Projected gradient descent for non-convex sparse spike estimation

Yann Traonmilin\textsuperscript{1,2}, Jean-François Aujol\textsuperscript{2} and Arthur Leclaire\textsuperscript{2}

Abstract—We propose a new algorithm for sparse spike estimation from Fourier measurements. Based on theoretical results on non-convex optimization techniques for off-the-grid sparse spike estimation, we present a projected gradient descent algorithm coupled with a back-projection initialization procedure. Our algorithm permits to estimate the positions of large numbers of Diracs in 2D from random Fourier measurements. We present, along with the algorithm, theoretical qualitative insights explaining the success of our algorithm. This opens a new direction for practical off-the-grid spike estimation with theoretical guarantees in imaging applications.

Index Terms—spike super-resolution, non-convex optimization, projected gradient descent

I. INTRODUCTION

In the space $\mathcal{M} = \mathcal{M}(\mathbb{R}^d)$ (respectively $\mathcal{M} = \mathcal{M}(\mathbb{T}^d)$) of finite signed measures over $\mathbb{R}^d$ (respectively the $d$-dimensional torus $\mathbb{T}^d$), we aim at recovering a superposition of impulsive sources $x_0 = \sum_{t_i} a_i \delta_{t_i} \in \mathcal{M}$ from the measurements

$$y = Ax_0 + e,$$

where $\delta_{t_i}$ is the Dirac measure at position $t_i$, the operator $A$ is a linear observation operator from $\mathcal{M}$ to $\mathbb{C}^m$, $y \in \mathbb{C}^m$ are the $m$ noisy measurements and $e$ is a finite energy observation noise. This inverse problem (called spike super-resolution [6], [2], [20], [10], [12]) models many imaging problems found in geophysics, microscopy, astronomy or even (compressive) machine learning [18]. Under a separation assumption on the positions of the Diracs, i.e when $x_0$ is in a set $\Sigma_k$, of sums of $k$ $\epsilon$-separated Diracs with bounded support, it has been shown that $x_0$ can be estimated by solving a non-convex problem as long as $A$ is an appropriately designed measurement process. This ideal non-convex minimization is:

$$x^* \in \arg \min_{x \in \Sigma_k} \| Ax - y \|^2_2,$$  \hspace{1cm} (1)

Recovery guarantees for this problem are of the form

$$\| x^* - x_0 \|_K \leq C \| e \|_2,$$  \hspace{1cm} (2)

where $\| \cdot \|_K$ is a kernel norm on $\mathcal{M}$ that measures distances in $\mathcal{M}$ at a given high resolution described by the kernel (in most of the literature $K$ is either a Fejér [6] or Gaussian kernel and $\| \sum_{t_i} a_i \delta_{t_i} \|^2 = \sum_{i,j} a_i a_j K(t_i - t_j)$ [17], [15]). Mathematically, (3) is guaranteed if the measurement operator $A$ has a restricted isometry property on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$ (the set of differences of elements of $\Sigma_{k,\epsilon}$) [15]. This property is typically obtained when the number of measurements is sufficient. For example, recovery guarantees are obtained when $m \geq 6d/\epsilon^2$ for regular low frequency Fourier measurements on the torus [6] and when $m \geq 6d(\log(k))^2 \log(kd/\epsilon)$ for random Fourier measurements on $\mathbb{R}^d$ [15].

Recent advances in this field proposed a convex relaxation of the problem in the space of measures [6], [10]. While giving theoretical recovery guarantees, these methods are not convex with respect to the parameters due to a polynomial root finding step. Moreover, they rely on a SDP relaxation of a dual formulation, thus squaring the size of the problem. Sliding Frank-Wolfe/conditional gradient methods were also proposed but suffer from increased complexity when the number of spikes increases [4], [11]. Methods based on structured low rank Toeplitz approximation are also difficult to extend to higher dimensions [9]. Other methods based on greedy heuristics (CL-OMP for compressive $k$-means [18]) have been proposed (very close in practice to the sliding Frank-Wolfe method) but they still lack theoretical justifications in this context even if some first theoretical results are emerging for some particular measurement methods [13].

