OFF-THE-GRID LEARNING OF SPARSE MIXTURES FROM A CONTINUOUS DICTIONARY

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

Abstract. We consider a general non-linear model where the signal is a finite mixture of an unknown, possibly increasing, number of features issued from a continuous dictionary parameterized by a real non-linear parameter. The signal is observed with Gaussian (possibly correlated) noise in either a continuous or a discrete setup. We propose an off-the-grid optimization method, that is, a method which does not use any discretization scheme on the parameter space, to estimate both the non-linear parameters of the features and the linear parameters of the mixture.

We use recent results on the geometry of off-the-grid methods to give minimal separation on the true underlying non-linear parameters such that interpolating certificate functions can be constructed. Using also tail bounds for suprema of Gaussian processes we bound the prediction error with high probability. Assuming that the certificate functions can be constructed, our prediction error bound is up to log −factors similar to the rates attained by the Lasso predictor in the linear regression model. We also establish convergence rates that quantify with high probability the quality of estimation for both the linear and the non-linear parameters.

Contents

Date: July 19, 2022.

1. Introduction

1.1. **Model and method.** Assume we observe a random element y of an Hilbert space and we consider a signal-plus-noise structure for the observation y, where the noise is distributed according to a centered Gaussian process. The signal is modeled as a mixture model, by a linear combination of at most K features of the form $\varphi(\theta)$ for some parameters $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}$ is an interval of parameters and φ is a smooth function defined on Θ and taking values in the Hilbert space. We denote by $(\varphi(\theta), \theta \in \Theta)$ the continuous dictionary.

In order to capture a great variety of examples, we shall assume there exists a Hilbert space H_T, endowed with the scalar product $\langle \cdot, \cdot \rangle_T$ and the norm $\|\cdot\|_T$, where T is a parameter belonging to $\mathbb{N}$, such that: the observed process y belongs to H_T; for all $\theta \in \Theta$, the feature $\varphi_T(\theta)$ (which may depend on T) belongs to H_T and is non degenerate, *i.e.* $\|\varphi_T(\theta)\|_T$ is finite and non zero; the noise process w_T, which might also depend on the parameter T is a centered Gaussian process belonging to H_T. In the next example, the parameter T is understood as an amount of information, and, for T large, the Hilbert space H_T can be seen as an approximation of a limit Hilbert space.

Example 1.1 (Observations on a regular grid). Consider a real-valued process y observed over a regular grid $t_1 < \cdots < t_T$ on $[0, 1]$, with $t_j = j/T$ and $T \in \mathbb{N}^*$, and the noise given by centered Gaussian random variables, say $G_1, \ldots, G_T$. Assuming that all the observations have the same weight amounts to considering y as an element of the Hilbert space $H_T = L^2(\lambda_T)$ of real valued functions defined in $\mathbb{R}$ and square integrable with respect to the uniform probability measure λ_T on $\{t_1, \ldots, t_T\}$: $\lambda_T = T^{-1} \sum_{j=1}^T \delta_{t_j}$, where δ_x denotes the Dirac mass at x. In this formalism, the noise $w_T \in H_T$ is given by $w_T(t) = \sum_{j=1}^T G_j \mathbf{1}_{\{t_j\}}(t)$, where $\mathbf{1}_A$ denotes the indicator function of an arbitrary set A. Now, for T large, one can approximate the measure λ_T by the Lebesgue measure on $[0, 1]$, say Leb. In various examples, it is also easier to compute the norms of the features and of their derivatives in the Hilbert space $L^2(\text{Leb})$. This amounts to seeing H_T as approximating Hilbert spaces of the fixed Hilbert space $L^2(\text{Leb})$.

Let us define the normalized function ϕ_T defined on Θ by:

$$\phi_T(\theta) = \frac{\varphi_T(\theta)}{\|\varphi_T(\theta)\|_T} \tag{1}$$

as well as the multivariate function Φ_T defined on Θ^K by:

$$\Phi_T(\vartheta) = (\phi_T(\theta_1), \ldots, \phi_T(\theta_K))^\top \quad \text{for} \quad \vartheta = (\theta_1, \cdots, \theta_K) \in \Theta^K.$$

We consider the model with unknown parameters $\beta^\star$ in $\mathbb{R}^K$ and $\vartheta^\star$ in Θ^K:

$$y = \beta^\star \Phi_T(\vartheta^\star) + w_T \quad \text{in } H_T. \tag{2}$$

We assume from now on that the unknown K dimensional vector $\beta^\star$ is sparse, *i.e* it has s non zero entries or, equivalently, $\beta^\star \in \mathcal{B}_0(s) = \left\{\beta \in \mathbb{R}^K,\ \|\beta\|_{\ell_0} = s\right\}$, where $\|\beta\|_{\ell_0}$ counts the number of non zero entries of the vector β. Let $S^\star$ be the support of $\beta^\star$:

$$S^\star = \operatorname{Supp}(\beta^\star) = \{k \in \{1, \cdots, K\},\ \beta_k^\star \neq 0\},$$

and call $s = \operatorname{Card} S^\star$ the sparsity parameter. We are interested in predicting observations and in recovering the unknown parameters. Let us denote in general by u_S the vector u in $\mathbb{R}^K$ restricted to the coordinates in S for any non-empty set $S \subseteq \{1, ..., K\}$. We estimate both the vector $\beta^\star_{S^\star}$ with unknown sparsity s and the vector $\vartheta^\star_{S^\star}$ with entries in some compact set Θ_T and containing the parameters of those functions from our continuous dictionary that appear in the mixture model. Note that when applying the same permutation on the coordinates of $\beta^\star$ and the coordinates of $\vartheta^\star$, we obtain the same model. Thus, the vectors $\beta^\star$ and $\vartheta^\star$ are defined up to such a joint permutation. Moreover, we have $\beta^\star\Phi_T(\vartheta^\star) = \beta^\star_{S^\star}\Phi_T(\vartheta^\star)_{S^\star}$, where, by definition, $\Phi_T(\vartheta^\star)_{S^\star} = \Phi_T(\vartheta^\star_{S^\star})$. Our model is linear and sparse in $\beta^\star$ but it is non-linear in $\vartheta^\star$.

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 1.1 (Admissible noise). *Let $T \in \mathbb{N}$. The noise process w_T belongs to H_T a.s., 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:*

$$\operatorname{Var}\left(\langle f, w_T\rangle_T\right) \leq \sigma^2\, \Delta_T\, \|f\|_T^2. \tag{3}$$

In our model, the parameter T may be understood as the amount of information that we have on the mixture, see Section 1.2 below. In the discrete case, see in particular Example 1.1, the amount of information grows as the frequency of the design points over which the process is observed increases; in the continuous case, it grows as the decay rate Δ_T of the noise variance decreases.

In order to recover the sparse vector $\beta^\star$ as well as the associated parameters $\vartheta^\star_{S^\star}$ (up to a permutation), we solve the following regularized optimization problem with a real tuning parameter $\kappa > 0$:

$$(\hat\beta, \hat\vartheta) \in \underset{\beta \in \mathbb{R}^K, \vartheta \in \Theta_T^K}{\operatorname{argmin}} \quad \frac{1}{2}\|y - \beta\Phi_T(\vartheta)\|_T^2 + \kappa\|\beta\|_{\ell_1}, \tag{4}$$

where the function Φ_T is assumed to be continuous and the set Θ_T on which the optimization of the non-linear parameters is performed is required to be a compact interval. Therefore the existence of at least a solution is guaranteed. The functional that we minimize in this problem is composed of a data fidelity term and a penalty term. The penalty is expressed with a ℓ_1-norm on the vector $\beta = (\beta_1, \ldots, \beta_K)$, *i.e* the sum of the absolute values of its coordinates: $\|\beta\|_{\ell_1} = \sum_{i=1}^K |\beta_i|$. This penalization is similar to that of the Lasso problem (also referred to as Basis pursuit) introduced in [48] and extensively studied since then (see [13] for a comprehensive survey). The optimization of the non-linear parameters is not performed on the whole set of parameters Θ but rather on a compact subset Θ_T indexed by the parameter T. Indeed, it may be necessary to restrict the set of parameters, *e.g.* in a finite mixture model where we consider a location parameter we can only recover those parameters within the support of the observations.

In the more general Beurling Lasso (BLasso) framework, one can rewrite the problem (4) in a measure setting. The actual solution $\left(\hat\beta = (\hat\beta_1, \ldots, \hat\beta_K), \hat\vartheta = (\hat\theta_1, \ldots, \hat\theta_K)\right)$ of (4) is then seen as the atomic measure $\hat\mu = \sum_{i=1}^K \hat\beta_i\, \delta_{\hat\theta_i}$, where the amplitudes and the locations of the Dirac masses correspond respectively to the linear coefficients in the mixture and the parameters of the features. The measure $\hat\mu$ is also a solution of the BLasso problem when the latter admits atomic solutions composed of less than K atoms. This is in particular the case in the model presented in Section 1.2.1 where $H_T = \mathbb{R}^T$ and $K \geq T$ according to [11]. However, to our knowledge, there are no such results when H_T is a general Hilbert space.

1.2. **Examples of admissible noises.** We give here various examples of discrete or continuous noise processes that satisfy our assumptions. They are frequently used in discrete regression models or continuous models like the Gaussian white noise model, see [50] or [31].

1.2.1. *Discrete model with unweighted observations.* Let $T \in \mathbb{N}^*$ and $H_T = \mathbb{R}^T$ endowed with the usual scalar product and the corresponding Euclidean norm $\|\cdot\|_T = \|\cdot\|_{\ell_2}$. If the noise w_T is a vector of T independent centered Gaussian random variables with variance σ^2, then Assumption 1.1 holds with an equality and $\Delta_T = 1$:

$$\mathrm{Var}(\langle f, w_T \rangle_T) = \sigma^2 \, \|f\|_{\ell_2}^2.$$

If w_T is a centered Gaussian vector of dimension T with each coordinate with variance σ^2, then Assumption 1.1 holds with Δ_T the spectral radius of the correlation matrix:

$$\mathrm{Var}(\langle f, w_T \rangle_T) \leq \sigma^2 \, \Delta_T \, \|f\|_{\ell_2}^2.$$

1.2.2. *Discrete model with weighted observations.* Assume the data set comes from the observations of a process y on a grid $t_1 < \cdots < t_T$ of size $T \in \mathbb{N}^*$. In this case, it might be pertinent to use a more general formalism (this is motivated by the case T large as in Example 1.1). In order to take into account possible different weights on the grid (which is legitimated when the grid is not regular), one can consider an atomic measure λ_T on $\mathbb{R}$ given by a (non-negative) linear combination of the Dirac masses on the grid, and the Hilbert space $H_T = L^2(\lambda_T)$ of real valued functions defined in $\mathbb{R}$ and square integrable with respect to the measure λ_T. We can then consider the noise given by the function $w_T = \sum_{j=1}^T G_j \mathbf{1}_{\{t_j\}}$ defined on $\mathbb{R}$, where $(G_1, \ldots, G_T)$ is a centered Gaussian vector with independent entries and common variance σ^2. In the particular case $\lambda_T(\mathrm{d}t) = \Delta_T \sum_{j=1}^T \delta_{t_j}(\mathrm{d}t)$ for some $\Delta_T > 0$ (in Example 1.1, $\Delta_T = 1/T$ and λ_T is a discrete approximation of the Lebesgue measure on $[0,1]$), we get that Assumption 1.1 holds with an equality:

$$\mathrm{Var}(\langle f, w_T \rangle_T) = \sigma^2 \Delta_T \, \|f\|_T^2.$$

In relation to the first model of Section 1.2.1, notice that in the present case $\|f\|_T = \Delta_T \|f\|_{\ell_2}$, where the right-hand side is understood as the ℓ_2 norm of the vector $(f(t_1), \ldots, f(t_T))$.

1.2.3. *Continuous model with truncated white noise or colored noise.* Consider the set $\mathcal{C} = \mathcal{C}([0,1], \mathbb{R})$ of $\mathbb{R}$-valued continuous functions defined on $[0,1]$, an orthonormal base $(\psi_k, k \in \mathbb{N})$ of $L^2 = L^2([0,1], \mathrm{Leb})$ of elements of $\mathcal{C}$, where Leb is the Lebesgue measure on $[0,1]$. We simply denote by $\langle \cdot, \cdot \rangle_{L^2}$ the corresponding scalar product. Let $p = (p_k, k \in \mathbb{N})$ be a sequence of non-negative real numbers and set $\mathrm{Supp}(p) = \{k \in \mathbb{N} : \; p_k > 0\}$ its support. Let H_T be the completion of the vector space generated by the base $(\psi_k, k \in \mathrm{Supp}(p))$ (which is also the completion of $\mathcal{C}$ if p is positive and bounded), with respect to the scalar product:

$$\langle f, g \rangle_T = \sum_{k \in \mathbb{N}} p_k \, \langle f, \psi_k \rangle_{L^2} \, \langle g, \psi_k \rangle_{L^2}.$$

Notice that the Hilbert space H_T does not depend on the parameter T unless p depends on T. Let us recall that if $p \equiv 1$, that is, the sequence p is constant equal to 1, then $H_T = L^2$.

Let $\xi = (\xi_k, k \in \mathbb{N})$ be a weight sequence of non-negative real numbers such that the sequence $p\,\xi = (p_k \, \xi_k, k \in \mathbb{N})$ is summable. Consider the noise $w_T = \sum_{k \in \mathrm{Supp}\,(p)} \sqrt{\xi_k} \, G_k \, \psi_k$, where $(G_k, k \in \mathbb{N})$ are independent centered Gaussian random variables with variance σ^2. Notice Assumption 1.1 holds as $\|w_T\|_T^2 = \sum_{k \in \mathbb{N}} p_k \, \xi_k \, G_k^2$ is a.s. finite and, with $\Delta_T = \sup_{\mathbb{N}} p\,\xi$:

$$\mathrm{Var}(\langle f, w_T \rangle_T) = \sigma^2 \sum_{k \in \mathbb{N}} p_k^2 \, \xi_k \, \langle f, \psi_k \rangle_{L^2}^2 \leq \sigma^2 \, \Delta_T \, \|f\|_T^2.$$

Notice that the noise w_T does not depend on the parameter T unless p or ξ depends on T.

The truncated white noise model corresponds to $p \equiv 1$ and $\xi = (\xi_k = \mathbf{1}_{\{k \leq T\}}, k \in \mathbb{N})$. In this case $\Delta_T = 1$ and $\|w_T\|_T^2$ is a.s. of order $\sigma^2 T$ by the strong law of large numbers. The white noise corresponds to the limit case $T = +\infty$, which does not satisfy the hypothesis as a.s. its L^2-norm is infinite. Let us mention that the bounds given in the main theorems in Section 2 rely on $\|w_T\|_T$ being finite and not on its value.

Consider again $p \equiv 1$. Thanks to the Karhunen-Loève's decomposition, the scaled Brownian motion $w_T = C_T B$, with B the Brownian motion on $[0, 1]$ and C_T a positive constant, corresponds to the base functions $\psi_k(t) = \sqrt{2} \sin((2k + 1)\pi t/2)$ for $t \in [0, 1]$ and the weights $\xi_k = 4C_T^2/(2k + 1)^2\pi^2$ for $k \in \mathbb{N}$, and $\sigma^2 = 1$. In this case, we have $\langle f, w_T \rangle_T = C_T \int_0^1 f(s)B(s)\, \mathrm{d}s$ for $f \in L^2$ and Assumption 1.1 holds with $\sigma^2 = 1$ and $\Delta_T = \sup_{\mathbb{N}} p\,\xi = 4C_T^2/\pi^2$.

1.3. **Previous work.** The model (2) in the particular case where $\vartheta^\star$ is supposed given and the observations depend linearly on a vector $\beta^\star$ has long been studied in the literature. Assume for simplicity that $H_T = \mathbb{R}^T$ is the T-dimensional Euclidean space, so that $\Phi_T \in \mathbb{R}^{K\times T}$ is a matrix whose entries are known and can be either random or deterministic, $y \in \mathbb{R}^T$ is an observed vector and $w_T \in \mathbb{R}^T$ is a vector of noise (often assumed Gaussian). Even when K is larger than T the estimation of $\beta^\star$ is still consistent provided the vector $\beta^\star$ is sparse and a null space property is verified by the matrix Φ_T, or some sufficient condition saying that the lines of Φ_T are not too colinear (see [51] for a complete overview). The Lasso estimator [48] or the Dantzig selector [15] are efficient to perfom such estimation and the quality of the estimation with respect to the dimension of the problem is now well known. The authors of [9] have given bounds for the prediction error for both estimators.

We consider here a highly non-linear extension of this model that consists in assuming that the matrix $\Phi_T = \Phi_T(\vartheta^\star)$ depends non-linearly on a parameter $\vartheta^\star$ to be estimated. In our model (2), Φ_T is composed of K row vectors belonging to a parametric family or by K features belonging to a continuous dictionary and the observed data y may be either a vector or a function. This model has proven to be relevant in many fields such as microscopy, astronomy, spectroscopy, imaging or signal processing.

When the observation y belongs to a finite-dimensional Hilbert space and the dimension K is fixed and small compared to T, the model received attention several decades ago and gave rise to separable least square problems and resolution methods such as variable projection (see [34, 33]). These papers mainly provided numerical methods but let us mention the consistency result in [35] for non-linear regression models.

On the contrary, when K is arbitrarily large many problems remain open. One of the natural ideas to estimate the underlying parameters could be to discretize the parameter space Θ and return to the study of a linear model. It would amount to considering a finite subfamily of $(\varphi(\theta), \theta \in \Theta)$ as in [46] and deal with overcomplete dictionary learning techniques (also referred to as sparse coding, see [40, 25]). In this case, sparse estimators for linear models such as the Lasso are available. However, in sparse spike deconvolution where the family $(\varphi(\theta), \theta \in \Theta)$ is a family of spikes parametrized by a location parameter, the authors of [27] have shown that in the presence of noise discretizing the space of parameters and solving a Lasso problem tends to produce clusters of spikes around the spikes one seeks to locate. That is why it is preferable to use off-the-grid methods. By off-the-grid, we mean that the methods employed do not use discretization schemes on the parameter set Θ. In [26], the authors show that in presence of a small noise, the BLasso only induces a slight perturbation of the spikes locations and amplitudes and does not produce clusters. The BLasso was introduced in [23] and has been studied in many papers since then mostly by the compressed sensing and super-resolution communities ([17], [5] among many others). It is basically an off-the-grid extension of the classical Lasso for continuous dictionary learning. The optimization problem is formulated as a convex minimization over the space of Radon measures. In the BLasso framework, the dimension K in (2) is infinite and the linear coefficients and non-linear parameters are encoded by an atomic measure made of weighted Dirac functions. By solving a minimization problem over Radon measures, the aim is to recover an atomic measure. It raises the question of whether such a solution exists. In [11] the question is answered by the affirmative when the observed data y belongs to a finite-dimensional Hilbert space H_T. When this is not the case, i.e. H_T is infinite dimensional, the question is open. In this paper, we avoid the problem by assuming a bound K on the number of functions in the mixture and restricting the space over which the BLasso is perfomed to the atomic measures with at most K atoms. The numerical methods used to solve the BLasso such as the Sliding Frank-Wolfe algorithm ([24] and [14, 32] for applications in spectroscopy and imaging), also called the alternating descent conditional gradient method (see [10]), and the conic particle gradient descent (see [21]), seek a solution directly in the space of Dirac mixtures. Hence, our formulation (4) is closer to the way algorithms proceed. Let us mention that other methods such as Orthogonal Matching Pursuit (see [28]) exist to tackle the problem of sparse learning from a continuous dictionary. Typically, the case of sparse spike deconvolution where the dictionary consists of Gaussian functions continuously parametrized by a location parameter is not included.

The study of the regression over a continuous dictionary in the framework of the BLasso has been quite specific to the dictionary considered. The literature first focused on the dictionary of complex exponential functions parametrized by their frequency ($\varphi(\theta) : t \mapsto \mathrm{e}^{i2\pi\langle t,\theta\rangle}, \theta \in \Theta$) where Θ is the d-dimensional torus (see [18]). In [12], a bound is given for the prediction error for this dictionary. The proof extends a previous result obtained in [47] for atomic norm denoising. What is particularly interesting is that the rates obtained for the prediction error almost reach the minimax rates achievable for linear models (see [42, 16]) provided that the frequencies are sufficiently separated. The separation condition between the non-linear parameters to estimate is inherent to the BLasso unless we assume the positivity of the linear parameters as in [44].

For results on a wider range of dictionaries, let us highlight the work of [26] that gives recovery and robustness to noise results for spike deconvolution. Let us also mention the recent work of [8] that generalizes some exact recovery results for a broader family of dictionaries as well as the paper [7] that gives robustness to noise guarantees for a family of shifted functions ($\varphi(\theta) = k(\cdot - \theta), \theta \in \Theta$) of a given specific function k. In a density model that is a mixture of shifted functions, [22] studies a modification of the BLasso by considering a weighted L^2 prediction error.

The case of non-translation invariant families remained for long intractable without very pessimistic separation conditions. In [41] the authors set a natural geometric framework to analyse the estimation problem. The separation condition between the parameters appears naturally in terms of a metric. In their paper, the design over which the observation are made is distributed according to a probability distribution. Their main result shows that in presence of noise the BLasso recovers a measure close to the one to be estimated with respect to a Wasserstein metric.

1.4. **Contributions.** This paper adresses the problem of learning sparse mixtures from a continuous dictionary for a wide variety of regression models within a common framework. Indeed, we tackle a wide range of possible dictionaries of sufficiently smooth features, observation schemes and Gaussian noises with various structures. The observations are supposed to belong to a Hilbert space H_T. Continuous observations over an interval of $\mathbb{R}$ as well as discrete observations at given design points are therefore included in our framework. Furthermore, the Hilbert structure and the mild assumption we make on the noise, encompass a wide range of Gaussian noises. In particular, our framework allows to take into account the case of correlated Gaussian noise processes.

One of the main results of this paper gives a high-probability bound for the prediction error:

$$\left\| \hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T,$$

where $(\hat{\beta}, \hat{\vartheta})$ is the solution of the optimization problem (4). Contrary to the BLasso optimization program over a set of measures whose result can be a diffuse measure, our formulation of the optimization problem has always a solution belonging to a finite set of values. Our prediction error bound matches (up to logarithmic factors and with high probability) that obtained in the linear case, that is when $\vartheta^\star$ is known and does not need to be estimated. We also give high-probability bounds on some loss functions comparing the estimators $\hat{\beta}$ and $\hat{\vartheta}$ given by (4) to the parameters $\beta^\star$ and $\vartheta^\star$, respectively. Our work extends results that were so far restricted to the specific case of a dictionary consisting of complex exponentials continuously parameterized by their frequencies (see [12, 47]). When the optimization problem produces a cluster of features to approximate an element of the mixture, we also show that there can be no compensation between the amplitudes of the features involved.

Following some work in compressed sensing and super-resolution ([18, 17] among others), our bounds rely on the existence of interpolating functions called "certificates" (see Assumptions 6.1 and 6.2) instead of relying on compatibility conditions or Restricted Eigenvalue conditions. We give an explicit way to construct such functions in the spirit of [41]. We show in this paper that such functions can be constructed provided the non-linear parameters belonging to Θ are well separated with respect to a Riemannian metric $\mathfrak{d}_T$ (defined in Section 4.1) associated to the kernel $\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T$. See Remark 8.2 for comments on the separation distance in the particular case of sparse spike deconvolution. The Riemannian metric appears naturally when it comes to tackle a wide variety of dictionaries. In addition, it leads to a lot of invariances in many quantities useful in the proofs. Typically, the Riemannian metrics $\mathfrak{d}_T$ and $\mathfrak{d}_T^{\mathrm{h}}$ associated respectively to the kernel $\mathcal{K}_T(\cdot, \cdot)$ and the warped kernel $\mathcal{K}_T^{\mathrm{h}} = \mathcal{K}_T(\mathrm{h}(\cdot), \mathrm{h}(\cdot))$ for some smooth enough diffeomorphism h are equal and we have $\mathfrak{d}_T(\theta, \theta') = \mathfrak{d}_T^{\mathrm{h}}(\mathrm{h}^{-1}(\theta), \mathrm{h}^{-1}(\theta'))$.

Our statistical results rely on tail bound inequalities for suprema of Gaussian processes: following [12], instead of using controls on $\|w_T\|_T$ as in the seminal works [26, 41], we used bounds, based on the noise structure from Assumption 1.1, on quantities of the form $\sup_{\Theta_T} \langle f(\theta), w_T \rangle_T$ for some H_T-valued functions f built from the dictionary $(\varphi_T(\theta), \theta \in \Theta)$ and its derivative. This approach is relevant as for some models the quantity $\|w_T\|_T$ may be very large, see for example the truncated white noise model from Section 1.2.3.

2. Main Results

Recall that we consider the model (2) that we can write in an equivalent way as, with $S^\star$ the support of the vector $\beta^\star$:

$$y = \sum_{j \in S^\star} \beta_j^\star \frac{\varphi_T(\theta_j^\star)}{\|\varphi_T(\theta_j^\star)\|_T} + w_T \quad \text{in } H_T.$$

The main theorem of this paper gives the behavior of the prediction error with respect to: the decay rate of the noise variance Δ_T, the parameter $T \in \mathbb{N}$, the sparsity $s \in \mathbb{N}^*$, the upper bound on the number of components in the mixed signal K and the intrinsic noise level σ. We shall consider assumptions on the regularity of the dictionary φ_T, on the parameter space Θ_T on which the optimization is performed and on the noise w_T. Using the features φ_T we build a kernel $\mathcal{K}_T$ on the space of parameters Θ and an associated Riemannian metric $\mathfrak{d}_T$, see Section 4, which is the intrinsic metric, rather than the usual Euclidean metric. More assumptions are necessary on the closeness of the kernel $\mathcal{K}_T$ and its derivatives defined in (30) to a limit kernel $\mathcal{K}_\infty$ and its derivatives.

