Simultaneous off-the-grid learning of mixtures issued from a continuous dictionary

Cristina Butucea, Jean-François Delmas, Anne Dutfoy, Clément Hardy

Submitted to Bernoulli

Simultaneous off-the-grid learning of mixtures issued from a continuous dictionary

CRISTINA BUTUCEA[1,a], JEAN-FRANÇOIS DELMAS[2,b], ANNE DUTFOY[3,d] and CLÉMENT HARDY[2,3,c]

[1]*CREST, ENSAE, IP Paris, France,* [a]*cristina.butucea@ensae.fr*
[2]*CERMICS, École des Ponts, France,* [b]*jean-francois.delmas@enpc.fr,* [c]*clement.hardy@enpc.fr*
[3]*EDF R&D, Palaiseau, France,* [d]*anne.dutfoy@edf.fr*

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 assuming that each signal has its active features included in a finite and sparse set. We formulate regularized optimization problem to estimate simultaneously the linear coefficients in the mixtures and the non-linear parameters of the features. The optimization problem is composed of a data fidelity term and a (ℓ_1, L^p)-penalty. We call its solution the Group-Nonlinear-Lasso and provide high probability bounds on the prediction error using 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. Furthermore, for $p = 2$ these prediction rates are faster than for $p = 1$ when all signals share most of the non-linear parameters.

Keywords: Continuous dictionary; group-nonlinear-lasso; 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.

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 Butucea et al. (2021).

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 H_T be a Hilbert space where the parameter $T \in \mathbb{N}$ accounts for the amount of information in the model. The Hilbert space H_T is endowed with the scalar product $\langle \cdot, \cdot \rangle_T$ and the norm $\|\cdot\|_T$.

The observations are a collection of random elements of H_T having a signal-plus-noise structure. The signal part is a mixture (linear combination) of at most K smooth features $\varphi_T(\theta)$ belonging to H_T and continuously parametrized by a real parameter $\theta \in \Theta \subseteq \mathbb{R}$. For example, consider the standardized Gaussian probability densities with mean θ or the Cauchy probability densities at location θ. We denote by $(\varphi_T(\theta), \theta \in \Theta)$ the continuous dictionary formed by all the features. 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 :

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

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 H_T.

An observation Y writes:

$$Y = \sum_{k=1}^{K} \beta_k^\star \cdot \phi_T(\theta_k^\star) + W_T \quad \text{in } H_T,$$

where the row vector $\beta^\star = (\beta_1^\star, \ldots, \beta_K^\star)$ is $s-$sparse with non-zero coordinates in the set $S^\star$ (with cardinality s) and $\theta_k^\star$ belongs to Θ for all k in $S^\star$. This can also be written as $Y = \beta^\star \cdot \Phi_T(\vartheta^\star) + W_T$ in H_T.

In this paper we consider a collection (either discrete or continuous) of such signals. We assume that all signals share s features from our continuous dictionary. We describe first the discrete case and then the continuous case. We will give a general setup including both cases after the following examples.

Example 1.1 (Discrete case). Let us assume that the process Y has been observed repeatedly n times. Thus, for i in $\{1, \ldots, n\}$, we observe:

$$Y(i) = \sum_{k=1}^{K} B_k^\star(i) \cdot \phi_T(\theta_k^\star) + W_T(i), \quad \text{in } H_T.$$

Let L_T be the set of H_T-valued square integrable functions f on $\{1, \ldots, n\}$ with:

$$\|f\|_{L_T}^2 := \frac{1}{n} \sum_{i=1}^{n} \|f(i)\|_T^2 < \infty.$$

We endow L_T with the scalar product $\langle f, g \rangle_{L_T} := \frac{1}{n} \sum_{i=1}^{n} \langle f(i), g(i) \rangle_T$, for all f, g in L_T. Thus, we obtain the Hilbert space $(L_T, \|\cdot\|_{L_T})$.

In our simultaneous analysis of the collection Y of n processes $Y(1), \ldots, Y(n)$, we assume that the matrix $B^\star$ with entries $B_{ik}^\star := B_k^\star(i)$ has $s-$sparse column structure in the sense that the set:

$$S^\star = \left\{ k \in \{1, \ldots, K\} : \frac{1}{n} \sum_{i=1}^{n} \|B_k^\star(i)\|_T^2 \neq 0 \right\}$$

has size s with $1 \leq s < K$. The model can be written:

$$Y = B^\star \cdot \Phi_T(\vartheta^\star) + W_T, \quad \text{in } L_T, \tag{1}$$

where the set $S^\star$ and its size s, the vectors $B_k^\star$ and the values $\theta_k^\star$ for k in $S^\star$ are unknown.

This model generalizes the multi-task regression model (and the Group-Lasso model) to a design matrix whose columns are not fully observed, but are issued from the continuous dictionary of features at unknown values $\theta_k^\star$ for k in $S^\star$. Note also that according to the choice of the Hilbert space H_T we get different non-linear regression models. For example, if H_T is $\mathbb{R}^m$ with its Euclidean norm we get a non-linear matrix regression model with unknown linear parameter $B^\star$ and unknown non-linear parameter $\vartheta^\star$ in the $n \times m$ design matrix $\Phi_T(\vartheta^\star)$. If H_T is the space of square integrable functions on a compact measure set $\mathcal{T}$, we rewrite the model (1) as the following multivariate functional data regression model, for $i = 1, \ldots, n$ and $t \in \mathcal{T}$:

$$Y(i,t) = B^\star(i) \cdot \Phi_T(\vartheta^\star, t) + W_T(i,t).$$

We may also need for practical reasons to associate to each observed process $Y(i)$ 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. In this context, we define on the space $\mathcal{Z} = \{1, \ldots, n\}$ the measure ν and $L_T = L^2(\nu, H_T)$ is the space of H_T-valued functions f such that:

$$\|f\|_{L_T}^2 := \int_{\mathcal{Z}} \|f(i)\|_T^2 \, \mathrm{d}\nu(i) < \infty.$$

Example 1.2 (Continuous case). Let us assume now that the process Y is observed continuously at z belonging to some set $\mathcal{Z}$:

$$Y(z) = \sum_{k=1}^{K} B_k^\star(z) \cdot \phi_T(\theta_k^\star) + W_T(z), \quad \text{in } H_T,$$

where the set $S^\star$ of indices k such that $B_k^\star$ is non-zero, the values $B_k^\star(z)$ and $\theta_k^\star$ for k in $S^\star$ are unknown. Such models are known as "function-on-scalar" models, referring to regression models where the linear coefficients depend on a time or spatial continuous parameter, see Barber, Reimherr and Schill (2017).

Let $(\mathcal{Z}, \mathcal{F}, \nu)$ be any measure space such that $0 < \nu(\mathcal{Z}) < +\infty$; we can take $\mathcal{Z}$ as a compact interval of $\mathbb{R}$ and ν as the Lebesgue measure on $\mathcal{Z}$. Here, L_T denotes the set of H_T-valued square integrable functions f on $\mathcal{Z}$ with:

$$\|f\|_{L_T}^2 := \int_{\mathcal{Z}} \|f(z)\|_T^2 \, \mathrm{d}\nu(z) < \infty.$$

Again we assume that the functional linear parameters share a sparse structure: the unknown set $S^\star$, which is then given by $\{k \in \{1, \ldots, K\} : \|B_k^\star\|_{L_T}^2 \neq 0\}$, is sparse with cardinality $s \ll K$.

