Support Localization and the Fisher Metric for off-the-grid Sparse Regularization

Clarice Poon
University of Bath

Nicolas Keriven
École Normale Supérieure

Gabriel Peyré
École Normale Supérieure

Abstract

Sparse regularization is a central technique for both machine learning (to achieve supervised features selection or unsupervised mixture learning) and imaging sciences (to achieve super-resolution). Existing performance guaranties assume a separation of the spikes based on an ad-hoc (usually Euclidean) minimum distance condition, which ignores the geometry of the problem. In this article, we study the BLASSO (i.e. the off-the-grid version of ℓ^1 LASSO regularization) and show that the Fisher-Rao distance is the natural way to ensure and quantify support recovery, since it preserves the invariance of the problem under reparameterization. We prove that under mild regularity and curvature conditions, stable support identification is achieved even in the presence of randomized sub-sampled observations (which is the case in compressed sensing or learning scenario). On deconvolution problems, which are translation invariant, this generalizes to the multi-dimensional setting existing results of the literature. For more complex translation-varying problems, such as Laplace transform inversion, this gives the first geometry-aware guarantees for sparse recovery.

1 Introduction

1.1 Sparse Regularization

In this work, we consider the general problem of estimating an unknown Radon measure $\mu_0 \in \mathcal{M}(\mathcal{X})$ defined over some metric space $\mathcal{X}$ (for instance $\mathcal{X} = \mathbb{R}^d$ for a possibly large d) from a few number m of randomized linear observations $y \in \mathbb{C}^m$, Let $\Phi : \mathcal{M}(\mathcal{X}) \mapsto \mathbb{C}^m$ be defined by

$$\Phi\mu \overset{\text{def.}}{=} \frac{1}{\sqrt{m}} \left(\int_{\mathcal{X}} \varphi_{\omega_k}(x) \mathrm{d}\mu(x) \right)_{k=1}^{m}, \tag{1}$$

where $(\omega_1, \ldots, \omega_m)$ are identically and independently distributed according to some probability distribution $\Lambda(\omega)$ on $\omega \in \Omega$, and for $\omega \in \Omega$, $\varphi_\omega : \mathcal{X} \to \mathbb{C}$ is a continuous function, denoted $\varphi_\omega \in \mathscr{C}(\mathcal{X})$. We further assume that $\varphi_\omega(x)$ is normalized, that is

$$\mathbb{E}_\omega[|\varphi_\omega(x)|^2] = 1, \qquad \forall x \in \mathcal{X}. \tag{2}$$

The observations are $y = \Phi\mu_0 + w$, where $w \in \mathbb{C}^m$ accounts for noise or modelling errors. Some representative examples of this setting include:

- *Off-the-grid compressed sensing:* off-the-grid compressed sensing, initially introduced in the special case of 1-D Fourier measurements on $\mathcal{X} = \mathbb{T} = \mathbb{R}/\mathbb{Z}$ by (Tang et al., 2013), corresponds exactly to measurements of the form (1). This is a "continuous" analogous of the celebrated compressed sensing line of works (Candès et al., 2006; Donoho, 2006).
- *Regression using a continuous dictionary:* given a set of m training samples $(\omega_k, y_k)_{k=1}^m$, one wants to predicts the values $y_k \in \mathbb{R}$ from the features $\omega_k \in \Omega$ using a continuous dictionary of functions $\omega \mapsto \varphi_\omega(x)$ (here $x \in \mathcal{X}$ parameterizes the dictionary), as $y_k \approx \int_{\mathcal{X}} \varphi_{\omega_k}(x)\mathrm{d}\mu(x)$. A typical example, studied for instance by Bach (2017) is the case of neural networks with a single hidden layer made of an infinite number of neurons, where $\Omega = \mathcal{X} = \mathbb{R}^p$ and one uses ridge functions of the form $\varphi_\omega(x) = \psi(\langle x, \omega \rangle)$, for instance using the ReLu non-linearity $\psi(u) = \max(u, 0)$.
- *Sketching mixtures:* the goal is estimate a (hopefully sparse) mixture of density probability distributions on some domain $\mathcal{T}$ of the form $\xi(t) = \sum_i a_i \xi_{x_i}(t)$ where the $(\xi_x)_{x\in\mathcal{X}}$ is a family of template densities, and $a_i \geqslant 0$, $\sum_i a_i = 1$. Introducing the measure $\mu_0 = \sum_i a_i \delta_{x_i}$, this mixture model is conveniently rewritten as $\xi(t) = \int_{\mathcal{X}} \xi_x(t) \mathrm{d}\mu_0(x)$. The most studied

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

example is the mixture of Gaussians, using (in 1-D for simplicity, $\mathcal{T} = \mathbb{R}$) as $\xi_x(t) \propto \sigma^{-1} e^{-\frac{(t-\tau)^2}{2\sigma^2}}$ where the parameter space is the mean and standard deviation $x = (\tau, \sigma) \in \mathcal{X} = \mathbb{R} \times \mathbb{R}^+$. In a typical machine learning scenario, one does not have direct access to ξ but rather to n i.i.d. samples $(t_1, \ldots, t_n) \in \mathcal{T}^n$ drawn from ξ. Instead of recording this (possibly huge, specially when $\mathcal{T}$ is high dimensional) set of data, following Gribonval et al. (2017), one computes "online" a small set $y \in \mathbb{C}^m$ of m sketches against sketching functions $\theta_\omega(t)$, that is, for $k = 1, \ldots, m$,

$$y_k \overset{\text{def.}}{=} \frac{1}{n} \sum_{j=1}^{n} \theta_{\omega_k}(t_j) \approx \int_{\mathcal{T}} \theta_{\omega_k}(t) \xi(t) \mathrm{d}t.$$

These sketches exactly have the form (1) when defining the functions $\varphi_\omega(x) \overset{\text{def.}}{=} \int_{\mathcal{T}} \theta_\omega(t) \xi_x(t) \mathrm{d}t$. A popular set of sketching functions, over $\mathcal{T} = \mathbb{R}^d$ are Fourier atoms $\theta_\omega(t) \overset{\text{def.}}{=} e^{\mathrm{i}\langle \omega,\, t \rangle}$, for which $\varphi_\cdot(x)$ is the characteristic functions of ξ_x, which can generally be computed in closed form.

BLASSO. In all these applications, and many more, one is actually interested in recovering a discrete and s-sparse measure μ_0 of the form $\mu_0 = \sum_{i=1}^{s} a_i \delta_{x_i}$ where $(x_i, a_i) \in \mathcal{X} \times \mathbb{C}$. An increasingly popular method to estimate such a sparse measure corresponds to solving a infinite-dimensional analogous of the Lasso regression problem

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} \frac{1}{2} \left\| \Phi\mu - y \right\|_2^2 + \lambda |\mu|(\mathcal{X}). \qquad (\mathcal{P}_\lambda(y))$$

Following De Castro and Gamboa (2012), we call this method the BLASSO (for Beurling-Lasso). Here $|\mu|(\mathcal{X})$ is the so-called total variation of the measure μ, and is defined as

$$|\mu|(\mathcal{X}) \overset{\text{def.}}{=} \sup \left\{ \mathrm{Re}\langle f,\, \mu \rangle \; ; \; f \in \mathscr{C}(\mathcal{X}), \|f\|_\infty \leqslant 1 \right\}.$$

Note that on unbounded $\mathcal{X}$, one needs to impose that f vanishes at infinity. If $\mathcal{X} = \{x_i\}_i$ is a finite space, then this corresponds to the classical finite-dimensional Lasso problem (Tibshirani, 1996), because $|\mu|(\mathcal{X}) = \|a\|_1 \overset{\text{def.}}{=} \sum_i |a_i|$ where $a_i = \mu(\{x_i\})$. Similarly, if $\mathcal{X}$ is possibly infinite but $\mu = \sum_i a_i \delta_{x_i}$, one also has that $|\mu|(\mathcal{X}) = \|a\|_1$.

