Estimation of off-the-grid sparse spikes with over-parametrized projected gradient descent: theory and application

Pierre-Jean Bénard, Yann Traonmilin, Jean-François Aujol, Emmanuel Soubies

Estimation of off-the-grid sparse spikes with over-parametrized projected gradient descent: theory and application

Pierre-Jean Bénard[1], **Yann Traonmilin**[1,*], **Jean-François Aujol**[1], **Emmanuel Soubies**[2]

[1] Université de Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400, Talence, France

[2] University of Toulouse, CNRS, IRIT, France

E-mail: * yann.traonmilin@math.u-bordeaux.fr

Abstract. In this article, we study the problem of recovering sparse spikes with over-parametrized projected descent. We first provide a theoretical study of approximate recovery with our chosen initialization method: Continuous Orthogonal Matching Pursuit without Sliding. Then we study the effect of over-parametrization on the gradient descent which highlights the benefits of the projection step. Finally, we show the improved calculation times of our algorithm compared to state-of-the-art model-based methods on realistic simulated microscopy data.

Keywords: off-the-grid sparse recovery, non-convex methods, over-parametrization, projected gradient descent, microscopy

1. Introduction

Off-the-grid super-resolution is the problem of recovering off-the-grid spikes from linear measurements. It has applications in many fields. A prime example is single molecule localization microscopy (SMLM) where positions and amplitudes of fluorescent molecules have to be recovered from their diffraction limited measurements. The signal of interest x_0 is modeled as a sum of K Dirac measures over $\mathbb{R}^d$:

$$x_0 = \sum_{i=1}^{K} a_i \delta_{t_i} \tag{1}$$

where $a = (a_1, \dots, a_K) \in \mathbb{R}^K$ are the amplitudes and $t = (t_1, \dots, t_K) \in \mathbb{R}^{K \times d}$ are the positions of the spikes. The vector $y \in \mathbb{C}^m$ of measurements is modeled as

$$y = Ax_0 + e, \tag{2}$$

where A is a linear operator that maps the space $\mathcal{M}$ of finite signed measures over $\mathbb{R}^d$ to $\mathbb{C}^m$ with m the number of observations and e a noise. For example, in SMLM with multi-angle total internal reflection fluorescence (MA-TIRF), the operator A is the measurement of the intensity of illuminated cells on a 3D grid (2D position plus incidence). Each measurement is the duality product $\langle x_0, \alpha_l \rangle$ between the spikes and the response of the system evaluated on a grid. For example, in SMLM each measurement corresponds to the duality product between the signal and a shifted version of the impulse response, i.e., $y_m = \langle x_0, h(\cdot - t_m) \rangle$ where h is the impulse response and t_m the position of the m-th camera pixel. In other fields, the operator A can be a finite number of Fourier measurements, i.e. each measurement is the duality product with a sinusoid at a given frequency, e.g. $\alpha_l(t) = e^{-j\langle \omega_l, t \rangle}$ for some frequencies ω_l.

A way to recover the true signal x_0 from (2) is to find the minimizer of a non-convex least-squares problem:

$$x^* \in \operatorname*{arg\,min}_{x \in \Sigma_{K,\epsilon}} \|Ax - y\|_2^2 \tag{3}$$

where

$$\Sigma_{K,\epsilon} := \left\{ \sum_{i=1}^{K} a_i \delta_{t_i} : a_i \in \mathbb{R}, t_i \in \mathbb{R}^d; \forall i, j \in \{1, \dots, K\}, i \neq j, \|t_i - t_j\|_2 > \epsilon, t_i \in \mathcal{D} \right\} \tag{4}$$

is a set modeling K spikes on a given domain $\mathcal{D} \subset \mathbb{R}^d$ with a separation constraint between spikes. Theoretical guarantees for the recovery of x_0 have been given in [26] for measurements operators A having a restricted isometry property (RIP). This property is verified for a sufficiently large number of random Fourier measurements (see Section 4 for details). Another approach is to consider the convex program in the space of measures

$$x^* \in \operatorname*{arg\,min}_{x \in \mathcal{M}} \|Ax - y\|_2^2 + \lambda \|x\|_{TV} \tag{5}$$

where $\|x\|_{TV}$ is the total variation for measures (i.e. the sum of absolute amplitudes for spikes).

In [12], using Fourier measurements up to a frequency f_c, sparse signals with a minimal distance between spikes of $2/f_c$ can be recovered exactly.

To solve (3), it has been shown [40, 38] that given an initialization sufficiently close to the true positions, an unconstrained gradient descent in the space of parameters converges to x^*. These results partly explain the success of greedy sliding continuous orthogonal matching pursuit (Sliding COMP, [29] called CL-OMP there) for the recovery of x^*. This algorithm adds a spike maximally correlated with the residual between observations and the current estimation at each iteration and performs a descent (the sliding step) on all parameters at each iteration. It was proposed in [7] (improving on [39]) to perform an over-parametrized continuous OMP without sliding (OP-COMP) followed by a projected gradient descent (PGD, projection on the separation constraint) to avoid the costly sliding step at each iteration of Sliding COMP. This method has been shown experimentally to be more efficient than Sliding COMP for large numbers of spikes in a synthetic random Fourier measurement setting. However, there is no theoretical study of the over-parametrized-COMP+ PGD method.

To solve the convex problem (5), Sliding Frank-Wolfe (SFW) [20] is the most efficient algorithm compared to the earlier dual methods [10, 17, 33], especially in higher dimension. It is the continuous version of the Frank-Wolfe algorithm [25]. In [20], Sliding Franke-Wolfe is guaranteed to converge to x^* if there exists a non-generate dual certificate, which is a difficult condition to check in practice (beyond the low-pass filtering case). The approach of SFW is to construct x^* by adding iteratively spikes and to perform a sliding step on the regularized function at each iteration. In practice the SFW is very close to Sliding COMP in terms of performed operations and recovery (see Section 5).

In this article, we focus on the non-convex formulation with COMP, and we propose a theoretical study of continuous orthogonal matching pursuit without sliding to provide insights in the success of the over-parametrized COMP + PGD method. We provide a fast implementation of this method and we evaluate its performance on synthetic SMLM data.

1.1. Contributions

In the following, we focus on the noiseless setting. We give the following contributions.

- In Section 2, we detail the implementation of Over-parametrized Continuous Orthogonal Matching Pursuit (without sliding) and Projected Gradient Descent. Compared to [7] where it was first introduced, we provide an accelerated implementation where we replace all gradient descents by their accelerated version using FISTA (Fast Iterative Shrinkage Thresholding Algorithm) with restart.
- In Section 3, we give theoretical guarantees for Continuous Orthogonal Matching Pursuit without sliding under a restricted isometry property (RIP) condition on

the measurement operator A. In general, COMP without sliding does not yield the true positions of x_0. However, we show that the output of the K-th first iterations of COMP without sliding still approximates the true positions with a precision depending on the quality of the RIP constant. Qualitatively, this shows that we can remove the sliding step from Sliding COMP and still be in a basin of attractions of x_0 if enough measurements are available. In practice, adding an over-parametrization permits to go beyond the quality of approximation given by our result at the K-th step.

- While over-parametrization permits to better localize spikes, we show experimentally in Section 4, that the Hessian of the considered over-parametrized functional becomes very ill-conditioned and even non-positive around the true solution. This shows the necessity of a projection step to reduce the number of spikes in order to estimate positions accurately in a computationally efficient way.
- Finally, in Section 5, we compare OP-COMP + PGD with Sliding COMP on two problems: spike estimation for microscope calibration and single molecule localization microscopy with data from the SMLM Challenge [35]. We show that OP-COMP + PGD estimates positions with accuracy at least as good as Sliding COMP with improved computation times. The computational improvement increases when the number K of illuminated molecules increases showing that we gain an order of magnitude with respect to K.

1.2. Related work

To solve directly the non-convex minimization problem (3), works are mainly focused on studying the success of descent algorithms in the space of parameters (amplitude and positions). It has been shown that, for operators having a restricted isometric property, we can estimate the size of the basin of attraction of the global minimum [38, 40]. This size is increasing explicitly with respect to the number of measurements through the RIP constant for e.g. random measurement operators. The greedy Sliding COMP has been used in the context of k-means clustering [30] or radar detection for off-the-grid targets [18]. Exact recovery of COMP has been studied for kernels of particular shape (e.g. exponential) [21, 22] (see Section 2). Our article focuses on approximate recovery. Still in the non-convex context, preconditioning with respect to amplitudes and positions to help local convergence for first-order methods has been studied in [16] when the number of particles is known.

The idea of projected gradient descent for non-convex problems has been used successfully in several domains such as sparse recovery (in the finite dimensional context), low rank matrix recovery [13] or in spectral compressed sensing [9]. In the specific case of off-the-grid sparse recovery, PGD (with a suboptimal initialization scheme) has been proposed in [39].

For convex methods solving (5), other than Sliding Frank-Wolfe, there is a whole body of works [10, 17, 33] based on solving the dual problem, which poses difficulties

in high dimensions, even if we can cite recent advances using methods with adaptive refining of a grid [23]. Concerning Sliding Frank-Wolfe, in the context of acoustic impulse response estimation, the idea of performing a single descent over all parameters after a greedy initialization is also used in [36].

Recently the idea of over-parametrization has emerged as a powerful tool to solve non-convex problems. Chizat and Bach [15] showed that for an infinite number of particles, the particle gradient descent converges to the global minimizer of a total-variation regularized problem lifting an original non-convex problem.

Our main tool for the study of OMP is the restricted isometry property, as introduced by Candès in compressed sensing [11]; it has many applications on inverse problems with, for examples, low-rank matrix factorization [14] and deep-learning [31]. In particular, in the finite dimensional case, the RIP guarantees the success of sparse recovery with OMP (e.g. [24]).

2. Sliding Continuous Orthogonal matching pursuit, Projected Gradient Descent and over-parametrized initialization: definitions and implementation

In this Section, we give the definition of COMP with or without sliding and of the Projected (accelerated) Gradient Descent.

2.1. (Sliding) COMP and Over-Parametrized COMP without sliding

Sliding Continuous OMP [30] is based on the discrete OMP algorithm (that is theoretically studied in [41]). Sliding COMP is described in Algorithm 1. We can choose to perform the sliding step or not, and even over-parametrize (parameter K) if we desire. We call over-parametrized COMP (OP-COMP) the simple execution of COMP without sliding with $K_{op} > K$ steps. It was observed in [7] that the complexity with respect to K with the sliding step is $O(K^2)$ and $O(K)$ without. From a computational point of view, this shows the interest of avoiding the sliding step for large number of spikes if possible.

Algorithm 1 (Sliding) Continuous Orthogonal Matching Pursuit algorithm (COMP).

procedure COMP($A, y, K,$ `is_sliding`)
  $r^{(0)} \leftarrow y$
  $t^{(0)} \leftarrow \{\}$
  for $k = 1 \to K$ **do**
    $t_{new} \leftarrow \arg\max_t \langle A\delta_t, r^{(k-1)} \rangle$ ▷ Add new spike
    $t^{(k)} \leftarrow t^{(k-1)} \cup \{t_{new}\}$
    $a^{(k)} \leftarrow \arg\min_a \|A \sum_{i=1}^{k} a_i \delta_{t_i^{(k)}} - y\|_2^2$
    if `is_sliding` **then**
      $a^{(k)}, t^{(k)} \leftarrow$ `descent`$(a^{(k)}, t^{(k)})$ ▷ Sliding step
    end if
    $r^{(k)} \leftarrow y - A \sum_{i=1}^{k} a_i^{(k)} \delta_{t_i^{(k)}}$
  end for
  return $a^{(K)}, t^{(K)}$
end procedure

In [21, 22], it is shown that COMP without sliding recovers exactly K spikes in K steps for exponential impulse response but these results cannot be applied for low-pass filter such as Gaussian filters which are more common in signal and image processing. In particular the Gaussian case is close to the example of microscopy (see Section 5). Our analysis of COMP without sliding in the next section relies on a restricted isometry condition which includes these two examples.

2.2. Our method: OP-COMP + Projected Gradient Descent

It was observed in [39, 7] that over-parametrization (i.e. OP-COMP) permits to ensure the approximate localization of all the spikes in the signal even without using the sliding step (see also Section 4 and related work for a discussion on very heavy over-parametrization). It is a way to go beyond the theoretical results from the next Section. Unfortunately, the Hessian of the considered functional becomes ill-conditioned and non-positive yielding very slow convergence of classical gradient descent to the true positions as we will see in Section 4. Hence the motivation to add a projection step in the descent.

A legitimate question to ask ourselves is when to stop the over-parameterization. In our implementation, we use a relative criterion based on the observation of the ground truth $y = Ax_0$ and the observation of our estimation Ax,

$$\frac{\|Ax - Ax_0\|_2}{\|Ax_0\|_2} \leq \epsilon. \tag{6}$$

This criterion gives us a good estimation on when to stop the over-parameterization as it guarantees that the over-parametrized solution reaches the same objective value as a K-spike estimation that estimates well true positions. Indeed, under a restricted isometry property assumption on A with constant γ with respect to a kernel norm $\|\cdot\|_h$

(see Definition 4), given $x \in \Sigma_{K,\epsilon}$, we have that $\|Ax - Ax_0\|_2 \leq \epsilon \|Ax_0\|_2$ implies

$$\sqrt{1-\gamma}\|x - x_0\|_h \leq \epsilon\sqrt{1+\gamma}\|x_0\|_h \qquad \text{i.e.} \qquad \sqrt{\frac{1-\gamma}{1+\gamma}}\frac{\|x - x_0\|_h}{\|x_0\|_h} \leq \epsilon. \tag{7}$$

This means that if A has a sufficiently good RIP constant (i.e. $\sqrt{(1-\gamma)/(1+\gamma)}$ is close to 1), we can have a good control on the relative error of the spikes only with their observations. In our experiments (see Section 5), we observe that a relative stopping criterion of $\epsilon = 5\%$ yields accurate estimations.

To reduce the number of spikes, we perform a heuristic projection P_ϵ based on the separation constraint at each descent step. If two spikes are too close from each other, their positions and amplitudes are merged, i.e. the projected descent is described by

$$(a^{(n+1)}, t^{(n+1)}) = P_\epsilon\left(\texttt{descent}(a^{(n)}, t^{(n)})\right) \tag{PGD}$$

where our actual fast implementation of $\texttt{descent}(a^{(n)}, t^{(n)})$ is described in Section 2.3.

The heuristic projection P_ϵ is described in Algorithm 2. An important step in our heuristic projection P_ϵ is that the merged positions are barycenters (using absolute value of amplitudes) of previous positions. This way, we ensure that the projection step yields a spike located in the convex hull of previous considered positions. As we suppose that for a small enough ϵ, the positions of all merged spikes belong in the same basin of attraction, the output of P_ϵ also belongs in this basin of attraction. Note that P_ϵ also performs a thresholding of very small amplitudes. It is possible to consider theoretical "projections" on the separation constraints, e.g. considering for some given norm on $\mathcal{M}$ the problem

$$\inf_{x \in \Sigma_{K,\epsilon}} \left\| x - \sum_{i=1}^{K} a_i \delta_{t_i} \right\|. \tag{8}$$

However, the choice of norm (kernel norm, Wasserstein distance) and the resolution of this problem is not straightforward. To show the convergence of the whole method, one would need to guarantee that the projection keeps positions in basins of attraction of the functional, which is out of the scope of this paper.

Algorithm 2 Projection on the separation constraint P_ϵ.

procedure P$(a, t, \epsilon, \texttt{threshold})$
  $\texttt{sort}(a_i, t_i)_{i=1,\dots,k}$ such that $|a_1| \geq \cdots \geq |a_k|$
  for $i = 1, \dots, k$ **do**
    if $a_i = 0$ **then**
      Skip iteration
    else if $a_i \leq \texttt{threshold}$ **then**
      $a_i \leftarrow 0$
    else
      Get $J = \{j, j \geq i, \|t_i - t_j\|_2 \leq \epsilon\}$
      $t_i \leftarrow \frac{1}{\sum_{j\in J} |a_j|} \sum_{j\in J} |a_j| t_j$
      $a_i \leftarrow \sum_{j\in J} a_j$
      $a_j \leftarrow 0$ for all $j \in J, j \neq i$
    end if
  end for
  return $(a_i, t_i)_{i\in I}$ with $I = \{i, a_i \neq 0\}$
end procedure

In the experiments, we will compare K step Sliding COMP with OP-COMP followed by accelerated PGD.

2.3. Implementation of descent algorithms with FISTA with restart

We describe the specific choices we made for each part of COMP (Algorithm 1).

- Calculating $\arg\max_t \langle A\delta_t, r^{(k-1)}\rangle$: we perform a descent on $-\langle A\delta_t, r^{(k-1)}\rangle$ using FISTA restart (see Algorithm 3) until convergence to an estimate $t_{\texttt{conv}}$. For spatial measurements on a grid with impulse response concentrated around spikes we initialize the descent with the minimum on the grid (as is done in the experimental part of this article). For Fourier measurements, it was observed that a random initialization on $\mathcal{D}$ was efficient [7].
- Calculating $\arg\min_a \|A\sum_{i=1}^k a_i \delta_{t_i^{(k)}} - y\|_2^2$: This can be solved using the conventional finite dimensional least-squares method. Indeed, we have

$$A\sum_{i=1}^{k} a_i \delta_{t_i} = Ba \tag{9}$$

  where B is the matrix whose columns are defined by $B_i = A\delta_{t_i}$.
- Calculating $\texttt{descent}(a^{(k)}, t^{(k)})$: as it is crucial to perform this step quickly given the dimension of the objects, we use the FISTA restart algorithm (also called monotone FISTA) [1] on the function

$$g(a, t) = \left\| A\sum_{i=1}^{k} a_i \delta_{t_i} - y \right\|_2^2. \tag{10}$$

Indeed the inertia introduced by FISTA can cause the following problem. Instead of decreasing towards the minimum of the basin of attraction, it can spiral around. This is called the rebound phenomenon. This bouncing effect slows the convergence of the system. To counter this effect, the FISTA restart method of resetting the inertia at the lowest stage of the rebound was introduced. While some papers work on an estimation of when to restart the inertia [2], we can simply look at when the energy of the system increases and reset the inertia. No theoretical works on this prove the acceleration of this method compared to FISTA, but it is very efficient in practice and simple to implement, [4, 5, 32].

Algorithm 3 FISTA restart for a function $f(z)$.

procedure FISTA RESTART(z, N, τ)
  $x^{(0)} \leftarrow z$
  $y^{(0)} \leftarrow z$
  $k \leftarrow 0$
  $n \leftarrow 0$
  repeat
    $k \leftarrow k + 1$
    $n \leftarrow n + 1$
    $x^{(k)} \leftarrow y^{(k-1)} - \tau \nabla f(y^{(k-1)})$
    $y^{(k)} \leftarrow x^{(k)} + \frac{n-1}{n+2}(x^{(k)} - x^{(k-1)})$
    if $f(x^{(k)}) > f(x^{(k-1)})$ **then**
      $n \leftarrow 0$
    end if
  until $k \geq N$
  return $x^{(K)}$
end procedure

We note that for all descent algorithms used in Sliding COMP and OP-COMP + PGD, we use the FISTA restart. In PGD, after each projection, if at least two spikes are merged, we reset the inertia of the system. Moreover, since we want our spikes to be in the domain $\mathcal{D}$, we perform another projection which clips the spikes outside of $\mathcal{D}$.

3. Theoretical study of COMP without sliding

We show in this section that with some additional hypotheses, it is not necessary to perform the descent step in Sliding COMP to obtain a controlled approximation of the positions of the spikes. We first proceed to introduce some facts and assumptions on the linear operator A and on Dirac measures.

3.1. Definitions

We first introduce the notion of kernel norm which permits to use the restricted isometry property in off-the-grid spike estimation (see e.g. [37]),

Definition 1 (Kernel, scalar product and norm)**.** *For finite signed measures over $\mathbb{R}^d$, the Hilbert structure indexed 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 ν_1 and ν_2,*

$$\langle \nu_1, \nu_2 \rangle_h = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(t_1, t_2) \mathrm{d}\nu_1(t_1) \mathrm{d}\nu_2(t_2). \tag{11}$$

We can consequently define

$$\|\nu_1\|_h^2 = \langle \nu_1, \nu_1 \rangle_h. \tag{12}$$

We have the relation

$$\|\nu_1 + \nu_2\|_h^2 = \|\nu_1\|_h^2 + 2\langle \nu_1, \nu_2 \rangle_h + \|\nu_2\|_h^2. \tag{13}$$

For the rest of the paper, we use the Gaussian kernel $h(t, s) = \mathrm{e}^{-\|t-s\|^2/(2\sigma^2)}$ taken from [26]. We have that h follows Assumptions 1 hereafter.

A fundamental object used in our study is the concept of dipole.

Definition 2 ((ξ-)Dipole and separation)**.** *A ξ-dipole (noted dipole for simplicity) is a measure $\pi_1 = a_1\delta_{t_1} - b_1\delta_{s_1}$ where $\|t_1 - s_1\|_2 \leq \xi$, with $a_1, b_1 \in \mathbb{R}$. Two dipoles $\pi_1 = a_1\delta_{t_1} - b_1\delta_{s_1}$ and $\pi_2 = a_2\delta_{t_2} - b_2\delta_{s_2}$ are ϵ-separated if their support are strictly ϵ-separated (with respect to the l^2-norm on $\mathbb{R}^d$), i.e. if $\|t_1 - t_2\|_2 > \epsilon$, $\|t_1 - s_2\|_2 > \epsilon$, $\|s_1 - t_2\|_2 > \epsilon$ and $\|s_1 - s_2\|_2 > \epsilon$.*

A typical assumption is that the kernel makes separated dipoles incoherent, i.e. their scalar product yields a small residue controlled by the kernel.

Assumption 1. *A Gaussian kernel h follows this assumption if there is a constant μ_h such that, for all two $\frac{\epsilon}{3}$-separated dipoles, $\langle \pi_1, \pi_2 \rangle_h \leq \mu_h \|\pi_1\|_h \|\pi_2\|_h$ (mutual coherence). Since we fix the kernel h for this paper, for simpler notations, we note $\mu := \mu_h$.*

With a Gaussian kernel h, we have the following properties.

