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

Claire Lacour[1], Pascal Massart[2], and Vincent Rivoirard[3,2]

[1]Univ Gustave Eiffel, Univ Paris Est Creteil, CNRS, LAMA UMR8050, F-77447, Marne-la-Vallée, France
[2]Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.
[3]CEREMADE, CNRS, Université Paris-Dauphine, Université PSL, 75016 PARIS, FRANCE

April 15, 2025

Abstract

This paper addresses the problem of model selection in the sequence model $Y = \theta + \varepsilon\xi$ when ξ is sub-Gaussian for non-euclidian loss-functions. In this model, the Penalized Comparison to Overfitting procedure is studied for the weighted 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 stated for Besov bodies $\mathcal{B}^s_{r,\infty}$. These results are applied to the functional model of nonparametric regression.

MSC Classification:
- 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 be typically 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 $\left(\widehat{f}_m\right)_{m\in\mathcal{M}}$ and then try to design some genuine data-driven procedure $\widehat{m} \in \mathcal{M}$ to produce a new estimator $\widehat{f}_{\widehat{m}}$. Considering some loss function ℓ, we measure the quality of each estimator $\widehat{f}_m$, through the quantity $\ell\big(f, \widehat{f}_m\big)$ and mathematical results on estimator selection are formulated in terms of upper bounds on $\ell\big(f, \widehat{f}_{\widehat{m}}\big)$ that allow to measure how far this quantity is from what is usually called the *oracle* risk $\inf_{m\in\mathcal{M}} \mathbb{E}\,\ell\big[\big(f, \widehat{f}_m\big)\big]$. These comparison inequalities are called oracle inequalities. The quadratic loss $\ell\big(f, \widehat{f}_m\big) = \left\|f - \widehat{f}_m\right\|^2$ 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 [BM01], [BM07], [Mas07], 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

$$g \in \mathcal{S} \longmapsto \mathbb{E}\left[L_n\left(g\right)\right]$$

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 $\widehat{f}_m$ of f is obtained by minimizing L_n over S_m ($\widehat{f}_m$ is called *minimum contrast estimator* or *empirical risk minimizer*). In the case where $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$, an empirical criterion L_n can be defined as an empirical mean

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

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

$$L_n\left(\widehat{f}_m\right) + \mathrm{pen}(m) \tag{1}$$

over $\mathcal{M}$. We can then define the selected model $S_{\widehat{m}}$ and the corresponding selected estimator $\widehat{f}_{\widehat{m}}$. Penalized criteria have been proposed in the early seventies by Akaike or Schwarz (see [Aka73] and [Sch78]) for penalized maximum log-likelihood in the density estimation framework and Mallows for penalized least squares regression (see [Mal73]). 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" (non negative) loss function ℓ attached to L_n through the simple definition

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

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 [Mas07] and references therein. Although it is not derived from an empirical contrast, we mention the model selection work associated with the Hellinger distance [Bir86], [Bar11]. 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-Euclidian loss functions, 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 [BH79, IH80, Sto82], an important advance was made by [Nem85] 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 works have been made by Oleg Lepski, who proposed the so-called famous Lepski-method for selecting a bandwidth of a kernel estimator. In [Lep91], he studied adaptive estimation under $\mathbb{L}_p$-loss, $1 \leq p \leq \infty$ over the collection of Hölder classes. Then [LMS97] 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 by [KLP01]. Then [GL08] 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 [GL11], [GL13], [GL14]. Nevertheless, the implementation of this method needs two steps of minimization, each requiring a thorough calibration. Note that the estimation under the $\mathbb{L}_1$-loss for bandwidth-selected kernel estimators has been investigated in [DL96], and the $\mathbb{L}_p$-aggregation of estimators has been studied by [Gol09].

In a very different spirit, another very popular method for adaptive estimation of functions is wavelet thresholding. In a series of papers, Donoho, Johnstone, Kerkyacharian and Picard have shown the power of wavelet thresholding for $\mathbb{L}_p$-estimation over the scale of Besov classes: see [DJKP95], [DJKP96],[DJKP97], [DJ98]. Refinements have been proposed in [Jud97], [KPT96], [HKP99], [JS05], 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 by [HKPT98] 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 [Joh19] 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 [ACE25] which considers Bayesian non-parametric 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 performances for non-Euclidian 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 Euclidian losses. We tackle this issue by considering the classical infinite sequence model and study a specific model selection procedure, called *PCO*, for weighted ℓ_p-loss functions, with $1 \leq p < \infty$. As explained by [Joh19] (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:

$$Y_\lambda = \theta_\lambda + \varepsilon \xi_\lambda, \qquad \lambda \in \Lambda. \tag{3}$$

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

$$\mathbb{P}(|\xi_\lambda| \geq t) \leq 2e^{-t^2/2}, \quad t \geq 0. \tag{4}$$

The previous definition refers, for instance, to Proposition 2.5.2 of [Ver18] with scale parameter $K_1 = \sqrt{2}$; note that fixing $K_1 = \sqrt{2}$ is not a restriction since, in our setting, we can replace the noise level ε with $K_1\varepsilon$ without loss of generality.

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 with a nonparametric problem. To connect our setting with nonparametric regression, we may have in mind that $N \propto \varepsilon^{-2}$ (see [DJKP95, Jud97, KP00] and Sections 3.2 and 4) but, unless specified, our results hold for any N.

For each m a subset of $\Lambda^{(N)}$, called model in the sequel, we set

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

In particular, $\widehat{\theta}^{(m)}_\lambda = 0$ for $\lambda \notin \Lambda^{(N)}$. For $1 \leq p < \infty$, and $w = (w_\lambda)_{\lambda \in \Lambda}$ a sequence of non negative weights, we denote $\ell_p(w)$ the weighted ℓ_p-norm on $\mathbb{R}^\Lambda$:

$$\|\vartheta\|^p_{\ell_p(w)} = \sum_{\lambda \in \Lambda} w_\lambda |\vartheta_\lambda|^p, \quad \vartheta \in \mathbb{R}^\Lambda. \tag{5}$$

In Model (3), we consider the risk associated with this weighted ℓ_p-norm. We assume in the sequel that $\|\theta\|_{\ell_p(w)} < \infty$.

Given a collection of models $\mathcal{M} \subset \mathcal{P}(\Lambda^{(N)})$, we wish to select $\widehat{m} \in \mathcal{M}$ in the best possible way. For this purpose, as explained previously, we rely on the *PCO criterion* introduced by [LMR17, VLMR23]. 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 $(\widehat{\theta}^{(m)})_{m\in\mathcal{M}}$. In our setting, it means that we have to consider $\widehat{\theta}^{(\Lambda^{(N)})}$ (see (7)). Then, adding as usual in the nonparametric setting a penalty term, we set

$$\widehat{m} = \operatorname*{arg\,min}_{m\in\mathcal{M}} \left\{ \|\widehat{\theta}^{(m)} - \widehat{\theta}^{(\Lambda^{(N)})}\|^p_{\ell_p(w)} + \mathrm{pen}(m) \right\}$$

and estimate θ by

$$\widetilde{\theta} = \widehat{\theta}^{(\widehat{m})}.$$

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

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

we obtain

$$\widehat{m} = \operatorname*{arg\,min}_{m\in\mathcal{M}} \mathrm{Crit}(m).$$

The heuristic of this approach is then different from the classical approach based on the contrast function. However, observe that if we take $p = 2$, the criterion function Crit corresponds to the one used in regression for various famous criteria such as Mallows's C_p [Mal73], AIC [Aka73] or BIC [Sch78] for instance. Two remarks are in order: Unlike Lepski type procedures, the derivation of the PCO estimate $\widetilde{\theta}$ involves only one minimization step, so its computational cost is much lower. Furthermore, if we take $\mathrm{pen}(m)$ of the form

$$\mathrm{pen}(m) = \sum_{\lambda\in m} w_\lambda t^p, \tag{6}$$

for some $t > 0$, we have:

$$\widehat{m} = \left\{ \lambda \in \Lambda^{(N)} : \quad |Y_\lambda| > t \right\}$$

and the PCO estimate corresponds to the thresholding estimate with threshold t:

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

Results of Sections 2, 3.2 and 4 show that we have to refine the definition of pen(m) given in (6) to obtain optimal results. This is described in the next subsection.

1.2 Contributions

As observed before, Lespki-type procedures achieve optimal properties in many settings but their computational cost is prohibitive. The main contributions of this paper consist in showing that under a convenient choice of pen(m), the PCO estimate, which is based on the simple $\ell_p(w)$-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 Theorem 2.1 that $\widetilde{\theta}$ satisfies a very general oracle inequality whatever the expression of the penalty term pen(m). For this purpose, we analyse the behavior of bias and variance terms of any estimate $\widehat{\theta}^{(m)}$. This first result shows how to choose pen(m), so that we obtain a more specific oracle inequality established in Theorem 2.5. As usual, these results depend on sharp concentration inequalities and our results rely on Theorem 2.3 involving sharp concentration of terms of the form $\sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p$ around their mean. The sharp tail bound involves the sum of two terms: a quadratic one, proportional to $\sqrt{x}$, which is classical, and a second one proportional to $x^{p/2}$. This last term is linear for $p = 2$ but it raises many technical difficulties otherwise in particular for $p > 2$. This result then reveals an elbow phenomenon depending on whether p is larger than 2 or not. We take into account this elbow to propose a more refined function pen(m) for the case $p > 2$ in Theorem 2.7. This result allows to deal with very large collections of models $\mathcal{M}$ when $p > 2$, which is crucial for the minimax setting.

Minimax rates of convergence for the estimate $\widetilde{\theta}$ are first derived when weights are constant and sequences θ have tails with a polynomial decreasing. We then consider the classical class of Besov bodies $\mathcal{B}^s_{r,\infty}(R)$ and study rates for any $r \geq 1$ and any $s > 1/r$ for the $\ell_p(w)$-risk. We prove in Theorem 3.2 that for $p \leq 2$, $\widetilde{\theta}$ is optimal under a suitable choice of the penalty function. For $p > 2$, $\widetilde{\theta}$ is also optimal if $r \geq p$ and if $r \leq p/(2s+1)$. For the case $p > 2$ and $p/(2s+1) < r < p$, the upper bound differs from the lower bound by a logarithmic term. To deal with the case $r < p$, we need to consider 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

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

(see Model (25)) and propose a PCO estimate of the function f based on wavelet representations. The generalization of results of Theorem 3.2 allows to obtain Theorem 4.1 that provides rates of our procedure on functional Besov spaces for the standard functional

$\mathbb{L}_p$-loss and to discuss optimality. The conclusions are similar to those of the sequential case. We also present the main steps of the methodology to derive functional minimax rates, which represents an interest per se.

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 cardinal of A, and $\mathcal{P}(A)$ the set of its subsets. The notation $\mathbb{N}$ is the set of non-negative integers: $\mathbb{N} = \{0, 1, 2, \cdots\}$. We denote by $u_\varepsilon \lesssim v_\varepsilon$ when there exists $0 < A < \infty$ such that $u_\varepsilon \leq A v_\varepsilon$ 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 (5). When weights w_λ are all equal to 1, we use the classical notation ℓ_p instead of $\ell_p(w)$. In the sequel, for short, we set $\|\cdot\|_p = \|\cdot\|_{\ell_p(w)}$. The functional 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 for the $\ell_p(w)$-risk 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. Along this section, we denote

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

We first derive in subsequent Theorem 2.1 a very general result which holds for any penalty function. Actually, by highlighting the key role of sums of the form $\sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p$, Theorem 2.1 allows to determine the ideal choice for the penalty $\mathrm{pen}(m)$. Concentrations of such sums around their mean are precisely studied in Theorem 2.3 and Corollary 2.4. This allows to refine Theorem 2.1, and sharp oracle inequalities are established in Theorems 2.5 and 2.7.

