

Is model selection possible for the ℓ_p -loss? PCO estimation for regression models

Claire Lacour^{1,a} , Pascal Massart^{2,b} and Vincent Rivoirard^{3,2,c} 

¹Univ Gustave Eiffel, Univ Paris Est Creteil, CNRS, LAMA UMR8050, F-77447, Marne-la-Vallée, France, ^aclaire.lacour@univ-eiffel.fr

²Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France, ^bpascal.massart@universite-paris-saclay.fr

³CEREMADE, CNRS, Université Paris-Dauphine, Université PSL, 75016 Paris, France, ^cvincent.rivoirard@dauphine.fr

Abstract. This paper addresses the problem of model selection in the sequence model $Y = \theta + \varepsilon\xi$, when ξ is sub-Gaussian, for non-Euclidean loss functions. In this model, the Penalized Comparison to Overfitting procedure is studied for the ℓ_p -loss, $p \geq 1$. Several oracle inequalities are derived from concentration inequalities for sub-Weibull variables. Using judicious collections of models and penalty terms, minimax rates of convergence are established for Besov bodies $\mathcal{B}_{r,\infty}^s$. These results are applied to the functional model of nonparametric regression.

Résumé. Cet article traite du problème de la sélection de modèles dans le modèle de suite $Y = \theta + \varepsilon\xi$, lorsque ξ est sous-gaussien, pour des fonctions de perte non euclidiennes. Dans ce modèle, on étudie la procédure PCO (comparaison pénalisée à l'estimateur surajusté) pour la perte ℓ_p , $p \geq 1$. Plusieurs inégalités oracles sont déduites d'inégalités de concentration pour des variables sous-Weibull. En choisissant judicieusement les collections de modèles et les termes de pénalité, on établit des vitesses de convergence minimax pour les espaces de Besov $\mathcal{B}_{r,\infty}^s$. Ces résultats sont appliqués au modèle fonctionnel de régression non paramétrique.

MSC2020 subject classifications: Primary 62G05; 62C20; secondary 62G08; 60E15

Keywords: Model selection; Oracle inequalities; Minimax rates; ℓ_p -loss; Sub-Gaussian sequence model; Nonparametric regression

1. Introduction

The problem of selecting a model from among several candidates is essential in statistics and machine learning, as well as in many application fields. In the most general sense, the aim of model selection is to construct data-driven criteria for selecting a model m from a given collection $\mathcal{M}$. In other words, if one observes some random variable $\xi^{(n)}$ (which can typically be a random vector of size n) with unknown distribution depending on some quantity f (the target) belonging to a set $\mathcal{S}$, a flexible approach to estimate f is to consider some collection of preliminary estimators $(\hat{f}_m)_{m \in \mathcal{M}}$ and then try to design some genuine data-driven procedure $\hat{m} \in \mathcal{M}$ to produce a new estimator $\hat{f}_{\hat{m}}$. Considering some loss function ℓ , we measure the quality of each estimator $\hat{f}_m$, through the quantity $\ell(f, \hat{f}_m)$ and mathematical results on estimator selection are formulated in terms of upper bounds on $\ell(f, \hat{f}_{\hat{m}})$ that allow us to measure how far this quantity is from what is usually called the *oracle risk* $\inf_{m \in \mathcal{M}} \mathbb{E} \ell(f, \hat{f}_m)$. These comparison inequalities are called oracle inequalities. The quadratic loss $\ell(f, \hat{f}_m) = \|f - \hat{f}_m\|^2$ is a standard choice, but of course other losses are considered in the literature.

Actually, in the classical model selection framework, developed and popularized by Birgé and Massart [5], [6], [41], the list of estimators and the loss function are intimately related in the sense that they derive from the same *contrast function* (also called *empirical risk* in the machine learning literature). More precisely, a contrast function L_n is a function on the set $\mathcal{S}$ depending on the observation $\xi^{(n)}$ in such a way that

$$g \in \mathcal{S} \mapsto \mathbb{E}[L_n(g)]$$

achieves a minimum at point f . Given some collection of subsets $(S_m)_{m \in \mathcal{M}}$ of $\mathcal{S}$, called models in the sequel, for every $m \in \mathcal{M}$, some estimator $\hat{f}_m$ of f is obtained by minimizing L_n over S_m ; $\hat{f}_m$ is then called *minimum contrast estimator* or

empirical risk minimizer. In the case where $\xi^{(n)} = (\xi_1, \dots, \xi_n)$, an empirical criterion L_n can be defined as an empirical mean

$$L_n(g) = P_n[L(g, \cdot)] := \frac{1}{n} \sum_{i=1}^n L(g, \xi_i),$$

which justifies the terminology of empirical risk. Empirical risk minimization includes maximum likelihood and least squares estimation. The penalized empirical risk selection procedure consists of considering some proper penalty function $\text{pen}: \mathcal{M} \rightarrow \mathbb{R}_+$ and taking $\hat{m}$ minimizing

$$(1) \quad L_n(\hat{f}_m) + \text{pen}(m)$$

over $\mathcal{M}$. We can then define the selected model $S_{\hat{m}}$ and the corresponding selected estimator $\hat{f}_{\hat{m}}$. Penalized criteria have been proposed in the early seventies by Akaike or Schwarz (see [2] and [44]) for penalized maximum log-likelihood in the density estimation framework and Mallows for penalized least squares regression (see [40]). In both cases the penalty functions are proportional to the number of parameters D_m of the corresponding model S_m . In such settings, the performance of model selection estimators can be studied for the “natural” (nonnegative) loss function ℓ attached to L_n through the simple definition

$$(2) \quad \ell(f, g) = \mathbb{E}[L_n(g)] - \mathbb{E}[L_n(f)], \quad g \in \mathcal{S}.$$

This approach has been successfully applied in many settings: density estimation or Poisson intensity estimation where ℓ is the Kullback-Leibler divergence or the $\mathbb{L}_2$ -loss, nonparametric regression for the $\mathbb{L}_2$ -loss, binary classification for the 0-1 loss or the Gaussian white noise model still for the $\mathbb{L}_2$ -loss. See [41] and references therein. Although it is not derived from an empirical contrast, we mention the model selection work associated with the Hellinger distance [4], [3]. This explains why, although successfully applied in many settings, model selection procedures have only been defined and analyzed for very specific loss functions. In particular, non-Euclidean loss functions, such as $\mathbb{L}_p$ -losses, have rarely been considered for model selection procedures. It is not the case for other classical nonparametric estimation procedures.

Although the use of $\mathbb{L}_p$ -loss in nonparametric statistics goes back at least to [7, 27, 45], an important advance was made by A. Nemirovsky [42] who established optimal rates of convergence in the problem of multivariate nonparametric regression for $\mathbb{L}_p$ -loss and functions belonging to $\mathbb{L}_q$ -Sobolev classes with possibly $p \neq q$. Some years later, fundamental work was carried out by O. Lepski, who proposed the well-known Lepski method for selecting a bandwidth of a kernel estimator. In [38], he studied adaptive estimation under $\mathbb{L}_p$ -loss, $1 \leq p \leq \infty$ over the collection of Hölder classes. Then [39] introduced a local bandwidth selection scheme to give kernel estimates which achieve optimal rates of convergence over Besov classes in the Gaussian white noise model, the anisotropic multivariate case being examined in [31]. Then Goldenshluger and Lepski [20] developed a powerful methodology for selecting a bandwidth of a kernel estimate, which works in a multitude of contexts and allows to establish oracle inequalities and to derive optimal rates of convergence: see [21–23]. In the same line [37] addressed the problem of adaptive minimax estimation in white Gaussian noise models on the anisotropic Nikol’skii classes. Note that the estimation under the $\mathbb{L}_1$ -loss for bandwidth-selected kernel estimators has been investigated in [11], and the $\mathbb{L}_p$ -aggregation of estimators has been studied in [19]. These works, whose list is not exhaustive, show that, unlike penalization methods, the use of non-Hilbertian norms for the class of methods based on the comparison of estimators has been widely studied.

In a very different spirit, another very popular method for adaptive estimation of functions is wavelet thresholding. In a series of papers, Donoho, Johnstone, Kerkycharian and Picard have shown the power of wavelet thresholding for $\mathbb{L}_p$ -estimation over the scale of Besov classes: see [14], [15], [16], [12]. Refinements have been proposed in [30], [33], [24], [29], to name but a few. It must be noted that these thresholding methods generally suffer from logarithmic losses in the rates of convergence. We refer the reader to the book [25] which details the construction of wavelet bases, their use in statistical estimation, $\mathbb{L}_p$ -minimax results as well as computational aspects. Johnstone’s recent book [28] addresses nonparametric function estimation by carefully studying the infinite Gaussian sequence model, with many results and thoughts about wavelet thresholding and adaptive minimaxity over ellipsoids.

We finally mention the forthcoming manuscript [1] which considers Bayesian nonparametric concentration rates of procedures based on Heavy-tailed and Horseshoe priors for ℓ_p -norms in the Gaussian white noise model.

The main goal of this paper is then to answer the following natural question: Is it possible to design a model selection procedure in the same spirit as in (1), and in particular resulting from one minimization step, so that it achieves optimal theoretical performance for non-Euclidean losses? In particular, we have in mind that classical model selection procedures are able to achieve optimal oracle properties and sharp minimax rates of convergence on classical functional spaces for Euclidean losses. We tackle this issue by considering the classical infinite sequence model and studying a specific model

selection procedure, called *PCO*, for ℓ_p -loss functions, with $1 \leq p < \infty$. As explained in [28] (see his Preface and Section 1.5), the Gaussian sequence model "captures many of the conceptual issues associated with non-parametric estimation".

1.1. The PCO estimation procedure for the sub-Gaussian sequence model

For Λ a countable set, we consider the following classical sequence model:

$$(3) \quad Y_\lambda = \theta_\lambda + \varepsilon \xi_\lambda, \quad \lambda \in \Lambda.$$

In this model, the noise level ε is assumed to be smaller than a constant, say 1, and $\varepsilon \rightarrow 0$ defines the asymptotic setting of our study. The ξ_λ 's are i.i.d. centered variables, satisfying the sub-Gaussian property, i.e.

$$(4) \quad \mathbb{E}(\exp(u\xi_\lambda)) \leq \exp(u^2/2), \quad \forall u \in \mathbb{R}.$$

The property (4) entails the following concentration inequality

$$\mathbb{P}(|\xi_\lambda| \geq t) \leq 2 \exp(-t^2/2), \quad \forall t \geq 0,$$

as well as bounds on $\mathbb{E}|\xi|^p$, see Lemma 5.1. This corresponds to the standard Gaussian tail decay, which justifies our choice of scaling parameter.

We aim at estimating the sequence $\theta = (\theta_\lambda)_{\lambda \in \Lambda}$ by using a finite number of observations, say $(Y_\lambda)_{\lambda \in \Lambda^{(N)}}$ where $\Lambda^{(N)}$ denotes the first N elements of Λ . The integer N may increase as ε decreases, so we actually face a nonparametric problem. To connect our setting with nonparametric regression, we may have in mind that $N \propto \varepsilon^{-2}$ (see [14, 30, 32] and Sections 3.2 and 4) but, unless specified, our results hold for any N .

In Model (3), for $1 \leq p < \infty$, we consider the risk associated with the standard ℓ_p -norm on $\mathbb{R}^\Lambda$. We assume in the sequel that $\|\theta\|_{\ell_p} < \infty$. For each subset m of $\Lambda^{(N)}$, called model in the sequel, we set

$$\hat{\theta}^{(m)} = (Y_\lambda \mathbb{1}_{\{\lambda \in m\}})_{\lambda \in \Lambda}.$$

In particular, $\hat{\theta}_\lambda^{(m)} = 0$ for $\lambda \notin \Lambda^{(N)}$. Given a collection of models $\mathcal{M} \subset \mathcal{P}(\Lambda^{(N)})$, we wish to select $\hat{m} \in \mathcal{M}$ in the best possible way. For this purpose, as explained previously, we rely on the *PCO criterion* introduced in [35, 46]. The heuristic of this approach is to build the goodness of fit criterion by using the estimator which has the smallest bias among the collection $(\hat{\theta}^{(m)})_{m \in \mathcal{M}}$. In our setting, it means that we have to consider $\hat{\theta}^{(\Lambda^{(N)})}$ (see (9)). Then, as is standard in the nonparametric setting, we introduce a penalty term and define

$$\hat{m} = \arg \min_{m \in \mathcal{M}} \left\{ \|\hat{\theta}^{(m)} - \hat{\theta}^{(\Lambda^{(N)})}\|_{\ell_p}^p + \text{pen}(m) \right\}$$

and estimate θ by

$$\tilde{\theta} = \hat{\theta}^{(\hat{m})}.$$

The idea of this methodology is to use $\|\hat{\theta}^{(m)} - \hat{\theta}^{(\Lambda^{(N)})}\|_{\ell_p}^p$ as a preliminary estimator of the bias of $\hat{\theta}^{(m)}$. The role of $\text{pen}(m)$ is then twofold: adjusting this preliminary step and taking into account the variance of $\hat{\theta}^{(m)}$. The estimator $\tilde{\theta}$ will be called the *Penalized Comparison to Overfitting* (abbreviated as *PCO*) in the sequel. This terminology is justified by the overfitting properties of $\hat{\theta}^{(\Lambda^{(N)})}$. Of course, setting for $m \in \mathcal{M}$,

$$(5) \quad \text{Crit}(m) = - \sum_{\lambda \in m} |Y_\lambda|^p + \text{pen}(m),$$

we obtain

$$\hat{m} = \arg \min_{m \in \mathcal{M}} \text{Crit}(m).$$

The heuristic of this approach thus differs from the classical approach based on the contrast function. However, note that for $p = 2$, the criterion function Crit corresponds to the one used in regression for well-known criteria such as Mallows's C_p [40], AIC [2] or BIC [44] for instance. Two remarks are in order: Unlike Lepski-type procedures, the derivation of the

PCO estimate $\tilde{\theta}$ involves only a single minimization step, so its implementation is more straightforward. Furthermore, if we take $\text{pen}(m)$ in the form

$$(6) \quad \text{pen}(m) = \text{card}(m)t^p,$$

for some $t > 0$, we have:

$$\hat{m} = \{\lambda \in \Lambda^{(N)} : |Y_\lambda| > t\}$$

and the PCO estimator reduces to the thresholding estimator with threshold t :

$$\tilde{\theta}_\lambda = \begin{cases} Y_\lambda \mathbb{1}_{\{|Y_\lambda| > t\}} & \text{if } \lambda \in \Lambda^{(N)}, \\ 0 & \text{if } \lambda \notin \Lambda^{(N)}. \end{cases}$$

The results of Sections 2, 3.2, and 4 show that the definition of $\text{pen}(m)$ in (6) must be refined to achieve optimal performance. This refinement is described in the next subsection.

1.2. Contributions

The main contribution of this paper is to show that, under a suitable choice of $\text{pen}(m)$, the PCO estimate, which is based on the simple ℓ_p -criterion Crit and whose computational cost is reasonable, is able to achieve optimal results in oracle and minimax settings for any value of $p \in [1, +\infty)$.

We first prove in Proposition 2.1 that $\tilde{\theta}$ satisfies a very general oracle inequality for ℓ_p -losses, regardless of the specific form of the penalty term $\text{pen}(m)$. To this end, we analyse the behavior of bias and variance terms of any estimator $\hat{\theta}^{(m)}$. This first result guides the choice of $\text{pen}(m)$, leading to a more specific oracle inequality established in Theorem 2.5. As usual, these results depend on sharp concentration inequalities. In our case, the argument is based on Theorem 2.3, which establishes sharp concentration for quantities of the form $\sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p$ around their expectation. The resulting tail bound consists of two terms: a classical quadratic term, proportional to $\sqrt{x}$, and a second term proportional to $x^{p/2}$. The second term is linear for $p = 2$ but it raises many technical difficulties otherwise, in particular for $p > 2$. This result highlights an "elbow phenomenon", depending on whether p exceeds 2 or not. We take into account this elbow to propose a refined penalty function $\text{pen}(m)$ for the case $p > 2$ in Theorem 2.8. This result allows us to deal with very large collections of models $\mathcal{M}$ when $p > 2$, which is crucial for the minimax setting. In Section 2.4, we extend these results to the case of weighted ℓ_p -norms, where the loss function is defined by

$$(7) \quad \|\vartheta\|_{\ell_p(w)}^p = \sum_{\lambda \in \Lambda} w_\lambda |\vartheta_\lambda|^p, \quad \vartheta \in \mathbb{R}^\Lambda,$$

with $w = (w_\lambda)_{\lambda \in \Lambda}$ denoting a sequence of positive weights.

Minimax rates of convergence for the estimate $\tilde{\theta}$ are first derived when the tails of sequences θ decay polynomially. We then consider the classical class of Besov bodies $\mathcal{B}_{r,\infty}^s(R)$ and study the rates for any $r \geq 1$, under minimax smoothness assumptions on s , for a suitable weighted ℓ_p -risk. We prove in Theorem 3.2 that for $p \leq 2$, $\tilde{\theta}$ is optimal under an appropriate choice of the penalty function. For $p > 2$, $\tilde{\theta}$ is also optimal if $r \geq p$ or if $r \leq p/(2s+1)$. In the case $p > 2$ and $p/(2s+1) < r < p$, the upper bound differs from the lower bound by a logarithmic factor. To deal with the case $r < p$, we need to consider a very large collection of models, in particular if $r \leq p/(2s+1)$. The last contribution of our paper consists in extending these last results to the functional framework. In Section 4, we consider the nonparametric regression model

$$X_i = f\left(\frac{i}{n}\right) + \sigma\eta_i, \quad 1 \leq i \leq n,$$

(see Model (29)) and propose a PCO estimate of the function f based on wavelet representations. The generalization of Theorem 3.2 yields Theorem 4.1, which provides rates for our procedure on functional Besov spaces under the standard functional $\mathbb{L}_p$ -loss and makes it possible to discuss optimality. The conclusions are similar to those obtained in the sequential setting. We also present the main steps of the methodology used to derive functional minimax rates, which are of independent interest.

1.3. Plan of the paper and notation

The paper is organized as follows. Section 2 is devoted to oracle results and the statement of concentration inequalities used in this paper. Section 3 presents the minimax rates of convergence achieved by the PCO estimator. Section 4 is devoted to the nonparametric regression model. Finally, Section 5 presents the proofs of the results.

For any set A we denote by $|A|$ the cardinality of A , and $\mathcal{P}(A)$ the set of its subsets. The notation $\mathbb{N}$ stands for the set of nonnegative integers: $\mathbb{N} = \{0, 1, 2, \dots\}$. We write $u_\varepsilon \lesssim v_\varepsilon$ if there exists a constant $0 < A < \infty$ such that $u_\varepsilon \leq Av_\varepsilon$ for all $\varepsilon > 0$. When $u_\varepsilon \lesssim v_\varepsilon$ and $v_\varepsilon \lesssim u_\varepsilon$ we write $u_\varepsilon \approx v_\varepsilon$. Remember that, for $1 \leq p \leq \infty$, $\ell_p(w)$ denotes the weighted ℓ_p -norm, defined in (7). When all weights w_λ are equal to 1, we use the classical notation ℓ_p instead of $\ell_p(w)$ and for short, we set $\|\cdot\|_p = \|\cdot\|_{\ell_p}$. Finally, the norm on the Banach space $\mathbb{L}_p(\mathbb{R})$ will be denoted by $\|\cdot\|_{\mathbb{L}_p}$.

2. Oracle approach and concentration inequalities

Given any $p \geq 1$, the goal of this section is to provide some optimality results in the oracle setting in Model (3). In particular, in the sequel, except in Theorem 2.3 (for the first point), we assume that the ξ_λ 's are i.i.d. centered sub-Gaussian variables. From Corollary 2.4, we will also assume that the distribution of the variables ξ_λ is not the Dirac mass at 0. Throughout this section, we denote

$$\sigma_q := (\mathbb{E}[|\xi_\lambda|^q])^{1/q}, \quad 1 \leq q < \infty.$$

Note that combining Assumption (4) and Lemma 5.1(ii), we have:

$$(8) \quad \sigma_q \leq 5\sqrt{q/2}.$$

For the ℓ_p -risk, we first establish, in Proposition 2.1, a general result that holds for an arbitrary penalty function. By emphasizing the central role played by sums of the form $\sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p$, Proposition 2.1 identifies the appropriate choice of the penalty term $\text{pen}(m)$. The concentration properties of these sums around their expectation are analyzed in detail in Theorem 2.3 and Corollary 2.4. Building on these results, we refine Proposition 2.1 and derive sharp oracle inequalities in Theorems 2.5 and 2.8. Extensions of these results for the $\ell_p(w)$ -risk are established in Section 2.4.

Before stating these results, we first observe that for any model m , we can easily compute the ℓ_p -distance between the estimator $\hat{\theta}^{(m)}$ and θ . Indeed, we have for $m \subset \Lambda^{(N)}$:

$$\|\hat{\theta}^{(m)} - \theta\|_p^p = \sum_{\lambda \in m} |Y_\lambda - \theta_\lambda|^p + \sum_{\lambda \notin m} |0 - \theta_\lambda|^p = V_p(m) + B_p(m),$$

with

$$(9) \quad B_p(m) := \sum_{\lambda \in \Lambda \setminus m} |\theta_\lambda|^p \quad \text{and} \quad V_p(m) := \varepsilon^p \sum_{\lambda \in m} |\xi_\lambda|^p.$$

In the last decomposition, $B_p(m)$ (resp. $V_p(m)$) can be viewed as an ℓ_p -bias term (resp. an ℓ_p -variance term). We now study $\tilde{\theta} = \hat{\theta}^{(\hat{m})}$ in the oracle setting.

2.1. A general oracle inequality

Recall that $\tilde{\theta} = \hat{\theta}^{(\hat{m})}$ with

$$\hat{\theta}^{(m)} = (Y_\lambda \mathbb{1}_{\{\lambda \in m\}})_{\lambda \in \Lambda}, \quad \hat{m} = \arg \min_{m \in \mathcal{M}} \left\{ - \sum_{\lambda \in m} |Y_\lambda|^p + \text{pen}(m) \right\}.$$

We obtain the following general oracle inequality.

Proposition 2.1. *If $p > 1$, for any $\alpha \in (0, 2)$, we have for any arbitrary $m \in \mathcal{M}$:*

$$\|\tilde{\theta} - \theta\|_p^p \leq M_{p,\alpha} \|\hat{\theta}^{(m)} - \theta\|_p^p + \frac{2}{\alpha} \left[(1 + \alpha) V_p(\hat{m}) - \text{pen}(\hat{m}) \right] - \frac{2}{\alpha} \left[(1 + \alpha) V_p(m) - \text{pen}(m) \right],$$

where $M_{p,\alpha}$ depends only on p and α . In particular, with $\alpha = 1$, we obtain:

$$(10) \quad \|\tilde{\theta} - \theta\|_p^p \leq M_p \|\hat{\theta}^{(m)} - \theta\|_p^p + 2 \left[2V_p(\hat{m}) - \text{pen}(\hat{m}) \right] - 2 \left[2V_p(m) - \text{pen}(m) \right],$$

where $M_p = M_{p,1}$ is given in Equation (36) in Section 5.2.2. If $p = 1$, the previous inequalities remain provided that the factor $\frac{2}{\alpha}$ is replaced by $\frac{1}{\alpha \wedge 1}$ in the right hand side of the first inequality.

The proof of Proposition 2.1 is provided in Section 5.2.2.

Remark 2.2. The expression of $M_{p,\alpha}$ is rather intricate (see the proof of Proposition 2.1) and the optimal choice of α depends on p . Nevertheless, numerical investigations indicate that $M_{p,1}$ is close to the minimum of the function $\alpha \mapsto M_{p,\alpha}$, and is in fact optimal when $p = 1$ and $p = 2$. For this reason, we focus on the case $\alpha = 1$ and state Inequality (10).

In light of the first result of Proposition 2.1, the optimality of the estimator $\tilde{\theta}$ among the collection $(\hat{\theta}^{(m)})_{m \in \mathcal{M}}$ will be achieved if we can construct a penalty term $\text{pen}(m)$ such that, for some constant $\alpha \in (0, 2)$, $\text{pen}(m)$ is close to $(1 + \alpha)V_p(m)$ for all $m \in \mathcal{M}$. However, $V_p(m)$ is not observable. Consequently, concentration inequalities are required to derive a suitable expression for $\text{pen}(m)$ within our procedure. Since

$$V_p(m) = \varepsilon^p \sum_{\lambda \in m} |\xi_\lambda|^p,$$

this amounts to analyzing sums of the form $\sum_{\lambda} |\xi_\lambda|^p$, where the ξ_λ 's are independent sub-Gaussian random variables.

2.2. Concentration inequalities

Let $\mathcal{I} \subset \Lambda$. We denote $D = \text{card}(\mathcal{I})$ and

$$Z := \sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p.$$

Using, for $r > 0$, the Orlicz norm of a random variable X , defined by

$$\|X\|_{\psi_r} = \inf \left\{ \eta > 0 : \mathbb{E} \left[\exp \left((|X|/\eta)^r \right) \right] \leq 2 \right\},$$

we set $b_\lambda = \left\| |\xi_\lambda|^p - \mathbb{E}|\xi_\lambda|^p \right\|_{\psi_{2/p}}$. We have the following result.