Lemma 3.1 (Consequences of choosing a Gaussian kernel)**.** *Let $h(t, s) = \mathrm{e}^{-\|t-s\|_2^2/(2\sigma^2)}$ a Gaussian kernel of variance $2\sigma^2$.*
Then,

(i) the kernel h is symmetrical with respect to 0 and translation invariant;

(ii) we have $h(t, t) = 1 = \max_{t \in \mathbb{R}^d, s \in \mathbb{R}^d} |h(t, s)|$;

(iii) if h verifies Assumption 1, then for any two $\frac{\epsilon}{3}$-separated spikes δ_t, δ_s,

$$\langle \delta_t, \delta_s \rangle_h = \mathrm{e}^{-\frac{\|t-s\|_2^2}{2\sigma^2}} \leq \mu \|\delta_t\|_h \|\delta_s\|_h = \mu. \tag{14}$$

The following Lemma is an approximate Pythagorean identity where the norm of the sum of dipoles is approximately equal to the sum of the norm of the dipoles.

Lemma 3.2. *(from [37]) Suppose for all two $\frac{\epsilon}{3}$-separated dipoles, $\langle \pi_1, \pi_2 \rangle_h \leq \mu \|\pi_1\|_h \|\pi_2\|_h$ (mutual coherence). Then for k, $\frac{\epsilon}{3}$-separated dipoles $\pi_1, \dots, \pi_k$ such that $\max_i \|\pi_i\|_h > 0$, we have*

$$1 - (k-1)\mu \leq \frac{\| \sum_{i=1,k} \pi_i \|_h^2}{\sum_{i=1,k} \|\pi_i\|_h^2} \leq 1 + (k-1)\mu. \tag{15}$$

To define the restricted isometry property (RIP), we need the notion of secant set.

Definition 3 (Secant set)**.** *The secant set $\mathcal{S}(\Sigma_{K,\epsilon})$ of the model set $\Sigma_{K,\epsilon}$ is defined as*

$$\mathcal{S}(\Sigma_{K,\epsilon}) := \{x - y, x \in \Sigma_{K,\epsilon}, y \in \Sigma_{K,\epsilon}\}. \tag{16}$$

Note: $\mathcal{S}(\Sigma_{K,\epsilon})$ can be written as $\Sigma_{K,\epsilon} - \Sigma_{K,\epsilon}$.

We can now introduce the Restricted Isometry Property (see [26]).

Definition 4 (RIP)**.** *The linear operator A has the γ-RIP on $\mathcal{S}(\Sigma_{K,\epsilon})$ with constant $0 < \gamma < 1$ if for all $x \in \mathcal{S}(\Sigma_{K,\epsilon})$:*

$$(1-\gamma)\|x\|_h^2 \leq \|Ax\|_2^2 \leq (1+\gamma)\|x\|_h^2. \tag{17}$$

Suppose that the operator A has the γ-RIP on $\mathcal{S}(\Sigma_{K,\epsilon})$ and consider ϵ-separated dipoles π_i. Then we have

$$(1-\gamma) \left\| \sum_{i=1}^{k} \pi_i \right\|_h^2 \leq \left\| A \sum_{i=1}^{k} \pi_i \right\|_2^2 \leq (1+\gamma) \left\| \sum_{i=1}^{k} \pi_i \right\|_h^2. \tag{18}$$

Finally we introduce the set $Z_{l,\xi,\epsilon}$ of ξ-concentrated dipoles, pairwise ϵ-separated,

$$Z_{l,\xi,\epsilon} = \Bigg\{ \sum_{i=1}^{l} \pi_i : \pi_i = a_i \delta_{t_i} - b_i \delta_{s_i};\ a, b \in \mathbb{R}^l;\ t_i \in \mathcal{D}; s_i \in \mathbb{R}^d;\ \forall i \in \{1, \dots, l\}, \\ \|t_i - s_i\|_2 \leq \xi;\ \forall i, j \in \{1, \dots, l\}, i \neq j, \pi_i \text{ and } \pi_j \text{ are } \epsilon\text{-separated.} \Bigg\} \tag{19}$$

3.2. Main theorem

Let us first state our main result.

Theorem 3.3. *Let $K \in \mathbb{N}$ with $K > 0$. Let $\epsilon > 0$. Let $x_{K,\epsilon} = \sum_{i=1}^{K} a_i \delta_{t_i} \in \Sigma_{K,\epsilon}$ where $\|a\|_\infty = |a_1| \geq |a_2| \geq \dots \geq |a_K|$. Denote $\alpha = |a_1|/|a_K|$. Assume that the linear operator A has the γ-RIP on $\mathcal{S}(\Sigma_{K,\frac{\epsilon}{3}})$ with a Gaussian kernel h with variance σ^2 following Assumption 1. We define $s_1^*, \dots, s_K^*$ the ordered K-th first output of COMP without sliding. Let*

$$\xi := \sqrt{2\sigma^2 \ln \left(\frac{1}{1 - (4K\alpha - 1)(\mu + \gamma)} \right)}. \tag{20}$$

Suppose

$$\mu + \gamma \leq \frac{1}{5K\left(4K\alpha - 1\right)} \tag{21}$$

and

$$\epsilon^2 > 18 \ln\left(\frac{10}{9}\right) \sigma^2. \tag{22}$$

Then for every $l \in \{0, \dots, K\}$*, there exists* $i \in \{0, \dots, K\}$*, such that* $\|s_l^* - t_i\|_2 \leq \xi$*.*

This result states that if the unknown signal follows a separation assumption, the operator A has a γ-RIP on such separated signals with respect to a kernel norm following the mutual coherence assumption, then we can bound the estimation error of the positions. The main hypothesis of this Theorem is the condition that $\mu+\gamma = O\left(\frac{1}{K^2}\right)$ which seems restrictive; it is also dependent on the ratio of amplitudes. We express in the following a second theorem where the amplitude are known and set to 1, yielding more favorable conditions. We attribute this restrictive conditions to the fact that we do not have exact orthogonality between spikes in this continuous setting compared to the discrete finite dimensional case ([24, Theorem 6.24]).

The hypothesis $\epsilon^2 > 18 \ln\left(\frac{10}{9}\right) \sigma^2$ is guaranteed if

$$\epsilon > 1.37\sigma \tag{23}$$

This inequality means that the minimum separation needed for a full recovery of the spikes' positions is larger than the standard deviation of the kernel h. This is in accordance with the idea that if two spikes are too close with respect to the kernel width, their observations are not separated enough to get recovered independently. Note that we did not look to achieve the best constants to keep the proof as simple as possible.

The localization error of the spikes is bounded by ξ which converges to 0 when $\mu + \gamma$ is sufficiently small. Hence, for very well conditioned operators A, the K step COMP without sliding has good recovery guarantees. Note that for a final descent on parameters to converge, the bound ξ must be smaller than the size of basins of attraction of g. In fact, using results on basins of attraction from [38] we just proved that there exists a linear operator (with possibly a very large number of measurements) where COMP with a final gradient descent converges to x_0. For example, with a linear operator composed of a large number of random Fourier measurements following the RIP, this result is true (see [38] for a precise study of the basins of attraction of (3)).

The proof of this theorem is an induction on the steps of COMP where approximate recovery at one step guarantees approximate recovery at the next steps. The main difficulty is to control all residuals generated by the fact that we only have approximate orthogonality between atoms (through the kernel norm).

3.3. Theorem for signals with fixed amplitudes

The previous Theorem 3.3 is a general case that applies for signals with variable amplitudes. For sparse signals of the form $x_0 = \sum_{i=1}^{K} \delta_{t_i}$, we give a theorem with weaker conditions.

Theorem 3.4. *Let $K \in \mathbb{N}$ with $K > 0$. Let $\epsilon > 0$. Let $x_{K,\epsilon} = \sum_{i=1}^{K} \delta_{t_i} \in \Sigma_{K,\epsilon}$. Assume that the linear operator A has the γ-RIP on $\mathcal{S}(\Sigma_{K,\frac{\epsilon}{3}})$ with a Gaussian kernel h with variance σ^2. We define $s_1^*, \dots, s_K^*$ the ordered K-th first output of COMP without sliding. Let*

$$\xi := \sqrt{2\sigma^2 \ln\left(\frac{1}{1-(4K-1)(\mu+\gamma)}\right)}. \tag{24}$$

Suppose

$$\mu + \gamma \leq \frac{1}{10\,(4K-1)} \tag{25}$$

and

$$\epsilon^2 > 18 \ln\left(\frac{10}{9}\right)\sigma^2. \tag{26}$$

Then for every $l \in \{0, \dots, K\}$, there exists $i \in \{0, \dots, K\}$, $\|s_l^* - t_i\|_2 \leq \xi$.

3.4. Relation between the RIP hypothesis and number of measurements

The major difference between the two theorems is the less restrictive upper bound on $\mu + \gamma$. Instead of having an upper bound of the order $\frac{1}{K^2}$, it is just of the order $\frac{1}{K}$ without dependency on the ratio of amplitudes. This means that the Gaussian kernel can be larger as it implies a less strict condition on the observation to observe a full recovery of the true signal.

In the context of random Gaussian measurements we can guarantee that the condition on γ is verified if the number of measurements m is of order $\frac{1}{\gamma^2}K^2 d$ times some logarithmic terms as demonstrated in [27]. In our case, this gives a number of measurements m needed to guarantee an upper-bound on γ of order $K^6 d$. For signals with fixed amplitudes, m must be of order of $K^4 d$.

For the discrete sparse recovery case, success of OMP is guaranteed (neglecting log terms) for $m = O(K)$. Thus an open question is to determine if our conditions on γ can be loosened to yield better recovery results with respect to the number of spikes K.

4. Ill-posedness of the over-parametrized problem

In this section, we show with simple experiments that, even if over-parametrization permits to minimize the energy (3) (see next Section), it leads to badly conditioned problem leading to slow and poor convergence of gradient descent. This illustrates the fact that performing a projection in the descent is a key part of our algorithm for a fast convergence. The code for these experiments is available for download at [8].

4.1. Presentation of the experiments

We place ourselves in the case of the recovery of one spike (note that this case is often very informative for the study of limits of super-resolution algorithms [19, 38]). Results on basins of attractions show that if the descent is initialized sufficiently close to the

observed spike, then under the RIP condition there will be convergence to the desired position (at a linear rate).

We place ourselves in 2D, i.e. $x_0 = \delta_t$ with $t = 0$. We observe this signal through random Fourier measurements (where the RIP with a Gaussian kernel is guaranteed theoretically). For this example, we take the number of measurements to be $m = 40$. Then we initialize our signal $x_{\texttt{init}}$ to be a sum of spikes positioned near the origin. This aims to recreate the output of OP-COMP.

For our experiments, we compare four cases.

- Case 1: we set $x_{\texttt{init}} = \delta_s$ where s is close to t, $\|t - s\| = 0.2$. In this case the number of true spikes is equal to the number of initialized spikes and there is no need for a projection.
- Case 2: we set $x_{\texttt{init}} = \frac{1}{2}\sum_{i=1}^{2} \delta_{t_i}$ where $t_1 = (0.2, 0)$ and $t_2 = (-0.2, 0)$.
- Case 3: we set $x_{\texttt{init}} = \frac{1}{5}\sum_{i=1}^{5} \delta_{s_i}$ where for all $i = 1, \dots, 5$, $\|t - s_i\|_2 = 0.2$ and each s_i is equally distant from the other s_i.
- Case 4: we set $x_{\texttt{init}} = \frac{1}{5}\sum_{i=1}^{5} \delta_{s_i}$ where for all $i = 1, \dots, 5$, $\|t - s_i\|_2 = 10^{-4}$ and each s_i is equally distant from the other s_i. The initialized spikes are very close to the true spike (same as case 3 but initialized very close to the solution).

For each case, we perform a simple gradient descent without projection for a large number of iterations and with a step selected with a line search.

A classical way to study the well posedness of first order algorithms is to calculate the Hessian H of the function g in (10) (calculated in [37, Proposition 2.2]). The conditioning and non-negativeness of H permits to show local convexity and gives information on convergence speed of gradient descent. We recall the expression of the condition number of H, noted $\kappa(H)$, with H a complex square matrix,

$$\kappa(H) = \frac{|\lambda_{min}(H)|}{|\lambda_{max}(H)|} \tag{27}$$

where $|\lambda_{min}(H)|$ and $|\lambda_{max}(H)|$ are respectively the moduli of the minimal and maximal eigenvalues of H.

4.2. Results and remarks

In Case 1, the estimated spike converges to the true spike, see Figure 1. The norm of the residue decreases to the numerical precision 10^{-14}. For the smallest eigenvalue of H, it is negative only for the first iteration (we set manually the initialization). But after its first iteration of Gradient Descent, it goes back to being positive and stabilizes at $\lambda_{min} \approx 2$, see Figure 2. For the condition number of H, it also stabilizes at $\kappa(H) \approx 25$. This value is fairly low compared to our following tests. The positiveness of λ_{min} indicates that H is positive definite at our estimated signal, i.e. it is in a basin of attraction.

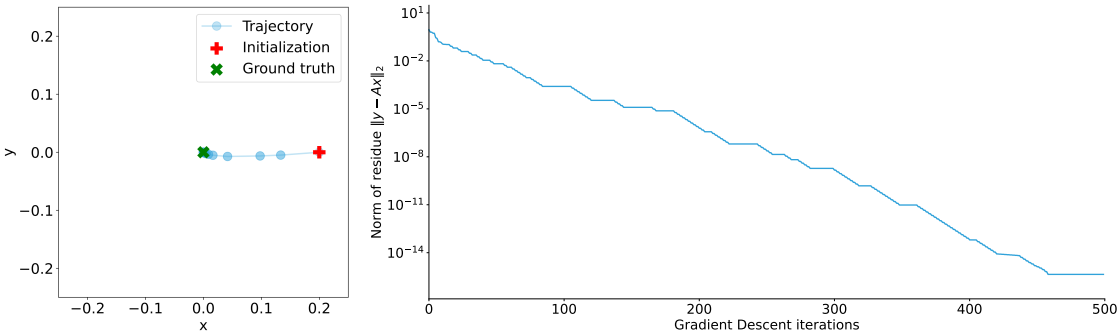

Figure 1: (left) Iterations of intermediate points $x_{\texttt{temp}}$ and trajectory from $x_{\texttt{init}} = \delta_s$ toward x_0 using Gradient Descent. (right) Norm of residue $\|y - Ax_{\texttt{temp}}\|_2$ at each iteration of Gradient Descent.

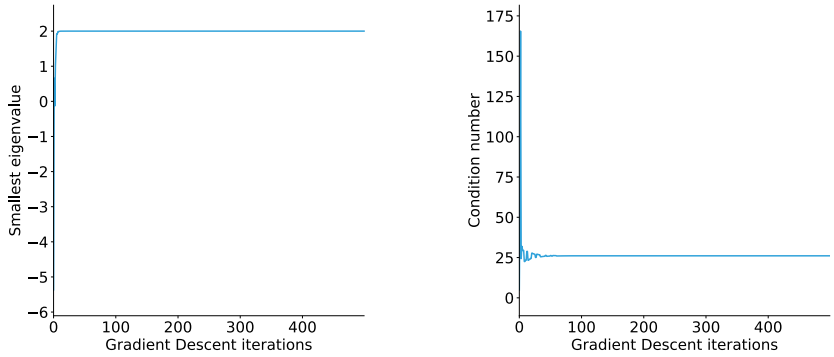

Figure 2: (left) Smallest eigenvalue and (right) Condition number of H at each iterations of the Gradient Descent for a one spike signal.

In Case 2, the trajectories of the spikes for the over-parameterized signal seem to converge towards the ground truth, Figure 3. However, when we observe the norm of the residue of the system, it decreases very slowly and does not have the same linear convergence rate as Case 1. The estimation error on positions is approximately 10^{-3} after 10 000 steps. When looking at the smallest eigenvalue and the conditioning of the system, we observe that at least one eigenvalue of the Hessian is negative, Figure 4. This indicates that the estimated over-parametrized functional is not convex around the global minimizer. When looking at the conditioning of the Hessian, while manageable far from the global optimum, it keeps increasing during convergence. This implies that the system is numerically unstable as it gets closer to the global minimizer.

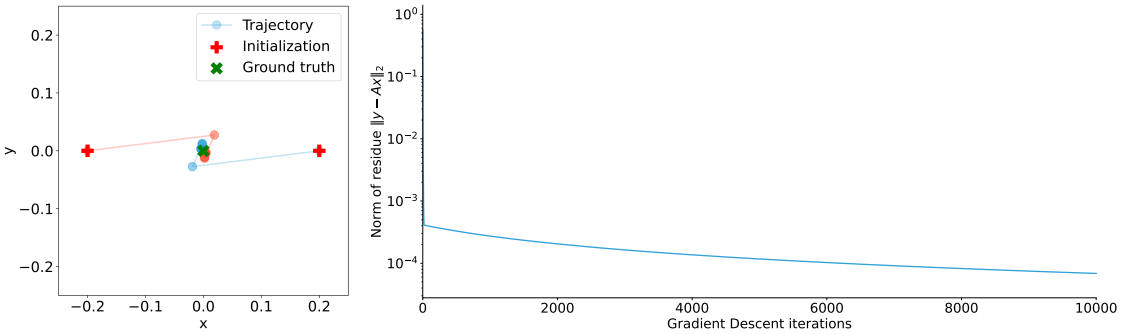

Figure 3: (left) Iterations of intermediate points $x_{\texttt{temp}}$ and trajectory from $x_{\texttt{init}} = \frac{1}{2}\delta_{s_1} + \frac{1}{2}\delta_{s_2}$ toward x_0 using Gradient Descent. (right) Norm of residue $\|y - Ax\|_2$ at each iteration of Gradient Descent.

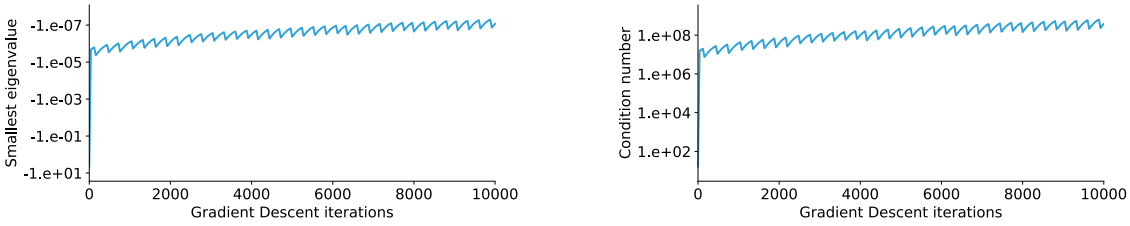

Figure 4: (left) Smallest eigenvalue and (right) Condition number of H at each iterations of the Gradient Descent for a two spikes over-parameterized signal.

In Case 3, we observe the same phenomena as in Case 2. The five spikes of the over-parameterized signal converge towards the solution. However, this convergence is very slow, Figure 5. For the smallest eigenvalue of the Hessian, it is still negative but gets closer to 0 through the iterations of Gradient Descent. As for the condition number of H, it increases and leads to computational instability (Figure 6). The estimation error on positions is approximately 10^{-3} after 10 000 steps.

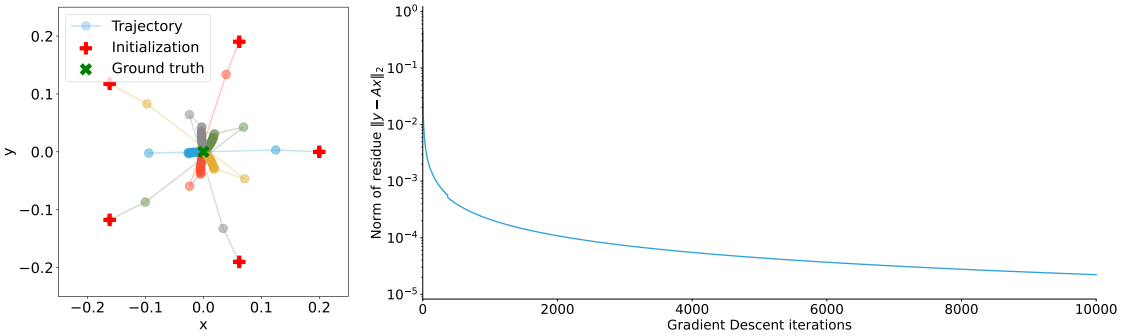

Figure 5: (left) Iterations of intermediate points $x_{\texttt{temp}}$ and trajectory from $x_{\texttt{init}} = \frac{1}{5}\sum_{i=1}^{5}\delta_{s_i}$ toward x_0 using Gradient Descent. (right) Norm of residue $\|y - Ax\|_2$ at each iteration of Gradient Descent.

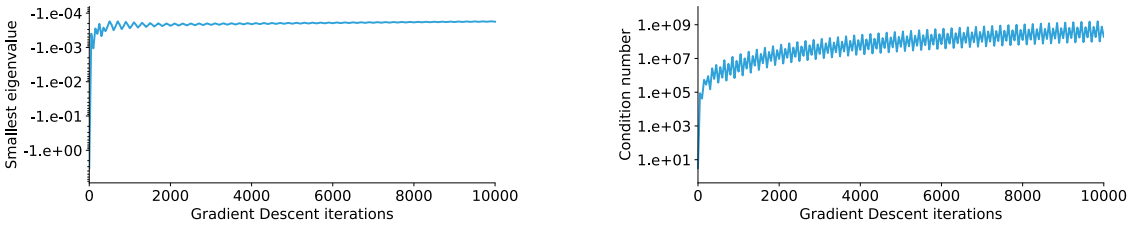

Figure 6: (left) Smallest eigenvalue and (right) Condition number of H at each iterations of the Gradient Descent for a five spike over-parameterized signal.

