Preconditioned Gradient Descent for Sketched Mixture Learning

Joseph Gabet, Maxime Ferreira Da Costa

Preconditioned Gradient Descent for Sketched Mixture Learning

Joseph Gabet and Maxime Ferreira Da Costa
CentraleSupélec, Université Paris–Saclay, CNRS, Laboratory of Signals and Systems, Gif-sur-Yvette, France
Emails: {joseph.gabet, maxime.ferreira}@centralesupelec.fr

***Abstract*—Sketching consists of reducing the dimensionality of data samples, generally by summarizing their essential features, such as retaining a small number of empirical moments. The reduced representation is called a sketch. In this paper, a Preconditioned Gradient Descent algorithm (PGD) is proposed to estimate the parameter of mixture models (MM) in arbitrary dimensions by minimizing the non-convex quadratic loss between the sketch and the characteristic function of an MM of varying parameters. Preconditioning is introduced to dynamically adapt the descent direction to the local landscape of the objective function, accelerating convergence, with no computational overhead per iteration compared with vanilla GD. An analysis of the linear convergence rate of PGD is conducted, and numerical simulations showcase the method's effectiveness, particularly when the weight of the classes is unbalanced or when a substantial number of data samples is available.**

I. Introduction

Clustering is a fundamental task in unsupervised machine learning. It aims to segment data into homogeneous classes, revealing intrinsic structures, patterns, or hidden relationships in data without prior knowledge of class labels or expected outcomes. Clustering is ubiquitous in applied science and engineering, with applications in areas like image segmentation, biological data analysis [1], content recommendations, and anomaly detection. Due to its flexibility in capturing various data shapes, the mixture model (MM) is commonly assumed to model clusters. It is a well-studied statistical prior in statistics and machine learning. When the class distribution is known, specific numerical methods are known to estimate the parameters of an MM from its empirical samples, such as expectation-maximization [2], [3], or hierarchical clustering [4]. However, alternative methods can provide comparative advantages such as a more comprehensive theoretical understanding or a better scaling with the ambient dimension and the dynamic range of the prior distribution [5].

More recently, estimating MM has been approached through *sketching* [6], [7], a scalable method, which retains in the first few empirical moments of the data distribution to reduce the ambient dimensionality. Then, the mixture parameters are estimated by minimizing the non-convex quadratic loss between the sketch and the characteristic function of an MM.

This work is supported in part by ANR PIA funding: ANR-20-IDEES-0002, AID funding: AID-2023639, and the Orange chair on "Sustainable 6G" held by CentraleSupélec.

A. Contributions and Organization of the Paper

In this study, we consider the sketching framework proposed in [6], and provide novel guarantees on the local geometry of the optimization landscape around the ground truth parameter. Furthermore, we adapt the Preconditioned Gradient Descent (PGD) algorithm [8] into sketching context and prove its linear convergence rate towards the ground truth under sufficiently provided a large number of samples. Additionally, the convergence guarantees are extended to the multidimensional case and to a broad class of convolution kernels.

The rest of the paper is organized as follows. The sketching method and the PGD algorithm are presented in Section II. Section III presents our main results. First, convergence guarantees are provided under infinitely many samples in Theorem 1, then the scaling law of the loss function is discussed under finitely many data samples. Section IV present a proof of Theorem 1. Numerical experiments are conducted in Section V, and a conclusion is drawn in Section VI.

B. Notation and Definitions

Vectors and matrices are denoted by boldface and capital boldface letters, respectively. Elements of vectors $\boldsymbol{x} \in \mathbb{C}^N$ with odd dimension $N = 2n+1$ are indexed between $-n$ and n, so that $\boldsymbol{x} = [x_{-n}, \dots, x_n]^\top$ for convenience. The notation a, b is used to denote the set of all integers from a to b, inclusive. Transpose and Hermitian transpose of a vector or a matrix $\boldsymbol{A}$ are denoted by $\boldsymbol{A}^\top$ and $\boldsymbol{A}^\mathsf{H}$, respectively. We denoted by $\mathbb{1}_d$ the all-one vector in $\mathbb{R}^d$. With a slight abuse of notation, we denote by $|\boldsymbol{a}|$, $|\boldsymbol{a}|^p$, the vector with entries equal the modulus, the pth power of the entries of vector $\boldsymbol{a}$, respectively. The entrywise multiplication product is denoted $\boldsymbol{a} \odot \boldsymbol{a}'$, and the d-dimensional torus is written $\mathbb{T}^d = (\mathbb{R}/\mathbb{Z})^d$.