Theorem 2.3. *Let $p \geq 1$. Assume that the ξ_λ 's are centered independent sub-Gaussian variables. There exist positive constants $d_{1,p}$ and $d_{2,p}$ only depending on p such that, for any $x > 0$,*

$$\mathbb{P} \left(|Z - \mathbb{E}(Z)| \geq d_{1,p} \|b\|_{\ell_2} \sqrt{x} + d_{2,p} \|b\|_{\ell_{1/(1-p/2)+}} x^{p/2} \right) \leq 2e^{-x}.$$

Moreover, if the random variables ξ_λ are identically distributed, then for every $x > 0$,

$$\mathbb{P} \left(|Z - \mathbb{E}[Z]| \geq c_{1,p} \sqrt{Dx} + c_{2,p} D^{(1-p/2)+} x^{p/2} \right) \leq 2e^{-x},$$

where the constants $c_{1,p}$ and $c_{2,p}$ depend only on p and on $\|\xi_\lambda\|_{\psi_2}$. More precisely,

$$c_{i,p} = d_{i,p} K_{2/p} \|\xi_\lambda\|_{\psi_2}^p, \quad i = 1, 2,$$

with

$$K_{2/p} = 1 + p(\log 2)^{-p/2} \Gamma(p/2), \quad d_{1,p} = 2eC(p), \quad d_{2,p} = \sqrt{2}e2^p (C(p)\mathbb{1}_{p>2} + 4e\mathbb{1}_{p \leq 2})$$

and

$$C(p) = \left[\sqrt{8}e^3 (2\pi)^{1/4} e^{1/24} (e^{2/e} p)^{p/2} \right] \mathbb{1}_{p>2} + \left[2^{p/2} (4e + 2(\log 2)^{p/2}) \right] \mathbb{1}_{p \leq 2}.$$

This result follows from concentration inequalities for sub-Weibull variables. Indeed, setting $X_\lambda = |\xi_\lambda|^p - \mathbb{E}|\xi_\lambda|^p$, the sub-Gaussianity property of ξ_λ implies that X_λ exhibits Weibull-type tail behavior:

$$\mathbb{P}(|X_\lambda| \geq t) \leq C \exp(-ct^{2/p}),$$

for some positive constant C . Random variables whose Orlicz norm $\|\cdot\|_{\psi_r}$ is finite are referred to as sub-Weibull variables. In the present setting, the Weibull parameter is $r = 2/p \leq 2$. Theorem 2.3 is then a direct consequence of Theorem 1 in [48] or Theorem 3.1 in [34] (see also Equation (3.6) therein), which provide explicit expressions for the constants $d_{1,p}$ and $d_{2,p}$. In the i.i.d. case, the quantities b_λ are identical and can be bounded by $\|\xi_\lambda\|^p_{\psi_{2/p}} = \|\xi_\lambda\|^p_{\psi_2}$, up to the multiplicative constant $K_{2/p}$ (see Lemma 5.3 and its proof).

In the Gaussian i.i.d. case, sharper constants are available for specific values of p . When $p = 1$, the Cirelson-Ibragimov-Sudakov inequality yields the bound with $c_{1,1} = \sqrt{2}$ and $c_{2,1} = 0$ (see Theorem 3.4 in [41]). For $p = 2$, one recovers the classical concentration inequality for chi-square variables, for which $c_{1,2} = c_{2,2} = 2$ is valid (see Lemma 1 in [36]). This is also consistent with the Hanson-Wright inequality in the case of the identity matrix, or with Bernstein-type inequalities for sub-exponential variables (see, e.g., [47]).

Observe that Theorem 2.3 can also be derived from earlier results, such as the moment inequalities established in [18] for the case $r > 1$. In particular, their corollary shows that our bound cannot be improved when $p < 2$. For $r < 1$, Theorem 6.2 of [26] provides a two-sided moment inequality that yields our tail estimate. Their bound cannot be improved: when the ξ_λ 's are Gaussian, the upper bound is achieved up to a multiplicative constant. Similar optimality considerations are discussed in [34].

From Theorem 2.3, we obtain the following corollary.

Corollary 2.4. *Let $p \geq 1$. Assume that the ξ_λ 's are i.i.d. centered sub-Gaussian variables. For any $x \geq 1$, with probability larger than $1 - 2\exp(-x)$,*

$$(11) \quad \frac{1}{2}\sigma_p^p D - \kappa_p D^{(1-\frac{p}{2})+x^{\frac{p}{2}}} \leq \sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p \leq \frac{3}{2}\sigma_p^p D + \kappa_p D^{(1-\frac{p}{2})+x^{\frac{p}{2}}},$$

where $\kappa_p = c_{2,p} + c_{1,p} \max(1, c_{1,p}/(2\sigma_p^p))$ is a positive constant only depending on p and σ_p and $\|\xi_\lambda\|_{\psi_2}$.

Proof. We obviously have $\mathbb{E}[Z] = D\sigma_p^p$. If $p \geq 2$, we use $2\sqrt{Dx} \leq \theta D + \theta^{-1}x$ with $\theta = \sigma_p^p/c_{1,p}$, and the inequality $x \leq x^{p/2}$. If $p < 2$ and $x \leq D$, we use the same bound for $\sqrt{Dx}$ with the same choice of θ , and use instead the inequality $x \leq D^{1-p/2}x^{p/2}$. Finally, if $p < 2$ and $x > D$, we directly write $\sqrt{Dx} \leq D^{1-p/2}x^{p/2}$. $\square$

Using (4) together with Lemma 5.1, we obtain

$$\kappa_p = K_{2/p} 6^{p/2} \left(d_{2,p} + d_{1,p} \max \left(1, d_{1,p} K_{2/p} \frac{6^{p/2}}{2\sigma_p^p} \right) \right).$$

The constants $d_{1,p}$, $d_{2,p}$ and $K_{2/p}$ are specified in Theorem 2.3 and depend only on p . Note that Corollary 2.4 remains valid if the constants $1/2$ and $3/2$ in (11) are replaced by $1 - \epsilon$ and $1 + \epsilon$, respectively, for any arbitrarily small $\epsilon > 0$. In this case, κ_p depends also on $\epsilon > 0$.

Using (8), the proof of Corollary 2.4 shows that for any $x \geq 1$, with probability larger than $1 - 2\exp(-x)$,

$$(12) \quad \frac{1}{2}(5\sqrt{p/2})^p D - \tilde{\kappa}_p D^{(1-\frac{p}{2})+x^{\frac{p}{2}}} \leq \sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p \leq \frac{3}{2}(5\sqrt{p/2})^p D + \tilde{\kappa}_p D^{(1-\frac{p}{2})+x^{\frac{p}{2}}},$$

with

$$(13) \quad \tilde{\kappa}_p = K_{2/p} 6^{p/2} \left(d_{2,p} + d_{1,p} \max \left(1, d_{1,p} K_{2/p} \frac{6^{p/2}}{2(5\sqrt{p/2})^p} \right) \right).$$

Using the explicit expressions of the constants $d_{1,p}$, $d_{2,p}$ and $K_{2/p}$ given in Theorem 2.3, we deduce that $\tilde{\kappa}_p$ only depends on p . In particular, it is independent of both σ_p and $\|\xi_\lambda\|_{\psi_2}$ and therefore does not depend on the distribution of the variables ξ_λ .

2.3. Refined oracle inequalities

In this section, we apply the general oracle inequality for losses of Proposition 2.1 with $\alpha = 1$ (see Remark 2.2) and use sharp concentration inequalities of Corollary 2.4 to derive suitable penalties. In view of Inequality (10) and Corollary 2.4, we take, for any $m \in \mathcal{M}$,

$$(14) \quad \text{pen}(m) := 2\varepsilon^p p(m),$$

with, for some $x_m \geq 1$,

$$(15) \quad p(m) = \frac{3}{2} \sigma_p^p |m| + \kappa_p 2^{\frac{(p-2)+}{2}} |m|^{(1-\frac{p}{2})_+} x_m^{\frac{p}{2}}.$$

Observe that:

$$p(m) = \begin{cases} \frac{3}{2} \sigma_p^p |m| + \kappa_p |m|^{1-\frac{p}{2}} x_m^{\frac{p}{2}} & \text{if } p \leq 2, \\ \frac{3}{2} \sigma_p^p |m| + \kappa_p 2^{\frac{p}{2}-1} x_m^{\frac{p}{2}} & \text{if } p \geq 2. \end{cases}$$

The factor x_m , depending on m , will be specified later. But note that, mimicking the computations of Section 1.1, the thresholding rule corresponds to the case where $p(m)$ is proportional to $|m|$, which is obtained for instance by taking x_m proportional to $|m|$ when $p \leq 2$ and in this case the threshold is proportional to the noise level ε as expected. However, subsequent results show that the resulting estimate is suboptimal in some situations by at least a logarithmic factor. When $p = 2$, $\text{pen}(m)$ corresponds to the penalty extensively used to derive oracle inequalities for model selection procedures on Hilbert spaces. See, for instance, oracle inequalities established in Theorems 4.2 and 4.18 of [41]. The extension of these results to the case $p \neq 2$ is provided by the following theorem.

Theorem 2.5. *Let $p \geq 1$ and let $(x_m)_{m \in \mathcal{M}}$ be a sequence of factors such that $x_m \geq 1$ for all $m \in \mathcal{M}$. We then consider the estimate $\tilde{\theta} = \hat{\theta}^{(\tilde{m})}$ associated with the penalty defined in (14) and (15). We have:*

$$(16) \quad \mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \leq \tilde{M}_p \inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right] + \text{pen}(m) \right\} + \check{M}_p \kappa_p \varepsilon^p R(\mathcal{M}),$$

with

$$(17) \quad R(\mathcal{M}) = \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} |m|^{(1-\frac{p}{2})_+} e^{-x_m}$$

and $\tilde{M}_p$ and $\check{M}_p$ are two constants only depending on p .

The proof of Theorem 2.5 is provided in Section 5.2.3.

Remark 2.6. Regarding the sub-Gaussian error distribution assumption. The statement of the concentration inequality we use (namely Corollary 2.4) might suggest that it is possible, if not to dispense with it entirely, then at least to relax it. That said, even in the case of the quadratic criterion, this assumption is minimal if one wishes to obtain general oracle inequalities without loss compared to the Gaussian case. One could obviously consider improvements in certain specific configurations within the family of models (as in [17]). However, it is proven in the same article that relaxing the integrability assumption on the errors can lead to results that are significantly worse than in the Gaussian case. Here, we have chosen to give priority to obtaining a general result valid for all possible model configurations without any degradation compared to the Gaussian case. This is what led us to formulate this sub-Gaussianity assumption on the errors.

Remark 2.7. Theorem 2.5 remains true if we replace the equality in (14) by the inequality:

$$\text{pen}(m) \geq 2\varepsilon^p p(m).$$

Furthermore, using (12) and (13), it is possible to choose an alternative penalty $\text{pen}(m)$ so that it does not depend on the distribution of the variables ξ_λ .

We now analyze the result of Theorem 2.5 in the asymptotic setting where $\varepsilon \rightarrow 0$ and $N \rightarrow +\infty$. The parameter σ_p , which is bounded (see (8)) is assumed to be positive and viewed as a constant.

First, let us discuss the choice of the factors x_m . We choose them so that $\tilde{\theta}$ performs well in the oracle setting, meaning that

$$(18) \quad \mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \leq C(\theta) \inf_{m \in \mathcal{M}} \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right],$$

with $C(\theta)$ remaining suitably controlled uniformly over θ by a quantity that is not too large. Now, observe that for any model m ,

$$\mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right] = B_p(m) + \mathbb{E} [V_p(m)] = \sum_{\lambda \notin m} |\theta_\lambda|^p + \varepsilon^p \sigma_p^p |m|$$

and $\text{pen}(m)$ can be compared to the second term of the right-hand side. In particular, Theorem 2.5 shows that $\tilde{\theta}$ performs well as soon as we have, for some constant Q_1 ,

$$(19) \quad |m|^{\left(1-\frac{p}{2}\right)+x_m^{\frac{p}{2}}} \leq Q_1 |m| \iff x_m \leq (Q_1 |m|)^{2/\max(2,p)}, \quad \text{for all } m \in \mathcal{M}$$

and

$$(20) \quad R(\mathcal{M}) \leq Q_2,$$

for Q_2 a constant not depending on ε and N . More precisely, under (19) and (20), we have

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \leq C_p \left(\inf_{m \in \mathcal{M}} \mathbb{E} [\|\hat{\theta}^{(m)} - \theta\|_p^p] + \kappa_p \varepsilon^p \right),$$

where C_p depends only on p (in particular, it does not depend on the distribution of the ξ_λ 's) and on Q_1, Q_2 . Hence, the bound coincides with (18) up to the additional term $\kappa_p \varepsilon^p$, which is negligible when the infimum in (16) is achieved for large models. Therefore, we aim to choose the factors x_m so that Conditions (19) and (20) are satisfied. On the one hand, Condition (19) means that the factors cannot be too large. On the other hand, the bound $R(\mathcal{M}) \leq Q_2$ can only hold if either $\mathcal{M}$ is not too large or the factors are sufficiently large (see (17)). For instance, consider a setting in which the number of models $m \in \mathcal{M}$ of size d is exponential in d , say $\exp(cd)$ for some constant $c > 0$ and for all $d \geq 1$. When $p \leq 2$, it is possible to choose the factors x_m so that Condition (19) is satisfied and (20) holds, for instance by taking

$$x_m = \tilde{c}|m| \quad \text{or} \quad x_m = \tilde{c} \log(|m|)$$

with $\tilde{c}$ large enough. However, when $p > 2$, if Condition (19) holds, we have $x_m \leq (Q_1 |m|)^{2/p}$ and

$$R(\mathcal{M}) = \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} |m|^{\left(1-\frac{p}{2}\right)+e^{-x_m}} = \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} e^{-x_m} \geq \sum_{d=1}^D \exp \left(cd - Q_1^{2/p} d^{2/p} \right),$$

with $D = \max_{m \in \mathcal{M}} |m|$ (assumed to be positive). The right-hand side becomes arbitrarily large as D increases, which occurs when the collection of models is large. Therefore, to obtain good performance for $\tilde{\theta}$ in the case $p > 2$, we need to modify the penalty to deal with large collections of models. Since the elbow occurs at $p = 2$, we use $p(m)$ with the value $p = 2$ and we introduce

$$(21) \quad p^\#(m) = \frac{3}{2} \sigma_2^2 |m| + \kappa_2 x_m.$$

We obtain the following theorem.

Theorem 2.8. *Let $p \geq 2$ and $q > 1$. Let $(x_m)_{m \in \mathcal{M}}$ be a sequence of factors such that $x_m \geq 1$ for all $m \in \mathcal{M}$. We consider the estimate $\tilde{\theta} = \hat{\theta}^{(\tilde{m})}$ associated with the penalty defined for $m \in \mathcal{M}$ by*

$$(22) \quad \text{pen}(m) := 2\varepsilon^p \min \left(p(m), (2q \log N)^{\frac{p}{2}-1} p^\#(m) \right),$$

where $p(m)$ and $p^\#(m)$ are defined in (15) and (21) respectively. Then

$$(23) \quad \mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \leq \tilde{M}_p \inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right] + \text{pen}(m) \right\} + \check{M}_p \left(N^{1-q} \|\theta\|_p^p + \varepsilon^p R^\#(\mathcal{M}) \right),$$

where $\tilde{M}_p$ (resp. $\check{M}_p$) is a constant only depending on p (resp. $p, \sigma_p, \sigma_2, \sigma_{2p}$) and

$$R^\#(\mathcal{M}) = N^{\frac{3-q}{2}} + (q \log N)^{\frac{p}{2}-1} R(\mathcal{M}),$$

where $R(\mathcal{M})$ is defined in (17).

The proof of Theorem 2.8 is provided in Section 5.2.4. Unfortunately, the use of $p^\#(m)$ requires to add the multiplicative logarithmic term $(2q \log N)^{\frac{p}{2}-1}$ which vanishes only if $p = 2$.

If $\log N$ is of the same order as $|\log(\varepsilon)|$, then, by choosing q sufficiently large, the second term on the right-hand side of (23) coincides with that in (16), up to the logarithmic factor $|\log(\varepsilon)|^{\frac{p}{2}-1}$. Two scenarios are then of interest.

1. If we can take x_m satisfying (19) and (20), we use

$$\text{pen}(m) \leq 2\varepsilon^p \text{p}(m)$$

and Theorem 2.8 gives:

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \lesssim \inf_{m \in \mathcal{M}} \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right] + \varepsilon^p |\log(\varepsilon)|^{\frac{p}{2}-1}.$$

2. If we can take, for all $m \in \mathcal{M}$, $x_m = \tilde{c}|m|$, for $\tilde{c}$ a constant, such that (20) is satisfied, we use

$$\text{pen}(m) \leq 2\varepsilon^p (2q \log N)^{\frac{p}{2}-1} \text{p}^\#(m) \leq 3(\sigma_2^2 + 2\kappa_2 \tilde{c}) \varepsilon^p (2q \log N)^{\frac{p}{2}-1} |m|$$

and Theorem 2.8 gives, by taking $\log N$ of the same order as $|\log(\varepsilon)|$ and q large enough:

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \lesssim |\log(\varepsilon)|^{\frac{p}{2}-1} \inf_{m \in \mathcal{M}} \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right] + \varepsilon^p |\log(\varepsilon)|^{\frac{p}{2}-1}.$$

In the first scenario, $\tilde{\theta}$ is optimal up to the term $\varepsilon^p |\log(\varepsilon)|^{\frac{p}{2}-1}$. In the second scenario, the additional logarithmic factor $|\log(\varepsilon)|^{\frac{p}{2}-1}$ is required in front of the leading term. In both cases, the condition $R(\mathcal{M}) \leq Q_2$ which depends on the chosen model collection $\mathcal{M}$, is crucial. In Section 3, devoted to the minimax framework, the collection $\mathcal{M}$ will be specified in order to investigate the optimality of $\tilde{\theta}$ under various regularity assumptions on θ .

2.4. Extensions to weighted ℓ_p -norms

In this section, we extend the previous results to the case where the loss function is the weighted ℓ_p -norm:

$$\|\vartheta\|_{\ell_p(w)}^p = \sum_{\lambda \in \Lambda} w_\lambda |\vartheta_\lambda|^p, \quad \vartheta \in \mathbb{R}^\Lambda,$$

where $w = (w_\lambda)_{\lambda \in \Lambda}$ is a sequence of positive weights. In Model (3), we consider the risk associated with this weighted ℓ_p -norm. We assume in the sequel that $\|\theta\|_{\ell_p(w)} < \infty$. We modify the criterion defined in (5) and we set

$$\text{Crit}_w(m) = - \sum_{\lambda \in m} w_\lambda |Y_\lambda|^p + \text{pen}_w(m),$$

where $\text{pen}_w: \mathcal{M} \rightarrow \mathbb{R}_+$ is a penalty function depending on w . We estimate θ by

$$\tilde{\theta} = \hat{\theta}^{(\hat{m})} \quad \text{with} \quad \hat{m} = \arg \min_{m \in \mathcal{M}} \text{Crit}_w(m).$$

As exemplified in Section 3.2.1, weights can be constant on some slices. Therefore, we consider the following partition of $\Lambda^{(N)}$:

$$(24) \quad \Lambda^{(N)} = \bigcup_{j \in \mathcal{J}} \Lambda_j$$

so that w_λ is constant for any $\lambda \in \Lambda_j$ with $w_\lambda = \omega_j$. For ease of notation, we omit the dependence of the Λ_j 's and $\mathcal{J}$ on N . We also assume that $\omega_j \neq \omega_{j'}$ if $j \neq j'$. Therefore, up to some permutation of elements of $\mathcal{J}$, the partition (24) is unique. Then, given (24), we consider for any model $m \in \mathcal{M}$

$$m_j = m \cap \Lambda_j, \quad j \in \mathcal{J},$$

so that $m = \bigcup_{j \in \mathcal{J}} m_j$ and we set

$$\mathcal{M}_j = \{m_j : m \in \mathcal{M}\}, \quad j \in \mathcal{J}.$$

Theorem 2.5 can be extended as follows.

Corollary 2.9. *Let $p \geq 1$ and let $(x_{m_j})_{m_j}$ be a sequence of factors such that $x_{m_j} \geq 1$ for all $m_j \in \mathcal{M}_j$ and $j \in \mathcal{J}$. We then consider the estimate $\tilde{\theta} = \hat{\theta}^{(\tilde{m})}$ associated with the penalty defined for any $m = \bigcup_{j \in \mathcal{J}} m_j \in \mathcal{M}$ by*

$$\text{pen}_w(m) := 2\varepsilon^p \sum_{j \in \mathcal{J}} \omega_j p(m_j),$$

where p is defined in (15). Then

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_{\ell_p(w)}^p \right] \leq \widetilde{M}_p \inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p \right] + \text{pen}_w(m) \right\} + \check{M}_p \kappa_p \varepsilon^p R_w(\mathcal{M}),$$

with

$$(25) \quad R_w(\mathcal{M}) = \sum_{j \in \mathcal{J}} \omega_j \sum_{\substack{m_j \in \mathcal{M}_j \\ m_j \neq \emptyset}} |m_j|^{\left(1 - \frac{p}{2}\right)_+} e^{-x_{m_j}},$$

and $\widetilde{M}_p$ and $\check{M}_p$ are two constants only depending on p .

Corollary 2.9 can be viewed as a generalization of Theorem 4.5 of [41] for $p \neq 2$. Theorem 2.8 can be extended as follows.

Corollary 2.10. *Let $p \geq 2$ and $q > 1$. Let $(x_{m_j})_{m_j}$ be a sequence of factors such that $x_{m_j} \geq 1$ for all $m_j \in \mathcal{M}_j$ and $j \in \mathcal{J}$. We then consider the estimate $\tilde{\theta} = \hat{\theta}^{(\tilde{m})}$ associated with the penalty defined for any $m = \bigcup_{j \in \mathcal{J}} m_j \in \mathcal{M}$ by*

$$\text{pen}_w(m) := 2\varepsilon^p \sum_{j \in \mathcal{J}} \omega_j \min \left(p(m_j), (2q \log N)^{\frac{p}{2}-1} p^\#(m_j) \right),$$

where p and $p^\#$ are defined in (15) and (21) respectively. Then

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_{\ell_p(w)}^p \right] \leq \widetilde{M}_p \inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p \right] + \text{pen}_w(m) \right\} + \check{M}_p \left(N^{1-q} \|\theta\|_{\ell_p(w)}^p + \varepsilon^p R_w^\#(\mathcal{M}) \right),$$

where $\widetilde{M}_p$ (resp. $\check{M}_p$) is a constant only depending on p (resp. $p, \sigma_p, \sigma_2, \sigma_{2p}$) and

$$R_w^\#(\mathcal{M}) = N^{\frac{1-q}{2}} \sum_{\lambda \in \Lambda(N)} w_\lambda + (q \log N)^{\frac{p}{2}-1} R_w(\mathcal{M}),$$

where $R_w(\mathcal{M})$ is defined in (25).

Corollaries 2.9 and 2.10 are proved in Section 5.2.5.

3. Minimax approach

The goal of this section is to prove the optimality of our procedure. For this purpose, we consider the minimax setting. Two cases are considered. In Section 3.1, we illustrate our results on a simple situation, namely the case of sequences θ whose tails decay polynomially. Then, in Section 3.2, we investigate the minimax rates of convergence of our procedure on Besov bodies.

3.1. Minimax rates for polynomially decaying tails of θ

In this section, we study minimax rates of convergence of our procedure when tails of θ decay polynomially. Therefore, assuming without loss of generality that Λ is the set of positive integers denoted $\mathbb{N}^*$, we introduce for $s > 0$ and $R > 0$, the set $\mathbb{B}_p^s(R)$ defined by

$$\mathbb{B}_p^s(R) := \left\{ \vartheta = (\vartheta_\lambda)_{\lambda \in \mathbb{N}^*}, \quad \sup_{k \in \mathbb{N}^*} k^s \left(\sum_{\lambda > k} |\vartheta_\lambda|^p \right)^{1/p} \leq R \right\}.$$

This functional class, allowing for an easy control of the bias term, is natural in our setting. In the sequel, θ is assumed to belong to $\mathbb{B}_p^s(R)$. Now, we take:

$$\mathcal{M} = \left\{ \{1, \dots, k\}, \quad 1 \leq k \leq N \right\}$$

and we apply Theorem 2.5 by plugging the value $x_m = a \log(|m|)$ (and $x_\emptyset = 0$) in $\text{pen}(m)$ with $a > 1 + (1 - \frac{p}{2})_+$. In this case, since for each value $d \in \mathbb{N}^*$ there is only one model $m \in \mathcal{M}$ with cardinality d , we obtain:

$$R(\mathcal{M}) = \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} |m|^{(1-\frac{p}{2})_+} e^{-x_m} \leq \sum_{d=1}^{+\infty} d^{(1-\frac{p}{2})_+} e^{-a \log(d)} < \infty.$$

Theorem 2.5 gives

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \leq \widetilde{M}_p \inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right] + \text{pen}(m) \right\} + \check{M}_p \kappa_p \varepsilon^p R(\mathcal{M}).$$

Since

$$\mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right] = B_p(m) + \mathbb{E}[V_p(m)]$$

we have

$$\text{pen}(m) \lesssim \varepsilon^p |m| \lesssim \mathbb{E}[V_p(m)]$$

and we finally obtain:

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \leq \widetilde{M}'_p \inf_{m \in \mathcal{M}} \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right],$$

for $\widetilde{M}'_p$ a constant only depending on p and κ_p . For a model $m = \{1, \dots, k\}$,

$$\mathbb{E}[V_p(m)] = \sigma_p^p k \varepsilon^p \quad \text{and} \quad B_p(m) = \sum_{i > k} |\theta_i|^p.$$

We obtain:

$$\sup_{\theta \in \mathbb{B}_p^s(R)} \mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \lesssim \inf_{1 \leq k \leq N} \left\{ R^p k^{-sp} + \varepsilon^p k \right\} \lesssim R^{\frac{p}{1+ps}} \varepsilon^{\frac{p^2 s}{1+ps}}$$

as soon as $N \geq R^p \varepsilon^{-p}$. In particular, when $p = 2$, we obtain the bound

$$\sup_{\theta \in \mathbb{B}_2^s(R)} \mathbb{E} \left[\|\tilde{\theta} - \theta\|_2^2 \right] \lesssim R^{\frac{2}{1+2s}} \varepsilon^{\frac{4s}{1+2s}}.$$

In particular, the rate in the right hand side is the optimal minimax rate on the class $\mathbb{B}_2^s(R)$, see [43].

