*Submitted to the Annals of Statistics*
arXiv: 1907.10592v2

# SUPERMIX: SPARSE REGULARIZATION FOR MIXTURES

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

*Abstract* This paper investigates the statistical estimation of a discrete mixing measure $\mu^0$ involved in a kernel mixture model. Using some recent advances in $\ell_1$-regularization over the space of measures, we introduce a "data fitting and regularization" convex program for estimating $\mu^0$ in a grid-less manner from a sample of mixture law, 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 $\mu^0$ 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 a sufficiently large sample size $n$, our estimator has exactly as many weighted Dirac masses as the target $\mu^0$, converging in amplitude and localization towards the true ones. Finally, we also introduce some tractable algorithms for solving this convex program based on "Sliding Frank-Wolfe" or "Conic Particle Gradient Descent".

Statistical performances of this estimator are investigated designing a so-called "dual certificate", which is appropriate to our setting. Some classical situations, as *e.g.* mixtures of super-smooth distributions (*e.g.* Gaussian distributions) or ordinary-smooth distributions (*e.g.* Laplace distributions), are discussed at the end of the paper.

*Preprint version of June 1, 2020*

## 1. Introduction.

1.1. *Mixture problems.* In this paper, we are interested in the estimation of a mixture distribution $\mu^0$ using some i.i.d. observations $\mathbf{X} := (X_1, \ldots, X_n) \in (\mathbb{R}^d)^n$ with the help of some $\ell_1$-regularization methods. More precisely, we consider the specific situation of a discrete distribution $\mu^0$ 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 $\delta_{t_k}$ is a Dirac mass at point $t_k \in \mathbb{R}^d$ while

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

*MSC 2010 subject classifications:* Primary: 62G05, 90C25; Secondary: 49M29

*Keywords and phrases:* Beurling Lasso; Mixture recovery; Dual certificate; Kernel approach; Super-resolution

*Acknowledgements:* S.G. benefited from the support of the ANR project Masdol funded by the *Agence Nationale pour la Recherche.*

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

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

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 $\varphi$ 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 $E$:*

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

*where $U^0 \in S^0$ is distributed according to $\mu^0$ (i.e., the mixing law (1)) and $E$ has a distribution of density $\varphi$ 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 [17, 22] for a complete overview.

1.2. *Previous works.* The main goal of this paper is to provide an estimation of the discrete mixture law $\mu^0$ 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 [30], 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, 15]). Several alternative methods have been proposed in this context. Some contributions extensively use the MLE point of view to derive consistent properties in general semi-parametric models, including the Gaussian case (see *e.g.* [28]), whereas some other ones developed some contrast functions in a semi-parametric framework: with symmetry and number of component assumptions in [5, 7], 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 [31]. The article [2] provides a general theoretical framework to analyze the convergence of the EM updates in a neighborhood of the MLE, and derives 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, ...

Our estimator will be any solution to a convex program and it does not require to know the number $K$ of components in the mixing law $\mu^0$. This estimator is based on ideas from super-resolution and "off-the-grid" methods [4, 8], where one aims at recovering a discrete measure from linear measurements. The so-called "*sparse deconvolution*" problem fits this framework since it concerns with estimating a target measure from the observation of some product of convolution between the target measure and known kernel as $f^0$ in (2). Note that in mixture models, we do not observe $f^0$ but rather a sample drawn from it, and standard strategies such that (1.15) in [8] cannot be invoked here. However, remark that one of the main advances has been the construction of the so-called "*dual certificate*" in [8] which is the key to demonstrate the success of discreteness inducing norm regularization (see *e.g.* [11, 8, 14, 12]).

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 $\mu^0$ (see Equation (1)) inspired by some recent results in $\ell_1$-regularization on the space of measures, sometimes referred to as super-resolution methods (see, *e.g.*, [11, 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. The stability result and the construction of the *dual certificate* given in [8] played a central role in our work to obtain the statistical recovery. However, these authors work on the torus and their construction provides periodic dual certificates which are not useful in our present framework. One important contribution of this paper is thus a novel dual certificate construction, interpolating phases/signs on $\mathbb{R}^d$ (and not the $d$-dimensional torus as in [8]). We also investigate the stability with respect to *sampling* of our estimation strategy, *i.e.* the ability of our procedure to recover the mixture when we compute $\hat{\mu}_n$ up to some i.i.d. observations $(X_i)_{i\in[n]}$ with $n \to +\infty$, which is a different problem from the stability issue studied in [8] that asserts the variation of the super-resolution solutions with respect to an $\ell_1$ norm control on the low-frequency data.

The minimized criterion requires to tune two parameters: a bandwidth parameter of the data fitting term denoted by $m \geq 1$ and an $\ell_1$-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 $\mu^0$ through the so-called "minimum separation" $\Delta$ 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. *Assume that the kernel $\varphi$ satisfies ($\mathcal{H}_\eta$) with $\eta = 4m$ (see Section 2.3 for a definition) for a bandwidth $m$ verifying*

$$m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1} \quad \text{where} \quad \Delta_+ = \min(\Delta, 1). \tag{3}$$

*Then, some quantity $\mathcal{C}_m(\varphi) > 0$ exists such that, setting*

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

*our estimator $\hat{\mu}_n$ 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 made explicit later on.*

*iii) Support stability property: if $\varphi$ satisfies the* Non-Degenerated Bandwidth *condition (NDB) (see Section 4.4 for a definition), for $n$ large enough, with an overwhelming probability, $\hat{\mu}_n$ 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 $n$ tends to infinity.*

Note that the constant $\mathcal{C}_m(\varphi)$ will depend on other quantities introduced later. It will be specified in Proposition 9.

These three results deserve several comments. First, *i*) indicates that when a set $A$ has enough mass w.r.t. the estimated measure $\hat{\mu}_n$, it includes 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 [14] and allows to derive the support stability. The last result *iii*) shows that, for large enough sample size, $\ell_1$-regularization successfully recovers the number of mixing components. The estimated weights on the Dirac masses then converge towards the true ones in amplitudes and localizations.

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 $\varphi$ and to the dimension $d$ of the ambient space. These results are applied to specific settings (super-smooth and ordinary-smooth mixtures).

1.4. *Outline.* This paper is organized as follows. Section 2 introduces some standard ingredients of $\ell_1$ regularization methods and gives a deterministic analysis of the exact recovery property of $\mu^0$ 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.* [11]). Tractable algorithms solving BLASSO when the observation is a sample from a mixing law are introduced in Section 3.3. 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 the companion supplementary material [10].

## 2. Assumptions, notation 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 notation 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 measure* $\mu \in \mathcal{M}(\mathbb{R}^d,\mathbb{R})$.

A standard argument proves that the total variation of $\mu$ 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 $\varphi$ 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 denote by $\mathcal{C}_0(\mathbb{R}^d,\mathbb{R})$ the space of continuous real valued functions *vanishing at infinity* on $\mathbb{R}^d$ and recall that $\mathcal{F}\left(L^1(\mathbb{R}^d)\right)$ is a dense subset of $\mathcal{C}_0(\mathbb{R}^d,\mathbb{R})$. We shall also introduce the convolution operator $\Phi$ 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.* [26, 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 $\varphi$ involved in (2), we will do the following assumption.