Previous Works. The BLASSO problem $(\mathcal{P}_\lambda(y))$ was initially proposed by De Castro and Gamboa (2012), see also Bredies and Pikkarainen (2013). The first sharp analysis of the solution of this problem is provided by Candès and Fernandez-Granda (2014) in the case of Fourier measurement on $\mathbb{T}^d$. They show that if the spikes are separated enough, then μ_0 is the unique solution of $(\mathcal{P}_\lambda(y))$ when $w = 0$ and $\lambda \to 0$. Robustness to noise under this separation condition is addressed in (Candès and Fernandez-Granda, 2013; Fernandez-Granda, 2013; Azais et al., 2015). A refined stability results is detailed by Duval and Peyré (2015) which shows that conditions based on minimum separation imply *support stability*, which means that when $\|w\|$ and $\|w\|/\lambda$ are small enough, then the solution of $(\mathcal{P}_\lambda(y))$ has the same number of Diracs as μ_0, and that both the amplitudes and positions of the spikes converges smoothly as $w \to 0$. These initial works have been extended by Tang et al. (2013) to the case of randomized compressive measurements of the form (1), when using Fourier sketching functions φ_ω. In all these results, the separation condition are given for the Euclidean cases, which is an ad-hoc choice which does not take into account the geometry of the problem, and gives vastly sub-optimal theories for spatially varying operators (such as data-dependent kernels in supervised learning, Gaussian mixture estimation and Laplace transform in imaging, see Section 1.2).

While this is not the topic of the present paper, note that for positive spikes, the separation condition is in some cases not needed, see for instance (Schiebinger et al., 2015; Denoyelle et al., 2017). It is important to note that efficient algorithms have been developed to solve $(\mathcal{P}_\lambda(y))$, among which SDP relaxations for Fourier measurements (Candès and Fernandez-Granda, 2013) and Frank-Wolfe (also known as conditional gradient) schemes (Bredies and Pikkarainen, 2013; Boyd et al., 2017). Note also that while we focus here on variational convex approaches, alternative methods exist, in particular greedy algorithms (Gribonval et al., 2017) and (for Fourier measurements) Prony-type approaches (Schmidt, 1986; Roy and Kailath, 1989). To the best of our knowledge, their theoretical analysis in the presence of noise is more involved, see however (Liao and Fannjiang, 2016) for an analysis of robustness to noise when a minimum separation holds.

1.2 The Fisher information metric

The empirical covariance operator is defined as $\hat{K}(x, x') \overset{\text{def.}}{=} \frac{1}{m} \sum_i \overline{\varphi_{\omega_i}(x)} \varphi_{\omega_i}(x')$ and the deterministic limit as $m \to +\infty$ is denoted K with

$$K(x, x') \overset{\text{def.}}{=} \int_\Omega \overline{\varphi_\omega(x)} \varphi_\omega(x') \mathrm{d}\Lambda(\omega). \qquad (3)$$

Note that many covariance kernels can be written under the form (3). By Bochner's theorem, this includes all translation-invariant kernels, for which possible features are $\varphi_\omega(x) = e^{\mathrm{i}\omega^\top x}$. The associated metric tensor is

$$\mathbf{H}_x \overset{\text{def.}}{=} \nabla_x \nabla_{x'} K(x, x) \in \mathbb{C}^{d \times d}. \qquad (4)$$

Throughout, we assume that $\mathbf{H}_x$ is positive definite for all $x \in \mathcal{X}$. Then, $\mathbf{H}$ naturally induces a distance between points in our parameter space $\mathcal{X}$. Given a piecewise smooth curve $\gamma : [0,1] \to \mathcal{X}$, the length $\ell_{\mathbf{H}}[\gamma]$ of γ is defined by $\ell_{\mathbf{H}}[\gamma] \overset{\text{def.}}{=} \int_0^1 \sqrt{\langle \mathbf{H}_{\gamma(t)} \gamma'(t), \, \gamma'(t) \rangle} \mathrm{d}t$. Given two points $x, x' \in \mathcal{X}$, the distance from x to x', induced by $\mathbf{H}$ is $d_{\mathbf{H}}(x,x') \overset{\text{def.}}{=} \inf_{\gamma \in \mathcal{F}} \ell_{\mathbf{H}}[\gamma]$ where $\mathcal{F}$ is the set of all piecewise smooth paths $\gamma : [0,1] \to \mathcal{X}$ with $\gamma(0) = x$ and $\gamma(1) = x'$.

The metric $\mathbf{H}$ is closely linked to the Fisher information matrix (Fisher, 1925) associated with Φ: This is clear when φ_ω are real-valued functions, since $f(x,\omega) \overset{\text{def.}}{=} |\varphi_\omega(x)|^2$ can be interpreted as a probability density function for the random variable ω conditional on parameter x, and the metric $\mathbf{H}_x$ is equal (up to rescaling) to its Fisher information matrix, since

$$\int \nabla \left(\log f(x,\omega)\right) \nabla \left(\log f(x,\omega)\right)^\top f(x,\omega) \mathrm{d}\Lambda(\omega)$$
$$= 4 \, \mathbb{E}_\omega [\overline{\nabla \varphi_\omega(x)} \nabla \varphi_\omega(x)^\top] = 4 \mathbf{H}_x .$$

The distance $d_{\mathbf{H}}$ is called the "Fisher-Rao" geodesic distance (Rao, 1945) and is used extensively in information geometry for estimation and learning problems on parametric families of distributions (Amari and Nagaoka, 2007). The Fisher-Rao is the unique Riemannian metric on a statistical manifold (Cencov, 2000) and it is invariant to reparameterization, which matches the invariance of the BLASSO problem ($\mathcal{P}_\lambda(y)$) to reparameterization of the space $\mathcal{X}$. Although $d_{\mathbf{H}}$ has been used in conjunction with kernel methods (see for instance Burges (1999)), to the best of our knowledge, it is the first time this metric is put forward to analyze the performance of off-the-grid sparse recovery problems. In the complex setting, we refer to the notion of the Fubini–Study metric instead (Facchi et al., 2010).

1.2.1 Examples

We detail some popular learning and imaging examples.

The Jackson kernel One of the first seminal result of super-resolution with sparse regularization was given by Candès and Fernandez-Granda (2014) for this kernel, which corresponds to discrete Fourier measurements on the torus. We give a multi-dimensional generalization of this result here. Let $f_c \in \mathbb{N}$, $\mathcal{X} \in \mathbb{T}^d$, $\Omega = \left\{ \omega \in \mathbb{Z}^d \, ; \, \|\omega\|_\infty \leqslant f_c \right\}$. Let $\varphi_\omega(x) \overset{\text{def.}}{=} e^{\mathrm{i}2\pi\omega^\top x}$ and $\Lambda(\omega) \propto \prod_{j=1}^d g(\omega_j)$ where $g(j) = \frac{1}{f_c} \sum_{k=\max(j-f_c,-f_c)}^{\min(j+f_c,f_c)} (1-|k/f_c|)(1-|(j-k)/f_c|)$. Note that this corresponds to sampling *discrete* Fourier frequencies. Then, the associated kernel is the Jackson kernel $K(x,x') = \prod_{i=1}^d \kappa(x_i - x_i')$, where $\kappa(x) \overset{\text{def.}}{=} \mathrm{sinc}^4_{f_c/2+1}(x)$ where $\mathrm{sinc}_s(x) \overset{\text{def.}}{=} s^{-1} \sin(\pi s x)/\sin(\pi x)$, which has a constant metric tensor $\mathbf{H}_x = C_{f_c} \mathrm{Id}$ and $d_{\mathbf{H}}(x,x') = \sqrt{C_{f_c}} \left\| x - x' \right\|_2$ is a scaled Euclidean metric (quotiented by the action of translation modulo 1 on $\mathbb{T}^d$), where $C_{f_c} = -\kappa''(0) = \frac{\pi^2 f_c(f_c+4)}{3}$.

