SIMULTANEOUS OFF-THE-GRID LEARNING OF MIXTURES ISSUED FROM A CONTINUOUS DICTIONARY

CRISTINA BUTUCEA, JEAN-FRANÇOIS DELMAS, ANNE DUTFOY, AND CLÉMENT HARDY

Abstract. In this paper we observe a set, possibly a continuum, of signals corrupted by noise. Each signal is a finite mixture of an unknown number of features belonging to a continuous dictionary. The continuous dictionary is parametrized by a real non-linear parameter. We shall assume that the signals share an underlying structure by saying that the union of active features in the whole dataset is finite.

We formulate regularized optimization problems to estimate simultaneously the linear coefficients in the mixtures and the non-linear parameters of the features. The optimization problems are composed of a data fidelity term and a (ℓ_1, L^p)-penalty. We prove high probability bounds on the prediction errors associated to our estimators. The proof is based on the existence of certificate functions. Following recent works on the geometry of off-the-grid methods, we show that such functions can be constructed provided the parameters of the active features are pairwise separated by a constant with respect to a Riemannian metric. When the number of signals is finite and the noise is assumed Gaussian, we give refinements of our results for $p = 1$ and $p = 2$ using tail bounds on suprema of Gaussian and χ^2 random processes. When $p = 2$, our prediction error reaches the rates obtained by the Group-Lasso estimator in the multi-task linear regression model.

MSC2020 subject classifictaion: Primary 62G08, secondary 62G05

Keywords: Continuous dictionary, Interpolating certificates, Mixture model, Multi-task learning, Non-linear regression model, Off-the-grid methods, Simultaneous recovery, Sparse spike deconvolution.

1. Introduction

Observing repeatedly the same process is very frequent nowadays, due to the abundance of data in all fields. Multi-task learning considers the simultaneous analysis of multiple datasets and produces an estimator for each dataset. Datasets can be either discrete-time (e.g. regression models) or continuous-time in our context. We assume that they bring information on the same underlying structure, but can also be contaminated at some extent by outliers.

We assume each process has a signal-plus-noise structure and that the signal is a mixture of features issued from a dictionary of smooth functions parametrized by some non-linear parameter (such as location, scale, etc.). Such mixtures can be seen e.g. in spectroscopy where each feature corresponds to a chemical component of the analyzed material, see [15].

We are interested in recovering simultaneously the signals, i.e. the linear weights in the mixture and the non-linear parameters of the features, by minimizing a weighted prediction risk penalized by the sum of the total energy of the weights that each feature has through the collection of all processes. The prediction risk may put more weight on prescribed signals of interest. We give high probability bounds on the weighted prediction risk that are analogous to the case of multi-task discrete linear regression models.

1.1. Model and method.

Let $(\mathcal{Z}, \mathcal{F}, \nu)$ be a measure space with ν a finite positive non-zero measure and let H_T be a Hilbert space where the parameter $T \in \mathbb{N}$ accounts for the increasing asymptotic information in the model. The Hilbert space H_T is endowed with the scalar product $\langle \cdot, \cdot \rangle_T$ and the norm $\|\cdot\|_T$. We shall consider the space $L_T = L^2(\nu, H_T)$, the set of H_T-valued strong measurable functions f defined on $(\mathcal{Z}, \mathcal{F}, \nu)$

Date: November 1, 2022.

such that $\|f\|_{L_T} = \sqrt{\int_{\mathcal{Z}} \|f(z)\|_T^2 \, \nu(\mathrm{d}z)}$ is finite. We then endow L_T with a scalar product noted $\langle \cdot, \cdot \rangle_{L_T}$ defined for any $f, g \in L_T$ by :

$$\langle f, g \rangle_{L_T} = \int \langle f(z), g(z) \rangle_T \, \nu(\mathrm{d}z).$$

The norm $\|\cdot\|_{L_T}$ is the natural norm associated with the scalar product and L_T is a Hilbert space, see [24, Section IV]. For $p \in [1, +\infty)$, we write $L^p(\nu, \mathbb{R}^K)$ for the space of $\mathbb{R}^K$-valued measurable function f defined on $(\mathcal{Z}, \mathcal{F}, \nu)$ such that

$$\|f\|_{L^p(\nu, \mathbb{R}^K)} = \left(\int_{\mathcal{Z}} \|f(z)\|_{\ell_2}^p \, \nu(\mathrm{d}z) \right)^{\frac{1}{p}}$$

is finite, where $\|\cdot\|_{\ell_2}$ is the usual Euclidean norm on $\mathbb{R}^K$. We simply write $L^p(\nu)$ for $L^p(\nu, \mathbb{R})$.

We assume we observe a random element Y of the Hilbert space L_T. For any $z \in \mathcal{Z}$, the element $Y(z) \in H_T$ has a signal-plus-noise structure. The signal part is a mixture (linear combination) of smooth features $\varphi_T(\theta)$ belonging to H_T and continuously parametrized by a real parameter $\theta \in \Theta \subseteq \mathbb{R}$. Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space, we note W_T the additional noise process defined on this space and assumed to be almost surely an element of L_T. We denote by $(\varphi_T(\theta), \theta \in \Theta)$ the continuous dictionary formed by all the features. For all $z \in \mathcal{Z}$, we note $\mathcal{Q}^\star(z)$ the finite set of the parameters of the active features appearing in $Y(z)$. We assume that the unknown number of active features s in the observation Y is bounded by a constant K, that is:

$$K \geq \mathrm{Card}(\bigcup_{z \in \mathcal{Z}} \mathcal{Q}^\star(z)) := s. \tag{1}$$

In the following we make a slight abuse of notation by writing $\mathcal{Q}^\star$ instead of $\bigcup_{z \in \mathcal{Z}} \mathcal{Q}^\star(z)$.

We consider features $\varphi_T(\theta)$ that are non degenerate, *i.e.* for any $\theta \in \Theta$, $\|\varphi_T(\theta)\|_T$ is finite and non-zero. Let us define the normalized function $\phi_T(\theta)$ for $\theta \in \Theta$ and its multivariate counterpart $\Phi_T(\vartheta)$ for $\vartheta = (\theta_1, \cdots, \theta_K) \in \Theta^K$ by :

$$\phi_T(\theta) = \frac{\varphi_T(\theta)}{\|\varphi_T(\theta)\|_T} \quad \text{and} \quad \Phi_T(\vartheta) = \begin{pmatrix} \phi_T(\theta_1) \\ \vdots \\ \phi_T(\theta_K) \end{pmatrix}.$$

We consider the model with unknown parameters $B^\star$ in $L^2(\nu, \mathbb{R}^K)$ and $\vartheta^\star$ in Θ^K:

$$Y = B^\star \Phi_T(\vartheta^\star) + W_T \quad \text{in } L_T. \tag{2}$$

In this work, we assume that the application $B^\star : \mathcal{Z} \to \mathbb{R}^K$ is $s-$ sparse that is,

$$1 \leq s < K \text{ with } s = \mathrm{Card}(S^\star) \text{ and } S^\star = \{k, \, \|B_k^\star\|_{L^2(\nu)} \neq 0 \, \}.$$

We remark that the model (2) is an extension of the model described in [14], as the latter amounts to taking ν as a Dirac measure. We gain in generality by letting the measure ν be any finite positive non-zero measure on $\mathcal{Z}$. By doing so, we can consider multiple mixture models.

Example 1.1 ($\mathcal{Z} = \{1, \cdots, n\}$). The framework presented above covers a large variety of multiple non-linear regression models. Assume we observe $n \in \mathbb{N}$ random elements of a Hilbert space. Assume that each element is a linear combination of features belonging to a continuous dictionary and is corrupted by a noise process. We encompass this model by indexing the n random elements, setting $\mathcal{Z}$ as the set of indices $\{1, \cdots, n\}$ and the measure ν as the counting measure on this set. The n observations in H are then $(Y(i), \, i = 1, \cdots, n)$.

We might be interested in associating to each observation $Y(i)$ a score indicating, for example, the reliability of the method of acquisition of the observed data. In this context, one can add the information to the model by assigning weights $\nu(i)$ to each process $Y(i)$ and average the prediction risk accordingly.

Example 1.2 ($\mathcal{Z}$ is a continuum). By letting $(\mathcal{Z}, \mathcal{F}, \nu)$ be any measure space such that $\nu(\mathcal{Z}) < +\infty$, we can take $\mathcal{Z}$ as a compact interval of $\mathbb{R}$ and ν as the Lebesgue measure on $\mathcal{Z}$. Hence, we generalize the "Function-on-Scalar" models that have many applications including in genomics (see [4]) by allowing the design matrix to be parametrized. The "Function-on-Scalar" models refer to regression models where the linear coefficients depend on a time or spatial continuous parameter. Thus, the observation $(Y(z), z \in \mathcal{Z})$ are longitudinal data.

In order to perform signal reconstruction, we are interested in recovering the application $B^\star$ with unknown sparsity s restricted to its support, that is $B^\star_{S^\star}$, and the associated parameters $\vartheta^\star_{S^\star}$ of the nonlinear parametric functions involved in the mixture model.

In order to recover the sparse application $B^\star$ as well as the associated parameters $\vartheta^\star_{S^\star}$ (up to a permutation) we solve a regularized optimization problem with a real tuning parameter $\kappa > 0$ and $p \in [1, 2]$:

$$(\hat{B}, \hat{\vartheta}) \in \underset{B \in L^2(\nu, \mathbb{R}^K), \vartheta \in \Theta_T^K}{\operatorname{argmin}} \quad \frac{1}{2\nu(\mathcal{Z})} \|Y - B\Phi_T(\vartheta)\|^2_{L_T} + \kappa \|B\|_{\ell_1, L^p(\nu)}, \tag{3}$$

where for $z \mapsto B(z) = (B_1(z), \ldots, B_K(z))$ in $L^2(\nu, \mathbb{R}^K)$:

$$\|B\|_{\ell_1, L^p(\nu)} = \sum_{k=1}^K \|B_k\|_{L^p(\nu)}.$$

The set Θ_T on which the optimization of the non-linear parameters is performed is required to be a compact interval and the function Φ_T is continuous. When $\mathcal{Z}$ is finite, the existence of at least a solution is therefore guaranteed. When $\mathcal{Z}$ is infinite (and $p \in (1, 2]$), we may use the following result whose proof is given in Section A.1.

Proposition 1.3. *Let $p \in (1, 2]$. Assume that the function $\theta \mapsto \phi_T(\theta)$ is continuous. Then, the minimization problem* (3) *over $L^2(\mathcal{Z}, \mathbb{R}^K) \times \Theta_T^K$, where Θ_T is a compact interval of $\mathbb{R}$, admits at least one solution.*

In the following, we shall assume that $p \in [1, 2]$. This will allow us to control norms of elements in the dual space $L^q(\nu)$ of $L^p(\nu)$, where $1/p + 1/q = 1$, using that $L^q(\nu) \subset L^p(\nu)$ as $p \leq q$.

In this paper, we aim at quantifying the quality of the prediction of $B^\star\Phi(\vartheta^\star)$ by $\hat{B}\Phi(\hat{\vartheta})$ for $\hat{B}$ and $\hat{\vartheta}$ given by (3), by providing an upper bound with high probability of the squared prediction error:

$$\hat{R}_T^2 = \frac{1}{\nu(\mathcal{Z})} \left\| B^\star\Phi(\vartheta^\star) - \hat{B}\Phi(\hat{\vartheta}) \right\|^2_{L_T}. \tag{4}$$

Example 1.4. Let us set as an example $\mathcal{Z} = \{1, \cdots, n\}$ and ν the counting measure $\sum_{i=1}^n \delta_i$. Assume the n observations belong to the Hilbert space $L^2(\lambda)$ for some measure λ (either discrete or continuous) on the Borel sigma field of $\mathbb{R}$. In this case, the squared prediction error becomes:

$$\hat{R}_T^2 = \frac{1}{n} \sum_{i=1}^n \left\| B^\star(i)\Phi(\vartheta^\star) - \hat{B}(i)\Phi(\hat{\vartheta}) \right\|^2_{L^2(\lambda)}.$$

1.2. **Previous work.**

Reconstructing from observations (that are discrete or continuous-time processes) signals that are linear combinations of features belonging to a continuous dictionary $(\varphi(\theta), \theta \in \Theta)$ has applications in many fields such as spectrocopy ([15]), microscopy ([23]), super-resolution ([18]) or spike deconvolution ([26]).

Most often, the Hilbert space H_T, to which the observations belong, is assumed to be of finite dimension and the dictionary of features is assumed finite of size K. Over the past two decades, the problem of retrieving a sparse vector in the framework of high dimensional regression models ($K \gg \dim(H_T)$) has generated a large number of works ([37], [7], [13], [16], [12] and references therein). The celebrated Lasso estimator, popularized by [37] and defined by an optimization problem composed of a data fidelity term and a ℓ_1 penalty, has been extensively studied and has proven to be efficient. In addition, its convex formulation makes its resolution easy to handle (see [5] for a resolution via fast iterative shrinkage-thresholding algorithms). Prediction error bounds and estimation bounds with respect to the ℓ_2 norm have been established for the Lasso under coherence assumptions on the finite dictionary. We refer to [38] for an overview of the coherence assumptions. It turns out that these rates have been proven minimax optimal in [35]. This means that one cannot find any estimator that achieves faster rates in expected value.

The prediction error bounds obtained for sparse high-dimensional linear models encompass the finite dictionary setting. We consider in this paper continuous dictionaries. As a consequence, the problem of recontruction is highly non-linear. It might be tempting to address this issue by discretizing the parameter space Θ and getting back to a finite dictionary. However, recent papers have advocated that taking a finite subfamily of a continuous dictionary and using a Lasso estimator to retrieve the linear coefficients of the mixture lead to some issues. In particular, the number of active features in the mixture tends to be overestimated, see [27].

A line of work has emerged around the reconstruction of signals that are mixtures of continuously parametrized features by solving a regularized minimization problem over a space of measures. Indeed, one can readily notice that a mixture of non-linear features $\sum_{k\in S^\star} \beta_k^\star \phi(\theta_k^\star)$ can be written as the application of the linear functional $\mu \mapsto \int \phi(\theta)\mu(\mathrm{d}\theta)$ to the atomic measure $\mu^\star = \sum_{k\in S^\star} \beta_k^\star \delta_{\theta_k^\star}$, where δ_x denotes a Dirac measure located in x. The Beurling Lasso (or BLasso) introduced in [22] has proven to be efficient to retrieve a sparse measure from its images through linear functionals. We stress that when $\dim(H_T) < +\infty$, there exists a solution to the BLasso made up of at most $\dim(H_T)$ Dirac measures. We refer to [9] and [25] for proofs of this result. For this reason, the BLasso has been used as a counterpart of the classical Lasso for continuous dictionaries. We remark that when H_T is infinite dimensional the BLasso may not have a priori an atomic solution. It makes its solutions difficult to interpret in our context. That is why we prefer in this paper to assume a bound K on the unknown number of features s in order to formulate (2). When only one element of H_T is observed (*i.e.* $\mathcal{Z}$ is reduced to a singleton and ν is a Dirac measure), this formulation is equivalent to that of the BLasso restricted to the set of atomic measures of at most K atoms. Efficient numerical methods to solve this problem are available such as modifications of the Frank-Wolfe algorithm ([23], [8]) or the Conic Gradient Particle Descent ([21]). We stress that these methods proceed by seeking a solution that is atomic.

It has been shown that under the assumption of the existence of certificate functions, the BLasso retrieves the exact number of features in a small noise regime ([18] for a specific dictionary and [26] in a more general framework). Regarding prediction error bounds, the research has first focused on mixtures of features issued from a dictionary of complex exponentials parametrized by their frequencies. Much progress has been done in super-resolution using the BLasso with this specific dictionary, see [18], [17] in this direction. In [10], the authors showed that the prediction error of the BLasso estimator in this specific case almost reached that of the Lasso estimator provided the frequencies are well separated. They adapted previous results from [6] and [36] for atomic norm denoising and they extended them to a more general case where the noise level is unknown and needs to be estimated. The authors of the present paper considered in [14] the model (2) when only one signal is considered ($\mathcal{Z}$ is a singleton and ν is a Dirac measure) and showed that when the one-dimensional parameters of the features are well separated, one can build estimators that lead to a nearly optimal prediction error bound. By nearly optimal, we mean that the prediction error bound obtained in [14] is of the same order (up to a logarithmic factor) as the minimax bounds obtained in the finite dictionary setting where only linear coefficients are to be retrieved. The result covers a large variety of dictionaries and noises. Let us specify that the separation is expressed with respect to a Riemannian metric following the insightful work of [34].

1.3. **Contributions.** We extend the work of [14] to encompass the case of multiple mixture models. Indeed, we let ν be any finite positive non-zero measure. In the framework of multiple high dimensional linear regressions (ℓ_1, ℓ_p)-norm penalties have been used to retrieve sparsity patterns among the signals. These penalties influence globally the estimations of the signals $(B(i)\Phi(\vartheta^\star),\ i \in \mathcal{Z})$. Let us mention the (ℓ_1, ℓ_2) mixed norm, used to define the Group-Lasso estimator introduced in [39] and that has received significant attention since then (see, [32], [3], [20],[29]). It was shown in [31] that the reconstruction of signals via the Group-Lasso estimator outperfom the reconstruction using the Lasso estimator when the signals share some sparsity pattern. Let us mention the work of [30] that provides consistency results and prediction error convergence rates for the general case (ℓ_1, ℓ_p) with $p \in [1, +\infty]$. Estimators obtained from regularized problems via mixed norms have been studied in the context of high dimensional multiple linear regression models but little has been done for the non-linear extension considered in (2). It is therefore natural to find counterpart estimators for the setting of continuous dictionaries. Let us highlight the work of [28] in which an extension of the BLasso has been proposed in order to address multiple mixture models. The authors extended the result of [26] to show exact support recovery results in the small noise regime. They used a penalty that is a convex combination of mixed norms on measures. We remark that when applied to atomic measures these norms reduce to the (ℓ_1, ℓ_1) and (ℓ_1, ℓ_2) norms on the weights of the Dirac measures.

In this paper, we prove a high-probability upper bound on the prediction error for estimators issued from an optimization problem regularized by a mixed norm $(\ell_1, L^p(\nu))$ with $p \in [1, 2]$ for a wide variety of dictionaries in the general framework where ν can be any finite positive measure. We give refinements of this result when the noise is assumed Gaussian and when the measure ν is discrete. These refined bounds on the prediction error use tail bounds on suprema of Gaussian and χ^2 processes. Our results rely on the existence of certificate functions, see Section 4. We also give sufficient conditions for their construction.

1.4. **Organization of the paper.** In Section 2, we formulate assumptions on the model and set some definitions. Section 3 presents the main results of this paper. We start by giving a high probability upper bound on the prediction error in the general case where the measure ν can be any finite measure. Then, we give refinements of this result when the measure ν is a finite weighted sum of Dirac measures and the noise process is assumed Gaussian. In Section 4, we present the assumptions on certificate functions that are used to state the high probability upper bound on the prediction error in Section 4.1. We give in Section 4.2 sufficient conditions to construct such functions. Section 5 is dedicated to the proof of the high probability upper bound on the prediction error in the most general framework and Sections 6-7 give proofs for refinements of this result when ν is a finite sum of weighted Dirac measures and the noise is Gaussian. Section 8 is dedicated to the proofs of the results stated in Section 4.2 on the existence of certificate functions.

1.5. **Notation.** We shall use for convenience the notation $\lesssim$ and write for two real quantities a and b, $a \lesssim b$ if there exists a positive finite constant C independent of the parameters s, K, T and the measure ν such that $a \leq C\, b$.

We also write for two quantities a, b that $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$.

2. Assumptions on the model

In this section, we briefly set some definitions and assumptions that are presented and discussed in more detail in [14, Sections 3, 4, 5].

2.1. **Regularity and non-degeneracy assumptions on the features.** Let be a fixed parameter $T \in \mathbb{N}$. The features $(\varphi_T(\theta), \theta \in \Theta)$ that form a continuous dictionary are elements of the Hilbert space $(H_T, \langle\cdot,\cdot\rangle_T)$. We shall integrate and differentiate those features with respect to their one-dimensional parameter belonging to the interval Θ of $\mathbb{R}$. To do so, we shall use the notions of Bochner integral and Fréchet derivative. We refer to [14, Section 3.1] for a short presentation of these objects. We recall that for any function $f : \Theta \mapsto H_T$ differentiable at $\theta \in \Theta$, we have for all $g \in H_T$ that:

$$\partial_\theta \langle f(\theta), g\rangle_T = \langle \partial_\theta f(\theta), g\rangle_T \,. \tag{5}$$

In addition, if f is Bochner integrable on Θ, then for all $g \in H_T$, we have that:

$$\int_\Theta \langle f(\theta), g\rangle_T \,\mathrm{d}\theta = \langle \int_\Theta f(\theta)\,\mathrm{d}\theta, g\rangle_T. \tag{6}$$

We shall require the features to satisfy the following regularity assumption.

Assumption 2.1 (Smoothness of φ_T)**.** *We assume that the function $\varphi_T : \Theta \to H_T$ is of class $\mathcal{C}^3$ and $\|\varphi_T(\theta)\|_T > 0$ on Θ.*

Assume that Assumption 2.1 holds. Recall that $\phi_T(\theta) = \varphi_T(\theta)/\|\varphi_T(\theta)\|_T$ for all $\theta \in \Theta$. We define the continuous function:

$$g_T(\theta) = \|\partial_\theta \phi_T(\theta)\|_T^2. \tag{7}$$

It will be convenient to assume the non-degeneracy of the function g_T.

Assumption 2.2 (Positivity of g_T)**.** *Assumption 2.1 holds and we have $g_T > 0$ on Θ.*

One can easily show that features are non-degenerate by checking that for any $\theta \in \Theta$ the elements $\varphi_T(\theta)$ and $\partial_\theta \varphi_T(\theta)$ of H_T are linearly independent, see [14, Lemma 3.1] in this direction.