The theorem is stated assuming the existence of certificate functions, see Assumptions 6.1 and 6.2. Sufficient conditions for their existence are given later in Section 7, in which Propositions 7.4 and 7.5 show that the limit kernel $\mathcal{K}_\infty$ must be uniformly bounded and have concavity properties. In this case, the existence of certificates stands provided the underlying non-linear parameters to be estimated are sufficiently separated according to the Riemannian metric $\mathfrak{d}_T$, see Condition *(iii)* in Propositions 7.4 and 7.5.

In the following result the parameter set Θ_T is a one dimensional compact interval. We note $|\Theta_T|_{\mathfrak{d}_T}$ its length with respect to the Riemannian metric $\mathfrak{d}_T$ on Θ^2 associated to the kernel $\mathcal{K}_T$.

Theorem 2.1. *Assume we observe the random element y of H_T under the regression model (2) with unknown parameters $\beta^\star$ 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:*

(i) ***Admissible noise:*** *The noise process w_T satisfies Assumption 1.1 for a noise level $\sigma > 0$ and a decay rate for the noise variance $\Delta_T > 0$.*
(ii) ***Regularity of the dictionary φ_T:*** *The dictionary function φ_T satisfies the smoothness conditions of Assumption 3.1. The function g_T defined in (15), satisfies the positivity condition of Assumption 3.2.*
(iii) ***Regularity of the limit kernel:*** *The kernel $\mathcal{K}_\infty$ and the functions g_∞ and h_∞, defined on an interval $\Theta_\infty \subset \Theta$, see (17) and (34), satisfy the smoothness conditions of Assumption 5.1.*
(iv) ***Proximity to the limit kernel:*** *The kernel $\mathcal{K}_T$ defined from the dictionary, see (30), is sufficiently close to the limit kernel $\mathcal{K}_\infty$ in the sense that Assumption 5.2 holds.*
(v) ***Existence of certificates:*** *The set of unknown parameters $\mathcal{Q}^\star = \{\theta_k^\star,\, k \in S^\star\}$, with $S^\star = \mathrm{Supp}(\beta^\star)$, satisfies Assumptions 6.1 and 6.2 with the same $r > 0$.*

Then, there exist finite positive constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ depending on the kernel $\mathcal{K}_\infty$ defined on Θ_∞ and on r such that for any $\tau > 0$ and a tuning parameter:

$$\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau},$$

we have the prediction error bound of the estimators $\hat{\beta}$ and $\hat{\vartheta}$ defined in (4) given by:

$$\left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T \leq \mathcal{C}_0 \sqrt{s}\, \kappa, \tag{5}$$

with probability larger than $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau \sqrt{\log \tau}} \vee \frac{1}{\tau} \right)$. Moreover, with the same probability, the difference of the ℓ_1-norms of $\hat{\beta}$ and $\beta^\star$ is bounded by:

$$\left| \| \hat{\beta} \|_{\ell_1} - \| \beta^\star \|_{\ell_1} \right| \leq \mathcal{C}_3\, \kappa\, s. \tag{6}$$

This result holds for both the continuous and discrete settings described in Section 1.2, covers a wide range of smooth dictionaries, and is proven under mild assumptions on the noise. We discuss in the next remark that the prediction error is, up to a logarithmic factor, almost optimal.

Remark 2.2 (Comparison with the Lasso estimator). Let us consider the model of Section 1.2.1 where the observation space is the Hilbert space $H_T = \mathbb{R}^T$ endowed with the Euclidean norm $\|\cdot\|_{\ell_2}$. The observation $y \in \mathbb{R}^T$ comes from the model (2) where the noise is a Gaussian vector with independent entries of variance σ^2. In this setting, the decay rate of the noise variance is fixed with $\Delta_T = 1$.

We first consider that the parameters $\vartheta^\star$ are known. In this case, the model becomes the classical high-dimensional regression model and the Lasso estimator $\hat{\beta}_L$ can be used to estimate $\beta^\star$ under coherence assumptions on the finite dictionary made of the rows of the matrix $\Phi^\star = \Phi_T(\vartheta^\star)$ (see [9]). The behavior of the Lasso estimator has been studied in the literature and its prediction risk tends to zero at the rate:

$$\frac{1}{T} \| (\hat{\beta}_L - \beta^\star) \Phi^\star \|_{\ell_2}^2 = \mathcal{O}\left(\frac{\sigma^2\, s\, \log(K)}{T} \right) \tag{7}$$

with high probability, larger than $1 - 1/K^\gamma$ for some positive constant $\gamma > 0$. Furthermore, in the case where $\beta^\star$ is an unknown s-sparse vector, $\vartheta^\star$ is known and $\Phi^\star$ verifies a coherence property, then the lower bounds of order $\sigma^2\, s\, \log(K/s)/T$ in expected value can be deduced from the more general bounds for group sparsity in [38] (see also [42]). The non-asymptotic prediction lower bounds for the prediction error given in [42] are:

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

where the infinimum is taken over all the estimators $\hat{\beta}$ (square integrable measurable functions of the obervation y) and for some constant $C > 0$ free of s and T. When the parameters $\vartheta^\star$ are unknown, Theorem 2.1 gives an upper bound for the prediction risk which is, up to a logarithmic factor, almost the best rate we could achieve even knowing the non-linear parameters $\vartheta^\star$. Consider the estimators in (4) where the Riemannian diameter of the set Θ_T is bounded by a constant free of T (this is the case of Example 5.1 below). By squaring (5) and then dividing it by T, we obtain from Theorem 2.1 with $\kappa = \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau}$ and $\tau = T^\gamma$ for some given $\gamma > 0$, that with high probability, larger than $1 - C/T^\gamma$:

$$\frac{1}{T}\left\|\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star)\right\|_{\ell_2}^2 = \mathcal{O}\left(\frac{\sigma^2\, s\, \log(T)}{T}\right). \tag{8}$$

Let us mention that [47] also obtained a similar prediction error (8) for the specific dictionary given by the complex exponential functions $(\varphi(\theta) : t \mapsto \mathrm{e}^{i2\pi t\theta}, \theta \in \Theta = [0, 2\pi])$; notice that the proof therein use the Parseval's identity for Fourier series as well as Markov-Bernstein type inequalities for trigonometric polynomials. Even if the structure of our proof is in the spirit of [47], our result is more general and does not rely on the convex setting of the BLasso approach.

Remark 2.3 (Proximity to the limit kernel). We comment on Condition *(iv)* on the proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$, which also appears as Conditions *(iv)*-*(v)* in Proposition 7.4 (and similarly as Condition *(iv)* in Proposition 7.5).

In the examples of Sections 3.3.2 and 3.3.4 on translation or scaling model with a continuum of observations, the parameter T does not play any role in the definition of $\mathcal{K}_T$, so that one can take $\mathcal{K}_\infty$ equal to $\mathcal{K}_T$. In this case, the proximity conditions on the kernels are trivially satisfied.

In the continuation of Example 1.1, the example from Section 8 is devoted to the sparse spike deconvolution, that is, to a mixture of Gaussian translation invariant features observed in a discrete regression model on a regular grid of size T. In this case, we built a family of models $(H_T, \varphi_T, w_T, \Theta_T)$ with a dictionary φ_T which does not depend on T and such that the kernel $\mathcal{K}_T$ and its derivatives converge to $\mathcal{K}_\infty$ (and also ρ_T from (36) converges to 1). In this setting, the proximity condition of Theorem 2.1 holds for T large enough, say T larger than some T_0 which depends on $\mathcal{K}_\infty$, see Assumption 5.2. The existence of the certificates, see Propositions 7.4 and 7.5, also requires a proximity criterion which is achieved for T large enough, say T larger than some T_1 which depends on $\mathcal{K}_\infty$ and is increasing with the sparsity parameter s (see for example Condition *(v)* in Proposition 7.4).

Remark 2.4 (On the dimension K, the upper bound of the sparsity). We remark that neither the bound on the prediction error nor the probability on which the bound holds, depends on the upper bound K on the sparsity s. Therefore, the value of K can be taken arbitrarily large. It is not surprising that K does not have any impact on the bound since the optimisation problem (4) could be formulated without any bound on the sparsity. Indeed, the problem (4) can be embedded in an optimization problem over a space of measures following the literature on the BLasso introduced in [23]. See also Remark 2.7.

The next theorem gives bounds on the differences between the parameters $\hat{\beta}$ given by the optimization problem (4) and the "true" parameters $\beta^\star$ for active features having their parameter $\hat{\theta}_\ell$ close, with respect to the Riemannian metric $\mathfrak{d}_T$, to a parameter $\theta_k^\star$, with k in $S^\star$. For $r > 0$ given by Assumptions 6.1 and 6.2, we define:

- The support of $\hat{\beta}$ given by the optimization problem (4): $\hat{S} = \mathrm{Supp}(\hat{\beta}) = \left\{\ell : \hat{\beta}_\ell \neq 0\right\}$.
- The near region $\tilde{S}(r)$ given by:

$$\tilde{S}(r) = \bigcup_{k\ \in S^\star} \tilde{S}_k(r) \quad \text{where} \quad \tilde{S}_k(r) = \left\{\ell \in \hat{S} : \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) \leq r\right\},$$

  which corresponds to the set of indices ℓ in the support of $\hat{\beta}$ such that the corresponding parameter $\hat{\theta}_\ell$ is close to one of the true parameter $\theta_k^\star$, for some $k \in S^\star$.

The set $\hat{S}\backslash\tilde{S}(r)$ is also called the far region. Notice that the sets $\tilde{S}_k(r)$ with $k \in S^\star$ are pairwise disjoint under Assumption 6.1, and that they can be empty. In what follows, we use the convention $\sum_\emptyset = 0$.

Theorem 2.5. *We consider the model in Theorem 2.1 and suppose that Assumptions (i)-(v) therein hold. Then, there exist finite positive constants $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, $\mathcal{C}_4$, $\mathcal{C}_5$ depending on $\mathcal{K}_\infty$ defined on Θ_∞ and on r such that for any $\tau > 0$ and a tuning parameter:*

$$\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau}$$

the estimator $\hat{\beta}$ defined in (4) satisfies the following bounds with probability larger than $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log \tau}} \vee \frac{1}{\tau} \right)$:

$$\sum_{k \in S^\star} \left| |\beta_k^\star| - \sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| \right| \leq \mathcal{C}_3 \, \kappa \, s, \quad \sum_{k \in S^\star} \left| \beta_k^\star - \sum_{\ell \in \tilde{S}_k(r)} \hat{\beta}_\ell \right| \leq \mathcal{C}_4 \, \kappa \, s \quad \text{and} \quad \left\| \hat{\beta}_{\tilde{S}(r)^c} \right\|_{\ell_1} \leq \mathcal{C}_5 \, \kappa \, s, \tag{9}$$

where for a subset S of $\mathcal{I} = \{1, \cdots, K\}$, the set S^c denotes the complementary set of S in $\mathcal{I}$, that is $\mathcal{I} \setminus S$.

Notice that each linear parameter $\beta_k^\star$ can be estimated by the sum of several linear coefficients $\hat{\beta}_\ell$ with $\ell \in \{1, \cdots, K\}$. The first two inequalities in (9) show that there can be no compensation between the estimators $\hat{\beta}_\ell$ that approximate the same $\beta_k^\star$ with $k \in S^\star$, meaning that there can be no large values of $\hat{\beta}_\ell$ having different signs that sum up to a possibly small (in absolute value) true $\beta_k^\star$. The second inequality in (9) gives the estimation rate of the linear parameters $\beta_k^\star$ with $k \in S^\star$. The last bound in (9) basically means that when an estimation $\hat{\theta}_\ell$ with $\ell \in \{1, \cdots, K\}$ is far from any parameter $\theta_k^\star$ with $k \in S^\star$, that is at a distance greater than r, the associated parameters $\hat{\beta}_\ell$ drop to zero if the tuning parameter κ is taken equal to its lower bound and the decay rate of the noise variance Δ_T drops to zero. Therefore, the contribution of the parameters $\hat{\theta}_\ell$ in the far region, that are not in $\tilde{S}(r)$, will drop to zero as well.

Remark 2.6 (Estimation rate for $\theta_k^\star$ with $k \in S^\star$). Under the assumptions of Theorem 2.5, we also have, with probability larger than $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log \tau}} \vee \frac{1}{\tau} \right)$, the bound:

$$\sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \left| \hat{\beta}_\ell \right| \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \leq \mathcal{C}_6 \, \kappa \, s. \tag{10}$$

This gives an estimation rate for the parameters $\theta_k^\star$ with $k \in S^\star$ when at least one estimator $\hat{\theta}_\ell$ given by the optimization problem (4) belongs to the near region $\tilde{S}_k(r)$, which is the Riemannian ball centered at $\theta_k^\star$ with radius r.

Remark 2.7 (Again on the dimension K). As in Theorem 2.1, we remark that neither the bounds nor the probability of the event on which the bounds hold depend on the upper bound K on the sparsity s.

If the optimization on ϑ in (4) is performed over a subset of Θ_T in which the coordinates of the considered vectors are at a distance greater than $2r$ pairwise with respect to the Riemannian metric $\mathfrak{d}_T$, then the sets $\tilde{S}_k(r)$ contain at most one element. However, by doing so, we introduce an upper bound on the dimension K whereas in Theorem 2.1 the dimension K can be arbitrarily large. Indeed, Θ_T is a compact set and therefore contains a finite number of balls of size $2r$.

Outline of the paper. In Section 3, we give the definition of the kernel $\mathcal{K}_T$ measuring the correlation between two elements in the continuous dictionary and we present the regularity assumptions on the function φ_T. Section 4 introduces the Riemannian geometry framework useful in our context. Section 5 defines the convergence (or closeness condition) of kernels $\mathcal{K}_T$ towards a limit kernel $\mathcal{K}_\infty$. Then, we require properties on the limit kernel $\mathcal{K}_\infty$ and propagate them to the kernels $\mathcal{K}_T$ thanks to this convergence. In Section 6, we present the assumptions on the existence of the so-called certificate functions used to state Theorems 2.1 and 2.5. We give sufficient conditions for the existence of certificate functions in Section 7. The example of sparse deconvolution in our regression model is fully detailed in Section 8. Then, Sections 9.1 and 9.2 are dedicated to the proofs of Theorems 2.1 and 2.5. The proofs of existence and explicit constructions of the certificates are detailed in Section 10.

3. Dictionary of features

We recall in the next section some basic results on the Fréchet derivative and the Bochner integral. Then, we present the regularity assumptions on the features $(\varphi_T(\theta), \theta \in \Theta)$ we shall consider.

3.1. **The Fréchet derivative and the Bochner integral.**

The Fréchet derivative and Bochner integrals are defined for Banach space valued functions, but we shall only consider the case of Hilbert space valued functions.

Let $(H, \langle\cdot,\cdot\rangle)$ be an Hilbert space and let Θ be an interval of $\mathbb{R}$. We note $\|\cdot\|$ the norm associated to the scalar product. A function f from Θ to H is Fréchet differentiable at $\theta \in \Theta$ if it is continuous at θ and there exists an element $\partial_\theta f \in H$ such that:

$$\lim_{h\to 0;\, \theta+h\in\Theta} \quad \left\| \frac{f(\theta+h) - f(\theta)}{h} - \partial_\theta f(\theta) \right\| = 0.$$

The derivative of f is the function $\partial_\theta f : \theta \mapsto \partial_\theta f(\theta)$ defined on Θ when it exists. We also define by recurrence the derivative $\partial_\theta^i f$ of order $i \in \mathbb{N}^*$ of f as the derivative of $\partial_\theta^{i-1} f$, with the convention that $\partial_\theta^0 f = f$, and say that f is of class $\mathcal{C}^i$ if the derivatives $\partial_\theta^j f$ exist and are continuous on Θ for $j \in \{0, \dots, i\}$. The standard differentiating rules for composition, addition and multiplication apply to the Fréchet derivative. We refer to [36] for a complete presentation of the subject. By definition, if f is differentiable at $\theta \in \Theta$, then we have for all $g \in H$ that:

$$\partial_\theta \langle f(\theta), g\rangle = \langle \partial_\theta f(\theta), g\rangle . \tag{11}$$

The Bochner integral extends the Lebesgue integral. We refer to [4, Chapter 1] and [3, Section 11.8] for further details on the Bochner integral. We endow the interval $\Theta \subset \mathbb{R}$ with its usual Borel sigma field inherited from the Borel sigma field on $\mathbb{R}$ and a measure μ. A function f from Θ to H is strongly measurable if it is the limit of simple functions or equivalently, see [3, Lemma 11.37], if the map $\theta \mapsto \langle f(\theta), g\rangle$ is measurable for all $g \in H$ and $f(\theta)$ lies for μ-almost every $\theta \in \Theta$ in a closed separable subspace of H. In particular if the function f is continuous, then it is strongly measurable, see [4, Corollary 1.1.2]. If f is strongly measurable, then the norm $\|f\|$ is a measurable function from Θ to $\mathbb{R}$, see [3, Lemma 11.39]. Then a function f defined on Θ (endowed with the Lebesgue measure) is Bochner integrable if and only if it is strongly measurable and if $\|f\|$ is integrable; in which case, we have $\left\|\int f(\theta)\,\mathrm{d}\theta\right\| \leq \int \|f(\theta)\|\,\mathrm{d}\theta$, see [3, Theorem 11.44] (which is easily extended from finite measure to σ-finite measure, see also [4, Theorem 1.1.4] in this direction). We remark that the fundamental theorem of calculus is still valid in this framework, see [4, Proposition 1.2.2]. In particular, if f is continuous and Bochner integrable on Θ and $\theta_0 \in \Theta$, then, we have:

$$F'(\theta) = f(\theta) \quad \text{where} \quad F(\theta) = \int_{\theta_0}^{\theta} f(q)\,\mathrm{d}q. \tag{12}$$

As a particular case of [3, Lemma 11.45], if f is Bochner integrable on Θ, then for all $g \in H$, we have that:

$$\int_\Theta \langle f(\theta), g\rangle\,\mathrm{d}\theta = \langle \int_\Theta f(\theta)\,\mathrm{d}\theta, g\rangle. \tag{13}$$

3.2. **Assumptions on the regularity of the features.**

Let $T \in \mathbb{N}$ be fixed. We consider the Hilbert space $(H_T, \langle\cdot,\cdot\rangle_T)$ and the features $(\varphi_T(\theta), \theta \in \Theta)$ which are elements of H_T. We shall consider the following regularity assumptions on the features.

Assumption 3.1 (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.*

Recall $\phi_T = \varphi_T/\|\varphi_T\|_T$ from (1) and notice that ϕ_T, and thus Φ_T, are continuous functions. Under Assumption 3.1, elementary calculations using (11) give:

$$\partial_\theta \phi_T(\theta) = \frac{\partial_\theta \varphi_T(\theta)}{\|\varphi_T(\theta)\|_T} - \frac{\varphi_T(\theta)\,\langle \varphi_T(\theta), \partial_\theta \varphi_T(\theta)\rangle_T}{\|\varphi_T(\theta)\|_T^3}, \tag{14}$$

and thus, we deduce that the function $g_T : \Theta \mapsto \mathbb{R}_+$ defined by:

$$g_T(\theta) = \|\partial_\theta \phi_T(\theta)\|_T^2 \tag{15}$$

is well defined and continuous.

We shall consider the following non-degeneracy assumption on the features.

Assumption 3.2 (Positivity of g_T)**.** *Assumption 3.1 holds and we have $g_T > 0$ on Θ.*

Even if Assumption 3.2 requires Assumption 3.1, in the following we shall stress when Assumption 3.1 is in force.

The next lemma gives a sufficient condition on φ_T for Assumption 3.2 to hold.

Lemma 3.1 (On the positivity of g_T)**.** *Suppose Assumption 3.1 holds. If the elements $\varphi_T(\theta)$ and $\partial_\theta \varphi_T(\theta)$ of H_T are linearly independent for all $\theta \in \Theta$ and $\|\partial_\theta \varphi_T(\theta)\|_T > 0$ for all $\theta \in \Theta$, then g_T is positive on Θ.*

Proof. For simplicity, we remove the subscript T, and for example write simply $\phi = \varphi/\|\varphi\|$. Recall that by Assumption 3.1 we have $\|\varphi(\theta)\| > 0$. Assume there exists $\theta \in \Theta$ such that $g(\theta) = 0$, that is $\partial_\theta \phi(\theta) = 0$. Since $\|\varphi(\theta)\| > 0$, we deduce from (14) that $\partial_\theta \varphi(\theta)\|\varphi(\theta)\|^2 - \varphi(\theta)\,\langle \varphi(\theta), \partial_\theta \varphi(\theta)\rangle = 0$. Then use that by assumption $\partial_\theta \varphi(\theta) \neq 0$ and $\|\varphi(\theta)\| > 0$, to get that $\varphi(\theta)$ and $\partial_\theta \varphi(\theta)$ are linearly dependent. In conclusion, we get that if $\varphi(\theta)$ and $\partial_\theta \varphi(\theta)$ are linearly independent, then $g(\theta) > 0$. □

3.3. **Examples of regular features.** The aim of this section of examples is to stress that a large variety of dictionaries of features and type of parameters verify Assumptions 3.1 and 3.2.

3.3.1. *Translation model with observations on a finite grid.* This model is an extension of Example 1.1 in the spirit of Section 1.2.2. Let $t_1 < \cdots < t_T$ be a grid on $\mathbb{R}$ of size $T \in \mathbb{N}$, λ_T an atomic measure whose support is the grid, and $H_T = L^2(\lambda_T)$. Consider the translation invariant dictionary:

$$(\varphi_T(\theta) = k(\cdot - \theta),\, \theta \in \Theta), \tag{16}$$

with $\Theta = \mathbb{R}$ and k is a real-valued $\mathcal{C}^3$ function defined on $\mathbb{R}$. Notice the dictionary does not depend on T. We now consider usual choices for the function k.

For the Gaussian function $k(t) = \mathrm{e}^{-t^2/2}$ and the Cauchy function $k(t) = 1/(1+t^2)$, we get that Assumption 3.1 holds and, using Lemma 3.1 that Assumption 3.2 is also satisfied provided respectively $T \geq 2$ and $T \geq 3$.

For the Shannon scaling function $k(t) = \mathrm{sinc}(t) = \sin(\pi t)/(\pi t)$, Assumption 3.1 holds provided that $\lambda_T((a+\mathbb{Z})^c) > 0$ for all $a \in \mathbb{R}$, that is the grid is not a subset of $a + \mathbb{Z}^*$ for some $a \in \mathbb{R}$. There is no easy way to write conditions on the grid, based on the use of Lemma 3.1, for Assumption 3.2 to hold (let us mention that $T \geq 2$ and $\min_{1 \leq i \leq T-1}(t_{i+1} - t_i) < 1/2$ is a sufficient condition for Assumption 3.2 to hold).

Eventually notice that the Laplace function $k(t) = \mathrm{e}^{-|t|}$ is not smooth enough for Assumption 3.1 to hold.

3.3.2. *Translation model with a continuum of observations.* Let $T \in \mathbb{N}$ (which does not play a role here) and $H_T = L^2(\mathrm{Leb})$, where Leb is the Lebesgue measure on $\mathbb{R}$. In this framework, the observation y defined in (2) is a continuum of observations. Consider the translation invariant dictionaries from Section 3.3.1, where k is either the Gaussian, the Cauchy or the Shannon scaling function. Notice that the Hilbert space and the dictionary do not depend on T. Then, it is easy to check that Assumptions 3.1 and 3.2 hold.

We see that this model, which can be seen as a continuous approximation (or limit) of the discrete models from Section 3.3.1 when T therein is large, is easier to handle than the corresponding discrete models.

3.3.3. *Translation model with a varying scaling parameter.* Let $T \in \mathbb{N}$, $H_T = L^2(\text{Leb})$, where Leb is the Lebesgue measure on $\mathbb{R}$, and consider the translation invariant dictionary scaled by $\overline{\sigma}_T > 0$ given by:

$$(\varphi_T(\theta) = k(\overline{\sigma}_T^{-1}(\cdot - \theta)),\, \theta \in \Theta),$$

with $\Theta = \mathbb{R}$ and k is a real-valued $\mathcal{C}^3$ function defined on $\mathbb{R}$. Contrary to Section 3.3.2, the features depend on T. Suppose that k is the Shannon scaling function (see Section 3.3.1) and consider the vector sub-space V_T given by the closure in H_T of the vector space spanned by the dictionary. According to [39, Theorem 3.5], the set V_T is the subset of H_T of all functions whose Fourier transform support is a subset of $[-\pi/\overline{\sigma}_T, \pi/\overline{\sigma}_T]$. Suppose that the sequence $(\overline{\sigma}_T, T \in \mathbb{N})$ is decreasing to 0. Then the sequence $(V_T, T \in \mathbb{N})$ is increasing and $\overline{\bigcup_{T\in\mathbb{N}} V_T} = H_T$. This model provides an example of translation models with possibly varying, but known, scaling parameter $\overline{\sigma}_T$.

3.3.4. *Scaling exponential model.* Let $T \in \mathbb{N}$ (which does not play a role here), $H_T = L^2(\text{Leb})$, where Leb is the Lebesgue measure on $\mathbb{R}_+$, and consider the scale invariant dictionary given by:

$$(\varphi_T(\theta) = k(\theta \cdot),\, \theta \in \Theta),$$

with $\Theta = \mathbb{R}_+^*$ and the exponential function $k : t \mapsto \mathrm{e}^{-t}$. This dictionary is used for example in fluorescence microscopy (see [24]). Clearly Assumption 3.1 holds as well as Assumption 3.2 as $g_T(\theta) = 1/(4\theta^2)$.

