}$ with $p \geq \max(1, (\frac{1}{1+2\alpha} + 2\alpha)^{-1})$.

We forget now the calculability issues and consider the maxiset of the original procedure proposed by Massart. Propositions 4, 5, and 6 lead then to the following result:

Proposition 8 *Let* $0 < \alpha < r$.

– *If for any* n, $\lambda_n = \lambda_0 \log(n)$ *with* λ_0 *large enough, then*

$$MS\big(\hat{s}^{(h)}_{\hat{m}}, \rho_\alpha\big) \not\subset MS\big(\hat{s}^{(l)}_{\hat{m}}, \rho_\alpha\big) \quad \text{and} \quad MS\big(\hat{s}^{(l)}_{\hat{m}}, \rho_\alpha\big) \not\subset MS\big(\hat{s}^{(h)}_{\hat{m}}, \rho_\alpha\big). \tag{3.6}$$

– *If for any* n, $\lambda_n = \lambda_0$ *or* $\lambda_n = \lambda_0 \log(n)$ *with* λ_0 *large enough, then*

$$MS\big(\hat{s}^{(ht)}_{\tilde{m}}, \rho_\alpha\big) \subsetneq MS\big(\hat{s}^{(h)}_{\hat{m}}, \rho_\alpha\big). \tag{3.7}$$

Hence within the maxiset framework, the estimator $\hat{s}^{(h)}_{\hat{m}}$ strictly outperforms the estimator $\hat{s}^{(ht)}_{\tilde{m}}$, while the estimators $\hat{s}^{(h)}_{\hat{m}}$ and $\hat{s}^{(l)}_{\hat{m}}$ are not comparable. Note that we did not consider the maxisets of the estimator $\hat{s}^{(s)}_{\hat{m}}$ in this section, as they are identical to the ones of the tractable estimator $\hat{s}^{(st)}_{\hat{m}}$. We summarize all those embeddings in Fig. 1 and Fig. 2: Fig. 1 represents these maxiset embeddings for the choice $\lambda_n = \lambda_0 \log(n)$, while Fig. 2 represents these maxiset embeddings for the choice $\lambda_n = \lambda_0$.

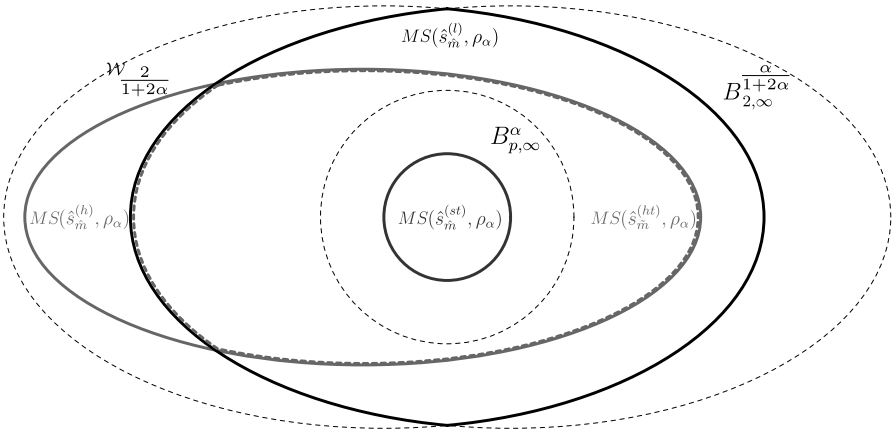

Fig. 1 Maxiset embeddings when $\lambda_n = \lambda_0 \log(n)$ and $\max(1, 2(\frac{1}{1+2\alpha} + 2\alpha)^{-1}) \le p \le 2$

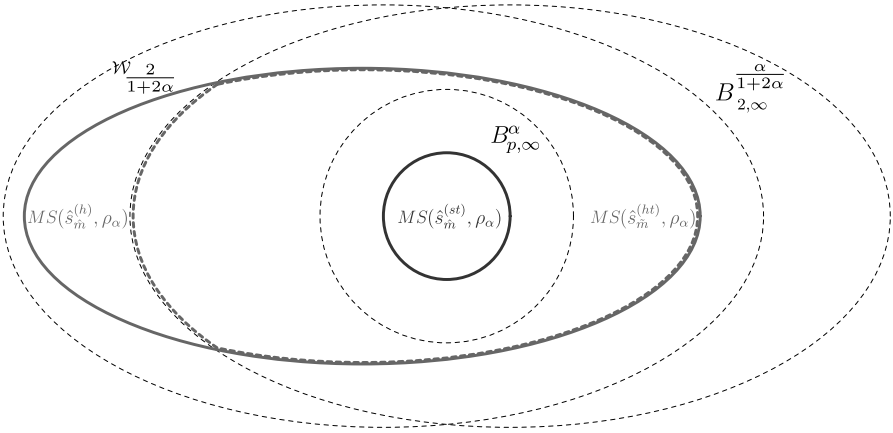

Fig. 2 Maxiset embeddings when $\lambda_n = \lambda_0$ and $\max(1, 2(\frac{1}{1+2\alpha} + 2\alpha)^{-1}) \le p \le 2$

4 Proofs

For any functions u and u' of $\mathbb{L}_2(\mathcal{D})$, we denote by $\langle u, u' \rangle$ the $\mathbb{L}_2$-scalar product between u and u':

$$\langle u, u' \rangle = \int_{\mathcal{D}} u(t)u'(t)\,dt.$$

We denote by C a constant whose value may change at each line.

4.1 Proof of Theorem 2

Without loss of generality, we assume that $n_0 = 1$. We start by constructing a different representation of the white noise model. For any model m, we define $\mathbf{W}_m$, the

projection of the noise on m, by

$$\mathbf{W}_m = \sum_{i=1}^{D_m} W_{e_i^m} e_i^m, \quad W_{e_i^m} = \int_{\mathcal{D}} e_i^m(t)\, dW_t,$$

where $\{e_i^m\}_{i=1}^{D_m}$ is any orthonormal basis of m. For any function $s \in m$, we have:

$$W_s = \int_{\mathcal{D}} s(t)\, dW_t = \sum_{i=1}^{D_m} \langle s, e_i^m \rangle W_{e_i^m} = \langle \mathbf{W}_m, s \rangle.$$

The key observation is now that with high probability, $\|\mathbf{W}_m\|^2$ can be controlled simultaneously over all models. More precisely, for any $m, m' \in \mathcal{M}_n$, we define the space $m + m'$ as the space spanned by the functions of m and m', and control the norm of $\|\mathbf{W}_{m+m'}\|^2$.

Lemma 1 *Let n be fixed and*

$$A_n = \Big\{ \sup_{m \in \mathcal{M}_n} \sup_{m' \in \mathcal{M}_n} \big\{ (D_m + D_{m'})^{-1} \|\mathbf{W}_{m+m'}\|^2 \big\} \le \lambda_n \Big\}.$$

Then under Assumption (2.6), we have $\mathbb{P}\{A_n\} \ge p$.