II. Sketched Mixture Learning

A. Problem Formulation

Given a d-dimension probability distribution with probability density function (PDF) g, we write $\mu(\boldsymbol{\theta}^\star)$ a g-mixture of r components parameterized by $\boldsymbol{\theta}^\star = [\boldsymbol{a}^{\star T}, \boldsymbol{\tau}^{\star T}]^\mathsf{T}$, where $\boldsymbol{a}^\star = [a_1^\star, \dots, a_r^\star]^\mathsf{T} \in [0,1]^r$ is the weight vector of the classes and $\boldsymbol{\tau}^\star = [\tau_{1,1}^\star, \dots, \tau_{1,d}^\star, \tau_{2,1}^\star \dots, \tau_{r,d}^\star]^\mathsf{T} \in \mathbb{R}^{dr}$ encodes the locations of the centroids. Thus the PDF $\mu(\boldsymbol{\theta})$ is given by

$$\mu(\boldsymbol{\theta}^\star)(\boldsymbol{\tau}) = \sum_{j=1}^{r} a_j^\star g(\boldsymbol{\tau} - \boldsymbol{\tau}_j^\star), \quad \boldsymbol{\tau} \in \mathbb{R}^d. \tag{1}$$

In this work, the mixture density function g is assumed to be known. Mixture learning is the task of inferring the parameter $\boldsymbol{\theta}^\star$ from m *i.i.d.* samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\}$ of the distribution $\mu(\boldsymbol{\theta}^\star)$.

In the paper, rather than directly relying on the empirical probability distribution to recover the mixture parameter $\boldsymbol{\theta}^\star$, we propose to minimize the Euclidean loss of n discrete samples of the empirical characteristic function of the observation $\boldsymbol{X}$. This sketching technique [6] maps the ambient dimension from $\mathbb{R}^{md} \to \mathbb{R}^N$, which allows fine control of the dimensionality of the optimization space, and yields substantial computational benefits when $md \gg n$. The empirical multidimensional characteristic function (ECF) $\Phi\{\boldsymbol{X}\} : \mathbb{R}^d \to \mathbb{C}$ of $\boldsymbol{X}$ reads

$$\Phi\{\boldsymbol{X}\}(\boldsymbol{u}) = \frac{1}{m} \sum_{i=1}^{m} e^{-2i\pi\langle \boldsymbol{u}, \boldsymbol{x}_i \rangle}, \quad \boldsymbol{u} \in \mathbb{R}^d, \tag{2}$$

which is an unbiased estimator for the characteristic function of the ground truth MM [9], that we denote $\overline{\Phi}\{\mu(\boldsymbol{\theta}^\star)\}$. Furthermore, the characteristic function $\widehat{g}$ of the PDF g, the expression for a MM $\mu(\boldsymbol{\theta})$ is given for all $\boldsymbol{\theta}$ by

$$\overline{\Phi}\{\mu(\boldsymbol{\theta})\}(\boldsymbol{u}) = \widehat{g}(\boldsymbol{u}) \odot \sum_{j=1}^{r} a_j e^{-2i\pi\langle \boldsymbol{u}, \boldsymbol{\tau}_j \rangle}, \quad \boldsymbol{u} \in \mathbb{R}^d. \tag{3}$$

For the purpose of data processing, one must evaluate the ECF for a finite number of moments according to a sketching scheme. Different sketching schemes have been studied in the literature to discriminate best the classes of the MM, such as uniform sampling and random sampling [6], [10]. For conciseness, we restrict our setup to uniform sampling schemes and let $N = 2n + 1$ be an odd number. We consider the acquisition of N^d samples of the ECF (2) taken over the centrally symmetric uniform sampling set $\Omega = [\![-\frac{n}{N}, \frac{n}{N}]\!]^d$.

In the sequel, the number of classes r is assumed to be known. The parameters $\boldsymbol{\theta}^\star$ are estimated by minimizing the quadratic loss $\mathcal{L}(\cdot)$ between the sketch and the characteristic function of a mixture parameterized by $\boldsymbol{\theta}$. That is

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \left\| [\Phi\{\boldsymbol{X}\}(\boldsymbol{u})]_{\boldsymbol{u}\in\Omega} - \left[\overline{\Phi}\{\mu(\boldsymbol{\theta})\}(\boldsymbol{u})\right]_{\boldsymbol{u}\in\Omega} \right\|_2^2. \tag{4}$$

Given the expression (3) of the characteristic function, Equation (4) amounts to finding a r-sparse combination of complex exponentials weighted by the moments of the shape function g that explains best in the quadratic sense the observations $[\Phi\{\boldsymbol{X}\}(\boldsymbol{u})]_{\boldsymbol{u}\in\Omega}$. This problem is also known as *line spectral estimation* or sparse super-resolution [11] in signal processing, and various methods exist to solve it (see *e.g.* [12]–[17], and reference therein). Of particular to the context of this paper, greedy approaches such as Orthogonal Matching Pursuit (OMP) algorithms [18], [19] provide a fast and scalable family of numerical methods for solving sparse inverse problems such as line spectral estimation. However, those methods rely on discretizing the parameter space, yielding a basis mismatch [20] and imperfect reconstruction, even with infinitely many distribution samples ($m = +\infty$).

To circumvent the limitations of greedy algorithms, we propose instead to study the refinement of first-order optimization iterates initialized at the output of the OMP algorithm. As the optimization landscape of the loss $\mathcal{L}(\boldsymbol{\theta})$ given in (4) is non-convex, global convergence guarantees could hardly be derived. Herein, we rely instead on a study of the *local geometry* of the loss around the ground truth $\boldsymbol{\theta}^\star$ to determine the width of its basin of attraction and the contraction rate to this minimum from an adaptive preconditioning of the descent direction. Further global convergence guarantees demand ensuring an initialization within the basin of attraction.