4. A Riemannian metric on the set of parameters

4.1. **On the Riemannian metric in dimension one.** Recall Θ is an interval of $\mathbb{R}$. 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 Θ by:

$$g_\mathcal{K}(\theta) = \partial^2_{x,y}\mathcal{K}(\theta, \theta) \tag{17}$$

is positive and locally bounded, where ∂_x (resp. ∂_y) denotes the usual derivative with respect to the first (resp. second) variable. Following [41], we define an intrinsic Riemannian metric, denoted $\mathfrak{d}_\mathcal{K}$, on the parameter set Θ using the function $g_\mathcal{K}$. One of the motivation to use the Riemannian metric is to work with intrinsic quantities related to the parameters which are invariant by reparametrization, such as the diameter of (subsets of) Θ. Since Θ is one-dimensional and connected, the Riemannian metric $\mathfrak{d}_\mathcal{K}(\theta, \theta')$ between $\theta, \theta' \in \Theta$ reduces to:

$$\mathfrak{d}_\mathcal{K}(\theta, \theta') = |G_\mathcal{K}(\theta) - G_\mathcal{K}(\theta')|, \tag{18}$$

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

Remark 4.1. We refer to [37] and [43] for a general presentation on Riemannian manifolds, and we give an immediate application in dimension one which entails in particular (18). Let Θ be a manifold (of dimension one). A path $\gamma : [0, 1] \to \Theta$ is an admissible path if it is continuous, piecewise continuously differentiable with non-vanishing derivative. Its length is given by $\mathcal{L}_\mathcal{K}(\gamma) = \int_0^1 |\dot{\gamma}_s| \sqrt{g_\mathcal{K}(\gamma_s)}\, \mathrm{d}s$, where $|\dot{\gamma}_s|$ is seen as the norm of the vector $\dot{\gamma}_s$ in the tangent space, and the scalar product on the tangent space at $\theta \in \Theta$ is given by $(u, v) \mapsto \langle u, g_\mathcal{K}(\theta) v\rangle$ with $\langle \cdot, \cdot \rangle$ the usual Euclidean scalar product. (In our case, the tangent vector space is $\mathbb{R}$ and the Euclidean scalar product reduces to the usual product). The Riemannian metric $\mathfrak{d}_\mathcal{K}$ between θ, θ' in Θ is then defined by:

$$\mathfrak{d}_\mathcal{K}(\theta, \theta') = \inf_\gamma \quad \mathcal{L}_\mathcal{K}(\gamma), \tag{19}$$

where the infinimum is taken over the admissible paths γ such that $\gamma_0 = \theta$ and $\gamma_1 = \theta'$. It is not hard to see that γ is a minimizing path, that is, $\mathfrak{d}_\mathcal{K}(\theta, \theta') = \mathcal{L}_\mathcal{K}(\gamma)$, if and only if γ is monotone (and thus $\gamma_s \in [\theta \wedge \theta', \theta \vee \theta']$ for all $s \in [0, 1]$). This is equivalent to say that the sign of $\dot{\gamma}_s$ is constant. Assume that

$g_\mathcal{K}$ is of class $\mathcal{C}^1$. The path γ is a geodesic if it is smooth with zero acceleration, that is, in dimension one for all $s \in (0,1)$:

$$\ddot{\gamma}_s + \frac{1}{2}\frac{g'_\mathcal{K}(\gamma_s)}{g_\mathcal{K}(\gamma_s)}\dot{\gamma}_s^2 = 0. \tag{20}$$

This is equivalent to $s \mapsto \dot{\gamma}_s \sqrt{g_\mathcal{K}(\gamma_s)}$ being constant, which implies that the geodesic is a minimizing path.

We now derive the equation of the geodesic path when $g_\mathcal{K}$ is of class $\mathcal{C}^1$. Recall $G_\mathcal{K}$ denotes the primitive of $\sqrt{g_\mathcal{K}}$. It is continuous increasing and thus induces a one-to-one map from Θ to its image. Set $a = G_\mathcal{K}(\theta)$ and $b = G_\mathcal{K}(\theta') - G_\mathcal{K}(\theta)$, so that the path $\gamma : [0,1] \to \Theta$ defined by $\gamma_s = G_\mathcal{K}^{-1}(a + bs)$ is a geodesic and minimizing path from θ to θ' with $\mathcal{L}_\mathcal{K}(\gamma) = \mathfrak{d}_\mathcal{K}(\theta, \theta')$.

Following [41], we introduce the covariant derivatives, see [2, Sections 3.6 and 5.6], which have elementary expressions as the set of parameters Θ is one-dimensional. For a smooth function f defined on Θ and taking values in an Hilbert space, say H, the covariant derivative $D_{i;\mathcal{K}}[f]$ of order $i \in \mathbb{N}$ is defined recursively by $D_{0;\mathcal{K}}[f] = f$ and for $i \in \mathbb{N}$, assuming that $g_\mathcal{K}$ is of class $\mathcal{C}^i$, and $\theta \in \Theta$:

$$D_{i+1;\mathcal{K}}[f](\theta) = g_\mathcal{K}(\theta)^{\frac{i}{2}} \partial_\theta \left(\frac{D_{i;\mathcal{K}}[f](\theta)}{g_\mathcal{K}(\theta)^{\frac{i}{2}}} \right). \tag{21}$$

In particular, we have for $f \in \mathcal{C}^2(\Theta, H)$ (and assuming that $g_\mathcal{K}$ is of class $\mathcal{C}^1$ for the last equality) that:

$$D_{0;\mathcal{K}}[f] = f, \quad D_{1;\mathcal{K}}[f] = \partial_\theta f, \quad D_{2;\mathcal{K}}[f] = \partial_\theta^2 f - \frac{1}{2}\frac{g'_\mathcal{K}}{g_\mathcal{K}}\partial_\theta f. \tag{22}$$

We shall also consider the following modification of the covariant derivative, for $i \in \mathbb{N}$:

$$\tilde{D}_{i;\mathcal{K}}[f](\theta) = g_\mathcal{K}(\theta)^{-i/2}\, D_{i;\mathcal{K}}[f](\theta). \tag{23}$$

We have $\tilde{D}_{0;\mathcal{K}}[f] = f$, and we deduce from (21) that for $i \in \mathbb{N}^*$, assuming that $g_\mathcal{K}$ is of class $\mathcal{C}^i$:

$$\tilde{D}_{i;\mathcal{K}} = \tilde{D}_{1;\mathcal{K}} \circ \tilde{D}_{i-1;\mathcal{K}} = \left(\tilde{D}_{1;\mathcal{K}}\right)^i, \tag{24}$$

so that $\tilde{D}_{1;\mathcal{K}}$ can be seen as a derivative operator.

We now give an elementary variant of the Taylor-Lagrange expansion using the previously defined Riemannian metric and covariant derivatives. Its proof can be found in the Appendix, Section A.3.

Lemma 4.2. *Assume $g_\mathcal{K}$ is positive and of class $\mathcal{C}^1$. Let f be a function defined on Θ taking values in an Hilbert space of class $\mathcal{C}^2$. Setting $f^{[i]} = \tilde{D}_{i;\mathcal{K}}[f]$ for $i \in \{1,2\}$, we have that for all $\theta, \theta_0 \in \Theta$:*

$$f(\theta) = f(\theta_0) + \operatorname{sign}(\theta - \theta_0)\, \mathfrak{d}_\mathcal{K}(\theta, \theta_0)\, f^{[1]}(\theta_0) + \mathfrak{d}_\mathcal{K}(\theta, \theta_0)^2 \int_0^1 (1-t) f^{[2]}(\gamma_t)\, \mathrm{d}t, \tag{25}$$

where γ is a geodesic path such that $\gamma_0 = \theta_0$, $\gamma_1 = \theta$ (and thus $\mathfrak{d}_\mathcal{K}(\theta, \theta_0) = \mathcal{L}_\mathcal{K}(\gamma)$).

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 $\theta, \theta' \in \Theta$:

$$D_{i+1,j;\mathcal{K}}[F](\theta, \theta') = g_\mathcal{K}(\theta)^{\frac{i}{2}} \partial_\theta \left(\frac{D_{i,j;\mathcal{K}}[F](\theta, \theta')}{g_\mathcal{K}(\theta)^{\frac{i}{2}}} \right) \quad \text{and} \quad D_{i,j;\mathcal{K}}[F](\theta, \theta') = D_{j,i;\mathcal{K}}[F](\theta', \theta). \tag{26}$$

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 $i, j \in \mathbb{N}$:

$$\tilde{D}_{i,j;\mathcal{K}}[F](\theta, \theta') = \frac{D_{i,j;\mathcal{K}}[F](\theta, \theta')}{g_{\mathcal{K}}(\theta)^{i/2}\, g_{\mathcal{K}}(\theta')^{j/2}}. \tag{27}$$

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 $\mathcal{C}^{\max(i,j)}$:

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

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

$$\mathcal{K}^{[i,j]} = \tilde{D}_{i,j;\mathcal{K}}[\mathcal{K}]. \tag{28}$$

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

$$\mathcal{K}^{[0,0]} = \mathcal{K} \quad \text{and} \quad \mathcal{K}^{[1,1]}(\theta, \theta) = 1. \tag{29}$$

4.2. **The kernel and the Riemannian metric associated to the dictionary of features.** Let $T \in \mathbb{N}$ be fixed and assume that Assumption 3.2 holds. We define the kernel $\mathcal{K}_T$ on Θ^2 by:

$$\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T = \frac{\langle \varphi_T(\theta), \varphi_T(\theta') \rangle_T}{\|\varphi_T(\theta)\|_T \|\varphi_T(\theta')\|_T}, \tag{30}$$

where we recall that $\phi_T = \varphi_T / \|\varphi_T\|_T$. When considering the kernel $\mathcal{K}_T$, we shall write g_T for $g_{\mathcal{K}_T}$, and similarly we shall use the notations $\tilde{D}_{i;T}$ and $\tilde{D}_{i,j;T}$ instead of $\tilde{D}_{i;\mathcal{K}_T}$ and $\tilde{D}_{i,j;\mathcal{K}_T}$. Recall the derivatives of the kernel $\mathcal{K}_T$ defined by (28). The next lemma insures in particular that the two definitions of g_T given by (15) and (17) are consistent, that is:

$$g_T(\theta) = \partial^2_{xy} \mathcal{K}_T(\theta, \theta) = \|\partial_\theta \phi_T(\theta)\|_T^2. \tag{31}$$

The proof of the next lemma can be found in the Appendix, Section A.3.

Lemma 4.3. *Let $T \in \mathbb{N}$ be fixed and assume that Assumptions 3.1 and 3.2 hold. Then, the symmetric kernel $\mathcal{K}_T$ is of class $\mathcal{C}^{3,3}$ on Θ^2 and for $i, j \in \{0, \dots, 3\}$ and $\theta, \theta' \in \Theta$, we have:*

$$\mathcal{K}_T^{[i,j]}(\theta, \theta') = \langle \tilde{D}_{i;T}[\phi_T](\theta), \tilde{D}_{j;T}[\phi_T](\theta') \rangle_T. \tag{32}$$

We also have:

$$\sup_{\Theta^2} |\mathcal{K}_T^{[0,0]}| \leq 1, \quad \mathcal{K}_T^{[0,0]}(\theta, \theta) = 1, \quad \mathcal{K}_T^{[1,0]}(\theta, \theta) = 0, \quad \mathcal{K}_T^{[2,0]}(\theta, \theta) = -1 \quad \textit{and} \quad \mathcal{K}_T^{[2,1]}(\theta, \theta) = 0. \tag{33}$$

5. Approximating the kernel associated to the dictionary

In the section we detail the assumptions guaranteeing the approximation of the kernel $\mathcal{K}_T$ (which is usually difficult to compute) by a kernel $\mathcal{K}_\infty$ (which is easier to handle). Both kernels are defined on Θ^2, however, we shall qualify the approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$ and properties of $\mathcal{K}_\infty$ on subsets of Θ, respectively Θ_T (which will be a compact interval) and Θ_∞ (which will be an interval possibly unbounded). We use notations from Section 4 and recall the definition of $g_{\mathcal{K}}$, resp. $\mathcal{K}^{[i,j]}$, given in (17), resp. in (28). Assuming the kernel $\mathcal{K}$ is of class $\mathcal{C}^{3,3}$ and using the notation (28), we also set for $\theta \in \Theta$:

$$h_{\mathcal{K}}(\theta) = \mathcal{K}^{[3,3]}(\theta, \theta). \tag{34}$$

For simplicity, for an expression A we write A_* for $A_{\mathcal{K}_*}$ where $*$ is equal to T or ∞. We first give a regularity assumption on the kernel $\mathcal{K}_\infty$.

Assumption 5.1 (Properties of the asymptotic kernel $\mathcal{K}_\infty$ and function h_∞)**.** *The symmetric kernel $\mathcal{K}_\infty$ defined on Θ^2 is of class $\mathcal{C}^{3,3}$, the function g_∞ defined by (17) 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:*

$$(35)\quad m_g := \inf_{\Theta_\infty} g_\infty > 0, \quad L_3 := \sup_{\Theta_\infty} h_\infty < +\infty, \quad \textit{and} \quad L_{i,j} := \sup_{\Theta_\infty^2} |\mathcal{K}_\infty^{[i,j]}| < +\infty \quad \textit{for all } i,j \in \{0,1,2\}.$$

Since Θ_T is compact, under Assumptions 3.2 and 5.1, we deduce that the constant ρ_T below is positive and finite, where:

$$(36)\qquad \rho_T = \max\left(\sup_{\Theta_T} \sqrt{\frac{g_T}{g_\infty}}, \sup_{\Theta_T} \sqrt{\frac{g_\infty}{g_T}} \right).$$

From the definition of the Riemannian metric given in (18) (see also (19)), we readily deduce that the metrics $\mathfrak{d}_T$ and $\mathfrak{d}_\infty$ are then strongly equivalent on Θ_T; more precisely we have that on Θ_T^2:

$$(37)\qquad \frac{1}{\rho_T}\, \mathfrak{d}_\infty \leq \mathfrak{d}_T \leq \rho_T\, \mathfrak{d}_\infty.$$

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

$$(38)\qquad \mathcal{V}_T = \max(\mathcal{V}_T^{(1)}, \mathcal{V}_T^{(2)}) \quad \text{with} \quad \mathcal{V}_T^{(1)} = \max_{i,j\in\{0,1,2\}} \sup_{\Theta_T^2} |\mathcal{K}_T^{[i,j]} - \mathcal{K}_\infty^{[i,j]}| \quad \text{and} \quad \mathcal{V}_T^{(2)} = \sup_{\Theta_T} |h_T - h_\infty|.$$

Let us recall that Assumption 3.2 implies regularity conditions on $\mathcal{K}_T$, see Lemma 4.3.

Assumption 5.2 (Quality of the approximation)**.** *Let $T \in \mathbb{N}$ be fixed. Assumptions 3.2 and 5.1 hold, the interval $\Theta_T \subset \Theta_\infty$ is a compact interval, and we have:*

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

Notice that if Assumption 3.2 holds, then Assumptions 5.1 and 5.2 hold trivially when one takes $\mathcal{K}_\infty = \mathcal{K}_T$ and $\Theta_\infty = \Theta_T$; notice also that $\rho_T = 1$ in this case. In the next example, the sequence of kernels $(\mathcal{K}_T, T \in \mathbb{N})$ converges to the kernel $\mathcal{K}_\infty$ as T goes to infinity, so that Assumption 5.2 holds for T large enough.

Example 5.1. We consider the example from Example 1.1 with the framework of Section 1.2.2. We assume that the process y is a function defined on $[0,1]$ which, for $T \in \mathbb{N}^*$ is observed through the regular grid $\{t_{j,T} = j/T :\ 1 \leq j \leq T\}$. The process y is seen as an element of the Hilbert space $H_T = L^2(\lambda_T)$, with the probability measure $\lambda_T = \Delta_T \sum_{j=1}^T \delta_{t_{j,T}}$ on $[0,1]$ with $\Delta_T = 1/T$. Let Θ be a compact interval of $\mathbb{R}$ and set $\Theta_T = \Theta_\infty = \Theta$. Consider a dictionary $(\varphi(\theta), \theta \in \Theta)$ independent of T, that is, $\varphi_T = \varphi$ for all $T \in \mathbb{N}^*$, and assume that the function $(\theta, t) \mapsto \varphi(\theta)(t)$ is defined on $\Theta \times [0,1]$ and of class $\mathcal{C}^{3,0}$. Assume that the dictionary satisfies the regularity assumptions of Assumption 3.2.

Let Leb be the Lebesgue measure on $[0,1]$, so that $(\lambda_T, T \in \mathbb{N}^*)$ converges weakly to Leb. Then, define the kernel $\mathcal{K}_\infty$ by (30) with φ_T replaced by φ (as the dictionary does not depend on T) and the scalar product $\langle\cdot,\cdot\rangle_T$ by the usual scalar product on $L^2(\text{Leb})$. Thanks to Lemma 4.3, we deduce that Assumption 5.1 on the properties of $\mathcal{K}_\infty$ is satisfied. Using the weak convergence of $(\lambda_T, T \in \mathbb{N}^*)$ to Leb, we deduce that $\lim_{T\to\infty} \partial_x^i \partial_y^j \mathcal{K}_T = \partial_x^i \partial_y^j \mathcal{K}_\infty$ uniformly on $[0,1]^2$ for all $i,j \in \{0,\ldots,3\}$. This implies that:

$$\lim_{T\to\infty} \mathcal{V}_T = 0 \quad \text{and} \quad \lim_{T\to\infty} \rho_T = 1.$$

Thus Assumption 5.2 holds for T large enough.

6. Certificates

In this section, we make assumptions on the existence of functions from Θ to $\mathbb{R}$ called certificates. These functions have interpolation properties that are corner stones in the proof of Theorem 2.1. The term "certificate" is inherited from the compressed sensing field were such functions were used to get rid of the Restricted Isometry Property condition (RIP) for exact recontruction of signals (see [20] for details on the RIP condition). In [19], the authors showed that is possible to reconstruct exactly a sparse signal from the observations of a finite number of Fourier coefficients by exhibiting a dual certificate. Many papers have followed this line of research since then (see *e.g.* [17, 18, 26]).

In sparse linear models the bounds for prediction error are proved using RIP, Restricted Eigenvalue or compatibility conditions (see [9, 51]). Among these assumptions, the compatibility conditions are the less restrictive. Indeed, the authors of [52] have shown that it is implied by both the RIP and the Restricted Eigenvalue. However, in many contexts even the weaker condition fails to hold. Typically the compatibility condition fails to hold in the context of super-resolution which aims at extracting the frequencies and amplitudes of a linear combination of complex exponentials from a small number of noisy time samples (see [12]).

In the papers [12] and [47], the authors achieve nearly optimal rates for the prediction error in the super-resolution framework using certificate functions. Their method and proof are however quite specific to complex exponentials and their certificates are trigonometric polynomials. The insightful paper of [26] builds certificates in a quite general setting for a one dimensional parameter set Θ. In [22], the authors exhibit certificate functions to deal with more general probability density models where Θ is multidimensional. However they are restricted to translation invariant dictionaries (16). The most general framework has been introduced in [41] where the Riemannian geometry is key to build in a natural way the so-called certificate functions. In fact a separation distance between the parameters to estimate is needed to build certificates and the Euclidean metric yields overly pessimistic minimum separation condition. In what follows we introduce new certificates, called derivative certificates, in order to control the prediction error.

We consider the following assumption in the spirit of [41]. We consider the setting where T may be finite. Let $T \in \mathbb{N}$, H_T be an Hilbert space and $(\varphi_T(\theta), \theta \in \Theta)$ a dictionary satisfying Assumptions 3.1 and 3.2, so that the kernel $\mathcal{K}_T$ is of class $\mathcal{C}^{3,3}$ on Θ^2. Recall the Riemannian metric $\mathfrak{d}_{\mathcal{K}_T}$ associated to $\mathcal{K}_T$, which we simply denote by $\mathfrak{d}_T$. We define 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 $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. For $r > 0$, the near region of $\mathcal{Q}^\star$ is the union of balls $\bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ and its far region is the complementary of the near region in Θ_T: $\Theta_T \setminus \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$. Sufficient conditions for the next assumption to hold are given in Section 7.

Assumption 6.1 (Interpolating certificate)**.** *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 3.1 and 3.2 on the dictionary $(\varphi_T(\theta), \theta \in \Theta)$, and Assumption 5.1 on the kernel $\mathcal{K}_\infty$, defined on Θ^2, hold. Suppose that $\mathfrak{d}_T(\theta, \theta') > 2r$ for all $\theta, \theta' \in \mathcal{Q}^\star \subset \Theta_T$, and that 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 application $v : \mathcal{Q}^\star \to \{-1, 1\}$ there exists an element $p \in H_T$ satisfying:*

- *(i) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $|\langle \phi_T(\theta), p\rangle_T| \leq 1 - C_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$.*
- *(ii) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $|\langle \phi_T(\theta), p\rangle_T - v(\theta^\star)| \leq C'_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$.*
- *(iii) 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 $|\langle \phi_T(\theta), p\rangle_T| \leq 1 - C_F$.*
- *(iv) We have $\|p\|_T \leq C_B \sqrt{s}$.*

The function $\eta : \theta \mapsto \langle \phi_T(\theta), p\rangle_T$ is the so-called "interpolating certificate" of the function v, as thanks to *(ii)* with $\theta = \theta^\star$, the function η coincides with the function v on $\mathcal{Q}^\star$. In addition, the interpolating certificate is required to have curvature properties in the near region and to be bounded by a constant strictly inferior to 1 in the far region. When r is sufficiently small (that is, $r \leq \sqrt{2/(C_N + C_N')}$) Conditions *(i)* and *(ii)* are equivalent to the fact that the function η is in-between two quadratic functions in the near region of $\mathcal{Q}^\star$: for all $\theta^\star \in \mathcal{Q}^\star$ such that $v(\theta^\star) = 1$ (resp. $v(\theta^\star) = -1$) and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $1 - C_N' \mathfrak{d}_T(\theta^\star, \theta)^2 \leq \eta(\theta) \leq 1 - C_N \mathfrak{d}_T(\theta^\star, \theta)^2$ (resp. $-1 + C_N \mathfrak{d}_T(\theta^\star, \theta)^2 \leq \eta(\theta) \leq -1 + C_N' \mathfrak{d}_T(\theta^\star, \theta)^2$).

Sufficient conditions for the next assumption to hold are also given in Section 7.

Assumption 6.2 (Interpolating derivative certificate)**.** *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 3.1 and 3.2 on the dictionary $(\varphi_T(\theta), \theta \in \Theta)$, and Assumption 5.1 on the kernel $\mathcal{K}_\infty$, defined on Θ^2, hold. Assume that $\mathfrak{d}_T(\theta, \theta') > 2r$ for all $\theta, \theta' \in \mathcal{Q}^\star \subset \Theta_T$ and that there exist finite positive constants c_N, c_F, c_B depending on r and $\mathcal{K}_\infty$, such that for any application $v : \mathcal{Q}^\star \to \{-1, 1\}$ there exists an element $q \in H_T$ satisfying:*

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

$$|\langle \phi_T(\theta), q\rangle_T - v(\theta^\star) \operatorname{sign}(\theta - \theta^\star)\mathfrak{d}_T(\theta, \theta^\star)| \leq c_N \, \mathfrak{d}_T(\theta^\star, \theta)^2.$$

(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 $|\langle \phi_T(\theta), q\rangle_T| \leq c_F$.

(iii) $||q||_T \leq c_B \sqrt{s}$.

The function $\theta \mapsto \langle \phi_T(\theta), q\rangle_T$ will be called an "interpolating derivative certificate" as it vanishes on $\mathcal{Q}^\star$. In addition, this function is required to decrease similarly to the function $\mathfrak{d}_T(\cdot, \theta^\star)$ near $\theta^\star$ and to be bounded in the far region of $\mathcal{Q}^\star$.

7. Sufficient conditions for the existence of certificates

In this section, we prove the existence of the certificate functions of Assumptions 6.1 and 6.2 provided that the parameters to be estimated are sufficiently separated in terms of the Riemannian metric. According to [45], the separation condition cannot be avoided to build certificate functions in general. It is however possible to remove this separation condition in some particular cases, see [44] for models with positive amplitudes.

In order to find sufficient conditions for the existence of the interpolating certificate functions of Assumption 6.1, we extend the construction from [41] to a non asymptotic setting. For the existence of the interpolating derivative certificate functions of Assumption 6.2, we generalize the proof of [17, Lemma 2.7] dedicated to the dictionary of complex exponential functions. The proofs for the existence of certificates given in Section 10 require boundedness and local concavity properties of the kernel $\mathcal{K}_T$. For practical application, they are deduced from the boundedness and local concavity properties of the kernel $\mathcal{K}_\infty$ and the quality of approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$ discussed in Section 5.

7.1. Boundedness and local concavity of the kernel $\mathcal{K}_T$.

In this work, we shall consider bounded kernels locally concave on the diagonal. More precisely, for $T \in \bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$ and $r > 0$, we define:

$$\varepsilon_T(r) = 1 - \sup \{|\mathcal{K}_T(\theta, \theta')|; \quad \theta, \theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta', \theta) \geq r\}, \tag{39}$$

$$\nu_T(r) = -\sup \left\{\mathcal{K}_T^{[0,2]}(\theta, \theta'); \quad \theta, \theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta', \theta) \leq r\right\}. \tag{40}$$

The fact that $\varepsilon_T(r)$ and $\nu_T(r)$ are positive depends on the function φ_T, the space H_T and the set Θ_T. Let us mention that in many examples the positiveness of $\varepsilon_\infty(r)$ and $\nu_\infty(r)$ is easy to check whereas the positiveness of $\varepsilon_T(r)$ and $\nu_T(r)$ might be more difficult to prove.