Before stating these results, we first observe that for any model m, we can easily compute the distance of the estimator $\widehat{\theta}^{(m)}$ with respect to θ for the norm $\ell_p(w)$. Indeed, we have for $m \subset \Lambda^{(N)}$:

$$\begin{aligned} \|\widehat{\theta}^{(m)} - \theta\|_p^p &= \sum_{\lambda \in m} w_\lambda |Y_\lambda - \theta_\lambda|^p + \sum_{\lambda \notin m} w_\lambda |0 - \theta_\lambda|^p \\ &= V_p(m) + B_p(m), \end{aligned}$$

with

$$B_p(m) := \sum_{\lambda \in \Lambda \setminus m} w_\lambda |\theta_\lambda|^p \quad \text{and} \quad V_p(m) := \varepsilon^p \sum_{\lambda \in m} w_\lambda |\xi_\lambda|^p. \tag{7}$$

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

2.1 A general oracle inequality

Recall that $\widetilde{\theta} = \widehat{\theta}^{(\widehat{m})}$ with

$$\widehat{\theta}^{(m)} = (Y_\lambda \mathbb{1}_{\{\lambda \in m\}})_{\lambda \in \Lambda}, \quad \widehat{m} = \operatorname*{arg\,min}_{m \in \mathcal{M}} \left\{ -\sum_{\lambda \in m} w_\lambda |Y_\lambda|^p + \mathrm{pen}(m) \right\}.$$

We obtain the following general oracle inequality for any $p \in [1, +\infty)$.

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

$$\|\widetilde{\theta} - \theta\|_p^p \le M_{p,\alpha} \|\widehat{\theta}^{(m)} - \theta\|_p^p + \frac{2}{\alpha}\Big[(1+\alpha)V_p(\widehat{m}) - \mathrm{pen}(\widehat{m})\Big] - \frac{2}{\alpha}\Big[(1+\alpha)V_p(m) - \mathrm{pen}(m)\Big],$$

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

$$\|\widetilde{\theta} - \theta\|_p^p \le M_p \|\widehat{\theta}^{(m)} - \theta\|_p^p + 2\Big[2V_p(\widehat{m}) - \mathrm{pen}(\widehat{m})\Big] - 2\Big[2V_p(m) - \mathrm{pen}(m)\Big], \tag{8}$$

where $M_p = M_{p,1}$ is given in Equation (31) (see the proof).
If $p = 1$, the previous inequalities are true by replacing $\frac{2}{\alpha}$ by $\frac{1}{\alpha \wedge 1}$ in the right hand side of the first inequality.

The proof of Theorem 2.1 is provided in Section 5.1.2.

Remark 2.2. *The value of $M_{p,\alpha}$ is rather intricate (see the proof of Theorem 2.1) and the best choice for α depends on p. Nevertheless, we numerically observe that $M_{p,1}$ is not far from the minimum of the function $\alpha \longmapsto M_{p,\alpha}$, and is even optimal for $p = 1$ and $p = 2$. This is why we focus on the case $\alpha = 1$ and state Inequality (8).*

In view of the first result of Theorem 2.1, optimality of the estimate $\widetilde{\theta}$ will be achieved among all estimates $(\widehat{\theta}^{(m)})_{m \in \mathcal{M}}$ if we are able to find $\mathrm{pen}(m)$ such that for a constant $\alpha \in (0, 2)$, $\mathrm{pen}(m)$ is close to $(1+\alpha)V_p(m)$ for all $m \in \mathcal{M}$. Note that $V_p(m)$ is not observable. Therefore, we need concentration inequalities to find the suitable expression of $\mathrm{pen}(m)$ that is involved in our procedure. Since $V_p(m) = \varepsilon^p \sum_{\lambda \in m} w_\lambda |\xi_\lambda|^p$, we need to study sums of $|\xi_\lambda|^p$ where the ξ_λ's are independent sub-Gaussian variables.

2.2 Concentration inequalities

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

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

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

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

we set $b_\lambda = \Big\| |\xi_\lambda|^p - \mathbb{E}|\xi_\lambda|^p \Big\|_{\psi_{2/p}}$. 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 $x > 0$,*

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

Moreover, if the ξ_λ's are also identically distributed, then, for any $x > 0$,

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

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

Note that for the Gaussian i.i.d. case, when $p = 1$, the Cirelson-Ibragimov-Sudakov inequality gives $c_{1,1} = \sqrt{2}$ and $c_{2,1} = 0$ (see Theorem 3.4 in [Mas07]). When $p = 2$, we retrieve the well-known inequality for chi-squared variables and $c_{1,2} = c_{2,2} = 2$ works: see Lemma 1 of [LM00]. This also matches with the Hanson-Wright inequality with identity matrix or Bernstein inequality for sub-exponential variables, see e.g. [Ver18].
In the general case $p \geq 1$, the result ensues from concentration theorems for sub-Weibull variables. Indeed, if we denote $X_\lambda = |\xi_\lambda|^p - \mathbb{E}|\xi_\lambda|^p$, we observe that the sub-Gaussianity property of ξ_λ entails that X_λ has a Weibull behavior:

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

for C a constant. Such a variable, with bounded Orlicz norm with function $e^{x^r} - 1$, is called a sub-Weibull variable. Note that the Weibull parameter here is $r = 2/p \leq 2$. Now our theorem directly follows from recent Theorem 1 of [ZW22], or Theorem 3.1 of [KC22] (see also their Equation (3.6)), both giving explicit formulas for $d_{1,p}$ and $d_{2,p}$. In the i.i.d case, all b_λ's are equal and bounded by $\||\xi_\lambda|^p\|_{\psi_{2/p}} = \|\xi_\lambda\|_{\psi_2}$, up to a universal constant.

Observe that Theorem 2.3 can also be deduced from older results, like moment bounds of [GK95] 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 [HMSO97] gives a two-sided moments inequality that leads to our tail result. Their bound cannot be improved meaning that if the ξ_λ's are Gaussian, the upper bound is achieved up to a constant. Same optimality considerations are raised by [KC22].

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 $1 - 2\exp(-x)$,*

$$\frac{1}{2}\sigma_p^p D - \kappa_p D^{\left(1-\frac{p}{2}\right)_+} x^{\frac{p}{2}} \leq \sum_{\lambda \in \mathcal{I}} |\xi_\lambda|^p \leq \frac{3}{2}\sigma_p^p D + \kappa_p D^{\left(1-\frac{p}{2}\right)_+} x^{\frac{p}{2}}, \tag{9}$$

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 θ, and this time $x \leq D^{1-p/2}x^{p/2}$. Finally, if $p < 2$ and $x > D$, we directly write $\sqrt{Dx} \leq D^{1-\frac{p}{2}}x^{\frac{p}{2}}$. □

Note that Corollary 2.4 holds by replacing 1/2 (resp. 3/2) in (9) by $(1-\epsilon)$ (resp. $(1+\epsilon)$) for ϵ arbitrary small (with κ_p depending on ϵ).

2.3 Refined oracle inequalities

In this section, we apply the general oracle inequality of Theorem 2.1 with $\alpha = 1$ (see Remark 2.2) and use sharp concentration inequalities of Corollary 2.4 to derive suitable penalties.

Typically, weights w_λ may not depend on λ or may be constant on some slices (see subsequent sections). Therefore, we consider the following partition of $\Lambda^{(N)}$:

$$\Lambda^{(N)} = \bigcup_{j \in \mathcal{J}} \Lambda_j \tag{10}$$

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 (10) is unique. Then, given (10), we consider for any model $m \in \mathcal{M}$

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

and we set

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

In view of Inequality (8) and Corollary 2.4, we take

$$\mathrm{pen}(m) := 2\varepsilon^p \sum_{j\in\mathcal{J}} \omega_j \mathrm{p}_j(m_j), \tag{11}$$

with, for some $x_{m_j} \geq 1$,

$$\mathrm{p}_j(m_j) = \frac{3}{2}\sigma_p^p |m_j| + \kappa_p 2^{\frac{(p-2)_+}{2}} |m_j|^{\left(1-\frac{p}{2}\right)_+} x_{m_j}^{\frac{p}{2}}. \tag{12}$$

Observe that:

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

The factor x_{m_j}, depending on m_j, will be specified later. But note that, mimicking the computations of Section 1.1, the thresholding rule corresponds to the case where $\mathrm{p}_j(m_j)$ is proportional to $|m_j|$, which is obtained for instance by taking x_{m_j} proportional to $|m_j|$ 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 $p = 2$, $\mathrm{pen}(m)$ 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, 4.5 and 4.18 of [Mas07]. The extension of these results to the case $p \neq 2$ is provided by the following theorem.

Theorem 2.5. *Let $p \geq 1$. We consider the estimate $\widetilde{\theta} = \widehat{\theta}^{(\widehat{m})}$ associated with the penalty defined in* (11) *and $\mathrm{p}_j(m_j)$ given in* (12)*. Then*

$$\mathbb{E}\Big[\|\widetilde{\theta} - \theta\|_p^p\Big] \leq \widetilde{M}_p \inf_{m\in\mathcal{M}} \Big\{\mathbb{E}\Big[\|\widehat{\theta}^{(m)} - \theta\|_p^p\Big] + \mathrm{pen}(m)\Big\} + \breve{M}_p \varepsilon^p R(\mathcal{M}) \tag{13}$$

with

$$R(\mathcal{M}) = \sum_{j\in\mathcal{J}} \omega_j \sum_{m_j\in\mathcal{M}_j, m_j\neq\emptyset} |m_j|^{\left(1-\frac{p}{2}\right)_+} e^{-x_{m_j}}, \tag{14}$$

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

The proof of Theorem 2.5 is provided in Section 5.1.3.