Proof The Cirelson–Ibragimov–Sudakov inequality (see [19], page 10) implies that for any $t > 0$, any $m \in \mathcal{M}_n$, and any $m' \in \mathcal{M}_n$,

$$\mathbb{P}\big\{ \|\mathbf{W}_{m+m'}\| \ge \mathbb{E}\big[\|\mathbf{W}_{m+m'}\|\big] + t \big\} \le e^{-\frac{t^2}{2}}.$$

Since

$$\mathbb{E}\big[\|\mathbf{W}_{m+m'}\|\big] \le \sqrt{\mathbb{E}\big[\|\mathbf{W}_{m+m'}\|^2\big]} \le \sqrt{D_m + D_{m'}},$$

with $t = \sqrt{\lambda_n (D_m + D_{m'})} - \sqrt{D_m + D_{m'}}$, we obtain

$$\mathbb{P}\big\{ \|\mathbf{W}_{m+m'}\|^2 \ge \lambda_n (D_m + D_{m'}) \big\} \le e^{-\frac{(\sqrt{\lambda_n}-1)^2 (D_m + D_{m'})}{2}}.$$

Assumption (2.6) thus implies that

$$\begin{aligned} 1 - \mathbb{P}\{A_n\} &\le \sum_{m \in \mathcal{M}_n} \sum_{m' \in \mathcal{M}_n} \mathbb{P}\big\{ \|\mathbf{W}_{m+m'}\|^2 \ge \lambda_n (D_m + D'_m) \big\} \\ &\le \sum_{m \in \mathcal{M}_n} \sum_{m' \in \mathcal{M}_n} e^{-\frac{(\sqrt{\lambda_n}-1)^2 (D_m + D_{m'})}{2}} \\ &\le \Big(\sum_{m \in \mathcal{M}_n} e^{-\frac{(\sqrt{\lambda_n}-1)^2 D_m}{2}} \Big)^2 \le 1 - p. \end{aligned}$$

□

We define $m_0(n)$ (denoted m_0 when there is no ambiguity), the model that minimizes a quantity close to $Q(s,n)$:

$$m_0(n) = \operatorname*{argmin}_{m\in\mathcal{M}_n}\left\{\|s_m - s\|^2 + \frac{\lambda_n}{Kn}D_m\right\},$$

where K is an absolute constant larger than 1 specified later. The proof of the theorem begins by a bound on $\|s_{m_0} - s\|^2$:

Lemma 2 *For any* $0 < \gamma < 1$,

$$\|s_{m_0} - s\|^2 \le \frac{\tilde{K} + 4\gamma^{-1}}{\tilde{K}\mathbb{P}\{A_n\}}\mathbb{E}\big[\|\hat{s}_{\hat{m}} - s\|^2\big] + \left(\frac{K(2\gamma^{-1}+1)}{\tilde{K}\mathbb{P}\{A_n\}} + \frac{2K\gamma\lambda_n}{\tilde{K}}\right)\frac{D_{m_0}}{Kn} \quad (4.1)$$

if the constant $\tilde{K} = K(1-\gamma) - 2\gamma^{-1} - 1$ *satisfies* $\tilde{K} > 0$.

Proof By definition,

$$\gamma_n(\hat{s}_{\hat{m}}) + \lambda_n\frac{D_{\hat{m}}}{n} \le \gamma_n(\hat{s}_{m_0}) + \lambda_n\frac{D_{m_0}}{n}.$$

Thus,

$$\begin{aligned}
\lambda_n\frac{D_{\hat{m}} - D_{m_0}}{n} &\le \gamma_n(\hat{s}_{m_0}) - \gamma_n(\hat{s}_{\hat{m}}) \\
&\le -2Y_n(\hat{s}_{m_0}) + \|\hat{s}_{m_0}\|^2 + 2Y_n(\hat{s}_{\hat{m}}) - \|\hat{s}_{\hat{m}}\|^2 \\
&\le -2\langle\hat{s}_{m_0}, s\rangle + \|\hat{s}_{m_0}\|^2 + 2\langle\hat{s}_{\hat{m}}, s\rangle - \|\hat{s}_{\hat{m}}\|^2 + \frac{2}{\sqrt{n}}W_{\hat{s}_{\hat{m}}-\hat{s}_{m_0}} \\
&\le \|\hat{s}_{m_0} - s\|^2 - \|\hat{s}_{\hat{m}} - s\|^2 + \frac{2}{\sqrt{n}}W_{\hat{s}_{\hat{m}}-\hat{s}_{m_0}}.
\end{aligned}$$

Let $0 < \gamma < 1$. As $\hat{s}_{\hat{m}} - \hat{s}_{m_0}$ is supported by the space $\hat{m} + m_0$ spanned by the functions of $\hat{m}$ and m_0, we obtain with the previous definition:

$$\begin{aligned}
\lambda_n\frac{D_{\hat{m}} - D_{m_0}}{n} &\le \|\hat{s}_{m_0} - s\|^2 - \|\hat{s}_{\hat{m}} - s\|^2 + \frac{2}{\sqrt{n}}\langle\mathbf{W}_{\hat{m}+m_0}, \hat{s}_{\hat{m}} - \hat{s}_{m_0}\rangle \\
&\le \|\hat{s}_{m_0} - s\|^2 - \|\hat{s}_{\hat{m}} - s\|^2 + \frac{\gamma}{n}\|\mathbf{W}_{\hat{m}+m_0}\|^2 \\
&\quad + \frac{2}{\gamma}\big(\|\hat{s}_{m_0} - s\|^2 + \|\hat{s}_{\hat{m}} - s\|^2\big) \\
&\le \left(\frac{2}{\gamma} + 1\right)\|\hat{s}_{m_0} - s\|^2 + \left(\frac{2}{\gamma} - 1\right)\|\hat{s}_{\hat{m}} - s\|^2 + \frac{\gamma}{n}\|\mathbf{W}_{\hat{m}+m_0}\|^2.
\end{aligned}$$

We multiply now by $\mathbf{1}_{A_n}$ to obtain

$$\begin{aligned}
\lambda_n\mathbf{1}_{A_n}\frac{D_{\hat{m}} - D_{m_0}}{n} \le \left(\frac{2}{\gamma} + 1\right)\mathbf{1}_{A_n}\|\hat{s}_{m_0} - s\|^2 + \left(\frac{2}{\gamma} - 1\right)\mathbf{1}_{A_n}\|\hat{s}_{\hat{m}} - s\|^2 \\
+ \mathbf{1}_{A_n}\frac{\gamma}{n}\|\mathbf{W}_{\hat{m}+m_0}\|^2.
\end{aligned}$$

Now using the definition of A_n and Lemma 1, this yields

$$\lambda_n \mathbf{1}_{A_n} \frac{D_{\hat{m}} - D_{m_0}}{n} \le \left(\frac{2}{\gamma} + 1\right) \mathbf{1}_{A_n} \|\hat{s}_{m_0} - s\|^2 + \left(\frac{2}{\gamma} - 1\right) \mathbf{1}_{A_n} \|\hat{s}_{\hat{m}} - s\|^2 + \gamma \lambda_n \mathbf{1}_{A_n} \frac{D_{\hat{m}} + D_{m_0}}{n},$$

and thus

$$(1-\gamma)\lambda_n \mathbf{1}_{A_n} \frac{D_{\hat{m}} - D_{m_0}}{n} \le \left(\frac{2}{\gamma} + 1\right) \mathbf{1}_{A_n} \|\hat{s}_{m_0} - s\|^2 + \left(\frac{2}{\gamma} - 1\right) \mathbf{1}_{A_n} \|\hat{s}_{\hat{m}} - s\|^2 + 2\gamma \lambda_n \mathbf{1}_{A_n} \frac{D_{m_0}}{n}.$$

One obtains

$$\lambda_n \mathbf{1}_{A_n} \frac{D_{\hat{m}} - D_{m_0}}{n} \le \frac{\frac{2}{\gamma} + 1}{1-\gamma} \mathbf{1}_{A_n} \|\hat{s}_{m_0} - s\|^2 + \frac{\frac{2}{\gamma} - 1}{1-\gamma} \mathbf{1}_{A_n} \|\hat{s}_{\hat{m}} - s\|^2 + \frac{2\gamma}{1-\gamma} \lambda_n \mathbf{1}_{A_n} \frac{D_{m_0}}{n}. \tag{4.2}$$

We derive now a bound on $\|s_{m_0} - s\|^2$. By definition,

$$\|s_{m_0} - s\|^2 + \lambda_n \frac{D_{m_0}}{Kn} \le \|s_{\hat{m}} - s\|^2 + \lambda_n \frac{D_{\hat{m}}}{Kn},$$

and thus

$$\|s_{m_0} - s\|^2 \le \|s_{\hat{m}} - s\|^2 + \lambda_n \frac{D_{\hat{m}} - D_{m_0}}{Kn}.$$