Algorithm 1 Preconditioned Gradient Descent

1: Initialize $\boldsymbol{\theta}_0$ using OMP; $k \leftarrow 0$.
2: **while** stopping criterion is not met **do**
3: Compute P_k as in (7)
4: $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - P_k \frac{\partial \mathcal{L}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}}$.
5: $k \leftarrow k + 1$
6: **Return** $\boldsymbol{\theta}_k$

B. Learning with Preconditioning

We propose refinements and new associated theoretical guarantees to pre-existing work. Given the heterogeneous nature of vector $\boldsymbol{\theta}$—it contains both locations and weights information which scale differently with N—, we use *preconditioning* to adapt the direction of descent of the first-order method to the local landscape of the cost function [8]. This is done by multiplying at step k the gradient by a matrix $\boldsymbol{P}_k$, which depends on the current estimate $\boldsymbol{\theta}_k$. The optimization procedure is detailed in Algorithm 1. We note that the algorithm doesn't leverage the implicit constraints $\sum_{j=1}^r a_j = 1$ and $a_j \geq 0$, which are unnecessary to establish the local convergence results.

As the computational complexity of Algorithm 1 is driven by that of the matrix-vector multiplication in the fourth line, we restrict our analysis to diagonal preconditioning matrices so that computing the preconditioned descent direction comes with a marginal additional computational cost compared with vanilla gradient descent.

In the sequel, we write $K : \mathbb{R}^d \to \mathbb{R}$ the discrete autocorrelation of $g(\cdot)$ defined by

$$K(\boldsymbol{\tau}) = \sum_{k\in\Omega} |\widehat{g}(\boldsymbol{k})|^2 e^{2i\pi\langle \boldsymbol{k}; \boldsymbol{\tau} \rangle}, \tag{5}$$

which is real-valued since $g(\cdot)$ is real, 1-periodic in every variable and infinitely differentiable.

For convenience, we present the expression of both gradients of the loss (4) with respect to amplitudes and positions, that on direct observations in the frequency domain, $\boldsymbol{x} = [\Phi\{\boldsymbol{X}\}(\boldsymbol{u})]_{\boldsymbol{u}\in\Omega}$,

$$\frac{d\mathcal{L}(\boldsymbol{\theta})}{da_j} = \sum_{l=1}^{r} a_l K(\boldsymbol{\tau}_j - \boldsymbol{\tau}_l) - \sum_{\boldsymbol{k}\in\Omega} \overline{\widehat{g}(\boldsymbol{k})} x_k e^{2i\pi\langle \boldsymbol{\tau}_j, \boldsymbol{k} \rangle} \tag{6a}$$

$$\frac{d\mathcal{L}(\boldsymbol{\theta})}{d\tau_{j,d_1}} = a_j \sum_{l=1}^{r} a_l \frac{\partial K}{\partial \tau_{\cdot,d_1}}(\boldsymbol{\tau}_j - \boldsymbol{\tau}_l) - \mathcal{R}\left(\sum_{\boldsymbol{k}\in\Omega} \overline{\widehat{g}(\boldsymbol{k})} x_k 2i\pi k_{d_1} e^{2i\pi\langle \boldsymbol{\tau}_j, \boldsymbol{k} \rangle} \right). \tag{6b}$$

Furthermore, we select the preconditionner,

$$\boldsymbol{P}_k = \operatorname{diag}\begin{pmatrix} \mathbb{1}_r \\ [-\nabla^2 K(\mathbf{0})]_{1,1}^{-1}|\boldsymbol{a}_k|^{-2} \\ \vdots \\ [-\nabla^2 K(\mathbf{0})]_{d,d}^{-1}|\boldsymbol{a}_k|^{-2} \end{pmatrix}. \tag{7}$$

III. Analysis of Algorithm 1

A. Metrics of Analysis

The *dynamic range* of the problem $\kappa > a^\star_{\max}/a^\star_{\min} > 1$ is defined as the ratio between the largest weight $a_{\max}$ and the smallest weight $a_{\min}$ of the classes of the ground truth MM $\mu(\boldsymbol{\theta}^\star)$. Additionally, the *minimal separation* between the classes, denoted $\Delta(\boldsymbol{\tau}^\star)$, is defined as the smallest possible ℓ_∞ distance over the torus $\mathbb{T}^d$ between two distinct centroids. That is $\Delta(\boldsymbol{\tau}^\star) = \min_{\boldsymbol{k}\in\mathbb{Z}^d} \min_{j\neq j'} \|\boldsymbol{\tau}_j - \boldsymbol{\tau}_{j'} - \boldsymbol{k}\|_\infty$. The dynamic range and the minimal separation are known to be quantities of interest to assess the stability of the line spectral estimation problem [21], [22].

