PROJECTED GRADIENT DESCENT FOR NON-CONVEX SPARSE SPIKE ESTIMATION

Yann Traonmilin, Jean-François Aujol, Arthur Leclaire

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.

ABSTRACT

We propose an algorithm to perform 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 simple projected descent algorithm coupled with an initialization procedure. Our algorithm permits to estimate the positions of large numbers of Diracs in 2d from random Fourier measurements. This opens the way for practical estimation of such signals for imaging applications as the algorithm scales well with respect to the dimensions of the problem. We present, along with the algorithm, theoretical qualitative insights explaining the success of our algorithm.

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

1. INTRODUCTION

In the space $\mathcal{M}(\mathbb{R}^d)$ (respectively $\mathcal{M}(\mathbb{T}^d)$) of finite signed measures over $\mathbb{R}^d$ (respectively the d-dimensional Torus $\mathbb{T}^d$), we aim at recovering $x_0 = \sum_{i=1,k} a_i \delta_{t_i}$ from the measurements

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

where δ_{t_i} is the Dirac measure at position t_i, the operator A is a linear observation operator, $y \in \mathbb{C}^m$ are the m noisy measurements and e is a finite energy observation noise. This inverse problem (called spike super-resolution [1, 2, 3, 4, 5]) models many problems found in geophysics, microscopy, astronomy or even (compressive) machine learning [6]. For sets $\Sigma_{k,\epsilon}$ of sums of k ϵ-separated Diracs with bounded support, it has been shown that x_0 can be well estimated by solving a non-convex problem as long as $x_0 \in \Sigma_{k,\epsilon}$ and A has a restricted isometry property on $\Sigma_{k,\epsilon} - \Sigma_{k,\epsilon}$ [7]. The ideal non-convex minimization we need to solve is:

$$x^* \in \underset{x \in \Sigma_{k,\epsilon}}{\operatorname{argmin}} \|Ax - y\|_2. \tag{2}$$

Recovery guarantees of this problem are of the form

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

where $\|\cdot\|_K$ is a kernel norm on $\mathcal{M}$ (in most of the literature K is either a Féjer or Gaussian kernel and $\|\sum_i a_i \delta_{t_i}\|_K^2 = \sum_{i,j} a_i a_j K(t_i - t_j)$) that measures distances in $\mathcal{M}$ at a given resolution described by the kernel. Such guarantees are obtained when $m \geq O(\frac{1}{\epsilon^d})$ for regular low frequency Fourier measurements [1] and when $m \geq O(k^2 d(log(k))^2 log(kd/\epsilon)$ for random Fourier measurements [7].

First advances in this field proposed a convex relaxation of the problem in the space of measures. 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 squaring the size of the problem (which becomes problematic as the dimension d increases). Other methods based on greedy heuristics (CL-OMP for compressive k-means [6]) have been proposed. Nevertheless, they still lack theoretical justifications in this context while having good scaling properties with respect to the number of parameters to estimate (amplitudes and position) even if some first theoretical results are emerging for some particular measurement methods [8].

In this paper, we propose a practical method to solve the non-convex minimization problem (2) for a large number of Diracs. Of course, at first sight, it is not possible to solve this problem efficiently. However, we describe qualitatively why our method succeeds. This justification relies on the separation assumption on x_0 and the assumption that there is 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 parameters to estimate. With this proof of concept, we are able to resolve many spikes in 2 dimensions, yielding to 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 [9, 10, 11, 12], relies on two steps :

- Overparametrized spectral initialization: we propose a

spectral initialization step for spike estimation that permits a good first estimation of the positions of the Diracs.

- Projected descent algorithm in the parameter space: from [13], 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.

Contributions. After recalling that the non-convex sparse spike estimation problem in the space of parameters is a smooth non convex constrained optimization problem,

- in Section 2, we describe precisely our projected descent algorithm and its implementation details;
- in Section 3, we propose an initialization scheme that is guaranteed to localize the Diracs at a given resolution;
- in Section 4, we showcase the advantages of the projection in the descent algorithm and its application to the estimation of many Diracs in 2 dimensions.