3.2. Minimax rates on Besov bodies

3.2.1. Setting

In this section, we wish to consider minimax rates on the very general class of Besov bodies. The latter is naturally associated with the wavelet framework. Therefore, we naturally adapt previous notations to this framework and, without loss of generality, we can rewrite the setting of Section 1.1 by assuming that

$$\Lambda = \bigcup_{j \geq -1} \left\{ \{j\} \times K_j \right\}, \quad K_j = \left\{ k \in \mathbb{N} : 0 \leq k < 2^j \right\}.$$

Such adaptations are also justified by the extensions of subsequent results to the regression framework studied in Section 4. For any $j \geq 0$, we have $|K_j| = 2^j$ and $|K_{-1}| = 1$ since $K_{-1} = \{0\}$. Now, our statistical model can be written as follows:

$$(26) \quad Y_{jk} = \theta_{jk} + \varepsilon \xi_{jk}, \quad j \geq -1, \quad k \in K_j.$$

In the sequel, we shall assume that the sequence $\theta = (\theta_{jk})_{(j,k) \in \Lambda}$ belongs to a Besov ball with smoothness s defined, as usual, by

$$\mathcal{B}_{r,\infty}^s(R) = \left\{ \vartheta = (\vartheta_{jk})_{(j,k) \in \Lambda} \in \mathbb{R}^\Lambda : \sup_{j \geq -1} 2^{j(s+\frac{1}{2}-\frac{1}{r})} \left(\sum_{k \in K_j} |\vartheta_{jk}|^r \right)^{1/r} \leq R \right\}$$

for $0 < s < \infty$, $1 \leq r < \infty$, $0 < R < \infty$ and

$$\mathcal{H}^s(R) = \mathcal{B}_{\infty,\infty}^s(R) = \left\{ \vartheta \in \mathbb{R}^\Lambda : \sup_{j \geq -1, k \in K_j} 2^{j(s+\frac{1}{2})} |\vartheta_{jk}| \leq R \right\}.$$

The sequence θ will be estimated by using the first N observations Y_{jk} , with $N = 2^{J+1}$ for some $J \geq 0$. It means that

$$\Lambda^{(N)} = \bigcup_{-1 \leq j \leq J} \Lambda_j, \quad \Lambda_j = \{j\} \times K_j.$$

We assume that the weights satisfy

$$(27) \quad w_{jk} = \omega_j = 2^{j(\frac{p}{2}-1)}, \quad (j,k) \in \Lambda,$$

and we still denote for any sequence ϑ in $\mathbb{R}^\Lambda$,

$$(28) \quad \|\vartheta\|_{\ell_p(w)}^p = \sum_{j=-1}^{+\infty} \sum_{k \in K_j} w_{jk} |\vartheta_{jk}|^p = \sum_{j=-1}^{+\infty} \omega_j \sum_{k \in K_j} |\vartheta_{jk}|^p$$

Such weights are naturally justified by the subsequent regression framework and strong relationships between the sequence norm $\|\cdot\|_{\ell_p(w)}$ and the classical functional $\mathbb{L}_p$ -norm. See Section 4.

3.2.2. Lower bounds

In this section, we still consider Model (26) but we assume that the ξ_{jk} 's are i.i.d. standard Gaussian variables. The lower bound for the $\ell_p(w)$ -risk is known for $s > 1/r$, see Theorems 7 and 9 of [14] (by taking, using their notation, $(\sigma, p, q) = (s, r, \infty)$, $(\sigma', p', q') = (0, p, p)$ and $C = R$), and Theorem 10 in [13]. Note that [37] obtained the same lower bounds in the Gaussian white noise model (except in the frontier case) under weaker assumption $s > 1/r - 1/p$.

Theorem 3.1. *Assume that $s > 1/r$ and $\varepsilon^{-2} \lesssim N$. Then, for ε small enough, we have:*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathcal{B}_{r,\infty}^s(R)} \mathbb{E} \left[\|\hat{\theta} - \theta\|_{\ell_p(w)}^p \right] \geq c \begin{cases} R^{\frac{p}{2s+1}} \varepsilon^{\frac{2ps}{2s+1}} & \text{if } r > \frac{p}{2s+1} \\ R^{\frac{p}{2s+1}} \varepsilon^{\frac{2ps}{2s+1}} |\log(\varepsilon)|^{\frac{ps}{2s+1}+1} & \text{if } r = \frac{p}{2s+1} \\ R^{\frac{p-2}{2s+1-\frac{2}{r}}} (\varepsilon^2 |\log(\varepsilon)|)^{\frac{p(s-\frac{1}{r}+\frac{1}{p})}{2s+1-\frac{2}{r}}} & \text{if } r < \frac{p}{2s+1} \end{cases}$$

where the infimum is taken over all estimators, i.e. measurable functions of $(Y_{jk})_{(j,k) \in \Lambda}$ and c is a positive constant depending on p , r and s .

The lower bounds reveal several zones according to the sign of $r - p/(2s+1)$. The case $r < p/(2s+1)$ will be called the *sparse case*. The dense case $r > p/(2s+1)$ can be decomposed into two cases: the case $r \geq p$ and the case $p/(2s+1) < r < p$ referred to respectively as the *homogeneous* and *intermediate cases*. Finally the case $r = p/(2s+1)$ will be called the *frontier case*. We refer to the zones for the upper bounds.

3.2.3. Upper bounds

We aim to derive the rate of convergence of our estimator over the class $\mathcal{B}_{r,\infty}^s(R)$ by considering a suitably chosen collection of models and applying Corollaries 2.9 and 2.10. Our collection of models is the union of three sub-collections $\mathcal{M}^H$, $\mathcal{M}^I$ and $\mathcal{M}^S$, each of them being the most suitable for homogeneous (H), intermediate (I) and sparse and frontier cases (S). For each sub-collection, we use a penalty with corresponding factor $x_{m_j}^H$, $x_{m_j}^I$ and $x_{m_j}^S$ (see below). The selection of the sub-collection and associated penalty is performed by our procedure automatically. We denote

$$\mathfrak{M} = \bigcup_{a \in \{H, I, S\}} \mathcal{M}^a \times \{a\}$$

with $\mathcal{M}^a$ defined as follows (recall that $m = \cup_{-1 \leq j \leq J} m_j$):

- strategy $a = H$: $m \in \mathcal{M}^H$ if there exists $L \in \{0, \dots, J\}$ such that
for all $-1 \leq j \leq J$, $m_j = \begin{cases} \Lambda_j & \text{if } -1 \leq j \leq L, \\ \emptyset & \text{if } j > L \end{cases}$
and $x_{m_j}^H = \frac{p}{2} \log |m_j|$ ($\log 0 = 0$);
- strategy $a = I$: $m \in \mathcal{M}^I$ if there exists $L \in \{0, \dots, J\}$ such that
for all $-1 \leq j \leq J$, $m_j = \begin{cases} \Lambda_j & \text{if } -1 \leq j \leq L-1 \\ m_{L+l} \subset \Lambda_{L+l} & \text{if } l = j-L \geq 0 \text{ with } |m_{L+l}| = \lfloor 2^{L+l} 2^{-lp/2} (l+1)^{-3} \rfloor, \end{cases}$
and $x_{m_j}^I = K |m_j| \left(1 + \log \left(\frac{2^j}{|m_j|} \right) \right)$ with K a constant only depending on p (see the proof);
- strategy $a = S$: $m \in \mathcal{M}^S$ (full collection) if
for all $-1 \leq j \leq J$, $m_j = \{j\} \times E$, $E \in \mathcal{P}(K_j)$,
and $x_{m_j}^S = (p+1) |m_j| j$.

For $(m, a) \in \mathfrak{M}$, having in mind Corollaries 2.9 and 2.10, denote

$$\text{pen}(m, a) = \text{pen}^a(m) = \begin{cases} 2\varepsilon^p \sum_{j=-1}^J \omega_j \mathfrak{p}_j(m_j, a) & \text{if } p \leq 2 \\ 2\varepsilon^p \sum_{j=-1}^J \omega_j \min \left(\mathfrak{p}_j(m_j, a), (2q \log N)^{\frac{p}{2}-1} \mathfrak{p}_j^\#(m_j, a) \right) & \text{if } p > 2 \end{cases}$$

with $q = p+1$ and

$$\begin{aligned} \mathfrak{p}_j(m_j, a) &= \mathfrak{p}_j^a(m_j) = \frac{3}{2} \sigma_p^p |m_j| + \kappa_p 2^{\frac{(p-2)+}{2}} |m_j|^{(1-\frac{p}{2})_+} (x_{m_j}^a)^{\frac{p}{2}}, \\ \mathfrak{p}_j^\#(m_j, a) &= \mathfrak{p}_j^{\#a}(m_j) = \frac{3}{2} \sigma_2^2 |m_j| + \kappa_2 x_{m_j}^a. \end{aligned}$$

We consider $(\hat{m}, \hat{a})$ which minimizes over all $(m, a) \in \mathfrak{M}$ the criterion

$$- \sum_{\lambda \in m} w_\lambda |Y_\lambda|^p + \text{pen}(m, a).$$

It can also be defined by minimizing over a the quantity $-\sum_{\lambda \in \hat{m}^a} w_\lambda |Y_\lambda|^p + \text{pen}^a(\hat{m}^a)$ where $\hat{m}^a$ minimizes over $\mathcal{M}^a$ the quantity $-\sum_{\lambda \in m} w_\lambda |Y_\lambda|^p + \text{pen}^a(m)$. Finally our PCO estimator is $\tilde{\theta} = \hat{\theta}(\hat{m}, \hat{a})$.

Theorem 3.2. Assume that $s > \max(1/r - 1/p, 0)$ and R the radius of the Besov ball belongs to the interval $[\varepsilon, \varepsilon^{-1}]$. We take the number of observations N such that $(\frac{R}{\varepsilon})^2 \leq N \leq (\frac{R}{\varepsilon})^\gamma$, with $\gamma \geq 2$. Then

$$\sup_{\theta \in \mathcal{B}_{r, \infty}^s(R)} \mathbb{E} \left[\|\tilde{\theta} - \theta\|_{\ell_p(w)}^p \right] \leq C \begin{cases} R^{\frac{p}{2s+1}} \varepsilon^{\frac{2ps}{2s+1}} & \text{if } r \geq p \\ R^{\frac{p}{2s+1}} \varepsilon^{\frac{2ps}{2s+1}} |\log(\varepsilon)|^{\frac{s(p-2)+}{2s+1}} & \text{if } \frac{p}{2s+1} < r < p \\ R^{\frac{p}{2s+1}} \varepsilon^{\frac{2ps}{2s+1}} |\log(\varepsilon)|^{\frac{ps}{2s+1}+1} & \text{if } r = \frac{p}{2s+1} \text{ and } s \geq 1/r \\ R^{\frac{p-2}{2s+1-\frac{2}{r}}} (\varepsilon^2 |\log(\varepsilon)|)^{\frac{p(s-\frac{1}{r}+\frac{1}{p})}{2s+1-\frac{2}{r}}} & \text{if } r < \frac{p}{2s+1} \text{ and } s \geq 1/r \end{cases}$$

for C a constant depending on p, σ_p, s, r and γ .

The proof of Theorem 3.2 is provided in Section 5.3. We remark that the radius R may either decrease to 0 or increase to $+\infty$ as $\varepsilon \rightarrow 0$. The lower bound $R \geq \varepsilon$ allows to have $N \geq 1$. The upper bound $R \leq \varepsilon^{-1}$ can be replaced by any upper bound $R \leq \varepsilon^{-u}$, with $u > 0$, so we have $\log(N) \approx |\log(\varepsilon)|$. Note that if $R \in [R_0, R_1]$ with known constants R_0 and R_1 with $0 < R_0 < R_1 < \infty$ then we can choose N not depending on R . Assumption $s \geq 1/r$ in the sparse and frontier cases can be removed by imposing $N \geq (R/\varepsilon)^{2/(1-2/p)}$ when $p > 2$; see Remark 5.9.

Theorem 3.2 shows that our procedure achieves the optimal rate in the homogeneous case $r \geq p$, in the sparse case $r < p/(2s+1)$ and in the frontier case $r = p/(2s+1)$. In the intermediate case $\frac{p}{2s+1} < r < p$, we have a logarithmic loss with exponent $\frac{s}{2s+1} (1 - \frac{2}{p})_+$ which is non-zero if and only if $p > 2$. When $p \leq 2$, the rate is optimal.

Many other adaptive procedures suffer from a logarithmic loss in at least one case. In particular, the estimators in [14] have a logarithmic loss, both in homogeneous and intermediate cases, with power $\frac{s}{2s+1}$. It is also the case in [39] (for

the study of the white noise model) or [31] (for the study of the multivariate white noise model) or [22] (for the study of the density model). The estimator proposed in [30], based on an involved combination of thresholding and Lepski-type procedures, reaches optimal rates in dense and sparse cases but not in the frontier case. Note that the adaptive procedure proposed in [37] for the Gaussian white noise model, using kernel estimators with varying bandwidths, achieves the optimal rate in all cases.

4. Nonparametric regression

In this section, we consider the following nonparametric regression model

$$(29) \quad X_i = f(t_i) + \sigma \eta_i, \quad 1 \leq i \leq n,$$

with η_i some i.i.d. centered sub-Gaussian variables modeling the noise and $\sigma > 0$ assumed to be known. In the previous expression, f is supported in $[0, 1]$ and the t_i 's are deterministic and equidistant : $t_i = i/n$, $i = 1, \dots, n$. We assume in the sequel that $\log_2(n)$ is an integer. Our goal is to estimate the function f by using observations $(X_i)_{i=1, \dots, n}$. The risk of any estimate will be evaluated by using the $\mathbb{L}_p$ -norm for some $1 \leq p < \infty$. For this purpose, we shall use results of Section 3.2.3 and in particular the weights (27) introduced in Section 3.2.1.

In this setting, we consider a decomposition of the signal f , assumed to be square-integrable, on a wavelet basis. The expansion of f is then of the form:

$$(30) \quad f = \sum_k \langle f, \phi_k \rangle \phi_k + \sum_{j=0}^{+\infty} \sum_k \langle f, \psi_{jk} \rangle \psi_{jk},$$

where ϕ_k is the translation of a father wavelet ϕ and ψ_{jk} is the dilation and translation of a mother wavelet ψ : for any x , we have:

$$\phi_k(x) = \phi(x - k), \quad \psi_{jk}(x) = 2^{j/2} \psi(2^j x - k).$$

The expansion (30) can of course be rewritten as

$$(31) \quad f = \sum_{j=-1}^{+\infty} \sum_{k \in K_j} \langle f, \varphi_{jk} \rangle \varphi_{jk},$$

with for any (j, k) , $\varphi_{jk} = \psi_{jk}$ if $j \geq 0$ and $\varphi_{jk} = \phi_k$ if $j = -1$. Considering the specific construction of Section 4 of [8] to obtain an orthonormal basis of $\mathbb{L}^2[0, 1]$, we can assume that the wavelets that generate the basis have a compact support which is then included into the interval $[A, B]$ for some $0 < A < B < \infty$ and is $M + 1$ times weakly differentiable, where M can be chosen by the practitioner. Finally, the construction of [8] shows that we can take

$$K_j = \{0, 1, 2, \dots, 2^j - 1\}, \quad j \geq 0$$

and $K_{-1} = \{0\}$.

Now, as in Section 3.2, we consider $\Lambda^{(N)}$ of size $N = 2^{J+1}$, for some $J \geq 0$, with

$$\Lambda^{(N)} = \bigcup_{-1 \leq j \leq J} \Lambda_j, \quad \Lambda_j = \{j\} \times K_j\}$$

and we set for any $(j, k) \in \Lambda^{(N)}$,

$$Y_{jk} = \frac{1}{n} \sum_{i=1}^n X_i \varphi_{jk}(t_i), \quad \theta_{jk} = \frac{1}{n} \sum_{i=1}^n f(t_i) \varphi_{jk}(t_i), \quad \xi_{jk} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i \varphi_{jk}(t_i).$$

Setting

$$\varepsilon = \frac{\sigma}{\sqrt{n}},$$

we have:

$$Y_{jk} = \theta_{jk} + \varepsilon \xi_{jk}, \quad (j, k) \in \Lambda^{(N)},$$

where the ξ_{jk} 's are centered. We obtain the model of Section 3.2.1, except that the ξ_{jk} 's are not independent and not identically distributed. Furthermore, our target is f and not $\theta = (\theta_{jk})_{(j,k) \in \Lambda(N)}$. However, we consider the PCO estimator $\tilde{\theta}$ introduced in Section 3.2.3 except that we modify the strategy I by taking $|m_{L+l}| = \lfloor 2^{L+l} 2^{-lp/2} (l+1)^{-3p/2} \rfloor$ when $p > 2$ and we slightly modify the value of κ_p (see equation (41) in Lemma 5.13 and equation (42)). We then define

$$\tilde{f} = \sum_{(j,k) \in \Lambda(N)} \tilde{\theta}_{jk} \varphi_{jk}$$

as an estimator of f . We study the upper bound of the $\mathbb{L}_p$ -risk of $\tilde{f}$ on the class of Besov spaces. We refer the reader to Section 9.2 of [25] for the definition of Besov spaces in terms of modulus of continuity. In the sequel, we use the characterization of Besov spaces through wavelet coefficients (see Corollary 9.1 of [25]): Under additional mild conditions on the wavelet functions ϕ and ψ , for $1 \leq r, q < \infty$ and $0 < s < M+1$, a function $g \in \mathbb{L}_r$ belongs to $\mathcal{B}_{r,q}^s$ if and only if

$$\sum_{j \geq -1} 2^{qj(s+\frac{1}{2}-\frac{1}{r})} \left(\sum_{k \in K_j} |\langle g, \psi_{jk} \rangle|^r \right)^{\frac{q}{r}} < \infty.$$

When $q = \infty$, this condition becomes

$$\sup_{j \geq -1} 2^{j(s+\frac{1}{2}-\frac{1}{r})} \left(\sum_{k \in K_j} |\langle g, \psi_{jk} \rangle|^r \right)^{\frac{1}{r}} < \infty.$$

We shall use this sequential characterization to define the radius of a Besov ball, which allows us to use the setting and notations of Section 3.2. We obtain the following result.

Theorem 4.1. *Let $1 \leq p, r < \infty$ and $s > 0$ such that $1/r < s < M+1$. We recall that $\varepsilon = \frac{\sigma}{\sqrt{n}}$. Assume that R , the radius of the Besov ball, belongs to an interval $[R_0, R_1]$, with $0 < R_0 < R_1 < +\infty$. We take the resolution level $N = 2^{J+1}$ such that $\varepsilon^{-2} \leq N \leq \varepsilon^{-\gamma}$, $\gamma \geq 2$. Then, we have:*

$$\sup_{f \in \mathcal{B}_{r,\infty}^s(R)} \mathbb{E} \left[\|\tilde{f} - f\|_{\mathbb{L}_p}^p \right] \leq C \begin{cases} \varepsilon^{\frac{2ps}{1+2s}} & \text{if } r \geq p \\ \varepsilon^{\frac{2ps}{1+2s}} |\log(\varepsilon)|^{\frac{s(p-2)+}{(1+2s)}} & \text{if } \frac{p}{2s+1} < r < p \\ \varepsilon^{\frac{2ps}{2s+1}} |\log(\varepsilon)|^{\frac{ps}{2s+1}+1} & \text{if } r = \frac{p}{2s+1} \\ (\varepsilon^2 |\log(\varepsilon)|)^{\frac{p(s-\frac{1}{r}+\frac{1}{p})}{2s+1-\frac{2}{r}}} & \text{if } r < \frac{p}{2s+1} \end{cases}$$

for C a constant depending on $p, s, r, \gamma, R_0, R_1$ and the wavelet functions ϕ and ψ .

Similarly to Theorem 3.2, Theorem 4.1 shows the optimality of our estimation procedure in the Gaussian case for the homogeneous case $r \geq p$, the frontier case $r = \frac{p}{2s+1}$, the sparse case $r < \frac{p}{2s+1}$ and the intermediate case $\frac{p}{2s+1} < r < p$ when $p \leq 2$. We refer the reader to [42] for the corresponding lower bounds; see also [12] and Theorem 4 of [14] devoted to the Gaussian case. When $p > 2$, for the intermediate case $\frac{p}{2s+1} < r < p$, we obtain the additional logarithmic term $|\log(\varepsilon)|^{\frac{s(p-2)}{(1+2s)}}$ similarly to Theorem 3.2.

To end this section, we give the main arguments of the proof of Theorem 4.1. The $\mathbb{L}_p$ -risk of $\tilde{f}$ is deduced from the following control:

$$(32) \quad \mathbb{E} \left[\|\tilde{f} - f\|_{\mathbb{L}_p}^p \right] \leq 2^{p-1} \left[\mathbb{E} \left[\left\| \sum_{(j,k) \in \Lambda(N)} (\tilde{\theta}_{jk} - \theta_{jk}) \varphi_{jk} \right\|_{\mathbb{L}_p}^p \right] + \left\| \sum_{(j,k) \in \Lambda(N)} \theta_{jk} \varphi_{jk} - f \right\|_{\mathbb{L}_p}^p \right]$$

and the application of Theorem 3.2 to bound the first term. However, several additional technical arguments are needed and we have to tackle several problems:

1. Concentration inequalities on the ξ_{jk} 's, which are now not i.i.d., were essential in previous sections. So, the question is the following: Do the noise variables ξ_{jk} have sufficiently nice concentration properties to apply Theorem 3.2?
2. How can we control the second term of the right hand side, which corresponds to an approximation term?
3. Can we connect the first term of the right hand side of (32) to the $\ell_p(w)$ -risk of $\tilde{\theta}$?

To address the first issue, we establish in the next proposition, proved in Section 5.4.1, that the $\xi_{j,k}$'s satisfy a result similar to that of Corollary 2.4. We use that the η_i 's are i.i.d. centered sub-Gaussian random variables.

Proposition 4.2. For any j and for any $\mathcal{I}_j \subset K_j$, we set

$$Z_j := \sum_{k \in \mathcal{I}_j} |\xi_{jk}|^p.$$

There exist positive constants c_φ , σ_p and κ'_p only depending on p and on the compactly supported father and mother wavelets ϕ and ψ such that for any $x \geq 1$,

$$\mathbb{P}\left(Z_j \geq \frac{3}{2}\sigma_p^p |\mathcal{I}_j| + \kappa'_p |\mathcal{I}_j|^{\left(1-\frac{p}{2}\right)+x\frac{p}{2}}\right) \leq c_\varphi e^{-x}.$$

Regarding the approximation term, we can prove the following result (see Section 5.4.2 for the proof).

Lemma 4.3. Assume that f belongs to the Besov set $\mathcal{B}_{r,\infty}^s(R)$ with $1/r < s < M+1$. Let $\theta_{jk} = \frac{1}{n} \sum_{i=1}^n f(t_i) \varphi_{jk}(t_i)$. Then, if $N \geq \frac{\varepsilon^{-2}}{|\log(\varepsilon)|}$,

$$\left\| \sum_{(j,k) \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \right\|_{\mathbb{L}_p}^p \leq C \begin{cases} R^p \varepsilon^{\frac{2ps}{1+2s}} & \text{if } r \geq \frac{p}{2s+1}, \\ R^p (\varepsilon^2 |\log(\varepsilon)|)^{\frac{p(s-\frac{1}{r}+\frac{1}{p})}{2s+1-\frac{2}{r}}} & \text{if } r < \frac{p}{2s+1}, \end{cases}$$

for C a constant depending on ϕ , ψ , s , r and p .

Finally, to address the third issue, we wish to compare the $\mathbb{L}_p$ -risk of $\tilde{f}$ (or rather the first terms of its decomposition) with the $\ell_p(w)$ -risk of $\tilde{\theta}$ when weights are those of Section 3.2.1 (see (27)). We first state the following lemma whose proof can be found in Section 5.4.3.