Remark 2.6. *Theorem 2.5 remains true if we replace the equality in* (11) *by the inequality:*

$$\mathrm{pen}(m) \geq 2\varepsilon^p \sum_{j\in\mathcal{J}} \omega_j \mathrm{p}_j(m_j).$$

Now let us discuss the choice of the factors x_{m_j}. We fix them in order $\widetilde{\theta}$ to be optimal in the oracle setting, meaning that

$$\mathbb{E}\Big[\|\widetilde{\theta}-\theta\|_p^p\Big] \lesssim \inf_{m\in\mathcal{M}} \mathbb{E}\Big[\|\widehat{\theta}^{(m)}-\theta\|_p^p\Big]. \tag{15}$$

Now, observe that for any model m,

$$\mathbb{E}\Big[\|\widehat{\theta}^{(m)}-\theta\|_p^p\Big] = B_p(m) + \sum_{j\in\mathcal{J}} \mathbb{E}\big[V_p(m_j)\big] = \sum_{\lambda\notin m} w_\lambda|\theta_\lambda|^p + \varepsilon^p\sigma_p^p \sum_{j\in\mathcal{J}} w_j|m_j|$$

and $\mathrm{pen}(m)$ can be compared to the second term of the right hand side. In particular, Theorem 2.5 shows that $\widetilde{\theta}$ is optimal as soon as we have

$$|m_j|^{\left(1-\frac{p}{2}\right)_+} x_{m_j}^{\frac{p}{2}} \lesssim |m_j| \iff x_{m_j} \lesssim |m_j|^{2/\max(2,p)}, \quad \text{for all } j\in\mathcal{J} \tag{16}$$

and

$$R(\mathcal{M}) < \infty. \tag{17}$$

More precisely, under (16) and (17), we obtain:

$$\mathbb{E}\Big[\|\widetilde{\theta}-\theta\|_p^p\Big] \lesssim \inf_{m\in\mathcal{M}} \mathbb{E}\Big[\|\widehat{\theta}^{(m)}-\theta\|_p^p\Big] + \varepsilon^p,$$

which corresponds to (15) up to the residual term ε^p. Therefore, we wish to fix the factors x_{m_j} so that Conditions (16) and (17) are satisfied. Condition (16) means that the factors cannot be too large. But the condition $R(\mathcal{M}) < \infty$ holds only if $\mathcal{M}$ is not too large or factors are large enough. In particular, to have (17), we need:

$$\sum_{m_j\in\mathcal{M}_j,\, m_j\neq\emptyset} |m_j|^{\left(1-\frac{p}{2}\right)_+} e^{-x_{m_j}} < \infty \tag{18}$$

(see (14)). Given $d \geq 1$, consider situations where the number of models of size d is exponential in d, say $\exp(cd)$ for $c > 0$. When $p \leq 2$, we can choose x_{m_j} satisfying (16) and such that (18) holds by taking

$$x_{m_j} = \widetilde{c}|m_j| \quad \text{or} \quad x_{m_j} = \widetilde{c}\log(|m_j|)$$

with $\widetilde{c}$ large enough. But, when $p > 2$, if (16) is verified, we have, for c_* a positive constant,

$$\sum_{m_j\in\mathcal{M}_j,\, m_j\neq\emptyset} |m_j|^{\left(1-\frac{p}{2}\right)_+} e^{-x_{m_j}} \geq \sum_{d=1}^{D_j} \exp(cd - c_* d^{2/p}),$$

with $D_j = \max_{m_j \in \mathcal{M}_j}(|m_j|)$ (assumed to be larger than 1). The right hand side becomes very large when D_j is large, which occurs when the model collection is large. Therefore, the optimality of $\widetilde{\theta}$ for the case $p > 2$ requires a modification of the penalty to deal with large collections of models. Since the elbow occurs at the value $p = 2$, we use $\mathrm{p}_j(m_j)$ with the value $p = 2$ and we introduce

$$\mathrm{p}_j^{\#}(m_j) = \frac{3}{2}\sigma_2^2|m_j| + \kappa_2 x_{m_j}. \tag{19}$$

We obtain the following theorem.

Theorem 2.7. *Let $q > 1$ and $p \geq 2$. Let $m \in \mathcal{M}$. For $j \in \mathcal{J}$ and $m_j \in \mathcal{M}_j$, we take $x_{m_j} \geq 1$ and we consider the estimate $\widetilde{\theta} = \widehat{\theta}^{(\widehat{m})}$ associated with the penalty*

$$\mathrm{pen}(m) := 2\varepsilon^p \sum_{j \in \mathcal{J}} \omega_j \min\Big(\mathrm{p}_j(m_j), (2q\log N)^{\frac{p}{2}-1}\mathrm{p}_j^{\#}(m_j)\Big), \tag{20}$$

where $\mathrm{p}_j(m_j)$ and $\mathrm{p}_j^{\#}(m_j)$ are defined in (12) *and* (19) *respectively. Then*

$$\mathbb{E}\Big[\|\widetilde{\theta} - \theta\|_p^p\Big] \leq \widetilde{M}_p \inf_{m \in \mathcal{M}} \Big\{\mathbb{E}\Big[\|\widehat{\theta}^{(m)} - \theta\|_p^p\Big] + \mathrm{pen}(m)\Big\} + \breve{M}_{p,q}\Big(N^{1-q}\|\theta\|_p^p + \varepsilon^p R^{\#}(\mathcal{M})\Big), \tag{21}$$

with $\widetilde{M}_p$ (resp. $\breve{M}_{p,q}$) a constant only depending on p (resp. p, σ_p and q) and

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

where $R(\mathcal{M})$ is defined in (14).

The proof of Theorem 2.7 is provided in Section 5.1.4. Unfortunately, the use of $\mathrm{p}_j^{\#}(m_j)$ 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 order of $|\log(\varepsilon)|$, by taking q large enough, the residuable term of the oracle inequality is the same as in (13) up to the logarithmic term $|\log(\varepsilon)|^{\frac{p}{2}-1}$. Two scenarios are then of interest.

1. If we can take x_{m_j} satisfying (16) such that $R(\mathcal{M}) < \infty$, we use

$$\mathrm{pen}(m) \leq 2\varepsilon^p \sum_{j \in \mathcal{J}} \omega_j \mathrm{p}_j(m_j)$$

and Theorem 2.7 gives:

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

2. If we can take $x_{m_j} \approx |m_j|$ such that $R(\mathcal{M}) < \infty$, we use

$$\mathrm{pen}(m) \lesssim |\log(\varepsilon)|^{\frac{p}{2}-1}\varepsilon^p \sum_{j\in\mathcal{J}} \omega_j \mathrm{p}_j^{\#}(m_j) \lesssim |\log(\varepsilon)|^{\frac{p}{2}-1}\varepsilon^p \sum_{j\in\mathcal{J}} w_j|m_j|$$

and Theorem 2.7 gives:

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

The first scenario provides optimality of $\widetilde{\theta}$ up to the residual term $\varepsilon^p|\log(\varepsilon)|^{\frac{p}{2}-1}$. An additional logarithmic factor $|\log(\varepsilon)|^{\frac{p}{2}-1}$ is required for the main term for the second scenario. In each case, the condition $R(\mathcal{M}) < \infty$, depending on the chosen collection $\mathcal{M}$ is crucial. In the next section devoted to the minimax setting, $\mathcal{M}$ will be specified in order to study optimality of $\widetilde{\theta}$ under different regularity conditions for θ.

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 next Section 3.1, we illustrate our results on a simple situation, namely the case of constant weights and sequences θ whose tails have a polynomial decreasing. Then, in Section 3.2, we investigate the minimax rates of convergence of our procedure on Besov bodies.

3.1 Case of constant weights

In this section, we assume that weights w_λ do not depend on λ and without loss of generality, we assume:

$$w_\lambda = 1, \quad \lambda \in \Lambda.$$

We study minimax rates of convergence of our procedure when tails of θ have a polynomial decreasing. 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

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

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:

$$\mathcal{M} = \Big\{ \{1,\dots,k\}, \quad 1 \le k \le N \Big\}$$

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

$$\begin{aligned} R(\mathcal{M}) &= \sum_{j=1}^{J} \omega_j \sum_{m_j \in \mathcal{M}_j,\, m_j \neq \emptyset} |m_j|^{\left(1-\frac{p}{2}\right)_+} e^{-x_{m_j}} \\ &\leq \sum_{d=1}^{+\infty} d^{\left(1-\frac{p}{2}\right)_+} e^{-a\log(d)} < \infty. \end{aligned}$$

In the previous inequality, we have used that $\mathcal{J} = \{1\}$ and $\Lambda_1 = \Lambda^{(N)}$; therefore, $m_j = m$ for any j and any $m \in \mathcal{M}$. Theorem 2.5 gives

$$\mathbb{E}\Big[\|\widetilde{\theta} - \theta\|_p^p\Big] \leq \widetilde{M}_p \inf_{m\in\mathcal{M}} \Big\{\mathbb{E}\Big[\|\widehat{\theta}^{(m)} - \theta\|_p^p\Big] + \mathrm{pen}(m)\Big\} + \breve{M}_p \varepsilon^p R(\mathcal{M}).$$

Since

$$\mathbb{E}\Big[\|\widehat{\theta}^{(m)} - \theta\|_p^p\Big] = B_p(m) + \mathbb{E}\big[V_p(m)\big]$$

we have

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

and we finally obtain:

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

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

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

We obtain:

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

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

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

In particular, the rate in the right hand side is the optimal minimax rate on the class $\mathbb{B}_2^s(R)$, see [Riv04].

3.2 Case of non-constant weights and 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

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

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 writes:

$$Y_{jk} = \theta_{jk} + \varepsilon \xi_{jk}, \qquad j \geq -1,\ k \in K_j. \tag{22}$$

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

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

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

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

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

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

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

$$\|\vartheta\|_p^p = \sum_{j=-1}^{+\infty} \sum_{k\in K_j} w_{jk} |\vartheta_{jk}|^p = \sum_{j=-1}^{+\infty} \omega_j \sum_{k\in K_j} |\vartheta_{jk}|^p, \tag{23}$$

but we assume that the weights satisfy

$$w_{jk} = \omega_j = 2^{j\left(\frac{p}{2}-1\right)}, \quad (j,k) \in \Lambda. \tag{24}$$

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

3.2.2 Lower bounds

In this section, we still consider Model (22) 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 [DJKP95] (by taking, using their notation, $(\sigma, p, q) = (s, r, \infty)$, $(\sigma', p', q') = (0, p, p)$ and $C = R$).

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

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

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 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

Our aim is to obtain the rate of convergence on the class $\mathcal{B}^s_{r,\infty}(R)$ of our estimator for some well-chosen collection of models. 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 homogenous (H), intermediate (I) and sparse and frontier cases (S). For each sub-collection, we use a penalty (see (11) and (20)) with corresponding factor $x^H_{m_j}$, $x^I_{m_j}$ and $x^S_{m_j}$ (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 $L \in \{0, \ldots, J\}$ such that

  for all $-1 \leq j \leq J$, $\quad m_j = \begin{cases} \Lambda_j & \text{if } -1 \leq j \leq L, \\ \emptyset & \text{if } j > L \end{cases}$

  and $x^H_{m_j} = \frac{p}{2} \log |m_j|$ ($\log 0 = 0$);

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

For $(m, a) \in \mathfrak{M}$, having in mind Theorems 2.5 and 2.7, denote

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

with $q = p + 1$ and

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

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

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

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

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

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

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

We remark that the radius R may decrease to 0 or increase to $+\infty$ when $\varepsilon \to 0$. 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.

The proof of Theorem 3.2 is provided in Section 5.2. 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. To the best of our knowledge, all other adaptive methods suffer from a logarithmic loss in at least one case, except the adaptive procedure proposed in [Jud97] based on an involved combination of thresholding and Lepski-type procedures. In particular, [DJKP95] have a logarithmic loss, both in homegeneous and intermediate cases, with power $\frac{s}{2s+1}$. It is also the case in [LMS97] (for the study of the white noise model) or [KLP01] (for the study of the multivariate white noise model) or [GL14] (for the study of the density model).

4 Nonparametric regression

In this section, we consider the following nonparametric regression model

$$X_i = f(t_i) + \sigma\eta_i, \qquad 1 \leq i \leq n, \tag{25}$$

with η_i some i.i.d. centered sub-Gaussian variables modelling the noise and $\sigma > 0$ assumed to be known. In the previous expression, f has support included into $[0, 1]$ and the t_i's are deterministic and equidistant : $t_i = i/n$, $i = 1, \ldots, n$. We assume in the sequel that $\log_2(n)$ is an integer. Our goal is to estimate the function f by using observations $(X_i)_{i=1,\ldots,n}$. 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 (24) introduced in Section 3.2.1.