2. THEORETICAL BACKGROUND AND ALGORITHM DESCRIPTION

2.1. Measurements and parameter space

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

$$(Ax)_l = \int_E \alpha_l(t)\,\mathrm{d}x(t) \tag{4}$$

where $(\alpha_l)_l$ is a collection (weighted) Fourier measurements: $\alpha_l(t) = c_l e^{-j\langle \omega_l, t\rangle}$ for some chosen frequencies $\omega_l \in \mathbb{R}^d$ and frequency dependent weights $c_l \in \mathbb{R}$. The model set of ϵ-separated Diracs with $\epsilon > 0$ is :

$$\Sigma_{k,\epsilon} := \left\{ \sum_{r=1,k} a_r \delta_{t_r} \,:\, a \in \mathbb{R}^k, t_r \in \mathbb{R}^d, \right. \\ \left. \forall r \neq l, \|t_r - t_l\|_2 \geq \epsilon, t_r \in \mathcal{B}_2(R) \right\}, \tag{5}$$

where $\mathcal{B}_2(R)$ is the L^2 ball of radius R centered in 0 in $\mathbb{R}^d$. We consider the following parametrization of $\Sigma_{k,\epsilon}$: $\sum_{i=1,k} a_i \delta_{t_i} = \phi(\theta)$ with $\theta = (a_1, .., a_k, t_1, .., t_k)$. We define

$$\Theta_{k,\epsilon} := \phi^{-1}(\Sigma_{k,\epsilon}), \tag{6}$$

and we consider the problem

$$\theta^* \in \arg\min_{\theta \in \Theta_{k,\epsilon}} g(\theta) = \arg\min_{\theta \in E} \|A\phi(\theta) - y\|_2. \tag{7}$$

Because the α_l are smooth, g is a smooth function. Note that performing the minimization (7) allows to recover the minima of the ideal minimization (2), yielding stable recovery guarantees under a restricted isometry assumption on A which is verified when $m \geq O(k^2 d(log(k))^2 log(kd/\epsilon))$ for adequately chosen Gaussian random Fourier measurements (on $\mathcal{M}(\mathbb{R}^d)$) and $m \geq O(\frac{1}{\epsilon^d})$ for regular Fourier measurements on $\mathcal{M}(\mathbb{T}^d)$. In [13], it has been shown that the simple gradient descent converges (without projection) to the global minimum of g as long as the initialization falls in 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.

2.2. Projected gradient descent in the parameter space

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

$$\theta_{n+1} = P_{\Theta_{k_{in},\epsilon}}(\theta_n - \tau \nabla g(\theta_n)) \tag{8}$$

where $P_{\Theta_{k_{in},\epsilon}}$ is a projection on the separation constraint, e.g. $P_{\Theta_{k_{in},\epsilon}}(\theta)$ could be defined naturally as a solution of the minimization problem $\inf_{\tilde{\theta} \in \Theta_{k_{in},\epsilon}} \|\phi(\tilde{\theta}) - \phi(\theta)\|_K$ (notice that there may be several solutions in $\Theta_{k_{in},\epsilon}$).

To avoid this non-convex minimization problem, we propose a heuristic (see Algorithm 1) for $P_{\Theta_{k_{in},\epsilon}}$ that consists in merging Diracs that are not ϵ-separated.

Input: List $\Theta = (a_i, t_i)_i$ of amplitudes and positions ordered by decreasing absolute amplitudes
for $i \geq 1$ **do**
  for $j > i$ **do**
    if $\|t_i - t_j\| < \epsilon$ **then**
      $t_i = (|a_i|t_i + |a_j|t_j)/(|a_i| + |a_j|)$;
      $a_i = a_i + a_j$;
      Remove (a_j, t_j) from Θ
    end
  end
end

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

Because we take a barycenter of the positions, if a set of Diracs that are at a distance at most ϵ' of a true position in x_0 is merged, the merged result will be within this distance. 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 next Section).

In practice, we implement the projected descent as follows.

- As suggested in [13], to avoid balancing problems between amplitudes and positions, we alternate descent steps between amplitudes and positions.
- To find the step size τ, we perform a line search to minimize the value of the function g.
- We start to project after a few iterations of the descent to increase the reduction of dimension of the projection.

From [13], 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.

3. OVERPARAMETRIZED SPECTRAL INITIALIZATION

Intuitively, as we measure the signal x at some frequencies ω_l, a natural way to recover an estimation of the signal is to back-project the irregular spectrum on a grid Γ that samples $\mathcal{B}_2(R)$ at a given precision ϵ_g (to be chosen later). Given a vector of Fourier measurements y at frequencies $(\omega_l)_{l=1,m}$, this can be done by calculating $z_\Gamma = B_\Gamma y$ where $z_\Gamma = \sum_{s_i \in \Gamma} z_{\Gamma,i} \delta_{s_i}$ and $z_{\Gamma,i} = \sum_l y_l d_l e^{j\langle \omega_l, s_i \rangle}$ for some weights d_l to be chosen in the next section (the $s_i \in \Gamma$ are the grid positions). We show in the noiseless case (in the supplementary material) that as the number of measurements increases and the grid for initialization gets finer, the original positions of Diracs get better approximated. Because the energy of Diracs is well localized by the initialization we then perform overparametrized hard thresholding of the initialization. We propose the following initialization

$$x_{init} = H_{k_{in}}(B_\Gamma y) \tag{9}$$

where for $|z_{\Gamma,j_1}| \geq |z_{\Gamma,j_2}| \geq ....|z_{\Gamma,j_n}|$, we have $H_{k_{in}}(z_\Gamma) = \sum_{i=1,k_{in}} z_{\Gamma,j_i} \delta_{s_{j_i}}$.

