Sparse Regularization for Mixture Problems

Y. De Castro[⋆], **S. Gadat**[∘], **C. Marteau**[•] **and C. Maugis-Rabusseau**[†].

[⋆] *CERMICS, École des Ponts ParisTech*
6 et 8 avenue Blaise Pascal, 77420 Champs-sur-Marne, France.

[∘] *Toulouse School of Economics, Université Toulouse 1 Capitole*
21 allée de Brienne, 31000 Toulouse, France.
Institut Universitaire de France.

[•] *Univ Lyon, Université Claude Bernard Lyon 1*
CNRS UMR 5208, Institut Camille Jordan
F-69622 Villeurbanne, France

[†] *Institut de Mathématiques de Toulouse; UMR5219*
Université de Toulouse; CNRS
INSA, F-31077 Toulouse, France

Abstract: This paper investigates the statistical estimation of a discrete mixing measure μ^0 involved in a kernel mixture model. Using some recent advances in ℓ_1-regularization over the space of measures, we introduce a "data fitting + regularization" convex program for estimating μ^0 in a grid-less manner, this method is referred to as Beurling-LASSO.

Our contribution is two-fold: we derive a lower bound on the bandwidth of our data fitting term depending only on the support of μ^0 and its so-called "minimum separation" to ensure quantitative support localization error bounds; and under a so-called "non-degenerate source condition" we derive a non-asymptotic support stability property. This latter shows that for sufficiently large sample size n, our estimator has exactly as many weighted Dirac masses as the target μ^0, converging in amplitude and localization towards the true ones.

The statistical performances of this estimator are investigated designing a so-called "dual certificate", which will be appropriate to our setting. Some classical situations, as *e.g.* Gaussian or ordinary smooth mixtures (*e.g.* Laplace distributions), are discussed at the end of the paper. We stress in particular that our method is completely adaptive w.r.t. the number of components involved in the mixture.

MSC 2010 subject classifications: Primary: 62G05, 90C25; Secondary: 49M29.
Keywords and phrases: Beurling Lasso; Mixture recovery; Dual certificate; Kernel approach; Super-resolution.

1. Introduction

1.1. Mixture problems

In this paper, we are interested in the estimation of a mixture distribution μ^0 using some i.i.d. observations $\mathbf{X} := (X_1, \dots, X_n) \in (\mathbb{R}^d)^n$ with the help of some ℓ_1-regularization methods. More precisely, we consider the specific situation of a discrete distribution μ^0 that is given by a finite sum of K components:

$$\mu^0 := \sum_{k=1}^{K} a_k^0 \delta_{t_k} \tag{1}$$

where the set of positive weights $(a_k^0)_{1\leq k\leq K}$ defines a discrete probability distribution, *i.e.* each δ_{t_k} is a Dirac mass at point $t_k \in \mathbb{R}^d$ while

$$\sum_{k=1}^{K} a_k^0 = 1 \qquad \text{and} \qquad \forall k \in [K] := \{1, \ldots, K\} : \quad a_k^0 > 0.$$

We denote by $S^0 := \{t_1, \ldots, t_K\}$ the support of the target distribution μ^0. This distribution is indirectly observed: we assume that our set of observations $\mathbf{X}$ in $\mathbb{R}^d$ satisfies

$$X_i \overset{\text{iid}}{\sim} \sum_{k=1}^{K} a_k^0 F_{t_k} , \quad \forall i \in [n] := \{1, \ldots, n\} ,$$

where $(F_t)_{t \in \mathbb{R}^d}$ is a family of *known* distributions on $\mathbb{R}^d$. Below, we consider the so-called location model where each distribution F_t has a density with respect to the Lebesgue measure on $\mathbb{R}^d$ given by the density function $\varphi(\cdot - t)$, where φ denotes a *known* density function. In this case, the density function f^0 of the data $\mathbf{X}$ can be written as a convolution, namely

$$f^0(x) = \sum_{k=1}^{K} a_k^0 \varphi(x - t_k) , \quad \forall x \in \mathbb{R}^d . \tag{2}$$

Remark 1. *Equation (2) has a simple interpretation in the context considered here: the law of one observation X_i is given by a sum of two independent random variables U^0 and ε:*

$$X_i \sim U^0 + \varepsilon ,$$

where $U^0 \in S^0$ is distributed according to μ^0 (i.e., the mixing law (1)) and ε is distributed with a distribution of density φ with respect to the Lebesgue measure on $\mathbb{R}^d$. In this context, recovering the distribution of U^0 from the sample $\mathbf{X}$ appears to be an inverse (deconvolution) problem. The main difference with former contributions (see, e.g. [23] for a comprehensive introduction) is that the probability measure associated to U^0 is discrete, which avoids classical regularization approaches.

Equation (2) is known in the literature as a *mixture model*. A mixture model allows to describe some practical situations where a population of interest is composed of K different sub-populations, each of them being associated to a proportion a_k^0 and to a location parameter t_k. Mixture models have been intensively investigated during the last decades and have been involved in several fields as biology, genetics, astronomy, among others. We refer to [22, 17] for a complete overview.

1.2. Previous works

The main goal of this paper is to provide an estimation of the discrete mixture law μ^0 introduced in (1). When the component number K is available, the maximum likelihood estimator (MLE) appears to be the most natural candidate. Although no analytic expression is available for the model (2), it can be numerically approximated. We mention for instance the well-known EM-algorithm and refer to [33], who established some of the most general convergence results known for the EM algorithm. However, the MLE (and the related EM-algorithm) does not always provide satisfactory results. First, the MLE suffers from several drawbacks (see, *e.g.*, [20]) such as non-uniqueness of the solution, and second, obtaining theoretical guarantees for the EM-algorithm is still a difficult question (see, *e.g.*, the recent contributions [2, 14]). Several alternative methods have been proposed in this context. Several contributions extensively use the MLE point of view to derive consistent properties in general semi-parametric models, including the Gaussian case (see *i.e.* [31]), whereas some other ones developed some contrast functions in a semi-parametric framework: with symmetry and number of component assumptions in [7, 5], or with a fixed number of component settings in [18] and a L^2 contrast. As a particular case, the Gaussian setting has attracted a lot of attention: a model selection strategy is developed in [21] and a specific analysis of the EM algorithm with two Gaussian components is provided in [34]. [2] provide a general theoretical framework to analyze the convergence of the EM updates in a neighborhood of the MLE, and derive some non-asymptotic bounds on the Euclidean error of sample-based EM iterates. Some of the aforementioned papers provide better results (for instance with parametric rates of convergence for the estimation of the weights a_k^0, see *e.g.* [24, 19]), but are obtained in more constrained settings: known fixed number of components (often $K = 2$), univariate case, ...

In super-resolution and "off-the-grid" methods [4, 8], recent works have addressed mixture models while assuming that the sampling law is known. For example, the authors of [25] study some dimension reduction techniques such as random "sketching" problems using "off-the-grid" minimization scheme. They prove convergence of random feature kernel towards the population kernel. We emphasize that the statistical estimation in terms of the sample size n has not been considered in the super-resolution research field. To the best of our knowledge, this paper is the first that bridges the gap between the recent "off-the-grid" sparse regularization methods and a sharp statistical study of this estimation procedure in terms of the sample size and the bandwidth of the data fitting term.

1.3. Contribution

In this paper, we propose an estimator $\hat{\mu}_n$ of the measure μ^0 (see Equation (1)) inspired by some recent results in ℓ_1-regularization on the space of measures, sometimes referred to as super-resolution methods (see, *e.g.*, [10, 8]). We investigate the statistical theoretical performances of $\hat{\mu}_n$. This estimator $\hat{\mu}_n$ is built according to the minimization of a criterion on the space of real measures on $\mathbb{R}^d$ and does not require any grid for its computation. This criterion requires to tune two parameters: a bandwidth parameter of the data fitting term denoted by $m \geq 1$ and an ℓ_1-regularization tuning parameter denoted by $\kappa > 0$ below. We prove that the bandwidth parameter m depends only on the intrinsic hardness of estimating the support S^0 of the target μ^0 through the so-called "minimum separation" Δ introduced in [8] that refers to the minimal distance between two spikes:

$$\Delta := \min_{k \neq \ell} \|t_k - t_\ell\|_2 \,.$$

We now assess briefly the performances of $\hat{\mu}_n$. We emphasize that a complete version is displayed in Theorem 10 (for points i) and ii)) and Theorem 11 (for point iii)) later on).

Theorem 1. *For any bandwidth m such that*

$$m \gtrsim K^{1/2} d^{7/3} \Delta^{-2} \,, \tag{3}$$

some quantities $\mathcal{C}_m(\varphi) > 0$ and ρ_n exist such that

$$\rho_n = \mathcal{O}\Big(\sqrt{\frac{m^d}{n}}\Big), \tag{4}$$

if the kernel φ satisfies ($\mathcal{H}_\eta$) with $\eta = 4m$ (see Section 2.3 for a definition) then $\hat{\mu}_n$ satisfies:

i) Spike detection property:

$$\forall A \subset \mathbb{R}^d, \quad \mathbb{E}[\hat{\mu}_n(A)] \gtrsim \rho_n \mathcal{C}_m(\varphi) \quad \Longrightarrow \quad \min_{k\in[K]} \inf_{t\in A} \|t - t_k\|_2^2 \lesssim \frac{1}{m^2}.$$

ii) Weight reconstruction property:

$$\forall k \in [K] : \qquad \mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right] \lesssim \rho_n \mathcal{C}_m(\varphi),$$

where $\mathbb{N}_k(\epsilon)$ denotes a region that contains t_k and $\epsilon = \epsilon_{n,m}(d)$ is explicited later on.

iii) Support stability property: if φ satisfies the Non-Degenerated Bandwidth *condition* (NDB) *(see Section 4.4 for a definition), for n large enough, with probability $1 - p_n$ where $\lim_{n\to+\infty} p_n = 0$, $\hat{\mu}_n$ can be written as*

$$\hat{\mu}_n = \sum_{k=1}^{\hat{K}} \hat{a}_k \delta_{\hat{t}_k} ,$$

with $\hat{K} = K$. Furthermore, $(\hat{a}_k, \hat{t}_k) \to (a_k^0, t_k)$ for all $k \in [K]$, as n grows.

These three results deserve several comments. The first result *i)* translates that when a set A has enough mass w.r.t. the estimated measure $\hat{\mu}_n$, it corresponds to a true spike with an accuracy of the order m^{-2}. The second result *ii)* provides some statistical guarantees on the mass set by $\hat{\mu}_n$ near a true spike t_k that converges to $\mu^0(\{t_k\}) = a_k^0$. Condition (NDB) is inspired from the so-called "non-degenerated source condition" (NDSC) introduced in [13] and allows to derive the support stability. The last result *iii)* shows that, for large enough sample size, ℓ_1-regularization successfully discovers the number of mixing components and the estimated weights on the Dirac masses converge towards the true ones in amplitudes and localization.

The bandwidth m has to be adjusted to avoid over and under-fitting. Condition (3) ensures that the target point is admissible for our convex program and it may be seen as a condition to avoid a large bias term and under-fitting. Condition (4) ensures that the sample size is sufficiently large with respect to the model size m and it might be seen as a condition to avoid over-fitting and therefore to upper-bound the variance of estimation.

Below, we will pay attention to the role of Fourier analysis of φ and to the dimension d of the ambient space. These results are applied to specific setting (Gaussian and Laplace mixtures).

1.4. Outline

This paper is organized as follows. Section 2 introduces some standard ingredients of ℓ_1 regularization methods and gives a deterministic analysis of the exact recovery property of μ^0 from f^0. Section 3 provides a description of the statistical estimator $\hat{\mu}_n$ derived from a deconvolution with a Beurling-LASSO strategy (BLASSO) (see *e.g.* [10]). Section 4 focuses on the statistical performances of our estimator whereas Section 5 details the rates of convergence for specific mixture models. The main proofs are gathered in Section 6 whereas the most technical ones are deferred to Appendices A and B.

2. Assumptions, notations and first results

This section gathers the main assumptions on the mixture model (2). Preliminary theoretical results in an "ideal" setting are stated in order to ease the understanding of the forthcoming paragraphs.

2.1. Functional framework

We introduce some notations that will be used all along the paper.

Definition 1 (Set $(\mathcal{M}(\mathbb{R}^d,\mathbb{R}),\|\cdot\|_1)$). *We denote by* $(\mathcal{M}(\mathbb{R}^d,\mathbb{R}),\|\cdot\|_1)$ *the space of* real valued *measures on* $\mathbb{R}^d$ *equipped with the total variation norm* $\|\cdot\|_1$*, which is defined as*

$$\|\mu\|_1 := \int_{\mathbb{R}^d} \mathrm{d}|\mu| \quad \forall \mu \in \mathcal{M}(\mathbb{R}^d,\mathbb{R})\,,$$

where $|\mu| = \mu^+ + \mu^-$ *and* $\mu = \mu^+ - \mu^-$ *is the Jordan decomposition associated to a given* $\mu \in \mathcal{M}(\mathbb{R}^d,\mathbb{R})$*.*

A standard argument proves that the total variation of μ is also described with the help of a variational relationship:

$$\|\mu\|_1 = \sup\left\{\int_{\mathbb{R}^d} f\mathrm{d}\mu \;: f \,\text{is}\, \mu-\text{measurable and }\; |f|\leq 1\right\}.$$

Recall that φ used in Equation (2) is a probability density function so that $\varphi \in L^1(\mathbb{R}^d)$.

Definition 2 (Fourier transform over $L^1(\mathbb{R}^d)$ and $\mathcal{M}(\mathbb{R}^d,\mathbb{R})$). *We denote by* $\mathcal{F}$ *the Fourier transform defined by:*

$$\forall x \in \mathbb{R}^d, \forall f \in L^1(\mathbb{R}^d), \quad \mathcal{F}[f](x) := \int_{\mathbb{R}^d} e^{-\imath x^\top \omega} f(\omega)\mathrm{d}\omega\,.$$

A standard approximation argument extends the Fourier transform to $\mathcal{M}(\mathbb{R}^d,\mathbb{R})$ *with:*

$$\forall x \in \mathbb{R}^d, \forall \mu \in \mathcal{M}(\mathbb{R}^d,\mathbb{R}), \quad \mathcal{F}[\mu](x) := \int_{\mathbb{R}^d} e^{-\imath x^\top \omega}\mathrm{d}\mu(\omega)\,.$$

We shall also introduce the convolution operator Φ as

$$\mu \mapsto \Phi(\mu) := \varphi \star \mu = \int_{\mathbb{R}^d} \varphi(\cdot - x)\mathrm{d}\mu(x)\,, \quad \mu \in \mathcal{M}(\mathbb{R}^d,\mathbb{R})\,, \tag{5}$$

and it holds equivalently that (see *e.g.* [27, Section 9.14]):

$$\forall \mu \in \mathcal{M}(\mathbb{R}^d,\mathbb{R})\,, \quad \mathcal{F}[\Phi(\mu)] = \mathcal{F}[\varphi]\mathcal{F}[\mu]\,. \tag{6}$$

Concerning the density φ involved in (2), we will do the following assumption.

The function φ is a *bounded continuous symmetric function of positive definite type.* $(\mathcal{H}_0)$

In particular, the positive definite type property involved in Assumption ($\mathcal{H}_0$) is equivalent to require that for any finite set of points $\{x_1, \ldots, x_n\} \in \mathbb{R}^d$ and for any $(z_1, \ldots, z_n) \in \mathbb{C}^n$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \varphi(x_i - x_j) z_i \bar{z}_j \geq 0.$$

In what follows, we consider $h : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ the function defined by $h(x, y) = \varphi(x - y)$ for all $x, y \in \mathbb{R}^d$. In such a case, Assumption ($\mathcal{H}_0$) entails that $h(\cdot, \cdot)$ is a bounded continuous symmetric positive definite kernel. By Bochner's theorem (see, *e.g.*, [27, Theorem 11.32]), φ is the inverse Fourier transform of a nonnegative measure Σ referred to as the *spectral measure*. The Fourier inversion theorem states that Σ has a nonnegative density $\sigma \geq 0$ with respect to the Lebesgue measure on $\mathbb{R}^d$ such that $\sigma \in L^1(\mathbb{R}^d)$. Hence, it holds from the preceding discussion that

$$\varphi = \mathcal{F}^{-1}[\sigma] \text{ for some nonnegative } \sigma \in L^1(\mathbb{R}^d)\,. \tag{7}$$

Below, the set of points where the Fourier transform of a function does not vanish will play an important role. We will denote this support by $\mathrm{Supp}(\sigma)$:

$$\mathrm{Supp}(\sigma) = \Big\{ \omega \in \mathbb{R}^d \,:\, \sigma(\omega) \neq 0 \Big\}\,.$$

Some examples of densities φ that satisfies ($\mathcal{H}_0$) will be given and discussed in the forthcoming sections. We emphasize that this assumption is not restrictive and concerns for instance Gaussian, Laplace or Cauchy distributions, this list being not exhaustive.

Additional notation. Given two real sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$) if there exists a constant $C > 0$ independent of n such that $a_n \leq b_n$ (resp. $a_n \geq b_n$) for all $n \in \mathbb{N}$. Similarly, we write $a_n \ll b_n$ if $a_n/b_n \to 0$ as $n \to +\infty$. The set $\mathbb{N}^*$ stands for $\mathbb{N} \setminus \{0\}$.

2.2. Exact Recovery of μ^0 from f^0 - Case $\mathrm{Supp}(\sigma) = \mathbb{R}^d$

In this paragraph, we are interested in an "ideal" problem where we are looking for μ^0 not from a sample $X_1, \ldots, X_n$ distributed according to Equation (2), but from the population law f^0 itself. Of course, this situation does not occur in practice since in concrete situations, we do not observe f^0 but an empirical version of it and we will have to preliminary use an estimation of f^0 before solving the deconvolution inverse problem. Nevertheless, this toy problem already provides the first ingredients for a better understanding of the difficulties that arise in the context we consider.

We stress that $f^0 := \Phi(\mu^0)$ where Φ is defined by (5). Hence, this paragraph concerns the recovery of μ^0 from its convolution by the kernel φ. We thus face an inverse (deconvolution) problem. Several solutions could be provided and a standard method would rely on Fourier inversion

$$\mu^0 = \mathcal{F}^{-1}\left[\mathcal{F}(f^0)\sigma^{-1}\right].$$

Here, we prove in a first step that this deconvolution problem can be efficiently solved using a ℓ_1-regularization approach. We will be interested in the convex program (8) given by:

$$\min_{\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}) \,:\, \Phi(\mu) = f^0} \|\mu\|_1\,. \tag{8}$$

In particular, we investigate under which conditions the solution set of (8) is the singleton $\{\mu^0\}$, that we referred to as the "Perfect Recovery" property. We introduce the set of admissible points to the program (8), denoted by $\mathcal{M}(f^0)$ and defined as:

$$\mathcal{M}(f^0) := \{\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}) : \Phi(\mu) = f^0\}.$$

In this context, some different assumptions on the kernel φ shall be used in our forthcoming results.

A first reasonable situation is when the spectral density $\sigma = \mathcal{F}(\varphi)$ has its support equal to $\mathbb{R}^d$ and in this case we denote $\sigma > 0$. This requirement can be summarized in the next assumption on the function φ:

$$\varphi = \mathcal{F}^{-1}[\sigma],\, \sigma(\omega) = \sigma(-\omega) \text{ a.e. with } \mathrm{Supp}(\sigma) = \mathbb{R}^d : \forall \omega \in \mathbb{R}^d \quad \sigma(\omega) > 0. \tag{$\mathcal{H}_\infty$}$$

Example 1. *It may be shown that the set of densities φ that satisfy both Assumptions ($\mathcal{H}_0$) and ($\mathcal{H}_\infty$) include the Gaussian, Laplace, $B_{2\ell+1}$-spline, inverse multi-quadrics, Matérn class (see, e.g., [29, top of page 2397]) examples.*

Under Assumptions ($\mathcal{H}_0$) and ($\mathcal{H}_\infty$), any target measure $\mu^0 \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$ is the only admissible point of the program (8).

Theorem 2 (Perfect Recovery under ($\mathcal{H}_0$) and ($\mathcal{H}_\infty$))**.** *Assume that the convolution kernel satisfies ($\mathcal{H}_0$) and ($\mathcal{H}_\infty$), then for any target μ^0 the program (8) has μ^0 as unique solution point:*

$$\mathcal{M}(f^0) = \{\mu^0\}.$$

We emphasize that the previous result also holds for measures μ^0 that are not necessarily discrete. The proof is postponed to Section 6.1.

2.3. The Super-resolution phenomenon

Theorem 2 entails that the measure μ^0 can be recovered as soon as the spectrum of f^0 is observed and as soon as its support is $\mathbb{R}^d$. Surprisingly, this latter assumption can be relaxed and reconstruction can be obtained in some specific situations. Such a phenomenon is associated to the super-resolution theory and has been popularized by [8] among others.

Of course, this reconstruction is feasible at the expense of an assumption on the Fourier transform of φ. For the sake of simplicity, we assume that the spectral density σ has a support that contains the hypercube $[-\eta, \eta]^d$ for some frequency threshold $\eta > 0$. This can be summarized in the following assumption on the function φ:

$$\varphi = \mathcal{F}^{-1}[\sigma],\, \sigma(\omega) = \sigma(-\omega) \text{ a.e. with } [-\eta, \eta]^d \subset \mathrm{Supp}(\sigma). \tag{$\mathcal{H}_\eta$}$$

Remark 2. *The densities φ that satisfy ($\mathcal{H}_\eta$) and for which* $\mathrm{Supp}(\sigma) = [-\eta, \eta]^d$ *act as "low pass filters". The convolution operator Φ described in (5) cancels all frequencies above η, see for instance (6). Of course, the larger η, the easier the inverse deconvolution problem.*

Under ($\mathcal{H}_0$) and ($\mathcal{H}_\eta$), the target measure $\mu^0 \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$ is not the only admissible point in $\mathcal{M}(f^0)$ to the program (8). We will need to ensure the existence of a specific function, called in what follows a *dual certificate*, that will entail that μ^0 is still the only solution of the program (8), referred to as "Perfect Recovery".

Theorem 3 (Dual Certificate for (8))**.** *Assume that the density φ satisfies* ($\mathcal{H}_0$) *and* ($\mathcal{H}_\eta$) *for some real $\eta > 0$. Assume that μ^0 and $S^0 = \{t_1, \ldots, t_K\}$ are given by Equation* (1) *and that a function $\mathcal{P}_\eta$ exists such that it satisfies the interpolation conditions:*

- $\forall t \in \{t_1, \ldots, t_K\} : \mathcal{P}_\eta(t) = 1$ *and* $\forall t \notin \{t_1, \ldots, t_K\} : |\mathcal{P}_\eta(t)| < 1$,

and the smoothness conditions:

- $\mathcal{P}_\eta \in \mathcal{C}_0(\mathbb{R}^d, \mathbb{R}) \cap L^1(\mathbb{R}^d)$,
- *the support of the Fourier transform $\mathcal{F}[\mathcal{P}_\eta]$ satisfies* $\mathrm{Supp}\,(\mathcal{F}[\mathcal{P}_\eta]) \subset [-\eta, \eta]^d$.

Then the program (8) *has μ^0 as unique solution point* (*Perfect Recovery*)*.*

The proof is postponed to Section 6.2. We will propose a construction of such a certificate $\mathcal{P}_\eta$ in Appendix B with some additional constraints. In particular, it will make it possible to address the more realistic statistical problem where only an empirical measure of the data is available.