The idea for Case 4 is to see what happens when we initialize the spikes closer to the solution (distance lower than 10^{-5}) than the final estimation obtained after all the iterations of a Gradient Descent from Case 3. In Case 4, the descent does not seem to bring our initialized spikes closer to x_0, see Figure 7. Between the first and last iterations of Gradient Descent, the residue's difference is in the order of 10^{-15} and the residual is of the order of 10^{-10}. There is a convergence of the iterates towards x_0 but it is very slow. At this scale, we attribute this behavior to the bad conditioning and non-convexity of the functional: the smallest eigenvalue is very small ($\lambda_{min} \approx -10^{-15}$) and does not get positive during any step of the Descent, see Figure 8.

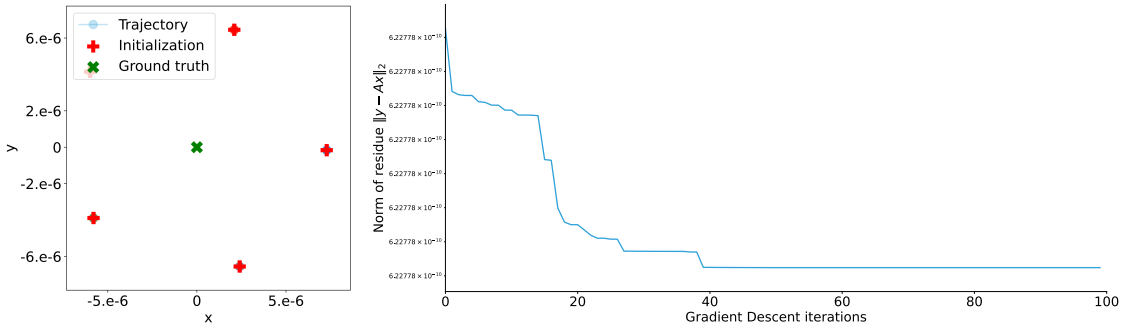

Figure 7: (left) Iterations of intermediate points $x_{\texttt{temp}}$ and trajectory from $x_{\texttt{init}} = \frac{1}{5}\sum_{i=1}^{5} \delta_{s_i}$ toward x_0 using Gradient Descent. (right) Norm of residue $\|y - Ax\|_2$ at each iteration of Gradient Descent.

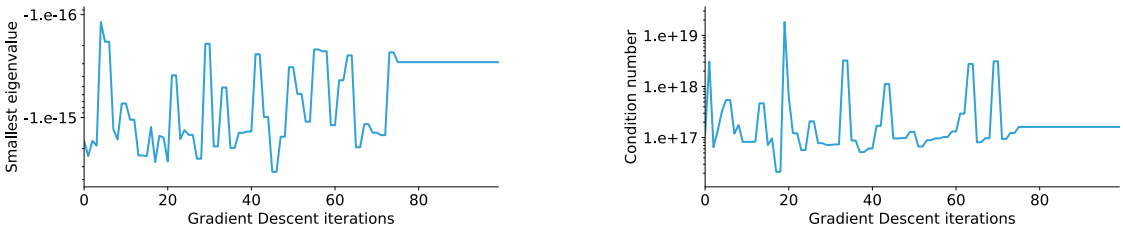

Figure 8: (left) Smallest eigenvalue and (right) Condition number of H at each iterations of the Gradient Descent for a five spikes over-parameterized signal.

We see that when the number of initialized spikes is larger than the number of true spikes, the signal does not converge towards the ground truth with good accuracy and in finite time. Gradient iterations do not improve the solution due to the ill-posedness of this situation. A solution for this problem is to reduce the number of spikes during the descent. In practice, the solution that we propose is to merge spikes close to each other (see Algorithm 2) to avoid dealing with ill conditioned problems and project the iterates into a convex basin of attraction of the global minimum. We must remark that the choice of the separation criterion for merging spikes is a trade-off between accuracy of the algorithm (resolving spikes close to each other) and computational speed (project as fast as possible in the descent).

5. Experiments

In this section, we apply our OP-COMP + PGD algorithm to two microscopy problems and compare its performances to the state-of-the-art method Sliding COMP. More precisely, we consider the problems of microscope calibration from fluorescent microbeads acquisitions (Section 5.4) and super-resolution reconstruction from SMLM data (Section 5.5). While both problems are related to the recovery of sparse spike signals, in the former microbeads usually follow a minimal separation distance in line with the assumptions of our theoretical results. The code for the following experiments is available for download at [8].

5.1. The MA-TIRF model

All reported experiments are performed with the multi-angle total internal reflection fluorescence (MA-TIRF) model, which we briefly describe here. The principle of TIRF is to illuminate the sample with an incident angle in the regime of total reflection [3]. In this regime, although all the incident light is reflected, an evanescent wave is created above the glass coverslip. It allows for the excitation of fluorophores within a thin layer of few hundred of nanometers at the basal surface of cells. MA-TIRF then simply consists in the consideration of a set of acquisitions for multiple TIRF angles. As the axial decay of the evanescent wave is related to the incident angle, MA-TIRF data

convey information about the axial position of fluorophores. Let $y \in \mathbb{R}^{N_1 \times N_2 \times K_{\texttt{angle}}}$ be our observation with N_1 and N_2 the number of pixels for the first two dimensions and $K_{\texttt{angle}}$ the number of incidence angles. Each entry of y is expressed as a weighted sum of the application of an impulse response $\alpha_{i,k}$, with $i \in \{1, \dots, N_1 N_2\}$, $k \in \{1, \dots, K_{\texttt{angle}}\}$, on each spike of $x_0 = \sum_{j=1}^{K} a_j \delta_{t_j}$ with $t_j \in \mathbb{R}^3$, i.e. $y_{i,k} = \sum_{j=1}^{K} a_j \alpha_{i,k}(t_j)$. The expression of $\alpha_{i,k}(t_j)$ for the MA-TIRF model is given by

$$\alpha_{i,k}(t_j) := \frac{\xi(t_{j,3}) \mathrm{e}^{-s_k t_{j,3}}}{2\pi\sigma_1\sigma_2} \int_{\Omega_i} \mathrm{e}^{-\left(\frac{(t_{j,1}-s_1)^2}{2\sigma_1^2} + \frac{(t_{j,2}-s_2)^2}{2\sigma_2^2}\right)} \mathrm{d}s_1 s_2 \tag{28}$$

where $\xi(t_{j,3}) = \left(\sum_{k=1}^{K_{\texttt{angle}}} \mathrm{e}^{-2 s_k t_{j,3}}\right)^{-1/2}$ and Ω_i the i-th camera pixel region. It corresponds to the combination of the evanescent excitation in the axial (third) dimension and the lateral (first two dimensions) convolution with the impulse response, which we approximate by a Gaussian kernel. For further details on this model, we refer the reader to [20]. To implement this continuous formulation, we discretize it to obtain

$$\alpha_{i,k}(t_j) = \frac{\xi(t_{j,3}) \mathrm{e}^{-s_k t_{j,3}}}{4} \Big(\operatorname{erf}(v_x^+) - \operatorname{erf}(v_x^-) \Big) \Big(\operatorname{erf}(v_y^+) - \operatorname{erf}(v_y^-) \Big) \tag{29}$$

with $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2} \mathrm{d}t$ the error function and

$$v_x^- = \frac{t_{j,1} - \Delta_{\mathrm{px}}/2 - x_i}{\sqrt{2}\sigma_1} \qquad v_x^+ = \frac{t_{j,1} + \Delta_{\mathrm{px}}/2 - x_i}{\sqrt{2}\sigma_1} \tag{30}$$

$$v_y^- = \frac{t_{j,2} - \Delta_{\mathrm{px}}/2 - y_i}{\sqrt{2}\sigma_2} \qquad v_y^+ = \frac{t_{j,2} + \Delta_{\mathrm{px}}/2 - y_i}{\sqrt{2}\sigma_2} \tag{31}$$

with (x_i, y_i) the coordinates of the ith pixel and Δ_{px} its width. This derivation from the continuous formulation is straightforward by considering $\Omega_i = [x_i - \Delta_{\mathrm{px}}/2, x_i + \Delta_{\mathrm{px}}/2] \times [y_i - \Delta_{\mathrm{px}}/2, y_i + \Delta_{\mathrm{px}}/2]$.

Finally, we recall that we place ourselves in the noiseless case, i.e. $y = Ax_0$ with y the observation, A the linear measurement operator following the MA-TIRF model and x_0 the original signal to recover.

5.2. Metrics to evaluate the performance of algorithms

To evaluate the performance of each algorithm, we take into account the computation time and the accuracy of the estimation. We introduce the true positive (TP), false positive (FP) and false negative (FN) notations. They label each ground truth (GT) and estimated spike differently according to whether a GT and an estimated spike are close to each other (TP), if a GT spike is alone (FN) and if an estimated spike is alone (FP). These notations are used to compute the Jaccard (Jac) index, the Recall (Rec) and the Precision (Pre) metrics [35]:

$$\mathrm{Jac} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP} + \#\mathrm{FN}} \qquad \mathrm{Rec} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}} \qquad \mathrm{Pre} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP}}. \tag{32}$$

These indexes and metrics allows us to evaluate the capability of an algorithm to give a good estimation. The scores are between 0 and 1, ranging the performance from worst to best. We note that to identify an estimated spike as TP, we check its distance to the closest GT spike noted $d_{\texttt{TP,GT}}$. We distinguish two cases in function of the smallest distance between two spikes of the ground truth noted $d_{\texttt{min}}$. Either the spikes of the ground truth are well separated from each other, i.e. $d_{\texttt{min}} \geq 40$nm, then $d_{\texttt{TP,GT}} = 20$nm. Either $d_{\texttt{min}} < 40$nm, then $d_{\texttt{TP,GT}} = \frac{1}{2} d_{\texttt{min}}$.

We also introduce the root mean squared error (RMSE) between the pairs of GT and estimated spikes (from the TP set) along each dimension,

$$\mathrm{RMSE}_{x_i} = \sqrt{\frac{1}{\#\mathrm{TP}} \sum_{j \in \mathrm{TP}} \left([x_j]_i - [x_{0,j}]_i\right)^2}. \tag{33}$$

This metric gives us an estimation on how close the estimated signal is to the ground truth in each dimension for every pair of GT and TP spikes.

5.3. Benchmarking algorithms and parameters

As previously introduced, there are two main differences between Sliding COMP and OP-COMP + PGD. The first one is the absence of the sliding step within OP-COMP, the second one is the addition of an over-parameterization step in OP-COMP and a descent on all parameters with projections performed in PGD. To test if both differences between the algorithms bring an added value, we introduce in our experiments the modified sliding COMP algorithm: COMP + GD. It is a hybrid algorithm between Sliding COMP and OP-COMP + PGD. For the initialization, it behaves as OP-COMP to the difference that it stops at exactly K iteration for a K-spikes signal. It then performs a Gradient Descent on all parameters but does not perform any projection. This algorithm permits to ensure that the over-parameterization with projection is a key part in improving the recovery of the positions of the spikes.

Regarding the projection step in OP-COMP + PGD, to set the projection distance `eps_dist`, we use an oracle: we compute beforehand the smallest distance between all the spikes in the true signal `min_dist` and set the projection distance `eps_dist` to be $0.75 \times$ `min_dist`. However, this user-defined parameter can be easily estimated from the observation of the dataset. In practice, if it is too large, we risk to project two estimated spikes that should not be projected. In the opposite case, if it is too small, PGD does not merge spikes that should be merged and slows down the computation due to the large number of spikes and the slow convergence of an ill-posed problem (see Section 4).

5.4. Microscope calibration from fluorescent beads acquisitions

A standard practice for calibration in fluorescent microscopy is to image fluorescent microbeads. They behave as point sources and the associated images thus constitute direct measures of the system impulse response. The process of calibration then consists

in a joint estimation of the impulse response and the microbeads position from these data. This is usually addressed with alternating methods and we focus here on the microbeads localization subproblem. Usually, such calibration samples are prepared such that microbeads follow a minimal separation distance, in line with the assumptions of theoretical results of the present work.

In this context, we synthesized fluorescent beads uniformly in the volume. We first set a uniform grid of points in our volume. Then, from each point $x = (x_1, x_2, x_3)$ of the grid, we randomly generated a spike $t = (t_1, t_2, t_3)$ within an ellipse centered at x, i.e.

$$\frac{(x_1 - t_1)^2}{(s_1/2)^2} + \frac{(x_2 - t_2)^2}{(s_2/2)^2} + \frac{(x_3 - t_3)^2}{(s_3/2)^2} \leq 1. \tag{34}$$

with s_1, s_2, s_3 the lengths of each axis of the ellipse. Regarding spikes amplitudes, they are sampled uniformly between 1 and 1.5. We show in Figure 9 a slice of the observation y along two different incident angles. We observe that there is separation between the beads with some responses still intersecting one another (i.e. we are not in the most favorable case in terms of separation).

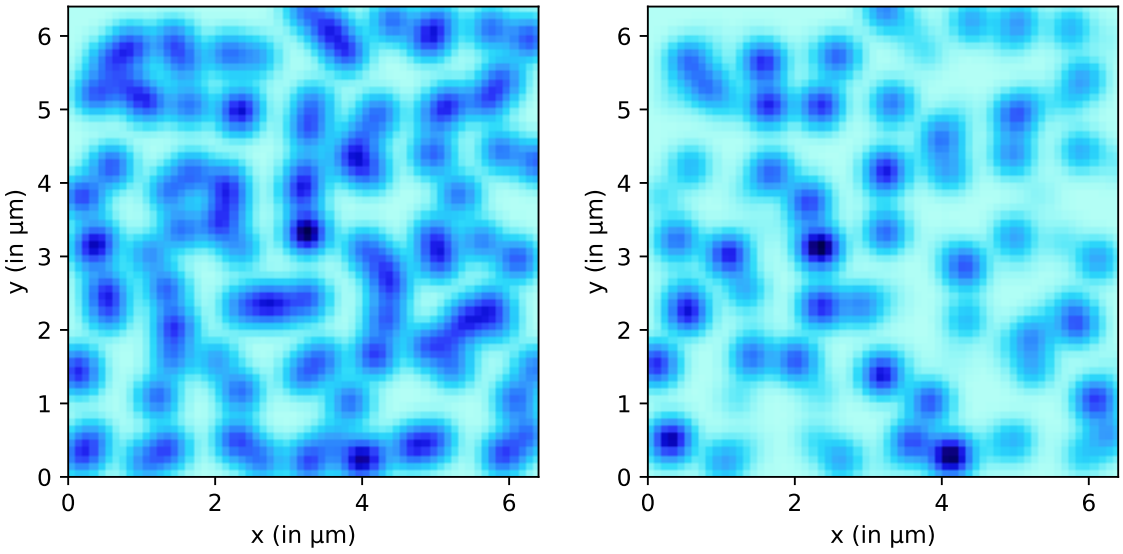

Figure 9: Observation of a signal composed of 98 spikes along two different incident angles through the MA-TIRF model.

We compare OP-COMP + PGD, COMP + GD and Sliding COMP on a signal composed of 98 spikes. The 3D localization is presented in Figure 10. Most of the spikes have been recovered by the three methods

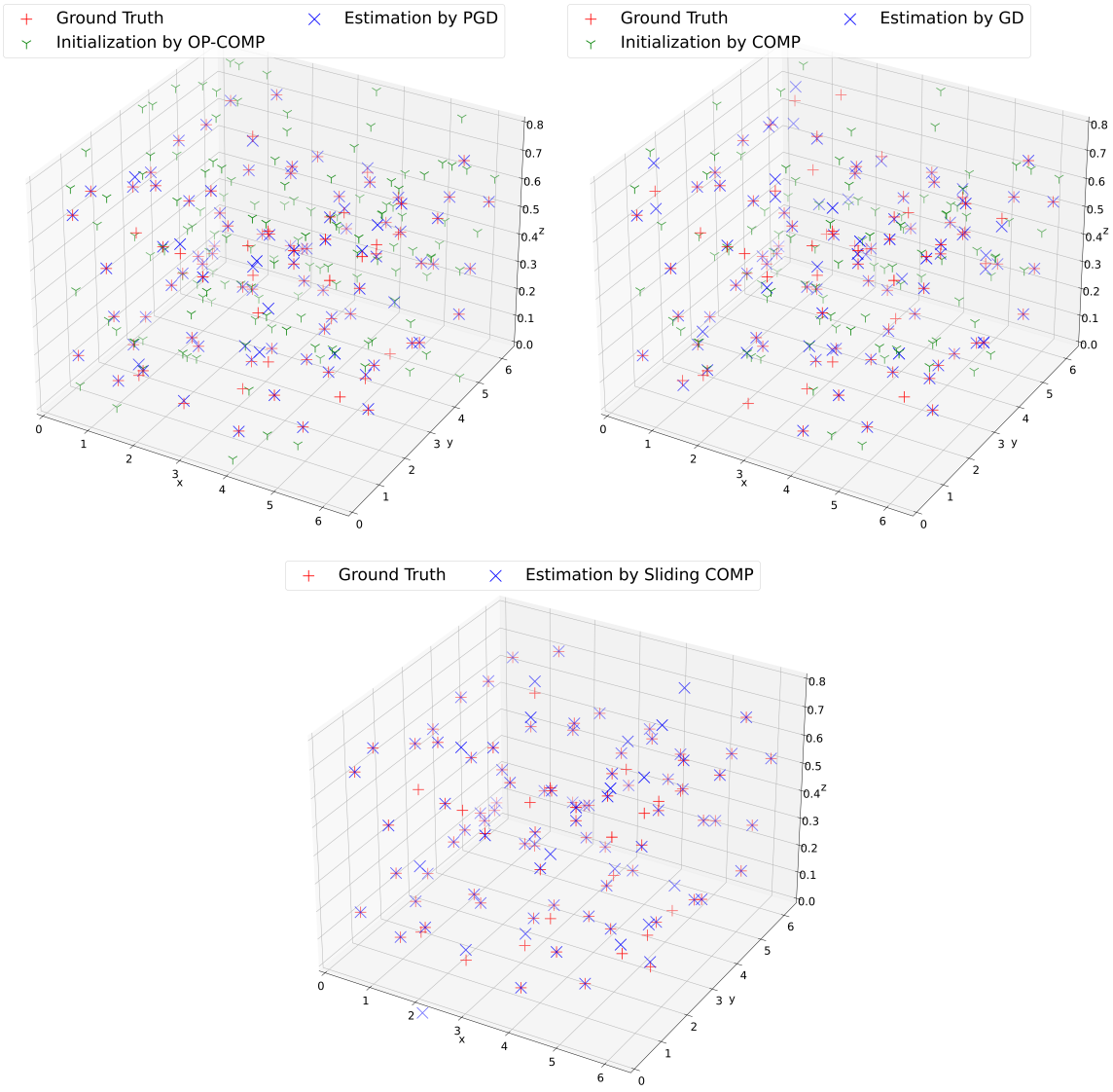

Figure 10: 3D Plots of a signal of size $K = 100$ and its estimation by OP-COMP (in green) and then PGD (in blue) (top left), its estimation by COMP (in green) and then GD (in blue) (top right) and its estimation by Sliding COMP (in blue) (bottom center).

We plot the residues $\|y - Ax\|_2$ at each step of all three algorithms in Figures 11, 12 and 13. At the end of OP-COMP, the initialized signal is composed of 135 spikes. After 10 000 steps and about 35 projections in PGD, the norm of the residue stops decreasing and we obtain a signal composed of 95 spikes. We observe that the norm of the residue decreases far less than with GD without projection. Compared to Sliding COMP, the norm of the residue of OP-COMP+PGD is similar.

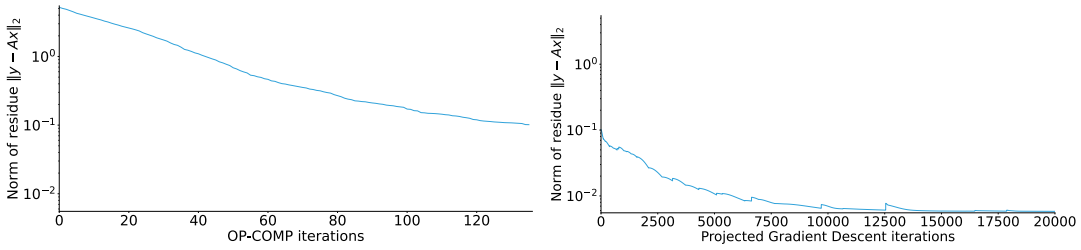

Figure 11: Norms of residue $\|y - Ax\|_2$ at each iteration of OP-COMP (left), Projected Gradient Descent (right) for a 98-spikes signal.

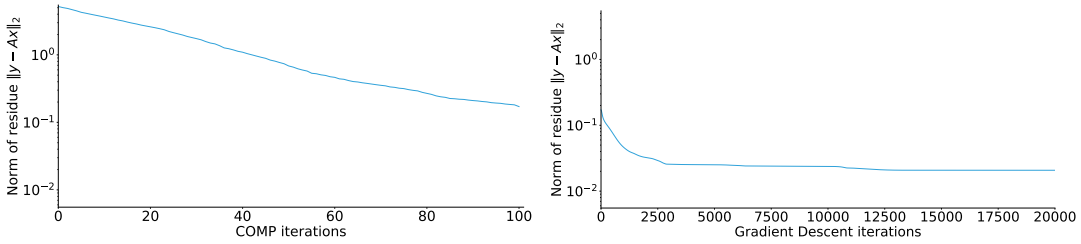

Figure 12: Norms of residue $\|y - Ax\|_2$ at each iteration of COMP (left), Gradient Descent (right) for a 98-spikes signal.

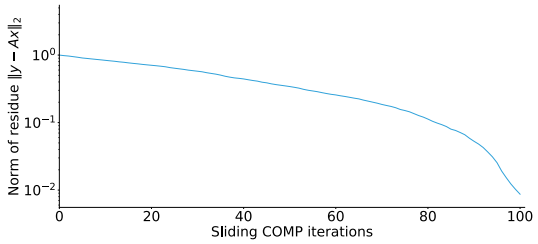

Figure 13: Norms of residue $\|y - Ax\|_2$ at each iteration of Sliding COMP for a 98-spikes signal.

