Blind Demixing via Wirtinger Flow with Random Initialization

Jialin Dong
dongjl@shanghaitech.edu.cn
ShanghaiTech University

Yuanming Shi
shiym@shanghaitech.edu.cn
ShanghaiTech University

Abstract

This paper concerns the problem of demixing a series of source signals from the sum of bilinear measurements. This problem spans diverse areas such as communication, imaging processing, machine learning, etc. However, semidefinite programming for blind demixing is prohibitive to large-scale problems due to high computational complexity and storage cost. Although several efficient algorithms have been developed recently that enjoy the benefits of fast convergence rates and even regularization free, they still call for spectral initialization. To find simple initialization approach that works equally well as spectral initialization, we propose to solve blind demixing problem via Wirtinger flow with *random initialization*, which yields a natural implementation. To reveal the efficiency of this algorithm, we provide the global convergence guarantee concerning randomly initialized Wirtinger flow for blind demixing. Specifically, it shows that with sufficient samples, the iterates of randomly initialized Wirtinger flow can enter a local region that enjoys strong convexity and strong smoothness within a few iterations at the first stage. At the second stage, iterates of randomly initialized Wirtinger flow further converge linearly to the ground truth.

1 INTRODUCTION

Suppose we are given an observation vector $\boldsymbol{y} \in \mathbb{C}^m$ in frequency domain generated from the sum of bilinear measurements of unknown vectors $\boldsymbol{x}_i^\natural \in \mathbb{C}^N$, $\boldsymbol{h}_i^\natural \in \mathbb{C}^K$, $i = 1, \cdots, s$, i.e.,

$$y_j = \sum_{i=1}^{s} \boldsymbol{b}_j^* \boldsymbol{h}_i^\natural \boldsymbol{x}_i^{\natural *} \boldsymbol{a}_{ij},\ 1 \le j \le m, \tag{1}$$

where $\{\boldsymbol{a}_{ij}\} \in \mathbb{C}^N$, $\{\boldsymbol{b}_j\} \in \mathbb{C}^K$ are design vectors. Here, the first K columns of the matrix $\boldsymbol{F}$ form the matrix $\boldsymbol{B} := [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_m]^* \in \mathbb{C}^{m \times K}$, where $\boldsymbol{F} \in \mathbb{C}^{m \times m}$ is the unitary discrete Fourier transform (DFT) matrix with $\boldsymbol{F}\boldsymbol{F}^* = \boldsymbol{I}_m$. Our goal is to recover $\{\boldsymbol{x}_i^\natural\}$ and $\{\boldsymbol{h}_i^\natural\}$ from the sum of bilinear measurements, which is known as *blind demixing* [1, 2].

This problem has spanned a wide scope of applications ranging from imaging processing [3, 4] and machine learning [5, 6] to communication [7, 8]. Specifically, by solving the blind demixing problem, both original images and corresponding convolutional kernels can be recovered from a blurred image [3]. This problem has also been exploited to demix and deconvolve calcium imaging recordings of neuronal ensembles [4]. Blind demixing also finds applications in machine learning [5, 6] where feature maps convolve with filters individually and added overall users to approximate the input vector. In the context of communication, this problem ensures to recover the source signal from distinguishing users without knowing channel vectors, thereby enabling low-latency communication [7].

Although blind demixing plays vital role in various areas, solving it is generally highly intractable. Moreover, some applications of blind demixing problem involves large-scale data, hence efficient algorithms with optimal guarantees are needed urgently to recover signal vectors from a single observation vector. Instead of computationally expensive convex method [1], a nonconvex algorithm, regularized gradient descent [9], has been recently proposed to efficiently solve blind demixing problem. This method, however, requires extra regularization and still yields conservative computational optimality guarantees. To elude the regularization and develop an algorithm with progressive computational guarantees, the least square estimation ap-