Remark 3. *The previous theorem can be extended to the case where the convolution kernel is bounded, continuous and symmetric positive definite. The proof is the same substituting $[-\eta, \eta]^d$ by the support Ω of its spectral density. Remark that since σ is nonzero, necessarily Ω has a nonempty interior.*

3. Off-The-Grid estimation via the Beurling-LASSO (BLASSO)

In this section, we consider the statistical situation where the density f^0 is not available and we deal instead with a sample $\mathbf{X} = (X_1, \ldots, X_n)$ of i.i.d. observations distributed with the density f^0. In this context, only the empirical measure

$$\hat{f}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \tag{9}$$

is available, and our aim is to recover μ^0 from $\hat{f}_n$. To this end, we use in this paper a *BLASSO* procedure (see *e.g.* [1]). Namely we deal with the following estimator $\hat{\mu}_n$ of the unknown discrete measure μ^0 defined as:

$$\hat{\mu}_n := \arg \min_{\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})} \left\{ C(\Phi\mu, \hat{f}_n) + \kappa \|\mu\|_1 \right\}, \tag{10}$$

where κ is a regularization parameter whose value will be made precise later on, and $C(\Phi\mu, \hat{f}_n)$ is a *data fidelity* term that depends on the sample $\mathbf{X}$. The purpose of the data fidelity term is to measure the *distance* between the target μ^0 and any candidate $\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$.

Some examples of possible cost $C : \mathbb{H} \times \mathcal{M}(\mathbb{R}^d, \mathbb{R}) \to \mathbb{R}$ are discussed in Section 3.1. Our goal is then to derive some theoretical results associated to this estimation procedure.

3.1. Kernel approach

3.1.1. RKHS functional structure

In order to design the data fidelity term, we need to define a space where we can compare the observations $\mathbf{X} = (X_1, \ldots, X_n)$ and any model $f = \varphi \star \mu = \Phi\mu$ for $\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$. In this work, we focus our attention on a kernel approach.

Reminders on RKHS The difficulty lies in the fact that the empirical law $\hat{f}_n$ introduced in (9) does not belong to $\mathcal{C}_0(\mathbb{R}^d, \mathbb{R})$. To compare the prediction $\Phi\mu$ with $\hat{f}_n$, we need to embed these quantities in the same space. We consider here a Reproducing Kernel Hilbert Space (RKHS) structure, which provides a lot of interesting properties and has been at the core of several investigations and applications in approximation theory [32], as well as in the statistical and machine learning communities, (see [29] and the references therein). We briefly recall the definition of such a space.

Definition 3. *Let $(\mathbb{L}, \|.\|_{\mathbb{L}})$ be a Hilbert space containing function from $\mathbb{R}^d$ to $\mathbb{R}$. The space $\mathbb{L}$ is said to be a RKHS if $\delta_x : f \mapsto f(x)$ are continuous for all $x \in \mathbb{R}^d$ from $(\mathbb{L}, \|.\|_{\mathbb{L}})$ to $(\mathbb{R}, |.|)$.*

The Riesz theorem leads to the existence of a function ℓ that satisfies the *representation* property:

$$\langle f, \ell(x, .)\rangle_{\mathbb{L}} = f(x) \quad \forall f \in \mathbb{L}, \quad \forall x \in \mathbb{R}^d. \tag{11}$$

The function ℓ is called the *reproducing kernel* associated to $\mathbb{L}$. Below, we consider a kernel ℓ such that $\ell(x, y) = \lambda(x - y)$ for all $x, y \in \mathbb{R}^d$ where λ satisfies $(\mathcal{H}_0)$. Again, the Bochner theorem yields the existence of a *nonnegative* measure $\Lambda \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$ such that λ is its inverse Fourier transform

$$\lambda = \mathcal{F}^{-1}(\Lambda), \quad \text{namely} \quad \forall x \in \mathbb{R}^d, \quad \lambda(x) = \int_{\mathbb{R}^d} e^{\imath x^\top \omega} \mathrm{d}\Lambda(\omega).$$

Moreover, since λ is continuous, Λ is then a bounded measure and the Mercer theorem (see *e.g.* [3]) proves that the RKHS $\mathbb{L}$ is exactly characterized by

$$\mathbb{L} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } \|f\|^2_{\mathbb{L}} = \int_{\mathbb{R}^d} \frac{|\mathcal{F}[f](t)|^2}{\mathcal{F}[\lambda](t)} \mathrm{d}t < +\infty \right\}. \tag{12}$$

Convolution in the RKHS The RKHS structure associated to the kernel λ entails a comparison between the empirical measure and any candidate $\Phi\mu$. Indeed, a convolution operator L similar to the one defined in Equation (5) can be associated to the RKHS as pointed out by the next result.

Proposition 4. *For any $\nu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$, the convolution $L\nu = \lambda \star \nu$ belongs to $\mathbb{L}$.*

The proof of Proposition 4 is deferred to Appendix A.1.

3.1.2. Data fidelity term

For any $\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$, both $L\hat{f}_n$ and $L \circ \Phi\mu$ belong to $\mathbb{L}$. Hence, one may use the following data fidelity term

$$\mathrm{C}_\lambda(\Phi\mu, \hat{f}_n) := \|L\hat{f}_n - L \circ \Phi\mu\|^2_{\mathbb{L}}, \quad \forall \mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}). \tag{13}$$

Example 2. *An important example is given by the sinus-cardinal kernel* sinc*. Given a frequency "cut-off" $1/\tau > 0$, one can consider the kernel*

$$\lambda_\tau(x) := \frac{1}{\tau^d} \lambda_{\text{sinc}}\left(\frac{x}{\tau}\right) \quad \textit{where} \quad \lambda_{\text{sinc}}(x) := \prod_{j=1}^d \frac{\sin(\pi x_j)}{\pi x_j} \quad \forall x \in \mathbb{R}^d.$$

Then, the spectral measure is given by

$$\mathrm{d}\Lambda_\tau(\omega) = \mathrm{d}\Lambda_{\mathrm{sinc}}(\omega\pi\tau) := \frac{1}{2^d}\prod_{j=1}^d \mathbb{1}_{[-1/\tau,1/\tau]}(\omega_j)\mathrm{d}\omega \quad \forall \omega \in \mathbb{R}^d.$$

In this particular case, we deduce that the convolution L is a low-pass filter with a frequency cut-off $1/\tau$ and the RKHS (denoted by $\mathbb{L}_\tau$) is given by:

$$\mathbb{L}_\tau = \left\{ f \text{ s.t. } \|f\|^2_{\mathbb{L}_\tau} = \frac{1}{2^d}\int_{B_\infty(1/\tau)} |\mathcal{F}[f]|^2 < +\infty \text{ and } \mathrm{Supp}(\mathcal{F}[f]) \subseteq B_\infty(1/\tau) \right\}, \qquad (14)$$

where $B_\infty(1/\tau)$ denotes the centered ℓ_∞ ball of radius $1/\tau$. The RKHS $\mathbb{L}_\tau$ then corresponds to the band-limited functions in $L^2(\mathbb{R}^d)$ whose Fourier transform vanishes for a frequency larger than $1/\tau$. In this context, our criterion becomes

$$C_{\lambda_\tau}(\Phi\mu, \hat{f}_n) = \frac{1}{2^d}\int_{[-1/\tau,1/\tau]^d} |\mathcal{F}[\Phi\mu - \hat{f}_n](\omega)|^2 \mathrm{d}\omega = \frac{1}{2^d}\int_{[-1/\tau,1/\tau]^d} |\sigma\mathcal{F}[\mu] - \mathcal{F}[\hat{f}_n](\omega)|^2 \mathrm{d}\omega\,,$$

and it may be checked that

$$C_{\lambda_\tau}(\Phi\mu, \hat{f}_n) = \frac{1}{2^d}\int_{\mathbb{R}^d} \Big|\lambda_\tau \star (\Phi\mu - \hat{f}_n)(x)\Big|^2 \mathrm{d}x.$$

This loss focuses on the L^2-error of $\Phi\mu - \hat{f}_n$ for frequencies in the Fourier domain $[-1/\tau, 1/\tau]^d$. In some sense, the kernel estimator $\lambda_\tau \star \hat{f}_n$ has a bandwidth τ that will prevent from over-fitting.

We stress that, as it is the case in the previous low-pass filter example, $C_{\lambda_\tau}(\Phi\mu, \hat{f}_n)$ may depend on a tuning parameter (the bandwidth τ in Example 2). For the ease of presentation, this parameter is not taken into account in the notation. However, its value will be discussed in Section 5.

3.1.3. Data-dependent computation

The next proposition entails that the criterion C_λ introduced in Equation (13) can be used in practice: it only depends on the candidate $\Phi\mu$, on the empirical sample $\mathbf{X}$ and on the kernel λ.

Proposition 5. *For all $\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$, we have:*

$$\begin{aligned} C_\lambda(\Phi\mu, \hat{f}_n) &= \|L\hat{f}_n - L \circ \Phi\mu\|^2_{\mathbb{L}} \\ &= \|L\hat{f}_n\|^2_{\mathbb{L}} + \int_{\mathbb{R}^d} \big[-\frac{2}{n}\sum_{i=1}^n \lambda(t - X_i)\big](\Phi\mu)(t)\mathrm{d}t + \int_{\mathbb{R}^d\times\mathbb{R}^d} \lambda(x-y)(\Phi\mu)(x)(\Phi\mu)(y)\mathrm{d}x\mathrm{d}y\,. \end{aligned}$$

We stress that $\|L\hat{f}_n\|^2_{\mathbb{L}}$ does no depend on μ and can be removed from the criterion when it is used in the optimization program (8). The proof of Proposition 5 is deferred to Appendix A.2.

3.2. Estimation by convex programming

Our estimator is defined as a (the) solution of the following optimization program with the data-fidelity term $C_\lambda(\Phi\mu, \hat{f}_n)$ introduced in (13). Hence, we consider the optimization problem:

$$\inf\left\{\frac{1}{2}\|L\hat{f}_n - L\circ\Phi\mu\|_{\mathbb{L}}^2 + \kappa\|\mu\|_1 \ :\ \mu\in\mathcal{M}(\mathbb{R}^d,\mathbb{R}) \text{ such that } \int_{\mathbb{R}^d}\mathrm{d}\mu = 1\right\}, \qquad (\mathbf{P}_\kappa)$$

where $\|\cdot\|_{\mathbb{L}}$ is the norm associated to the RKHS generated by λ (see Section 3.1) and κ is a tuning parameter whose value will be made precise later on. We emphasize that $(\mathbf{P}_\kappa)$ is a convex programming optimization problem (convex function to be minimized on a convex constrained set). The estimator $\hat{\mu}_n$ is then any solution of

$$\hat{\mu}_n \in \arg\min_{\mu\in\mathcal{M}(\mathbb{R}^d,\mathbb{R})\,:\,\int_{\mathbb{R}^d}\mathrm{d}\mu=1}\left\{\frac{1}{2}\|L\hat{f}_n - L\circ\Phi\mu\|_{\mathbb{L}}^2 + \kappa\|\mu\|_1\right\}. \qquad (15)$$

Remark 4. *Our theoretical analysis shows that we can obtain the same statistical guarantees when dropping the constraint $\int_{\mathbb{R}^d}\mathrm{d}\mu = 1$, except for the support stability with a large sample size, see Section 4.4. In this latter case, we are able to derive the same results when we jointly associate the optimization to the constraint $\int_{\mathbb{R}^d}\mathrm{d}\mu = 1$. We refer to Remark 9 for further details.*

Algorithmic issues related to the computation of (15) are out of the scope of this paper and will not be discussed here. Determining an algorithm for computing an approximation of $\hat{\mu}_n$ will be at the core of forthcoming contribution. However, we point out that this "off-the-grid" methodology has already been intensively discussed in the literature (we mention, among others, [4, 30, 13, 1, 11]) although the considered framework of our paper is different. In particular, we refer to greedy methods that provide heuristically good performances such as "*Sliding Frank Wolfe*" [12] (also known as conditional gradient) or "*Continuous Orthogonal Matching Pursuit*" [15]. These methods can be deployed to approximately solve (15).

Super-resolution is the ability to recover a discrete measure on the torus from some Fourier coefficients (recall that the Pontryagin's dual of the torus is $\mathbb{Z}^d$) while we want to recover a discrete measure on $\mathbb{R}^d$ from some Fourier transform over $\mathbb{R}^d$ (recall that the Pontryagin's dual of $\mathbb{R}^d$ is $\mathbb{R}^d$). In particular the dual of $(\mathbf{P}_\kappa)$ does not involve a set of fixed degree trigonometric polynomials as in super-resolution but inverse Fourier transform of some tempered distribution. Hence, new theoretical guarantees are necessary in order to properly define the estimator $\hat{\mu}_n$. This is the objective of the next theorem. In this view, we consider primal variables $\mu\in\mathcal{M}(\mathbb{R}^d,\mathbb{R})$ and $z\in\mathbb{L}$ and introduce the dual variables $c\in\mathbb{L}$ and $\rho\in\mathbb{R}$ as well as the following Lagrangian:

$$\mathcal{L}(\mu,z,c,\rho) := \frac{1}{2}\|L\hat{f}_n - z\|_{\mathbb{L}}^2 + \kappa\|\mu\|_1 - \langle c, L\circ\Phi\mu - z\rangle_{\mathbb{L}} + \rho\left(\int_{\mathbb{R}^d}\mathrm{d}\mu - 1\right). \qquad (16)$$

It is immediate to check that if $\int_{\mathbb{R}^d}\mathrm{d}\mu \neq 1$, then the supremum of $\mathcal{L}(\mu,z,c,\rho)$ over ρ is $+\infty$. Such fact also holds when $z \neq L\circ\Phi\mu$ while considering the supremum of $\mathcal{L}(\mu,z,c,\rho)$ when c varies. Therefore, the primal expression coincides with the supremum in the dual variables, namely

$$\inf_{\mu,z}\sup_{c,\rho}\mathcal{L}(\mu,z,c,\rho) = \inf_{\mu\in\mathcal{M}(\mathbb{R}^d,\mathbb{R})\,:\,\int_{\mathbb{R}^d}\mathrm{d}\mu=1}\left\{\frac{1}{2}\|L\hat{f}_n - L\circ\Phi\mu\|_{\mathbb{L}}^2 + \kappa\|\mu\|_1\right\} \iff (\mathbf{P}_\kappa).$$

In the meantime, the dual program of $(\mathbf{P}_\kappa)$ is given by

$$\sup_{(c,\rho)\in\mathbb{L}\times\mathbb{R}}\ \inf_{(\mu,z)\in\mathcal{M}(\mathbb{R}^d,\mathbb{R})\times\mathbb{L}} \mathcal{L}(\mu,z,c,\rho)\,. \tag{$\mathbf{P}^*_\kappa$}$$

Theorem 6 (Primal-dual formulation and strong duality)**.** *The following statements are true.*

i) The primal problem $(\mathbf{P}_\kappa)$ has at least one solution and it holds that $\hat z_n := L\circ\Phi\hat\mu_n$ and $\hat m_n := \|\hat\mu_n\|_1$ are uniquely defined (they do not depend on the choice of the solution $\hat\mu_n$).

ii) The dual program of $(\mathbf{P}_\kappa)$, given by $(\mathbf{P}^*_\kappa)$ satisfies

$$\frac{\|L\hat f_n\|_{\mathbb{L}}^2}{2} - \inf\Big\{\rho + \frac{1}{2}\|L\hat f_n - c\|_{\mathbb{L}}^2 \ :\ (c,\rho)\in\mathbb{L}\times\mathbb{R}\ s.t.\ \|\Phi c-\rho\|_\infty\le\kappa\Big\} \iff (\mathbf{P}^*_\kappa),$$

and there is no duality gap. A unique pair $(\hat c,\hat\rho)\in\mathbb{L}\times\mathbb{R}$ exists such that

$$(\hat c,\hat\rho) = \arg\min_{(c,\rho)\in\mathbb{L}\times\mathbb{R}:\ \|\Phi c-\rho\|_\infty\le\kappa} \Big\{\rho + \frac{1}{2}\|L\hat f_n - c\|_{\mathbb{L}}^2\Big\}\,. \tag{$\mathcal{D}_\kappa$}$$

iii) Any solution $\hat\mu_n$ to $(\mathbf{P}_\kappa)$ satisfies the support inclusion:

$$\mathrm{Supp}(\hat\mu_n) \subseteq \Big\{x\in\mathbb{R}^d \ :\ \big|\Phi\hat c-\hat\rho\,\big|(x)=\kappa\Big\}\,,$$

where $(\hat c,\hat\rho)$ is the unique solution to $(\mathcal{D}_\kappa)$. Moreover, for all solutions $\hat\mu_n$ to $(\mathbf{P}_\kappa)$:

$$\int_{\mathbb{R}^d}\Big(\frac{\Phi\hat c-\hat\rho}{\kappa}\Big)\,\mathrm{d}\hat\mu_n = \|\hat\mu_n\|_1\,,$$

namely $(\Phi\hat c-\hat\rho)/\kappa$ is a sub-gradient of the total variation norm at point $\hat\mu_n$.

iv) If $d=1$ and if at least one of the spectral measures Λ or σ has a bounded support, then $\{x\in\mathbb{R} \ :\ \big|\Phi\hat c-\hat\rho\,\big|(x)=\kappa\}$ is discrete with no accumulation point, any primal solution $\hat\mu_n$ have an (at most countable) discrete support $\hat S\subset\mathbb{R}$ with no accumulation point:

$$\hat\mu_n = \sum_{t\in\hat S}\hat a_t\delta_t \quad \textit{where} \quad \sum_{t\in\hat S}\hat a_t = 1. \tag{17}$$

The proof of this result can be found in Section 6.3.

It is generally numerically admitted, see for instance [8, Page 939], that the extrema of the dual polynomial $\Phi\hat c-\hat\rho$ are located in a discrete set, so that any solution to $(\mathbf{P}_\kappa)$ has a discrete support by using *iii*). However, this issue remains an open question. In practice, all solvers of $(\mathbf{P}_\kappa)$ lead to discrete solutions: greedy methods are discrete by construction, and L^1-regularization methods empirically lead to discrete solutions, see *e.g.* [8]. Furthermore, as presented in Theorem 11, our theoretical result shows that for large enough n, the support stability property holds. In this case, the solution of $(\mathbf{P}_\kappa)$ is discrete with $\hat K = K$ atoms.

Example 3. *Observe that the low-pass filter defined in Example 2 satisfies the requirements of iv) in Theorem 6 and we deduce that in dimension $d=1$, all solutions $\hat\mu_n$ are of the form (17).*

4. Statistical recovery of μ^0

This section provides some theoretical results for $\hat{\mu}_n$, built as the solution of ($\mathbf{P}_\kappa$). Contrary to ℓ_1-regularization in high-dimensions, standard RIP or REC compatibility conditions do not hold in our context, and all the cornerstone results of high-dimensional statistics cannot be used here. In our situation, the statistical analysis is performed using a "dual certificate" $\mathcal{P}_m$ as in super-resolution, see [1, 4, 8, 13] for instance. The construction and the key properties satisfied by $\mathcal{P}_m$ are detailed in Section 4.1. However, our framework is quite different from super-resolution and we had to address two issues: build a dual certificate on the space $\mathbb{R}^d$ and adapt its "frequency cut-off" (namely $4m$ in *iii*) of Theorem 7) to the sample size n and the tail of the kernel. This latter point is addressed in Section 5.

4.1. Strong dual certificate

Let $S^0 = \{t_1, \ldots, t_K\}$ be a fixed set of points in $\mathbb{R}^d$ and define $\Delta := \min_{k\neq\ell} \|t_k - t_\ell\|_2$. For any $m \in \mathbb{N}^*$, we consider the function $p_m^{\alpha,\beta}$ parameterized by a vector α and a matrix β of coefficents, defined as:

$$p_m^{\alpha,\beta}(t) = \sum_{k=1}^{K} \left\{\alpha_k \psi_m(t - t_k) + \langle \beta_k, \nabla \psi_m(t - t_k)\rangle\right\}, \quad \forall t \in \mathbb{R}^d, \tag{18}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)^T$, $\beta = (\beta_k^i)_{1\leq k\leq K, 1\leq i\leq d}$ with

$$\psi_m(.) = \psi^4(m.) \text{ with } \forall u = (u^1, \ldots, u^d) \in \mathbb{R}^d \quad \psi(u) = \prod_{j=1}^{d} \mathrm{sinc}(u^j) \text{ and } \mathrm{sinc}(x) = \frac{\sin(x)}{x}. \tag{19}$$

One important feature of ψ_m is its ability to interpolate 1 at the origin, while being positive and compactly supported in the Fourier domain. We then state the next result, which is of primary importance for the statistical accuracy of our procedure.

Theorem 7 (Strong dual certificate)**.** *Let be given a set of K points $S^0 = \{t_1, \ldots, t_K\}$ in $\mathbb{R}^d$ such that $\Delta := \min_{k\neq\ell} \|t_k - t_\ell\|_2$. Then, the following properties hold:*

- *i) A function $\mathcal{P}_m$ defined by $\mathcal{P}_m(t) = [p_m^{\alpha,\beta}(t)]^2$ exists with $m \gtrsim K^{1/2} d^{7/3} \Delta^{-2}$ such that*

$$\forall k \in [K],\ \mathcal{P}_m(t_k) = 1 \qquad \text{and} \qquad 0 \leq \mathcal{P}_m \leq 1$$

  and

$$\mathcal{P}_m(t) = 1 \Longleftrightarrow t \in S^0 = \{t_1, \ldots, t_K\}.$$

- *ii) A universal pair (υ, γ) independent from n, m and d exists such that for*

$$\epsilon = \frac{\upsilon}{m d^{3/2}}$$

  – *<u>Near region:</u> If we define*

$$\mathbb{N}(\epsilon) := \bigcup_{k=1}^{K} \mathbb{N}_k(\epsilon) \text{ where } \mathbb{N}_k(\epsilon) := \{t : \|t - t_k\|_2 \leq \epsilon\},$$

  a positive constant $\mathcal{C}$ exists such that:

$$\forall t \in \mathbb{N}_k(\epsilon) : \qquad 0 \leq \mathcal{P}_m(t) \leq 1 - \mathcal{C} m^2 \|t - t_k\|_2^2.$$

– <u>*Far region:*</u>

$$\forall t \in \mathbb{F}(\epsilon) := \mathbb{R}^d \setminus \mathbb{N}(\epsilon) : \qquad 0 \leq \mathcal{P}_m(t) \leq 1 - \gamma \frac{\upsilon}{d^4}.$$

- *iii) The support of the Fourier transform of $\mathcal{P}_m$ is growing linearly with m:*

$$\text{Supp}(\mathcal{F}[\mathcal{P}_m]) \subset [-4m, 4m]^d \qquad \text{and} \qquad \|\mathcal{P}_m\|_2 \lesssim K^2 m^{-d/2}.$$

- *iv) If $(\mathcal{H}_\eta)$ holds with $\eta = 4m$, then an element $c_{0,m} \in \mathbb{L}$ exists such that $\mathcal{P}_m = \Phi c_{0,m}$.*

The proof of this result is proposed in Appendix B. This construction is inspired by the one given in [8], which has been adapted to our specific setting. We emphasize that the size of the spectrum of $\mathcal{P}_m$ increases linearly with m, while the effect of the number of points K, the dimension d, and the spacing Δ is translated in the initial constraint $m \gtrsim K^{1/2} d^{7/3} \Delta^{-2}$.

We also state a complementary result, that will be useful for the proof of Theorem 10, *iii*).