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

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

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

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.*, [26, Theorem 11.32]), $\varphi$ is the inverse Fourier transform of a nonnegative measure $\Sigma$ referred to as the *spectral measure.* The Fourier inversion theorem states that $\Sigma$ 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 $\mathrm{Supp}(\sigma)$:

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

Some examples of densities $\varphi$ 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 $\mu^0$ 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 $\mu^0$ not from a sample $X_1,\dots,X_n$ 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 $\Phi$ is defined by (5). Hence, this paragraph concerns the recovery of $\mu^0$ from its convolution by the kernel $\varphi$. We thus face an inverse (deconvolution) problem. Several solutions could be provided and a standard method would rely on Fourier inversion

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

where $\sigma$ is given by (7).

Here, we prove in a first step that this deconvolution problem can be efficiently solved using a $\ell_1$-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 $\varphi$ 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 $\varphi$:

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

Example 1. *It may be shown that the set of densities $\varphi$ 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., [27, 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 $\mu^0$ the program (8) has $\mu^0$ as unique solution point:*

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

We emphasize that the previous result also holds for measures $\mu^0$ that are not necessarily discrete. The proof is given in Section 3.1 of [10].

2.3. *The Super-resolution phenomenon.* Theorem 2 entails that the measure $\mu^0$ 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 $\varphi$. For the sake of simplicity, we assume that the spectral density $\sigma$ has a support that contains the hypercube $[-\eta,\eta]^d$ for some frequency threshold $\eta>0$:

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

Remark 2. *The densities $\varphi$ that satisfy ($\mathcal{H}_\eta$) and for which $\operatorname{Supp}(\sigma)=[-\eta,\eta]^d$ act as "low pass filters". The convolution operator $\Phi$ described in (5) cancels all frequencies above $\eta$, see for instance (6). Of course, the larger $\eta$, 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 $\mu^0$ is still the only solution of the program (8).

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

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

*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* $\operatorname{Supp}(\mathcal{F}[\mathcal{P}_\eta])\subset[-\eta,\eta]^d$.

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

The proof is given in Section 3.2 of [10]. A construction of such a certificate $\mathcal{P}_\eta$ is presented in Section 6 of [10] 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 $\Omega$ of its spectral density. Remark that since $\sigma$ is nonzero, necessarily $\Omega$ 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 $\mathbf{X}=(X_1,\dots,X_n)$ of i.i.d. observations distributed with the density $f^0$. In this context, only the empirical measure

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

is available, and our aim is to recover $\mu^0$ 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 $\mu^0$ defined as:

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

where $\kappa$ 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 $\mu^0$ and any candidate $\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$.

Some examples of possible cost functions $C : \mathbb{H} \times \mathcal{M}(\mathbb{R}^d, \mathbb{R}) \to \mathbb{R}$ 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 [29], as well as in the statistical and machine learning communities, (see [27] 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 $\ell$ 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 $\ell$ is called the *reproducing kernel* associated to $\mathbb{L}$. Below, we consider a kernel $\ell$ such that $\ell(x, y) = \lambda(x - y)$ for all $x, y \in \mathbb{R}^d$ where $\lambda$ 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 $\lambda$ 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 $\lambda$ is continuous, $\Lambda$ is then a bounded measure and the Mercer theorem (see *e.g.* [3]) proves that the RKHS $\mathbb{L}$ is exactly characterized by

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

with dot product

$$\forall f, g \in \mathbb{L}\,, \quad \langle f, g\rangle_{\mathbb{L}} = \int_{\mathbb{R}^d} \frac{\overline{\mathcal{F}[f]}(t)\mathcal{F}[g](t)}{\mathcal{F}[\lambda](t)} \mathrm{d}t\,.$$

*Convolution in the RKHS.* The RKHS structure associated to the kernel $\lambda$ 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 given in Section 2.1 of [10].

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

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

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*

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

*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:*

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

*where $B_\infty(1/\tau)$ denotes the centered $\ell_\infty$ 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*

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

*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 $\tau$ 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 $\tau$ 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 $C_\lambda$ introduced in Equation (13) can be used in practice giving a useful expression to compute it.

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

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

We stress that $\|L\hat{f}_n\|_{\mathbb{L}}^2$ does no depend on $\mu$ and can be removed from the criterion when it is used in the program (8). The proof of Proposition 5 is given in Section 2.2 of [10].

3.2. *Estimation by convex programming.* Our estimator is defined as a 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:

$$(\mathbf{P}_\kappa) \qquad \inf_{\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})} \left\{ \frac{1}{2} \|L\hat{f}_n - L \circ \Phi\mu\|_{\mathbb{L}}^2 + \kappa \|\mu\|_1 \right\}, \,,$$

where $\|\cdot\|_{\mathbb{L}}$ is the norm associated to the RKHS generated by $\lambda$ (see Section 3.1) and $\kappa > 0$ 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

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

Algorithmic issues related to the computation of (15) are sketched in Section 3.3 and discussed in depth in Section 1 of [10].

REMARK 4. *The tuning parameter $\kappa > 0$ needs to be chosen carefully. First note that it may depend on the choice of the frequency cut-off $1/\tau$ in $\lambda_\tau$, which is the bandwidth feature map (see Remark 2 for a definition). Our analysis shows that $\tau = 1/4m$ as in (28), and $m$ is a standard nonparametric bandwidth in mixture models for which rates are given in Section 5. The main message being that it depends only on the regularity of $\varphi$ and on the sample size $n$ for $n$ large enough. From a practical view point, it is not excluded to use a Cross-Validation scheme as it heuristically performs well for $L^1$-based methods such as LASSO. In this case, the score function can be chosen to be the data fidelity term $\|L\hat{f}_n - L \circ \Phi\mu_{\mathrm{cv}}\|_{\mathbb{L}}^2$ evaluated on the validation set. From a theoretical view point, one may choose $\kappa$ as in Equation (24). Then, Equation (25) results in*

$$\kappa \geq \frac{\rho_n}{\mathcal{C}_m(\varphi, \lambda)} \,,$$

*and these quantities depend only on the regularity of $\varphi$ and the sample size $n$ for $n$ large enough, as studied in Section 5.*

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 aim 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 variable $c \in \mathbb{L}$ as well as the following Lagrangian:

$$\mathcal{L}(\mu, z, c) := \frac{1}{2}\|L\hat{f}_n - z\|_{\mathbb{L}}^2 + \kappa\|\mu\|_1 - \langle c, L \circ \Phi\mu - z\rangle_{\mathbb{L}}\,. \tag{16}$$

It is immediate to check that if $z \neq L \circ \Phi\mu$, then the supremum of $\mathcal{L}(\mu, z, c)$ over $c$ is $+\infty$. Therefore, the primal expression coincides with the supremum in the dual variables, namely

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

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

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

Theorem 6 (Primal-Dual programs, strong duality). *The following statements are true.*