Lemma 4.4. Let $1 \leq p < \infty$. For any function g , we have:

$$\|g\|_{\mathbb{L}_p} \leq C \|g\|_{\mathcal{B}_{p,p \wedge 2}^0},$$

for C a constant and with

$$\|g\|_{\mathcal{B}_{p,p \wedge 2}^0}^{p \wedge 2} := \sum_{j \geq -1} 2^{j(p \wedge 2)(\frac{1}{2}-\frac{1}{p})} \left(\sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^p \right)^{\frac{p \wedge 2}{p}}.$$

Now, we naturally distinguish two cases.

- If $1 \leq p \leq 2$, using Lemma 4.4, we have:

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{(j,k) \in \Lambda^{(N)}} (\tilde{\theta}_{jk} - \theta_{jk}) \varphi_{jk} \right\|_{\mathbb{L}_p}^p \right] &\lesssim \mathbb{E} \left[\left\| \sum_{(j,k) \in \Lambda^{(N)}} (\tilde{\theta}_{jk} - \theta_{jk}) \varphi_{jk} \right\|_{\mathcal{B}_{p,p}^0}^p \right] \\ (33) \quad &\lesssim \mathbb{E} \left[\sum_{j=-1}^J 2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\tilde{\theta}_{jk} - \theta_{jk}|^p \right] \leq \mathbb{E} \|\tilde{\theta} - \theta\|_{\ell_p(w)}^p, \end{aligned}$$

where the $\|\cdot\|_{\ell_p(w)}$ -norm is defined in (28) with weights defined in (27). Therefore, results of Section 3.2.3 can be applied.

- If $2 < p < \infty$, still using Lemma 4.4, we have:

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{(j,k) \in \Lambda^{(N)}} (\tilde{\theta}_{jk} - \theta_{jk}) \varphi_{jk} \right\|_{\mathbb{L}_p}^p \right] &\lesssim \mathbb{E} \left[\left\| \sum_{(j,k) \in \Lambda^{(N)}} (\tilde{\theta}_{jk} - \theta_{jk}) \varphi_{jk} \right\|_{\mathcal{B}_{p,2}^0}^p \right] \\ &\lesssim \mathbb{E} \left[\left(\sum_{j=-1}^J 2^{j(1-\frac{2}{p})} \left(\sum_{k \in K_j} |\tilde{\theta}_{jk} - \theta_{jk}|^p \right)^{\frac{2}{p}} \right)^{\frac{p}{2}} \right] \\ (34) \quad &\lesssim \left[\sum_{j=-1}^J \left(\mathbb{E} \left[2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\tilde{\theta}_{jk} - \theta_{jk}|^p \right] \right)^{\frac{2}{p}} \right]^{\frac{p}{2}}, \end{aligned}$$

by using the generalized Minkowski inequality. Since we cannot insert the sum in j before taking the power $2/p$, the control of the $\mathbb{L}_p$ -risk by the $\ell_p(w)$ -one is not immediate. This last issue is addressed in Section 5.4.

These arguments and technical complements of Section 5.4 allow us to prove Theorem 4.1.

5. Proofs

5.1. Inequalities for sub-Gaussian variables

In this subsection, we collect several results concerning sub-Gaussian random variables. Although some of these results are classical, we provide the details for completeness.

Lemma 5.1. *Let $A > 0$. Let ξ be a real-valued centered random variable such that for all real u , we have:*

$$\mathbb{E}(\exp(u\xi)) \leq \exp(A^2 u^2).$$

Then, we obtain:

- (i) For any $t \geq 0$, $\mathbb{P}(|\xi| \geq t) \leq 2 \exp(-t^2/(4A^2)).$
- (ii) For any $p \geq 1$, $\sigma_p := (\mathbb{E}|\xi|^p)^{1/p} \leq 2A(3p)^{1/p} \sqrt{p/2} \leq 5A\sqrt{p}.$
- (iii) $\|\xi\|_{\psi_2} \leq 2A\sqrt{3}.$

Proof. (i) For any $t \geq 0$,

$$\mathbb{P}(\xi \geq t) \leq e^{-ut} \mathbb{E}(\exp(u\xi)) \leq \exp(-ut + A^2 u^2) \leq \exp(-t^2/(4A^2)),$$

by taking $u = t/(2A^2)$.

(ii) Using the previous inequality, we have:

$$(35) \quad \mathbb{E}|\xi|^p = \int_0^\infty \mathbb{P}(|\xi| \geq t) p t^{p-1} dt \leq 2p \int_0^\infty \exp(-t^2/(4A^2)) t^{p-1} dt = (2A)^p p \Gamma(p/2),$$

by using the change of variable $v = t^2/(4A^2)$. The bound comes from $\Gamma(x) \leq 3x^x$ for $x \geq 1/2$.

(iii) We recall that $\|\xi\|_{\psi_2} = \inf \left\{ t > 0 : \mathbb{E}[\exp((|\xi|/t)^2)] \leq 2 \right\}$. For any $t > 0$, we have the Taylor expansion:

$$\mathbb{E}[\exp((|\xi|/t)^2)] = \mathbb{E} \left(\sum_{k=0}^{\infty} \frac{(\xi^2/t^2)^k}{k!} \right) = 1 + \sum_{k=1}^{\infty} \frac{\mathbb{E}(\xi^{2k})}{t^{2k} k!}.$$

Using Inequality (35) for $p = 2k$ gives:

$$\mathbb{E}[\exp((|\xi|/t)^2)] \leq 1 + \sum_{k=1}^{\infty} \frac{(2A)^{2k} 2k \Gamma(k)}{t^{2k} k!} = 1 + 2 \sum_{k=1}^{\infty} \frac{(2A)^{2k}}{t^{2k}},$$

which is finite if $t > 2A$. Then $\mathbb{E}[\exp((|\xi|/t)^2)] \leq 1 + 2(2A/t)^2 \frac{1}{1-(2A/t)^2}$ and this quantity is bounded by 2 if $\frac{2}{(t/2A)^2 - 1} \leq 1 \Leftrightarrow t^2 \geq 3 \times 4A^2$. □

The following lemma provides an alternative upper bound of moments of sub-Gaussian random variables in terms of the Orlicz norm.

Lemma 5.2. *Let ξ be a sub-Gaussian random variable. Then, we have: $(\mathbb{E}|\xi|^p)^{1/p} \leq 2\sqrt{p}\|\xi\|_{\psi_2}$.*

Proof of Lemma 5.2. The definition of the Orlicz norm and the Markov inequality give $\mathbb{P}(|\xi| \geq t) \leq 2e^{-t^2/\|\xi\|_{\psi_2}^2}$. Then

$$\mathbb{E}|\xi|^p = \int_0^\infty \mathbb{P}(|\xi| \geq t) p t^{p-1} dt \leq 2p \int_0^\infty \exp(-t^2/\|\xi\|_{\psi_2}^2) t^{p-1} dt = \|\xi\|_{\psi_2}^p p \Gamma(p/2),$$

by using the change of variable $v = t^2/\|\xi\|_{\psi_2}^2$. The bound comes from $\Gamma(u/2) \leq 2.6(u/2)^{\frac{u}{2}} \leq (2\sqrt{u})^u/u$, for $u \geq 1$. □

Lemma 5.3. For any $r > 0$ and any random variable Y such that $\|Y\|_{\psi_r} < \infty$, we have:

$$\|Y - \mathbb{E}(Y)\|_{\psi_r} \leq K_r \|Y\|_{\psi_r} \quad \text{with } K_r = 1 + \frac{2(\log 2)^{-1/r} \Gamma(1/r)}{r}.$$

Proof. It is sufficient to prove the result for Y such that $\|Y\|_{\psi_r} = 1$. Note that such a variable satisfies $\mathbb{E}[\exp(|Y|^r)] \leq 2$. The triangle inequality gives

$$\|Y - \mathbb{E}(Y)\|_{\psi_r} \leq \|Y\|_{\psi_r} + \|\mathbb{E}(Y)\|_{\psi_r} = 1 + \|\mathbb{E}(Y)\|_{\psi_r}.$$

If $Y = a$, for a a deterministic constant, we have:

$$\|Y\|_{\psi_r} = \inf \left\{ t > 0 : \mathbb{E} \left[\exp \left((|a|/t)^r \right) \right] \leq 2 \right\} = |a|(\log 2)^{-1/r}.$$

Thus $\|Y - \mathbb{E}(Y)\|_{\psi_r} \leq 1 + |\mathbb{E}(Y)|(\log 2)^{-1/r}$ and it remains to bound $|\mathbb{E}(Y)|$. We can write

$$\begin{aligned} |\mathbb{E}(Y)| &\leq \mathbb{E}|Y| = \int_0^\infty \mathbb{P}(|Y| > u) du = \int_0^\infty \mathbb{P}(\exp(|Y|^r) > \exp(u^r)) du \\ &\leq \int_0^\infty \mathbb{E}[\exp(|Y|^r)] \exp(-u^r) du \leq 2 \int_0^\infty \exp(-u^r) du = \frac{2}{r} \Gamma(1/r). \end{aligned}$$

Thus $\|Y - \mathbb{E}(Y)\|_{\psi_r} \leq 1 + (\log 2)^{-1/r} (2/r) \Gamma(1/r)$. □

5.2. Proofs of oracle results of Section 2

5.2.1. Technical lemmas

Lemma 5.4. Let $p \geq 1$. Let $K > 0$ and let a, b be two real numbers such that $|a| \geq K|b|$. Then, for any $\alpha > 0$,

$$\begin{aligned} ||a + b|^p - |a|^p - \alpha|b|^p| &\leq C_{1p}(\alpha, K)|a|^p \\ \text{and } |a + b|^p - \alpha|a|^p - |b|^p| &\leq C_{2p}(\alpha, K)|a|^p \end{aligned}$$

where $C_{1p}(\alpha, K)$ and $C_{2p}(\alpha, K)$ are positive constants depending on p, α, K such that $\lim_{K \rightarrow \infty} C_{1p}(\alpha, K) = 0$.

Proof. The case $a = 0$ is obvious, so we assume $a \neq 0$. Denoting $x = b/a$, it is sufficient to study the function $g_\alpha(x) = |1 + x|^p - 1 - \alpha|x|^p$ on $[-K^{-1}, K^{-1}]$. Since it is a continuous function on a compact set, it is bounded and we denote by $C_{1p}(\alpha, K)$ the maximum of $|g_\alpha|$ on $[-K^{-1}, K^{-1}]$. Moreover $g_\alpha(0) = 0$ so that $\lim_{K \rightarrow \infty} C_{1p}(\alpha, K) = 0$. In the same way, the continuous function $|1 + x|^p - \alpha - |x|^p$ is bounded on $[-K^{-1}, K^{-1}]$ and we denote by $C_{2p}(\alpha, K)$ its bound. □

Lemma 5.5. For any $p > 1$, for any $\alpha > 0$, there exists $C(\alpha, p)$ such that for any $x > 0$ and $y > 0$,

$$(x + y)^p \leq (1 + \alpha)x^p + C(\alpha, p)y^p.$$

We have, when $\alpha \rightarrow 0$, $C(\alpha, p) \sim \left(\frac{\alpha}{p-1}\right)^{1-p} \rightarrow +\infty$.

Proof. We prove that

$$C(\alpha, p) = \left(1 - (1 + \alpha)^{\frac{-1}{p-1}}\right)^{-(p-1)}$$

by studying the function $t \mapsto C(\alpha, p) + (1 + \alpha)t^p - (t + 1)^p$. □

5.2.2. Proof of Proposition 2.1

Let us denote

$$J(m) = \sum_{\lambda \in \Lambda \setminus m} |\theta_\lambda|^p + \sum_{\lambda \in \Lambda^{(N)} \setminus m} (\alpha \varepsilon^p |\xi_\lambda|^p - |Y_\lambda|^p).$$

Setting

$$C_1 = \varepsilon^p \sum_{\lambda \in \Lambda^{(N)}} |\xi_\lambda|^p \quad \text{and} \quad C_2 = \sum_{\lambda \in \Lambda^{(N)}} |Y_\lambda|^p,$$

we can write

$$\begin{aligned} J(m) &= B_p(m) + \alpha(C_1 - V_p(m)) - (C_2 - \sum_{\lambda \in m} |Y_\lambda|^p) \\ &= \|\hat{\theta}^{(m)} - \theta\|_p^p - (1 + \alpha)V_p(m) + \sum_{\lambda \in m} |Y_\lambda|^p + \alpha C_1 - C_2. \end{aligned}$$

But the definition of $\hat{m}$ gives

$$-\sum_{\lambda \in \hat{m}} |Y_\lambda|^p + \text{pen}(\hat{m}) \leq -\sum_{\lambda \in m} |Y_\lambda|^p + \text{pen}(m).$$

Then, since C_1 and C_2 do not depend on m ,

$$\|\hat{\theta}^{(\hat{m})} - \theta\|_p^p - (1 + \alpha)V_p(\hat{m}) - J(\hat{m}) + \text{pen}(\hat{m}) \leq \|\hat{\theta}^{(m)} - \theta\|_p^p - (1 + \alpha)V_p(m) - J(m) + \text{pen}(m).$$

Thus

$$\|\hat{\theta}^{(\hat{m})} - \theta\|_p^p \leq \|\hat{\theta}^{(m)} - \theta\|_p^p + \left[(1 + \alpha)V_p(\hat{m}) - \text{pen}(\hat{m}) \right] - \left[(1 + \alpha)V_p(m) - \text{pen}(m) \right] + J(\hat{m}) - J(m)$$

and it is sufficient to control $J(\hat{m}) - J(m)$. Let us denote

$$S_\lambda = |\theta_\lambda|^p + \alpha \varepsilon^p |\xi_\lambda|^p - |Y_\lambda|^p,$$

so that, with $m^c = \Lambda^{(N)} \setminus m$ and $\hat{m}^c = \Lambda^{(N)} \setminus \hat{m}$,

$$\begin{aligned} J(\hat{m}) - J(m) &= \sum_{\lambda \in \hat{m}^c} S_\lambda - \sum_{\lambda \in m^c} S_\lambda = \left(\sum_{\lambda \in \hat{m}^c \cap m} S_\lambda + \sum_{\lambda \in \hat{m}^c \cap m^c} S_\lambda \right) - \left(\sum_{\lambda \in m^c \cap \hat{m}^c} S_\lambda + \sum_{\lambda \in m^c \cap \hat{m}} S_\lambda \right) \\ &= \sum_{\lambda \in \hat{m}^c \cap m} S_\lambda - \sum_{\lambda \in m^c \cap \hat{m}} S_\lambda. \end{aligned}$$

5.2.2.1. Case $p > 1$. We first deal with the second term. We have:

$$\begin{aligned} - \sum_{\lambda \in m^c \cap \hat{m}} S_\lambda &= \sum_{\lambda \in m^c \cap \hat{m}} \left[|Y_\lambda|^p - |\theta_\lambda|^p - \alpha \varepsilon^p |\xi_\lambda|^p \right] = \sum_{\lambda \in m^c \cap \hat{m}} \left[|\theta_\lambda + \varepsilon \xi_\lambda|^p - |\theta_\lambda|^p - \alpha \varepsilon^p |\xi_\lambda|^p \right] \\ &\leq \sum_{\lambda \in m^c \cap \hat{m}} \left[\left(1 + \frac{\alpha}{2} \right) |\varepsilon \xi_\lambda|^p + C(\alpha/2, p) |\theta_\lambda|^p - |\theta_\lambda|^p - \alpha \varepsilon^p |\xi_\lambda|^p \right], \end{aligned}$$

by using notations of Lemma 5.5. Finally,

$$\begin{aligned} - \sum_{\lambda \in m^c \cap \hat{m}} S_\lambda &\leq \left(1 - \frac{\alpha}{2} \right) \sum_{\lambda \in \hat{m}} |\varepsilon \xi_\lambda|^p + \left(C(\alpha/2, p) - 1 \right) \sum_{\lambda \in m^c} |\theta_\lambda|^p \\ &\leq \left(1 - \frac{\alpha}{2} \right) V_p(\hat{m}) + \left(C(\alpha/2, p) - 1 \right) B_p(m). \end{aligned}$$

We now deal with the first term, namely $\sum_{\lambda \in \hat{m}^c \cap m} S_\lambda$. Let $K > 0$.

1. We assume that $|\theta_\lambda| \geq K \varepsilon |\xi_\lambda|$. Then, applying Lemma 5.4, we have

$$|S_\lambda| = \left| |\theta_\lambda|^p + \alpha \varepsilon^p |\xi_\lambda|^p - |Y_\lambda|^p \right| \leq C_{1p}(\alpha, K) |\theta_\lambda|^p \leq \left(1 - \frac{\alpha}{2} \right) |\theta_\lambda|^p,$$

choosing $K \equiv K_{p, \alpha}$ large enough.

2. We assume that $|\theta_\lambda| < K\varepsilon|\xi_\lambda|$. Then

$$|S_\lambda| \leq \left| |\theta_\lambda|^p + \alpha\varepsilon^p|\xi_\lambda|^p - |Y_\lambda|^p \right| \leq C_{2p}(\alpha, K^{-1})|\varepsilon\xi_\lambda|^p.$$

using again Lemma 5.4 with $|\varepsilon\xi_\lambda| \geq K^{-1}|\theta_\lambda|$.

Finally,

$$\begin{aligned} \sum_{\lambda \in \widehat{m}^c \cap m} S_\lambda &\leq \left(1 - \frac{\alpha}{2}\right) \sum_{\lambda \in \widehat{m}^c} |\theta_\lambda|^p + C_{2p}(\alpha, K_{p,\alpha}^{-1}) \sum_{\lambda \in m} |\varepsilon\xi_\lambda|^p \\ &\leq \left(1 - \frac{\alpha}{2}\right) B_p(\widehat{m}) + C_{2p}(\alpha, K_{p,\alpha}^{-1}) V_p(m). \end{aligned}$$

We obtain

$$J(\widehat{m}) - J(m) \leq \left(1 - \frac{\alpha}{2}\right) \|\widehat{\theta}^{(m)} - \theta\|_p^p + C'(\alpha, p) \|\widehat{\theta}^{(m)} - \theta\|_p^p$$

with $C'(\alpha, p) = \max\left(C\left(\frac{\alpha}{2}, p\right) - 1, C_{2p}(\alpha, K_{p,\alpha}^{-1})\right)$. Thus

$$\frac{\alpha}{2} \|\widehat{\theta}^{(m)} - \theta\|_p^p \leq (1 + C'(\alpha, p)) \|\widehat{\theta}^{(m)} - \theta\|_p^p + [(1 + \alpha)V_p(\widehat{m}) - \text{pen}(\widehat{m})] - [(1 + \alpha)V_p(m) - \text{pen}(m)]$$

and

$$\|\widehat{\theta}^{(m)} - \theta\|_p^p \leq \frac{2}{\alpha} (1 + C'(\alpha, p)) \|\widehat{\theta}^{(m)} - \theta\|_p^p + \frac{2}{\alpha} [(1 + \alpha)V_p(\widehat{m}) - \text{pen}(\widehat{m})] - \frac{2}{\alpha} [(1 + \alpha)V_p(m) - \text{pen}(m)].$$

Thus the result is proved with

$$(36) \quad M_{p,\alpha} = \frac{2}{\alpha} \left(1 + \max\left(C\left(\frac{\alpha}{2}, p\right) - 1, C_{2p}(\alpha, K_{p,\alpha}^{-1})\right)\right)$$

where $K_{p,\alpha}$ is such that $C_{1p}(\alpha, K_{p,\alpha}) \leq 1 - \alpha/2$ and where $C_{1p}, C_{2p}, C(\cdot, p)$ are defined in Lemmas 5.4 and 5.5.

5.2.2.2. Case $p = 1$. In this case

$$S_\lambda = |\theta_\lambda| + \alpha\varepsilon|\xi_\lambda| - |Y_\lambda| \geq (\alpha - 1)\varepsilon|\xi_\lambda|.$$

Note also that

$$S_\lambda = |\theta_\lambda| + \alpha\varepsilon|\xi_\lambda| - |\theta_\lambda + \varepsilon\xi_\lambda| \leq (\alpha + 1)\varepsilon|\xi_\lambda|.$$

Then, if $\alpha \geq 1$,

$$\begin{aligned} J(\widehat{m}) - J(m) &= \sum_{\lambda \in \widehat{m}^c} S_\lambda - \sum_{\lambda \in m^c} S_\lambda \leq \sum_{\lambda \in \Lambda^{(N)}} S_\lambda - \sum_{\lambda \in m^c} S_\lambda = \sum_{\lambda \in m} S_\lambda \\ &\leq \sum_{\lambda \in m} (\alpha + 1)\varepsilon|\xi_\lambda| \leq (\alpha + 1) \|\widehat{\theta}^{(m)} - \theta\|_1 \end{aligned}$$

and then

$$\|\widehat{\theta} - \theta\|_p^p \leq (2 + \alpha) \|\widehat{\theta}^{(m)} - \theta\|_p^p + [(1 + \alpha)V_p(\widehat{m}) - \text{pen}(\widehat{m})] - [(1 + \alpha)V_p(m) - \text{pen}(m)].$$

Now, if $0 < \alpha < 1$,

$$\begin{aligned} J(\widehat{m}) - J(m) &= \sum_{\lambda \in \widehat{m}^c \cap m} S_\lambda - \sum_{\lambda \in m^c \cap \widehat{m}} S_\lambda \leq \sum_{\lambda \in \widehat{m}^c \cap m} (\alpha + 1)\varepsilon|\xi_\lambda| + \sum_{\lambda \in m^c \cap \widehat{m}} (1 - \alpha)\varepsilon|\xi_\lambda| \\ &\leq (\alpha + 1) \|\widehat{\theta}^{(m)} - \theta\|_1 + (1 - \alpha) \|\widehat{\theta}^{(m)} - \theta\|_1 \end{aligned}$$

and then

$$\alpha \|\widehat{\theta} - \theta\|_p^p \leq (2 + \alpha) \|\widehat{\theta}^{(m)} - \theta\|_p^p + [(1 + \alpha)V_p(\widehat{m}) - \text{pen}(\widehat{m})] - [(1 + \alpha)V_p(m) - \text{pen}(m)].$$

Here $M_{p,\alpha} = M_{1,\alpha} = \max(2 + \alpha, 1 + 2/\alpha)$.

5.2.3. Proof of Theorem 2.5

We only consider the case $p > 1$, the case $p = 1$ is similar.