Corollary 8. *Let be given a set of K points $S^0 = \{t_1, \ldots, t_K\}$ such that $\Delta := \min_{k \neq \ell} \|t_k - t_\ell\|_2$. Then, for any $k \in [K]$, a function $\mathcal{Q}_m^k$ exists such that*

$$\forall i \in [K] \qquad \mathcal{Q}_m^k(t_i) = \delta_i(k) \quad \text{and} \quad 0 \leq \mathcal{Q}_m^k \leq 1,$$

and a universal couple of constants (υ, γ) exists such that the function $\mathcal{Q}_m^k$ satisfies for $\epsilon = \frac{\upsilon}{md^{3/2}}$:

i) Near region $\mathbb{N}_k(\epsilon)$: a positive constant $\widetilde{\mathcal{C}}$ exists such that:

$$\forall t \in \mathbb{R}^d \qquad \|t - t_k\|_2 \leq \epsilon \Longrightarrow 0 \leq \mathcal{Q}_m^k(t) \leq 1 - \widetilde{\mathcal{C}} m^2 \|t - t_k\|_2^2,$$

ii) Near region $\mathbb{N}(\epsilon) \setminus \mathbb{N}_k(\epsilon)$:

$$\forall i \neq k \qquad \|t - t_i\|_2 \leq \epsilon \Longrightarrow |\mathcal{Q}_m^k(t)| \leq \widetilde{\mathcal{C}} m^2 \|t - t_i\|_2^2.$$

iii) Far region $\mathbb{F}(\epsilon)$:

$$\forall t \in \mathbb{F}(\epsilon), \ 0 \leq \mathcal{Q}_m^k(t) \leq 1 - \gamma \frac{\upsilon}{d^4}.$$

iv) A $c_{k,m} \in \mathbb{L}$ exists such that $\mathcal{Q}_m^k = \Phi c_{k,m}$ and

$$\|c_{k,m}\|_{\mathbb{L}} \lesssim \frac{K^2 m^{-d/2}}{\sqrt{\inf\limits_{\|t\|_\infty \leq 4m} \left\{\sigma^2(t) \mathcal{F}[\lambda](t)\right\}}}.$$

Proofs of $i), ii), iii)$ are similar to those of Theorem 7 and are omitted: the construction of $\mathcal{Q}_m^k$ obeys the same rules as the construction of $\mathcal{P}_m$ (the interpolation conditions only differ at points $t_i, i \neq k$ and are switched from 1 to 0). Regarding now $iv)$, the upper bound of the $\|.\|_{\mathbb{L}}$ norm uses the same arguments as the ones given below in the proof of $ii)$, Proposition 9.

4.2. Bregman divergence $D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0)$

Below, the statistical loss between $\hat{\mu}_n$ and μ^0 will be obtained in terms of the Bregman divergence associated to the dual certificate $\mathcal{P}_m$ obtained in Theorem 7. This divergence is defined by:

$$D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) := \|\hat{\mu}_n\|_1 - \|\mu^0\|_1 - \int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) \geq 0\,. \tag{20}$$

We also introduce the term Γ_n defined as

$$\Gamma_n = L\hat{f}_n - L \circ \Phi\mu^0,$$

which models the difference between the target $f^0 = \Phi\mu^0$ and its empirical counterpart $\hat{f}_n$ in the RKHS. The next proposition provides a control in expectation of the Bregman divergence between $\hat{\mu}_n$ and μ^0. Some similar results could be obtained for an upper bound with large probability.

Proposition 9. *Let $\mathcal{P}_m = \Phi c_{0,m}$ the dual certificate obtained in Theorem 7. Let $(\rho_n)_{n\in\mathbb{N}^*}$ be a sequence such that $\mathbb{E}[\|\Gamma_n\|_{\mathbb{L}}^2] \leq \rho_n^2$ for all $n \in \mathbb{N}^*$. If κ is chosen such that $\kappa = \rho_n/\|c_{0,m}\|_{\mathbb{L}}$ and if $\hat{\mu}_n$ is defined in ($\mathbf{P}_\kappa$), then:*

i) For any integer n:

$$\mathbb{E}\left[D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0)\right] \leq \frac{3\sqrt{2}}{2}\rho_n \|c_{0,m}\|_{\mathbb{L}}\,,$$

ii) $c_{0,m} \in \mathbb{L}$ satisfies

$$\|c_{0,m}\|_{\mathbb{L}} := \sqrt{\frac{\|\mathcal{P}_m\|_2^2}{\inf\limits_{\|t\|_\infty \leq 4m} \left\{\sigma^2(t)\mathcal{F}[\lambda](t)\right\}}} \lesssim \underbrace{\frac{K^2 m^{-d/2}}{\sqrt{\inf\limits_{\|t\|_\infty \leq 4m} \left\{\sigma^2(t)\mathcal{F}[\lambda](t)\right\}}}}_{:=\mathcal{C}_m(\varphi,\lambda)}\,. \tag{21}$$

The proof of Proposition 9 is postponed to Section 6.4. The previous results indicate that the Bregman divergence between our estimator $\hat{\mu}_n$ and the target measure μ^0 depends, up to some constants, on three main quantities:

- The variance of the empirical measure through the operator L quantified by ρ_n,
- The Fourier transform σ of the convolution kernel φ over the interval $[-4m; 4m]^d$. This term measures the ill-posedness of the inverse problem, which is associated to the difficulty to recover μ^0 with indirect observations (here $f^0 = \Phi\mu^0$ and we need to invert Φ),
- The structure of the RKHS used to smooth the problem identified through the kernel λ.

Remark 5. *We will derive from Proposition 9 some explicit convergence rates each time we will consider specific situation, i.e. as soon as the quantities involved in Equation (21) are made precise on some concrete examples. These rates will depend on the tuning parameter m for solving the optimization problem ($\mathbf{P}_\kappa$), and on the choice of the kernel λ. Some examples will be discussed in Section 5. Indeed, κ is related to m through the relationship $\kappa = \rho_n/\|c_{0,m}\|_{\mathbb{L}}$. Similarly, we will see in some specific situations (see Section 5) that the kernel λ is also linked to m in a transparent way. We stress that according to Proposition 9, $m \gtrsim K^{1/2}d^{7/3}\Delta^{-2}$. Such a condition will be satisfied provided m is allowed to go to infinity with n whereas K, Δ, d are kept fixed.*

Remark 6. *The upper bound proposed in Proposition 9 only uses items (iii) and (iv) of Theorem 7. An enhanced control on the performances of $\hat{\mu}_n$ is provided in the next section. Alternative features will be also proposed with a slightly different dual certificate.*

4.3. Statistical recovery of far and near regions

The next result sheds light on the performance of the BLASSO estimator introduced in Equation (10). The goodness-of-fit reconstruction of the mixture distribution μ^0 by $\hat{\mu}_n$ is translated by the statistical properties of the computed weights of $\hat{\mu}_n$ around the spikes of μ^0 (the support points $S^0 = \{t_1, \ldots, t_K\}$), which will define a family of K *near regions*, as well as the behaviour of $\hat{\mu}_n$ in the complementary set, the *far region*. The sets $\mathbb{F}(\epsilon)$ and $\mathbb{N}(\epsilon)$ have already been introduced in Theorem 7. Our result takes advantage on the previous bounds and on i) and ii) of Theorem 7.

Theorem 10. *Let $m \gtrsim K^{1/2} d^{7/3} \Delta^{-2}$ and let $\mathcal{P}_m$ be a dual certificate given in Theorem 7. Assume that $\hat{\mu}_n$ is the BLASSO estimator given by ($\mathbf{P}_\kappa$) with $\kappa = \kappa_n$ chosen in Proposition 9. Let $\mathcal{C}_m(\varphi, \lambda)$ be the quantity introduced in Proposition 9, $\hat{\mu}_n = \hat{\mu}_n^+ - \hat{\mu}_n^-$ the Jordan decomposition of $\hat{\mu}_n$. A universal couple of constants (γ, υ) exists such that, if*

$$\epsilon = \frac{\upsilon}{m d^{3/2}}, \tag{22}$$

i) Far region and negative part:

$$\mathbb{E}\left[\hat{\mu}_n^-(\mathbb{R}^d)\right] \leq \frac{3\sqrt{2}}{2} \rho_n \mathcal{C}_m(\varphi, \lambda) \ \text{and}\ \mathbb{E}\left[\hat{\mu}_n^+(\mathbb{F}(\epsilon))\right] \leq \frac{3\sqrt{2}}{2} \frac{d^4}{\gamma \upsilon} \rho_n \mathcal{C}_m(\varphi, \lambda).$$

ii) Near region (spike detection): a positive constant $\mathcal{C}$ exists such that

$$\forall A \subset \mathbb{R}^d, \quad \mathbb{E}[\hat{\mu}_n^+(A)] > \frac{3\sqrt{2}}{2} \frac{d^4}{\gamma \upsilon} \rho_n \mathcal{C}_m(\varphi, \lambda) \quad \Longrightarrow \quad \min_{k \in [K]} \inf_{t \in A} \|t - t_k\|_2^2 \leq \frac{\gamma \upsilon}{\mathcal{C} d^4 m^2}.$$

iii) Near region (weight reconstruction): for any $k \in [K]$:

$$\mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right] \lesssim \rho_n \mathcal{C}_m(\varphi, \lambda).$$

The proof of this important result is deferred to Section 6.5.

Remark 7. *It can be shown in specific situations (see, e.g., iv) of Theorem 6) that the solution of ($\mathbf{P}_\kappa$) is indeed a discrete measure that can be written as*

$$\hat{\mu}_n = \sum_{t \in \hat{S}} \hat{a}_t \delta_t.$$

In such a case, the relevance of the locations $\hat{S}$ of the reconstructed spikes $\hat{a}_t$ can be derived from the results of Theorem 10. A discussion is given in some specific cases in Section 5.

4.4. Support stability for large sample size

Assume that $\lambda = \lambda_\tau$ with $\tau = 1/(4m)$ for some bandwidth parameter $m \geq 1$. In this section, we drop the constraint $\int \mathrm{d}\mu = 1$ in the definition of BLASSO (see Remark 9 below). We introduce $\mathcal{P}_0 := \Phi c_0$ the "minimal norm certificate" (see *e.g.* [13]), which is defined by:

$$c_0 = \arg\min\{\|c\|_{\mathbb{L}}^2 \ : \ c \in \mathbb{L} \quad \text{s.t.} \quad \|\Phi c\|_\infty \leq 1 \text{ and } (\Phi c)(t_k) = 1\,,\ k \in [K]\}\,,$$

when it exists.

We say that the support $S^0 = \{t_1, \ldots, t_K\}$ of μ^0 satisfies the *Non-Degenerate Bandwidth* condition (NDB) if there exists $0 < q < 1$, $r > 0$ and $\rho > 0$ such that:

$$\mathcal{P}_0 \text{ exists}\,, \quad \forall t \in \mathbb{F}(r)\,,\ |\mathcal{P}_0(t)| < 1 - q\,, \quad \forall t \in \mathbb{N}(r)\,,\ \nabla^2 \mathcal{P}_0(t) \prec \rho\, \mathrm{Id}_d\,. \tag{NDB}$$

We then have the support stability result for large values of n.

Theorem 11. *Let $m \gtrsim K^{1/2} d^{7/3} \Delta^{-2}$ and $\mathcal{C}_m(\varphi)$ be the quantity introduced in Proposition 9. Let $\hat{\mu}_n$ be the BLASSO estimator given by ($\mathbf{P}_\kappa$) (without the constraint $\int \mathrm{d}\mu = 1$) with a tuning parameter $\kappa = \kappa_n$ such that $\|\Gamma_n\|_{\mathbb{L}} = o_{\mathbb{P}}(\kappa_n)$ and $\kappa_n = o(1)$.*

Then, under (NDB), for n large enough and with probability $1 - e_n$ (and $e_n > 0$ arbitrarily small), it holds that $\hat{\mu}_n$ has K spikes with exactly one spike $\hat{t}_k$ in each region $\mathbb{N}_k(r)$. These spikes converge to the true ones, and so do the amplitudes $\hat{a}_k$.

The proof can be found in Section 6.6.

Remark 8. *If the kernel φ is such that $\sigma = 1$ on $[-1/\tau, 1/\tau]^d$, then our certificate $\mathcal{P}_m$ is called the* **vanishing derivatives pre-certificate** *by [13, Section 4, Page 1335]. According to Theorem 7, we know that $\|\mathcal{P}_m\|_\infty \leq 1$. In this case, vanishing derivatives pre-certificate and certificate of minimal norm coincides so that $\mathcal{P}_m$ is the minimal norm certificate.*

Remark 9. *If we keep the constraint $\int \mathrm{d}\mu = 1$ active, then one can prove, studying the dual programs, that the "minimal norm certificate" is given by*

$$(c_0, \rho_0) = \arg\min\{\|c\|_{\mathbb{L}}^2 \ : \ (c, \rho) \in \mathbb{L} \times \mathbb{R}\ s.t.\ \|\Phi c - \rho\|_\infty \leq 1 \ and\ (\Phi c - \rho)(t_k) = 1\,,\ k \in [K]\}\,,$$

namely $c_0 = 0$, $\rho_0 = 1$ and $\mathcal{P}_0 = 1$. In this case, (NDB) never holds. Indeed, a negative result can be found in [13, Proposition 8] that shows that the dual polynomials converges to the minimal norm certificate $\mathcal{P}_0$. Since $\mathcal{P}_0 = 1$, it may suggest that the BLASSO with the constraint $\int \mathrm{d}\mu = 1$ may be extremely ill-conditioned for small values of κ, namely large sample sizes. This discussion leads to the conclusion that, at least theoretically, it would be better to drop the constraint $\int \mathrm{d}\mu = 1$ for large sample size.

5. Rates of convergence for some usual mixture models

5.1. Frequency cut-off and sinc kernel

In this section, we describe the consequences of Theorem 10 for some mixture models with classical densities φ. For this purpose, we will consider the sinus-cardinal kernel sinc with a frequency cut-off $1/\tau$, which is introduced in Example 2. As a band-limited function λ_τ, we have that

$$\|t\|_\infty \geq \frac{1}{\tau} \Longrightarrow \mathcal{F}[\lambda_\tau](t) = 0.$$

Hence, to obtain a tractable version of Theorem 10 with $\mathcal{C}_m(\varphi,\lambda) < +\infty$ (see Equation (22)) we are led to consider τ such that

$$\frac{1}{\tau} = 4m.$$

In that case, $\mathcal{F}[\lambda_\tau]$ is a constant function over its support and the term $\mathcal{C}_m(\varphi,\lambda_\tau)$ involved in Proposition 9 and Theorem 10 appears to be equal to

$$\mathcal{C}_m(\varphi,\lambda_\tau) = \frac{K^2 m^{-d/2} 2^{d/2}}{\inf_{\|t\|_\infty \leq 4m} \sigma(t)}.$$

To make use of Theorem 10, we also need an explicit expression of $(\rho_n)_{n\in\mathbb{N}^*}$, which itself strongly depends on the kernel λ_τ. In this context, some straightforward and standard computations yield

$$\begin{aligned}
\mathbb{E}\left[\|\Gamma_n\|_{\mathbb{L}}^2\right] &= \mathbb{E}\left[\|\hat{f}_n - f^0\|_{\mathbb{L}}^2\right], \\
&= \mathbb{E}\left[\int_{\|t\|_\infty \leq 1/\tau} \left|\mathcal{F}[\hat{f}_n](t) - \mathcal{F}[f^0](t)\right|^2 dt\right], \\
&= \int_{\|t\|_\infty \leq 1/\tau} \mathrm{Var}(\mathcal{F}[\hat{f}_n](t))\mathrm{d}t \leq \frac{1}{n\tau^d}.
\end{aligned}$$

This provides a natural choice for the sequence $(\rho_n)_{n\in\mathbb{N}^*}$ as

$$\forall n \in \mathbb{N}^* \qquad \rho_n = \frac{1}{\sqrt{n\tau^d}} = \frac{2^d m^{d/2}}{\sqrt{n}}.$$

Therefore, the statistical rate obtained in Theorem 10 satisfies

$$\rho_n \mathcal{C}_m(\varphi,\lambda_\tau) \leq \frac{K^2 2^{3d/2}}{\sqrt{n} \inf_{\|t\|_\infty \leq 4m} \{\sigma(t)\}}. \tag{23}$$

We should understand the previous inequality as an upper bound that translates a tradeoff between the sharpness of the window where spikes are located (given by $\epsilon = \mathcal{O}(1/(md^{3/2}))$ in (22)) and the associated statistical ability to recover a such targeted accuracy (given by the bound $\rho_n \mathcal{C}_m(\varphi,\lambda_\tau)$ on the Bregman divergence). A careful inspection of the previous tradeoff leads to the following conclusion: the window size ϵ is improved for large values of m but the statistical variability is then degraded according to the decrease rate of the Fourier transform σ of φ, which typically translates an inverse problem phenomenon.

Finally, we emphasize that the dimensionality effect is not only involved in the term $2^{3d/2}$ of Equation (23) but is also hidden in the constraint

$$m \gtrsim K^{1/2} d^{7/3} \Delta^{-2},$$

used to build our dual certificate in Theorem 7. By the way, we stress that at the end, the only tuning parameter involved in ($\mathbf{P}_\kappa$) appears to be m.

We now focus our attention to some specific and classical examples in mixture models:

- the Gaussian case, which is an example of *super-smooth inverse problems* with an exponential decrease of the Fourier transform at large frequencies.
- the case of *ordinary-smooth inverse problems* which encompasses multivariate Laplace distributions, Gamma distributions, double exponentials among others.

5.2. *Multivariate Gaussian mixtures: an example of super-smooth inverse problem*

We describe the behavior of our estimator when φ denotes the density of the standard d-dimensional Gaussian distribution $\varphi : u \longmapsto (2\pi)^{-d/2} e^{-\|u\|^2/2}$. In that case, the Fourier transform is

$$\mathcal{F}[\varphi](t) := \sigma(t) = e^{-\frac{\|t\|_2^2}{2}} \quad \forall t \in \mathbb{R}^d,$$

which entails

$$\inf_{\|t\|_\infty \le 4m} \sigma(t) = (2\pi)^{-d/2} e^{-8dm^2}.$$

A direct application of Theorem 10 leads to the following bounds.

Proposition 12. *Let $\mathcal{P}_m$ be a dual certificate given in Theorem 7 and let $\hat{\mu}_n$ be the BLASSO estimator given by ($\mathbf{P}_\kappa$) with $\kappa = \kappa_n$ chosen as in Proposition 9, then up to universal constants (independent from n, d, K and m):*

i) Far region and negative part: if $\epsilon = \mathcal{O}(\frac{1}{md^{3/2}})$, then:

$$\mathbb{E}\left[\hat{\mu}_n^-(\mathbb{R}^d)\right] \lesssim \frac{K^2(4\sqrt{\pi})^d e^{8dm^2}}{\sqrt{n}} \quad \text{and} \quad \mathbb{E}\left[\hat{\mu}_n^+(\mathbb{F}(\epsilon))\right] \lesssim \frac{K^2 d^4 (4\sqrt{\pi})^d e^{8dm^2}}{\sqrt{n}}.$$

ii) Near region (spike detection): a couple of constants $(c, \mathcal{C})$ exists such that

$$\forall A \subset \mathbb{R}^d, \quad \mathbb{E}[\hat{\mu}_n^+(A)] > c\frac{d^4(4\sqrt{\pi})^d K^2 e^{8dm^2}}{\sqrt{n}} \quad \Longrightarrow \quad \min_{k\in[K]} \inf_{t\in A} \|t - t_k\|_2^2 \le \frac{1}{\mathcal{C} d^4 m^2}.$$

iii) Near region (weight reconstruction): for any $k \in [K]$:

$$\mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right] \lesssim \frac{(4\sqrt{\pi})^d K^2 e^{8dm^2}}{\sqrt{n}}.$$

The proof (left to the reader) is a straightforward application of Theorem 10 and of the previous computations. Below, we shall discuss on some important consequences of Proposition 12:

- **Quantitative considerations** When the dimension d is kept fixed (as the number of components K and the minimal value for the spacings between the spikes Δ), the statistical ability of the BLASSO estimator $\hat{\mu}_n$ is driven by the term $e^{8dm^2}/\sqrt{n}$. In particular, this sequence converges to 0 provided that the following condition holds:

$$e^{8dm^2} \ll \sqrt{n} \quad \text{i.e.} \quad m = \mathcal{O}\left(\sqrt{\frac{\log(n)}{d}}\right) \quad \text{and} \quad m \longrightarrow +\infty \quad \text{as } n \longrightarrow +\infty. \tag{24}$$

  In other words, the maximal admissible value for m is $\sqrt{\frac{\log(n)}{8d}}$. In particular, if we consider $m = \sqrt{\frac{\kappa}{8}\frac{\log(n)}{d}}$ for κ small enough, we observe that

$$\mathbb{E}\left[\hat{\mu}_n^-(\mathbb{R}^d)\right] + \mathbb{E}\left[\hat{\mu}_n^+(\mathbb{F}(\epsilon_n))\right] \lesssim n^{\kappa-1/2}.$$

The counterpart of this admissible size for m is a slow rate for ϵ_n:

$$\epsilon_n = \mathcal{O}\left(m^{-1}d^{-3/2}\right) = \frac{\kappa^{-1/2}}{d\sqrt{\log n}}.$$

Therefore, the size of the near regions recovered at an almost parametric rate $n^{-1/2}$ are of the order $\log(n)^{-1/2}$.

- **Nature of the result** Item i) of Proposition 12 indicates that the mass set by $\hat{\mu}_n$ both on the negative part and on the far region tends to 0 as the sample size n grows under Condition (24) (see below). Our estimator is consistent: the mass allowed on the near region will be close to 1 as soon as n is large enough. At this step, we stress that the parameter m plays the role of a accuracy index: if m is constant, the mass of the near region converges to 1 at a parametric rate... but this near region is in this case not really informative. On the opposite hand, if m is close to the limit admissible value expressed in (24), the near region is very close to the support of the measure μ^0 but the convergence of the associated mass will be quite slow.

- **Case of dimension 1 and number of spikes detection** According to Item ii) of Proposition 12, any set with a sufficiently large mass is close to a true spike (a_k^0, t_k) for some $k \in [K]$. We stress that in the specific situation where $d = 1$, $\hat{\mu}_n$ is a discrete measure (see Theorem 6), namely that

$$\hat{\mu}_n = \sum_{\hat{t} \in \hat{S}} \hat{a}_{\hat{t}} \delta_{\hat{t}}.$$

In such a case, we get from Proposition 12 that if a reconstructed spike $(\hat{a}_{\hat{t}}, \hat{t})$ is large enough, it is in some sense close to a true spike. More formally, if $m = \mathcal{O}(\sqrt{\kappa \log(n)})$ and $\hat{t} \in \hat{S}$, then

$$\hat{a}_{\hat{t}} \gtrsim K^2 n^{-1/2+\kappa} \Longrightarrow \inf_{k \in [K]} |\hat{t} - t_k| \lesssim \frac{1}{\sqrt{\kappa \log(n)}}.$$

In particular, the BLASSO estimator $\hat{\mu}_n$ provides a lower bound on the number of true spikes. Once again, the value of m is critical in such a case. In particular, according to (24), we cannot expect more than a logarithmic precision.

- **Importance of the mixture parameters** It is also interesting to pay attention to the effect of the number of components K, the size of the minimal spacing Δ and of the dimension d on the statistical accuracy of our method. In the Gaussian case, the rate is of the order $K^2 C^d e^{8dm^2} n^{-1/2}$ but an important effect is hidden in the constraint brought by Theorem 7:

$$m \gtrsim K^{1/2} d^{7/3} \Delta^{-2}.$$

In particular, the behavior of our estimator is seriously damaged in the Gaussian situation when $(\Delta^{-1} \vee K \vee d) \to +\infty$ since in that case we obtain a rate of the order

$$e^{d^{17/3} K \Delta^{-2}} n^{-1/2}.$$

We observe that d, K and Δ^{-1} cannot increase faster than a power of a logarithm of the number of observations: $d^{17/3}K\Delta^{-2} \ll \log(n)$. We will observe in Section 5.3 that a such hard constraint disappears in more favorable cases with smaller degrees of ill-posedness.