*i) The primal problem* $(\mathbf{P}_\kappa)$ *has at least one solution and it holds that*

$$\hat{z}_n := L \circ \Phi\hat{\mu}_n \quad \textit{and} \quad \hat{m}_n := \|\hat{\mu}_n\|_1 \quad \textit{are uniquely defined,}$$

*hence, 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*

$$\frac{\|L\hat{f}_n\|_{\mathbb{L}}^2}{2} - \inf\left\{\frac{1}{2}\|L\hat{f}_n - c\|_{\mathbb{L}}^2 \ :\ c \in \mathbb{L}\ s.t.\ \|\Phi c\|_\infty \leq \kappa\right\} \iff (\mathbf{P}_\kappa^*),$$

*and there is no duality gap (strong duality holds). Furthermore, it has a unique solution*

$$\hat{c}_n = L\hat{f}_n - \hat{z}_n\,.$$

*iii) Any solution* $\hat{\mu}_n$ *to* $(\mathbf{P}_\kappa)$ *satisfies*

$$\mathrm{Supp}(\hat{\mu}_n) \subseteq \left\{x \in \mathbb{R}^d \ :\ |\hat{\eta}_n|(x) = 1\right\} \quad \textit{and} \quad \int_{\mathbb{R}^d}\hat{\eta}_n\,\mathrm{d}\hat{\mu}_n = \hat{m}_n\,,$$

*where*

$$\hat{\eta}_n := \frac{\Phi\hat{c}_n}{\kappa} = \frac{1}{\kappa}\Phi(L\hat{f}_n - z_n)\,, \tag{17}$$

*i.e. it 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* $\Lambda$ *or* $\sigma$ *has a bounded support, then* $\{x \in \mathbb{R} : |\hat{\eta}_n|(x) = 1\}$ *is discrete with no accumulation point, any primal solution* $\hat{\mu}_n$ *has an (at most countable) discrete support* $\hat{S} \subset \mathbb{R}$ *with no accumulation point:*

$$\hat{\mu}_n = \sum_{t\in\hat{S}}\hat{a}_t\delta_t\,. \tag{18}$$

The proof of this result can be found in Section 4 of [10].

It is generally numerically admitted, see for instance [8, Page 939], that the extrema of the dual polynomial $\hat{\eta}_n = \Phi \hat{c}_n$ 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 $n$ and under the so-called (NDB) condition, 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: we deduce that when $d = 1$, all solutions $\hat{\mu}_n$ are of the form* (18)*.*

3.3. *Tractable Algorithms for BLASSO Mixture Models.* Available algorithms for solving (15) with "*off-the-grid*" methodology can be roughly divided into two categories: greedy methods and Riemannian descent methods. We emphasize that if the BLASSO has been studied in the past decade, the formulation (15) has two new important features. First the observation is a sample from a mixing law. Second, the data fidelity term has been tuned to incorporate a low pass filter kernel $\lambda$. For both methods, we refer to the supplementary material [10] for further details and references.

**Algorithm 1** Sliding Frank Wolfe Algorithm (SFW) for BLASSO Mixture Models

```
Initialize with $\hat{\mu}^{(0)} = 0$
while the stopping criterion is not met do
    $\hat{\mu}^{(k)} = \sum_{i=1}^{N^{(k)}} a_i^{(k)} \delta_{t_i^{(k)}}$, $a_i^{(k)} \in \mathbb{R}$, $t_i^{(k)} \in \mathbb{R}^d$ where $N^{(k)} = |Supp(\hat{\mu}^{(k)})|$ and find $t_\star^{(k)}$ such that
        $t_\star^{(k)} \in \arg\max_{t \in \mathbb{R}^d} |\eta^{(k)}(t)|$   where   $\eta^{(k)} = -\frac{\nabla \mathrm{F}(\hat{\mu}^{(k)})}{\kappa}$
    if $|\eta^{(k)}(t_\star^{(k)})| \leq 1$ then
        $\hat{\mu}^{(k)}$ is an exact solution Stop
    else
        Find $\hat{\mu}^{(k+\frac{1}{2})} = \sum_{i=1}^{N^{(k)}} a_i^{(k+\frac{1}{2})} \delta_{t_i^{(k)}} + a_i^{(k+\frac{1}{2})} \delta_{t_\star^{(k)}}$ such that
(LASSO Step)        $a^{(k+\frac{1}{2})} \in \arg\min_{a \in \mathbb{R}^{N^{(k)}+1}} F_{N^{(k)}+1}(a, t^{(k+\frac{1}{2})}) + \kappa \|a\|_1$
where $t^{(k+\frac{1}{2})} := (t_1^{(k)}, \ldots, t_{N^{(k)}}^{(k)}, t_\star^{(k)})$ is kept fixed.
        Obtain $\hat{\mu}^{(k+1)} = \sum_{i=1}^{N^{(k)}+1} a_i^{(k+1)} \delta_{t_i^{(k+1)}}$ such that
(19)        $(a^{(k+1)}, t^{(k+1)}) \in \arg\min_{(a,t) \in \mathbb{R}^{N^{(k)}+1} \times (\mathbb{R}^d)^{N^{(k)}+1}} F_{N^{(k)}+1}(a, t) + \kappa \|a\|_1$
        using a non-convex solver initialized with $(a^{(k+\frac{1}{2})}, t^{(k+\frac{1}{2})})$.
        Eventually remove zero amplitudes Dirac masses from $\hat{\mu}^{(k+1)}$.
    end if
end while
```

*Greedy method: the Sliding Frank-Wolfe algorithm (SFW).* The Frank-Wolfe algorithm is an interesting avenue for solving *differentiable convex* programs on *weakly compact convex* sets, see [13] and references therein for further details, which can be adapted to compute approximate solutions of the BLASSO Mixture Models (15) with a supplementary *sliding* step. For a measure $\mu_{a,t}$ that may be decomposed into a finite sum of Dirac masses, we define $F_N$ the data-fitting term:

$$\mu_{a,t} := \sum_{i=1}^{N} a_i \delta_{t_i} \quad \text{and} \quad \mathrm{F}_N(a,t) := \mathrm{F}(\mu_{a,t}) = \frac{1}{2}\|L\hat{f}_n - \sum_{i=1}^{N} a_i L \circ \Phi\delta_{t_i}\|_{\mathbb{L}}^2 \,. \tag{20}$$

The SFW method is then described in Algorithm 1. It is a greedy method that recursively builds

$$\eta_\mu := -\frac{\nabla \mathrm{F}(\mu)}{\kappa} = \frac{1}{\kappa}\Phi(L\hat{f}_n - L\circ\Phi\mu)\,,$$

see Line 3 of Algorithm 1.