3.1. Ideal spectral initialization and sampling

Let B the operator from $\mathbb{C}^m$ to $\mathcal{M}$ defined for $z = By$ by

$$z(t) = \sum_{l=1,m} d_l y_l e^{j\langle \omega_l, t \rangle} \tag{10}$$

We call z an ideal spectral initialization because $z_\Gamma = S_\Gamma z$ where S_Γ is the sampling on the grid Γ: let a measure x defined by its density χ, i.e. $\mathrm{d}x(t) = \chi(t)\,\mathrm{d}t$. 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 an infinitely differentiable density as it is a finite sum of complex exponentials.

We first show that for the right choice of weights d_l, the energy of $z = BAx$ (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.

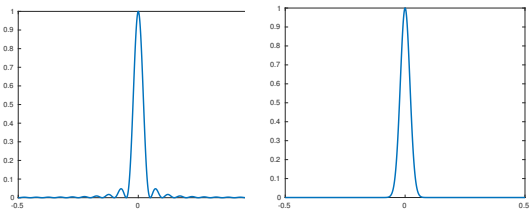

Fig. 1. Kernels used for a separation $\epsilon = 0.1$. Left: On the torus the Féjer kernel of maximum frequency $\frac{2}{\epsilon}$. Right: on $\mathbb{R}$, the Gaussian kernel of parameter $\sigma = \frac{1}{500\epsilon}$

We show the following results for the Féjer and Gaussian kernel as they are typically used in the literature for deterministic [1] and random [7] Fourier sampling.

Lemma 3.1. *On $\mathcal{M}(\mathbb{T}^d)$, we choose A such that $(\omega_l)_{l=1,m}$ is a regular sampling of $[-\omega_{max}, \omega_{max}]^d$ with $\omega_l \in 2\pi.\mathbb{Z}^d$. In (10), take $d_l = \hat{K_f}(\omega_l)/((2\pi)^d c_l)$ where $\hat{K_f}$ is the Fourier transform of the Féjer kernel K_f on the Torus whose Fourier spectrum support is $(\omega_l)_{l=1,m}$, then*

$$z(t) = \sum_{i=1,k} a_i K_f(t - t_i). \tag{11}$$

A sampling of frequencies with maximum frequency $\omega_{max} \geq O(\frac{1}{\epsilon})$ guarantees recovery with methods based on convex relaxations, the associated Féjer kernel is concentrated around 0 as seen in Figure 1. Also, this result is valid for any kernel having its spectrum supported on the ω_l.

For random Fourier sampling, we look at the expected value of z and control its variance with respect to the distribution of the ω_l.

Lemma 3.2. *On $\mathcal{M}(\mathbb{R}^d)$, we choose A such that the ω_l are m i.i.d random variables with a Gaussian distribution with density $G(\omega_r) = \frac{\sigma^d}{(\sqrt{2\pi})^d} e^{-\frac{\sigma^2}{2}\|\omega_r\|_2^2}$. Let $K_g(t) = e^{-\frac{\|t\|_2^2}{2\sigma^2}}$. In (10), take $d_l = 1/(mc_l)$ then*

$$\mathbb{E}(z(t)) = \sum_{i=1,k} a_i K_g(t - t_i) \tag{12}$$

$$\mathbb{E}(|z(t) - E(z(t))|^2) = -\frac{1}{m}|E(z(t))|^2 + \frac{1}{m}\|x_0\|_{K_g}^2 \tag{13}$$

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

Similarly to the regular sampling, the energy of the expected value of z is concentrated around the positions t_i (see Figure 1). In [7] the frequency distribution scales as the inverse of the kernel precision, i.e. the kernel parameter σ of the kernel $h(t) = e^{-\|t\|_2^2/(2\sigma^2)}$ is chosen as $O(1/\epsilon)$. This lemma is valid for any distribution of frequencies having a Fourier transform (which defines the kernel).

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

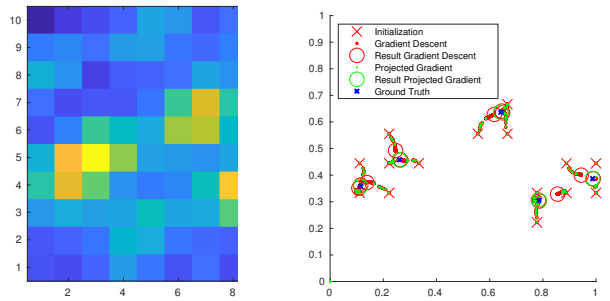

Fig. 2. Result for a few spikes in 2d. Left: back-projection of measurements on a grid. Right: Initialization, gradient descent and projected gradient descent trajectories.