By multiplying by $\mathbf{1}_{A_n}$ and plugging in the bound (4.2), we have:

$$\begin{aligned}
\mathbf{1}_{A_n} \|s_{m_0} - s\|^2 &\le \mathbf{1}_{A_n} \|s_{\hat{m}} - s\|^2 + \lambda_n \mathbf{1}_{A_n} \frac{D_{\hat{m}} - D_{m_0}}{Kn} \\
&\le \mathbf{1}_{A_n} \|s_{\hat{m}} - s\|^2 + \frac{\frac{2}{\gamma} + 1}{K(1-\gamma)} \mathbf{1}_{A_n} \|\hat{s}_{m_0} - s\|^2 \\
&\quad + \frac{\frac{2}{\gamma} - 1}{K(1-\gamma)} \mathbf{1}_{A_n} \|\hat{s}_{\hat{m}} - s\|^2 + \frac{2\gamma}{K(1-\gamma)} \lambda_n \mathbf{1}_{A_n} \frac{D_{m_0}}{n} \\
&\le \left(1 + \frac{\frac{2}{\gamma} - 1}{K(1-\gamma)}\right) \mathbf{1}_{A_n} \|\hat{s}_{\hat{m}} - s\|^2 \\
&\quad + \frac{\frac{2}{\gamma} + 1}{K(1-\gamma)} \mathbf{1}_{A_n} \left(\|s_{m_0} - s\|^2 + \frac{1}{n} \|W_{m_0}\|^2\right) \\
&\quad + \frac{2\gamma}{K(1-\gamma)} \lambda_n \mathbf{1}_{A_n} \frac{D_{m_0}}{n}
\end{aligned}$$

$$\leq \left(1+\frac{\frac{2}{\gamma}-1}{K(1-\gamma)}\right)\|\hat{s}_{\hat{m}}-s\|^2+\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}\mathbf{1}_{A_n}\|s_{m_0}-s\|^2$$
$$+\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}\frac{1}{n}\|W_{m_0}\|^2+\frac{2\gamma}{K(1-\gamma)}\lambda_n\mathbf{1}_{A_n}\frac{D_{m_0}}{n},$$

and thus

$$\left(1-\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}\right)\mathbf{1}_{A_n}\|s_{m_0}-s\|^2$$
$$\leq \left(1+\frac{\frac{2}{\gamma}-1}{K(1-\gamma)}\right)\|\hat{s}_{\hat{m}}-s\|^2$$
$$+\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}\frac{1}{n}\|W_{m_0}\|^2+\frac{2\gamma}{K(1-\gamma)}\lambda_n\mathbf{1}_{A_n}\frac{D_{m_0}}{n}.$$

Taking the expectation on both sides yields:

$$\left(1-\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}\right)\mathbb{P}\{A_n\}\|s_{m_0}-s\|^2$$
$$\leq \left(1+\frac{\frac{2}{\gamma}-1}{K(1-\gamma)}\right)\mathbb{E}\big[\|\hat{s}_{\hat{m}}-s\|^2\big]$$
$$+\left(\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}+\frac{2\gamma}{K(1-\gamma)}\mathbb{P}\{A_n\}\lambda_n\right)\frac{D_{m_0}}{n},$$

and thus as soon as $1-\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}>0$,

$$\|s_{m_0}-s\|^2\leq \frac{1+\frac{\frac{2}{\gamma}-1}{K(1-\gamma)}}{\big(1-\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}\big)\mathbb{P}\{A_n\}}\mathbb{E}\big[\|\hat{s}_{\hat{m}}-s\|^2\big]$$
$$+\frac{\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}+\frac{2\gamma}{K(1-\gamma)}\mathbb{P}\{A_n\}\lambda_n}{\big(1-\frac{\frac{2}{\gamma}+1}{K(1-\gamma)}\big)\mathbb{P}\{A_n\}}\frac{D_{m_0}}{n}$$
$$\leq \frac{K(1-\gamma)+\frac{2}{\gamma}-1}{\big(K(1-\gamma)-\frac{2}{\gamma}-1\big)\mathbb{P}\{A_n\}}\mathbb{E}\big[\|\hat{s}_{\hat{m}}-s\|^2\big]$$
$$+\frac{\frac{2}{\gamma}+1+2\gamma\mathbb{P}\{A_n\}\lambda_n}{\big(K(1-\gamma)-\frac{2}{\gamma}-1\big)\mathbb{P}\{A_n\}}\frac{D_{m_0}}{n}$$
$$\leq \frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_n\}}\mathbb{E}\big[\|\hat{s}_{\hat{m}}-s\|^2\big]+\frac{\frac{2}{\gamma}+1+2\gamma\mathbb{P}\{A_n\}\lambda_n}{\tilde{K}\mathbb{P}\{A_n\}}\frac{D_{m_0}}{n},$$

which yields:

$$\|s_{m_0} - s\|^2 \le \frac{\tilde{K} + \frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_n\}}\mathbb{E}\big[\|\hat{s}_{\hat{m}} - s\|^2\big] + \left(\frac{K\left(\frac{2}{\gamma}+1\right)}{\tilde{K}\mathbb{P}\{A_n\}} + \frac{2K\gamma\lambda_n}{\tilde{K}}\right)\frac{D_{m_0}}{Kn}$$

with $\tilde{K} = K(1-\gamma) - \frac{2}{\gamma} - 1$. □

Now, let us specify the constants. We set

$$g(\delta, \alpha_0) = \inf_{\alpha\in(0,\alpha_0]} \inf_{x\in[\frac{1}{2},1-\delta]} \left\{x^{\frac{2\alpha}{2\alpha+1}} - x\right\} = (1-\delta)^{\frac{2\alpha_0}{2\alpha_0+1}} - 1 + \delta \in (0,1).$$

Then we put

$$\gamma = \frac{1}{8}g(\delta,\alpha_0) \quad \text{and} \quad K = \frac{\frac{2}{\gamma}+1}{\frac{1}{2}-\gamma}.$$

This implies $\tilde{K} = \frac{K}{2}$, and assumptions of the previous lemma are satisfied. We consider now the dependency of m_0 on n and prove by induction the following lemma:

Lemma 3 *If there exists $C_1 > 0$ such that for any n,*

$$\mathbb{E}\big[\|\hat{s}_{\hat{m}(n/2)} - s\|^2\big] \le C_1\left(\frac{2\lambda_{n/2}}{n}\right)^{\frac{2\alpha}{2\alpha+1}},$$

then, provided $\lambda_1 \ge \Upsilon(\delta, p, \alpha_0)$, where

$$\Upsilon(\delta, p, \alpha_0) = \frac{8}{pg(\delta,\alpha_0)}\left(\frac{16}{g(\delta,\alpha_0)} + 1\right), \tag{4.3}$$

there exists a constant C_2 such that for any n,

$$\|s_{m_0(n)} - s\|^2 + \lambda_n\frac{D_{m_0(n)}}{Kn} \le C_2\left(\frac{\lambda_n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