In this paper, we propose a practical method to solve the non-convex minimization problem (2) for a large number of Diracs in imaging problems. Of course, at first sight, it is not possible to solve this problem efficiently. However, we justify qualitatively why our method succeeds. This justification relies on the separation assumption on $x_0$ and the assumption that there are enough measurements of $x_0$. We also give numerical experiments validating the method. One of the main practical advantages of our method is its ability to perform off-the-grid spike estimation from random Fourier measurements with a good scaling with respect to the number of spikes. With this proof of concept, we can estimate many spikes in two dimensions from compressive measurements, yielding potential applications in fields such as astronomy or microscopy where the sum of spikes model is relevant.

Our method, following insights from the literature on non-convex optimization for low-dimensional models [22], [19], [5], [7], [8], relies on two steps:

- Overparametrized initialization by hard-thresholded back-projection: we propose an initialization step that permits a good first estimation of the positions of the Diracs.
- Projected gradient descent algorithm in the parameter space: the idea of projected gradient descent for low-dimensional model recovery has shown its benefits in the finite dimensional case [3], [14]. We adapt this idea to sparse spike recovery. From [21], the global minimizer of (2) can be recovered by unconstrained gradient descent as long as the initialization lies in an explicit basin of attraction of the
global minimizer. It was also shown that projecting on the
separation constraint improves the control on the Hessian
of the function we minimize. However, no practical way to
perform a projection and no implementation were proposed.

Contributions. After recalling the context of non-convex
sparse spike estimation, we propose a new practical projected
gradient descent algorithm for sparse spike estimation. The
simple gradient descent is already used as a refinement step in
sparse spike estimation, we propose a new practical projected
initialization approximates well the Diracs positions;

In Section II, we describe our practical projected gradient
descent algorithm and its implementation details;

In Section III, we give a grid based initialization using hard-
thresholded back-projection. A qualitative analysis shows
that when the number of measurements is large enough, our
initialization approximates well the Diracs positions;

In Section IV, we show the practical benefit of the projection
in the descent algorithm and its application to the estimation
of large number of Diracs in 2 dimensions.

II. THEORETICAL BACKGROUND AND ALGORITHM
DESCRIPTION

A. Measurements and parameter space

The operator $A$ is a linear operator modeling $m$
measurements in $\mathbb{C}^m$ (Im$A \subset \mathbb{C}^m$) on the space of measures on a
domain $E$ (either $E = \mathbb{R}^d$ or $E = \mathbb{T}^d$) defined by:

\[ \forall l, 1, \ldots, m, \quad (Ax)_l = \int_E a_l(t) \, dx(t), \]  

(4)

where $(a_l)_{l=1}^m$ is a collection of (weighted) Fourier
measurements: $a_l(t) = c_l e^{-i \omega_l t}$ for some chosen
frequencies $\omega_l \in \mathbb{R}$ and frequency dependent weights $c_l \in \mathbb{R}$ (the $c_l$
are mostly of theoretical interest for the study of recovery
 guarantees [15] but can be set to 1 in practice). The model set of $\epsilon$-separated Diracs with $\epsilon > 0$ is:

\[ \Sigma_{k, \epsilon} := \left\{ \sum_{r=1}^k a_r \delta_{t_r} : a \in \mathbb{R}^k, t_r \in B_2(R), \forall r \neq l, \|t_r - t_l\|_2 \geq \epsilon \right\}, \]  

(5)

where $B_2(R) = \{ t \in \mathbb{R}^d : \|t\|_2 \leq R \}$ is the $l^2$ ball of radius $R$
centered in $0$ in $\mathbb{R}^d$. We consider the following parametrization
of $\Sigma_{k, \epsilon}$: for any $\theta = (a_1, \ldots, a_k, t_1, \ldots, t_k) \in \mathbb{R}^{k(d+1)}$, we define

\[ \phi(\theta) = \sum_{i=1}^k a_i \delta_{t_i}, \]  

and we set

\[ \Theta_{k, \epsilon} := \phi^{-1}(\Sigma_{k, \epsilon}), \]  

(6)

the reciprocal image of $\Sigma_{k, \epsilon}$ by $\phi$. Note that any parametrization
of elements of $\Sigma_{k, \epsilon}$ is invariant by permutation of the
positions. This is not a problem in practice for the convergence
of descent algorithms. We define the parametrized functional

\[ g(\theta) := \|A\phi(\theta) - y\|_2^2 \]  