**Algorithm 2** Conic Particle Gradient Descent Algorithm for BLASSO Mixture Models

```
Choose two gradient step sizes $\alpha, \beta > 0$ and the number of Particles $N \geq 1$.
Define $N$ Particles weights-locations $(r_i^{(0)}, t_i^{(0)})_{i=1}^N$ representing the *initial measure*
    $$\hat{\mu}^{(0)} := \frac{1}{N}\sum_{i=1}^{N} a_i^{(0)} \delta_{t_i^{(0)}}\,,$$
    where $a_i^{(0)} := (r_i^{(0)})^2$.
while stopping criterion is not met do
    For all $i = 1, \dots, N$ update (*mirror descent step for $r$* associated to the KL divergence over $\mathbb{R}_+^d$)
        $$r_i^{(k+1)} = r_i^{(k)} \exp\Big(2\,\alpha\,\kappa\,\big(\eta^{(k)}(t_i^{(k)}) - 1\big)\Big)$$
        $$t_i^{(k+1)} = t_i^{(k)} + \beta\,\kappa\,\nabla\eta^{(k)}(t_i^{(k)})$$
    where $\eta^{(k)} = -\frac{\nabla \mathrm{F}(\hat{\mu}^{(k)})}{\kappa}$, $\hat{\mu}^{(k)} := \frac{1}{N}\sum_{i=1}^{N} a_i^{(k)}\delta_{t_i^{(k)}}$ and $a_i^{(k)} = (r_i^{(k)})^2$.
end while
```

*Conic Particle Gradient Descent (CPGD).* Conic Particle Gradient Descent [9] is an alternative promising avenue for solving BLASSO for Mixture Models (15). The idea is still to discretize a positive measure into a system of particles, *i.e.* a sum of $N$ Dirac masses following (20) with $a_i = r_i^2$ and use a mean-field approximation in the Wasserstein space jointly associated with a Riemannian gradient descent with the conic metric. We refer to [9] and the references therein for further details. This method may be shown to be rapid, with a $\log(\epsilon^{-1})$ cost instead of $\epsilon^{-1/2}$ for standard convex programs. Adapted to the BLASSO for Mixture Models, we derive in Algorithm 2 a version of the Conic Particle Gradient Descent of [9, Algorithm 1] and we implemented this algorithm for Mixture Models in Figure 1.

More precisely, Figure 1 is a *proof-of-concept* and CPGD for Mixture Models would be investigated in future work. One may see that this method uncovers the right number of targets Dirac masses and their locations as some particules cluster around three poles. Some of particules vanishes and do not detect the support. Notice that a *soft-thresholding* effect tends to zero the small amplitudes as it may standardly be shown in $L^1$ regularization.

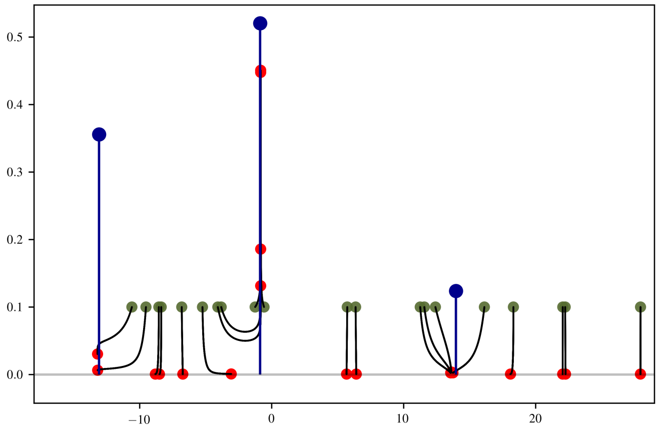

Figure 1. *Conic Particle Gradient Descent for BLASSO Mixture Models. We consider the mixing law* $\mu^0$ *made by three Dirac masses (in blue) at location* $(-13.1, -0.9, 14.0)$ *(chosen at random) and amplitudes* $(0.36, 0.52, 0.12)$*(chosen at random). We draw* $n = 200$ *iid samples with respect to the mixture with density* $f^0 = \mu^0 \star \varphi$ *where* $\varphi$ *is the pdf of standard Gaussian. Then we start CPGD for BLASSO (with parameters* $\kappa = 0.01$ *and* $\tau = 0.1$*) with* 20 *particules (in green) located at random and we run* 2,500 *gradient steps (with parameters* $\alpha = 0.05$ *and* $\beta = 1$*) as in Algorithm 2. The final locations* $(t_i)$ *and weights* $(a_i)$ *are given in red (for readability we represented* $(t_i, 2 * a_i)$*).*

**4. Statistical recovery of $\mu^0$.** This section provides some theoretical results for $\hat{\mu}_n$, built as the solution of ($\mathbf{P}_\kappa$). Contrary to $\ell_1$-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, 14] 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 $\mathbb{R}^d$ 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 $\alpha$ and a matrix $\beta$ of coefficients, defined as:

$$p_m^{\alpha,\beta}(t) = \sum_{k=1}^{K} \left\{ \alpha_k \psi_m(t - t_k) + \langle \beta_k, \nabla \psi_m(t - t_k) \rangle \right\}, \quad \forall t \in \mathbb{R}^d, \tag{21}$$

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

$$\psi_m(.) = \psi^4(m.) \text{ with } \forall u = (u^1, \ldots, u^d) \in \mathbb{R}^d \quad \psi(u) = \prod_{j=1}^{d} \operatorname{sinc}(u^j) \text{ and } \operatorname{sinc}(x) = \frac{\sin(x)}{x}. \tag{22}$$

One important feature of $\psi_m$ 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 $S^0 = \{t_1, \dots, t_K\}$ in $\mathbb{R}^d$ with $\Delta := \min_{k \neq \ell} \|t_k - t_\ell\|_2$ and $\Delta_+ = \min(\Delta, 1)$. 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 $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$ such that*

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

*and*

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

- *ii) A universal pair $(\upsilon, \gamma)$ independent from $n, m$ and $d$ exists such that for $\epsilon = \frac{\upsilon}{md}$ :*
  - *Near region: 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:*

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

  - *Far region:*

$$\forall t \in \mathbb{F}(\epsilon) := \mathbb{R}^d \setminus \mathbb{N}(\epsilon) : \qquad 0 \leq \mathcal{P}_m(t) \leq 1 - \gamma \frac{\upsilon^2}{d^3}.$$

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

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

- *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 Section 6 of [10]. This construction is inspired from 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 $\Delta$ between the location parameters $\{t_1, \dots, t_K\}$ is translated in the initial constraint $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$.

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 $S^0 = \{t_1, \dots, t_K\}$ such that $\Delta := \min_{k \neq \ell} \|t_k - t_\ell\|_2$. Let $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$. 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 $(\upsilon, \gamma)$ exists such that the function $\mathcal{Q}_m^k$ satisfies for $\epsilon = \frac{\upsilon}{md}$:*

*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* $\mathbb{F}(\epsilon)$*:*

$$\forall t \in \mathbb{F}(\epsilon),\ 0 \leq \mathcal{Q}_m^k(t) \leq 1 - \gamma \frac{\upsilon^2}{d^3}.$$

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

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).

4.2. *Bregman divergence* $D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0)$. Below, the statistical loss between $\hat{\mu}_n$ and $\mu^0$ 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:

$$D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) := \|\hat{\mu}_n\|_1 - \|\mu^0\|_1 - \int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) \geq 0\,. \tag{23}$$

We also introduce the term $\Gamma_n$ 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 result provides a control between $\hat{\mu}_n$ and $\mu^0$ with the Bregman divergence.