Furthermore, we measure the contraction rate in terms of *worst relative distance* of any parameter to the ground truth. This allows a homogeneous control of the parameter estimate across all the classes and transcends certain properties of the problem at hand, such as the number of source points or the dynamic range. Mathematically, we define the diagonal matrix $\boldsymbol{S}$ as

$$\boldsymbol{S} = \operatorname{diag}\begin{pmatrix} \boldsymbol{a}^{\star -1} \\ \sqrt{[-\nabla^2 K(\mathbf{0})]_{1,1}}\,\mathbb{1}_r \\ \vdots \\ \sqrt{[-\nabla^2 K(\mathbf{0})]_{d,d}}\,\mathbb{1}_r \end{pmatrix},$$

and study the contraction rate of the sequence $\|\boldsymbol{S}\left(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\right)\|_\infty$. Specifically, in one-dimensional settings, the error metric reads

$$\|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty = \max_j \left\{ \frac{|a_{k,j} - a_j^\star|}{|a_j^\star|}; \sqrt{-K''(0)}|\tau_{k,j} - \tau_k^\star| \right\}.$$

B. Asymptotic Convergence Guarantees

In this section, we study the convergence of Algorithm 1 towards the global minimum of the loss function (4) when the characteristic function of $\mu(\boldsymbol{\theta}^\star)$ is observed, that is $m \to \infty$. To that end, we introduce the notion of admissible density, for which we will establish a guarantee of convergence.

Definition 1 (Admissible density): A density g is said to be *admissible* if there exist functions h_0, $\{h_1^i\}_{1\leq i\leq d}$, $\{h_2^{i,j}\}_{1\leq i,j\leq d}$ of $\mathbb{T}^d \to \mathbb{R}$, that are positive, summable, and

- decreasing by positive coordinates : $\mathbf{0} \leq \boldsymbol{x} \leq \boldsymbol{y} \implies h(\boldsymbol{x}) \geq h(\boldsymbol{y})$;
- absolutely dominating the derivatives: $\forall \boldsymbol{u}$, $h_0(\boldsymbol{u}) \geq |K(\boldsymbol{u})|$, $h_1^i(\boldsymbol{u}) \geq |\frac{\partial K(\boldsymbol{u})}{\partial u_i}|$, $h_2^{i,j}(\boldsymbol{u}) \geq |\frac{\partial K(\boldsymbol{u})}{\partial u_i \partial u_j}|$,

and if there exists a constant $C_g > 0$, depending only on g and independent on N such that $\int_{\mathbb{T}^d} h_0 \leq C_g$, $\int_{\mathbb{T}^d} h_1^i \leq \left(-[\nabla^2 K(\mathbf{0})]_{i,i}\right)^{\frac{1}{2}} C_g$, and $\int_{\mathbb{T}^d} h_2^{i,j} \leq \left([\nabla^2 K(\mathbf{0})]_{i,i}[\nabla^2 K(\mathbf{0})]_{j,j}\right)^{\frac{1}{2}} C_g$.

Note the admissibility conditions in Definition 1 are mild and satisfied by most densities of interest, such as the Gaussian. The next theorem, whose proof is presented in Section IV, guarantees the linear convergence of Algorithm 1 as a function of the dynamic range and the minimal separation, provided a close enough initialization and under an admissibility assumption of the mixture shape g.

Theorem 1 (Asymptotic convergence of PGD): Suppose that g is admissible in the sense of Definition 1. Assume $n \geq 2$, an infinite number of samples ($m = +\infty$), then there exists a constant $C_g > 0$, depending only on g, such that if

$$\gamma := C_g \left(\frac{N\Delta(\boldsymbol{\tau}^\star)}{2}\right)^{-d} \frac{a^\star_{\max}}{a^\star_{\min}} < \frac{1}{2} \tag{8a}$$

$$4 \leq \max_{1\leq p\leq d} \left\{ \sqrt{[-\nabla^2 K(\mathbf{0})]_{p,p}} \right\} \Delta(\boldsymbol{\tau}^\star), \tag{8b}$$

and if the initial point satisfies

$$\|\boldsymbol{S}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star)\|_\infty \leq 1 - \sqrt{\frac{2}{3}} \tag{8c}$$

then $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}}$ converges towards $\boldsymbol{\theta}^\star$ and

$$\|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty \leq \left(\frac{1}{2} + \gamma\right)^k \|\boldsymbol{S}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star)\|_\infty. \tag{9}$$

C. Scaling Law under a Finite Number of Samples

In practice, the characteristic function is estimated from finitely many samples $m < +\infty$, and the ECF is subject to stochastic noise. A benefit of stretching from finite many samples resides in the boundedness of the ECF, which allows the control of its pointwise deviation from the characteristic function as follows.

Theorem 2 (Pointwise concentration of the ECF):

$$\forall \boldsymbol{u} \in \mathbb{R}^d,\ \mathbb{P}\left(\left|\Phi\{\boldsymbol{X}\}(\boldsymbol{u}) - \mathbb{E}\left(e^{-2i\pi\langle \boldsymbol{u}, X\rangle}\right)\right| \geq t\right) \leq 2e^{\frac{-t^2 m}{4}} \tag{10}$$

The proof of Theorem (2) is a direct consequence of Hoeffding's inequality [23] applied to the real and imaginary part of the difference in (10). The deviation between the ECF and the ground truth characteristic function can be uniformly controlled via the union bound, yielding for any $t \geq 0$

$$\mathbb{P}\left(\left\|[\Phi\{\boldsymbol{X}\}(\boldsymbol{u})]_{\boldsymbol{u}\in\Omega} - \left[\overline{\Phi}\{\mu(\boldsymbol{\theta})\}(\boldsymbol{u})\right]_{\boldsymbol{u}\in\Omega}\right\|_\infty \geq t\right) \leq 2N^d e^{\frac{-t^2 m}{4}}. \tag{11}$$

Equation (10) suggests $m = \Omega(d\log(N))$ samples are for the stochastic noise to vanish in the asymptotic $N \to \infty$.