(7)

and consider the problem

\[ \theta^* \in \arg \min_{\theta \in \Theta_{k, \epsilon}} g(\theta). \]  

(8)

Since the $a_i$ are smooth, $g$ is a smooth function. Note that
performing the minimization (8) allows to recover the minima
of the ideal minimization (2), yielding stable recovery guar-
antees under a restricted isometry assumption on $A$ which is
verified when $m \geq O(k^2d(\log(k))^2 \log(kd/\epsilon))$ for adequately
chosen Gaussian random Fourier measurements (on $\mathbb{R}^d$) and
$m \geq O(\frac{d}{\epsilon})$ for regular Fourier measurements on $\mathbb{T}^d$. In [21],
it has been shown that the simple gradient descent converges
(without projection) to the global minimum of $g$ as long as the
initialization falls into an explicit basin of attraction of this
global minimum. It was also shown that the projection on the
separation constraint improves the control on the Hessian on
$g$ and subsequently the convergence of the descent algorithm.

B. Projected gradient descent in the parameter space

For a user-defined initial number of Diracs $k_{in}$, we consider
the following iterations:

\[ \theta_{n+1} = P_{\Theta_{k_{in}, \epsilon}}(\theta_n - \tau_n \nabla g(\theta_n)) \]  

(9)

where $P_{\Theta_{k_{in}, \epsilon}}$ is a projection on the separation constraint,
(notice that there may be several solutions in $\Theta_{k_{in}, \epsilon}$) and $\tau_n$
is the step size at iteration $n$. The projection $P_{\Theta_{k_{in}, \epsilon}}(\theta)$ could
be defined naturally as a solution of the minimization problem
$\inf_{\theta \in \Theta_{k_{in}, \epsilon}} \|\phi(\theta) - \phi(\theta)\|_K$. Unfortunately this optimization is
not convex. We propose instead a heuristic (see Algorithm 1)
for $P_{\Theta_{k_{in}, \epsilon}}$ that consists in merging Diracs that are not $\epsilon$-
separated.

Input: List $\Theta = (a_i, t_i)_i$ of amplitudes and positions
ordered by decreasing absolute amplitudes

\[ \text{for } i \geq 1 \text{ do} \]

\[ \text{if } \|t_i - t_j\| < \epsilon \text{ then} \]

\[ a_i = a_i + a_j; \]

\[ t_i = \frac{|a_i| t_i + |a_j| t_j}{|a_i| + |a_j|}; \]

\[ \text{Remove } (a_j, t_j) \text{ from } \Theta \]

Output: List $\Theta = (a_i, t_i)_i$ of amplitudes and positions
of projected spikes

Algorithm 1: Heuristic for the projection $P_{\Theta_{k_{in}, \epsilon}}$

Since we take the barycenter of the positions, if a set of
Diracs that are at a distance at most $\epsilon$ of a true position in $x_0$
is merged, the merged result will be within this distance. We
use the ordering by decreasing amplitude to avoid that a low
amplitude spike pulls a meaningful high-amplitude spike away
from a true position. After this projection step, we pursue the
descent with the remaining number of Diracs. Note that we
overparametrize with $k_{in}$ the number of Diracs in the descent
to ensure the recovery of all positions in $x_0$ (see also the
next section). In practice, we implement the projected gradient
descent as follows.

As suggested in [21], to avoid balancing problems between
amplitudes and positions, we alternate descent steps between
amplitudes and positions.
To find the step size $\tau_n$, we perform a line search to minimize the value of the function $g$.

We start to project after a few iterations (20 iterations in our experiments) of the gradient descent so that spikes have already started clustering together towards the solution.

From [21], this algorithm will converge as soon as the initialization falls into a basin of attraction of global minimum of $g$. The basins of attraction get larger as the number of measurements increases (up to a fundamental limit depending on the separation $\epsilon$ and the amplitudes in $x_0$).