To compare these estimations in terms of localization, we use the metrics we introduced previously (see Section 5.2) and the norm of the residue of each estimation. Results with computation times are reported in Table 1. We conclude that for solving (3), OP-COMP + PGD performs better than its counterpart with no over-parameterization nor projection.

	OP-COMP + PGD	COMP + GD	Sliding COMP
Computation time	129 minutes	115 minutes	272 minutes
Norm of residue of estimation	5.84×10^{-3}	2.08×10^{-2}	8.69×10^{-3}
Jaccard Index	0.771	0.567	0.664
Recall	0.857	0.727	0.806
Precision	0.884	0.720	0.790
RMSE (x_1, x_2, x_3) in nm	$(2.70, 1.68, 2.05)$	$(3.39, 3.56, 4.83)$	$(1.43, 1.58, 1.55)$

Table 1: Table of comparison between OP-COMP + PGD, COMP + GD and Sliding COMP on computation time, norm of residue, Jaccard index, Recall and Precision metrics, and on RMSE along each dimensions.

We observe that in terms of computation time, Sliding COMP is largely outclassed by both OP-COMP + PGD and COMP+GD. For all three Jaccard, Recall and Precision metrics, we note that OP-COMP + PGD performs better than both COMP + GD and Sliding COMP.

We conclude that in the setting where we need to estimate a large number of spikes in the signal with a (relatively) good separation between each pair of spikes, OP-COMP + PGD is a well suited algorithm to recover such signal as it offers a lower computation time and better performances when compared to the state of the art Sliding COMP.

5.5. SMLM localization problem

Single molecule localization microscopy is currently one of the most powerful super-resolution technique in fluorescence microscopy with the potential to reach a resolution of up to ten nanometers [35]. As opposed to conventional fluorescence microscopy where all fluorescent molecules are activated and imaged at the same time, it proceeds by sequentially activating and localizing sparse subsets of these molecules. The density of molecules activated on each SMLM frame comes as a trade-off between the difficulty of the localization problem (easier for low densities) and the temporal resolution in live imaging (better for high densities).

Three-dimensional SMLM can be achieved by using a modified impulse response whose shape varies axially (third dimension). These include for instance the popular astigmatism [28] and double helix [34] responses. To encode the depth of molecules through axial variations, such exotic impulse responses come with a larger lateral support than conventional responses, thus increasing the difficulty of the localization

problem. An alternative is to exploit MA-TIRF excitations to encode the depth of molecules while keeping a conventional, narrower, response. This approach has been numerically investigated and compared to the use of astigmatism and double helix responses in [20].

In this work, we consider the simulated microtubules structures proposed in [35]. From the available ground truth positions of the $K_{\texttt{tot}}$ simulated molecules, we generate SMLM (with MA-TIRF) frames, each containing a subset of $K \ll K_{\texttt{tot}}$ molecules. In the state-of-the art reference [20], K varies between 5 to 15. As K gets larger, the probability of two molecules being close from each other increases and the localization problem becomes harder. In our experiments, we evaluate the performance of OP-COMP + PGD and Sliding COMP over 100 batches where a batch corresponds to an observation of a signal composed of $K \in \{5, 10, 15, 30\}$ spikes (i.e., an SMLM frame). Regarding parameters of model (28), we consider the setting described in [20]. In particular, we use the number of pixels of the grid detector $N_1 = N_2 = 64$, the number of different angles $K_{\texttt{angle}} = 4$ and the excitation wavelength $\lambda_l = 660$nm. The latter is directly related to the decay of the evanescent wave of the TIRF illumination. Moreover, following [20], the variances σ_1 and σ_2 of the Gaussian filter used in model (28) have been set to $\sigma_1 = \sigma_2 = 0.42\lambda_l/\text{NA}$ with NA $= 1.49$ the numerical aperture of the system. This corresponds to a fitting of the Gaussian filter model to the experimentally measured filter of the SMLM challenge [35]. In this way, λ_l is also related to the spread of the impulse response. A smaller wavelength leads to a narrower response and allows for a better separability of the molecules' observations.

In this setting, as opposed to the calibration problem presented in Section 5.4, the separation assumption with respect to the shape of the measurements is not completely verified: some spikes from the ground truth are very close to each other and the linear measurements operator A does not give enough information to distinguish them. With these shortcomings in mind, a direct application of OP-COMP without sliding + PGD does not work as well as in the previous section. In these experiments, we add a small sliding step at each iteration for both OP-COMP and COMP. In comparison to Sliding COMP where several hundred iterations are performed in every sliding step, we only performs 10 to 20 iterations of descent. This number is low enough to not slow down the whole algorithms while still improving the global convergence.

Finally, to perform an adequate comparison, the number of iterations in the final (projected) gradient descent is set to provide similar performance in terms of recovery of positions when possible. When comparing execution times of OP-COMP + PGD, COMP + GD and Sliding COMP in Figure 14 (in logarithmic scale) and their relative differences in Table 2, we observe that for $K \in \{5, 10\}$ (i.e. a small number of spikes), our OP-COMP + PGD algorithm and COMP + GD are slower than Sliding COMP. But as the number of spikes per batch grows, we notice an increasing time gain for OP-COMP + PGD compared to Sliding COMP as predicted by the theoretical computational complexity of the algorithms. We gain 10% of computation time for $K = 15$ up to 75% for $K = 30$. Indeed, as the number of spikes gets larger, the descent step at each

iteration of the Sliding COMP takes longer to compute.

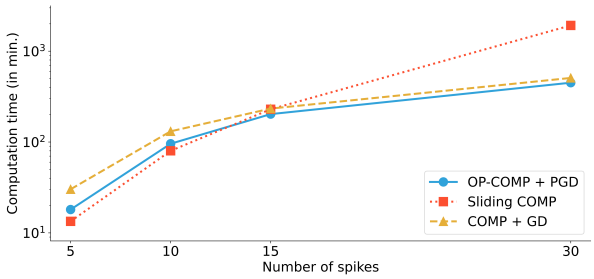

Figure 14: Computation time in *logarithmic scale* for OP-COMP + PGD, COMP + GD and Sliding COMP on a 100 batches of sizes $K = 5,\ 10,\ 15$ and 30.

	Number of spikes K			
Relative difference from Sliding COMP to	5	10	15	30
OP-COMP + PGD (in %)	34.8	19.6	−11.6	−76.6
COMP + GD (in %)	125.5	63.9	1.7	−73.6

Table 2: Relative differences in computation time in percentage between Sliding COMP and both methods OP-COMP + PGD and COMP + GD on a 100 batches of sizes $K = 5,\ 10,\ 15$ and 30.

The recovery results in terms of Jaccard index, Recall and Precision metrics are given in Figure 15.

We observe that the different indices and metrics are very similar for all three methods for $K = 5,\ 10$ and 15. This means that even if our method OP-COMP + PGD and COMP + GD are faster than Sliding COMP, there is no significant degradation of the quality of the estimation.

For the case of $K = 30$, we notice a drop in performance for every index and metrics for all methods. This result is to be expected as the size of our observation y as some spikes become very close. Both OP-COMP + PGD and Sliding COMP have similar metrics in this setting for all three performance measures. However, for COMP + GD, we notice a drop in performances compared to OP-COMP + PGD. This validates the benefit of having an over-parameterized initialization and a projection in the final Gradient Descent.

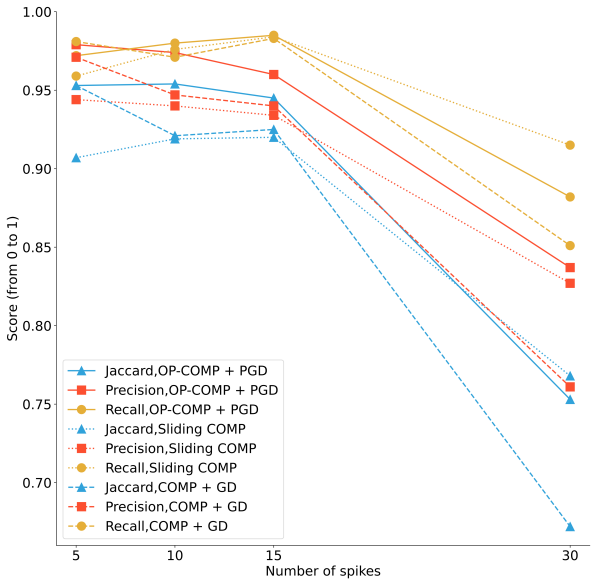

Figure 15: Score of Jaccard index, Recall and Precision metrics for OP-COMP + PGD, COMP + GD, and Sliding COMP on a 100 batches of sizes $K = 5,\ 10,\ 15$ and 30.

For the RMSE metric for each dimension, we see that the distance between GT and TP spikes is small for all three algorithms (Figure 16). We note that the RMSE in the third dimension (the same associated to the different angles of incidence in MA-TIRF) is lower than the RMSE for the first two dimensions.

An expected, we also observe that the RMSE increases with the number of spikes K as the observation y limits our accuracy and the probability of having two ground truth spikes close to each other increases with more spikes per batch.

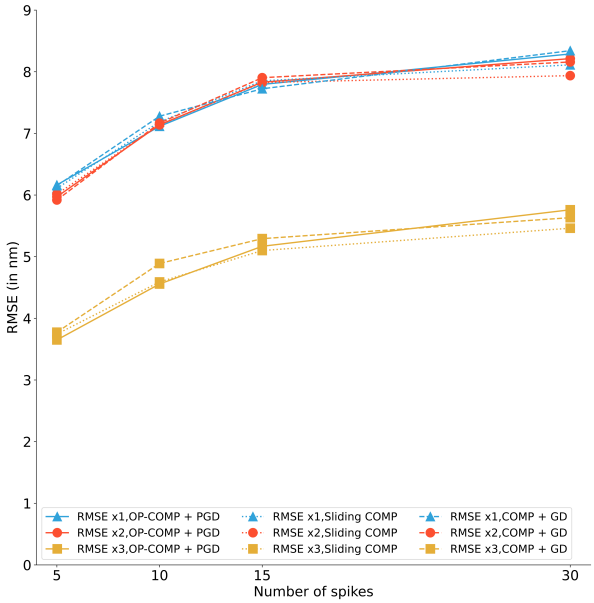

Figure 16: RMSE in each dimension for OP-COMP + PGD, COMP + GD, and Sliding COMP on a 100 batches of sizes $K = 5,\ 10,\ 15$ and 30.

We also combine all batches from our tests with $K = 15$ in a single plot in the same style as in [20] to obtain Figure 17. We note that all three estimations are close to our initial microtubules. This supports our RMSE results presented earlier as for all three algorithms, their RMSE are very similar and close to 0 in each dimension.

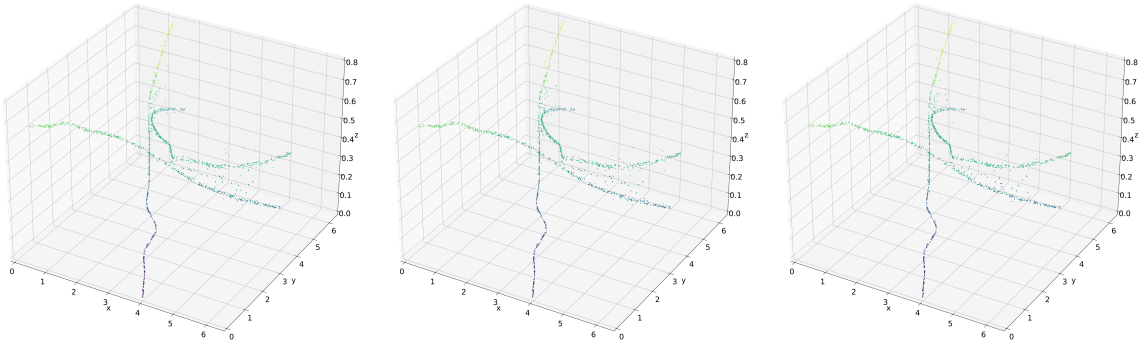

Figure 17: 3D Plot of estimations of a signal of size $K = 15 \times 100$ by OP-COMP + PGD (left) by COMP + GD (center) and by Sliding COMP (right) with a color gradient for the third dimension and a size associated to the amplitude of each spike.

In conclusion to this subsection, even though the signals recovered do not perfectly fit the central separation assumption (very small distances between spikes for a low quality observation), OP-COMP + PGD does perform as well as the state-of-the-art Sliding COMP in classical metrics. As the number of spikes in a signal gets bigger, OP-COMP + PGD gets faster than Sliding COMP. We provide in Section D an SMLM

experiment with 50 spikes with much smaller wavelength (which are unfortunately not available in real set up). This experiment shows that, should the resolution of instruments improve, our algorithm will provide a consequent improvement on computation times.

6. Conclusion

In this article, we showed, for an off-the-grid K sparse signal with separation condition, that with a sufficiently good RIP constant of the linear measurement operator, we can estimate a solution in K iterations up to a controlled accuracy for the non-convex problem.

We also showed that in comparison to Sliding COMP, our method OP-COMP + PGD becomes really efficient when the number of spikes is large. In bad and good settings (depending on the separation between spikes of a signal), OP-COMP + PGD has at least comparable precision to Sliding COMP. We also showed that the over-parametrization step and the projection in the descent bring a non-negligible added value to our method as removing these steps decreases the precision.

For future works, we can investigate the same convergence guarantees for noisy observations. A possible extension of our theorem could be to extend the setting to kernels beyond the Gaussian case. Moreover, is it possible to get less strict bound for the γ RIP as in the discrete case? Also, we can focus on determining conditions on estimating a projection distance instead on relying on an oracle projection distance.

Acknowledgments

This work was supported by the French National Research Agency (ANR) under reference ANR-20-CE40-0001 (EFFIREG project). Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr).

A. Appendix: Proofs of technical results

In this appendix, we detail the proofs of the technical results we need to prove our main theorem.

We start with the proof of Lemma 3.1,

Proof of Lemma 3.1. We have that $h(t,s) = \mathrm{e}^{-\|t-s\|_2^2/(2\sigma^2)}$,

(i) We can write $h(t,s) = \rho(\|t-s\|_2)$ where $\rho(x) = \mathrm{e}^{-x^2/(2\sigma^2)}$ for $t, s \in \mathbb{R}^d$, i.e. h is symmetrical with respect to 0 and translation invariant.

(ii) We have that $h(t,t) = \rho(0) = 1$. However, $-x^2/(2\sigma^2) \leq 0$ for $x \in \mathbb{R}$ with equality when $x = 0$, i.e. $h(t,t) = \max_{t\in\mathbb{R}^d, s\in\mathbb{R}^d} |h(t,s)|$

(iii) For any two spikes $\delta_{t_0}, \delta_{s_0}$ such that $\|t_0 - s_0\|_2 = \epsilon/3$, from the definition of the scalar product with respect to h, we have $\langle \delta_{t_0}, \delta_{s_0} \rangle_h = \mathrm{e}^{-\|t_0 - s_0\|_2^2/(2\sigma^2)} = \mathrm{e}^{-(\epsilon/3)^2/(2\sigma^2)}$. Moreover, from Assumption 1 and (ii), $\langle \delta_{t_0}, \delta_{s_0} \rangle_h \leq \mu$. By combining both equations, we have

$$\langle \delta_{t_0}, \delta_{s_0} \rangle_h = \mathrm{e}^{-\frac{(\epsilon/3)^2}{2\sigma^2}} \leq \mu \|\delta_t\|_h \|\delta_s\|_h = \mu. \tag{35}$$

Taking t, s such that $\|t - s\|_2 \leq \epsilon/3 = \|t_0 - s_0\|_2$, we have

$$\langle \delta_t, \delta_s \rangle_h \leq \mathrm{e}^{-\frac{(\epsilon/3)^2}{2\sigma^2}} \leq \mu. \tag{36}$$

□

Lemma A.1 (Bound on the norm of a dipole)**.** *Let $\pi = a\delta_t - b\delta_s$ be a dipole and h a kernel. Then, we have,*

$$\|\pi\|_h \leq |a - b| + |a| \|\delta_s - \delta_t\|_h. \tag{37}$$

Proof. Remark that $a\delta_t - b\delta_s = a(\delta_t - \delta_s) + (a - b)\delta_s$. With the triangle inequality,

$$\|\pi\|_h \leq \|(a - b)\delta_s\|_h + \|a(\delta_t - \delta_s)\|_h = |a - b| + |a| \|\delta_t - \delta_s\|_h. \tag{38}$$

□

Lemma A.2 (Norm of a dipole for a Gaussian kernel)**.** *Let $\pi = a\delta_t - b\delta_s$ be a ξ_0-dipole such that $0 \leq \xi_0 \leq \xi$ and h be a Gaussian kernel of variance $2\sigma^2$. Then, we have the following properties:*

$$(i) \qquad \|\pi\|_h^2 = (a - b)^2 + 2ab(1 - \mathrm{e}^{-\frac{\xi_0^2}{2\sigma^2}}), \tag{39}$$

$$(ii) \qquad \|\pi\|_h^2 \leq (a - b)^2 + 2|ab|(1 - \mathrm{e}^{-\frac{\xi^2}{2\sigma^2}}). \tag{40}$$

Proof. From $\|\pi\|_h^2$, we develop to get (39),

$$\|\pi\|_h^2 = \|a\delta_t - b\delta_s\|_h^2 \tag{41}$$

$$= a^2 \|\delta_t\|_h^2 + b^2 \|\delta_s\|_h^2 - 2ab\langle \delta_t, \delta_s \rangle_h \tag{42}$$

$$= a^2 + b^2 - 2ab\mathrm{e}^{-\frac{\|t-s\|_2^2}{2\sigma^2}} \tag{43}$$

$$= a^2 + b^2 - 2ab + 2ab - 2ab\mathrm{e}^{-\frac{\|t-s\|_2^2}{2\sigma^2}} \tag{44}$$

$$\|\pi\|_h^2 = (a - b)^2 + 2ab(1 - \mathrm{e}^{-\frac{\xi_0^2}{2\sigma^2}}). \tag{45}$$

Then, since $\xi_0 \leq \xi$, $\mathrm{e}^{-\xi_0^2/(2\sigma^2)} \geq \mathrm{e}^{-\xi^2/(2\sigma^2)}$ and

$$2ab(1 - \mathrm{e}^{-\frac{\xi_0^2}{2\sigma^2}}) \leq 2|a||b|(1 - \mathrm{e}^{-\frac{\xi^2}{2\sigma^2}}). \tag{46}$$

We inject (45) into (46) to get (40). □

Lemma A.3 (Bound scalar product of two spikes)**.** *Let δ_s, δ_t be two Dirac measures with $s, t \in \mathbb{R}^d$. Let A be a linear operator with the γ-RIP on $\mathcal{S}(\Sigma_{K,\epsilon})$. Then,*

$$(1+\gamma)\langle \delta_s, \delta_t \rangle_h - 2\gamma \leq \langle A\delta_s, A\delta_t \rangle \leq (1-\gamma)\langle \delta_s, \delta_t \rangle_h + 2\gamma, \tag{47}$$

and

$$|\langle A\delta_s, A\delta_t \rangle| \leq (1-\gamma)\langle \delta_s, \delta_t \rangle_h + 2\gamma. \tag{48}$$

Proof. Let us start by developing $\langle A\delta_s, A\delta_t \rangle$,

$$\langle A\delta_s, A\delta_t \rangle = \frac{1}{2}\left(\|A\delta_s\|_2^2 + \|A\delta_t\|_2^2 - \|A(\delta_s - \delta_t)\|_2^2 \right) \tag{49}$$

The γ-RIP and Lemma 3.1 imply

$$\begin{aligned}
\langle A\delta_s, A\delta_t \rangle &\geq \frac{1}{2}\left((1-\gamma)\|\delta_s\|_h^2 + (1-\gamma)\|\delta_t\|_h^2 - (1+\gamma)\|(\delta_s - \delta_t)\|_h^2 \right) && (50)\\
&= (1-\gamma) - \frac{1}{2}(1+\gamma)\|(\delta_s - \delta_t)\|_h^2 && (51)\\
&= 1 - \gamma - \frac{1}{2}(1+\gamma)\left(\|\delta_s\|_h^2 + \|\delta_t\|_h^2 - 2\langle \delta_s, \delta_t \rangle_h \right) && (52)\\
&= 1 - \gamma - (1+\gamma)(1 - \langle \delta_s, \delta_t \rangle_h) && (53)\\
&= (1+\gamma)\langle \delta_s, \delta_t \rangle_h - 2\gamma. && (54)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\langle A\delta_s, A\delta_t \rangle &\leq \frac{1}{2}\left((1+\gamma)\|\delta_s\|_h^2 + (1+\gamma)\|\delta_t\|_h^2 - (1-\gamma)\|(\delta_s - \delta_t)\|_h^2 \right) && (55)\\
&= (1+\gamma) - \frac{1}{2}(1-\gamma)\|(\delta_s - \delta_t)\|_h^2 && (56)\\
&= 1 + \gamma - \frac{1}{2}(1-\gamma)\left(\|\delta_s\|_h^2 + \|\delta_t\|_h^2 - 2\langle \delta_s, \delta_t \rangle_h \right) && (57)\\
&= 1 + \gamma - (1-\gamma)(1 - \langle \delta_s, \delta_t \rangle_h) && (58)\\
&= (1-\gamma)\langle \delta_s, \delta_t \rangle_h + 2\gamma. && (59)
\end{aligned}$$

We have just shown the first result. Now, to bound $|\langle A\delta_s, A\delta_t \rangle|$, we compare the upper and lower bounds. If $(1+\gamma)\langle \delta_s, \delta_t \rangle_h - 2\gamma \geq 0$, then

$$|(1-\gamma)\langle \delta_s, \delta_t \rangle_h + 2\gamma| \geq |(1+\gamma)\langle \delta_s, \delta_t \rangle_h - 2\gamma|. \tag{60}$$