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

$$f = \sum_k \langle f, \phi_k\rangle\phi_k + \sum_{j=0}^{+\infty}\sum_k \langle f, \psi_{jk}\rangle\psi_{jk}, \tag{26}$$

where ϕ_k is the translation of a father wavelet ϕ and ψ_{jk} is the dilation and translation 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 (26) can be of course rewritten as

$$f = \sum_{j=-1}^{+\infty} \sum_{k \in K_j} \langle f, \varphi_{jk} \rangle \varphi_{jk}, \tag{27}$$

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 [CDV93] 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 [CDV93] shows that we can take

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

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

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

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 $\widetilde{\theta}$ the PCO estimate of Section 3.2.3 (except that me 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 set

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

to estimate f. We study the upper bound of the $\mathbb{L}_p$-risk of $\widetilde{f}$ on the class of Besov spaces. We refer the reader to Section 9.2 of [HKPT98] 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 [HKPT98]): 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 $\mathcal{B}^s_{r,q}$ if and only if

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

When $q = \infty$, this condition becomes

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

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:*

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

for C a constant depending on p, s, r, γ, 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 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 [Nem85] for corresponding lower bounds, see also [DJ98] and Theorem 4 of [DJKP95]. 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 main arguments of the proof of Theorem 4.1. The $\mathbb{L}_p$-risk of $\widetilde{f}$ is deduced from the following control:

$$\mathbb{E}\Big[\|\widetilde{f} - f\|^p_{\mathbb{L}_p} \Big] \leq 2^{p-1} \left[\mathbb{E}\left[\Big\| \sum_{(j,k) \in \Lambda^{(N)}} (\widetilde{\theta}_{jk} - \theta_{jk}) \varphi_{jk} \Big\|^p_{\mathbb{L}_p} \right] + \Big\| \sum_{(j,k) \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \Big\|^p_{\mathbb{L}_p} \right] \quad (28)$$

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 (28) to the $\ell_p(w)$-risk of $\widetilde{\theta}$?

To address the first issue, we establish in the next proposition, proved in Section 5.3.1, that the $\xi_{j,k}$'s satisfy a result similar to the result 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 the compactly father and mother wavelets ϕ and ψ such that for any $x \geq 1$,

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

Regarding the approximation term, we can prove the following result (see Section 5.3.2 for the proof).

Lemma 4.3. *Assume that f belongs to the Besov set $\mathcal{B}^s_{r,\infty}(R)$ 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 $N \geq \frac{\varepsilon^{-2}}{|\log(\varepsilon)|}$,*

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

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 $\widetilde{f}$ (or rather the first terms of its decomposition) with the $\ell_p(w)$-risk of $\widetilde{\theta}$ when weights are those of Section 3.2.1 (see (24)). We first state the following lemma whose proof can be found in Section 5.3.3.

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

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

for C a constant and with

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

Now, we naturally distinguish two cases.

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

$$\begin{aligned}\mathbb{E}\Big[\Big\|\sum_{(j,k)\in\Lambda^{(N)}}(\widetilde{\theta}_{jk}-\theta_{jk})\varphi_{jk}\Big\|^p_{\mathbb{L}_p}\Big] &\lesssim \mathbb{E}\Big[\Big\|\sum_{(j,k)\in\Lambda^{(N)}}(\widetilde{\theta}_{jk}-\theta_{jk})\varphi_{jk}\Big\|^p_{\mathcal{B}^0_{p,p}}\Big] \\ &\lesssim \mathbb{E}\Big[\sum_{j=-1}^{J} 2^{j(\frac{p}{2}-1)}\sum_{k\in K_j}\big|\widetilde{\theta}_{jk}-\theta_{jk}\big|^p\Big] \leq \mathbb{E}\|\widetilde{\theta}-\theta\|^p_p, \quad (29)\end{aligned}$$

  with the $\|\cdot\|_p$-norm is defined in (23) with weights defined in (24). Therefore, results of Section 3.2.3 can be applied.

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

$$\begin{aligned}\mathbb{E}\Big[\Big\|\sum_{(j,k)\in\Lambda^{(N)}}(\widetilde{\theta}_{jk}-\theta_{jk})\varphi_{jk}\Big\|^p_{\mathbb{L}_p}\Big] &\lesssim \mathbb{E}\Big[\Big\|\sum_{(j,k)\in\Lambda^{(N)}}(\widetilde{\theta}_{jk}-\theta_{jk})\varphi_{jk}\Big\|^p_{\mathcal{B}^0_{p,2}}\Big] \\ &\lesssim \mathbb{E}\left[\left(\sum_{j=-1}^{J} 2^{j(1-\frac{2}{p})}\Big(\sum_{k\in K_j}\big|\widetilde{\theta}_{jk}-\theta_{jk}\big|^p\Big)^{\frac{2}{p}}\right)^{\frac{p}{2}}\right] \\ &\lesssim \left[\sum_{j=-1}^{J}\left(\mathbb{E}\Big[2^{j(\frac{p}{2}-1)}\sum_{k\in K_j}\big|\widetilde{\theta}_{jk}-\theta_{jk}\big|^p\Big]\right)^{\frac{2}{p}}\right]^{\frac{p}{2}}, \quad (30)\end{aligned}$$

  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 5.3.

These arguments and technical complements of Section 5.3 allow to prove Theorem 4.1.

5 Proofs

5.1 Proofs of oracle results of Section 2

5.1.1 Technical lemmas

Lemma 5.1. *Let $p \geq 1$. Let $K > 0$ and a, b two reals such that $|a| \geq K|b|$. Then, for any $\alpha > 0$,*

$$||a+b|^p - |a|^p - \alpha|b|^p| \leq C_{1p}(\alpha, K)|a|^p$$

and

$$||a+b|^p - \alpha|a|^p - |b|^p| \leq C_{2p}(\alpha, K)|a|^p$$

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

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 $\lim_{K\to\infty} C_{1p}(\alpha, K) = 0$. 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 5.2. *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 $\alpha \to 0$,

$$C(\alpha, p) \sim \Big(\frac{\alpha}{p-1}\Big)^{1-p} \to +\infty.$$

Proof. We prove that

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

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

5.1.2 Proof of Theorem 2.1

Let us denote

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

Setting

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

we can write

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

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

$$-\sum_{\lambda \in \widehat{m}} w_\lambda |Y_\lambda|^p + \mathrm{pen}(\widehat{m}) \leq -\sum_{\lambda \in m} w_\lambda |Y_\lambda|^p + \mathrm{pen}(m).$$

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

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

Thus

$$\|\widehat{\theta}^{(\widehat{m})} - \theta\|_p^p \leq \|\widehat{\theta}^{(m)} - \theta\|_p^p + \Big[(1+\alpha)V_p(\widehat{m}) - \mathrm{pen}(\widehat{m})\Big] - \Big[(1+\alpha)V_p(m) - \mathrm{pen}(m)\Big] + J(\widehat{m}) - J(m)$$

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

$$S_\lambda = w_\lambda\big(|\theta_\lambda|^p + \alpha\varepsilon^p|\xi_\lambda|^p - |Y_\lambda|^p\big)$$

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

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

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

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

by using notations of Lemma 5.2. Finally,

$$-\sum_{\lambda\in m^c\cap\widehat{m}} S_\lambda \le \Big(1-\frac{\alpha}{2}\Big)\sum_{\lambda\in\widehat{m}} w_\lambda|\varepsilon\xi_\lambda|^p + \Big(C(\alpha/2,p)-1\Big)\sum_{\lambda\in m^c} w_\lambda|\theta_\lambda|^p$$
$$\le \Big(1-\frac{\alpha}{2}\Big)V_p(\widehat{m}) + \Big(C(\alpha/2,p)-1\Big)B_p(m).$$

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

1. We assume that $|\theta_\lambda|\ge K\varepsilon|\xi_\lambda|$. Then, applying Lemma 5.1, we have

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

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

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

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

   using again Lemma 5.1 with $|\varepsilon\xi_\lambda|\ge K^{-1}|\theta_\lambda|$.

Finally,

$$\sum_{\lambda\in\widehat{m}^c\cap m} S_\lambda \le \Big(1-\frac{\alpha}{2}\Big)\sum_{\lambda\in\widehat{m}^c} w_\lambda|\theta_\lambda|^p + C_{2p}(\alpha,K_{p,\alpha}^{-1})\sum_{\lambda\in m} w_\lambda|\varepsilon\xi_\lambda|^p$$
$$\le \Big(1-\frac{\alpha}{2}\Big)B_p(\widehat{m}) + C_{2p}(\alpha,K_{p,\alpha}^{-1})V_p(m).$$

We obtain

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

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

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

and

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

Thus the result is proved with

$$M_{p,\alpha} = \frac{2}{\alpha}\left(1 + \max\left(C\left(\frac{\alpha}{2}, p\right) - 1, C_{2p}(\alpha, K_{p,\alpha}^{-1})\right)\right) \tag{31}$$

where $K_{p,\alpha}$ is such that $C_{1p}(\alpha, K_{p,\alpha}) \leq 1 - \alpha/2$ and where $C_{1p}, C_{2p}, C(., p)$ are defined in Lemmas 5.1 and 5.2.

Case $p = 1$. In this case

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

Note also that

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

Then, if $\alpha \geq 1$,

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

and then

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

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

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

and then

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

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

5.1.3 Proof of Theorem 2.5

We only consider the case $p > 1$, the case $p = 1$ is similar.
Starting from Theorem 2.1, we have:

$$\begin{aligned}\|\widetilde{\theta} - \theta\|_p^p &\leq M_p \|\widehat{\theta}^{(m)} - \theta\|_p^p + 2\Big[2V_p(\widehat{m}) - \mathrm{pen}(\widehat{m})\Big] - 2\Big[2V_p(m) - \mathrm{pen}(m)\Big] \\ &\leq M_p \|\widehat{\theta}^{(m)} - \theta\|_p^p + 2\mathrm{pen}(m) + 2\Big[2V_p(\widehat{m}) - \mathrm{pen}(\widehat{m})\Big] \\ &\leq M_p \|\widehat{\theta}^{(m)} - \theta\|_p^p + 2\mathrm{pen}(m) + 4\varepsilon^p \sum_{j\in\mathcal{J}} \omega_j \sum_{m_j\in\mathcal{M}_j} \Big[Z(m_j) - \mathrm{p}_j(m_j)\Big]_+,\end{aligned}$$

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

$$\mathrm{p}_j(m_j) = \frac{3}{2}|m_j|\sigma_p^p + \kappa_p 2^{\frac{(p-2)_+}{2}} |m_j|^{\left(1-\frac{p}{2}\right)_+} x_{m_j}^{p/2}.$$

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

$$\|\widetilde{\theta} - \theta\|_p^p \leq M_p \|\widehat{\theta}^{(m)} - \theta\|_p^p + 2\mathrm{pen}(m) + 4\varepsilon^p \sum_{j\in\mathcal{J}} \omega_j \sum_{m_j\in\mathcal{M}_j, m_j\neq\emptyset} \Big[Z(m_j) - \mathrm{p}_j(m_j)\Big]_+.$$

Taking the expectation yields

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

with

$$R = \sum_{j\in\mathcal{J}} \omega_j \sum_{\substack{m_j\in\mathcal{M}_j \\ m_j\neq\emptyset}} \mathbb{E}\Big[Z(m_j) - \mathrm{p}_j(m_j)\Big]_+ = \sum_{j\in\mathcal{J}} \omega_j \sum_{\substack{m_j\in\mathcal{M}_j \\ m_j\neq\emptyset}} \int_0^\infty \mathbb{P}\Big(Z(m_j) - \mathrm{p}_j(m_j) > u\Big) du.$$