III. OVERPARAMETERIZED INITIALIZATION BY HARD-THRESHOLDED BACK-PROJECTION

The idea of using back-projection of measurements was used for non-convex optimization in the context of phase recovery [22] and blind deconvolution [5] with so-called spectral initialization techniques, where a leading eigenvector of a matrix constructed with back-projections is used as initialization. As we measure the signal $x_0$ at some frequencies $\omega_i$, a way to recover an estimation of the signal is to back-project the whole irregular spectrum on a grid $\Gamma$ that samples $B_2(R)$ at a given precision $\epsilon_g$ (to be chosen later). For Fourier measurements $y$ at frequencies $(\omega_i)_{i=1,m}$, we calculate $z_\Gamma = B_\Gamma y$ where $B_\Gamma$ is the linear operator back-projecting the Fourier measurements on a grid in the spatial domain $\mathbb{R}^d$:

$$z_\Gamma = B_\Gamma y := \sum_{s_i \in \Gamma} z_{\Gamma,i} \delta_{s_i},$$

where the $s_i \in \Gamma$ are the grid positions and

$$z_{\Gamma,i} = \sum_l y(l) e^{i(\omega_l,s_i)}$$

(10) (11)

for some appropriate weights $d_l$ to be chosen in the next section. We can show in the noiseless case with Lemmas III.1 and III.2 that when the number of measurements increases and the grid for initialization gets finer, the original positions of Diracs get better approximated. All “on-the-grid” methods such as least-squares estimation or the LASSO are ways to back-project measurements. As we aim at a fast algorithm, and since the energy of Diracs is well localized by the initialization we then perform overparametrized hard thresholding of the back-projection. We propose the initialization $\theta_{init}$ defined by

$$\phi(\theta_{init}) := x_{init} = H_{\epsilon_{init}}(B_\Gamma y)$$

(12)

where for $|z_{\Gamma,j_1}| \geq |z_{\Gamma,j_2}| \geq \ldots \geq |z_{\Gamma,j_n}|$, we have $H_{\epsilon_{init}}(z_\Gamma) = \sum_{i=1}^n z_{\Gamma,j_i} \delta_{s_i}.$

Ideal back-projection and sampling: In the context of Diracs recovery our initialization by hard-thresholded back-projection is a sampling of an ideal back-projection. Let $B$ the operator from $\mathbb{C}^m$ to $\mathcal{M}$ defined for $z = By$ by

$$z(t) = \sum_{l=1}^m y(l) e^{i(\omega_l,t)}$$

We call $z$ an ideal back-projection because $z_{\Gamma} = S_{\Gamma} z$ where $S_{\Gamma}$ is the sampling on a grid $\Gamma$: for a measure $x$ with a continuous density $\chi$ (i.e. $dx(t) = \chi(t) dt$), we define the sampling operation $S_{\Gamma}(x) = \sum_{t_i \in \Gamma} \chi(t_i) \delta_{t_i}$. Also, the measure $z = By$ has a smooth density as it is a finite sum of complex exponentials.

We first show (proofs are in the supplementary material) that for the right choice of weights $d_l$, the energy of $z = B A x$ (where $x = \sum_{i=1}^k a_i \delta_{t_i}$) is localized around the positions $t_i$ in both the regular Fourier sampling on the torus case, and the random Fourier sampling on $\mathbb{R}^d$ case. We show the following results for the Fejér and Gaussian kernels as they are typically used in the literature for deterministic [6] and random [15] Fourier sampling.

$$z(t) = \sum_{i=1}^k a_i K_f(t - t_i).$$

(14)

This immediate lemma states that on the torus, measuring low frequencies is equivalent to measuring a low-pass filtered signal in the time domain. For example, the low-pass Fejér filter is shown Figure 1. Also, this result holds for any kernel with spectrum supported on the $\omega_l$. Sampling frequencies regularly with maximum frequency $\omega_{\text{max}} \geq O(1/\epsilon)$ guarantees recovery with convex relaxation methods. For random Fourier sampling, we look at the expected value of $z$ and control its variance with respect to the distribution of the $\omega_l$.

Lemma III.2. On $\mathcal{M}(\mathbb{R}^d)$, we choose $A$ such that $(\omega_l)_{l=1,m}$ are $m$ i.i.d random variables with a Gaussian distribution with density $G(\omega) = \frac{e^{-\|\omega\|_2^2}}{(2\pi)^d}$. Let $K_\sigma(t) = e^{-\frac{|t|^2}{2\sigma^2}}$.