The Gaussian kernel Let $\Sigma \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix, $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Omega = \mathbb{R}^d$. Let $\varphi_\omega(x) = e^{\mathrm{i}\omega^\top x}$ and $\Lambda(\omega) = \mathcal{N}(0, \Sigma^{-1})$, the centered Gaussian distribution with covariance Σ^{-1}. This can be interpreted as sampling *continuous* Fourier frequencies. Then, the associated kernel is $K(x,x') = e^{-\frac{1}{2}\left\| x - x' \right\|^2_{\Sigma^{-1}}}$ where $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$, with constant metric $\mathbf{H}_x = \Sigma^{-1}$, and $d_{\mathbf{H}}(x,x') = \|x - x'\|_{\Sigma^{-1}}$. In Section 3, we also detail how to exploit this kernel for Gaussian Mixture Model (GMM) estimation with the BLASSO.

The Laplace transform Let $\bar\alpha = (\alpha_j) \in \mathbb{R}^d_+$, $\mathcal{X} \subseteq (0, +\infty)^d$ and $\Omega = \mathbb{R}^d_+$. A (sampled) Laplace transform is defined by setting $\varphi_\omega(x) = \prod_{i=1}^d \sqrt{\frac{2(x_i+\alpha_i)}{\alpha_i}} e^{-\langle x, \, \omega \rangle}$ and $\Lambda(\omega) = \prod_{j=1}^d (2\alpha_j) e^{-\langle 2\bar\alpha, \, \omega \rangle}$. Then, $K(x,x') = \prod_{i=1}^d \kappa(x_i + \alpha_i, x_i' + \alpha_i)$ where $\kappa(a,b) = \frac{2\sqrt{ab}}{a+b}$, with metric $\mathbf{H}_x$ as the diagonal matrix with diagonal $\left((2(x_i + \alpha_i))^{-2} \right)_{i=1}^d$ and distance $d_{\mathbf{H}}(x,x') = \sqrt{\sum_i \left| \log\left(\frac{x_i + \alpha_i}{x_i' + \alpha_i} \right) \right|^2}$. We remark that this kernel, associated to the Laplace transform (which should not be confused with the translation-invariant Laplace kernel $\exp(-\left\| x - x' \right\|)$) appears in some microscopy imaging technique, see for instance Boulanger et al. (2014). Unlike the previous examples, it is not translation-invariant, and therefore the metric $\mathbf{H}_x$ is not constant. Our results show that the corresponding Fisher metric is the natural way to impose the separation condition in super-resolution.

1.3 Contributions.

Our main contribution is Theorem 1, which states that if the sought after spikes positions X_0 are sufficiently separated with respect to the Fisher distance $d_{\mathbf{H}}$, then the solution to ($\mathcal{P}_\lambda(y)$) is *support stable* (that is, the solution of the BLASSO is formed of exactly s Diracs) provided that the number of random noisy measurements m is, up to log factors and under the assumption of random signs of the amplitudes a_0, linear in s, and the noise level $\|w\|$ is less than $1/s$. In the case of translation invariant kernels, this generalizes existing results to a large class of multi-dimensional kernels, and also provides for the first time a quantitative bounds on the impact of the noise and sub-sampling on the spikes positions and amplitudes errors. For non-translation kernels, this provides for the first time a meaningful support recovery guarantee, a typical example being the Laplace kernel (see Section 1.2).

2 Key concepts

Notation for derivatives. Given $f \in \mathscr{C}^\infty(\mathcal{X})$, by interpreting the r^{th} derivative as a multilinear map: $\nabla^r f : (\mathbb{C}^d)^r \to \mathbb{C}$, so given $Q \overset{\text{def.}}{=} \{q_\ell\}_{\ell=1}^r \in (\mathbb{C}^d)^r$,

$$\nabla^r f[Q] = \sum_{i_1,\cdots,i_r} \partial_{i_1}\cdots\partial_{i_r} f(x) q_{1,i_1}\cdots q_{r,i_r}.$$

and we define the r^{th} normalized derivative of f as

$$\mathrm{D}_r\left[f\right](x)[Q] \overset{\text{def.}}{=} \nabla^r f(x)[\{\mathbf{H}_x^{-\frac{1}{2}} q_i\}_{i=1}^r]$$

with norm $\|\mathrm{D}_r\left[f\right](x)\| \overset{\text{def.}}{=} \sup_{\forall \ell, \|q_\ell\| \leqslant 1} |\mathrm{D}_r\left[f\right](x)[Q]|$. For $i, j \in \{0, 1, 2\}$, let $K^{(ij)}(x, x')$ be a "bi"-multilinear map, defined for $Q \in (\mathbb{C}^d)^i$ and $V \in (\mathbb{C}^d)^j$ as

$$[Q]K^{(ij)}(x,x')[V] \overset{\text{def.}}{=} \mathbb{E}[\overline{\mathrm{D}_i\left[\varphi_\omega\right](x)[Q]}\mathrm{D}_j\left[\varphi_\omega\right](x')[V]]$$

and $\left\|K^{(ij)}(x,x')\right\| \overset{\text{def.}}{=} \sup_{Q,V}\left\|[Q]K^{(ij)}(x,x')[V]\right\|$ where the supremum is defined over all $Q \overset{\text{def.}}{=} \{q_\ell\}_{\ell=1}^i$, $V \overset{\text{def.}}{=} \{v_\ell\}_{\ell=1}^j$ with $\|q_\ell\| \leqslant 1$, $\|v_\ell\| \leqslant 1$. Note that $\mathrm{D}_2\left[f\right](x)$ and $K^{(02)}(x, x')$ can also be interpreted as a matrix in $\mathbb{C}^{d\times d}$, and we have the normalization $K^{(02)}(x, x) = -\mathrm{Id}$ for all x.

2.1 Admissible kernel and separation

In previous studies on the recovery properties of $(\mathcal{P}_\lambda(y))$ (Candès and Fernandez-Granda, 2014; Bhaskar et al., 2013; Bendory et al., 2016; Duval and Peyré, 2015; Fernandez-Granda, 2016), recovery bounds are attained in the context of K being *admissible* and a separation condition on the underlying positions $\{x_j\}_j$. Namely, given $X = \{x_j\}_j$, that $\min_{i\neq j} d_{\mathbf{H}}(x_i, x_j)$ is sufficiently large with respect to the decay properties of K. For example, in the case where Φ corresponds to Fourier sampling on a grid, up to frequency f_c, this separation condition is $\min_{j\neq\ell}\|x_j - x_\ell\|_2 \gtrsim 1/f_c$. In fact, if $\mathrm{sign}(a_j)$ can take arbitrary values in $\{+1, -1\}$, this separation condition is a necessary to ensure exact recovery for the BLASSO (Tang, 2015).

Following the aforementioned works, we introduce the notion of an admissible kernel.

Definition 1 *A kernel K will be said* admissible *with respect to $\mathcal{K} \overset{\text{def.}}{=} \{r_{\text{near}}, \Delta, \varepsilon_i, B_{ij}, s_{\max}\}$, where $0 < r_{\text{near}} < \Delta/4$ is a neighborhood size, $\varepsilon_0 \in (0, 1)$, $\varepsilon_2 \in (0, r_{\text{near}}^{-2})$ are respectively a distance to 1 and a curvature, $\Delta > 0$ is a minimal separation, $B_{ij} > 0$ for $i, j = 0, \ldots, 2$ are some constants and $s_{\max} \in \mathbb{N}^*$ is a maximal sparsity level, if*

1. ***Uniform bounds**: For $(i, j) \in \{(0, 0), (1, 0)\}$, $\sup_{x,x'\in\mathcal{X}}\|K^{(ij)}(x, x')\| \leqslant B_{ij}$; for $(i, j) \in \{(0, 2), (1, 1), (1, 2)\}$ and all x, x' such that $d_{\mathbf{H}}(x, x') \leqslant r_{\text{near}}$ or $d_{\mathbf{H}}(x, x') > \Delta/4$, $\|K^{(ij)}(x, x')\| \leqslant B_{ij}$; and finally, $\sup_{x\in\mathcal{X}}\left\|K^{(22)}(x, x)\right\| \leqslant B_{22}$.*
2. ***Neighborhood of each point**: For all $x \in \mathcal{X}$, $K(x, x) = 1$ and for all $x, x' \in \mathcal{X}$ with $d_{\mathbf{H}}(x, x') \leqslant r_{\text{near}}$, $\mathrm{Re}\left(K^{(02)}(x, x')\right) \preccurlyeq -\varepsilon_2\mathrm{Id}$ and $\left\|\mathrm{Im}\left(K^{(02)}(x, x')\right)\right\| \leqslant c\varepsilon_2$, where $c \overset{\text{def.}}{=} \frac{1}{2}\sqrt{\frac{2-\varepsilon_2 r_{\text{near}}^2}{\varepsilon_2 r_{\text{near}}^2}}$ and for $d_{\mathbf{H}}(x, x') \geqslant r_{\text{near}}$, $|K(x, x')| \leqslant 1 - \varepsilon_0$.*
3. ***Separation**: For $d_{\mathbf{H}}(x, x') \geqslant \Delta/4$, for all $i, j \in \{0, \ldots, 2\}$ with $i + j \leqslant 3$, $\|K^{(ij)}(x, x')\| \leqslant \frac{h}{s_{\max}}$, where $h \overset{\text{def.}}{=} \min_{i\in\{0,2\}}\left(\frac{\varepsilon_i}{32B_{1i}+32}, \frac{5\varepsilon_2}{16B_{12}+24}\right)$.*