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

proach was proposed in [2]:

$$\mathscr{P}: \underset{\{\boldsymbol{h}_i\},\{\boldsymbol{x}_i\}}{\text{minimize}} \ f(\boldsymbol{h},\boldsymbol{x}) := \sum_{j=1}^{m}\Big|\sum_{i=1}^{s}\boldsymbol{b}_j^*\boldsymbol{h}_i\boldsymbol{x}_i^*\boldsymbol{a}_{ij} - y_j\Big|^2, \quad (2)$$

which is solved via Wirtinger flow with spectral initialization by harnessing the benefits of low computational complexity, regularization-free, fast convergence rate with aggressive step size and computational optimality guarantees [2]. Nevertheless, this algorithm with theoretical guarantees depends on the spectral initialization by computing the largest singular vector of the data matrix [2]. To find a natural implementation for the practitioners, we propose to solve the blind demixing problem via randomly initialized Wirtinger flow. Our goal, in this paper, is to confirm the efficiency of this algorithm with theoretical guarantees. Specifically, we shall verify that the iterates of randomly initialized Wirtinger flow are able to achieve a local region that enjoys strong convexity and strong smoothness within a few iteration at the first stage. At the next stage, we further show that the iterates of the randomly initialized Wirtinger flow linearly converges to the ground truth.

1.1 State-of-the-Art Algorithms

Despite the general intractability of blind demixing, it can be effectively solved by several algorithms under the proper statistical models. In particular, semidefinite programming was developed in [1] to solve the blind demixing problem by lifting the bilinear model into the matrix space. However, it is computationally prohibitive for solving large-scale problem due to the high computation and storage cost. To address this issue, the nonconvex algorithm, e.g., regularized gradient descent with spectral initialization [9], was further developed to optimize the variables in the natural space. Nevertheless, the theoretical guarantees for the regularized gradient [9] provide pessimistic convergence rate and require carefully-designed initialization. The Riemannian trust-region optimization algorithm without regularization was further proposed in [7] to improve the convergence rate. However, the second-order algorithm brings unique challenges in providing statistical guarantees. Recently, theoretical guarantees concerning regularization-freed Wirtinger flow with spectral initialization for blind demixing was provided in [2]. However, this regularization-freed method still calls for spectral initialization. In this paper, we aim to explore the random initialization strategy for natural implementation and theoretically verify the efficiency of the randomly initialized Wirtinger flow.

Based on the random initialization strategy, a line of research studies the benign global landscapes and aims to design generic saddle-point escaping algorithms, e.g., noisy stochastic gradient descent [10], trust-region method [11], perturbed gradient descent [12]. With sufficient sample size, these algorithms are guaranteed to converge globally for phase retrieval [11], matrix recovery [13], matrix sensing [13], robust PCA [14] and shallow neural networks [15] where all local minima are as good as global and all the saddle points are strict. However, the theoretical results developed in [10, 11, 12, 13, 14, 15] are fairly general and may yield pessimistic convergence rate guarantees. Moreover, these saddle-point escaping algorithms are more complicated for implementation than the natural vanilla gradient descent or Wirtinger flow. To advance the theoretical analysis for gradient descent with random initialization, the fast global convergence guarantee concerning randomly initialized gradient descent for phase retrieval has been recently provided in [16].

1.2 Wirtinger Flow with Random Initialization

In this paper, we solve the blind demixing problem via randomly initialized Wirtinger flow by harnessing the benefits of computational efficiency, regularization-free and careful initialization free. In this paper, our main contribution is to provide the global convergence guarantee for the randomly initialized Wirtinger flow.