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

$$\begin{aligned}&\int_0^\infty \mathbb{P}\Big(Z(m_j) - \mathrm{p}_j(m_j) > u\Big) du \\ &\quad= \int_0^\infty \mathbb{P}\Big(Z(m_j) - \frac{3}{2}\sigma_p^p|m_j| - C_j 2^{\frac{(p-2)_+}{2}} x_{m_j}^{\frac{p}{2}} > u\Big) du \\ &\quad= \int_0^\infty \mathbb{P}\Big(Z(m_j) - \frac{3}{2}\sigma_p^p|m_j| - C_j 2^{\frac{(p-2)_+}{2}} x_{m_j}^{\frac{p}{2}} > C_j 2^{\frac{(p-2)_+}{2}} v^{\frac{p}{2}}\Big) \left[C_j 2^{\frac{(p-2)_+}{2}} \frac{p}{2} v^{\frac{p}{2}-1}\right] dv \\ &\quad\leq \int_0^\infty \mathbb{P}\Big(Z(m_j) > \frac{3}{2}\sigma_p^p|m_j| + C_j (x_{m_j} + v)^{\frac{p}{2}}\Big) \left[C_j 2^{\frac{(p-2)_+}{2}} \frac{p}{2} v^{\frac{p}{2}-1}\right] dv\end{aligned}$$

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

$$\begin{aligned}\int_0^\infty \mathbb{P}\Big(Z(m_j) - \mathrm{p}_j(m_j) > u\Big)du \leq & 2\int_0^\infty e^{-(x_{m_j}+v)}\left[C_j 2^{\frac{(p-2)_+}{2}}\frac{p}{2}v^{\frac{p}{2}-1}\right]dv \\ \leq & C(p)C_j e^{-x_{m_j}} = C(p)\kappa_p |m_j|^{(1-p/2)_+}e^{-x_{m_j}},\end{aligned}$$

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

$$R \leq C(p)\kappa_p \sum_{j\in\mathcal{J}}\omega_j \sum_{m_j\in\mathcal{M}_j, m_j\neq\emptyset} |m_j|^{(1-p/2)_+}e^{-x_{m_j}}.$$

5.1.4 Proof of Theorem 2.7

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

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

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

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

First,

$$\begin{aligned}\mathbb{E}\Big[\|\widetilde{\theta}-\theta\|_p^p \mathbb{1}_{\Omega(N,q)^c}\Big] &= \mathbb{E}\Big[\sum_{\lambda\notin\widehat{m}} w_\lambda |\theta_\lambda|^p \mathbb{1}_{\Omega(N,q)^c}\Big] + \varepsilon^p \mathbb{E}\Big[\sum_{\lambda\in\widehat{m}} w_\lambda |\xi_\lambda|^p \mathbb{1}_{\Omega(N,q)^c}\Big] \\ &\leq 2N^{1-q}\|\theta\|_p^p + \varepsilon^p\sqrt{2}N^{(1-q)/2}\sigma_{2p}^p \sum_{\lambda\in\Lambda} w_\lambda.\end{aligned}$$

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

$$\begin{aligned}\|\widetilde{\theta}-\theta\|_p^p &\leq M_p\|\widehat{\theta}^{(m)}-\theta\|_p^p + 2\Big[2V_p(\widehat{m}) - \mathrm{pen}(\widehat{m})\Big] - 2\Big[2V_p(m) - \mathrm{pen}(m)\Big] \\ &\leq M_p\|\widehat{\theta}^{(m)}-\theta\|_p^p + 2\mathrm{pen}(m) + 2\Big[2V_p(\widehat{m}) - \mathrm{pen}(\widehat{m})\Big] \\ &\leq M_p\|\widehat{\theta}^{(m)}-\theta\|_p^p + 2\mathrm{pen}(m) + 4\varepsilon^p \sum_{j\in\mathcal{J}}\omega_j \sum_{m_j\in\mathcal{M}_j}\Big[Z(m_j) - P_j(m_j)\Big]_+,\end{aligned}$$

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

$$P_j(m_j) = \min\Big(\mathrm{p}_j^1(m_j); \mathrm{p}_j^2(m_j)\Big)$$

with

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

Then, taking the expectation

$$\mathbb{E}\Big[\|\widetilde{\theta}-\theta\|_p^p \mathbb{1}_{\Omega(N,q)}\Big] \leq M_p \mathbb{E}\Big[\|\widehat{\theta}^{(m)}-\theta\|_p^p \mathbb{1}_{\Omega(N,q)}\Big] + 2\mathrm{pen}(m) + 4\varepsilon^p R$$

with

$$R = \sum_{j\in\mathcal{J}} \omega_j \sum_{\substack{m_j\in\mathcal{M}_j \\ m_j\neq\emptyset}} \mathbb{E}\Big[\big(Z(m_j) - P_j(m_j)\big)_+ \mathbb{1}_{\Omega(N,q)}\Big].$$

It remains to control the term R. We have:

$$\begin{aligned} R &= \sum_{j\in\mathcal{J}} \omega_j \sum_{\substack{m_j\in\mathcal{M}_j \\ m_j\neq\emptyset}} \int_0^\infty \mathbb{P}\Big(Z(m_j)\mathbb{1}_{\Omega(N,q)} - \min\big(\mathrm{p}_j^1(m_j);\mathrm{p}_j^2(m_j)\big)\mathbb{1}_{\Omega(N,q)} > u\Big)du \\ &\leq \sum_{j\in\mathcal{J}} \omega_j \sum_{\substack{m_j\in\mathcal{M}_j \\ m_j\neq\emptyset}} \Bigg[\int_0^\infty \mathbb{P}\Big(Z(m_j) - \mathrm{p}_j^1(m_j) > u\Big)du \\ &\qquad\qquad + \int_0^\infty \mathbb{P}\Big(\big\{Z(m_j) - \mathrm{p}_j^2(m_j) > u\big\} \cap \Omega(N,q)\Big)du\Bigg]. \end{aligned}$$

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

$$\int_0^\infty \mathbb{P}\Big(Z(m_j) - \mathrm{p}_j^1(m_j) > u\Big)du \leq C(p)|m_j|^{(1-p/2)_+} e^{-x_{m_j}}.$$

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

$$\begin{aligned} Z(m_j) - \mathrm{p}_j^2(m_j) &= \sum_{\lambda\in m_j} |\xi_\lambda|^p - (2q\log N)^{\frac{p}{2}-1}\mathrm{p}_j^{\#}(m_j) \\ &\leq \big(\sqrt{2q\log N}\big)^{p-2} \sum_{\lambda\in m_j} \xi_\lambda^2 - (2q\log N)^{\frac{p}{2}-1}\Big(\frac{3}{2}\sigma_2^2|m_j| + \kappa_2 x_{m_j}\Big) \\ &\leq (2q\log N)^{\frac{p}{2}-1}\Big(\sum_{\lambda\in m_j} \xi_\lambda^2 - \Big(\frac{3}{2}\sigma_2^2|m_j| + \kappa_2 x_{m_j}\Big)\Big). \end{aligned}$$

Therefore,

$$\int_0^\infty \mathbb{P}\Big(\big\{Z(m_j) - \mathrm{p}_j^2(m_j) > u\big\} \cap \Omega(N,q)\Big)du \ \leq (2q\log N)^{\frac{p}{2}-1} C(2) e^{-x_{m_j}}.$$

We obtain, for $p \geq 2$,

$$R \leq \max\big(C(p); C(2)\big) \sum_{j \in \mathcal{J}} \omega_j \sum_{\substack{m_j \in \mathcal{M}_j \\ m_j \neq \emptyset}} \Big[1 + (2q \log N)^{\frac{p}{2}-1}\Big] e^{-x_{m_j}}.$$

Finally, for $\breve{M}_{p,q}$ a constant only depending on p and q, we have:

$$\mathbb{E}\Big[\|\widetilde{\theta} - \theta\|_p^p\Big] \leq M_p \mathbb{E}\Big[\|\widehat{\theta}^{(m)} - \theta\|_p^p\Big] + 2\mathrm{pen}(m) + \breve{M}_{p,q}\Big(N^{1-q}\|\theta\|_p^p + \varepsilon^p R(\mathcal{M})\Big),$$

with

$$R(\mathcal{M}) = N^{(1-q)/2} \sum_{\lambda \in \Lambda} w_\lambda + (\log N)^{\frac{p}{2}-1} \sum_{j \in \mathcal{J}} \omega_j \sum_{\substack{m_j \in \mathcal{M}_j \\ m_j \neq \emptyset}} e^{-x_{m_j}}.$$

5.2 Proof of Theorem 3.2

We first prove that $\|\theta\|_p < \infty$.

Lemma 5.3. *Assume that $\theta \in \mathcal{B}^s_{r,\infty}(R)$ with $s > \frac{1}{r}$. Then, there exists C, only depending on s, r and p, such that*

$$\|\theta\|_p^p \leq C R^p.$$

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,

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

since $s > \frac{1}{r}$. Now, assume that $p < r$. Hölder's inequality implies that

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

and

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

Finally, in both cases,

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

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 Theorem 2.5 to the collection $\mathfrak{M}$ and penalty $\mathfrak{pen}$ and we obtain

$$\mathbb{E}\|\widehat{\theta}^{(\widehat{m}^{\widehat{a}})} - \theta\|_p^p \leq \widetilde{M}_p \inf_{(m,a)\in\mathfrak{M}} \Big\{\mathbb{E}\|\widehat{\theta}^{(m)} - \theta\|_p^p + \mathfrak{pen}(m,a)\Big\} + \breve{M}_p \varepsilon^p R(\mathfrak{M}). \tag{32}$$

When $p > 2$, we apply Theorem 2.7 to the collection $\mathfrak{M}$ and penalty $\mathfrak{pen}$. This gives

$$\mathbb{E}\|\widehat{\theta}^{(\widehat{m}^{\widehat{a}})} - \theta\|_p^p \leq \widetilde{M}_p \inf_{(m,a)\in\mathfrak{M}} \Big\{\mathbb{E}\|\widehat{\theta}^{(m)} - \theta\|_p^p + \mathfrak{pen}(m,a)\Big\} + \breve{M}_{p,q}\Big(N^{1-q}\|\theta\|_p^p + \varepsilon^p R^{\#}(\mathfrak{M})\Big),$$

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

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

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

We will prove in the following Sections 5.2.1, 5.2.2 , 5.2.3 that for each $a \in \{H, I, S\}$, $R(\mathcal{M}^a)$ 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 5.3, the remaining term is bounded as follows (for $q \geq p+1$):

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

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.

$$\inf_{(m,a)\in\mathfrak{M}} \left\{ \mathbb{E}\|\widehat{\theta}^{(m)} - \theta\|_p^p + \mathfrak{pen}(m,a) \right\} = \min_{a\in\{H,I,S\}} \inf_{m\in\mathcal{M}^a} \left\{ \mathbb{E}\|\widehat{\theta}^{(m)} - \theta\|_p^p + \mathrm{pen}^a(m) \right\}.$$

In the following Sections 5.2.1 (homogeneous case), 5.2.2 (intermediate case), 5.2.3 (sparse and frontier cases), we bound for each $a \in \{H, I, S\}$ the quantities $R(\mathcal{M}^a)$ and we prove that

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

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 we drop the upperscript a for ease of notation.