Notice that (33) for $T \in \mathbb{N}$ and Assumption 5.1 for $T = \infty$, and the continuity of $\mathcal{K}_T$ and $\mathcal{K}_T^{[0,2]}$ give that:

$$\lim_{r \to 0^+} \varepsilon_T(r) = 0 \quad \text{and} \quad \lim_{r \to 0^+} \nu_T(r) = 1. \tag{41}$$

Recall ρ_T and $\mathcal{V}_T$ defined in (36) and (38). The next lemmas state that if $\varepsilon_\infty(r/\rho_T)$ (resp. $\nu_\infty(r\rho_T)$) is positive and if the approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$ is good, *i.e.* $\mathcal{V}_T$ is small, then $\varepsilon_T(r)$ (resp. $\nu_T(r)$) is also positive.

Lemma 7.1. *Let $T \in \mathbb{N}$. Suppose Assumptions 3.1, 3.2 and 5.1 hold. Then we have for $r > 0$:*

$$\varepsilon_T(r) \geq \varepsilon_\infty(r/\rho_T) - \mathcal{V}_T \quad \textit{and} \quad \nu_T(r) \geq \nu_\infty(r\rho_T) - \mathcal{V}_T.$$

Proof. As Assumptions 3.2 and 5.1 hold, recall that $\mathfrak{d}_\infty/\rho_T \leq \mathfrak{d}_T \leq \rho_T \mathfrak{d}_\infty$ on Θ_T^2, see (37).

Let $\theta, \theta' \in \Theta_T$ such that $\mathfrak{d}_T(\theta', \theta) \geq r$. We have $\mathfrak{d}_\infty(\theta', \theta) \geq r/\rho_T$. We get from the definition of $\mathcal{V}_T$ that:

$$|\mathcal{K}_T(\theta, \theta')| \leq |\mathcal{K}_\infty(\theta, \theta')| + \mathcal{V}_T \leq 1 - \varepsilon_\infty(r/\rho_T) + \mathcal{V}_T.$$

Then, use (39) to get $\varepsilon_T(r) \geq \varepsilon_\infty(r/\rho_T) - \mathcal{V}_T$. We also have $\mathfrak{d}_\infty(\theta', \theta) \leq r\rho_T$. We deduce that:

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

Finally, using (40), we obtain $\nu_T(r) \geq \nu_\infty(r\rho_T) - \mathcal{V}_T$. □

When we require in addition of the assumptions of Lemma 7.1 that $\varepsilon_\infty(r/\rho_T) \wedge \nu_\infty(r\rho_T) > \mathcal{V}_T \geq 0$, then we have $\varepsilon_T(r) > 0$ and $\nu_T(r) > 0$.

7.2. **Separation conditions for the non-linear parameters.** In what follows, we measure the interferences (or the overlap) between the features in the mixture through a quantity δ_T introduced in [41] and defined below. Let $T \in \bar{\mathbb{N}}$, $\delta > 0$ and $s \in \mathbb{N}^*$. 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:

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

Using the convention $\inf \emptyset = +\infty$, we set for $u > 0$:

$$\delta_T(u, s) = \inf \Big\{\delta > 0 : \max_{1 \leq \ell \leq s} \sum_{k=1, k\neq \ell}^{s} |\mathcal{K}_T^{[i,j]}(\theta_\ell, \theta_k)| \leq u \\ \text{for all } (i,j) \in \{0,1\} \times \{0,1,2\} \text{ and } (\theta_1, \cdots, \theta_s) \in \Theta_{T,\delta}^s\Big\}. \tag{42}$$

The quantity $\delta_T(u, s)$ is the minimum distance (with respect to the Riemannian metric $\mathfrak{d}_T$) between s parameters so that the coherence of the associated dictionary is bounded by u. The notion of coherence between the features in the definition of $\delta_T(u, s)$ is quite similar to the one used in compressed sensing (see [30, Section 5]). A standard problem in compressed sensing is to retrieve the vector $\beta^\star$ when the multivariate function $\Phi_T(\vartheta^\star)$ is known in the discrete setting of Example 1.1 or Section 1.2.1. In this framework, the matrix $\Phi_T(\vartheta^\star)$, whose rows correspond to the K discretized functions in the dictionary, is known. The coherence is defined as $\max_{1\leq k\neq \ell \leq K} |\mathcal{K}_T(\theta_k^\star, \theta_\ell^\star)|$. Usually, the smaller the coherence, the easier it is to retrieve the parameter $\beta^\star$. The Babel function, introduced in [49], is even closer to our measure of overlap. We refer to [41] for a discussion on this function.

Remark 7.2 (Rewriting the separation condition with operator norm). We shall stress that the definition of δ_T in (42) is related to the operator norm $\|\cdot\|_{\text{op}}$ associated to the ℓ_∞ norm on $\mathbb{R}^s$. We restate (42) using this operator norm $\|\cdot\|_{\text{op}}$, and leave the interested reader to check that another choice of operator norm does not improve the bounds on the certificates. 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:

$$\mathcal{K}_T^{[i,j]}(\vartheta) = \Big(\mathcal{K}_T^{[i,j]}(\theta_k, \theta_\ell)\Big)_{1\leq k,\ell \leq s}. \tag{43}$$

Let I be the identity matrix of size $s \times s$. For $i = 0$ or $i = 1$, since the diagonal coefficients of $\mathcal{K}_T^{[i,i]}(\vartheta)$ are equal to 1, see (29), we get:

$$\left\| I - \mathcal{K}_T^{[i,i]}(\vartheta) \right\|_{\text{op}} = \max_{1 \leq k \leq s} \sum_{\ell \neq k} |\mathcal{K}_T^{[i,i]}(\theta_k, \theta_\ell)|.$$

Since the diagonal coefficients of $\mathcal{K}_T^{[1,0]}(\vartheta)$, $\mathcal{K}_T^{[0,1]}(\vartheta)$ and $\mathcal{K}_T^{[1,2]}(\vartheta)$ are zero, see (33), we also get:

$$\left\| \mathcal{K}_T^{[1,0]}(\vartheta) \right\|_{\text{op}} = \max_{1 \leq k \leq s} \sum_{\ell \neq k} |\mathcal{K}_T^{[1,0]}(\theta_k, \theta_\ell)| \quad \text{and} \quad \left\| \mathcal{K}_T^{[1,2]}(\vartheta) \right\|_{\text{op}} = \max_{1 \leq k \leq s} \sum_{\ell \neq k} |\mathcal{K}_T^{[1,2]}(\theta_k, \theta_\ell)|$$

and by symmetry, with $\|\cdot\|_{\text{op}}^*$ for the operator norm associated to the ℓ_1 norm:

$$\left\| \mathcal{K}_T^{[0,1]}(\vartheta) \right\|_{\text{op}} = \left\| \mathcal{K}_T^{[1,0]\top}(\vartheta) \right\|_{\text{op}} = \left\| \mathcal{K}_T^{[1,0]}(\vartheta) \right\|_{\text{op}}^* = \max_{1 \leq \ell \leq s} \sum_{k \neq \ell} |\mathcal{K}_T^{[1,0]}(\theta_k, \theta_\ell)| = \max_{1 \leq k \leq s} \sum_{\ell \neq k} |\mathcal{K}_T^{[0,1]}(\theta_k, \theta_\ell)|.$$

Since the diagonal coefficients of $\mathcal{K}_T^{[2,0]}(\vartheta)$ are equal to -1, see (33), we also get:

$$\left\| I + \mathcal{K}_T^{[2,0]}(\vartheta) \right\|_{\text{op}} = \max_{1 \leq k \leq s} \sum_{\ell \neq k} |\mathcal{K}_T^{[2,0]}(\theta_k, \theta_\ell)|.$$

Thus, we have:

$$\delta_T(u, s) = \inf \left\{ \delta > 0 : \; A_{T,\ell_\infty}(\vartheta) \leq u, \vartheta \in \Theta_{T,\delta}^s \right\}, \tag{44}$$

where:

$$A_{T,\ell_\infty}(\vartheta) = \tag{45}$$
$$\max \left(\left\| I - \mathcal{K}_T^{[0,0]}(\vartheta) \right\|_{\text{op}}, \left\| I - \mathcal{K}_T^{[1,1]}(\vartheta) \right\|_{\text{op}}, \left\| I + \mathcal{K}_T^{[2,0]}(\vartheta) \right\|_{\text{op}}, \left\| \mathcal{K}_T^{[1,0]}(\vartheta) \right\|_{\text{op}}, \left\| \mathcal{K}_T^{[0,1]}(\vartheta) \right\|_{\text{op}}, \left\| \mathcal{K}_T^{[1,2]}(\vartheta) \right\|_{\text{op}} \right).$$

Lemma 7.3 below enables us to compare the separation distance at T fixed and at the limit case where $T = +\infty$. Recall that the constant ρ_T is defined in (36).

Lemma 7.3. *Let $T \in \bar{\mathbb{N}}$ and $s \in \mathbb{N}^*$. Suppose Assumptions 3.1, 3.2 and 5.1 hold. Then, for $u > 0$ and with:*

$$u_T(s) = u + (s-1)\mathcal{V}_T,$$

we have:

$$\delta_T(u_T(s), s) \leq \rho_T \, \delta_\infty(u, s) \quad \textit{and} \quad \Theta_{T, \rho_T \, \delta_\infty(u,s)}^s \subseteq \Theta_{T, \delta_T(u_T(s), s)}^s.$$

Proof. Since Assumptions 3.2 and 5.1 hold, we have from (37) that $\mathfrak{d}_T \leq \rho_T \, \mathfrak{d}_\infty$ on Θ_T^2. Hence for any $\delta > 0$, we have the inclusion $\Theta_{T,\rho_T \, \delta}^s \subseteq \Theta_{\infty,\delta}^s$. In particular, we have for $u > 0$ that $\Theta_{T,\rho_T \, \delta_\infty(u,s)}^s \subseteq \Theta_{\infty,\delta_\infty(u,s)}^s$. Using the triangle inequality and the definition of $\mathcal{V}_T$ in (38), we have that for $(i,j) \in \{0,1\} \times \{0,1,2\}$ and $(\theta_1, \cdots, \theta_s) \in \Theta_T^s$:

$$\sum_{k=1, k \neq \ell}^{s} |\mathcal{K}_T^{[i,j]}(\theta_\ell, \theta_k)| \leq \sum_{k=1, k \neq \ell}^{s} \left(|\mathcal{K}_\infty^{[i,j]}(\theta_\ell, \theta_k)| + \mathcal{V}_T \right).$$

Then, the inclusion $\Theta_{T,\rho_T \, \delta_\infty(u,s)}^s \subseteq \Theta_{\infty,\delta_\infty(u,s)}^s$ gives that for all $(i,j) \in \{0,1\} \times \{0,1,2\}$ and $(\theta_1, \cdots, \theta_s) \in \Theta_{T,\rho_T \, \delta_\infty(u,s)}^s$:

$$\sum_{k=1, k \neq \ell}^{s} |\mathcal{K}_T^{[i,j]}(\theta_\ell, \theta_k)| \leq u + (s-1)\mathcal{V}_T.$$

With $u_T(s) = u + (s-1)\mathcal{V}_T$, we deduce that $\delta_T(u_T(s), s) \leq \rho_T \, \delta_\infty(u,s)$, which proves the inclusion $\Theta_{T,\rho_T \, \delta_\infty(u,s)}^s \subseteq \Theta_{T,\delta_T(u_T(s),s)}^s$. $\square$

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

$$
\begin{aligned}
H^{(1)}_\infty(r,\rho) &= \frac{1}{2} \wedge L_{2,0} \wedge L_{2,1} \wedge \frac{\nu_\infty(\rho r)}{10} \wedge \frac{\varepsilon_\infty(r/\rho)}{10},\\
H^{(2)}_\infty(r,\rho) &= \frac{1}{6} \wedge \frac{8\,\varepsilon_\infty(r/\rho)}{10(5+2L_{1,0})} \wedge \frac{8\,\nu_\infty(\rho r)}{9(2L_{2,0}+2L_{2,1}+4)},
\end{aligned}
\tag{46}
$$

where the constants involved are defined in (35). By recalling the behaviors of $\varepsilon_\infty(r)$ and $\nu_\infty(r)$ when r goes down to zero from (41), we have for $\rho \geq 1$:

$$
\lim_{r\to 0^+} H^{(1)}_\infty(r,\rho) = 0 \quad \text{and} \quad \lim_{r\to 0^+} H^{(2)}_\infty(r,\rho) = 0.
$$

We state the first main result of this section whose proof is given in Section 10.

Proposition 7.4 (Interpolating certificate)**.** *Let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $\rho \geq 1$ and $r > 0$. We assume that:*

(i) ***Regularity of the dictionary φ_T:*** *Assumptions 3.1 and 3.2 hold.*

(ii) ***Regularity of the limit kernel $\mathcal{K}_\infty$:*** *Assumption 5.1 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 $u_\infty \in \left(0, H^{(2)}_\infty(r,\rho)\right)$ 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 $\mathcal{K}_\infty$:***

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

Then, with the positive constants:

$$
C_N = \frac{\nu_\infty(\rho r)}{180}, \quad C'_N = \frac{5}{8}L_{2,0} + \frac{1}{8}L_{2,1} + \frac{1}{2}, \quad C_B = 2 \quad \text{and} \quad C_F = \frac{\varepsilon_\infty(r/\rho)}{10} \leq 1,
\tag{47}
$$

Assumption 6.1 holds (with the same r) for any subset $\mathcal{Q}^\star = \{\theta^\star_i, 1 \leq i \leq s\}$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$:

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

Note that *(i)* concerns the dictionary φ_T, *(ii)* and *(iii)* the limit kernel $\mathcal{K}_\infty$ and the set of parameters, and *(iv)* and *(v)* the regime for the parameters s and T. Notice that if $\mathcal{K}_\infty$ is chosen equal to $\mathcal{K}_T$, then $\mathcal{V}_T = 0$ and $\rho_T = 1$, and also *(iv)* and *(v)* hold and ρ can be chosen equal to 1.

We now give the second main result of this section whose proof is given in Section 10.2.

Proposition 7.5 (Interpolating derivative certificate)**.** *Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. We assume that:*

(i) ***Regularity of the dictionary φ_T:*** *Assumptions 3.1 and 3.2 hold.*

(ii) ***Regularity of the limit kernel $\mathcal{K}_\infty$:*** *Assumption 5.1 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:

$$
c_N = \frac{1}{8}L_{2,0} + \frac{5}{8}L_{2,1} + \frac{7}{8}, \quad c_B = 2 \quad \text{and} \quad c_F = \frac{5}{4}L_{1,0} + \frac{7}{4},
\tag{48}
$$

Assumption 6.2 holds for any $r > 0$ and any subset $\mathcal{Q}^\star = \{\theta^\star_i, 1 \leq i \leq s\}$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$:

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

Let us briefly indicate how the certificates are constructed in Section 10 using the features of the dictionary. Let $\alpha = (\alpha_1, \ldots, \alpha_s)$ and $\xi = (\xi_1, \ldots, \xi_s)$ be elements of $\mathbb{R}^s$. Let $p_{\alpha,\xi} \in H_T$ be defined by:

$$p_{\alpha,\xi} = \sum_{k=1}^{s} \alpha_k \phi_T(\theta_k^\star) + \sum_{k=1}^{s} \xi_k \, \phi_T^{[1]}(\theta_k^\star),$$

where $\phi_T^{[1]}$ denotes the derivative $\tilde{D}_{1;T}[\phi_T]$. Using (32) in Lemma 4.3, set the interpolating real-valued function $\eta_{\alpha,\xi}$ defined on Θ by:

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

By Assumption 3.2 on the regularity of φ_T and the positivity of g_T and Lemma 4.3, we get that the function $\eta_{\alpha,\xi}$ is of class $\mathcal{C}^3$ on Θ, and using (24), we get that:

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

We show in Section 10 that for any function $v : \mathcal{Q}^\star \to \{-1, 1\}$ there exists a unique choice of α and ξ such that $\eta_{\alpha,\xi}$ becomes an interpolating certificate, that is, $\eta_{\alpha,\xi} = v$ and $\eta_{\alpha,\xi}^{[1]} = 0$ on $\mathcal{Q}^\star$, and $p_{\alpha,\xi}$ satisfies Points *(i)*-*(iv)* of Assumption 6.1.

Moreover, for any function $v : \mathcal{Q}^\star \to \{-1, 1\}$ there exists another unique choice of α and ξ such that $\eta_{\alpha,\xi}$ is an interpolating derivative certificate, that is, $\eta_{\alpha,\xi} = 0$ and $\eta_{\alpha,\xi}^{[1]} = v$ on $\mathcal{Q}^\star$, and $p_{\alpha,\xi}$ satisfies Points *(i)*-*(iii)* of Assumption 6.2.

8. Sparse spike deconvolution

We develop here in full details the particular example of a mixture of Gaussian features observed in a discrete regression model with regular design. In particular, we check the numerous but not very restrictive assumptions, and we illustrate that our general and more restrictive sufficient conditions for the existence of certificates can turn simpler and far less restrictive on concrete examples. The model is presented in Section 8.1, where we also check the first assumptions. The technical Section 8.2 on the existence of the certificates allows to point out the separation distance in (55) and with the simpler expression in (56). This separation distance is usually very pessimistic, but one can rely on numerical estimations to be more realistic, see Remark 8.2 in this direction. Eventually, we apply to this context our main Theorem 2.1 in Section 8.3 as Corollary 8.3 and illustrate a particular choice of the tuning parameter in Remark 8.4 in the spirit of [47, 12] established for the specific dictionary of complex exponentials.

8.1. **Model and first assumptions of Theorem 2.1.** Consider a real-valued process y observed over a regular grid $t_1 < \cdots < t_T$ of a symmetric interval $[a_T, b_T]$, with $T \geq 2$, $a_T = -b_T > 0$, $t_j = a_T + j\Delta_T$ for $j = 1, \ldots, T$ and grid step:

$$\Delta_T = \frac{b_T - a_T}{T} \cdot$$

Assuming that all the observations have the same weight amounts to considering y as an element of the Hilbert space $H_T = L^2(\lambda_T)$ of real valued functions defined in $\mathbb{R}$ and square integrable with respect to the atomic measure λ_T on $\{t_1, \ldots, t_T\}$:

$$\lambda_T(\mathrm{d}t) = \Delta_T \sum_{j=1}^{T} \delta_{t_j}(\mathrm{d}t).$$

We consider a noise process $w_T(t) = \sum_{j=1}^T G_j \mathbf{1}_{\{t_j=t\}}$ for $t \in \mathbb{R}$, where $(G_1, \ldots, G_T)$ is a centered Gaussian vector such that, for some noise level $\sigma_1 > 0$:

$$\mathbb{E}[G_j^2] = \sigma_1^2 \quad \text{and} \quad |\mathbb{E}[G_j G_i]| \leq \sigma_1^2/T \quad \text{for } j \neq i \text{ in } \{1,\ldots,\mathrm{T}\}.$$

Thus, the norm of the noise $\|w_T\|_T$ is finite almost surely, and for any $f \in L^2(\lambda_T)$ we have:

$$\mathrm{Var}(\langle f, w_T\rangle_T) = \mathrm{Var}\Big(\Delta_T \sum_{j=1}^T f(t_j) G_j\Big) \leq 2\sigma_1^2 \Delta_T \|f\|_T^2.$$

Hence, Assumption 1.1 on the noise is satisfied with $\sigma^2 = 2\sigma_1^2$. (Notice that if the random variables $G_1, \ldots, G_T$ are independent, then $\mathrm{Var}(\langle f, w_T\rangle_T) = \sigma^2 \Delta_T \|f\|_T^2$ with $\sigma^2 = \sigma_1^2$.) This gives that Point *(i)* of Theorem 2.1 holds.

We consider the dictionary given by the translation model of Section 3.3.1 with Gaussian features and fixed scaling parameter $\sigma_0 > 0$, that is the dictionary does not depend on T and is given by:

$$\Big(\varphi(\theta) = k\Big(\frac{\cdot - \theta}{\sigma_0}\Big), \theta \in \Theta\Big) \quad \text{with} \quad k(t) = \mathrm{e}^{-t^2/2} \quad \text{and} \quad \Theta = \mathbb{R}.$$

Thus, the signal $\beta^\star \Phi(\vartheta^\star)$ in model (2) can indeed be written as the convolution product of the function k and an atomic measure. It is elementary to check that Assumption 3.1 on the regularity of the features holds. Furthermore, the functions $\varphi(\theta)$ and $\partial_\theta \varphi(\theta)$ are linearly independent $\lambda_T - a.e$ for all $\theta \in \Theta$ as $T \geq 2$. Hence the function g_T is positive on Θ by Lemma 3.1 and thus Assumption 3.2 holds. This gives that Point *(ii)* of Theorem 2.1 holds.

We now define the limit kernel $\mathcal{K}_\infty$. To do so, we shall assume that $(b_T, T \geq 2)$ is a sequence of positive numbers, such that:

(49) $$\lim_{T\to\infty} b_T = +\infty \quad \text{and} \quad \lim_{T\to\infty} \Delta_T = 0.$$

This in particular implies that the sequence of measures $(\lambda_T, T \geq 2)$ converges with respect to the vague topology towards the Lebesgue measure, say λ_∞, on $\Theta_\infty = \mathbb{R}$. We also consider the Hilbert space $H_\infty = L^2(\lambda_\infty)$ endowed with its usual scalar product denoted $\langle \cdot, \cdot \rangle_\infty$ and corresponding norm denoted $\|\cdot\|_\infty$ (not to be confused with the supremum norm!). Note that the kernel $\mathcal{K}_T$ and the associated quantities such as ε_T and ν_T defined in (39) and (40), respectively, or the uniform bounds on $\mathcal{K}_T^{[i,j]}$, are difficult to calculate. However the uniform bounds on $\Theta_\infty = \mathbb{R}$ for the kernel $\mathcal{K}_\infty$, defined by (30) with T replaced by ∞, are easily computed. Elementary calculations give for $\theta, \theta' \in \Theta$:

$$\|\varphi(\theta)\|_\infty^2 = \sqrt{\pi}\,\sigma_0, \quad \phi_\infty(\theta) = \frac{1}{\pi^{\frac{1}{4}}\sqrt{\sigma_0}}\varphi(\theta), \quad \mathcal{K}_\infty(\theta,\theta') = k\Big(\frac{\theta - \theta'}{\sqrt{2}\,\sigma_0}\Big) \quad \text{and} \quad g_\infty(\theta) = \frac{1}{2\sigma_0^2}.$$

In particular, we have $g'_\infty(\theta) = 0$. The Riemannian metric is equal to the Euclidean distance up to a multiplicative factor, for all $\theta, \theta' \in \Theta_\infty = \mathbb{R}$:

(50) $$\mathfrak{d}_\infty(\theta, \theta') = \frac{|\theta - \theta'|}{\sqrt{2}\,\sigma_0}.$$

We see that $\mathcal{K}_\infty$ is of class $\mathcal{C}^{\infty,\infty}$ and that:

(51) $$\mathcal{K}_\infty^{[i,j]}(\theta,\theta') = (-1)^j\, k^{(i+j)}\Big(\frac{\theta - \theta'}{\sqrt{2}\,\sigma_0}\Big) \quad \text{and} \quad k^{(i)}(t) = P_i(t)\, k(t),$$

where we give for convenience the formulae for some of the polynomials P_i:

$$P_1(t) = -t, \quad P_2(t) = -1 + t^2, \quad P_3(t) = 3t - t^3, \quad P_4(t) = 3 - 6t^2 + t^4, \quad P_6(t) = -15 + 45t^2 - 15t^4 + t^6.$$

Then, we explicitly compute the constants $L_{i,j}$ for $i, j \in \{0, \cdots, 2\}$ and L_3 defined in (35):

$$m_g = (2\sigma_0^2)^{-1}, \quad L_{0,0} = 1, \quad L_{1,0} = L_{0,1} = \mathrm{e}^{-1/2}, \quad L_{1,1} = L_{2,0} = L_{0,2} = 1,$$
$$L_{2,1} = L_{1,2} = \sqrt{18 - 6\sqrt{6}}\, \mathrm{e}^{\sqrt{3/2}-3/2} \leq \sqrt{2}, \quad L_{2,2} = 3 \quad \text{and} \quad L_3 = 15.$$

Notice the constants $L_{i,j}$ and L_3 do not depend on the scaling factor σ_0. Thus Assumption 5.1 holds. This gives that Point (*iii*) of Theorem 2.1 holds.

We now check the proximity of the kernel $\mathcal{K}_T$ to the limit kernel $\mathcal{K}_\infty$. The support of λ_T is spread over the window $[a_T, b_T]$ where the signal is observed. Hence it is legitimate to look for the location parameters on a smaller subset of this window, and thus restrict the optimization (4) to the compact set:

$$\Theta_T = [(1-\epsilon)a_T, (1-\epsilon)b_T] \subset [a_T, b_T] \quad \text{with a given shrinkage } \epsilon > 0.$$

The proof of the next lemma is given in Section A.5. Recall ρ_T and $\mathcal{V}_T$ defined in (36) and (38). Set:

$$\gamma_T = 2\Delta_T\, \sigma_0^{-1} + \sqrt{\pi}\, \mathrm{e}^{-\epsilon^2 b_T^2/2\sigma_0^2}.$$

Lemma 8.1. *There exist finite positive universal constants c_0, c_1 and c_2, such that $\gamma_T < c_0$ implies:*

$$\mathcal{V}_T \leq c_1\gamma_T \quad \textit{and} \quad |1-\rho_T| \leq c_2\gamma_T. \tag{52}$$

This implies that Assumption 5.2 holds for T such that $\gamma_T \leq c_0$ and $c_1\gamma_T \leq 3$, which holds for T large enough thanks to (49). Thus Point (*iv*) of Theorem 2.1 holds for T large enough.