2.2. **The kernel and its Riemannian derivatives.** In this section, we introduce a function on Θ^2, called kernel, that will quantify the correlation between two features in the dictionary. We shall derive from this kernel a Riemannian metric on the parameter space Θ following [34]. This metric will be in particular invariant to a reparametrization of the parameter space.

2.2.1. *Kernel space and associated Riemannian metric.* We shall set a few bases on the notion of kernel and refer to [14] for further details.

We call kernel a real-valued function defined on Θ^2. Let $\mathcal{K}$ be a symmetric kernel of class $\mathcal{C}^2$ such that the function $g_\mathcal{K}$ defined on the one-dimensional and connected set Θ by:

$$g_\mathcal{K}(\theta) = \partial^2_{x,y}\mathcal{K}(\theta,\theta) \tag{8}$$

is positive and locally bounded, where ∂_x (resp. ∂_y) denotes the usual derivative with respect to the first (resp. second) variable.

We derive from the kernel $\mathcal{K}$ the metric $\mathfrak{d}_\mathcal{K}(\theta,\theta')$ between $\theta, \theta' \in \Theta$ by:

$$\mathfrak{d}_\mathcal{K}(\theta,\theta') = |G_\mathcal{K}(\theta) - G_\mathcal{K}(\theta')|, \tag{9}$$

where $G_\mathcal{K}$ is a primitive of $\sqrt{g_\mathcal{K}}$. We refer to [14, Remark 4.1] for details on the connection with Riemannian metrics.

We shall need to differentiate the kernel $\mathcal{K}$ on the manifold $(\Theta, g_\mathcal{K})$. We shall use the covariant derivatives that generalize the classical directional derivative of vector fields on a manifold. Since we only consider the case of a one-dimensional parameter space, the covariant derivatives reduce to simple expressions.

For a real-valued function F defined on Θ^2, we say that F is of class $\mathcal{C}^{0,0}$ on Θ^2 if it is continuous on Θ^2, and of class $\mathcal{C}^{i,j}$ on Θ^2, with $i, j \in \mathbb{N}$, as soon as: F is of class $\mathcal{C}^{0,0}$, and if $i \geq 1$ then the function $\theta \mapsto F(\theta,\theta')$ is of class $\mathcal{C}^i$ on Θ and its derivative $\partial_x F$ is of class $\mathcal{C}^{i-1,j}$ on Θ^2, and if $j \geq 1$ the function $\theta' \mapsto F(\theta,\theta')$ is of class $\mathcal{C}^j$ on Θ and its derivative $\partial_y F$ is of class $\mathcal{C}^{i,j-1}$ on Θ^2. For a real-valued symmetric function F

defined on Θ^2 of class $\mathcal{C}^{i,j}$, we define the covariant derivatives $D_{i,j;\mathcal{K}}[F]$ of order $(i,j) \in \mathbb{N}^2$ recursively by $D_{0,0;\mathcal{K}}[F] = F$ and for $i, j \in \mathbb{N}$, assuming that $g_\mathcal{K}$ is of class $\mathcal{C}^{\max(i,j)}$, and $\theta, \theta' \in \Theta$:

$$D_{i+1,j;\mathcal{K}}[F](\theta,\theta') = g_\mathcal{K}(\theta)^{\frac{i}{2}} \partial_\theta \left(\frac{D_{i,j;\mathcal{K}}[F](\theta,\theta')}{g_\mathcal{K}(\theta)^{\frac{i}{2}}} \right) \quad \text{and} \quad D_{i,j;\mathcal{K}}[F](\theta,\theta') = D_{j,i;\mathcal{K}}[F](\theta',\theta). \tag{10}$$

In particular, we have $D_{0,0;\mathcal{K}}[F] = F$, $D_{1,0;\mathcal{K}} = \partial_x F$, $D_{0,1;\mathcal{K}} = \partial_y F$ and $D_{1,1;\mathcal{K}} = \partial^2_{xy} F$. We shall also consider the following modification of the covariant derivative, for $i, j \in \mathbb{N}$:

$$\tilde{D}_{i,j;\mathcal{K}}[F](\theta,\theta') = \frac{D_{i,j;\mathcal{K}}[F](\theta,\theta')}{g_\mathcal{K}(\theta)^{i/2}\, g_\mathcal{K}(\theta')^{j/2}}. \tag{11}$$

We have $\tilde{D}_{1,0;\mathcal{K}} \circ \tilde{D}_{0,1;\mathcal{K}} = \tilde{D}_{0,1;\mathcal{K}} \circ \tilde{D}_{1,0;\mathcal{K}}$ and for $i, j \in \mathbb{N}$, assuming that $g_\mathcal{K}$ is of class $\mathcal{C}^{\max(i,j)}$:

$$\tilde{D}_{i,j;\mathcal{K}} = \left(\tilde{D}_{1,0;\mathcal{K}}\right)^i \circ \left(\tilde{D}_{0,1;\mathcal{K}}\right)^j .$$

The definitions of covariant derivatives and their modifications cover the case of 1-dimensional functions defined on Θ. For any smooth function f defined on Θ, we shall note $D_{i;\mathcal{K}}[f]$ (resp. $\tilde{D}_{i;\mathcal{K}}[f]$) for $D_{i,0;\mathcal{K}}[F]$ (resp. $\tilde{D}_{i,0;\mathcal{K}}[F]$) where $F : (\theta,\theta') \mapsto f(\theta)$.

For $i, j \in \mathbb{N}$, if $\mathcal{K}$ is of class $\mathcal{C}^{i\vee 1, j\vee 1}$, then we consider the real-valued function defined on Θ^2 by:

$$\mathcal{K}^{[i,j]} = \tilde{D}_{i,j;\mathcal{K}}[\mathcal{K}]. \tag{12}$$

In particular, when $\mathcal{K}$ is of class $\mathcal{C}^2$, we have:

$$\mathcal{K}^{[0,0]} = \mathcal{K} \quad \text{and} \quad \mathcal{K}^{[1,1]}(\theta,\theta) = 1. \tag{13}$$

2.2.2. *The kernel associated to the dictionary of features.* Let $T \in \mathbb{N}$ be fixed and assume that Assumption 2.2 holds. We associate to the dictionary of features $(\varphi_T(\theta), \theta \in \Theta)$ a kernel $\mathcal{K}_T$ on Θ^2 defined by:

$$\mathcal{K}_T(\theta,\theta') = \langle \phi_T(\theta), \phi_T(\theta')\rangle_T = \frac{\langle \varphi_T(\theta), \varphi_T(\theta')\rangle_T}{\|\varphi_T(\theta)\|_T \|\varphi_T(\theta')\|_T}. \tag{14}$$

In the following, for an expression A we will often replace the notation $A_{\mathcal{K}_*}$ by A_* where $*$ is T or ∞.

We remark that under Assumptions 2.1 and 2.2 the definitions (7) and (8) are consistent by Lemma [14, Lemma 4.3]. Furthermore, we have that the kernel $\mathcal{K}_T$ is of class $\mathcal{C}^{3,3}$ on Θ^2 and for $i, j \in \{0, \dots, 3\}$ and for any $\theta, \theta' \in \Theta$:

$$\mathcal{K}_T^{[i,j]}(\theta,\theta') = \langle \tilde{D}_{i;T}[\phi_T](\theta), \tilde{D}_{j;T}[\phi_T](\theta')\rangle_T, \tag{15}$$

$$\sup_{\Theta^2} |\mathcal{K}_T^{[0,0]}| \leq 1, \quad \mathcal{K}_T^{[0,0]}(\theta,\theta) = 1, \quad \mathcal{K}_T^{[1,0]}(\theta,\theta) = 0, \quad \mathcal{K}_T^{[2,0]}(\theta,\theta) = -1 \quad \text{and} \quad \mathcal{K}_T^{[2,1]}(\theta,\theta) = 0. \tag{16}$$

In practice, the kernel $\mathcal{K}_T$ may be difficult to handle. It might be convenient to approximate $\mathcal{K}_T$ by a kernel $\mathcal{K}_\infty$ for which some assumptions will be easier to check, see [14, Section 8] in this direction. We shall give some properties that an approximating kernel $\mathcal{K}_\infty$ must verify. Then we shall define a quantity measuring the precision of the approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$ over some compact set $\Theta_T \subseteq \Theta$.

Let us first define for a kernel $\mathcal{K}$ of class $\mathcal{C}^{3,3}$ the function on Θ:

$$h_\mathcal{K}(\theta) = \mathcal{K}^{[3,3]}(\theta,\theta). \tag{17}$$

The following assumption gathers the properties that an approximating kernel $\mathcal{K}_\infty$ must sastify.

Assumption 2.3 (Properties of the asymptotic kernel $\mathcal{K}_\infty$)**.** *The symmetric kernel $\mathcal{K}_\infty$ defined on Θ^2 is of class $\mathcal{C}^{3,3}$, the function g_∞ defined by (8) on Θ is positive and locally bounded (as well as of class $\mathcal{C}^2$), and we have $\mathcal{K}_\infty(\theta,\theta) = -\mathcal{K}_\infty^{[2,0]}(\theta,\theta) = 1$ for $\theta \in \Theta$. The set $\Theta_\infty \subseteq \Theta$ is an interval and we have:*

$$m_g := \inf_{\Theta_\infty} g_\infty > 0, \quad L_3 := \sup_{\Theta_\infty} h_\infty < +\infty, \quad \textit{and} \quad L_{i,j} := \sup_{\Theta_\infty^2} |\mathcal{K}_\infty^{[i,j]}| < +\infty \quad \textit{for all } i,j \in \{0,1,2\}. \tag{18}$$

We stress that the interval Θ_∞ is possibly unbounded contrary to the set Θ_T which is compact.

Under assumption 2.3, we derive from the kernel $\mathcal{K}_\infty$ the Riemannian metric $\mathfrak{d}_\infty$ as in (9). One can show that the metrics $\mathfrak{d}_T$ and $\mathfrak{d}_\infty$ are strongly equivalent on the compact set Θ_T^2. Indeed, we have:

$$\frac{1}{\rho_T}\,\mathfrak{d}_\infty \leq \mathfrak{d}_T \leq \rho_T\,\mathfrak{d}_\infty, \tag{19}$$

where ρ_T is a finite positive constant defined by:

$$\rho_T = \max\left(\sup_{\Theta_T}\sqrt{\frac{g_T}{g_\infty}}, \sup_{\Theta_T}\sqrt{\frac{g_\infty}{g_T}}\right). \tag{20}$$

We then give an assumption on the quality of approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$. We set:

$$\mathcal{V}_T = \max(\mathcal{V}_T^{(1)}, \mathcal{V}_T^{(2)}) \quad \text{with} \quad \mathcal{V}_T^{(1)} = \max_{i,j\in\{0,1,2\}} \sup_{\Theta_T^2} |\mathcal{K}_T^{[i,j]} - \mathcal{K}_\infty^{[i,j]}| \quad \text{and} \quad \mathcal{V}_T^{(2)} = \sup_{\Theta_T} |h_T - h_\infty|. \tag{21}$$

Assumption 2.4 (Quality of the approximation)**.** *Let $T \in \mathbb{N}$ be fixed. Assumptions 2.2 and 2.3 hold, the interval $\Theta_T \subset \Theta_\infty$ is a compact interval, and we have:*

$$\mathcal{V}_T \leq L_{2,2} \wedge L_3.$$

3. Main Results

3.1. General bound on the prediction error.

The main goal of this paper is to bound the prediction error (4) associated to the estimators defined in (3). We first give a bound that holds with a controled probability in the general case where the penalty of the optimization problem (3) is the norm $\|\cdot\|_{\ell_1,L^p(\nu)}$ with $p \in [1,2]$. The bound will be expressed as a function of the tuning parameter κ, the sparsity s, the mass of the measure ν and the parameter of the penalty p. It will stand on an event whose probability is bounded from below by tails of distributions of random variables defined by taking the supremum over the compact set Θ_T and the norm $\|\cdot\|_{L^q(\nu)}$ of real-valued processes indexed on $\mathcal{Z} \times \Theta_T$ of the form:

$$X(z,\theta) = \langle W_T(z), g(\theta)\rangle_T\,,$$

for some smooth functions $g : \Theta_T \to H_T$ related to the dictionary of features and where q is the conjugate of p in the sense that $1/q + 1/p = 1$.

The assumptions on the regularity of the dictionary, the regularity of the limit kernel and the proximity to the limit kernel are the same as those from [14, Theorem 2.1]. Regarding the noise, we only require that it belongs almost surely to $L^q(\nu, H_T)$. We highlight that the Theorem below is proven under the existence of certificate functions. Those certificates generalize that of [14, Theorem 2.1]. (In particular, they reduce to those in [14] when ν is a Dirac measure.) A construction of certificates has been proposed in [28] for the case where ν is the counting measure. Our construction is slightly different and covers the general case where ν can be any finite positive measure, see Remark 8.4. We shall give in Section 4.2 sufficient conditions for their existence. It turns out that we can construct such certificates provided the elements of the set $\mathcal{Q}^\star$ defined in (1) are pairwise separated with respect to a Riemannian metric. We remark that the separation does not depend on the space $(\mathcal{Z}, \mathcal{F}, \nu)$. In particular, in the example where $\mathcal{Z}$ is a finite set of cardinal n, increasing n does not improve or deteriorate the separation.

We state the main result of this paper that is proved in Section 5.

Theorem 3.1. *Let $T \in \mathbb{N}$. Let be $p \in [1,2]$ and $q \in [2,+\infty]$ such that $1/p + 1/q = 1$. When $p = 1$, we assume that $\mathcal{Z}$ is finite. Assume we observe the random element Y of L_T under the regression model (2) with a noise W_T belonging to $L^q(\nu, H_T)$ almost surely and unknown parameters $B^\star \in L^2(\nu, \mathbb{R}^K)$ and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_K^\star)$ a vector with entries in Θ_T (compact interval of $\mathbb{R}$). Let us suppose that the following assumptions hold :*

- (i) ***Regularity of the dictionary φ_T:*** *The dictionary function φ_T satisfies the smoothness conditions 2.1 . The function g_T satisfies the positivity condition 2.2.*
- (ii) ***Regularity of the limit kernel:*** *The kernel $\mathcal{K}_\infty$ and the functions g_∞ and h_∞, defined on an interval $\Theta_\infty \subset \Theta$, satisfy the smoothness conditions of Assumption 2.3.*
- (iii) ***Proximity to the limit kernel:*** *The kernel $\mathcal{K}_T$ defined from the dictionary is sufficiently close to the limit kernel $\mathcal{K}_\infty$ in the sense that Assumption 2.4 holds.*
- (iv) ***Existence of certificates:*** *The non-empty set of unknown parameters $\mathcal{Q}^\star = \{\theta_k^\star, k \in S^\star\}$, with $S^\star = \{k, \|B_k^\star\|_{L^2(\nu)} \neq 0 \}$, satisfies Assumptions 4.1 and 4.2 with the same $r > 0$.*

Then, there exist finite positive constants $\mathcal{C}$, $\mathcal{C}_0$ depending on r and on the kernel $\mathcal{K}_\infty$ defined on Θ_∞ such that we have the prediction error bound of the estimators $\hat{B}$ and $\hat{\vartheta}$ defined for a tuning parameter $\kappa > 0$ (in (3)) given by:

$$\frac{1}{\sqrt{\nu(\mathcal{Z})}} \left\| \hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star) \right\|_{L_T} \leq \mathcal{C}_0 \sqrt{s}\, \nu(\mathcal{Z})^{\frac{1}{p}} \kappa, \tag{22}$$

with probability larger than

$$1 - \sum_{i=0}^{2} \mathbb{P}\left(M_i > \mathcal{C}\, \kappa\, \nu(\mathcal{Z})\right), \tag{23}$$

where M_i is defined by:

$$M_i = \sup_{\theta \in \Theta_T} \left\| \left\langle W_T, \phi_T^{[i]}(\theta) \right\rangle_T \right\|_{L^q(\nu)}, \quad \textit{for } i = 0, 1, 2. \tag{24}$$

Remark 3.2 (On the choice of κ). We typically choose κ in (22) as small as possible giving a global bound on the prediction risk small, such that the event on which the bound stands occurs with a sufficiently large probability.

Remark 3.3 (On the dimension K). The bound K on the sparsity s does not appear neither in the upper bound on the prediction error (22) nor in the lower bound on the probability (23). Thus, it can be taken arbitrarily large. This was already the case in [14] where $\mathcal{Z}$ is a singleton and ν is a Dirac measure, see Remark 2.4 therein.

3.2. **Explicit bounds for Gaussian noise and finite number of signals.** It is not straightforward to establish tail bounds for the random variables M_i defined in Theorem 3.1. However, if the noise process for fixed z in $\mathcal{Z}$ is centered Gaussian, for the cases $p = q = 2$ and $p = 1$ together with $q = +\infty$, this can be done using Rice formulae (see [2] for a complete overview of Rice formulae). We shall then deduce bounds for arbitrary values of conjugate pairs (p,q) in $[1,2] \times [2,\infty]$ using interpolation inequalities.

We will give an explicit lower bound for the probability (23). The lower bound will depend on the parameter T and the number of signals $n = \mathrm{Card}(\mathcal{Z})$ assumed to be finite here. Thus, we will be able to give a convergence rate towards zero for the prediction error with respect to these parameters.

In order to use tail bounds for the random variables M_j from Theorem 3.1, we state additional assumptions on the noise W_T. As in [14], we make the following assumption on the noise process W_T, where the decay rate $\Delta_T > 0$ controls the noise variance decay as the parameter T grows and $\sigma > 0$ is the intrinsic noise level.

Assumption 3.1 (Admissible noise)**.** *Let $T \in \mathbb{N}$. Assume that the set $\mathcal{Z}$ is finite. The processes $(W_T(z),\, z \in \mathcal{Z})$ are independent copies of a noise process w_T. The noise process w_T belongs to H_T almost surely and, there exist a noise level $\sigma > 0$ and a decay rate $\Delta_T > 0$ such that for all $f \in H_T$ the random variable $\langle f, w_T\rangle_T$ is a centered Gaussian random variable satisfying:*

$$\operatorname{Var}\left(\langle f, w_T\rangle_T\right) \leq \sigma^2\, \Delta_T\, \|f\|_T^2. \tag{25}$$

3.2.1. *The case $p = 2$ and $\mathcal{Z}$ finite.* We state a corollary of Theorem 3.1 for the specific case where ν is an atomic measure composed of n atoms and the penalty of the optimization problem (3) is a mixed $(\ell_1, L^2(\nu))$ norm. The proof is given in Section 6.

We denote by $|\Theta_T|_{\mathfrak{d}_T}$ the diameter of the interval Θ_T with respect to the Riemannian metric $\mathfrak{d}_T$ associated to the kernel $\mathcal{K}_T$ and defined in (9).

Corollary 3.4. *Let $T \in \mathbb{N}$. We fix $p = q = 2$. We assume that* $\operatorname{Card}(\mathcal{Z}) = n < +\infty$ *and that the measure ν is $\nu = \sum_{z\in\mathcal{Z}} a_z\delta_z$ where δ_z denotes a Dirac measure located in $z \in \mathcal{Z}$ and $(a_z, z \in \mathcal{Z})$ are non-negative real numbers. Assume we observe the random element Y of L_T under the regression model (2) with unknown parameters $B^\star$ in $L^2(\nu, \mathbb{R}^K)$ (which can be identified with $\mathbb{R}^{n\times K}$) and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_K^\star)$ a vector with entries in Θ_T, a compact interval of $\mathbb{R}$, such that Points (i)-(iv) of Theorem 3.1 are satisfied and the noise process W_T satisfies Assumption 3.1 for a noise level $\sigma > 0$ and a decay rate for the noise variance $\Delta_T > 0$.*

Then, there exist finite positive constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$, depending on the kernel $\mathcal{K}_\infty$ defined on Θ_∞ and on r such that for any $\tau > 1$ and a tuning parameter:

$$\kappa \geq \mathcal{C}_1\sigma\sqrt{\frac{\|a\|_{\ell_\infty}\Delta_T\, n}{\nu(\mathcal{Z})^2}}\left(1 + \sqrt{1 + \frac{\log(\tau)}{n}}\right),$$

where $\|a\|_{\ell_\infty} = \max_{z\in\mathcal{Z}} |a_z|$, we have the following prediction error bound of the estimators $\hat{B}$ and $\hat{\vartheta}$ defined in (3):

$$\frac{1}{\sqrt{\nu(\mathcal{Z})}}\left\|\hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star)\right\|_{L_T} \leq \mathcal{C}_0\,\sqrt{s\,\nu(\mathcal{Z})}\,\kappa, \tag{26}$$

with probability larger than $1 - \mathcal{C}_2\left(\frac{1}{\tau} + \frac{|\Theta_T|_{\mathfrak{d}_T}F(n)}{\sqrt{\tau}}\right)$ with a sequence $F(n) \asymp \sqrt{n}\,\mathrm{e}^{-n/2}$.