Hence, we generalize the "function-on-scalar" models that have many applications (*e.g.* in genomics, see Barber, Reimherr and Schill (2017)) by allowing the design matrix to be parametrized.

In all generality, let $(\mathcal{Z}, \mathcal{F}, \nu)$ be a measure space with ν a finite positive non-zero measure. We consider the space $L_T = L^2(\nu, H_T)$, the set of H_T-valued strongly measurable functions f defined on

$(\mathcal{Z}, \mathcal{F}, \nu)$ 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 :

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

The norm $\|\cdot\|_{L_T}$ is the natural norm associated with the scalar product and $(L_T, \|\cdot\|_{L_T})$ is an Hilbert space, see (Diestel and Uhl, 1977, 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 observe a random element Y of the Hilbert space L_T. We consider the model with unknown parameters $B^\star$ in $L^2(\nu, \mathbb{R}^K)$ and $\vartheta^\star$ in Θ^K:

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

Here, we assume that the mapping $B^\star : \mathcal{Z} \to \mathbb{R}^K$ is $s-$ sparse that is,

$$1 \le s < K \quad \text{with} \quad s = \mathrm{Card}(S^\star) \quad \text{and} \quad S^\star = \{k \in \{1, \dots, K\} : \left\| B_k^\star \right\|_{L^2(\nu)} \neq 0 \}.$$

The set $S^\star$ and the parameters $B^\star$ and $\vartheta^\star$ are unknown. Thus, the sparsity s is unknown, but an upper bound K on this value is supposed available. The value K is used as a maximal size of our parameters and to write the optimization problem that we solve here after in order to build estimators, but it does not appear in the rates we obtain later. In order to perform signal reconstruction, we are interested in recovering the sparse mapping $B^\star$ restricted to its support $S^\star$, that is $B^\star_{S^\star}$, and the associated parameters $\vartheta^\star_{S^\star}$ of the nonlinear parametric functions involved in the mixture model.

We remark that the model (2) is an extension of the model described in Butucea et al. (2022), as the latter amounts to taking $\mathcal{Z}$ a singleton (or ν a Dirac measure). We gain in generality by letting the measure ν be any finite positive non-zero measure on $\mathcal{Z}$, see Section 1.3 for further comments. By doing so, the observation $(Y(z), z \in \mathcal{Z})$ can be applied *e.g.* to longitudinal data and to multiple mixture models.

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

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

where for $z \mapsto B(z) = (B_1(z), \dots, B_K(z))$ 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 for any p in $[1,2]$. When $\mathcal{Z}$ is infinite and p in $(1,2]$, we may use the following result whose proof (based on the reflexivity of $L^p(\nu)$ which is not valid for $p=1$) is given in Section E.1 of the supplementary material Butucea et al. (2023).

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.*

The estimator $\hat{\vartheta}$ defined in (3) is called off-the-grid as it does not depend on any discretization scheme applied to the parameter space Θ. This approach differs from previous works in which the parameter space is discretized and the dictionary used to approximate the signals is therefore finite, see Tang, Bhaskar and Recht (2013) in this direction.

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:

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

To understand the normalization by $\nu(\mathcal{Z})$, consider the previous example of a finite collection of processes: $\mathcal{Z} = \{1,\cdots,n\}$ and ν the counting measure $\sum_{i=1}^n \delta_i$. Assume the n observations belong to the Hilbert space $H_T = L^2(\lambda)$ for some measure λ (either discrete or continuous) on the Borel sigma field of $\mathbb{R}$. In this case, the squared prediction error becomes:

$$\hat{R}_T^2 = \frac{1}{n}\sum_{i=1}^n \left\|B^\star(i)\Phi(\vartheta^\star) - \hat{B}(i)\Phi(\hat{\vartheta})\right\|_{L^2(\lambda)}^2. \tag{5}$$

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 super-resolution (Candès and Fernandez-Granda (2014)), spike deconvolution (Duval and Peyré (2015)), microscopy (Denoyelle et al. (2020)) or spectroscopy (Butucea et al. (2021)).

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 (Tibshirani (1996), Bickel, Ritov and Tsybakov (2009), Bunea, Tsybakov and Wegkamp (2007), Candes and Tao (2007), Bühlmann and van de Geer (2011) and references therein). The celebrated Lasso estimator, popularized by Tibshirani (1996) 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 Beck and Teboulle (2009) 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 van de Geer and Bühlmann (2009) for an overview of the coherence assumptions. It turns out that these rates have been proven mini-max optimal in Raskutti, Wainwright and Yu (2011). This means that one cannot find any estimator that achieves faster rates in expected value when estimating the worst possible parameters.

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.

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 de Castro and Gamboa (2012) 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 Boyer et al. (2019) and Duval (2021) 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 optimization problem over the space of measures may not have a solution which is an atomic measure. 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) and to solve a different optimization problem (3) producing an atomic measure as a solution. 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 (Denoyelle et al. (2020), Boyd, Schiebinger and Recht (2017)) or the Conic Gradient Particle Descent (Chizat (2021)). 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 (Candès and Fernandez-Granda (2014) for a specific dictionary and Duval and Peyré (2015) 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 Candès and Fernandez-Granda (2014), Candès and Fernandez-Granda (2013) in this direction. In Boyer, De Castro and Salmon (2017), 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 Bhaskar, Tang and Recht (2013) and Tang, Bhaskar and Recht (2015) 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 Butucea et al. (2022) 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 non-linear 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 Butucea et al. (2022) 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 Poon, Keriven and Peyré (2021).

1.3. Contributions

We extend the work of Butucea et al. (2022) to encompass the case of multiple (a discrete or continuous collection of) mixture models. In this prior work, we studied a method to reconstruct efficiently a single signal and illustrate it for various examples of observation spaces, dictionaries and noise settings. Here, our goal is to reconstruct more generally a set (possibly a continuum) of signals. Of course, when dealing with a finite set of signals, one could reconstruct each signal individually using the method employed in our previous work. However, we show here that the simultaneous reconstruction with

$p = 2$ outperforms individual reconstruction when all signals share most of the non-linear parameters. To obtain this enhancement, we introduce an optimization problem with a mixed-norm penalty, develop novel certificates, derive tail bound inequalities for the supremum of χ^2 processes, and substantially expand the proof presented in Butucea et al. (2022) that only covers the case where the measure ν is a Dirac distribution.