8.2. **Existence of certificates.** We keep the model and the notations from Section 8.1. In order to get the prediction error from Theorem 2.1, we only need to check that Point (*iv*) therein on the existence of the certificates holds. To check the existence of the certificates, we can use Propositions 7.4 and 7.5, and check that all the hypotheses required in those two propositions hold.

We first concentrate on the hypotheses of Proposition 7.4. Assumption (*i*) on the regularity of the dictionary holds, see Section 8.1.

We recall that $L_{0,2} = 1$ and thus $1/\sqrt{2L_{0,2}} = 1/\sqrt{2} > 1/2$. Recall $\varepsilon_\infty(r)$ and $\nu_\infty(r)$ defined in (39) and (40), and thanks to the explicit form of the Riemannan metric, we get for $r \in (0,1)$:

$$\varepsilon_\infty(r) = 1 - \mathrm{e}^{-r^2/2} > 0 \quad \text{and} \quad \nu_\infty(r) = \left(1 - r^2\right)\mathrm{e}^{-r^2/2}.$$

This and the regularity of the kernel $\mathcal{K}_\infty$ from Section 8.1 imply that Assumption (*ii*) holds for all $r \in (0, 1/(\rho \vee \sqrt{2}))$.

We obtain from (51) that $\lim_{q\to\infty} \sup_{|\theta-\theta'|\geq q} |\mathcal{K}_\infty^{[i,j]}(\theta,\theta')| = 0$ for all $i, j \in \{0,1,2\}$. Thus, we deduce from the definition (42) of δ_∞ that $\delta_\infty(u,s)$ is finite for all $s \in \mathbb{N}^*$ and $u > 0$. This implies that Assumption (*iii*) on the separation of the parameters holds.

To simplify, we set $\rho = 2$ (but we could take any value of $\rho > 1$). We deduce from Lemma 8.1, that for T large enough $\rho_T \leq \rho = 2$, and thus Assumption (*iv*) on the closeness of the metrics $\mathfrak{d}_T$ and $\mathfrak{d}_\infty$ holds.

Recall the definition of $H_\infty^{(1)}$ and $H_\infty^{(2)}$ from (46). To get the smallest separation distance, we also set:

$$r = \underset{0\leq r'\leq 1/2}{\operatorname{argmax}}\ H_\infty^{(2)}(r', \rho) \approx 0.49. \tag{53}$$

Notice that the function is not *a priori* monotone in ρ. We have $\varepsilon_\infty(r/2) \approx 2.9\times 10^{-2}$, $\nu_\infty(2r) \approx 3.7\times 10^{-2}$, $H_\infty^{(1)}(r,2) \approx 2.9\times 10^{-3}$ and $H_\infty^{(2)}(r,2) \approx 3.7\times 10^{-3}$. Again in order to get a "small" separation distance, we choose u_∞ close to $H_\infty^{(2)}(r,2)$, say $u_\infty = \eta_0 H_\infty^{(2)}(r,2)$ for some $\eta_0 < 1$ close to 1. For simplicity set

$\eta_0 = 9/10$. Thanks to hypothesis (49), we get $\lim_{T\to\infty} \gamma_T = 0$ and Lemma 8.1 implies that for T large enough, depending on σ_0, ϵ and the sparsity parameter s, we have:

$$\rho_T \leq 2, \quad \mathcal{V}_T \leq H^{(1)}_\infty(r,2) \quad \text{and} \quad (s-1)\mathcal{V}_T \leq (1-\eta_0) H^{(2)}_\infty(r,2), \tag{54}$$

and thus Assumption *(v)* on the proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$ holds.

Thus, the assumptions of Proposition 7.4 are satisfied, and we deduce that Assumption 6.1 holds with, thanks to (47):

$$C_N \approx 2 \times 10^{-4}, \quad C'_N \approx 1.3, \quad C_B = 2 \quad \text{and} \quad C_F \approx 2.9 \times 10^{-3}.$$

We now concentrate on the hypotheses of Proposition 7.5. Assumptions *(i)*-*(iii)* clearly hold for the same reasons as Assumptions *(i)*-*(iii)* of Proposition 7.4.

Again in order to get a "small" separation distance, there is no need to choose u'_∞ larger that u_∞, and for this reason we take $u'_\infty = u_\infty$. We deduce from (54) that for T large enough, depending on σ_0, ϵ and the sparsity parameter s:

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

and thus Assumption *(iv)* on the proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$ holds.

Thus, the assumptions of Proposition 7.5 are satisfied, and we deduce, thanks to (48), that Assumption 6.2 holds with the same value of r given by (53):

$$c_N \approx 1.9, \quad c_B = 2, \quad \text{and} \quad c_F \approx 2.6.$$

In conclusion, we get that Assumptions 6.1 and 6.2 hold for T large enough, and thus Point *(v)* of Theorem 2.1 holds for T large enough and $\mathcal{Q}^\star$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$ the distance $\mathfrak{d}_T(\theta,\theta')$ is larger than the separation distance:

$$2\max(r,\, \rho_T\, \delta_\infty(u_\infty, s),\, \rho_T\, \delta_\infty(u'_\infty, s)). \tag{55}$$

Notice that since $u_\infty = u'_\infty$, $\rho_T \mathfrak{d}_T(\theta,\theta') \geq \mathfrak{d}_\infty(\theta,\theta')$ and $\rho_T \leq 2$, we deduce from (50), that a slightly stronger condition is to assume that $|\theta - \theta'|$ is larger than:

$$\sqrt{2}\, \sigma_0 \max(1, 4\delta_\infty(u_\infty, s)). \tag{56}$$

Remark 8.2 (On the separation distance (55)). The separation distance (55) is a non-decreasing function of s. We now provide an upper bound. Let $(i,j) \in \{0,1\}\times\{0,1,2\}$. By considering the kernel $\mathcal{K}_T$ and its derivative given by (51) and the bound $M = \max_{0\leq i\leq 3} \sup |P_i| \sqrt{k}$, we deduce that $|\mathcal{K}^{[i,j]}_\infty(\theta,\theta')| \leq M\, \mathrm{e}^{-\mathfrak{d}_\infty(\theta,\theta')^2/2}$ for all $\theta,\theta' \in \Theta$. We easily obtain that for $\vartheta = (\theta_1, \cdots, \theta_s) \in \Theta^s_{\infty,\delta}$ with $\delta > 0$:

$$\max_{1\leq \ell\leq s} \sum_{k=1, k\neq \ell}^{s} |\mathcal{K}^{[i,j]}_\infty(\theta_\ell, \theta_k)| \leq \psi_s(\delta) \quad \text{with} \quad \psi_s(\delta) = 2M \int_0^{s/2+1} \mathrm{e}^{-t^2\delta^2/4}\, \mathrm{d}t.$$

The function ψ_s is decreasing and one to one from $\mathbb{R}_+$ to $(0, M(s+2)]$. Setting $\psi_s^{-1}(u) = 0$ for $u > M(s+2)$, we deduce from (42) that for $u > 0$:

$$\delta_\infty(u,s) \leq \psi_s^{-1}(u).$$

Since the map $s \mapsto \psi_s(\delta)$ is increasing with limit $\psi_\infty(\delta) = 2\sqrt{\pi}\, M/\delta$, we deduce that for $s \in \mathbb{N}^*$:

$$\delta_\infty(u,s) \leq \frac{2\sqrt{\pi}\, M}{u},$$

so that the separation distance (55) (or (56)) can be bounded uniformly in s for given r and $u_\infty = u'_\infty$.

In fact, we shall illustrate for $s = 2$ that the separation distance (55) is largely overestimated. We can compute $\delta_\infty(u, s)$ thanks to its expression (44). For $s = 2$ and with the values chosen in this section for $u_\infty = u'_\infty$, we obtain $\delta_\infty(u_\infty, 2) \approx 4.5$. We deduce that the separation distance (55) expressed with respect to the metric $\mathfrak{d}_T$ is approximately $9\,\rho_T$ (which gives $13\,\sigma_0\rho_T^2$ in terms of the Euclidean metric), which is unconveniently large. However, a detailed numerical approach (using the very certificates provided in the proof of Propositions 7.4 and 7.5) with T large so that the kernel $\mathcal{K}_T$ is indeed well approximated by $\mathcal{K}_\infty$ (and thus $\rho_T \approx 1$), gives that one can take for $s = 2$ the separation distance with respect to the Euclidean metric equal to $3.1 \times \sigma_0$ (that is approximately equal to 2.2 with respect to the metric $\mathfrak{d}_\infty$), which is much more realistic. Therefore, the theoretical separation distance (55) is in general largely overestimated.

8.3. **Prediction error.** We keep the model and the notations from Section 8.1 and the values chosen in Section 8.2. We deduce from Theorem 2.1 the following result.

Corollary 8.3. *For T large enough, depending on σ_0, ϵ and the sparsity parameter s, such that (54) holds and for all $\theta \neq \theta' \in \mathcal{Q}^\star = \{\theta^\star_k,\, k \in S^\star\}$, with $S^\star = \mathrm{Supp}(\beta^\star)$ such that $|\theta - \theta'|$ is larger than the separation parameter $\sqrt{2}\,\sigma_0 \max(1, 4\delta_\infty(u_\infty, s))$ given by (56), then, with some universal finite constants $\mathcal{C}_0, ..., \mathcal{C}_3 > 0$, for any $\tau > 0$ and a tuning parameter:*

$$\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log(\tau)}, \tag{57}$$

we have the prediction error bound of the estimators $\hat{\beta}$ and $\hat{\vartheta}$ defined in (4) given by:

$$\sqrt{\Delta_T}\, \left\| \hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_{\ell_2} \leq \mathcal{C}_0 \sqrt{s}\, \kappa, \tag{58}$$

with probability larger than $1 - \mathcal{C}_2 \left(\frac{\sqrt{2} b_T}{\sigma_0 \tau \sqrt{\log(\tau)}} \vee \frac{1}{\tau} \right)$. Moreover, with the same probability, we have that $\left| \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} \right| \leq \mathcal{C}_3 \kappa s$ as well as the inequalities (9) of Theorem 2.5.

The values of the universal constants $\mathcal{C}_i$, $i = 0, \ldots, 3$, can be given explicitly and they are large, but they could be improved numerically.

Remark 8.4 (A particular choice of the tuning parameter). Let $\gamma > 0$ and $\gamma' \geq \gamma$ such that $1 > \gamma' - \gamma$. Set $\tau = T^{\gamma'}$, $b_T = \sigma_0 T^{\gamma' - \gamma} \sqrt{\log(T)}$ and $\kappa = \mathcal{C}_1 \sigma \sqrt{\Delta_T \log(\tau)}$ (which corresponds to the equality in (57)). Then, we get under the assumptions of Corollary 8.3 (and thus T large enough) that:

$$\frac{1}{\sqrt{T}} \left\| \hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_{\ell_2} \leq \mathcal{C}''_0\, \sigma \sqrt{s \frac{\log T}{T}},$$

with probability larger than $1 - \mathcal{C}''_2 / T^\gamma$ where $\mathcal{C}''_0 = \sqrt{\gamma'}\, \mathcal{C}_0\, \mathcal{C}_1$ and $\mathcal{C}''_2 = \sqrt{2/\gamma'}\, \mathcal{C}_2$. Hence, we obtain a similar prediction error bound as the one given in Remark 2.2, see (8). Notice however that in the model and references given in Remark 2.2 , the Riemannian diameter of the parameter set Θ_T is bounded by a constant free of T, whereas in this section it grows (sublinearly) with T without degrading the prediction error bound.

9. Proofs of Theorems 2.1 and 2.5

9.1. **Proof of Theorem 2.1.** Let us bound the prediction error $\hat{R}_T := \left\| \hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T$. By definition (4) of $\hat{\beta}$ and $\hat{\vartheta}$ for the tuning parameter κ, we have:

$$\frac{1}{2} \left\| y - \hat{\beta}\Phi_T(\hat{\vartheta}) \right\|_T^2 + \kappa \|\hat{\beta}\|_{\ell_1} \leq \frac{1}{2} \|y - \beta^\star \Phi_T(\vartheta^\star)\|_T^2 + \kappa \|\beta^\star\|_{\ell_1}.$$

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

$$\hat{\Upsilon}(f) = \left\langle \hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star), f \right\rangle_T .$$

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

$$\frac{1}{2}\hat{R}_T^2 \leq \hat{\Upsilon}(w_T) + \kappa\left(\|\beta^\star\|_{\ell_1} - \|\hat{\beta}\|_{\ell_1}\right). \tag{59}$$

Next, we shall expand the two terms on the right hand side of (59) according to $\hat{\beta}_\ell$ close to some $\beta^\star_k$ or not. In the rest of the proof, we fix $r > 0$ so that Assumptions 6.1 and 6.2, are verified by $\mathcal{Q}^\star$. In particular, for all $k \neq k'$ in $S^\star = \{k'' \in \{1, \cdots, K\}, \beta^\star_{k''} \neq 0\}$ we have $\mathfrak{d}_T(\theta^\star_k, \theta^\star_{k'}) > 2r$.

Recall the definitions given in Section 2 of the sets of indices $\hat{S}$, $\tilde{S}_k(r)$ and $\tilde{S}(r)$ for $k \in S^\star$. Since the closed balls $\mathcal{B}_T(\theta^\star_k, r)$ 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:

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

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 $\theta^\star_k$ to be estimated. We use a Taylor-type expansion with the Riemannian metric $\mathfrak{d}_T$ for the function $\phi_T(\theta)$ around the elements of $\mathcal{Q}^\star$. By Assumption 3.1, the function ϕ_T is twice continuously differentiable with respect to the variable θ and the function g_T defined in (15) is positive on Θ_T and of class $\mathcal{C}^1$ by Assumption 3.2. We set in this section $\tilde{D}_{i;T}[\phi_T] = \phi_T^{[i]}$ for $i = 0, 1, 2$. According to Lemma 4.2, we have for any $\theta^\star_k$ and $\hat{\theta}_\ell$ in Θ_T:

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

where $\gamma^{(k\ell)}$ is a distance realizing geodesic path belonging to Θ_T such that $\gamma_0^{(k\ell)} = \theta^\star_k$, $\gamma_1^{(k\ell)} = \hat{\theta}_\ell$ and $\mathfrak{d}_T(\hat{\theta}_\ell, \theta^\star_k) = \mathcal{L}_T(\gamma^{(k\ell)})$. Hence we obtain:

$$\begin{aligned}
\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star) = &\sum_{k\in S^\star} I_{0,k}(r)\,\phi_T(\theta^\star_k) + \sum_{k\in S^\star} I_{1,k}(r)\,\phi_T^{[1]}(\theta^\star_k) + \sum_{k\in\tilde{S}(r)^c} \hat{\beta}_k\,\phi_T(\hat{\theta}_k)\\
&+ \sum_{k\in S^\star}\left(\sum_{\ell\in\tilde{S}_k(r)} \hat{\beta}_\ell\,\mathfrak{d}_T(\hat{\theta}_\ell, \theta^\star_k)^2 \int_0^1 (1-s)\phi_T^{[2]}(\gamma_s^{(k\ell)})\,\mathrm{d}s\right),
\end{aligned} \tag{60}$$

with

$$I_{0,k}(r) = \Big(\sum_{\ell\in\tilde{S}_k(r)} \hat{\beta}_\ell\Big) - \beta^\star_k \quad \text{and} \quad I_{1,k}(r) = \sum_{\ell\in\tilde{S}_k(r)} \hat{\beta}_\ell\,\operatorname{sign}(\hat{\theta}_\ell - \theta^\star_k)\,\mathfrak{d}_T(\hat{\theta}_\ell, \theta^\star_k).$$

Let us introduce some notations in order to bound the different terms of the expansion above:

$$I_0(r) = \sum_{k\in S^\star} |I_{0,k}(r)| \quad \text{and} \quad I_1(r) = \sum_{k\in S^\star} |I_{1,k}(r)|, \tag{61}$$

$$I_{2,k}(r) = \sum_{\ell\in \tilde{S}_k(r)} \left|\hat{\beta}_\ell\right| \mathfrak{d}_T(\hat{\theta}_\ell, \theta^\star_k)^2 \quad \text{and} \quad I_2(r) = \sum_{k\in S^\star} I_{2,k}(r), \tag{62}$$

$$I_3(r) = \sum_{\ell\in \tilde{S}(r)^c} \left|\hat{\beta}_\ell\right| = \left\|\hat{\beta}_{\tilde{S}(r)^c}\right\|_{\ell_1}, \tag{63}$$

and we omit the dependence in r when there is no ambiguity.

We bound the difference $\|\beta^\star\|_{\ell_1} - \|\hat{\beta}\|_{\ell_1}$ by noticing that:

$$\|\beta^\star\|_{\ell_1} - \|\hat{\beta}\|_{\ell_1} = \sum_{k\in S^\star}\Big(|\beta^\star_k| - \sum_{\ell\in\tilde{S}_k(r)}|\hat{\beta}_\ell|\Big) - \sum_{k\in\tilde{S}(r)^c}\left|\hat{\beta}_k\right| \leq \sum_{k\in S^\star}\Big|\beta^\star_k - \sum_{\ell\in\tilde{S}_k(r)}\hat{\beta}_\ell\Big| = I_0. \tag{64}$$

In the next lemma, we give an upper bound of I_0. Recall the constants C'_N and C_F from Assumption 6.1.

Lemma 9.1. *Under the assumptions of Theorem 2.1 and with the element $p_1 \in H_T$ from Assumption 6.1 associated to the function $v : \mathcal{Q}^\star \to \{-1,1\}$ defined by:*

$$v(\theta^\star_k) = \operatorname{sign}(I_{0,k}) \quad \textit{for all } k\in S^\star,$$

we get that:

$$I_0 \leq C'_N I_2 + (1 - C_F) I_3 + |\hat{\Upsilon}(p_1)|. \tag{65}$$

Proof. Let $v \in \{-1,1\}^s$ with entries $v_k = v(\theta^\star_k)$ so that:

$$I_0 = \sum_{k\in S^\star}|I_{0,k}| = \sum_{k\in S^\star} v_k I_{0,k} = \sum_{k\in S^\star} v_k\Big(\Big(\sum_{\ell\in\tilde{S}_k(r)}\hat{\beta}_\ell\Big) - \beta^\star_k\Big).$$

Let p_1 be an element of H_T from Assumption 6.1 associated to the application v such that properties (i)-(iv) therein hold. By adding and substracting $\sum_{k\in S^\star}\sum_{\ell\in\tilde{S}_k(r)}\hat{\beta}_\ell\left\langle\phi_T(\hat{\theta}_\ell), p_1\right\rangle_T$ to I_0 and using the property (ii) satisfied by the element p_1, that is, $\langle\phi_T(\theta^\star_k), p_1\rangle_T = v_k$ for all $k\in S^\star$, we obtain:

$$I_0 = \sum_{k\in S^\star}\sum_{\ell\in\tilde{S}_k(r)}\hat{\beta}_\ell\left(v_k - \left\langle\phi_T(\hat{\theta}_\ell), p_1\right\rangle_T\right) + \left\langle\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star), p_1\right\rangle_T - \sum_{\ell\in\tilde{S}(r)^c}\hat{\beta}_\ell\left\langle\phi_T(\hat{\theta}_\ell), p_1\right\rangle_T.$$

We deduce that:

$$I_0 \leq \sum_{k\in S^\star}\sum_{\ell\in\tilde{S}_k(r)}|\hat{\beta}_\ell|\left|v_k - \left\langle\phi_T(\hat{\theta}_\ell), p_1\right\rangle_T\right| + |\hat{\Upsilon}(p_1)| + \sum_{\ell\in\tilde{S}(r)^c}|\hat{\beta}_\ell|\left|\left\langle\phi_T(\hat{\theta}_\ell), p_1\right\rangle_T\right|.$$

Notice that for $\ell\in\tilde{S}(r)^c$, $\hat{\theta}_\ell \notin \bigcup_{k\in S^\star}\mathcal{B}_T(\theta^\star_k, r)$. Then, by using the properties (ii) and (iii) from Assumption 6.1, we get that (65) holds with the constants C'_N and C_F from Assumption 6.1. □

In the next lemma, we give an upper bound of I_1. Recall the constants c_N and c_F from Assumption 6.2.

Lemma 9.2. *Under the assumptions of Theorem 2.1 and with the element $q_0 \in H_T$ from Assumption 6.2 associated to the function $v : \mathcal{Q}^\star \to \{-1,1\}$ defined by:*

$$v(\theta^\star_k) = \operatorname{sign}(I_{1,k}) \quad \textit{for all } k\in S^\star,$$

we get that:

$$I_1 \leq c_N I_2 + c_F I_3 + |\hat{\Upsilon}(q_0)|. \tag{66}$$

Proof. Let $v \in \{-1,1\}^s$ with entries $v_k = v(\theta_k^\star)$ so that:

$$I_1 = \sum_{k\in S^\star} |I_{1,k}| = \sum_{k\in S^\star} v_k I_{1,k} = \sum_{k\in S^\star} \sum_{\ell\in \tilde{S}_k(r)} \hat{\beta}_\ell \, v_k \, \mathrm{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star).$$

Let $q_0 \in H_T$ from Assumption 6.2 associated to the application v such that properties (*i*)-(*iii*) therein hold. By adding and substracting $\sum_{\ell\in\tilde{S}(r)} \hat{\beta}_\ell \left\langle \phi_T(\hat{\theta}_\ell), q_0 \right\rangle_T = \left\langle \hat{\beta}\Phi_T(\hat{\vartheta}), q_0 \right\rangle_T - \sum_{\ell\in\tilde{S}(r)^c} \hat{\beta}_\ell \left\langle \phi_T(\hat{\theta}_\ell), q_0 \right\rangle_T$ to I_1 and using the triangle inequality, we obtain:

$$I_1 \le \sum_{k\in S^\star} \sum_{\ell\in \tilde{S}_k(r)} |\hat{\beta}_\ell| \left| v_k \, \mathrm{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) - \left\langle \phi_T(\hat{\theta}_\ell), q_0 \right\rangle_T \right|$$
$$+ \sum_{\ell\in\tilde{S}(r)^c} |\hat{\beta}_\ell| \left| \left\langle \phi_T(\hat{\theta}_\ell), q_0 \right\rangle_T \right| + \left| \left\langle \hat{\beta}\Phi_T(\hat{\vartheta}), q_0 \right\rangle_T \right|.$$

The property (*i*) of Assumption 6.2 gives that $\langle \phi_T(\theta_k^\star), q_0 \rangle_T = 0$ for all $k \in S^\star$. This implies that $\langle \beta^\star \Phi_T(\vartheta^\star), q_0 \rangle_T = 0$. Then, by using the definition of I_2 and I_3 from (62)-(63) and the properties (*i*) and (*ii*) of Assumption 6.2, we obtain:

$$I_1 \le c_N I_2 + c_F I_3 + \left| \left\langle \hat{\beta}\Phi_T(\hat{\vartheta}), q_0 \right\rangle_T \right| = c_N I_2 + c_F I_3 + |\hat{\Upsilon}(q_0)|,$$

with the constants c_N and c_F from Assumption 6.2. □

We consider the following suprema of Gaussian processes for $i = 0,1,2$:

$$M_i = \sup_{\theta\in\Theta_T} \left| \left\langle w_T, \phi_T^{[i]}(\theta) \right\rangle_T \right|.$$

By using the expansion (60) and the bounds (66) and (65) for the second inequality, we obtain:

$$|\hat{\Upsilon}(w_T)| \le (I_0 + I_3) M_0 + I_1 M_1 + I_2\, 2^{-1}\, M_2 \tag{67}$$
$$\le (C_N' I_2 + (2 - C_F) I_3 + |\hat{\Upsilon}(p_1)|) M_0 + (c_N I_2 + c_F I_3 + |\hat{\Upsilon}(q_0)|) M_1 + I_2\, 2^{-1}\, M_2. \tag{68}$$

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

$$D_B = \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} - \hat{\Upsilon}(p_0), \tag{69}$$

where p_0 is the element of H_T given by the Assumption 6.1 associated to the application $v : \mathcal{Q}^\star \to \{-1,1\}$ given by:

$$v(\theta_k^\star) = \mathrm{sign}(\beta_k^\star) \quad \text{for all } k \in S^\star. \tag{70}$$

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

Lemma 9.3. *Under the assumptions of Theorem 2.1 and with the constants C_N and C_F of Assumption 6.1, we get that:*

$$D_B \ge C_N I_2 + C_F I_3. \tag{71}$$

Proof. By definition (69) of D_B we have:

$$D_B = \sum_{k\in\hat{S}} |\hat{\beta}_k| - \hat{\beta}_k \left\langle \phi_T(\hat{\theta}_k), p_0 \right\rangle_T - \left(\sum_{k\in S^\star} |\beta_k^\star| - \beta_k^\star \langle \phi_T(\theta_k^\star), p_0 \rangle_T \right).$$

By using the interpolating properties of the element p_0 of H_T from Assumption 6.1 associated to the function v defined in (70), we have $\sum_{k\in S^\star} |\beta^\star_k| - \beta^\star_k \langle \phi_T(\theta^\star_k), p_0\rangle_T = 0$. Hence, we deduce that:

$$\begin{aligned} D_B &= \sum_{k\in\hat{S}} |\hat{\beta}_k| - \hat{\beta}_k \left\langle \phi_T(\hat{\theta}_k), p_0 \right\rangle_T \\ &\geq \sum_{k\in\hat{S}} |\hat{\beta}_k| - |\hat{\beta}_k| \left| \left\langle \phi_T(\hat{\theta}_k), p_0 \right\rangle_T \right| \\ &= \sum_{\ell\in\tilde{S}(r)} |\hat{\beta}_\ell| \left(1 - \left|\left\langle \phi_T(\hat{\theta}_\ell), p_0\right\rangle_T\right|\right) + \sum_{k\in\tilde{S}(r)^c} |\hat{\beta}_k| \left(1 - \left|\left\langle \phi_T(\hat{\theta}_k), p_0\right\rangle_T\right|\right). \end{aligned}$$

Thanks to properties *(i)* and *(iii)* of Assumption 6.1 and the definitions (62) and (63) of I_2 and I_3, we obtain:

$$D_B \geq \sum_{k\in S^\star} \sum_{\ell\in\tilde{S}_k(r)} C_N |\hat{\beta}_\ell| \mathfrak{d}_T(\hat{\theta}_\ell, \theta^\star_k)^2 + \sum_{k\in\tilde{S}(r)^c} C_F |\hat{\beta}_k| = C_N I_2 + C_F I_3,$$

where the constants C_N and C_F are that of Assumption 6.1. □

We now give an upper bound of the Bregman divergence.

