The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem

Yann Traonmilin, Jean-François Aujol

The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem

Yann Traonmilin[1,2,*] and Jean-François Aujol[2]

[1]CNRS,

[2]Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251,F-33400 Talence, France.

Abstract

The sparse spike estimation problem consists in estimating a number of off-the-grid impulsive sources from under-determined linear measurements. Information theoretic results ensure that the minimization of a non-convex functional is able to recover the spikes for adequately chosen measurements (deterministic or random). To solve this problem, methods inspired from the case of finite dimensional sparse estimation where a convex program is used have been proposed. Also greedy heuristics have shown nice practical results. However, little is known on the ideal non-convex minimization to perform. In this article, we study the shape of the global minimum of this non-convex functional: we give an explicit basin of attraction of the global minimum that shows that the non-convex problem becomes easier as the number of measurements grows. This has important consequences for methods involving descent algorithms (such as the greedy heuristic) and it gives insights for potential improvements of such descent methods.

1 Introduction

1.1 Context

Sums of sparse off-the-grid spikes can be used to model impulsive sources in signal processing (e.g. in astronomy, microscopy,...). Estimating such signals from a finite number of measurements is known as the super-resolution problem [8]. In the space $\mathcal{M}$ of finite signed measure over $\mathbb{R}^d$, we aim at recovering $x_0 = \sum_{i=1,k} a_i \delta_{t_i}$ from the measurements

$$y = Ax_0 + e, \tag{1}$$

where δ_{t_i} is the Dirac measure at position t_i, the operator A is a linear observation operator, $y \in \mathbb{C}^m$ are the m noisy measurements and e is a finite energy observation noise. Recent works have shown that it is possible to estimate spikes from a finite

*Contact author : yann.traonmilin@math.u-bordeaux.fr

number of adequately chosen Fourier measurements as long as their locations are sufficiently separated, using convex minimization based variational methods in the space of measures [7, 1, 24, 11, 13]. Other general studies on inverse problems have shown that an ideal non-convex method (unfortunately computationally inefficient) can be used to recover these signals as long as the linear measurement operator has a restricted isometry property (RIP) [4]. In the case of super-resolution, adequately chosen random compressive measurements have been shown to meet the sufficient RIP conditions for separated spikes, thus guaranteeing the success of the ideal non-convex decoder [18]. These RIP results are based on an adequate kernel metric on $\mathcal{M}$. It must be noted that, according to the work of [4], the success of the convex decoders as described in [7] for regular Fourier sampling implies a (lower) restricted isometry property of A with respect to such a kernel metric (and not with the natural total variation metric: in this case no RIP is possible with finite regular Fourier measurements, see e.g. [5]). Greedy heuristics have also been proposed to approach the non-convex minimization problem and they have shown good practical utility [19, 20, 25].

While giving theoretical recovery guarantees, the convex-based method is non-convex in the space of parameters (amplitudes and locations) due to a polynomial root finding step. Also, it is difficult to implement in dimensions larger than one in practice [12]. Greedy heuristics based on orthogonal matching pursuit are implemented in higher dimension (they can practically be used up to $d = 50$), but they still miss theoretical recovery guarantees [19]. It would be possible to overcome the limitations of such methods if it were possible to perform the ideal non-convex minimization:

$$x^* \in \underset{x \in \Sigma}{\operatorname{argmin}} \|Ax - y\|_2 \tag{2}$$

where Σ is a low-dimensional set modeling the separation constraints on the k Diracs. While simple in its formulation, properties of this minimization procedure have not yet been thoroughly studied.

In this article, as a first important step towards the understanding of the non-convex sparse spike estimation problem (2), we study its formulation in the parameter space (the space of amplitudes and locations of the Diracs). We observe that a smooth non-convex optimization can be performed. We place ourselves in a context where the number of measurements, either deterministic or random, guarantees the success of the ideal non-convex decoder with respect to a kernel metric $\|\cdot\|_h$, i.e. when we can ensure that:

$$\|x^* - x_0\|_h \leq C\|e\|_2, \tag{3}$$

where C is an absolute constant with respect to e and $x_0 \in \Sigma_{k,\epsilon}$, the set $\Sigma_{k,\epsilon}$ is the set of sums of k spikes separated by ϵ on a given bounded domain. Qualitatively, the kernel metric can be viewed as a measure of the energy at a given resolution set by a kernel h (see Section 2.3).

The bound (3) is guaranteed by a restricted isometry property of A defined using such kernel metric [18]. This RIP setting is verified in the deterministic (see Section 2.3) and random Fourier measurement contexts [18].We link this RIP of measurement operators with the conditioning of the Hessian of the global minimum, and we give an explicit basin of attraction of the global minimum. This study has direct consequences for the theoretical study of greedy approaches. Indeed a basin of attraction permits to give recovery guarantees for descent methods (the initialization must fall within the basin), since the gradient descent is a step in the iterations of the greedy approach.

1.2 Parametrization of the model set Σ

Let $\Sigma \subset \mathcal{M}$ a model set (union of subspaces) and $x_0 \in \Sigma$. Let $f(x) = \|Ax - y\|_2$.

Definition 1.1 (Local minimum in Σ)**.** *The point x is a local minimum of f in Σ if there is $\epsilon > 0$ such that for any $x' \in \Sigma$ such that $\|x - x'\| \leq \epsilon$, we have $f(x) \leq f(x')$.*

Definition 1.2 (Parametrization of Σ)**.** *A parametrization of Σ is a function such that $\Sigma \subset \phi(\mathbb{R}^d) = \{\phi(\theta) : \theta \in \mathbb{R}^d\}$.*

We consider the problem

$$\theta^* \in \arg\min_{\theta \in E} g(\theta) = \arg\min_{\theta \in E} \|A\phi(\theta) - y\|_2. \tag{4}$$

where $E = \mathbb{R}^{k(d+1)}$ or $E = \Theta := \phi^{-1}(\Sigma)$ and $g(\theta) = f(\phi(\theta))$.

In the following, we consider the model of separated Diracs:

$$\begin{aligned} \Sigma = \Sigma_{k,\epsilon} := \{\phi(\theta) = \sum_{r=1,k} a_r \delta_{t_r} : \theta = (a, t_1, .., t_k) \in \mathbb{R}^{k(d+1)}, a \in \mathbb{R}^k, t_r \in \mathbb{R}^d, \\ \forall r \neq l, \|t_r - t_l\|_2 > \epsilon, t_r \in \mathcal{B}_2(R)\}, \end{aligned} \tag{5}$$

where

$$\mathcal{B}_2(R) = \{t \in \mathbb{R}^d : \|t\|_2 \leq R\}. \tag{6}$$

Note that, in this paper, the Dirac distribution could be supported on any compact set. We use $\mathcal{B}_2(R)$ for the sake of simplicity. For $t_r \in \mathbb{R}^d$, we write $t_r = (t_{r,j})_{j=1,d}$.

We consider the following Parametrization of $\Sigma_{k,\epsilon}$: $\sum_{i=1,k} a_i \delta_{t_i} = \phi(\theta)$ with $\theta = (a_1, .., a_k, t_1, .., t_k)$. We define

$$\Theta_{k,\epsilon} := \phi^{-1}(\Sigma_{k,\epsilon}). \tag{7}$$

Note that when $E = \Theta_{k,\epsilon}$, performing minimization (4) allows to recover the minima of the ideal minimization (2), yielding stable recovery guarantees. Hence we are particularly interested in this case. When $E = \mathbb{R}^{k(d+1)}$, we speak about unconstrained minimization for minimization (4).

The objective of this paper is to study the shape of the basin of attraction of the global minimum of (4) when $E = \Theta_{k,\epsilon}$.

1.3 Basin of attraction and descent algorithms

In this work, we are interested in minimizing g defined in (4). Since g is a smooth function, a classical method to minimize g is to consider a fixed step gradient descent. The algorithm is the following. Consider an initial point $\theta_0 \in \mathbb{R}^d$ and a step size $\tau > 0$. We define by recursion the sequence θ_n by

$$\theta_{n+1} = \theta_n - \tau \nabla g(\theta_n) \tag{8}$$

We now give the definition of basin of attraction that we will use in this paper.

Definition 1.3 (Basin of attraction)**.** *We say that a set $\Lambda \subset \mathbb{R}^d$ is a basin of attraction of g if there exists $\theta^* \in \Lambda$ and $\tau > 0$, such that if $\theta_0 \in \Lambda$ then the sequence θ_n defined by (8) converges to θ^*.*

This definition of basin of attraction is related to the following classical optimization result (see e.g. [10]):

Proposition 1.1. *Assume g to be a smooth coercive convex function, whose gradient is L Lipschitz. Let $\theta_0 \in \mathbb{R}^d$. Then, if $\tau < \frac{1}{L}$, there exists $\theta^* \in \mathbb{R}^d$ such that the sequence θ_n defined by (8) converges to θ^*.*

An immediate consequence of the previous proposition is the following corollary.

Corollary 1.1. *Assume g to be a smooth function. Assume that g has a minimizer $\theta^* \in \mathbb{R}^d$. Assume that there exists an open set $\Lambda \subset \mathbb{R}^d$ such that $\theta^* \in \Lambda$, g is convex on Λ with L Lipschitz gradient. Then, if $\theta_0 \in \Lambda$ and $\tau < \frac{1}{L}$, the sequence θ_n defined by (8) converges to θ^*.*

Remark 1.1. *Assume that g is $\mathcal{C}^2$. Let $\lambda_{\max}(t)$ the largest eigenvalue of the Hessian matrix of $g(t)$. Let $\Theta \subset \mathbb{R}^d$ an open set. If there exists $L > 0$ such that for all t in Θ, $\lambda_{\max}(t) \leq L$, then g has a L Lipschitz gradient in Θ.*

1.4 Related work

While original for the sparse spike estimation problem, it must be noted that the study of non-convex optimization schemes for linear inverse problems has gained attraction recently for different kinds of low-dimensional models. For low-rank matrix estimation, a smooth Parametrization of the problem is possible and it has been shown that a RIP guarantees the absence of spurious minima [27, 2]. In [26], a model for phase recovery with alternated projections and smart initialization is considered. Conditions on the number of measurements guarantee the success of the technique. In the area of blind deconvolution and bi-convex programming, recent works have exploited similar ideas [21, 6].

In the case of super-resolution, the idea of gradient descent has been studied in an asymptotic regime ($k \to \infty$) in [9] with theoretical conditions based on Wasserstein gradient flow for the initialization. In our case, we study the particular super-resolution problem with a fixed number of impulsions and we place ourselves in conditions when stable recovery is guaranteed, leading to explicit conditions on the initialization.

The objective of this article is to investigate to what extent these ideas can be applied to the theoretical study of the case of spike super-resolution estimation.

The question of projected gradient descent raised in the last Section has been explored for general low-dimensional models [3]. It has been shown that the RIP guarantees the convergence of such algorithms with an ideal (often non practical) projection. Approached projected gradient descents have also been studied and shown to be successful for some particular applications [16]. The spikes super-resolution problem adds the Parametrization step to these problems.