Remark 3.5 (Comparison to the Group-Lasso estimator)*.* Assume that the Hilbert space $H_T = \mathbb{R}^T$ is endowed with the Euclidean scalar product and Euclidean norm $\|\cdot\|_{\ell_2}$. Let $\mathcal{Z} = \{1, \cdots, n\}$ and let ν be the counting measure on $\mathcal{Z}$, *i.e.* $\nu = \sum_{k=1}^n \delta_k$. Notice that in this setting $L_T = L^2(\nu, H_T)$ is of finite dimension and can be identified with $\mathbb{R}^{n\times T}$. Assume that the observation $Y \in L_T$ comes from the model (2) where for any $i \in \{1, \cdots, n\}$, $W_T(i)$ is a Gaussian vector in $\mathbb{R}^T$ with independent entries of variance σ^2. Assume also that the Gaussian vectors $(W_T(i),\, 1 \leq i \leq n)$ are independent. Thus, Assumption 3.1 holds with an equality in (25) and

$$\Delta_T = 1.$$

We first consider that the parameters $\vartheta^\star$ are known. In this case, the model becomes the classical high-dimensional multiple linear regression model and the Group-Lasso estimator $\hat{B}_L$ can be used to estimate $B^\star$ under coherence assumptions on the finite dictionary made of the rows of the matrix $\Phi^\star = \Phi_T(\vartheta^\star) \in \mathbb{R}^{K\times T}$ (see [7]). The authors of [31] showed that the prediction error associated to the Group-Lasso estimator satisfies the bound:

$$\frac{1}{n\,T}\sum_{i=1}^n \|(\hat{B}_L(i) - B^\star(i))\Phi^\star\|_{\ell_2}^2 \lesssim \frac{\sigma^2\, s}{T}\left(1 + \frac{\log(K)}{n}\right), \tag{27}$$

with high probability, larger than $1 - 1/K^\gamma$ for some positive constant $\gamma > 0$. Furthermore, in the case where $B^\star$ is an unknown s-sparse application, $\vartheta^\star$ is known and $\Phi^\star$ verifies a coherence property, then lower bounds of order $\sigma^2\, s(1+\log(K/s)/n)/T$ in expected value can be established. The non-asymptotic prediction lower bounds for the prediction error given in [31] are for $2s < K$:

$$\inf_{\hat{B}} \sup_{B^\star\ s-\text{sparse}} \mathbb{E}\left[\frac{1}{n\,T}\sum_{i=1}^{n} \|(\hat{B}(i)-B^\star(i))\Phi^\star\|_{\ell_2}^2\right] \geq C\cdot \frac{\sigma^2\, s}{T}\left(1+\frac{\log(K/s)}{n}\right),$$

where the infinimum is taken over all the estimators $\hat{B}$ (measurable functions of the obervation Y taking their values in $L^2(\nu,\mathbb{R}^K)$) and for some constant $C>0$ free of s, K, n and T.

When the linear coefficients $B^\star$ and the parameters $\vartheta^\star$ are unknown, Corollary 3.4 gives an upper bound for the prediction risk which is similar to that of the linear case. Consider the estimators from (3) with $p=2$. Assume that the Riemannian diameter of the set Θ_T is bounded by a constant free of T. By squaring (26) and then dividing it by T, we obtain from Corollary 3.4 with:

$$\kappa = \mathcal{C}_1\sigma\sqrt{\frac{1}{n}}\left(1+\sqrt{1+\frac{\log(\tau)}{n}}\right) \qquad \text{and } \tau = T^\gamma \quad \text{for some given } \gamma>0,$$

that with high probability, larger than $1-C'/T^\gamma - C''F(n)/T^{\gamma/2}$:

$$\frac{1}{n\,T}\sum_{i=1}^{n}\left\|\hat{B}(i)\Phi_T(\hat{\vartheta}) - B^\star(i)\Phi_T(\vartheta^\star)\right\|_{\ell_2}^2 \lesssim \frac{\sigma^2\, s}{T}\left(1+\frac{\log(T)}{n}\right). \tag{28}$$

We identify two regimes depending on the ratio $\log(T)/n$. Indeed, when $\log(T)/n \gg 1$ the bound (28) behaves as $\frac{\sigma^2\, s\log(T)}{n\,T}$ and stands with probability that converges towards 1 at the rate $F(n)/T^{\gamma/2}$. On the contrary, when $\log(T)/n \ll 1$ the bound (28) is of order $\frac{\sigma^2\, s}{T}$ and stands with probality that converges towards 1 at the rate $1/T^\gamma$.

3.2.2. *The case $p=1$ and $\mathcal{Z}$ finite.* We apply Theorem 3.1 to the particular case $p=1$. It turns out that for $q=+\infty$, tail bounds for the random variables M_j with $j=0,1,2$ can be established from Rice formulae for smooth Gaussian processes. The following Corollary is proved in Section 7.

Corollary 3.6. *Let $T\in\mathbb{N}$. We fix $p=1, q=+\infty$. We assume that $\mathrm{Card}(\mathcal{Z}) = n < +\infty$ and that the measure ν is $\nu=\sum_{z\in\mathcal{Z}} a_z\delta_z$ where δ_z denotes a Dirac measure located in $z\in\mathcal{Z}$ and $(a_z, z\in\mathcal{Z})$ are non-negative real numbers. Assume we observe the random element Y of L_T under the regression model (2) with unknown parameters $B^\star$ in $L^2(\nu,\mathbb{R}^K)$ (which can be identified with $\mathbb{R}^{n\times K}$) and $\vartheta^\star = (\theta_1^\star,\cdots,\theta_K^\star)$ a vector with entries in Θ_T, a compact interval of $\mathbb{R}$, such that Points (i)-(iv) of Theorem 3.1 are satisfied and the noise process W_T satisfies Assumption 3.1 for a noise level $\sigma>0$ and a decay rate for the noise variance $\Delta_T>0$.*

Then, there exist finite positive constants $\mathcal{C}_0$, $\mathcal{C}_3$, $\mathcal{C}_4$, depending on the kernel $\mathcal{K}_\infty$ defined on Θ_∞ and on r such that for any $\tau>1$ and a tuning parameter:

$$\kappa \geq \mathcal{C}_3\sigma\sqrt{\Delta_T\log(\tau)}/\nu(\mathcal{Z}),$$

we have the following prediction error bound of the estimators $\hat{B}$ and $\hat{\vartheta}$ defined in (3):

$$\frac{1}{\sqrt{\nu(\mathcal{Z})}}\left\|\hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star)\right\|_{L_T} \leq \mathcal{C}_0\sqrt{s}\,\nu(\mathcal{Z})\,\kappa, \tag{29}$$

with probability larger than $1-\mathcal{C}_4\, n\left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log\tau}}\vee\frac{1}{\tau}\right)$.

Remark 3.7. When the measure ν is composed of one atom, that is $n=1$. This result covers that of [14, Theorem 2.1].

Remark 3.8 (Comparison to other estimators). Let us set $H_T = \mathbb{R}^T$, $\mathcal{Z} = \{1, \cdots, n\}$, ν the counting measure and W_T as in Remark 3.5 and assume that the Riemannian diameter of the set Θ_T is bounded by a constant free of T. We recall that in this case $\Delta_T = 1$. By considering the estimators built from the optimization problem (3) with $p = 1$ and applying Corollary 3.6, we get with:

$$\kappa = \mathcal{C}_3 \sigma \sqrt{\Delta_T \log \tau} / n \quad \text{and } \tau = T^{\gamma/2} \quad \text{for some given} \quad \gamma > 1,$$

that, with probability, larger than $1 - C\, n/T^{\gamma/2}$:

$$\frac{1}{n\,T} \sum_{i=1}^{n} \left\| \hat{B}(i)\Phi_T(\hat{\vartheta}) - B^\star(i)\Phi_T(\vartheta^\star) \right\|_{\ell_2}^2 \lesssim \frac{\sigma^2\, s \log(T)}{T} \cdot \tag{30}$$

We note that this simultaneous estimation procedure gives the same result as estimating separately n signals as in [14] under the assumption that each signal has sparsity s. Individual estimation can be better for those signals with smaller sparsity than the global one we use here.

In Remark 3.5, we showed that by taking $p = 2$ in the optimization problem (3) defining the estimators $\hat{B}$ and $\hat{\vartheta}$, we obtain the bound (28) for a well chosen tuning parameter κ. When n and T are sufficiently large, we remark that the bound (30) is larger than the bound (28) established for the estimators from Corollary 3.4 and stands with a smaller probability.

3.2.3. *Arbitrary value of p in [1,2].* For the cases $p = 2$ and $p = 1$ we established tail bounds for the random variables M_i for $i = 0, 1, 2$ in Corollaries 3.4 and 3.6, respectively. We recall that these random variables are obtained by taking the supremum over the set Θ_T and the $L^q(\nu)$ norm of real-valued processes indexed on $\mathcal{Z} \times \Theta_T$. For the case $p = 1, q = +\infty$ we used a Rice formula for suprema of smooth Gaussian processes, see [14, Lemma A.2]. For the case $p = q = 2$ we used a Rice formula for suprema of chi-squared processes; see Lemma A.1. Unfortunately, in the more general case where $p \in [1, 2]$ and $q \in [2, +\infty]$, such formulae seem out of reach. However, we may use the log-convexity of L^q-norms and use the controls we obtained for the cases $p = 1$ and $p = 2$. Indeed for any $f \in L^\infty(\nu)$ and $q \in [2, +\infty]$, we have the inequality:

$$\|f\|_{L^q} \leq \|f\|_{L^2}^{\frac{2}{q}} \|f\|_{L^\infty}^{\frac{q-2}{q}} .$$

Hence, we readily deduce the following inclusion for any bound $M \geq 0$:

$$\{\|f\|_{L^q} > M\} \subset \{\|f\|_{L^\infty} > M\} \cup \{\|f\|_{L^2} > M\}.$$

4. Certificates

We present the certificate functions whose existence is required in Theorem 3.1. Such functions were introduced for exact reconstruction of signals, see [19], [18], [26]. Exact recovery results for the simultaneous reconstruction of signals via the Group-BLasso were proved in [28] using an extension of the certificates from [26]. In [34], sufficient conditions for the existence of certificate functions were proved for a wide variety of dictionaries. The authors showed that certificates can be built provided the parameters of the features to be retrieved are well separated with respect to a Riemannian metric. This result requires some assumptions on the kernel associated to the dictionary. In particular, the kernel must be local concave on its diagonal, strictly inferior to 1 outside the diagonal and smooth. Their construction was used in [14] to establish prediction error bounds under similar assumptions on the dictionary but for a one-dimensional parameter space Θ.

In this paper, we extend the notion of certificates for our context of multiple reconstructions of signals, following the work of [28]. Let us emphasize that we use a different contruction than [28], see Remark 8.4.

4.1. **Assumptions on the certificates.** In this section, we introduce the assumptions on the certificates. We will give later in Section 4.2 an explicit construction and sufficient conditions for these assumptions to hold.

Let $T \in \mathbb{N}$. We denote the closed ball centered at $\theta \in \Theta_T$ with radius r by:

$$\mathcal{B}_T(\theta, r) = \{\theta' \in \Theta_T, : \mathfrak{d}_T(\theta, \theta') \leq r\} \subseteq \Theta_T.$$

Let $r > 0$ and let $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. We call near region of $\mathcal{Q}^\star$ the union of balls $\bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ and far region the set Θ_T minus the near region: $\Theta_T \setminus \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$.

Assumption 4.1 (Interpolating certificate)**.** *Let $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$, let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $r > 0$ and $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. Suppose Assumptions 2.1 and 2.2 on the dictionary $(\varphi_T(\theta),\ \theta \in \Theta)$ and Assumption 2.3 on $\mathcal{K}_\infty$ hold. Suppose that $\mathfrak{d}_T(\theta, \theta') > 2r$ for all $\theta, \theta' \in \mathcal{Q}^\star \subset \Theta_T$. There exist finite positive constants C_N, C'_N, C_F, C_B with $C_F < 1$, depending on r and $\mathcal{K}_\infty$, such that for any measurable application $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\cdot, \theta^\star)\|_{L^q(\nu)} = 1$, there exists an element $P \in L^q(\nu, H_T)$ satisfying:*

- *(i) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $\|\langle \phi_T(\theta), P\rangle_T\|_{L^q(\nu)} \leq 1 - C_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$.*
- *(ii) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $\|\langle \phi_T(\theta), P\rangle_T - V(\cdot, \theta^\star)\|_{L^q(\nu)} \leq C'_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$.*
- *(iii) For all θ in Θ_T, $\theta \notin \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ (far region), we have $\|\langle \phi_T(\theta), P\rangle_T\|_{L^q(\nu)} \leq 1 - C_F$.*
- *(iv) We have $\|P\|_{L_T} \leq C_B \sqrt{s}\, \nu(\mathcal{Z})^{\frac{1}{2p} - \frac{1}{2q}}$.*

We call "interpolating certificate" the real-valued functions defined on $\mathcal{Z} \times \Theta$ by $(z, \theta) \mapsto \langle \phi_T(\theta), P(z)\rangle_T$ where P is an element of $L^q(\nu, H_T)$ satisfying Points *(i)* – *(iv)* from 4.1.

We emphasize the interpolating properties of those certificates by noticing that for any $\theta^\star \in \mathcal{Q}^\star$ we have from Point *(ii)* for ν-almost every $z \in \mathcal{Z}$ that:

$$\langle \phi_T(\theta^\star), P(z)\rangle_T = V(z, \theta^\star).$$

In order to establish prediction error bounds another type of certificate functions having different interpolating properties will be needed, see [17], [36], [10] in this direction.

Assumption 4.2 (Interpolating derivative certificate)**.** *Let $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$, let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $r > 0$ and $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. Suppose Assumption 2.1 and 2.2 on the dictionary $(\varphi_T(\theta),\ \theta \in \Theta)$ and Assumption 2.3 on $\mathcal{K}_\infty$ hold. Suppose that $\mathfrak{d}_T(\theta, \theta') > 2r$ for all $\theta, \theta' \in \mathcal{Q}^\star \subset \Theta_T$. There exist finite positive constants c_N, c_F, c_B depending on r and $\mathcal{K}_\infty$ such that for any measurable application $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\cdot, \theta^\star)\|_{L^q(\nu)} = 1$, there exists an element $Q \in L^q(\nu, H_T)$ satisfying:*

- *(i) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have:*
$$\|\langle \phi_T(\theta), Q\rangle_T - V(\cdot, \theta^\star)\ \mathrm{sign}(\theta - \theta^\star)\, \mathfrak{d}_T(\theta, \theta^\star)\|_{L^q(\nu)} \leq c_N\, \mathfrak{d}_T(\theta^\star, \theta)^2.$$
- *(ii) For all θ in Θ_T and $\theta \notin \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ (far region), we have $\|\langle \phi_T(\theta), Q\rangle_T\|_{L^q(\nu)} \leq c_F$.*
- *(iii) We have $\|Q\|_{L_T} \leq c_B \sqrt{s}\, \nu(\mathcal{Z})^{\frac{1}{2p} - \frac{1}{2q}}$.*

We call "interpolating derivative certificate" the real-valued functions defined on $\mathcal{Z} \times \Theta$ by $(z, \theta) \mapsto \langle \phi_T(\theta), Q(z)\rangle_T$ where Q is an element of $L^q(\nu, H_T)$ satisfying Points *(i)* – *(iii)* from 4.2.

We remark that for any $\theta^\star \in \mathcal{Q}^\star$ we deduce from Point *(i)* for ν-almost every $z \in \mathcal{Z}$:

$$\langle \phi_T(\theta^\star), Q(z)\rangle_T = 0.$$

Let us remark that when ν is a Dirac measure, the norm $\|\cdot\|_{L^q(\nu)}$ reduces to an absolute value and Assumptions 4.1 and 4.2 correspond to Assumptions 6.1 and 6.2 of [14].

In the following, we shall often write by a slight abuse of notation $f(\theta)$ for $f(\cdot,\theta)$ when considering a function f from $\mathcal{Z}\times\Theta$ to $\mathbb{R}$.

4.2. **Construction of the certificates.** We give in this section sufficient conditions for Assumptions 4.1 and 4.2 to hold. These assumptions rely on the existence of real-valued functions defined on $\mathcal{Z}\times\Theta$ called certificates and of the form:

$$(z,\theta)\mapsto\langle\phi_T(\theta),P(z)\rangle_T\,,$$

where P is an element of $L^q(\nu,H_T)$ satisfying some properties.

We shall follow the construction from [34, Theorem 2] for interpolating certificates and generalize the contruction of [17, Lemma 2.7] for interpolating derivative certificates. In [17, Lemma 2.7], the authors consider certificates that are trigonometric polynomials whereas we are interested here in a more general framework. Furthermore, we remark that the constructions aforementioned only cover the case where ν is a Dirac measure whereas ν can be any finite positive measure in our framework.

Once built, we will then show that our certificates satisfy the properties required in Assumptions 4.1 and 4.2. The proofs of the results of this section will closely follow the proofs of [14, Propositions 7.4 and 7.5] that cover the case where ν is a Dirac measure (*i.e.* only one signal is considered).

Similarly to [14], we shall consider bounded kernels locally concave on the diagonal. We shall also require the kernels to be strictly less than 1 outside their diagonal. In order to state these properties clearly, we define for $T\in\bar{\mathbb{N}}=\mathbb{N}\cup\{\infty\}$ and $r>0$:

$$\varepsilon_T(r)=1-\sup\left\{|\mathcal{K}_T(\theta,\theta')|;\quad\theta,\theta'\in\Theta_T\text{ such that }\mathfrak{d}_T(\theta',\theta)\geq r\right\},\tag{31}$$

$$\nu_T(r)=-\sup\left\{\mathcal{K}_T^{[0,2]}(\theta,\theta');\quad\theta,\theta'\in\Theta_T\text{ such that }\mathfrak{d}_T(\theta',\theta)\leq r\right\}.\tag{32}$$

The quantities $\varepsilon_T(r)$ and $\nu_T(r)$ defined from the considered kernel $\mathcal{K}_T$ and the set Θ_T will have to be positive for some $r>0$. The positivity may be difficult to show when $T\in\mathbb{N}$. In order to show the positivity of $\varepsilon_T(r)$ and $\nu_T(r)$, one can rather show the positivity of $\varepsilon_\infty(r)$ and $\nu_\infty(r)$ derived from an approximating kernel easier to handle and use [14, Lemma 7.1].

We define the set $\Theta_{T,\delta}^s\subset\Theta_T^s$ of vector of parameters of dimension $s\in\mathbb{N}^*$ and separation $\delta>0$ as:

$$\Theta_{T,\delta}^s=\Big\{(\theta_1,\cdots,\theta_s)\in\Theta_T^s:\ \mathfrak{d}_T(\theta_\ell,\theta_k)>\delta\text{ for all distinct }k,\ell\in\{1,\ldots,s\}\Big\}.\tag{33}$$

Let us define for $i,j=0,1,2$ (assuming the kernel $\mathcal{K}_T$ is smooth enough) and $\vartheta=(\theta_1,\ldots,\theta_s)\in\Theta_T^s$ the $s\times s$ matrix:

$$\mathcal{K}_T^{[i,j]}(\vartheta)=\Big(\mathcal{K}_T^{[i,j]}(\theta_k,\theta_\ell)\Big)_{1\leq k,\ell\leq s}.\tag{34}$$

Let I be the identity matrix of size $s\times s$.

Using the convention $\inf\emptyset=+\infty$, We define:

$$\delta_T(u,s)=\inf\Big\{\delta>0:\ A_{T,\ell_\infty}(\vartheta)\leq u,\vartheta\in\Theta_{T,\delta}^s\Big\},\tag{35}$$

where:

$$A_{T,\ell_\infty}(\vartheta)=\max\Bigg(\left\|I-\mathcal{K}_T^{[0,0]}(\vartheta)\right\|_{\text{op},\ell_\infty},\left\|I-\mathcal{K}_T^{[1,1]}(\vartheta)\right\|_{\text{op},\ell_\infty},\left\|I+\mathcal{K}_T^{[2,0]}(\vartheta)\right\|_{\text{op},\ell_\infty},\left\|\mathcal{K}_T^{[1,0]}(\vartheta)\right\|_{\text{op},\ell_\infty},$$
$$\left\|\mathcal{K}_T^{[0,1]}(\vartheta)\right\|_{\text{op},\ell_\infty},\left\|\mathcal{K}_T^{[1,2]}(\vartheta)\right\|_{\text{op},\ell_\infty}\Bigg),\tag{36}$$

and $\|\cdot\|_{\text{op},\ell_\infty}$ denotes the operator norm associated to the sup-norm $\|\cdot\|_{\ell_\infty}$, that is for a matrix $A \in \mathbb{R}^{s\times s}$,

$$\|A\|_{\text{op},\ell_\infty} = \sup_{x\in\mathbb{R}^s, \|x\|_{\ell_\infty}\leq 1} \|Ax\|_{\ell_\infty}.$$

We define quantities which depend on $\mathcal{K}_\infty$, Θ_∞ and on real parameters $r > 0$ and $\rho \geq 1$:

$$\begin{aligned} H_\infty^{(1)}(r,\rho) &= \frac{1}{2} \wedge L_{2,0} \wedge L_{2,1} \wedge \frac{\nu_\infty(\rho r)}{10} \wedge \frac{\varepsilon_\infty(r/\rho)}{10}, \\ H_\infty^{(2)}(r,\rho) &= \frac{1}{6} \wedge \frac{8\,\varepsilon_\infty(r/\rho)}{10(5+2L_{1,0})} \wedge \frac{8\,\nu_\infty(\rho r)}{9(2L_{2,0}+2L_{2,1}+4)}, \end{aligned} \tag{37}$$

where the constants $L_{i,j}$ are defined in (18).

We give sufficient conditions for Assumption 4.1 to hold. The proof of the following result is given in Section 8.1.

Proposition 4.1 (Interpolating certificate). *Let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $\rho \geq 1$, $r > 0$ and $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$. We assume that:*

(i) ***Regularity of the dictionary φ_T:*** *Assumptions 2.1 and 2.2 hold.*

(ii) ***Regularity of the limit kernel $\mathcal{K}_\infty$:*** *Assumption 2.3 holds, we have $r \in \left(0, 1/\sqrt{2L_{2,0}}\right)$, and also $\varepsilon_\infty(r/\rho) > 0$ and $\nu_\infty(\rho r) > 0$.*

(iii) **Separation of the non-linear parameters:** *There exists $u_\infty \in \left(0, H_\infty^{(2)}(r,\rho)\right)$ such that:*

$$\delta_\infty(u_\infty, s) < +\infty.$$

(iv) ***Closeness of the metrics $\mathfrak{d}_T$ and $\mathfrak{d}_\infty$:*** *We have $\rho_T \leq \rho$.*

(v) ***Proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$:***

$$\mathcal{V}_T \leq H_\infty^{(1)}(r,\rho) \quad \text{and} \quad (s-1)\mathcal{V}_T \leq H_\infty^{(2)}(r,\rho) - u_\infty.$$

Then, with the positive constants:

$$C_N = \frac{\nu_\infty(\rho r)}{180}, \quad C'_N = \frac{5}{8}L_{2,0} + \frac{1}{8}L_{2,1} + \frac{1}{2}, \quad C_B = 2 \quad \text{and} \quad C_F = \frac{\varepsilon_\infty(r/\rho)}{10} \leq 1, \tag{38}$$

Assumption 4.1 holds (with the same r) for any subset $\mathcal{Q}^\star = \{\theta_i^\star, 1 \leq i \leq s\}$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$:

$$\mathfrak{d}_T(\theta, \theta') > 2\max(r, \rho_T\, \delta_\infty(u_\infty, s)).$$

We state a second result that gives sufficient conditions for Assumption 4.2 to hold. The proof is given in Section 8.2.