Lemma 9.4. *Under the assumptions of Theorem 2.1, we have:*

$$\begin{aligned} \kappa D_B \leq I_2 \left(C'_N M_0 + c_N M_1 + 2^{-1} M_2\right) + I_3 \left((2 - C_F) M_0 + c_F M_1\right) \\ + |\hat{\Upsilon}(p_1)| M_0 + |\hat{\Upsilon}(q_0)| M_1 + \kappa |\hat{\Upsilon}(p_0)|. \end{aligned} \tag{72}$$

Proof. Recall that $\mathcal{Q}^\star \subset \Theta_T$. We deduce from (59) that:

$$\kappa(||\hat{\beta}||_{\ell_1} - ||\beta^\star||_{\ell_1}) \leq \hat{\Upsilon}(w_T) - \frac{1}{2}\left\|\beta^\star \Phi_T(\vartheta^\star) - \hat{\beta}\Phi_T(\hat{\vartheta})\right\|_T^2 \leq \hat{\Upsilon}(w_T). \tag{73}$$

Using (69), we obtain:

$$\kappa D_B \leq |\hat{\Upsilon}(w_T)| + \kappa |\hat{\Upsilon}(p_0)|.$$

Then, use (68) to get (72). □

By combining the upper and lower bounds (71) and (72), we deduce that:

$$\begin{aligned} I_2 \left(C_N - \frac{1}{\kappa}\left(C'_N M_0 + c_N M_1 + 2^{-1} M_2\right)\right) + I_3 \left(C_F - \frac{1}{\kappa}\left((2 - C_F) M_0 + c_F M_1\right)\right) \\ \leq \frac{1}{\kappa} |\hat{\Upsilon}(p_1)| M_0 + \frac{1}{\kappa} |\hat{\Upsilon}(q_0)| M_1 + |\hat{\Upsilon}(p_0)|. \end{aligned} \tag{74}$$

We define the events:

$$\mathcal{A}_i = \{M_i \leq \mathcal{C}\,\kappa\}, \quad \text{for } i \in \{0,1,2\} \quad \text{and} \quad \mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2, \tag{75}$$

where:

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

(We shall prove in (84) that the event $\mathcal{A}$ occurs with high probability.) We get from Inequality (74), that on the event $\mathcal{A}$:

$$C_N I_2 + C_F I_3 \leq 2\mathcal{C}' \left(|\hat{\Upsilon}(p_1)| + |\hat{\Upsilon}(q_0)| + |\hat{\Upsilon}(p_0)|\right) \quad \text{with} \quad \mathcal{C}' = \mathcal{C} \vee 1. \tag{76}$$

By reinjecting (64), (68), (65) and (66) in (59) one gets:

$$\frac{1}{2}\hat{R}_T^2 \leq I_2(C'_N M_0 + c_N M_1 + 2^{-1}M_2 + \kappa C'_N) + I_3((2-C_F)M_0 + c_F M_1 + \kappa(1-C_F)) + |\hat{\Upsilon}(p_1)|(M_0+\kappa) + |\hat{\Upsilon}(q_0)|M_1.$$

Using (76), we obtain an upper bound for the prediction error on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq C\,\kappa\,(|\hat{\Upsilon}(p_0)| + |\hat{\Upsilon}(p_1)| + |\hat{\Upsilon}(q_0)|), \tag{77}$$

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 $f \in H_T$:

$$|\hat{\Upsilon}(f)| \leq \hat{R}_T \|f\|_T. \tag{78}$$

Using Assumption 6.1 *(iv)* for p_0 and p_1, and Assumption 6.2 *(iii)* for q_0, we get:

$$\|p_0\|_T \leq C_B\sqrt{s}, \quad \|p_1\|_T \leq C_B\sqrt{s} \quad \text{and} \quad \|q_0\|_T \leq c_B\sqrt{s}. \tag{79}$$

Plugging this in (77), we get that on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq \mathcal{C}_0\,\kappa \hat{R}_T\,\sqrt{s} \quad \text{with} \quad \mathcal{C}_0 = (c_B + 2C_B)C. \tag{80}$$

This gives (5).

The proof of (6) is postponed to Section 9.2 and will be easily deduced from the first and third inequalities in (9).

To complete the proof of Theorem 2.1 we shall give a lower bound for the probability of the event $\mathcal{A}$ defined in (75). For $i = 0, 1, 2$ and $\theta \in \Theta$, set $X_i(\theta) = \left\langle w_T, \phi_T^{[i]}(\theta)\right\rangle_T$ a real centered Gaussian process with continuously differentiable sample paths, so that its supremum is $M_i = \sup_{\Theta_T}|X_i|$.

We first consider $i = 0$. We have, thanks to (32) and (29) for the second part:

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

Recall Assumption 1.1 on the noise w_T holds. We deduce from Lemma A.1 with $C_1 = C_2 = 1$ that:

$$\mathbb{P}(\mathcal{A}_0^c) = \mathbb{P}\left(\sup_{\Theta_T}|X_0| > \mathcal{C}\,\kappa\right) \leq c_0\left(\sigma\frac{|\Theta_T|_{\mathfrak{d}_T}\sqrt{\Delta_T}}{\mathcal{C}\,\kappa} \vee 1\right)\,\mathrm{e}^{-(\mathcal{C}\,\kappa)^2/(4\sigma^2\Delta_T)}, \tag{81}$$

where $|\Theta_T|_{\mathfrak{d}_T}$ denotes the diameter of the set Θ_T with respect to the metric $\mathfrak{d}_T$ and $c_0 = 3$.

We consider $i = 1$. Thanks to (32), we get:

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

Recall $L_{2,2}$ and $\mathcal{V}_T$ are defined in (35) and (38). Since Assumptions 5.1 and 5.2 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.1 with $C_1 = 1$ and $C_2 = \sqrt{2L_{2,2}}$ and taking $c_1 = 2\sqrt{2L_{2,2}} + 1$, that:

$$\mathbb{P}(\mathcal{A}_1^c) = \mathbb{P}\left(\sup_{\Theta_T}|X_1| > \mathcal{C}\,\kappa\right) \leq c_1\left(\sigma\frac{|\Theta_T|_{\mathfrak{d}_T}\sqrt{\Delta_T}}{\mathcal{C}\,\kappa} \vee 1\right)\,\mathrm{e}^{-(\mathcal{C}\,\kappa)^2/(4\sigma^2\Delta_T)}. \tag{82}$$

We consider $i = 2$. Thanks to (32), we get:

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

Recall the definition of the function h_∞ given in (34) and the constants $L_{2,2}$, L_3, $\mathcal{V}_T$ defined in (35) and (38). Using also Assumption 5.2 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.1 with $C_1 = \sqrt{2L_{2,2}}$ and $C_2 = \sqrt{2L_3}$ and taking $c_2 = 2\sqrt{2L_3} + 1$, that:

$$\mathbb{P}\left(\mathcal{A}_2^c\right) = \mathbb{P}\left(\sup_{\Theta_T} |X_2| > \mathcal{C}\,\kappa\right) \leq c_2 \left(\sigma \frac{|\Theta_T|_{\mathfrak{d}_T}\sqrt{\Delta_T}}{\mathcal{C}\,\kappa} \vee 1\right) \mathrm{e}^{-(\mathcal{C}\,\kappa)^2/(8\sigma^2\Delta_T L_{2,2})}. \tag{83}$$

Since $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2$, we deduce from (81), (82) and (83) that:

$$\mathbb{P}\left(\mathcal{A}^c\right) = \mathbb{P}\left(\mathcal{A}_0^c \cup \mathcal{A}_1^c \cup \mathcal{A}_2^c\right) \leq \mathcal{C}_2' \left(\sigma \frac{|\Theta_T|_{\mathfrak{d}_T}\sqrt{\Delta_T}}{\mathcal{C}\kappa} \vee 1\right) \mathrm{e}^{-\kappa^2/(\mathcal{C}_1^2\,\sigma^2\Delta_T)},$$

with the finite positive constants:

$$\mathcal{C}_1 = \frac{2}{\mathcal{C}}\left(1 \vee \sqrt{2L_{2,2}}\right) \quad \text{and} \quad \mathcal{C}_2' = c_0 + c_1 + c_2.$$

By taking $\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau}$, for any positive constant $\tau > 0$, we get:

$$\mathbb{P}\left(\mathcal{A}_0^c \cup \mathcal{A}_1^c \cup \mathcal{A}_2^c\right) \leq \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log \tau}} \vee \frac{1}{\tau}\right) \quad \text{with} \quad \mathcal{C}_2 = \mathcal{C}_2' \left(\frac{1}{\mathcal{C}\mathcal{C}_1} \vee 1\right). \tag{84}$$

This completes the proof of the theorem.

9.2. **Proof of Theorem 2.5 and of Equation (6).** We keep notations from Section 9.1. Recall that Assumptions (i)-(v) of Theorem 2.1 are in force. We shall first provide an upper bound of I_i for $i = 0, 1, 2, 3$. We deduce from (78), (79) and (80), that, on the event $\mathcal{A}$:

$$|\hat{\Upsilon}(p_0)| \leq \mathcal{C}_0 C_B\,\kappa\,s, \quad |\hat{\Upsilon}(p_1)| \leq \mathcal{C}_0 C_B\,\kappa\,s \quad \text{and} \quad |\hat{\Upsilon}(q_0)| \leq \mathcal{C}_0 c_B\,\kappa\,s.$$

Then, we obtain from (76) that, on the event $\mathcal{A}$:

$$I_3 \leq \mathcal{C}_5\,\kappa\,s \quad \text{and} \quad I_2 \leq \mathcal{C}_6\,\kappa\,s \quad \text{with} \quad \mathcal{C}_5 = 2\frac{\mathcal{C}'}{C_F}\mathcal{C}_0(c_B + 2C_B) \quad \text{and} \quad \mathcal{C}_6 = \frac{C_F}{C_N}\mathcal{C}_5. \tag{85}$$

This gives the third inequality in (9), as well as Inequality (10) in Remark 2.6. We also deduce from (65) that, on the event $\mathcal{A}$:

$$I_0 \leq \mathcal{C}_4\,\kappa\,s \quad \text{with} \quad \mathcal{C}_4 = C_N'\mathcal{C}_6 + (1 - C_F)\mathcal{C}_5 + \mathcal{C}_0 C_B. \tag{86}$$

This gives the second inequality in (9).

We now establish the first inequality in (9). We deduce from (59) that:

$$\kappa(\|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1}) \leq \hat{\Upsilon}(w_T). \tag{87}$$

Then, using the bounds (86) and (85) on I_0, I_2 and I_3, we deduce from (67) and (66) that, on the event $\mathcal{A}$:

$$|\hat{\Upsilon}(w_T)| \leq \mathcal{C}_7\,s\,\kappa^2 \quad \text{with} \quad \mathcal{C}_7 = \mathcal{C}\left(\mathcal{C}_4 + \mathcal{C}_5(1 + c_F) + \mathcal{C}_6(1 + c_N) + \mathcal{C}_0 c_B\right). \tag{88}$$

Thus, (87) and (88) imply that, on the event $\mathcal{A}$:

$$\|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} \leq \mathcal{C}_7\,s\,\kappa. \tag{89}$$

Then, use (64) and (86) to deduce that, on the event $\mathcal{A}$:

$$\left|\|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1}\right| \le (\mathcal{C}_4 \vee \mathcal{C}_7)\, s\, \kappa.$$

This proves (6) (we shall take $\mathcal{C}_3 = \mathcal{C}_7 + 2\mathcal{C}_4$, see below). Let $\mathcal{I}^+$ (resp. $\mathcal{I}^-$) be the set of indices $k \in S^\star$ such that the quantity $\left(\sum_{\ell\in\tilde{S}_k(r)} |\hat{\beta}_\ell|\right) - |\beta^\star_k|$ is non negative (resp. negative). We have the following decomposition:

$$\begin{aligned}\sum_{k\in S^\star}\Big|\sum_{\ell\in\tilde{S}_k(r)}|\hat{\beta}_\ell| - |\beta^\star_k|\Big| &= \sum_{k\in\mathcal{I}^+}\Big(\sum_{\ell\in\tilde{S}_k(r)}|\hat{\beta}_\ell| - |\beta^\star_k|\Big) + \sum_{k\in\mathcal{I}^-}\Big(|\beta^\star_k| - \sum_{\ell\in\tilde{S}_k(r)}|\hat{\beta}_\ell|\Big)\\ &\le \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} + 2\sum_{k\in\mathcal{I}^-}\Big(|\beta^\star_k| - \sum_{\ell\in\tilde{S}_k(r)}|\hat{\beta}_\ell|\Big)\\ &\le \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} + 2I_0.\end{aligned} \tag{90}$$

Then, use (86) and (89) to obtain the first inequality (9) with $\mathcal{C}_3 = \mathcal{C}_7 + 2\mathcal{C}_4$. This ends the proof of Theorem 2.5.

10. Construction of certificate functions

10.1. **Proof of Proposition 7.4 (Construction of an interpolating certificate).** This section is devoted to the proof of Proposition 7.4. We closely follow the proof of [41] taking into account the approximation of the kernel $\mathcal{K}_T$ by the kernel $\mathcal{K}_\infty$, which is measured through the quantity $\mathcal{V}_T$ defined in (38).

Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. Recall Assumptions 3.2 (and thus 3.1 on the regularity of φ_T) and 5.1 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 $u_\infty \in \left(0, H^{(2)}_\infty(r,\rho)\right)$ such that *(ii)*, *(iii)*, *(iv)* and *(v)* of Proposition 7.4 hold. We denote by $\|\cdot\|_{\text{op}}$ the operator norm associated to the ℓ_∞ norm on $\mathbb{R}^s$.

By assumption $\delta_\infty(u_\infty, s)$ is finite. Let $\vartheta^\star = (\theta^\star_1, \dots, \theta^\star_s) \in \Theta^s_{T,2\rho_T\,\delta_\infty(u_\infty,s)}$. We note $\mathcal{Q}^\star = \{\theta^\star_i,\, 1 \le i \le s\}$ the set of parameters of cardinal s. By Lemma 7.3, we have:

$$\Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)} \subseteq \Theta^s_{T,\delta_T(u_T(s),s)} \quad \text{where} \quad u_T(s) = u_\infty + (s-1)\mathcal{V}_T.$$

Hence we have:

$$\vartheta^\star \in \Theta^s_{T,\delta_T(u_T(s),s)}. \tag{91}$$

Set

$$\Gamma^{[i,j]} = \mathcal{K}_T^{[i,j]}(\vartheta^\star) \quad \text{and} \quad \Gamma = \begin{pmatrix}\Gamma^{[0,0]} & \Gamma^{[1,0]\top}\\ \Gamma^{[1,0]} & \Gamma^{[1,1]}\end{pmatrix}. \tag{92}$$

We deduce from (44) and (91) that:

$$\left\|I - \Gamma^{[0,0]}\right\|_{\text{op}} \le u_T(s),\quad \left\|I - \Gamma^{[1,1]}\right\|_{\text{op}} \le u_T(s),\quad \left\|\Gamma^{[1,0]}\right\|_{\text{op}} \le u_T(s) \quad \text{and} \quad \left\|\Gamma^{[1,0]\top}\right\|_{\text{op}} \le u_T(s). \tag{93}$$

For simplicity, for an expression A we write A_T for $A_{\mathcal{K}_T}$. Using this convention, recall the definition of the derivative operator $\tilde{D}_{i;T}$ and write $\phi^{[1]}_T$ for $\tilde{D}_{1;T}[\phi_T]$.

Let $\alpha = (\alpha_1, \dots, \alpha_s)^\top$ and $\xi = (\xi_1, \dots, \xi_s)^\top$ be elements of $\mathbb{R}^s$. Let $p_{\alpha,\xi}$ be an element of H_T defined by:

$$p_{\alpha,\xi} = \sum_{k=1}^s \alpha_k \phi_T(\theta^\star_k) + \sum_{k=1}^s \xi_k\, \phi^{[1]}_T(\theta^\star_k), \tag{94}$$

and, using (32) in Lemma 4.3, set the interpolating real-valued function $\eta_{\alpha,\xi}$ defined on Θ by:

$$\eta_{\alpha,\xi}(\theta) = \langle \phi_T(\theta), p_{\alpha,\xi}\rangle_T = \sum_{k=1}^{s} \alpha_k \, \mathcal{K}_T(\theta, \theta_k^\star) + \sum_{k=1}^{s} \xi_k \, \mathcal{K}_T^{[0,1]}(\theta, \theta_k^\star). \tag{95}$$

By Assumption 3.2 on the regularity of φ_T and the positivity of g_T and Lemma 4.3, we get that the function $\eta_{\alpha,\xi}$ is of class $\mathcal{C}^3$ on Θ, and using (24), we get that:

$$\eta_{\alpha,\xi}^{[1]} := \tilde{D}_{1;T}[\eta_{\alpha,\xi}](\theta) = \sum_{k=1}^{s} \alpha_k \, \mathcal{K}_T^{[1,0]}(\theta, \theta_k^\star) + \sum_{k=1}^{s} \xi_k \, \mathcal{K}_T^{[1,1]}(\theta, \theta_k^\star). \tag{96}$$

We give a preliminary technical lemma.

Lemma 10.1. *Let $v = (v_1, \cdots, v_s)^\top \in \{-1, 1\}^s$ be a sign vector. Assume that (93) holds with $u_T(s) < 1/2$. Under Assumption 3.2, there exist unique $\alpha, \xi \in \mathbb{R}^s$ such that:*

$$\eta_{\alpha,\xi}(\theta_k^\star) = v_k \in \{-1, 1\} \quad \text{and} \quad \eta_{\alpha,\xi}^{[1]}(\theta_k^\star) = 0 \quad \text{for} \quad 1 \leq k \leq s. \tag{97}$$

Furthermore, we have:

$$\|\alpha\|_{\ell_\infty} \leq \frac{1 - u_T(s)}{1 - 2u_T(s)}, \quad \|\alpha - v\|_{\ell_\infty} \leq \frac{u_T(s)}{1 - 2u_T(s)} \quad \text{and} \quad \|\xi\|_{\ell_\infty} \leq \frac{u_T(s)}{1 - 2u_T(s)}. \tag{98}$$

Proof of Lemma 10.1 . Thanks to (32), (29) and (96), we have:

$$\left(\eta_{\alpha,\xi}(\theta_1^\star), \ldots, \eta_{\alpha,\xi}(\theta_s^\star), \eta_{\alpha,\xi}^{[1]}(\theta_1^\star), \ldots, \eta_{\alpha,\xi}^{[1]}(\theta_s^\star)\right)^\top = \Gamma \begin{pmatrix} \alpha \\ \xi \end{pmatrix}.$$

Thus, solving (97) is equivalent to solving,

$$\Gamma \begin{pmatrix} \alpha \\ \xi \end{pmatrix} = \begin{pmatrix} v \\ 0_s \end{pmatrix}, \tag{99}$$

with 0_s the vector of size s with all its components equal to zero.

We first show that Γ is non singular so that α and ξ exist and are uniquely defined. Using Lemma A.3 based on the Schur complement, Γ has an inverse provided that $\Gamma^{[1,1]}$ and $\Gamma_{SC} := \Gamma^{[0,0]} - \Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}$ are non singular. 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 $\|M^{-1}\|_{\text{op}} \leq \left(1 - \|I - M\|_{\text{op}}\right)^{-1}$.

Recall that by assumption $u_T(s) \leq 1/2$. Then, the second inequality in (93) imply that $\left\|I - \Gamma^{[1,1]}\right\|_{\text{op}} < 1$ and thus $\Gamma^{[1,1]}$ is non singular. We now prove that Γ_{SC} is also non singular. Using the triangle inequality we have:

$$\begin{aligned} \|I - \Gamma_{SC}\|_{\text{op}} &= \left\|I - \Gamma^{[0,0]} + \Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\right\|_{\text{op}} \\ &\leq \left\|I - \Gamma^{[0,0]}\right\|_{\text{op}} + \left\|\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\right\|_{\text{op}}. \end{aligned}$$

Let us bound the terms on the right hand side of the inequality above. To bound $\left\|\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\right\|_{\text{op}}$ notice that:

$$\left\|\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\right\|_{\text{op}} \leq ||\Gamma^{[1,0]}||_{\text{op}} \left\|\Gamma^{[1,0]\top}\right\|_{\text{op}} \left\|[\Gamma^{[1,1]}]^{-1}\right\|_{\text{op}}.$$

We have, thanks to (93) for the second inequality:

$$\left\|[\Gamma^{[1,1]}]^{-1}\right\|_{\text{op}} \leq \frac{1}{1 - \left\|I - \Gamma^{[1,1]}\right\|_{\text{op}}} \leq \frac{1}{1 - u_T(s)}. \tag{100}$$

Using (93), we get:

$$||I - \Gamma_{SC}||_{\text{op}} \le u_T(s) + \frac{u_T(s)^2}{1 - u_T(s)} = \frac{u_T(s)}{1 - u_T(s)}.$$

By assumption, we have $u_T(s) \le H_\infty^{(2)}(r, \rho) < 1/2$. Hence, we have $\frac{u_T(s)}{1-u_T(s)} < 1$ and thus, Γ_{SC} is non singular. Furthermore, we get:

$$||\Gamma_{SC}^{-1}||_{\text{op}} \le \frac{1}{1 - \|I - \Gamma_{SC}\|_{\text{op}}} \le \frac{1 - u_T(s)}{1 - 2u_T(s)}. \tag{101}$$

As the matrices $\Gamma^{[1,1]}$ and Γ_{SC} are non singular, we deduce that the matrix Γ is non singular.

We now give bounds related to α and ξ. The Lemma A.3 on the Schur complement gives also that:

$$\alpha = \Gamma_{SC}^{-1} v \quad \text{and} \quad \xi = -[\Gamma^{[1,1]}]^{-1} \Gamma^{[1,0]} \Gamma_{SC}^{-1} v.$$

Hence, we deduce that:

$$\begin{aligned}
\|\alpha\|_{\ell_\infty} &\le \left\|\Gamma_{SC}^{-1}\right\|_{\text{op}} \|v\|_{\ell_\infty} \le \frac{1 - u_T(s)}{1 - 2u_T(s)}, \\
\|\xi\|_{\ell_\infty} &\le \left\|[\Gamma^{[1,1]}]^{-1} \Gamma^{[1,0]} \Gamma_{SC}^{-1}\right\|_{\text{op}} \|v\|_{\ell_\infty} \le \left\|[\Gamma^{[1,1]}]^{-1}\right\|_{\text{op}} \left\|\Gamma^{[1,0]}\right\|_{\text{op}} \left\|\Gamma_{SC}^{-1}\right\|_{\text{op}} \le \frac{u_T(s)}{1 - 2u_T(s)}, \\
\|\alpha - v\|_{\ell_\infty} &\le \left\|(\Gamma_{SC}^{-1} - I)\right\|_{\text{op}} \|v\|_{\ell_\infty} \le \|\Gamma_{SC} - I\|_{\text{op}} \left\|\Gamma_{SC}^{-1}\right\|_{\text{op}} \le \frac{u_T(s)}{1 - 2u_T(s)}.
\end{aligned}$$

This finishes the proof. □

We now fix a sign vector $v = (v_1, \cdots, v_s)^\top \in \{-1, 1\}^s$ and consider $p_{\alpha,\xi}$ and $\eta_{\alpha,\xi}$ with α and ξ characterized by (97) from Lemma 10.1. 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 6.1 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 $|\eta_{\alpha,\xi}(\theta)| \le 1 - C_F$. Let $\theta_\ell^\star$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. Since $\vartheta^\star \in \Theta^s_{T, 2\rho_T \delta_\infty(u_\infty, s)}$, we have, by the triangle inequality that for any $k \neq \ell$:

$$2\rho_T\, \delta_\infty(u_\infty, s) < \mathfrak{d}_T(\theta_\ell^\star, \theta_k^\star) \le \mathfrak{d}_T(\theta_\ell^\star, \theta) + \mathfrak{d}_T(\theta, \theta_k^\star) \le 2\mathfrak{d}_T(\theta, \theta_k^\star).$$

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 that $\Theta^s_{T, \rho_T \delta_\infty(u_\infty, s)} \subseteq \Theta^s_{T, \delta_T(u_T(s), s)}$ and thus:

$$\vartheta^\star_{\ell,\theta} \in \Theta^s_{T, \delta_T(u_T(s), s)}. \tag{102}$$

We denote by $\Gamma_{\ell,\theta}$ (resp. $\Gamma^{[i,j]}_{\ell,\theta}$) the matrix Γ (resp. $\Gamma^{[i,j]}$) in (92) where $\vartheta^\star$ has been replaced by $\vartheta^\star_{\ell,\theta}$. Notice the upper bounds (93) also hold for $\Gamma_{\ell,\theta}$ because of (102). Recall we have Equalities (33) on the diagonal of the kernel $\mathcal{K}_T$ and its derivatives. Elementary calculations give with $\eta_{\alpha,\xi}$ from Lemma 10.1 that:

$$\eta_{\alpha,\xi}(\theta) = e_\ell^\top \left(\Gamma^{[0,0]}_{\ell,\theta} - I\right) \alpha + \mathcal{K}_T(\theta, \theta_\ell^\star)\alpha_\ell + e_\ell^\top \Gamma^{[1,0]\top}_{\ell,\theta}\, \xi + \mathcal{K}_T^{[0,1]}(\theta, \theta_\ell^\star)\xi_\ell. \tag{103}$$

We deduce that:

$$|\eta_{\alpha,\xi}(\theta)| \le \left\|\Gamma^{[0,0]}_{\ell,\theta} - I\right\|_{\text{op}} \|\alpha\|_{\ell_\infty} + \|\alpha\|_{\ell_\infty} |\mathcal{K}_T(\theta, \theta_\ell^\star)| + \left\|\Gamma^{[1,0]\top}_{\ell,\theta}\right\|_{\text{op}} \|\xi\|_{\ell_\infty} + |\mathcal{K}_T^{[0,1]}(\theta, \theta_\ell^\star)| \|\xi\|_{\ell_\infty}. \tag{104}$$

Since θ belongs to the "far region", we have by definition of $\varepsilon_T(r)$ given in (39) that:

$$|\mathcal{K}_T(\theta, \theta_\ell^\star)| \le 1 - \varepsilon_T(r). \tag{105}$$

The triangle inequality, the definitions (38) of $\mathcal{V}_T$ and (35) of $L_{1,0}$, give:

$$|\mathcal{K}_T^{[0,1]}(\theta, \theta_\ell^\star)| \le L_{0,1} + \mathcal{V}_T. \tag{106}$$

Then, using (93) (which holds for $\Gamma_{\ell,\theta}$ thanks to (102)), we get that:

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

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

$$\varepsilon_T(r) \geq \varepsilon_\infty(r/\rho_T) - \mathcal{V}_T \geq \varepsilon_\infty(r/\rho) - \mathcal{V}_T. \tag{107}$$

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

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

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

$$u_T(s) \leq \frac{8}{10\,(5+2L_{1,0})}\varepsilon_\infty(r/\rho). \tag{108}$$

The assumption $\mathcal{V}_T \leq H_\infty^{(1)}(r,\rho)$ gives $\mathcal{V}_T \leq \varepsilon_\infty(r/\rho)/10$. Hence, we have $|\eta_{\alpha,\xi}(\theta)| \leq 1 - \frac{\varepsilon_\infty(r/\rho)}{10}$. Thus, Property *(iii)* from Assumption 6.1 holds with $C_F = \varepsilon_\infty(r/\rho)/10$.