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* $\kappa$ *is chosen such that*

$$\kappa = \frac{\rho_n}{\|c_{0,m}\|_{\mathbb{L}}} \tag{24}$$

*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*

$$\|c_{0,m}\|_{\mathbb{L}} \leq \sqrt{\frac{\|\mathcal{P}_m\|_2^2}{\inf\limits_{\|t\|_\infty \leq 4m}\left\{\sigma^2(t)\mathcal{F}[\lambda](t)\right\}}} \lesssim \underbrace{\frac{K^2 m^{-d/2}}{\sqrt{\inf\limits_{\|t\|_\infty \leq 4m}\left\{\sigma^2(t)\mathcal{F}[\lambda](t)\right\}}}}_{:=\mathcal{C}_m(\varphi,\lambda)}\,. \tag{25}$$

The proof of Proposition 9 is postponed to Section 6.1. The previous results indicate that the Bregman divergence between our estimator $\hat{\mu}_n$ and the target measure $\mu^0$ depends, up to some constants, on three main quantities:

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

Remark 5. *By using similar arguments to prove item ii) of Proposition 9, we can complete item (iv) of Corollary 8 as follows: A $c_{k,m} \in \mathbb{L}$ exists such that $\mathcal{Q}_m^k = \Phi c_{k,m}$ and*

$$\|c_{k,m}\|_{\mathbb{L}} \lesssim \frac{K^2 m^{-d/2}}{\sqrt{\inf_{\|t\|_\infty \leq 4m} \left\{\sigma^2(t)\mathcal{F}[\lambda](t)\right\}}}. \tag{26}$$

Remark 6. *We will derive from Proposition 9 some explicit convergence rates in each specific situation, i.e. as soon as the quantities involved in Equation (25) 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 $\lambda$. Some examples will be discussed in Section 5. Indeed, $\kappa$ is related to $m$ through the relationship $\kappa = \rho_n/\|c_{0,m}\|_{\mathbb{L}}$. Similarly, we will see in Section 5 that the kernel $\lambda$ is also linked to $m$ in a transparent way. We stress that according to Proposition 9, $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$. Such a condition will be satisfied provided $m$ is allowed to go to infinity with $n$ whereas $K, \Delta, d$ are kept fixed.*

Remark 7. *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 the alternative certificate $\mathcal{Q}_m$ introduced in Corollary 8 .*

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 $\mu^0$ by $\hat{\mu}_n$ is translated by the statistical properties of the computed weights of $\hat{\mu}_n$ around the spikes of $\mu^0$ (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 $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$ 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 $(\gamma, \upsilon)$ exists such that, if*

$$\epsilon = \frac{\upsilon}{md}, \tag{27}$$

*i) Far region and negative part:*

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

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

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

*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 6.2.

REMARK 8. *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.* We introduce $\mathcal{P}_0 := \Phi c_0$ the "minimal norm certificate" (see *e.g.* [14]), which is defined by:

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

when it exists.

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

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

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

THEOREM 11. *Let the triple $\lambda, \varphi, \mu^0$ be such that* (NDB) *holds. Let $r_\kappa \in (0, \frac{1}{2})$ and set $\kappa_n = \sqrt{\lambda(0)}\, n^{-r_\kappa}$. Let $\hat{\mu}_n$ be the BLASSO estimator* ($\mathbf{P}_\kappa$) *with a tuning parameter $\kappa = \kappa_n$.*

*Then for $n$ large enough, and with probability at least $1 - Ce^{-n^{\frac{1}{2}-r_\kappa}}$ for a universal constant $C > 0$, 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 $\hat{a}_k$, as $n$ tends to infinity.*

The proof can be found in Section 5 of [10]. We emphasize that $C$ is independent from the dimension $d$, from the RKHS used $\mathbb{L}$ or the location of the spikes for example.

REMARK 9. *In Theorem 11, note that the data fidelity kernel $\lambda$ is fixed but in practice, the bandwidth of $\lambda$ often depends on the sample size $n$. Theorem 11 suggests the heuristics that the data fidelity kernel $\lambda = \lambda_n$ may depend on $n$ and it might be such that $\kappa_n = \sqrt{\lambda_n(0)}\, n^{-r_\kappa}$ vanishes as $n$ tends to infinity.*

REMARK 10. *Assume that the mixing kernel $\varphi$ is such that $\varphi \star \lambda = \psi_m \star \lambda$ where $\psi_m$ is defined by* (22)*, assume that the data fidelity kernel is such that $\lambda = \lambda_{1/(4m)}$ and assume that $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$. Then our certificate $\mathcal{P}_m$ is called the* **vanishing derivatives pre-certificate** *by [14, 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 coincide so that $\mathcal{P}_m$ is the minimal norm certificate $\mathcal{P}_0$ appearing in* (NDB)*, and Theorem 7 shows that* (NDB) *holds.*

## 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 $\varphi$. 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 $\lambda_\tau$, 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 (25)) we are led to consider $\tau$ such that

$$\frac{1}{\tau} = 4m. \tag{28}$$

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 $\lambda_\tau$. In this context, some straightforward and standard computations yield

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

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

$$\rho_n \mathcal{C}_m(\varphi, \lambda_\tau) \leq \frac{K^2 2^{3d/2}}{\sqrt{n} \times \inf\limits_{\|t\|_\infty \leq 4m} \sigma(t)}. \tag{29}$$

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 $\epsilon = \mathcal{O}(1/(md))$ in (27)) 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 $\epsilon$ is improved for large values of $m$ but the statistical variability is then degraded according to the decrease rate of the Fourier transform $\sigma$ of $\varphi$, 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 (29) but is also hidden in the constraint

$$m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1},$$

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 case of severely ill-posed inverse problems with an exponential decrease of the Fourier transform for large frequencies, which corresponds to super-smooth distributions. We emphasize that this class contains the standard benchmark of the Gaussian case, which will be discussed in details.
- the case of *mildly ill-posed inverse problems* which encompasses multivariate Laplace distributions, Gamma distributions, double exponentials among others.

### 5.2. *Super-smooth mixture deconvolution and Gaussian case.*

5.2.1. *Description of the distributions.* We consider in this paragraph the statistically hard situation of the general family of mixing distribution $\varphi$ with an exponential decrease of the Fourier transform. More precisely, we assume that the spectral density $\sigma$ of $\varphi$ satisfies:

$$(\mathcal{H}_{\alpha,\beta}^{supersmooth}) \qquad \exists j \in \mathbb{N}^\star \quad s.t. \quad \mathcal{F}[\varphi](t) = \sigma(t) = e^{-\alpha \|t\|_j^\beta} \quad \forall t \in \mathbb{R}^d, \alpha > 0, \beta > 0.$$

where for any $j \in \mathbb{N}^\star$, $\|.\|_j$ denotes the $\ell^j$-norm. The assumption ($\mathcal{H}_{\alpha,\beta}^{supersmooth}$) includes obviously the Gaussian distribution but also many other distributions as suggested by the list of examples displayed below (among others).