5.2.1 Proof of Theorem 3.2: homogeneous case

In this section we assume that $r \geq p$. 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 $\mathcal{B}^s_{r,\infty}(R)$ we can prove that $B_p(m) \lesssim R^p 2^{-Lps}$. Indeed the bias verifies

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

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 $\theta \in \mathcal{B}^s_{r,\infty}(R)$, 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

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

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

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

with

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

and

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

Thus, $\mathrm{p}_j(m_j) = 0$ if $j > L$; and for $j \leq L$:

$$\mathrm{p}_j(m_j) = \frac{3}{2}\sigma_p^p 2^j + \kappa_p 2^{\frac{(p-2)_+}{2}} (2^j)^{\left(1-\frac{p}{2}\right)_+} (Kj \log 2)^{\frac{p}{2}} \leq C(p,\sigma_p) 2^j,$$

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

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

which provides

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

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

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

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

5.2.2 Proof of Theorem 3.2: intermediate case

In this section, we assume that $\frac{p}{2s+1} < r < p$. Let us now consider the following model, inspired from [Mas07]: 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 cardinal $\lfloor 2^{L+l} A(l) \rfloor$ 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

$$x_{m_j} := K|m_j| \left(1 + \log\left(\frac{2^j}{|m_j|}\right)\right) \tag{33}$$

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:

$$\begin{cases} x_{m_j} = K|m_j| & \text{for the first type,} \\ x_{m_j} \approx K|m_j|\Big(1 - \log A(j-L)\Big) & \text{for the second type.} \end{cases} \tag{34}$$

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 5.2, we have to bound $R(\mathcal{M})$, that is to show that the term $\sum_{j=-1}^{J} \omega_j \sum_{m_j \in \mathcal{M}_j, m_j \neq \emptyset} |m_j|^{\left(1-\frac{p}{2}\right)_+} e^{-x_{m_j}}$ 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 $\mathcal{M}_j(L) = \big\{m_j : \ m \in \mathcal{M}(L)\big\}$,

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

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

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

for $K > 0$. Furthermore,

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

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

$$\log|\mathcal{M}_j(L)| \ \leq \ \log\binom{2^{L+l}}{|m_{L+l}|} \leq |m_{L+l}| \log\left(\frac{e2^{L+l}}{|m_{L+l}|}\right)$$

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

$$\begin{aligned} \log|\mathcal{M}_j(L)| &\leq 2^L 2^{-lb}(l+1)^{-3} \log\Big(e2^{L+l}\lfloor 2^{L+l} A(l)\rfloor^{-1}\Big) \\ &\leq 2^L 2^{-lb}(l+1)^{-3} \log\left(\frac{e}{2^{-lp/2}(l+1)^{-3} - 2^{-(L+l)}}\right) \\ &\leq C 2^L 2^{-lb}(l+1)^{-2}, \end{aligned}$$

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

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

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 $C - Kp\log(2)/2 < 0$,

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

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

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

with $\widetilde{K}$ 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,

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

with K' a positive constant. Finally, we have proved that that $R(\mathcal{M})$ is bounded by $\log N$ up to a constant. It means that in (32), the last term $\breve{M}_p \varepsilon^p R(\mathfrak{M})$ is bounded by $\varepsilon^p |\log(\varepsilon)|$, which is negligible when compared to the rate.

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

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

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

$$\inf_{m\in\mathcal{M}}\left\{\mathbb{E}\|\widehat{\theta}^{(m)}-\theta\|_p^p+\mathrm{pen}(m)\right\}\le\inf_L\left\{\sup_{m\in\mathcal{M}(L)}\mathbb{E}[V_p(m)]+\inf_{m\in\mathcal{M}(L)}B_p(m)+\sup_{m\in\mathcal{M}(L)}\mathrm{pen}(m)\right\}. \tag{35}$$

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

$$\begin{aligned}\mathbb{E}[V_p(m)] &\le \varepsilon^p\sigma_p^p\left(\sum_{j<L}2^{j(p/2-1)}2^j+\sum_{l\ge 0}2^{(L+l)(\frac{p}{2}-1)}2^L2^{-l(p/2-1)}(l+1)^{-3}\right)\\ &\lesssim \varepsilon^p\sigma_p^p\left(2^{Lp/2}+2^{L(p/2-1+1)}\right)\lesssim\varepsilon^p\sigma_p^p2^{Lp/2}\end{aligned}$$

Therefore

$$\sup_{m\in\mathcal{M}(L)}\mathbb{E}[V_p(m)]\lesssim\varepsilon^p\sigma_p^p2^{Lp/2}.$$

Moreover we can prove the following lemma (see Section 5.2.4)

Lemma 5.4. *If $\theta\in\mathcal{B}^s_{r,\infty}(R)$ and $\frac{p}{2s+1}<r$*

$$\inf_{m\in\mathcal{M}(L)}B_p(m)\lesssim R^p2^{-spL}.$$

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

case $p>2$: We have

$$\mathrm{p}_j(m_j)\le(2q\log N)^{\frac{p}{2}-1}\big(\frac{3}{2}\sigma_2^2|m_j|+\kappa_2x_{m_j}\big)$$

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

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

Finally equation (35) provides in the case $p > 2$

$$\inf_{m\in\mathcal{M}}\Big\{\mathbb{E}\|\widehat{\theta}^{(m)}-\theta\|_p^p+\mathrm{pen}(m)\Big\}\lesssim\inf_{L}\Big\{\varepsilon^p 2^{Lp/2}+R^p 2^{-spL}+\varepsilon^p(\log N)^{\frac{p}{2}-1}2^{Lp/2}\Big\}$$

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

$$\inf_{m\in\mathcal{M}}\Big\{\mathbb{E}\|\widehat{\theta}^{(m)}-\theta\|_p^p+\mathrm{pen}(m)\Big\}\lesssim R^{\frac{p}{2s+1}}(\varepsilon^{2p}|\log(\varepsilon)|^{p-2})^{\frac{s}{2s+1}}.$$

case $p\le 2$: We have

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

We just have to deal with the following term:

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

since $3-p/2>1$. Finally equation(35) provides in the case $p\le 2$

$$\inf_{m\in\mathcal{M}}\Big\{\mathbb{E}\|\widehat{\theta}^{(m)}-\theta\|_p^p+\mathrm{pen}(m)\Big\}\lesssim\inf_{L}\Big\{\varepsilon^p 2^{Lp/2}+R^p 2^{-spL}+\varepsilon^p 2^{Lp/2}\Big\}\lesssim R^{\frac{p}{2s+1}}\varepsilon^{\frac{2sp}{2s+1}}.$$

5.2.3 Proof of Theorem 3.2: sparse and frontier case

In this section we assume that $r\le\frac{p}{2s+1}$. Since $s>1/r$, it only occurs when $p>2$ (since $p\ge r(2s+1)\Rightarrow p>2+r>2$).

Recall that our model collection is defined by $\mathcal{M}_j=\{\{j\}\times E,\ E\in\mathcal{P}(K_j)\}$ for any $j\ge -1$. 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 5.2 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

$$\begin{aligned}R(\mathcal{M})&=\sum_{j=-1}^{J}\omega_j\sum_{d=1}^{2^j}\sum_{m_j\in\mathcal{M}_j,|m_j|=d}e^{-x_{m_j}}\\&=\sum_{j=-1}^{J}\omega_j\sum_{d=1}^{2^j}\mathrm{card}\{m_j\in\mathcal{M}_j,|m_j|=d\}e^{-Kdj}.\end{aligned}$$

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

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

The following remark will be useful in Section 5.3.4.

Remark 5.5. *We also have:*

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

It remains to control the term $\inf_{m\in\mathcal{M}}\left\{\mathbb{E}\|\widehat{\theta}^{(m)}-\theta\|_p^p + \mathrm{pen}(m)\right\}$. Taking inspiration from the various works of Donoho, Johnstone, Kerkyacharian 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 $s > 1/r$,

$$0 < \beta < \min\Big(\frac{1}{2}, s'\Big).$$

Moreover, since $R\varepsilon^{-1} \geq 1$,

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

Then, the model $\check{m} = \cup_j \check{m}_j$ belongs to $\mathcal{M}$ (even if $\check{m}$ depends on θ). It satisfies the following property, proved in Section 5.2.5.

Proposition 5.6. *Assume that $r < p/(2s+1)$. There exists a positive constant C (depending on s, r, p) such that*

$$\sup_{\theta \in \mathcal{B}^s_{r,\infty}(R)} \mathbb{E}\|\widehat{\theta}^{(\check{m})} - \theta\|_p^p \leq CR^{p(1-2\beta)} |\log\varepsilon|^{p\beta}\, \varepsilon^{2p\beta}$$

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

$$\sup_{\theta \in \mathcal{B}^s_{r,\infty}(R)} \mathbb{E}\|\widehat{\theta}^{(\check{m})} - \theta\|_p^p \leq CR^{p(1-2\beta)} |\log\varepsilon|^{p\beta+1}\, \varepsilon^{2p\beta}.$$

The following remark will be useful in Section 5.3.4.

Remark 5.7. *We also have, for $r < p/(2s+1)$,*

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

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 $m = \check{m}$, the term

$$\mathrm{pen}(m) = 2\varepsilon^p \sum_{j=-1}^{J} \omega_j \min\Big(\mathrm{p}_j(m_j), (2q\log N)^{\frac{p}{2}-1} \mathrm{p}_j^{\#}(m_j)\Big)$$

with

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

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

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

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

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

For $j_1 \le j \le j_0$, 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 5.6. 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 $\log(N) \lesssim |\log(\varepsilon)|$,

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

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

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

) Finally, reminding Proposition 5.6, we obtain

$$\inf_{m\in\mathcal{M}} \left\{ \mathbb{E}\|\widehat{\theta}^{(m)} - \theta\|_p^p + \mathrm{pen}(m) \right\} \le \mathbb{E}\|\widehat{\theta}^{(\check{m})} - \theta\|_p^p + \mathrm{pen}(\check{m}) \lesssim R^{p(1-2\beta)}(\varepsilon\sqrt{|\log\varepsilon|})^{2\beta p}.$$

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

5.2.4 Proof of Lemma 5.4

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

$$|\theta_{j,(1)}| \ge |\theta_{j,(2)}| \ge \cdots \ge$$

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 5.8. *If $a_{(1)} \geq \cdots \geq a_{(N)} \geq 0$, 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 5.8, we can write

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

Since $\theta \in \mathcal{B}^s_{r,\infty}(R)$, 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

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

The series converges because $\frac{p}{2r} - \frac{1}{2} - s < 0 \Leftrightarrow \frac{p}{2s+1} < r$. Lemma 5.4 is proved.

5.2.5 Proof of Proposition 5.6

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

$$\|\widehat{\theta} - \theta\|_p^p = \sum_j \sum_{k \in K_j} \omega_j |\widehat{\theta}_{jk} - \theta_{jk}|^p = A_1 + A_2 + A_3,$$

with

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

We bound each sum.

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

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

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

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

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

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

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

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

We can split the last term in the following way:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

But remember that

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

Then

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

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 $p - r = 2p\beta$,

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

5.3 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 give the proof of intermediary technical results stated in Section 4.

5.3.1 Proof of Proposition 4.2

The proof of Proposition 4.2 needs following lemmas. The first one recalls classical facts about Orlicz norms.