Otherwise consider the case $(1+\gamma)\langle \delta_s, \delta_t \rangle_h - 2\gamma \leq 0$. Then as $\gamma < 1$, we have $(1-\gamma)\langle \delta_s, \delta_t \rangle_h + 2\gamma > 0$. By taking the difference between the two absolute values, we obtain

$$\begin{aligned}
&|(1-\gamma)\langle \delta_s, \delta_t \rangle_h + 2\gamma| - |(1+\gamma)\langle \delta_s, \delta_t \rangle_h - 2\gamma|\\
&\qquad = (1-\gamma)\langle \delta_s, \delta_t \rangle_h + 2\gamma - (-1) \times \left((1+\gamma)\langle \delta_s, \delta_t \rangle_h - 2\gamma \right) && (61)\\
&\qquad = (1-\gamma)\langle \delta_s, \delta_t \rangle_h + 2\gamma + (1+\gamma)\langle \delta_s, \delta_t \rangle_h - 2\gamma && (62)\\
&\qquad = 2\langle \delta_s, \delta_t \rangle_h \geq 0. && (63)
\end{aligned}$$

We conclude

$$|\langle A\delta_s, A\delta_t\rangle| \leq \max\left(|(1-\gamma)\langle\delta_s,\delta_t\rangle_h + 2\gamma|, |(1+\gamma)\langle\delta_s,\delta_t\rangle_h - 2\gamma|\right) \tag{64}$$
$$= (1-\gamma)\langle\delta_s,\delta_t\rangle_h + 2\gamma. \tag{65}$$

□

Lemma A.4 (Bound on the scalar products of two dipoles)**.** *Let π_1, π_2 be two ϵ-separated dipoles. Let A be a linear operator with the γ-RIP on $\mathcal{S}(\Sigma_{K,\epsilon})$. Then we have,*

$$(1+\gamma)\langle\pi_1,\pi_2\rangle_h - \gamma(\|\pi_1\|_h^2 + \|\pi_2\|_h^2) \leq \langle A\pi_1, A\pi_2\rangle \leq (1-\gamma)\langle\pi_1,\pi_2\rangle_h + \gamma(\|\pi_1\|_h^2 + \|\pi_2\|_h^2) \tag{66}$$

and

$$|\langle A\pi_1, A\pi_2\rangle| \leq |\langle\pi_1,\pi_2\rangle_h| + \gamma\left|\|\pi_1\|_h^2 + \|\pi_2\|_h^2 - \langle\pi_1,\pi_2\rangle_h\right|. \tag{67}$$

Proof. With the γ-RIP, we have,

$$\langle A\pi_1, A\pi_2\rangle = \frac{1}{2}\left(\|A\pi_1\|_2^2 + \|A\pi_2\|_2^2 - \|A(\pi_1-\pi_2)\|_2^2\right) \tag{68}$$
$$\geq \frac{1}{2}\Big((1-\gamma)(\|\pi_1\|_h^2 + \|\pi_2\|_h^2) - (1+\gamma)\|\pi_1-\pi_2\|_h^2\Big). \tag{69}$$

We develop the norm of the difference between two dipoles,

$$\|\pi_1-\pi_2\|_h^2 = \|\pi_1\|_h^2 + \|\pi_2\|_h^2 - 2\langle\pi_1,\pi_2\rangle_h. \tag{70}$$

By injecting in our previous inequality (68), we get,

$$\langle A\pi_1, A\pi_2\rangle \geq \frac{1}{2}\Big((1-\gamma)\left(\|\pi_1\|_h^2 + \|\pi_2\|_h^2\right) - (1+\gamma)\left(\|\pi_1\|_h^2 + \|\pi_2\|_h^2 - 2\langle\pi_1,\pi_2\rangle_h\right)\Big) \tag{71}$$
$$= (1+\gamma)\langle\pi_1,\pi_2\rangle_h - \gamma(\|\pi_1\|_h^2 + \|\pi_2\|_h^2). \tag{72}$$

Similarly,

$$\langle A\pi_1, A\pi_2\rangle \leq \frac{1}{2}\left((1+\gamma)(\|\pi_1\|_h^2 + \|\pi_2\|_h^2) - (1-\gamma)\|\pi_1-\pi_2\|_h^2\right). \tag{73}$$

We inject in this result (70) and obtain

$$\langle A\pi_1, A\pi_2\rangle \leq \frac{1}{2}\Big((1+\gamma)(\|\pi_1\|_h^2 + \|\pi_2\|_h^2) - (1-\gamma)(\|\pi_1\|_h^2 + \|\pi_2\|_h^2 - 2\langle\pi_1,\pi_2\rangle_h)\Big) \tag{74}$$
$$= (1-\gamma)\langle\pi_1,\pi_2\rangle_h + \gamma(\|\pi_1\|_h^2 + \|\pi_2\|_h^2). \tag{75}$$

The previous bounds (72) and (75), combined and rewritten, give

$$\gamma\left(\langle\pi_1,\pi_2\rangle_h - \|\pi_1\|_h^2 - \|\pi_2\|_h^2\right) \leq \langle A\pi_1, A\pi_2\rangle - \langle\pi_1,\pi_2\rangle_h \leq \gamma\left(\|\pi_1\|_h^2 + \|\pi_2\|_h^2 - \langle\pi_1,\pi_2\rangle_h\right). \tag{76}$$

The bounds are symmetrical, giving us,

$$|\langle A\pi_1, A\pi_2\rangle - \langle\pi_1,\pi_2\rangle_h| \leq \gamma\left|\|\pi_1\|_h^2 + \|\pi_2\|_h^2 - \langle\pi_1,\pi_2\rangle_h\right|. \tag{77}$$

Using the triangle inequality, we get

$$|\langle A\pi_1, A\pi_2\rangle| - |\langle\pi_1,\pi_2\rangle_h| \leq |\langle A\pi_1, A\pi_2\rangle - \langle\pi_1,\pi_2\rangle_h|. \tag{78}$$

This implies

$$|\langle A\pi_1, A\pi_2\rangle| \leq |\langle \pi_1, \pi_2\rangle_h| + \gamma \left| \|\pi_1\|_h^2 + \|\pi_2\|_h^2 - \langle \pi_1, \pi_2\rangle_h \right|. \tag{79}$$

□

B. Appendix: Proofs for Theorem 3.3

In this appendix, we detail the proof of our main theorem. We start by technical Lemmas.

Lemma B.1. *Let δ_s a Dirac measure and $z_l = \sum_{i=1}^l \pi_i \in Z_{l,\xi,\frac{\epsilon}{3}}$ such that $\forall i = 1, \dots, l$, δ_s and π_i are $\frac{\epsilon}{3}$-separated and $\xi < \epsilon/3$. Assume that the linear operator A has the γ-RIP on $\mathcal{S}(\Sigma_{K,\frac{\epsilon}{3}})$. Denote $D = \max_{i\in\{1,\dots,l\}} \|\pi_i\|_h / \|a\|_\infty$ where $a \in \mathbb{R}^K$. Then,*

$$\frac{\|Az_l\|_2^2}{\|a\|_\infty^2} \leq l(1+\gamma)(1+(l-1)\mu)D^2. \tag{80}$$

Furthermore, we have,

$$\frac{|\langle A\delta_s, Az_l\rangle|}{\|a\|_\infty} \leq \kappa_l := (1+\gamma)\mu l D + \gamma l \left(1 + D^2\right). \tag{81}$$

Proof. With the γ-RIP,

$$\|Az_l\|_2^2 = \|A\sum_{i=1}^l \pi_i\|_2^2 \leq (1+\gamma)\|\sum_{i=1}^l \pi_i\|_h^2. \tag{82}$$

From Lemma 3.2, we get

$$\frac{\|Az_l\|_2^2}{\|a\|_\infty^2} \leq (1+\gamma)(1+(l-1)\mu)\sum_{i=1}^l \frac{\|\pi_i\|_h^2}{\|a\|_\infty^2} \leq l(1+\gamma)(1+(l-1)\mu)D^2. \tag{83}$$

This concludes the first part of the proof.

We now bound $|\langle A\delta_s, Az_l\rangle|$. With the triangle inequality,

$$\frac{|\langle A\delta_s, Az_l\rangle|}{\|a\|_\infty} \leq \sum_{i=1}^l |\langle A\delta_s, A\frac{\pi_i}{\|a\|_\infty}\rangle|. \tag{84}$$

Then, using the result (67) from Lemma A.4,

$$\frac{|\langle A\delta_s, Az_l\rangle|}{\|a\|_\infty} \leq \sum_{i=1}^l \left(\left| \langle \delta_s, \frac{\pi_i}{\|a\|_\infty}\rangle_h \right| + \gamma \left| \|\delta_s\|_h^2 + \frac{\|\pi_i\|_h^2}{\|a\|_\infty^2} - \langle \delta_s, \frac{\pi_i}{\|a\|_\infty}\rangle_h \right| \right). \tag{85}$$

As δ_s and $\pi_i, i = 1, \dots, l$ are $\epsilon/3$-separated, we have $\langle \delta_s, \pi_i\rangle_h \leq \mu\|\delta_s\|_h\|\pi_i\|_h = \mu\|\pi_i\|_h$. We inject this into our inequality (85) to obtain

$$\frac{|\langle A\delta_s, Az_l\rangle|}{\|a\|_\infty} \leq \sum_{i=1}^l \left(\mu\frac{\|\pi_i\|_h}{\|a\|_\infty} + \gamma\left(1 + \frac{\|\pi_i\|_h^2}{\|a\|_\infty^2} + \mu\frac{\|\pi_i\|_h}{\|a\|_\infty}\right)\right). \tag{86}$$

After rearranging the terms in (86), we get

$$\frac{|\langle A\delta_s, Az_l\rangle|}{\|a\|_\infty} \leq \sum_{i=1}^{l}\left((1+\gamma)\mu\frac{\|\pi_i\|_h}{\|a\|_\infty} + \gamma\left(1+\frac{\|\pi_i\|_h^2}{\|a\|_\infty^2}\right)\right) \tag{87}$$

$$\leq (1+\gamma)\mu l D + \gamma l\left(1+D^2\right). \tag{88}$$

□

The following lemma gives a condition on ξ to bound the scalar product of a ξ-separated dipole.

Lemma B.2 (Condition on ξ)**.** *Let $k+l=K$ with $k,l,K\in\mathbb{N}$ and $K>0$. Let $\alpha\geq 1$ and*

$$\xi \geq \sqrt{2\sigma^2\ln\left(\frac{1}{1-(4K\alpha-1)(\mu+\gamma)}\right)}, \tag{89}$$

with $\mu,\gamma\in\mathbb{R}_*^+$ such that $\mu+\gamma<\frac{1}{4K\alpha-1}$. Consider κ_l and D defined in Lemma B.1 and suppose $D<1$. Then

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < \frac{1-((2k-1)\alpha-1)(1-\gamma)\mu-(4k\alpha-1)\gamma-2\kappa_l\alpha}{1-\gamma}. \tag{90}$$

Proof. Using the hypothesis that $D<1$, we obtain $\kappa_l<(1+\gamma)\mu l+2\gamma l$. Hence (90) is verified if

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < \frac{1-((2k-1)\alpha-1)(1-\gamma)\mu-(4k\alpha-1)\gamma-2\alpha(1+\gamma)\mu l-4\alpha\gamma l}{1-\gamma} \tag{91}$$

$$= \frac{1-[((2k-1)\alpha-1)(1-\gamma)+2\alpha(1+\gamma)l]\mu-[(4k+4l)\alpha-1]\gamma}{1-\gamma}. \tag{92}$$

Since we suppose that $0<\gamma<1$, this inequality is verified if

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < 1-[((2k-1)\alpha-1)(1-\gamma)+2\alpha(1+\gamma)l]\mu-[(4k+4l)\alpha-1]\gamma. \tag{93}$$

Using our hypothesis on γ, we can bound some values of our previous inequality,

$$1-\gamma<1 \qquad \text{and} \qquad 2\alpha(1+\gamma)<4\alpha. \tag{94}$$

Thus, (93) is verified if the following inequality is true,

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < 1-[(2k-1)\alpha-1+4\alpha l]\mu-[(4k+4l)\alpha-1]\gamma. \tag{95}$$

Finally the previous inequality is verified if the following inequalities are true,

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < 1-[4k\alpha+4l\alpha-1]\mu-[4(k+l)\alpha-1]\gamma \tag{96}$$

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < 1-[4(k+l)\alpha-1](\mu+\gamma). \tag{97}$$

Using our hypothesis that $k+l=K$, we obtain the condition

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} \leq 1-(4K\alpha-1)(\mu+\gamma); \tag{98}$$

which is exactly (89) (by taking the natural log). □

The following lemma gives a bound on μ by rewriting some conditions on $\mu+\gamma$ and κ_l.

Lemma B.3 (Condition on γ and μ)**.** *Let $k+l=K$ with $k,l,K\in\mathbb{N}$, $K>0$. Let $a\in\mathbb{R}^k$ and $\alpha\geq 1$. Let $\kappa_l := (1+\gamma)\mu l D+\gamma l(1+D^2)$ and suppose $D<1$. Suppose that*

$$\gamma+\mu<\frac{1}{4K\alpha-1}\qquad \text{with}\qquad \gamma,\mu\in\mathbb{R}^+. \tag{99}$$

Then

$$\mu<\frac{1-(4k\alpha-1)\gamma-2\kappa_l\alpha}{(2k\alpha-1)(1-\gamma)}. \tag{100}$$

Proof. Since $D<1$, we have

$$\kappa_l=(1+\gamma)\mu l D+\gamma l(1+D^2)<(1+\gamma)\mu l+2\gamma l \tag{101}$$

We deduce that our conclusion (100) is verified if

$$\mu<\frac{1-(4k\alpha-1)\gamma-2\alpha((1+\gamma)\mu l+2\gamma l)}{(2k\alpha-1)(1-\gamma)}. \tag{102}$$

Because $0<\gamma<1$, we have that (102) is guaranteed if

$$(2k\alpha-1)\mu<1-(4k\alpha-1)\gamma-2\alpha((1+\gamma)\mu l+2\gamma l); \tag{103}$$

which, since $1+\gamma<2$, is in turn guaranteed if

$$(2k\alpha-1)\mu<1-(4k\alpha-1)\gamma-4\alpha l(\mu+\gamma). \tag{104}$$

By rearranging the terms in this inequality, we obtain,

$$((2k+4l)\alpha-1)\mu+(4(k+l)\alpha-1)\gamma<1. \tag{105}$$

Since $0\leq k,l\leq K$ and $k+l=K$, the above inequality is guaranteed if

$$(4K\alpha-1)\mu+(4K\alpha-1)\gamma<1, \tag{106}$$

which is equivalent to the hypothesis $\mu+\gamma<1/(4K\alpha-1)$. □

The following lemma gives a numerical bound on the norm of the difference between the amplitudes of the true signal and their estimation by a least-squares estimation.

Lemma B.4 (Bound for estimated amplitudes)**.** *Let $y=\sum_{i=1}^k a_i\delta_{t_i}\in\Sigma_{K,\epsilon}$, and $s_1,\dots,s_l\in\mathbb{R}^d$, with $l,k\in\mathbb{N}$ such that $k+l=K\geq 2$. Suppose $\|s_i-t_i\|_2\leq\xi<\epsilon/3$ and $\max_{i\neq j}|\langle\delta_{s_i},\delta_{s_j}\rangle|\leq\mu$ for all $1\leq i,j\leq l$. Suppose $\mu+\gamma\leq\frac{1}{5K(4K\alpha-1)}$ with $\alpha\geq 1$. Assume that the linear operator A has the γ-RIP on $\mathcal{S}(\Sigma_{K,\frac{\epsilon}{3}})$. Let*

$$\xi:=\sqrt{2\sigma^2\ln\left(\frac{1}{(1-(4K\alpha-1)(\mu+\gamma))}\right)}. \tag{107}$$

and let

$$b \in \arg\min_{\tilde{b}\in\mathbb{R}^l} \|A(\sum_{i=1}^{l} \tilde{b}_i\delta_{s_i} - a_i\delta_{t_i}) - A\sum_{i=l+1}^{K} a_i\delta_{t_i}\|_2^2. \tag{108}$$

Then

$$\frac{\|b - a_{1:l}\|_2}{\|a\|_\infty} \leq 0.77. \tag{109}$$

Proof. For this proof, we note the upper-bound of $\mu + \gamma \leq \frac{1}{\beta K(4K-1)}$ with $\beta = 5$ for simpler notations.

Let $\pi = b_i\delta_{s_i} - a_i\delta_{t_i}$. With Lemma A.1, we have $\|\pi_i\|_h = \|b_i\delta_{s_i} - a_i\delta_{t_i}\|_h = \|(b_i-a_i)\delta_{s_i} - a_i(\delta_{t_i}-\delta_{s_i})\|_h \leq \|(b_i-a_i)\delta_{s_i}\|_h + |a_i|\|(\delta_{t_i}-\delta_{s_i})\|_h = |b_i-a_i| + |a_i|\|\delta_{t_i}-\delta_{s_i}\|_h$.

Consider the matrix $B = (A\delta_{s_1}, \dots, A\delta_{s_l})$ with B^H its conjugate transpose, we have $b = (B^HB)^{-1}B^Hy$ is the solution of the least squares estimation (108) with $y = A\sum_{i=1}^K a_i\delta_{t_i}$.

The norm $\|(B^HB)^{-1}\|_2^2$ is linked to the smallest eigenvalue of B^HB that we note $\lambda_{min}(B^HB)$. Indeed, we have

$$\|(B^HB)^{-1}\|_{op}^2 = \frac{1}{|\lambda_{min}(B^HB)|^2}. \tag{110}$$

To find a bound on $\lambda_{min}(B^HB)$, we use the Gershgorin circle theorem [6]. There exists $1 \leq i \leq l$, such that,

$$\left|\lambda_{min}(B^HB) - \|A\delta_{s_i}\|_2^2\right| \leq \sum_{j\neq i} \left|\langle A\delta_{s_i}, A\delta_{s_j}\rangle\right|. \tag{111}$$

With the triangle inequality, we have

$$\left|\lambda_{min}(B^HB) - \|A\delta_{s_i}\|_2^2\right| \geq \|A\delta_{s_i}\|_2^2 - |\lambda_{min}(B^HB)|. \tag{112}$$

Multiplying by -1 and injecting (111), we get

$$\left|\lambda_{min}(B^HB)\right| - \|A\delta_{s_i}\|_2^2 \geq -\sum_{j\neq i} \left|\langle A\delta_{s_i}, A\delta_{s_j}\rangle\right| \tag{113}$$

$$\left|\lambda_{min}(B^HB)\right| \geq \|A\delta_{s_i}\|_2^2 - \sum_{j\neq i} \left|\langle A\delta_{s_i}, A\delta_{s_j}\rangle\right| \tag{114}$$

$$\left|\lambda_{min}(B^HB)\right| \geq \|A\delta_{s_i}\|_2^2 - (l-1)\max_{i\neq j} \left|\langle A\delta_{s_i}, A\delta_{s_j}\rangle\right|. \tag{115}$$

Using Lemma A.3, we have $\left|\langle A\delta_{s_i}, A\delta_{s_j}\rangle\right| \leq (1-\gamma)\langle\delta_{s_i}, \delta_{s_j}\rangle + 2\gamma$. Thus, using the RIP, we have

$$|\lambda_{min}(B^HB)| \geq (1-\gamma)\|\delta_{s_i}\|_h^2 - 2(l-1)\gamma - (l-1)(1-\gamma)\max_{i\neq j} |\langle\delta_{s_i}, \delta_{s_j}\rangle| \tag{116}$$

$$= 1 - (2l-1)\gamma - (l-1)(1-\gamma)\max_{i\neq j} |\langle\delta_{s_i}, \delta_{s_j}\rangle|. \tag{117}$$

Using the mutual coherence (see (14)) and the hypothesis on $\gamma + \mu$, we obtain

$$|\lambda_{min}(B^H B)| \geq 1 - (2l-1)\gamma - (l-1)(1-\gamma)\mu \tag{118}$$
$$\geq 1 - (2l-1)\gamma - (l-1)\mu \tag{119}$$
$$\geq 1 - 2K\gamma - K\mu \geq 1 - 2K(\gamma+\mu) \tag{120}$$
$$\geq 1 - \frac{2K}{\beta K(4K-1)} \geq 1 - \frac{2}{\beta(4K-1)}. \tag{121}$$

As $K \geq 2$, we conclude with

$$|\lambda_{min}(B^H B)| \geq 1 - \frac{2}{7\beta}. \tag{122}$$

Now that we have a bound on the lowest eigenvalue of $B^H B$, we calculate and bound $\|b - a_{1:l}\|_2^2$. We have

$$b - a_{1:l} = (B^H B)^{-1} B^H y - a_{1:l} = (B^H B)^{-1}(B^H y - B^H B a_{1:l}). \tag{123}$$

On the one hand, we have

$$B^H y = \begin{bmatrix} \overline{A\delta_{s_1}} \\ \hline \vdots \\ \hline \overline{A\delta_{s_l}} \end{bmatrix} A \sum_{j=1}^{K} a_j \delta_{t_j} \quad \text{i.e.} \quad \left[B^H y\right]_i = \sum_{j=1}^{K} a_j \langle A\delta_{s_i}, A\delta_{t_j}\rangle. \tag{124}$$

On the other hand,

$$B^H B a_{1:l} = \begin{bmatrix} \langle A\delta_{s_1}, A\delta_{s_1}\rangle & \cdots & \langle A\delta_{s_1}, A\delta_{s_l}\rangle \\ \vdots & \ddots & \vdots \\ \langle A\delta_{s_l}, A\delta_{s_1}\rangle & \cdots & \langle A\delta_{s_l}, A\delta_{s_l}\rangle \end{bmatrix} \begin{bmatrix} a_1 \\ \hline \vdots \\ \hline a_l \end{bmatrix}. \tag{125}$$

Hence,

$$\left[B^H B a_{1:l}\right]_i = \sum_{j=1}^{l} a_j \langle A\delta_{s_i}, A\delta_{s_j}\rangle. \tag{126}$$