1.5 Contributions and organization of the paper

After a precise description of the setting, the definition of the kernel metric of interest and the associated restricted isometry for the spike estimation problem at the beginning of Section 2, this article gives the following original results:

1. A bound on the conditioning of the Hessian at a global minimum of the minimization in the parameter space is given in Section 2. This bound shows that the better RIP constants are (RIP constants improve with respect to the number of measurements), the better the non-convex minimization problem behaves. It also shows that there is a basin of attraction of the global optimum where no separation constraints are needed (for descent algorithms with an initialization close to the minimum, separation constraints can be discarded)

2. An explicit shape of the basin of attraction of global minima is given in Section 3. The size of the basin of attraction increases when the RIP constant gets better.

To conclude, we discuss the role of the separation constraint in descent algorithms in Section 4, and we explain why enforcing a separation might improve them.

2 Conditioning of the Hessian

This section is devoted to the study of the Hessian matrix of g. In particular, we provide a bound on the conditioning of the Hessian at a global minimum of the minimization in the parameter space.

2.1 Notations

The operator A is a linear operator modeling m measurements in $\mathbb{C}^m$ ($\mathrm{Im}A \subset \mathbb{C}^m$) on the space of measures on $\mathbb{R}^d$ defined by: for $l = 1, m$,

$$(Au)_l = \int_{\mathbb{R}^d} \alpha_l(t) \, \mathrm{d}u(t) \tag{9}$$

where $(\alpha_l)_l$ is a collection of functions in $\mathcal{C}^2(\mathcal{B}_2(R))$ (twice continuously differentiable functions on $\mathcal{B}_2(R)$ defined in (6)).

Notice that the integral used in (9) is in fact a duality product between a function in $\mathcal{C}^2(\mathcal{B}_2(R))$ and a finite signed measure over $\mathbb{R}^d$. As the α_l are in $\mathcal{C}^2(\mathcal{B}_2(R))$, we can similarly apply A to distributions of order 1 and 2.

While a lot of results for spike super-resolution are expressed on the Torus $\mathbb{T}$, we prefer the setting of Diracs with bounded support on $\mathbb{R}^d$ which is often closer to the physics of the considered phenomenom. However, our work is directly extended to the Torus setting by replacing $\mathbb{R}^d$ by $\mathbb{T}$ and $\mathcal{B}^2(R)$ by $\mathcal{T}$.

In $\mathbb{C}^m$, we consider the Hermitian product $\langle u, v\rangle = \sum u_i \bar{v}_i$. An example of such measurement operator is the Fourier sampling: $(Au)_l = \frac{1}{\sqrt{m}} \int_{\mathbb{R}^d} e^{-j\langle \omega_l, t\rangle} \,\mathrm{d}u(t)$ for some chosen $\omega_l \in \mathbb{R}^d$.

Let $x = \sum_{i=1,k} a_i \delta_{t_i}$. By linearity of A, we have

$$(Ax)_l = \sum_{i=1}^{k} (A\delta_{t_i})_l = \sum_{i=1}^{k} a_i \alpha_l(t_i). \tag{10}$$

With $g(\theta) = f(\phi(\theta)) = \|A\phi(\theta) - y\|_2^2$, we get:

$$g(\theta) = \sum_{l=1}^{m} \left| \sum_{i=1}^{k} a_i \alpha_l(t_i) - y_l \right|^2. \tag{11}$$

In the following, the notion of directional derivative will be important.

Definition 2.1 (Directional derivatives)**.** *Let f be a C^1 function, and $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$. Then we can define the directional derivative of f in direction v by:*

$$f'_v(t) := \langle v, \nabla f(t)\rangle = \lim_{h\to 0^+} \frac{f(t+hv) - f(t)}{h} \tag{12}$$

Let f be a $\mathcal{C}^2$ function, and $(v_1, v_2) \in \mathbb{R}^{2d}$ such that $\|v_1\|_2 = \|v_2\|_2 = 1$. Then we can define the second order directional derivative of f in directions v_1 and v_2 by:

$$f''_{v_1,v_2}(t) := \langle v_1, \nabla^2 f(t) v_2\rangle \tag{13}$$

Notice that of course $f''_{v_1,v_2}(t) = f''_{v_2,v_1}(t)$. If $v_1 = v_2$, we write $f''_{v_1}(t) := f''_{v_1,v_1}(t)$

In particular, they permit to introduce derivatives of Dirac measures supported on $\mathbb{R}^d$.

Definition 2.2 (Directional derivatives of Dirac)**.** *Let $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$. The distribution $\delta'_{t_0,v}$ is defined by $\int_{\mathbb{R}} f(t)\,\mathrm{d}\delta'_{t_0,v}(t) = -f'_v(t_0)$. It is the limit of $\nu_\eta = -\frac{\delta_{t_0+\eta v} - \delta_{t_0}}{\eta}$ for $\eta \to 0^+$ in the distributional sense : for all $h \in \mathcal{C}^1(\mathbb{R}^d)$, $\int_{\mathbb{R}} h(t)\,\mathrm{d}\nu_\eta(t) \to_{\eta\to 0^+} \int_{\mathbb{R}} h(t)\,\mathrm{d}\delta'_{t_0,v}(t)$.*

Similarly, The distribution $\delta''_{t_0,v}$ is defined by $\int_{\mathbb{R}} f(t)\,\mathrm{d}\delta''_{t_0,v}(t) = f''_v(t_0)$ for $f \in \mathcal{C}^2(\mathbb{R}^d)$ and the distribution δ''_{t_0,v_1,v_2} is defined by $\int_{\mathbb{R}} f(t)\,\mathrm{d}\delta''_{t_0,v_1,v_2}(t) = f''_{v_1,v_2}(t_0)$ for $f \in \mathcal{C}^2(\mathbb{R}^d)$ where f''_{v_1,v_2} is the derivative of f in direction v_1 chained with the derivative of f in direction v_2.

When $v = e_i$ is a vector of the canonical basis of $\mathbb{R}^d$, we write $\delta'_{t_0,i} = \delta'_{t_0,e_i}$ and $\delta''_{t_0,i} = \delta''_{t_0,e_i,e_i}$.

Finally, we note $\mathrm{rint}\mathcal{B}_2(R)$ the relative interior of $\mathcal{B}_2(R)$.

We now have the necessary tools to start the study of the Hessian of g.

2.2 Gradient and Hessian of the objective function g

We calculate the gradient and Hessian of g in the two following propositions. We start with the gradient of g.

Proposition 2.1. *For any $\theta \in \mathbb{R}^{2k}$, we have:*

$$\frac{\partial g(\theta)}{\partial a_r} = 2\mathcal{R}e\langle A\delta_{t_r}, A\phi(\theta) - y\rangle, \tag{14}$$

$$\frac{\partial g(\theta)}{\partial t_{r,j}} = -2a_r\mathcal{R}e\langle A\delta'_{t_r,j}, A\phi(\theta) - y\rangle. \tag{15}$$

Proof. See Appendix A.1. □

The next proposition gives the values of the Hessian matrix of g which has a simple expression with the use of derivatives of Diracs.

Proposition 2.2. *For any $\theta \in \mathbb{R}^{k(d+1)}$*

$$H_{1,r,s} = \frac{\partial^2 g(\theta)}{\partial a_r \partial a_s} = 2\mathcal{R}e\langle A\delta_{t_r}, A\delta_{t_s}\rangle. \tag{16}$$

$$\begin{aligned} H_{2,r,j_1,s,j_2} = \frac{\partial^2 g(\theta)}{\partial t_{r,j_1}\partial t_{s,j_2}} &= 2a_r a_s\mathcal{R}e\langle A\delta'_{t_r,j_1}, A\delta'_{t_s,j_2}\rangle \\ &\quad + \mathbf{1}(r=s)2a_r\mathcal{R}e\langle A\delta''_{t_r,j_1,j_2}, A\phi(\theta) - y\rangle. \end{aligned} \tag{17}$$

$$H_{12,r,s,j} = \frac{\partial^2 g(\theta)}{\partial a_r \partial t_{s,j}} = -2a_s\mathcal{R}e\langle A\delta_{t_r}, A\delta'_{t_s,j}\rangle - \mathbf{1}(r=s)2\mathcal{R}e\langle A\delta'_{t_0,j}, A\phi(\theta) - y\rangle. \tag{18}$$

Hence the Hessian can be decomposed as the sum of two matrices $H = G + F$ with

$$\begin{aligned} G_{1,r,s} &= 2\mathcal{R}e\langle A\delta_{t_r}, A\delta_{t_s}\rangle, \\ G_{2,r,j_1,s,j_2} &= 2a_r a_s\mathcal{R}e\langle A\delta'_{t_r,j_1}, A\delta'_{t_s,j_2}\rangle, \\ G_{12,r,s,j} &= -2a_s\mathcal{R}e\langle A\delta_{t_r}, A\delta'_{t_s,j}\rangle. \end{aligned} \tag{19}$$

and

$$\begin{aligned} F_{1,r,s} &= 0, \\ F_{2,r,j_1,s,j_2} &= \mathbf{1}(r=s)2a_r\mathcal{R}e\langle A\delta''_{t_r,j_1,j_2}, A\phi(\theta) - y\rangle, \\ F_{12,r,s,j} &= -\mathbf{1}(r=s)2\mathcal{R}e\langle A\delta'_{t_0,j}, A\phi(\theta) - y\rangle. \end{aligned} \tag{20}$$

Proof. See Appendix A.1. □

2.3 Kernel, dipoles and the RIP

In order to be able to build an operator A with a RIP, we define a reproducible kernel Hilbert space (RKHS) structure on the space of measures as in [18], see also [23]. The natural metric on the space of finite signed measures, the total variation of measures, is not well suited for a RIP analysis of the spikes super-resolution problems, as it does not measure the spacing between Diracs. When using the RIP, fundamental objects appear in the calculations: dipoles of Diracs. In this section we show that the typical RIP implies a RIP on dipoles and their generalization.

Definition 2.3 (Kernel, scalar product and norm). *For finite signed measures over $\mathbb{R}^d$, the Hilbert structure induced by a kernel h (a smooth function from $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$) is defined by the following scalar product between 2 measures π_1, π_2*

$$\langle \pi_1, \pi_2 \rangle_h = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(t_1, t_2) \, \mathrm{d}\pi_1(t_1) \, \mathrm{d}\pi_2(t_2). \tag{21}$$

We can consequently define

$$\|\pi_1\|_h^2 = \langle \pi_1, \pi_1 \rangle_h. \tag{22}$$

We have the relation

$$\|\pi_1 + \pi_2\|_h^2 = \|\pi_1\|_h^2 + 2\langle \pi_1, \pi_2 \rangle_h + \|\pi_2\|_h^2. \tag{23}$$

Measuring distances with the help of $\|\cdot\|_h$ can be viewed as measuring distances at a given resolution set by h. Typically we use Gaussian kernels where the sharper the kernel is, the more accurate it is.

The next definition is taken from [18].

Definition 2.4 ((ϵ-)Dipole, separation). *An ϵ-dipole (noted dipole for simplicity) is a measure $\pi = a_1\delta_{t_1} - a_2\delta_{t_2}$ where $\|t_1 - t_2\|_2 \leq \epsilon$. Two dipoles $\pi_1 = a_1\delta_{t_1} - a_2\delta_{t_2}$ and $\pi_2 = a_3\delta_{t_3} - a_4\delta_{t_4}$ are ϵ-separated if their support are strictly ϵ-separated (with respect to the ℓ^2-norm on $\mathbb{R}^d$), i.e. if $\|t_1 - t_3\|_2 > \epsilon$, $\|t_2 - t_3\|_2 > \epsilon$ and $\|t_1 - t_4\|_2 > \epsilon$ and $\|t_2 - t_4\|_2 > \epsilon$.*

Compared to [18], we need to introduce a new definition.