Lemma 5.9. *Let ξ be a sub-Gaussian random variable.*

1. $(\mathbb{E}|\xi|^p)^{1/p} \leq 2\sqrt{p} \|\xi\|_{\psi_2}$.

2. $\||\xi|^p\|_{\psi_{2/p}} = \|\xi\|_{\psi_2}^p$.

3. *Let $X = |\xi|^p - \mathbb{E}|\xi|^p$. There exists C_1 a positive constant only depending on p such that $\|X\|_{\psi_{2/p}} \leq C_1 \|\xi\|_{\psi_2}^p$.*

Proof of Lemma 5.9. 1. See [Ver18] Proposition 2.5.2 (sub-Gaussian properties) and its proof, as well as Definition 2.5.5.

2. This comes directly from the definitions of the Orlicz norms.

3. We can prove that for any variable Y we have $\|Y - \mathbb{E}Y\|_{\psi_{2/p}} \leq C_1 \|Y\|_{\psi_{2/p}}$ similarly as Lemma 2.6.6 in [Ver18] (it uses the triangular inequality and the fact that $\mathbb{E}|Y| \leq C(p)\|Y\|_{\psi_{2/p}}$). Then we take $Y = |\xi|^p$ and we use the previous point.

□

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 5.10. *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 $\lambda \in I$, $(\mathbb{E}|\xi_\lambda|^p)^{1/p} \leq 2\sqrt{p}\tau$. Moreover, denoting*

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

we have

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

with $\sigma_p := 2\sqrt{p}\tau$ and $D = card(\mathcal{I})$. Furthermore, for any $x \geq 1$, with probability larger than $1 - 2\exp(-x)$,

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

where κ_p is a positive constant only depending on p and σ_p.

Proof of Lemma 5.10. Using the first point of Lemma 5.9 and our assumption on uniform subgaussianity, we have:

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

Now we apply Theorem 2.3. Recall that $b_\lambda = \|X_\lambda\|_{\psi_{2/p}}$ with $X_\lambda = |\xi_\lambda|^p - \mathbb{E}|\xi_\lambda|^p$. The third point of Lemma 5.9 gives:

$$\|b\|_{\ell_q}^q = \sum_{\lambda \in \mathcal{I}} \|X_\lambda\|_{\psi_{2/p}}^q \leq \sum_{\lambda \in \mathcal{I}} (C_1 \|\xi_\lambda\|_{\psi_2}^p)^q \leq (C_1 \tau^p)^q |\mathcal{I}|$$

so that $\|b\|_{\ell_q} \leq C_1 \tau^p D^{1/q}$. Theorem 2.3 gives that with probability larger than $1 - 2e^{-x}$,

$$\begin{aligned} |Z - \mathbb{E}(Z)| &\leq d_{1,p}\|b\|_{\ell_2}\sqrt{x} + d_{2,p}\|b\|_{\ell_{1/(1-p/2)_+}} x^{p/2} \\ &\leq d_{1,p}C_1\tau^p\sqrt{Dx} + d_{2,p}C_1\tau^p D^{(1-p/2)_+} x^{p/2} \\ &\leq d'_{1,p}\sigma_p^p\sqrt{Dx} + d'_{2,p}\sigma_p^p D^{(1-p/2)_+} x^{p/2} \end{aligned}$$

with $d'_{ip} = d_{ip}C_1/(2\sqrt{p})^p$. Recalling that $\mathbb{E}(Z) \leq \sigma_p^p D$, we can obtain with probability larger than $1 - 2\exp(-x)$,

$$Z \leq \frac{3}{2}\sigma_p^p D + \kappa_p D^{\left(1-\frac{p}{2}\right)_+} x^{\frac{p}{2}},$$

with the same proof as the one of Corollary 2.4. □

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 $t \in \mathbb{R}$,

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

(see Proposition 2.5.2 of [Ver18], actually $c = 2\|\eta_1\|_{\psi_2}$) 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 contant 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 [Ver18], 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

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

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 $\mathcal{I}_j$

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

where the set $(\mathcal{I}_{j\ell})_\ell$'s 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 5.10, with $\tau = \sup_{jk} \|\xi_{jk}\|_{\psi_2}$. Let $x \geq 1$. For any $1 \leq \ell \leq K$

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

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

$$\begin{aligned} \sum_{k \in \mathcal{I}_j} |\xi_{jk}|^p = \sum_{\ell=1}^{K} \sum_{k \in \mathcal{I}_{j\ell}} |\xi_{jk}|^p &< \frac{3}{2} \sigma_p^p \sum_{\ell=1}^{K} |\mathcal{I}_{j\ell}| + \kappa_p \sum_{\ell=1}^{K} |\mathcal{I}_{j\ell}|^{\left(1 - \frac{p}{2}\right)_+} x^{\frac{p}{2}} \\ &< \frac{3}{2} \sigma_p^p |\mathcal{I}_j| + \kappa_p K^{1 - \left(1 - \frac{p}{2}\right)_+} \left(\sum_{\ell=1}^{K} |\mathcal{I}_{j\ell}| \right)^{\left(1 - \frac{p}{2}\right)_+} x^{\frac{p}{2}} \end{aligned}$$

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

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

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

5.3.2 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 [Dau92] ensures that ψ is orthogonal to polynomials of degree less or equal to M. Therefore Assumption 1 of [DJ97] is satisfied. We then use the following result.

Proposition 5.11 (Proposition 2 of [DJ97]). *We assume that ϕ and ψ are C^{M+1} and that $f \in \mathcal{B}^s_{r,\infty}(R)$, with $1/r < s < M+1$. Then, if $p \geq r$,*

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

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

Now we consider the three cases:
- If $r < \frac{p}{2s+1}$, since $s > 1/r$,

$$\begin{aligned}
\Big\| \sum_{\lambda \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \Big\|^p_{\mathbb{L}_p} &\lesssim R^p N^{-p(s-\frac{1}{r}+\frac{1}{p})} \\
&\lesssim R^p \big(|\log(\varepsilon)| \varepsilon^2 \big)^{p(s-\frac{1}{r}+\frac{1}{p})} \\
&\lesssim R^p \big(|\log(\varepsilon)| \varepsilon^2 \big)^{p \frac{s-\frac{1}{r}+\frac{1}{p}}{2s+1-\frac{2}{r}}}.
\end{aligned}$$

- If $\frac{p}{2s+1} \leq r \leq p$, observe that

$$\begin{aligned}
r \geq \frac{p}{2s+1} &\iff p \leq r(1+2s) \iff -1/p \leq -1/(r(1+2s)) \\
&\iff 1/r - 1/p \leq 1/r(1 - 1/(1+2s)) \\
&\iff 1/r - 1/p \leq 1/r \times 2s/(1+2s).
\end{aligned}$$

Therefore, using $s > 1/r$, we have:

$$1/r - 1/p < 2s^2/(1+2s),$$

which means that

$$s - 1/r + 1/p > s/(1+2s)$$

and

$$\begin{aligned}
\Big\| \sum_{\lambda \in \Lambda^{(N)}} \theta_{jk} \varphi_{jk} - f \Big\|^p_{\mathbb{L}_p} &\lesssim R^p \big(|\log(\varepsilon)| \varepsilon^2 \big)^{p(s-\frac{1}{r}+\frac{1}{p})} \\
&\lesssim R^p \varepsilon^{\frac{2ps}{1+2s}}. \qquad (36)
\end{aligned}$$

- If $r \geq p$, we have, since functions are compactly supported,

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

using (36).

5.3.3 Proof of Lemma 4.4

In the sequel, we assume that $\|g\|^{p \wedge 2}_{\mathcal{B}^0_{p,p \wedge 2}} < \infty$.

Section 9.2 of Daubechies (1992) shows that for $1 < p < \infty$

$$g \in \mathbb{L}_p \iff \Big[\sum_{j \geq -1} \sum_{k \in K_j} \big| \langle g, \varphi_{jk} \rangle \big|^2 \varphi^2_{jk}(\cdot) \Big]^{\frac{1}{2}} \in \mathbb{L}_p \tag{37}$$

$$\iff \Big[\sum_{j \geq -1} \sum_{k \in K_j} \big| \langle g, \varphi_{jk} \rangle \big|^2 2^j 1_{[2^{-j}k, 2^{-j}(k+1)]}(\cdot) \Big]^{\frac{1}{2}} \in \mathbb{L}_p \tag{38}$$

Case $1 < p \leq 2$**.** We use (37). Since $\| \cdot \|_{\ell_2} \leq \| \cdot \|_{\ell_p}$,

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

and

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

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

Case $p \geq 2$**.** We use (38) We recall the generalized Minkowski inequality: Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ two σ-finite measure spaces and F a measurable function. Then, for any $q \in [1, +\infty)$,

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

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

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

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

Case $p = 1$. Finally, we deal with the case $p = 1$. We have

$$\|g\|_{\mathcal{B}^0_{p,p\wedge 2}}^{p\wedge 2} = \sum_{j\geq -1} 2^{-j/2} \sum_{k\in K_j} \big|\langle g, \varphi_{jk}\rangle\big|$$

and, with $g = \sum_{j\geq -1} \sum_{k\in K_j} \langle g, \varphi_{jk}\rangle \varphi_{jk}$,

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

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

5.3.4 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 [DJ97] and the assumption $f \in \mathcal{B}^s_{r,\infty}(R)$. It uses that $\log_2(n)$ is an integer.

Lemma 5.12. *If f belongs to the Besov set $\mathcal{B}^s_{r,\infty}(R)$, then we have*

$$\sum_{k\in K_j} |\theta_{jk}|^r \leq CR^r 2^{-jr(s+1/2-1/r)}, \tag{39}$$

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 (28) combined with Proposition 4.2, Theorem 3.2, Lemmas 5.12 and 4.3 and Inequality (29).

When $p > 2$, Inequality (29) does not hold, but Inequality (30) shows that we only have to bound

$$\left[\sum_{j=-1}^{J}\left(\mathbb{E}\Big[2^{j(\frac{p}{2}-1)}\sum_{k\in K_j}\big|\widehat{\theta}_{jk}^{(\widehat{m})}-\theta_{jk}\big|^p\Big]\right)^{\frac{2}{p}}\right]^{\frac{p}{2}} \tag{40}$$

to conclude.