Let m be an arbitrary model in $\mathcal{M}$. Starting from Proposition 2.1, we have:

$$\begin{aligned} \|\tilde{\theta} - \theta\|_p^p &\leq M_p \|\hat{\theta}^{(m)} - \theta\|_p^p + 2 \left[2V_p(\hat{m}) - \text{pen}(\hat{m}) \right] - 2 \left[2V_p(m) - \text{pen}(m) \right] \\ &\leq M_p \|\hat{\theta}^{(m)} - \theta\|_p^p + 2\text{pen}(m) + 2 \left[2V_p(\hat{m}) - \text{pen}(\hat{m}) \right] \\ &\leq M_p \|\hat{\theta}^{(m)} - \theta\|_p^p + 2\text{pen}(m) + 4\varepsilon^p \sum_{m \in \mathcal{M}} \left[Z(m) - p(m) \right]_+, \end{aligned}$$

where $Z(m) = \sum_{\lambda \in m} |\xi_\lambda|^p$ and

$$p(m) = \frac{3}{2} \sigma_p^p |m| + \kappa_p 2^{\frac{(p-2)_+}{2}} |m|^{(1-\frac{p}{2})_+} x_m^{\frac{p}{2}}.$$

Note that $[Z(\emptyset) - p(\emptyset)]_+ = [-p(\emptyset)]_+ = 0$, so we have:

$$\|\tilde{\theta} - \theta\|_p^p \leq M_p \|\hat{\theta}^{(m)} - \theta\|_p^p + 2\text{pen}(m) + 4\varepsilon^p \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} \left[Z(m) - p(m) \right]_+.$$

Taking the expectation yields

$$\mathbb{E} \|\tilde{\theta} - \theta\|_p^p \leq M_p \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|_p^p + 2\text{pen}(m) + 4\varepsilon^p R,$$

with

$$R = \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} \mathbb{E} \left[Z(m) - p(m) \right]_+ = \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} \int_0^\infty \mathbb{P} \left(Z(m) - p(m) > u \right) du.$$

It remains to control the term R . For the next computation, we denote $\mathbf{a}_{m,p} = \kappa_p |m|^{(1-p/2)_+}$. Using the change of variable $u = \mathbf{a}_{m,p} 2^{\frac{(p-2)_+}{2}} v^{p/2}$, we have:

$$\begin{aligned} \int_0^\infty \mathbb{P} \left(Z(m) - p(m) > u \right) du &= \int_0^\infty \mathbb{P} \left(Z(m) - \frac{3}{2} \sigma_p^p |m| - \mathbf{a}_{m,p} 2^{\frac{(p-2)_+}{2}} x_m^{\frac{p}{2}} > u \right) du \\ &= \int_0^\infty \mathbb{P} \left(Z(m) - \frac{3}{2} \sigma_p^p |m| - \mathbf{a}_{m,p} 2^{\frac{(p-2)_+}{2}} x_m^{\frac{p}{2}} > \mathbf{a}_{m,p} 2^{\frac{(p-2)_+}{2}} v^{\frac{p}{2}} \right) \\ &\quad \times \left[\mathbf{a}_{m,p} 2^{\frac{(p-2)_+}{2}} \frac{p}{2} v^{\frac{p}{2}-1} \right] dv \\ &\leq \int_0^\infty \mathbb{P} \left(Z(m) > \frac{3}{2} \sigma_p^p |m| + \mathbf{a}_{m,p} (x_m + v)^{\frac{p}{2}} \right) \left[\mathbf{a}_{m,p} 2^{\frac{(p-2)_+}{2}} \frac{p}{2} v^{\frac{p}{2}-1} \right] dv \end{aligned}$$

since $2^{\frac{(p-2)_+}{2}} (a^{p/2} + b^{p/2}) \geq (a + b)^{p/2}$. Corollary 2.4 gives

$$\int_0^\infty \mathbb{P} \left(Z(m) - p(m) > u \right) du \leq 2 \int_0^\infty e^{-(x_m+v)} \left[\mathbf{a}_{m,p} 2^{\frac{(p-2)_+}{2}} \frac{p}{2} v^{\frac{p}{2}-1} \right] dv \leq C(p) \mathbf{a}_{m,p} e^{-x_m},$$

for $C(p)$ a constant only depending on p . Finally,

$$R \leq C(p) \kappa_p \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} |m|^{(1-p/2)_+} e^{-x_m}.$$

5.2.4. Proof of Theorem 2.8

We denote for $q > 1$,

$$\Omega(N, q) := \bigcap_{\lambda \in \Lambda^{(N)}} \left\{ |\xi_\lambda| \leq \sqrt{2q \log N} \right\}.$$

Recall that our sub-Gaussian assumption implies $\mathbb{P}(|\xi_\lambda| > t) \leq 2e^{-t^2/2}$ so we have $\mathbb{P}(\Omega(N, q)) \geq 1 - 2N^{1-q}$. Now,

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] = \mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \mathbf{1}_{\Omega(N, q)} \right] + \mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \mathbf{1}_{\Omega(N, q)^c} \right].$$

First,

$$\begin{aligned} \mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \mathbf{1}_{\Omega(N, q)^c} \right] &= \mathbb{E} \left[\sum_{\lambda \notin \hat{m}} |\theta_\lambda|^p \mathbf{1}_{\Omega(N, q)^c} \right] + \varepsilon^p \mathbb{E} \left[\sum_{\lambda \in \hat{m}} |\xi_\lambda|^p \mathbf{1}_{\Omega(N, q)^c} \right] \\ &\leq 2N^{1-q} \|\theta\|_p^p + \varepsilon^p \sqrt{2} N^{\frac{3-q}{2}} \sigma_{2p}^p. \end{aligned}$$

Then, starting from Proposition 2.1, we first have on $\Omega(N, q)$, for any $m \in \mathcal{M}$,

$$\begin{aligned} \|\tilde{\theta} - \theta\|_p^p &\leq M_p \|\hat{\theta}^{(m)} - \theta\|_p^p + 2 \left[2V_p(\hat{m}) - \text{pen}(\hat{m}) \right] - 2 \left[2V_p(m) - \text{pen}(m) \right] \\ &\leq M_p \|\hat{\theta}^{(m)} - \theta\|_p^p + 2\text{pen}(m) + 2 \left[2V_p(\hat{m}) - \text{pen}(\hat{m}) \right] \\ &\leq M_p \|\hat{\theta}^{(m)} - \theta\|_p^p + 2\text{pen}(m) + 4\varepsilon^p \sum_{m \in \mathcal{M}} \left[Z(m) - P(m) \right]_+, \end{aligned}$$

where $Z(m) = \sum_{\lambda \in m} |\xi_\lambda|^p$ and

$$P(m) = \min(p^1(m), p^2(m)),$$

with

$$p^1(m) = p(m), \quad p^2(m) = (2q \log N)^{\frac{p}{2}-1} p^\#(m).$$

Then, taking the expectation

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \mathbf{1}_{\Omega(N, q)} \right] \leq M_p \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \mathbf{1}_{\Omega(N, q)} \right] + 2\text{pen}(m) + 4\varepsilon^p R,$$

with

$$R = \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} \mathbb{E} \left[(Z(m) - P(m))_+ \mathbf{1}_{\Omega(N, q)} \right].$$

It remains to control the term R . We have:

$$\begin{aligned} R &= \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} \int_0^\infty \mathbb{P} \left(Z(m) \mathbf{1}_{\Omega(N, q)} - \min(p^1(m); p^2(m)) \mathbf{1}_{\Omega(N, q)} > u \right) du \\ &\leq \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} \left[\int_0^\infty \mathbb{P} \left(Z(m) - p^1(m) > u \right) du + \int_0^\infty \mathbb{P} \left(\{Z(m) - p^2(m) > u\} \cap \Omega(N, q) \right) du \right]. \end{aligned}$$

With the same computation as in the proof of Theorem 2.5, we have

$$\int_0^\infty \mathbb{P} \left(Z(m) - p^1(m) > u \right) du \leq C(p) \kappa_p |m|^{(1-p/2)+} e^{-x_m}.$$

On $\Omega(N, q)$, for $p \geq 2$,

$$Z(m) - p^2(m) = \sum_{\lambda \in m} |\xi_\lambda|^p - (2q \log N)^{\frac{p}{2}-1} p^\#(m)$$

$$\begin{aligned}
&\leq (\sqrt{2q \log N})^{p-2} \sum_{\lambda \in m} \xi_\lambda^2 - (2q \log N)^{\frac{p}{2}-1} \left(\frac{3}{2} \sigma_2^2 |m| + \kappa_2 x_m \right) \\
&\leq (2q \log N)^{\frac{p}{2}-1} \left(\sum_{\lambda \in m} \xi_\lambda^2 - \left(\frac{3}{2} \sigma_2^2 |m| + \kappa_2 x_m \right) \right).
\end{aligned}$$

Therefore,

$$\int_0^\infty \mathbb{P} \left(\{Z(m) - p^2(m) > u\} \cap \Omega(N, q) \right) du \leq (2q \log N)^{\frac{p}{2}-1} C(2) \kappa_2 e^{-x_m}.$$

We obtain, for $p \geq 2$,

$$R \leq \max(C(p) \kappa_p, C(2) \kappa_2) \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} \left[1 + (2q \log N)^{\frac{p}{2}-1} \right] e^{-x_m}.$$

Finally, for $\check{M}_p$ a constant only depending on $p, \sigma_p, \sigma_2, \sigma_{2p}$, we have:

$$\mathbb{E} \left[\|\tilde{\theta} - \theta\|_p^p \right] \leq M_p \mathbb{E} \left[\|\hat{\theta}^{(m)} - \theta\|_p^p \right] + 2\text{pen}(m) + \check{M}_p \left(N^{1-q} \|\theta\|_p^p + \varepsilon^p R^\#(\mathcal{M}) \right),$$

with

$$R^\#(\mathcal{M}) = N^{\frac{3-q}{2}} + (q \log N)^{\frac{p}{2}-1} \sum_{\substack{m \in \mathcal{M} \\ m \neq \emptyset}} e^{-x_m}.$$

5.2.5. Proof of Corollaries 2.9 and 2.10

We prove Corollary 2.9. We have:

$$\begin{aligned}
\text{Crit}_w(m) &= - \sum_{\lambda \in m} w_\lambda |Y_\lambda|^p + \text{pen}_w(m) \\
&= \sum_{j \in \mathcal{J}} \omega_j \left(- \sum_{\lambda \in m_j} |Y_\lambda|^p + 2\varepsilon^p p_j(m_j) \right).
\end{aligned}$$

Since the $\mathcal{M}_j$'s are disjoint and the weights ω_j 's are positive, we have:

$$\tilde{\theta} = \hat{\theta}^{(\hat{m})}, \quad \hat{m} = \bigcup_{j \in \mathcal{J}} \hat{m}_j,$$

with

$$\hat{m}_j = \arg \min_{m_j \in \mathcal{M}_j} \left\{ - \sum_{\lambda \in m_j} |Y_\lambda|^p + 2\varepsilon^p p_j(m_j) \right\}.$$

We can then apply Theorem 2.5 to each $\mathcal{M}_j, j \in \mathcal{J}$ and we obtain:

$$\begin{aligned}
\mathbb{E} \left[\|\tilde{\theta} - \theta\|_{\ell_p(w)}^p \right] &= \sum_{j \in \mathcal{J}} \omega_j \mathbb{E} \left[\sum_{\lambda \in \hat{m}_j} |\hat{\theta}_\lambda^{(\hat{m}_j)} - \theta_\lambda|^p \right] \\
&\leq \sum_{j \in \mathcal{J}} \omega_j \left(\widetilde{M}_p \inf_{m_j \in \mathcal{M}_j} \left\{ \mathbb{E} \left[\sum_{\lambda \in m_j} |\hat{\theta}_\lambda^{(m_j)} - \theta_\lambda|^p \right] + \text{pen}(m_j) \right\} + \check{M}_p \kappa_p \varepsilon^p R(\mathcal{M}_j) \right) \\
&= \widetilde{M}_p \inf_{m \in \mathcal{M}} \left\{ \sum_{j \in \mathcal{J}} \omega_j \left(\mathbb{E} \left[\sum_{\lambda \in m_j} |\hat{\theta}_\lambda^{(m_j)} - \theta_\lambda|^p \right] + \text{pen}(m_j) \right) \right\} + \check{M}_p \kappa_p \varepsilon^p \sum_{j \in \mathcal{J}} \omega_j R(\mathcal{M}_j),
\end{aligned}$$

which gives the result of Corollary 2.9. The proof of Corollary 2.10 is similar.

5.3. Proof of Theorem 3.2

We first prove that $\|\theta\|_{\ell_p(w)} < \infty$.

Lemma 5.6. *Assume that $\theta \in \mathcal{B}_{r,\infty}^s(R)$ with $s > \max(\frac{1}{r} - \frac{1}{p}, 0)$. Then, there exists C , only depending on s , r and p , such that*

$$\|\theta\|_{\ell_p(w)}^p \leq CR^p.$$

Proof. Let $j \geq -1$ be fixed. Assume first that $p \geq r$. Then

$$\left(\sum_{k \in K_j} |\theta_{jk}|^p \right)^{\frac{1}{p}} \leq \left(\sum_{k \in K_j} |\theta_{jk}|^r \right)^{\frac{1}{r}} \leq R 2^{-j(s + \frac{1}{2} - \frac{1}{r})}.$$

Therefore,

$$2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\theta_{jk}|^p \leq R^p 2^{j(\frac{p}{2}-1-ps-\frac{p}{2}+\frac{p}{r})} \leq R^p 2^{-jp(s-\frac{1}{r}+\frac{1}{p})}.$$

Now, assume that $p < r$. Hölder's inequality implies that

$$\sum_{k \in K_j} |\theta_{jk}|^p \leq \left(\sum_{k \in K_j} |\theta_{jk}|^r \right)^{\frac{p}{r}} 2^{j(1-\frac{p}{r})}$$

and

$$2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\theta_{jk}|^p \leq R^p 2^{j(\frac{p}{2}-1-ps-\frac{p}{2}+\frac{p}{r}+1-\frac{p}{r})} \leq R^p 2^{-jsp}.$$

Finally, in both cases, since $s > \max(\frac{1}{r} - \frac{1}{p}, 0)$,

$$\|\theta\|_{\ell_p(w)}^p = \sum_{j=-1}^{+\infty} 2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\theta_{jk}|^p \leq CR^p,$$

with C only depending on s , r and p . □

We distinguish the cases $p \leq 2$ and $p > 2$. When $p \leq 2$, we apply Corollary 2.9 to the collection $\mathfrak{M}$ and penalty pen and we obtain

$$(37) \quad \mathbb{E} \|\widehat{\theta}^{(\widehat{m}^a)} - \theta\|_{\ell_p(w)}^p \leq \widetilde{M}_p \inf_{(m,a) \in \mathfrak{M}} \left\{ \mathbb{E} \|\widehat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m, a) \right\} + \check{M}_p \varepsilon^p R_w(\mathfrak{M}).$$

When $p > 2$, we apply Corollary 2.10 to the collection $\mathfrak{M}$ and penalty pen . This gives

$$\mathbb{E} \|\widehat{\theta}^{(\widehat{m}^a)} - \theta\|_{\ell_p(w)}^p \leq \widetilde{M}_p \inf_{(m,a) \in \mathfrak{M}} \left\{ \mathbb{E} \|\widehat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m, a) \right\} + \check{M}_p \left(N^{1-q} \|\theta\|_{\ell_p(w)}^p + \varepsilon^p R_w^\#(\mathfrak{M}) \right),$$

with $R_w^\#(\mathfrak{M}) = N^{(1-q)/2} \sum_{\lambda \in \Lambda(N)} w_\lambda + (\log N)^{\frac{p}{2}-1} R_w(\mathfrak{M})$.

Let us first analyse the remaining term $N^{1-q} \|\theta\|_{\ell_p(w)}^p + \varepsilon^p R_w^\#(\mathfrak{M})$. In both cases observe that

$$R_w(\mathfrak{M}) = \sum_{j=-1}^J \omega_j \sum_{\substack{(m_j, a) \in \mathcal{M}_j \times \{H, I, S\} \\ m_j \neq \emptyset}} |m_j|^{(1-\frac{p}{2})_+} e^{-x_{m_j}^a} = R_w(\mathcal{M}^H) + R_w(\mathcal{M}^I) + R_w(\mathcal{M}^S).$$

We will prove in the following Sections 5.3.1, 5.3.2, 5.3.3 that for each $a \in \{H, I, S\}$, $R_w(\mathcal{M}^a)$ is bounded by a constant, except in the intermediate case for $p < 2$ where the bound is $\log N$ up to a constant. Note that

$$\sum_{\lambda} w_{\lambda} = \sum_{j=-1}^J \omega_j 2^j \lesssim 2^{Jp/2} \lesssim N^{p/2}$$

and, using Lemma 5.6, the remaining term is bounded as follows (for $q \geq p+1$):

$$\begin{aligned} N^{1-q} \|\theta\|_{\ell_p(w)}^p + \varepsilon^p R_w^{\#}(\mathfrak{M}) &\lesssim N^{1-q} R^p + N^{\frac{1+p-q}{2}} \varepsilon^p + \varepsilon^p (\log N)^{\frac{p}{2}-1} R_w(\mathfrak{M}) \\ &\lesssim \left(\frac{R}{\varepsilon}\right)^{2(1-q)} R^p + \left(\frac{R}{\varepsilon}\right)^{1+p-q} \varepsilon^p + \varepsilon^p (\log N)^{\frac{p}{2}-1} R_w(\mathfrak{M}) \\ &\lesssim \left(\frac{R}{\varepsilon}\right)^{2-2q+p} \varepsilon^p + \left(\frac{R}{\varepsilon}\right)^{1+p-q} \varepsilon^p + \varepsilon^p (\log N)^{\frac{p}{2}-1} R_w(\mathfrak{M}). \end{aligned}$$

Since $R \geq \varepsilon$, taking $q = p+1$ allows to show that this term is negligible. Indeed, we have that

$$\varepsilon^p (\log N)^{\frac{p}{2}} \lesssim R^{\frac{p}{2s+1}} \varepsilon^{\frac{2ps}{2s+1}} \iff (\log N)^{\frac{p}{2}} \lesssim \left(\frac{R}{\varepsilon}\right)^{\frac{p}{2s+1}},$$

which is true. Thus the remaining term is negligible compared to the faster rate given in Theorem 3.2, and *a fortiori* to the other rates.

Now, we consider the main term, i.e.

$$\inf_{(m,a) \in \mathfrak{M}} \left\{ \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m, a) \right\} = \min_{a \in \{H, I, S\}} \inf_{m \in \mathcal{M}^a} \left\{ \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}^a(m) \right\}.$$

In the following Sections 5.3.1 (homogeneous case), 5.3.2 (intermediate case), 5.3.3 (sparse and frontier cases), we bound for each $a \in \{H, I, S\}$ the quantities $R(\mathcal{M}^a)$ and we prove that

$$\inf_{m \in \mathcal{M}^a} \left\{ \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}^a(m) \right\} \leq \begin{cases} v_H(\varepsilon) & \text{if } a = H \text{ and } r \geq p \\ v_I(\varepsilon) & \text{if } a = I \text{ and } \frac{p}{2s+1} < r < p \\ v_S(\varepsilon) & \text{if } a = S \text{ and } r \leq \frac{p}{2s+1} \end{cases}$$

where the $v_a(\varepsilon)$'s are the rates given in Theorem 3.2. This completes the proof. As each subsection deals with a different case, from now on, we drop the upperscript a for ease of notation. We also adapt the definitions of B_p and V_p to the weighted $\ell_p(w)$ -norm:

$$B_p(m) := \sum_{\lambda \in \Lambda \setminus m} w_{\lambda} |\theta_{\lambda}|^p \quad \text{and} \quad V_p(m) := \varepsilon^p \sum_{\lambda \in m} w_{\lambda} |\xi_{\lambda}|^p.$$

5.3.1. Proof of Theorem 3.2: homogeneous case

In this section we assume that $r \geq p$ and $s > 1/r - 1/p$. Let us recall our sub-collection of models. A model $m = \cup_{j=-1}^J m_j \in \mathcal{M} = \mathcal{M}^H$ if for some $0 \leq L \leq J$

$$\begin{aligned} \forall j \leq L, \quad m_j &= \Lambda_j = \{j\} \times K_j, \\ \forall j > L, \quad m_j &= \emptyset. \end{aligned}$$

Note that for any $m \in \mathcal{M}$,

$$\mathbb{E}[V_p(m)] = \varepsilon^p \sigma_p^p \sum_{j \geq -1} \omega_j |m_j| = \varepsilon^p \sigma_p^p \sum_{j=-1}^L \omega_j 2^j \leq C(p, \sigma_p) \varepsilon^p 2^{Lp/2},$$

with $C(p, \sigma_p)$ a constant only depending on p and σ_p . If θ belongs to $\mathcal{B}_{r,\infty}^s(R)$ we can prove that $B_p(m) \lesssim R^p 2^{-Lps}$. Indeed the bias verifies

$$B_p(m) = \sum_{(j,k) \notin m} \omega_j |\theta_{jk}|^p = \sum_{j > L, k \in K_j} 2^{j(\frac{p}{2}-1)} |\theta_{jk}|^p.$$

Since $r \geq p$, Hölder's inequality gives for any set E ,

$$\sum_{k \in E} |\theta_{jk}|^p \leq |E|^{(1-p/r)} \left(\sum_k |\theta_{jk}|^r \right)^{p/r}.$$

When $\theta \in \mathcal{B}_{r,\infty}^s(R)$, that yields $\sum_{k \in E} |\theta_{jk}|^p \leq |E|^{(1-p/r)} R^p 2^{-pj(s+\frac{1}{2}-\frac{1}{r})}$. Then, if $E = K_j$ with cardinal 2^j , it gives $\sum_{k \in K_j} |\theta_{jk}|^p \leq R^p 2^{j(-ps+1-\frac{p}{2})}$, and

$$B_p(m) = \sum_{j > L} 2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\theta_{jk}|^p \leq C(p, s) R^p 2^{-pLs},$$

with $C(p, s)$ a constant only depending on p and s . For our PCO procedure we have chosen $x_{m_j} = K \log |m_j|$ (with $\log 0 = 0$ and $K = p/2$) so that

$$\text{pen}^H(m) = \text{pen}(m) = \begin{cases} 2\varepsilon^p \sum_{j=-1}^J \omega_j p_j(m_j) & \text{if } p \leq 2 \\ 2\varepsilon^p \sum_{j=-1}^J \omega_j \min \left(p_j(m_j), (2q \log N)^{\frac{p}{2}-1} p_j^\#(m_j) \right) & \text{if } p > 2 \end{cases}$$

with

$$p_j(m_j) = \frac{3}{2} \sigma_p^p |m_j| + \kappa_p 2^{\frac{(p-2)+}{2}} |m_j|^{(1-\frac{p}{2})_+} + (K \log |m_j|)^{\frac{p}{2}}$$

and

$$p_j^\#(m_j) = \frac{3}{2} \sigma_2^2 |m_j| + \kappa_2 K \log |m_j|.$$

Thus, $p_j(m_j) = 0$ if $j > L$; and for $j \leq L$:

$$p_j(m_j) = \frac{3}{2} \sigma_p^p 2^j + \kappa_p 2^{\frac{(p-2)+}{2}} (2^j)^{(1-\frac{p}{2})_+} + (K j \log 2)^{\frac{p}{2}} \leq C(p, \sigma_p) 2^j,$$

with $C(p, \sigma_p)$ a constant only depending on p and σ_p . Then

$$\begin{aligned} \mathbb{E} \|\widehat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) &\leq \mathbb{E}[V_p(m)] + B_p(m) + \varepsilon^p \sum_{j=-1}^L \omega_j C(p, \sigma_p) 2^j \\ &\lesssim \varepsilon^p 2^{Lp/2} + R^p 2^{-Lps} + \varepsilon^p \sum_{j=-1}^L 2^{j\frac{p}{2}} \lesssim \varepsilon^p 2^{Lp/2} + R^p 2^{-Lps} \end{aligned}$$

which provides

$$\inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \|\widehat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) \right\} \leq C R^{\frac{p}{2s+1}} \varepsilon^{\frac{2sp}{2s+1}},$$

for C a constant, choosing L such that $2^L \approx (R/\varepsilon)^{2/(2s+1)}$ (possible since $2^J = N/2 \gtrsim (R/\varepsilon)^2$). Moreover, we compute

$$\begin{aligned} R_w(\mathcal{M}) &= \sum_{j=-1}^J \omega_j \sum_{m_j \in \mathcal{M}_j, m_j \neq \emptyset} |m_j|^{(1-p/2)_+} e^{-K \log |m_j|} \\ &= \sum_{j=-1}^J 2^{j(\frac{p}{2}-1)} \sum_{m_j \in \mathcal{M}_j} \mathbb{1}_{j \leq L} 2^{j(1-\frac{p}{2})_+} 2^{-jK} \\ &\leq \sum_{j \geq -1} 2^{j(\frac{p}{2}-1)} 2^{j[(1-\frac{p}{2})_+ - K]} = \sum_{j \geq -1} 2^{j[(\frac{p}{2}-1)_+ - K]} < \infty \end{aligned}$$

as soon as $K > (\frac{p}{2} - 1)_+$.

5.3.2. Proof of Theorem 3.2: intermediate case

In this section, we assume that $\frac{p}{2s+1} < r < p$ and $s > 1/r - 1/p$. Let us now consider the following model, inspired by [41]: m belongs to $\mathcal{M}(L)$ if

$$m_j = \begin{cases} \Lambda_j & \text{if } -1 \leq j \leq L-1 \\ m_{L+l} \subset \Lambda_{L+l} & \text{if } l = j - L \geq 0 \text{ with } |m_{L+l}| = \lfloor 2^{L+l} A(l) \rfloor \end{cases}$$

with $A(l) := 2^{-lp/2}(l+1)^{-3}$. At the end $\mathcal{M} = \bigcup_{L=0}^J \mathcal{M}(L)$.

Note that the cardinality $\lfloor 2^{L+l} A(l) \rfloor$ is equal to 0 when $2^{l(p/2-1)}(l+1)^3 > 2^L$. Then if $p \geq 2$, $m_j = \emptyset$ as soon as $j > L + l_{\max}$, with $l_{\max}$ such that

$$2^L 2^{-l_{\max}(p/2-1)} (l_{\max} + 1)^{-3} \approx 1.$$

Therefore, $l_{\max}$ is of order $L/(p/2-1)$ when $p > 2$ and of order $2^{L/3}$ if $p = 2$. When $p < 2$, we only have $l \leq J - L$.

To apply our model selection strategy, we set

$$(38) \quad x_{m_j} := K|m_j| \left(1 + \log \left(\frac{2^j}{|m_j|} \right) \right)$$

with K large enough (see later). Observe that at each level j , we consider two types of models. Either the model m_j is the whole slice $\{j\} \times K_j$, or it is a strict subset of this slice and in this case, it means that there exists $L \leq j$ such that

$$|m_j| = \lfloor 2^j A(j-L) \rfloor = \lfloor 2^j 2^{-(j-L)p/2} (j-L+1)^{-3} \rfloor.$$

Our choice of the factor x_{m_j} automatically adapts to both types of models:

$$(39) \quad \begin{cases} x_{m_j} = K|m_j| & \text{for the first type,} \\ x_{m_j} \approx K|m_j| \left(1 - \log A(j-L) \right) & \text{for the second type.} \end{cases}$$

In particular, for the first type $x_{m_j} = K2^j$ and for the second type x_{m_j} is of the same order as $K'|m_j| \times (j-L)$.