Using (123) and combining (124) and (126), we obtain,

$$\|b - a_{1:l}\|_2^2 = \|(B^H B)^{-1}(B^H y - B^H B a_{1:l})\|_2^2 \tag{127}$$
$$\leq \sum_{i=1}^{l} \left| \sum_{j=1}^{K} a_j \langle A\delta_{s_i}, A\delta_{t_j}\rangle - \sum_{j=1}^{l} a_j \langle A\delta_{s_i}, A\delta_{s_j}\rangle \right|^2 \frac{1}{|\lambda_{min}(B^H B)|^2} \tag{128}$$
$$= \sum_{i=1}^{l} \left| \sum_{j=1}^{l} a_j \langle A\delta_{s_i}, A(\delta_{t_j} - \delta_{s_j})\rangle + \sum_{j=l+1}^{K} a_j \langle A\delta_{s_i}, A\delta_{t_j}\rangle \right|^2 \frac{1}{|\lambda_{min}(B^H B)|^2}. \tag{129}$$

With the triangle inequality,

$$\|b - a_{1:l}\|_2^2 \leq \sum_{i=1}^{l} \left(\sum_{j=1}^{l} |a_j| |\langle A\delta_{s_i}, A(\delta_{t_j} - \delta_{s_j})\rangle| + \sum_{j=l+1}^{K} |a_j| |\langle A\delta_{s_i}, A\delta_{t_j}\rangle| \right)^2 \frac{1}{|\lambda_{min}(B^H B)|^2}. \tag{130}$$

To find an upper bound to this equation, let us bound each term separately.
Firstly, for all $1 \leq i \leq l, l+1 \leq j \leq K$, with Lemma A.3, we have,

$$|\langle A\delta_{s_i}, A\delta_{t_j}\rangle| \leq (1-\gamma)\langle \delta_{s_i}, \delta_{t_j}\rangle_h + 2\gamma \leq (1-\gamma)\mu + 2\gamma. \tag{131}$$

Secondly, for all $1 \leq i \leq l, 1 \leq j \leq l$, we separate the cases where $i = j$ and $i \neq j$. We use the Cauchy-Schwarz inequality and Lemma A.4,

$$|\langle A\delta_{s_i}, A(\delta_{t_j} - \delta_{s_j})\rangle| \leq \begin{cases} \|A\delta_{s_i}\|_2 \|A(\delta_{t_j} - \delta_{s_j})\|_2, & i = j \\ \left|\langle \delta_{s_i}, \delta_{t_j} - \delta_{s_j}\rangle_h\right| & \\ \quad +\gamma\left|\|\delta_{s_i}\|_h^2 + \|\delta_{t_j} - \delta_{s_j}\|_h^2 - \langle \delta_{s_i}, \delta_{t_j} - \delta_{s_j}\rangle_h\right|, & i \neq j \end{cases} \tag{132}$$

Using the property that $|\langle \pi_1, \pi_2\rangle| \leq \mu\|\pi_1\|_h\|\pi_2\|_h$ for π_1, π_2 $\frac{\epsilon}{3}$-separated dipoles from Assumption 1, the Cauchy-Schwarz inequality, and $\|A\pi\|_2 \leq \sqrt{1+\gamma}\|\pi\|_h$ with the RIP, we get

$$|\langle A\delta_{s_i}, A(\delta_{t_j} - \delta_{s_j})\rangle| \leq \begin{cases} (1+\gamma)\|\delta_{t_j} - \delta_{s_j}\|_h, & i = j \\ \mu\|\delta_{t_j} - \delta_{s_j}\|_h & \\ \quad +\gamma\left(1 + \|\delta_{t_j} - \delta_{s_j}\|_h^2 + \left|\langle \delta_{s_i}, \delta_{t_j} - \delta_{s_j}\rangle_h\right|\right), & i \neq j \end{cases} \tag{133}$$

$$\leq \begin{cases} (1+\gamma)\|\delta_{t_j} - \delta_{s_j}\|_h, & i = j \\ \mu\|\delta_{t_j} - \delta_{s_j}\|_h & \\ \quad +\gamma\left(1 + \|\delta_{t_j} - \delta_{s_j}\|_h^2 + \mu\|\delta_{t_j} - \delta_{s_j}\|_h\right), & i \neq j \end{cases} \tag{134}$$

$$= \begin{cases} (1+\gamma)\|\delta_{t_j} - \delta_{s_j}\|_h, & i = j \\ \mu(1+\gamma)\|\delta_{t_j} - \delta_{s_j}\|_h + \gamma\left(1 + \|\delta_{t_j} - \delta_{s_j}\|_h^2\right), & i \neq j. \end{cases} \tag{135}$$

We use the fact that $\|s_i - t_i\|_2 \leq \xi = \sqrt{2\sigma^2 \ln\left(\frac{1}{(1-(4K\alpha-1)(\mu+\gamma))}\right)}$ from hypothesis (107),

$$|\langle A\delta_{s_i}, A(\delta_{t_j} - \delta_{s_j})\rangle| \leq \begin{cases} (1+\gamma)\sqrt{2(1 - \mathrm{e}^{-\frac{\xi^2}{2\sigma^2}})}, & i = j \\ \mu(1+\gamma)\sqrt{2(1 - \mathrm{e}^{-\frac{\xi^2}{2\sigma^2}})} + \gamma + 2\gamma(1 - \mathrm{e}^{-\frac{\xi^2}{2\sigma^2}}), & i \neq j \end{cases} \tag{136}$$

$$\leq \begin{cases} (1+\gamma)\sqrt{2(4K\alpha-1)(\mu+\gamma)}, & i = j \\ \mu(1+\gamma)\sqrt{2(4K\alpha-1)(\mu+\gamma)} & \\ \quad +\gamma + 2\gamma(4K\alpha-1)(\mu+\gamma), & i \neq j. \end{cases} \tag{137}$$

With the hypothesis on $\mu + \gamma$, we have,

$$|\langle A\delta_{s_i}, A(\delta_{t_j} - \delta_{s_j})\rangle| \leq \begin{cases} (1+\gamma)\sqrt{\frac{2}{\beta K}}, & i = j \\ \mu(1+\gamma)\sqrt{\frac{2}{\beta K}} + (1 + \frac{2}{\beta K})\gamma, & i \neq j. \end{cases} \tag{138}$$

By injecting the bounds (131) and (138) in (130), we get

$$\|b - a_{1:l}\|_2^2 \leq \frac{1}{|\lambda_{min}(B^H B)|^2} \sum_{i=1}^{l} \Bigg[(1+\gamma)\sqrt{\frac{2}{\beta K}}|a_i| + \left(\mu(1+\gamma)\sqrt{\frac{2}{\beta K}} + (1+\frac{2}{\beta K})\gamma\right) \sum_{\substack{j=1 \\ i\neq j}}^{l} |a_j| + \Big((1-\gamma)\mu + 2\gamma\Big) \sum_{j=l+1}^{K} |a_j| \Bigg]^2. \tag{139}$$

By dividing both sides by $\|a\|_\infty^2$, we obtain the following upper-bound,

$$\frac{\|b - a_{1:l}\|_2^2}{\|a\|_\infty^2} \leq \frac{l}{|\lambda_{min}(B^H B)|^2} \Bigg[(1+\gamma)\sqrt{\frac{2}{\beta K}} + \left(\mu(1+\gamma)\sqrt{\frac{2}{\beta K}} + (1+\frac{2}{\beta K})\gamma\right)(l-1) + \Big((1-\gamma)\mu + 2\gamma\Big)(K-l) \Bigg]^2. \tag{140}$$

Taking the square root on both sides of this inequality, we get

$$\frac{\|b - a_{1:l}\|_2}{\|a\|_\infty} \leq \frac{\sqrt{l}}{|\lambda_{min}(B^H B)|} \Bigg[(1+\gamma)\sqrt{\frac{2}{\beta K}} + \left(\mu(1+\gamma)\sqrt{\frac{2}{\beta K}} + (1+\frac{2}{\beta K})\gamma\right)(l-1) + \Big((1-\gamma)\mu + 2\gamma\Big)(K-l) \Bigg]. \tag{141}$$

We take the max between $\mu(1+\gamma)\sqrt{\frac{2}{\beta K}} + (1+\frac{2}{\beta K})\gamma$ and $(1-\gamma)\mu + 2\gamma$ to obtain

$$\frac{\|b - a_{1:l}\|_2}{\|a\|_\infty} \leq \frac{\sqrt{l}}{|\lambda_{min}(B^H B)|} \Bigg[(1+\gamma)\sqrt{\frac{2}{\beta K}} + \max\left(\mu(1+\gamma)\sqrt{\frac{2}{\beta K}} + (1+\frac{2}{\beta K})\gamma, (1-\gamma)\mu + 2\gamma\right)(K-1) \Bigg]. \tag{142}$$

Remarking that

$$c_1 := \mu(1+\gamma)\sqrt{\frac{2}{\beta K}} + \left(1+\frac{2}{\beta K}\right)\gamma \leq 2\mu\sqrt{\frac{2}{\beta K}} + \left(1+\frac{2}{\beta K}\right)\gamma \leq 2(\mu+\gamma) \tag{143}$$

and

$$c_2 := (1-\gamma)\mu + 2\gamma \leq 2(\mu+\gamma), \tag{144}$$

we have $\max(c_1, c_2) \leq 2(\mu+\gamma)$ and

$$\frac{\|b - a_{1:l}\|_2}{\|a\|_\infty} \leq \frac{1}{|\lambda_{min}(B^H B)|}\sqrt{l}\left((1+\gamma)\sqrt{\frac{2}{\beta K}} + 2(\mu+\gamma)K\right) \tag{145}$$

$$\leq \frac{1}{|\lambda_{min}(B^H B)|}\sqrt{K}\left((1+\gamma)\sqrt{\frac{2}{\beta K}} + \frac{2}{\beta(4K\alpha - 1)}\right) \tag{146}$$

$$\leq \frac{1}{|\lambda_{min}(B^H B)|}\left((1+\gamma)\sqrt{\frac{2}{\beta}} + \frac{2\sqrt{K}}{\beta(4K-1)}\right). \tag{147}$$

For $K \geq 2$, we can bound γ by

$$\gamma \leq \mu + \gamma \leq \frac{1}{\beta K(4K\alpha - 1)} \leq \frac{1}{14\beta}. \tag{148}$$

We get, for $K \geq 2$,

$$\frac{\|b - a_{1:l}\|_2}{\|a\|_\infty} \leq \frac{\left(1 + \frac{1}{14\beta}\right)\sqrt{\frac{2}{\beta}} + \frac{2\sqrt{K}}{\beta(4K-1)}}{|\lambda_{min}(B^H B)|} \leq \frac{\left(1 + \frac{1}{14\beta}\right)\sqrt{\frac{2}{\beta}} + \frac{2\sqrt{2}}{7\beta}}{|\lambda_{min}(B^H B)|}. \tag{149}$$

Using the bound (122) on $\lambda_{min}(B^H B)$, we get

$$\frac{\|b - a_{1:l}\|_2}{\|a\|_\infty} \leq \frac{1}{1 - \frac{2}{7\beta}} \left(\left(1 + \frac{1}{14\beta}\right)\sqrt{\frac{2}{\beta}} + \frac{2\sqrt{2}}{7\beta}\right). \tag{150}$$

By choosing $\beta = 5$, we obtain numerically that

$$\frac{\|b - a_{1:l}\|_2}{\|a\|_\infty} \leq 0.77\,. \tag{151}$$

□

The next proposition states that for any ϵ-separated signal plus a residue, the output given by the maximum of the correlation between this signal and a single spike is in the ξ-radius of a spike of the signal.

Proposition B.1. *Let $k, l, K \in \mathbb{N}$ such that $k+l = K \geq 2$. Let $x_{k,\epsilon} = \sum_{i=1}^{k} a_i \delta_{t_i} \in \Sigma_{K,\epsilon}$ where $\|a\|_\infty = |a_1| \geq |a_2| \geq \ldots \geq |a_K|$. Denote $\alpha = |a_1|/|a_K|$. Assume that the linear operator A has the γ-RIP on $\mathcal{S}(\Sigma_{K,\frac{\epsilon}{3}})$ with a Gaussian kernel h of variance σ^2.*
Let

$$\xi := \sqrt{2\sigma^2 \ln\left(\frac{1}{(1 - (4K\alpha - 1)(\mu + \gamma))}\right)}. \tag{152}$$

Let $X = x_{k,\epsilon} + z_l$ where $z_l = \sum_{i=k+1}^{K} a_i \delta_{t_i} - b_i \delta_{s_i} \in Z_{l,\xi,\frac{\epsilon}{3}}$ and $s^* \in \arg\max_{\tilde{s} \in \mathbb{R}^d} |\langle A\delta_{\tilde{s}}, AX\rangle|$.
Suppose

$$\mu + \gamma < \frac{1}{5K\,(4K\alpha - 1)} \tag{153}$$

and

$$\epsilon^2 > 18\sigma^2 \ln\left(\frac{10}{9}\right). \tag{154}$$

Then, there exists $i_0 \in \{1, \ldots, k\}$ such that $\|s^* - t_{i_0}\|_2 < \xi < \frac{\epsilon}{3}$.

Proof. **Preliminary bounds.**

- With our hypotheses (152) and (153), we have, for $K \geq 2$

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} = (1 - (4K\alpha - 1)(\mu + \gamma)) > 1 - \frac{1}{5K} \geq \frac{9}{10}. \tag{155}$$

We obtain

$$2\left(1-\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}}\right) \leq 2\left(1-\frac{9}{10}\right) = \frac{1}{5}. \tag{156}$$

Using Lemma A.2, we have

$$\frac{\|a_i\delta_{t_i} - b_i\delta_{s_i}\|_h}{\|a\|_\infty} \leq \frac{|a_i|\|\delta_{t_i} - \delta_{s_i}\|_h + |b_i - a_i|}{\|a\|_\infty}. \tag{157}$$

On one hand, $|a_i|/\|a\|_\infty \leq 1$. On the other hand, applying Lemma B.4 guarantees that $\|b - a_{1:l}\|_2/\|a\|_\infty \leq 0.77$. We deduce that $|b_i - a_i|/\|a\|_\infty \leq 0.77$. Combined with (157), we obtain

$$\frac{\|a_i\delta_{t_i} - b_i\delta_{s_i}\|_h}{\|a\|_\infty} \leq \|\delta_{t_i} - \delta_{s_i}\|_h + 0.77 \leq \frac{1}{5} + 0.77 < 1. \tag{158}$$

Hence,

$$D := \max_{i\in\{k+1,\dots,K\}} \frac{\|a_i\delta_{t_i} - b_i\delta_{s_i}\|_h}{\|a\|_\infty} < 1. \tag{159}$$

This inequality (158) will be useful to use Lemma B.2 and Lemma B.3 later in the proof.

- From (153), we deduce that the hypothesis

$$\mu + \gamma \leq \frac{1}{4K\alpha - 1}. \tag{160}$$

of Lemma B.3 is verified.

- We have that (155) implies:

$$\frac{1}{1 - (4K\alpha - 1)(\mu + \gamma)} < \frac{10}{9} \tag{161}$$

and

$$\xi = \sqrt{2\sigma^2 \ln\left(\frac{1}{1 - 4(4K\alpha - 1)(\mu + \gamma)}\right)} < \sqrt{2\sigma^2 \ln\left(\frac{10}{9}\right)} = \frac{1}{3}\sqrt{18\sigma^2 \ln\left(\frac{10}{9}\right)}. \tag{162}$$

Using the hypothesis (154), we obtain

$$\xi < \frac{\epsilon}{3}. \tag{163}$$

Main proof of Proposition B.1. We divide $\mathbb{R}^d$ into three sets:

- $E_1 = \{s \in \mathbb{R}^d : \exists i_0 \in \{1,\dots,k\} \;/\; \|s - t_{i_0}\|_2 < \xi\}$;
- $E_2 = \{s \in \mathbb{R}^d : \exists i_0 \in \{1,\dots,k\} \;/\; \xi \leq \|s - t_{i_0}\|_2 < \frac{\epsilon}{3}\}$;
- $E_3 = \{s \in \mathbb{R}^d : \forall t_i, \|s - t_i\|_2 \geq \frac{\epsilon}{3}\}$.

We note that E_1, E_2 and E_3 are pairwise disjoint and $E_1 \cup E_2 \cup E_3 = \mathbb{R}^n$. Moreover, since $\xi < \frac{\epsilon}{3}$, the set E_2 is non-empty.

a) Construction of the proof Our goal is to show that if $s^* \in \underset{\tilde{s}\in E_1\cup E_2\cup E_3}{\arg\max} |\langle A\delta_{\tilde{s}}, AX\rangle|$ then $s^* \in E_1$. To proceed, we separate the proof in two parts.
First, we choose a specific element $s_1 \in E_1$ and any $s_2 \in E_2$. Then, by showing that $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_2}, AX\rangle|$, we rule out the possibility that $s^* \in E_2$.
Then, we choose a specific element $s_1 \in E_1$ and any element $s_3 \in E_3$. By showing that $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_3}, AX\rangle|$, we rule out the possibility that $s^* \in E_3$.
Finally, since $s^* \notin E_2 \cup E_3$ and that $E_1 \cup E_2 \cup E_3 = \mathbb{R}^n$, then $s^* \in E_1$ i.e. $\exists i_0 \,/\, \|s^* - t_{i_0}\|_2 < \xi$.

b) Comparison between $s_1 \in E_1$ and $s_2 \in E_2$ Let $s_1 \in E_1$ and i_0 such that $\|s_1 - t_{i_0}\|_2 \leq \xi$. As s^* does not change if we multiply X by -1, we suppose that $a_0 \geq 0$ without loss of generality. We have

$$
\begin{aligned}
|\langle A\delta_{s_1}, AX\rangle| &\geq \langle A\delta_{s_1}, AX\rangle && (164)\\
&= \langle A\delta_{s_1}, Aa_{i_0}\delta_{t_{i_0}} + A\sum_{i=1;i\neq i_0}^{k} a_i\delta_{t_i} + Az_l\rangle && (165)\\
&= \langle A\delta_{s_1}, Aa_{i_0}\delta_{t_{i_0}}\rangle + \langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} a_i\delta_{t_i}\rangle + \langle A\delta_{s_1}, Az_l\rangle. && (166)
\end{aligned}
$$

We focus on the term $\langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} a_i\delta_{t_i}\rangle$. We split the sum in two terms, with $I = \{i \neq i_0 : a_i \geq 0\}$ and $J = \{i \neq i_0 : a_i < 0\}$ to get,

$$
\begin{aligned}
\langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} a_i\delta_{t_i}\rangle &= \sum_{i=1;i\neq i_0}^{k} a_i\langle A\delta_{s_1}, A\delta_{t_i}\rangle && (167)\\
&= \sum_{i\in I} |a_i|\langle A\delta_{s_1}, A\delta_{t_i}\rangle - \sum_{i\in J} |a_i|\langle A\delta_{s_1}, A\delta_{t_i}\rangle. && (168)
\end{aligned}
$$

Using Lemma A.3, we get

$$
\begin{aligned}
\langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} a_i\delta_{t_i}\rangle &\geq (1+\gamma)\sum_{i\in I} |a_i|\langle \delta_{s_1}, \delta_{t_i}\rangle_h - 2\gamma\sum_{i\in I}|a_i|\\
&\quad - (1-\gamma)\sum_{i\in J} |a_i|\langle \delta_{s_1}, \delta_{t_i}\rangle_h - 2\gamma\sum_{i\in J}|a_i|\\
&= (1+\gamma)\sum_{i\in I} |a_i|\langle \delta_{s_1}, \delta_{t_i}\rangle_h - 2\gamma\sum_{i\in I}|a_i|\\
&\quad - (1-\gamma)\sum_{i\in J} |a_i| e^{-\frac{\|s_1-t_i\|_2^2}{2\sigma^2}} - 2\gamma\sum_{i\in J}|a_i|.
\end{aligned} \tag{169}
$$

Since we have $\|t_{i_0} - s_1\|_2 \leq \xi < \frac{\epsilon}{3}$ and the t_i are pairwise ϵ-separated, we use the triangle inequality to get for $i \neq i_0$,

$$
\|t_{i_0} - s_1\|_2 + \|s_1 - t_i\|_2 \geq \|t_{i_0} - t_i\|_2 \quad \text{i.e.} \quad \frac{\epsilon}{3} + \|s_1 - t_i\|_2 > \epsilon \quad \text{so that} \quad \|s_1 - t_i\|_2 > \frac{2\epsilon}{3}. \tag{170}
$$

This means that for $i \neq i_0$, with Lemma 3.1 (iii), we have $\mathrm{e}^{-\|s_1-t_i\|_2^2/(2\sigma^2)} \leq \mu$. By injecting this result into (169), we obtain

$$\begin{aligned}\langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} a_i\delta_{t_i}\rangle \geq (1+\gamma)\sum_{i\in I}|a_i|\langle\delta_{s_1},\delta_{t_i}\rangle_h - 2\gamma\sum_{i\in I}|a_i| \\ - (1-\gamma)\mu\sum_{i\in J}|a_i| - 2\gamma\sum_{i\in J}|a_i|.\end{aligned} \tag{171}$$

Since $\forall i \in I, \langle\delta_{s_1},\delta_{t_i}\rangle_h \geq 0$ and $|I|+|J| = k-1$, we have that

$$\langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} a_i\delta_{t_i}\rangle \geq -(1-\gamma)\mu\sum_{i\in J}|a_i| - 2\gamma\sum_{i=1;i\neq i_0}^{k}|a_i|. \tag{172}$$

In the worst case, all the amplitudes a_i are negative (i.e. $|J| = k-1$), this implies

$$\langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} a_i\delta_{t_i}\rangle \geq -\left((1-\gamma)\mu + 2\gamma\right)\sum_{i=1;i\neq i_0}^{k}|a_i|. \tag{173}$$