Proposition 4.2 (Interpolating derivative certificate). *Let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$ and $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$. We assume that:*

(i) ***Regularity of the dictionary φ_T:*** *Assumptions 2.1 and 2.2 hold.*

(ii) ***Regularity of the limit kernel $\mathcal{K}_\infty$:*** *Assumption 2.3 holds.*

(iii) ***Separation of the non-linear parameters:*** *There exists $u'_\infty \in (0, 1/6)$, such that:*

$$\delta_\infty(u'_\infty, s) < +\infty.$$

(iv) ***Proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$:*** *We have:*

$$\mathcal{V}_T \leq 1 \quad \text{and} \quad (s-1)\mathcal{V}_T + u'_\infty \leq 1/6.$$

Then, with the positive constants:

$$c_N = \frac{1}{8}L_{2,0} + \frac{5}{8}L_{2,1} + \frac{7}{8}, \quad c_B = 2 \quad \text{and} \quad c_F = \frac{5}{4}L_{1,0} + \frac{7}{4}, \tag{39}$$

Assumption 4.2 holds for any $r > 0$ and any subset $\mathcal{Q}^\star = \{\theta_i^\star,\, 1 \leq i \leq s\}$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$:

$$\mathfrak{d}_T(\theta, \theta') > 2 \max(r, \rho_T\, \delta_\infty(u'_\infty, s)).$$

The assumptions of Proposition 4.1 (resp. 4.2) are identical to those of [14, Proposition 7.4] (resp. [14, Proposition 7.5]). It is not surprising since those results are based on the same construction of certificates. In order to build a certificate $\eta : (z, \theta) \mapsto \mathbb{R}$ satisfying Assumption 4.1 or 4.2, we shall build for every element $z \in \mathcal{Z}$ certificate functions $\eta_z(\theta) \mapsto \mathbb{R}$ following the same construction as in [14] and set $\eta(z, \theta) = \eta_z(\theta)$. The functions η_z will be coupled through interpolated values on $\mathcal{Q}^\star$.

5. Proof of Theorem 3.1

In this section, we shall prove Theorem 3.1. We closely follow the proof of [14, Theorem 2.1] and extend it to the case of a measure ν that is not necessarily a Dirac measure. We decompose the risk over values of estimated non-linear parameters $\hat{\theta}_\ell$ in a neighborhood of the true values $\theta_k^\star$ and those which are far away. Linear functionals of the noise depending on some $\theta \in\ \Theta_T$ appear in the bounds and we use tail bounds on the suprema of these functionals over all possible values of θ.

Let us bound the prediction error

$$\hat{R}_T := \frac{1}{\sqrt{\nu(\mathcal{Z})}} \left\| \hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star) \right\|_{L_T}.$$

The predicition error corresponds to the integration on $\mathcal{Z}$ of the prediction error studied in [14, Theorem 2.1].

By definition (3) of $\hat{B}$ and $\hat{\vartheta}$ for the tuning parameter κ, we have:

$$\frac{1}{2\nu(\mathcal{Z})} \left\| Y - \hat{B}\Phi_T(\hat{\vartheta}) \right\|_{L_T}^2 + \kappa \|\hat{B}\|_{\ell_1, L^p(\nu)} \leq \frac{1}{2\nu(\mathcal{Z})} \|Y - B^\star \Phi_T(\vartheta^\star)\|_{L_T}^2 + \kappa \|B^\star\|_{\ell_1, L^p(\nu)}. \tag{40}$$

We define the application $\hat{\Upsilon}$ from L_T to $\mathbb{R}$ by:

$$\hat{\Upsilon}(F) = \left\langle \hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star), F \right\rangle_{L_T}.$$

This gives, by rearranging terms and using the equation of the model $Y = B^\star \Phi_T(\vartheta^\star) + W_T$, that:

$$\frac{1}{2} \hat{R}_T^2 \leq \frac{1}{\nu(\mathcal{Z})} \hat{\Upsilon}(W_T) + \kappa \left(\|B^\star\|_{\ell_1, L^p(\nu)} - \|\hat{B}\|_{\ell_1, L^p(\nu)} \right). \tag{41}$$

Next, we shall expand the two terms on the right-hand side of (41). In the rest of the proof, we fix $r > 0$ so that Assumptions 4.1 and 4.2 are verified for $\mathcal{Q}^\star$. In particular, for all $k \neq k'$ in the support $S^\star = \{k,\, \|B_k^\star\|_{L^2(\nu)} \neq 0\ \}$ we have $\mathfrak{d}_T(\theta_k^\star, \theta_{k'}^\star) > 2r$.

We give the definitions of the sets of indices $\hat{S}$, $\tilde{S}_k(r)$ and $\tilde{S}(r)$ for $k \in S^\star$:

- $\hat{S} = \left\{ \ell : \left\| \hat{B}_\ell \right\|_{L^p(\nu)} \neq 0 \right\}$ the support set of $\hat{B}$ given by the optimization problem (3);
- $\tilde{S}_k(r) = \left\{ \ell \in \hat{S} : \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) \leq r \right\}$ the set of indices ℓ in the support of $\hat{B}$ associated to the active parametric functions having $\hat{\theta}_\ell$ close to the true parameter $\theta_k^\star$, for a fixed k in $S^\star$;
- $\tilde{S}(r) = \bigcup_{k \in S^\star} \tilde{S}_k(r)$ the set of indices ℓ in the support of $\hat{B}$ associated to the active parametric functions having $\hat{\theta}_\ell$ close to any true parameter $\theta_k^\star$, for some k in $S^\star$.

Since the closed balls $\mathcal{B}_T(\theta_k^\star, r)$ with $k \in S^\star$ are pairwise disjoint, the sets $\tilde{S}_k(r)$, for $k \in S^\star$, are also pairwise disjoint and one can write the following decomposition with $\tilde{S}(r)^c = \{1, \cdots, K\} \setminus \tilde{S}(r)$:

$$\begin{aligned}\hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star) &= \sum_{k=1}^{K} \hat{B}_k \phi_T(\hat{\theta}_k) - \sum_{k\in S^\star} B_k^\star \phi_T(\theta_k^\star)\\ &= \sum_{k\in S^\star, \tilde{S}_k(r)\neq\emptyset} \sum_{\ell\in\tilde{S}_k(r)} \hat{B}_\ell \phi_T(\hat{\theta}_\ell) + \sum_{k\in\tilde{S}(r)^c} \hat{B}_k \phi_T(\hat{\theta}_k) - \sum_{k\in S^\star} B_k^\star \phi_T(\theta_k^\star).\end{aligned}$$

This decomposition groups the elements of the predicted mixture according to the proximity of the estimated parameter $\hat{\theta}_\ell$ to a true underlying parameter $\theta_k^\star$ to be estimated. We use a Taylor-type expansion with the Riemannian distance $\mathfrak{d}_T$ for the function $\phi_T(\theta)$ around the elements of $\mathcal{Q}^\star$. By Assumption, the function ϕ_T is twice continuously differentiable with respect to the variable θ and the function g_T is positive on Θ_T. We recall that $\tilde{D}_{i;T}[\phi_T] = \phi_T^{[i]}$ for $i = 0, 1, 2$. According to [14, Lemma 4.2], we have for any $\theta_k^\star$ and $\hat{\theta}_\ell$ in Θ_T:

$$\phi_T(\hat{\theta}_\ell) = \phi_T(\theta_k^\star) + \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)\, \phi_T^{[1]}(\theta_k^\star) + \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \int_0^1 (1-s)\phi_T^{[2]}(\gamma_s^{(k\ell)})\, \mathrm{d}s,$$

where $\gamma^{(k\ell)}$ is a distance realizing geodesic path belonging to Θ_T such that $\gamma_0^{(k\ell)} = \theta_k^\star$, $\gamma_1^{(k\ell)} = \hat{\theta}_\ell$ and $\mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) = \int_0^1 |\dot{\gamma}_s^{(k\ell)}| \sqrt{g_T(\gamma_s^{(k\ell)})}\mathrm{d}s$.

Hence we obtain:

$$\begin{aligned}\hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star) = &\sum_{k\in S^\star} I_{0,k}(r)\, \phi_T(\theta_k^\star) + \sum_{k\in S^\star} I_{1,k}(r)\, \phi_T^{[1]}(\theta_k^\star) + \sum_{k\in\tilde{S}(r)^c} \hat{B}_k\, \phi_T(\hat{\theta}_k)\\ &+ \sum_{k\in S^\star} \left(\sum_{\ell\in\tilde{S}_k(r)} \hat{B}_\ell\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \int_0^1 (1-s)\phi_T^{[2]}(\gamma_s^{(k\ell)})\, \mathrm{d}s \right),\end{aligned} \tag{42}$$

with

$$I_{0,k}(r) = \left(\sum_{\ell\in\tilde{S}_k(r)} \hat{B}_\ell \right) - B_k^\star \quad \text{and} \quad I_{1,k}(r) = \sum_{\ell\in\tilde{S}_k(r)} \hat{B}_\ell \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star). \tag{43}$$

The functions $I_{0,k}(r)$ and $I_{1,k}(r)$ belong to $L^2(\nu)$. We shall omit the dependence in r when there is no ambiguity. In particular, we write $I_{0,k}(z)$ for $I_{0,k}(r)(z)$. Let us introduce some notations in order to bound the different terms of the expansion above:

$$I_0(r) = \sum_{k\in S^\star} \|I_{0,k}(r)\|_{L^p(\nu)} \qquad \text{and} \qquad I_1(r) = \sum_{k\in S^\star} \|I_{1,k}(r)\|_{L^p(\nu)}, \tag{44}$$

$$I_{2,k}(r) = \sum_{\ell\in\tilde{S}_k(r)} \left\|\hat{B}_\ell\right\|_{L^p(\nu)} \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \qquad \text{and} \qquad I_2(r) = \sum_{k\in S^\star} I_{2,k}(r), \tag{45}$$

$$I_3(r) = \sum_{\ell\in\tilde{S}(r)^c} \left\|\hat{B}_\ell\right\|_{L^p(\nu)} = \left\|\hat{B}_{\tilde{S}(r)^c}\right\|_{\ell_1, L^p(\nu)}, \tag{46}$$

where $\hat{B}_{\tilde{S}(r)^c}$ denotes the restriction of the vector-valued application $\hat{B}$ to its components in the set of indices $\tilde{S}(r)^c$. We recall that we omit the dependence in r when there is no ambiguity. These quantities are generalizations of the real numbers I_i, where $i = 0, \cdots, 3$, defined in the proof of [14, Theorem 2.1] as they correspond here to sums of $L^p(\nu)$ norms instead of sums of absolute values.

We bound the difference $\|B^\star\|_{\ell_1,L^p(\nu)} - \left\|\hat{B}\right\|_{\ell_1,L^p(\nu)}$ by noticing that:

$$\|B^\star\|_{\ell_1,L^p(\nu)} - \left\|\hat{B}\right\|_{\ell_1,L^p(\nu)} = \sum_{k\in S^\star}\left(\|B_k^\star\|_{L^p(\nu)} - \sum_{\ell\in\tilde{S}_k(r)}\left\|\hat{B}_\ell\right\|_{L^p(\nu)}\right) - \sum_{k\in\tilde{S}(r)^c}\left\|\hat{B}_k\right\|_{L^p(\nu)} \leq I_0. \tag{47}$$

In the next lemma, we give an upper bound of I_0. Recall the constants C'_N and C_F from Assumption 4.1.

Let $f \in L^2(\nu)$, we define the application $v : L^2(\nu) \to L^2(\nu)$ such that for any $z \in \mathcal{Z}$:

$$v(f)(z) = \begin{cases} \operatorname{sign}(f(z))\,\frac{|f(z)|^{p-1}}{\|f\|^{p-1}_{L^p(\nu)}} & \text{if} \quad \|f\|_{L^p(\nu)} > 0, \\ \nu(\mathcal{Z})^{-\frac{1}{q}} & \text{otherwise,} \end{cases} \tag{48}$$

so that $\|v(f)\|_{L^q(\nu)} = 1$.

Lemma 5.1. *Under the assumptions of Theorem 3.1 and with the element $P_1 \in H_T$ from Assumption 4.1 associated to the function $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ defined by:*

$$V(z, \theta_k^\star) = v(I_{0,k})(z), \tag{49}$$

we get that:

$$I_0 \leq C'_N I_2 + (1 - C_F) I_3 + |\hat{\Upsilon}(P_1)|. \tag{50}$$

Proof. We have $\|I_{0,k}\|_{L^p(\nu)} = \|I_{0,k}\|^p_{L^p(\nu)}/\|I_{0,k}\|^{p-1}_{L^p(\nu)}$ and therefore:

$$I_0 := \sum_{k\in S^\star}\|I_{0,k}\|_{L^p(\nu)} = \sum_{k\in S^\star}\int V(z,\theta_k^\star)\left(\left(\sum_{\ell\in\tilde{S}_k(r)}\hat{B}_\ell(z)\right) - B_k^\star(z)\right)\nu(\mathrm{d}z).$$

Let P_1 be an element of L_T from Assumption 4.1 associated to the function V such that properties $(i)-(iv)$ therein hold. By adding and substracting $\sum_{k\in S^\star}\sum_{\ell\in\tilde{S}_k(r)}\left\langle \hat{B}_\ell\phi_T(\hat{\theta}_\ell), P_1\right\rangle_{L_T}$ to I_0 and using the property (ii) satisfied by the element P_1, that is, $\langle\phi_T(\theta_k^\star), P_1(z)\rangle_T = V(z,\theta_k^\star)$ for all $k \in S^\star$ and ν-almost every $z \in \mathcal{Z}$, we obtain:

$$I_0 = \sum_{k\in S^\star}\sum_{\ell\in\tilde{S}_k(r)}\int \hat{B}_\ell(z)\left(V(z,\theta_k^\star) - \left\langle\phi_T(\hat{\theta}_\ell), P_1(z)\right\rangle_T\right)\nu(\mathrm{d}z) + \hat{\Upsilon}(P_1) - \sum_{\ell\in\tilde{S}(r)^c}\left\langle\hat{B}_\ell\phi_T(\hat{\theta}_\ell), P_1\right\rangle_{L_T}.$$

We deduce, using Hölder's inequality, that:

$$I_0 \leq \sum_{k\in S^\star}\sum_{\ell\in\tilde{S}_k(r)}\left\|\hat{B}_\ell\right\|_{L^p(\nu)}\left\|V(\theta_k^\star) - \left\langle\phi_T(\hat{\theta}_\ell), P_1\right\rangle_T\right\|_{L^q(\nu)} + |\hat{\Upsilon}(P_1)| + \sum_{\ell\in\tilde{S}(r)^c}\left\|\hat{B}_\ell\right\|_{L^p(\nu)}\left\|\left\langle\phi_T(\hat{\theta}_\ell), P_1\right\rangle_T\right\|_{L^q(\nu)}.$$

Notice that $\hat{\theta}_\ell \notin \bigcup_{k\in S^\star}\mathcal{B}_T(\theta_k^\star, r)$ for $\ell \in \tilde{S}(r)^c$. Then, by using the properties (ii) and (iii) from Assumption 4.1, we get that (50) holds with the constants C'_N and C_F from Assumption 4.1. □

In the next lemma, we give an upper bound of I_1. Recall the constants c_N and c_F from Assumption 4.2. Recall the application v defined in (48).

Lemma 5.2. *Under the assumptions of Theorem 3.1 and with the element $Q_0 \in L_T$ from Assumption 4.2 associated to the function $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ defined by:*

$$V(z, \theta_k^\star) = v(I_{1,k})(z), \tag{51}$$

we get that:

$$I_1 \leq c_N I_2 + c_F I_3 + |\hat{\Upsilon}(Q_0)|. \tag{52}$$

Proof. We have writing $I_{1,k}(z)$ for $I_{1,k}(r)(z)$:

$$I_1 = \sum_{k \in S^\star} \|I_{1,k}\|_{L^p(\nu)} = \sum_{k \in S^\star} \int V(z, \theta_k^\star) I_{1,k}(z) \nu(\mathrm{d}z).$$

Let Q_0 be an element of L_T from Assumption 4.2 associated to the function V such that properties $(i) - (iii)$ therein hold. By adding and substracting $\sum_{\ell \in \tilde{S}(r)} \left\langle \hat{B}_\ell \phi_T(\hat{\theta}_\ell), Q_0 \right\rangle_{L_T} = \left\langle \hat{B} \Phi_T(\hat{\vartheta}), Q_0 \right\rangle_{L_T} - \sum_{\ell \in \tilde{S}(r)^c} \left\langle \hat{B}_\ell \phi_T(\hat{\theta}_\ell), Q_0 \right\rangle_{L_T}$ to I_1 and using the triangle inequality, we obtain:

$$\begin{aligned} I_1 \leq \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \int |\hat{B}_\ell(z)| \left| V(z, \theta_k^\star) \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star) \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) - \left\langle \phi_T(\hat{\theta}_\ell), Q_0(z) \right\rangle_T \right| \nu(\mathrm{d}z) \\ + \sum_{\ell \in \tilde{S}(r)^c} \left| \left\langle \hat{B}_\ell \phi_T(\hat{\theta}_\ell), Q_0 \right\rangle_{L_T} \right| + \left| \left\langle \hat{B} \Phi_T(\hat{\vartheta}), Q_0 \right\rangle_{L_T} \right|. \end{aligned}$$

The property (i) of Assumption 4.2 gives that $\langle \phi_T(\theta_k^\star), Q_0(z) \rangle_T = 0$ for all $k \in S^\star$ and $\nu-$almost every $z \in \mathcal{Z}$. This implies that $\langle B^\star \Phi_T(\vartheta^\star), Q_0 \rangle_{L_T} = 0$. Then, by using the definition of I_2 and I_3 from (45)-(46) and the properties (i) and (ii) of Assumption 4.2, we obtain:

$$I_1 \leq c_N I_2 + c_F I_3 + \left| \left\langle \hat{B} \Phi_T(\hat{\vartheta}), Q_0 \right\rangle_{L_T} \right| = c_N I_2 + c_F I_3 + |\hat{\Upsilon}(Q_0)|,$$

with the constants c_N and c_F from Assumption 4.2. □

We consider the following random variables for $j = 0, 1, 2$:

$$M_j = \sup_{\theta \in \Theta_T} \left\| \left\langle W_T, \phi_T^{[j]}(\theta) \right\rangle_T \right\|_{L^q(\nu)}. \tag{53}$$

By using the expansion (42), Hölder's inequality and the bounds (52) and (50) for the second inequality, we obtain:

$$\begin{aligned} |\hat{\Upsilon}(W_T)| &\leq (I_0 + I_3) M_0 + I_1 M_1 + I_2 \, 2^{-1} \, M_2 \qquad (54) \\ &\leq (C'_N I_2 + (2 - C_F) I_3 + |\hat{\Upsilon}(P_1)|) M_0 + (c_N I_2 + c_F I_3 + |\hat{\Upsilon}(Q_0)|) M_1 + I_2 \, 2^{-1} \, M_2. \qquad (55) \end{aligned}$$

At this point, one needs to bound I_2 and I_3. In order to do so, we bound from above and from below the Bregman divergence D_B defined by:

$$D_B = \|\hat{B}\|_{\ell_1, L^p(\nu)} - \|B^\star\|_{\ell_1, L^p(\nu)} - \hat{\Upsilon}(P_0), \tag{56}$$

where P_0 is the element given by Assumption 4.1 associated to the function V given by:

$$V(z, \theta_k^\star) = \operatorname{sign}(B_k^\star(z)) \frac{|B_k^\star(z)|^{p-1}}{\|B_k^\star\|_{L_p(\nu)}^{p-1}} \quad \text{for all } k \in S^\star. \tag{57}$$

The next lemma gives a lower bound of the Bregman divergence.