We let here ν be any finite positive non-zero measure. In the framework of multiple high dimensional linear regressions (ℓ_1, ℓ_p)-mixed 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 Yuan and Lin (2006) and that has received significant attention since then (see, Nardi and Rinaldo (2008), Bach (2008), Chesneau and Hebiri (2008), Huang and Zhang (2010)). It was shown in Lounici et al. (2011) that the reconstruction of signals via the Group-Lasso estimator outperfoms the reconstruction using the Lasso estimator when the signals share some sparsity pattern. Let us mention the work of Liu and Zhang (2008) 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 Golbabaee and Poon (2022) in which an extension of the BLasso has been proposed in order to address multiple mixture models. The authors extended the work of Duval and Peyré (2015) 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. Group-Nonlinear-Lasso vs. Group-Lasso on a grid

Our main objective is to reconstruct signals that are linear combinations of features continuously parametrized. This problem has been long handled by discretizing the parameter space Θ and using a finite dictionary to approximate the signals as suggested in Tang, Bhaskar and Recht (2013). In this way, the problem is reduced to a (high-dimensional) linear model which has been extensively studied in the literature. 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 approximating mixture lead to some issues. In particular, the number of active features in the mixture tends to be overestimated, see Duval and Peyré (2017) in the context of reconstructing a single signal. This phenomenon can also be observed in our more general multi-task setting.

To illustrate this, we conduct a short numerical experiment. We consider a scenario where we have $n = 100$ noisy signals observed at 100 equally spaced points between -10 and 10; all signals share an underlying structure that consists of two spikes with unknown locations and varying amplitudes (refer to Figure 1 for a visual representation of such a signal). In Figure 2, we compare the performance of the Group-Nonlinear-Lasso (with $p = 2$ and $K = 50$) to that of the Group-Lasso in reconstructing

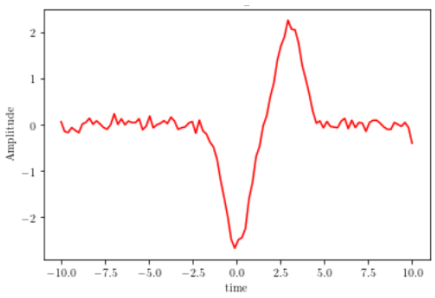

Figure 1. Signal in $H_T = \mathbb{R}^T$ with $T = 100$, mixture of two Gaussian-shaped spikes with $\theta_1^\star = 0$ and $\theta_2^\star = 3$ and amplitudes in [-10,10] uniformly distributed, corrupted by i.i.d. centered Gaussian r. v. with $\sigma = 0.1$.

these signals. In the Group-Lasso approach we give three examples of regular grids of non-linear parameters with different grid steps and such that the two spikes are always located at half distance of two consecutive points on the grid.

The Group-Nonlinear-Lasso outperforms the Group-Lasso in terms of prediction error, regardless of the penalty strength. In addition, the Group-Nonlinear-Lasso accurately identifies the two spikes, while the Group-Lasso approach incorrectly detects four spikes, even when we refine the grid.

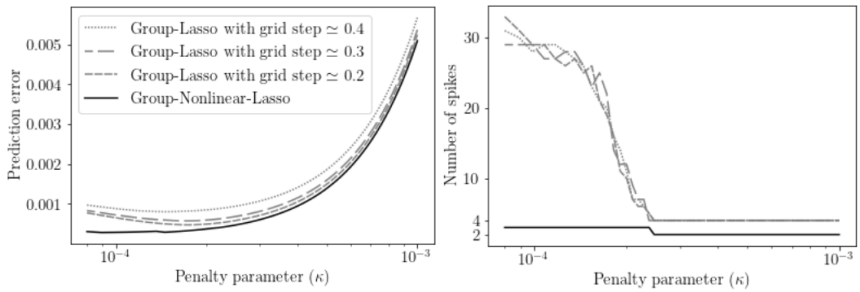

Figure 2. Prediction error $\hat{R}_T^2 = \|Y - \hat{Y}\|_{\ell_2}^2/(nT)$ given in (5), with $\hat{Y}$ denoting the reconstructed signals, and number of spikes obtained with the Group-Nonlinear-Lasso and the Group-Lasso approaches. These quantities are represented as functions of the penalty parameter κ.

All the figures contained in this section can be reproduced using the code available online at https://github.com/ClementHardy/PySFW.

1.5. Organization of the paper and notation

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.

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 $b \lesssim a$.

We write $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

2. Assumptions on the model

2.1. Regularity and non-degeneracy assumptions on the features

Let $T \in \mathbb{N}$ be a fixed parameter. 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 recall that for any function $f : \Theta \mapsto H_T$ differentiable at $\theta \in \Theta$, we have for all $g \in H_T$ that:

$$\partial_\theta \langle f(\theta), g \rangle_T = \langle \partial_\theta f(\theta), g \rangle_T \,.$$

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

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

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 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:

$$g_T(\theta) = \|\partial_\theta \phi_T(\theta)\|_T^2 \,. \tag{6}$$

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 (Butucea et al., 2022, 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 Poon, Keriven and Peyré (2021). This metric will be in particular invariant to a reparametrization of the parameter space.

2.2.1. Kernel space and associated Riemannian metric

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

$$g_{\mathcal{K}}(\theta) = \partial^2_{x,y}\mathcal{K}(\theta,\theta) \tag{7}$$

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:

$$\mathfrak{d}_{\mathcal{K}}(\theta,\theta') = |G_{\mathcal{K}}(\theta) - G_{\mathcal{K}}(\theta')|, \tag{8}$$

where $G_{\mathcal{K}}$ is a primitive of $\sqrt{g_{\mathcal{K}}}$.

We need to differentiate the kernel $\mathcal{K}$ on the manifold $(\Theta, g_{\mathcal{K}})$. We 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 $C^{0,0}$ on Θ^2 if it is continuous on Θ^2, and of class $C^{i,j}$ on Θ^2, with $i,j \in \mathbb{N}$, as soon as: F is of class $C^{0,0}$, and if $i \geq 1$ then the function $\theta \mapsto F(\theta,\theta')$ is of class C^i on Θ and its derivative $\partial_x F$ is of class $C^{i-1,j}$ on Θ^2, and if $j \geq 1$ the function $\theta' \mapsto F(\theta,\theta')$ is of class C^j on Θ and its derivative $\partial_y F$ is of class $C^{i,j-1}$ on Θ^2. For a real-valued symmetric function F defined on Θ^2 of class $C^{i,j}$, 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 $C^{i\vee j}$, and $\theta,\theta' \in \Theta$:

$$D_{i+1,j;\mathcal{K}}[F](\theta,\theta') = g_{\mathcal{K}}(\theta)^{\frac{i}{2}}\partial_\theta\left(\frac{D_{i,j;\mathcal{K}}[F](\theta,\theta')}{g_{\mathcal{K}}(\theta)^{\frac{i}{2}}}\right) \quad \text{and} \quad D_{i,j;\mathcal{K}}[F](\theta,\theta') = D_{j,i;\mathcal{K}}[F](\theta',\theta). \tag{9}$$

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

$$\tilde{D}_{i,j;\mathcal{K}}[F](\theta,\theta') = \frac{D_{i,j;\mathcal{K}}[F](\theta,\theta')}{g_{\mathcal{K}}(\theta)^{i/2}\, g_{\mathcal{K}}(\theta')^{j/2}}. \tag{10}$$