Additionally, there exists $C_{\mathbf{H}} \geqslant 0$ such that for $d_{\mathbf{H}}(x, x_0) \leqslant r_{\text{near}}$: $\left\|\mathrm{Id} - \mathbf{H}_{x_0}^{-\frac{1}{2}}\mathbf{H}_x^{\frac{1}{2}}\right\| \leqslant C_{\mathbf{H}} d_{\mathbf{H}}(x, x_0)$. We also denote $d_{\mathbf{H}}(X, X_0) = \sqrt{\sum_i d_{\mathbf{H}}(x_i, x_{0,i})^2}$ and $B \overset{\text{def.}}{=} \sum_{i+j\leqslant 3} B_{ij}$ and $\varepsilon \overset{\text{def.}}{=} \min\{\varepsilon_0, \varepsilon_2\}$.

Intuitively, these three conditions express the following facts: 1) the kernel and its derivatives are uniformly bounded, 2) near $x = x'$, the kernel has negative curvature, and otherwise it is strictly less than 1, and 3) for x and x' sufficiently separated, the kernel and all its derivatives have a small value.

2.2 Almost bounded random features

Ideally, we would like our features and its derivatives to be uniformly bounded for all ω. However this may not be the case: think of $e^{i\omega^\top x}$ where the support of the distribution Λ is not bounded. Hence our results will be dependent on the probability that the derivatives are greater than some value T decays sufficiently quickly as T increases. In the following, for $r \in \{0, 1, 2, 3\}$, $L_r(\omega) \overset{\text{def.}}{=} \sup_{x\in\mathcal{X}}\|\mathrm{D}_r\left[\varphi_\omega\right](x)\|$, and let F_r be such that $\mathbb{P}_\omega\left(L_r(\omega) > t\right) \leqslant F_r(t)$.

2.3 Key assumptions

Our main result will be valid under the following assumptions.

I. On the domain and limit kernel Let $\mathcal{X}$ be a compact domain with radius $R_{\mathcal{X}} \overset{\text{def.}}{=} \sup_{x,x'\in\mathcal{X}} d_{\mathbf{H}}(x, x')$. Assume the kernel is admissible wrt $\mathcal{K} \overset{\text{def.}}{=} \{r_{\text{near}}, \Delta, \varepsilon_i, B_{ij}, s_{\max}\}$.

II. Assumption on the underlying signal For $s \leqslant s_{\max}$, let $a_0 \in \mathbb{C}^s$ and let $X_0 \overset{\text{def.}}{=} (x_{0,j})_{j=0}^s$ be such that $d_{\mathbf{H}}(x_{0,i}, x_{0,j}) \geqslant \Delta$ for $i \neq j$. The underlying measure is assumed to be $\mu_0 = \sum_{j=1}^s a_{0,j}\delta_{x_{0,j}}$.

III. Assumption on the sampling complexity For $\rho > 0$, suppose that $m \in \mathbb{N}$ and $\{\bar{L}_i\}_{i=0}^3 \in \mathbb{R}_+^4$ are chosen such that

$$\begin{gathered}\sum_{j=0}^{3} F_j(\bar{L}_j) \leqslant \frac{\rho}{m}, \qquad \text{and} \\ \max_{j=0}^{3}\{\bar{L}_j^2 \sum_{i=0}^{3} F_i(\bar{L}_i) + 2\int_{\bar{L}_j}^{\infty} tF_j(t)\mathrm{d}t\} \leqslant \frac{\varepsilon}{m},\end{gathered} \tag{5}$$

and either one of the following hold:

$$m \gtrsim C \cdot s \cdot \log\left(N^d/\rho\right) \log\left(s/\rho\right), \tag{6}$$

$$\text{or} \quad m \gtrsim C \cdot s^{3/2} \cdot \log\left(N^d/\rho\right), \tag{7}$$

where $C \overset{\text{def.}}{=} \varepsilon^{-2}(\bar{L}_2^2 B_{11} + \bar{L}_1^2 B_{22} + (B_0 + B_2)\bar{L}_{01}^2)$, $N \overset{\text{def.}}{=} \mathbb{L}_3 d R_{\mathcal{X}}(r_{\text{near}}\varepsilon)^{-1}$ and $\mathbb{L}_r = \max_{i=1}^r \bar{L}_i$.

Remark 1 *Our main theorem presents support stability guarantees under the sampling complexity rate* (6) *if* $\operatorname{sign}(a_0) = (a_{0,i}/|a_{0,i}|)_{i=1}^s$ *forms a Steinhaus sequence, that is, iid uniformly distributed on the complex unit circle. This assumption has been used before in compressed sensing (Candès and Romberg, 2007; Tang et al., 2013) to achieve optimal complexity (see also Foucart and Rauhut (2013), Chap. 14). As noted in previous works, this random signs assumption is likely to be a proof artefact, however achieving optimal complexity without it may require more involved arguments (Candes and Plan, 2011). When the signs are arbitrary, we prove our results under* (7). *Although this* $s^{3/2}$ *scaling is still sub-optimal in* s*, we remark it improves upon the previous theoretical rate of* s^2 *(up to log factors) (Li and Chi, 2017).*

Remark 2 *The assumption on the choice of* $\bar{L}_r$ *ensures that with high probability,* $\mathrm{D}_r[\varphi_\omega](x)$ *is uniformly bounded up to* $r = 3$*. In general,* $\{\bar{L}_r\}$ *depend on* m*, through* (5)*. However, in all our examples: either a)* $\sup_{x\in\mathcal{X}} \|\mathrm{D}_r[\varphi_\omega](x)\|$ *are already uniformly bounded, in which case* $\bar{L}_i$ *can be chosen independently of* ρ *and* m *(for instance this is the case for the Jackson kernel); or b) the* $F_r(t)$ *are exponentially decaying, in which case we can show that* $\bar{L}_r = \mathcal{O}(\log(m/\rho)^p)$ *for some* $p > 0$*, which only incurs additional logarithmic terms on the bounds* (6) *and* (7)*. This is the case for the Gaussian or Laplace transform kernel.*

3 Main result

Our main theorem below states quantitative exact support recovery bounds under a minimum separation condition according to $d_{\mathbf{H}}$.