IV. Proof of Theorem 1

We start the proof with the following lemma, which relates the contraction of the iterates sequence with the conditioning of the Hessian over the segment between $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}_k$.

Lemma 1 (Contraction of the iterates):
Let $\mathcal{S}_k = \{\boldsymbol{\theta}^\star + u(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star) \mid u \in [0,1]\}$, we have

$$\|\boldsymbol{S}(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^\star)\|_\infty \leq \max_{\boldsymbol{\theta}\in\mathcal{S}_k} \{\|I - \boldsymbol{S}\boldsymbol{P}_k\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{S}^{-1}\|_\infty\} \|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty. \tag{12}$$

The rest of the proof focuses on bounding the quantity $\|\boldsymbol{I} - \boldsymbol{S}\boldsymbol{P}_k\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{S}^{-1}\|_\infty$.

A. Hessian Decomposition

Starting from the expression (6) of the gradient of the loss, the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta}) = \nabla^2\mathcal{L}(\boldsymbol{\theta})$ is given by (*c.f.* [24])

$$\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{M}(\boldsymbol{a})^{\mathsf{H}}\boldsymbol{D}(\boldsymbol{\tau})\boldsymbol{M}(\boldsymbol{a}) + \boldsymbol{E}(\boldsymbol{\theta}), \tag{13}$$

where the matrices $\boldsymbol{M}(\boldsymbol{a})$ and $\boldsymbol{D}(\boldsymbol{\tau})$ are given by

$$\boldsymbol{M}(\boldsymbol{a}) = \operatorname{diag}\begin{pmatrix} \mathbb{1}_r \\ \sqrt{-[\nabla^2 K(0)]_{1,1}}\boldsymbol{a} \\ \vdots \\ \sqrt{-[\nabla^2 K(0)]_{d,d}}\boldsymbol{a} \end{pmatrix} \tag{14}$$

$$\boldsymbol{D}(\boldsymbol{\tau}) = \begin{pmatrix} \boldsymbol{D}_0(\boldsymbol{\tau}) & \boldsymbol{D}_1^1(\boldsymbol{\tau}) & \dots & \boldsymbol{D}_1^d(\boldsymbol{\tau}) \\ \boldsymbol{D}_1^1(\boldsymbol{\tau})^{\mathsf{H}} & \boldsymbol{D}_2^{1,1}(\boldsymbol{\tau}) & \dots & \boldsymbol{D}_2^{1,d}(\boldsymbol{\tau}) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{D}_1^d(\boldsymbol{\tau})^{\mathsf{H}} & \boldsymbol{D}_2^{1,d}(\boldsymbol{\tau})^{\mathsf{H}} & \dots & \boldsymbol{D}_2^{d,d}(\boldsymbol{\tau}) \end{pmatrix}, \tag{15}$$

and where the generic term of each block $\boldsymbol{D}_p(\boldsymbol{\tau})$ is given by

$$[\boldsymbol{D}_0(\boldsymbol{\tau})]_{i,j} = K(\boldsymbol{\tau}_i - \boldsymbol{\tau}_j) \tag{16a}$$

$$[\boldsymbol{D}_1^{d_1}(\boldsymbol{\tau})]_{i,j} = \left(-[\nabla^2 K(\boldsymbol{0})]_{d_1}\right)^{-\frac{1}{2}} \frac{\partial K(\boldsymbol{\tau}_i - \boldsymbol{\tau}_j)}{\partial w_{d_1}} \tag{16b}$$

$$[\boldsymbol{D}_2^{d_1,d_2}(\boldsymbol{\tau})]_{i,j} = \left(\nabla^2 K(\boldsymbol{0})_{d_1}\nabla^2 K(\boldsymbol{0})_{d_2}\right)^{-\frac{1}{2}} \frac{\partial^2 K(\boldsymbol{\tau}_i - \boldsymbol{\tau}_j)}{\partial w_{d_1}\partial w_{d_2}}. \tag{16c}$$

The term $\boldsymbol{E}(\boldsymbol{\theta})$ can be interpreted as a perturbation term that vanishes at $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$. By the triangle inequality, we have $\|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{S}^{-1} - \boldsymbol{I}\|_\infty \le \|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{M}(\boldsymbol{a})^{\mathsf{H}}\boldsymbol{D}(\boldsymbol{\tau})\boldsymbol{M}(\boldsymbol{a})\boldsymbol{S}^{-1} - \boldsymbol{I}\|_\infty + \|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{E}(\boldsymbol{\theta})\boldsymbol{S}^{-1}\|_\infty$. We provide the following lemma on $\|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{E}(\boldsymbol{\theta})\boldsymbol{S}^{-1}\|_\infty$, and skip its proof conciseness, and focus on the sequel by bounding the remaining term.