As explained in Section 5.3, we have to bound $R_w(\mathcal{M})$, that is to show that the term $\sum_{j=-1}^J \omega_j \sum_{m_j \in \mathcal{M}_j, m_j \neq \emptyset} |m_j|^{(1-\frac{p}{2})_+} e^{-x_{m_j}}$ is bounded. In the sequel, for the sake of simplicity, we set

$$b := \frac{p}{2} - 1.$$

Considering the two types of models, we have, with $\mathcal{M}_j(L) = \{m_j : m \in \mathcal{M}(L)\}$,

$$\begin{aligned} R_w(\mathcal{M}) &\leq \sum_{j=-1}^J \omega_j (2^j)^{(-b)_+} e^{-K2^j} + \sum_{L=0}^J \sum_{j=-1}^J \omega_j |\mathcal{M}_j(L)| |m_j|^{(-b)_+} e^{-K|m_j| (1 + \log (2^{(j-L)p/2} (j-L+1)^3))} \mathbf{1}_{j \geq L} \\ &=: T_1 + T_2. \end{aligned}$$

Since $\omega_j = 2^{j(\frac{p}{2}-1)} = 2^{jb}$,

$$T_1 \leq \sum_{j=-1}^{+\infty} 2^{jb} (2^j)^{(-b)_+} e^{-K2^j} < \infty,$$

for $K > 0$. Furthermore,

$$\begin{aligned} T_2 &\leq \sum_{L=0}^J \sum_{j=L}^J 2^{jb} |\mathcal{M}_j(L)| (2^j)^{(-b)_+} e^{-K|m_j| (1 + \log (2^{(j-L)p/2} (j-L+1)^3))} \\ &\leq \sum_{L=0}^J \sum_{l=0}^{J-L} 2^{(L+l)b_+} |\mathcal{M}_{L+l}(L)| e^{-K2^L 2^{-lb} (l+1)^{-3} (1 + \log (2^{lp/2} (l+1)^3))}. \end{aligned}$$

For $j \geq L$, the complexity of the collection at level $j = L + l$ is

$$\log |\mathcal{M}_j(L)| \leq \log \left(\frac{2^{L+l}}{|m_{L+l}|} \right) \leq |m_{L+l}| \log \left(\frac{e 2^{L+l}}{|m_{L+l}|} \right)$$

where we have used the bound $\log \binom{c}{d} \leq d \log \left(\frac{ec}{d} \right)$. Then

$$\begin{aligned} \log |\mathcal{M}_j(L)| &\leq 2^L 2^{-lb} (l+1)^{-3} \log \left(e 2^{L+l} \lfloor 2^{L+l} A(l) \rfloor^{-1} \right) \\ &\leq 2^L 2^{-lb} (l+1)^{-3} \log \left(\frac{e}{2^{-lp/2} (l+1)^{-3} - 2^{-(L+l)}} \right) \\ &\leq C 2^L 2^{-lb} (l+1)^{-2}, \end{aligned}$$

with C a constant only depending on p . Then we have

$$T_2 \leq \sum_{L=0}^J \sum_{l=0}^{J-L} 2^{(L+l)b_+} \exp \left((C - Kp \log(2)/2) 2^L 2^{-lb} (l+1)^{-2} \right).$$

Here we distinguish two cases.

Case $p \geq 2$: Recall that, if $l \geq 0$ and $p \geq 2$, $m_{L+l} = \emptyset$ if $l > l_{\max}$. Then we have, for K large enough such that $C - Kp \log(2)/2 < 0$,

$$\begin{aligned} T_2 &\leq \sum_{L=0}^J \sum_{l=0}^{l_{\max}} 2^{(L+l)b} \exp \left((C - Kp \log(2)/2) 2^L 2^{-lb} (l+1)^{-2} \right) \\ &\leq \sum_{L=0}^J \sum_{l=0}^{l_{\max}} 2^{(L+l)b} \exp \left((C - Kp \log(2)/2) 2^L 2^{-l_{\max}b} (l_{\max} + 1)^{-2} \right) \\ &\leq \sum_{L=0}^J \sum_{l=0}^{l_{\max}} 2^{(L+l)b} \exp \left((C - Kp \log(2)/2) (l_{\max} + 1) \right). \end{aligned}$$

We have used that

$$2^L 2^{-l_{\max}b} (l_{\max} + 1)^{-3} \approx 1.$$

For $p > 2$, $l_{\max} \approx L/b$, and for K constant large enough

$$T_2 \leq \sum_{L=0}^J 2^{Lb} \exp(-\tilde{K}L)$$

with $\tilde{K}$ as large as desired and $T_2 < \infty$.

For $p = 2$, $l_{\max} \approx 2^{L/3}$, and for K constant large enough $T_2 < \infty$.

Case $p < 2$: The function $l \mapsto 2^{-l(p/2-1)}(l+1)^{-2}$ is increasing except on a compact interval. Therefore, for K constant large enough,

$$\begin{aligned} T_2 &\leq \sum_{L=0}^J \sum_{l=0}^{J-L} \exp \left((C - Kp \log(2)/2) 2^L 2^{-lb} (l+1)^{-2} \right) \\ &\leq \sum_{L=0}^J \sum_{l=0}^{J-L} \exp(-K' 2^L) \lesssim J = \log_2(N/2), \end{aligned}$$

with K' a positive constant. Finally, we have proved that $R_w(\mathcal{M})$ is bounded by $\log N$ up to a constant. It means that in (37), the last term $\tilde{M}_p \varepsilon^p R_w(\mathfrak{M})$ is bounded by $\varepsilon^p |\log(\varepsilon)|$, which is negligible when compared to the rate.

It remains to bound $\inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \|\widehat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) \right\}$, with $\text{pen}(m) = 2\varepsilon^p \sum_j \omega_j p_j(m_j)$ and, with a slight abuse of notation,

$$p_j(m_j) = \begin{cases} \frac{3}{2} \sigma_p^p |m_j| + \kappa_p |m_j|^{1-\frac{p}{2}} x_{m_j}^{\frac{p}{2}} & \text{if } p \leq 2 \\ (2q \log N)^{\frac{p}{2}-1} \left(\frac{3}{2} \sigma_2^2 |m_j| + \kappa_2 x_{m_j} \right) & \text{if } p > 2 \end{cases}$$

where $q = p + 1$. Since $\mathcal{M} = \bigcup_{L=0}^J \mathcal{M}(L)$, we can write

$$(40) \quad \inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \|\widehat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) \right\} \leq \inf_L \left\{ \sup_{m \in \mathcal{M}(L)} \mathbb{E}[V_p(m)] + \inf_{m \in \mathcal{M}(L)} B_p(m) + \sup_{m \in \mathcal{M}(L)} \text{pen}(m) \right\}.$$

Let us study the three terms in the right hand side. Since $\omega_j = 2^{j(p/2-1)}$, we have for any $m \in \mathcal{M}(L)$,

$$\begin{aligned} \mathbb{E}[V_p(m)] &\leq \varepsilon^p \sigma_p^p \left(\sum_{j < L} 2^{j(p/2-1)} 2^j + \sum_{l \geq 0} 2^{(L+l)(\frac{p}{2}-1)} 2^L 2^{-l(p/2-1)} (l+1)^{-3} \right) \\ &\lesssim \varepsilon^p \sigma_p^p \left(2^{Lp/2} + 2^{L(p/2-1+1)} \right) \lesssim \varepsilon^p \sigma_p^p 2^{Lp/2} \end{aligned}$$

Therefore

$$\sup_{m \in \mathcal{M}(L)} \mathbb{E}[V_p(m)] \lesssim \varepsilon^p \sigma_p^p 2^{Lp/2}.$$

Moreover we can prove the following lemma (see Section 5.3.4)

Lemma 5.7. *If $\theta \in \mathcal{B}_{r,\infty}^s(R)$ and $\frac{p}{2s+1} < r$*

$$\inf_{m \in \mathcal{M}(L)} B_p(m) \lesssim R^p 2^{-spL}.$$

For the last term $\sup_{m \in \mathcal{M}(L)} \text{pen}(m)$, we distinguish two cases.

case $p > 2$: We have

$$p_j(m_j) \leq (2q \log N)^{\frac{p}{2}-1} \left(\frac{3}{2} \sigma_2^2 |m_j| + \kappa_2 x_{m_j} \right)$$

and $\sum_{j=1}^J \omega_j p_j(m_j)$ is bounded by (up to a constant):

$$\begin{aligned} &(\log N)^{\frac{p}{2}-1} \left[\sum_{j < L} 2^{jb} |m_j| + \sum_{j \geq L} 2^{jb} 2^{(1+b)L} 2^{-jb} (j-L+1)^{-3} ((j-L+1) + \log(j-L+1)) \right] \\ &\lesssim (\log N)^{\frac{p}{2}-1} 2^{Lp/2}. \end{aligned}$$

Finally equation (40) provides in the case $p > 2$

$$\inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \|\widehat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) \right\} \lesssim \inf_L \left\{ \varepsilon^p 2^{Lp/2} + R^p 2^{-spL} + \varepsilon^p (\log N)^{\frac{p}{2}-1} 2^{Lp/2} \right\}$$

Using that $\log(N) \lesssim |\log(\varepsilon)|$, and considering an L such that $2^L \approx \left(R^2 \varepsilon^{-2} |\log(\varepsilon)|^{\frac{2}{p}-1} \right)^{\frac{1}{2s+1}}$ (possible since $2^J \gtrsim R^2 \varepsilon^{-2}$), we obtain

$$\inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \|\widehat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) \right\} \lesssim R^{\frac{p}{2s+1}} (\varepsilon^{2p} |\log(\varepsilon)|^{p-2})^{\frac{s}{2s+1}}.$$

case $p \leq 2$: We have

$$p_j(m_j) = \frac{3}{2} \sigma_p^p |m_j| + \kappa_p |m_j|^{(1-\frac{p}{2})} x_{m_j}^{\frac{p}{2}}.$$

We just have to deal with the following term:

$$\begin{aligned} \sum_{j=1}^J \omega_j |m_j|^{(1-\frac{p}{2})} x_{m_j}^{\frac{p}{2}} &\lesssim \sum_{j < L} 2^{jb} |m_j| + \sum_{j \geq L} \omega_j |m_j| ((j-L+1) + \log(j-L+1))^{p/2} \\ &\lesssim 2^{Lp/2} + \sum_{j \geq L} 2^{jb} 2^{(1+b)L} 2^{-jb} (j-L+1)^{-3} ((j-L+1) + \log(j-L+1))^{p/2} \\ &\lesssim 2^{Lp/2}, \end{aligned}$$

since $3 - p/2 > 1$. Finally equation(40) provides in the case $p \leq 2$

$$\inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) \right\} \lesssim \inf_L \left\{ \varepsilon^p 2^{Lp/2} + R^p 2^{-spL} + \varepsilon^p 2^{Lp/2} \right\} \lesssim R^{\frac{p}{2s+1}} \varepsilon^{\frac{2sp}{2s+1}}.$$

5.3.3. Proof of Theorem 3.2: sparse and frontier case

In this section we assume that $r \leq \frac{p}{2s+1}$ and $s \geq 1/r$. Observe that $\frac{1}{r} \geq \frac{2s+1}{p}$ and hence $\frac{2s}{p} \leq \frac{1}{r} - \frac{1}{p} < s$ so that necessarily $p > 2$.

Recall that our model collection is defined by $\mathcal{M}_j = \{\{j\} \times E, E \in \mathcal{P}(K_j)\}$ for any $j \geq -1$. For this collection we choose $x_{m_j} = K|m_j||j|$ with $K = p+1 > 2 + (\frac{p}{2}-1)\log(2)$. As required in Section 5.3 let us bound $R(\mathcal{M}) = \sum_{j=-1}^J \omega_j \sum_{m_j \in \mathcal{M}_j, m_j \neq \emptyset} e^{-x_{m_j}}$. We can write

$$R_w(\mathcal{M}) = \sum_{j=-1}^J \omega_j \sum_{d=1}^{2^j} \sum_{m_j \in \mathcal{M}_j, |m_j|=d} e^{-x_{m_j}} = \sum_{j=-1}^J \omega_j \sum_{d=1}^{2^j} \text{card}\{m_j \in \mathcal{M}_j, |m_j|=d\} e^{-Kdj}.$$

Now we use that

$$\log \binom{2^j}{d} \leq d \left(1 + \log \left(\frac{2^j}{d} \right) \right) \leq 2dj$$

to state

$$\begin{aligned} R_w(\mathcal{M}) &\leq \sum_{j=-1}^J \omega_j \sum_{d=1}^{2^j} e^{2dj} e^{-Kdj} \leq \sum_{j=-1}^J \omega_j \sum_{d=1}^{\infty} e^{(2-K)dj} \\ &\leq \sum_{j=-1}^J \omega_j \frac{e^{(2-K)j}}{1 - e^{(2-K)j}} \lesssim \sum_{j=-1}^{\infty} e^{((\frac{p}{2}-1)\log(2)+2-K)j} < \infty. \end{aligned}$$

The following remark will be useful in Section 5.4.4.

Remark 5.8. We also have:

$$\begin{aligned} R_2^{2/p} &:= \sum_{j=-1}^J \left(\omega_j \sum_{m_j \in \mathcal{M}_j, m_j \neq \emptyset} e^{-x_{m_j}} \right)^{2/p} = \sum_{j=-1}^J \left(\omega_j \sum_{d=1}^{2^j} \sum_{m_j \in \mathcal{M}_j, |m_j|=d} e^{-Kdj} \right)^{2/p} \\ &\lesssim \sum_{j=-1}^J \left(e^{(2+(\frac{p}{2}-1)\log(2)-K)j} \right)^{2/p} < \infty \end{aligned}$$

It remains to control the term $\inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) \right\}$. Taking inspiration from the various works of Donoho, Johnstone, Kerkycharian and Picard, we now define

$$\check{m}_j = \begin{cases} \{j\} \times K_j & \text{if } j < j_1 \\ \{j\} \times \left\{ k \in K_j, |\theta_{jk}| > \varepsilon \sqrt{j} \right\} & \text{if } j_1 \leq j \leq j_0 \\ \emptyset & \text{if } j > j_0 \end{cases}$$

where $j_1 = j_1(\varepsilon)$ and $j_0 = j_0(\varepsilon)$ are defined by

$$2^{j_1} \approx (R^{-1}\varepsilon|\log \varepsilon|^{\frac{1}{2}})^{4\beta-2}, \quad 2^{j_0} \approx (R^{-1}\varepsilon|\log \varepsilon|^{\frac{1}{2}})^{-2\beta/s'}$$

with

$$s' = s - \frac{1}{r} + \frac{1}{p}, \quad \beta = \frac{s'}{2s+1-\frac{2}{r}}$$

so that

$$p - r + \frac{2\beta}{s'}r \left(s + \frac{1}{2} - \frac{p}{2r} \right) = 2p\beta.$$

Observe that since $p > 2$ and $s > 1/r - 1/p$, the quantities s' and β are positive and $\beta < 1/2$. Moreover one can see that $-2\beta/s' \leq 4\beta - 2$. Thus, since $R\varepsilon^{-1} \geq 1$,

$$1 \lesssim 2^{j_1} \lesssim 2^{j_0}.$$

In addition, inequality $s \geq 1/r$ gives $\beta \leq s'$ and

$$2^{j_0} \ll (R^{-1}\varepsilon)^{-2\beta/s'} = (R\varepsilon^{-1})^{2\beta/s'} \leq (R\varepsilon^{-1})^2 \lesssim N/2 = 2^J.$$

Remark 5.9. In this proof, the assumption $s \geq 1/r$ is used only in the previous line. If, instead, we assume only that $s > 1/r - 1/p$, the inequality $2^{j_0} \lesssim N/2 = 2^J$ remains valid, provided that $(R/\varepsilon)^{2/(1-2/p)} \lesssim N$.

Then, the model $\tilde{m} = \cup_j \tilde{m}_j$ belongs to $\mathcal{M}$ (even if $\tilde{m}$ depends on θ). It satisfies the following property, proved in Section 5.3.5.

Proposition 5.10. Assume that $r < p/(2s+1)$. There exists a positive constant C (depending on s, r, p) such that

$$\sup_{\theta \in \mathcal{B}_{r,\infty}^s(R)} \mathbb{E} \|\hat{\theta}^{(\tilde{m})} - \theta\|_{\ell_p(w)}^p \leq CR^{p(1-2\beta)} |\log \varepsilon|^{p\beta} \varepsilon^{2p\beta}$$

Moreover, if $r = p/(2s+1)$,

$$\sup_{\theta \in \mathcal{B}_{r,\infty}^s(R)} \mathbb{E} \|\hat{\theta}^{(\tilde{m})} - \theta\|_{\ell_p(w)}^p \leq CR^{p(1-2\beta)} |\log \varepsilon|^{p\beta+1} \varepsilon^{2p\beta}.$$

The following remark will be useful in Section 5.4.4.

Remark 5.11. We also have, for $r < p/(2s+1)$,

$$\sup_{\theta \in \mathcal{B}_{r,\infty}^s(R)} \sum_j \left(2^{j(\frac{p}{2}-1)} \sum_k \mathbb{E} |(\hat{\theta}^{(\tilde{m})} - \theta)_{jk}|^p \right)^{\frac{2}{p}} \leq CR^{2(1-2\beta)} |\log \varepsilon|^{2\beta} \varepsilon^{4\beta},$$

and for $r = p/(2s+1)$, the right hand side is replaced by $CR^{2(1-2\beta)} |\log \varepsilon|^{2\beta+1} \varepsilon^{4\beta}$.

Let us now bound, for $m = \tilde{m}$, the term

$$\text{pen}(m) = 2\varepsilon^p \sum_{j=-1}^J \omega_j \min \left(p_j(m_j), (2q \log N)^{\frac{p}{2}-1} p_j^\#(m_j) \right)$$

with

$$\begin{cases} p_j(m_j) = \frac{3}{2} \sigma_p^p |m_j| + \kappa_p 2^{\frac{(p-2)}{2}} (Kj |m_j|)^{\frac{p}{2}} \\ p_j^\#(m_j) = \frac{3}{2} \sigma_2^2 |m_j| + \kappa_2 Kj |m_j|. \end{cases}$$

Note that for $j > j_0$, $|m_j| = 0$ so that $\text{pen}(m) \leq 2\varepsilon^p (2q \log N)^{\frac{p}{2}-1} \sum_{j=-1}^{j_0} \omega_j p_j^\#(m_j)$. For $j < j_1$, $|m_j| = 2^j$ so that, since $|\log(R^{-1}\varepsilon)| \leq 2|\log \varepsilon|$,

$$\varepsilon^p \sum_{j=-1}^{j_1-1} \omega_j p_j^\#(\tilde{m}_j) \lesssim \varepsilon^p \sum_{j=-1}^{j_1-1} \omega_j |\tilde{m}_j| j \leq \varepsilon^p \sum_{j=-1}^{j_1-1} j 2^{jp/2} \lesssim \varepsilon^p 2^{j_1 p/2} j_1 \lesssim R^{p(1-2\beta)} \varepsilon^{2\beta p} |\log \varepsilon|^{p\beta-p/2+1}.$$

Then, since $\log(N) \lesssim |\log(\varepsilon)|$,

$$(\log N)^{\frac{p}{2}-1} \varepsilon^p \sum_{j=-1}^{j_1-1} \omega_j p_j^\#(\check{m}_j) \lesssim R^{p(1-2\beta)} \varepsilon^{2\beta p} |\log \varepsilon|^{p\beta}.$$

For $j_1 \leq j \leq j_0$, we write $|\check{m}_j| = \sum_{k \in K_j} \mathbb{1}_{|\theta_{jk}| > \varepsilon \sqrt{j}}$, thus

$$\varepsilon^p \sum_{j=j_1}^{j_0} \omega_j |\check{m}_j| j = \varepsilon^p \sum_{j=j_1}^{j_0} \omega_j j \sum_{k \in K_j} \mathbb{1}_{|\theta_{jk}| > \varepsilon \sqrt{j}}.$$

To control this term we use the same method and bounds as used for term A_{31} in the proof of Proposition 5.10. This gives

$$\begin{aligned} \varepsilon^p \sum_{j=j_1}^{j_0} \omega_j |\check{m}_j| j &\lesssim R^r \varepsilon^{p-r} j_0^{1-r/2} 2^{-j_0 r(s+\frac{1}{2}-\frac{p}{2r})} \\ &\lesssim R^r \varepsilon^{p-r} |\log \varepsilon|^{1-r/2} (R^{-1} \varepsilon |\log \varepsilon|^{\frac{1}{2}})^{\frac{2\beta}{s} r(s+\frac{1}{2}-\frac{p}{2r})} \\ &\lesssim R^{p-2p\beta} |\log \varepsilon|^{1-\frac{r}{2}+\frac{\beta}{s} r(s+\frac{1}{2}-\frac{p}{2r})} \varepsilon^{2p\beta}. \end{aligned}$$

Then, with $\log(N) \lesssim |\log(\varepsilon)|$,

$$(\log N)^{\frac{p}{2}-1} \varepsilon^p \sum_{j=j_1}^{j_0} \omega_j p_j^\#(\check{m}_j) \lesssim R^{p-2p\beta} \varepsilon^{2\beta p} |\log \varepsilon|^{\frac{p}{2}-\frac{r}{2}+\frac{\beta}{s} r(s+\frac{1}{2}-\frac{p}{2r})} \lesssim R^{p(1-2\beta)} \varepsilon^{2\beta p} |\log \varepsilon|^{p\beta}.$$

(In the frontier case $r = p/(2s+1)$, we obtain the bound

$$(\log N)^{\frac{p}{2}-1} \varepsilon^p \sum_{j=j_1}^{j_0} \omega_j p_j^\#(\check{m}_j) \lesssim R^{p(1-2\beta)} \varepsilon^{2\beta p} |\log \varepsilon|^{p/2-1+2-r/2}.$$

) Finally, recalling Proposition 5.10, we obtain

$$\inf_{m \in \mathcal{M}} \left\{ \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|_{\ell_p(w)}^p + \text{pen}(m) \right\} \leq \mathbb{E} \|\hat{\theta}^{(\check{m})} - \theta\|_{\ell_p(w)}^p + \text{pen}(\check{m}) \lesssim R^{p(1-2\beta)} (\varepsilon \sqrt{|\log \varepsilon|})^{2\beta p}.$$

(In the frontier case $r = p/(2s+1)$, we obtain the bound $R^{p(1-2\beta)} \varepsilon^{2\beta p} |\log \varepsilon|^{1+p\beta}$.)

5.3.4. Proof of Lemma 5.7

We sort the θ_{jk} 's in the following way: for any j we denote

$$|\theta_{j,(1)}| \geq |\theta_{j,(2)}| \geq \dots$$

Then

$$\inf_{m \in \mathcal{M}(L)} B_p(m) = \inf_{m \in \mathcal{M}(L)} \sum_{(j,k) \notin m} \omega_j |\theta_{jk}|^p = \sum_{l \geq 0} 2^{(L+l)(\frac{p}{2}-1)} \sum_{k > \lfloor 2^{L+l} A(l) \rfloor} |\theta_{L+l,(k)}|^p.$$

To bound the last term we use the following lemma.

Lemma 5.12. *If $a_{(1)} \geq \dots \geq a_{(N)} \geq 0$, then for any $0 < r \leq p$ and any $0 \leq n \leq N-1$, we have*

$$\sum_{k=n+1}^N a_{(k)}^p \leq \left(\sum_{i=1}^N a_i^r \right)^{p/r} (n+1)^{1-p/r}.$$

Proof. Let $a = a_{(n+1)}$. Then $a \leq a_{(j)}$ for any $j \leq n+1$ and then $(n+1)a^r \leq \sum_i a_i^r$. Therefore,

$$\sum_{k=n+1}^N a_{(k)}^p = \sum_{k=n+1}^N a_{(k)}^{p-r} a_{(k)}^r \leq a^{p-r} \sum_{k=n+1}^N a_{(k)}^r \leq \left(\frac{\sum_{i=1}^N a_i^r}{n+1} \right)^{p/r-1} \sum_{i=1}^N a_i^r.$$

Using Lemma 5.12, we can write

$$\begin{aligned} \inf_{m \in \mathcal{M}(L)} B_p(m) &\leq \sum_{l \geq 0} 2^{(L+l)(\frac{p}{2}-1)} \left(\sum_k |\theta_{L+l,k}|^r \right)^{\frac{p}{r}} ([2^{L+l} A(l)] + 1)^{1-\frac{p}{r}} \\ &\leq \sum_{l \geq 0} 2^{(L+l)(\frac{p}{2}-1)} \left(\sum_k |\theta_{L+l,k}|^r \right)^{\frac{p}{r}} (2^{L+l} 2^{-lp/2} (l+1)^{-3})^{1-\frac{p}{r}} \\ &\leq 2^{L(\frac{p}{2}-\frac{p}{r})} \sum_{l \geq 0} 2^{l(\frac{p}{2}-1)} \left(\sum_k |\theta_{L+l,k}|^r \right)^{\frac{p}{r}} (2^{l(1-p/2)} (l+1)^{-3})^{1-\frac{p}{r}}. \end{aligned}$$

Since $\theta \in \mathcal{B}_{r,\infty}^s(R)$, for all l

$$\left(\sum_k |\theta_{L+l,k}|^r \right)^{1/r} \leq R 2^{-(s+\frac{1}{2}-\frac{1}{r})(L+l)} \Rightarrow \left(\sum_k |\theta_{L+l,k}|^r \right)^{p/r} \leq R^p 2^{-(sp+\frac{p}{2}-\frac{p}{r})(L+l)}.$$

Finally

$$\inf_{m \in \mathcal{M}(L)} B_p(m) \leq R^p 2^{-spL} \sum_{l \geq 0} 2^{lp(\frac{p}{2}-\frac{1}{2}-s)} (l+1)^{3\frac{p}{r}-3} \leq C R^p 2^{-spL}.$$

The series converges because $\frac{p}{2r} - \frac{1}{2} - s < 0 \Leftrightarrow \frac{p}{2s+1} < r$. Lemma 5.7 is proved.