- **Position of our result on Gaussian mixture models**
To conclude this discussion, we would like to recall that the BLASSO estimator $\hat{\mu}_n$ depends on m. This parameter plays the role of a precision filter and only provides a quantification of the performances of our method. This is one of the main differences with the classical super-resolution theory where in general m is fixed and constrained by the experiment. We should point out that many works have studied statistical estimation in Gaussian mixture models with a semi-parametric point of view (see, *e.g.* [31], [5]). These investigations are often reduced to the two-component case (K=2): we refer to [7], [19] or [18] among others. The general case ($K \in \mathbb{N}^*$) has been for instance addressed in [21] using a model selection point of view: the selection of K is achieved through the minimization of a criterion penalized by the number of components. We also refer to [6] where a Lasso-type estimator is built for mixture model using a discretization of the possible values of t_k. However, this last approach is limited by some constraints on the Gram matrix involved in the model that do not allow to consider situations where Δ is small: in [6], the minimal separation between two spikes has to satisfy $\Delta \geq \Delta_0 > 0$, *i.e.* has to be lower bounded by a positive constant Δ_0, which depends on the mixing distribution φ. We emphasize that in our work, we only need an upper bound on K and a lower bound on Δ or at least to assume that these quantities are fixed w.r.t. n. According to Proposition 12, our constraint expressed on these parameters already allows to cover a large number of interesting situations.

5.3. Ordinary smooth distributions

General result Gaussian distributions belong to the class of super-smooth distributions, oppositely, ordinary smooth ones are described through a polynomial decrease of their Fourier transform.

Hypothesis 1 (($\mathcal{H}_\beta^{smooth}$) on the spectral density σ)**.** *The function φ satisfies*

$$\mathcal{F}[\varphi] = \sigma \quad \textit{and} \quad \|x\|_2^{-\beta} \lesssim \sigma(x) \lesssim \|x\|_2^{-\beta} \quad \textit{when} \quad \|x\|_2 \to +\infty. \qquad (\mathcal{H}_\beta^{smooth})$$

We refer to [16] and the references therein for an extended description of the class of distributions involved by ($\mathcal{H}_\beta^{smooth}$) and some statistical consequences in the situation of standard non-parametric deconvolution. For our purpose, it is straightforward to verify that

$$\inf_{\|t\|_\infty \leq 4m} \sigma(t) \leq \inf_{\|t\|_2 \leq 4m\sqrt{d}} \sigma(t) \lesssim [\sqrt{d}m]^{-\beta}.$$

In that case, we obtain that

$$\rho_n \mathcal{C}_m(\varphi, \lambda_\tau) \lesssim \frac{K^2 2^{3d/2} m^\beta d^{\beta/2}}{\sqrt{n}}.$$

We then deduce the following result (which is a direct application of Theorem 10).

Proposition 13. *Assume that φ is ordinary smooth and satisfies $(\mathcal{H}_\beta^{smooth})$. Let $\mathcal{P}_m$ be the dual certificate given in Theorem 7 and let $\hat{\mu}_n$ be the BLASSO estimator given by $(\mathbf{P}_\kappa)$ with $\kappa = \kappa_n$ chosen as in Proposition 9, then up to universal constants (independent from n, d, K and m):*

i) Far region and negative part: if $\epsilon \propto \frac{1}{md^{3/2}}$, then:

$$\mathbb{E}\left[\hat{\mu}_n^-(\mathbb{R}^d)\right] \lesssim \frac{K^2 2^{3d/2} m^\beta d^{\beta/2}}{\sqrt{n}} \quad \text{and} \quad \mathbb{E}\left[\hat{\mu}_n^+(\mathbb{F}(\epsilon))\right] \lesssim \frac{K^2 d^4 2^{3d/2} m^\beta d^{\beta/2}}{\sqrt{n}}.$$

ii) Near region (spike detection): a couple of constants $(c, \mathcal{C})$ exists such that

$$\forall A \subset \mathbb{R}^d, \quad \mathbb{E}[\hat{\mu}_n^+(A)] > c\frac{d^4 2^{3d/2} K^2 m^\beta d^{\beta/2}}{\sqrt{n}} \quad \Longrightarrow \quad \min_{k \in [K]} \inf_{t \in A} \|t - t_k\|_2^2 \leq \frac{1}{\mathcal{C} d^4 m^2}.$$

iii) Near region (weight reconstruction): for any $k \in [K]$:

$$\mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right] \lesssim \frac{2^{3d/2} K^2 m^\beta d^{\beta/2}}{\sqrt{n}}.$$

The proof of this proposition is omitted, and we only comment on the consequences of this result for ordinary smooth mixtures. We obtain a consistent estimation with the BLASSO estimator $\hat{\mu}_n$ when m is chosen such that

$$m_n = n^\delta \quad \text{with} \quad \delta < \frac{1}{2\beta} \quad \text{as} \quad n \to +\infty.$$

For this settings, the size of ϵ_n is $\epsilon_n \propto d^{-3/2} n^{-\delta}$. Again, when $K \vee d \vee \Delta^{-1}$ is allowed to grow towards $+\infty$, the effect of these quantities is translated by a rate of the order

$$\frac{K^{2+\beta/2} \Delta^{-2\beta} 2^{3d/2}}{\sqrt{n}}.$$

In particular, the dimension should not increase faster than $\frac{\log n}{3 \log 2}$. In the same way, the minimal size of spacings to permit a consistent estimation should not be smaller than $n^{-1/(4\beta)}$. In particular, this indicates that a polynomial accuracy is possible (see e.g. Item *ii*) of Proposition 13). This emphasized the strong role played by the mixture density φ in our analysis.

Multivariate Laplace distributions The multivariate distribution is described by :

$$\sigma(x) = \frac{2}{2 + \|x\|_2^2}.$$

We obtain here an ordinary smooth density with $\beta = 2$. The minimal spacing for a discoverable spike is therefore of the order $n^{-1/4}$ while the constraint on the dimension is not affected by the value of β. Concerning the number of components K, its value should not exceed $n^{1/6}$ and the smallest size of the window ϵ_n is $n^{-1/4}$.

Tensor product of Laplace distributions Another interesting case is the situation where φ is given by a tensor product of standard Laplace univariate distributions:

$$\varphi(x) = \frac{1}{2^d} e^{-\sum_{j=1}^d |x_j|} \quad \text{and} \quad \mathcal{F}[\varphi](x) := \sigma(x) = \prod_{j=1}^d \frac{1}{1 + x_j^2} \quad \forall x \in \mathbb{R}^d.$$

In that case, $\beta = 2d$ and the previous comments apply: the maximal value of m is $n^{1/4d}$ with an optimal size of the window of the order $n^{-1/(4d)}$ whereas d should not be larger than $\log n$.

6. Proof of the Main Results

6.1. Perfect recovery

This paragraph is dedicated to the proof of the perfect recovery property under ($\mathcal{H}_0$) and ($\mathcal{H}_\infty$).

Proof of Theorem 2. Remark first that under ($\mathcal{H}_0$) and ($\mathcal{H}_\infty$), the RKHS, denoted by $\mathbb{H}$, generated by the kernel $h(.,.) = \varphi(. - .)$ is dense in $\mathcal{C}_0(\mathbb{R}^d, \mathbb{R})$ with respect to the uniform norm, see [9, Proposition 5.6] for instance. Furthermore, using [29, Proposition 2], we can show that ($\mathcal{H}_\infty$) implies that the embedding Φ is injective onto $\mathbb{H}$. This means that we have identifiability of μ from the knowledge of $\Phi\mu$. More precisely, denote $f^0 := \Phi\mu^0$, we deduce that if it holds $\|f^0 - \Phi\mu\|_{\mathbb{H}}^2 = 0$ then one has $\mu = \mu^0$. □

6.2. Perfect recovery with a dual certificate

This paragraph is dedicated to the proof of Theorem 3, which entails the perfect recovery property under a less restrictive assumption on φ.

Proof of Theorem 3. Let

$$\hat{\mu} \in \arg \min_{\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}) \,:\, \Phi\mu = f^0} \|\mu\|_1 .$$

Step 1: Support inclusion. We observe that both $\hat{\mu}$ and μ^0 belong to $\mathcal{M}(f^0)$ so that $\Phi\hat{\mu} = \Phi\mu^0$. Hence, considering the Fourier transform on both sides and using (6), we get that $\sigma\mathcal{F}(\hat{\mu}) = \sigma\mathcal{F}(\mu^0)$ which is equivalent to $\mathcal{F}(\hat{\mu}) = \mathcal{F}(\mu^0)$ on the support of σ. Now, Assumption ($\mathcal{H}_\eta$) yields:

$$(\mathcal{F}(\hat{\mu}) - \mathcal{F}(\mu^0))\mathbb{1}_{[-\eta,\eta]^d} = 0 . \tag{25}$$

Denote by $q_\eta := \mathcal{F}(\mathcal{P}_\eta)$ the Fourier transform of $\mathcal{P}_\eta$. By assumption, the support of q_η is included in $[-\eta, \eta]^d$ and from (25) we get that:

$$\int_{\mathbb{R}^d} q_\eta \mathcal{F}(\hat{\mu}) = \int_{\mathbb{R}^d} q_\eta \mathcal{F}(\mu^0) .$$

Since $\mathcal{P}_\eta \in L^1(\mathbb{R}^d)$, the Riemann-Lebesgue lemma shows that q_η is continuous. Recall also that q_η has a compact support so we deduce that $q_\eta \in L^1(\mathbb{R}^d)$. By Fourier inversion theorem, we have

$$\int_{\mathbb{R}^d} \mathcal{F}(q_\eta) \mathrm{d}\hat{\mu} = \int_{\mathbb{R}^d} q_\eta \mathcal{F}(\hat{\mu}) = \int_{\mathbb{R}^d} q_\eta \mathcal{F}(\mu^0) = \int_{\mathbb{R}^d} \mathcal{F}(q_\eta) \mathrm{d}\mu^0,$$

namely

$$\int_{\mathbb{R}^d} \mathcal{P}_\eta \mathrm{d}\hat{\mu} = \int_{\mathbb{R}^d} \mathcal{P}_\eta \mathrm{d}\mu^0 .$$

Remark that $\mathcal{P}_\eta$ satisfies

$$\int_{\mathbb{R}^d} \mathcal{P}_\eta \mathrm{d}\mu^0 = \|\mu^0\|_1 ,$$

and the Hölder inequality leads to

$$\int_{\mathbb{R}^d} \mathcal{P}_\eta \mathrm{d}\hat{\mu} \leq \|\mathcal{P}_\eta\|_\infty \|\hat{\mu}\|_1 = \|\hat{\mu}\|_1 .$$

From the definition of $\hat{\mu}$, one also has $\|\hat{\mu}\|_1 \leq \|\mu^0\|_1$. Putting everything together, we deduce that

$$\|\hat{\mu}\|_1 = \int_{\mathbb{R}^d} \mathcal{P}_\eta \mathrm{d}\hat{\mu} = \|\mu^0\|_1 \,.$$

Since $\mathcal{P}_\eta$ is continuous and strictly lower than one outside of the support of μ^0, we deduce from the above equality that the support of $\hat{\mu}$ is included in the support of μ^0:

$$\mathrm{Supp}(\hat{\mu}) \subset \mathrm{Supp}(\mu^0) = S^0 \,.$$

<u>Step 2: Identifiability and conclusion.</u> We prove that $\{\varphi(\cdot - t_1), \ldots, \varphi(\cdot - t_K)\}$ spans a vector subspace of $\mathcal{C}_0(\mathbb{R}^d, \mathbb{R}) \cap L^1(\mathbb{R}^d)$ of dimension K. This proof is standard and relies on a Vandermonde argument. We assume first that K coefficients $x_1, \ldots, x_K \in \mathbb{R}$ exist such that:

$$\sum_{k=1}^{K} x_k \varphi(\cdot - t_k) = 0 \,.$$

Applying the Fourier transform and using (7), we deduce that:

$$\sigma(u) \sum_{k=1}^{K} x_k e^{\imath u^\top t_k} = 0 \,, \quad \forall u \in \mathbb{R}^d \,.$$

Since σ is nonzero, there exists an open set $\Omega \subseteq \mathbb{R}^d$ such that $\sigma > 0$ on Ω. We deduce that:

$$\sum_{k=1}^{K} x_k e^{\imath u^\top t_k} = 0 \,, \quad \forall u \in \Omega \,.$$

Now, we can choose some points u_j in Ω so that the Fourier matrix with entries $(e^{\imath u_j^\top t_k^0})_{kj}$ is invertible. It implies that $x_k = 0$ and $\{\varphi(\cdot - t_1), \ldots, \varphi(\cdot - t_K)\}$ spans a subspace of dimension K.

We now conclude the proof. We know from Step 1 that

$$\hat{\mu} = \sum_{k=1}^{K} x_k \delta_{t_k} .$$

Since $\hat{\mu}$ and μ^0 belong to $\mathcal{M}(f^0)$, then

$$\sum_{k=1}^{K} x_k \varphi(. - t_k) = \sum_{k=1}^{K} a_k^0 \varphi(. - t_k),$$

which in turn implies that $x_k = a_k^0$ for all $k \in \{1, \ldots, K\}$, namely $\hat{\mu} = \mu^0$. □

6.3. Primal-Dual problems and duality gap

This paragraph is dedicated to the proof of the equality between the optimal primal value ($\mathbf{P}_\kappa$) and the optimal dual value ($\mathbf{P}_\kappa^*$), which implies the no duality gap result.

Proof of Theorem 6. We consider some primal variables $\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$ and $z \in \mathbb{L}$ and introduce the dual variables $c \in \mathbb{L}$ and $\rho \in \mathbb{R}$. The Lagrangian is given in Equation (16) and we consider the dual problem ($\mathbf{P}_\kappa^*$).

Proof of i). The existence of some solutions $\hat{\mu}_n$ to the primal problem ($\mathbf{P}_\kappa$) is obtained with the help of a standard argument: we prove that the primal objective function is a proper lower semi-continuous (for the weak-* topology) convex function on the Banach space $\mathcal{M}(\mathbb{R}^d, \mathbb{R})$.

We now consider the "invariant property" related to the solutions of ($\mathbf{P}_\kappa$). The norm $\|\cdot\|_{\mathbb{L}}$ satisfies

$$\forall a, b \in \mathbb{L}\,, \quad \frac{\|a\|_{\mathbb{L}}^2 + \|b\|_{\mathbb{L}}^2}{2} - \Big\|\frac{a+b}{2}\Big\|_{\mathbb{L}}^2 = \frac{\|a-b\|_{\mathbb{L}}^2}{4}\,. \tag{26}$$

Now consider two primal solutions μ_1 and μ_2 and define $\tilde{\mu} = (\mu_1 + \mu_2)/2$. Using (26) and the triangle inequality for $\|.\|_1$, one has:

$$\begin{aligned}
\frac{1}{2}\|L\hat{f}_n - L\circ\Phi\tilde{\mu}\|_{\mathbb{L}}^2 + \kappa\|\tilde{\mu}\|_1 &\leq \frac{1}{2}\|L\hat{f}_n - L\circ\Phi\tilde{\mu}\|_{\mathbb{L}}^2 + \kappa\frac{\|\mu_1\|_1 + \|\mu_2\|_1}{2} \\
&\leq \frac{\frac{1}{2}\|L\hat{f}_n - L\circ\Phi\mu_1\|_{\mathbb{L}}^2 + \kappa\|\mu_1\|_1}{2} \\
&\quad + \frac{\frac{1}{2}\|L\hat{f}_n - L\circ\Phi\mu_2\|_{\mathbb{L}}^2 + \kappa\|\mu_2\|_1}{2} \\
&\quad - \frac{1}{8}\|L\circ\Phi\mu_1 - L\circ\Phi\mu_2\|_{\mathbb{L}}^2\,.
\end{aligned}$$

But, remind that:

$$\frac{1}{2}\|L\hat{f}_n - L\circ\Phi\mu_1\|_{\mathbb{L}}^2 + \kappa\|\mu_1\|_1 = \frac{1}{2}\|L\hat{f}_n - L\circ\Phi\mu_2\|_{\mathbb{L}}^2 + \kappa\|\mu_2\|_1 = \min\left\{\frac{1}{2}\|L\hat{f}_n - L\circ\Phi\mu\|_{\mathbb{L}}^2 + \kappa\|\mu\|_1\right\}.$$

We then conclude that $\tilde{\mu}$ is also a solution to the primal problem and that $L\circ\Phi\mu_1 = L\circ\Phi\mu_2$. We can repeat this argument for any pair of primal solutions so that the quantity $\hat{z}_n := L\circ\Phi\hat{\mu}_n$ is uniquely defined and does not depend on the choice of the primal solution point $\hat{\mu}_n$. It also implies that $\hat{m}_n := \|\hat{\mu}_n\|_1$ is uniquely defined (does not depend on the choice of the primal solution point).

Proof of ii). We shall write the dual program ($\mathbf{P}_\kappa^*$) as follows: consider dual variables (c, ρ) and write:

$$\inf_{\mu,z} \mathcal{L}(\mu, z, c, \rho) = \inf_{\mu,z}\Big\{\underbrace{\frac{1}{2}\|L\hat{f}_n - z\|_{\mathbb{L}}^2 + \langle c, z\rangle_{\mathbb{L}}}_{①} + \underbrace{\kappa\|\mu\|_1 - \langle c, L\circ\Phi\mu\rangle_{\mathbb{L}} + \rho\int_{\mathbb{R}^d}\mathrm{d}\mu - \rho}_{②}\Big\}\,,$$

and the previous infimum appears to be splitted in terms of the influence of z and μ. Optimizing in z the first term ① leads to $z = L\hat{f}_n - c$ so that:

$$\inf_z ① = \langle c, L\hat{f}_n\rangle_{\mathbb{L}} - \frac{1}{2}\|c\|_{\mathbb{L}}^2 = \frac{1}{2}\big(\|L\hat{f}_n\|_{\mathbb{L}}^2 - \|L\hat{f}_n - c\|_{\mathbb{L}}^2\big)\,.$$

The second term ② is more intricate. Observe that:

$$\begin{aligned}
\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\Big|c(s)\varphi(s-u)\mathrm{d}\mu(u)\Big|\mathrm{d}s &\leq \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|c\|_\infty\varphi(s-u)\mathrm{d}|\mu|(u)\mathrm{d}s\,, \\
&= \|c\|_\infty\|\varphi\|_1\|\mu\|_1 = \|c\|_\infty\|\mu\|_1 < \infty\,,
\end{aligned}$$

and the Fubini yields:

$$
\begin{aligned}
\langle c, (L\circ\Phi)\mu\rangle_{\mathbb{L}} &= \langle c(\cdot), \int_{\mathbb{R}^d}\lambda(\cdot-s)(\Phi\mu)(s)\mathrm{d}s\rangle_{\mathbb{L}} = \int_{\mathbb{R}^d}\langle c(\cdot),\lambda(\cdot-s)\rangle_{\mathbb{L}}(\Phi\mu)(s)\mathrm{d}s\,,\\
&= \int_{\mathbb{R}^d} c(s)(\Phi\mu)(s)\mathrm{d}s = \int_{\mathbb{R}^d} c(s)\Big(\int_{\mathbb{R}^d}\varphi(s-u)\mathrm{d}\mu(u)\Big)\mathrm{d}s\,,\\
&= \int_{\mathbb{R}^d}\Big(\int_{\mathbb{R}^d} c(s)\varphi(u-s)\mathrm{d}s\Big)\mathrm{d}\mu(u)\,,\\
&= \int_{\mathbb{R}^d}\Phi c\,\mathrm{d}\mu\,.
\end{aligned}
\tag{27}
$$

We deduce that:

$$
② = \kappa\|\mu\|_1 - \int_{\mathbb{R}^d}(\Phi c-\rho)\,\mathrm{d}\mu\,.
$$

We use the L^1-L^∞ Hölder inequality, namely $\int_{\mathbb{R}^d}(\Phi c-\rho)\,\mathrm{d}\mu \le \|\Phi c-\rho\|_\infty\|\mu\|_1$, which yields:

$$
\inf_\mu ② \ge \inf_\mu[\kappa-\|\Phi c-\rho\|_\infty]\|\mu\|_1 = [\kappa-\|\Phi c-\rho\|_\infty]\inf_\mu\|\mu\|_1\,.
$$

Hence, we conclude that:

$$
\inf_\mu ② = -\boldsymbol{I}_{\{\|\Phi c-\rho\|_\infty\le\kappa\}}(c,\rho)\,,
\tag{28}
$$

where $\boldsymbol{I}_{\{\|\Phi c-\rho\|_\infty\le\kappa\}}(c,\rho)$ is the constraint $\|\Phi c-\rho\|_\infty\le\kappa$, namely it is 0 if (c,ρ) are such that $\|\Phi c-\rho\|_\infty\le\kappa$ and ∞ otherwise. Finally, we obtain that for a fixed pair of dual variables (c,ρ):

$$
\inf_{\mu,z}\mathcal{L}(\mu,z,c,\rho) = \frac{1}{2}\Big(\|L\hat f_n\|_{\mathbb{L}}^2-\|L\hat f_n-c\|_{\mathbb{L}}^2\Big)-\rho-\boldsymbol{I}_{\{\|\Phi c-\rho\|_\infty\le\kappa\}}(c,\rho)\,.
$$

Hence, the dual problem ($\mathbf{P}_\kappa^*$) shall be written as

$$
\begin{aligned}
(\mathbf{P}_\kappa^*) \iff \sup_{c,\rho}\inf_{\mu,z}\mathcal{L}(\mu,z,c,\rho) &= \sup_{c,\rho}\Big\{\frac{1}{2}\Big(\|L\hat f_n\|_{\mathbb{L}}^2-\|L\hat f_n-c\|_{\mathbb{L}}^2\Big)-\rho-\boldsymbol{I}_{\{\|\Phi c-\rho\|_\infty\le\kappa\}}(c,\rho)\Big\}\,,\\
&= \frac{\|L\hat f_n\|_{\mathbb{L}}^2}{2} - \inf_{(c,\rho)\,:\ \|\Phi c-\rho\|_\infty\le\kappa}\Big\{\rho+\frac{1}{2}\|L\hat f_n-c\|_{\mathbb{L}}^2\Big\}\,.
\end{aligned}
$$

Here again the dual objective function is lower semi-continuous and coercive on the Hilbert space $\mathbb{L}\times\mathbb{R}$ so a pair of minimizers $(\hat c,\hat\rho)$ exists. Again, Inequality (26) implies the uniqueness of $\hat c$, which in turn implies the uniqueness of $\hat\rho$.