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 $C^{i\vee j}$:

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

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 $C^{i\vee 1, j\vee 1}$, then we consider the real-valued function defined on Θ^2 by:

$$\mathcal{K}^{[i,j]} = \tilde{D}_{i,j;\mathcal{K}}[\mathcal{K}]. \tag{11}$$

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

$$\mathcal{K}^{[0,0]} = \mathcal{K} \quad \text{and} \quad \mathcal{K}^{[1,1]}(\theta,\theta) = 1. \tag{12}$$

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:

$$\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T = \frac{\langle \varphi_T(\theta), \varphi_T(\theta') \rangle_T}{\|\varphi_T(\theta)\|_T \|\varphi_T(\theta')\|_T}. \tag{13}$$

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

We remark that under Assumptions 2.1 and 2.2 the definitions (6) and (7) are consistent by (Butucea et al., 2022, Lemma 4.3). Furthermore, we have that the kernel $\mathcal{K}_T$ is of class $C^{3,3}$ on Θ^2 and for $i, j \in \{0, \dots, 3\}$ and for any $\theta, \theta' \in \Theta$:

$$\mathcal{K}_T^{[i,j]}(\theta, \theta') = \langle \tilde{D}_{i;T}[\phi_T](\theta), \tilde{D}_{j;T}[\phi_T](\theta') \rangle_T, \tag{14}$$

$$\sup_{\Theta^2} |\mathcal{K}_T^{[0,0]}| \leq 1, \quad \mathcal{K}_T^{[0,0]}(\theta, \theta) = 1, \quad \mathcal{K}_T^{[1,0]}(\theta, \theta) = 0, \quad \mathcal{K}_T^{[2,0]}(\theta, \theta) = -1 \quad \text{and} \quad \mathcal{K}_T^{[2,1]}(\theta, \theta) = 0. \tag{15}$$

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. We give necessary conditions that an approximating kernel $\mathcal{K}_\infty$ must verify. Then we 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 $C^{3,3}$ the function on Θ:

$$h_{\mathcal{K}}(\theta) = \mathcal{K}^{[3,3]}(\theta, \theta). \tag{16}$$

We also and simply write for a real-valued function f on Θ of class C^i:

$$f^{[i]} = \tilde{D}_{i;T}[f].$$

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

Assumption 2.3 (Necessary conditions on the asymptotic kernel $\mathcal{K}_\infty$). *The symmetric kernel $\mathcal{K}_\infty$ defined on Θ^2 is of class $C^{3,3}$, the function g_∞ defined by (7) on Θ is positive and locally bounded (as well as of class 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:*

$$m_g := \inf_{\Theta_\infty} g_\infty > 0, \quad L_3 := \sup_{\Theta_\infty} h_\infty < +\infty, \quad \text{and} \quad L_{i,j} := \sup_{\Theta_\infty^2} |\mathcal{K}_\infty^{[i,j]}| < +\infty \quad \text{for all } i, j \in \{0, 1, 2\}. \tag{17}$$

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 (8). 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:

$$\mathfrak{d}_\infty / \rho_T \leq \mathfrak{d}_T \leq \rho_T \, \mathfrak{d}_\infty, \tag{18}$$

where ρ_T is a finite positive constant defined by:

$$\rho_T = \max\left(\sup_{\Theta_T} \sqrt{\frac{g_T}{g_\infty}}, \sup_{\Theta_T} \sqrt{\frac{g_\infty}{g_T}} \right). \tag{19}$$

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

$$\mathcal{V}_T = \max(\mathcal{V}_T^{(1)}, \mathcal{V}_T^{(2)}) \quad \text{with} \quad \mathcal{V}_T^{(1)} = \max_{i,j\in\{0,1,2\}} \sup_{\Theta_T^2} |\mathcal{K}_T^{[i,j]} - \mathcal{K}_\infty^{[i,j]}| \quad \text{and} \quad \mathcal{V}_T^{(2)} = \sup_{\Theta_T} |h_T - h_\infty|. \tag{20}$$

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 is 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 stands 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 (Butucea et al., 2022, 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 those of (Butucea et al., 2022, Theorem 2.1). (In particular, they reduce to those in Butucea et al. (2022) when ν is a Dirac measure.) A construction of certificates has been proposed in Golbabaee and Poon (2022) 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 D.4. of the supplementary material. We shall give in Section 4.2 sufficient conditions for their existence. For all $z \in \mathcal{Z}$, we note $Q^\star(z)$ 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:

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

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

It turns out that we can construct such certificates provided the elements of the set $Q^\star$ defined above 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 cardinality 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 $\vartheta^\star = \left(\theta_1^\star, \cdots, \theta_K^\star\right)$ 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 $Q^\star = \{\theta_k^\star,\ k \in S^\star\}$, with $S^\star = \{k,\ \left\|B_k^\star\right\|_{L^2(\nu)} \neq 0\ \}$, satisfies Assumptions 4.1 and 4.2 with the same $r > 0$.*

Then, there exist finite positive constants C, 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:

$$\frac{1}{\nu(\mathcal{Z})}\left\|\hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star)\right\|_{L_T}^2 \leq C_0\, s\, \nu(\mathcal{Z})^{\frac{2}{p}}\, \kappa^2, \tag{21}$$

with probability larger than

$$1 - \sum_{i=0}^{2} \mathbb{P}\left(\frac{1}{\nu(\mathcal{Z})} M_i > C\,\kappa\right), \tag{22}$$

where M_i is defined by:

$$M_i = \sup_{\theta \in \Theta_T} \left\|\left\langle W_T, \phi_T^{[i]}(\theta)\right\rangle_T\right\|_{L^q(\nu)}, \quad \text{for } i = 0,1,2. \tag{23}$$

We show in Section 3.2 below that the random variables M_i can be bounded explicitly with high probability when a finite number of signals is observed and the noise (assumed Gaussian) satisfies Assumption 3.1. Giving explicit bounds in the case where an infinite number of signals (possibly a continuum) is observed is beyond the scope of this paper and could be an avenue for future work.

Remark 3.2 (On the choice of κ). We typically choose κ in (21) 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 appears neither in the upper bound on the prediction error (21) nor in the lower bound on the probability (22). Thus, it can be taken arbitrarily large. This was already the case in Butucea et al. (2022) 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 Azaïs and Wschebor (2009) for a complete overview of Rice formulae).

We will give an explicit lower bound for the probability (22). 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_i, $i \in \{0, 1, 2\}$, from Theorem 3.1, we state additional assumptions on the noise W_T. 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:*

$$\mathrm{Var}\left(\langle f, w_T \rangle_T\right) \leq \sigma^2 \, \Delta_T \, \|f\|_T^2 . \tag{24}$$

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 B of the supplementary material.

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 (8).