Remember that $\widehat{m} = \arg\min_{m\in\mathcal{M}} Crit(m)$ with

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

and $\mathrm{pen}(m)$ has the form $\mathrm{pen}(m) = 2\varepsilon^p \sum_{j=1}^J \omega_j P_j(m_j)$ (see (20)). This can be rewritten

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

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

$$m_j \longmapsto Crit_j(m_j) := \sum_{k\in K_j}|Y_{jk}|^p - 2\varepsilon^p P_j(m_j).$$

Let m_j be some model in $\mathcal{M}_j$. Replacing for $j' \neq j$, $w_{j'}$ by 0 we obtain directly from Theorem 2.1, for any model m_j,

$$\|\widetilde{\theta}_{j\cdot} - \theta_{j\cdot}\|_p^p \leq M_p \|\widehat{\theta}^{(m_j)} - \theta_{j\cdot}\|_p^p + 2\Big[2V_p(\widehat{m}_j) - \mathrm{pen}(\widehat{m}_j)\Big] - 2\Big[2V_p(m_j) - \mathrm{pen}(m_j)\Big],$$

which means

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

With $Z(m_j) = \sum_{k\in m_j} |\xi_{jk}|^p$, mimicking the proof of Theorem 2.7, by taking q large enough,

$$\sum_{j=-1}^{J}\Big(2^{j(\frac{p}{2}-1)}\sum_{k\in K_j}\mathbb{E}\Big[\big|\widetilde{\theta}_{jk}-\theta_{jk}\big|^p\Big]\Big)^{\frac{2}{p}} \lesssim \sum_{j\geq -1}\Big(2^{j(\frac{p}{2}-1)}\sum_{k\in K_j}\mathbb{E}\Big[\big|\widehat{\theta}_k^{(m_j)}-\theta_{jk}\big|^p\Big]+\varepsilon^p 2^{j(\frac{p}{2}-1)}P_j(m_j)\Big)^{\frac{2}{p}}$$
$$+(\log N)^{1-\frac{2}{p}}\varepsilon^2\sum_{j=-1}^{J}\Big(2^{j(\frac{p}{2}-1)}\sum_{\substack{m_j\in\mathcal{M}_j\\ m_j\neq\emptyset}} e^{-x_{m_j}}\Big)^{\frac{2}{p}}+O(\varepsilon^2).$$

We have

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

with

$$B_p(m_j) := 2^{j(\frac{p}{2}-1)}\sum_{k\notin m_j}|\theta_{jk}|^p \quad\text{and}\quad V_p(m_j) := \varepsilon^p\sum_{k\in m_j}2^{j(\frac{p}{2}-1)}|\xi_{jk}|^p. \tag{41}$$

This gives

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

and

$$\begin{aligned}
A(m) &:= \sum_{j=-1}^{J}\Big(2^{j(\frac{p}{2}-1)}\sum_{k\in K_j}\mathbb{E}\Big[\big|\widehat{\theta}_k^{(m_j)}-\theta_{jk}\big|^p\Big]+\varepsilon^p 2^{j(\frac{p}{2}-1)}P_j(m_j)\Big)^{\frac{2}{p}}\\
&\lesssim \sum_{j=-1}^{J}\Big(B_p(m_j)+\varepsilon^p\sigma_p^p 2^{j(\frac{p}{2}-1)}|m_j|+\varepsilon^p 2^{j(\frac{p}{2}-1)}P_j(m_j)\Big)^{\frac{2}{p}}\\
&\lesssim \sum_{j=-1}^{J}B_p^{\frac{2}{p}}(m_j)+\sum_{j=-1}^{J}\Big(\varepsilon^p\sigma_p^p 2^{j(\frac{p}{2}-1)}|m_j|\Big)^{\frac{2}{p}}+\sum_{j=-1}^{J}\Big(\varepsilon^p 2^{j(\frac{p}{2}-1)}P_j(m_j)\Big)^{\frac{2}{p}}\\
&=: \widetilde{B}_p(m)+\widetilde{V}_p(m)+\widetilde{P}_p(m).
\end{aligned}$$

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

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

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 $L \leq J$

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

Then we can prove that $\widetilde{V}_p(m) \lesssim \varepsilon^2 2^L$, $\widetilde{B}_p(m) \lesssim R^2 2^{-Ls}$ for $\theta \in \mathcal{B}^s_{r,\infty}(R)$ and $\widetilde{P}_p(m) \leq \varepsilon^2 2^L$. Moreover $\widetilde{R}(\mathcal{M}) \lesssim (\log N)^{1-\frac{2}{p}} \varepsilon^2$. 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 $\mathcal{M} = \bigcup_{L=0}^{J} \mathcal{M}(L)$ and m belongs to $\mathcal{M}(L)$ if

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

but this time $\widetilde{A}(l) := 2^{-lp/2}(l+1)^{-3p/2}$. We follow Section 5.2.2 step by step, keeping in mind that $p > 2$ and $b = \frac{p}{2} - 1$. We have $\widetilde{R}(\mathcal{M}) = (\log N)^{1-\frac{2}{p}} \varepsilon^2 (T_1 + T_2)$ 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

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

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 $\widetilde{R}(\mathcal{M}) \lesssim (\log N)^{1-\frac{2}{p}}\varepsilon^2$. We now deal with $\widetilde{V}_p(m)$: for any $m \in \mathcal{M}(L)$,

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

For the study of $\widetilde{B}_p(m)$, we follow the proof of Lemma 5.4 *mutatis mutandis*:

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

Since $\theta \in \mathcal{B}^s_{r,\infty}(R)$, 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

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

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

It remains to control $\widetilde{P}_p(m) = \sum_{j=-1}^{J}\left(\varepsilon^p 2^{jb} P_j(m_j)\right)^{2/p}$ with P_j given through (20). 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 $m \in \mathcal{M}(L)$,

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

We conclude

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

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 5.5 gives $\widetilde{R}(\mathcal{M}) \lesssim (\log N)^{1-\frac{2}{p}} \varepsilon^2$ and Remark 5.7 provides

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

Finally, following the outlines of the end of Section 5.2.3, we can prove

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

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

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

This ends the proof.

Acknowledgements

We are grateful to Radosław Adamczak for his advices about concentration inequalities for sub-Weibull variables.

References

[ACE25] Sergios Agapiou, Ismaël Castillo, and Paul Egels. Heavy-tailed and Horseshoe priors for regression and sparse Besov rates. 2025. Manuscript in preparation.

[Aka73] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory (Tsahkadsor, 1971), pages 267–281. Akad. Kiadó, Budapest, 1973.

[Bar11] Yannick Baraud. Estimator selection with respect to Hellinger-type risks. Probab. Theory Related Fields, 151(1-2):353–401, 2011.

[BH79] Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. Z. Wahrsch. Verw. Gebiete, 47(2):119–137, 1979.

[Bir86] Lucien Birgé. On estimating a density using Hellinger distance and some other strange facts. Probab. Theory Relat. Fields, 71(2):271–291, 1986.

[BM01] Lucien Birgé and Pascal Massart. Gaussian model selection. J. Eur. Math. Soc. (JEMS), 3(3):203–268, 2001.

[BM07] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. Probab. Theory Related Fields, 138(1-2):33–73, 2007.

[CDV93] Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. Appl. Comput. Harmon. Anal., 1(1):54–81, 1993.

[Dau92] Ingrid Daubechies. Ten lectures on wavelets, volume 61 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[DJ97] Bernard Delyon and Anatoli Juditsky. On the computation of wavelet coefficients. J. Approx. Theory, 88(1):47–79, 1997.

[DJ98] David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. Ann. Statist., 26(3):879–921, 1998.

[DJKP95] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? Journal of the Royal Statistical Society: Series B (Methodological), 57(2):301–337, 1995.

[DJKP96] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. Ann. Statist., 24(2):508–539, 1996.

[DJKP97] David L Donoho, Iain M Johnstone, G Kerkyacharian, and Dominique Picard. Universal near minimaxity of wavelet shrinkage. In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics, pages 183–218. Springer, 1997.

[DL96] Luc Devroye and Gábor Lugosi. A universally acceptable smoothing factor for kernel density estimates. Ann. Statist., 24(6):2499–2512, 1996.

[GK95] Efim D Gluskin and Stanisław Kwapień. Tail and moment estimates for sums of independent random variables with logarithmically concave tails. Studia Mathematica, 3(114):303–309, 1995.

[GL08] Alexander Goldenshluger and Oleg Lepski. Universal pointwise selection rule in multivariate function estimation. Bernoulli, 14(4):1150–1190, 2008.

[GL11] Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. Ann. Statist., 39(3):1608–1632, 2011.

[GL13] A. V. Goldenshluger and O. V. Lepski. General selection rule from a family of linear estimators. Theory Probab. Appl., 57(2):209–226, 2013.

[GL14] A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on R^d. Probab. Theory Related Fields, 159(3-4):479–543, 2014.

[Gol09] Alexander Goldenshluger. A universal procedure for aggregating estimators. Ann. Statist., 37(1):542–568, 2009.

[HKP99] Peter Hall, Gérard Kerkyacharian, and Dominique Picard. On the minimax optimality of block thresholded wavelet estimators. Statistica Sinica, pages 33–49, 1999.

[HKPT98] Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. Wavelets, approximation, and statistical applications, volume 129 of Lecture Notes in Statistics. Springer-Verlag, New York, 1998.

[HMSO97] Paweł Hitczenko, Stephen J Montgomery-Smith, and Krzysztof Oleszkiewicz. Moment inequalities for sums of certain independent symmetric random variables. Studia Math, 123(1):15–42, 1997.

[IH80] Il'dar A. Ibragimov and Rafail Z. Hasminskii. An estimate of the density of a distribution. Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 98:61–85, 161–162, 166, 1980. Studies in mathematical statistics, IV.

[Joh19] Ian M. Johnstone. Gaussian estimation: sequence and wavelet models. Draft version available on the author's webpage, 2019.

[JS05] Iain M. Johnstone and Bernard W. Silverman. Empirical Bayes selection of wavelet thresholds. Ann. Statist., 33(4):1700–1752, 2005.

[Jud97] Anatoli Juditsky. Wavelet estimators: adapting to unknown smoothness. Mathematical Methods of Statistics, 6(1):1–25, 1997.

[KC22] Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. Information and Inference: A Journal of the IMA, 11(4):1389–1456, 2022.

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

[KP00] Gérard Kerkyacharian and Dominique Picard. Thresholding algorithms, maxisets and well-concentrated bases. Test, 9:283–344, 2000.

[KPT96] Gérard Kerkyacharian, Dominique Picard, and Karine Tribouley. L^p adaptive density estimation. Bernoulli, 2(3):229–247, 1996.

[Lep91] Oleg V Lepski. Asymptotically minimax adaptive estimation. i. upper bounds. optimally adaptive estimates. Teoriya Veroyatnostei i ee Primeneniya, 36(4):645–659, 1991.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. Ann. Statist., pages 1302–1338, 2000.

[LMR17] Claire Lacour, Pascal Massart, and Vincent Rivoirard. Estimator selection: a new method with applications to kernel density estimation. Sankhya A, 79(2):298–335, 2017.

[LMS97] Oleg V Lepski, Enno Mammen, and Vladimir G Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. The Annals of Statistics, pages 929–947, 1997.

[Mal73] Colin L. Mallows. Some comments on Cp. Technometrics, 15(4):661–675, 1973.

[Mas07] Pascal Massart. Concentration Inequalities and Model Selection (Saint-Flour Summer School on Probability Theory 2003, ed. J. Picard). Springer, 2007.

[Nem85] Arkadi S. Nemirovskiy. Nonparametric estimation of smooth regression functions. Soviet J. Comput. Systems Sci., 23(6):1–11, 1985.

[Riv04] Vincent Rivoirard. Maxisets for linear procedures. Statistics & probability letters, 67(3):267–275, 2004.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. Ann. Statist., 6(2):461–464, 1978.

[Sto82] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. Ann. Statist., 10(4):1040–1053, 1982.

[Ver18] Roman Vershynin. High-Dimensional Probability: An introduction with applications in data science, volume 47. Cambridge University Press, 2018.

[VLMR23] Suzanne Varet, Claire Lacour, Pascal Massart, and Vincent Rivoirard. Numerical performance of penalized comparison to overfitting for multivariate kernel density estimation. ESAIM Probab. Stat., 27:621–667, 2023.

[ZW22] Huiming Zhang and Haoyu Wei. Sharper sub-Weibull concentrations. Mathematics, 10(13):2252, 2022.