- *The multivariate Cauchy distribution.* For a dispersion parameter $\alpha$, $\varphi$ is defined by:

$$\varphi(x) = \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{1}{2})\pi^{\frac{d}{2}}\sqrt{\alpha}\{1+\alpha^{-1}\|x\|_2^2\}^{\frac{d+1}{2}}} \quad \forall x \in \mathbb{R}^d \qquad \text{and} \qquad \sigma(t) = e^{-\sqrt{\alpha}\|t\|_2}, \quad \forall t \in \mathbb{R}^d.$$

- *The tensor product of univariate Cauchy distribution.* An alternative example is:

$$\varphi(x) = \frac{1}{\pi^d}\prod_{j=1}^d \left(\frac{\alpha}{x_j^2+\alpha^2}\right) \quad \forall x = (x_1 \ldots x_d)^T \in \mathbb{R}^d \quad \text{and} \quad \sigma(t) = e^{-\alpha\|t\|_1}, \quad \forall t \in \mathbb{R}^d.$$

- *The multivariate Gaussian distribution.* A standard benchmark study of the Gaussian law:

$$\varphi : x \longmapsto (2\pi)^{-d/2} e^{-\|x\|^2/2} \qquad \text{and} \qquad \sigma(t) = e^{-\frac{\|t\|_2^2}{2}}, \quad \forall t \in \mathbb{R}^d.$$

5.2.2. *General recovery result.* In the situations covered by assumption ($\mathcal{H}_{\alpha,\beta}^{supersmooth}$), we shall observe that $\|t\|_j \leq d^{1/j}\|t\|_\infty$ and we verify that:

$$\inf_{\|t\|_\infty \leq 4m} \sigma(t) = e^{-\alpha(4d^{1/j}m)^\beta}.$$

In that case, we obtain that

$$\rho_n \mathcal{C}_m(\varphi, \lambda_\tau) \lesssim K^2 2^{3d/2} \times \frac{e^{\alpha(4d^{1/j}m)^\beta}}{\sqrt{n}}.$$

A straightforward application of Theorem 10 leads to the following result.

PROPOSITION 12. *Assume that $\varphi$ satisfies $(\mathcal{H}_{\alpha,\beta}^{supersmooth})$. Let $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$. 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 some universal constants (independent from $n, d, K$ and $m$):*

*i) Far region and negative part: if $\epsilon = \mathcal{O}\left(\frac{1}{md}\right)$, then:*

$$\mathbb{E}\left[\hat{\mu}_n^-(\mathbb{R}^d)\right] \lesssim K^2 2^{3d/2} \times \frac{e^{\alpha(4d^{1/j}m)^\beta}}{\sqrt{n}} \quad \textit{and} \quad \mathbb{E}\left[\hat{\mu}_n^+(\mathbb{F}(\epsilon))\right] \lesssim K^2 d^3 2^{3d/2} \times \frac{e^{\alpha(4d^{1/j}m)^\beta}}{\sqrt{n}}.$$

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

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

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

$$\mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right] \lesssim 2^{3d/2} K^2 \times \frac{e^{\alpha(4d^{1/j}m)^\beta}}{\sqrt{n}}.$$

According to the results displayed in Proposition 12, our estimation procedure $\hat{\mu}_n$ leads to a consistent estimation as soon as $m$ is chosen as

$$m = \left(\frac{\delta \log n}{\alpha}\right)^{1/\beta} \frac{1}{4d^{1/j}} \quad \text{with} \quad \delta \in \left]0, \frac{1}{2}\right[.$$

In such a case,

$$\max\left(\mathbb{E}\left[\hat{\mu}_n^-(\mathbb{R}^d)\right]\ ,\ \mathbb{E}\left[\hat{\mu}_n^+(\mathbb{F}(\epsilon))\right]\ ,\ \mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right]\right) \lesssim n^{-\frac{1}{2}+\delta},$$

and every set $A$ such that $\mathbb{E}[\hat{\mu}_n^+(A)] \gtrsim n^{-\frac{1}{2}+\delta}$ is at least at a logarithmic distance $(\mathcal{O}(m^{-2}))$ of a true spike.

We observe that as it is commonly observed in severely-ill conditioned inverse problems, we can expect only logarithmic rates of convergence. This logarithmic limitation in the super-smooth situation has been intensively discussed in the literature and we refer among others to [16]. To make the situation more explicit, we illustrate it in the Gaussian mixture model.

5.2.3. *Multivariate Gaussian mixtures.* As a specific case of super-smooth distribution with $\beta = j = 2$ and $\alpha = 1/2$, Proposition 12 holds and we obtain that if $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$ and if $\epsilon = \mathcal{O}(\frac{1}{md})$, then the weights of the far region and of the negative parts are upper bounded by:

$$\mathbb{E}\left[\hat{\mu}_n^-(\mathbb{R}^d)\right] \lesssim K^2 2^{3d/2} \times \frac{e^{8dm^2}}{\sqrt{n}} \quad \text{and} \quad \mathbb{E}\left[\hat{\mu}_n^+(\mathbb{F}(\epsilon))\right] \lesssim K^2 d^3 2^{3d/2} \times \frac{e^{8dm^2}}{\sqrt{n}}. \tag{30}$$

Similarly, a couple of constants $(c, \mathcal{C})$ exists such that:

$$\forall A \subset \mathbb{R}^d, \quad \mathbb{E}[\hat{\mu}_n^+(A)] > cd^3 2^{3d/2} K^2 \times \frac{e^{8dm^2}}{\sqrt{n}} \quad \Longrightarrow \quad \min_{k\in[K]} \inf_{t\in A} \|t - t_k\|_2^2 \leq \frac{1}{\mathcal{C} d^3 m^2}, \tag{31}$$

whereas the weights recovery is ensured by the following inequality: for any $k \in [K]$:

$$\mathbb{E}\left[|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))|\right] \lesssim 2^{3d/2} K^2 \times \frac{e^{8dm^2}}{\sqrt{n}}.$$

- **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 $\Delta$), 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:

$$e^{8dm^2} \ll \sqrt{n} \quad \text{i.e.} \quad m = \mathcal{O}\left(\sqrt{\frac{\log(n)}{d}}\right) \quad \text{and} \quad m \longrightarrow +\infty \quad \text{as } n \longrightarrow +\infty. \tag{32}$$

  In other words, the maximal admissible value for $m$ is $\sqrt{\frac{\log(n)}{16d}}$. In particular, if we consider $m = \sqrt{\frac{\delta}{16}\frac{\log(n)}{d}}$ for $\delta$ small enough, we observe that

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

  The counterpart of this admissible size for $m$ is a slow rate for $\epsilon_n$:

$$\epsilon_n = \mathcal{O}\left(\frac{1}{md}\right) = \frac{\delta^{-1/2}}{\sqrt{d \log n}},$$

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

- **Nature of the results** Item *i)* of Proposition 12 and Equation (30) both indicate that the mass set by $\hat{\mu}_n$ on the negative part and on the far region tends to 0 as the sample size $n$ grows under Condition (32). 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 an 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 (32), Item *ii)* of Proposition 12 and Equation (31) translate the fact that the near region is close to the support of the measure $\mu^0$ 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 and Equation (31), 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