To prove that there is no duality gap, we use the Slater condition: we remark that a feasible point (c°,ρ°) exists in the interior of the constrained set $\{\|\Phi c-\rho\|_\infty\le\kappa\}$. Now, the generalized Slater condition shall be used (see *e.g.* [26]). Indeed, given any nonzero $c\in\mathbb{L}\subseteq\mathcal{C}_0(\mathbb{R}^d,\mathbb{R})$, note that the convolution operator satisfies $\|\Phi c\|_\infty\le\|c\|_\infty$. Hence, we set $\rho^\circ=0$ and $c^\circ=\kappa c/(2\|c\|_\infty)$ and these points are in the interior of the constrained set. The generalized Slater condition implies that strong duality holds, and there is no duality gap:

$$
(\mathbf{P}_\kappa) = (\mathbf{P}_\kappa^*).
$$

Proof of *iii)*. We consider the unique pair $(\hat c,\hat\rho)$ solution to

$$
(\hat c,\hat\rho) = \arg\min_{(c,\rho)\in\mathbb{L}\times\mathbb{R}\,:\ \|\Phi c-\rho\|_\infty\le\kappa}\Big\{\rho+\frac{1}{2}\|L\hat f_n-c\|_{\mathbb{L}}^2\Big\}\,,
$$

and the strong duality implies that:

$$0 = \kappa\|\hat{\mu}\|_1 - \langle \hat{c}, L \circ \Phi\hat{\mu}\rangle_{\mathbb{L}} + \rho \int_{\mathbb{R}^d} \mathrm{d}\hat{\mu} = \kappa\|\hat{\mu}\|_1 - \int_{\mathbb{R}^d} (\Phi\hat{c} - \hat{\rho})\, \mathrm{d}\hat{\mu}\,.$$

Since $\Phi\hat{c} - \hat{\rho}$ is continuous, we verify, using the argument of Lemma A.1 in [10], that:

$$\mathrm{Supp}(\hat{\mu}) \subseteq \left\{ x \in \mathbb{R}^d \ : \ \left|\Phi\hat{c} - \hat{\rho}\right|(x) = \kappa \right\},$$

where we recall that $\Phi\hat{c} - \hat{\rho} \in L^\infty(\mathbb{R}^d)$ is such that its supremum norm is less than κ.
Proof of *iv)*. The last point is a consequence of the Schwartz-Paley-Wiener theorem (see *e.g.* Theorem XVI, chapter VII in [28, Page 272]). Indeed, note that $\Phi\hat{c}$ is a continuous function whose inverse Fourier transform has a support included in the support of $\sigma \times \Lambda$. By assumption, this latter is bounded and one may apply the Schwartz-Paley-Wiener Theorem: we deduce that $\Phi\hat{c}$ can be extended to complex values $\mathbb{C}^d$ into an analytic entire function of exponential type. In particular, $\Phi\hat{c} - \hat{\rho} \pm \kappa$ has isolated zeros one the real line, which concludes the proof. ☐

6.4. Analysis of the Bregman divergence

This paragraph is devoted to the statistical analysis of the Bregman divergence whose definition is recalled below:

$$D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) := \|\hat{\mu}_n\|_1 - \|\mu^0\|_1 - \int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) \geq 0\,.$$

Proof of Proposition 9. According to the definition of $\hat{\mu}_n$ as the minimum of our variational criterion (see Equation (15)), we know that:

$$\|L\hat{f}_n - L \circ \Phi\hat{\mu}_n\|_{\mathbb{L}}^2 + \kappa\|\hat{\mu}_n\|_1 \leq \|L\hat{f}_n - L \circ \Phi\mu^0\|_{\mathbb{L}}^2 + \kappa\|\mu^0\|_1.$$

Proof of *i)*. With our notation $\Gamma_n = L\hat{f}_n - L \circ \Phi\mu^0$ introduced in Section 4.2, we deduce that:

$$\|L\hat{f}_n - L \circ \Phi\hat{\mu}_n\|_{\mathbb{L}}^2 + \kappa\|\hat{\mu}_n\|_1 \leq \|\Gamma_n\|_{\mathbb{L}}^2 + \kappa\|\mu^0\|_1.$$

Using now $\mathcal{P}_m$ obtained in Theorem 7, we deduce that

$$\|L\hat{f}_n - L \circ \Phi\hat{\mu}_n\|_{\mathbb{L}}^2 + \kappa \left[\|\hat{\mu}_n\|_1 - \|\mu^0\|_1 - \int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) \right] + \kappa \int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) \leq \|\Gamma_n\|_{\mathbb{L}}^2. \tag{29}$$

Hence, we deduce the following upper bound on the Bregman divergence:

$$\|L\hat{f}_n - L \circ \Phi\hat{\mu}_n\|_{\mathbb{L}}^2 + \kappa D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) + \kappa \int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu) \leq \|\Gamma_n\|_{\mathbb{L}}^2. \tag{30}$$

According to Theorem 7, $\mathcal{P}_m = \Phi c_{0,m}$ for some $c_{0,m} \in \mathbb{L}$. In particular, we get as in (27):

$$\begin{aligned}\int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) &= \langle \mathcal{P}_m, \hat{\mu}_n - \mu^0 \rangle_{L^2(\mathbb{R}^d)}, \\ &= \langle \Phi c_{0,m}, \hat{\mu}_n - \mu^0 \rangle_{L^2(\mathbb{R}^d)}, \\ &= \langle c_{0,m}, \Phi(\hat{\mu}_n - \mu^0) \rangle_{L^2(\mathbb{R}^d)},\end{aligned}$$

where the last equality comes from the self-adjoint property of Φ in $L^2(\mathbb{R}^d)$. The reproducing kernel relationship yields:

$$\begin{aligned}\int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) &= \int_{\mathbb{R}^d} \langle c_{0,m}, \lambda(t-.)\rangle_{\mathbb{L}} \Phi(\hat{\mu}_n - \mu^0)(t)dt, \\ &= \langle c_{0,m}, L \circ \Phi(\hat{\mu}_n - \mu^0)\rangle_{\mathbb{L}}, \\ &= \langle c_{0,m}, L \circ \Phi\hat{\mu}_n - L\hat{f}_n + \Gamma_n\rangle_{\mathbb{L}}. \end{aligned} \tag{31}$$

Gathering (30) and (31), we deduce that:

$$\|L\hat{f}_n - L\circ\Phi\hat{\mu}_n\|_{\mathbb{L}}^2 + \kappa D_{\mathcal{P}_m}(\hat{\mu}_n,\mu^0) + \kappa\langle c_{0,m}, L\circ\Phi\hat{\mu}_n - L\hat{f}_n\rangle_{\mathbb{L}} + \kappa\langle c_{0,m},\Gamma_n\rangle_{\mathbb{L}} \leq \|\Gamma_n\|_{\mathbb{L}}^2 .$$

Using now a straightforward computation with $\|.\|_{\mathbb{L}}$, we conclude that:

$$\left\|L\hat{f}_n - L\circ\Phi\hat{\mu}_n - \frac{\kappa}{2}c_{0,m}\right\|_{\mathbb{L}}^2 + \kappa D_{\mathcal{P}_m}(\hat{\mu}_n,\mu^0) \leq \left\|\Gamma_n - \frac{\kappa}{2}c_{0,m}\right\|_{\mathbb{L}}^2 .$$

Since the first term of the left hand side is positive, the previous inequality leads to:

$$D_{\mathcal{P}_m}(\hat{\mu}_n,\mu^0) \leq \frac{3}{2\kappa}\|\Gamma_n\|_{\mathbb{L}}^2 + \frac{3\kappa}{4}\|c_{0,m}\|_{\mathbb{L}}^2, \tag{32}$$

where we have used a triangle inequality for the right hand side. We now consider a sequence $(\rho_n)_{n\in\mathbb{N}^*}$ such that $\mathbb{E}[\|\Gamma_n\|_{\mathbb{L}}^2] \leq \rho_n^2$ for all $n \in \mathbb{N}^*$ and we choose:

$$\kappa = \sqrt{2}\rho_n/\|c_{0,m}\|_{\mathbb{L}}.$$

Then we deduce from (32) that:

$$\mathbb{E}[D_{\mathcal{P}_m}(\hat{\mu}_n,\mu^0)] \leq \frac{3\sqrt{2}}{2}\rho_n \times \|c_{0,m}\|_{\mathbb{L}}. \tag{33}$$

Proof of *ii*). We now derive an upper bound on $\|c_{0,m}\|_{\mathbb{L}}$. Recall that according to ($\mathcal{H}_0$) and in particular (12) we have:

$$\|g\|_{\mathbb{L}}^2 = \int_{\mathbb{R}^d} \frac{|\mathcal{F}[g](t)|^2}{\mathcal{F}[\lambda](t)}dt \quad \forall g \in \mathbb{L}.$$

Since φ is symmetric and $\Phi^\star = \Phi$, we have according to Theorem 7 that:

$$\begin{aligned}\|\mathcal{P}_m\|_2^2 &= \|\Phi c_{0,m}\|_2^2, \\ &= \int_{\mathbb{R}^d} |\mathcal{F}[\varphi](t)|^2 |\mathcal{F}[c_{0,m}](t)|^2 dt, \\ &= \int_{\mathbb{R}^d} |\mathcal{F}[\varphi](t)|^2 \mathcal{F}[\lambda](t) \times \frac{|\mathcal{F}[c_{0,m}](t)|^2}{\mathcal{F}[\lambda](t)}dt, \\ &\geq \inf_{\|t\|_\infty \leq 4m} \left\{|\mathcal{F}[\varphi](t)|^2 \mathcal{F}[\lambda](t)\right\} \|c_{0,m}\|_{\mathbb{L}}^2. \end{aligned} \tag{34}$$

Indeed, *iii*) of Theorem 7 entails that the support of the Fourier transform of $\mathcal{P}_m$ is contained in $[-4m,4m]^d$. This embedding, together with ($\mathcal{H}_\infty$) entails:

$$\mathrm{Supp}(\mathcal{F}[\mathcal{P}_m]) \subset [-4m,4m]^d,$$

which provides the last inequality. The inequality (34) can be rewritten as:

$$\|c_{0,m}\|_{\mathbb{L}}^2 \leq \frac{\|\mathcal{P}_m\|_2^2}{\inf_{\|t\|_\infty\leq 4m}\{|\mathcal{F}[\varphi](t)|^2\mathcal{F}[\lambda](t)\}}. \tag{35}$$

We use (33), (35) and observe that $|\mathcal{F}[\varphi]| = \sigma$ to conclude the proof. □

6.5. Near and Far region estimations

In this paragraph, we provide the main result of the paper that establishes the statistical accuracy of our BLASSO estimation.

Proof of Theorem 10. Proof of $i)$ In a first time, we provide a lower bound on the Bregman divergence. This bound takes advantage on the properties of the dual certificate associated to Theorem 7. First remark that

$$\begin{aligned} \int \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) &= \int \mathcal{P}_m \mathrm{d}\hat{\mu}_n - \sum_{k=1}^{K} a_k^0 \mathcal{P}_m(t_k) \\ &\leq \|\hat{\mu}_n\|_1 - \|\mu^0\|_1, \end{aligned}$$

since $\mathcal{P}_m(t_k) = 1$ for all $k \in [K]$. This inequality yields the positiveness of the Bregman divergence:

$$D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) \geq 0.$$

Now, using similar arguments and the Borel's decomposition $\hat{\mu}_n = \hat{\mu}_n^+ - \hat{\mu}_n^-$, we obtain

$$\begin{aligned} D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) &= \|\hat{\mu}_n\|_1 - \|\mu^0\|_1 - \int \mathcal{P}_m \mathrm{d}\hat{\mu}_n + \int \mathcal{P}_m \mathrm{d}\mu^0, \\ &= \|\hat{\mu}_n\|_1 - \int \mathcal{P}_m \mathrm{d}\hat{\mu}_n, \\ &= \int \mathrm{d}\hat{\mu}_n^+ + \int \mathrm{d}\hat{\mu}_n^- - \int \mathcal{P}_m \mathrm{d}\hat{\mu}_n^+ + \int \mathcal{P}_m \mathrm{d}\hat{\mu}_n^-, \\ &= \int (1 - \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^+ + \int (1 + \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^-. \end{aligned}$$

Proposition 9 then implies that:

$$\mathbb{E}\left[\int (1 - \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^+ + \int (1 + \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^-\right] \leq \frac{3\sqrt{2}}{2} \rho_n \mathcal{C}_m(\varphi, \lambda). \tag{36}$$

Weight of the negative part. Since the dual certificate $\mathcal{P}_m$ is always positive, we have

$$\mu_n^-(\mathbb{R}^d) = \int \mathrm{d}\hat{\mu}_n^- \leq \int (1 + \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^- \leq \frac{3\sqrt{2}}{2} \rho_n \mathcal{C}_m(\varphi, \lambda). \tag{37}$$

Moreover, according to item $ii)$ of Theorem 7,

$$1 - \mathcal{P}_m(t) \geq \gamma \frac{\upsilon}{d^4} \quad \forall t \in \mathbb{F}(\epsilon).$$

Therefore, we obtain that:

$$\hat{\mu}_n^+(\mathbb{F}(\epsilon)) \leq \frac{d^4}{\gamma \upsilon} \int_{\mathbb{F}(\epsilon)} (1 - \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^+ \leq \frac{d^4}{\gamma \upsilon} \int (1 - \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^+. \tag{38}$$

Finally, the first part of $i)$ of Theorem 10 is a direct consequence of (36)-(38).
Weight of the far region. We consider γ such that $d^4 \geq \gamma \upsilon$ and we know that in the far region:

$$(1 - \mathcal{P}_m) \mathbb{1}_{\mathbb{F}(\epsilon)} \geq \frac{\gamma \upsilon}{d^4} \mathbb{1}_{\mathbb{F}(\epsilon)}.$$

Thus,

$$\begin{aligned}
D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) &= \int (1-\mathcal{P}_m)\mathrm{d}\hat{\mu}_n^+ + \int (1+\mathcal{P}_m)\mathrm{d}\hat{\mu}_n^- \\
&\geq \int_{\mathbb{F}(\epsilon)} \frac{\gamma\upsilon}{d^4}\mathrm{d}\hat{\mu}_n^+ + \int_{\mathbb{F}(\epsilon)} 1\mathrm{d}\hat{\mu}_n^- \\
&\geq \frac{\gamma\upsilon}{d^4}\int_{\mathbb{F}(\epsilon)} \mathrm{d}\hat{\mu}_n^+ + \int_{\mathbb{F}(\epsilon)} \mathrm{d}\hat{\mu}_n^- \\
&\geq \frac{\gamma\upsilon}{d^4}\left(\int_{\mathbb{F}(\epsilon)} \mathrm{d}\hat{\mu}_n^+ + \int_{\mathbb{F}(\epsilon)} \mathrm{d}\hat{\mu}_n^-\right) \\
&\geq \frac{\gamma\upsilon}{d^4}|\hat{\mu}_n|(\mathbb{F}(\epsilon)).
\end{aligned}$$

We then conclude, using the previous expectation upper bound, that:

$$\mathbb{E}[|\hat{\mu}_n|(\mathbb{F}(\epsilon))] \leq \frac{d^4}{\gamma\upsilon}\frac{3\sqrt{2}}{2}\rho_n\mathcal{C}_m(\varphi,\lambda).$$

Proof of *ii*). Thanks to Theorem 7, we have:

$$1-\mathcal{P}_m(t) \geq \left[\mathcal{C}m^2 \min_{k\in[K]} \|t-t_k\|_2^2 \wedge \frac{\gamma\upsilon}{d^4}\right] \quad \forall t\in\mathbb{R}^d.$$

Then, for any subset $A\subset\mathbb{R}^d$,

$$\begin{aligned}
D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) &\geq \int (1-\mathcal{P}_m)\mathrm{d}\mu_n^+ \\
&\geq \int_A (1-\mathcal{P}_m)\mathrm{d}\mu_n^+, \\
&\geq \left[\mathcal{C}m^2 \min_{t\in A}\min_{k\in[K]} \|t-t_k\|^2 \wedge \frac{\gamma\upsilon}{d^4}\right]\hat{\mu}_n^+(A). \qquad (39)
\end{aligned}$$

Equations (36) and (39) lead to:

$$\left[\mathcal{C}m^2 \min_{t\in A}\min_{k\in[K]} \|t-t_k\|^2 \wedge \frac{\gamma\upsilon}{d^4}\right]\mathbb{E}[\hat{\mu}_n^+(A)] \leq \frac{3\sqrt{2}}{2}\rho_n\mathcal{C}_m(\varphi,\lambda).$$

Then,

$$\mathbb{E}[\hat{\mu}_n^+(A)] \geq \frac{3\sqrt{2}}{2}\rho_n\mathcal{C}_m(\varphi,\lambda)\frac{d^4}{\gamma\upsilon} \Rightarrow \min_{t\in A}\min_{k\in[K]} \|t-t_k\|_2^2 \leq \frac{\gamma\upsilon}{d^4m^2\mathcal{C}}.$$

Proof of *iii*). The idea of this proof is close to the one of [1, Theorem 2.1]. We consider the function $\mathcal{Q}_m^k$ given by Corollary 8 that interpolates 1 at t_k and 0 on the other points of the support of μ^0. From the construction of $\mathcal{Q}_m^k$, we have that:

$$a_k^0 = \int \mathcal{Q}_m^k \mathrm{d}\mu^0.$$

We then use the decomposition:

$$
\begin{aligned}
|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))| &= |a_k^0 - \int \mathcal{Q}_m^k \mathrm{d}\hat{\mu}_n + \int \mathcal{Q}_m^k \mathrm{d}\hat{\mu}_n - \int_{\mathbb{N}_k(\epsilon)} \mathrm{d}\hat{\mu}_n| \\
&\leq \underbrace{|\int \mathcal{Q}_m^k \mathrm{d}(\mu^0 - \hat{\mu}_n)|}_{:=A} + \underbrace{\int_{\mathbb{N}_k(\epsilon)} |\mathcal{Q}_m^k - 1|\mathrm{d}|\hat{\mu}_n|}_{:=B} \\
&\quad + \underbrace{\int_{\mathbb{N}(\epsilon)\setminus\mathbb{N}_k(\epsilon)} |\mathcal{Q}_m^k|\mathrm{d}|\hat{\mu}_n|}_{:=C} + \underbrace{\int_{\mathbb{F}(\epsilon)} |\mathcal{Q}_m^k|\mathrm{d}|\hat{\mu}_n|}_{:=D}. \qquad (40)
\end{aligned}
$$

<u>Study of $B + C + D$.</u> On the set $\mathbb{F}(\epsilon)$, we use that $\mathcal{Q}_m^k \leq 1 - \gamma\frac{\upsilon}{d^4}$ so that:

$$
D \leq \int_{\mathbb{F}(\epsilon)} (1 - \gamma\frac{\upsilon}{d^4})\mathrm{d}|\hat{\mu}_n| \leq \diamond \int_{\mathbb{F}(\epsilon)} (1 - \mathcal{Q}_m^k)\mathrm{d}|\hat{\mu}_n|,
$$

where

$$
\diamond = \frac{\left(1 - \gamma\frac{\upsilon}{d^4}\right)}{\gamma\frac{\upsilon}{d^4}}.
$$

For the term C, we use the upper bound satisfied by $\mathcal{Q}_m^k$ in $\bigcup_{i\neq k} \mathbb{N}_i(\epsilon)$ and obtain that:

$$
\begin{aligned}
\int_{\mathbb{N}(\epsilon)\setminus\mathbb{N}_k(\epsilon)} |\mathcal{Q}_m^k|\mathrm{d}|\hat{\mu}_n| &\leq \widetilde{\mathcal{C}}m^2 \int_{\mathbb{N}(\epsilon)\setminus\mathbb{N}_k(\epsilon)} \min_{i\neq k} \|t - t_i\|_2^2 \mathrm{d}|\hat{\mu}_n|(t) \\
&\leq \frac{\widetilde{\mathcal{C}}}{\mathcal{C}} \int_{\mathbb{N}(\epsilon)\setminus\mathbb{N}_k(\epsilon)} (1 - \mathcal{P}_m)\mathrm{d}|\hat{\mu}_n|.
\end{aligned}
$$

Finally, for B, we use that on the set $\mathbb{N}_k(\epsilon)$, we have $|\mathcal{Q}_m^k - 1| \leq \widetilde{\mathcal{C}}m^2\|t - t_k\|_2^2$. Therefore, we have:

$$
B \leq \frac{\widetilde{\mathcal{C}}}{\mathcal{C}} \int_{\mathbb{N}_k(\epsilon)} (1 - \mathcal{P}_m)\mathrm{d}|\hat{\mu}_n|.
$$

We then conclude that:

$$
\begin{aligned}
B + C + D &\leq \left(\frac{\widetilde{\mathcal{C}}}{\mathcal{C}} \vee \diamond\right) \int_{\mathbb{R}^d} (1 - \mathcal{P}_m)(t)\mathrm{d}|\hat{\mu}_n|(t) \\
&\leq \left(\frac{\widetilde{\mathcal{C}}}{\mathcal{C}} \vee \diamond\right) \left[\int_{\mathbb{R}^d} (1 - \mathcal{P}_m)(t)\mathrm{d}\hat{\mu}_n^+(t) + \int_{\mathbb{R}^d} (1 + \mathcal{P}_m)(t)\mathrm{d}\hat{\mu}_n^-(t)\right] \\
&\leq \left(\frac{\widetilde{\mathcal{C}}}{\mathcal{C}} \vee \diamond\right) D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0). \qquad (41)
\end{aligned}
$$

<u>Study of A.</u> We use that $\mathcal{Q}_m^k$ may be written as:

$$
\mathcal{Q}_m^k = \Phi c_{k,m}, \quad \text{where} \quad c_{k,m} \in \mathbb{L}.
$$

Since Φ is self-adjoint in L^2, we shall write that:

$$\begin{aligned} A = |\int \mathcal{Q}_m^k \mathrm{d}(\mu^0 - \hat{\mu}_n)| &= |\langle \mathcal{Q}_m^k, \hat{\mu}_n - \mu^0 \rangle_{L^2}| \\ &= |\langle c_{k,m}, \Phi(\hat{\mu}_n - \mu^0) \rangle_{L^2}| \\ &= |\langle c_{k,m}, L \circ \Phi \hat{\mu}_n - L\hat{f}_n + \Gamma_n \rangle_{\mathbb{L}}| \\ &\leq \|c_{k,m}\|_{\mathbb{L}} [\|L \circ \Phi \hat{\mu}_n - L\hat{f}_n\|_{\mathbb{L}} + \|\Gamma_n\|_{\mathbb{L}}], \end{aligned}$$

where we used the Cauchy-Schwarz inequality and the triangle inequality in the last line. We then use (29) and obtain:

$$\|L\hat{f}_n - L \circ \Phi \hat{\mu}_n\|_{\mathbb{L}}^2 + \kappa D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) + \kappa \langle c_{0,m}, L \circ \Phi \hat{\mu}_n - L\hat{f}_n \rangle_{\mathbb{L}} + \kappa \langle c_{0,m}, \Gamma_n \rangle_{\mathbb{L}} \leq \|\Gamma_n\|_{\mathbb{L}}^2 .$$

Since we have obtained the positiveness of the Bregman divergence, we then conclude that:

$$\|L\hat{f}_n - L \circ \Phi \hat{\mu}_n\|_{\mathbb{L}}^2 + \kappa \langle c_{0,m}, L \circ \Phi \hat{\mu}_n - L\hat{f}_n \rangle_{\mathbb{L}} \leq \|\Gamma_n\|_{\mathbb{L}}^2 - \kappa \langle c_{0,m}, \Gamma_n \rangle_{\mathbb{L}} .$$

The Cauchy-Schwarz inequality yields:

$$\|L\hat{f}_n - L \circ \Phi \hat{\mu}_n\|_{\mathbb{L}}^2 - \kappa \|c_{0,m}\|_{\mathbb{L}} \|L \circ \Phi \hat{\mu}_n - L\hat{f}_n\|_{\mathbb{L}} \leq \|\Gamma_n\|_{\mathbb{L}}^2 + \kappa \|c_{0,m}\|_{\mathbb{L}} \|\Gamma_n\|_{\mathbb{L}} .$$

This inequality holds for any value of κ and we choose:

$$\kappa = \frac{\|L\hat{f}_n - L \circ \Phi \hat{\mu}_n\|_{\mathbb{L}}}{2\|c_{0,m}\|_{\mathbb{L}}}.$$

Using this value of κ, we then obtain:

$$\frac{\|L\hat{f}_n - L \circ \Phi \hat{\mu}_n\|_{\mathbb{L}}^2}{2} \leq \|\Gamma_n\|_{\mathbb{L}}^2 + \|\Gamma_n\|_{\mathbb{L}} \|L\hat{f}_n - L \circ \Phi \hat{\mu}_n\|_{\mathbb{L}} .$$