Lemma 5.3. *Under the assumptions of Theorem 3.1 and with the constants C_N and C_F of Assumption 4.1, we get that:*

$$D_B \geq C_N I_2 + C_F I_3. \tag{58}$$

Proof. By definition (56) of D_B we have:

$$D_B = \sum_{k\in\hat{S}} \left(\left\| \hat{B}_k \right\|_{L^p(\nu)} - \left\langle \hat{B}_k \phi_T(\hat{\theta}_k), P_0 \right\rangle_{L_T} \right) - \sum_{k\in S^\star} \left(\| B_k^\star \|_{L^p(\nu)} - \langle B_k^\star \phi_T(\theta_k^\star), P_0 \rangle_{L_T} \right).$$

By using the interpolating properties of P_0 from Assumption 4.1 associated to V defined in (57), we have $\sum\limits_{k\in S^\star} \| B_k^\star \|_{L^p(\nu)} - \langle B_k^\star \phi_T(\theta_k^\star), P_0 \rangle_{L_T} = 0$. Hence, we deduce that:

$$\begin{aligned}
D_B &= \sum_{k\in\hat{S}} \left\| \hat{B}_k \right\|_{L^p(\nu)} - \left\langle \hat{B}_k \phi_T(\hat{\theta}_k), P_0 \right\rangle_{L_T} \\
&\geq \sum_{k\in\hat{S}} \left\| \hat{B}_k \right\|_{L^p(\nu)} - \left| \left\langle \hat{B}_k \phi_T(\hat{\theta}_k), P_0 \right\rangle_{L_T} \right| \\
&\geq \sum_{k\in\hat{S}} \left\| \hat{B}_k \right\|_{L^p(\nu)} - \left\| \hat{B}_k \right\|_{L^p(\nu)} \left\| \left\langle \phi_T(\hat{\theta}_k), P_0 \right\rangle_T \right\|_{L^q(\nu)} \\
&\geq \sum_{\ell\in\tilde{S}(r)} \left\| \hat{B}_\ell \right\|_{L^p(\nu)} \left(1 - \left\| \left\langle \phi_T(\hat{\theta}_\ell), P_0 \right\rangle_T \right\|_{L^q(\nu)} \right) + \sum_{k\in\tilde{S}(r)^c} \left\| \hat{B}_k \right\|_{L^p(\nu)} \left(1 - \left\| \left\langle \phi_T(\hat{\theta}_k), P_0 \right\rangle_T \right\|_{L^q(\nu)} \right).
\end{aligned}$$

Thanks to properties *(i)* and *(iii)* of Assumption 4.1 and the definitions (45) and (46) of I_2 and I_3, we obtain:

$$D_B \geq \sum_{k\in S^\star} \sum_{\ell\in\tilde{S}_k(r)} C_N \left\| \hat{B}_\ell \right\|_{L^p(\nu)} \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 + \sum_{k\in\tilde{S}(r)^c} C_F \left\| \hat{B}_k \right\|_{L^p(\nu)} \geq C_N I_2 + C_F I_3,$$

where the constants C_N and C_F are that of Assumption 4.1. □

We now give an upper bound of the Bregman divergence.

Lemma 5.4. *Under the assumptions of Theorem 3.1, we have:*

$$\begin{aligned}
\kappa\,\nu(\mathcal{Z})\, D_B \leq I_2 \left(C'_N M_0 + c_N M_1 + 2^{-1} M_2 \right) + I_3 \left((2 - C_F) M_0 + c_F M_1 \right) \\
+ |\hat{\Upsilon}(P_1)| M_0 + |\hat{\Upsilon}(Q_0)| M_1 + \kappa\,\nu(\mathcal{Z})\, |\hat{\Upsilon}(P_0)|.
\end{aligned} \tag{59}$$

Proof. Recall that $\mathcal{Q}^\star \subset \Theta_T$. We deduce from (41) that:

$$\kappa(\|\hat{B}\|_{\ell_1, L^p(\nu)} - \|B^\star\|_{\ell_1, L^p(\nu)}) \leq \frac{1}{\nu(\mathcal{Z})} \hat{\Upsilon}(W_T) - \frac{1}{2} \hat{R}_T^2 \leq \frac{1}{\nu(\mathcal{Z})} \hat{\Upsilon}(W_T).$$

Together with (56), we obtain:

$$\kappa D_B \leq \frac{1}{\nu(\mathcal{Z})} |\hat{\Upsilon}(W_T)| + \kappa |\hat{\Upsilon}(P_0)|.$$

Then, use (55) to get (59). □

By combining the upper and lower bounds (58) and (59), we deduce that:

$$\begin{aligned}
I_2 \left(C_N - \frac{1}{\kappa\,\nu(\mathcal{Z})} \left(C'_N M_0 + c_N M_1 + 2^{-1} M_2 \right) \right) + I_3 \left(C_F - \frac{1}{\kappa\,\nu(\mathcal{Z})} \left((2 - C_F) M_0 + c_F M_1 \right) \right) \\
\leq \frac{1}{\kappa\,\nu(\mathcal{Z})} |\hat{\Upsilon}(P_1)| M_0 + \frac{1}{\kappa\,\nu(\mathcal{Z})} |\hat{\Upsilon}(Q_0)| M_1 + |\hat{\Upsilon}(P_0)|.
\end{aligned} \tag{60}$$

We define the events:

$$\mathcal{A}_i = \{M_i \leq \mathcal{C}\,\kappa\,\nu(\mathcal{Z})\}\,, \quad \text{for } i \in \{0,1,2\} \quad \text{and} \quad \mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2, \tag{61}$$

where:

$$\mathcal{C} = \frac{C_F}{2(2 - C_F + c_F)} \wedge \frac{C_N}{2(C'_N + c_N + 2^{-1})}.$$

We get from Inequality (60), that on the event $\mathcal{A}$:

$$C_N I_2 + C_F I_3 \leq 2\mathcal{C}' \left(|\hat{\Upsilon}(P_1)| + |\hat{\Upsilon}(Q_0)| + |\hat{\Upsilon}(P_0)|\right) \quad \text{with} \quad \mathcal{C}' = \mathcal{C} \vee 1. \tag{62}$$

By reinjecting (47), (55), (50) and (52) in (41) one gets:

$$\frac{1}{2}\hat{R}_T^2 \leq I_2 \left(\frac{C'_N M_0 + c_N M_1 + 2^{-1} M_2}{\nu(\mathcal{Z})} + \kappa C'_N\right) + I_3 \left(\frac{(2 - C_F) M_0 + c_F M_1}{\nu(\mathcal{Z})} + \kappa(1 - C_F)\right)$$
$$+ |\hat{\Upsilon}(P_1)| \left(\frac{M_0}{\nu(\mathcal{Z})} + \kappa\right) + |\hat{\Upsilon}(Q_0)| \frac{M_1}{\nu(\mathcal{Z})}.$$

Using (62), we obtain an upper bound for the prediction error on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq C\,\kappa\,(|\hat{\Upsilon}(P_0)| + |\hat{\Upsilon}(P_1)| + |\hat{\Upsilon}(Q_0)|), \tag{63}$$

with

$$C = 4\mathcal{C}' \left(1 + \frac{\mathcal{C}'}{C_N}(2C'_N + c_N + 1) + \frac{\mathcal{C}'}{C_F}(3 - 2C_F + c_F)\right).$$

Using the Cauchy-Schwarz inequality and the definition of $\hat{\Upsilon}$, we get that for $f \in L_T$:

$$|\hat{\Upsilon}(f)| \leq \hat{R}_T \sqrt{\nu(\mathcal{Z})}\, \|f\|_{L_T}. \tag{64}$$

Using Assumption 4.1 *(iv)* for P_0 and P_1, and Assumption 4.2 *(iii)* for Q_0, we get:

$$\|P_0\|_{L_T} \leq C_B \sqrt{s}\nu(\mathcal{Z})^{1/2p - 1/2q}, \quad \|P_1\|_{L_T} \leq C_B \sqrt{s}\nu(\mathcal{Z})^{1/2p-1/2q} \quad \text{and} \quad \|Q_0\|_{L_T} \leq c_B \sqrt{s}\nu(\mathcal{Z})^{1/2p-1/2q}. \tag{65}$$

Plugging this in (63), we get that on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq \mathcal{C}_0\,\kappa \hat{R}_T \sqrt{s}\,\nu(\mathcal{Z})^{\frac{1}{p}} \quad \text{with} \quad \mathcal{C}_0 = (c_B + 2C_B)C. \tag{66}$$

We obtain (22) on the event $\mathcal{A}$ defined in (61).

6. Proof of Corollary 3.4

This section is dedicated to the proof of Corollary 3.4. We shall apply Theorem 3.1 in the particular case $p = 2$ and $q = 2$. Recall that the measure ν is a sum of n weighted Dirac measures. All the assumptions of Theorem 3.1 are in force. We shall only give tail bounds for the quantities M_j with $j = 0, 1, 2$ defined in (24).

For $j = 0, 1, 2$ and $\theta \in \Theta_T$, we set $X_j(\theta) = \left\| \left\langle W_T, \phi_T^{[j]}(\theta)\right\rangle_T \right\|_{L^2(\nu)}$. Notice that $M_j = \sup_{\Theta_T} X_j$ and that the process X_j^2 is a χ^2 process.

We first consider $j = 0$. Using (13) and (15), we have that:

$$\|\phi_T(\theta)\|_T^2 = 1 \quad \text{and} \quad \left\|\phi_T^{[1]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[1,1]}(\theta, \theta) = 1.$$

We define two functions f_n and g_n on $\mathbb{R}$ by:

$$f_n(x) = \mathrm{e}^{-x(1 - 2\sqrt{\frac{n}{x}})} \quad \text{and} \quad g_n(x) = \frac{x^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}\,\mathrm{e}^{-x/2}, \tag{67}$$

where Γ denotes the gamma function. Notice that both functions are decreasing on $[n, +\infty)$.

We set:

$$A = \frac{\mathcal{C}^2\, \nu(\mathcal{Z})^2}{\sigma^2 \|a\|_{\ell_\infty} \Delta_T}\cdot$$

Recall Assumption 3.1 on the noise holds. We deduce from Lemma A.2 with $C_1 = C_2 = 1$ and $u = \mathcal{C}^2\, \kappa^2\, \nu(\mathcal{Z})^2$, that for $\kappa \geq \sqrt{(n+1)/A}$:

$$\mathbb{P}\left(M_0^2 > \mathcal{C}^2\, \kappa^2\, \nu(\mathcal{Z})^2\right) \leq f_n\left(\kappa^2\, A\right) + \frac{4|\Theta_T|_{\mathfrak{d}_T}}{2^{n/2}} g_n\left(\kappa^2 A\right), \tag{68}$$

where $|\Theta_T|_{\mathfrak{d}_T}$ denotes the diameter of the set Θ_T with respect to the metric $\mathfrak{d}_T$.

We consider $j = 1$. We have by (13) and (15) that:

$$\left\|\phi_T^{[1]}(\theta)\right\|_T^2 = 1 \quad \text{and} \quad \left\|\tilde{D}_{1;T}[\phi_T^{[1]}](\theta)\right\|_T^2 = \left\|\phi_T^{[2]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[2,2]}(\theta, \theta).$$

Recall $L_{2,2}$ and $\mathcal{V}_T$ are defined in (18) and (21). Since Assumptions 2.3 and 2.4 hold, we get that for $\theta \in \Theta_T$:

$$\mathcal{K}_T^{[2,2]}(\theta, \theta) \leq L_{2,2} + \mathcal{V}_T \leq 2L_{2,2}.$$

We deduce from Lemma A.2 with $C_1 = 1$, $C_2 = \sqrt{2L_{2,2}}$ and $u = \mathcal{C}^2\, \kappa^2\, \nu(\mathcal{Z})^2$, that for $\kappa \geq \sqrt{(n+1)/A}$:

$$\mathbb{P}\left(M_1^2 > \mathcal{C}^2\, \kappa^2\, \nu(\mathcal{Z})^2\right) \leq f_n\left(\kappa^2 A\right) + \frac{4\sqrt{2L_{2,2}}\, |\Theta_T|_{\mathfrak{d}_T}}{2^{n/2}} g_n\left(\kappa^2 A\right). \tag{69}$$

We consider $j = 2$. We have by (15) that:

$$\left\|\phi_T^{[2]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[2,2]}(\theta, \theta) \quad \text{and} \quad \left\|\tilde{D}_{1;T}[\phi_T^{[2]}](\theta)\right\|_T^2 = \left\|\phi_T^{[3]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[3,3]}(\theta, \theta).$$

Recall the definition of the function h_∞ from (17) and the constants $L_{2,2}$, L_3, $\mathcal{V}_T$ defined in (18) and (21). Using also Assumption 2.4 so that $\mathcal{V}_T \leq L_{2,2} \wedge L_3$, we get that for all $\theta \in \Theta_T$:

$$\mathcal{K}_T^{[2,2]}(\theta, \theta) \leq L_{2,2} + \mathcal{V}_T \leq 2L_{2,2} \quad \text{and} \quad \mathcal{K}_T^{[3,3]}(\theta, \theta) \leq L_3 + \mathcal{V}_T \leq 2\, L_3.$$

We deduce from Lemma A.2 with $C_1 = \sqrt{2L_{2,2}}$, $C_2 = \sqrt{2L_3}$ and $u = \mathcal{C}^2\, \kappa^2\, \nu(\mathcal{Z})^2$, that for

$$\kappa \geq \sqrt{2\, L_{2,2}\, (n+1)/A},$$

we have:

$$\mathbb{P}\left(M_2^2 > \mathcal{C}^2\, \kappa^2\, \nu(\mathcal{Z})^2\right) \leq f_n\left(\frac{\kappa^2\, A}{2L_{2,2}}\right) + \frac{4\sqrt{L_3}\, |\Theta_T|_{\mathfrak{d}_T}}{\sqrt{L_{2,2}}\, 2^{n/2}} g_n\left(\frac{\kappa^2\, A}{2L_{2,2}}\right). \tag{70}$$

Wet set:

$$B = \frac{\mathcal{C}_1'^2\, \nu(\mathcal{Z})^2}{\sigma^2 \|a\|_{\ell_\infty} \Delta_T} \quad \text{with} \quad \mathcal{C}_1' = \sqrt{\frac{\mathcal{C}^2}{2L_{2,2} \vee 1}}. \tag{71}$$

We deduce from (68), (69) and (70) that for $\kappa \geq \sqrt{(n+1)/B}$:

$$\sum_{j=0}^{2} \mathbb{P}(M_j > \mathcal{C}\, \kappa\, \nu(\mathcal{Z})) \leq 3\left(f_n\left(\kappa^2 B\right) + \frac{\mathcal{C}_2'|\Theta_T|_{\mathfrak{d}_T}}{2^{n/2}} g_n\left(\kappa^2 B\right)\right), \tag{72}$$

where the constant $\mathcal{C}_2'$ is finite positive and defined by:

$$\mathcal{C}_2' = 4\left(1 \vee \sqrt{2L_{2,2}} \vee \frac{\sqrt{L_3}}{\sqrt{L_{2,2}}}\right).$$

Recall that the functions f_n and g_n are decreasing on $[n, +\infty)$. We get the following asymptotically-equivalent functions (up to a multiplicative constant) for $f_n(c\,n)$ and $g_n(c\,n)$ and some positive constant c:

$$\begin{aligned} &f_n(c\,n) = \mathrm{e}^{-n(c-2\sqrt{c})} \\ &g_n(c\,n)/2^{\frac{n}{2}} \asymp \mathrm{e}^{-\frac{n}{2}(c-\log(c)-1)+\frac{1}{2}\log(n)} \lesssim \mathrm{e}^{-\frac{n}{2}(c-\log(c)-3/2)}. \end{aligned} \tag{73}$$

Indeed, we use that $\Gamma(n/2) \asymp \mathrm{e}^{\frac{n}{2}\log(\frac{n}{2})-\frac{n}{2}-\frac{1}{2}\log(n)}$. Thus, the constant c determines which of the two terms $f_n(c\,n)$ and $g_n(c\,n)/2^{\frac{n}{2}}$ is dominant.

By solving a second order inequality, we give a lower bound on the tuning parameter κ so that the first right hand term of (72) is bounded by $1/\tau$ for some $\tau > 1$.

Indeed for $\tau > 1$ and $\kappa \geq \sqrt{(n+1)/B}\left(1+\sqrt{1+\frac{\log(\tau)}{n}}\right)$ we have:

$$f_n(\kappa^2 B) \leq \frac{1}{\tau}.$$

We also have:

$$g_n(\kappa^2 B) \leq g_n\left(n\left(1+\sqrt{1+\frac{\log(\tau)}{n}}\right)^2\right) \leq g_n\,(n)\,\mathrm{e}^{-n/2}/\sqrt{\tau},$$

where we used that g_n is decreasing on $[n, +\infty)$ for the first inequality and that $\log(1+x) \leq x$ for the second. So that, we get:

$$\sum_{j=0}^{2} \mathbb{P}(M_j > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})) \leq \frac{3}{\tau} + \frac{3\,\mathcal{C}_2'|\Theta_T|_{\mathfrak{d}_T}}{\sqrt{\tau}\,2^{\frac{n}{2}}} g_n\,(n)\,\mathrm{e}^{-n/2}. \tag{74}$$

Then, by using (73), we deduce an asymptotical equivalence up to a multiplicative constant: $F(n) := g_n(n)\,\mathrm{e}^{-n/2}/2^{n/2} \asymp \mathrm{e}^{-n/2+\log(n)/2}$.

Finally, using the definition of B given in (71), when

$$\kappa \geq \mathcal{C}_1\sigma\sqrt{\frac{\|a\|_{\ell_\infty}\Delta_T\,n}{\nu(\mathcal{Z})^2}}\left(1+\sqrt{1+\frac{\log(\tau)}{n}}\right)$$

we get by Theorem 3.1 that the bound (22) stands with probability larger than $1-\mathcal{C}_2\left(\frac{1}{\tau}+\frac{|\Theta_T|_{\mathfrak{d}_T}F(n)}{\sqrt{\tau}}\right)$, where:

$$\mathcal{C}_1 = \sqrt{2}/\mathcal{C}_1' \quad \text{and} \quad \mathcal{C}_2 = 3(1 \vee \mathcal{C}_2').$$

This completes the proof of the corollary.

7. Proof of Corollary 3.6

In this section, we prove Corollary 3.6. We shall apply Theorem 3.1 in the particular case $p = 1$ and $q = +\infty$. Recall that the measure ν is a sum of n weighted Dirac measures. All the assumptions of Theorem 3.1 are in force. We shall only give tail bounds for the quantities M_j with $j = 0, 1, 2$ defined by $M_j = \sup_{\Theta_T} X_j$ where $X_j(\theta) = \left\|\left\langle W_T, \phi_T^{[j]}(\theta)\right\rangle_T\right\|_{L^\infty(\nu)}$.

Using Assumption 3.1, we get for any $j = 0, 1, 2$ that:

$$\mathbb{P}(M_j > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})) \leq \sum_{z\in\mathcal{Z}} \mathbb{P}\left(\sup_{\Theta_T} \left\langle W_T(z), \phi_T^{[j]}(\theta)\right\rangle_T > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})\right) \leq n\mathbb{P}\left(\sup_{\Theta_T} \left\langle w_T, \phi_T^{[j]}(\theta)\right\rangle_T > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})\right).$$

We use [14, Lemma A.2] that establishes a tail bound for suprema of smooth Gaussian processes and similar arguments as those developed in the proof of [14, Theorem 2.1] to get tail bounds on $\sup_{\Theta_T} \left\langle w_T, \phi_T^{[j]}(\theta)\right\rangle_T$ for $j = 0, 1, 2$. We obtain for any $\tau > 1$ and $\kappa \geq \mathcal{C}_3\sigma\sqrt{\Delta_T \log\tau}/\nu(\mathcal{Z})$ with $\mathcal{C}_3 = \frac{2}{\mathcal{C}}\,(1 \vee \sqrt{2L_{2,2}})$:

$$\mathbb{P}\left(\sup_{\Theta_T} \left\langle w_T, \phi_T^{[j]}(\theta)\right\rangle_T > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})\right) \leq \mathcal{C}_4' \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log\tau}} \vee \frac{1}{\tau}\right),$$

where $\mathcal{C}_4'$ is a positive constant depending on r and $\mathcal{K}_\infty$ defined in [14, Eq. (84)]. We get:

$$\sum_{j=0}^{2} \mathbb{P}(M_j > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})) \leq 3\,\mathcal{C}_4'\,n \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log\tau}} \vee \frac{1}{\tau}\right).$$

Therefore, we obtain by Theorem 3.1 that (22) stands with probability larger than $1 - \mathcal{C}_4\,n\,\left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log\tau}} \vee \frac{1}{\tau}\right)$ with $\mathcal{C}_4 = 3\,\mathcal{C}_4'$ provided the tuning parameter in (3) satisfies $\kappa \geq \mathcal{C}_3\sigma\sqrt{\Delta_T\log\tau}/\nu(\mathcal{Z})$.

8. Proofs for the construction of certificates

This section is devoted to the proof of Propositions 4.1 and 4.2. We shall first introduce norms that will be useful later in the proof. Then, we shall closely follow the proofs of [14, Propositions 7.4 and 7.5].

Let $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$, let $m, n \in \mathbb{N}$. We define a norm $\|\cdot\|_{*,q}$ on $L^q(\nu, \mathbb{R}^n)$ by:

$$\|f\|_{*,q} = \max_{1\leq k\leq n} \|f_k\|_{L^q(\nu)}.$$

We shall also define a norm on any matrix $A \in \mathbb{R}^{n\times m}$ by:

$$\|A\|_{\mathrm{op},*,q} = \sup_{\substack{f\,\in\, L^q(\nu,\mathbb{R}^m)\\ \|f\|_{*,q}\leq 1}} \|Af\|_{*,q}.$$

Recall the definition of the operator norm associated to the ℓ_∞ sup-norm defined for any matrix $A \in \mathbb{R}^{n\times m}$ by:

$$\|A\|_{\mathrm{op},\ell_\infty} = \max_{1\leq k\leq n} \sum_{1\leq \ell\leq m} |A_{k,\ell}|.$$

We have the following elementary result.

Lemma 8.1. *We have the equality on matrix norms on* $\mathbb{R}^{n\times m}$*:*

$$\|\cdot\|_{\mathrm{op},\ell_\infty} = \|\cdot\|_{\mathrm{op},*,q}.$$

Proof. Let be $A \in \mathbb{R}^{n\times m}$. We have by definition and the triangle inequality for any $f \in L^q(\nu, \mathbb{R}^m)$:

$$\|Af\|_{*,q} = \max_{1\leq k\leq n} \left\| \sum_{\ell=1}^{m} A_{k,\ell} f_\ell \right\|_{L^q(\nu)} \leq \max_{1\leq k\leq n} \sum_{\ell=1}^{m} |A_{k,\ell}| \|f_\ell\|_{L^q(\nu)}.$$

Hence for any $f \in L^q(\nu, \mathbb{R}^m)$ such that $\|f\|_{*,q} \leq 1$, we have:

$$\|Af\|_{*,q} \leq \max_{1\leq k\leq n} \sum_{1\leq \ell\leq m} |A_{k,\ell}| = \|A\|_{\mathrm{op},\ell_\infty}.$$

Therefore, we have the bound $\|A\|_{\mathrm{op},*,q} \leq \|A\|_{\mathrm{op},\ell_\infty}$.