Proof By using $\mathcal{M}_{n/2} \subset \mathcal{M}_n$ and (4.1), for any $\beta \in [0,1]$, if we set

$$A = \|s_{m_0(n)} - s\|^2 + \lambda_n\frac{D_{m_0(n)}}{Kn},$$

we have

$$\begin{aligned} A &\le \|s_{m_0(n/2)} - s\|^2 + \lambda_n\frac{D_{m_0(n/2)}}{Kn} \\ &\le \beta\|s_{m_0(n/2)} - s\|^2 + (1-\beta)\|s_{m_0(n/2)} - s\|^2 + \frac{\lambda_n}{2\lambda_{n/2}}\lambda_{n/2}\frac{2D_{m_0(n/2)}}{Kn} \end{aligned}$$

$$\leq \beta \frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\mathbb{E}\big[\|\hat{s}_{\hat{m}(n/2)}-s\|^2\big]+(1-\beta)\|s_{m_0(n/2)}-s\|^2$$
$$+\left(\beta\left(\frac{K\left(\frac{2}{\gamma}+1\right)}{\tilde{K}\mathbb{P}\{A_{n/2}\}\lambda_{n/2}}+\frac{2K\gamma}{\tilde{K}}\right)+\frac{\lambda_n}{2\lambda_{n/2}}\right)\frac{2\lambda_{n/2}D_{m_0(n/2)}}{Kn}.$$

As $\lambda_n \leq 2\lambda_{n/2}$, there exists $\beta_n \in [0,1]$ such that

$$1-\beta_n=\beta_n\left(\frac{K\left(\frac{2}{\gamma}+1\right)}{\tilde{K}\mathbb{P}\{A_{n/2}\}\lambda_{n/2}}+\frac{2K\gamma}{\tilde{K}}\right)+\frac{\lambda_n}{2\lambda_{n/2}},$$

so that

$$A \leq \beta_n \frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\mathbb{E}\big[\|\hat{s}_{\hat{m}(n/2)}-s\|^2\big]$$
$$+(1-\beta_n)\left(\|s_{m_0(n/2)}-s\|^2+\frac{2\lambda_{n/2}D_{m_0(n/2)}}{Kn}\right).$$

The induction can now be started. We assume now that for all $n' \leq n-1$,

$$\|s_{m_0(n')}-s\|^2+\lambda_{n'}\frac{D_{m_0(n')}}{Kn'} \leq C_2\left(\frac{\lambda_{n'}}{n'}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

By assumption,

$$\mathbb{E}\big[\|\hat{s}_{\hat{m}(n/2)}-s\|^2\big] \leq C_1\left(\frac{2\lambda_{n/2}}{n}\right)^{\frac{2\alpha}{2\alpha+1}},$$

so that

$$A \leq \beta_n \frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}C_1\left(\frac{2\lambda_{n/2}}{n}\right)^{\frac{2\alpha}{2\alpha+1}}+(1-\beta_n)C_2\left(\frac{2\lambda_{n/2}}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$$
$$\leq \left(\beta_n \frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\frac{C_1}{C_2}+1-\beta_n\right)\left(\frac{2\lambda_{n/2}}{\lambda_n}\right)^{\frac{2\alpha}{2\alpha+1}}C_2\left(\frac{\lambda_n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

So, we have to prove that

$$\left(\beta_n \frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\frac{C_1}{C_2}+1-\beta_n\right)\left(\frac{2\lambda_{n/2}}{\lambda_n}\right)^{\frac{2\alpha}{2\alpha+1}} \leq 1$$

or equivalently,

$$\left(\beta_n\left(\frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\frac{C_1}{C_2}+\frac{K\left(\frac{2}{\gamma}+1\right)}{\tilde{K}\mathbb{P}\{A_{n/2}\}\lambda_{n/2}}+\frac{2K\gamma}{\tilde{K}}\right)+\frac{\lambda_n}{2\lambda_{n/2}}\right)\left(\frac{2\lambda_{n/2}}{\lambda_n}\right)^{\frac{2\alpha}{2\alpha+1}} \leq 1.$$

This condition can be rewritten as

$$\beta_n\left(\frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\frac{C_1}{C_2}+\frac{K\left(\frac{2}{\gamma}+1\right)}{\tilde{K}\mathbb{P}\{A_{n/2}\}\lambda_{n/2}}+\frac{2K\gamma}{\tilde{K}}\right)\left(\frac{2\lambda_{n/2}}{\lambda_n}\right)^{\frac{2\alpha}{2\alpha+1}}\leq 1-\left(\frac{\lambda_n}{2\lambda_{n/2}}\right)^{\frac{1}{2\alpha+1}}$$

or

$$\lambda_{n/2}\geq\frac{K\left(\frac{2}{\gamma}+1\right)}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\left[\frac{1}{\beta_n}\left(\left(\frac{\lambda_n}{2\lambda_{n/2}}\right)^{\frac{2\alpha}{2\alpha+1}}-\frac{\lambda_n}{2\lambda_{n/2}}\right)-\frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\frac{C_1}{C_2}-\frac{2K\gamma}{\tilde{K}}\right]^{-1}$$

provided the right member is positive. Under the very mild assumption $2(1-\delta)\lambda_{n/2}\geq\lambda_n\geq\lambda_{n/2}$, it is sufficient to ensure that (4.3) is true. Indeed, $\lambda_{n/2}\geq\lambda_1$, and using values of the constants, we have:

$$\begin{aligned}
&\frac{K\left(\frac{2}{\gamma}+1\right)}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\left[\frac{1}{\beta_n}\left(\left(\frac{\lambda_n}{2\lambda_{n/2}}\right)^{\frac{2\alpha}{2\alpha+1}}-\frac{\lambda_n}{2\lambda_{n/2}}\right)-\frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}\mathbb{P}\{A_{n/2}\}}\frac{C_1}{C_2}-\frac{2K\gamma}{\tilde{K}}\right]^{-1}\\
&\quad\leq\frac{2\left(\frac{2}{\gamma}+1\right)}{p}\left[\frac{g(\delta,\alpha_0)}{2}-\frac{\tilde{K}+\frac{4}{\gamma}}{\tilde{K}p}\frac{C_1}{C_2}\right]^{-1}\\
&\quad\leq\frac{8\left(\frac{2}{\gamma}+1\right)}{pg(\delta,\alpha_0)}\\
&\quad\leq\frac{8}{pg(\delta,\alpha_0)}\left(\frac{16}{g(\delta,\alpha_0)}+1\right)
\end{aligned}$$

if

$$C_2\geq\frac{4\tilde{K}+\frac{16}{\gamma}}{\tilde{K}p}\frac{C_1}{g(\delta,\alpha_0)}.$$

□

Finally, Theorem 2 follows from the previous lemma that gives the following inequality:

$$\begin{aligned}
\frac{Q(s,n)}{K}&\leq\inf_{m\in\mathcal{M}_n}\left\{\|s-s_m\|^2+\frac{\lambda_n}{Kn}D_m\right\}\\
&\leq\|s_{m_0(n)}-s\|^2+\frac{\lambda_n}{Kn}D_{m_0(n)}\\
&\leq C_2\left(\frac{\lambda_n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}.
\end{aligned}$$

4.2 Proofs of Theorem 3 and Proposition 1

Theorem 2 implies that for any $s\in MS(\hat{s}_{\hat{m}},\rho_\alpha)$,

$$\sup_n\left\{\rho_{n,\alpha}^{-2}Q(s,n)\right\}<\infty,$$

or, equivalently, there exists $C > 0$ such that for any n,

$$\inf_{m\in\mathcal{M}_n}\left\{\|s_m - s\|^2 + \frac{\lambda_n}{n}D_m\right\} \le C\rho_{n,\alpha}^2. \tag{4.4}$$

By the definition of V_n, any function s_m with $m \in \mathcal{M}_n$ belongs to V_n, and thus Inequality (4.4) implies:

$$\|P_{V_n}s - s\|^2 \le C\rho_{n,\alpha}^2, \tag{4.5}$$

that is, $s \in \mathcal{L}_V^\alpha$. By definition, $\widetilde{\mathcal{M}}$ is a larger collection than $\mathcal{M}_n$, and thus Inequality (4.4) also implies that for any n,

$$\inf_{m\in\widetilde{\mathcal{M}}}\left\{\|s_m - s\|^2 + \frac{\lambda_n}{n}D_m\right\} \le C\rho_{n,\alpha}^2,$$

which turns out to be a characterization of $\mathcal{A}^\alpha_{\widetilde{\mathcal{M}}}$ when $\rho_{n,\alpha} = (\frac{\lambda_n}{n})^{\frac{\alpha}{2\alpha+1}}$ as a consequence of the following lemma:

Lemma 4 *Under assumptions of Theorem 3,*

$$\sup_n\left\{\left(\frac{\lambda_n}{n}\right)^{-\frac{2\alpha}{2\alpha+1}}\inf_{m\in\widetilde{\mathcal{M}}}\left\{\|s_m - s\|^2 + \frac{\lambda_n}{n}D_m\right\}\right\} < \infty \Leftrightarrow s \in \mathcal{A}^\alpha_{\widetilde{\mathcal{M}}}. \tag{4.6}$$

Proof We set

$$\tilde{m}(n) = \arg\min_{m\in\widetilde{\mathcal{M}}}\left\{\|s_m - s\|^2 + \frac{\lambda_n}{n}D_m\right\}.$$

First, let us assume that for any n,

$$\|s_{\tilde{m}(n)} - s\|^2 + \frac{\lambda_n}{n}D_{\tilde{m}(n)} \le C_1\left(\frac{\lambda_n}{n}\right)^{\frac{2\alpha}{2\alpha+1}},$$

where C_1 is a constant. Then

$$D_{\tilde{m}(n)} \le C_1\left(\frac{\lambda_n}{n}\right)^{-\frac{1}{1+2\alpha}}.$$

Using $\lambda_n \le \lambda_{2n} \le 2\lambda_n$, for $M \in \mathbb{N}^*$, as soon as $M \ge C_1(\lambda_1)^{-\frac{1}{1+2\alpha}}$, there exists $n \in \mathbb{N}^*$ such that

$$C_1\left(\frac{\lambda_n}{n}\right)^{-\frac{1}{1+2\alpha}} \le M < C_1\left(\frac{\lambda_{2n}}{2n}\right)^{-\frac{1}{1+2\alpha}} \le C_1 2^{\frac{1}{1+2\alpha}}\left(\frac{\lambda_n}{n}\right)^{-\frac{1}{1+2\alpha}}. \tag{4.7}$$

Then

$$\begin{aligned}\inf_{\{m\in\widetilde{\mathcal{M}}:\ D_m\le M\}}\|s_m-s\|^2 &\le \inf_{\{m\in\widetilde{\mathcal{M}}:\ D_m\le M\}}\left\{\|s_m-s\|^2+\frac{\lambda_n}{n}D_m\right\}\\ &\le \inf_{\{m\in\widetilde{\mathcal{M}}:\ D_m\le C_1(\frac{\lambda_n}{n})^{-\frac{1}{1+2\alpha}}\}}\left\{\|s_m-s\|^2+\frac{\lambda_n}{n}D_m\right\}\\ &\le C_1\left(\frac{\lambda_n}{n}\right)^{\frac{2\alpha}{1+2\alpha}}\\ &\le C_1^{2\alpha+1}2^{\frac{2\alpha}{1+2\alpha}}M^{-2\alpha}.\end{aligned}$$

Conversely, assume that there exists $\tilde{C}_1$ satisfying

$$\inf_{\{m\in\widetilde{\mathcal{M}}:\ D_m\le M\}}\|s_m-s\|^2\le\tilde{C}_1M^{-2\alpha}.$$

Then for any $T>0$,

$$\begin{aligned}\inf_{m\in\widetilde{\mathcal{M}}}\left\{\|s_m-s\|^2+T^2D_m\right\} &= \inf_{M\in\mathbb{N}^*}\inf_{\{m\in\widetilde{\mathcal{M}}:\ D_m=M\}}\left\{\|s_m-s\|^2+T^2M\right\}\\ &\le \inf_{M\in\mathbb{N}^*}\left\{\tilde{C}_1M^{-2\alpha}+T^2M\right\}\\ &\le \inf_{x\in\mathbb{R}_+^*}\left\{\tilde{C}_1x^{-2\alpha}+T^2(x+1)\right\}\\ &\le \tilde{C}_1\left(\frac{T^2}{2\alpha\tilde{C}_1}\right)^{\frac{2\alpha}{1+2\alpha}}+T^2\left(\left(\frac{T^2}{2\alpha\tilde{C}_1}\right)^{-\frac{1}{1+2\alpha}}+1\right)\\ &\le C_1(T^2)^{\frac{2\alpha}{1+2\alpha}},\end{aligned}$$

where C_1 is a constant. □

We have proved so far that $MS(\hat{s}_{\hat{m}},\rho_\alpha)\subset\mathcal{L}_V^\alpha\cap\mathcal{A}^\alpha_{\widetilde{\mathcal{M}}}$. It remains to prove the converse inclusion. Corollary 1 and the previous lemma imply that it suffices to prove that inequalities (4.5) and (4.6) imply inequality (4.4) (possibly with a different constant C).

Let $s\in\mathcal{L}_V^\alpha\cap\mathcal{A}^\alpha_{\widetilde{\mathcal{M}}}$. By inequality (4.6), for every n, there exists a model $m\in\widetilde{\mathcal{M}}$ such that

$$\|s_m-s\|^2+\frac{\lambda_n}{n}D_m\le C\rho_{n,\alpha}^2.$$

By the definition of $\widetilde{\mathcal{M}}$, there exists k such that $m\in\mathcal{M}_k'$.

If $k\le n$, then $m\in\mathcal{M}_n$, and thus

$$\inf_{m\in\mathcal{M}_n}\left\{\|s_m-s\|^2+\frac{\lambda_n}{n}D_m\right\}\le C\rho_{n,\alpha}^2.$$

Otherwise $k > n$, and let $m' \in \mathcal{M}$ be the model such that $\mathcal{I}_m = \mathcal{I}_{m'} \cap \mathcal{J}_k$ as defined in Sect. 2.2. We define $m'' \in \mathcal{M}_n$ by its index set $\mathcal{I}_{m''} = \mathcal{I}_{m'} \cap \mathcal{J}_n$. Note that $m'' \subset m$ and $s_m - s_{m''} \in V_n^\perp$, so

$$\begin{aligned}\|s_{m''} - s\|^2 + \frac{\lambda_n}{n} D_{m''} &= \|s_{m''} - s_m\|^2 + \|s_m - s\|^2 + \frac{\lambda_n}{n} D_{m''} \\ &\le \|P_{V_n} s - s\|^2 + \|s_m - s\|^2 + \frac{\lambda_n}{n} D_m \\ &\le C\rho_{n,\alpha}^2.\end{aligned}$$

Theorem 3 is proved.

The proof of Proposition 1 relies on the definition of $\widetilde{\mathrm{pen}}_n(m)$. Recall that for any model $m \in \mathcal{M}_n'$ there is a model $\tilde{m} \in \mathcal{M}$ such that

$$m = \mathrm{span}\{\varphi_i : i \in \mathcal{I}_{\tilde{m}} \cap \mathcal{J}_n\},$$

and that

$$\widetilde{\mathrm{pen}}_n(m) = \frac{\lambda_n}{n} D_{\tilde{m}}.$$

One deduces

$$\|s_m - s\|^2 + \widetilde{\mathrm{pen}}_n(m) = \|s_m - s\|^2 + \frac{\lambda_n}{n} D_{\tilde{m}} \ge \|s_{\tilde{m}} - s\|^2 + \frac{\lambda_n}{n} D_{\tilde{m}},$$

and thus

$$\inf_{m \in \mathcal{M}_n} \left\{ \|s_m - s\|^2 + \widetilde{\mathrm{pen}}_n(m) \right\} \le C\rho_{n,\alpha}^2 \implies \inf_{m \in \mathcal{M}} \left\{ \|s_m - s\|^2 + \frac{\lambda_n}{n} D_m \right\} \le C\rho_{n,\alpha}^2.$$

Mimicking the proof of Theorem 3, one obtains Proposition 1.