Theorem 1 *Let* $\rho > 0$*, suppose that* K *is admissible, and that* a_0*,* X_0*,* m *and* $\bar{L}_i$ *satisfy the assumptions of Section 2.3. Let* $\mathcal{D}_{\lambda_0,c_0} \overset{\text{def.}}{=} \{(\lambda, w) \in \mathbb{R}_+ \times \mathbb{C}^m\ ;\ \lambda \leqslant \lambda_0,\ \|w\| \leqslant c_0\lambda\}$ *where* $c_0 \sim \min\left(\frac{\varepsilon_0}{\bar{L}_0}, \frac{\varepsilon_2}{\bar{L}_2}\right)$ *and* $\lambda_0 \sim D/s$ *with*

$$D \overset{\text{def.}}{=} \underline{a} \min\left(r_{\text{near}}\sqrt{s},\ \tfrac{\varepsilon\sqrt{s}}{\mathbb{L}_2^2\|a\|},\ \tfrac{\varepsilon}{C_{\mathbf{H}}(B+\mathbb{L}_2^2)}\right) \tag{8}$$

where $\underline{a} = \min\{|a_{0,i}|^2, |a_{0,i}|^{-2}\}$*. Suppose that either* $\operatorname{sign}(a_0)$ *is a Steinhaus sequence and* m *satisfies* (6) *or* $\operatorname{sign}(a_0)$ *is an arbitrary sign sequence and* m *satisfies* (7)*. Then, with probability at least* $1 - \rho$*,*

(i) for all $v \overset{\text{def.}}{=} (\lambda, w) \in \mathcal{D}_{\lambda_0,c_0}$*,* $(\mathcal{P}_\lambda(y))$ *has a unique solution which consists of exactly* s *spikes. Moreover, up to a permutation of indices, the solution can be written as* $\sum_{i=1}^s a_i^v \delta_{x_i^v}$*.*

(ii) The mapping $v \in \mathcal{D}_{\lambda_0,c_0} \mapsto (a^v, X^v)$ *is* $\mathscr{C}^1$ *and we have the error bound*

$$\|a^v - a_0\| + d_{\mathbf{H}}(X^v, X_0) \leqslant \tfrac{\sqrt{s}(\lambda+\|w\|)}{\min_i|a_{0,i}|} \tag{9}$$

We detail below the values relating to the sampling complexity corresponding to each of the examples detailed in Section 1.2.1. The corresponding proofs can be found in Section F of the appendix.

Discrete Fourier sampling The Fejer kernel of order $f_c \geqslant 128$ is admissible with $\Delta = \mathcal{O}(\sqrt{d}\sqrt[4]{s_{\max}})$, $r_{\text{near}} = 1/(8\sqrt{2})$, $\varepsilon_0 = 0.00097$, $\varepsilon_2 = 0.941$, $B_{01} = \mathcal{O}(d)$, $B_{11} = B_{02} = B_{12} = \mathcal{O}(1)$ and $B_{22} = \mathcal{O}(d)$. Moreover, $\bar{L}_r = \mathcal{O}(d^{r/2})$. Hence, up to logarithmic terms, Thm. 1 is applicable with $m = \mathcal{O}(sd^3)$ when the random signs assumption holds, and $m = \mathcal{O}(s^{\frac{3}{2}}d^3)$ in the general case, with guaranteed support stability when $\lambda = \mathcal{O}(s^{-1}d^{-2})$, $\|w\| = \mathcal{O}(s^{-1}d^{-3})$. Note that our choice of Δ imposes that $\|x_i - x_j\|_2 \gtrsim \sqrt{d}s_{\max}^{1/4}/f_c$ whereas the previous result of Candès and Fernandez-Granda (2014) requires $\|x_i - x_j\|_\infty \gtrsim C_d/f_c$ with no dependency in $s_{\max}$, however, their proof would imply that the constant C_d grows exponentially in d. Since we are interested in having a general theory in arbitrary dimension, we have opted to present a polynomial dependency on $s_{\max}$.

Continuous Gaussian Fourier sampling In the appendix we prove that the kernel is admissible with $\Delta = \mathcal{O}\left(\sqrt{\log s_{\max}}\right)$, $r_{\text{near}} = 1/\sqrt{2}$, $\varepsilon_0 = 1 - e^{-\frac{1}{4}}$, $\varepsilon_2 = e^{-\frac{1}{4}}/2$, $B_{ij} = \mathcal{O}(1)$ for $i + j \leqslant 3$, $B_{22} = \mathcal{O}(d)$ and $\bar{L}_r = \left(d + \log\left(\frac{dm}{\rho}\right)^2\right)^{\frac{r}{2}}$ (as mentioned before, the dependence in m only incurs additional logarithmic factors in (6) and (7)). Hence, up to log factors, the sample complexity and noise level for the application of Thm. 1 is the same as for the Fejér kernel.

Laplace sampling The associated kernel is admissible with $\Delta = \mathcal{O}\left(d + \log(ds_{\max})\right)$, $r_{\text{near}} = 0.2$, $\varepsilon_0 = 0.005$, $\varepsilon_2 = 1.52$, $B_{ij} = \mathcal{O}(1)$ for $i + j \leqslant 3$ and $B_{22} = \mathcal{O}(d)$. Define $\bar{R}_{\mathcal{X}} = \left(1 + \frac{R_{\mathcal{X}}}{\min_i \alpha_i}\right)^d$ (where we recall that $R_{\mathcal{X}}$ is the radius of $\mathcal{X}$). Assuming for simplicity that all α_j are distinct, we can set $\bar{L}_r = \bar{R}_{\mathcal{X}}(R_{\mathcal{X}} + \|\alpha\|_\infty)^r \left(\sqrt{d} + \max_i \frac{1}{\alpha_i} \log\left(\frac{d\beta_i m \bar{R}_{\mathcal{X}}}{\rho \alpha_i}\right)\right)^r$ Hence, choosing $\alpha_i \sim d$, we have that $\bar{R}_{\mathcal{X}} = (1)$ and up to log factors, (6) is $\mathcal{O}(sd^7)$ and (7) is $\mathcal{O}(s^{3/2}d^7)$, and support stability is guaranteed when $\lambda = \mathcal{O}(s^{-1}d^{-3})$ and $\|w\| = \mathcal{O}(s^{-1}d^{-5})$. Note that despite the stronger dependency on d, for practical applications (microscopy), one is typically only interested in the low dimensional setting of $d = 2, 3$.

Gaussian mixture learning Consider n datapoints $z_1, \ldots, z_n \in \mathbb{R}^d$ drawn *iid* from a mixture of Gaussians $\sum_i a_{0,i}\mathcal{N}(x_{0,i}, \Sigma)$ with means $x_{0,i} \in \mathcal{X} \subset \mathbb{R}^d$ and known covariance Σ, where $\mathcal{X}$ is bounded. Consider the following procedure:

- draw ω_j *iid* from $\mathcal{N}(0, \Sigma^{-1}/d)$ (the $1/d$ normalization is necessary to avoid an exponential dependency in d later on)
- compute the generalized moments $y = \frac{1}{\sqrt{m}} \sum_{i=1}^n (e^{i\langle \omega_j, x_i\rangle})_{j=1}^m$
- solve the BLASSO with features $\varphi_\omega(x) = e^{i\langle \omega, x\rangle} e^{-\frac{1}{2}\|\omega\|_\Sigma^2}$, to obtain a distribution $\tilde{\mu}$

Then, as described in the introduction, we can interpret y as noisy Fourier measurements of $\mu_0 = \sum_i a_{0,i}\delta_{x_{0,i}}$ *in the space of means* $\mathcal{X}$, where the "noise" w corresponds to using the empirical average over the z_i instead of a true integration. It is easily bounded with probability $1-\rho$ by $\|w\| \leqslant \mathcal{O}\left(\sqrt{\frac{\log(1/\rho)}{n}}\right)$, by a simple application of Hoeffding's inequality (Gribonval et al., 2017).

The associated kernel is the Gaussian kernel with covariance $(2 + d)\Sigma$ and hence, our result states that, if $\|x_i - x_j\|_{\Sigma^{-1}} \geqslant \sqrt{d \log s}$, and the number of measurements and sample complexity satisfy, up to logarithmic terms, $m = \mathcal{O}\left(s^{\frac{3}{2}} d^3\right)$, $n = \mathcal{O}\left(s^2 d^6 / \min_i |a_{0,i}|^2\right)$ and $\lambda_0 = \mathcal{O}\left(\frac{\min_i |a_{0,i}|}{\sqrt{s} d^2 \|a_0\|_2}\right)$, then, with probability $1 - \rho$ on both samples z_j and frequencies ω_j, the distribution $\tilde{\mu}$ is formed of exactly s Diracs, and their positions and weights converge to the means and weights of the GMM. Let us give a few remarks on this result.