Let us show that, in fact, we have an equality between those two norms. We set

$$k^\star = \arg \max_{1 \le k \le n} \sum_{\ell=1}^{m} |A_{k,\ell}|$$

and we define $f^\star$ so that for almost every $z \in \mathcal{Z}$, $f^\star(z) = \nu(\mathcal{Z})^{-1/q}(\operatorname{sign}(A_{k^\star,1}), \cdots, \operatorname{sign}(A_{k^\star,q}))$. We have $\|f\|_{*,q} = 1$ and $\|Af\|_{*,q} = \|A\|_{\mathrm{op},\ell_\infty}$. Thus, we have $\|A\|_{\mathrm{op},*,q} \ge \|A\|_{\mathrm{op},\ell_\infty}$. Therefore we obtain the equality $\|\cdot\|_{\mathrm{op},\ell_\infty} = \|\cdot\|_{\mathrm{op},*,q}$. □

Since the norm $\|\cdot\|_{\mathrm{op},*,q}$ does not depend on q, we note $\|\cdot\|_{\mathrm{op},*}$ instead of $\|\cdot\|_{\mathrm{op},*,q}$.

Lemma 8.2. *Let $x \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times m}$ and $f \in L^q(\nu, \mathbb{R}^m)$. We have the following inequalities:*

$$\left\|x^\top f\right\|_{L^q(\nu)} \le \|x\|_{\ell_1} \|f\|_{*,q} \quad \text{and} \quad \|Af\|_{*,q} \le \|A\|_{\mathrm{op},*} \|f\|_{*,q}.$$

Proof. This is clear since $\left\|x^\top f\right\|_{L^q(\nu)} \le \sum_{\ell=1}^m |x_\ell| \|f_\ell\|_{L^q(\nu)} \le \|x\|_{\ell_1} \|f\|_{*,q}$.

□

For a function $f : \mathcal{Z} \times \Theta \to \mathbb{R}$, we note for any $z \in \mathcal{Z}$, $f(z)$ (resp. any $\theta \in \Theta$, $f(\theta)$) the function $f(z, \cdot) : \theta \mapsto f(z, \theta)$ (resp. $f(\cdot, \theta) : z \mapsto f(z, \theta)$). The context in which we shall use this notation will be clear so that there is no confusion.

8.1. **Proof of Proposition 4.1(Construction of an interpolating certificate).** Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. Recall Assumptions 2.2 (and thus 2.1 on the regularity of φ_T) and 2.3 on the regularity of the asymptotic kernel $\mathcal{K}_\infty$ are in force. Let $\rho \ge 1$, let $r \in \left(0, 1/\sqrt{2L_{0,2}}\right)$ and $u_\infty \in \left(0, H_\infty^{(2)}(r, \rho)\right)$ such that *(ii)*, *(iii)*, *(iv)* and *(v)* of Proposition 4.1 hold. Recall the definitions (33) and (35) of $\Theta^s_{T,\delta}$ and δ_∞. By assumption $\delta_\infty(u_\infty, s)$ is finite. Let $\vartheta^\star = (\theta_1^\star, \ldots, \theta_s^\star) \in \Theta^s_{T,(2\rho_T\, \delta_\infty(u_\infty,s))\vee(2r)}$. We note $\mathcal{Q}^\star = \{\theta_i^\star, 1 \le i \le s\}$ the set of cardinal s. Let $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\theta^\star)\|_{L^q(\nu)} = 1$. Let $\alpha, \xi \in L^q(\nu, \mathbb{R}^s)$. We define the function $P_{\alpha,\xi}$ on $\mathcal{Z}$ as:

$$P_{\alpha,\xi}(z) = \sum_{k=1}^{s} \alpha_k(z)\phi_T(\theta_k^\star) + \sum_{k=1}^{s} \xi_k(z)\, \tilde{D}_{1,T}[\phi_T](\theta_k^\star), \tag{75}$$

which belongs to H_T. Recall the definition (14) of the kernel $\mathcal{K}_T$. Using (15), we define the corresponding certificate function on $\mathcal{Z} \times \Theta$ by:

$$\eta_{\alpha,\xi}(z, \theta) = \langle \phi_T(\theta), P_{\alpha,\xi}(z)\rangle_T = \sum_{k=1}^{s} \alpha_k(z)\, \mathcal{K}_T(\theta, \theta_k^\star) + \sum_{k=1}^{s} \xi_k(z)\, \mathcal{K}_T^{[0,1]}(\theta, \theta_k^\star). \tag{76}$$

Notice that the function η is twice continuously differentiable on Θ with respect to its second variable θ due to Assumption 2.1. By Assumption 2.2 on the regularity of φ_T and the positivity of g_T and (15), we get that for almost every $z \in \mathcal{Z}$ the function $\theta \mapsto \eta_{\alpha,\xi}(z, \theta)$ is of class $\mathcal{C}^3$ on Θ, and that:

$$\tilde{D}_{1;T}[\eta_{\alpha,\xi}(z)](\theta) = \sum_{k=1}^{s} \alpha_k(z)\, \mathcal{K}_T^{[1,0]}(\theta, \theta_k^\star) + \sum_{k=1}^{s} \xi_k(z)\, \mathcal{K}_T^{[1,1]}(\theta, \theta_k^\star). \tag{77}$$

We give a preliminary technical lemma. Set:

$$\Gamma = \begin{pmatrix} \Gamma^{[0,0]} & \Gamma^{[1,0]\top} \\ \Gamma^{[1,0]} & \Gamma^{[1,1]} \end{pmatrix}, \qquad \text{for } \Gamma^{[i,j]} = \mathcal{K}_T^{[i,j]}(\vartheta^\star). \tag{78}$$

As we have $\mathcal{V}(T) \le \inf_{\Theta_\infty} g_\infty$, by Lemma 7.3 of [14] we have that:

$$\Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)} \subseteq \Theta^s_{T,\delta_T(u_T(s),s)} \tag{79}$$

where $u_T(s) = u_\infty + (s-1)\mathcal{V}_1(T)$. Hence we have:

$$(\theta_i^\star, 1 \leq i \leq s) \in \Theta^s_{T,\delta_T(u_T(s),s)}. \tag{80}$$

We deduce from (35), (36), (80) and Lemma 8.1 that:

$$\left\| I - \Gamma^{[0,0]} \right\|_{\text{op},*} \leq u_T(s), \quad \left\| I - \Gamma^{[1,1]} \right\|_{\text{op},*} \leq u_T(s), \quad \left\| \Gamma^{[1,0]} \right\|_{\text{op},*} \leq u_T(s) \quad \text{and} \quad \left\| \Gamma^{[1,0]\top} \right\|_{\text{op},*} \leq u_T(s). \tag{81}$$

We shall write for any $z \in \mathcal{Z}$:

$$\overline{V}(z) = (V(z,\theta_1^\star), \cdots, V(z,\theta_s^\star))^\top. \tag{82}$$

Lemma 8.3. *Let be $1 \leq p \leq q \leq +\infty$ such that $1/p + 1/q = 1$. Let $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ be a measurable application such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\cdot,\theta^\star)\|_{L^q(\nu)} = 1$. Assume that (81) holds. Assume also that $u_T(s) < 1/2$. Then, there exist $\alpha, \xi \in L^q(\nu, \mathbb{R}^s)$ such that:*

$$\eta_{\alpha,\xi}(z,\theta_k^\star) = V(z,\theta_k^\star) \quad \textit{for} \quad 1 \leq k \leq s, \ \textit{for } \nu - \textit{almost every } z, \tag{83}$$

$$\tilde{D}_{1,T}[\eta_{\alpha,\xi}(z)](\theta_k^\star) = 0 \quad \textit{for} \quad 1 \leq k \leq s, \ \textit{for } \nu - \textit{almost every } z, \tag{84}$$

and we have also that:

$$\|\alpha\|_{*,q} \leq \frac{1 - u_T(s)}{1 - 2u_T(s)}, \quad \|\xi\|_{*,q} \leq \frac{u_T(s)}{1 - 2u_T(s)}, \quad \left\|\alpha - \overline{V}\right\|_{*,q} \leq \frac{u_T(s)}{1 - 2u_T(s)},$$

and

$$\|\alpha\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q} \frac{1 - u_T(s)}{1 - 2u_T(s)}, \ \|\xi\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q} \frac{u_T(s)}{1 - 2u_T(s)}, \ \left\|\alpha - \overline{V}\right\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q} \frac{u_T(s)}{1 - 2u_T(s)}.$$

Remark 8.4. The construction of interpolating certificates is different from the one introduced in [28] where ν is the counting measure and $q = 2$. Indeed, in [28] the application ξ is constant and α and ξ solve (83) and $\nabla \|\eta_{\alpha,\xi}(\cdot,\theta_k^\star)\|^2_{L^2(\nu)} = 0$ for $1 \leq k \leq s$ instead of (84).

Proof. Let $z \in \mathcal{Z}$ such that (83) and (84) are satisfied. By [14, Lemma 10.1], we obtain that:

$$\alpha(z) = \Gamma_{SC}^{-1} \overline{V}(z) \quad \text{and} \quad \xi(z) = -[\Gamma^{[1,1]}]^{-1} \Gamma^{[1,0]} \Gamma_{SC}^{-1} \overline{V}(z).$$

where $\Gamma_{SC} = \Gamma^{[0,0]} - \Gamma^{[1,0]\top} [\Gamma^{[1,1]}]^{-1} \Gamma^{[1,0]}$ and:

$$\|I - \Gamma_{SC}\|_{\text{op},*} = \|I - \Gamma_{SC}\|_{\text{op},\ell_\infty} \leq \frac{u_T(s)}{1 - u_T(s)}, \quad \|\Gamma_{SC}^{-1}\|_{\text{op},*} = \|\Gamma_{SC}^{-1}\|_{\text{op},\ell_\infty} \leq \frac{1 - u_T(s)}{1 - 2u_T(s)}. \tag{85}$$

We recall that if M is a matrix such that, $\|I - M\|_{\text{op},*} < 1$, then M is non-singular, $M^{-1} = \sum_{i \geq 0} (I - M)^i$ and $\|M^{-1}\|_{\text{op},*} \leq \Big(1 - \|I - M\|_{\text{op},*}\Big)^{-1}$. Using (81), (85), the fact that $\left\|\overline{V}\right\|_{*,q} = 1$ and Lemma 8.2, we get:

$$\|\alpha\|_{*,q} \leq \|\Gamma_{SC}^{-1}\|_{\text{op},*} \|\overline{V}\|_{*,q} \leq \frac{1 - u_T(s)}{1 - 2u_T(s)},$$

$$\|\xi\|_{*,q} \leq \left\| [\Gamma^{[1,1]}]^{-1} \Gamma^{[1,0]} \Gamma_{SC}^{-1} \right\|_{\text{op},*} \|\overline{V}\|_{*,q} \leq \left\| [\Gamma^{[1,1]}]^{-1} \right\|_{\text{op},*} \left\| \Gamma^{[1,0]} \right\|_{\text{op},*} \|\Gamma_{SC}^{-1}\|_{\text{op},*} \leq \frac{u_T(s)}{1 - 2u_T(s)},$$

$$\left\|\alpha - \overline{V}\right\|_{*,q} \leq \left\|(\Gamma_{SC}^{-1} - I)\right\|_{\text{op},*} \left\|\overline{V}\right\|_{*,q} \leq \|\Gamma_{SC} - I\|_{\text{op},*} \|\Gamma_{SC}^{-1}\|_{\text{op},*} \leq \frac{u_T(s)}{1 - 2u_T(s)}.$$

Then use that for any $f \in L^q(\nu)$, we have

$$\|f\|_{L^p(\nu)} \leq \nu(\mathcal{Z})^{1/p-1/q} \|f\|_{L^q(\nu)} \tag{86}$$

by Hölder's inequality as $p \leq q$, to obtain the upper bound on the norm $\|\cdot\|_{*,p}$. This finishes the proof. □

We fix $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$ we have $\|V(\theta^\star)\|_{L^q(\nu)} = 1$ and we consider $P_{\alpha,\xi}$ and $\eta_{\alpha,\xi}$ with α and ξ characterized by (83) and (84) from Lemma 8.3. Let $e_\ell \in \mathbb{R}^s$ be the vector with all the entries equal to zero but the ℓ-th which is equal to 1.

Proof of *(iii)* **from Assumption 4.1** with $C_F = \varepsilon_\infty(r/\rho)/10$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta, \mathcal{Q}^\star) > r$ (far region). It is enough to prove that $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - C_F$. Let $\theta^\star_\ell$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. Since $\vartheta^\star \in \Theta^s_{T,2\rho_T\delta_\infty(u_\infty,s)}$, we have, by the triangle inequality that for any $k \neq \ell$:

$$2\rho_T\,\delta_\infty(u_\infty, s) < \mathfrak{d}_T(\theta^\star_\ell, \theta^\star_k) \leq \mathfrak{d}_T(\theta^\star_\ell, \theta) + \mathfrak{d}_T(\theta, \theta^\star_k) \leq 2\mathfrak{d}_T(\theta, \theta^\star_k).$$

Hence, we have $\vartheta^\star_{\ell,\theta} \in \Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)}$, where $\vartheta^\star_{\ell,\theta}$ denotes the vector $\vartheta^\star$ whose ℓ-th coordinate has been replaced by θ. Then, we obtain from Lemma 7.3 of [14] that $\Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)} \subseteq \Theta^s_{T,\delta_T(u_T(s),s)}$ and thus:

$$\vartheta^\star_{\ell,\theta} \in \Theta^s_{T,\delta_T(u_T(s),s)}. \tag{87}$$

We denote by $\Gamma_{\ell,\theta}$ (resp. $\Gamma^{[i,j]}_{\ell,\theta}$) the matrix Γ (resp. $\Gamma^{[i,j]}$) in (78) where $\vartheta^\star$ has been replaced by $\vartheta^\star_{\ell,\theta}$. Notice the upper bounds (81) also hold for $\Gamma_{\ell,\theta}$ because of (87). Recall we have that for any $\theta \in \Theta$, $\mathcal{K}_T(\theta,\theta) = 1$ and $\mathcal{K}^{[0,1]}_T(\theta,\theta) = 0$. Elementary calculations give with $\eta_{\alpha,\xi}$ from Lemma 8.3 that:

$$\eta_{\alpha,\xi}(z,\theta) = e_\ell^\top \left(\Gamma^{[0,0]}_{\ell,\theta} - I\right)\alpha(z) + \mathcal{K}_T(\theta,\theta^\star_\ell)\alpha_\ell(z) + e_\ell^\top \Gamma^{[1,0]\top}_{\ell,\theta}\xi(z) + \mathcal{K}^{[0,1]}_T(\theta,\theta^\star_\ell)\xi_\ell(z). \tag{88}$$

By taking the norm $\|\cdot\|_{L^q(\nu)}$ in (88) and using the triangle inequality we get:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq \left\|\Gamma^{[0,0]}_{\ell,\theta} - I\right\|_{\text{op},*}\|\alpha\|_{*,q} + \|\alpha\|_{*,q}|\mathcal{K}_T(\theta,\theta^\star_\ell)| + \left\|\Gamma^{[1,0]\top}_{\ell,\theta}\right\|_{\text{op},*}\|\xi\|_{*,q} + |\mathcal{K}^{[0,1]}_T(\theta,\theta^\star_\ell)|\|\xi\|_{*,q}. \tag{89}$$

Since θ belongs to the "far region", we have by definition of $\varepsilon_T(r)$ given in (31) that:

$$|\mathcal{K}_T(\theta,\theta^\star_\ell)| \leq 1 - \varepsilon_T(r). \tag{90}$$

The triangle inequality and the definitions (21) of $\mathcal{V}_T$ and (18) of $L_{1,0}$ give:

$$|\mathcal{K}^{[0,1]}_T(\theta,\theta^\star_\ell)| \leq L_{1,0} + \mathcal{V}_T. \tag{91}$$

Then, using (81) (which holds for $\Gamma_{\ell,\theta}$ thanks to (87)), we get that:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - \varepsilon_T(r) + \frac{u_T(s)}{1 - 2u_T(s)}\left(2 + L_{1,0} + \mathcal{V}_T\right).$$

Notice that the function $r \mapsto \varepsilon_\infty(r)$ is increasing. Since $\rho_T \leq \rho$, we get by Lemma 7.1 of [14] that:

$$\varepsilon_T(r) \geq \varepsilon_\infty(r/\rho_T) - \mathcal{V}_T \geq \varepsilon_\infty(r/\rho) - \mathcal{V}_T. \tag{92}$$

By assumption, we have $u_T(s) \leq H^{(2)}_\infty(r,\rho) \leq 1/4$. Hence, we have $\frac{1}{1-2u_T(s)} \leq 2$. We also have $\mathcal{V}_T \leq 1/2$ as $\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho)$. Therefore, we get:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - \varepsilon_\infty(r/\rho) + \mathcal{V}_T + u_T(s)\left(5 + 2L_{1,0}\right).$$

The assumption $u_T(s) \leq H^{(2)}_\infty(r,\rho)$ gives:

$$u_T(s) \leq \frac{8}{10\,(5 + 2L_{1,0})}\varepsilon_\infty(r/\rho). \tag{93}$$

The assumption $\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho)$ gives $\mathcal{V}_T \leq \varepsilon_\infty(r/\rho)/10$. Hence, we have $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - \frac{\varepsilon_\infty(r/\rho)}{10}$. Thus, Property *(iii)* from Assumption 4.1 holds with $C_F = \varepsilon_\infty(r/\rho)/10$.

Proof of *(i)* **from Assumption 4.1** with $C_N = \nu_\infty(\rho r)/180$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta, \mathcal{Q}^\star) \leq r$. Let $\ell \in \{1, \cdots, s\}$ such that $\theta \in \mathcal{B}_T(\theta^\star_\ell, r)$ ("near region"). Thus, it is enough to prove that $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq$

$1 - C_N\, \mathfrak{d}_T(\theta^\star_\ell, \theta)^2$. This will be done by using Lemma A.3 to obtain a quadratic decay on $\eta_{\alpha,\xi}$ from a bound on its second Riemannian derivative.

Recall that the function $\eta_{\alpha,\xi}$ is twice continuously differentiable with respect to its second variable. Differentiating (77) and using that $\mathcal{K}_T^{[2,0]}(\theta,\theta) = -1$ and $\mathcal{K}_T^{[2,1]}(\theta,\theta) = 0$, we deduce that for almost every $z \in \mathcal{Z}$:

$$\tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)](\theta) = e_\ell^\top (I + \Gamma^{[2,0]}_{\ell,\theta})\alpha(z) + \mathcal{K}_T^{[2,0]}(\theta, \theta^\star_\ell) e_\ell^\top \alpha(z) + e_\ell^\top \Gamma^{[2,1]}_{\ell,\theta} \xi(z) + \mathcal{K}_T^{[2,1]}(\theta, \theta^\star_\ell) e_\ell^\top \xi(z). \tag{94}$$

We get:

$$\begin{aligned} \tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)](\theta) - V(z,\theta^\star_\ell)\mathcal{K}_T^{[2,0]}(\theta,\theta^\star_\ell) = e_\ell^\top (I + \Gamma^{[2,0]}_{\ell,\theta})\alpha(z) + \mathcal{K}_T^{[2,0]}(\theta, \theta^\star_\ell) e_\ell^\top (\alpha(z) - \overline{V}(z)) \\ + e_\ell^\top \Gamma^{[2,1]}_{\ell,\theta} \xi(z) + \mathcal{K}_T^{[2,1]}(\theta, \theta^\star_\ell) e_\ell^\top \xi(z). \end{aligned} \tag{95}$$

The triangle inequality and the definition of $\mathcal{V}_T$ give:

$$|\mathcal{K}_T^{[2,0]}(\theta,\theta^\star_\ell)| \leq L_{2,0} + \mathcal{V}_T \quad \text{and} \quad |\mathcal{K}_T^{[2,1]}(\theta,\theta^\star_\ell)| \leq L_{2,1} + \mathcal{V}_T, \tag{96}$$

where $L_{2,0}$ and $L_{1,2}$ are defined in (18). We deduce from (87), the definition of δ_T in (35) and (36) that:

$$\left\| I + \Gamma^{[2,0]}_{\ell,\theta} \right\|_{\mathrm{op},*} \leq u_T(s) \quad \text{and} \quad \left\| \Gamma^{[2,1]}_{\ell,\theta} \right\|_{\mathrm{op},*} \leq u_T(s). \tag{97}$$

We deduce from (95) that:

$$\begin{aligned} \left\| \tilde{D}_{2;T}[\eta_{\alpha,\xi}](\theta) - V_\ell(z)\mathcal{K}_T^{[2,0]}(\theta,\theta^\star_\ell) \right\|_{L^q(\nu)} &\leq \|\alpha\|_{*,q} \left\| I + \Gamma^{[2,0]}_{\ell,\theta} \right\|_{\mathrm{op},*} + \left\| \alpha - \overline{V} \right\|_{*,q} (L_{2,0} + \mathcal{V}_T) \\ &\quad + \|\xi\|_{*,q} \left(\left\| \Gamma^{[2,1]}_{\ell,\theta} \right\|_{\mathrm{op},*} + L_{2,1} + \mathcal{V}_T \right) \\ &\leq \frac{u_T(s)}{1 - 2u_T(s)} (1 + L_{2,0} + L_{2,1} + 2\mathcal{V}_T). \end{aligned}$$

By assumption, we have $u_T(s) \leq H^{(2)}_\infty(r,\rho) \leq 1/6$. Hence, we have $\frac{1}{1-2u_T(s)} \leq 2$. Furthermore, we have by assumption that $\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho) \leq 1/2$ and $u_T(s) \leq H^{(2)}_\infty(r,\rho)$. In particular, we have:

$$u_T(s) \leq \frac{8}{9(2L_{2,0} + 2L_{2,1} + 4)} \nu_\infty(\rho r).$$

Therefore, we obtain:

$$\left\| \tilde{D}_{2;T}[\eta_{\alpha,\xi}](\theta) - V(z,\theta^\star_\ell)\mathcal{K}_T^{[2,0]}(\theta,\theta^\star_\ell) \right\|_{L^q(\nu)} \leq \frac{8}{9} \nu_\infty(\rho r). \tag{98}$$

We now check that the hypotheses of Lemma A.3-(*ii*) hold in order to obtain a quadratic decay on $\theta \mapsto \|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)}$ from the bound (98). First recall that for almost every $z \in \mathcal{Z}$, $\theta \mapsto \eta_{\alpha,\xi}(z,\theta)$ is twice continuously differentiable and have the interpolation properties (83). By the triangle inequality and since by assumption $\mathcal{V}_T \leq L_{2,0}$, we have:

$$\sup_{\Theta_T^2} |\mathcal{K}_T^{[2,0]}| \leq L_{2,0} + \mathcal{V}_T \leq 2L_{2,0}.$$

Then, Lemma 7.1 of [14] ensures that for any θ, θ' in Θ_T such that $\mathfrak{d}_T(\theta,\theta') \leq r$ we have:

$$-\mathcal{K}_T^{[2,0]}(\theta,\theta') \geq \nu_\infty(r\rho_T) - \mathcal{V}_T \geq \nu_\infty(\rho r) - \mathcal{V}_T \geq \frac{9}{10}\nu_\infty(\rho r),$$

where we used that the function $r \mapsto \nu_\infty(r)$ is decreasing and $\rho_T \leq \rho$ for the second inequality and that $\mathcal{V}_T \leq H_\infty^{(1)}(r,\rho) \leq \nu_\infty(\rho r)/10$ for the last inequality.