4.3 Proof of Proposition 2

In the same spirit as in the proof of Theorem 2, for any n, we write:

$$m_0(n) = \arg\min_{m \in \mathcal{M}} \left\{ \|s_m - s\|^2 + \frac{\mathrm{pen}(m)}{4} \right\} = \arg\min_{m \in \mathcal{M}} \left\{ \|s_m - s\|^2 + \frac{\lambda_n D_m}{4n} \right\} \quad (4.8)$$

(we have set $K = 4$), and

$$\hat{m}(n) = \arg\min_{m \in \mathcal{M}} \left\{ -\|\hat{s}_m\|^2 + \mathrm{pen}(m) \right\} = \arg\min_{m \in \mathcal{M}} \left\{ -\|\hat{s}_m\|^2 + \frac{\lambda_n D_m}{n} \right\}. \quad (4.9)$$

In the nested case, Lemma 2 becomes the following much stronger lemma:

Lemma 5 *For any n, almost surely*

$$\|s_{m_0(n)} - s\|^2 \le \|\hat{s}_{\hat{m}(n)} - s\|^2. \quad (4.10)$$

Proof As the models are embedded, either $\hat{m}(n) \subset m_0(n)$ or $m_0(n) \subset \hat{m}(n)$.

In the first case, $\|s_{m_0(n)} - s\|^2 \leq \|s_{\hat{m}(n)} - s\|^2 \leq \|\hat{s}_{\hat{m}(n)} - s\|^2$, and thus (4.10) holds.

Otherwise, by construction,

$$\begin{cases} \|s_{m_0(n)} - s\|^2 + \frac{\lambda_n D_{m_0(n)}}{4n} \leq \|s_{\hat{m}(n)} - s\|^2 + \frac{\lambda_n D_{\hat{m}(n)}}{4n}, \\ -\|\hat{s}_{\hat{m}(n)}\|^2 + \frac{\lambda_n D_{\hat{m}(n)}}{n} \leq -\|\hat{s}_{m_0(n)}\|^2 + \frac{\lambda_n D_{m_0(n)}}{n}, \end{cases}$$

and thus as $m_0(n) \subset \hat{m}(n)$,

$$\begin{cases} \|s_{\hat{m}(n)\setminus m_0(n)}\|^2 \leq \frac{\lambda_n D_{\hat{m}(n)}}{4n} - \frac{\lambda_n D_{m_0(n)}}{4n}, \\ \frac{\lambda_n D_{\hat{m}(n)}}{n} - \frac{\lambda_n D_{m_0(n)}}{n} \leq \|\hat{s}_{\hat{m}(n)\setminus m_0(n)}\|^2. \end{cases}$$

Combining these two inequalities yields:

$$\begin{aligned} \|s_{\hat{m}(n)\setminus m_0(n)}\|^2 &\leq \frac{1}{4}\|\hat{s}_{\hat{m}(n)\setminus m_0(n)}\|^2 \\ &\leq \frac{1}{2}\big(\|\hat{s}_{\hat{m}(n)\setminus m_0(n)} - s_{\hat{m}(n)\setminus m_0(n)}\|^2 + \|s_{\hat{m}(n)\setminus m_0(n)}\|^2\big), \end{aligned}$$

and thus

$$\|s_{\hat{m}(n)\setminus m_0(n)}\|^2 \leq \|\hat{s}_{\hat{m}(n)\setminus m_0(n)} - s_{\hat{m}(n)\setminus m_0(n)}\|^2.$$

Now, (4.10) holds as

$$\begin{aligned} \|s_{m_0(n)} - s\|^2 &= \|s_{\hat{m}(n)} - s\|^2 + \|s_{\hat{m}(n)\setminus m_0(n)}\|^2 \\ &\leq \|s_{\hat{m}(n)} - s\|^2 + \|\hat{s}_{\hat{m}(n)\setminus m_0(n)} - s_{\hat{m}(n)\setminus m_0(n)}\|^2 \\ &\leq \|s_{\hat{m}(n)} - s\|^2 + \|\hat{s}_{\hat{m}(n)} - s_{\hat{m}(n)}\|^2 = \|\hat{s}_{\hat{m}(n)} - s\|^2. \end{aligned}$$ □

Now we can conclude the proof of Proposition 2 with an induction similar to the one used in the proof of Lemma 3. Indeed, let

$$A = \|s_{m_0(n)} - s\|^2 + \frac{\lambda_n D_{m_0(n)}}{4n},$$

$$\begin{aligned} A &\leq \|s_{m_0(n/2)} - s\|^2 + \frac{\lambda_n D_{m_0(n/2)}}{4n} \\ &\leq \beta_n \mathbb{E}\big(\|\hat{s}_{\hat{m}(n/2)} - s\|^2\big) + (1-\beta_n)\|s_{m_0(n/2)} - s\|^2 + \frac{\lambda_n}{2\lambda_{n/2}} \frac{\lambda_{n/2} D_{m_0(n/2)}}{4(n/2)}. \end{aligned}$$

The choice $\beta_n = 1 - \frac{\lambda_n}{2\lambda_{n/2}}$ is such that $\delta \leq \beta_n \leq \frac{1}{2}$, and it implies:

$$A \leq \beta_n \mathbb{E}\big(\|\hat{s}_{\hat{m}(n/2)} - s\|^2\big) + (1-\beta_n)\left(\|s_{m_0(n/2)} - s\|^2 + \frac{\lambda_{n/2} D_{m_0(n/2)}}{4(n/2)}\right).$$

Now using almost the same induction as in Theorem 2, we obtain:

$$A \le \beta_n C_1^2 \left(\frac{2\lambda_{n/2}}{n}\right)^{\frac{2\alpha}{1+2\alpha}} + (1-\beta_n) C_2 \left(\frac{2\lambda_{n/2}}{n}\right)^{\frac{2\alpha}{1+2\alpha}}$$
$$\le \left(\frac{2\lambda_{n/2}}{\lambda_n}\right)^{\frac{2\alpha}{1+2\alpha}} \left(C_1^2 \beta_n C_2^{-1} + (1-\beta_n)\right) C_2 \left(\frac{\lambda_n}{n}\right)^{\frac{2\alpha}{1+2\alpha}},$$

where C_1 is a constant. It thus suffices to verify that

$$\left(\frac{2\lambda_{n/2}}{\lambda_n}\right)^{\frac{2\alpha}{1+2\alpha}} \left(C_1^2 \beta_n C_2^{-1} + (1-\beta_n)\right) \le 1,$$

which is the case as soon as $C_2 \ge \frac{C_1^2}{2g(\delta,\alpha)}$.