Definition 2.5 (Generalized dipole). *A generalized dipole ν is either a dipole or a distribution of order 1 of the form $a_1\delta_t + a_2\delta'_{t,v}$. Two generalized dipoles are ϵ-separated if their support are strictly ϵ-separated (with respect to the ℓ^2-norm on $\mathbb{R}^d$).*

In this article we use regular, symmetrical, translation invariant kernels. Most recent developments to non translation invariant kernels [22] could be considered to generalize this work, but they are out of the scope of this article for the sake of simplicity.

Assumption 2.1. *A kernel h follows this assumption if*

- *$h \in \mathcal{C}^2(\mathbb{R}^d)$.*

- *h is symmetrical with respect to* 0*, translation invariant, i.e. we can write* $h(t_1, t_2) = \rho(\|t_1 - t_2\|_2)$ *where* $\rho \in \mathcal{C}^2(\mathbb{R})$*.*
- *there is a constant* μ_h *such that, for all two* ϵ*-separated dipoles,* $\langle \nu_1, \nu_2 \rangle_h \leq \mu_h \|\nu_1\|_h \|\nu_2\|_h$ *(mutual coherence).*
- $h(0) = 1 = \max_{t \in \mathbb{R}^d} h(t)$*,* $h'_v(0) = \rho'(0) = 0$*, and* $h''_v(0) = \rho''(0) \leq 0$*.*

Note that the assumption that $h \in \mathcal{C}^2$ guarantees the existence of integrals with respect to finite signed measures and distribution of order 1 with bounded supports.

Example The now almost canonical well behaved kernel is the Gaussian kernel. From [18], for $\epsilon = 1$, using $h_0(t) = e^{-t^2/(2\sigma_k^2)}$ with $\sigma_k^2 = \frac{1}{2.4log(2k-1)+24}$, we have that h_0 follows Assumption 2.1 with $\mu_{h_0} = \frac{3}{4(k-1)}$.

We have the following properties.

Lemma 2.1. *Let h be a kernel meeting Assumption 2.1. We have the following properties for any $t \in \mathbb{R}$:*

$$\|\delta_t\|_h^2 = \rho(0) = 1 \tag{24}$$

$$\langle \delta_t, \delta'_{t,v} \rangle_h = -\rho'(0) = 0 \tag{25}$$

$$\|\delta'_{t,v}\|_h^2 = |\rho''(0)| \tag{26}$$

Proof. See Appendix A.2. □

From [18, Lemma 6.5], we have the following Lemma:

Lemma 2.2. *Suppose for all two ϵ-separated dipoles, $\langle \pi_1, \pi_2 \rangle_h \leq \mu \|\pi_1\|_h \|\pi_2\|_h$ (mutual coherence). Then for k, ϵ-separated dipoles $\pi_1, ...\pi_k$ such that $\max_i \|\pi_i\| > 0$, we have*

$$1 - (k-1)\mu \leq \frac{\|\sum_{i=1,k} \pi_i\|_h^2}{\sum_{i=1,k} \|\pi_i\|_h^2} \leq 1 + (k-1)\mu \tag{27}$$

We can generalize the previous result to generalized dipoles.

Lemma 2.3. *Let two ϵ-separated* ***generalized*** *dipoles ν_1, ν_2. Suppose for all two ϵ-separated dipoles π_1, π_2, $\langle \pi_1, \pi_2 \rangle_h \leq \mu \|\pi_1\|_h \|\pi_2\|_h$ (mutual coherence). Then we have:*

$$\langle \nu_1, \nu_2 \rangle_h \leq \mu \|\nu_1\|_h \|\nu_2\|_h \tag{28}$$

Proof. See Appendix A.2. □

A consequence of the previous result is the following Lemma:

Lemma 2.4. *Suppose for all two ϵ-separated generalized dipoles, $\langle \nu_1, \nu_2 \rangle_h \leq \mu \|\nu_1\|_h \|\nu_2\|_h$ (mutual coherence). Then for k ϵ-separated generalized dipoles $\nu_1, ...\nu_k$ such that $\max_i \|\nu_i\| > 0$, we have*

$$1 - (k-1)\mu \leq \frac{\|\sum_{i=1,k} \nu_i\|_h^2}{\sum_{i=1,k} \|\nu_i\|_h^2} \leq 1 + (k-1)\mu \tag{29}$$

Proof. See Appendix A.2. ◻

We are now able to define the Restricted Isometry Property (RIP). The secant set of the model set Σ is $\Sigma - \Sigma := \{x - y : x \in \Sigma, y \in \Sigma\}$.

Definition 2.6 (RIP)**.** *A has the RIP on $\Sigma - \Sigma$ with respect to $\|\cdot\|$ with constant γ if for all $x \in \Sigma - \Sigma$:*

$$(1-\gamma)\|x\|^2 \leq \|Ax\|_2^2 \leq (1+\gamma)\|Ax\|^2. \tag{30}$$

In the following we will suppose that A has RIP γ on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$ with respect to $\|\cdot\|_h$. i.e. for $\sum_{r=1,k} a_r\delta_{t_r} - \sum_{r=1,k} b_r\delta_{s_r} \in \Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$, we have

$$\begin{aligned}(1-\gamma)\left\|\sum_{r=1,k}(a_r\delta_{t_r} - b_r\delta_{s_r})\right\|_h^2 &\leq \left\|A\sum_{r=1,k}(a_r\delta_{t_r} - b_r\delta_{s_r})\right\|_2^2 \\ &\leq (1+\gamma)\left\|\sum_{r=1,k} a_r\delta_{t_r} - b_r\delta_{s_r}\right\|_h^2.\end{aligned} \tag{31}$$

From [18], with a Gaussian kernel h it is possible to build a random A with RIP constant γ. With this choice of A, the ideal minimization (2) yields a stable and robust estimation of x_0 with respect to the $\|\cdot\|_h$.

In [7], stable recovery for ϵ-separated Diracs on is guaranteed on the Torus with the metric $\|K_{hi} * \cdot\|_{L^1}$ where $K_{hi}*$ is the convolution with a Fejér kernel. From [4, IV.A], this guarantees a lower RIP with respect to this metric. Thanks to the inclusion of L^p-spaces on the Torus ($L^p \subset L^q$ if $1 \leq q \leq p \leq \infty$), there exists a kernel metric $\|\cdot\|_{h_K}$ (apply Lemma A.1 from the Annex on the Fejér kernel) that lower bounds $\|K_{hi} * \cdot\|_{L^1}$ for sums of Diracs. This guarantees the existence of a lower RIP with respect to a kernel metric for the conventional deterministic spike super-resolution setting.

The RIP on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$ implies a RIP on ϵ-separated generalized dipoles.

Lemma 2.5 (RIP on generalized dipoles)**.** *Suppose A has the RIP on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$ with constant γ. Let $(\nu_r)_{r=1,k}$, k ϵ-separated dipoles supported in $rint\mathcal{B}_2(R)$, we have*

$$(1-\gamma)\left\|\sum_{r=1,k}\nu_r\right\|_h^2 \leq \left\|A(\sum_{r=1,k}\nu_r)\right\|_2^2 \leq (1+\gamma)\left\|\sum_{r=1,k}\nu_r\right\|_h^2 \tag{32}$$

Proof. See Appendix A.2. ◻

Finally, we will need a last estimate. To state it, we need first to introduce the following definition:

Definition 2.7. *Let A such that the α_l are in $\mathcal{C}^2(\mathcal{B}_2(R))$. We define*

$$D_{A,R} := \sup_{1\leq l\leq m, v\in\mathcal{B}_2(1), w\in\mathcal{B}_2(1), t\in\mathcal{B}_2(R)} |\alpha''_{l,v,w}(t)|_2. \tag{33}$$

The constant $D_{A,R}$ is finite, and it is thus a bound of the directional second derivatives of the α_l over $\mathcal{B}_2(R)$.

Lemma 2.6. *Let A such that the α_l are in $\mathcal{C}^2(\mathcal{B}_2(R))$. Then, for any $t \in \mathcal{B}_2(R)$, with directions v_1, v_2, we have*

$$\|A\delta''_{t_0,v_1,v_2}\|_2 \leq \sqrt{m} D_{A,R}. \tag{34}$$

where $D_{A,R}$ is defined in Equation (33).

Proof. See Appendix A.2. □

2.4 Control of the conditioning of the Hessian with the restricted isometry property

We can now give a lower (resp. upper) bound for the highest (resp. lowest) eigenvalues of the Hessian matrix H of g (computed in Proposition 2.2).

Theorem 2.1 (Control of the Hessian)**.** *Let $\theta = (a_1, .., a_k, t_1, ..t_k) \in \Theta_{k,\epsilon}$ with $t \in rint\mathcal{B}_2(R)$ and $\theta^* \in \Theta_{k,\epsilon}$ a minimizer of (4). Suppose h follows Assumption 2.1. Let H the Hessian of g at θ. Suppose A has RIP γ on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$. We have*

$$\sup_{\|u\|_2=1} u^T H u \leq 2(1+\gamma)(1+(k-1)\mu)\max(1, (a_r^2|\rho''(0)|)_{r=1,l}) + \xi; \tag{35}$$

$$\inf_{\|u\|_2=1} u^T H u \geq 2(1-\gamma)(1-(k-1)\mu)\min(1, (a_r^2|\rho''(0)|)_{r=1,l}) - \xi \tag{36}$$

where $\xi = 2(d+1)\max(\max_r |a_r|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})(\|A\phi(\theta) - A\phi(\theta^*)\|_2 + \|e\|_2)$, the constant $D_{A,R}$ is defined in (33) and e is the finite energy measurement noise.

Proof. See Appendix A.3. □

Remark 2.1. *Notice that, in the noiseless case, (36) ensures in particular that g has a positive Hessian matrix in θ^*. Moreover, if $\min_r |a_r| > 0$, there exists a neighbourhood of θ^*, in which g remains convex. We will give an explicit size for this neighbourhood in the next section. Notice also that (35) gives an upper bound on the Lipschitz constant of the gradient of g. This implies the existence of a basin of attraction (see Definition 1.3) with a uniform bound for the step size.*

Remark 2.2. *With the method to choose A from [18, Lemma 6.5], for any γ and $m \gtrsim k^2 d polylog(k,d)/\gamma^2$, we can find A that has RIP with high probability with kernel h_0 having the right properties.*

We can control the conditioning of the Hessian matrix $\kappa(H)$ at a global minimum as the term $\|A\phi(\theta) - A\phi(\theta^*)\|_2$ vanishes in the control from Theorem 2.1. Particularly, in the noiseless case we have the following Corollary. The lower bound is useful to confirm the dependency on the ratio of amplitudes when it converges to $+\infty$. For this next result, we make the additional assumption that $\min_r |a_r| > 0$. In practice, this amounts to assuming that when estimating the Diracs, we do not over-estimate their number (which will often be the case, in particular in the presence of noise). When the number of Diracs is overestimated, the minimizers of (4) are points that are not isolated, the notion of basin of attraction would have to be generalized to a basin of attraction of a set of minimizers (when $a_r = 0$, $g(\theta)$ does not depend on t_r), which is out of the scope of this article for clarity purpose.