Set $\delta = \frac{8}{9}\nu_\infty(\rho r)$, $\varepsilon = \frac{9}{10}\nu_\infty(\rho r)$, $L = 2L_{2,0}$. As $r < L^{-\frac{1}{2}}$ and $\delta < \varepsilon$, we apply Lemma A.3-(ii) and get for $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - \frac{\nu_\infty(\rho r)}{180}\,\mathfrak{d}_T(\theta,\theta_\ell^\star)^2.$$

Proof of (ii) from Assumption 4.1 with $C_N' = (5L_{2,0}+L_{2,1}+4)/8$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta,\mathcal{Q}^\star) \leq r$. Let $\ell \in \{1,\cdots,s\}$ such that $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$ ("near region"). We shall prove that $\|\eta_{\alpha,\xi}(\theta) - V(\theta_\ell^\star)\|_{L^q(\nu)} \leq C_N'\,\mathfrak{d}_T(\theta_\ell^\star,\theta)^2$.

Let us consider the function $f : (z,\theta) \to \eta_{\alpha,\xi}(z,\theta) - V(z,\theta_\ell^\star)$. We will bound $\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)}$ on $\mathcal{B}_T(\theta_\ell^\star, r)$ and apply Lemma A.3-(i) on f to prove the the inequality of property (ii). Notice that for almost every $z \in \mathcal{Z}$, the map $\theta \mapsto f(z,\theta)$ is twice continuously differentiable. By construction, see (83), we have for almost every $z \in \mathcal{Z}$ that $\tilde{D}_{2;T}[f(z)] = \tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)]$, $f(z,\theta_\ell^\star) = 0$ and $\tilde{D}_{1;T}[f(z)](\theta_\ell^\star) = 0$. We deduce from (94) and the bounds (96) that:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \|\alpha\|_{*,q}\left\|I + \Gamma_{\ell,\theta}^{[2,0]}\right\|_{\mathrm{op},*} + \|\alpha\|_{*,q}(L_{2,0}+\mathcal{V}_T) + \|\xi\|_{*,q}\left\|\Gamma_{\ell,\theta}^{[2,1]}\right\|_{\mathrm{op},*} + \|\xi\|_{*,q}(L_{2,1}+\mathcal{V}_T).$$

Using (97), and the bounds on α and ξ from Lemma 8.3, we get:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \frac{1-u_T(s)}{1-2u_T(s)}(L_{2,0}+\mathcal{V}_T+u_T(s)) + \frac{u_T(s)}{1-2u_T(s)}(L_{2,1}+\mathcal{V}_T+u_T(s)).$$

Since $u_T(s) \leq H_\infty^{(2)}(r,\rho) \leq 1/6$ and $\mathcal{V}_T \leq H_\infty^{(1)}(r,\rho) \leq 1/2$, we get:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \frac{5}{4}L_{2,0} + \frac{1}{4}L_{2,1} + 1.$$

We get thanks to Lemma A.3-(i) on the function f that for any $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$:

$$\|\eta_{\alpha,\xi}(\theta) - V(\theta_\ell^\star)\|_{L^q(\nu)} \leq \frac{1}{8}\,(5L_{2,0}+L_{1,2}+4)\,\mathfrak{d}_T(\theta,\theta_\ell^\star)^2.$$

Proof of (iv) from Assumption 4.1 with $C_B = 2$. Recall the definition of $P_{\alpha,\xi}$ in (75). Elementary calculations give using the definitions of $\Gamma^{[0,0]}$ and $\Gamma^{[1,1]}$ in (78):

$$\begin{aligned}
\|P_{\alpha,\xi}\|_{L_T}^2 &\leq 2\left\|\sum_{k=1}^s \alpha_k(z)\phi_T(\theta_k^\star)\right\|_{L_T}^2 + 2\left\|\sum_{k=1}^s \xi_k(z)\,\phi_T^{[1]}(\theta_k^\star)\right\|_{L_T}^2 \\
&= 2\sum_{1\leq k,\ell\leq s} \mathcal{K}_T(\theta_k^\star,\theta_\ell^\star)\int \alpha_k(z)\alpha_\ell(z)\nu(\mathrm{d}z) + 2\sum_{1\leq k,\ell\leq s} \mathcal{K}_T^{[1,1]}(\theta_k^\star,\theta_\ell^\star)\int \xi_k(z)\xi_\ell(z)\nu(\mathrm{d}z) \\
&\leq 2\|\alpha\|_{*,q}\|\alpha\|_{*,p}\sum_{1\leq k,\ell\leq s}|\mathcal{K}_T(\theta_k^\star,\theta_\ell^\star)| + 2\|\xi\|_{*,q}\|\xi\|_{*,p}\sum_{1\leq k,\ell\leq s}|\mathcal{K}_T^{[1,1]}(\theta_k^\star,\theta_\ell^\star)| \\
&\leq 2s\,\|\alpha\|_{*,q}\|\alpha\|_{*,p}\left\|\Gamma^{[0,0]}\right\|_{\mathrm{op},*} + 2s\,\|\xi\|_{*,q}\|\xi\|_{*,p}\left\|\Gamma^{[1,1]}\right\|_{\mathrm{op},*}.
\end{aligned}$$

Using that $\|I\|_{\mathrm{op},*} = 1$ and (81), we get that:

$$\left\|\Gamma^{[0,0]}\right\|_{\mathrm{op},*} \leq 1 + u_T(s) \quad \text{and} \quad \left\|\Gamma^{[1,1]}\right\|_{\mathrm{op},*} \leq 1 + u_T(s).$$

By assumption we have $u_T(s) \leq H_\infty^{(2)}(r,\rho) \leq \frac{1}{6}$. Using (86), we deduce that:

$$\|P_{\alpha,\xi}\|_{L_T}^2 \leq 2(1+u_T(s))\frac{(1-u_T(s))^2+u_T(s)^2}{(1-2u_T(s))^2}\nu(\mathcal{Z})^{1/p-1/q}s \leq 4\, s\, \nu(\mathcal{Z})^{1/p-1/q}.$$

This gives:

$$\|P_{\alpha,\xi}\|_{L_T} \leq 2\sqrt{s}\,\nu(\mathcal{Z})^{1/2p-1/2q}. \tag{99}$$

We proved that *(i)*-*(iv)* from Assumption 4.1 stand. By assumption we also have that for all $\theta \neq \theta' \in \mathcal{Q}^\star : \mathfrak{d}_T(\theta,\theta') > 2\,r$, therefore Assumption 4.1 holds. This finishes the proof of Proposition 4.1.

8.2. **Proof of Proposition 4.2 (Construction of an interpolating derivative certificate).** This section is devoted to the proof of Proposition 4.2. We shall closely follow the proof of [14, Proposition 7.5].

Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. Recall Assumptions 2.2 (and thus 2.1 on the regularity of φ_T) and 2.3 on the regularity of the asymptotic kernel $\mathcal{K}_\infty$ are in force. Let $r > 0$ and $u'_\infty \in (0,1/6)$ such that *(iii)* and *(iv)* of Proposition 4.2 hold. Recall the definitions (33) and (35) of $\Theta_{T,\delta}^s$ and δ_∞. By assumption $\delta_\infty(u'_\infty, s)$ is finite. Let $\vartheta^\star = (\theta_1^\star,\dots,\theta_s^\star) \in \Theta^s_{T,(2\rho_T\,\delta_\infty(u'_\infty,s))\vee(2r)}$. We note $\mathcal{Q}^\star = \{\theta_i^\star,\, 1\leq i\leq s\}$ the set of cardinal s.

Let $V : \mathcal{Z}\times\mathcal{Q}^\star \to \mathbb{R}$ be such that $\|V(\theta^\star)\|_{L^q(\nu)} = 1$ for any $\theta^\star \in \mathcal{Q}^\star$. Recall the notation $\overline{V}$ defined in (82). Let $\alpha,\xi \in L^q(\nu,\mathbb{R}^s)$. We consider the real-valued function $\eta_{\alpha,\xi}$ defined on $\mathcal{Z}\times\Theta$ by (76).

Recall the definition of $\mathcal{V}_T$ from (21) and define $u'_T(s) = u'_\infty + (s-1)\mathcal{V}_T$. Thanks to (79) and (80), we get that (81) holds with $u_T(s)$ replaced by $u'_T(s)$.

Lemma 8.5. *Let be $1 \leq p \leq q \leq +\infty$ such that $1/p + 1/q = 1$. Let $V : \mathcal{Z}\times\mathcal{Q}^\star \to \mathbb{R}$ be a measurable application such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\cdot,\theta^\star)\|_{L^q(\nu)} = 1$. Assume that we have (81) with $u_T(s)$ replaced by $u'_T(s) < 1/2$. Then, there exist $\alpha,\xi \in L^q(\nu,\mathbb{R}^s)$ such that:*

$$\eta_{\alpha,\xi}(z,\theta_k^\star) = 0 \quad \text{for} \quad 1\leq k\leq s,\ \text{for } \nu-\text{almost every } z, \tag{100}$$

$$\tilde{D}_{1,T}[\eta_{\alpha,\xi}(z)](\theta_k^\star) = V(z,\theta_k^\star) \quad \text{for} \quad 1\leq k\leq s,\ \text{for } \nu-\text{almost every } z, \tag{101}$$

and we also have:

$$\|\alpha\|_{*,q} \leq \frac{u'_T(s)}{1-2u'_T(s)}, \quad \|\xi\|_{*,q} \leq \frac{1-u'_T(s)}{1-2u'_T(s)}, \tag{102}$$

and

$$\|\alpha\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q}\frac{u'_T(s)}{1-2u'_T(s)}, \quad \|\xi\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q}\frac{1-u'_T(s)}{1-2u'_T(s)}. \tag{103}$$

Proof. Let $z \in \mathcal{Z}$ such that (100) and (101) are satisfied. Using the notations from Section 8.1, we obtain by [14, Lemma 10.2] that:

$$\alpha(z) = -\Gamma_{SC}^{-1}\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\overline{V}(z) \quad \text{and} \quad \xi(z) = \left(I + [\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\Gamma_{SC}^{-1}\Gamma^{[1,0]\top}\right)[\Gamma^{[1,1]}]^{-1}\overline{V}(z).$$

Using (81), (85) and the fact that $\|\overline{V}\|_{*,q} = 1$, we readily obtain (102). We then obtain the controls (103) using (86).

□

We fix $V : \mathcal{Z}\times\mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$ we have $\|V(\theta^\star)\|_{L^q(\nu)} = 1$ and we consider $P_{\alpha,\xi}$ and $\eta_{\alpha,\xi}$ given by (75) and (76), with α and ξ given by Lemma 8.5.

Proof of *(i)* from Assumption 4.2 with $c_N = (L_{0,2} + L_{2,1} + 7)/8$. We define the function $f : (z,\theta) \mapsto \eta_{\alpha,\xi}(z,\theta) - V(z,\theta_\ell^\star)\,\mathrm{sign}(\theta-\theta_\ell^\star)\mathfrak{d}_T(\theta,\theta_\ell^\star)$ on $\mathcal{Z}\times\Theta$. To prove the Property *(i)*, we will bound $\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)}$ on Θ and apply Lemma A.3-*(i)*. Recall $\mathfrak{d}_T(\theta,\theta_\ell^\star) = |G_T(\theta) - G_T(\theta_\ell^\star)|$ with G_T a primitive of $\sqrt{g_T}$, and thus

$f(z,\theta) = \eta_{\alpha,\xi}(z,\theta) - V(z,\theta_\ell^\star)(G_T(\theta) - G_T(\theta_\ell^\star))$. We deduce that for ν-almost every $z \in \mathcal{Z}$ the function f is twice continuously differentiable with respect to its second variable on Θ; and elementary calculations give that $\tilde{D}_{2;T}[f(z)](\theta) = \tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)](\theta)$ for any $\theta \in \Theta$ and for ν-almost every $z \in \mathcal{Z}$ as $\tilde{D}_{1;T}[G_T] = 1$ and $\tilde{D}_{2;T}[G_T] = 0$.

Let $\theta \in \Theta_T$ and let $\theta_\ell^\star$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. Recall the notations $\Gamma_{\ell,\theta}$ (resp. $\Gamma_{\ell,\theta}^{[i,j]}$) and $\vartheta_{\ell,\theta}^\star$ defined after (87). Since for ν-almost every $z \in \mathcal{Z}$ we have $\tilde{D}_{2;T}[f(z)] = \tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)]$, we deduce from (94) that:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \left\|I + \Gamma_{\ell,\theta}^{[2,0]}\right\|_{\mathrm{op},*} \|\alpha\|_{*,q} + \|\alpha\|_{*,q} |\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)| + \|\xi\|_{*,q} \left\|\Gamma_{\ell,\theta}^{[2,1]}\right\|_{\mathrm{op},*} + \|\xi\|_{*,q} |\mathcal{K}_T^{[2,1]}(\theta,\theta_\ell^\star)|.$$

Notice that (87) holds with $u_T(s)$ replaced by $u'_T(s)$. Using (96) and (97) and the bounds (102) on α and ξ from Lemma 8.5, we get:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \frac{u'_T(s)}{1-2u'_T(s)}(L_{2,0} + \mathcal{V}_T + u'_T(s)) + \frac{1-u'_T(s)}{1-2u'_T(s)}(L_{2,1} + \mathcal{V}_T + u'_T(s)).$$

By assumption, we have $u'_T(s) \leq 1/6$ and $\mathcal{V}_T \leq 1$. Hence, we obtain:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \frac{1}{4}L_{2,0} + \frac{5}{4}L_{2,1} + \frac{7}{4}.$$

Since we have for almost every $z \in \mathcal{Z}$, $f(z,\theta_\ell^\star) = 0$ and $\tilde{D}_{1;T}[f(z)](\theta_\ell^\star) = \tilde{D}_{1;T}[\eta_{\alpha,\xi}(z)](\theta_\ell^\star) - V(z,\theta_\ell^\star) = 0$, using Lemma A.3 *(i)*, we get, with $c_N = (L_{2,0} + 5L_{2,1} + 7)/8$:

$$\|\eta_{\alpha,\xi}(\theta) - V(\theta_\ell^\star)\,\mathrm{sign}(\theta - \theta_\ell^\star)\,\mathfrak{d}_T(\theta,\theta_\ell^\star)\|_{L^q(\nu)} = \|f(\theta)\|_{L^q(\nu)} \leq c_N\,\mathfrak{d}_T(\theta,\theta_\ell^\star)^2.$$

Proof of *(ii)* from Assumption 4.2 with $c_F = (5L_{1,0} + 7)/4$. Let $\theta \in \Theta_T$, we shall prove that $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq c_F$. Let $\theta_\ell^\star$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. We deduce from (89) on the upper bound of $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)}$, using (81), the inequality from (16), (91) and the bounds (102) on α and ξ from Lemma 8.5 that:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq \frac{u'_T(s)}{1-2u'_T(s)}\,(1 + u'_T(s)) + \frac{1-u'_T(s)}{1-2u'_T(s)}\,(L_{1,0} + \mathcal{V}_T + u'_T(s))\,.$$

Since $u'_T(s) \leq 1/6$ and $\mathcal{V}_T \leq 1$, we obtain:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq \frac{5}{4}L_{1,0} + \frac{7}{4}.$$

Proof of *(iii)* from Assumption 4.2 with $c_B = 2$. Using very similar arguments as in the proof of (99) (taking care that the upper bound of the norms $\|\cdot\|_{*,q}$ and $\|\cdot\|_{*,p}$ of α and ξ are given by (102) and (103)) we also get $\|P_{\alpha,\xi}\|_{L_T} \leq 2\sqrt{s}\,\nu(\mathcal{Z})^{1/2p-1/2q}$.

We proved that *(i)*-*(ii)* from Assumption 4.2 stand for any $\theta \in \Theta_T$. Hence Assumption 4.2 holds for any positive r such that for all $\theta \neq \theta' \in \mathcal{Q}^\star : \mathfrak{d}_T(\theta,\theta') > 2\,r$. This finishes the proof of Proposition 4.2.

References

[1] Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis*. Springer, Berlin, third edition, 2006. A hitchhiker's guide. 34

[2] Jean-Marc Azaïs and Mario Wschebor. *Level sets and extrema of random processes and fields*. John Wiley & Sons, Inc., Hoboken, NJ, 2009. 9

[3] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008. 5

[4] Rina Foygel Barber, Matthew Reimherr, and Thomas Schill. The function-on-scalar LASSO with applications to longitudinal GWAS. *Electron. J. Stat.*, 11(1):1351–1389, 2017. 3

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009. 4

[6] Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Trans. Signal Process.*, 61(23):5987–5999, 2013. 4

[7] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. 3, 10

[8] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM J. Optim.*, 27(2):616–639, 2017. 4

[9] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric de Gournay, and Pierre Weiss. On representer theorems and convex regularization. *SIAM J. Optim.*, 29(2):1260–1281, 2019. 4

[10] Claire Boyer, Yohann De Castro, and Joseph Salmon. Adapting to unknown noise level in sparse deconvolution. *Inf. Inference*, 6(3):310–348, 2017. 4, 13

[11] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011. 33

[12] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications. 3

[13] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007. 3

[14] Cristina Butucea, Jean-François Delmas, Anne Dutfoy, and Clément Hardy. Off-the-grid learning of sparse mixtures from a continuous dictionary. 2022. 2, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 24, 25, 26, 27, 28, 30, 34

[15] Cristina Butucea, Jean-François Delmas, Anne Dutfoy, and Clément Hardy. Modeling infra-red spectra: an algorithm for an automatic and simultaneous analysis. In *In Proceedings of the 31st European Safety and Reliability Conference*, pages 3359–3366, 2021. 1, 3

[16] Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007. 3

[17] Emmanuel J. Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *J. Fourier Anal. Appl.*, 19(6):1229–1254, 2013. 4, 13, 14

[18] Emmanuel J. Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Comm. Pure Appl. Math.*, 67(6):906–956, 2014. 3, 4, 12

[19] Emmanuel J. Candès and Yaniv Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inform. Theory*, 57(11):7235–7254, 2011. 12

[20] Ch. Chesneau and M. Hebiri. Some theoretical results on the grouped variables Lasso. *Math. Methods Statist.*, 17(4):317–326, 2008. 5

[21] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, pages 1–46, 2021. 4

[22] Yohann de Castro and Fabrice Gamboa. Exact reconstruction using Beurling minimal extrapolation. *J. Math. Anal. Appl.*, 395(1):336–354, 2012. 4

[23] Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 42, 2020. 3, 4

[24] J. Diestel and J. J. Uhl, Jr. *Vector measures*. Mathematical Surveys, No. 15. American Mathematical Society, Providence, R.I., 1977. With a foreword by B. J. Pettis. 2, 33

[25] Vincent Duval. An epigraphical approach to the representer theorem. *J. Convex Anal.*, 28(3):819–836, 2021. 4

[26] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Found. Comput. Math.*, 15(5):1315–1355, 2015. 3, 4, 5, 12

[27] Vincent Duval and Gabriel Peyré. Sparse regularization on thin grids I: the Lasso. *Inverse Problems*, 33(5):055008, 29, 2017. 4

[28] Mohammad Golbabaee and Clarice Poon. An off-the-grid approach to multi-compartment magnetic resonance fingerprinting. *arXiv preprint arXiv:2011.11193*, 2020. 5, 8, 12, 26

[29] Junzhou Huang and Tong Zhang. The benefit of group sparsity. *Ann. Statist.*, 38(4):1978–2004, 2010. 5

[30] Han Liu and Jian Zhang. On the ℓ_1-ℓ_q regularized regression. *arXiv preprint arXiv:0802.1517*, 2008. 5

[31] Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011. 5, 10, 11

[32] Yuval Nardi and Alessandro Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.*, 2:605–633, 2008. 5

[33] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, 39(1):1–47, 2011. 34

[34] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*, 2021. 4, 6, 12, 14, 36

[35] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011. 4

[36] Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Near minimax line spectral estimation. *IEEE Trans. Inform. Theory*, 61(1):499–512, 2015. 4, 13

[37] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. 3, 4

[38] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009. 4

[39] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006. 5

Appendix A. Auxiliary Lemmas

In this section, we provide the proofs of the intermediate results.