Now, we define $\square_n = \|L\hat{f}_n - L \circ \Phi \hat{\mu}_n\|_{\mathbb{L}} \|\Gamma_n\|_{\mathbb{L}}^{-1}$ and remark that:

$$\frac{\square_n^2}{2} \leq 1 + \square_n.$$

This last inequality implies that $\square_n \leq 1 + \sqrt{3}$, which leads to:

$$\|L\hat{f}_n - L \circ \Phi \hat{\mu}_n\|_{\mathbb{L}} \leq (1 + \sqrt{3}) \|\Gamma_n\|_{\mathbb{L}}.$$

We then come back to A and write that:

$$A \leq (2 + \sqrt{3}) \|c_{k,m}\|_{\mathbb{L}} \|\Gamma_n\|_{\mathbb{L}}. \tag{42}$$

Final bound. We use Equations (42) and (41) in the decomposition given in Equation (40) and obtain that:

$$\mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right] \lesssim \rho_n \left(\|c_{k,m}\|_{\mathbb{L}} + \|c_{0,m}\|_{\mathbb{L}} \right).$$

Finally, we conclude the proof using $iv)$ of Corollary 8 and $ii)$ of Theorem 10:

$$\mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right] \lesssim \rho_n \frac{K^2 m^{-d/2}}{\sqrt{\inf_{\|t\|_\infty \leq 4m} \left\{ \sigma^2(t) \mathcal{F}[\lambda](t) \right\}}}.$$

$\square$

6.6. Support stability

We follow the ideas of [13] for the proof of Theorem 11. Consider the convex program

$$\inf\Big\{\|\mu\|_1 \;:\; \mu\in\mathcal{M}(\mathbb{R}^d;\mathbb{R}) \text{ s.t. } L\circ\Phi\mu = L\circ\Phi\mu^0\Big\} \tag{$\mathbf{P}_0$}$$

whose Lagrangian expression is, for all $\mu\in\mathcal{M}(\mathbb{R}^d,\mathbb{R})$, $c\in\mathbb{L}$,

$$\begin{aligned}\mathcal{L}(\mu,c) &= \|\mu\|_1 + \langle c, L\circ\Phi(\mu^0-\mu)\rangle_{\mathbb{L}}\,,\\ &= \|\mu\|_1 + \langle c, L\circ\Phi\mu^0\rangle_{\mathbb{L}} - \int \Phi c\,\mathrm{d}\mu\,,\\ &= \|\mu\|_1 - \int \Phi c\,\mathrm{d}\mu + \int \Phi c\,\mathrm{d}\mu^0\,,\end{aligned}$$

using (27) in the last equation. Now, Equation (28) yields that the dual program is:

$$\sup\Big\{\int_{\mathbb{R}^d}\Phi c\,\mathrm{d}\mu^0 \;:\; c\in\mathbb{L} \text{ s.t. } \|\Phi c\|_\infty\le 1\Big\}\,.$$

Note also that the objective function of the dual program satisfies:

$$\int_{\mathbb{R}^d}\Phi c\,\mathrm{d}\mu^0 = \langle c, L\circ\Phi\mu^0\rangle_{\mathbb{L}}\,,$$

which gives the following equivalent formulation of the dual:

$$\sup\Big\{\langle c, L\circ\Phi\mu^0\rangle_{\mathbb{L}} \;:\; c\in\mathbb{L} \text{ s.t. } \|\Phi c\|_\infty\le 1\Big\}\,. \tag{$\mathbf{D}_0$}$$

Note that the dual certificate $\mathcal{P}_m$ exists, then we know that μ^0 is a solution to ($\mathbf{P}_0$) by Theorem 3. As in Section 6.3, we use the Slater condition to prove that there is no duality gap: we remark that a feasible point c exists in the interior of the set $\{\|\Phi c\|_\infty\le 1\}$. Now, the generalized Slater condition shall be used (see *e.g.* [26]). We get that any solution c to ($\mathbf{D}_0$) satisfies that Φc is a sub-gradient of the total variation norm at point μ^0. Under condition (NDB), we know that $\mathcal{P}_0 := \Phi c_0$ is a solution to ($\mathbf{D}_0$).

Consider also the following convex program:

$$\inf_{\mu\in\mathcal{M}(\mathbb{R}^d,\mathbb{R})}\Big\{\frac{1}{2}\|L\circ\Phi\mu^0 - L\circ\Phi\mu\|_{\mathbb{L}}^2 + \kappa\|\mu\|_1\Big\}\,, \tag{$\mathbf{P}_\kappa(\Phi\mu^0)$}$$

which is the same as the one used in Section 6.3 and Theorem 6, exchanging $L\hat{f}_n$ by $L\circ\Phi\mu^0$ and dropping the constraint $\int_{\mathbb{R}^d}\mathrm{d}\mu = 1$. Following the arguments used in Section 6.3, one may prove that there is no duality gap and the dual program is given by:

$$\frac{\|L\circ\Phi\mu^0\|_{\mathbb{L}}^2}{2} - \kappa\inf\Big\{\frac{\kappa}{2}\Big\|\frac{L\circ\Phi\mu^0}{k} - c\Big\|_{\mathbb{L}}^2 \;:\; c \text{ s.t. } \|\Phi c\|_\infty\le 1\Big\}. \tag{$\mathbf{D}_\kappa(\Phi\mu^0)$}$$

We denote by c_κ the solution to ($\mathbf{D}_\kappa(\Phi\mu^0)$) (unicity can be proven by (26)) and by $\mathcal{P}_\kappa := \Phi c_\kappa$ the dual polynomial. Its gradient is denoted by $\nabla\mathcal{P}_\kappa$, and its Hessian is denoted by $\nabla^2\mathcal{P}_\kappa$. We first state the next lemma.

Lemma 14. *If c_0 exists, then $\|c_\kappa - c_0\|_{\mathbb{L}} \to 0$, $\nabla \mathcal{P}_\kappa \to \nabla \mathcal{P}_0$ uniformly, and $\nabla^2 \mathcal{P}_\kappa \to \nabla^2 \mathcal{P}_0$ uniformly as $\kappa \to 0$.*

Proof. Since c_κ is a solution to ($\mathbf{D}_\kappa(\Phi\mu^0)$), it holds that:

$$\frac{\kappa}{2}\Big\|\frac{L\circ\Phi\mu^0}{k} - c_\kappa\Big\|_{\mathbb{L}}^2 \leq \frac{\kappa}{2}\Big\|\frac{L\circ\Phi\mu^0}{k} - c_0\Big\|_{\mathbb{L}}^2,$$

leading to:

$$\langle c_\kappa, L\circ\Phi\mu^0\rangle_{\mathbb{L}} - \frac{\kappa}{2}\|c_\kappa\|_{\mathbb{L}}^2 \geq \langle c_0, L\circ\Phi\mu^0\rangle_{\mathbb{L}} - \frac{\kappa}{2}\|c_0\|_{\mathbb{L}}^2\,, \tag{43}$$

and c_0 being a solution to ($\mathbf{D}_0$) implies that:

$$\langle c_\kappa, L\circ\Phi\mu^0\rangle_{\mathbb{L}} \leq \langle c_0, L\circ\Phi\mu^0\rangle_{\mathbb{L}}\,.$$

We deduce that $\|c_\kappa\|_{\mathbb{L}} \leq \|c_0\|_{\mathbb{L}}$. Closed unit balls of Hilbert spaces being weakly sequentially compact we deduce that given $\kappa_n \to 0$, one shall extract a subsequence such that c_{κ_n} weakly converge to some $c^\star \in \mathbb{L}$. Taking the limit as $\kappa \to 0$ in (43), we obtain that:

$$\langle c^\star, L\circ\Phi\mu^0\rangle_{\mathbb{L}} \geq \langle c_0, L\circ\Phi\mu^0\rangle_{\mathbb{L}}\,.$$

Note that Φc_{κ_n} converges weakly to $\Phi c^\star$ so that:

$$\|\Phi c^\star\|_\infty \leq \liminf_n \|\Phi c_{\kappa_n}\|_\infty \leq 1$$

We deduce that $c^\star \in \mathbb{L}$ is a solution to ($\mathbf{D}_0$) and hence:

$$\|\Phi c^\star\|_\infty \leq 1 \text{ and } (\Phi c^\star)(t_k) = 1\,,\ k \in [K]\,.$$

Furthermore, $c^\star$ is the solution of minimal norm since:

$$\|c^\star\|_{\mathbb{L}} \leq \liminf_n \|c_{\kappa_n}\|_{\mathbb{L}} \leq \|c^0\|_{\mathbb{L}}\,.$$

The solution of minimal norm is unique by strict coercivity of the norm $\|\cdot\|_{\mathbb{L}}$, see (26). We deduce that $c^\star = c_0$, $\|c_{\kappa_n}\|_{\mathbb{L}} \to \|c_0\|_{\mathbb{L}}$, and $c_{\kappa_n} \to c_0$ strongly in $\mathbb{L}$. Note that it implies that $\lim_{\kappa\to 0}\|c_\kappa - c_0\|_{\mathbb{L}} = 0$, since otherwise one can extract a subsequence c_{κ_n} such that $\|c_{\kappa_n} - c_0\|_{\mathbb{L}} > \varepsilon$, and by the above argument, one can extract a sequence such that $c_{\kappa_n} \to c_0$.
Now, the Cauchy-Schwarz inequality yields:

$$\forall t \in \mathbb{R}^d, \quad \|\nabla^2\mathcal{P}_\kappa(t) - \nabla^2\mathcal{P}_0(t)\|_\infty \leq (\sup_{i,j}\|\partial^2\varphi/(\partial x_i \partial x_j)\|_{\mathbb{L}})\|c_\kappa - c_0\|_{\mathbb{L}}\,,$$

which proves the uniform convergence. The same computation gives the uniform convergence of the functions and their gradients. □

We denote by $c_{\kappa,n}$ the dual solution of ($\mathbf{P}_\kappa$) when we drop the constraint $\int \mathrm{d}\mu = 1$, namely:

$$\frac{\|L\hat{f}_n\|_{\mathbb{L}}^2}{2} - \kappa \inf\Big\{\frac{\kappa}{2}\Big\|\frac{L\hat{f}_n}{\kappa} - c\Big\|_{\mathbb{L}}^2 \ :\ c \text{ s.t. } \|\Phi c\|_\infty \leq 1\Big\} \tag{$\mathbf{D}_\kappa(\hat{f}_n)$}$$

and $\mathcal{P}_{\kappa,n} = \Phi c_{\kappa,n}$. The primal solution is denoted by $\hat{\mu}_n$.

Lemma 15. *If κ and $\|\Gamma_n\|_{\mathbb{L}}/\kappa$ are sufficiently small, any solution $\hat{\mu}_n$ has support of size $\hat{K} = K$ with one and only spike in each near region $\mathbb{N}_k(r)$ for $k \in [K]$.*

Proof. Note that ($\mathbf{D}_\kappa(\hat{f}_n)$) and ($\mathbf{D}_\kappa(\Phi\mu^0)$) are projection onto a closed convex set. We deduce that

$$\|c_{\kappa,n} - c_\kappa\|_{\mathbb{L}} \leq \frac{\|\Gamma_n\|_{\mathbb{L}}}{\kappa}\,,$$

and that $\|\nabla^2\mathcal{P}_\kappa - \nabla^2\mathcal{P}_{\kappa,n}\|_\infty = \mathcal{O}(\frac{\|\Gamma_n\|_{\mathbb{L}}}{\kappa})$ (the same result holds for the functions and their gradients). Under (NDB), we know that there exists $0 < q < 1$, $r > 0$ and $\rho > 0$ such that $\nabla^2\mathcal{P}_0 \prec \rho \mathrm{Id}_d$ on $\mathbb{N}(r)$ and $|\mathcal{P}_0| < 1 - q$ on $\mathbb{F}(r)$. We deduce that, for sufficiently small κ and $\|\Gamma_n\|_{\mathbb{L}}/\kappa$, $\mathcal{P}_{\kappa,n}$ is such that $\nabla^2\mathcal{P}_{\kappa,n} \prec (\rho/2)\mathrm{Id}_d$ on $\mathbb{N}(r)$ and $|\mathcal{P}_{\kappa,n}| < 1 - q/2$ on $\mathbb{F}(r)$. We deduce that at most 1 point in each $\mathbb{N}_k(r)$ is such that $\mathcal{P}_{\kappa,n}(\hat{t}_k) = 1$.

But, since μ^0 is the unique solution of ($\mathbf{P}_0$) (see Theorem 3), we deduce that $\hat{\mu}_n$ converges to μ^0 in the weak-*topology as κ and $\|\Gamma_n\|_{\mathbb{L}}/\kappa$ go to zero. Hence, it holds that $\hat{\mu}_n(\mathbb{N}_k(r)) \to \mu^0(\mathbb{N}_k(r)) = a_k^0$. In particular, $\hat{\mu}_n$ has one spike in $\mathbb{N}_k(r)$. □

We set $\kappa = \kappa_n$ such that $\|\Gamma_n\|_{\mathbb{L}} = o_{\mathbb{P}}(\kappa_n)$ and $\kappa_n = o(1)$. In this case, the requirements of the aforementioned lemma are met and, collecting the pieces, we have that a sequence $(p_n)_{n\geq 1}$ exist such that $\lim_{n\to+\infty} p_n = 0$ and for which the desired result holds.

References

[1] J.-M. Azaïs, Y. De Castro, and F. Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.

[2] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, New York, 2004.

[4] B. N. Bhaskar, G. Tang, and B. Recht. Atomic norm denoising with applications to line spectral estimation. Technical report, University of Wisconsin-Madison, 2012.

[5] L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *The Annals of Statistics*, 34(3):1204–1232, 2006.

[6] F. Bunea, A. B. Tsybakov, M. H. Wegkamp, and A. Barbu. Spades and mixture models. *The Annals of Statistics*, 38(4):2525–2558, 2010.

[7] C. Butucea and P. Vandekerkhove. Semiparametric mixtures of symmetric distributions. *Scand. J. Stat.*, 41(1):227–239, 2014.

[8] E. J. Candès and C. Fernandez-Granda. Towards a Mathematical Theory of Super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

[9] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9(Jul):1615–1646, 2008.

[10] Y. De Castro and F. Gamboa. Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

[11] Y. De Castro, F. Gamboa, D. Henrion, and J.-B. Lasserre. Exact solutions to Super Resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630, 2017.

[12] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies. The sliding frank-wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 2019.

[13] V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, pages 1–41, 2015.

[14] R. Dwivedi, N. Ho, K. Khamaru, M. I. Jordan, M. J. Wainwright, and B. Yu. Singularity, Misspecification, and the Convergence Rate of EM. *arXiv preprint arXiv:1810.00828*, 2018.

[15] C. Elvira, R. Gribonval, C. Soussen, and C. Herzet. When does OMP achieves support recovery with continuous dictionaries? *arXiv preprint arXiv:1904.06311*, 2019.

[16] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19:1257–1272, 1991.

[17] S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

[18] S. Gadat, J. Kahn, C. Marteau, and C. Maugis-Rabusseau. Parameter recovery in two-component contamination mixtures: the l2 strategy. *To appear in Annales de l'Institut Henri Poincaré*, 2018.

[19] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46:2844–2870, 2018.

[20] L. Le Cam. Maximum likelihood: an introduction. *International Statistical Review*, 58(2):153–171, 1990.

[21] C. Maugis-Rabusseau and B. Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.*, 15:41–68, 2011.

[22] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley series in Probability and Statistics, 2000.

[23] A. Meister. *Deconvolution problems in nonparametric statistics*, volume 193 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 2009.

[24] X. Nguyen et al. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[25] C. Poon, N. Keriven, and G. Peyré. Support localization and the Fisher Metric for off-the-grid Sparse Regularization. *arXiv preprint arXiv:1810.03340*, 2018.

[26] R. T. Rockafellar. *Conjugate duality and optimization*, volume 16. Siam, 1974.

[27] W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, Inc., New York, 1991.

[28] L. Schwartz and I. de mathématique (Strasbourg). *Théorie des distributions*, volume 2. Hermann Paris, 1957.

[29] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[30] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *Information Theory, IEEE Transactions on*, 59(11):7465–7490, 2013.

[31] A. W. van der Vaart. Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.*, 24(2):862–878, 1996.

[32] G. Wahba. *Spline Models for Observational Data*. SIAM publisher: Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[33] C. J. Wu et al. On the convergence properties of the em algorithm. *The Annals of statistics*, 11(1):95–103, 1983.

[34] J. Xu, D. J. Hsu, and A. Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

Appendix A: Proofs

A.1. Convolution in the RKHS

Proof of Proposition 4. Consider $\mathcal{A} : f \longmapsto x \mapsto \int \ell(x-y)f(y)dy$, $\mathcal{A}$ is a self-adjoint operator. We denote by $(w_k)_{k\geq 1}$ the non-negative eigenvalues of $\mathcal{A}$ and $(\psi_k)_{k\geq 1}$ the associated eigenvectors. We shall remark that the following equality holds:

$$\ell(x,y) = \lambda(x-y) = \sum_{k\geq 1} w_k \psi_k(x)\psi_k(y),$$

while $\mathbb{L}$ corresponds to the next Hilbert space

$$\mathbb{L} = \Big\{ f = \sum_{k\geq 1} c_k(f)\psi_k \, : \, \sum_{k\geq 1} \frac{c_k(f)^2}{w_k} < +\infty \Big\} \qquad \text{and} \qquad < f, g >_{\mathbb{L}} = \sum_{k\geq 1} \frac{c_k(f)c_k(g)}{w_k}.$$

We now consider a non-negative measure ν and we remark that

$$\begin{aligned} L\nu(x) &= \lambda \star \nu(x) \\ &= \int \lambda(x-y)\nu(y)dy \\ &= \int \sum_{k\geq 1} w_k\psi_k(x)\psi_k(y)\nu(y)dy \\ &= \sum_{k\geq 1} \Big[w_k \int \psi_k(y)\nu(y)dy \Big] \psi_k(x). \end{aligned}$$

We observe that the coefficients of $L\nu$ are $c_k(L\nu) = w_k \int \psi_k(y)\nu(y)dy$. We shall remark that

$$\|L\nu\|_{\mathbb{L}}^2 = \sum_{k\geq 1} \frac{w_k^2 \left[\int \psi_k(y)\nu(y)dy \right]^2}{w_k} = \sum_{k\geq 1} w_k \left[\int \psi_k(y)\nu(y)dy \right]^2.$$

The Jensen inequality yields

$$\|L\nu\|_{\mathbb{L}}^2 \leq \sum_{k\geq 1} w_k \int \psi_k^2(y)\nu(y)dy = \int \sum_{k\geq 1} w_k\psi_k(y)^2\nu(y)dy,$$

where the last equality comes from the Tonelli Theorem. We then observe that

$$\|L\nu\|_{\mathbb{L}}^2 \leq \int \ell(y,y)\nu(y)dy = \lambda(0) < +\infty,$$

giving the result. □

A.2. Computation of the data-fidelity terms

Proof of Proposition 5. Recall that $\Phi\mu \in \mathcal{C}_0(\mathbb{R}^d, \mathbb{R}) \cap L^1(\mathbb{R}^d)$. Now, given $f \in \mathcal{C}_0(\mathbb{R}^d, \mathbb{R}) \cap L^1(\mathbb{R}^d)$, one can consider the measure μ with signed density function f and we may define:

$$\forall f \in \mathcal{C}_0(\mathbb{R}^d, \mathbb{R}) \cap L^1(\mathbb{R}^d), \quad Lf := \lambda \star f = \int_{\mathbb{R}^d} \lambda(\cdot - t)f(t)\mathrm{d}t.$$

The embedding L allows to compare $\hat{f}_n$ with $\Phi\mu$ in $\mathcal{C}_0(\mathbb{R}^d,\mathbb{R})\cap L^1(\mathbb{R}^d)$. One has:

$$
\begin{aligned}
&\|Lf^0 - Lf\|_{\mathbb{L}}^2 - \|Lf^0\|_{\mathbb{L}}^2 = -2\langle Lf^0, Lf\rangle_{\mathbb{L}} + \|Lf\|_{\mathbb{L}}^2\\
&\quad = -2\langle Lf^0, \int_{\mathbb{R}^d} \ell(\cdot,t)f(t)\mathrm{d}t\rangle_{\mathbb{L}} + \|Lf\|_{\mathbb{L}}^2\\
&\quad = -2\int_{\mathbb{R}^d} \langle Lf^0, \ell(\cdot,t)\rangle_{\mathbb{L}} f(t)\mathrm{d}t + \|Lf\|_{\mathbb{L}}^2\\
&\quad = -2\int_{\mathbb{R}^d} Lf^0(t)f(t)\mathrm{d}t + \|Lf\|_{\mathbb{L}}^2\\
&\quad = \int_{\mathbb{R}^d}\Big(-2\int_{\mathbb{R}^d}\lambda(t-x)f^0(x)\mathrm{d}x\Big)f(t)\mathrm{d}t + \int_{\mathbb{R}^d\times\mathbb{R}^d}\lambda(x-y)f(x)f(y)\mathrm{d}x\mathrm{d}y\,.
\end{aligned}
$$

Replacing f^0, which is unknown, by the empirical measure $\hat{f}_n$ in the previous equation leads to the following criterion:

$$
\mathrm{C}_\lambda(f) := \int_{\mathbb{R}^d}\big[-\frac{2}{n}\sum_{i=1}^n \lambda(t-X_i)\big]f(t)\mathrm{d}t + \int_{\mathbb{R}^d\times\mathbb{R}^d}\lambda(x-y)f(x)f(y)\mathrm{d}x\mathrm{d}y\,.
$$

In particular, for all $\mu\in\mathcal{M}(\mathbb{R}^d,\mathbb{R})$ note that $\Phi\mu\in\mathcal{C}_0(\mathbb{R}^d,\mathbb{R})\cap L^1(\mathbb{R}^d)$ and introduce the criterion:

$$
\mathrm{C}_\lambda(\Phi\mu) = \int_{\mathbb{R}^d}\big[-\frac{2}{n}\sum_{i=1}^n \lambda(t-X_i)\big](\Phi\mu)(t)\mathrm{d}t + \int_{\mathbb{R}^d\times\mathbb{R}^d}\lambda(x-y)(\Phi\mu)(x)(\Phi\mu)(y)\mathrm{d}x\mathrm{d}y\,,
$$

which will be investigated in this paper. Note that it holds

$$
\begin{aligned}
&\|L\hat{f}_n - L\circ\Phi\mu\|_{\mathbb{L}}^2 - \|L\hat{f}_n\|_{\mathbb{L}}^2\\
&= \int_{\mathbb{R}^d}\big[-\frac{2}{n}\sum_{i=1}^n \lambda(t-X_i)\big](\Phi\mu)(t)\mathrm{d}t + \int_{\mathbb{R}^d\times\mathbb{R}^d}\lambda(x-y)(\Phi\mu)(x)(\Phi\mu)(y)\mathrm{d}x\mathrm{d}y\,,
\end{aligned}
$$

as claimed. □

Appendix B: Construction of a dual certificate (proof of Theorem 7)

For a given set of points $S^0=\{t_1,\ldots,t_K\}$, a vector α and a matrix β, we recall that $\Delta := \min_{k\neq\ell}\|t_k-t_\ell\|_2$ and

$$
p_m^{\alpha,\beta}(t) = \sum_{k=1}^K \{\alpha_k\psi_m(t-t_k) + \langle\beta_k, \nabla\psi_m(t-t_k)\rangle\}\,,\quad \forall t\in\mathbb{R}^d, \tag{44}
$$

For the sake of convenience, we omit the dependency in α and β of the previous function and simply denote it by p_m. We are interested in the existence of a set of coefficients (α,β) such that p_m defined in (44) satisfies the two interpolation conditions:

$$
\forall k\in\{1,\ldots,K\}\qquad p_m(t_k)=1 \quad\text{and}\quad \nabla p_m(t_k)=0. \tag{45}
$$

We establish the first result.