Corollary 3.4. *Let $T \in \mathbb{N}$. We fix $p = q = 2$. 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 $\vartheta^\star = \left(\theta_1^\star, \cdots, \theta_K^\star\right)$ 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 C_0, C_1, 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:

$$\kappa \geq C_1 \sigma \sqrt{\frac{\|a\|_{\ell_\infty} \Delta_T \, n}{\nu(\mathcal{Z})^2}} \left(1 + \sqrt{1 + \frac{\log(\tau)}{n}}\right),$$

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):

$$\frac{1}{\nu(\mathcal{Z})} \left\| \hat{B} \Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star) \right\|_{L_T}^2 \leq C_0 \, s \, \nu(\mathcal{Z}) \, \kappa^2, \tag{25}$$

with probability larger than $1 - C_2 \left(\frac{1}{\tau} + \frac{|\Theta_T|_{\mathfrak{d}_T} F(n)}{\sqrt{\tau}}\right)$ 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 $(W_T(i),\ 1 \le i \le n)$ are independent. Thus, Assumption 3.1 holds with an equality in (24) 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 Bickel, Ritov and Tsybakov (2009)). The authors of Lounici et al. (2011) showed that the prediction error associated to the Group-Lasso estimator satisfies the bound:

$$\frac{1}{nT}\sum_{i=1}^{n} \|(\hat{B}_L(i) - B^\star(i))\Phi^\star\|_{\ell_2}^2 \lesssim \frac{\sigma^2 s}{T}\left(1 + \frac{\log(K)}{n}\right), \tag{26}$$

with high probability, larger than $1 - 1/K^\gamma$ for some positive constant $\gamma > 0$ (note that the roles of T and n are reversed in their paper). Furthermore, in the case where $B^\star$ is an unknown s-sparse mapping, $\vartheta^\star$ 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 Lounici et al. (2011) are for $2s < K$:

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

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 dividing (25) by T, we obtain from Corollary 3.4 with:

$$\kappa = C_1\sigma\sqrt{\frac{1}{n}}\left(1 + \sqrt{1 + \frac{\log(\tau)}{n}}\right) \quad \text{and} \quad \tau = T^\gamma \quad \text{for some given } \gamma > 0,$$

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

$$\frac{1}{nT}\sum_{i=1}^{n} \left\|\hat{B}(i)\Phi_T(\hat{\vartheta}) - B^\star(i)\Phi_T(\vartheta^\star)\right\|_{\ell_2}^2 \lesssim \frac{\sigma^2 s}{T}\left(1 + \frac{\log(T)}{n}\right). \tag{27}$$

We identify two regimes depending on the ratio $\log(T)/n$. Indeed, when $\log(T)/n \gg 1$ the bound (27) behaves as $\frac{\sigma^2 s \log(T)}{nT}$ 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 (27) 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 C of the supplementary material.

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 $\vartheta^\star = \left(\theta_1^\star, \cdots, \theta_K^\star\right)$ 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 C_0, C_3, 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:

$$\kappa \geq C_3 \sigma \sqrt{\Delta_T \log(\tau)}/\nu(\mathcal{Z}),$$

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

$$\frac{1}{\nu(\mathcal{Z})}\left\|\hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star)\right\|_{L_T}^2 \leq C_0\, s\, \nu(\mathcal{Z})^2\, \kappa^2, \tag{28}$$

with probability larger than $1 - C_4\, n \left(\frac{|\Theta_T|_{\mathfrak{b}_T}}{\tau\sqrt{\log\tau}} \vee \frac{1}{\tau}\right)$.

Remark 3.7. When the measure ν is composed of one atom, that is $n = 1$. This result covers that of (Butucea et al., 2022, 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:

$$\kappa = C_3\sigma\sqrt{\Delta_T\ \log\tau}/n \quad \text{ and } \tau = T^{\gamma/2} \quad \text{ for some given } \quad \gamma > 1,$$

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

$$\frac{1}{nT}\sum_{i=1}^{n}\left\|\hat{B}(i)\Phi_T(\hat{\vartheta}) - B^\star(i)\Phi_T(\vartheta^\star)\right\|_{\ell_2}^2 \lesssim \frac{\sigma^2\, s\, \log(T)}{T}. \tag{29}$$

We note that this simultaneous estimation procedure gives the same predictors as estimating separately n signals according to Butucea et al. (2022), provided that the design matrix has size K large enough. Separate estimation and aggregation of the bounds give the following bound $\sigma^2\bar{s}\log(T)/T$ instead of (29), where $\bar{s}$ is the average sparsity of the n signals. The latter bound is smaller, but is of the order $\sigma^2 s\log(T)/T$ when all signals share most of the non-linear parameters.

Remark 3.9 (Comparison for $p = 2$ and $p = 1$). 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 (27) for a well chosen

tuning parameter κ. When n and T are sufficiently large, we remark that the bound (29) (obtained when $p = 1$) is larger than the bound (27) (obtained when $p = 2$) established for the estimators from Corollary 3.4 and stands with a smaller probability.

Furthermore, separate estimation of each signal as in Butucea et al. (2022) and aggregation of the bounds give the following bound $\sigma^2 \bar{s} \log(T)/T$ instead of (27) (obtained with $p = 2$), where $\bar{s}$ is the average sparsity of the n signals. Provided $\log(T)/n$ is large, this bound is always larger than (27), but the bounds are of the same order when all signals have disjoint sets of non-linear parameters.

In conclusion, the Group-Nonlinear-Lasso for $p = 2$ provides faster prediction rates than for $p = 1$ when all signals share most of the non-linear parameters.

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 Candès and Plan (2011), Candès and Fernandez-Granda (2014), Duval and Peyré (2015). Exact recovery results for the simultaneous reconstruction of signals via the Group-Nonlinear-Lasso were proved in Golbabaee and Poon (2022) using an extension of the certificates from Duval and Peyré (2015). In Poon, Keriven and Peyré (2021), 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 Butucea et al. (2022) 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 Golbabaee and Poon (2022). Let us emphasize that we use a different contruction than Golbabaee and Poon (2022), see Remark D.4 of the supplementary material.

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

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 $Q^\star$ be a subset of Θ_T containing s values. 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 $\theta, \theta' \in Q^\star \subset \Theta_T$. 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 mapping $V : \mathcal{Z} \times Q^\star \to \mathbb{R}$ such that for any $\theta^\star \in Q^\star$, $\left\|V(\cdot, \theta^\star)\right\|_{L^q(\nu)} = 1$, there exists an element $P \in L^q(\nu, H_T)$ satisfying:*

(i) For all $\theta^\star \in Q^\star$ 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* $\theta^\star \in Q^\star$ *and* $\theta \in \mathcal{B}_T(\theta^\star, r)$, *we have* $\left\|\langle \phi_T(\theta), P\rangle_T - V(\cdot,\theta^\star)\right\|_{L^q(\nu)} \leq C'_N\, \mathfrak{d}_T(\theta^\star,\theta)^2$.

(iii) *For all* θ *in* Θ_T, $\theta \notin \bigcup_{\theta^\star \in Q^\star} \mathcal{B}_T(\theta^\star, r)$ *(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 $(z,\theta) \mapsto \langle \phi_T(\theta), P(z)\rangle_T$ defined on $\mathcal{Z}\times\Theta$ where P is an element of $L^q(\nu, H_T)$ satisfying Points *(i)* – *(iv)* from 4.1.