Proof of *(i)* **from Assumption 6.1** 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 $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$ ("near region"). Thus, it is enough to prove that $|\eta_{\alpha,\xi}(\theta)| \leq 1 - C_N\, \mathfrak{d}_T(\theta_\ell^\star, \theta)^2$. This will be done by using Lemma A.4 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. Set $\eta_{\alpha,\xi}^{[2]} = \tilde{D}_{2;T}[\eta_{\alpha,\xi}]$. Differentiating (96) and using that $\mathcal{K}_T^{[2,0]}(\theta,\theta) = -1$ and $\mathcal{K}_T^{[2,1]}(\theta,\theta) = 0$, see (33), we deduce that:

$$\eta_{\alpha,\xi}^{[2]}(\theta) = e_\ell^\top(I + \Gamma_{\ell,\theta}^{[2,0]})\alpha + \mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)e_\ell^\top\alpha + e_\ell^\top\Gamma_{\ell,\theta}^{[2,1]}\xi + \mathcal{K}_T^{[2,1]}(\theta,\theta_\ell^\star)e_\ell^\top\xi. \tag{109}$$

Since $v = (v_1, \cdots, v_s)^\top \in \{-1,1\}^s$ is a sign vector, we get:

$$\eta_{\alpha,\xi}^{[2]}(\theta) - v_\ell\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star) = e_\ell^\top(I + \Gamma_{\ell,\theta}^{[2,0]})\alpha + \mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)e_\ell^\top(\alpha - v) + e_\ell^\top\Gamma_{\ell,\theta}^{[2,1]}\xi + \mathcal{K}_T^{[2,1]}(\theta,\theta_\ell^\star)e_\ell^\top\xi. \tag{110}$$

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

$$|\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)| \leq L_{2,0} + \mathcal{V}_T \quad \text{and} \quad |\mathcal{K}_T^{[2,1]}(\theta,\theta_\ell^\star)| \leq L_{2,1} + \mathcal{V}_T, \tag{111}$$

where $L_{2,0}$ and $L_{1,2}$ are defined in (35). We deduce from (102), the definition of δ_T in (44) and (45) that:

$$\left\|I + \Gamma_{\ell,\theta}^{[2,0]}\right\|_{\text{op}} \leq u_T(s) \quad \text{and} \quad \left\|\Gamma_{\ell,\theta}^{[2,1]}\right\|_{\text{op}} \leq u_T(s). \tag{112}$$

We deduce from (110) that:

$$\begin{aligned}|\eta_{\alpha,\xi}^{[2]}(\theta) - v_\ell\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)| &\leq \|\alpha\|_{\ell_\infty}\left\|I + \Gamma_{\ell,\theta}^{[2,0]}\right\|_{\text{op}} + \|\alpha - v\|_{\ell_\infty}(L_{2,0} + \mathcal{V}_T) + \|\xi\|_{\ell_\infty}\left(\left\|\Gamma_{\ell,\theta}^{[2,1]}\right\|_{\text{op}} + L_{2,1} + \mathcal{V}_T\right)\\ &\leq \frac{u_T(s)}{1-2u_T(s)}(1 + L_{2,0} + L_{2,1} + 2\mathcal{V}_T).\end{aligned}$$

By assumption, we have $u_T(s) \leq H_\infty^{(2)}(r) \leq 1/4$. Hence, we have $\frac{1}{1-2u_T(s)} \leq 2$. Furthermore, we have by assumption $\mathcal{V}_T \leq H_\infty^{(1)}(r) \leq 1/2$ and $u_T(s) \leq H_\infty^{(2)}(r)$. 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:

$$|\eta_{\alpha,\xi}^{[2]}(\theta) - v_\ell\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)| \leq \frac{8}{9}\nu_\infty(\rho r). \tag{113}$$

We now check that the hypotheses of Lemma A.4-(ii) hold in order to obtain a quadratic decay on $\eta_{\alpha,\xi}$ from the bound (113). First recall that $\eta_{\alpha,\xi}$ is twice continuously differentiable and have the interpolation properties (97). 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 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 that the function $r \mapsto \nu_\infty(r)$ is decreasing and $\rho_T \leq \rho$ for the second inequality and that $\mathcal{V}_T \leq H_\infty^{(1)}(r) \leq \nu_\infty(\rho r)/10$ 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.4-(ii) and get for $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$:

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

Proof of (ii) from Assumption 6.1 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 $|\eta_{\alpha,\xi}(\theta) - v_\ell| \leq C'_N\, \mathfrak{d}_T(\theta_\ell^\star,\theta)^2$.

Let us consider the function $f : \theta \to \eta_{\alpha,\xi}(\theta) - v_\ell$. We will bound the second covariant derivative $f^{[2]} = \tilde{D}_{2;T}[f]$ of f and apply Lemma A.4-(i) on f to prove the property (ii) for $\eta_{\alpha,\xi}$. Notice that f is twice continuously differentiable. By construction, see (97), we have $f(\theta_\ell^\star) = 0$ and $f^{[1]}(\theta_\ell^\star) = 0$. Since $f^{[2]} = \eta_{\alpha,\xi}^{[2]}$, we deduce from (109), the bounds (111) that:

$$|f^{[2]}(\theta)| \leq \|\alpha\|_{\ell_\infty} \left\| I + \Gamma_{\ell,\theta}^{[2,0]} \right\|_{\text{op}} + \|\alpha\|_{\ell_\infty}(L_{2,0} + \mathcal{V}_T) + \|\xi\|_{\ell_\infty} \left\| \Gamma_{\ell,\theta}^{[2,1]} \right\|_{\text{op}} + \|\xi\|_{\ell_\infty}(L_{2,1} + \mathcal{V}_T).$$

Using (112), and the bounds on α and ξ from Lemma 10.1, we get:

$$|f^{[2]}(\theta)| \leq \frac{1 - u_T(s)}{1 - 2u_T(s)}(L_{2,0} + \mathcal{V}_T + u_T(s)) + \frac{u_T(s)}{1 - 2u_T(s)}(L_{2,1} + \mathcal{V}_T + u_T(s)).$$

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

$$|f^{[2]}(\theta)| \leq \frac{5}{4}L_{2,0} + \frac{1}{4}L_{2,1} + 1.$$

We get thanks to Lemma A.4-(i) on the function f that for any $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$:

$$|\eta_{\alpha,\xi}(\theta) - v_\ell| \leq \frac{1}{8}\left(5L_{2,0} + L_{1,2} + 4\right)\, \mathfrak{d}_T(\theta,\theta_\ell^\star)^2.$$

Proof of (iv) from Assumption 6.1 with $C_B = 2$. Recall the definition of $p_{\alpha,\xi}$ in (94). Elementary calculations give using the definitions of $\Gamma^{[0,0]}$, $\Gamma^{[1,1]}$ and $\Gamma^{[1,1]}$ in (92):

$$\begin{aligned}
\|p_{\alpha,\xi}\|_T^2 &\leq 2\left\| \sum_{k=1}^s \alpha_k \phi_T(\theta_k^\star) \right\|_T^2 + 2\left\| \sum_{k=1}^s \xi_k\, \phi_T^{[1]}(\theta_k^\star) \right\|_T^2 \\
&= 2\alpha^\top \Gamma^{[0,0]}\alpha + 2\xi^\top \Gamma^{[1,1]}\xi \\
&\leq 2\|\alpha\|_{\ell_1}\|\alpha\|_{\ell_\infty} \left\|\Gamma^{[0,0]}\right\|_{\text{op}} + 2\|\xi\|_{\ell_1}\|\xi\|_{\ell_\infty} \left\|\Gamma^{[1,1]}\right\|_{\text{op}}.
\end{aligned}$$

Using that $\|I\|_{\text{op}} = 1$ and (93), we get that:

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

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

$$\|p_{\alpha,\xi}\|_T^2 \leq 2(1 + u_T(s)) \frac{(1 - u_T(s))^2 + u_T(s)^2}{(1 - 2u_T(s))^2} s \leq 4s.$$

This gives:

$$\|p_{\alpha,\xi}\|_T \leq 2\sqrt{s}. \tag{114}$$

We proved that *(i)*-*(iv)* from Assumption 6.1 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 6.1 holds.
This finishes the proof of Proposition 7.4.

10.2. **Proof of Proposition 7.5 (Construction of an interpolating derivative certificate).** This section is devoted to the proof of Proposition 7.5 and is close to Section 10.1. Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. Recall Assumptions 3.2 (and thus 3.1 on the regularity of φ_T) and 5.1 on the regularity of the limit kernel $\mathcal{K}_\infty$ are in force. Set $u'_\infty \in (0, 1/6)$. We denote by $\|\cdot\|_{\text{op}}$ the operator norm associated to the ℓ_∞ norm on $\mathbb{R}^s$. By assumption $\delta_\infty(u'_\infty, s)$ is finite. Let $\vartheta^\star = (\theta_1^\star, \dots, \theta_s^\star) \in \Theta^s_{T, 2\rho_T\, \delta_\infty(u'_\infty, s)}$. We note $\mathcal{Q}^\star = \{\theta_i^\star,\, 1 \leq i \leq s\}$ the set of parameters of cardinal s. Let $\alpha = (\alpha_1, \dots, \alpha_s)^\top$ and $\xi = (\xi_1, \dots, \xi_s)^\top$ be elements of $\mathbb{R}^s$. Recall $p_{\alpha,\xi}$, $\eta_{\alpha,\xi}$ and $\eta^{[1]}_{\alpha,\xi} = \tilde{D}_{1;T}[\eta_{\alpha,\xi}]$ given by (94), (95) and (96).

The next lemma is similar to Lemma 10.1, but notice that in Lemma 10.2 the function $\eta_{\alpha,\xi}$ vanished on $\mathcal{Q}^\star$ and has a derivative that interpolates a sign vector, whereas in Lemma 10.1 it is the opposite.

Recall the definition of $\mathcal{V}_T$ from (38) and define $u'_T(s) = u'_\infty + (s-1)\mathcal{V}_T$. We remark that (93) holds with $u_T(s)$ replaced by $u'_T(s)$ because of (91).

Lemma 10.2. *Let $v = (v_1, \cdots, v_s)^\top \in \{-1, 1\}^s$ be a sign vector. Assume that (93) holds with $u_T(s)$ replaced by $u'_T(s) < 1/2$. Under Assumption 3.2, there exist unique $\alpha, \xi \in \mathbb{R}^s$ such that:*

$$\eta_{\alpha,\xi}(\theta_k^\star) = 0 \quad \textit{and} \quad \eta^{[1]}_{\alpha,\xi}(\theta_k^\star) = v_k \quad \textit{for} \quad 1 \leq k \leq s. \tag{115}$$

Furthermore, we have:

$$\|\alpha\|_{\ell_\infty} \leq \frac{u'_T(s)}{1 - 2u'_T(s)} \quad \textit{and} \quad \|\xi\|_{\ell_\infty} \leq \frac{1 - u'_T(s)}{1 - 2u'_T(s)}. \tag{116}$$

Proof. Thus, with 0_s the vector of size s with all its components equal to zero and Γ defined by (92), Equation (115) is equivalent to:

$$\Gamma \begin{pmatrix} \alpha \\ \xi \end{pmatrix} = \begin{pmatrix} 0_s \\ v \end{pmatrix}. \tag{117}$$

According to the proof of Lemma 10.1, the matrices $\Gamma_{SC} = \Gamma^{[0,0]} - \Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}$, $\Gamma^{[1,1]}$ and Γ are non singular. Thus the vectors α and ξ exist and are uniquely determined by (117). From Lemma A.3, we deduce that:

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

Using (101), (93) and (100) and replacing $u_T(s)$ by $u'_T(s)$, we easily obtain the inequalities (116). □

We fix the sign vector $v = (v_1, \cdots, v_s)^\top \in \{-1, 1\}^s$ and consider $p_{\alpha,\xi}$ and $\eta_{\alpha,\xi}$ given by (94) and (95), with α and ξ given by Lemma 10.2.

Proof of *(i)* **from Assumption 6.2** with $c_N = (L_{0,2} + L_{2,1} + 7)/8$. We define the function $f : \theta \mapsto \eta_{\alpha,\xi}(\theta) - v_\ell \operatorname{sign}(\theta - \theta_\ell^\star)\mathfrak{d}_T(\theta, \theta_\ell^\star)$ on Θ. To prove the Property *(i)*, we will bound the second covariant derivative of f, that is $f^{[2]} := \tilde{D}_{2;T}[f]$, and apply Lemma A.4-*(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 $f(\theta) = \eta_{\alpha,\xi}(\theta) - G_T(\theta) + G_T(\theta_\ell^\star)$. We deduce that f is twice continuously differentiable on Θ; and elementary calculations give $f^{[2]} = \eta_{\alpha,\xi}^{[2]}$.

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 $\vartheta_{\ell,\theta}^\star$ from the proof of Proposition 7.4. Since $f^{[2]} = \eta_{\alpha,\xi}^{[2]}$, we deduce from (109) that:

$$|f^{[2]}(\theta)| \leq \left\| I + \Gamma_{\ell,\theta}^{[2,0]} \right\|_{\text{op}} \|\alpha\|_{\ell_\infty} + \|\alpha\|_{\ell_\infty} |\mathcal{K}_T^{[2,0]}(\theta, \theta_\ell^\star)| + \|\xi\|_{\ell_\infty} \left\| \Gamma_{\ell,\theta}^{[2,1]} \right\|_{\text{op}} + \|\xi\|_{\ell_\infty} |\mathcal{K}_T^{[2,1]}(\theta, \theta_\ell^\star)|. \tag{118}$$

Notice that (102) holds with $u_T(s)$ replaced by $u_T'(s)$. Using (111) and (112) and the bounds (116) on α and ξ from Lemma 10.2, we get:

$$|f^{[2]}(\theta)| \leq \frac{u_T'(s)}{1 - 2u_T'(s)}(L_{2,0} + \mathcal{V}_T + u_T'(s)) + \frac{1 - u_T'(s)}{1 - 2u_T'(s)}(L_{2,1} + \mathcal{V}_T + u_T'(s)).$$

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

$$|f^{[2]}(\theta)| \leq \frac{1}{4}L_{2,0} + \frac{5}{4}L_{2,1} + \frac{7}{4}.$$

Since $f(\theta_\ell^\star) = 0$ and $f^{[1]}(\theta_\ell^\star) = 0$ as well, using Lemma A.4 *(i)*, we get, with $c_N = (L_{2,0} + 5L_{2,1} + 7)/8$:

$$\left|\eta_{\alpha,\xi}(\theta) - v_\ell \operatorname{sign}(\theta - \theta_\ell^\star)\, \mathfrak{d}_T(\theta, \theta_\ell^\star)\right| = |f(\theta)| \leq c_N\, \mathfrak{d}_T(\theta, \theta_\ell^\star)^2.$$

Proof of *(ii)* **from Assumption 6.2** with $c_F = (5L_{1,0} + 7)/4$. Let $\theta \in \Theta_T$, we shall prove that $|\eta_{\alpha,\xi}(\theta)| \leq c_F$. 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 (103) that:

$$|\eta_{\alpha,\xi}(\theta)| \leq \|\alpha\|_{\ell_\infty} \left\| \Gamma_{\ell,\theta}^{[0,0]} - I \right\|_{\text{op}} + \|\alpha\|_{\ell_\infty} |\mathcal{K}_T(\theta, \theta_\ell^\star)| + \|\xi\|_{\ell_\infty} \left\| \Gamma_{\ell,\theta}^{[1,0]\top} \right\|_{\text{op}} + \|\xi\|_{\ell_\infty} |\mathcal{K}_T^{[0,1]}(\theta, \theta_\ell^\star)|.$$

Using (93), (33), (106) and the bounds (116) on α and ξ from Lemma 10.2, we get:

$$|\eta_{\alpha,\xi}(\theta)| \leq \frac{u_T'(s)}{1 - 2u_T'(s)}\left(1 + u_T'(s)\right) + \frac{1 - u_T'(s)}{1 - 2u_T'(s)}\left(L_{1,0} + \mathcal{V}_T + u_T'(s)\right).$$

By assumption, we have $u_T'(s) \leq 1/6$, and thus $\frac{1}{1-2u_T'(s)} \leq 3/2$. Since $\mathcal{V}_T \leq 1$, we obtain:

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

Proof of *(iii)* **from Assumption 6.2** with $c_B = 2$. Using very similar arguments as in the proof of (114) (taking care that the upper bound of the ℓ_∞ norm of α and ξ are given by (116)) we also get $\|p_{\alpha,\xi}\|_T \leq 2\sqrt{s}$.

We proved that *(i)*-*(ii)* from Assumption 6.2 stand for any $\theta \in \Theta_T$. Hence Assumption 6.2 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 7.5.

References

[1] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York, 1992. Reprint of the 1972 edition. 42, 46

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008. With a foreword by Paul Van Dooren. 14

[3] Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis*. Springer, Berlin, third edition, 2006. A hitchhiker's guide. 11

[4] Wolfgang Arendt, Charles J. K. Batty, Matthias Hieber, and Frank Neubrander. *Vector-valued Laplace transforms and Cauchy problems*, volume 96 of *Monographs in Mathematics*. Birkhäuser/Springer Basel AG, Basel, second edition, 2011. 11

[5] Jean-Marc Azaïs, Yohann de Castro, and Fabrice Gamboa. Spike detection from inaccurate samplings. *Appl. Comput. Harmon. Anal.*, 38(2):177–195, 2015. 5

[6] Jean-Marc Azaïs and Mario Wschebor. *Level sets and extrema of random processes and fields*. John Wiley & Sons, Inc., Hoboken, NJ, 2009. 42

[7] Brett Bernstein and Carlos Fernandez-Granda. Deconvolution of point sources: a sampling theorem and robustness guarantees. *Comm. Pure Appl. Math.*, 72(6):1152–1230, 2019. 6

[8] Brett Bernstein, Sheng Liu, Chrysa Papadaniil, and Carlos Fernandez-Granda. Sparse recovery beyond compressed sensing: separable nonlinear inverse problems. *IEEE Trans. Inform. Theory*, 66(9):5904–5926, 2020. 6

[9] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. 5, 8, 17

[10] 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. 6

[11] 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, 6

[12] Claire Boyer, Yohann De Castro, and Joseph Salmon. Adapting to unknown noise level in sparse deconvolution. *Inf. Inference*, 6(3):310–348, 2017. 6, 7, 17, 22

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

[14] 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. 6

[15] Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007. 5

[16] Emmanuel J. Candès and Mark A. Davenport. How well can we estimate a sparse vector? *Appl. Comput. Harmon. Anal.*, 34(2):317–323, 2013. 6

[17] Emmanuel J. Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *J. Fourier Anal. Appl.*, 19(6):1229–1254, 2013. 5, 7, 17, 18

[18] Emmanuel J. Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Comm. Pure Appl. Math.*, 67(6):906–956, 2014. 6, 7, 17

[19] Emmanuel J. Candès and Yaniv Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inform. Theory*, 57(11):7235–7254, 2011. 17

[20] Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005. 17

[21] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, pages 1–46, 2021. 6

[22] Y. De Castro, S. Gadat, C. Marteau, and C. Maugis-Rabusseau. SuperMix: sparse regularization for mixtures. *Ann. Statist.*, 49(3):1779–1809, 2021. 6, 17

[23] Yohann de Castro and Fabrice Gamboa. Exact reconstruction using Beurling minimal extrapolation. *J. Math. Anal. Appl.*, 395(1):336–354, 2012. 5, 9

[24] 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. 6, 13

[25] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006. 5

[26] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Found. Comput. Math.*, 15(5):1315–1355, 2015. 5, 6, 7, 17

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

[28] Clément Elvira, Rémi Gribonval, Charles Soussen, and Cédric Herzet. When does OMP achieve exact recovery with continuous dictionaries? *Appl. Comput. Harmon. Anal.*, 51:374–413, 2021. 6
[29] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition, 2015. 42
[30] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013. 19
[31] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016. 4
[32] Mohammad Golbabaee and Clarice Poon. An off-the-grid approach to multi-compartment magnetic resonance fingerprinting. *arXiv preprint arXiv:2011.11193*, 2020. 6
[33] G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM J. Numer. Anal.*, 10:413–432, 1973. 5
[34] Linda Kaufman. A variable projection method for solving separable nonlinear least squares problems. *Nordisk Tidskr. Informationsbehandling (BIT)*, 15(1):49–57, 1975. 5
[35] Alois Kneip and Theo Gasser. Convergence and consistency results for self-modeling nonlinear regression. *Ann. Statist.*, 16(1):82–112, 1988. 5
[36] Serge Lang. *Real and functional analysis*, volume 142 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition, 1993. 11
[37] John M. Lee. *Introduction to Riemannian manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, Cham, 2018. Second edition of [MR1468735]. 13
[38] 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. 8
[39] Stéphane Mallat. *A wavelet tour of signal processing : the sparse way*. Elsevier/Academic Press, Amsterdam, third edition, 2009. With contributions from Gabriel Peyré. 13
[40] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997. 5
[41] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*, 2021. 6, 7, 13, 14, 17, 18, 19, 33, 45
[42] 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. 6, 8
[43] Takashi Sakai. *Riemannian geometry*, volume 149 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1996. Translated from the 1992 Japanese original by the author. 13
[44] Geoffrey Schiebinger, Elina Robeva, and Benjamin Recht. Superresolution without separation. *Inf. Inference*, 7(1):1–30, 2018. 6, 18
[45] Gongguo Tang. Resolution limits for atomic decompositions via markov-bernstein type inequalities. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 548–552, 2015. 18
[46] Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Sparse recovery over continuous dictionaries-just discretize. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2013. 5
[47] Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Near minimax line spectral estimation. *IEEE Trans. Inform. Theory*, 61(1):499–512, 2015. 6, 7, 9, 17, 22
[48] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. 3, 5
[49] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004. 19
[50] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. 4
[51] Sara van de Geer. *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, [Cham], 2016. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. 5, 17
[52] 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. 17

Appendix A. Auxiliary Lemmas

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

A.1. **Tail bound inequalities for suprema of Gaussian processes.** In order to prove Theorems 2.1 and 2.5, we provide in Lemma A.1 a bound with high probability of the supremum of a Gaussian process given by $\theta \mapsto \langle w_T, h(\theta)\rangle_T$, where w_T is a noise process and h is a function from Θ, an interval of $\mathbb{R}$, to the Hilbert space $(H_T, \langle\cdot,\cdot\rangle_T)$. The next lemma is in the spirit of [6, Proposition 4.1] (where one assumes that the Gaussian process has unitary variance); its proof is given at the end of this section and relies on Lemma A.2.

We denote by $\mathfrak{d}_T$ the Riemannian metric associated to the kernel $\mathcal{K}_T$, see also Section 4.2. Recall definitions (21) and (23) and set $f^{[1]}(\theta) = \tilde{D}_{1,T}[f](\theta) = \partial_\theta f(\theta)/\sqrt{g_T(\theta)}$ with g_T defined in (31).