Going back to $|\langle A\delta_{s_1}, AX\rangle|$, we use Lemma B.1 with the upper bound $\kappa_l \geq |\langle A\delta_{s_1}, Az_l\rangle|/\|a\|_\infty$ defined in (81). Using Lemma A.3, with (166) and (173), we obtain with $a_{i_0} \geq 0$,

$$\begin{aligned}|\langle A\delta_{s_1}, AX\rangle| &\geq (1+\gamma)|a_{i_0}|\langle\delta_{s_1},\delta_{t_{i_0}}\rangle_h - 2|a_{i_0}|\gamma - \left((1-\gamma)\mu+2\gamma\right)\sum_{i=1;i\neq i_0}^{k}|a_i| - \|a\|_\infty\kappa_l \\ &= (1+\gamma)|a_{i_0}|\mathrm{e}^{-\frac{\|s_1-t_{i_0}\|_2^2}{2\sigma^2}} - (1-\gamma)\mu\sum_{i=1;i\neq i_0}^{k}|a_i| - 2\gamma\sum_{i=1}^{k}|a_i| - \|a\|_\infty\kappa_l.\end{aligned} \tag{174}$$

Now that we have a lower bound for $|\langle A\delta_{s_1}, AX\rangle|$, let us calculate an upper bound for $|\langle A\delta_{s_2}, AX\rangle|$,

$$|\langle A\delta_{s_2}, AX\rangle| = \left|\sum_{i=1}^{k} a_i\langle A\delta_{s_2}, A\delta_{t_i}\rangle + \langle A\delta_{s_2}, Az\rangle\right| \leq \sum_{i=1}^{k}|a_i\langle A\delta_{s_2}, A\delta_{t_i}\rangle| + |\langle A\delta_{s_2}, Az\rangle|\,. \tag{175}$$

There is j_0 such that $\xi \leq \|t_{j_0} - s^*\|_2 < \epsilon/3$. With Lemma B.1, we get an upper bound for $|\langle A\delta_{s_2}, Az\rangle|/\|a\|_\infty$,

$$|\langle A\delta_{s_2}, AX\rangle| \leq |a_{j_0}||\langle A\delta_{s_2}, A\delta_{t_{j_0}}\rangle| + \sum_{i=1;i\neq j_0}^{k}|a_i||\langle A\delta_{s_2}, A\delta_{t_i}\rangle| + \|a\|_\infty\kappa_l. \tag{176}$$

We use the upper bound given in Lemma A.3 to get,

$$\begin{aligned}|\langle A\delta_{s_2}, AX\rangle| \quad &\leq (1-\gamma)|a_{j_0}|\langle \delta_{s_2}, \delta_{t_{j_0}}\rangle_h + 2\gamma|a_{j_0}| \\ &\quad + (1-\gamma)\sum_{i=1;i\neq j_0}^{k}|a_i|\langle \delta_{s_2}, \delta_{t_i}\rangle_h + 2\gamma\sum_{i=1;i\neq j_0}^{k}|a_i| + \|a\|_\infty\kappa_l \end{aligned} \tag{177}$$

$$\begin{aligned}|\langle A\delta_{s_2}, AX\rangle| \quad &\leq (1-\gamma)|a_{j_0}|\mathrm{e}^{-\frac{\|s_2-t_{j_0}\|_2^2}{2\sigma^2}} + 2\gamma|a_{j_0}| \\ &\quad + (1-\gamma)\sum_{i=1;i\neq j_0}^{k}|a_i|\mathrm{e}^{-\frac{\|s_2-t_i\|_2^2}{2\sigma^2}} + 2\gamma\sum_{i=1;i\neq j_0}^{k}|a_i| + \|a\|_\infty\kappa_l. \end{aligned} \tag{178}$$

Since $\|t_{j_0} - s_2\|_2 \leq \frac{\epsilon}{3} < \frac{\epsilon}{2}$, with the triangle inequality, we get, for $i \neq j_0$,

$$\|t_{j_0} - s_2\|_2 + \|s_2 - t_i\|_2 \geq \|t_{j_0} - t_i\|_2 \quad \text{i.e.} \quad \frac{\epsilon}{2} + \|s_2 - t_i\|_2 > \epsilon \quad \text{so that} \quad \|s_2 - t_i\|_2 > \frac{\epsilon}{2}. \tag{179}$$

By using this property in (178) and by the fact that, as in (14), for all $i \neq j_0$, $\mathrm{e}^{-\|s_2-t_{j_0}\|_2^2/(2\sigma^2)} \leq \mu$, we obtain

$$|\langle A\delta_{s_2}, AX\rangle| \leq (1-\gamma)|a_{j_0}|\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} + (1-\gamma)\mu\sum_{i=1;i\neq j_0}^{k}|a_i| + 2\gamma\sum_{i=1}^{k}|a_i| + \|a\|_\infty\kappa_l. \tag{180}$$

A sufficient condition to prove $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_2}, AX\rangle|$ is

$$\begin{aligned}&|\langle A\delta_{s_1}, AX\rangle| \geq (1+\gamma)|a_{i_0}|\mathrm{e}^{-\frac{\|s_1-t_{i_0}\|_2^2}{2\sigma^2}} - (1-\gamma)\mu\sum_{i=1;i\neq i_0}^{k}|a_i| - 2\gamma\sum_{i=1}^{k}|a_i| - \|a\|_\infty\kappa_l > \\ &\qquad (1-\gamma)|a_{j_0}|\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} + (1-\gamma)\mu\sum_{i=1;i\neq j_0}^{k}|a_i| + 2\gamma\sum_{i=1}^{k}|a_i| + \|a\|_\infty\kappa_l \geq |\langle A\delta_{s_2}, AX\rangle|. \end{aligned} \tag{181}$$

Consider any $s_1 \in E_1$, if for all $s_2 \in E_2$, the previous inequality (181) is verified, then $s^* \notin E_2$. Now take $s_1 = t_1 = t_{i_0} \in E_1$ with $i_0 = 1$ and, we rewrite the middle inequality in (181) as

$$\begin{aligned}(1-\gamma)|a_{j_0}|\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < (1+\gamma)|a_1| &- (1-\gamma)\mu\left(\sum_{i=2}^{k}|a_i| + \sum_{i=1;i\neq j_0}^{k}|a_i|\right) \\ &- 4\gamma\sum_{i=1}^{k}|a_i| - 2\|a\|_\infty\kappa_l. \end{aligned} \tag{182}$$

By dividing both sides by $(1-\gamma)|a_{j_0}|$, we get that the inequality above is verified if the one following is verified,

$$\begin{aligned}\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < \frac{1}{(1-\gamma)|a_{j_0}|}\Bigg[|a_1| &- (1-\gamma)\mu\left(\sum_{i=2}^{k}|a_i| + \sum_{i=1;i\neq j_0}^{k}|a_i|\right) \\ &- \gamma\left(4\left(\sum_{i=1}^{k}|a_i|\right) - |a_1|\right) - 2\|a\|_\infty\kappa_l\Bigg]. \end{aligned} \tag{183}$$

If we rewrite $\sum_{i=1}^k |a_i|$ as $\|a\|_1$, we obtain

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} = \frac{1}{1-\gamma}\left[\frac{|a_1|}{|a_{j_0}|} - (1-\gamma)\mu\left(\frac{\|a\|_1 - |a_1|}{|a_{j_0}|} + \frac{\|a\|_1 - |a_{j_0}|}{|a_{j_0}|}\right) - \frac{4\|a\|_1 - |a_1|}{|a_{j_0}|}\gamma - 2\frac{\|a\|_\infty}{|a_{j_0}|}\kappa_l\right]. \tag{184}$$

With $\alpha = |a_1|/|a_k| = \|a\|_\infty / \min_i |a_i|$ then $(\|a\|_1 - |a_1|)/|a_{j_0}| \leq (k-1)\alpha$, $(\|a\|_1 - |a_{j_0}|)/|a_{j_0}| \leq k\alpha - 1$ and $(4\|a\|_1 - |a_1|)/|a_{j_0}| \leq (4k-1)\alpha$. We deduce that the above inequality is verified if

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < \frac{1 - ((2k-1)\alpha - 1)(1-\gamma)\mu - (4k\alpha - 1)\gamma - 2\kappa_l\alpha}{1-\gamma}. \tag{185}$$

With Lemma B.2, since (152) is verified and we have shown with (159) that $D < 1$, we have that (185) is verified and in turn $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_2}, AX\rangle|$. We deduce that $|\langle A\delta_{s^*}, AX\rangle| \geq |\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_2}, AX\rangle|$.

c) Comparison between $s_1 \in E_1$ and $s_3 \in E_3$ For this case, we use the same lower bound (174) for $|\langle A\delta_{s_1}, AX\rangle|$. Let $s_3 \in E_3$,

$$|\langle A\delta_{s_3}, AX\rangle| = \left|\sum_{i=1}^k a_i \langle A\delta_{s_3}, A\delta_{t_i}\rangle + \langle A\delta_{s_3}, Az_l\rangle\right| \leq \sum_{i=1}^k |a_i\langle A\delta_{s_3}, A\delta_{t_i}\rangle| + |\langle A\delta_{s_3}, Az_l\rangle|. \tag{186}$$

We use Lemma B.1 and the upper bound κ_l defined in (81) to get

$$|\langle A\delta_{s_3}, AX\rangle| \leq \sum_{i=1}^k |a_i||\langle A\delta_{s_3}, A\delta_{t_i}\rangle| + \|a\|_\infty \kappa_l. \tag{187}$$

Using Lemma A.3,

$$|\langle A\delta_{s_3}, AX\rangle| \leq (1-\gamma)\sum_{i=1}^k |a_i|\langle \delta_{s_3}, \delta_{t_i}\rangle_h + 2\gamma\sum_{i=1}^k |a_i| + \|a\|_\infty\kappa_l. \tag{188}$$

Since $s_3 \in E_3$, we have that $\forall i \in \{1,\dots,K\}, \|s_3 - t_i\|_2 \geq \frac{\epsilon}{3}$ and with Lemma 3.1 (iii), we obtain

$$|\langle A\delta_{s_3}, AX\rangle| \leq (1-\gamma)\mu\sum_{i=1}^k |a_i| + 2\gamma\sum_{i=1}^k |a_i| + \|a\|_\infty\kappa_l. \tag{189}$$

A sufficient condition to prove $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_3}, AX\rangle|$ is

$$(1+\gamma)|a_{i_0}|\mathrm{e}^{-\frac{\|s_1 - t_{i_0}\|_2^2}{2\sigma^2}} - (1-\gamma)\mu\sum_{i=i;i\neq i_0}^k |a_i| - 2\gamma\sum_{i=1}^k |a_i| - \|a\|_\infty\kappa_l > (1-\gamma)\mu\sum_{i=1}^k |a_i| + 2\gamma\sum_{i=1}^k |a_i| + \|a\|_\infty\kappa_l. \tag{190}$$

Consider any $s_1 \in E_1$. If for all $s_3 \in E_3$, the previous inequality (190) is verified, then $s^* \notin E_3$. Now take $s_1 = t_1 = t_{i_0} \in E_1$ with $i_0 = 1$, we rewrite (190) as

$$(1+\gamma)|a_1| - 2\kappa_l \|a\|_\infty > (1-\gamma)\mu \left(2\sum_{i=1}^{k} |a_i| - |a_1| \right) + 4\gamma \sum_{i=1}^{k} |a_i|. \tag{191}$$

By isolating μ, we get

$$\mu < \frac{|a_1| - (4\sum_{i=1}^{k} |a_i| - |a_1|)\gamma - 2\kappa_l \|a\|_\infty}{(2\sum_{i=1}^{k} |a_i| - |a_1|)(1-\gamma)} = \frac{1 - (4\|a\|_1/|a_1| - 1)\gamma - 2\kappa_l \|a\|_\infty/|a_1|}{(2\|a\|_1/|a_1| - 1)(1-\gamma)}. \tag{192}$$

This is verified if

$$\mu < \frac{1 - (4k\alpha - 1)\gamma - 2\kappa_l \alpha}{(2k\alpha - 1)(1-\gamma)}. \tag{193}$$

As shown in Lemma B.3 thanks to (160) and the fact that $D < 1$ (from (159)), we have that (193) is true. We deduce that $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_3}, AX\rangle|$ and $|\langle A\delta_{s^*}, AX\rangle| \geq |\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_3}, AX\rangle|$.

d) Conclusion: We have shown that for all $s \in E_2 \cup E_3$, there is $s_1 \in E_1$ such that $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_s, AX\rangle|$. This shows that s^* is necessarily in E_1 as $E_1 \cap (E_2 \cup E_3) = \emptyset$ i.e. there exists an i_0 such that $\|s^* - t_{i_0}\|_2 < \xi$. □

We can now prove Theorem 3.3 by induction.

Proof for Theorem (3.3). Let us define the set

$$\mathcal{X}_{K,l}^{\epsilon,\xi} := \Bigg\{ x_{k,\epsilon} + z_{l,\xi,\frac{\epsilon}{3}} : k = K - l;\ x_{k,\epsilon} = \sum_{i=1}^{k} a_i \delta_{t_i} \in \Sigma_{K,\epsilon};$$
$$z_{l,\xi,\frac{\epsilon}{3}} = \sum_{i=k+1}^{k+l} a_i \delta_{t_i} - b_i \delta_{s_i} \in Z_{l,\xi,\frac{\epsilon}{3}};\ \{t_i, i = 1,\dots,K\} \text{ pairwise } \epsilon\text{-separated} \Bigg\}. \tag{194}$$

Let us define $X_{K,0} := x_{K,\epsilon}$ and, for $l \in \{1,\dots,K-1\}$, $X_{K,l} = x_{K,\epsilon} - \sum_{i=1}^{l} b_i \delta_{s_i^*}$, where b_i is the amplitude output by OP-COMP at step i. Our goal is to show by induction that for all $l = 0,\dots,K$, we have that $X_{K,l} \in \mathcal{X}_{K,l}^{\epsilon,\xi}$.

a) Induction assumption For every $l \in \{0,\dots,K\}, X_{K,l} \in \mathcal{X}_{K,l}^{\epsilon,\xi}$.
Let $\mathcal{P}(l)$: $X_{K,l} \in \mathcal{X}_{K,l}^{\epsilon,\xi}$. We give a proof by induction on l.

b) $\mathcal{P}(0)$ We have $X_{K,0} = x_{K,\epsilon} \in \Sigma_{K,\epsilon}$. Using the definition of $\mathcal{X}_{K,l}^{\epsilon,\xi}$, we have that $X_{K,0} = x_{K,\epsilon} \in \mathcal{X}_{K,0}^{\epsilon,\xi}$ with $z_{l,\xi,\frac{\epsilon}{3}} = 0$, $\mathcal{P}(0)$ is true.

c) Induction step Let $l \in \{0, \dots, K-1\}$ and $k = K - l$. Assume $\mathcal{P}(l)$. We show $\mathcal{P}(l+1)$.

To ease our notations, we consider a permutation $\tilde{t}$ of the positions t (and the corresponding permutation $\tilde{a}$ of the amplitudes a) such that for the first l outputs of COMP without sliding $s_1^*, \dots, s_l^*$, we have $\forall i = 1, \dots, l+1, \|\tilde{t}_{K-l+i} - s_i^*\|_2 < \xi$. The hypotheses of Proposition B.1 are verified for $X_{K,l}$, so there exists $i_0 \in \{1, \dots, k-l\}$ such that $\|s_{l+1}^* - \tilde{t}_{i_0}\|_2 < \xi < \frac{\epsilon}{3}$ and

$$X_{K,l+1} = x_{K,\epsilon} - \sum_{j=1}^{l+1} b_j \delta_{s_j^*} = \sum_{i=1}^{K} \tilde{a}\delta_{\tilde{t}_i} - \sum_{j=1}^{l+1} b_j \delta_{s_j^*} \tag{195}$$

$$= \underbrace{\sum_{i=1; i\neq i_0}^{k} \tilde{a}_i \delta_{\tilde{t}_i}}_{=x_{k-1,\epsilon} \in \Sigma_{k-1,\epsilon}} + \underbrace{\sum_{i=K-l+1}^{K} \tilde{a}_i \delta_{\tilde{t}_i} - \sum_{j=1}^{l} b_j \delta_{s_j^*} + \tilde{a}_i \delta_{\tilde{t}_{i_0}} - b_{l+1}\delta_{s_{l+1}^*}}_{=y \in Z_{l,\xi,\frac{\epsilon}{3}}}. \tag{196}$$

To show that $y + \tilde{a}_i \delta_{\tilde{t}_{i_0}} - b_{l+1}\delta_{s_{l+1}^*} \in Z_{l+1,\xi,\frac{\epsilon}{3}}$, as we have ensured that $\pi_{l+1} := \tilde{a}_i \delta_{\tilde{t}_{i_0}} - b_{l+1}\delta_{s_{l+1}^*}$ is a ξ-concentrated dipole, we just need to make sure that π_{l+1} is $\frac{\epsilon}{3}$-separated with every $\pi_i := \tilde{a}_{K-l+i}\delta_{\tilde{t}_{K-l+i}} - b_i \delta_{s_i^*}$ for $i = 1, \dots, l$.

d) $\frac{\epsilon}{3}$-separated dipoles Using the reverse triangle inequality we have for all $i \in \{1, \dots, l\}$, $\|\tilde{t}_i - s_{l+1}^*\|_2 \geq \|\tilde{t}_i - \tilde{t}_{i_0}\|_2 - \|\tilde{t}_{i_0} - s_{l+1}^*\|_2 \geq \epsilon - \xi > \epsilon - \frac{\epsilon}{3} > \frac{\epsilon}{3}$.

To control distances $\|s_{l+1}^* - s_i^*\|_2$ for $i = 1, \dots, l$, we use that $\|\tilde{t}_{K-l+i} - s_i^*\|_2 \leq \xi < \frac{\epsilon}{3}$ (from $\mathcal{P}(l)$). Starting from $\|\tilde{t}_{i_0} - \tilde{t}_{K-l+i}\|_2 \geq \epsilon$, we have

$$\|\tilde{t}_{i_0} - \tilde{t}_{K-l+i}\|_2 \geq \epsilon \tag{197}$$

$$\text{i.e.} \quad \|\tilde{t}_{i_0} - s_{l+1}^* + s_{l+1}^* - s_i^* + s_i^* - \tilde{t}_{K-l+i}\|_2 \geq \epsilon \tag{198}$$

$$\text{so that} \quad \|\tilde{t}_{i_0} - s_{l+1}^*\|_2 + \|s_{l+1}^* - s_i^*\|_2 + \|s_i^* - \tilde{t}_{K-l+i}\|_2 \geq \epsilon \tag{199}$$

$$\text{hence} \quad \frac{\epsilon}{3} + \|s_{l+1}^* - s_i^*\|_2 + \frac{\epsilon}{3} \geq \epsilon \tag{200}$$

$$\text{so} \quad \|s_{l+1}^* - s_i^*\|_2 > \frac{\epsilon}{3}. \tag{201}$$

We showed that $\{s_i^*, i = 1, \dots, l\}$ are $\frac{\epsilon}{3}$-separated with s_{l+1}^*. Since the supports of $\{\pi_i, i = 1, \dots, l\}$ are $\frac{\epsilon}{3}$-separated with the support of π_{l+1}, we have that $y + \tilde{a}_i \delta_{\tilde{t}_{i_0}} - b_{l+1}\delta_{s_{l+1}^*} \in Z_{l+1,\xi,\frac{\epsilon}{3}}$.

Finally, by setting $z_{l+1,\xi,\frac{\epsilon}{3}} = y + \tilde{a}_i \delta_{\tilde{t}_{i_0}} - b_{l+1}\delta_{s_{l+1}^*}$, we get

$$X_{K,l+1} = x_{k-1,\epsilon} + z_{l+1,\xi,\frac{\epsilon}{3}} \in \mathcal{X}_{K,l+1}^{\epsilon,\xi}. \tag{202}$$

That is, the statement $\mathcal{P}(l+1)$ also holds true, establishing the induction step.

e) Conclusion Since both the base case and the induction step have been proven as true, by induction, the statement $\mathcal{P}(l)$ holds for every $l \in \{0, \dots, K\}$ which implies that for every $l \in \{0, \dots, K\}$ there is $i \in \{0, \dots, K\}$, $\|s_l^* - t_i\|_2 \leq \xi$. □

C. Appendix: Proofs for Theorem 3.4 (result without amplitudes)

The proof is essentially the same as Theorem 3.3. As the induction is same, we just update the induction step in the following Proposition.

Proposition C.1. *Let $k, l, K \in \mathbb{N}$ such that $k+l = K \geq 2$. Let $x_{k,\epsilon} = \sum_{i=1}^{k} \delta_{t_i} \in \Sigma_{K,\epsilon}$. Assume that the linear operator A has the γ-RIP on $\mathcal{S}(\Sigma_{K,\frac{\epsilon}{3}})$ with a Gaussian kernel h of variance σ^2. Let*

$$\xi := \sqrt{2\sigma^2 \ln\left(\frac{1}{(1-(4K-1)(\mu+\gamma))}\right)}. \tag{203}$$

Let $X = x_{k,\epsilon} + z_l$ where $z_l = \sum_{i=k+1}^{K} \delta_{t_i} - \delta_{s_i} \in Z_{l,\xi,\frac{\epsilon}{3}}$ and $s^* \in \underset{\tilde{s}\in\mathbb{R}^d}{\arg\max} |\langle A\delta_{\tilde{s}}, AX\rangle|$. Suppose

$$\mu + \gamma < \frac{1}{10\,(4K-1)} \tag{204}$$

and

$$\epsilon^2 > 18\sigma^2 \ln\left(\frac{10}{9}\right). \tag{205}$$

Then, there exists $i_0 \in \{1, \dots, k\}$ such that $\|s^* - t_{i_0}\|_2 < \xi < \frac{\epsilon}{3}$.

This proof is similar to the one of Proposition B.1. The elements that differs are the less strict hypothesis.