On model selection. Besides convexity (with respect to the distribution of means) of the BLASSO, which is not the case of classical likelihood- or moments-based methods for learning GMM, the most striking feature of our approach is probably the support stability: with a sample complexity that is polynomial in s and d, the BLASSO yields *exactly* the right number of components for the GMM. Despite the huge literature on model selection for GMM, to our knowledge, this is one of the only result which is *non-asymptotic* in sample complexity, as opposed to many approaches (Roeder and Wasserman, 1997; Huang et al., 2013) which guarantee that the selected number of components approaches the correct one when the number of samples grows to infinity.

On separation condition. Our separation condition of $\sqrt{d \log s}$ is, up to the logarithmic term, similar to the $\sqrt{d}$ found in the seminal work by Dasgupta (1999). This was later improved by different methods (Dasgupta and Schulman, 2000; Vempala and Wang, 2004), until the most recent results on the topic (Moitra and Valianty, 2010) show that it is possible to learn a GMM with *no* separation condition, provided the sample complexity is exponential in s, which is a necessary condition (Moitra and Valianty, 2010). As mentioned in the introduction, similar results exist for the BLASSO: Denoyelle et al. (2017) showed that in one dimension, one can identify s positive spikes with no separation, provided the noise level is exponentially small with s. Hence learning GMM with the BLASSO and no separation condition may be feasible, which we leave for future work, however we note that the multi-dimensional case is still largely an open problem (Poon and Peyré, 2017).

On known covariance. An important path for future work is to handle arbitrary covariance. When the components all share the same mean and have diagonal covariance, the Fisher metric is related, up to a change of variables, to the Laplace transform kernel case treated earlier. When both means and covariance vary, in one dimension, the Fisher metric is related to the Poincaré half-plane metric (Costa et al., 2015). In the general case, it does not have a closed-form expression. We leave the treatment of these cases for future work.

4 Sketch of proof

4.1 Background on dual certificates

Our approach to establishing that the solutions to $(\mathcal{P}_\lambda(y))$ are support stable is via the study of the associated dual solutions in accordance to the framework introduced in Duval and Peyré (2015). We first recall some of their key ideas. In order to study the support stability properties of $(\mathcal{P}_\lambda(y))$ in the small noise regime, we consider the limit problem as $\lambda \to 0$ and $\|w\| \to 0$, that is

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

The dual of $(\mathcal{P}_\lambda(y))$ and $(\mathcal{P}_0(y))$ are

$$\min_p \left\{ \|y/\lambda - p\|_2^2 \ ; \ \|\Phi^* p\|_\infty \leqslant 1 \right\} \qquad (\mathcal{D}_\lambda(y))$$

$$\max_p \left\{ \langle y, \, p\rangle \ ; \ \|\Phi^* p\|_\infty \leqslant 1 \right\}. \qquad (\mathcal{D}_0(y))$$

Any solution μ_λ of $(\mathcal{P}_\lambda(y))$ to related to the (unique) solution p_λ of $(\mathcal{D}_\lambda(y))$ by $-p_\lambda = \frac{1}{\lambda}(\Phi\mu_\lambda - y)$ and writing $\eta_\lambda \overset{\text{def.}}{=} \Phi^* p_\lambda$, $\langle \eta_\lambda, \, \mu_\lambda \rangle = |\mu_\lambda|(\mathcal{X})$. Note that $\mathrm{Supp}(\mu_\lambda) \subseteq \{x \in \mathcal{X} \ ; \ |\Phi^* p_\lambda(x)| = 1\}$, so η_λ "certifies" the support of μ_λ and is often referred to as a *dual certificate*. Furthermore, by defining the minimal norm certificate η_0 as $\eta_0 \overset{\text{def.}}{=} \Phi^* p_0$ where

$$p_0 = \mathrm{argmin}\left\{ \|p\|_2 \ ; \ p \text{ is a solution to } (\mathcal{D}_0(y)) \right\} \quad (10)$$

one can show that p_λ converges as $\lambda \to 0$ to p_0 and hence η_λ converges to $\eta_0 \overset{\text{def.}}{=} \Phi^* p_0$ in L^∞. When λ and $\|w\|$ are sufficiently small, solutions to $(\mathcal{P}_\lambda(y))$ are support stable provided that η_0 (called the minimal norm certificate) is *nondegenerate*, that is $\eta_0(x_i) = \mathrm{sign}(a_i)$ for $i = 1, \ldots, s$ and $\nabla^2 |\eta_0|^2 (x_i)$ is negative definite. This is proven to be an almost sharp condition for support stability, since Duval and Peyré (2017) provided explicit examples where $|\eta_0(x)| = 1$ for some $x \notin \{x_i\}_i$ implies that $(\mathcal{P}_\lambda(y))$ recovers more than s spikes under arbitrarily small noise.

Pre-certificates In practice, the minimal norm certificate η_0 is hard to compute and analyse due to the nonlinear ℓ^∞ constraint in (10). So, one often introduces a proxy which can be computed in closed form by solving an linear system associated to the following least squares problem: $\eta_{X,a} \overset{\text{def.}}{=} \Phi^* p_{X,a}$ where

$$p_{X,a} \overset{\text{def.}}{=} \mathrm{argmin}\{\|p\|_2 \ ; \ (\Phi^* p)(x_i) = \mathrm{sign}(a_i), \; \nabla(\Phi^* p)(x_i) = 0\}. \quad (11)$$

Note that if $\eta_{X,a}$ satisfies $\|\eta_{X,a}\|_\infty \leqslant 1$, then $\eta_{X,a} = \eta_0$.

Computation of $\eta_{X,a}$ For $x \in \mathcal{X}$, let $\varphi(x) \overset{\text{def.}}{=} \frac{1}{\sqrt{m}} \left(\varphi_{\omega_k}(x)\right)_{k=1}^m$. For $X = \{x_i\}_{i=1}^s$ we define $\Gamma_X : \mathbb{C}^{s(d+1)} \to \mathbb{C}^m$ as $\Gamma_X([\alpha, \beta]) \overset{\text{def.}}{=} \sum_{i=1}^s \alpha_i \varphi(x_i) + \nabla\varphi(x_i)^\top \beta_i$ where $\nabla\varphi \in \mathbb{C}^{m \times d}$. Then, the minimizer of (11) is $p_{X,a} = \Gamma_X^{*,\dagger} \binom{\mathrm{sign}(a)}{\mathbf{0}_{sd}}$. Furthermore, when Γ_X is full rank, we can write $\hat{\eta}_{X,a}(x) \overset{\text{def.}}{=} \sum_i \hat{\alpha}_i \hat{K}(x_i, x) + \langle \hat{\beta}_i, \, \nabla_1 \hat{K}(x_i, x)\rangle$, where $\hat{\alpha}_i \in \mathbb{C}$, $\hat{\beta}_i \in \mathbb{C}^d$ are such that $\binom{\hat{\alpha}}{\hat{\beta}} = (\Gamma_X^* \Gamma_X)^{-1} \binom{\mathrm{sign}(a)}{0_{sd}}$, and the hat notation refers to the fact that we are using sub-sampled measurements. The limit precertificate is defined as $\eta_{X,a}(x) \overset{\text{def.}}{=} \sum_i \alpha_i K(x_i, x) + \langle \beta_i, \, \nabla_1 K(x_i, x)\rangle$, where $\binom{\alpha}{\beta} = (\mathbb{E}[\Gamma_X^* \Gamma_X])^{-1} \binom{\mathrm{sign}(a)}{0_{sd}}$.

The key to establishing our recovery results is to show that $\hat{\eta}_{X,a}$ is nondegenerate. In this paper, we will actually prove a stronger notion of nondegeneracy:

Definition 2 *Let $a \in \mathbb{C}^s$, $X = \{x_i\}_{i=1}^s \in \mathcal{X}^s$ for some $s \in \mathbb{N}$, and $\varepsilon_0, \varepsilon_2, r > 0$. We say that $\eta \in \mathscr{C}^1(\mathcal{X})$ is $(\varepsilon_0, \varepsilon_2)$-nondegenerate with respect to a, X and r if for all i, $\eta(x_i) = \mathrm{sign}(a_i)$, $\nabla\eta(x_i) = 0$ and*

$$\forall\, x \in \mathcal{X}^{\text{far}}, \ |\eta(x)| \leqslant 1 - \varepsilon_0$$
$$\forall\, x \in \mathcal{X}_j^{\text{near}}, |\eta(x)| \leqslant 1 - \varepsilon_2 d_{\mathbf{H}}(x, x_j)^2$$

where $\mathcal{X}_j^{\text{near}} \overset{\text{def.}}{=} \{x \in \mathcal{X} \ ; \ d_{\mathbf{H}}(x_i, x) \leqslant r\}$ and $\mathcal{X}^{\text{far}} \overset{\text{def.}}{=} \mathcal{X} \setminus \bigcup_{j=1}^s \mathcal{X}_j^{\text{near}}$.