Lemma A.1. *Let $T \in \mathbb{N}$ be fixed. Suppose that Assumptions 3.1 and 3.2 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 $\theta \in \Theta_T$:*

$$\|h(\theta)\|_T \leq C_1 \quad \textit{and} \quad \left\|h^{[1]}(\theta)\right\|_T \leq C_2. \tag{119}$$

Let w_T be an H_T-valued Gaussian noise such that Assumption 1.1 holds, and consider the Gaussian process $X = (X(\theta) = \langle h(\theta), w_T\rangle_T, \theta \in \Theta)$. Then, we have for $u > 0$:

$$\mathbb{P}\left(\sup_{\theta\in\Theta_T} |X(\theta)| \geq u\right) \leq c \cdot \left(\sigma \frac{|\Theta_T|\sqrt{\Delta_T}}{u} \vee 1\right) \mathrm{e}^{-u^2/(4\sigma^2\Delta_T C_1^2)}, \tag{120}$$

where $|\Theta_T|$ denotes the Riemannian length of the interval Θ_T and $c = 2C_2 + 1$.

We first state a technical lemma.

Lemma A.2. *Let $I \subset \mathbb{R}$ be an interval. Assume that $X = (X(\theta),\, \theta \in I)$ is a real centered Gaussian process with Lipschitz sample paths. Then, for all $u > 0$ and an arbitrary $\theta_0 \in I$, we have:*

$$\mathbb{P}\left(\sup_I X \geq u\right) \leq \frac{1}{u}\int_I \sqrt{\mathrm{Var}(X'(\theta))}\, \mathrm{e}^{-u^2/(4\mathrm{Var}(X(\theta)))}\, \mathrm{d}\theta + \frac{1}{2}\mathrm{e}^{-u^2/(2\mathrm{Var}(X(\theta_0)))}. \tag{121}$$

Proof. We first start with a general remark on Lipschitz functions on $\mathbb{R}$. Let f be a real-valued Lipschitz function defined on an interval $I \subset \mathbb{R}$. Let $b > a$ and set $f_{a,b} = \min(\max(f, a), b)$. The function $f_{a,b}$ is also Lipschitz and, thanks to [29, Theorem 3.3 p107], we get that $f'_{a,b} = f' = 0$ a.e. on $\{x \in I :\ f(x) = a \text{ or } b\}$ and thus $f'_{a,b} = f'\, \mathbf{1}_{\{f\in(a,b)\}}$ a.e. on I. We deduce that:

$$\sup f_{a,b} - \inf f_{a,b} \leq \int_I |f'_{a,b}(x)|\,\mathrm{d}x = \int_I |f'(x)|\,\mathbf{1}_{\{f(x)\in(a,b)\}}\,\mathrm{d}x.$$

Using this inequality, we obtain that for any $x_0 \in I$:

$$\int_a^b \mathbf{1}_{\{\sup_I f > t\}}\,\mathrm{d}t = \int_a^b \mathbf{1}_{\{\sup_I f_{a,b} > t\}}\,\mathrm{d}t = \sup f_{a,b} - a \leq (b-a)\mathbf{1}_{\{f(x_0)\geq a\}} + \int_I |f'(x)|\,\mathbf{1}_{\{f(x)\in(a,b)\}}\,\mathrm{d}x. \tag{122}$$

Then, applying Inequality (122) to the function X and taking the expectation, we get, with $M = \sup_I X$, $a = u > 0$, $b = u + \varepsilon$, $\varepsilon > 0$ and $x_0 = \theta_0$:

$$\int_u^{u+\varepsilon} \mathbb{P}(M \geq t)\,\mathrm{d}t \leq \varepsilon\mathbb{P}(X(\theta_0) \geq u) + \int_I \mathbb{E}\left[|X'(\theta)|\mathbf{1}_{\{u<X(\theta)<u+\varepsilon\}}\right]\,\mathrm{d}\theta. \tag{123}$$

The random variable $X(\theta_0)$ is a centered Gaussian variable and therefore we have:

$$\mathbb{P}\left(X(\theta_0) \geq u\right) = \int_u^{+\infty} \frac{\mathrm{e}^{-x^2/(2\mathrm{Var}(X(\theta_0)))}}{\sqrt{2\pi\mathrm{Var}(X(\theta_0))}}\,\mathrm{d}x \leq \frac{1}{2}\,\mathrm{e}^{-u^2/2\mathrm{Var}(X(\theta_0))}, \tag{124}$$

where we used for the inequality that $\int_u^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t \leq \frac{\sqrt{\pi}}{2}\,\mathrm{e}^{-u^2}$ holds for $u > 0$, see [1, Formula 7.1.13]. Notice that (124) 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 (123). Since (X', X) is also a Gaussian process, we can write:

$$X'(\theta) = \alpha_\theta X(\theta) + \beta_\theta G,$$

where G is a standard Gaussian random variable independent of $X(\theta)$ 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 $\mathrm{Var}(X(\theta)) = 0$. We get $|X'(\theta)| \leq |\alpha_\theta X(\theta)| + |\beta_\theta|\,|G|$. Since G is independent of $X(\theta)$ and $u > 0$, we deduce that:

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

Letting ε goes to 0 in (123), using (124) the right continuity of the cdf of M and the monotonicity of the density $p_{X(\theta)}(u)$ of the law of $X(\theta)$, we deduce that:

$$\mathbb{P}(M \geq u) \leq \frac{1}{2}\,\mathrm{e}^{-u^2/2\mathrm{Var}(X(\theta_0))} + \int_I \left(|\alpha_\theta| u + \sqrt{\frac{2}{\pi}}\,|\beta_\theta|\right) p_{X(\theta)}(u)\,\mathrm{d}\theta, \tag{125}$$

where by convention $p_{X(\theta)}(u)$ is taken equal to 0 if $\mathrm{Var}(X(\theta)) = 0$. We now bound the second term of the right-hand side of (125) in two steps. Using that $\beta_\theta^2 \leq \mathrm{Var}(X'(\theta))$ and the inequality $\mathrm{e}^{-x^2} \leq \mathrm{e}^{-x^2/2}/\sqrt{2}\,x$ for $x > 0$, we get that:

$$\sqrt{\frac{2}{\pi}}\,|\beta_\theta|\,p_{X(\theta)}(u) \leq \frac{1}{\pi}\,\frac{\sqrt{\mathrm{Var}(X'(\theta))}}{u}\,\mathrm{e}^{-u^2/4\mathrm{Var}(X(\theta))}. \tag{126}$$

Thanks to the Cauchy-Schwarz inequality, we get $|\alpha_\theta| \leq \sqrt{\mathrm{Var}(X'(\theta))}/\sqrt{\mathrm{Var}(X(\theta))}$. Using also the inequality $\mathrm{e}^{-x^2} \leq 3\,\mathrm{e}^{-x^2/2}/4x^2$ for $x > 0$, we get that:

$$|\alpha_\theta| u\, p_{X(\theta)}(u) \leq \frac{3}{4}\sqrt{\frac{2}{\pi}}\,\frac{\sqrt{\mathrm{Var}(X'(\theta))}}{u}\,\mathrm{e}^{-u^2/4\mathrm{Var}(X(\theta))}. \tag{127}$$

Notice that (126) and (127) hold also if $\mathrm{Var}(X(\theta)) = 0$. Using that $\frac{3}{4}\sqrt{\frac{2}{\pi}} + \frac{1}{\pi} \simeq 0.92 \leq 1$, we deduce (121) from (125), (126) and (127). □

Proof of Lemma A.1. We first consider the case $\Theta_T = [\theta_0, \theta_1]$ and let $\gamma : [0,1] \to [\theta_0, \theta_1]$ be a minimizing path with respect to the Riemannian metric $\mathfrak{d}_T$ (see Remark 4.1); in particular we have $|\gamma'(s)|\sqrt{g_T(\gamma(s))} = \mathfrak{d}_T(\theta_0, \theta_1)$. Thanks to (11), the Gaussian process $\tilde{X} = (\tilde{X}(s) = X(\gamma(s)), s \in [0,1])$ is of class $\mathcal{C}^1$ on $s \in [0,1]$, with derivative $\tilde{X}'(s) = \gamma'(s)\,X'(\gamma(s)) = \gamma'(s)\,\langle \partial_\theta h(\gamma(s)), w_T\rangle_T$. Then, according to Lemma A.2, Inequality (121) holds. By Assumption 1.1, we have for all $\theta \in \Theta_T$:

$$\mathrm{Var}(X(\theta))) \leq \sigma^2 \Delta_T \|h(\theta)\|_T^2 \leq \sigma^2 \Delta_T C_1^2 \quad \text{and} \quad \frac{\mathrm{Var}(X'(\theta)))}{g_T(\theta)} \leq \sigma^2 \Delta_T \left\|h^{[1]}(\theta)\right\|_T^2 \leq \sigma^2 \Delta_T C_2^2.$$

Plugging those bounds in Inequality (121) with $|\gamma'(s)|\sqrt{g_T(\gamma(s))} = \mathfrak{d}_T(\theta_0, \theta_1)$, we obtain:

$$\begin{aligned}\mathbb{P}\left(\sup_{[\theta_0,\theta_1]} X \geq u\right) &\leq \frac{1}{u}\sqrt{\sigma^2\Delta_T}C_2\,\mathrm{e}^{-u^2/(4\sigma^2\Delta_T C_1^2)} \int_0^1 |\gamma'(s)|\sqrt{g_T(\gamma(s))}\,\mathrm{d}s + \frac{1}{2}\,\mathrm{e}^{-u^2/(2\sigma^2\Delta_T C_1^2)} \\ &\leq \left(C_2 + \frac{1}{2}\right)\left(\sigma\frac{\mathfrak{d}_T(\theta_0,\theta_1)\sqrt{\Delta_T}}{u} \vee 1\right)\mathrm{e}^{-u^2/(4\sigma^2\Delta_T C_1^2)}.\end{aligned}$$

Since $\mathbb{P}\left(\sup_{[\theta_0,\theta_1]} |X| \geq u\right) \leq 2\,\mathbb{P}\left(\sup_{[\theta_0,\theta_1]} X \geq u\right)$, we obtain that (120) holds for Θ_T a bounded closed interval. Then, use monotone convergence and the continuity of X to get (120) for any interval Θ_T. □

A.2. **Schur complement.** The following Lemma is a classical result on the Schur complement.

Lemma A.3 (Schur complement). *Let $M \in \mathbb{R}^{n\times n}$ be a matrix composed of blocks $A \in \mathbb{R}^{(n-k)\times(n-k)}$, $B \in \mathbb{R}^{(n-k)\times k}$, $C \in \mathbb{R}^{k\times(n-k)}$, $D \in \mathbb{R}^{k\times k}$:*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

Assume that D and $S_1 = A - BD^{-1}C$ are non singular. Then, the system:

$$M \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}. \tag{128}$$

with $x \in \mathbb{R}^{n-k}$, $y \in \mathbb{R}^k$, $a \in \mathbb{R}^{n-k}$ and $b \in \mathbb{R}^k$, has a unique solution given by:

$$x = S_1^{-1}a - S_1^{-1}BD^{-1}b \quad \text{and} \quad y = D^{-1}b - D^{-1}CS_1^{-1}a + D^{-1}CS_1^{-1}BD^{-1}b.$$

A.3. **Proofs of Lemmas in Section 4.**

Proof of Lemma 4.2. For simplicity, we remove the subscript $\mathcal{K}$ and for example write $f^{[1]} = \tilde{D}_1[f] = D_1[f]/\sqrt{g}$. Recall that G, a primitive of $\sqrt{g}$, is continuous increasing and thus induces a one-to-one map from Θ to its image. Following Remark 4.1, we consider the minimizing path $\gamma : [0,1] \to \Theta$ from θ_0 to θ defined by $\gamma_s = G^{-1}(as+b)$, with $b = G(\theta_0)$ and $a = G(\theta) - G(\theta_0)$. Thus, we have $\mathcal{L}(\gamma) = \mathfrak{d}(\theta,\theta_0)$. The minimizing path from θ_0 to θ has constant speed thus equal to $\mathfrak{d}(\theta_0,\theta)$. From the explicit expression of γ, we get in fact that $\dot{\gamma}_t\sqrt{g(\gamma_t)} = A$ for $t \in [0,1]$, where $A = \operatorname{sign}(\theta-\theta_0)\,\mathfrak{d}(\theta,\theta_0)$. Thus, we have:

$$f(\theta) - f(\theta_0) = f(\gamma_1) - f(\gamma_0) = \int_0^1 \dot{\gamma}_t\, f'(\gamma_t)\; \mathrm{d}t = A\int_0^1 \tilde{D}_1[f](\gamma_t)\; \mathrm{d}t = A\int_0^1 f^{[1]}(\gamma_t)\; \mathrm{d}t, \tag{129}$$

where we used (12) and that the derivative of $f\circ\gamma_t$ is $\dot{\gamma}_t\, f'\circ\gamma_t$ for the second equality and the definition of $\tilde{D}_1[f]$ as well as the equality $\dot{\gamma}_t\sqrt{g(\gamma_t)} = A$ for the last.

Using (129) for f and θ replaced by $f^{[1]}$ and $\gamma(t)$ for some $t \in [0,1]$, we get thanks to (24) that:

$$f^{[1]}(\gamma_t) = f^{[1]}(\theta_0) + \tilde{A}\int_0^1 f^{[2]}(\tilde{\gamma}_s)\; \mathrm{d}s,$$

where $\tilde{\gamma}$ is a geodesic from θ_0 to γ_t and $\tilde{A} = \dot{\tilde{\gamma}}_s\sqrt{g(\tilde{\gamma}_s)}$. Since γ is itself a geodesic, we deduce that $\tilde{\gamma}_s = \gamma_{st}$, and thus $\tilde{A} = tA$. Plugging this in (129), we get:

$$f(\theta) - f(\theta_0) = A\, f^{[1]}(\theta_0) + A^2\int_{[0,1]^2} f^{[2]}(\gamma_{st})\; t\,\mathrm{d}t\,\mathrm{d}s = A\, f^{[1]}(\theta_0) + A^2\int_0^1 (1-r)\, f^{[2]}(\gamma_r)\; \mathrm{d}r.$$

This gives (25). □

Proof of Lemma 4.3. Recall that by Assumption 3.2 the function ϕ_T is $\mathcal{C}^3$. According to (11), we have that for any $i,j \in \{0,\dots,3\}$ and any $\theta,\theta' \in \Theta$:

$$\partial^{i,j}_{\theta,\theta'}\,\langle\phi_T(\theta),\phi_T(\theta')\rangle_T = \left\langle \partial^i_\theta\phi_T(\theta), \partial^j_{\theta'}\phi_T(\theta')\right\rangle_T. \tag{130}$$

This and (21), (23), (26) and (27) readily imply (32). The first equality of (33) comes from Cauchy-Schwarz's inequality. The second is clear. We also have:

$$\langle\partial_\theta\phi_T(\theta),\phi_T(\theta)\rangle_T = \frac{1}{2}\,\partial_\theta\|\phi_T(\theta)\|^2 = 0 \tag{131}$$

Since the right hand-side is also equal to $\sqrt{g_T(\theta)}\,\mathcal{K}_T^{[1,0]}(\theta,\theta)$ thanks to (32), we get the third equality of (33). Taking the derivative with respect to θ in (131) yields $g_T(\theta) = \langle \partial_\theta \phi_T(\theta), \partial_\theta \phi_T(\theta)\rangle = -\langle \partial_\theta^2 \phi_T(\theta), \phi_T(\theta)\rangle$. Thanks to (22), we get $\partial_\theta^2 \phi_T = g_T \tilde{D}_{2,T}[\phi_T] + (1/2g_T) g_T' \partial_\theta \phi_T$. Using (32) and (131) again, we deduce that $\langle \partial_\theta^2 \phi_T(\theta), \phi_T(\theta)\rangle = g_T(\theta)\,\mathcal{K}_T^{[2,0]}(\theta,\theta)$. This gives the fourth equality of (33). Eventually, we deduce from (32), (22) and (23) that:

$$g_T(\theta)^{3/2}\,\mathcal{K}_T^{[2,1]}(\theta,\theta) = \langle \partial_\theta^2 \phi_T(\theta), \partial_\theta \phi_T(\theta)\rangle - \frac{1}{2}\frac{g_T'(\theta)}{g_T(\theta)}\langle \partial_\theta \phi_T(\theta), \partial_\theta \phi_T(\theta)\rangle.$$

Then, use that $g_T'(\theta) = 2\langle \partial_\theta^2 \phi_T(\theta), \partial_\theta \phi_T(\theta)\rangle$ to deduce that $\mathcal{K}_T^{[2,1]}(\theta,\theta) = 0$. □

A.4. **Control on f from its derivatives $f^{[2]}$.** The proof of the next lemma is similar to the proof of [41, Lemma 2] and is left to the reader. Recall from (33) that $\mathcal{K}_T^{[2,0]}(\theta,\theta) = -1$ on Θ.

Lemma A.4. *Suppose Assumptions 3.1 and 3.2 on the dictionary hold. Let f be a real valued function defined on an interval Θ of class $\mathcal{C}^2$. Let $\theta_0 \in \Theta$. Set for $i = 1, 2$, $f^{[i]} = \tilde{D}_{i;T}[f]$ (see (23)).*

(i) Assume $f(\theta_0) = 0$, $f^{[1]}(\theta_0) = 0$ and that there exist $\delta > 0$ and $r > 0$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$:

$$|f^{[2]}(\theta)| \leq 2\delta. \tag{132}$$

Then, we have $|f(\theta)| \leq \delta\,\mathfrak{d}_T(\theta,\theta_0)^2$, for any $\theta \in \mathcal{B}_T(\theta_0, r)$.

(ii) Let $\Theta_T \subset \Theta$ be an interval and suppose that $L \geq \sup_{\Theta_T^2} |\mathcal{K}_T^{[2,0]}|$ is finite and there exist $\varepsilon > 0$ and $r \in (0, L^{-\frac{1}{2}})$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$, $-\mathcal{K}_T^{[2,0]}(\theta,\theta_0) \geq \varepsilon$. Assume that $\mathcal{B}_T(\theta_0, r) \subset \Theta_T$, $f(\theta_0) = v \in \{-1; 1\}$, $f^{[1]}(\theta_0) = 0$ and that there exists $\delta \in (0, \varepsilon)$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$:

$$|f^{[2]}(\theta) - v\mathcal{K}_T^{[2,0]}(\theta,\theta_0)| \leq \delta. \tag{133}$$

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

A.5. **Proof of Lemma 8.1.** We keep the notations from Section 8.1. In order to prove that the constants c_0, c_1 and c_2 do not depend on the scaling factor σ_0, we shall rewrite ρ_T and $\mathcal{V}_T$ defined in (36) and (38) using a change of scale. To do so, we define $\varphi^0(\theta) = k(\cdot - \theta)$ for $\theta \in \Theta$; the grid $t_1^0, \cdots, t_T^0$ where $t_j^0 = t_j/\sigma_0$; the Hilbert space $L^2(\lambda_T^0)$ with $\lambda_T^0 = \Delta_T\,\sigma_0^{-1}\sum_{j=1}^T \delta_{t_j^0}$, endowed with its natural scalar product noted $\langle\cdot,\cdot\rangle_{\lambda_T^0}$ and norm $\|\cdot\|_{\lambda_T^0}$; the parameter space $\Theta_T^0 = [a_T(1-\epsilon)\sigma_0^{-1}, b_T(1-\epsilon)\sigma_0^{-1}]$. Since the scaling factor σ_0 is fixed, the measures $(\lambda_T^0, T \geq 2)$ converge vaguely towards the Lebesgue measure λ_∞ on $\mathbb{R}$. We shall also consider another kernel:

$$\mathcal{K}_T^0(\theta,\theta') = \left\langle \phi_T^0(\theta), \phi_T^0(\theta')\right\rangle_{\lambda_T^0} \quad \text{with} \quad \phi_T^0 = \varphi^0/\left\|\varphi^0\right\|_{\lambda_T^0},$$

and the limit kernel $\mathcal{K}_\infty^0(\theta,\theta') = \left\langle \phi_\infty^0(\theta), \phi_\infty^0(\theta')\right\rangle_\infty$ with $\phi_\infty^0 = \varphi^0/\left\|\varphi^0\right\|_\infty$. For any $T \in \mathbb{N}\cup\{+\infty\}$, the kernel $\mathcal{K}_T^0$ is of class $\mathcal{C}^{3,3}$ on Θ^2 and for $i, j \in \{0, \ldots, 3\}$ and $\theta, \theta' \in \Theta$, we have:

$$\mathcal{K}_T^{[i,j]}(\theta,\theta') = \mathcal{K}_T^{0[i,j]}\left(\frac{\theta}{\sigma_0}, \frac{\theta'}{\sigma_0}\right) \quad \text{and} \quad \frac{1}{\sigma_0^2} g_{\mathcal{K}_T^0}\left(\frac{\theta}{\sigma_0}\right) = g_{\mathcal{K}_T}(\theta).$$

We can now rewrite ρ_T and $\mathcal{V}_T$ by using a change of scale and we get:

$$\rho_T = \max\left(\sup_{\Theta_T^0}\sqrt{\frac{g_{\mathcal{K}_T^0}}{g_{\mathcal{K}_\infty^0}}}, \sup_{\Theta_T^0}\sqrt{\frac{g_{\mathcal{K}_\infty^0}}{g_{\mathcal{K}_T^0}}}\right),$$

and

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

Thus, bounding ρ_T and $\mathcal{V}_T$ amounts to controling the proximity between the kernels $\mathcal{K}_T^0$ and $\mathcal{K}_\infty^0$.

First, we provide an upper bound for any $i, j \in \{0, \cdots, 3\}$ of:

$$B_{i,j}(T) = \sup_{\theta,\theta' \in \Theta_T^0} \left| \left\langle \partial_\theta^i \varphi^0(\theta), \partial_\theta^j \varphi^0(\theta') \right\rangle_{\lambda_T^0} - \left\langle \partial_\theta^i \varphi^0(\theta), \partial_\theta^j \varphi^0(\theta') \right\rangle_\infty \right|. \tag{134}$$

Notice that:

$$\partial_\theta^i \partial_t^j \varphi^0(\theta, t) = (-1)^j \, k^{(i+j)}(\theta - t).$$

Recall the polynomials P_i defined as $k^{(i)} = P_i \, k$. And set $M = \max_{0 \leq i \leq 4} \sup |P_i| \sqrt{k}$. It is elementary to get that for $\theta, \theta' \in \mathbb{R}$:

$$\left| (\Delta_T/\sigma_0) \sum_{k=1}^{T} \partial_\theta^i \varphi^0(\theta, t_k^0) \partial_\theta^j \varphi^0(\theta', t_k^0) - \int_{a_T/\sigma_0}^{b_T/\sigma_0} \partial_\theta^i \varphi^0(\theta, t) \partial_\theta^j \varphi^0(\theta', t) \, \mathrm{d}t \right| \leq 4\sqrt{\pi} \, \Delta_T M^2 \sigma_0^{-1}.$$

We have for $\theta, \theta' \in \Theta_T^0$ that:

$$\begin{aligned} \left| \int_{\mathbb{R} \setminus [a_T/\sigma_0, b_T/\sigma_0]} \partial_\theta^i \varphi^0(\theta, t) \partial_\theta^j \varphi^0(\theta', t) \, \mathrm{d}t \right| &\leq \left| \int_{b_T/\sigma_0}^{+\infty} \partial_\theta^i \varphi^0(\theta, t) \partial_\theta^j \varphi^0(\theta', t) \, \mathrm{d}t \right| + \left| \int_{-\infty}^{a_T/\sigma_0} \partial_\theta^i \varphi^0(\theta, t) \partial_\theta^j \varphi^0(\theta', t) \, \mathrm{d}t \right| \\ &\leq 2M^2 \int_{\epsilon b_T/\sigma_0}^{+\infty} k(t) \, \mathrm{d}t \\ &\leq 2\sqrt{\pi} \, M^2 \, \mathrm{e}^{-\epsilon^2 b_T^2/2\sigma_0^2}, \end{aligned}$$

where we used that $2\int_u^{+\infty} \mathrm{e}^{-t^2} \, \mathrm{d}t \leq \sqrt{\pi} \, \mathrm{e}^{-u^2}$ for $u > 0$, see formula 7.1.13 in [1]. We deduce that:

$$B_{i,j}(T) \leq 4\sqrt{\pi} \, \Delta_T M^2 \sigma_0^{-1} + 2\sqrt{\pi} \, M^2 \, \mathrm{e}^{-\epsilon^2 b_T^2/2\sigma_0^2} \leq 2\sqrt{\pi} \, M^2 \gamma_T,$$

with $\gamma_T = 2\Delta_T \sigma_0^{-1} + \sqrt{\pi} \, \mathrm{e}^{-\epsilon^2 b_T^2/2\sigma_0^2}$.

Similar arguments as above yield that:

$$\sup_{\theta \in \Theta_T^0} \left| \left\| \varphi^0(\theta) \right\|_{\lambda_T^0}^2 - \left\| \varphi^0(\theta) \right\|_\infty^2 \right| \leq \gamma_T.$$

so that $\left\| \varphi^0(\theta) \right\|_{\lambda_T^0}^2 \geq \sqrt{\pi} - \gamma_T$ for all $\theta \in \Theta_T^0$. It is then easy to deduce that $\sup_{\Theta_T^0} |g_{\mathcal{K}_T^0} - g_{\mathcal{K}_\infty^0}|$ is bounded by a constant times γ_T when γ_T is smaller than a universal finite constant. Up to taking γ_T smaller than some universal finite constant, this and the fact that $g_{\mathcal{K}_\infty^0} = 1/2$ give the second part of (52). Then use formulae for the derivatives of the kernels, see (30) and (23), to get the first part of (52).

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