In (13), take $d_l = 1/(\epsilon m \epsilon)$ then

$$E(z(t)) = \sum_{i=1}^k a_i K_\sigma(t - t_i).$$

(15)

$$E((z(t) - E(z(t)))^2) = \frac{1}{m} E(z(t))^2 + \frac{1}{m} \|x_0\|^2_{K_\sigma}$$

(16)

where $\|x_0\|^2_{K_\sigma}$ is the norm associated with the kernel $K_\sigma$.

Similarly to the regular sampling, the energy of the expected value of $z$ is concentrated around the positions $t_i$ (see Figure 1). In [15] the frequency distribution scales as the inverse of the kernel precision, i.e. the kernel parameter $\sigma$ of the kernel $h(t) = e^{-\|t\|^2/(2\sigma^2)}$ is chosen as $O(1/\epsilon)$. The control of the variance shows that when the number of measurements increases, the back-projection of these measurements to the space of measures are closer to the ideal initialization which is the expected value of $z$. In practice we set the number of...
measurements using a rule \( m = \mu kd \) with a user defined multiplicative parameter \( \mu \) that does not depend on the dimension of the problem. The quality of the initialization is thus directly linked to \( \mu \). Finally the following lemma makes sure that as the grid gets finer we recover all the energy of the ideal back-projection that lies within the domain sampled by the grid.

**Lemma III.3.** Let \( z_d = (z_\Gamma \lambda_i)_{1 \leq i \leq 1} \) where \( \Gamma \) is a grid with step size \( \epsilon_g \). Then \( \| e_g \|_{L^2} \approx \epsilon_g \rightarrow 0 \| z \|_{L^2}(\mathcal{B}_d(R))^{-1} \).

We considered the noiseless case. The noisy case just adds a noise term with energy controlled by the noise energy level \( \| e \|_2 \) because \( B_\Gamma \) is a Fourier back-projection.

### IV. Numerical Experiments

We first run the algorithm on few Diracs in 2d to illustrate the added benefit of the projection. We then show results with many Diracs in 2d to show the computational feasibility of projected gradient descent for imaging applications. We perform the experiments in the noiseless case with a stopping criterion based on the value of function \( g \) and leave the study of the impact of the noise for future work. The Matlab code used to generate these experiments is available at [1].

**Illustration with few Diracs:** As a first proof of concept we run the algorithm with the recovery of 5 Diracs in 2 dimensions from \( m = 120 \) Gaussian random measurements. The trajectories of 500 iterations of the gradient descent and projected gradient descent are represented in Figure 2. We observe that while the gradient descent with overparametrized initialization might converge with a large number of iterations, the projection step greatly accelerates the convergence.

**Estimation of 100 Diracs in 2d:** We recover 100 Diracs, with a separation 0.01 on the square \([0, 1] \times [0, 1]\) from \( m = 2000 \) compressive Gaussian measurements (we would need \( \approx 10000 \) regular measurements to obtain a separation 0.01). In practice, the grid \( \Gamma \) must be fine enough to overparametrize the number of Diracs with a good sampling of the ideal back-projection. If \( \epsilon_g \) is too small, the number of initial Diracs needed to sample the energy gets larger, leading to an increased cost in the first iterations of the gradient descent. In this example we use \( \epsilon_g = \epsilon \) and use \( k_{ini} = 4k \). We observe in Figure 3 that with these parameters all the Diracs positions are well estimated after 184 iterations (convergence criterion met) of our algorithm. Similarly to our first example, we observe that spikes that are not separated in the back-projection on the grid are well estimated by our algorithm.

**A. Complexity**

The cost of our algorithm is the sum of the cost of the initialization and the cost of the projected gradient descent. The back-projection on the grid scales as \( O((1/\epsilon_g)^d) \) (irregular Fourier transform on a grid), but it is done only once and fast transform techniques could be investigated. With our strategy, this cost seems unavoidable as we want to localize Diracs off-the-grid with a separation \( \epsilon \) (doing the same on the grid would have this exponential scaling with respect to the dimension and the separation). Our algorithm stays tractable when the dimension \( d \) of the domain of the positions of the Diracs is not too large, which is the case in 2d or 3d imaging applications. This cost \( O((1/\epsilon_g)^d) \) would also be impossible to avoid in an eventual visualization of the full recovered image over a precise grid on \( \mathbb{R}^d \). Note also that in practice our proposed initialization could be replaced by any satisfying overparametrized initialization, i.e. any state-of-the-art on-the-grid estimation technique could benefit from an added projected gradient descent.