4.4 Space Embeddings

In this paragraph we provide many embedding properties between the functional spaces considered in Sect. 3. Let us recall the following definitions:

$$\mathcal{B}^{\alpha}_{p,\infty} = \left\{ s \in \mathbb{L}_2\big([0,1]\big) : \sup_{J\in\mathbb{N}} 2^{J(\alpha-\frac{1}{p}+\frac{1}{2})p} \sum_{k=0}^{2^j-1} |\beta_{jk}|^p < \infty \right\};$$

$$\mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} = \left\{ s \in \mathbb{L}_2\big([0,1]\big) : \sup_{J\in\mathbb{N}} 2^{\frac{2J\alpha}{1+2\alpha}} \sum_{j\ge J} \sum_{k=0}^{2^j-1} \beta_{jk}^2 < \infty \right\};$$

$$\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} = \left\{ s \in \mathbb{L}_2\big([0,1]\big) : \sup_{J\in\mathbb{N}} 2^{2J\alpha} \sum_{j\ge J} \sum_{k=\lfloor 2^J (j-J+1)^{-\theta} \rfloor}^{2^j} |\beta_j|^2_{(k)} < \infty \right\};$$

$$\mathcal{W}_{\frac{2}{1+2\alpha}} = \left\{ s \in \mathbb{L}_2\big([0,1]\big) : \sup_{u>0} u^{\frac{2}{1+2\alpha}} \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \mathbf{1}_{|\beta_{jk}|>u} < \infty \right\}.$$

4.4.1 Space Embeddings: Part I

$$\bigcup_{p\ge 1, p>\frac{2}{1+2\alpha}} \mathcal{B}^{\alpha}_{p,\infty} \overset{\text{(i)}}{\subsetneq} \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} \overset{\text{(ii)}}{\subsetneq} \mathcal{W}_{\frac{2}{1+2\alpha}}.$$

Proof of (i) Let s belong to $B^{\alpha}_{p,\infty}$ with $p \ge 1$ and $p > \frac{2}{1+2\alpha}$, and, for any scale $j \in \mathbb{N}$, let us denote by $(|\beta_j|_{(k)})_k$ the sequence of the non-decreasing reordered wavelet coefficients of any level j. Then there exists a non-negative constant C such that for any $j \in \mathbb{N}$,

$$\sum_{k=1}^{2^j} |\beta_j|^p_{(k)} \le C 2^{-jp(\alpha+1/2-1/p)}.$$

Fix $J \in \mathbb{N}$. If $p < 2$, according to Lemma 4.16 of [19], for all j larger than J,

$$\sum_{k=\lfloor 2^J(j-J+1)^{-\theta}\rfloor+1}^{2^j} |\beta_j|_{(k)}^2 \le C^{2/p}\, 2^{-2j(\alpha+1/2-1/p)}\big(\lfloor 2^J(j-J+1)^{-\theta}\rfloor\big)^{1-2/p}$$
$$\le C^{2/p}\, 2^{-2J\alpha}2^{-2(j-J)(\alpha+1/2-1/p)}(j-J+1)^{\theta(2/p-1)}.$$

Summing over the indices j larger than J yields:

$$\sum_{j\ge J}\sum_{k=\lfloor 2^J(j-J+1)^{-\theta}\rfloor}^{2^j} |\beta_j|_{(k)}^2 \le C^{2/p}2^{-2J\alpha}\sum_{j'\ge 0} 2^{-2j'(\alpha+1/2-1/p)}(j'+1)^{\theta(2/p-1)},$$

and thus

$$\sup_{J\ge 0} 2^{2J\alpha}\sum_{j\ge J}\sum_{k=\lfloor 2^J(j-J+1)^{-\theta}\rfloor}^{2^j} |\beta_j|_{(k)}^2$$
$$\le C^{2/p}\sum_{j'\ge 0} 2^{-2j'(\alpha+1/2-1/p)}(j'+1)^{\theta(2/p-1)} < \infty.$$

So s belongs to $\mathcal{A}^\alpha_{\mathcal{M}^{(h)}}$.

For the case $p = 2$,

$$\sum_{j\ge J}\sum_{k=\lfloor 2^J(j-J+1)^{-\theta}\rfloor}^{2^j} |\beta_j|_{(k)}^2 \le \sum_{j\ge J}\sum_{k=1}^{2^j} |\beta_j|_{(k)}^2 \le \sum_{j\ge J} C2^{-2j\alpha} \le C\frac{2^{-2J\alpha}}{1-2^{-2\alpha}}.$$

Thus,

$$\sup_{J\in\mathbb{N}} 2^{2J\alpha}\sum_{j\ge J}\sum_{k=\lfloor 2^J(j-J+1)^{-\theta}\rfloor}^{2^j} |\beta_j|_{(k)}^2 < \infty.$$

So s also belongs to $\mathcal{A}^\alpha_{\mathcal{M}^{(h)}}$.

We conclude that for any $p \ge 1$ satisfying $p > \frac{2}{1+2\alpha}$, $B^\alpha_{p,\infty} \subseteq \mathcal{A}^\alpha_{\mathcal{M}^{(h)}}$.

Let us now prove the strict inclusion by considering the function s_0 defined as follows:

$$s_0 = \sum_{j\ge 0}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} = \sum_{j\ge 0} 2^{-\sqrt{j}}\psi_{j,0}.$$

For any (α', p) such that $\alpha' > \max(\frac{1}{p}-\frac{1}{2}, 0)$,

$$2^{(\alpha'-\frac{1}{p}+\frac{1}{2})pj}\sum_{k=0}^{2^j-1}|\beta_{j,k}|^p = 2^{(\alpha'-\frac{1}{p}+\frac{1}{2})pj}2^{-\sqrt{j}p}$$

and thus goes to $+\infty$ when j goes to $+\infty$. This implies that s_0 does not belong to $\mathcal{B}^{\alpha}_{p,\infty}$ for any $p > \frac{2}{1+2\alpha}$.

Now for any $J \in \mathbb{N}$,

$$2^{2J\alpha} \sum_{j \geq J} \sum_{k \geq \lfloor 2^J (j-J+1)^{-\theta} \rfloor}^{2^j} |\beta_j|^2_{(k)} = 2^{2J\alpha} \sum_{j \geq \min\{j' \geq J : 2^J (j'-J+1)^{-\theta} < 1\}} 2^{-2\sqrt{j}} \leq 2^{2J\alpha} \sum_{j \geq 2^{J/\theta}+J} 2^{-2\sqrt{j}},$$

which implies:

$$\sup_{J \geq 0} 2^{2J\alpha} \sum_{j \geq J} \sum_{k \geq \lfloor 2^J (j-J+1)^{-\theta} \rfloor}^{2^j} |\beta_j|^2_{(k)} < \infty,$$

and thus $s_0 \in \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}$. Hence (i) is proved. □

Proof of (ii) There is no doubt that $\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} \subseteq \mathcal{W}_{\frac{2}{1+2\alpha}}$, since $\mathcal{W}_{\frac{2}{1+2\alpha}} = \mathcal{A}^{\alpha}_{\mathcal{M}^{(l)}}$. The strict inclusion is a direct consequence of (iv), just below. □

4.4.2 Space Embeddings: Part II

$$\bigcup_{p \geq \max(1, \frac{2}{(1+2\alpha)^{-1}+2\alpha})} \mathcal{B}^{\alpha}_{p,\infty} \overset{\text{(iii)}}{\subseteq} \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} \overset{\text{(iv)}}{\subsetneq} \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{W}_{\frac{2}{1+2\alpha}}.$$

Proof of (iii) Let $\alpha > 0$ and $p \geq 1$ satisfying $p \geq 2((1+2\alpha)^{-1} + 2\alpha)^{-1}$. Using the classical Besov embeddings $\mathcal{B}^{\alpha}_{p,\infty} \subseteq \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty}$, and, according to (i), we have $\mathcal{B}^{\alpha}_{p,\infty} \subsetneq \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}$. Hence $\mathcal{B}^{\alpha}_{p,\infty} \subseteq \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}$, and (iii) is proved. □

Proof of (iv) We already know that $\mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} \subseteq \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{W}_{\frac{2}{1+2\alpha}}$. The strict inclusion is a direct consequence of (iv) proved in the next subsection. □

4.4.3 A Non-embedded Case

$$\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} \overset{\text{(v)}}{\not\subset} \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{W}_{\frac{2}{1+2\alpha}} \quad \text{and} \quad \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{W}_{\frac{2}{1+2\alpha}} \overset{\text{(vi)}}{\not\subset} \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}.$$

Proof of (v) Let us consider the function $s_0 \in \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}$ defined in the proof of (i). We already know that it does not belong to $\mathcal{B}^{\alpha'}_{p,\infty}$ for any (α', p) satisfying $\alpha' > \max(\frac{1}{p} - \frac{1}{2}, 0)$. As a consequence for the case $(\alpha', p) = (\frac{\alpha}{1+2\alpha}, 2)$, where $\alpha > 0$, we deduce that s_0 does not belong to $\mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty}$.