Lemma 2: If g is admissible in the sense of Definition 1, and if $\|\boldsymbol{S}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_\infty < 1$, then there exists a constant $C_g' > 0$, depending only on g such that

$$\|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{E}(\boldsymbol{\theta})\boldsymbol{S}^{-1}\|_\infty \le C_g' \left(N\Delta(\boldsymbol{\tau})\right)^{-d} \frac{\boldsymbol{a}_{\max}^\star}{\boldsymbol{a}_{\min}^\star} \frac{\|\boldsymbol{S}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_\infty}{\left(1 - \|\boldsymbol{S}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_\infty\right)^2}. \tag{17}$$

B. Bound on $\|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{M}(\boldsymbol{a})^{\mathsf{H}}\boldsymbol{D}(\boldsymbol{\tau})\boldsymbol{M}(\boldsymbol{a})\boldsymbol{S}^{-1} - \boldsymbol{I}\|_\infty$

For all j, we have the identities [8]

$$\frac{|a_j^\star|}{|a_{k,j}|} \le \frac{|a_j^\star|}{|a_j^\star| - |a_{k,j} - a_j^\star|} \le \frac{1}{1 - \|\boldsymbol{S}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_\infty} \tag{18a}$$

$$\max\{|a_j^\star|, |a_j|\} \le 1 + \|\boldsymbol{S}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_\infty \tag{18b}$$

From a direct calculation, exploiting the diagonal structure of the matrices $\boldsymbol{S}$ and $\boldsymbol{M}(\boldsymbol{a})$, and Equations (18), we have

$$\begin{aligned}
&\|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{M}(\boldsymbol{a})^{\mathsf{H}}\boldsymbol{D}(\boldsymbol{\tau})\boldsymbol{M}(\boldsymbol{a})\boldsymbol{S}^{-1} - \boldsymbol{I}\|_\infty \\
&\quad \le \max_j \left\{\frac{1}{a_j^\star}; \frac{|a_j|}{|a_{k,j}|^2}\right\} \|\boldsymbol{D}(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty \max_j \left\{|a_j^\star|; |a_j|\right\} \\
&\qquad + \max_j \left\{\left|1 - \frac{a_j^2}{|a_{k,j}|^2}\right|\right\} \\
&\quad \le \frac{a_{\max}^\star}{a_{\min}^\star} \frac{1 + \|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty}{(1 - \|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty)^2} \|\boldsymbol{D}(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty \\
&\qquad + \frac{1}{(1 - \|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty)^2} - 1.
\end{aligned} \tag{19}$$

We now aim at bounding $\|\boldsymbol{D}(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty$. To proceed, we exploit the block decomposition (15) and the Hermitian structure of sub-blocks (16) to obtain

$$\begin{aligned}
\|\boldsymbol{D}(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty \le \max_i \Big\{ &\|\boldsymbol{D}_0(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty + \sum_{j=1}^d \|\boldsymbol{D}_1^j(\boldsymbol{\tau})\|_\infty, \\
&\|\boldsymbol{D}_1^i(\boldsymbol{\tau})\|_\infty + \|\boldsymbol{D}_2^{i,i}(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty + \sum_{\substack{1 \le j \le d \\ i \ne j}} \|\boldsymbol{D}_2^{i,j}(\boldsymbol{\tau})\|_\infty \Big\}.
\end{aligned} \tag{20}$$

A key element to control the right-hand side of (20) via the generic terms (16) is to bound the summation of the autocorrelation function K and its derivatives at the points $\{\boldsymbol{\tau}_i - \boldsymbol{\tau}_j\}$, which the following lemma proposes.

Lemma 3 (Summation bounds on the autocorrelation): If g is admissible in the sense of Definition 1 then

$$\begin{aligned}
\max_{i,j} \big\{ &\|\boldsymbol{D}_0(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty, \|\boldsymbol{D}_1^i(\boldsymbol{\tau})\|_\infty, \\
&\|\boldsymbol{D}_2^{i,j}(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty, \|\boldsymbol{D}_2^{i,j}(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty \Big\} \le \left(N\Delta(\boldsymbol{\tau})\right)^{-d} C_g.
\end{aligned} \tag{21}$$

Equation (20) and Lemma 3 immediately imply $\|\boldsymbol{D}(\boldsymbol{\tau}) - \boldsymbol{I}\|_\infty \le (d+1)\left(N\Delta(\boldsymbol{\tau})\right)^{-d} C_g$.

C. End of the proof

First of all, Lemma 2, Equation (19) and Lemma 3 yield

$$\begin{aligned}
\|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{S}^{-1} - \boldsymbol{I}\|_\infty \le &\frac{1}{(1 - \|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty)^2} - 1 \\
&+ C_g'' \left(N\Delta(\boldsymbol{\tau})\right)^{-d} \frac{a_{\max}^\star}{a_{\min}^\star} (1 + \|\boldsymbol{S}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_\infty)^2.
\end{aligned} \tag{22}$$

We conclude the proof of Theorem 1 by induction. Assume the conditions (8b) and (8c) hold, which is true of $k = 0$. This ensures $\|\boldsymbol{\tau}_0 - \boldsymbol{\tau}^\star\| \le \frac{\Delta(\boldsymbol{\tau}^\star)}{4}$, thus $\Delta(\boldsymbol{\tau}) \ge \frac{1}{2}\Delta(\boldsymbol{\tau}^\star)$. Thus, the inequality (22) becomes

$$\begin{aligned}
&\max_{\boldsymbol{\theta} \in \mathcal{S}_k} \|\boldsymbol{S}\boldsymbol{P}_k\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{S}^{-1} - \boldsymbol{I}\|_\infty \\
&\quad \le \frac{1}{2} + C_g \left(\frac{N\Delta(\boldsymbol{\tau}^\star)}{2}\right)^{-d} \frac{a_{\max}^\star}{a_{\min}^\star} \qquad (23) \\
&\quad \le \frac{1}{2} + \gamma < 1, \qquad (24)
\end{aligned}$$

with $C_g = C_g''\left(2 - \sqrt{\frac{2}{3}}\right)^2$. We conclude with Lemma 1 that $\|\boldsymbol{S}(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^\star)\|_\infty \le \left(\frac{1}{2} + \gamma\right)\|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty$ which proves the desired statement. ∎