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

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 $m = \mathcal{O}(\sqrt{\delta \log(n)})$ and $\hat{t} \in \hat{S}$, then

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

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 (32), 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 $\Delta$ 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:

  $$m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}.$$

  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, taking the minimal value of $m$ satisfying the previous contraint, we obtain a rate of the order

  $$e^{d^4 K \Delta^{-2}} n^{-1/2}.$$

  We observe that $d$, $K$ and $\Delta^{-1}$ cannot increase faster than a power of $\log(n)$: $d^4 K \Delta^{-2} \ll \log(n)$.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.* [28], [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 $\Delta$ 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 $\Delta_0$, which depends on the mixing distribution $\varphi$. We emphasize that in our work, we only need an upper bound on $K$ and a lower bound on $\Delta$ 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.* Ordinary smooth distributions investigated in this section are described through a polynomial decrease of their Fourier transform. The corresponding deconvolution problem is then said to be mildly ill-posed. In this section, we assume that the density $\varphi$ satisfies

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

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 (see also the end of this section for two examples). 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 $\varphi$ is ordinary smooth and satisfies $(\mathcal{H}_\beta^{smooth})$. Consider $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$. 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 $\epsilon = \mathcal{O}\left(\frac{1}{md}\right)$, then:*

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

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

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

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

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

The proof of this proposition is omitted, and we only comment on the consequences of this result for ordinary smooth mixtures. Provided $K, d$ and $\Delta$ are bounded (or fixed), we obtain a consistent estimation with the BLASSO estimator $\hat{\mu}_n$ when $m$ is chosen such that

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

In such a case, $\epsilon_n = \mathcal{O}(d^{-1} n^{-\delta})$. Now, if $K \vee d \vee \Delta^{-1}$ is allowed to grow towards $+\infty$, setting $m \sim \sqrt{K} d^{3/2} \Delta_+^{-1}$ (the minimal value satisfying the constraint (4)) leads to a bound of order

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

In particular, the maximal order for the dimension is $\mathcal{O}(\log(n))$ as $n \to +\infty$. In the same way, the minimal size of spacings to permit a consistent estimation should not be smaller than $n^{-1/(2\beta)}$. 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 $\varphi$ in our analysis. We present below two specific examples of ordinary smooth mixture density.

*Multivariate Laplace distributions.* In such a case:

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

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 $\beta$. Concerning the number of components $K$, its value should not exceed $n^{1/6}$ and the smallest size of the window $\epsilon_n$ is $n^{-1/4}$.

*Tensor product of Laplace distributions.* Another interesting case is the situation where $\varphi$ 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 be at least of order $\mathcal{O}(\log(n))$.

## 6. Proof of the Main Results.

6.1. *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:

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

<u>Proof of *i*).</u> 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

(33)

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

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

$$\|L\hat{f}_n - L \circ \Phi\hat{\mu}_n\|_{\mathbb{L}}^2 + \kappa D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) + \kappa\int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu) \leq \|\Gamma_n\|_{\mathbb{L}}^2. \tag{34}$$

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

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

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

$$\begin{aligned}\int_{\mathbb{R}^d} \mathcal{P}_m \mathrm{d}(\hat{\mu}_n - \mu^0) &= \int_{\mathbb{R}^d} \langle c_{0,m}, \lambda(t - .) \rangle_{\mathbb{L}} \Phi(\hat{\mu}_n - \mu^0)(t) dt, \\ &= \langle c_{0,m}, L \circ \Phi(\hat{\mu}_n - \mu^0) \rangle_{\mathbb{L}}, \\ &= \langle c_{0,m}, L \circ \Phi \hat{\mu}_n - L\hat{f}_n + \Gamma_n \rangle_{\mathbb{L}}. \end{aligned} \tag{35}$$

Gathering (34) and (35), 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:

$$D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) \leq \frac{3}{2\kappa} \|\Gamma_n\|_{\mathbb{L}}^2 + \frac{3\kappa}{4} \|c_{0,m}\|_{\mathbb{L}}^2, \tag{36}$$

where we have used $\|a + b\|_{\mathbb{L}}^2 \leq 1.5\|a\|_{\mathbb{L}}^2 + 3\|b\|_{\mathbb{L}}^2$ with $a = \Gamma_n$ and $b = -\kappa c_{0,m}/2$ 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 (36) that:

$$\mathbb{E}[D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0)] \leq \frac{3\sqrt{2}}{2} \rho_n \times \|c_{0,m}\|_{\mathbb{L}}. \tag{37}$$

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

$$\begin{aligned}\|\mathcal{P}_m\|_2^2 &= \|\Phi c_{0,m}\|_2^2, \\ &= \int_{\mathbb{R}^d} |\mathcal{F}[\varphi](t)|^2 |\mathcal{F}[c_{0,m}](t)|^2 dt, \\ &= \int_{\mathbb{R}^d} |\mathcal{F}[\varphi](t)|^2 \mathcal{F}[\lambda](t) \times \frac{|\mathcal{F}[c_{0,m}](t)|^2}{\mathcal{F}[\lambda](t)} dt, \\ &\geq \inf_{\|t\|_\infty \leq 4m} \left\{ |\mathcal{F}[\varphi](t)|^2 \mathcal{F}[\lambda](t) \right\} \|c_{0,m}\|_{\mathbb{L}}^2. \end{aligned} \tag{38}$$

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 (38) can be rewritten as:

$$\|c_{0,m}\|_{\mathbb{L}}^2 \leq \frac{\|\mathcal{P}_m\|_2^2}{\inf_{\|t\|_\infty \leq 4m} \{|\mathcal{F}[\varphi](t)|^2 \mathcal{F}[\lambda](t)\}}. \tag{39}$$

We use (37), (39) and observe that $|\mathcal{F}[\varphi]| = \sigma$ to conclude the proof. □

6.2. *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 $k$. 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:

$$\mathbb{E}\left[\int (1 - \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^+ + \int (1 + \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^-\right] \leq \frac{3\sqrt{2}}{2} \rho_n \mathcal{C}_m(\varphi, \lambda). \tag{40}$$

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

$$\mu_n^-(\mathbb{R}^d) = \int \mathrm{d}\hat{\mu}_n^- \leq \int (1 + \mathcal{P}_m) \mathrm{d}\hat{\mu}_n^- \leq \frac{3\sqrt{2}}{2} \rho_n \mathcal{C}_m(\varphi, \lambda). \tag{41}$$

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

$$1 - \mathcal{P}_m(t) \geq \gamma \frac{\upsilon^2}{d^3} \quad \forall t \in \mathbb{F}(\epsilon).$$

Therefore, we obtain that:

$$\hat{\mu}_n^+(\mathbb{F}(\epsilon)) \le \frac{d^3}{\gamma \upsilon^2}\int_{\mathbb{F}(\epsilon)}(1-\mathcal{P}_m)\mathrm{d}\hat{\mu}_n^+ \le \frac{d^3}{\gamma \upsilon^2}\int(1-\mathcal{P}_m)\mathrm{d}\hat{\mu}_n^+. \tag{42}$$