5.3.5. Proof of Proposition 5.10

We denote by $\hat{\theta}$ the estimator $\hat{\theta}^{(\tilde{m})}$. First observe that

$$\|\hat{\theta} - \theta\|_{\ell_p(w)}^p = \sum_j \sum_{k \in K_j} \omega_j |\hat{\theta}_{jk} - \theta_{jk}|^p = A_1 + A_2 + A_3,$$

with

$$A_1 := \sum_{j > j_0} \sum_{k \in K_j} \omega_j |\hat{\theta}_{jk} - \theta_{jk}|^p, \quad A_2 := \sum_{j < j_1} \sum_{k \in K_j} \omega_j |\hat{\theta}_{jk} - \theta_{jk}|^p, \quad A_3 := \sum_{j=j_1}^{j_0} \sum_{k \in K_j} \omega_j |\hat{\theta}_{jk} - \theta_{jk}|^p.$$

We bound each sum.

Since $\hat{\theta}_{jk} = 0$ for $j > j_0$,

$$A_1 = \sum_{j > j_0} \sum_{k \in K_j} \omega_j |\hat{\theta}_{jk} - \theta_{jk}|^p = \sum_{j > j_0} \omega_j \sum_{k \in K_j} |\theta_{jk}|^p.$$

Since $r \leq p$, we have $\sum_k |\theta_{jk}|^p \leq (\sum_k |\theta_{jk}|^r)^{p/r}$. Then, if $\theta \in \mathcal{B}_{r,\infty}^s(R)$

$$A_1 \leq R^p \sum_{j > j_0} 2^{j(\frac{p}{2}-1)} 2^{-jp(s+\frac{1}{2}-\frac{1}{r})} \leq R^p \sum_{j > j_0} 2^{-jp(s+\frac{1}{p}-\frac{1}{r})} \leq R^p 2^{-j_0 p s'}.$$

With the value of j_0 this gives

$$A_1 \leq R^{p(1-2\beta)} (\varepsilon |\log \varepsilon|^{\frac{1}{2}})^{2p\beta}.$$

Let us compute

$$A_2 = \sum_{j < j_1} \sum_{k \in K_j} \omega_j |\hat{\theta}_{jk} - \theta_{jk}|^p = \sum_{j < j_1} \omega_j \sum_{k \in K_j} |Y_{jk} - \theta_{jk}|^p = \varepsilon^p \sum_{j < j_1} 2^{j(p/2-1)} \sum_k |\xi_{jk}|^p.$$

Then $\mathbb{E}[A_2] = \varepsilon^p \sum_{j < j_1} 2^{jp/2} \sigma_p^p \lesssim \varepsilon^p 2^{j_1 p/2}$. With the value of j_1 this gives, since $\beta < 1/2$,

$$\mathbb{E}[A_2] \lesssim R^{p(1-2\beta)} \varepsilon^{2p\beta} |\log \varepsilon|^{p\beta-p/2} \lesssim R^{p(1-2\beta)} \varepsilon^{2p\beta}.$$

We can split the last term in the following way:

$$A_3 = \sum_{j=j_1}^{j_0} \sum_{k \in K_j} \omega_j |\hat{\theta}_{jk} - \theta_{jk}|^p = \underbrace{\sum_{j=j_1}^{j_0} \sum_k \omega_j |Y_{jk} - \theta_{jk}|^p \mathbb{1}_{|\theta_{jk}| > \varepsilon \sqrt{j}}}_{A_{31}} + \underbrace{\sum_{j=j_1}^{j_0} \sum_k \omega_j |\theta_{jk}|^p \mathbb{1}_{|\theta_{jk}| \leq \varepsilon \sqrt{j}}}_{A_{32}}.$$

Let us bound the expectation of the term A_{31} :

$$\begin{aligned} \mathbb{E}[A_{31}] &= \sum_{j=j_1}^{j_0} \sum_k \omega_j \mathbb{E} |Y_{jk} - \theta_{jk}|^p \mathbb{1}_{|\theta_{jk}| > \varepsilon \sqrt{j}} = \sum_{j=j_1}^{j_0} \sum_k \omega_j \varepsilon^p \sigma_p^p \mathbb{1}_{|\theta_{jk}| > \varepsilon \sqrt{j}} \\ &\leq \varepsilon^p \sigma_p^p \sum_{j=j_1}^{j_0} \sum_k \omega_j (\varepsilon \sqrt{j})^{-r} |\theta_{jk}|^r \end{aligned}$$

using that $\mathbb{1}_{|\theta_{jk}| > \varepsilon \sqrt{j}} \leq (|\theta_{jk}| / \varepsilon \sqrt{j})^r$. Now recall that θ belongs to $\mathcal{B}_{r,\infty}^s(R)$. Then

$$\begin{aligned} \mathbb{E}[A_{31}] &\leq \sigma_p^p \varepsilon^{p-r} \sum_{j=j_1}^{j_0} \omega_j j^{-r/2} \sum_k |\theta_{jk}|^r \leq R^r \sigma_p^p \varepsilon^{p-r} \sum_{j=j_1}^{j_0} 2^{j(p/2-1)} j^{-r/2} 2^{-jr(s+\frac{1}{2}-\frac{1}{r})} \\ &\lesssim R^r \varepsilon^{p-r} j_1^{-r/2} \sum_{j=j_1}^{j_0} 2^{-jr(s+\frac{1}{2}-\frac{p}{2r})}. \end{aligned}$$

In the sparse case, we have $s + \frac{1}{2} - \frac{p}{2r} < 0$. Thus, with the definition of j_0 :

$$\begin{aligned} \mathbb{E}[A_{31}] &\lesssim R^r \varepsilon^{p-r} j_1^{-r/2} 2^{-j_0 r(s+\frac{1}{2}-\frac{p}{2r})} \\ &\lesssim R^r \varepsilon^{p-r} |\log \varepsilon|^{-r/2} (R^{-1} \varepsilon |\log \varepsilon|^{\frac{1}{2}})^{\frac{2\beta}{s'} r(s+\frac{1}{2}-\frac{p}{2r})}. \end{aligned}$$

Note that in the frontier case $s + \frac{1}{2} - \frac{p}{2r} = 0$, we obtain

$$\mathbb{E}[A_{31}] \lesssim R^r \varepsilon^{p-r} j_1^{-r/2} j_0 \lesssim R^r \varepsilon^{p-r} |\log(\varepsilon)|^{1-r/2}.$$

It remains to control the term A_{32} , which is deterministic. Since $p - r > 0$

$$A_{32} := \sum_{j=j_1}^{j_0} \sum_k \omega_j |\theta_{jk}|^p \mathbb{1}_{|\theta_{jk}| \leq \varepsilon \sqrt{j}} \leq \sum_{j=j_1}^{j_0} \sum_k \omega_j |\theta_{jk}|^p (\varepsilon \sqrt{j} / |\theta_{jk}|)^{p-r}.$$

Now recall that θ belongs to $\mathcal{B}_{r,\infty}^s(R)$. Then

$$A_{32} \lesssim \varepsilon^{p-r} \sum_{j=j_1}^{j_0} j^{(p-r)/2} \omega_j \sum_k |\theta_{jk}|^r \lesssim R^r \varepsilon^{p-r} j_0^{(p-r)/2} \sum_{j=j_1}^{j_0} 2^{-jr(s+\frac{1}{2}-\frac{p}{2r})}.$$

In the sparse case, we have $s + \frac{1}{2} - \frac{p}{2r} < 0$. Thus, with the definition of j_0 :

$$A_{32} \lesssim R^r \varepsilon^{p-r} j_0^{(p-r)/2} 2^{-j_0 r(s+\frac{1}{2}-\frac{p}{2r})} \lesssim R^r \varepsilon^{p-r} |\log \varepsilon|^{\frac{p-r}{2}} (R^{-1} \varepsilon |\log \varepsilon|^{\frac{1}{2}})^{\frac{2\beta}{s'} r(s+\frac{1}{2}-\frac{p}{2r})}.$$

But remember that

$$p - r + \frac{2\beta}{s'} r \left(s + \frac{1}{2} - \frac{p}{2r} \right) = 2p\beta.$$

Then

$$\mathbb{E}[A_{31} + A_{32}] \lesssim R^{p-2p\beta} (\varepsilon |\log \varepsilon|^{\frac{1}{2}})^{2\beta p},$$

which concludes the proof for $r < p/(2s+1)$.

In the frontier case $s + \frac{1}{2} - \frac{p}{2r} = 0$, we obtain

$$A_{32} \lesssim R^r \varepsilon^{p-r} j_0^{(p-r)/2} j_0 \lesssim R^r \varepsilon^{p-r} |\log(\varepsilon)|^{1+(p-r)/2}$$

and then, using $p - r = 2p\beta$,

$$\mathbb{E}[A_{31} + A_{32}] \lesssim R^{p-2p\beta} \varepsilon^{2\beta p} |\log \varepsilon|^{1+\beta p}.$$

5.4. Proofs of results of Section 4

This section is devoted to the proofs of results of Section 4 and in particular to the proof of Theorem 4.1. We first prove the intermediate technical results stated in Section 4.

5.4.1. Proof of Proposition 4.2

In Section 4, we are faced with non-identically distributed variables. In this case, we use the following result, derived from Theorem 2.3.

Lemma 5.13. *Let $p \geq 1$. Assume that the ξ_λ 's are centered independent sub-Gaussian variables. We assume that there exists a positive constant τ such that for any $\lambda \in \mathcal{I}$, $\|\xi_\lambda\|_{\psi_2} \leq \tau$. Then, for all $\lambda \in \mathcal{I}$, $(\mathbb{E}|\xi_\lambda|^p)^{1/p} \leq 2\sqrt{p}\tau$. Moreover, denoting*

$$Z := \sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p, \quad \text{and} \quad \sigma_p := 2\sqrt{p}\tau$$

we have

$$\mathbb{E}(Z) \leq \sigma_p^p D,$$

with $D = \text{card}(\mathcal{I})$. Furthermore, for any $x \geq 1$, with probability larger than $1 - 2\exp(-x)$,

$$\sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p < \frac{3}{2} \sigma_p^p D + \kappa'_p D^{(1-\frac{p}{2})_+} x^{\frac{p}{2}},$$

where κ'_p is a positive constant only depending on p and τ .

Proof of Lemma 5.13. Using Lemma 5.2 and our assumption on uniform subgaussianity, we have:

$$\mathbb{E}(Z) = \sum_{\lambda \in \mathcal{I}} \mathbb{E}(|\xi_\lambda|^p) \leq \sum_{\lambda \in \mathcal{I}} (2\sqrt{p}\|\xi_\lambda\|_{\psi_2})^p \leq (2\sqrt{p}\tau)^p |\mathcal{I}| = \sigma_p^p D.$$

Now we apply Theorem 2.3. Recall that $b_\lambda = \|X_\lambda\|_{\psi_{2/p}}$ with $X_\lambda = |\xi_\lambda|^p - \mathbb{E}|\xi_\lambda|^p$. Lemma 5.3 gives:

$$\|b\|_{\ell_q}^q = \sum_{\lambda \in \mathcal{I}} \|X_\lambda\|_{\psi_{2/p}}^q \leq \sum_{\lambda \in \mathcal{I}} (K_{2/p} \|\xi_\lambda\|_{\psi_2}^p)^q \leq (K_{2/p} \tau^p)^q |\mathcal{I}|$$

so that $\|b\|_{\ell_q} \leq K_{2/p} \tau^p D^{1/q}$. Theorem 2.3 gives that with probability larger than $1 - 2e^{-x}$,

$$\begin{aligned} |Z - \mathbb{E}(Z)| &\leq d_{1,p} \|b\|_{\ell_2} \sqrt{x} + d_{2,p} \|b\|_{\ell_{1/(1-p/2)_+}} x^{p/2} \\ &\leq d_{1,p} K_{2/p} \tau^p \sqrt{Dx} + d_{2,p} K_{2/p} \tau^p D^{(1-p/2)_+} x^{p/2} \\ &\leq c'_{1,p} \sqrt{Dx} + c'_{2,p} D^{(1-p/2)_+} x^{p/2} \end{aligned}$$

with $c'_{i,p} = d_{i,p} K_{2/p} \tau^p$. Recalling that $\mathbb{E}(Z) \leq \sigma_p^p D$, we can obtain with probability larger than $1 - 2\exp(-x)$,

$$Z \leq \frac{3}{2} \sigma_p^p D + \kappa'_p D^{(1-\frac{p}{2})_+} x^{\frac{p}{2}},$$

with the same proof as the one of Corollary 2.4 replacing $c_{i,p}$ by $c'_{i,p} = d_{i,p} K_{2/p} \tau^p$. Here

$$(41) \quad \kappa'_p = c'_{2,p} + c'_{1,p} \max(1, c'_{1,p}/(2\sigma_p^p)) = K_{2/p} \tau^p \left[d_{2,p} + d_{1,p} \max\left(1, \frac{d_{1,p} K_{2/p}}{2(2\sqrt{p})^p}\right) \right].$$

The constants $d_{1,p}$, $d_{2,p}$ and $K_{2/p}$ are specified in Theorem 2.3 and depend only on p . $\square$

We now prove Proposition 4.2. Let us fix $j \geq -1$ and $k \in K_j$. Remember that

$$\xi_{jk} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i \varphi_{jk}(t_i).$$

The ξ_{jk} 's are centered sub-Gaussian random variables. Indeed, since the η_i 's are i.i.d. centered sub-Gaussian random variables, there exists a constant c such that for any $t \in \mathbb{R}$,

$$\mathbb{E}[\exp(t\eta_i)] \leq \exp(ct^2), \quad i = 1, \dots, n$$

(see Proposition 2.5.2 of [47]) and

$$\mathbb{E}[\exp(t\xi_{jk})] \leq \exp\left(ct^2 \frac{\sum_{i=1}^n \varphi_{jk}^2(t_i)}{n}\right), \quad t \in \mathbb{R}.$$

It remains to prove that $\frac{\sum_{i=1}^n \varphi_{jk}^2(t_i)}{n}$ is bounded by a constant independent of n and (j, k) .

In the sequel, we assume that $j \geq 0$ so that $\varphi_{jk} = \psi_{jk}$. Remember that the father and mother wavelets ϕ and ψ are assumed to be supported by the compact interval $[A, B]$. Therefore, if $\psi_{jk}(t_i) \neq 0$ then

$$A \leq 2^j \frac{i}{n} - k \leq B$$

and the size of the set of i 's such that $\psi_{jk}(t_i) \neq 0$ is not larger than $n2^{-j}$ up to a constant only depending on ψ . This yields that

$$\sum_{i=1}^n \frac{\psi_{jk}^2(t_i)}{n} = \frac{2^j}{n} \sum_{i=1}^n \psi^2\left(2^j \frac{i}{n} - k\right)$$

is bounded by a constant only depending on ψ . The proof for the case $j = -1$ is similar. This shows that the ξ_{jk} 's are sub-Gaussian variables. Moreover, using property 4 of Proposition 2.5.2 of [47], this ensures the existence of a positive constant τ only depending on ϕ , ψ and $\|\eta_1\|_{\psi_2}$ such that for all $j \geq -1$ and for all $k \in K_j$, $\|\xi_{jk}\|_{\psi_2} \leq \tau$.

We wish to apply Theorem 2.3. However, the ξ_{jk} 's are not independent. But, still assuming that $j \geq 0$, we have:

$$\xi_{j,k} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i \psi_{j,k}(t_i) = \frac{1}{\sqrt{n}} \sum_{i \in M_{jk}} \eta_i 2^{j/2} \psi(2^j t_i - k),$$

where

$$M_{jk} = \{i : A \leq 2^j t_i - k \leq B\} = \{i : 2^j t_i - B \leq k \leq 2^j t_i - A\}.$$

Finally, if $\xi_{j,k}$ depends on (t_i, η_i) , it means that $i \in M_{jk}$. Now, we take $k' > k + B - A$. If $i \in M_{jk}$ then $k \geq 2^j t_i - B$, which yields that $k' > 2^j t_i - A$ meaning that $i \notin M_{jk'}$. Therefore $\xi_{j,k'}$ does not depend on (t_i, η_i) . We conclude that if $k' > k + B - A$ then $\xi_{j,k}$ and $\xi_{j,k'}$ are independent. Now, we build a deterministic partition of $\mathcal{I}_j$

$$\mathcal{I}_j = \mathcal{I}_{j1} \cup \dots \cup \mathcal{I}_{jK},$$

where the sets $\mathcal{I}_{j\ell}$ are built so that if for some given ℓ , if we take two distinct elements k and k' of $\mathcal{I}_{j\ell}$, then $|k - k'| > B - A$. Therefore $\xi_{j,k}$ and $\xi_{j,k'}$ are independent. In particular, we can take K , the size of the partition, of order $B - A$ and the size of $\mathcal{I}_{j\ell}$ is smaller than $|\mathcal{I}_j|$.

Now, we apply Lemma 5.13, with $\tau = \sup_{jk} \|\xi_{jk}\|_{\psi_2}$. Let $x \geq 1$. For any $1 \leq \ell \leq K$

$$\mathbb{P}\left(\sum_{k \in \mathcal{I}_{j\ell}} |\xi_{jk}|^p \geq \frac{3}{2} \sigma_p^p |\mathcal{I}_{j\ell}| + \kappa'_p |\mathcal{I}_{j\ell}|^{(1-\frac{p}{2})} x^{\frac{p}{2}}\right) \leq 2e^{-x}.$$

Then, with probability larger than $1 - 2Ke^{-x}$

$$\sum_{k \in \mathcal{I}_j} |\xi_{jk}|^p = \sum_{\ell=1}^K \sum_{k \in \mathcal{I}_{j\ell}} |\xi_{jk}|^p < \frac{3}{2} \sigma_p^p \sum_{\ell=1}^K |\mathcal{I}_{j\ell}| + \kappa'_p \sum_{\ell=1}^K |\mathcal{I}_{j\ell}|^{(1-\frac{p}{2})} x^{\frac{p}{2}}$$

$$< \frac{3}{2} \sigma_p^p |\mathcal{I}_j| + \kappa'_p K^{1-(1-\frac{p}{2})_+} \left(\sum_{\ell=1}^K |\mathcal{I}_{j\ell}| \right)^{(1-\frac{p}{2})_+} x^{\frac{p}{2}}$$

using the concavity of $x \mapsto x^{(1-\frac{p}{2})_+}$. This gives

$$(42) \quad \mathbb{P} \left(\sum_{k \in \mathcal{I}_j} |\xi_{jk}|^p \geq \frac{3}{2} \sigma_p^p |\mathcal{I}_j| + \kappa'_p K^{1-(1-\frac{p}{2})_+} |\mathcal{I}_j|^{(1-\frac{p}{2})_+} x^{\frac{p}{2}} \right) \leq 2K e^{-x}.$$

The proof for the case $j = -1$ is similar.

5.4.2. Proof of Lemma 4.3

Observe that, since ϕ and ψ are assumed to be C^{M+1} and compactly supported, then Corollary 5.5.2 of [9] ensures that ψ is orthogonal to polynomials of degree less than or equal to M . Therefore Assumption 1 of [10] is satisfied. We then use the following result.

Proposition 5.14 (Proposition 2 of [10]). *We assume that ϕ and ψ are C^{M+1} and that $f \in \mathcal{B}_{r,\infty}^s(R)$, with $1/r < s < M+1$. Then, if $p \geq r$,*

$$\left\| \sum_{(j,k) \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \right\|_{\mathbb{L}_p} \lesssim R N^{-(s-1/r+1/p)},$$

where N is the cardinality of $\Lambda^{(N)}$.

Now we consider the three cases:

- If $r < \frac{p}{2s+1}$, since $s > 1/r$,

$$\begin{aligned} \left\| \sum_{(j,k) \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \right\|_{\mathbb{L}_p}^p &\lesssim R^p N^{-p(s-\frac{1}{r}+\frac{1}{p})} \lesssim R^p (|\log(\varepsilon)|\varepsilon^2)^{p(s-\frac{1}{r}+\frac{1}{p})} \\ &\lesssim R^p (|\log(\varepsilon)|\varepsilon^2)^{p \frac{s-\frac{1}{r}+\frac{1}{p}}{2s+1-\frac{2}{r}}}. \end{aligned}$$

- If $\frac{p}{2s+1} \leq r \leq p$, the inequalities $\frac{p}{2s+1} \leq r$ and $s > 1/r$ imply $s - 1/r + 1/p > s/(1+2s)$, so that

$$(43) \quad \left\| \sum_{(j,k) \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \right\|_{\mathbb{L}_p}^p \lesssim R^p (|\log(\varepsilon)|\varepsilon^2)^{p(s-\frac{1}{r}+\frac{1}{p})} \lesssim R^p \varepsilon^{\frac{2ps}{1+2s}}.$$

- If $r \geq p$, we have, since functions are compactly supported, and using (43)

$$\left\| \sum_{(j,k) \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \right\|_{\mathbb{L}_p} \lesssim \left\| \sum_{(j,k) \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \right\|_{\mathbb{L}_r} \lesssim R \varepsilon^{\frac{2s}{1+2s}}.$$

5.4.3. Proof of Lemma 4.4

In the sequel, we assume that $\|g\|_{\mathcal{B}_{p,p \wedge 2}^{p \wedge 2}}^{p \wedge 2} < \infty$. Section 9.2 of [9] shows that for $1 < p < \infty$

$$(44) \quad g \in \mathbb{L}_p \iff \left[\sum_{j \geq -1} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^2 \varphi_{jk}^2(\cdot) \right]^{\frac{1}{2}} \in \mathbb{L}_p$$

$$(45) \quad \iff \left[\sum_{j \geq -1} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^2 2^j 1_{[2^{-j}k, 2^{-j}(k+1)]}(\cdot) \right]^{\frac{1}{2}} \in \mathbb{L}_p$$

Case $1 < p \leq 2$. We use (44). Since $\|\cdot\|_{\ell_2} \leq \|\cdot\|_{\ell_p}$,

$$\left[\sum_{j \geq -1} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^2 \varphi_{jk}^2 \right]^{\frac{p}{2}} \leq \sum_{j \geq -1} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^p |\varphi_{jk}|^p$$

and

$$\begin{aligned} \|g\|_{\mathbb{L}_p}^p &\lesssim \int \left[\sum_{j \geq -1} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^2 \varphi_{jk}^2(x) \right]^{\frac{p}{2}} dx \leq \sum_{j \geq -1} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^p \int |\varphi_{jk}(x)|^p dx \\ &\lesssim \sum_{j \geq -1} 2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^p = \|g\|_{\mathcal{B}_{p,p}^0}^p < \infty. \end{aligned}$$

This shows that $\|g\|_{\mathbb{L}_p} \lesssim \|g\|_{\mathcal{B}_{p,p \wedge 2}^0} < \infty$.

Case $p \geq 2$. We use (45). We recall the generalized Minkowski inequality: Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be two σ -finite measure spaces, and F a measurable function. Then, for any $q \in [1, +\infty)$,

$$\left(\int_X \left(\int_Y |F(x, y)| d\nu(y) \right)^q d\mu(x) \right)^{\frac{1}{q}} \leq \int_Y \left(\int_X |F(x, y)|^q d\mu(x) \right)^{\frac{1}{q}} d\nu(y).$$

We apply this inequality with μ the Lebesgue measure and ν the counting measure. We take $q = p/2$ and $F = \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^2 2^j 1_{[2^{-j}k, 2^{-j}(k+1)]}$. We have

$$\begin{aligned} \|g\|_{\mathbb{L}_p}^2 &\lesssim \left(\int \left(\sum_{j \geq -1} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^2 2^j 1_{[2^{-j}k, 2^{-j}(k+1)]}(x) \right)^{\frac{p}{2}} dx \right)^{\frac{2}{p}} \\ &\leq \sum_{j \geq -1} \left(\int \left(\sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^2 2^j 1_{[2^{-j}k, 2^{-j}(k+1)]}(x) \right)^{\frac{p}{2}} dx \right)^{\frac{2}{p}} \\ &= \sum_{j \geq -1} \left(\int \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^p 2^{j\frac{p}{2}} 1_{[2^{-j}k, 2^{-j}(k+1)]}(x) dx \right)^{\frac{2}{p}} \leq \sum_{j \geq -1} \left(\sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^p 2^{j(\frac{p}{2}-1)} \right)^{\frac{2}{p}} \\ &\leq \sum_{j \geq -1} 2^{j2(\frac{1}{2}-\frac{1}{p})} \left(\sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|^p \right)^{\frac{2}{p}} = \|g\|_{\mathcal{B}_{p,2}^0}^2 < \infty. \end{aligned}$$

This shows that $\|g\|_{\mathbb{L}_p} \lesssim \|g\|_{\mathcal{B}_{p,p \wedge 2}^0} < \infty$.

Case $p = 1$. Finally, we deal with the case $p = 1$. We have $\|g\|_{\mathcal{B}_{p,p \wedge 2}^0}^{p \wedge 2} = \sum_{j \geq -1} 2^{-j/2} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle|$ and, with $g = \sum_{j \geq -1} \sum_{k \in K_j} \langle g, \varphi_{jk} \rangle \varphi_{jk}$,

$$\begin{aligned} \|g\|_{\mathbb{L}_1} &= \int \left| \sum_{j \geq -1} \sum_{k \in K_j} \langle g, \varphi_{jk} \rangle \varphi_{jk}(x) \right| dx \leq \sum_{j \geq -1} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle| \int |\varphi_{jk}(x)| dx \\ &\lesssim \sum_{j \geq -1} 2^{-j/2} \sum_{k \in K_j} |\langle g, \varphi_{jk} \rangle| = \|g\|_{\mathcal{B}_{1,1}^0} < \infty. \end{aligned}$$

This shows that $\|g\|_{\mathbb{L}_1} \lesssim \|g\|_{\mathcal{B}_{1,1}^0} < \infty$.