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

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

In order to establish prediction error bounds another type of certificate functions having different interpolating properties will be needed, see Candès and Fernandez-Granda (2013), Tang, Bhaskar and Recht (2015), Boyer, De Castro and Salmon (2017) 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* $Q^\star$ *be a subset of* Θ_T *containing* s *values. 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* $\theta,\theta' \in Q^\star \subset \Theta_T$. *There exist finite positive constants* c_N, c_F, c_B *depending on* r *and* $\mathcal{K}_\infty$ *such that for any measurable mapping* $V : \mathcal{Z}\times Q^\star \to \mathbb{R}$ *such that for any* $\theta^\star \in Q^\star$, $\left\|V(\cdot,\theta^\star)\right\|_{L^q(\nu)} = 1$, *there exists an element* $Q \in L^q(\nu, H_T)$ *satisfying:*

(i) *For all* $\theta^\star \in Q^\star$ *and* $\theta \in \mathcal{B}_T(\theta^\star, r)$, *we have:*

$$\left\|\langle \phi_T(\theta), Q\rangle_T - V(\cdot,\theta^\star)\ \operatorname{sign}(\theta-\theta^\star)\, \mathfrak{d}_T(\theta,\theta^\star)\right\|_{L^q(\nu)} \leq c_N\, \mathfrak{d}_T(\theta^\star,\theta)^2.$$

(ii) *For all* θ *in* Θ_T *and* $\theta \notin \bigcup_{\theta^\star \in Q^\star} \mathcal{B}_T(\theta^\star, r)$ *(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 $\theta^\star \in Q^\star$ we deduce from Point *(i)* for ν-almost every $z \in \mathcal{Z}$:

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

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 Butucea et al. (2022).

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 (Poon, Keriven and Peyré, 2021, Theorem 2) for interpolating certificates and generalize the contruction of (Candès and Fernandez-Granda, 2013, Lemma 2.7) for interpolating derivative certificates. In (Candès and Fernandez-Granda, 2013, 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 generalize the proofs of (Butucea et al., 2022, Propositions 7.4 and 7.5) in order to cover the case where ν is a finite measure instead of a Dirac measure (*i.e.* only one signal).

We 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 $r > 0$:

$$\varepsilon_T(r) = 1 - \sup\{|\mathcal{K}_T(\theta, \theta')|; \quad \theta, \theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta', \theta) \geq r\}, \tag{30}$$

$$\nu_T(r) = -\sup\left\{\mathcal{K}_T^{[0,2]}(\theta, \theta'); \quad \theta, \theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta', \theta) \leq r\right\}. \tag{31}$$

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 (Butucea et al., 2022, 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:

$$\Theta_{T,\delta}^s = \left\{(\theta_1, \cdots, \theta_s) \in \Theta_T^s : \ \mathfrak{d}_T(\theta_\ell, \theta_k) > \delta \text{ for all distinct } k, \ell \in \{1, \ldots, s\}\right\}. \tag{32}$$

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:

$$\mathcal{K}_T^{[i,j]}(\vartheta) = \left(\mathcal{K}_T^{[i,j]}(\theta_k, \theta_\ell)\right)_{1 \leq k, \ell \leq s}. \tag{33}$$

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

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

$$\delta_T(u, s) = \inf\left\{\delta > 0 : \ A_{T,\ell_\infty}(\vartheta) \leq u, \vartheta \in \Theta_{T,\delta}^s\right\}, \tag{34}$$

where:

$$A_{T,\ell_\infty}(\vartheta) = \max\Bigg(\left\|I - \mathcal{K}_T^{[0,0]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \left\|I - \mathcal{K}_T^{[1,1]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \left\|I + \mathcal{K}_T^{[2,0]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty},$$
$$\left\|\mathcal{K}_T^{[1,0]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \left\|\mathcal{K}_T^{[0,1]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \left\|\mathcal{K}_T^{[1,2]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}\Bigg), \tag{35}$$

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

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

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

$$\begin{aligned} H_\infty^{(1)}(r,\rho) &= \frac{1}{2} \wedge L_{2,0} \wedge L_{2,1} \wedge \frac{\nu_\infty(\rho r)}{10} \wedge \frac{\varepsilon_\infty(r/\rho)}{10}, \\ H_\infty^{(2)}(r,\rho) &= \frac{1}{6} \wedge \frac{8\,\varepsilon_\infty(r/\rho)}{10(5+2L_{1,0})} \wedge \frac{8\,\nu_\infty(\rho r)}{9(2L_{2,0}+2L_{2,1}+4)}, \end{aligned} \tag{36}$$

where the constants $L_{i,j}$ are defined in (17).

We give sufficient conditions for Assumption 4.1 to hold. The proof of the following result is given in Section D.1 of the supplementary material.

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:

$$C_N = \frac{\nu_\infty(\rho r)}{180}, \quad C_N' = \frac{5}{8}L_{2,0} + \frac{1}{8}L_{2,1} + \frac{1}{2}, \quad C_B = 2 \quad \text{and} \quad C_F = \frac{\varepsilon_\infty(r/\rho)}{10} \leq 1, \tag{37}$$

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

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

We state a second result that gives sufficient conditions for Assumption 4.2 to hold. The proof is given in Section D.2 of the supplementary material.

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:

$$c_N = \frac{1}{8}L_{2,0} + \frac{5}{8}L_{2,1} + \frac{7}{8}, \quad c_B = 2 \quad \textit{and} \quad c_F = \frac{5}{4}L_{1,0} + \frac{7}{4}, \tag{38}$$

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

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

The assumptions of Proposition 4.1 (resp. 4.2) are identical to those of (Butucea et al., 2022, Proposition 7.4) (resp. (Butucea et al., 2022, 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 Butucea et al. (2022) and set $\eta(z, \theta) = \eta_z(\theta)$. The functions η_z will be coupled through interpolated values on $Q^\star$.

5. Proof of Theorem 3.1

In this section, we sketch the proof of Theorem 3.1. We extend the proof of (Butucea et al., 2022, Theorem 2.1) to the case of a finite measure ν that is not necessarily a Dirac measure. When compared to the (now) standard proofs for Group-Lasso, this proof has the major difficulty that the design matrix is not observed but parametrized by an unknown sparse large vector $\vartheta^\star$. A Taylor expansion of second order using the metric induced by the statistical model at hand is used. Moreover, the coherence assumptions are here replaced by the existence of interpolating functions, the so called certificates.

We decompose the risk over values of estimated non-linear parameters $\hat{\theta}_\ell$ which are 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 probabilistic tail bounds on the suprema of these functionals over all possible values of θ. Let us bound the squared prediction error:

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

The predicition error corresponds to the integration on $\mathcal{Z}$ of the prediction error for one signal.