Finally, the first part of $i)$ of Theorem 10 is a direct consequence of (40)-(42).
Weight of the far region. We consider $\gamma$ such that $d^3 \ge \gamma\upsilon^2$ and we know that in the far region:

$$(1-\mathcal{P}_m)\mathbb{1}_{\mathbb{F}(\epsilon)} \ge \frac{\gamma\upsilon^2}{d^3}\mathbb{1}_{\mathbb{F}(\epsilon)}.$$

Thus,

$$\begin{aligned} D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) &= \int(1-\mathcal{P}_m)\mathrm{d}\hat{\mu}_n^+ + \int(1+\mathcal{P}_m)\mathrm{d}\hat{\mu}_n^- \\ &\ge \int_{\mathbb{F}(\epsilon)}\frac{\gamma\upsilon^2}{d^3}\mathrm{d}\hat{\mu}_n^+ + \int_{\mathbb{F}(\epsilon)}1\mathrm{d}\hat{\mu}_n^- \\ &\ge \frac{\gamma\upsilon^2}{d^3}\int_{\mathbb{F}(\epsilon)}\mathrm{d}\hat{\mu}_n^+ + \int_{\mathbb{F}(\epsilon)}\mathrm{d}\hat{\mu}_n^- \\ &\ge \frac{\gamma\upsilon^2}{d^3}\left(\int_{\mathbb{F}(\epsilon)}\mathrm{d}\hat{\mu}_n^+ + \int_{\mathbb{F}(\epsilon)}\mathrm{d}\hat{\mu}_n^-\right) \\ &\ge \frac{\gamma\upsilon^2}{d^3}|\hat{\mu}_n|(\mathbb{F}(\epsilon)). \end{aligned}$$

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

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

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

$$1-\mathcal{P}_m(t) \ge \left[\mathcal{C}m^2\min_{k\in[K]}\|t-t_k\|_2^2 \wedge \frac{\gamma\upsilon^2}{d^3}\right] \quad \forall t\in\mathbb{R}^d.$$

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

$$\begin{aligned} D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0) &\ge \int(1-\mathcal{P}_m)\mathrm{d}\mu_n^+ \\ &\ge \int_A(1-\mathcal{P}_m)\mathrm{d}\mu_n^+, \\ &\ge \left[\mathcal{C}m^2\min_{t\in A}\min_{k\in[K]}\|t-t_k\|^2 \wedge \frac{\gamma\upsilon^2}{d^3}\right]\hat{\mu}_n^+(A). \end{aligned} \tag{43}$$

Equations (40) and (43) lead to:

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

Then,

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

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 $\mu^0$. 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:

$$\begin{aligned}
|a_k^0 - \hat{\mu}_n(\mathbb{N}_k(\epsilon))| &= |a_k^0 - \int \mathcal{Q}_m^k \mathrm{d}\hat{\mu}_n + \int \mathcal{Q}_m^k \mathrm{d}\hat{\mu}_n - \int_{\mathbb{N}_k(\epsilon)} \mathrm{d}\hat{\mu}_n| \\
&\leq \underbrace{|\int \mathcal{Q}_m^k \mathrm{d}(\mu^0 - \hat{\mu}_n)|}_{:=A} + \underbrace{\int_{\mathbb{N}_k(\epsilon)} |\mathcal{Q}_m^k - 1| \mathrm{d}|\hat{\mu}_n|}_{:=B} \\
&\quad + \underbrace{\int_{\mathbb{N}(\epsilon)\setminus\mathbb{N}_k(\epsilon)} |\mathcal{Q}_m^k| \mathrm{d}|\hat{\mu}_n|}_{:=C} + \underbrace{\int_{\mathbb{F}(\epsilon)} |\mathcal{Q}_m^k| \mathrm{d}|\hat{\mu}_n|}_{:=D}. \qquad (44)
\end{aligned}$$

Study of $B + C + D$. On the set $\mathbb{F}(\epsilon)$, we use that $\mathcal{Q}_m^k \leq 1 - \gamma\frac{\upsilon^2}{d^3}$ so that:

$$D \leq \int_{\mathbb{F}(\epsilon)} (1 - \gamma\frac{\upsilon^2}{d^3}) \mathrm{d}|\hat{\mu}_n| \leq \diamond \int_{\mathbb{F}(\epsilon)} (1 - \mathcal{Q}_m^k) \mathrm{d}|\hat{\mu}_n| \quad \text{where} \quad \diamond = \frac{\left(1 - \gamma\frac{\upsilon^2}{d^3}\right)}{\gamma\frac{\upsilon^2}{d^3}}.$$

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:

$$\begin{aligned}
B + C + D &\leq \left(\frac{\widetilde{\mathcal{C}}}{\mathcal{C}} \vee \diamond\right) \int_{\mathbb{R}^d} (1 - \mathcal{P}_m)(t) \mathrm{d}|\hat{\mu}_n|(t) \\
&\leq \left(\frac{\widetilde{\mathcal{C}}}{\mathcal{C}} \vee \diamond\right) \left[\int_{\mathbb{R}^d} (1 - \mathcal{P}_m)(t) \mathrm{d}\hat{\mu}_n^+(t) + \int_{\mathbb{R}^d} (1 + \mathcal{P}_m)(t) \mathrm{d}\hat{\mu}_n^-(t)\right] \\
&\leq \left(\frac{\widetilde{\mathcal{C}}}{\mathcal{C}} \vee \diamond\right) D_{\mathcal{P}_m}(\hat{\mu}_n, \mu^0). \qquad (45)
\end{aligned}$$

Study of $A$. 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 $\Phi$ 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 (33) 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 $\kappa$ 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 $\kappa$, 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:

$$A \leq (2 + \sqrt{3}) \|c_{k,m}\|_{\mathbb{L}} \|\Gamma_n\|_{\mathbb{L}}. \tag{46}$$

Final bound. We use Equations (46) and (45) in the decomposition given in Equation (44) 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 Equation (26) and *ii*) of Proposition 9:

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

□

# 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] L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

[10] Y. De Castro, S. Gadat, C. Marteau, and C. Maugis-Rabusseau. SuperMix: Sparse Regularization for Mixtures. *Supplementary material*, 2019.

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

[12] 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.

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

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

[15] 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.

[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. *Annales de l'Institut Henri Poincaré*, 56:1391–1418, 2020.

[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. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1341–1350, 4 2019.

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

[27] 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.

[28] A. W. van der Vaart. Efficient maximum likelihood estimation in semiparametric mixture models. *The Annals of Statistics*, 24(2):862–878, 1996.

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

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

[31] 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.

[⋆]Institut Camille Jordan, CNRS UMR 5208
École Centrale de Lyon
F-69134 Écully, France.

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

[°] Toulouse School of Economics, CNRS UMR 5314
Université Toulouse 1 Capitole
Esplanade de l'Université, Toulouse, France.
[‡] Institut Universitaire de France.

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