Our proof proceeds in three steps:

1. Show that under admissibility of the kernel and sufficient separation, the limit precertificate η_{X_0,a_0} is non-degenerate (see Theorem 2).
2. Show that this non-degeneracy transfers to $\hat{\eta}_{X,a}$ when m is large enough, a is close to a_0 and X is close to X_0. This is the purpose of Section 4.3.
3. As discussed, nondegeneracy of $\hat{\eta}_{X_0,a_0}$ automatically guarantees support stability when $(\lambda, w) \in \mathcal{D}_{\lambda_0,c_0}$ for λ_0 and c_0 *sufficiently small.* To conclude we simply need to quantify λ_0 and c_0. This is the purpose of Section 4.4. In particular, given (λ, w), we construct a candidate solution by means of (a quantitative version of) the Implicit Function Theorem, and show that it is indeed a true solution using the previous results.

4.2 Non-degeneracy of the limit certificate

Our first result shows that the "limit precertificate" η_{X_0,a_0} is nondegenerate:

Theorem 2 *Assume the kernel is admissible wrt $\mathcal{K}$ (see Definition 1). Then, for $s \leqslant s_{\max}$, for all $a_0 \in \mathbb{C}^s$ and $X = \{x_j\}_{j=1}^s \in \mathcal{X}^s$ such that $d_{\mathbf{H}}(x_i, x_j) \geqslant \Delta$, the function η_{X_0,a_0} is $(\frac{\varepsilon_0}{2}, \frac{\varepsilon_2}{2})$-nondegenerate with respect to a_0, X and r_{near}.*

The proof of this result can be found in Appendix B and is a generalization of the arguments of Candès and Fernandez-Granda (2014) (see also Bendory et al. (2016)). We remark that unlike previous works which focus on translation invariant kernels, the Fisher metric provides a natural way to understand the required separation between the points in X_0 and thus open up the possibility of analysing more complex problems such as Laplace transform inversion.

4.3 The randomized setting

For the remainder of this paper, we consider solutions of $(\mathcal{P}_\lambda(y))$ given $y = \Phi\mu_{a_0,X_0} + w$ for some fixed $a_0 \in \mathbb{C}^s$ and $X_0 \in \mathcal{X}^s$. The following result shows that $\hat{\eta}_X$ is nondegenerate for all X close to X_0:

Theorem 3 *Let $\rho > 0$. Under the assumptions of Section 2.3, and assuming that either m satisfies* (6) *and* $\operatorname{sign}(a_0)$ *is a Steinhaus sequence, or m satisfies* (7) *and* $\operatorname{sign}(a_0)$ *is an arbitrary sign sequence, with probability at least $1-\rho$: for all $a \in \mathbb{C}^s$ and $X \in \mathcal{X}^s$ such that*

$$d_{\mathbf{H}}(X, X_0) \lesssim \min\left(r_{\text{near}}, \frac{\varepsilon}{C_{\mathbf{H}}\sqrt{s}\max\left(B, \bar{L}_{12}\bar{L}_r\right)}\right), \quad (12)$$

and $\|a - a_0\| \lesssim \frac{\varepsilon}{B}\min_i |a_{0,i}|$, Γ_X is full rank and $\hat{\eta}_{X,a}$ is $(\frac{\varepsilon_0}{8}, \frac{\varepsilon_2}{8})$-nondegenerate with respect to a, X and r_{near}.

The proof of this result is given in Appendix D. We simply make a remark on the proof here: We first prove that $\hat{\eta}_{X_0,a_0}$ is nondegenerate by bounding variations between η_{X_0,a_0} and $\hat{\eta}_{X_0,a_0}$. The proof of this fact is a generalization of the arguments in Tang et al. (2013) to the multidimensional and general operator case. We then exploit the fact the φ is smooth and hence, $\Gamma_X^*\Gamma_X$ satisfies certain Lipschitz properties with respect to X, to bound the local variation between $\hat{\eta}_{X,a}$ and $\hat{\eta}_{X_0,a_0}$.

4.4 Quantitative support recovery

This final section concludes the proof of Theorem 1 by quantifying the regions for λ and $\|w\|$ on which support stability is guaranteed.

Solution of the noisy BLASSO. Let $\Phi_X : \mathbb{C}^s \to \mathbb{C}^m$ be defined by $\Phi_X a = \sum_{i=1}^s a_i\varphi(x_i)$. Recall that $\mu_{a,X} = \sum_i a_i\delta_{x_i}$ is a solution to the BLASSO with $y = \Phi\mu_{a_0,X_0} + w$ if and only if $\hat{\eta}_\lambda = \Phi^* p_\lambda$, with $p_\lambda = \frac{1}{\lambda}(y - \Phi_X a)$, satisfies $\|\hat{\eta}_\lambda\|_\infty \leqslant 1$ and $\hat{\eta}(x_j) = \operatorname{sign}(a_j)$. In that case, p_λ is the *unique* solution to the dual of the BLASSO. Moreover, if $|\hat{\eta}_\lambda(x)| < 1$ for $x \neq x_i$ and Φ_X is full rank (which follows by Theorem 3), then $\mu_{a,X}$ is also the unique solution of the primal.

Construction of a solution Following Denoyelle et al. (2017), we define the function $f : \mathbb{R}^{2s} \times \mathcal{X}^s \times \mathbb{R}_+ \times \mathbb{R}^{2m} \to \mathbb{R}^{2s} \times \mathbb{C}^{sd}$ by

$$f(u,v) \stackrel{\text{def.}}{=} \begin{pmatrix} \operatorname{Re}\left(\Phi_X^*(z)\right) \\ \operatorname{Im}\left(\Phi_X^*(z)\right) \\ (\Phi_X^{(1)})^*(z) \end{pmatrix} + \lambda \begin{pmatrix} \operatorname{Re}\left(\operatorname{sign}(a)\right) \\ \operatorname{Im}\left(\operatorname{sign}(a)\right) \\ 0_{sd} \end{pmatrix}$$

where $u \stackrel{\text{def.}}{=} (a_{\mathrm{r}}, a_{\mathrm{i}}, X)$, $v \stackrel{\text{def.}}{=} (\lambda, w_{\mathrm{r}}, w_{\mathrm{i}})$, $a \stackrel{\text{def.}}{=} a_{\mathrm{r}} + \mathrm{i}a_{\mathrm{i}}$, $w \stackrel{\text{def.}}{=} w_{\mathrm{r}} + \mathrm{i}w_{\mathrm{i}}$ and $z \stackrel{\text{def.}}{=} (\Phi_X a - \Phi_{X_0}a_0 - w)$. Observe that having $f(u,v) = 0$ ensures the existence of $\hat{\eta}_\lambda$ defined as above that satisfies $\hat{\eta}_\lambda(x_i) = \operatorname{sign}(a_{0,i})$ and $\nabla\hat{\eta}_\lambda(x_i) = 0$. We will use it to construct a nondegenerate solution to $\mathcal{D}_\lambda(y)$ for small λ and $\|w\|$. Now, f is continuously differentiable, with explicit forms of $\partial_v f(u,v)$ and $\partial_u f(u,v)$ given in (E.2) and (E.3) in the appendix, and in particular, $\partial_u f(u_0, 0)$ is invertible and $f(u_0, 0) = 0$. Hence, by the Implicit Function Theorem, there exists a neighbourhood V of 0 in $\mathbb{R} \times \mathbb{R}^{2m}$, a neighbourhood U of u_0 in $\mathbb{R}^{2s} \times \mathcal{X}^s$ and a differentiable function $g : V \to U$ such that for all $(u,v) \in U \times V$, $f(u,v) = 0$ if and only if $u = g(v)$. So, to establish support stability for $(\mathcal{P}_\lambda(y))$, we simply need to estimate the size of the neighbourhood V on which g is well defined, and given $(\lambda, w) \in V$, for $(a, Z) = g((\lambda, w))$, to check that the associated certificate $\hat{\eta}_{\lambda,w} \stackrel{\text{def.}}{=} \Phi^* p_{\lambda,w}$ with $p_{\lambda,w} \stackrel{\text{def.}}{=} \frac{1}{\lambda}\left(\Phi_X a - \Phi_{X_0}a_0 - w\right)$ is nondegenerate.