Corollary 2.1. *Let $x_0 = \sum_{r=1,k} a_r \delta_{t_r} \in \Sigma_{k,\epsilon} = \phi(\theta_0)$ and $e = 0$. Suppose h follows Assumption 2.1. Let H the Hessian of g at θ_0. Suppose A has RIP γ on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$, and that $\min_r |a_r| > 0$. We have*

$$\begin{aligned}\frac{(1-\gamma)\max(1,(a_r^2|\rho''(0)|)_{r=1,l})}{(1+\gamma)\min(1,(a_r^2|\rho''(0)|)_{r=1,l})} &\leq \kappa(H) \\ &\leq \frac{(1+\gamma)(1+(k-1)\mu)\max(1,(a_r^2|\rho''(0)|)_{r=1,l})}{(1-\gamma)(1-(k-1)\mu)\min(1,(a_r^2|\rho''(0)|)_{r=1,l})}\end{aligned} \tag{37}$$

Proof. See Appendix A.3. □

It is easy to see that for a noise e with small enough energy (i.e. such that ξ is strictly lower than $2(1-\gamma)(1-(k-1)\mu)\min(1,(a_r^2|\rho''(0)|)_{r=1,l})$, if $\min_r |a_r| > 0$, then the Hessian at a global minimum is strictly positive. Of course, this may require a very small noise since the ratio of amplitudes at the global minimum can be large.

Remark 2.3. *We remark that for a same maximal ratio of amplitudes in θ^*, a better conditioning bound is achieved when $\max_{r=1,l} a_r^2|\rho''(0)| \geq 1 \geq \min_{r=1,l} a_r^2|\rho''(0)|$. We attribute this to the fact that we estimate amplitudes and locations at the same time. The amplitudes must be appropriately scaled to match the variations of g with respect to locations. Intuitively, alternate descent with respect to amplitudes and locations might be better than the classical gradient descent for easily setting the descent step.*

Remark 2.4. *As g is $\mathcal{C}^2$, ensuring the strict positivity of the Hessian at the global minimum guarantees the* existence *of a basin of attraction as emphasized in Section 1.3. In the next Section, we give an explicit formulation of a basin of attraction.*

3 Explicit basin of attraction of the global minimum

Let $\theta_1 \in \mathbb{R}^d$. Can we guarantee, for some notion of distance d, that $d(\theta_1, \theta_0) \leq C$ and $\theta_1 \neq \theta_0$, with C an explicit constant, implies $\nabla g(\theta_1) \neq 0$? The following theorems show that it is in fact the case. With a strong RIP assumption, we can give an explicit basin of attraction of the global minimum for minimization (4) without separation constraints.

3.1 Uniform control of the Hessian

In the noiseless case, a global minimum θ^* of the constrained minimization of g over $\Theta_{k,\epsilon}$ is also a global minimum of the unconstrained minimization because $g(\theta^*) = 0$. In the presence of noise, we can no longer guarantee that the minimizer of the constrained problem θ^* is a global minimum of the unconstrained problem. However, the shape of the constraint guarantees that it is a local minimum (see next Lemma).

Lemma 3.1. *Suppose $\theta^* = (a_1, .., a_k, t_1, .., t_k)$ is a result of constrained minimization (4) with $t_i \in rint\mathcal{B}_2(R)$. Then θ^* is a local minimum of g.*

Proof. let $\theta^* = (a_1, .., a_k, t_1, .., t_k)$. As for all $i \neq j$, $|t_i - t_j| > \epsilon$, there exists $\eta > 0$ such that for all $\theta = (b_1, .., b_k, s_1, .., s_k)$ such that $|s_i - t_i| < \eta$, we have $\theta \in \Theta_{k,\epsilon}$. Hence, $\theta^* + B_\infty(\eta) \subset \Theta_{k,\epsilon}$, and $\theta^* \in \arg\min_{\theta \in \theta^* + B_\infty(\eta)} g(\theta)$. □

Hence we can still calculate a basin of attraction of θ^* (for the unconstrained minimization). The expression of the basin in the next Section is a direct consequence of the following Theorem that uniformly control the Hessian of g in an explicit neighbourhood of θ^*.

Theorem 3.1. *Suppose A has RIP γ on $\Sigma_{k,\frac{\epsilon}{2}} - \Sigma_{k,\frac{\epsilon}{2}}$ and that h follows Assumption 2.1 and has mutual coherence constant μ on $\frac{\epsilon}{2}$-separated dipoles. Let $\theta^* = (a_1, .., a_k, t_1, .., t_k) \in \Theta_{k,\epsilon}$ be a result of constrained minimization (4) such that $t_i \in rint\mathcal{B}_2(R)$. Suppose $0 < |a_1| \leq |a_2| ... \leq |a_k|$. Let $0 \leq \beta \leq 1$ and*

$$\begin{aligned}\Lambda_{\theta^*,\beta,\frac{\epsilon}{4}} = \{\theta = (b_1, .., b_k, s_1, .., s_k) :&\\ sign(b_i) = sign(a_i), \frac{\|b_i\delta_{s_i} - a_i\delta_{t_i}\|_h}{|a_i|} < \beta, \sup |s_j - t_j| < \frac{\epsilon}{4}\}&\end{aligned} \tag{38}$$

If $\theta \in \Lambda_{\theta^*,\beta,\frac{\epsilon}{4}}$, then H the Hessian of g at θ has the following bounds :

$$\sup_{\|u\|_2=1} u^T H u \leq 2(1+\gamma)(1+(k-1)\mu)\max(1, (|a_k|(1+\beta))^2|\rho''(0)|)) + \xi; \tag{39}$$

$$\inf_{\|u\|_2=1} u^T H u \geq 2(1-\gamma)(1-(k-1)\mu)\min(1, (|a_1|(1-\beta))^2|\rho''(0)|) - \xi \tag{40}$$

where $\xi = 2(d+1)\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})(\sup_{\theta \in \Lambda_{\theta^*,\beta,\frac{\epsilon}{4}}} \|A\phi(\theta) - A\phi(\theta^*)\|_2 + \|e\|_2)$, the constant $D_{A,R}$ is given in (33) and e is the finite energy measurement noise.

Proof. See Appendix A.4. □

Remark 3.1. *We observe that we require a stronger RIP than the usual one on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$ to guarantee that unconstrained minimization converges in the basin of attraction $\Lambda_{\theta^*,\beta,\frac{\epsilon}{4}}$.*

When the separation constraint is added for the basin of attraction (we look for potential critical points in $\Sigma_{k,\epsilon}$), we can provide better bounds. We will discuss what we could expect from constrained descent algorithms in Section 4.

Theorem 3.2. *Suppose A has RIP γ on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$ and that h follows Assumption 2.1 and has mutual coherence constant μ on ϵ-separated dipoles. Suppose $0 < |a_1| \leq |a_2| ... \leq |a_k|$.*

Let $\theta^* = (a_1, ..., a_k, t_1, ..t_k) \in \Theta_{k,\epsilon}$ be a result of constrained minimization (4) such that $t_i \in rint\mathcal{B}_2(R)$. Let $0 \leq \beta \leq 1$ and

$$\begin{aligned}\Lambda_{\theta^*,\beta,\infty} = \{\theta = (b_1, ..b_k, s_1, ..s_k) :&\\ sign(b_i) = sign(a_i), \frac{\|b_i\delta_{s_i} - a_i\delta_{t_i}\|_h}{|a_i|} < \beta\}&\end{aligned} \tag{41}$$

Then for $\theta \in \Theta_{k,\epsilon} \cap \Lambda_{\theta^*,\beta,\infty}$, then H the Hessian of g at θ has the following bounds:

$$\sup_{\|u\|_2=1} u^T H u \leq 2(1+\gamma)(1+(k-1)\mu)\max(1,(|a_k|(1+\beta))^2|\rho''(0)|)) + \xi; \tag{42}$$

$$\inf_{\|u\|_2=1} u^T H u \geq 2(1-\gamma)(1-(k-1)\mu)\min(1,(|a_1|(1-\beta))^2|\rho''(0)|) - \xi \tag{43}$$

where $\xi = 2(d+1)\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})(\sup_{\theta\in\Lambda_{\theta^*,\beta,\infty}} \|A\phi(\theta)-A\phi(\theta^*)\|_2 + \|e\|_2)$, the constant $D_{A,R}$ is given in (33) and e is the finite energy measurement noise.

Proof. See Appendix A.4. □

3.2 Explicit basin of attraction in the noiseless and noisy case

With the help of this uniform control of the Hessian we give an explicit (yet suboptimal) basin of attraction.

Corollary 3.1 (of Theorem 3.1, noiseless case)**.** *Under the hypotheses of Theorem 3.1, let $\theta^* \in \Theta_{k,\epsilon}$ be a result of constrained minimization (4).*

Take $\beta \leq \beta_{max}$ where $\beta_{max} := \min\left(\frac{1}{2}, \frac{(1-\gamma)(1-(k-1)\mu)\min(1,|a_1|^2|\rho''(0)|/4)}{k(d+1)|a_k|\sqrt{1+\gamma}\max(|a_k|\sqrt{m}D_{A,R},\sqrt{1+\gamma}\sqrt{|\rho''(0)|})}\right)$. Then the set $\Lambda_{\theta^*,\beta,\epsilon/4}$ is a basin of attraction of θ^*.

Proof. See Appendix A.4. □

The parameter β controls the distance between a parameter and the optimal parameter as $\|b\delta_s - a\delta_t\|_h^2 = (b-a)^2 + 2ab(1-\rho(\|s-t\|_2)) \geq \min((b-a)^2, 2|ab|(1-\rho(\|s-t\|_2))$ when $sign(a) = sign(b)$. When the RIP constant γ decreases (and generally as the number of measurement increases), the size of the basin of attraction increases. When the mutual coherence constant μ decreases, the basin of attraction also increases. Finally, we note that the smaller β is, the smaller is the upper bound on the operator norm of the Hessian.

When the noise contaminating the measurements is small enough, we have similar results with a smaller basin of attraction.

Corollary 3.2 (of Theorem 3.1, noisy case)**.** *Under the hypotheses of Theorem 3.1, let $\theta^* \in \Theta_{k,\epsilon}$ be a result of constrained minimization (4). Suppose $\|e\| \leq k|a_k|\beta_{max}$ where $\beta_{max} := \frac{1}{2}\min\left(\frac{1}{2}, \frac{(1-\gamma)(1-(k-1)\mu)\min(1,|a_1|^2|\rho''(0)|/4)}{k(d+1)|a_k|(1+\sqrt{1+\gamma})\max(|a_k|\sqrt{m}D_{A,R},\sqrt{1+\gamma}\sqrt{|\rho''(0)|})}\right)$.*

Take $\beta \leq \beta_{max}$. Then the set $\Lambda_{\theta^*,\beta,\epsilon/4}$ is a basin of attraction of θ^*.