Proof. **Preliminary bounds.**

- With our hypotheses (203) and (204), we have, for $K \geq 2$,

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} = 1 - (4K-1)(\mu+\gamma) > 1 - \frac{1}{10} = \frac{9}{10}. \tag{206}$$

  We obtain

$$2\left(1 - \mathrm{e}^{-\frac{\xi^2}{2\sigma^2}}\right) \leq 2\left(1 - \frac{9}{10}\right) = \frac{1}{5}. \tag{207}$$

  We can bound $\|\delta_{t_i} - \delta_{s_i}\|_h$ for all $i \in \{k+1, \dots, K\}$,

$$\|\delta_{t_i} - \delta_{s_i}\|_h = \|\delta_{t_i}\|_h + \|\delta_{s_i}\|_h - 2\langle \delta_{t_i}, \delta_{s_i}\rangle_h = 2(1 - \mathrm{e}^{\frac{\|t_i - s_i\|_2^2}{2\sigma^2}}) \tag{208}$$

  so that
$$\|\delta_{t_i} - \delta_{s_i}\|_h < 2\left(1 - \mathrm{e}^{-\frac{\xi^2}{2\sigma^2}}\right) = \frac{1}{5}. \tag{209}$$

  Hence,

$$D := \max_{i\in\{k+1,\dots,K\}} \|\delta_{t_i} - \delta_{s_i}\|_h < 1. \tag{210}$$

  This inequality (210) will be useful to use Lemma B.2 and Lemma B.3 later in the proof.

- From (204), we deduce that the hypothesis

$$\mu + \gamma \leq \frac{1}{4K\alpha - 1}. \tag{211}$$

  of Lemma B.3 is verified.

- We have that (206) implies $\left(1-(4K-1)(\mu+\gamma)\right)^{-1} < \frac{10}{9}$, so that

$$\xi = \sqrt{2\sigma^2 \ln\left(\frac{1}{1-(4K-1)(\mu+\gamma)}\right)} < \sqrt{2\sigma^2 \ln\left(\frac{10}{9}\right)} = \frac{1}{3}\sqrt{18\sigma^2 \ln\left(\frac{10}{9}\right)}. \tag{212}$$

  Using the hypothesis (205), we obtain $\xi < \frac{\epsilon}{3}$.

Main proof of Proposition C.1. We divide $\mathbb{R}^d$ into three sets:

- $E_1 = \{s \in \mathbb{R}^d : \exists i_0 \in \{1,\dots,k\} \,/\, \|s - t_{i_0}\|_2 < \xi\}$;
- $E_2 = \{s \in \mathbb{R}^d : \exists i_0 \in \{1,\dots,k\} \,/\, \xi \leq \|s - t_{i_0}\|_2 < \frac{\epsilon}{3}\}$;
- $E_3 = \{s \in \mathbb{R}^d : \forall t_i, \|s - t_i\|_2 \geq \frac{\epsilon}{3}\}$.

We note that E_1, E_2 and E_3 are pairwise disjoint and $E_1 \cup E_2 \cup E_3 = \mathbb{R}^n$. Moreover, since $\xi < \frac{\epsilon}{3}$, the set E_2 is non-empty.

b) Comparison between $s_1 \in E_1$ and $s_2 \in E_2$ Let $s_1 \in E_1$ and i_0 such that $\|s_1 - t_{i_0}\|_2 \leq \xi$. We have

$$|\langle A\delta_{s_1}, AX\rangle| \quad \geq \langle A\delta_{s_1}, AX\rangle = \langle A\delta_{s_1}, A\delta_{t_{i_0}} + A\sum_{i=1;i\neq i_0}^{k} \delta_{t_i} + Az_l\rangle \tag{213}$$

$$= \langle A\delta_{s_1}, A\delta_{t_{i_0}}\rangle + \langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} \delta_{t_i}\rangle + \langle A\delta_{s_1}, Az_l\rangle. \tag{214}$$

Using Lemma A.3, we get

$$\langle A\delta_{s_1}, A\sum_{i=1;i\neq i_0}^{k} \delta_{t_i}\rangle \geq (1+\gamma)\sum_{i=1;i\neq i_0}^{k} \langle \delta_{s_1}, \delta_{t_i}\rangle_h - \sum_{i=1;i\neq i_0}^{k} 2\gamma \geq -2\gamma(k-1). \tag{215}$$

Going back to $|\langle A\delta_{s_1}, AX\rangle|$, we use Lemma B.1 with the upper bound $\kappa_l \geq |\langle A\delta_{s_1}, Az_l\rangle|$ defined in (81). Using Lemma A.3, with (214) and (215), we obtain,

$$|\langle A\delta_{s_1}, AX\rangle| \geq (1+\gamma)\langle \delta_{s_1}, \delta_{t_{i_0}}\rangle_h - 2\gamma - 2\gamma(k-1) - \kappa_l \tag{216}$$

$$= (1+\gamma)\mathrm{e}^{-\frac{\|s_1 - t_{i_0}\|_2^2}{2\sigma^2}} - 2\gamma k - \kappa_l. \tag{217}$$

Now that we have a lower bound for $|\langle A\delta_{s_1}, AX\rangle|$, let us get an upper bound for $|\langle A\delta_{s_2}, AX\rangle|$,

$$|\langle A\delta_{s_2}, AX\rangle| = \left|\sum_{i=1}^{k}\langle A\delta_{s_2}, A\delta_{t_i}\rangle + \langle A\delta_{s_2}, Az\rangle\right| \leq \sum_{i=1}^{k} |\langle A\delta_{s_2}, A\delta_{t_i}\rangle| + |\langle A\delta_{s_2}, Az\rangle|. \tag{218}$$

There is j_0 such that $\xi \leq \|t_{j_0} - s^*\|_2 < \frac{\epsilon}{3}$. With Lemma B.1, we get an upper bound for $|\langle A\delta_{s_2}, Az\rangle|$,

$$|\langle A\delta_{s_2}, AX\rangle| \leq |\langle A\delta_{s_2}, A\delta_{t_{j_0}}\rangle| + \sum_{i=1;i\neq j_0}^{k} \langle A\delta_{s_2}, A\delta_{t_i}\rangle| + \kappa_l. \tag{219}$$

We use the upper bound given in Lemma A.3 to get,

$$\begin{aligned}|\langle A\delta_{s_2}, AX\rangle| \leq & (1-\gamma)\langle \delta_{s_2}, \delta_{t_{j_0}}\rangle_h + 2\gamma + (1-\gamma)\sum_{i=1;i\neq j_0}^{k} \langle \delta_{s_2}, \delta_{t_i}\rangle_h \\ & + 2\gamma \sum_{i=1;i\neq j_0}^{k} 1+\kappa_l \\ \leq & (1-\gamma)\mathrm{e}^{-\frac{\|s_2-t_{j_0}\|_2^2}{2\sigma^2}} + 2\gamma + (1-\gamma)\sum_{i=1;i\neq j_0}^{k} \mathrm{e}^{-\frac{\|s_2-t_i\|_2^2}{2\sigma^2}} \\ & + 2\gamma(k-1) + \kappa_l.\end{aligned} \tag{220}$$

Since $\|t_{j_0} - s_2\|_2 \leq \frac{\epsilon}{3}$, with the triangle inequality, we get, for $i \neq j_0$,

$$\|t_{j_0}-s_2\|_2+\|s_2-t_i\|_2 \geq \|t_{j_0}-t_i\|_2 \quad \text{i.e.} \quad \frac{\epsilon}{3}+\|s_2-t_i\|_2 > \epsilon \quad \text{so that} \quad \|s_2-t_i\|_2 > \frac{2\epsilon}{3} > \frac{\epsilon}{3}. \tag{221}$$

By using this property in (220) and by the fact that, as in (14), $\forall i \neq j_0, \mathrm{e}^{-\|s_2-t_i\|_2^2/(2\sigma^2)} \leq \mu$, we obtain

$$|\langle A\delta_{s_2}, AX\rangle| \leq (1-\gamma)\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} + (1-\gamma)\mu k + 2\gamma k + \kappa_l. \tag{222}$$

A sufficient condition to prove $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_2}, AX\rangle|$ is

$$\begin{aligned}|\langle A\delta_{s_1}, AX\rangle| \geq (1+\gamma)\mathrm{e}^{-\frac{\|s_1-t_{i_0}\|_2^2}{2\sigma^2}} - 2\gamma k - \kappa_l > \\ (1-\gamma)\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} + (1-\gamma)\mu k + 2\gamma k + \kappa_l \geq |\langle A\delta_{s_2}, AX\rangle|.\end{aligned} \tag{223}$$

Consider any $s_1 \in E_1$, if for all $s_2 \in E_2$, the previous inequality (223) is verified, then $s^* \notin E_2$. Now we take $s_1 = t_1 = t_{i_0} \in E_1$ with $i_0 = 1$, we rewrite the middle inequality in (223) as

$$(1-\gamma)\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < (1+\gamma) - (1-\gamma)\mu k - 4\gamma k - 2\kappa_l. \tag{224}$$

By dividing both sides by $1-\gamma$ and rearranging some terms, we get

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < \frac{1}{1-\gamma}\Big(1 - (1-\gamma)\mu k - (4k-1)\gamma - 2\kappa_l\Big). \tag{225}$$

With $\alpha \geq 1$, we deduce that the above inequality is verified if

$$\mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} < \frac{1 - ((2k-1)\alpha - 1)(1-\gamma)\mu - (4k\alpha - 1)\gamma - 2\kappa_l\alpha}{1-\gamma}. \tag{226}$$

With Lemma B.2, since (203) is verified and we have shown with (210) that $D < 1$, we get that (226) is verified and in turn $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_2}, AX\rangle|$. We deduce that $|\langle A\delta_{s^*}, AX\rangle| \geq |\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_2}, AX\rangle|$.

c) Comparison between $s_1 \in E_1$ and $s_3 \in E_3$ For this case, we use the same lower bound (217) for $|\langle A\delta_{s_1}, AX\rangle|$. Let $s_3 \in E_3$,

$$|\langle A\delta_{s_3}, AX\rangle| = \left|\sum_{i=1}^{k}\langle A\delta_{s_3}, A\delta_{t_i}\rangle + \langle A\delta_{s_3}, Az_l\rangle\right| \leq \sum_{i=1}^{k}|\langle A\delta_{s_3}, A\delta_{t_i}\rangle| + |\langle A\delta_{s_3}, Az_l\rangle|\,. \quad (227)$$

We use Lemma B.1 and the upper bound κ_l defined in (81) to get

$$|\langle A\delta_{s_3}, AX\rangle| \leq \sum_{i=1}^{k}\langle A\delta_{s_3}, A\delta_{t_i}\rangle| + \kappa_l. \quad (228)$$

Using Lemma A.3,

$$|\langle A\delta_{s_3}, AX\rangle| \leq (1-\gamma)\sum_{i=1}^{k}\langle \delta_{s_3}, \delta_{t_i}\rangle_h + \sum_{i=1}^{k} 2\gamma + \kappa_l = (1-\gamma)\sum_{i=1}^{k} \mathrm{e}^{-\frac{\|s_3-t_i\|_2^2}{2\sigma^2}} + 2\gamma k + \kappa_l. \quad (229)$$

Since $s_3 \in E_3$, we have that $\forall i \in \{1,\dots,K\}, \|s_3 - t_i\|_2 \geq \frac{\epsilon}{3}$ and with the mutual coherence in Assumption 1, we obtain

$$|\langle A\delta_{s_3}, AX\rangle| \leq (1-\gamma)\mu k + 2\gamma k + \kappa_l. \quad (230)$$

A sufficient condition to prove $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_3}, AX\rangle|$ is

$$(1+\gamma)\mathrm{e}^{-\frac{\|s_1-t_{i_0}\|_2^2}{2\sigma^2}} - 2\gamma k - \kappa_l > (1-\gamma)\mu k + 2\gamma k + \kappa_l. \quad (231)$$

Consider any $s_1 \in E_1$. If for all $s_3 \in E_3$, the previous inequality (231) is verified, then $s^* \notin E_3$. Now take $i_0 = 1$ and $s_1 = t_{i_0} \in E_1$, we rewrite (231) as

$$(1+\gamma) > (1-\gamma)\mu k + 4\gamma k + 2\kappa_l. \quad (232)$$

With $\alpha \geq 1$, this is verified if

$$\mu < \frac{1-(4k\alpha-1)\gamma - 2\kappa_l\alpha}{(2k\alpha-1)(1-\gamma)}. \quad (233)$$

As shown in Lemma B.3 thanks to (211) and the fact that $D < 1$ (from (210)), we get that (233) is true. We deduce that $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_3}, AX\rangle|$ and $|\langle A\delta_{s^*}, AX\rangle| \geq |\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s_3}, AX\rangle|$.

d) Conclusion We have shown that for all $s \in E_2 \cup E_3$, there is $s_1 \in E_1$ such that $|\langle A\delta_{s_1}, AX\rangle| > |\langle A\delta_{s}, AX\rangle|$. This shows that s^* is necessarily in E_1 as $E_1 \cap (E_2 \cup E_3) = \emptyset$ i.e. there exists an i_0 such that $\|s^* - t_{i_0}\|_2 < \xi$. □

D. Appendix: Towards improved precision in SMLM, recovering signals with a large number of spikes

We compare all three OP-COMP + PGD, COMP + GD and Sliding COMP algorithms on a batch with $K = 50$ spikes. Since our method scales well with a larger number of spikes, our goal is to reduce the number of batches by increasing the number of spikes K in each batch. However, in the classical setting, we note that recovering spikes from their observation is limited by the quality of said observation.

The idea is to decrease the width of the observation kernel. For the MA-TIRF model, this can be achieved by decreasing the excitation wavelength λ_l. We choose to change $\lambda_l = 660$nm in the classical setting to $\lambda_l = 110$nm, the adapted grid is finer with 192×192 points on the same domain.

Even though these settings (low variance and fine grid) are uncommon and very hard to set up in a practical sense, we anticipate such advancements in acquisition methods and the potential improvements given by OP-COMP + PGD algorithm over Sliding COMP.

When comparing the estimated signal from OP-COMP + PGD, COMP + GD and Sliding COMP, we plot both ground truth and estimated signal, see Figure 18. We observe that the estimated signals are close to the ground truth with all three methods.

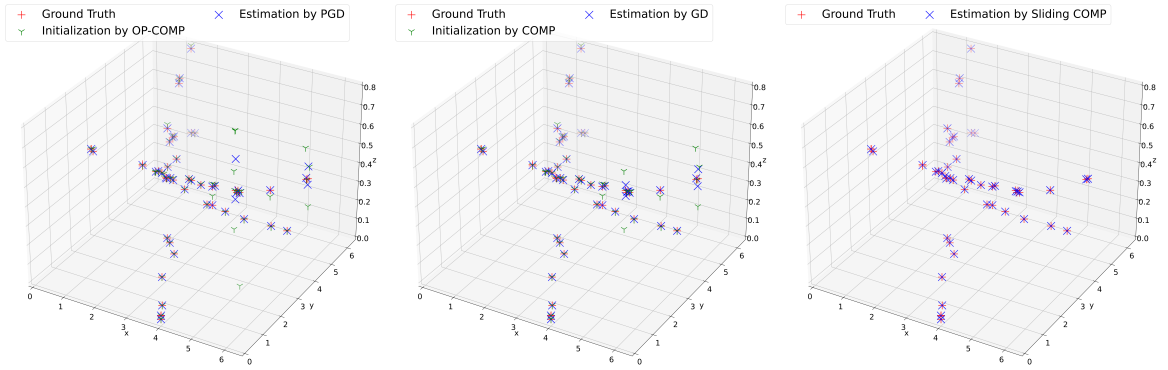

Figure 18: 3D Plots of a signal of size $K = 50$ and its estimation by OP-COMP (in green) and then PGD (in blue) (left), its estimation by COMP (in green) and then GD (in blue) (center) and its estimation by Sliding COMP (in blue) (right).

To compare these estimations, we use the same metrics we introduced previously and the norm of the residue of each estimation. We get the following table,

	OP-COMP + PGD	COMP + GD	Sliding COMP
Computation time	333 minutes	327 minutes	537 minutes
Norm of residue of estimation	3.58×10^{-4}	1.09×10^{-3}	2.46×10^{-2}
Jaccard Index	0.839	0.776	0.741
Recall	0.940	0.882	0.860
Precision	0.887	0.843	0.865
RMSE (x_1, x_2, x_3) in nm	$(3.07, 2.70, 1.50)$	$(3.17, 2.65, 2.47)$	$(4.36, 2.80, 2.07)$

Table 3: Table of comparison between OP-COMP + PGD, COMP + GD and Sliding COMP on computation time, norm of residue, Jaccard, Recall and Precision metrics and on RMSE of each dimensions.

We observe in Table 3 that all metrics of OP-COMP + PGD are better compared to Sliding COMP with a 40% improvement in computation time. We obtain faster a result closer to the ground truth in both RMSE and index terms. When comparing OP-COMP + PGD with COMP + GD, the over-parametrization with projection is more efficient as we have significant gains in all three metrics (Jaccard, Recall and Precision).

References

[1] T. Alamo, D. Limon, and P. Krupa. Restart fista with global linear convergence. In *2019 18th European Control Conference (ECC)*, pages 1969–1974, 2019.

[2] J.-F. Aujol, C. Dossal, H. Labarrière, and A. Rondepierre. FISTA restart using an automatic estimation of the growth parameter. working paper or preprint, May 2022.

[3] D. Axelrod. Cell-substrate contacts illuminated by total internal reflection fluorescence. *The Journal of cell biology*, 89(1):141–145, 1981.

[4] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[6] H. E. Bell. Gershgorin's theorem and the zeros of polynomials. *The American Mathematical Monthly*, 72(3):292–295, 1965.

[7] P.-J. Bénard, Y. Traonmilin, and J.-F. Aujol. Fast off-the-grid sparse recovery with over-parametrized projected gradient descent. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 2206–2210. IEEE, 2022.

[8] P.-J. Bénard, Y. Traonmilin, and J.-F. Aujol. Code for the experiments. https://github.com/pjbenard/opCOMP_PGD_microscopy, 2023. [Online].

[9] J.-F. Cai, T. Wang, and K. Wei. Spectral compressed sensing via projected gradient descent. *SIAM Journal on Optimization*, 28(3):2625–2653, 2018.

[10] E. Candes and J. Romberg. l1-magic: Recovery of sparse signals via convex programming. *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf*, 4(14):16, 2005.

[11] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[12] 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.

[13] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[14] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[15] L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1-2):487–532, 2022.

[16] M. F. Da Costa and Y. Chi. Local geometry of nonconvex spike deconvolution from low-pass measurements. *IEEE Journal on Selected Areas in Information Theory*, 2023.

[17] Y. De Castro, F. Gamboa, D. Henrion, and J.-B. Lasserre. Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630, 2016.

[18] G. M. de Galland, T. Feuillen, L. Vandendorpe, and L. Jacques. Sparse factorization-based detection of off-the-grid moving targets using fmcw radars. In *ICASSP 2021*, pages 4575–4579. IEEE, 2021.

[19] V. Debarnot and P. Weiss. Blind inverse problems with isolated spikes. *Information and Inference: A Journal of the IMA*, 12(1):26–71, 2023.

[20] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies. The Sliding Frank-Wolfe Algorithm and its Application to Super-Resolution Microscopy. *Inverse Problems*, 2019.

[21] C. Elvira, R. Gribonval, C. Soussen, and C. Herzet. OMP and continuous dictionaries: Is k-step recovery possible? In *ICASSP*, pages 5546–5550. IEEE, 2019.

[22] C. Elvira, R. Gribonval, C. Soussen, and C. Herzet. When does OMP achieve exact recovery with continuous dictionaries? *Applied and Comp. Harmonic Analysis*, 51:39, 2021.

[23] A. Flinth, F. de Gournay, and P. Weiss. Grid is good: Adaptive refinement algorithms for off-the-grid total variation minimization, 2023.

[24] S. Foucart and H. Rauhut. *An invitation to compressive sensing*. Springer, 2013.

[25] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[26] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Compressive Statistical Learning with Random Feature Moments. *Mathematical Statistics and Learning*, 3(2):113–164, 2021.

[27] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Statistical learning guarantees for compressive clustering and compressive mixture modeling. *Math. Stat. Learn., In press*, 3(2):165–257, 2021.

[28] B. Huang, W. Wang, M. Bates, and X. Zhuang. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science*, 2008.

[29] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez. Sketching for large-scale learning of mixture models. *Information and Inference: A Journal of the IMA*, 7(3):447–508, 2018.

[30] N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval. Compressive k-means. In *2017 IEEE ICASSP*, pages 6369–6373, 2017.

[31] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020.

[32] B. O'donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15:715–732, 2015.

[33] M. Q. Pham, L. Duval, C. Chaux, and J.-C. Pesquet. A primal-dual proximal algorithm for sparse template-based adaptive filtering: Application to seismic multiple removal. *IEEE Transactions on Signal Processing*, 62(16):4256–4269, 2014.

[34] S. Rama Prasanna Pavani, M. A Thompson, J. S Biteen, S. Lord, N. Liu, R. Twieg, R. Piestun, and W. Moerner. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proceedings of the National Academy of Sciences of the United States of America*, 106:2995–9, 03 2009.

[35] D. Sage, T.-A. Pham, H. Babcock, T. Lukes, T. Pengo, J. Chao, R. Velmurugan, A. Herbert, A. Agrawal, S. Colabrese, et al. Super-resolution fight club: assessment of 2d and 3d single-molecule localization microscopy software. *Nature methods*, 16(5):387–395, 2019.

[36] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy. Gridless 3d recovery of image sources from room impulse responses. *IEEE Signal Processing Letters*, 29:2427–2431, 2022.

[37] Y. Traonmilin and J.-F. Aujol. The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem. *Inverse Problems*, 2020.

[38] Y. Traonmilin, J.-F. Aujol, P.-J. Bénard, and A. Leclaire. On strong basins of attractions for non-convex sparse spike estimation: upper and lower bounds. working paper or preprint, Mar. 2023.

[39] Y. Traonmilin, J.-F. Aujol, and A. Leclaire. Projected gradient descent for non-convex sparse spike estimation. *IEEE Signal Processing Letters*, 27:1110–1114, 2020.

[40] Y. Traonmilin, J.-F. Aujol, and A. Leclaire. The basins of attraction of the global minimizers of non-convex inverse problems with low-dimensional models in infinite dimension. *Information and Inference: A Journal of the IMA*, 12(1):113–156, 2023.

[41] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.