Indeed, one can prove (see Theorem E.1) that with probability at least $1-\rho$, V contains the ball $B_r(0)$ with radius $r \sim \frac{1}{\sqrt{s}}\min\left(\frac{\min\{r_{\text{near}}, (C_{\mathbf{H}}B)^{-1}\}}{\min_i|a_{0,i}|}, \frac{1}{\bar{L}_{01}\bar{L}_{12}(1+\|a_0\|)}\right)$ and given any $v \in B_r(0)$, $(a, X) = g(v)$ indeed satisfy the error bound (9).

Checking that the candidate solution is a true solution It remains to check that $g(\lambda, w)$ defines a valid certificate and is non-degenerate (and hence, $\sum_i a_i\delta_{x_i}$ is the unique solution to $(\mathcal{P}_\lambda(y))$) provided that λ, w satisfy (8). Given $(\lambda, w) \in V$, let $(a, X) = g((\lambda, w))$. Define $\hat{\eta}_{\lambda,w} \stackrel{\text{def.}}{=} \frac{-1}{\lambda}\Phi^*(\Phi_X a - \Phi_{X_0}a_0 - w)$ and one can show (see Lemma E.1) that

$$\hat{\eta}_{\lambda,w} = \hat{\eta}_{X,a} + \varphi(\cdot)^\top\Pi_X\frac{w}{\lambda} + \frac{1}{\lambda}\varphi(\cdot)^\top\Pi_X\Phi_{X_0}a_0$$

where Π_X is the orthogonal projection onto $\operatorname{Im}(\Gamma_X)^\perp$. Note that since we have the error bound (9), our choice of λ and $\|w\|$ ensures that (12) holds and hence, Theorem 3 implies that $\hat{\eta}_{X,a}$ is nondegenerate with probability at least $1-\rho$. To conclude, it is suffices to bound the two remaining terms so that $\hat{\eta}_{\lambda,w}$ remains non-degenerate. Under $\bar{E}$, $\|\mathrm{D}_r[\varphi_\omega](\cdot)\| \leqslant \bar{L}_r$, and for any $z \in \mathbb{C}^m$, $\left\|\mathrm{D}_r\left[\varphi^\top z\right]\cdot\right\| \leqslant \bar{L}_r\|z\|$. Therefore, since Π_X is a projection, we have $\left\|\mathrm{D}_r\left[\varphi(\cdot)^\top\Pi_X\frac{w}{\lambda}\right]\right\| \lesssim \varepsilon_r$ when $\|w\|/\lambda \lesssim \varepsilon_r/\bar{L}_r$. Finally, since $\Phi_{X_0}a_0 = \sum_{j=1}^s a_{0,j}\varphi(x_{0,j})$, by Taylor expansion of $\varphi(x_{0,j})$ around x_j and applying Π_X (see Lemma E.2 for this computation), we have

$$\|\Pi_X\Phi_{X_0}a_0\| \leqslant \bar{L}_2\|a_0\|_\infty d_{\mathbf{H}}(X, X_0)^2.$$

Since g satisfies (9) our choice of $\lambda_0 = \mathcal{O}(s^{-1})$ ensures that we can upper bound this by $\bar{L}_2\|a_0\|_\infty\frac{s\left(\lambda+\|w\|^2/\lambda\right)}{\min|a_{0,i}|^2} \lesssim \varepsilon$ and consequently, $\frac{1}{\lambda}\left\|\mathrm{D}_r\left[\varphi(\cdot)^\top\Pi_X\Phi_{X_0}a_0\right]\right\| \lesssim \varepsilon_r$.

Acknowledgements

We would like to thank Ben Adcock for a helpful conversation regarding the stochastic gradient bounds. This work was partly funded by the CFM-ENS chair "Modèles et Sciences des données" and the European Research Council, NORIA project.

References

S.-i. Amari and H. Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

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

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

T. Bendory, S. Dekel, and A. Feuer. Robust recovery of stream of pulses using convex optimization. *Journal of mathematical analysis and applications*, 442(2):511–536, 2016.

B. N. Bhaskar, G. Tang, and B. Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

J. Boulanger, C. Gueudry, D. Münch, B. Cinquin, P. Paul-Gilloteaux, S. Bardin, C. Guérin, F. Senger, L. Blanchoin, and J. Salamero. Fast high-resolution 3D total internal reflection fluorescence microscopy by incidence angle scanning and azimuthal averaging. *Proceedings of the National Academy of Sciences*, 111(48):17164–17169, 2014.

N. Boyd, G. Schiebinger, and B. Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.

K. Bredies and H. K. Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

C. J. Burges. Geometry and invariance in kernel based methods. 1999.

E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

E. J. Candès and C. Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

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.

E. J. Candes and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

N. N. Cencov. *Statistical decision rules and optimal inference*. Number 53. American Mathematical Soc., 2000.

S. I. Costa, S. A. Santos, and J. E. Strapasson. Fisher information distance: A geometrical reading. In *Discrete Applied Mathematics*, volume 197, pages 59–69. Elsevier B.V., 2015.

S. Dasgupta. Learning mixtures of Gaussians. In *IEEE 51st Annual Symposium on Foundations of Computer Science*, number May, 1999.

S. Dasgupta and L. J. Schulman. A Two-Round Variant of EM for Gaussian Mixtures. *Uncertainty in Artificial Intelligence*, pages 152–159, 2000.

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

Q. Denoyelle, V. Duval, and G. Peyré. Support recovery for sparse super-resolution of positive measures. *to appear in Journal of Fourier Analysis and Applications*, 2017.

D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

V. Duval and G. Peyré. Sparse spikes super-resolution on thin grids I: the LASSO. *Inverse Problems*, 33(5):055008, 2017.

P. Facchi, R. Kulkarni, V. Man'ko, G. Marmo, E. Sudarshan, and F. Ventriglia. Classical and quantum fisher information in the geometrical formulation of quantum mechanics. *Physics Letters A*, 374(48):4801–4803, 2010.

C. Fernandez-Granda. Support detection in super-resolution. *Proc. Proceedings of the 10th International Conference on Sampling Theory and Applications*, pages 145–148, 2013.

C. Fernandez-Granda. Super-resolution of point sources via convex programming. *Information and Inference: A Journal of the IMA*, 5(3):251–303, 2016.

R. A. Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press, 1925.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer New York, NY, 2013.

R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Compressive statistical learning with random feature moments. *arXiv preprint arXiv:1706.07180*, 2017.

T. Huang, H. Peng, and K. Zhang. Model Selection for Gaussian Mixture Models. *Statistica Sinica*, pages 1–27, 2013.

Y. Li and Y. Chi. Stable separation and super-resolution of mixture models. *Applied and Computational Harmonic Analysis*, 2017.

W. Liao and A. Fannjiang. MUSIC for single-snapshot spectral estimation: Stability and super-resolution. *Applied and Computational Harmonic Analysis*, 40(1):33–67, 2016.

A. Moitra and G. Valianty. Settling the polynomial learnability of mixtures of Gaussians. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 93–102, 2010.

C. Poon and G. Peyré. Multi-dimensional Sparse Super-resolution. pages 1–42, 2017.

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normal. *JASA*, 92:894–902, 1997.

R. Roy and T. Kailath. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on acoustics, speech, and signal processing*, 37(7):984–995, 1989.

G. Schiebinger, E. Robeva, and B. Recht. Superresolution without separation. *arXiv preprint arXiv:1506.03144*, 2015.

R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

G. Tang. Resolution limits for atomic decompositions via markov-bernstein type inequalities. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 548–552. IEEE, 2015.

G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.