Proof. See Appendix A.4. □

4 Towards new descent algorithms for SR estimation?

We have shown that, given an appropriate measurement operator for separated Diracs, a good initialization is sufficient to guarantee the success of a simple gradient descent. Moreover the gradient descent can be used as a refinement step in the greedy heuristic based on orthogonal matching pursuit [19]. If we could guarantee that greedily estimating Diracs, we fall within the basin of attraction, we would have a full non-convex optimization technique with guarantees of convergence to a global minimum.

In other works [14, 15], it has been shown that discretization (on grids) of convex methods have a tendency to produce spurious spikes at Dirac locations. Our results seem to indicate that merging spikes that are close to each other when performing a gradient descent might break the barrier between continuous and discrete methods.

Theorem 3.2 brings another question as the Hessian of g is more easily controlled in $\Theta_{k,\epsilon}$. More generally, can we build a simple descent algorithm that stays in $\Theta_{k,\epsilon}$ to get larger basins of attraction? Consider the problem for $d = 1$ in the noiseless case for the sake of clarity. We want to use the following descent algorithm:

$$\theta_{i+1} = P_{\Theta_{k,\epsilon}}(\theta_i - \tau \nabla g(\theta_i)) \tag{44}$$

Where $P_{\Theta_{k,\epsilon}}$ is a projection onto the separation constraint. Notice that since $\Theta_{k,\epsilon}$ is not a convex set, we cannot easily define the orthogonal projection onto it (it may not even exists).

If we suppose that the gradient descent step decreases g (i.e. $g(\theta_i - \tau\nabla g(\theta_i)) < g(\theta_i)$), is it possible to guarantee that applying projection step keeps decreasing g? Consider:

$$P_{\Theta_{k,\epsilon}}(\theta) \in \arg\min_{\tilde{\theta}\in\Theta_{k,\epsilon}} \left| \|A\phi(\tilde{\theta}) - y\|_2 - \|A\phi(\theta) - y\|_2 \right| \tag{45}$$

First consider the following Lemma:

Lemma 4.1. *Let $d = 1$. Let $\theta_0, \theta_1 \in \Theta_{k,\epsilon}$. Let $g(\theta) = \|A\phi(\theta) - A\phi(\theta_0)\|$. Then for all α such that $0 = g(\theta_0) \leq \alpha \leq g(\theta_1)$, there exists $\theta^* \in \Theta_{k,\epsilon}$ such that $g(\theta^*) = \alpha$.*

Proof. See Appendix A.5. □

Lemma 4.1 essentially guarantees that is is possible to continuously map the interval $[0, g(\theta_1)]$ by g with elements of $\Theta_{k,\epsilon}$. Hence, at a step $i+1$, we have

$$|g(\theta_{i+1}) - g(\theta_i)| = |g(\theta_i - \tau\nabla g(\theta_i)) - g(\theta_i)|. \tag{46}$$

The projection $P_{\Theta_{k,\epsilon}}$ defined by (45) is not easy to calculate (in fact, it is a similar optimization problem as the main problem). Other more "natural" projections on $\Theta_{k,\epsilon}$ could be defined as :

$$P_{\Theta_{k,\epsilon}}(\theta) \in \phi^{-1}(\arg\inf_{x\in\Sigma_{k,\epsilon}} \|Ax - A\phi(\theta)\|_2) \tag{47}$$

or

$$P_{\Theta_{k,\epsilon}}(\theta) \in \phi^{-1}(\arg\inf_{x\in\Sigma_{k,\epsilon}} \|x - \phi(\theta)\|_h). \tag{48}$$

However they suffer from the same calculability drawback. This suggests to build a new family of heuristic algorithms of spike estimation where we propose heuristics to approach the projection of $\hat{\theta}_{i+1}$ on $\Theta_{k,\epsilon}$. Recovery guarantees would be obtained by guaranteeing that the projection heuristic does not increase the value of g by too much compared to the gradient descent step.

A Annex

A.1 Proofs for Section 2.2

Proof of Proposition 2.1.

$$
\begin{aligned}
\frac{\partial g(\theta)}{\partial a_r} &= \frac{\partial}{\partial a_r} \sum_{l=1}^{m} \left| \sum_{i=1}^{k} a_i \alpha_l(t_i) - y_l \right|^2 \\
&= \sum_{l=1}^{m} 2\mathcal{R}e \left(\alpha_l(t_{r,j}) \left(\overline{\sum_{i=1}^{k} a_i \alpha_l(t_i) - y_l} \right) \right) \\
&= 2\mathcal{R}e \langle A\delta_{t_r}, A\phi(\theta) - y \rangle.
\end{aligned}
\tag{49}
$$

Similarly,

$$
\begin{aligned}
\frac{\partial g(\theta)}{\partial t_{r,j}} &= \frac{\partial}{\partial t_r} \sum_{l=1}^{m} \left| \sum_{i=1}^{k} a_i \alpha_l(t_i) - y_l \right|^2 \\
&= \sum_{l=1}^{m} 2\mathcal{R}e \left(a_r \partial_j \alpha_l(t_r) \left(\overline{\sum_{i=1}^{k} a_i \alpha_l(t_i) - y_l} \right) \right). \\
&= -2a_r \mathcal{R}e \langle A\delta'_{t_0,j}, A\phi(\theta) - y \rangle
\end{aligned}
\tag{50}
$$

□

Proof of Proposition 2.2. For $H_{1,r,s}$,

$$
\begin{aligned}
\frac{\partial^2 g(\theta)}{\partial a_r \partial a_s} &= \frac{\partial}{\partial a_s} \sum_{l=1}^{m} 2\mathcal{R}e \left(\alpha_l(t_r) \left(\overline{\sum_{i=1}^{k} a_i \alpha_l(t_i) - y_l} \right) \right) \\
&= \sum_{l=1}^{m} 2\mathcal{R}e \left(\alpha_l(t_r) \left(\overline{\alpha_l(t_s)} \right) \right).
\end{aligned}
\tag{51}
$$

For H_{2,r,j_1,s,j_2},

$$\begin{aligned}\frac{\partial^2 g(\theta)}{\partial t_{r,j_1}\partial t_{s,j_2}} &= \frac{\partial}{\partial t_{s,j_1}}\sum_{l=1}^{m} 2\mathcal{R}e\left(a_r\partial_{j_1}\alpha_l(t_r)\left(\overline{\sum_{i=1}^{k} a_i\alpha_l(t_i)-y_l}\right)\right)\\ &= \sum_{l=1}^{m} 2\mathcal{R}e\left(a_r\partial_{j_1}\alpha_l(t_r)\left(\overline{a_s\partial_{j_2}\alpha_l(t_s)}\right)\right) \\ &+ \mathbf{1}(r=s)\sum_{l=1}^{m} 2\mathcal{R}e\left(a_r\partial_{j_2}\partial_{j_1}\alpha_l(t_r)\left(\overline{\sum_{i=1}^{k} a_i\alpha_l(t_i)-y_l}\right)\right).\end{aligned} \tag{52}$$

For $H_{12,r,s,j}$

$$\begin{aligned}\frac{\partial^2 g(\theta)}{\partial a_r\partial t_{s,j}} &= \frac{\partial}{\partial t_{s,j}}\sum_{l=1}^{m} 2\mathcal{R}e(\alpha_l(t_r))\left(\overline{\sum_{i=1}^{k} a_i\alpha_l(t_i)-y_l}\right)\\ &= \sum_{l=1}^{m} 2\mathcal{R}e\left(\alpha_l(t_r)\left(\overline{a_s\partial_{j}\alpha_l(t_s)}\right)\right) \\ &+ \mathbf{1}(r=s)\sum_{l=1}^{m} 2\mathcal{R}e\left(\partial_{j}\alpha_l(t_r)\left(\overline{\sum_{i=1}^{k} a_i\alpha_l(t_i)-y_l}\right)\right).\end{aligned} \tag{53}$$

□

A.2 Proofs for Section 2.3

Proof of Lemma 2.1. Using the assumption,

$$\|\delta_t\|_h^2 = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \rho(\|t_1-t_2\|_2)\,\mathrm{d}\delta_t(t_1)\,\mathrm{d}\delta_t(t_2)) = \int_{\mathbb{R}^d}\rho(\|t_1-t\|_2)\,\mathrm{d}\delta_t(t_1) = \rho(0). \tag{54}$$

With the fact that $\int_{\mathbb{R}} f(t)\,\mathrm{d}\delta'_{t_0,v}(t)\,\mathrm{d}t = -f'_v(t_0)$, we have

$$\begin{aligned}\langle \delta_t, \delta'_t\rangle_h = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \rho(\|t_1-t_2\|_2)\,\mathrm{d}\delta_t(t_1)\,\mathrm{d}\delta'_{t,v}(t_2) &= \int_{\mathbb{R}^d}\rho(\|t-t_2\|_2)\,\mathrm{d}\delta'_{t,v}(t_2)\\ &= \int_{\mathbb{R}^d}\rho(\|t_2-t\|_2)\,\mathrm{d}\delta'_{t,v}(t_2).\end{aligned} \tag{55}$$

Hence $\langle \delta_t, \delta'_t\rangle_h = -\lim_{\eta\to 0^+}\frac{\rho(\eta\|v\|)-\rho(0)}{\eta} = -\rho'(0) = 0$.

Let $f : t \to f(t) = \rho(\|t\|_2)$.

$$\|\delta'_t\|_h^2 = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \rho(\|t_1-t_2\|_2)\,\mathrm{d}\delta'_{t,v}(t_1)\,\mathrm{d}\delta'_{t,v}(t_2) = -\int_{\mathbb{R}^d} f'_v(t-t_2)\,\mathrm{d}\delta'_{t,v}(t_2). \tag{56}$$

Hence $\|\delta'_{t,v}\|_h^2 = -f''_v(0) = |\rho''(0)|$. □

Proof of Lemma 2.3. Let two ϵ-separated generalized dipole ν_1, ν_2. The ν_i are the limit (in the distributional sense) of a family of ϵ-separated dipole $\nu_i^{\eta_i}$ for $\eta_i \to 0^+$. With the hypothesis, we have

$$\langle \nu_1^{\eta_1}, \nu_2^{\eta_2} \rangle_h \leq \mu \|\nu_1^{\eta_1}\|_h \|\nu_2^{\eta_2}\|_h \tag{57}$$

Furthermore,

$$\langle \nu_1^{\eta_1}, \nu_2^{\eta_2} \rangle_h = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \rho(\|t_1 - t_2\|_2) \, \mathrm{d}\nu_1^{\eta_1}(t_1) \, \mathrm{d}\nu_2^{\eta_2}(t_2) \tag{58}$$

Remark that by construction $f_{\eta_1}(t_2) := \int_{\mathbb{R}^d} \rho(\|t_1 - t_2\|_2) \, \mathrm{d}\nu_1^{\eta_1}(t_1) \to_{\eta_1 \to 0^+} f(t_2) := \int_{\mathbb{R}^d} \rho(\|t_1 - t_2\|_2) \, \mathrm{d}\nu_1(t_1) < +\infty$ where f_{η_1} is in$\mathcal{C}^2$ and f is $\mathcal{C}^1$ thanks to the assumption on h and ρ. Hence by boundedness of the integrals and the dominated convergence theorem, for any η_2

$$\langle \nu_1^{\eta_1}, \nu_2^{\eta_2} \rangle_h \to_{\eta_1 \to 0^+} \langle \nu_1, \nu_2^{\eta_2} \rangle_h. \tag{59}$$