Proposition 16. *If m is chosen such that $m \gtrsim \frac{K^{1/3}d}{\Delta^{4/3}}$ then (α, β) exists such that (45) holds and:*

- *$i)$ The supremum norm is upper bounded by:*

$$\|\alpha - \mathbf{1}\|_\infty \vee \left(\sup_{1 \leq k \leq K} \|\beta_k\|_\infty \right) \lesssim \frac{Kd^3}{m^3 \Delta^4}.$$

- *$ii)$ The Euclidean norm is upper bounded by:*

$$\|\alpha - \mathbf{1}\|_2 \vee \sqrt{\sum_{k=1}^{K} \|\beta_k\|_2^2} \lesssim \frac{K^{3/2}d^3}{m^3 \Delta^4}.$$

Proof. The proofs of $i)$ and $ii)$ are divided into three steps.
Step 1: Matricial formulation of (45). The certificate p_m should satisfy the following properties:

$$\forall i \in \{1, \ldots, K\} : \left\{ \begin{array}{c} p_m(t_i) = 1 \\ \nabla p_m(t_i) = 0 \end{array} \right.$$

$$\iff \left\{ \begin{array}{l} \alpha_i + \sum_{k \neq i} \alpha_k \psi_m(t_i - t_k) + \sum_{k=1}^{K} \sum_{v=1}^{d} \beta_k^v \partial_v(\psi_m)(t_i - t_k) = 1 \\ \sum_{k=1}^{K} \alpha_k \partial_u(\psi_m)(t_i - t_k) + \sum_{k=1}^{K} \sum_{v=1}^{d} \beta_k^v \partial_{u,v}^2(\psi_m)(t_i - t_k) = 0 \end{array} \right. \quad \forall u \in [d], \forall i \in [K].$$

We can organize the above equations to obtain a linear system of $K(d+1)$ equations with $K(d+1)$ parameters. We denote them by $\alpha = (\alpha_1, \ldots, \alpha_K)^T$ and $\beta = (\beta_1^1, \ldots, \beta_1^d, \beta_2^1, \ldots, \beta_2^d, \ldots, \beta_K^1, \ldots, \beta_K^d)^T$. The above equations can be rewritten as:

$$(M_0 + H) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix},$$

with:

$$M_0 = \begin{pmatrix} I_K & 0 \\ 0 & -\frac{4}{3}m^2 I_{K \times d} \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} A_m & D_m \\ D_m^T & B_m \end{pmatrix},$$

where the matrix A_m acts on the coefficients α as:

$$(A_m)_{i,j} = \mathbb{1}_{i \neq k} \psi_m(t_i - t_k),$$

while D_m describes the effect of the partial derivatives of ψ_m on α and β as:

$$(D_m)_{i,(k,v)} = \partial_v \psi_m(t_i - t_k).$$

Finally, the squared matrix B_m is given by:

$$\forall (i,u),(k,v) : \qquad (B_m)_{(i,u),(k,v)} = \mathbb{1}_{(i,u) \neq (k,v)} \partial_{u,v}^2(\psi_m)(t_i - t_k).$$

Then, we remark that:

$$M_0^{-1} = \begin{pmatrix} I_K & 0 \\ 0 & -\frac{3}{4m^2} I_{K \times d} \end{pmatrix},$$

and

$$M_0 + H = M_0(I_{K(d+1)} + \tilde{H}),$$

where:

$$\tilde{H} = M_0^{-1} H = \begin{pmatrix} A_m & D_m \\ -\frac{3}{4m^2} D_m^T & -\frac{3}{4m^2} B_m \end{pmatrix}.$$

Below, we will use that when $M_0 + H$ is invertible, then

$$\begin{aligned}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} &= (M_0 + H)^{-1} \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix} = M_0^{-1} \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix} + \left((\mathbf{I} + \tilde{H})^{-1} - \mathbf{I} \right) M_0^{-1} \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix} + (\mathbf{I} + \tilde{H})^{-1} - \mathbf{I}) \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix},\end{aligned}$$

where $\mathbf{I} = I_{K(d+1)}$. Using the definition of $\|.\|_\infty$ and of the operator norm, we obtain that:

$$\|\alpha - \mathbf{1}\|_\infty \vee \|\beta\|_\infty = \left\| \begin{pmatrix} \alpha - \mathbf{1} \\ \beta \end{pmatrix} \right\|_\infty \leq \left\| (\mathbf{I} + \tilde{H})^{-1} - \mathbf{I}) \right\|_\infty \left\| \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix} \right\|_\infty = \left\| (\mathbf{I} + \tilde{H})^{-1} - \mathbf{I}) \right\|_\infty.$$

In a similar way, we also have:

$$\left\| \begin{pmatrix} \alpha - \mathbf{1} \\ \beta \end{pmatrix} \right\|_2 \leq \left\| (\mathbf{I} + \tilde{H})^{-1} - \mathbf{I}) \right\|_2 \left\| \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix} \right\|_2 = \sqrt{K} \left\| (\mathbf{I} + \tilde{H})^{-1} - \mathbf{I}) \right\|_2.$$

Hence, since for $\tilde{H}$ small enough (*i.e.* for a sufficiently small norm) we have

$$(\mathbf{I} + \tilde{H})^{-1} = \sum_{k \geq 0} (-\tilde{H})^k = \mathbf{I} + \sum_{k \geq 1} (-\tilde{H})^k,$$

we deduce that for $\tilde{H}$ small enough:

$$\|\alpha - \mathbf{1}\|_\infty \vee \|\beta\|_\infty \leq \left\| \sum_{k \geq 1} (-\tilde{H})^k \right\|_\infty \leq \left\| \tilde{H} \right\|_\infty \sum_{k \geq 0} \left\| \tilde{H} \right\|_\infty^k, \tag{46}$$

while

$$\left\| \begin{pmatrix} \alpha - \mathbf{1} \\ \beta \end{pmatrix} \right\|_2 \leq \sqrt{K} \left\| \tilde{H} \right\|_2 \sum_{k \geq 0} \left\| \tilde{H} \right\|_2^k. \tag{47}$$

Step 2: Computation of $\|\tilde{H}\|_\infty$ and proof of i). Thanks to the definition of the ∞-norm,

$$\begin{aligned}\|\tilde{H}\|_\infty &= \max_{1 \leq i \leq K} \left\{ \sum_{j=1}^K |(A_m)_{i,j}| + \sum_{(k,v)} |(D_m)_{i,(k,v)}| \right\} \\ &\vee \frac{3 \max_{(i,u)} \left\{ \sum_{j=1}^K |(D_m)_{j,(i,u)}| + \sum_{(k,v)} |(B_m)|_{(i,u),(k,v)}| \right\}}{4m^2} \\ &\leq (\|A_m\|_\infty + \|D_m\|_\infty) \vee \frac{3 \left(\|D_m^T\|_\infty + \|B_m^T\|_\infty \right)}{4m^2}.\end{aligned}$$

- In a first time, let $i \in \{1, \dots, K\}$ and consider an upper bound of $\|A_m\|_\infty$:

$$\sum_{j=1}^K |(A_m)_{ij}| = \sum_{j=1}^K |\psi_m(t_i - t_j)| \mathbb{1}_{i \neq j}.$$

Applying $i)$ of Lemma 19, we obtain that:

$$\|A_m\|_\infty \lesssim \frac{Kd^2}{m^4\Delta^4}. \tag{48}$$

With a similar argument, we know that:

$$\sum_{(k,v)} |(D_m)_{i,(k,v)}| = \sum_{k=1}^{K}\sum_{v=1}^{d} |\partial_v(\psi_m)(t_i - t_k)|.$$

Then, we apply $ii)$ of Lemma 19 and use that when $i = k$, then $\nabla\psi_m(t_i - t_k) = \nabla\psi_m(0) = 0$, we deduce that:

$$\|D_m\|_\infty \lesssim \frac{Kd^3}{m^3\Delta^4}. \tag{49}$$

• Following the same ideas, for any pair (i,u) with $i \in \{1,\ldots,K\}$ and $u \in \{1,\ldots,d\}$, we have:

$$\frac{3\sum_{j=1}^{K} |(D_m)_{j,(i,u)}|}{4m^2} = \frac{3}{4m^2}\sum_{j=1}^{K} |\partial_u(\psi_m)(t_i - t_j)|.$$

Again, $ii)$ of Lemma 19 yields:

$$\frac{3\|D_m^T\|_\infty}{4m^2} \lesssim \frac{K}{m^2}\frac{d^2}{m^3\Delta^4} \lesssim \frac{Kd^2}{m^5\Delta^4}. \tag{50}$$

Finally, we can write that:

$$\begin{aligned}\frac{3\sum_{(k,v)} |(B_m)|_{(i,u),(k,v)}|}{4m^2} &= \frac{3}{4m^2}\sum_{k=1}^{K}\sum_{v=1}^{d} |\partial^2_{u,v}(\psi_m)(t_i - t_k)|\mathbb{1}_{(i,u)\neq(k,v)} \\ &\lesssim m^{-2}\left[\sum_{k\neq i}\sum_{v=1}^{d} |\partial^2_{u,v}(\psi_m)(t_i - t_k)| + \sum_{v\neq u} |\partial^2_{u,v}(\psi_m)(0)|\right].\end{aligned}$$

Now, using that when $u \neq v, \partial_{u,v}(\psi_m)(0) = 0$ and $iii)$ of Lemma 19, we obtain:

$$\frac{3\|B_m^T\|_\infty}{4m^2} \lesssim \frac{Kd^3}{m^4\Delta^4}. \tag{51}$$

Gathering (48), (49), (50) and (51), we deduce that:

$$\|\tilde{H}\|_\infty \lesssim \frac{Kd^3}{m^3\Delta^4}, \tag{52}$$

and we can choose m large enough such that $\|\tilde{H}\|_\infty \leq 1/2$, it is possible as soon as:

$$m \gtrsim \frac{K^{1/3}d}{\Delta^{4/3}}.$$

In this case, the matrix $M_0 + H$ is invertible and we obtain $i)$, *i.e.*, when $m \gtrsim \frac{K^{1/3}d}{\Delta^{4/3}}$:

$$\|\alpha - \mathbf{1}\|_\infty \vee \|\beta\|_\infty \lesssim \frac{Kd^3}{m^3\Delta^4}. \tag{53}$$

Step 3: Computation of $\|\tilde{H}\|_2$ and proof of $ii)$.
We use the definition of the $\|.\|_2$ matricial norm related to the spectral radius ρ:

$$\begin{aligned}\|\tilde{H}\|_2 &= \sqrt{\rho(\tilde{H}^T\tilde{H})}\\ &\leq \sqrt{\|\tilde{H}^T\tilde{H}\|_\infty}\\ &\leq \sqrt{\|\tilde{H}^T\|_\infty\|\tilde{H}\|_\infty}\\ &\leq \sqrt{\|\tilde{H}\|_1\|\tilde{H}\|_\infty}.\end{aligned}$$

Now, remark from the definition of $\|.\|_1$ that:

$$\|\tilde{H}\|_1 \leq \left(\|A_m\|_1 + \frac{3\|D_m^T\|_1}{4m^2}\right) \vee \left(\|D_m\|_1 + \frac{3\|B_m\|_1}{4m^2}\right).$$

The matrices A_m and B_m are symmetric so that (48) and (51) lead to:

$$\|A_m\|_1 = \|A_m\|_\infty \lesssim \frac{Kd^2}{m^4\Delta^4},$$

and

$$\frac{3\|B_m\|_1}{4m^2} = \frac{3\|B_m\|_\infty}{4m^2} \lesssim \frac{Kd^3}{m^4\Delta^4}.$$

In the meantime, we use (49) and the dual relationship between $\|.\|_1$ and $\|.\|_\infty$ to deduce that:

$$\frac{3\|D_m^T\|_1}{4m^2} = \frac{3\|D_m\|_\infty}{4m^2} \lesssim \frac{Kd^3}{m^5\Delta^4}.$$

A similar argument with (50) yields:

$$\|D_m\|_1 = \|D_m^T\|_\infty \lesssim \frac{Kd^2}{m^3\Delta^4}.$$

We then deduce that $\tilde{H}$ satisfies an upper bound equivalent to (52): $\|\tilde{H}\|_1 \lesssim \frac{Kd^3}{m^3\Delta^4}$, so that:

$$\|\tilde{H}\|_2 \lesssim \frac{Kd^3}{m^3\Delta^4}.$$

Using (47), we deduce that:

$$\left\|\begin{pmatrix}\alpha - \mathbf{1}\\ \beta\end{pmatrix}\right\|_2 \lesssim \frac{K^{3/2}d^3}{m^3\Delta^4}.$$

□

Thanks to the previous proposition, we are now ready to prove Theorem 7.

Proof of Theorem 7. We define an integer m that will be chosen large enough below and consider $\mathcal{P}_m = p_m^2$.

Proof of $i)$ and $ii)$: From Proposition 16, we know that if m satisfies $m \geq C\frac{K^{1/3}d}{\Delta^{4/3}}$, for a constant C large enough independent from K, Δ and d, then $\mathcal{P}_m$ satisfies the interpolation properties:

$$0 \leq \mathcal{P}_m \leq 1 \qquad \text{with} \qquad \mathcal{P}_m(t) = 1 \Longleftrightarrow t \in \{t_1, \ldots, t_K\}.$$

Our strategy relies on a study of the variations of $\mathcal{P}_m$ near each support points $\{t_1, \ldots, t_K\}$, whose union defines the *near region*, and far from these support points, which is then the *far region*.

Near region Let $\epsilon \in \left(0, \frac{\Delta}{2}\right)$ a parameter whose value will be made precise later on. The near-region $\mathbb{N}(\epsilon)$ is the union of K intervals that are defined by:

$$\mathbb{N}(\epsilon) = \bigcup_{i=1}^{K} \{t \in \mathbb{R}^d,\ \|t - t_i\|_2 \leq \epsilon\} := \bigcup_{i=1}^{K} \mathbb{N}_i(\epsilon).$$

The far region is therefore given by:

$$\mathbb{F}(\epsilon) = \mathbb{R}^d \setminus \mathbb{N}(\epsilon).$$

Let $i \in \{1, \ldots, K\}$ be fixed, the function p_m involves a sum over k and we consider two cases:

- If $k \neq i$, then, for all $t \in \mathbb{N}_i(\epsilon)$, $\xi_{t,i,k}$ exists such that

  $$\psi_m(t - t_k) = \psi_m(t_i - t_k) + \langle (t - t_i), \nabla \psi_m(t_i - t_k) \rangle + \frac{1}{2}(t - t_i)^T D^2 \psi_m(\xi_{t,i,k} - t_k)(t - t_i),$$

  with $\|\xi_{t,i,k} - t_i\|_2 \leq \|t - t_i\|_2$. Moreover, for any $u \in \{1, \ldots, d\}$, a $\tilde{\xi}^u_{t,i,k}$ exists such that:

  $$\begin{aligned}\partial_u(\psi_m)(t - t_k) &= \partial_u(\psi_m)(t_i - t_k) + \langle (t - t_i), (\partial_{u,v}(\psi_m)(t_i - t_k))_v \rangle \\ &\quad + \frac{1}{2}(t - t_i)^T D^2\{\partial_u(\psi_m)\}(\tilde{\xi}^u_{t,i,k} - t_k)(t - t_i),\end{aligned}$$

  with $\|\tilde{\xi}^u_{t,i,k} - t_i\|_2 \leq \|t - t_i\|_2$.
- If $k = i$, since $\nabla \psi_m(0) = 0$ and $D^3(\psi_m)(0) = 0$, for all $t \in \mathbb{N}_i(\epsilon)$, a $\xi_{t,i,i}$ exists such that:

  $$\begin{aligned}&\psi_m(t - t_i) = \psi_m(0) + \frac{1}{2}(t - t_i)^T D^2(\psi_m)(0)(t - t_i) \\ &+ \frac{1}{24} \underbrace{\sum_{1 \leq u_1, u_2, u_3, u_4 \leq d} (t^{u_1} - t_i^{u_1})(t^{u_2} - t_i^{u_2})(t^{u_3} - t_i^{u_3})(t^{u_4} - t_i^{u_4}) \partial_{u_1,u_2,u_3,u_4}(\psi_m)(\xi_{t,i,i} - t_i)}_{:=(t-t_i)^T A(\xi_{t,i,i} - t_i)(t - t_i)}\end{aligned}$$

  with $\|\xi_{t,i,i} - t_i\|_2 \leq \|t - t_i\|_2$. We also have that for any $u \in \{1, \ldots, d\}$, the existence of $\tilde{\xi}^u_{t,i,i}$ such that:

  $$\partial_u \psi_m(t - t_i) = \partial_u \psi_m(0) + \langle t - t_i, (\partial_{u,v}(\psi_m)(0))_v \rangle + \frac{1}{2}(t - t_i)^T D^2(\partial_u(\psi_m))(\tilde{\xi}^u_{t,i,i} - t_i)(t - t_i),$$

  with $\|\tilde{\xi}^u_{t,i,i} - t_i\|_2 \leq \|t - t_i\|_2$.

Hence, for all $t \in \mathbb{N}_i(\epsilon)$, we can use the previous Taylor formulas and obtain that:

$$\begin{aligned}
p_m(t) &= \sum_{k=1}^{K} \left[\alpha_k \psi_m(t - t_k) + \langle \beta_k, \nabla \psi_m(t - t_k) \rangle\right], \\
&= \alpha_i \psi_m(t - t_i) + \langle \beta_i, \nabla \psi_m(t - t_i) \rangle + \sum_{k \neq i} \alpha_k \psi_m(t - t_k) + \sum_{k \neq i} \langle \beta_k, \nabla \psi_m(t - t_k) \rangle \\
&= \alpha_i \left[\psi_m(0) + \frac{1}{2}(t - t_i)^T D^2(\psi_m)(0)(t - t_i) + \frac{1}{24}(t - t_i)^T A(\xi_{t,i,i} - t_i)(t - t_i)\right] \\
&\quad + \left\langle \beta_i, \nabla \psi_m(0) + D^2(\psi_m)(0)(t - t_i) + \frac{1}{2}\left((t - t_i)^T D^2 \partial_u(\psi_m)(\tilde{\xi}^u_{t,i,i} - t_i)(t - t_i)\right)_u \right\rangle \\
&\quad + \sum_{k \neq i} \alpha_k \left[\psi_m(t_i - t_k) + \langle t - t_i, \nabla \psi_m(t_i - t_k) \rangle + \frac{1}{2}(t - t_i)^T D^2(\psi_m)(\xi_{t,i,k} - t_k)(t - t_i)\right] \\
&\quad + \sum_{k \neq i} \left\langle \beta_k, \nabla \psi_m(t_i - t_k) + D^2(\psi_m)(t_i - t_k)(t - t_i) + \frac{1}{2}\left((t - t_i)^T D^2 \partial_u(\psi_m)(\tilde{\xi}_{t,i,k} - t_k)(t - t_i)\right)_u \right\rangle.
\end{aligned}$$

These terms can be re-arranged as follows:

$$\begin{aligned}
p_m(t) &= \sum_{k=1}^{K} \left[\alpha_k \psi_m(t_i - t_k) + \langle \beta_k, \nabla \psi_m(t_i - t_k)\rangle\right] \\
&\quad + \left\langle D^2(\psi_m)(0)\beta_i + \sum_{k\neq i} \alpha_k \nabla \psi_m(t_i - t_k) + D^2(\psi_m)(t_i - t_k)\beta_k, (t - t_i)\right\rangle \\
&\quad + \frac{(t-t_i)^T}{2}\left[\alpha_i D^2(\psi_m)(0) + \sum_{k\neq i} \alpha_k D^2(\psi_m)(\xi_{t,i,k} - t_k)\right. \\
&\quad \left. + \frac{\alpha_i}{12} A(\xi_{t,i,i} - t_i) + \sum_{k=1}^{K}\sum_{u=1}^{d} \beta_i^u D^2(\partial_u \psi_m)(\tilde{\xi}_{t,i,k}^u - t_i)\right](t - t_i) \\
&= C_0 + \langle C_1, t - t_i\rangle + \frac{1}{2}(t - t_i)^T C_2(t)(t - t_i).
\end{aligned}$$

Of course, the construction of Proposition 16 yields

$$C_0 = \sum_{k=1}^{K} \left[\alpha_k \psi_m(t_i - t_k) + \langle \beta_k, \nabla \psi_m(t_i - t_k)\rangle\right] = p_m(t_i) = 1,$$

and

$$C_1 = \sum_{k\neq i} \alpha_k \nabla \psi_m(t_i - t_k) + \sum_{k=1}^{K} D^2(\psi_m)(t_i - t_k)\beta_k = \nabla p_m(t_i) = 0,$$

thanks to the constraints expressed on the function p_m. Hence, for all $t \in \mathbb{N}_i(\epsilon)$ we have

$$p_m(t) = 1 + \frac{1}{2}(t - t_i)^T C_2(t)(t - t_i).$$

In the following, we prove that C_2 is a negative matrix and bounded from below. Thanks to Lemma 19, we can compute the first term of C_2 and we have

$$D^2(\psi_m)(0) = -\frac{4m^2}{3} I_d,$$

so that

$$\frac{1}{2}(t - t_i)^T D^2(\psi_m)(0)(t - t_i) = -\frac{2m^2}{3}\|t - t_i\|_2^2. \tag{54}$$

The second term of C_2 may be upper bounded with the help of the spectral radius of $D^2(\psi_m)(\xi_{t,i,k} - t_k)$: (denoted by $\rho(M)$ for any squared symmetric matrix M):

$$\frac{1}{2}(t - t_i)^T \sum_{k\neq i} \alpha_k D^2(\psi_m)(\xi_{t,i,k} - t_k)(t - t_i) \leq \|\alpha\|_\infty \|t - t_i\|_2^2 \sum_{k\neq i} \rho\left(D^2(\psi_m)(\xi_{t,i,k} - t_k)\right).$$

To handle this last term, we use the fact that in the near region $\mathbb{N}_i(\epsilon)$, $\|\xi_{t,i,k} - t_k\|_2$ is far from 0. Using the triangle inequality, since $\epsilon < \frac{\Delta}{2}$, we have:

$$\|\xi_{t,i,k} - t_k\|_2 \geq \|t_i - t_k\|_2 - \|\xi_{t,i,k} - t_i\|_2 \geq \|t_i - t_k\|_2 - \|t - t_i\|_2 \geq \Delta - \epsilon \geq \frac{\Delta}{2}.$$

Using the inequality $\rho(M) \leq \|M\|_\infty$ for any symmetric matrix, Proposition 16 and *iii*) of Lemma 19, we obtain that:

$$\begin{aligned}
\frac{1}{2}(t-t_i)^T \sum_{k\neq i} \alpha_k D^2(\psi_m)(\xi_{t,i,k}-t_k)(t-t_i) &\lesssim K\|\alpha\|_\infty \left(d \times \frac{d^2}{m^2\Delta^4}\right)\|t-t_i\|_2^2 \\
&\lesssim \frac{Kd^3}{m^2\Delta^4}\|t-t_i\|_2^2. \qquad (55)
\end{aligned}$$

The third term of C_2 is described by the matrix:

$$\frac{\alpha_i}{12}\left(A(\xi_{t,i,i}-t_i)\right)_{u,v} = \frac{\alpha_i}{12}\left(\sum_{p=1}^d \sum_{q=1}^d (\xi_{t,i,i}^p - t_i^p)(\xi_{t,i,i}^q - t_i^q)\partial_{u,v,p,q}\psi_m(\xi_{t,i,i}-t_i)\right)_{u,v}.$$