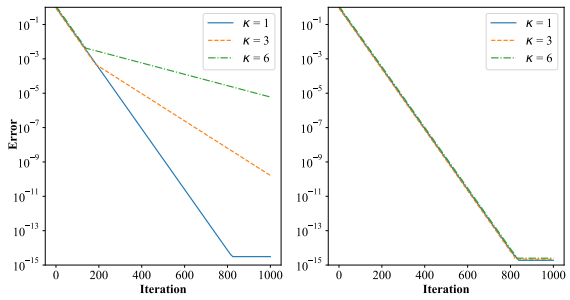

Figure 1. Convergence rates of the iterate sequence of preconditioned GD with fixed step size (left) and adaptive step size (right) towards the ground truth for a 1D Gaussian mixture with variance 1 and centers -2 and 2. The amplitudes are generated to satisfy different values of κ.

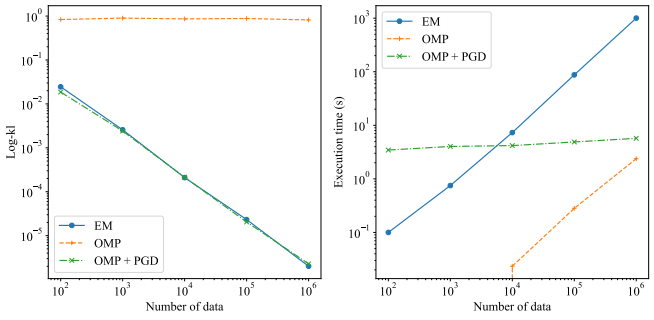

Figure 2. Kullback-Leibler divergence (left) and runtime in seconds (right) per number of samples drawn, displayed in a log-log scale for EM, OMP, and OMP+PGD methods applied to a 2-component 1D Gaussian mixture (with variances of 1, centroids at 0 and 3, and proportions of 0.7 and 0.3, respectively). The results are averaged over 10 random trials.

V. Experimental results

A. Linear convergence rate

This section conducts a numerical study of the convergence rate of Algorithm 1 compared to a vanilla gradient descent algorithm, where the preconditioning matrix $\boldsymbol{P}$ doesn't vary with the iterate k. A one-dimensional homoscedastic Gaussian mixture with sufficient separation is considered, while the dynamic range $\kappa = \frac{a^\star_{\max}}{a^\star_{\min}}$ varies. Figure 1 pictures the error sequence of the iterates $\|\boldsymbol{S}(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\|_\infty$ as k varies. Figure 1 confirms linear convergence of Algorithm 1. Furthermore, we note that adaptive preconditioning remains unaffected by the dynamic scale, in contrast to the fixed step where performance degrades under similar conditions, which corroborates with Theorem 1 for a large enough minimal separation $N\Delta(\boldsymbol{\tau})$, and reinforces the theoretical underpinnings of our approach.

B. Stochastic noise recovery

We evaluate the performance of our proposed approach, which combines Orthogonal Matching Pursuit (OMP) with Preconditioned Gradient Descent (PGD), denoted as OMP+PGD, against two other clustering methods: OMP alone, and Expectation-Maximization (EM) [3].

We select a Gaussian convolution kernel. The left-hand side graph of Figure 2, shows the Kullback-Leibler divergence (KL) between the reconstructing mixture and the ground truth as a function of the number of samples m for all the considered algorithms. OMP + PGD appears to achieve similar performance as EM and converges as m increases, while OMP is not able to converge to the ground truth due to the mismatch introduced by the discretization. The right-side graph presents the algorithms' execution time. As expected, the execution time increases polynomially with m for all three algorithms. However, sketching in OMP+PDG allows a reduction in the ambient dimension before optimization, yielding better scalability.

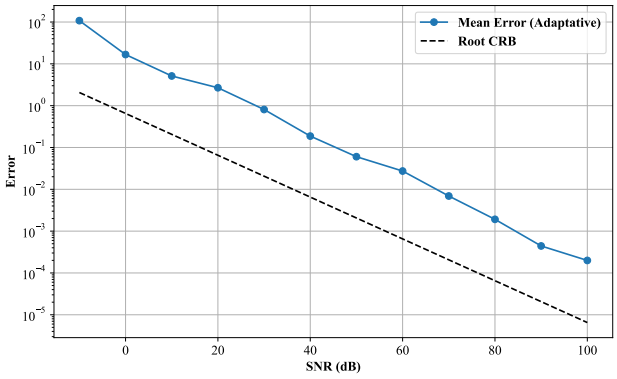

Figure 3. Error $\|\boldsymbol{S}(\boldsymbol{\theta}_\infty - \boldsymbol{\theta}^\star)\|_\infty$ of Algorithm 1 at convergence as a function of the SNR. g is the Gaussian kernel with $r = 2$ classes. The dynamic range is set to $\kappa = 3$, and the results are averaged over 100 random trials.