Moreover, by construction of $\nu_2^{\eta_2}$,

$$\begin{aligned} \langle \nu_1, \nu_2^{\eta_2} \rangle_h = \int_{\mathbb{R}^d} f(t_2) \, \mathrm{d}\nu_2^{\eta_2}(t_2) \to_{\eta_2 \to 0^+} & \int_{\mathbb{R}^d} f(t_2) \, \mathrm{d}\nu_2(t_2) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \rho(\|t_1 - t_2\|_2) \, \mathrm{d}\nu_1(t_1) \, \mathrm{d}\nu_2(t_2), \end{aligned} \tag{60}$$

and

$$\langle \nu_1^{\eta_1}, \nu_2^{\eta_2} \rangle_h \to_{\eta_1 \to 0^+, \eta_2 \to 0^+} \langle \nu_1, \nu_2 \rangle_h. \tag{61}$$

Let $\nu = a\delta_t + b\delta'_{t,v}$ with $\|v\|_2 = 1$ and $\nu^\eta = a\delta_t - b\frac{\delta_{t+\eta v} - \delta_t}{\eta} = \left(a + \frac{b}{\eta}\right)\delta_t - b\frac{\delta_{t+\eta v}}{\eta}$. We have $\|\nu\|_h^2 = a^2 + b^2 |\rho''(0)|$ (with Lemma 2.1) and

$$\begin{aligned} \|\nu^\eta\|_h^2 &= \left(a + \frac{b}{\eta}\right)^2 + \left(\frac{b}{\eta}\right)^2 - 2\left(a + \frac{b}{\eta}\right)\frac{b}{\eta}\rho(\eta) \\ &= a^2 + 2\left(\frac{b}{\eta}\right)^2 + 2\frac{ab}{\eta} - 2\frac{ab}{\eta}\rho(\eta) - 2\left(\frac{b}{\eta}\right)^2 \rho(\eta) \\ &= a^2 + 2\frac{ab}{\eta}(1 - \rho(\eta)) + 2\frac{b^2}{\eta^2}(1 - \rho(\eta)). \end{aligned} \tag{62}$$

But $\frac{1-\rho(\eta)}{\eta} = \frac{\rho(0)-\rho(\eta)}{\eta} \to -\rho'(0)$ when $\eta \to 0^+$, and $\rho'(0) = 0$.

Moreover, $\rho(\eta) = h(0) + \eta\rho'(0) + \frac{\eta^2}{2}\rho''(0) + o(\eta^2) = 1 - \frac{\eta^2}{2}|\rho''(0)| + o(\eta^2)$. Hence $\frac{1-\rho(\eta)}{\eta^2} \to_{\eta \to 0^+} \frac{1}{2}|\rho''(0)|$.

We thus deduce that $\|\nu^\eta\|_h^2 \to a^2 + b^2|\rho''(0)| = \|\nu\|_h$ when $\eta \to 0^+$.

Hence we can take the limit $\eta_1, \eta_2 \to 0$ in Equation (57) to get the result. □

Proof of Lemma 2.4. Using Lemma 2.3, and the same proof as in Lemma 2.2, we get the result. □

Proof of Lemma 2.5. Let $\nu_r = a_r\delta_{t_r} + b_r\delta'_{t_r,v}$ the ϵ-separated generalized dipoles. Similarly to Lemma 2.3, take $\nu_r^\eta = (a+\frac{b}{\eta})\delta_{t_r} - b\frac{\delta_{t_r+\eta v}}{\eta}$. For sufficiently small η the ν_r^η are ϵ-separated dipoles, hence $\sum \nu_r^\eta \in \Sigma - \Sigma$ and

$$(1-\gamma)\left\|\sum_{r=1,k}\nu_r^\eta\right\|_h^2 \leq \left\|A(\sum_{r=1,k}\nu_r^\eta)\right\|_2^2 \leq (1+\gamma)\left\|\sum_{r=1,k}\nu_r^\eta\right\|_h^2. \tag{63}$$

Now remark that $g_1(\eta) = \|\sum_{r=1,k}\nu_r^\eta\|_h^2$ and $g_2(\eta) = \|A(\sum_{r=1,k}\nu_r^\eta)\|_2^2$ are continuous functions of η that converge to $\|\sum_{r=1,k}(a_r\delta_{t_r} + b_r\delta'_{t_r,v})\|_h^2$ and $\|A(\sum_{r=1,k}(a_r\delta_{t_r} + b_r\delta'_{t_r,v}))\|_2^2$ when $\eta\to 0$:

- For g_1, use the same proof as in Lemma 2.3 with the linearity of the limit.
- For g_2:

$$\begin{aligned}
g_2(\eta) &= \sum_{l=1,m}\left|\sum_{r=1,k}\int \alpha_l(t)(a_r \,\mathrm{d}\delta_{t_r}(t) - \frac{b_r}{\eta}(\mathrm{d}\delta_{t_r+\eta v}(t) - \mathrm{d}\delta_{t_r}(t)))\right|^2\\
&= \sum_{l=1,m}\left|\sum_{r=1,k}\left(\alpha_l(t_r)a_r - \frac{b_r}{\eta}(\alpha_l(t_r+\eta v) - \alpha_l(t_r))\right)\right|^2\\
&\to_{\eta\to 0^+} \sum_{l=1,m}\left|\sum_{r=1,k}\left(\alpha_l(t_r)a_r - b_r(\alpha_l)'_v(t_r)\right)\right|^2\\
&= \sum_{l=1,m}\left|\sum_{r=1,k}\int \alpha_l(t)(a_r\,\mathrm{d}\delta_{t_r}(t) + b_r\,\mathrm{d}\delta'_{t_r,v}(t))\right|^2 = \left\|A(\sum_{r=1,k}a_r\delta_{t_r} + b_r\delta'_{t_r,v})\right\|_2^2.
\end{aligned} \tag{64}$$

Taking the limit of Equation (63) for $\eta \to 0$ yields the result.

□

Proof of Lemma 2.6. We have

$$\begin{aligned}
\|A\delta''_{t_0,v_1,v_2}\|_2^2 &= \sum_{l=1,m}|(A\delta''_{t_0,v_1,v_2})_l|_2^2\\
&= \sum_{l=1,m}|\alpha''_{l,t_0,v_1,v_2}(t_0)|_2^2\\
&\leq m \sup_{l=1,m;t\in\mathcal{B}_2(R)}|\alpha''_{l,t,v_1,v_2}(t)|_2^2 \leq mD_{A,R}^2
\end{aligned} \tag{65}$$

where $D_{A,R}$ is given in (33), i.e. $D_{A,R}$ is the supremum of directional second derivatives of the α_l over $\mathcal{B}_2(R)$. We have $D_{A,R} < +\infty$ because the α_l are supposed to be in $\mathcal{C}^2(\mathcal{B}_2(R))$.

□

Lemma A.1. *Let K be a symmetrical convolution kernel in $\mathcal{C}^2$ and $h_K : (t_1, t_2) \to h_K(t_1, t_2) = [K*K](t_1 - t_2)$ (the convolution of K by itself) then for any $x \in \Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$, we have*

$$\|x\|_{h_K}^2 = \|K * x\|_{L^2}^2. \tag{66}$$

Proof of Lemma A.1. Write $x = \sum a_i \delta_{t_i}$ and use the symmetry of K:

$$\begin{aligned}\|K * x\|_{L^2}^2 = \int \left|\sum a_i K(t-t_i)\right|^2 \mathrm{d}t &= \sum_{i,j} a_i a_j \int K(t-t_i)K(t-t_j)\,\mathrm{d}t \\ &= \sum_{i,j} a_i a_j \int K(t)K(t+t_i-t_j)\,\mathrm{d}t \\ &= \sum_{i,j} a_i a_j [K*K](t_i - t_j) = \|x\|_{h_K}^2.\end{aligned} \tag{67}$$

□

A.3 Proofs for Section 2.4

We will use the following Lemma on directional derivatives of Diracs.

Lemma A.2. *Let $u, t_0 \in \mathbb{R}^d$. Suppose $u \neq 0$. Then , $\sum_{i=1,d} u_i \delta'_{t_0,j} = \|u\|_2 \delta'_{t_0,\frac{u}{\|u\|_2}}$*

Proof. Let f a function in $\mathcal{C}^2(\mathbb{R}^d)$, we have $\int_{t\in\mathbb{R}^d} f(t) \sum_{i=1,d} u_i \,\mathrm{d}\delta'_{t_0,i}(t) = -\sum_{i=1,d} u_i \partial_i f(t_0) = -\langle u_i, \nabla f(t_0)\rangle = -\|u\|_2 f'_{\frac{u}{\|u\|_2}}(t_0)$. Hence, $\sum_{i=1,d} u_i \delta'_{t_0,i} = \|u\|_2 \delta'_{t_r,\frac{u}{\|u\|_2}}$ □

To prove Theorem 2.1, we control first the eigenvalues of G in the decomposition $H = G + F$.

Lemma A.3. *Suppose h follows Assumption 2.1. Let $\theta = (a_1, .., a_k, t_1, ..t_k) \in \Theta_{k,\epsilon}$ with $t \in rint\mathcal{B}_2(R)$. Let H the Hessian of g at θ. Suppose A has RIP γ on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$. We have*

$$\sup_{\|u\|_2=1} u^T G u \leq 2(1+\gamma)(1+(k-1)\mu)\max(1, (a_r^2|\rho''(0)|)_{r=1,l}); \tag{68}$$

$$\inf_{\|u\|_2=1} u^T G u \geq 2(1-\gamma)(1-(k-1)\mu)\min(1, (a_r^2|\rho''(0)|)_{r=1,l}). \tag{69}$$

where G is defined in Proposition 2.2.