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

$$\frac{1}{2\nu(\mathcal{Z})}\left\|Y - \hat{B}\Phi_T(\hat{\vartheta})\right\|_{L_T}^2 + \kappa\|\hat{B}\|_{\ell_1, L^p(\nu)} \le \frac{1}{2\nu(\mathcal{Z})}\left\|Y - B^\star\Phi_T(\vartheta^\star)\right\|_{L_T}^2 + \kappa\|B^\star\|_{\ell_1, L^p(\nu)}. \tag{39}$$

We define the linear mapping $\hat{\Upsilon}$ 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:

$$\frac{1}{2}\hat{R}_T^2 \le \frac{1}{\nu(\mathcal{Z})}\hat{\Upsilon}(W_T) + \kappa\left(\|B^\star\|_{\ell_1, L^p(\nu)} - \|\hat{B}\|_{\ell_1, L^p(\nu)}\right). \tag{40}$$

Next, we shall expand the two terms on the right-hand side of (40). Recall the subset $Q^\star = \{\theta_k^\star :\ k \in S^\star\}$ of Θ_T is the set of the active non-linear parameters of the model. In the rest of the proof, we fix $r > 0$ so that Assumptions 4.1 and 4.2 are verified for $Q^\star$ and bound them from above.

In particular, for all $k \neq k'$ in the support $S^\star = \{k, \|B_k^\star\|_{L^2(\nu)} \neq 0\}$, we have $\mathfrak{d}_T(\theta_k^\star, \theta_{k'}^\star) > 2r$.
Let us define the following sets of indices:

- $\hat{S} = \left\{\ell : \|\hat{B}_\ell\|_{L^p(\nu)} \neq 0\right\}$ 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 Riemann distance $\mathfrak{d}_T$ for the function $\phi_T(\theta)$ around the elements of $Q^\star$. 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 the notation $\phi_T^{[i]} = \tilde{D}_{i;T}[\phi_T]$ for $i \in \{0, 1, 2\}$. According to (Butucea et al., 2022, 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:

$$\begin{aligned}\hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star) = \sum_{k \in S^\star} I_{0,k}(r)\, \phi_T(\theta_k^\star) + \sum_{k \in S^\star} I_{1,k}(r)\, \phi_T^{[1]}(\theta_k^\star) + \sum_{\ell \in \tilde{S}(r)^c} \hat{B}_\ell\, \phi_T(\hat{\theta}_\ell) \\ + \sum_{k \in S^\star} \left(\sum_{\ell \in \tilde{S}_k(r)} \hat{B}_\ell\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \int_0^1 (1-s)\phi_T^{[2]}(\gamma_s^{(k\ell)})\, \mathrm{d}s \right), \quad (41)\end{aligned}$$

with

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

We note that $I_{0,k}(r)$ and $I_{1,k}(r)$ are functions of z that belong to $L^2(\nu)$. We shall omit the dependence in r and in z when there is no ambiguity. Let us moreover denote by:

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

$$I_{2,k}(r) = \sum_{\ell \in \tilde{S}_k(r)} \left\|\hat{B}_\ell\right\|_{L^p(\nu)} \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \qquad \text{and} \qquad I_2(r) = \sum_{k \in S^\star} I_{2,k}(r), \tag{42}$$

$$I_3(r) = \sum_{\ell \in \tilde{S}(r)^c} \left\|\hat{B}_\ell\right\|_{L^p(\nu)} = \left\|\hat{B}_{\tilde{S}(r)^c}\right\|_{\ell_1, L^p(\nu)}, \tag{43}$$

where $\hat{B}_{\tilde{S}(r)^c}$ denotes the restriction of the vector-valued mapping $\hat{B}$ to its components in the set of indices $\tilde{S}(r)^c$. Again, we omit the dependence in r when there is no ambiguity.

Let us bound now the difference $\left\|B^\star\right\|_{\ell_1, L^p(\nu)} - \left\|\hat{B}\right\|_{\ell_1, L^p(\nu)}$, see (40), by using Lemma A.1 of the supplementary material:

$$\begin{aligned}\left\|B^\star\right\|_{\ell_1, L^p(\nu)} - \left\|\hat{B}\right\|_{\ell_1, L^p(\nu)} &= \sum_{k \in S^\star} \left(\left\|B_k^\star\right\|_{L^p(\nu)} - \sum_{\ell \in \tilde{S}_k(r)} \left\|\hat{B}_\ell\right\|_{L^p(\nu)}\right) - \sum_{k \in \tilde{S}(r)^c} \left\|\hat{B}_k\right\|_{L^p(\nu)} \\ &\leq I_0 \leq C_N' I_2 + (1 - C_F) I_3 + |\hat{\Upsilon}(P_1)|, \end{aligned} \tag{44}$$

where the positive constants C_N', $C_F < 1$ are given in Assumption 4.1 and $P_1 \in H_T$ corresponds to the certificate P therein with V given in Lemma A.1 of the supplementary material.

We give next an upper bound for $|\hat{\Upsilon}(W_T)| = |\langle \hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star), W_T \rangle_{L_T}|$ in (40). First, we use the expansion (41) and Hölder's inequality and we get as an example for the first term:

$$\begin{aligned}|\langle \sum_{k \in S^\star} I_{0,k}(r)\, \phi_T(\theta_k^\star), W_T \rangle_{L_T}| &\leq \sum_{k \in S^\star} |\langle I_{0,k}(r)\, \phi_T(\theta_k^\star), W_T \rangle_{L_T}| \\ &\leq \sum_{k \in S^\star} \|I_{0,k}(r)\|_{L^p(\nu)} \cdot \|\langle \phi_T(\theta_k^\star), W_T \rangle_T\|_{L^q(\nu)} \\ &\leq I_0(r) \cdot \sup_{\theta \in \Theta_T} \|\langle \phi_T(\theta), W_T \rangle_T\|_{L^q(\nu)} = I_0 \cdot M_0, \end{aligned}$$

where the random variables M_i for $i \in \{0, 1, 2\}$ are defined in (23). We proceed similarly for the remaining terms to get that:

$$\begin{aligned}|\hat{\Upsilon}(W_T)| &\leq (I_0 + I_3) M_0 + I_1 M_1 + I_2\, 2^{-1} M_2 \\ &\leq (C_N' I_2 + (2 - C_F) I_3 + |\hat{\Upsilon}(P_1)|) M_0 + (c_N I_2 + c_F I_3 + |\hat{\Upsilon}(Q_0)|) M_1 + I_2\, 2^{-1} M_2, \end{aligned} \tag{45}$$

where we also applied Lemmas A.1 and A.2 of the supplementary material, with the positive constants C_N', C_F, c_N, c_F given in Assumptions 4.1 and 4.2 and $Q_0 \in L_T$ corresponds to the derivative certificate Q in Assumption 4.2 with V given in Lemma A.2 of the supplementary material. By reinjecting (44) and (45) in (40) one gets:

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

We define the events:

$$\mathcal{A}_i = \left\{\frac{1}{\nu(\mathcal{Z})} M_i \leq C\kappa\right\}, \quad \text{for } i \in \{0, 1, 2\} \quad \text{and} \quad \mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2, \tag{46}$$

where: $C = \frac{C_F}{2(2-C_F+c_F)} \wedge \frac{C_N}{2(C'_N+c_N+2^{-1})}$. Using Lemma A.5 of the supplementary material, we obtain an upper bound for the prediction error on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq C'' \kappa \left(|\hat{\Upsilon}(P_0)| + |\hat{\Upsilon}(P_1)| + |\hat{\Upsilon}(Q_0)|\right), \tag{47}$$

with $P_0 \in H_T$ corresponding to the certificate P in Assumption 4.1 with V given by (50) of the supplementary material and: $C'' = 4C' \left(1 + \frac{C'}{C_N}(2C'_N + c_N + 1) + \frac{C'}{C_F}(3 - 2C_F + c_F)\right)$ and $C' = C \vee 1$. Using the Cauchy-Schwarz inequality and the definition of $\hat{\Upsilon}$, we get that for $f \in L_T$: $|\hat{\Upsilon}(f)| \leq \hat{R}_T \sqrt{\nu(\mathcal{Z})}\, \|f\|_{L_T}$.