C. Signal noise recovery

Finally, we assess the robustness of Algorithm 1 to the stochastic noise caused by the finite number of samples m. Let $\boldsymbol{w} = [\Phi\{\boldsymbol{X}\}(\boldsymbol{u})]_{\boldsymbol{u}\in\Omega} - \left[\overline{\Phi}\{\mu(\boldsymbol{\theta}^\star)\}(\boldsymbol{u})\right]_{\boldsymbol{u}\in\Omega}$ be the stochastic noise, which is controlled with high-probability in ℓ_∞-norm in Equation (11). We define the signal-to-noise ratio as $\mathrm{SNR} = \|\Phi\{\mu(\boldsymbol{\theta}^\star)\}\|_2^2 / \|\boldsymbol{w}\|_2^2$. Figure 3 shows the error estimate on the parameters of OMP+PDG to estimate a Gaussian mixture with two classes as a function of the SNR. The results are benchmarked against the statistical Cramér-Rao lower bound [25]. Although the empirical error does not achieve the CRB, the statistical error is of the same decay rate. Thus, the proposed method is robust to the stochastic noise induced by a finite number empirical samples.

VI. Conclusion

We have presented a preconditioned gradient descent algorithm to identify the components of a mixture model from the empirical moments of the observation through a sketching technique. Local convergence guarantees are provided in arbitrary dimensions and for a very broad class of kernels in Theorem 1, under a good enough initial point, and in the asymptotic number of distribution samples. In particular, we prove the linear contraction rate of the iterate sequence. The analysis amounts to line spectral estimation, bridging classical learning and signal processing.

Future research directions include a study of the convergence in a stochastic framework, where a limited number of samples is assumed.

References

[1] Y. Zhao and G. Karypis, "Data clustering in life sciences", *Molecular biotechnology*, vol. 31, pp. 55–80, 2005.

[2] T. K. Moon, "The expectation-maximization algorithm", *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[3] F. Dellaert, "The expectation maximization algorithm", *College of Computing, Georgia Institute of Technology*, 2002.

[4] J. Goldberger and S. Roweis, "Hierarchical clustering of a mixture model", *Advances in neural information processing systems*, vol. 17, 2004.

[5] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*. M. Dekker New York, 1988, vol. 38.

[6] R. G. Nicolas Keriven Anthony Bourrier and P. Pérez, "Sketching for large-scale learning of mixture models", *Information and Inference*, vol. 7, no. 3, pp. 447–508, 2018.

[7] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin, "Statistical learning guarantees for compressive clustering and compressive mixture modeling", *Mathematical Statistics and Learning*, vol. 3, no. 2, pp. 165–257, 2021.

[8] M. Ferreira Da Costa and Y. Chi, "Local geometry of nonconvex spike deconvolution from low-pass measurements", *IEEE Journal on Selected Areas in Information Theory*, vol. 4, pp. 1–15, 2023.

[9] A. Feuerverger and R. A. Mureika, "The empirical characteristic function and its applications", *Annals of Statistics*, vol. 5, no. 1, pp. 88–97, 1977.

[10] R. Gribonval, A. Chatalic, N. Keriven, V. Schellekens, L. Jacques, and P. Schniter, "Sketching data sets for large-scale learning: Keeping only what you need", *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 12–36, 2021.

[11] G. Tang, B. N. Bhaskar, and B. Recht, "Near minimax line spectral estimation", *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 499–512, 2014.

[12] R. Prony, "Essai experimental", *J. de l'Ecole Polytechnique*, vol. 2, p. 929, 1795.

[13] R. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[14] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques", *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[15] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution", *Communications on pure and applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[16] Y. Chi and M. Ferreira Da Costa, "Harnessing sparsity over the continuum: Atomic norm minimization for superresolution", *IEEE Signal Processing Magazine*, vol. 37, no. 2, pp. 39–57, 2020.

[17] P.-J. Bénard, Y. Traonmilin, J.-F. Aujol, and E. Soubies, "Estimation of off-the-grid sparse spikes with over-parametrized projected gradient descent: Theory and application", *hal-04220523*, 2023.

[18] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit", *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[19] C. Soussen, R. Gribonval, J. Idier, and C. Herzet, "Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares", *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 3158–3174, 2013.

[20] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing", *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, 2011.

[21] A. Moitra, "Super-resolution, extremal functions and the condition number of Vandermonde matrices", in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, 2015, pp. 821–830.

[22] M. Ferreira Da Costa, "The condition number of weighted non-harmonic Fourier matrices with applications to super-resolution", *hal-04261330*, Oct. 2023, preprint.

[23] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press), English. Oxford, UK: Oxford University Press, 2016, p. 496.

[24] Y. Traonmilin and J.-F. Aujol, "The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem", *Inverse Problems*, vol. 36, no. 4, p. 045 003, 2020.

[25] L. L. Scharf and L. McWhorter, "Geometry of the Cramér-Rao bound", *Signal Processing*, vol. 31, no. 3, pp. 301–311, 1993.