Proof. Let $u \in \mathbb{R}^{k(d+1)}$ such that $\|u\|_2 = 1$. We index u as follows: $u_r \in \mathbb{R}$ for $r = 1, k$. $u_r \in \mathbb{R}^d$ for $r = k+1, 2k$ (it follows the indexing of H and G we used). Remark that

$$
\begin{aligned}
u^T G u = &\sum_{r,s=1,k} u_r u_s G_{1,r,s} + \sum_{r=k+1,2k; j_1=1,d; s=k+1,2k; j_2=1,d} u_{r,j_1} u_{s,j_2} G_{2,r,j_1,s,j_2} \\
&+ \sum_{r=1,k; s=k+1,2k; j=1,d} u_r u_{s,j} G_{12,r,s,j} + \sum_{r=k+1,2k; j=1,d; s=1,k} u_{r,j} u_s G_{21,r,j,s} \\
=&2 \sum_{r,s=1,k} \mathcal{R}e\langle A u_r \delta_{t_r}, A u_s \delta_{t_s}\rangle \\
&+ 2 \sum_{r=k+1,2k; j_1=1,d; s=k+1,2k; j_2=1,d} \mathcal{R}e\langle A u_{r,j_1} a_{r-k} \delta'_{t_{r-k},j_1}, A u_{s,j_2} a_{s-k} \delta'_{t_{s-k},j_2}\rangle \\
&-2 \sum_{r=1,k; s=k+1,2k; j=1,d} \mathcal{R}e\langle A u_r \delta_{t_r}, A u_{s,j} a_{s-k} \delta'_{t_{s-k},j}\rangle \\
&-2 \sum_{r=k+1,2k; j=1,d; s=1,k} \mathcal{R}e\langle A u_{r,j} a_{r-k} \delta'_{t_{r-k},j}, A u_s \delta_{t_s}\rangle
\end{aligned}
\tag{70}
$$

Thus we have

$$
\begin{aligned}
u^T G u =&2 \left\| A \sum_{r=1,k} u_r \delta_{t_r} \right\|_2^2 + 2 \left\| A \sum_{r=k+1,2k; j=1,d} u_{r,j} a_{r-k} \delta'_{t_{r-k},j} \right\|_2^2 \\
&-2\mathcal{R}e \left\langle A \sum_{r=1,k} u_r \delta_{t_r}, A \sum_{r=k+1,2k; j=1,d} u_{r,j} a_{r-k} \delta'_{t_{r-k},j} \right\rangle \\
&-2\mathcal{R}e \left\langle A \sum_{r=k+1,2k; j=1,d} u_{r,j} a_{r-k} \delta'_{t_{r-k},j}, A \sum_{r=1,k} u_r \delta_{t_r} \right\rangle \\
=&2 \left\| A \left(\sum_{r=1,k} u_r \delta_{t_r} - \sum_{r=k+1,2k; j=1,d} u_{r,j} a_{r-k} \delta'_{t_{r-k},j} \right) \right\|_2^2 \\
=&2 \left\| A \left(\sum_{r=1,k} u_r \delta_{t_r} - a_r \sum_{j=1,d} u_{r+k,j} \delta'_{t_r,j} \right) \right\|_2^2 .
\end{aligned}
\tag{71}
$$

Using Lemma A.2, we have $\sum_{j=1,d} w_j \delta'_{t_r,j} = \|w\|_2 \delta'_{t_r, \frac{w}{\|w\|_2}}$ and

$$
u^T G u = 2 \left\| A \sum_{r=1,k} (u_r \delta_{t_r} - a_r \|u_{r+k}\|_2 \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}}) \right\|_2^2 . \tag{72}
$$

We use the lower RIP in Lemma 2.5,

$$u^T G u \geq 2(1-\gamma) \left\| \sum_{r=1,k} (u_r \delta_{t_r} - a_r \|u_{r+k}\|_2 \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}}) \right\|_h^2 . \tag{73}$$

Then the hypothesis on $\|\cdot\|_h$ and Lemma 2.4 yields

$$\begin{aligned} &\| \sum_{r=1,k} (u_r \delta_{t_r} - a_r \|u_{r+k}\|_2 \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}})\|_h^2 \\ &\qquad \geq (1-(k-1)\mu) \sum_{r=1,k} \|u_r \delta_{t_r} - a_r \|u_{r+k}\|_2 \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}} \|_h^2 \end{aligned} \tag{74}$$

and

$$\begin{aligned} u^T G u &\geq 2(1-\gamma)(1-(k-1)\mu) \sum_{r=1,k} \|u_r \delta_{t_r} - a_r \|u_{r+k}\|_2 \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}} \|_h^2 \\ &\geq 2(1-\gamma)(1-(k-1)\mu) \sum_{r=1,k} \Bigg(|u_r|^2 - 2 a_r u_r \|u_{k+r}\|_2 \langle \delta_{t_r}, \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}} \rangle h \\ &+ a_r^2 \|u_{k+r}\|_2^2 \|\delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}} \|_h^2 \Bigg) . \end{aligned} \tag{75}$$

Then using Lemma 2.1:

$$\begin{aligned} u^T G u &\geq 2(1-\gamma)(1-(k-1)\mu) \sum_{r=1,k} \left(|u_r|^2 + a_r^2 \|u_{k+r}\|_2^2 |\rho''(0)| \right) \\ &\geq 2(1-\gamma)(1-(k-1)\mu) \inf_{\|u\|_2=1} \sum_{r=1,k} \left(|u_r|^2 + \|u_{k+r}\|_2^2 a_r^2 |\rho''(0)| \right) . \\ &= 2(1-\gamma)(1-(k-1)\mu) \min(1, (a_r^2 |\rho''(0)|)_{r=1,l}). \end{aligned} \tag{76}$$

Similarly, using the upper RIP in Lemma 2.5:

$$u^T G u \leq 2(1+\gamma) \| \sum_{r=1,k} (u_r \delta_{t_r} + a_r \|u_{r+k}\|_2 \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}}) \|_2^2 . \tag{77}$$

Then the hypothesis on $\|\cdot\|_h$ yields (Lemma 2.4)

$$\begin{aligned} &\| \sum_{r=1,k} (u_r \delta_{t_r} + u_{r+k} a_r \delta'_{t_r}) \|_2^2 \\ &\qquad \leq (1+(k-1)\mu) \sum_{r=1,k} \|u_r \delta_{t_r} + a_r \|u_{r+k}\|_2 \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}} \|_2^2 \end{aligned} \tag{78}$$

and

$$u^T G u \leq 2(1+\gamma)(1+(k-1)\mu) \sum_{r=1,k} \|u_r \delta_{t_r} + a_r \|u_{r+k}\|_2 \delta'_{t_r, \frac{u_{r+k}}{\|u_{r+k}\|_2}} \|_2^2. \tag{79}$$

Then using Lemma 2.1:

$$\begin{aligned} u^T G u &\leq 2(1+\gamma)(1+(k-1)\mu) \sum_{r=1,k} \left(|u_r|^2 + a_r^2 \|u_{k+r}\|_2^2 |\rho''(0)|\right) \\ &\leq 2(1+\gamma)(1+(k-1)\mu) \sup_{\|u\|_2=1} \sum_{r=1,k} \left(|u_r|^2 + \|u_{k+r}\|_2^2 a_r^2 |\rho''(0)|\right) \\ &= 2(1+\gamma)(1+(k-1)\mu) \max(1, (a_r^2|\rho''(0)|)_{r=1,l}). \end{aligned} \tag{80}$$

□

Proof of Theorem 2.1. Let θ^* a minimizer of (4). Consider H the Hessian of g at θ. We recall that $H = G + F$ (see Proposition 2.2). Using Lemma A.3, we just need to bound the operator norm of F and then to combine it with the bounds on the eigenvalues of G to get bounds on eigenvalues of $H = G + F$.

We use Lemma 2.6, the Cauchy-Schwartz and triangle inequalities. We have $\|A\delta''_{t_r,j_1,j_2}\|_2 \leq \sqrt{m} D_{A,R}$ and

$$\begin{aligned} |F_{2,r,j_1,s,j_2}| &\leq \mathbf{1}(r=s)|2a_r| \|A\delta''_{t_r,j_1,j_2}\|_2 \|A\phi(\theta) - y\|_2. \\ &\leq \mathbf{1}(r=s) 2|a_r| \sqrt{m} D_{A,R} \|A\phi(\theta) - y\|_2. \\ &\leq \mathbf{1}(r=s) 2|a_r| \sqrt{m} D_{A,R} \|A\phi(\theta) - A\phi(\theta^*) + A\phi(\theta^*) - y\|_2. \\ &\leq \mathbf{1}(r=s) 2|a_r| \sqrt{m} D_{A,R} (\|A\phi(\theta) - A\phi(\theta^*)\|_2 + \|e\|_2). \end{aligned} \tag{81}$$

Similarly, with Lemma 2.5,

$$\begin{aligned} F_{12,r,s,j} &\leq \mathbf{1}(r=s) 2\sqrt{1+\gamma} \|\delta'_{t_0,j}\|_h \|A\phi(\theta) - y\|_2 \\ &\leq \mathbf{1}(r=s) 2\sqrt{1+\gamma}\sqrt{|\rho''(0)|}(\|A\phi(\theta) - A\phi(\theta^*)\|_2 + \|e\|_2). \end{aligned} \tag{82}$$

Let $\|\cdot\|_{op}$ be the ℓ^2 operator norm of a matrix. With Gerschgorin circle theorem [17], we have

$$\|F\|_{op} \leq \max_l \|F_{l,:}\|_1 \tag{83}$$

where $F_{l,:}$ is the l-th row of F. We get

$$\begin{aligned} \|F\|_{op} &\leq \max(d \max_{r,s,j} |F_{12,r,s,j}|, \max_{r,s,j} |F_{12,r,s,j}| + d \max_{r,j_1,s,j_2} |F_{2,r,j_1,s,j_2}|) \\ &\leq 2(d+1) \max(\max_{r,s,j} |F_{12,r,s,j}|, \max_{r,j_1,s,j_2} |F_{2,r,j_1,s,j_2}|)(\|A\phi(\theta) - A\phi(\theta^*)\|_2 + \|e\|_2) \\ &\leq 2(d+1) \max(\max_r |a_r| \sqrt{m} D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})(\|A\phi(\theta) - A\phi(\theta^*)\|_2 + \|e\|_2) \end{aligned} \tag{84}$$

Hence, using Weyl's perturbation inequalities on $H = G + F$, i.e. $\lambda_{min}(H) \geq \lambda_{min}(G) - \lambda_{max}(F)$ and $\lambda_{max}(H) \leq \lambda_{max}(G) + \lambda_{max}(F)$, we get the result. □

Proof of Corollary 2.1. First, observe that at θ_0, $F = 0$.

The upper bound is a direct consequence of Theorem A.3.

We show the result in the case $\max(1, (a_r^2|\rho''(0)|)_{r=1,l}) \neq 1$ and $\min(1, (a_r^2|\rho''(0)|)_{r=1,l}) \neq 1$ (the proof is similar in the other case). For the lower bound let $v \in \mathbb{R}^{k(d+1)}$ and $i_0 = \arg\max_{r=1,l}(a_r^2|\rho''(0)|)$, set $\|v_{i_0}\|_2 = 1$ and $v_j = 0$ for $j \neq i_0$.

$$\sup_{\|u\|_2=1} u^T H u \geq v^T H v \geq 2(1-\gamma)\max(1, (a_r^2|\rho''(0)|)_{r=1,l}). \tag{85}$$

Similarly, let $v \in \mathbb{R}^{k(d+1)}$ and $i_0 = \arg\min((a_r^2|\rho''(0)|)_{r=1,l})$, $\|v_{i_0}\| = 1$ and $v_j = 0$ for $j \neq i_0$.

$$\inf_{\|u\|_2=1} u^T H u \leq 2(1+\gamma)\min(1, (a_r^2|\rho''(0)|)_{r=1,l}). \tag{86}$$

□

A.4 Proofs for Section 3

Proof of Theorem 3.1. Let $\theta^* = (a_1, ..., a_k, t_1, ..t_k) \in \Theta_{k,\epsilon}$ the global minimum of g and $\theta = (b_1, ..., b_k, s_1, ..s_k) \in \Lambda_{\theta^*,\beta,\frac{\epsilon}{4}}$. Hence for all j, $|s_j - t_j| \leq \frac{\epsilon}{4}$. Hence for $i \neq j$ we have $|s_i - s_j| = |s_i - t_i + t_i - t_j + t_j - s_j| \geq |t_i - t_j| - |t_i - s_i| - |t_j - s_j| > \epsilon - 2\epsilon/4 = \epsilon/2$ and $\phi(\theta) \in \Sigma_{k,\frac{\epsilon}{2}}$.

We use Theorem 2.1 to get the bound on the min and max eigenvalues of the Hessian.