The cost of the projected gradient descent is \( O(n_{ini} C_\gamma) \) where \( C_\gamma \) is the cost the calculation of the gradient. This cost is of the order of the calculation of the \( m \) Fourier measurements for the current number of Diracs in the descent (close to \( k \) after a few iterations). For the experiment with 100 spikes, our algorithm successfully completed the estimation of the spikes in 5.9 minutes (Matlab implementation) while the CL-OMP algorithm proposed by [16] (their Matlab implementation) took 30.5 minutes to complete the successfull estimation of all the spikes on a laptop for office purpose.

### V. Conclusion

We gave a practical algorithm to perform off-the-grid sparse spike estimation. This proof-of-concept shows that it is possible to estimate efficiently a large number of Diracs in imaging applications with some strong theoretical insights of success guarantees. Future research directions are:

- Full theoretical convergence proof of the algorithm with sufficient conditions on the number of measurements. The main question is to know if it is possible to have a convergence guarantee without the computational cost \( O((1/\epsilon_g)^d) \)
- Investigate other methods for reducing the number of parameters after the back-projection on a grid and accelerate the descent (quasi-Newton schemes).
- Study the algorithm stability to noise and modeling error with respect to the number of measurements.
REFERENCES

[1] Matlab code for "projected gradient descent for non-convex sparse spike estimation". yanntraonmilin.wordpress.com/code.

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

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

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

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

[6] 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.

[7] 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.

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

[9] L. Condat and A. Hirabayashi. Cadzow denoising upgraded: A new projection method for the recovery of dirac pulses from noisy linear measurements. 2015.

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

[11] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies. The sliding frank–wolfe algorithm and its application to super-resolution microscopy. Inverse Problems, 36(1):014001, 2019.

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

[13] C. Elvira, R. Gribonval, C. Soussen, and C. Herzet. Omp and continuous dictionaries: Is k-step recovery possible? In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5546–5550. IEEE, 2019.

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

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

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

[17] 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.

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

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

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

[21] Y. Traonmilin and J.-F. Aujol. The basins of attraction of the global minimizers of the non-convex sparse spikes estimation problem. Inverse Problems, 2019.

[22] I. Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. IEEE Transactions on Information Theory, 2018.
Supplementary material for: Projected gradient descent for non-convex sparse spike estimation

Yann Traonmilin$^{1,2}$, Jean-François Aujol$^2$ and Arthur Leclaire$^2$

$^1$CNRS, $^2$Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400 Talence, France.
I. Definitions and properties

For \( x_0 = \sum_{i=1}^{d} a_i \delta_{t_i} \), we have \( y_l = (Ax_0)_l = \sum_{i=1}^{d} a_i e^{-j\langle \omega_l, t_i \rangle} \).
The Fourier transform of a function \( f \) supported on \( \mathbb{R}^d \) is

\[
\hat{f}(\omega) = \int_{\mathbb{R}^d} e^{-j\langle \omega, t \rangle} f(t) dt
\]  

(1)

Let \( E = \mathbb{R}^d \). The Gaussian function \( G \) defined by \( G(u) = \frac{d}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|u\|^2} \), Then, \( \hat{G}(v) = e^{-\frac{1}{2}\|v\|^2} \).

II. Proofs

Proof of Lemma III.1. Using the fact that the \( \omega_l \) sample exactly the support of the Féjer Kernel and that, by definition,

\[
K_f(t) = \sum_{l=1, m} \hat{K}(\omega_l)e^j\langle \omega_l, t \rangle,
\]  

(2)

we have

\[
z(t) = \sum_{l=1, m} d_l(\sum_{i=1, k} a_i c_i e^{-j\langle \omega_l, t_i \rangle})e^j\langle \omega_l, t \rangle
\]

\[
= \sum_{i=1, k} a_i \sum_{l=1, m} d_l e^j\langle \omega_l, t - t_i \rangle
\]

\[
= \sum_{i=1, k} a_i K_f(t - t_i).
\]  

(3)

\[
\square
\]