A.1. **Proof of Proposition 1.3.** We prove the optimization problem (3) is well posed. Denote the objective function of (3) by $F(B, \vartheta)$, that is the penalized risk. Then, we have:

$$\inf_{B\in L^2(\nu,\mathbb{R}^K),\vartheta\in\Theta_T^K} F(B,\vartheta) \le F(0,\vartheta^\star) = \frac{1}{2\nu(\mathcal{Z})}\|Y\|_{L_T}^2.$$

By Minkowski inequality, we have that $\|\cdot\|_{L^p(\nu,\mathbb{R}^K)} \le \|\cdot\|_{\ell_1,L^p(\nu)}$. Indeed, we have for any $B \in L^2(\nu,\mathbb{R}^K)$:

$$\|B\|_{L^p(\nu,\mathbb{R}^K)} := \left\| \left(\sum_{k=1}^K B_k^2\right)^{\frac{1}{2}} \right\|_{L^p(\nu)} \le \left\| \sum_{k=1}^K |B_k| \right\|_{L^p(\nu)} \le \|B\|_{\ell_1,L^p(\nu)}.$$

Therefore, the minimization of F over B can be restricted to the centered closed ball $\mathcal{B}_0$ in $L^p(\nu,\mathbb{R}^K)$ of radius $\|Y\|_{L_T}^2/(\kappa 2\nu(\mathcal{Z}))$. We recall that the space $L^p(\nu,\mathbb{R}^K)$ is a reflexive Banach space whose dual is $L^q(\nu,\mathbb{R}^K)$ with $1/p+1/q=1$, see [24, Theorem 1 p.98]. By Kakutani Theorem, the closed balls of $L^p(\nu,\mathbb{R}^K)$ are therefore compact with respect to the weak topology, see [11, Theorem 3.17]. In particular (3) amounts to minimizing F over the compact set $\mathcal{B}_0 \times \Theta_T^K$.

We show that the objective function is lower semi-continuous (lsc). Recall that a convex strongly continuous (that is, continuous with respect to the strong topology) real valued function defined on a Banach space is weakly lsc (that is, lsc with respect to the weak topology), see [11, Corollary 3.9]. For any $B \in L^2(\nu,\mathbb{R}^K)$, we have:

$$\begin{aligned}\|B\|_{\ell_1,L^p(\nu)} &\le K^{\frac{1}{q}} \left(\sum_{k=1}^K \|B_k\|_{L^p(\nu)}^p\right)^{\frac{1}{p}} = K^{\frac{1}{q}} \left(\int \|B(z)\|_{\ell_p}^p \nu(\mathrm{d}z)\right)^{\frac{1}{p}} \\ &\le K^{\frac{1}{2}} \left(\int \|B(z)\|_{\ell_2}^p \nu(\mathrm{d}z)\right)^{\frac{1}{p}} = K^{\frac{1}{2}} \cdot \|B\|_{L^p(\nu,\mathbb{R}^K)},\end{aligned} \tag{104}$$

where we used Hölder's inequality. We deduce that the function $B \mapsto \|B\|_{\ell_1,L^p(\nu)}$ is strongly continuous. Since it is also convex, we get it is weakly lsc.

Recall the space $(L_T, \|\cdot\|_{L_T})$ is a Hilbert space, see [24, Section IV]. The function $X \mapsto \|Y - X\|_{L_T}$ defined on L_T is weakly lsc as it is strongly continuous and convex. Then, since the function $\vartheta \mapsto \Phi(\vartheta)$ is continuous, we deduce that the function $(B,\vartheta) \mapsto B\Phi(\vartheta)$ is continuous from $L^p(\nu,\mathbb{R}^K)\times\mathbb{R}^K$ to L_T with respect to the product topology of the weak topology on $L^p(\nu,\mathbb{R}^K)$ and the usual topology on $\mathbb{R}^K$. Since the composition of a continuous function by a lsc function is a lsc function, we deduce that the function

$(B,\vartheta) \mapsto \|Y - B\Phi(\vartheta)\|_{L_T}$ is lsc (with respect to product topology of the weak topology on $L^p(\nu, \mathbb{R}^K)$ and the usual topology on $\mathbb{R}^K$).

In conclusion, the objective function $(B,\vartheta) \mapsto F(B,\vartheta)$ is lsc (with respect to the product topology of the weak topology on $L^p(\nu, \mathbb{R}^K)$ and the usual topology on $\mathbb{R}^K$). Then, we conclude using that a lsc function on a compact set attains a minimum value, see [1, Theorem 2.43].

A.2. **Tail bound for suprema of χ^2 processes.** We give a tail bound for suprema of weighted χ^2 processes indexed on an interval $I \subset \mathbb{R}$.

Lemma A.1. *Let $I \subset \mathbb{R}$ be a bounded interval. Assume that $X = (X(\theta),\, \theta \in I)$ is a real centered Gaussian process with Lipschitz sample paths. Consider the process $Y = \sum_{i=1}^n X_i^2$ where $(X_i,\, 1 \leq i \leq n)$ are independent copies of X. Then, for an arbitrary $\theta_0 \in I$ and for all $u > n \sup_{\theta\in I} \mathrm{Var}(X(\theta))$, we have:*

(105)
$$\mathbb{P}\left(\sup_I Y > u\right) \leq \mathrm{e}^{-\frac{u}{\mathrm{Var}X(\theta_0)}\left(1-2\sqrt{\frac{n\mathrm{Var}X(\theta_0)}{u}}\right)} + 4\int_I \frac{\sqrt{\mathrm{Var}(X'(\theta))}}{2^{n/2}\Gamma(n/2)\sqrt{u}} \left(\frac{u}{\mathrm{Var}(X(\theta))}\right)^{(n+1)/2} \mathrm{e}^{-\frac{u}{2\mathrm{Var}(X(\theta))}}\, \mathrm{d}\theta.$$

Proof. Recall that I is a bounded interval. Hence, the process Y defined on I has Lipschitz sample paths. Then, applying Inequality (122) from [14] to the process Y and taking the expectation, we get, with $M = \sup_I Y$, $a = u > 0$, $b = u + \varepsilon$, $\varepsilon > 0$ and $x_0 = \theta_0$:

$$\int_u^{u+\varepsilon} \mathbb{P}(M \geq t)\, \mathrm{d}t \leq \varepsilon \mathbb{P}(Y(\theta_0) \geq u) + \int_I \mathbb{E}\left[|Y'(\theta)| \mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] \mathrm{d}\theta. \tag{106}$$

The random variable $Y(\theta_0)$ is a standard χ^2 variable of degree n and therefore we have by [33, Lemma 11] for $u > n\mathrm{Var}(X(\theta_0))$:

$$\mathbb{P}\left(Y(\theta_0) \geq u\right) \leq \mathrm{e}^{-\frac{u}{\mathrm{Var}(X(\theta_0))}\left(1-2\sqrt{\frac{n\mathrm{Var}(X(\theta_0))}{u}}\right)}. \tag{107}$$

Notice that (107) trivially holds if $\mathrm{Var}(X(\theta_0)) = 0$ as $u > 0$.

We now give a bound of the second term in the right-hand side of (106). Since (X_i', X_i) are independent Gaussian processes for $i = 1, \cdots, n$, we can write for a given $\theta \in I$:

$$X_i'(\theta) = \alpha_\theta X_i(\theta) + \beta_\theta G_i,$$

where $(G_i,\, 1 \leq i \ \leq n)$ are independent standard Gaussian random variables independent of the variables $(X_i(\theta),\, 1 \leq i \leq n)$ and:

$$\alpha_\theta = \frac{\mathbb{E}[X'(\theta)X(\theta)]}{\mathrm{Var}(X(\theta))} \quad \text{and} \quad \beta_\theta^2 = \mathrm{Var}(X'(\theta)) - \alpha_\theta^2 \mathrm{Var}(X(\theta)),$$

with the convention that $\alpha_\theta = 0$ if $\mathrm{Var}(X(\theta)) = 0$. Since $Y' = 2\sum_{i=1}^n X_i' X_i$ *a.e.*, we get that:

$$\mathbb{E}\left[|Y'(\theta)|\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] \leq 2|\alpha_\theta| \mathbb{E}\left[Y(\theta)\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] + 2|\beta_\theta| \mathbb{E}\left[\left|\sum_{i=1}^n X_i(\theta) G_i\right| \mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right].$$

Since the variables $(G_i,\, 1 \leq i \ \leq n)$ and $(X_i(\theta),\, 1 \leq i \leq n)$ are independent, the variable $Z = \sum_{i=1}^n X_i(\theta) G_i$ contidionally to the variables $(X_i(\theta),\, 1 \leq i \leq n)$ is a standard Gaussian random variable of variance $Y(\theta)$. This implies that:

$$\mathbb{E}\left[\left|\sum_{i=1}^n X_i(\theta) G_i\right| \mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] = \sqrt{\frac{2}{\pi}}\, \mathbb{E}\left[\sqrt{Y(\theta)}\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right].$$

We deduce that:

$$\mathbb{E}\left[|Y'(\theta)|\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] \leq 2\left(|\alpha_\theta|(u+\varepsilon)+\sqrt{\frac{2}{\pi}}\,|\beta_\theta|\sqrt{u+\varepsilon}\right)\mathbb{P}(u<Y(\theta)<u+\varepsilon),$$

The random variable $Y(\theta)$ is distributed as a χ^2 variable and has a density:

$$p_{Y(\theta)}(u)=\frac{u^{n/2-1}}{2^{n/2}\Gamma(n/2)}\left(\frac{1}{\operatorname{Var}(X(\theta))}\right)^{n/2}\mathrm{e}^{-\frac{u}{2\operatorname{Var}(X(\theta))}},$$

where by convention $p_{Y(\theta)}(u)$ is taken equal to 0 if $\operatorname{Var}(X(\theta))=0$ and where Γ denotes the gamma function.

Letting ε goes to 0 in (106), using (107), the right continuity of the cdf of M and the monotonicity of the density $p_{Y(\theta)}(u)$ of $Y(\theta)$ on $[n\operatorname{Var}X(\theta),+\infty[$, we deduce that for $u>n\sup_{\theta\in I}\operatorname{Var}(X(\theta))$:

$$\mathbb{P}(M\geq u)\leq \mathrm{e}^{-\frac{u}{\operatorname{Var}X(\theta_0)}\left(1-2\sqrt{\frac{n\operatorname{Var}X(\theta_0)}{u}}\right)}+2\int_I\left(|\alpha_\theta|u+\sqrt{\frac{2}{\pi}}\,|\beta_\theta|\sqrt{u}\right)p_{Y(\theta)}(u)\,\mathrm{d}\theta. \tag{108}$$

We now bound the second term of the right-hand side of (108) in two steps. Using that $\beta_\theta^2\leq\operatorname{Var}(X'(\theta))$, we get that:

$$\sqrt{\frac{2}{\pi}}\,|\beta_\theta|\,\sqrt{u}\,p_{Y(\theta)}(u)\leq\frac{1}{\sqrt{\pi}}\,\frac{\sqrt{\operatorname{Var}(X'(\theta))}}{2^{(n-1)/2}\Gamma(n/2)\sqrt{u}}\left(\frac{u}{\operatorname{Var}(X(\theta))}\right)^{n/2}\mathrm{e}^{-\frac{u}{2\operatorname{Var}(X(\theta))}}. \tag{109}$$

Thanks to the Cauchy-Schwarz inequality, we get $|\alpha_\theta|\leq\sqrt{\operatorname{Var}(X'(\theta))}/\sqrt{\operatorname{Var}(X(\theta))}$. We get that:

$$|\alpha_\theta|u\,p_{Y(\theta)}(u)\leq\frac{\sqrt{\operatorname{Var}(X'(\theta))}}{2^{n/2}\Gamma(n/2)\,\sqrt{u}}\left(\frac{u}{\operatorname{Var}(X(\theta))}\right)^{(n+1)/2}\mathrm{e}^{-\frac{u}{2\operatorname{Var}(X(\theta))}}. \tag{110}$$

Notice that (109) and (110) hold also if $\operatorname{Var}(X(\theta))=0$. Using that $\sqrt{\frac{2}{\pi}}+1\simeq 1.8\leq 2$ and that $u\geq\sup_{\theta\in I}\operatorname{Var}(X(\theta))$, we deduce (105) from (108), (109) and (110). □

Recall the functions f_n and g_n defined by (67).

Lemma A.2. *Let $T\in\mathbb{N}$ and $n\in\mathbb{N}^*$ be fixed. Let be $\mathcal{Z}=\{1,\cdots,n\}$. Suppose that Assumptions 2.1 and 2.2 hold. Let h be a function of class $\mathcal{C}^1$ from Θ_T to H_T, with Θ_T a sub-interval of Θ. Assume there exist finite constants C_1 and C_2 such that for all $\theta\in\Theta_T$:*

$$\|h(\theta)\|_T\leq C_1\quad\text{and}\quad\left\|\tilde{D}_{1;T}[h](\theta)\right\|_T\leq C_2. \tag{111}$$

Let $(W_T(z),z\in\mathcal{Z})$ be H_T-valued noise processes such that Assumption 3.1 holds. Let $a=(a_1,\cdots,a_n)$ be a sequence of nonnegative real numbers.
Set for any z in the set $\mathcal{Z}$ of cardinal n, $X(z)=(X(z,\theta)=\langle h(\theta),W_T(z)\rangle_T,\ \theta\in\Theta)$ and $Y=\sum_{z\in\mathcal{Z}}a_zX(z)^2$.
Then, we have for $u\geq(n+1)\|a\|_{\ell_\infty}\sigma^2\Delta_TC_1^2$:

$$\mathbb{P}\left(\sup_{\theta\in\Theta_T}Y(\theta)>u\right)\leq f_n\left(\frac{u}{\sigma^2\|a\|_{\ell_\infty}\Delta_TC_1^2}\right)+\frac{4\,C_2\,|\Theta_T|_{\mathfrak{d}_T}}{C_1\,2^{n/2}}g_n\left(\frac{u}{\sigma^2\|a\|_{\ell_\infty}\Delta_TC_1^2}\right), \tag{112}$$

where $|\Theta_T|_{\mathfrak{d}_T}$ denotes the diameter of the interval Θ_T with respect to the metric $\mathfrak{d}_T$, $\|a\|_{\ell_\infty}=\max_{z\in\mathcal{Z}}|a_z|$ and Γ denotes the classical gamma function.

Proof. First we notice that:

$$\mathbb{P}\left(\sup_{\Theta_T}Y>u\right)\leq\mathbb{P}\left(\sup_{\Theta_T}Z>u/\|a\|_{\ell_\infty}\right), \tag{113}$$

where $Z = \sum_{z\in\mathcal{Z}} X(z)^2$. We shall apply Lemma A.1 to the process Z.

Recall that the Gaussian processes $X(z)$ with $z \in \mathcal{Z}$ are independent with the same distribution as a process denoted $X = (X(\theta), \theta \in \Theta_T)$. The process X has Lipschitz sample paths on Θ_T and $X'(\theta) = \langle \partial_\theta h(\theta), w_T\rangle_T$ for *a.e.* $\theta \in \Theta_T$. By Assumption 3.1, we have for all $\theta \in \Theta_T$ and $z \in \mathcal{Z}$:

$$\operatorname{Var}(X(\theta)) \leq \sigma^2 \Delta_T \|h(\theta)\|_T^2 \leq \sigma^2 \Delta_T C_1^2. \tag{114}$$

We first consider the case where $\Theta_T = [\theta_{\min}, \theta_{\max}]$ is a compact interval with $\theta_{\min} < \theta_{\max}$. Then, according to Lemma A.1, Inequality (105) holds with Y replaced by Z for $u > n\,\sigma^2\Delta_T C_1^2$.

Notice that the function $x \mapsto x^{\frac{n+1}{2}} \mathrm{e}^{-x/2}$ is decreasing on $[n+1, +\infty)$ and that the function $x \mapsto \mathrm{e}^{-x\left(1-2\sqrt{\frac{n}{x}}\right)}$ is decreasing on $[n, +\infty)$. Then, plugging (114) in Inequality (105), we obtain for $u > (n+1)\,\sigma^2\Delta_T C_1^2$:

$$\begin{aligned}\mathbb{P}\left(\sup_{\Theta_T} Z > u\right) &\leq \mathrm{e}^{-\frac{u}{\sigma^2\Delta_T C_1^2}\left(1-2\sqrt{\frac{n\sigma^2\Delta_T C_1^2}{u}}\right)} \\ &\quad + \frac{4}{2^{n/2}\Gamma(n/2)\sqrt{u}}\left(\frac{u}{\sigma^2\Delta_T C_1^2}\right)^{(n+1)/2} \mathrm{e}^{-\frac{u}{2\sigma^2\Delta_T C_1^2}} \int_{\Theta_T} \sqrt{\operatorname{Var}(X'(\theta))}\,\mathrm{d}\theta.\end{aligned} \tag{115}$$

There exists a geodesic $\gamma : [0,1] \mapsto \Theta_T$ such that $\gamma_0 = \theta_{\min}$, $\gamma_1 = \theta_{\max}$ and $\mathfrak{d}_T(\theta_{\min}, \theta_{\max}) = \int_0^1 |\dot\gamma_t|\,\sqrt{g_T(\gamma_t)}\,\mathrm{d}t$. Hence, a change of variable gives:

$$\int_{\Theta_T} \sqrt{\operatorname{Var}(X'(\theta))}\,\mathrm{d}\theta = \int_0^1 |\dot\gamma_t| \sqrt{g_T(\gamma_t)\cdot\frac{\operatorname{Var}(X'(\gamma_t))}{g_T(\gamma_t)}}\,\mathrm{d}t. \tag{116}$$

By Assumption 3.1, we have for all $\theta \in \Theta_T$:

$$\frac{\operatorname{Var}(X'(\theta)))}{g_T(\theta)} \leq \sigma^2\Delta_T \left\|\tilde{D}_{1;T}[h](\theta)\right\|_T^2 \leq \sigma^2\Delta_T C_2^2.$$

Using this bound in (116), we get:

$$\int_{\Theta_T} \sqrt{\operatorname{Var}(X'(\theta))}\,\mathrm{d}\theta \leq C_2\,\sigma\,\sqrt{\Delta_T}\,|\Theta_T|_{\mathfrak{d}_T}, \tag{117}$$

where $|\Theta_T|_{\mathfrak{d}_T}$ is the diameter of the interval Θ_T with respect to the metric $\mathfrak{d}_T$.

Combining (115), (117) and (113), we finally obtain (112) for Θ_T a bounded closed interval. Then, use monotone convergence and the continuity of Z to get (112) for any interval Θ_T. □

A.3. **Technical lemma.** We consider functions $\eta : \mathcal{Z} \times \Theta \mapsto \mathbb{R}$ and bound the quantities $\|\eta(\theta)\|_{L^q(\nu)}$ on some regions of Θ under some assumptions on the second covariant derivative of η with respect to θ. The following Lemma extends [34, Lemma 2]. The proof is similar, as the latter covers the case where ν is a Dirac measure and $\|\cdot\|_{L^q(\nu)}$ reduces to $|\cdot|$.

Lemma A.3. *Let $q \in [1, +\infty]$. Suppose Assumption 2.2 holds. Consider a function $\eta : \mathcal{Z} \times \Theta$ twice continuously differentiable with respect to its second variable and $\theta_0 \in \Theta_T$.*

(i) Assume that for ν-almost every $z \in \mathcal{Z}$ we have $\eta(z, \theta_0) = 0$ and $\tilde{D}_{1;T}[\eta(z)](\theta_0) = 0$, and that there exist $\delta > 0$ and $r > 0$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$ we have:

$$\left\|\tilde{D}_{2;T}[\eta](\theta)\right\|_{L^q(\nu)} \leq \delta. \tag{118}$$

Then, we have $\|\eta(\theta)\|_{L^q(\nu)} < (\delta/2)\, \mathfrak{d}_T(\theta, \theta_0)^2$*, for any* $\theta \in \mathcal{B}_T(\theta_0, r)$*.*

(ii) Assume now that for ν*-almost every* $z \in \mathcal{Z}$*,* $\eta(z, \theta_0) = V(z)$ *and* $\tilde{D}_{1;T}[\eta(z)](\theta_0) = 0$ *where* $V \in L^q(\nu)$ *with* $\|V\|_{L^q(\nu)} = 1$*. Assume there exists a finite positive constant* L *such that* $\sup_{\theta_0, \theta \in \Theta_T} |\mathcal{K}_T^{[0,2]}(\theta_0, \theta)| \leq L$ *and there exist* $\varepsilon > 0$ *and* $r \in (0, L^{-\frac{1}{2}})$ *such that for any* $\theta \in \mathcal{B}_T(\theta_0, r)$*,* $-\mathcal{K}_T^{[0,2]}(\theta_0, \theta) \geq \varepsilon$*. Suppose that for any* $\theta \in \mathcal{B}_T(\theta_0, r)$ *and* $\delta < \varepsilon$*:*

$$\left\| \tilde{D}_{2;T}[\eta](\theta) - V \mathcal{K}_T^{[0,2]}(\theta_0, \theta) \right\|_{L^q(\nu)} \leq \delta. \tag{119}$$

Then, we have $\|\eta(\theta)\|_{L^q(\nu)} \leq 1 - \frac{(\varepsilon - \delta)}{2} \mathfrak{d}_T(\theta, \theta_0)^2$*, for any* $\theta \in \mathcal{B}_T(\theta_0, r)$*.*

Cristina Butucea, CREST, ENSAE, IP Paris
Email address: cristina.butucea@ensae.fr

Jean-François Delmas, CERMICS, École des Ponts, France
Email address: jean-francois.delmas@enpc.fr

Anne Dutfoy, EDF R&D, Palaiseau, France
Email address: anne.dutfoy@edf.fr

Clément Hardy, CERMICS École des Ponts and EDF R&D Palaiseau, France
Email address: clement.hardy@enpc.fr