We then notice that:

$$\begin{aligned}
\|a\delta_t - b\delta_s\|_h^2 &= a^2\|\delta_t\|_h^2 + b^2\|\delta_t\|_h^2 - 2ab\langle\delta_t, \delta_s\rangle_h \\
&= a^2 + b^2 - 2ab\rho(\|s-t\|_2) \\
&= (b-a)^2 + 2ab(1 - \rho(\|s-t\|_2))
\end{aligned} \tag{87}$$

Hence we see that $\|a\delta_t - b\delta_s\|_h/|a| \leq \beta$ is equivalent to

$$(b-a)^2 + 2ab(1 - \rho(\|s-t\|_2)) \leq a^2\beta^2 \tag{88}$$

and it therefore implies that $||b| - |a|| = |b - a| \leq |a|\beta$ when $sign(a) = sign(b)$. We thus deduce that $|a|(1-\beta) \leq |b| \leq |a|(1+\beta)$. Hence, for any r, we get the following inequality:

$$|a_1|(1-\beta) \leq |a_r|(1-\beta) \leq |b_r| \leq |a_r|(1+\beta) \leq |a_k|(1+\beta) \tag{89}$$

We can then plug these inequalities into the one of Theorem 2.1.

Finally we notice the fact that $\sup_{\theta\in\Lambda_{\theta^*,\beta,\epsilon/4}} \|A\phi(\theta) - A\phi(\theta^*)\|_2$ exists because $\Lambda_{\theta^*,\beta,\epsilon/4}$ is bounded.

□

Proof of Theorem 3.2. This is a direct consequence of Theorem 2.1. The proof follows the same lines as the one of Theorem 3.1. □

Proof of Corollary 3.1. The set $\Lambda = \Lambda_{\theta^*,\beta,\epsilon/4}$ is an open set where the Hessian of g at Λ is positive as long as $\xi \leq 2(1-\gamma)(1-(k-1)\mu)\min(1,(|a_1|(1-\beta))^2|\rho''(0)|)$ with Theorem 3.1.

In this case g is convex on Λ. Theorem 3.1 also gives a uniform bound for the operator norm of the Hessian: $\|H\|_{op} \leq 2(1+\gamma)(1+(k-1)\mu)\max(1,(|a_k|(1+\beta))^2|\rho''(0)|)+\xi$ and g has Lipschitz gradient. We thus deduce from Corollary 1.1 that Λ is a basin of attraction.

Hence we just need to show that $\xi \leq 2(1-\gamma)(1-(k-1)\mu)\min(1,(|a_1|(1-\beta))^2|\rho''(0)|)$. Let $\theta \in \Lambda$, we have, with the RIP hypothesis

$$\begin{aligned}
\xi(\theta) &:= 2(d+1)\max(\max_r |a_r|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})\|A\phi(\theta)-A\phi(\theta^*)\|_2 \\
&\leq 2(d+1)\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})k\sup_i \|A(b_i\delta_{s_i}-a_i\delta_{t_i})\|_h \\
&\leq 2(d+1)\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})k\sqrt{1+\gamma}\sup_i \|b_i\delta_{s_i}-a_i\delta_{t_i}\|_h \\
&\leq 2(d+1)\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})k\sqrt{1+\gamma}\beta|a_k|
\end{aligned} \tag{90}$$

where we wrote $\theta^* = \sum_i a_i\delta_{t_i}$ and $\theta = \sum_i b_i\delta_{s_i}$ such that $|s_i-t_i| \leq \epsilon/4$. The fact that $\beta \leq 1/2$ implies

$$\frac{\xi(\theta)}{\min(1,|a_1|^2(1-\beta)^2|\rho''(0)|)} \leq \frac{2k(d+1)|a_k|\sqrt{1+\gamma}\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})\beta}{\min(1,|a_1|^2|\rho''(0)|/4)} \tag{91}$$

Hence using the hypothesis that $\beta \leq \frac{(1-\gamma)(1-(k-1)\mu)\min(1,|a_1|^2|\rho''(0)|/4)}{k(d+1)|a_k|\sqrt{1+\gamma}\max(|a_k|\sqrt{m}D_{A,R},\sqrt{1+\gamma}\sqrt{|\rho''(0)|})}$ we have

$$\xi(\theta) \leq 2(1-\gamma)(1-(k-1)\mu)\min(1,(|a_1|(1-\beta))^2|\rho''(0)|). \tag{92}$$

□

Proof of Corollary 3.2. The set $\Lambda = \Lambda_{\theta^*,\beta,\epsilon/4}$ is an open set where the Hessian of g at Λ is positive as long as $\xi \leq 2(1-\gamma)(1-(k-1)\mu)\min(1,(|a_1|(1-\beta))^2|\rho''(0)|)$ with Theorem 3.1.

In this case g is convex on Λ. Theorem 3.1 also gives a uniform bound for the operator norm of the Hessian: $\|H\|_{op} \leq 2(1+\gamma)(1+(k-1)\mu)\max(1,(|a_k|(1+\beta))^2|\rho''(0)|)+\xi$ and g has Lipschitz gradient. We thus deduce from Corollary 1.1 that Λ is a basin of attraction.

Hence we just need to show that $\xi \leq 2(1-\gamma)(1-(k-1)\mu)\min(1,(|a_1|(1-\beta))^2|\rho''(0)|)$.

Let $\theta \in \Lambda$, we have, with the RIP hypothesis,

$$\begin{aligned}
\xi(\theta) &:= 2(d+1)\max(\max_r |a_r|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})(\|A\phi(\theta) - A\phi(\theta^*)\|_2 + \|e\|_2) \\
&\leq 2(d+1)\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})(k \sup_i \|A(b_i\delta_{s_i} - a_i\delta_{t_i})\|_h + \|e\|_2) \\
&\leq 2(d+1)\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})(k\sqrt{1+\gamma} \sup_i \|b_i\delta_{s_i} - a_i\delta_{t_i}\|_h + \|e\|_2) \\
&\leq 2(d+1)\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})(k\sqrt{1+\gamma}\beta|a_k| + \|e\|_2)
\end{aligned} \tag{93}$$

where we wrote $\theta^* = \sum_i a_i\delta_{t_i}$ and $\theta = \sum_i b_i\delta_{s_i}$ such that $|s_i - t_i| \leq \epsilon/4$. The fact that $\beta \leq 1/2$ and $\|e\|_2 \leq k\beta|a_k|$ implies

$$\frac{\xi(\theta)}{\min(1, |a_1|^2(1-\beta)^2|\rho''(0)|} \leq \frac{2k(d+1)|a_k|(1+\sqrt{1+\gamma})\max(|a_k|\sqrt{m}D_{A,R}, \sqrt{1+\gamma}\sqrt{|\rho''(0)|})\beta}{\min(1, |a_1|^2|\rho''(0)|/4)} \tag{94}$$

Hence using the hypothesis that $\beta \leq \frac{(1-\gamma)(1-(k-1)\mu)\min(1,|a_1|^2|\rho''(0)|/4)}{k(d+1)|a_k|(1+\sqrt{1+\gamma})\max(|a_k|\sqrt{m}D_{A,R},\sqrt{1+\gamma}\sqrt{|\rho''(0)|})}$, we have

$$\xi(\theta) \leq 2(1-\gamma)(1-(k-1)\mu)\min(1, (|a_1|(1-\beta))^2|\rho''(0)|) \tag{95}$$

□

A.5 Proofs for Section 4

Proof of Lemma 4.1. Remark that $g(\theta)$ does not depend on the ordering of the positions. Reorder $\theta_0 = (a,t)$ and $\theta_1 = (b,s)$ such that $t_1 < t_2 ... < t_k$ and $s_1 < s_2 ... < s_k$. Consider the function $g_1(\lambda) = g(\theta_\lambda)$ with $\theta_\lambda = (1-\lambda)\theta_0 + \lambda\theta_1$. Remark that g_1 is a continuous function of λ taking values $g_1(0) = g(\theta_0)$ and $g_1(1) = g(\theta_1)$. Hence, with the intermediate value theorem, there is λ such that $g(\theta_\lambda) = g_1(\lambda) = \alpha$. Moreover, denoting $\theta_\lambda = (a_\lambda, t_\lambda)$, we have, using the sorting of t and s, for $1 \leq i < k$,

$$\begin{aligned}
|t_{\lambda,i+1} - t_{\lambda,i}| &= |(1-\lambda)t_{i+1} + \lambda s_{i+1} - (1-\lambda)t_i - \lambda s_i| \\
&= (1-\lambda)|t_{i+1} - t_i| + \lambda|s_{i+1} - s_i| > (1-\lambda)\epsilon + \lambda\epsilon = \epsilon.
\end{aligned} \tag{96}$$

Hence $\theta_\lambda \in \Theta_{k,\epsilon}$. □

References

[1] 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.

[2] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[3] T. Blumensath. Sampling and reconstructing signals from a union of linear subspaces. *IEEE Transactions on Information Theory*, 57(7):4660–4671, 2011.

[4] A. Bourrier, M. Davies, T. Peleg, P. Perez, and R. Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *Information Theory, IEEE Transactions on*, 60(12):7928–7946, 2014.

[5] C. Boyer, Y. De Castro, and J. Salmon. Adapting to unknown noise level in sparse deconvolution. *Information and Inference: A Journal of the IMA*, 6(3):310–348, 2017.

[6] V. Cambareri and L. Jacques. Through the haze: a non-convex approach to blind gain calibration for linear random sensing models. *Information and Inference: A Journal of the IMA*, 2018.

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

[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 and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

[10] P. G. Ciarlet, B. Miara, and J.-M. Thomas. *Introduction to numerical linear algebra and optimisation.* Cambridge University Press, 1989.

[11] Y. De Castro, F. Gamboa, D. Henrion, and J.-B. Lasserre. Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *arXiv preprint arXiv:1502.02436*, 2015.

[12] V. Duval, P. Catala, and G. Peyré. A low-rank approach to off-the-grid sparse deconvolution. *preprint*, 2017.

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

[14] V. Duval and G. Peyré. Sparse regularization on thin grids i: the lasso. *Inverse Problems*, 33(5):055008, 2017.

[15] V. Duval and G. Peyré. Sparse spikes super-resolution on thin grids ii: the continuous basis pursuit. *Inverse Problems*, 33(9):095008, 2017.

[16] M. Golbabaee and M. E. Davies. Inexact gradient projection and fast data driven compressed sensing. *IEEE Transactions on Information Theory*, 2018.

[17] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

[18] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Compressive Statistical Learning with Random Feature Moments. *Preprint*, 2017.

[19] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez. Sketching for Large-Scale Learning of Mixture Models. Preprint, 2016.

[20] N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval. Compressive k-means. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 6369–6373. IEEE, 2017.

[21] S. Ling and T. Strohmer. Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing. *Information and Inference: A Journal of the IMA*, 2017.

[22] C. Poon, N. Keriven, and G. Peyré. Support localization and the fisher metric for off-the-grid sparse regularization. *arXiv preprint arXiv:1810.03340*, 2018.

[23] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

[24] 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.

[25] Y. Traonmilin, N. Keriven, R. Gribonval, and G. Blanchard. Spikes super-resolution with random fourier sampling. In *SPARS workshop 2017*, 2017.

[26] I. Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *IEEE Transactions on Information Theory*, 2018.

[27] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.