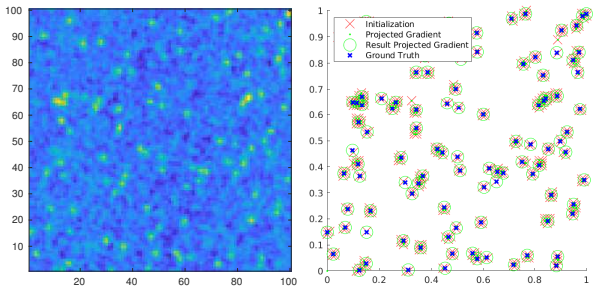

Fig. 3. Result for 100 spikes in 2d. Left: back-projection of measurements on a grid. Right: Initialization, and projected gradient descent trajectories.

the number of measurements using a rule $m = \alpha k d$ with a multiplicative parameter α. The quality of the initialization is thus directly linked to α.

Finally the following Lemma makes sure that as the grid gets finer we recover all the energy of the ideal spectral initialization that lies within the domain sampled by the grid .

Lemma 3.3. *Let* $z_d = (z_{\Gamma,i})_{t_i \in \Gamma}$. *Then* $\| \sqrt{\epsilon_g^d} z_d \|_2^2 \to_{\epsilon_g \to 0} \|z\|^2_{L^2(\mathcal{B}_2(R))}$.

Note that the noisy case just adds a noise term with energy controlled by the noise energy level $\|e\|_2$ because B_Γ is a Fourier back projection.

4. NUMERICAL EXPERIMENTS

We first illustrate the algorithm on few Diracs in 2d then we show some results with many Diracs in 2d to show that the projected descent scales well. We perform the experiments in the noiseless case and leave the study of the impact of the noise for future work. The Matlab code used to generate these experiments is available at yanntraonmilin.wordpress.com/code.

4.1. 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.

4.2. Estimation of many Diracs

We recover 100 Diracs, with a separation 0.01 on the square $[0,1] \times [0,1]$ from $m = 2000$ measurements. In practice, the grid Γ must be fine enough to overparametrize the number of Diracs with a good sampling of the ideal spectral initialization. If ϵ_g is too small, the number of initial Diracs needed to sample the energy gets bigger, leading to an increased cost in the first iterations of the gradient descent. In this example we use $\epsilon_g = \epsilon$ and use $k_{in} = 4k$. We observe in Figure 3 that with these parameters all the Diracs positions are well estimated after 500 iterations of our algorithm.

4.3. Complexity

The cost of our algorithm is the sum of the cost of the initialization and the cost of the projected descent algorithm. Of course, 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. With our strategy, this cost seems unavoidable as we want to localize Diracs off-the-grid at a precision of the order of ϵ (doing the same on the grid would have this exponential scaling with respect to the dimension and the separation). The cost of the projected gradient descent is essentially $O(n_{it} C_\nabla)$ where C_∇ 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 (which is close to k after a few iterations).

5. CONCLUSION

We gave a practical algorithm to perform off-the-grid sparse spike estimation. This proof-of-concept show that it is possible to build a method able to estimate efficiently a large number of parameters 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.
- Investigate other methods for reducing the number of parameters after the back-projection on a grid.
- Investigate quasi-Newton schemes to accelerate the descent (some Hessian information is available).
- Study the algorithm stability to noise and modeling error with respect to the number of measurements.

6. REFERENCES

[1] Emmanuel J Candès and Carlos Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[2] Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5987–5999, 2013.

[3] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht, "Compressed sensing off the grid," *IEEE transactions on information theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

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

[5] Vincent Duval and Gabriel Peyré, "Exact support recovery for sparse spikes deconvolution," *Foundations of Computational Mathematics*, vol. 15, no. 5, pp. 1315–1355, 2015.

[6] Nicolas Keriven, Nicolas Tremblay, Yann Traonmilin, and Rémi Gribonval, "Compressive k-means," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6369–6373.

[7] Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin, "Compressive Statistical Learning with Random Feature Moments," *Preprint*, 2017.

[8] Clément Elvira, Rémi Gribonval, Charles Soussen, and Cedric Herzet, "Omp and continuous dictionaries: Is k-step recovery possible?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5546–5550.

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

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

[11] Valerio Cambareri and Laurent 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.

[12] Lenaic Chizat, "Sparse optimization on measures with over-parameterized gradient descent," *arXiv preprint arXiv:1907.10300*, 2019.

[13] Yann Traonmilin and Jean-François Aujol, "The basins of attraction of the global minimizers of the non-convex sparse spikes estimation problem," *arXiv preprint arXiv:1811.12000*, 2018.