Moreover, we immediately deduce that $\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} \not\subset \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{W}_{\frac{2}{1+2\alpha}}$. □

Proof of (vi) Let $s_1 \in \mathbb{L}^2([0,1])$ whose wavelet expansion is given by

$$s_1 = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \beta_{jk} \psi_{jk}.$$

We set

$$\beta_{jk} = \begin{cases} 2^{-\frac{j}{2}} & \text{if } k < 2^{\frac{j}{1+2\alpha}}, \\ 0 & \text{otherwise.} \end{cases}$$

We are going to prove that $s_1 \in \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{W}_{\frac{2}{1+2\alpha}}$, while $s_1 \notin \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}$. Summing at a given scale j yields:

$$\sum_{k=0}^{2^j-1} \beta_{jk}^2 = 2^{\frac{j}{1+2\alpha}} 2^{-j} = 2^{-\frac{2\alpha j}{1+2\alpha}},$$

and thus $s_1 \in B^{\frac{\alpha}{1+2\alpha}}_{2,\infty}$.

Let $0 < u < 1$ and j_u be the real number such that $2^{j_u} = u^{-2}$. Then

$$\begin{aligned} u^{\frac{2}{1+2\alpha}} \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \mathbf{1}_{|\beta_{jk}|>u} &= u^{\frac{2}{1+2\alpha}} \sum_{j<j_u} \sum_{k=0}^{2^j-1} \mathbf{1}_{|\beta_{jk}|>u} \\ &= u^{\frac{2}{1+2\alpha}} \sum_{j<j_u} 2^{\frac{j}{1+2\alpha}} \\ &\le 2^{\frac{1}{1+2\alpha}} \left(2^{\frac{1}{1+2\alpha}} - 1\right)^{-1}. \end{aligned}$$

So

$$\sup_{u>0} u^{\frac{2}{1+2\alpha}} \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \mathbf{1}_{|\beta_{jk}|>u} < \infty,$$

and $s_1 \in \mathcal{W}_{\frac{2}{1+2\alpha}}$.

Let us now prove that s_1 does not belong to $\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}$. Fix $J \in \mathbb{N}$ large enough. Then

$$\begin{aligned} E_J &= \sum_{j \ge J} \sum_{k=\lfloor 2^J (j-J+1)^{-\theta} \rfloor}^{2^j-1} |\beta_j|^2_{(k)} \\ &= \sum_{j \ge J} \max\left(0, 2^{j/(2\alpha+1)} - \frac{2^J}{(j-J+1)^{\theta}}\right) 2^{-j}. \end{aligned}$$

Let $J^\star$ be the real number such that $2^{\frac{J^\star}{1+2\alpha}} = \frac{2^J}{(J^\star - J+1)^\theta}$.

From $J^\star = (2\alpha+1)J - (2\alpha+1)\theta \log_2(J^\star - J + 1)$, one thus deduces $J^\star \le (2\alpha+1)J$, which implies $J^\star \ge (2\alpha+1)J - (2\alpha+1)\theta \log_2(2\alpha J + 1)$, and finally $J^\star \le (2\alpha+1)J - (2\alpha+1)\theta \log_2(2\alpha J + 1 - (2\alpha+1)\theta \log_2(2\alpha J+1))$. So,

$$\begin{aligned} E_J &= \sum_{j>J^\star} \left(2^{j/(2\alpha+1)} - \frac{2^J}{(j-J+1)^\theta}\right) 2^{-j} \\ &\ge \sum_{j>J^\star} \left(2^{j/(2\alpha+1)} - 2^{J^\star/(2\alpha+1)}\right) 2^{-j} \\ &\ge C\ 2^{-2J^\star\alpha/(2\alpha+1)} \\ &\ge C\ (\log)^{2\alpha\theta}\ 2^{-2J\alpha}. \end{aligned}$$

So,

$$\sup_{J\ge 0} 2^{2J\alpha} \sum_{j\ge J} \sum_{k\ge \lfloor 2^J (j-J+1)^{-\theta} \rfloor}^{2^j-1} |\beta_j|^2_{(k)} = \infty.$$

This implies that $s_1 \notin \mathcal{A}^\alpha_{\mathcal{M}^{(h)}}$. Finally (vi) is proved. □

Acknowledgements We warmly thank the anonymous referees for their careful reading and their remarks which allowed us to improve the paper.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, Tsahkadsor, 1971, pp. 267–281. Akadémiai Kiadó, Budapest (1973)
2. Autin, F.: Maxiset for density estimation on $\mathbb{R}$. Math. Methods Stat. **15**(2), 123–145 (2006)
3. Autin, F.: Maxisets for μ-thresholding rules. Test **17**(2), 332–349 (2008)
4. Autin, F., Picard, D., Rivoirard, V.: Large variance Gaussian priors in Bayesian nonparametric estimation: a maxiset approach. Math. Methods Stat. **15**(4), 349–373 (2006)
5. Baraud, Y.: Model selection for regression on a fixed design. Probab. Theory Relat. Fields **117**(4), 467–493 (2000)
6. Baraud, Y.: Model selection for regression on a random design. ESAIM Probab. Stat. **6**, 127–146 (2002)
7. Barron, A., Birgé, L., Massart, P.: Risk bounds for model selection via penalization. Probab. Theory Relat. Fields **113**(3), 301–413 (1999)
8. Bertin, K., Rivoirard, V.: Maxiset in sup-norm for kernel estimators. Test (2009). Doi:10.1007/s11749-008-0109-7
9. Birgé, L., Massart, P.: An adaptive compression algorithm in Besov spaces. Constr. Approx. **16**(1), 1–36 (2000)
10. Birgé, L., Massart, P.: Gaussian model selection. J. Eur. Math. Soc. (JEMS) **3**(3), 203–268 (2001)
11. Birgé, L., Massart, P.: Minimal penalties for Gaussian model selection. Probab. Theory Relat. Fields **138**(1–2), 33–73 (2007)
12. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: a survey of some recent advances. ESAIM Probab. Stat. **9**, 323–375 (2005)
13. Cohen, A., DeVore, R.A., Kerkyacharian, G., Picard, D.: Maximal spaces with given rate of convergence for thresholding algorithms. Appl. Comput. Harmon. Anal. **11**(2), 167–191 (2001)

14. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
15. Kerkyacharian, G., Picard, D.: Thresholding algorithms, maxisets and well-concentrated bases. Test **9**(2), 283–344 (2000)
16. Loubes, J.-M., Ludeña, C.: Adaptive complexity regularization for linear inverse problems. Electron. J. Stat. **2**, 661–677 (2008)
17. Loubes, J.-M., Ludeña, C.: Penalized estimators for non linear inverse problems. ESAIM Probab. Stat. (2008). DOI:10.1051/ps:2008024
18. Mallows, C.L.: Some comments on C_p. Technometrics **15**, 661–675 (1973)
19. Massart, P.: Concentration inequalities and model selection. In: Lectures on Probability Theory and Statistics, Saint-Flour, 2003. Lecture Notes in Math., vol. 1896. Springer, Berlin (2007)
20. Meyer, Y.: Ondelettes et opérateurs. I. Hermann, Paris (1990)
21. Rivoirard, V.: Maxisets for linear procedures. Stat. Probab. Lett. **67**(3), 267–275 (2004)
22. Rivoirard, V.: Bayesian modeling of sparse sequences and maxisets for Bayes rules. Math. Methods Stat. **14**(3), 346–376 (2005)
23. Rivoirard, V., Tribouley, K.: The maxiset point of view for estimating integrated quadratic functionals. Stat. Sin. **18**(1), 255–279 (2008)
24. Nussbaum, M.: Asymptotic equivalence of density estimation and Gaussian white noise. Ann. Stat. **24**(6), 2399–2430 (1996)