Using that $\|\text{sinc}'\|_\infty \vee \|\text{sinc}^{(2)}\|_\infty \vee \|\text{sinc}^{(3)}\|_\infty \vee \|\text{sinc}^{(4)}\|_\infty \leq 1/2$, we obtain that $\partial_{u,v,p,q}\psi_m(\xi_{t,i,i}-t_i) \lesssim m^4$. Therefore, for any (u,v), we have:

$$\begin{aligned}
\left|\frac{\alpha_i}{12}A(\xi_{t,i,i}-t_i)_{u,v}\right| &\leq \|\alpha\|_\infty \sum_{p=1}^d\sum_{q=1}^d \left|\xi_{t,i,i}^p - t_i^p\right|\left|\xi_{t,i,i}^q - t_i^q\right| \left|\partial_{u,v,p,q}\psi_m(\xi_{t,i,i}-t_i)\right| \\
&\lesssim \|\alpha\|_\infty m^4 \sum_{p=1}^d \left|\xi_{t,i,i}^p - t_i^p\right| \sum_{q=1}^d \left|\xi_{t,i,i}^q - t_i^q\right| \\
&\lesssim d^2 m^4 \|\alpha\|_\infty \epsilon^2,
\end{aligned}$$

where the last line comes from the Cauchy-Schwarz inequality. Again, the inequality $\rho(M) \leq \|M\|_\infty$ and Proposition 16 yield:

$$\left|\frac{1}{2}(t-t_i)^T \frac{\alpha_i}{12} A(\xi_{t,i,i}-t_i)(t-t_i)\right| \lesssim d^3 m^4 \|\alpha\|_\infty \epsilon^2 \|t-t_i\|_2^2 \lesssim d^3 m^4 \epsilon^2 \|t-t_i\|_2^2. \qquad (56)$$

The last term of C_2 is studied into two steps. We first consider the situation when $k \neq i$: the triangle inequality, *iv*) of Lemma 19 and the inequality $\rho(M) \leq \|M\|_\infty$ yield:

$$\begin{aligned}
\rho\left(\sum_{k\neq i}\sum_{u=1}^d \beta_i^u D^2(\partial_u\psi_m)(\tilde{\xi}_{t,i,k}^u - t_i)\right) &\leq K\|\beta\|_\infty d \sup_{1\leq u\leq d} \rho\left(D^2(\partial_u\psi_m)(\tilde{\xi}_{t,i,k}^u - t_i)\right) \\
&\lesssim K\|\beta\|_\infty d \times \left(d \times \frac{d^2}{m\Delta^4}\right) \\
&\lesssim K\|\beta\|_\infty \frac{d^4}{m\Delta^4},
\end{aligned}$$

because each term involved in $D^2(\partial_u\psi_m)(\tilde{\xi}_{t,i,k}^u - t_i)$ is upper bounded by $\frac{d^2}{m\Delta^4}$ thanks to *iv*) of Lemma 19. Hence, we deduce from Proposition 16 that:

$$\rho\left(\sum_{k\neq i}\sum_{u=1}^d \beta_i^u D^2(\partial_u\psi_m)(\tilde{\xi}_{t,i,k}^u - t_i)\right) \lesssim \frac{K^2 d^7}{m^4\Delta^8}. \qquad (57)$$

We now consider the situation where $k = i$: for any pair (u, v):

$$\sum_{w=1}^{d} \beta_i^w \partial_{u,v,w}(\psi_m)(\tilde{\xi}_{t,i,i}^u - t_i) \lesssim d\|\beta\|_\infty m^3[m\epsilon + (m\epsilon)^3] \lesssim \frac{Kd^4}{\Delta^4}(m\epsilon + (m\epsilon)^3),$$

where we used $iv)$ of Lemma 18, $\nabla\psi_m(0) = 0$, $D^3\psi_m(0) = 0$ and $\|m(\tilde{\xi}_{t,i,i}^u - t_i)\|_2 \leq m\epsilon$ and $i)$ of Proposition 16.
We then conclude that

$$\begin{aligned}
\rho\left(\sum_{u=1}^{d} \beta_i^u D^2(\partial_u\psi_m)(\tilde{\xi}_{t,i,i}^u - t_i)\right) &\lesssim \frac{K^2d^7}{m^4\Delta^8} + \frac{Kd^4}{\Delta^4}(m\epsilon + (m\epsilon)^3) \\
&\lesssim \frac{Kd^4}{\Delta^4}\left[\frac{Kd^3}{m^4\Delta^4} + (m\epsilon + (m\epsilon)^3)\right] \\
&\lesssim \frac{Kd^4}{\Delta^4}\left[m^{-1} + (m\epsilon + (m\epsilon)^3)\right], \qquad (58)
\end{aligned}$$

where the last line comes from the writing $m^4 = m \times m^3$ and the constraint $m \gtrsim K^{1/3}d\Delta^{-4/3}$ introduced in Proposition 16.

Since $t \in \mathbb{N}_i(\epsilon)$, we deduce that $\|t - t_i\|_2 < \epsilon$. We choose

$$\epsilon = \frac{\delta}{m}. \qquad (59)$$

We now plug Equations (54), (55), (56), (57) and (58) in $C_2(t)$ and deduce that a constant $\square$ exists such that

$$\frac{1}{2}(t-t_i)^T C_2(t)(t-t_i) \leq m^2\|t-t_i\|_2^2 \left[\underbrace{-\frac{2}{3}\alpha_i + \square\left[\frac{Kd^3}{m^4\Delta^4} + \alpha_i d^3 m^2\epsilon^2 + \frac{Kd^4}{m^2\Delta^4}[m^{-1} + m\epsilon + (m\epsilon)^3]\right]}_{:=A_{\epsilon,m}}\right].$$

Then, we choose ϵ and m such that $A_{\epsilon,m} \leq -\frac{\alpha_i}{3}$, a careful inspection of the above terms prove that a sufficiently small υ and large enough C (both independent of d, K and Δ) exist such that

$$\epsilon \leq \frac{\upsilon}{md^{3/2}} \quad \text{and} \quad m = C\frac{K^{1/2}d^{5/4}}{\Delta^2} \Longrightarrow \frac{1}{2}(t-t_i)^T C_2(t)(t-t_i) \leq -\frac{\alpha_i m^2}{3}\|t-t_i\|_2^2. \qquad (60)$$

Far region $\mathbb{F}(\epsilon)$ The value of δ in (59) being fixed, we are looking for a value of $\eta > 0$ such that

$$t \in \mathbb{F}(\epsilon) \Rightarrow |p_m(t)| < 1 - \eta.$$

The definition of p_m and the Cauchy-Schwarz inequality yield

$$|p_m(t)| \leq \sum_{k=1}^{K} |\alpha_k||\psi_m(t-t_k)| + \sum_{k=1}^{K} \|\beta_k\|_2\|\nabla\psi_m(t-t_k)\|_2.$$

We consider the second term of the right hand side with the help of Lemma 18 and Proposition 16:

$$\sum_{k=1}^{K} \|\beta_k\|_2\|\nabla\psi_m(t-t_k)\|_2 \lesssim \sum_{k=1}^{K} \|\beta_k\|_2 m\|\nabla\psi\|_\infty|\psi(m(t-t_k))^3| \lesssim \frac{K^{5/2}d^3}{m^2\Delta^4}.$$

Hence, we obtain that a constant $\check{C}$ exists such that

$$|p_m(t)| \leq \sum_{k=1}^{K} |\alpha_k|\, |\psi_m(t-t_k)| + \check{C}\frac{K^2 d^3}{m^2\Delta^4}.$$

Let $t \in \mathbb{F}(\epsilon)$ and $t_{i^\star}$ the closest point of t in the set $\{t_1,\ldots,t_K\}$, the triangle inequality shows that $\forall k \neq i^\star$, we have $\|t-t_k\|_2 > \frac{\Delta}{2}$. Hence, since $\|\alpha\|_\infty$ is upper bounded by a universal constant (see Proposition 16), we deduce from $i)$ Lemma 19 that

$$\sum_{k\neq i^\star} |\alpha_k||\psi_m(t-t_k)| \lesssim \frac{Kd^2}{m^4\Delta^4}.$$

In the same time, the last term that involves $i^\star$ is upper bounded by

$$|\alpha_{i^\star}||\psi_m(t-t_{i^\star})| \leq \|\alpha\|_\infty \max_{\|x\|_2>\kappa\frac{d^{-3/2}}{m}} |\psi_m(x)| \leq \left(1+\frac{C_0 K d^3}{m^3\Delta^4}\right) \max_{\|y\|_2>\kappa d^{-3/2}} \psi^4(y),$$

where C_0 is a large enough universal constant. Using that

$$|g(x)| = \frac{|\sin(x)|}{|x|} \leq (1-x^2/12)\mathbb{1}_{|x|\leq 2} + \frac{1}{2}\mathbb{1}_{|x|\geq 2},$$

and the fact that when $\|y\|_2 \geq \upsilon d^{-3/2}$, then the absolute value of one of the coordinate of y is greater than υd^{-2}, we deduce that

$$|\alpha_{i^\star}||\psi_m(t-t_{i^\star})| \leq \left(1+\frac{C_0 K d^3}{m^3\Delta^4}\right)\left[\left(1-\frac{\upsilon^2}{12d^4}\right)\vee\frac{1}{2}\right]^4 \leq \left(1+\frac{C_0 K d^3}{m^3\Delta^4}\right)(1-\eta)^4,$$

where $\eta \asymp \upsilon^2 d^{-4}$. This entails the desired result as soon as m is chosen such that

$$m \gtrsim \frac{K^{1/3}d^{7/3}}{\Delta^{4/3}}.$$

It is easy to check that in this case, a small enough υ exists (independent of d, K, m and Δ) such that:

$$m \gtrsim \frac{K^{1/3}d^{7/3}}{\Delta^{4/3}} \quad \text{and} \quad t\in\mathbb{F}\left(\frac{\upsilon d^{-3/2}}{m}\right) \Longrightarrow |p_m(t)| \leq 1-\frac{\upsilon}{d^4}. \tag{61}$$

Conclusion of the interpolation To accomodate with conditions (60) and (61), we consider an integer m such that $m \gtrsim \sqrt{K}d^{7/3}\Delta^{-2}$ and $\epsilon = \upsilon m^{-1}d^{-3/2}$. We deduce that p_m satisfies in the far region $\mathbb{F}(\epsilon)$:

$$\forall t\in\mathbb{F}(\epsilon) \qquad -\left(1-\frac{\upsilon}{2d^4}\right) \leq p_m(t) \leq \left(1-\frac{\upsilon}{2d^4}\right),$$

while in the near region we have:

$$\forall i\in\{1,\ldots,K\} \quad \forall t\in\mathbb{N}_i(\epsilon) \qquad 0\leq p_m(t)\leq 1-\mathcal{C}m^2\|t-t_i\|^2.$$

We then consider $\mathcal{P}_m = p_m^2$ and obtain that satisfies both the constraints and the interpolation conditions in the statement of Theorem 7. We then obtain $i)$ and $ii)$.

Proof of *iii*):

Remark first that p_m is a linear combination of shifted sinus cardinal functions and derivatives of sinus cardinal functions up to the power 4 used in ψ. Moreover, it is straightforward to check that

$$\mathcal{F}[\psi^4] = \mathcal{F}[\psi^2] \star \mathcal{F}[\psi^2] = \mathcal{F}[\psi] \star \mathcal{F}[\psi] \star \mathcal{F}[\psi] \star \mathcal{F}[\psi].$$

Therefore, the Fourier transform of ψ^4 has a compact support of size $[-2,2]^d$ since the Fourier transform of the sinus cardinal is the rectangular indicator function of $[-1/2,1/2]$. Using the effect on the Fourier transform of scaling and shifting a function we deduce that the Fourier transform of p_m has a compact support, which size varies linearly with m:

$$Supp(\mathcal{F}[p_m]) \subset [-2m, 2m]^d.$$

Since $\mathcal{P}_m = p_m^2$, we have $\mathcal{F}[\mathcal{P}_m] = \mathcal{F}[p_m] \star \mathcal{F}[p_m]$ so that

$$Supp(\mathcal{F}[\mathcal{P}_m]) \subset [-4m, 4m]^d.$$

We now compute an upper bound of $\|\mathcal{P}_m\|_2$: the isometry property entails the several inequalities:

$$\begin{aligned}
\|\mathcal{P}_m\|_2 &= \|\mathcal{F}[\mathcal{P}_m]\|_2 \\
&= \|\mathcal{F}[p_m] \star \mathcal{F}[p_m]\|_2 \\
&\leq \|\mathcal{F}[p_m]\|_2 \|\mathcal{F}[p_m]\|_1,
\end{aligned}$$

where we used the standard inequality $\|g \star h\|_2 \leq \|g\|_2 \|h\|_1$.

Now, the triangle inequality yields

$$\begin{aligned}
\|\mathcal{F}[p_m]\|_2 &= \left\| \sum_{k=1}^{K} \alpha_k \mathcal{F}[\psi_m(.-t_k)] + \mathcal{F}[\langle \beta_k, \nabla \psi_m(.-t_k)\rangle] \right\|_2 \\
&\leq \sum_{k=1}^{K} |\alpha_k| \, \|\mathcal{F}[\psi_m(.-t_k)]\|_2 + \|\mathcal{F}[\langle \beta_k, \nabla \psi_m(.-t_k)\rangle]\|_2 \\
&\leq K \sup_{1\leq k\leq K} \left(|\alpha_k| \|\mathcal{F}[\psi_m(.-t_k)]\|_2 + |\beta_k|_2 \|\mathcal{F}[\nabla \psi_m(.-t_k)]\|_2 \right) \\
&\leq K \left(\|\alpha\|_\infty \|\mathcal{F}[\psi_m]\|_2 + \sup_{1\leq k\leq K} \|\beta_k\|_2 \left\| \sqrt{\sum_{i=1}^{d} \mathcal{F}[\partial_i \psi_m(.-t_k)]^2} \right\|_2 \right),
\end{aligned}$$

where the last line comes from the Cauchy-Schwarz inequality.

We then deduce that

$$\|\mathcal{F}[p_m]\|_2 \leq K \left(\|\alpha\|_\infty \|\mathcal{F}[\psi_m]\|_2 + \sup_{1\leq k\leq K} \|\beta_k\|_2 \| |\mathcal{F}[\nabla \psi_m]|_2 \|_2 \right),$$

where $|\mathcal{F}[\nabla \psi_m]|_2$ refers to the Euclidean norm of the d-dimensional vector $\mathcal{F}[\nabla \psi_m]$ Now, remark that a dilatation by a ratio m yields on L^2 norms:

$$\|\mathcal{F}[\psi_m]\|_2 \lesssim m^{-d/2} \qquad \text{and} \qquad \| |\mathcal{F}[\nabla \psi_m]|_2 \|_2 \lesssim d m^{-d/2}.$$

We use a similar argument and obtain that

$$\|\mathcal{F}[p_m]\|_1 \leq K\left(\|\alpha\|_\infty\|\mathcal{F}[\psi_m]\|_1 + \sup_{1\leq k\leq K}\|\beta_k\|_2\left\|\sqrt{\sum_{i=1}^d \mathcal{F}[\partial_i\psi_m(.-t_k)]^2}\right\|_1\right)$$

In the meantime, the effect of this dilatation on the L^1 norms is managed by:

$$\|\mathcal{F}[\psi_m]\|_1 = \int |\mathcal{F}[\psi_m](\xi)|\mathrm{d}\xi \leq m^{-d}\|\mathcal{F}[\psi]\|_\infty |Supp(\mathcal{F}[\psi_m])| \lesssim m^{-d}\|\mathcal{F}[\psi]\|_\infty m^d \lesssim 1,$$

and with a same argument we obtain that: $\||\mathcal{F}[\nabla\psi_m]|_2\|_1 \lesssim d$. Using our choice of m, we then obtain that

$$\|\mathcal{P}_m\|_2 \lesssim K^2 m^{-d/2}.$$

Proof of *iv*): The last point is a simple consequence of the convolution kernel induced by Φ. Since φ satisfies $(\mathcal{H}_{4m})$, then $\forall \xi \in [-4m, 4m]^d$, we have $\sigma(\xi) \neq 0$. Hence, we can define $c_{0,m}$ through its Fourier transform:

$$\forall \xi \in \mathbb{R}^d \qquad \mathcal{F}[c_{0,m}](\xi) = \frac{\mathcal{F}[\mathcal{P}_m](\xi)}{\sigma(\xi)}\mathbb{1}_{\xi\in Supp(\mathcal{F}[\mathcal{P}_m])}. \tag{62}$$

Moreover, the Fourier transform of $c_{0,m}$ is naturally compact, which entails that $c_{0,m} \in \mathbb{L}$. □

Some useful properties of the sinus cardinal function are detailed in the following basic lemma.

Lemma 17. *If $g(x) = sinc(x)$, then for any $x \in \mathbb{R}$:*

i)
$$g'(x) = \frac{x\cos x - \sin x}{x^2} \quad \textit{and} \quad \|g'\|_\infty \leq \frac{1}{2}.$$

ii)
$$g''(x) = -\frac{(x^2-2)\sin x + 2x\cos x}{x^3} \quad \textit{and} \quad \|g''\|_\infty \leq \frac{1}{2}.$$

iii)
$$g^{(3)}(x) = \frac{3(x^2-2)\sin x - x(x^2-6)\cos x}{x^4} \quad \textit{and} \quad \|g^{(3)}\|_\infty \leq \frac{1}{2}.$$

iii)
$$g^{(4)}(x) = \frac{4x(x^2-6)\cos x + (x^4 - 12x^2 + 24)\sin x}{x^5} \quad \textit{and} \quad \|g^{(4)}\|_\infty \leq \frac{1}{2}.$$

Some additional ingredients on ψ_m are detailed below. In the sequel, we will use the shortcut of notation ∂_u instead of $\partial_u^{|u|}\psi$ for any multi-index u.

Lemma 18. *Let ψ_m be the function defined in (19). Then*

- *i)* $\psi_m(0) = g^4(0)^d = 1$.
- *ii)* $\nabla\psi_m(0) = 0$ *and*
$$\nabla\psi_m(x) = 4m\psi^3(mx)\nabla\psi(mx).$$
- *iii)* $D^2\psi_m(0) = -\frac{4}{3}m^2 I_d$ *and*
$$(D^2\psi_m(x))_{i,j} = 4m^2[\psi^3\partial^2_{i,j} + 3\psi^2\partial_i\partial_j](mx).$$

- *iv)* $(D^3\psi_m)(0) = 0$ *and*

$$(D^3\psi_m(x))_{i,j,k} = 4m^3[\psi^3\partial^3_{i,j,k} + 6\psi\partial_i\partial_j\partial_k + 3\psi^2[\partial^2_{i,j}\partial_k + \partial^2_{i,k}\partial_j + \partial^2_{j,k}\partial_i]](mx)$$

- *v) Finally*

$$(D^4\psi_m)(x)_{i,j,k,l} \quad = \quad 4m^4[\psi^3\partial^4_{i,j,k,l} + 3\psi^2\square_{i,j,k,l} + 6\psi\tilde{\square}_{i,j,k,l} + 6\check{\square}_{i,j,k,l}](mx),$$

with

$$\square_{i,j,k,l} = \partial_i\partial^3_{j,k,l} + \partial_j\partial^3_{i,k,l} + \partial_k\partial^3_{i,j,l} + \partial_l\partial^3_{i,j,k} + \partial^2_{i,j}\partial^2_{k,l} + \partial^2_{i,k}\partial^2_{j,l} + \partial^2_{i,l}\partial^2_{j,k},$$

$$\tilde{\square}_{i,j,k,l} = \partial^2_{i,j}\partial_k\partial_l + \partial^2_{i,k}\partial_j\partial_l + \partial^2_{i,l}\partial_k\partial_j + \partial^2_{j,k}\partial_i\partial_l + \partial^2_{j,l}\partial_i\partial_k + \partial^2_{k,l}\partial_i\partial_j$$

and

$$\check{\square}_{i,j,k,l} = \partial_i\partial_j\partial_k\partial_l$$

Several bounds on the successive derivatives of ψ_m are given in the following lemma.

Lemma 19. *For any pair* (i,j) *such that* $i \neq j$*:*

- *i)* $|\psi_m(t_i - t_j)| \lesssim \frac{d^2}{m^4\Delta^4}$.
- *ii)* $|\partial_u\psi_m(t_i - t_j)| \lesssim \frac{d^2}{m^3\Delta^4}$
- *iii)* $|\partial^2_{u,v}\psi_m(t_i - t_j)| \lesssim \frac{d^2}{m^2\Delta^4}$.
- *iv)* $|\partial^3_{u,v,w}\psi_m(t_i - t_j)| \lesssim \frac{d^2}{m\Delta^4}$.

Proof. In what follows, we deliberately choose to omit the multiplicative constants since the rest of the paragraph will be managed in the same way.
Point *i*): we use $|\mathrm{sinc}(x)| \leq |x|^{-1}$ and remark that $\|t_i - t_j\|_2 \geq \Delta$ so that

$$\sum_{\ell=1}^{d}(t_i^\ell - t_j^\ell)^2 \geq \Delta^2.$$

We then deduce that

$$\psi_m(t_i - t_j) = \prod_{\ell=1}^{d}\mathrm{sinc}(m(t_i^\ell - t_j^\ell))^4 \leq \frac{1}{m^4(\Delta^2/d)^2}$$

because one coordinate ℓ_0 exists such that $|t_i^{\ell_0} - t_j^{\ell_0}|^2 \geq \Delta^2 d^{-1}$.
Point *ii*): we use Lemma 17, Lemma 18 and

$$\partial_u\psi(t) = g'(t^u)\prod_{\ell\neq u} g(t^\ell),$$

associated with $|g(x)| \vee |g'(x)| \lesssim \frac{1}{|x|}$. It yields

$$|\partial_u\psi_m(t_i - t_j)| \lesssim m\frac{d^{1/2}}{m\Delta}\left(\frac{d^{1/2}}{m\Delta}\right)^3 \lesssim \frac{d^2}{m^3\Delta^4}.$$

We then obtain *ii*).

Point *iii*): we still use Lemma 17 and Lemma 18, the fact that

$$\partial^2_{u,v}\psi(t) = \mathbb{1}_{u\neq v} g'(t^u)g'(t^v)\prod_{\ell\neq u,\ell\neq v} g(t^\ell) + \mathbb{1}_{u=v} g''(t^u)\prod_{\ell\neq u} g(t^\ell)$$

and $|g(x)| \vee |g'(x)| \vee |g''(x)| \lesssim \frac{1}{|x|}$. It leads to

$$|\partial^2_{u,v}\psi_m(t_i - t_j)| \lesssim m^2\left[\frac{d^{1/2}}{m\Delta}\frac{d^{3/2}}{(m\Delta)^3} + \frac{d}{(m\Delta)^2}\frac{d}{(m\Delta)^2}\right] \lesssim \frac{d^2}{m^2\Delta^4}.$$

Point *iv*): the proof follows the same lines with the help of the previous lemmas, we check that

$$|\partial^3_{u,v,w}\psi_m(t_i - t_j)| \quad\lesssim\quad m^3\left[\frac{d^{1/2}}{m\Delta}\frac{d^{3/2}}{(m\Delta)^3} + \frac{d^{1/2}}{m\Delta}\frac{d^{1/2}}{m\Delta}\frac{d}{(m\Delta)^2} + \frac{d^{3/2}}{(m\Delta)^3}\frac{d^{1/2}}{m\Delta}\right] \lesssim \frac{d^2}{m\Delta^4}.$$

□