Using Assumption 4.1 *(iv)* for P_i with $i = 1, 2$, and Assumption 4.2 *(iii)* for Q_0, we get: $\|P_i\|_{L_T} \leq C_B \sqrt{s}\nu(\mathcal{Z})^{1/2p-1/2q}$ and $\|Q_0\|_{L_T} \leq c_B \sqrt{s}\nu(\mathcal{Z})^{1/2p-1/2q}$. Plugging this in (47), we get that on the event $\mathcal{A}$: $\hat{R}_T^2 \leq \sqrt{C_0}\, \kappa \hat{R}_T \sqrt{s}\, \nu(\mathcal{Z})^{\frac{1}{p}}$ with $C_0 = (c_B + 2C_B)^2 C''^2$. We obtain (21) on the event $\mathcal{A}$ defined in (46) whose probability writes as in (22).

Acknowledgments

This work was partially supported by the ANRT grant N°2019/1260 and the grant Investissements d'Avenir (ANR11-IDEX0003/Labex Ecodec/ANR-11-LABX-00). The authors are grateful to the associate editor and the referees for their useful comments.

Supplementary Material

All proofs not included in this text can be found in the supplementary material Butucea et al. (2023).

References

AZAÏS, J.-M. and WSCHEBOR, M. (2009). *Level Sets and Extrema of Random Processes and Fields*. John Wiley & Sons, Inc., Hoboken, NJ.

BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225.

BARBER, R. F., REIMHERR, M. and SCHILL, T. (2017). The function-on-scalar lasso with applications to longitudinal GWAS. *Electron. J. Stat.* **11** 1351–1389.

BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202.

BHASKAR, B. N., TANG, G. and RECHT, B. (2013). Atomic norm denoising with applications to line spectral estimation. *IEEE Trans. Signal Process.* **61** 5987–5999.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* **37** 1705–1732.

BOYD, N., SCHIEBINGER, G. and RECHT, B. (2017). The alternating descent conditional gradient method for sparse inverse problems. *SIAM J. Optim.* **27** 616–639.

BOYER, C., DE CASTRO, Y. and SALMON, J. (2017). Adapting to unknown noise level in sparse deconvolution. *Inf. Inference* **6** 310–348.

BOYER, C., CHAMBOLLE, A., DE CASTRO, Y., DUVAL, V., DE GOURNAY, F. and WEISS, P. (2019). On representer theorems and convex regularization. *SIAM J. Optim.* **29** 1260–1281.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data. Springer Series in Statistics*. Springer, Heidelberg Methods, theory and applications.

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the lasso. *Electron. J. Stat.* **1** 169–194.

BUTUCEA, C., DELMAS, J.-F., DUTFOY, A. and HARDY, C. (2021). Modeling infra-red spectra: an algorithm for an automatic and simultaneous analysis. In *In Proceedings of the 31st European Safety and Reliability Conference* 3359–3366.

BUTUCEA, C., DELMAS, J.-F., DUTFOY, A. and HARDY, C. (2022). Off-the-grid learning of sparse mixtures from a continuous dictionary. *arXiv preprint arXiv:2207.00171*.

BUTUCEA, C., DELMAS, J.-F., DUTFOY, A. and HARDY, C. (2023). Supplement to "Simultaneous off-the-grid learning of sparse mixtures from a continuous dictionary".

CANDÈS, E. J. and FERNANDEZ-GRANDA, C. (2013). Super-resolution from noisy data. *J. Fourier Anal. Appl.* **19** 1229–1254.

CANDÈS, E. J. and FERNANDEZ-GRANDA, C. (2014). Towards a mathematical theory of super-resolution. *Comm. Pure Appl. Math.* **67** 906–956.

CANDÈS, E. J. and PLAN, Y. (2011). A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inform. Theory* **57** 7235–7254.

CANDES, E. and TAO, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351.

CHESNEAU, C. and HEBIRI, M. (2008). Some theoretical results on the grouped variables lasso. *Math. Methods Statist.* **17** 317–326.

CHIZAT, L. (2021). Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming* 1–46.

DE CASTRO, Y. and GAMBOA, F. (2012). Exact reconstruction using beurling minimal extrapolation. *J. Math. Anal. Appl.* **395** 336–354.

DENOYELLE, Q., DUVAL, V., PEYRÉ, G. and SOUBIES, E. (2020). The sliding frank-wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems* **36** 014001, 42.

DIESTEL, J. and UHL, J. J. JR. (1977). *Vector Measures*. *Mathematical Surveys, No. 15*. American Mathematical Society, Providence, R.I. With a foreword by B. J. Pettis.

DUVAL, V. (2021). An epigraphical approach to the representer theorem. *J. Convex Anal.* **28** 819–836.

DUVAL, V. and PEYRÉ, G. (2015). Exact support recovery for sparse spikes deconvolution. *Found. Comput. Math.* **15** 1315–1355.

DUVAL, V. and PEYRÉ, G. (2017). Sparse regularization on thin grids I: the lasso. *Inverse Problems* **33** 055008, 29.

GOLBABAEE, M. and POON, C. (2022). An off-the-grid approach to multi-compartment magnetic resonance fingerprinting. *Inverse Problems* **38** Paper No. 085002, 31.

HUANG, J. and ZHANG, T. (2010). The benefit of group sparsity. *Ann. Statist.* **38** 1978–2004.

LIU, H. and ZHANG, J. (2008). On the ℓ_1-ℓ_q regularized regression. *arXiv preprint arXiv:0802.1517*.

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204.

NARDI, Y. and RINALDO, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.* **2** 605–633.

POON, C., KERIVEN, N. and PEYRÉ, G. (2021). The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994.

TANG, G., BHASKAR, B. N. and RECHT, B. (2013). Sparse recovery over continuous dictionaries-just discretize. In *2013 Asilomar Conference on Signals, Systems and Computers* 1043–1047. IEEE.

TANG, G., BHASKAR, B. N. and RECHT, B. (2015). Near minimax line spectral estimation. *IEEE Trans. Inform. Theory* **61** 499–512.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.* **3** 1360–1392.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67.