Proof of Lemma III.2. Using the properties of the Fourier transform \( \hat{G} \) of the Gaussian function \( G \) and \( E \) being the expectation with respect to the distribution of the frequencies \( \omega_l \), we have

\[
E(z(t)) = \sum_{i=1, k} a_i \sum_{l=1, m} d_l c_i E(e^j\langle \omega_l, t - t_i \rangle)
\]

\[
= \sum_{i=1, k} a_i \sum_{l=1, m} d_l \hat{G}(t_i - t)
\]

\[
= \sum_{i=1, k} a_i K_f(t_i - t).
\]  

(4)

Using the formula \( E(|z(t) - E(z(t))|^2) = E(|z(t)|^2) - |E(z(t))|^2 \) and the previous Lemma, we just need to calculate

\[
E(|z(t)|^2) = E\left( \left| \sum_{i=1, k} a_i \sum_{l=1, m} d_l c_i e^j\langle \omega_l, t - t_i \rangle \right|^2 \right)
\]

\[
= E\left( \sum_{i=1, k} a_i \sum_{l=1, m} d_l c_i e^j\langle \omega_l, t - t_i \rangle \sum_{i_2=1, k} a_{i_2} \sum_{l=1, m} d_l c_{i_2} e^{-j\langle \omega_l, t - t_{i_2} \rangle} \right)
\]

\[
= \sum_{i=1, k} \sum_{i_2=1, k} a_i a_{i_2} \sum_{l=1, m} d_l d_{i_2} \sum_{i_2=1, k} a_{i_2} \sum_{l=1, m} d_l c_{i_2} E\left( e^j\langle \omega_l, t - t_{i_2} \rangle e^{-j\langle \omega_l, t - t_{i_2} \rangle} \right)
\]  

(5)

For \( r = l \),

\[
E\left( e^j\langle \omega_l, t - t_{i_1} \rangle e^{-j\langle \omega_l, t - t_{i_2} \rangle} \right) = \hat{G}(t_{i_2} - t_{i_1})
\]  

(6)

For \( r \neq l \),

\[
E\left( e^j\langle \omega_l, t - t_{i_1} \rangle e^{-j\langle \omega_l, t - t_{i_2} \rangle} \right) = E\left( e^j\langle \omega_l, t - t_{i_1} \rangle \right) E\left( e^{-j\langle \omega_l, t - t_{i_2} \rangle} \right)
\]

\[
= \hat{G}(t - t_{i_1}) \hat{G}(t - t_{i_2})
\]  

(7)
Hence
\[
\mathbb{E}(|z(t)|^2) = \sum_{i_1=1,k} a_{i_1} a_{i_2} \left( \sum_{l=1, m} |d_l c_l|^2 \tilde{G}(t_{i_2} - t_{i_1}) + \sum_{l \neq r} d_l c_l \tilde{c}_r \tilde{G}(t - t_{i_1}) \tilde{G}(t - t_{i_2}) \right)
\]
\[
= \sum_{i_1=1,k} a_{i_1} a_{i_2} \left( \sum_{l=1, m} \frac{1}{m} K_g(t_{i_2} - t_{i_1}) + \sum_{l \neq r} \frac{1}{m} K_g(t - t_{i_1}) K_g(t - t_{i_2}) \right)
\]
\[
= \left(1 - \frac{1}{m}\right) \left( \sum_{i=1,k} a_i K_g(t - t_i) \right)^2 + \frac{1}{m} \sum_{i_1=1,k} \sum_{i_2=1,k} a_{i_1} a_{i_2} \frac{1}{m} K_g(t_{i_2} - t_{i_1})
\]
\[
= \left(1 - \frac{1}{m}\right) |E(z(t))|^2 + \frac{1}{m} \|x_0\|^2_{K_g}
\]  

(8)

Going back to the variance,
\[
\mathbb{E}(|z(t) - E(z(t))|^2) = \frac{1}{m} |E(z(t))|^2 + \frac{1}{m} \|x_0\|^2_{K_g}
\]  

(9)

Proof of Lemma III.3..
\[
\|\sqrt{\epsilon_g} d_zd\|_2 = \epsilon_g \sum_{i \in \Gamma} |z(t_i)|^2
\]  

(10)

As $z$ is continuous on $B_2(R)$, it is integrable over $B_2(R)$ and this sum is approximating the Riemann integral of $t \to |z(t)|^2$ over $B_2(R)$ as the grid step size $\epsilon_g$ of $\Gamma$ converges to 0.