5.4.4. End of the proof of Theorem 4.1

The proof of Theorem 4.1 uses the following lemma which is a consequence of Proposition 2 of [10] and the assumption $f \in \mathcal{B}_{r,\infty}^s(R)$. It uses that $\log_2(n)$ is an integer.

Lemma 5.15. *If f belongs to the Besov set $\mathcal{B}_{r,\infty}^s(R)$, then we have*

$$(46) \quad \sum_{k \in K_j} |\theta_{jk}|^r \leq CR^r 2^{-jr(s+1/2-1/r)},$$

for C a constant only depending on ϕ, ψ, s and r .

Now, when $p \leq 2$, the proof of Theorem 4.1 follows from (32) combined with Proposition 4.2, Theorem 3.2, Lemmas 5.15 and 4.3 and Inequality (33).

When $p > 2$, Inequality (33) does not hold, but Inequality (34) shows that we only have to bound

$$(47) \quad \left[\sum_{j=-1}^J \left(\mathbb{E} \left[2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\hat{\theta}_{jk}^{(\hat{m})} - \theta_{jk}|^p \right] \right)^{\frac{2}{p}} \right]^{\frac{p}{2}}$$

to conclude. Remember that $\hat{m} = \arg \min_{m \in \mathcal{M}} \text{Crit}_w(m)$ with

$$\text{Crit}_w(m) = - \sum_{\lambda \in m} w_\lambda |Y_\lambda|^p + \text{pen}(m)$$

and $\text{pen}(m)$ has the form $\text{pen}(m) = 2\varepsilon^p \sum_{j=-1}^J \omega_j P_j(m_j)$ (see (22)). This can be rewritten

$$\text{Crit}_w(m) = - \sum_{j=-1}^J 2^{j(\frac{p}{2}-1)} \left[\sum_{k \in K_j} |Y_{jk}|^p - 2\varepsilon^p P_j(m_j) \right],$$

so that $\hat{m}$ is the union of disjoint sets $\hat{m}_j$ obtained by maximizing

$$m_j \mapsto \text{Crit}_j(m_j) := \sum_{k \in K_j} |Y_{jk}|^p - 2\varepsilon^p P_j(m_j).$$

Let m_j be some model in $\mathcal{M}_j$. We obtain directly from Proposition 2.1, for any model m_j ,

$$\|\tilde{\theta}_j - \theta_j\|_p^p \leq M_p \|\hat{\theta}^{(m_j)} - \theta_j\|_p^p + 2 \left[2V_p(\hat{m}_j) - \text{pen}(\hat{m}_j) \right] - 2 \left[2V_p(m_j) - \text{pen}(m_j) \right],$$

which means

$$\sum_{k \in K_j} |\tilde{\theta}_{jk} - \theta_{jk}|^p \leq M_p \sum_{k \in K_j} |\hat{\theta}_k^{(m_j)} - \theta_{jk}|^p + 4\varepsilon^p \left[\sum_{k \in \hat{m}_j} |\xi_{jk}|^p - P_j(\hat{m}_j) \right] - 4\varepsilon^p \left[\sum_{k \in m_j} |\xi_{jk}|^p - P_j(m_j) \right].$$

With $Z(m_j) = \sum_{k \in m_j} |\xi_{jk}|^p$, mimicking the proof of Theorem 2.8, by taking q large enough,

$$\begin{aligned} \sum_{j=-1}^J \left(2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} \mathbb{E} \left[|\tilde{\theta}_{jk} - \theta_{jk}|^p \right] \right)^{\frac{2}{p}} &\lesssim \sum_{j \geq -1} \left(2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} \mathbb{E} \left[|\hat{\theta}_k^{(m_j)} - \theta_{jk}|^p \right] + \varepsilon^p 2^{j(\frac{p}{2}-1)} P_j(m_j) \right)^{\frac{2}{p}} \\ &\quad + (\log N)^{1-\frac{2}{p}} \varepsilon^2 \sum_{j=-1}^J \left(2^{j(\frac{p}{2}-1)} \sum_{\substack{m_j \in \mathcal{M}_j \\ m_j \neq \emptyset}} e^{-x_{m_j}} \right)^{\frac{2}{p}} + O(\varepsilon^2). \end{aligned}$$

We have

$$\begin{aligned} 2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} |\hat{\theta}_k^{(m_j)} - \theta_{jk}|^p &= 2^{j(\frac{p}{2}-1)} \sum_{k \in m_j} |Y_{jk} - \theta_{jk}|^p + 2^{j(\frac{p}{2}-1)} \sum_{k \notin m_j} |0 - \theta_{jk}|^p \\ &=: V_p(m_j) + B_p(m_j), \end{aligned}$$

with

$$(48) \quad B_p(m_j) := 2^{j(\frac{p}{2}-1)} \sum_{k \notin m_j} |\theta_{jk}|^p \quad \text{and} \quad V_p(m_j) := \varepsilon^p \sum_{k \in m_j} 2^{j(\frac{p}{2}-1)} |\xi_{jk}|^p.$$

This gives $\mathbb{E}[V_p(m_j)] \leq \varepsilon^p \sigma_p^p 2^{j(\frac{p}{2}-1)} |m_j|$ and

$$A(m) := \sum_{j=-1}^J \left(2^{j(\frac{p}{2}-1)} \sum_{k \in K_j} \mathbb{E} \left[|\hat{\theta}_k^{(m_j)} - \theta_{jk}|^p \right] + \varepsilon^p 2^{j(\frac{p}{2}-1)} P_j(m_j) \right)^{\frac{2}{p}}$$

$$\begin{aligned}
&\lesssim \sum_{j=-1}^J \left(B_p(m_j) + \varepsilon^p \sigma_p^p 2^{j(\frac{p}{2}-1)} |m_j| + \varepsilon^p 2^{j(\frac{p}{2}-1)} P_j(m_j) \right)^{\frac{2}{p}} \\
&\lesssim \sum_{j=-1}^J B_p^{\frac{2}{p}}(m_j) + \sum_{j=-1}^J \left(\varepsilon^p \sigma_p^p 2^{j(\frac{p}{2}-1)} |m_j| \right)^{\frac{2}{p}} + \sum_{j=-1}^J \left(\varepsilon^p 2^{j(\frac{p}{2}-1)} P_j(m_j) \right)^{\frac{2}{p}} \\
&=: \tilde{B}_p(m) + \tilde{V}_p(m) + \tilde{P}_p(m).
\end{aligned}$$

Then we have to control for each collection $\mathcal{M}$ (Homogeneous, Intermediate, Sparse) the terms $\tilde{B}_p(m)$, $\tilde{V}_p(m)$, $\tilde{P}_p(m)$ for $m \in \mathcal{M}$, as well as

$$\tilde{R}(\mathcal{M}) := (\log N)^{1-\frac{2}{p}} \varepsilon^2 \sum_{j=-1}^J \left(2^{j(\frac{p}{2}-1)} \sum_{\substack{m_j \in \mathcal{M}_j \\ m_j \neq \emptyset}} e^{-x_{m_j}} \right)^{\frac{2}{p}}.$$

Thus, the proof is similar to that of Theorem 3.2, but with the sum in j in a different position, i.e. terms of type $\sum_j z_j$ are replaced by $(\sum_j z_j^{2/p})^{p/2}$.

- In the Homogeneous case, recall that a model $m = \cup_{j=-1}^J m_j$ belongs to $\mathcal{M}$ if for some $L \leq J$

$$\forall j \leq L, m_j = \{j\} \times K_j, \quad \forall j > L, m_j = \emptyset.$$

Then we can prove that $\tilde{V}_p(m) \lesssim \varepsilon^2 2^L$, $\tilde{B}_p(m) \lesssim R^2 2^{-2Ls}$ for $\theta \in \mathcal{B}_{r,\infty}^s(R)$ and $\tilde{P}_p(m) \leq \varepsilon^2 2^L$. Moreover $\tilde{R}(\mathcal{M}) \lesssim (\log N)^{1-\frac{2}{p}} \varepsilon^2$. This allows us to conclude.

- We now deal with the Intermediate case ($\frac{p}{2s+1} < r < p$). For this case, we slightly modify the previous collection of models: we still have $\mathcal{M} = \cup_{L=0}^J \mathcal{M}(L)$ and m belongs to $\mathcal{M}(L)$ if

$$m_j = \begin{cases} \Lambda_j & \text{if } -1 \leq j \leq L-1 \\ m_{L+l} \subset \Lambda_{L+l} & \text{if } l = j-L \geq 0 \text{ with } |m_{L+l}| = \lfloor 2^{L+l} \tilde{A}(l) \rfloor \end{cases}$$

but this time $\tilde{A}(l) := 2^{-lp/2} (l+1)^{-3p/2}$. We follow Section 5.3.2 step by step, keeping in mind that $p > 2$ and $b = \frac{p}{2} - 1$.

We have $\tilde{R}(\mathcal{M}) = (\log N)^{1-\frac{2}{p}} \varepsilon^2 (T_1 + T_2)$ with $T_1 = \sum_{j=-1}^J \left[2^{jb} |m_j|^{(-b)+} e^{-K|m_j|} \right]^{\frac{2}{p}} < \infty$ and

$$\begin{aligned}
T_2 &\leq \sum_{L=0}^J \sum_{j=L}^J \left[2^{jb} |\mathcal{M}_j(L)| e^{-K|m_j|} \left(1 + \log \left(2^{(j-L)p/2} (j-L+1)^{3p/2} \right) \right) \right]^{\frac{2}{p}} \\
&\leq \sum_{L=0}^J \sum_{l=0}^{J-L} \left[2^{(L+l)b} |\mathcal{M}_{L+l}(L)| e^{-K 2^L 2^{-lb} (l+1)^{-3p/2}} \left(1 + \log \left(2^{lp/2} (l+1)^{3p/2} \right) \right) \right]^{\frac{2}{p}},
\end{aligned}$$

with $\log |\mathcal{M}_j(L)| \leq C 2^L 2^{-lb} (l+1)^{-p}$ for $j \geq L$. Then, for K large enough

$$\begin{aligned}
T_2 &\leq \sum_{L=0}^J \sum_{l=0}^{l_{\max}} \left[2^{(L+l)b} \exp \left((C - Kp \log(2)/2) 2^L 2^{-lb} (l+1)^{-p} \right) \right]^{\frac{2}{p}} \\
&\leq \sum_{L=0}^J \sum_{l=0}^{l_{\max}} 2^{2(L+l)b/p} \exp \left(\frac{2}{p} (C - Kp \log(2)/2) 2^L 2^{-l_{\max}b} (l_{\max} + 1)^{-p} \right) \\
&\leq \sum_{L=0}^J \sum_{l=0}^{l_{\max}} 2^{2(L+l)b/p} \exp \left(\frac{2}{p} (C - Kp \log(2)/2) (l_{\max} + 1)^{p/2} \right) \lesssim \sum_{L=0}^J 2^{2Lb/p} \exp(-\tilde{K}L) < \infty
\end{aligned}$$

where we have used that $2^L 2^{-l_{\max} b} (l_{\max} + 1)^{-3p/2} \approx 1$ and $l_{\max} \approx L/b$. Thus $\tilde{R}(\mathcal{M}) \lesssim (\log N)^{1-\frac{2}{p}} \varepsilon^2$. We now deal with $\tilde{V}_p(m)$: for any $m \in \mathcal{M}(L)$,

$$\begin{aligned} \tilde{V}_p(m) &= \sum_{j=-1}^J \left(\varepsilon^p \sigma_p^p 2^{j(\frac{p}{2}-1)} |m_j| \right)^{\frac{2}{p}} \\ &\leq \varepsilon^2 \sigma_p^2 \left(\sum_{j < L} \left(2^{j(p/2-1)} 2^j \right)^{\frac{2}{p}} + \sum_{l \geq 0} \left(2^{(L+l)(\frac{p}{2}-1)} 2^L 2^{-l(p/2-1)} (l+1)^{-3p/2} \right)^{\frac{2}{p}} \right) \lesssim \varepsilon^2 2^L. \end{aligned}$$

For the study of $\tilde{B}_p(m)$, we follow the proof of Lemma 5.7 *mutatis mutandis*:

$$\begin{aligned} \inf_{m \in \mathcal{M}(L)} \tilde{B}_p(m) &= \sum_{l \geq 0} \left(2^{(L+l)(\frac{p}{2}-1)} \sum_{k > \lfloor 2^{L+l} \tilde{A}(l) \rfloor} |\theta_{L+l,k}|^p \right)^{\frac{2}{p}} \\ &\leq \sum_{l \geq 0} \left(2^{(L+l)(\frac{p}{2}-1)} \left(\sum_k |\theta_{L+l,k}|^r \right)^{\frac{p}{r}} (\lfloor 2^{L+l} \tilde{A}(l) \rfloor + 1)^{1-\frac{p}{r}} \right)^{\frac{2}{p}} \\ &\leq \sum_{l \geq 0} \left(2^{(L+l)(\frac{p}{2}-1)} \left(\sum_k |\theta_{L+l,k}|^r \right)^{\frac{p}{r}} (2^L 2^{-l(\frac{p}{2}-1)} (l+1)^{-3p/2})^{1-\frac{p}{r}} \right)^{\frac{2}{p}} \\ &\leq \left(2^{L(\frac{p}{2}-\frac{p}{r})} \right)^{\frac{2}{p}} \sum_{l \geq 0} \left(2^{l(\frac{p}{2}-1)\frac{p}{r}} (l+1)^{\frac{3p}{2}(\frac{p}{r}-1)} \left(\sum_k |\theta_{L+l,k}|^r \right)^{p/r} \right)^{\frac{2}{p}}. \end{aligned}$$

Since $\theta \in \mathcal{B}_{r,\infty}^s(R)$, for all l : $(\sum_k |\theta_{L+l,k}|^r)^{p/r} \leq R^p 2^{-(sp+\frac{p}{2}-\frac{p}{r})(L+l)}$. Finally

$$\begin{aligned} \inf_{m \in \mathcal{M}(L)} \tilde{B}_p(m) &\leq R^2 2^{L(1-\frac{2}{r})} \sum_{l \geq 0} 2^{l(\frac{p}{2}-1)\frac{2}{r}} (l+1)^{3(\frac{p}{r}-1)} 2^{-(2s+1-\frac{2}{r})(L+l)} \\ &\leq R^2 2^{-2sL} \sum_{l \geq 0} 2^{2l(\frac{p}{2r}-\frac{1}{2}-s)} (l+1)^{3(\frac{p}{r}-1)} \lesssim R^2 2^{-2sL} \end{aligned}$$

since the series converges because $\frac{p}{2r} - \frac{1}{2} - s < 0 \Leftrightarrow \frac{p}{2s+1} < r$.

It remains to control $\tilde{P}_p(m) = \sum_{j=-1}^J \left(\varepsilon^p 2^{jb} P_j(m_j) \right)^{2/p}$ with P_j given through (22). We have

$$P_j(m_j) \leq (2q \log N)^{\frac{p}{2}-1} \left(\frac{3}{2} \sigma_2^2 |m_j| + \kappa_2 x_{m_j} \right)$$

so that, for any $m \in \mathcal{M}(L)$,

$$\begin{aligned} \tilde{P}_p(m) &\lesssim \varepsilon^2 (\log N)^{1-\frac{2}{p}} \left[\sum_{j < L} \left(2^{jb} |m_j| \right)^{\frac{2}{p}} + \sum_{j \geq L} \left(2^{jb} 2^{(1+b)L} 2^{-jb} (j-L+1)^{-\frac{3p}{2}} ((j-L+1) + \log(j-L+1)) \right)^{\frac{2}{p}} \right] \\ &\lesssim \varepsilon^2 (\log N)^{1-\frac{2}{p}} \left[2^L + 2^L \sum_{l \geq 0} (l+1)^{-3} (l+1 + \log(l+1))^{\frac{2}{p}} \right] \lesssim \varepsilon^2 (\log N)^{1-\frac{2}{p}} 2^L. \end{aligned}$$

We conclude

$$\begin{aligned} \inf_{m \in \mathcal{M}} A(m) &\leq \inf_L \left\{ \inf_{m \in \mathcal{M}(L)} \tilde{B}_p(m) + \sup_{m \in \mathcal{M}(L)} \tilde{V}_p(m) + \sup_{m \in \mathcal{M}(L)} \tilde{P}_p(m) \right\} \\ &\lesssim \inf_L \left\{ R^2 2^{-2sL} + \varepsilon^2 (\log N)^{1-\frac{2}{p}} 2^L \right\} \\ &\lesssim R^{\frac{2}{1+2s}} \varepsilon^{\frac{4s}{1+2s}} |\log \varepsilon|^{\frac{2s(p-2)}{p(1+2s)}}, \end{aligned}$$

by taking L such that

$$2^L \approx (\varepsilon^{-1} R)^{\frac{2}{1+2s}} |\log \varepsilon|^{-\frac{(p-2)}{p(1+2s)}}.$$

• In the Sparse case, Remark 5.8 gives $\tilde{R}(\mathcal{M}) \lesssim (\log N)^{1-\frac{2}{p}} \varepsilon^2$ and Remark 5.11 provides

$$\sup_{\theta \in \mathcal{B}_{r,\infty}^s(R)} \left(\tilde{B}_p(m) + \tilde{V}_p(m) \right) \lesssim R^{2(1-2\beta)} |\log \varepsilon|^{2\beta} \varepsilon^{4\beta}.$$

Finally, following the outlines of the end of Section 5.3.3, we can prove

$$\tilde{P}_p(m) \lesssim \left(R^{p(1-2\beta)} \varepsilon^{2\beta p} |\log \varepsilon|^{p\beta} \right)^{2/p}.$$

In the frontier case, $r = p/(2s+1)$, we obtain

$$\sup_{\theta \in \mathcal{B}_{r,\infty}^s(R)} \left(\tilde{B}_p(m) + \tilde{V}_p(m) \right) + \tilde{P}_p(m) \lesssim R^{2(1-2\beta)} |\log \varepsilon|^{2\beta+1} \varepsilon^{4\beta}.$$

This ends the proof.

Acknowledgments

We are grateful to Professor Radosław Adamczak for his advice about concentration inequalities for sub-Weibull variables. We thank the anonymous referees for valuable comments and suggestions leading to improvements of the article.

References

- [1] AGAPIOU, S., CASTILLO, I. and EGELS, P. (2025). Heavy-tailed and Horseshoe priors for regression and sparse Besov rates. arXiv:2505.15543.
- [2] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)* 267–281. Akad. Kiadó, Budapest. [MR483125](#)
- [3] BARAUD, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields* **151** 353–401. [https://doi.org/10.1007/s00440-010-0302-y](#) [MR2834722](#)
- [4] BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields* **71** 271–291. [https://doi.org/10.1007/BF00332312](#) [MR816706](#)
- [5] BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** 203–268. [https://doi.org/10.1007/s100970100031](#) [MR1848946](#)
- [6] BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. [https://doi.org/10.1007/s00440-006-0011-8](#) [MR2288064](#)
- [7] BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137. [https://doi.org/10.1007/BF00535278](#) [MR523165](#)
- [8] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81. [https://doi.org/10.1006/acha.1993.1005](#) [MR1256527](#)
- [9] DAUBECHIES, I. (1992). *Ten lectures on wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics **61**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. [https://doi.org/10.1137/1.9781611970104](#) [MR1162107](#)
- [10] DELYON, B. and JUDITSKY, A. (1997). On the computation of wavelet coefficients. *J. Approx. Theory* **88** 47–79. [https://doi.org/10.1006/jath.1996.3008](#) [MR1426459](#)
- [11] DEVROYE, L. and LUGOSI, G. (1996). A universally acceptable smoothing factor for kernel density estimates. *Ann. Statist.* **24** 2499–2512. [https://doi.org/10.1214/aos/1032181164](#) [MR1425963](#)
- [12] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921.
- [13] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1993). Wavelet Shrinkage: Asymptopia? Technical Report No. EFS NSF 419, Department of Statistics, Stanford University. Stanford Digital Repository.
- [14] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological)* **57** 301–337.
- [15] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539.
- [16] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1997). Universal near minimaxity of wavelet shrinkage. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* 183–218. Springer.
- [17] GAO, C., HAN, F. and ZHANG, C. H. (2020). On estimation of isotonic piecewise constant signals. *Annals of Statistics* **48** 629–654.
- [18] GLUSKIN, E. D. and KWAPIEŃ, S. (1995). Tail and moment estimates for sums of independent random variables with logarithmically concave tails. *Studia Mathematica* **114** 303–309. [https://doi.org/10.4064/sm-114-3-303-309](#)
- [19] GOLDENSHLUGER, A. (2009). A universal procedure for aggregating estimators. *Ann. Statist.* **37** 542–568. [https://doi.org/10.1214/00-AOS576](#) [MR2488362](#)

- [20] GOLDENSHLUGER, A. and LEPSKI, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150–1190. <https://doi.org/10.3150/08-BEJ144> MR2543590 (2010m:62097)
- [21] GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. <https://doi.org/10.1214/11-AOS883> MR2850214
- [22] GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on R^d . *Probab. Theory Related Fields* **159** 479–543. <https://doi.org/10.1007/s00440-013-0512-1> MR3230001
- [23] GOLDENSHLUGER, A. V. and LEPSKI, O. V. (2013). General selection rule from a family of linear estimators. *Theory Probab. Appl.* **57** 209–226. <https://doi.org/10.1137/S0040585X97985923> MR3201652
- [24] HALL, P., KERKYACHARIAN, G. and PICARD, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica* 33–49.
- [25] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, approximation, and statistical applications. Lecture Notes in Statistics* **129**. Springer-Verlag, New York. <https://doi.org/10.1007/978-1-4612-2222-4> MR1618204
- [26] HITCZENKO, P., MONTGOMERY-SMITH, S. J. and OLESZKIEWICZ, K. (1997). Moment inequalities for sums of certain independent symmetric random variables. *Studia Math* **123** 15–42.
- [27] IBRAGIMOV, I. A. and HASMINSKI, R. Z. (1980). An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **98** 61–85, 161–162, 166. Studies in mathematical statistics, IV. MR591862
- [28] JOHNSTONE, I. M. (2019). Gaussian Estimation: Sequence and Wavelet Models. Draft manuscript, available from the author’s webpage.
- [29] JOHNSTONE, I. M. and SILVERMAN, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752. <https://doi.org/10.1214/009053605000000345> MR2166560
- [30] JUDITSKY, A. (1997). Wavelet estimators: adapting to unknown smoothness. *Mathematical Methods of Statistics* **6** 1–25.
- [31] KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121** 137–170. <https://doi.org/10.1007/PL00008800> MR1863916
- [32] KERKYACHARIAN, G. and PICARD, D. (2000). Thresholding algorithms, maxisets and well-concentrated bases. *Test* **9** 283–344.
- [33] KERKYACHARIAN, G., PICARD, D. and TRIBOULEY, K. (1996). L^p adaptive density estimation. *Bernoulli* **2** 229–247.
- [34] KUCHIBHOTLA, A. K. and CHAKRABORTY, A. (2022). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA* **11** 1389–1456.
- [35] LACOUR, C., MASSART, P. and RIVOIRARD, V. (2017). Estimator selection: a new method with applications to kernel density estimation. *Sankhya A* **79** 298–335. <https://doi.org/10.1007/s13171-017-0107-5> MR3707423
- [36] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. <https://doi.org/10.1214/aos/1015957395>
- [37] LEPSKI, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *The Annals of Statistics* **43** 1178 – 1242. <https://doi.org/10.1214/14-AOS1306>
- [38] LEPSKI, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teoriya Veroyatnostei i ee Primeneniya* **36** 645–659.
- [39] 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. Statist.* **25** 929–947. <https://doi.org/10.1214/aos/1069362731>
- [40] MALLOWS, C. L. (1973). Some Comments on Cp. *Technometrics* **15** 661–675.
- [41] MASSART, P. (2007). *Concentration Inequalities and Model Selection (Saint-Flour Summer School on Probability Theory 2003, ed. J. Picard)*. Springer.
- [42] NEMIROVSKIY, A. S. (1985). Nonparametric estimation of smooth regression functions. *Soviet J. Comput. Systems Sci.* **23** 1–11. MR844292
- [43] RIVOIRARD, V. (2004). Maxisets for linear procedures. *Statistics & probability letters* **67** 267–275.
- [44] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR468014
- [45] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR673642
- [46] VARET, S., LACOUR, C., MASSART, P. and RIVOIRARD, V. (2023). Numerical performance of penalized comparison to overfitting for multivariate kernel density estimation. *ESAIM Probab. Stat.* **27** 621–667. <https://doi.org/10.1051/ps/2022018> MR4602387
- [47] VERSHYNIN, R. (2018). *High-Dimensional Probability: An introduction with applications in data science* **47**. Cambridge University Press.
- [48] ZHANG, H. and WEI, H. (2022). Sharper sub-Weibull concentrations. *Mathematics* **10** 2252.