Wirtinger flow with random initialization is an iterative algorithm with vanilla gradient descent update procedure, i.e., without regularization. Specifically, the gradient step of Wirtinger flow is represented by the notion of Wirtinger derivatives [17], i.e., the derivatives of real-valued functions over complex variables. To simplify the presentation, we denote $f(\boldsymbol{z}) := f(\boldsymbol{h},\boldsymbol{x})$, where

$$\boldsymbol{z} = \begin{bmatrix}\boldsymbol{z}_1\\ \cdots \\ \boldsymbol{z}_s\end{bmatrix} \in \mathbb{C}^{s(N+K)} \text{ with } \boldsymbol{z}_i = \begin{bmatrix}\boldsymbol{h}_i\\ \boldsymbol{x}_i\end{bmatrix} \in \mathbb{C}^{N+K}. \quad (3)$$

Furthermore, we define the discrepancy between the estimate $\boldsymbol{z}$ and the ground truth $\boldsymbol{z}^\natural$ as the distance function, given as

$$\text{dist}(\boldsymbol{z},\boldsymbol{z}^\natural) = \left(\sum_{i=1}^{s}\text{dist}^2(\boldsymbol{z}_i,\boldsymbol{z}_i^\natural)\right)^{1/2}, \quad (4)$$

where

$$\text{dist}^2(\boldsymbol{z}_i,\boldsymbol{z}_i^\natural) = \min_{\alpha_i\in\mathbb{C}}(\|\frac{1}{\alpha_i}\boldsymbol{h}_i - \boldsymbol{h}_i^\natural\|_2^2 + \|\alpha_i\boldsymbol{x}_i - \boldsymbol{x}_i^\natural\|_2^2)/d_i$$

for $i = 1,\cdots,s$. Here, $d_i = \|\boldsymbol{h}_i^\natural\|_2^2 + \|\boldsymbol{x}_i^\natural\|_2^2$ and each α_i is the alignment parameter.

Let $\boldsymbol{A}^*$ and $\boldsymbol{a}^*$ denote the conjugate transpose of matrix $\boldsymbol{A}$ vector $\boldsymbol{a}$ respectively. For each $i = 1,\cdots,s$,

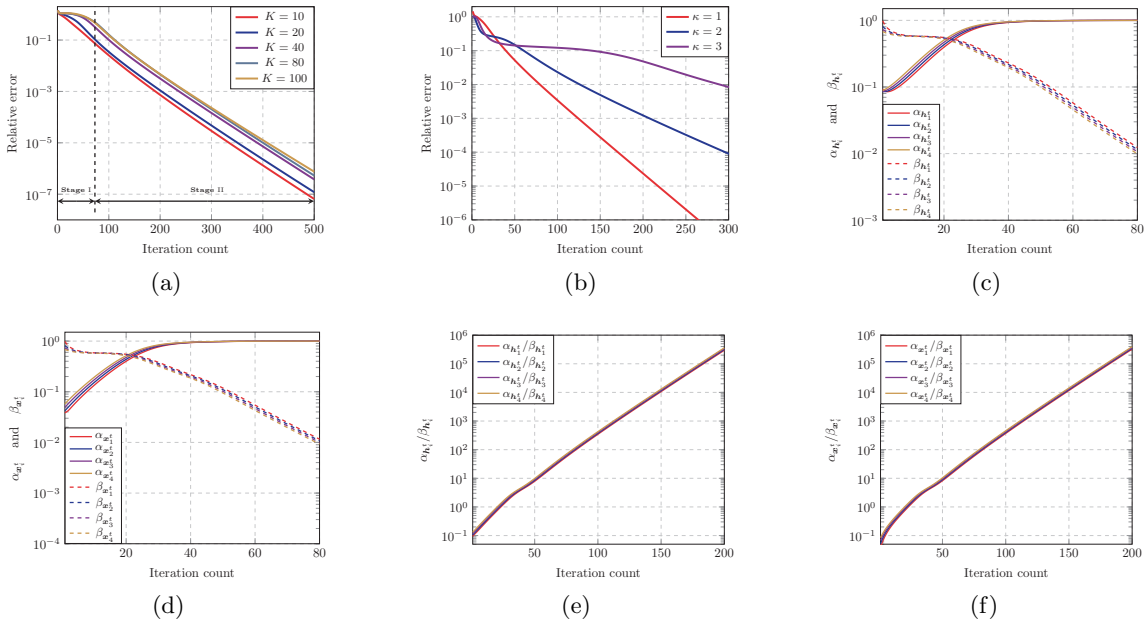

Figure 1: Numerical results.

$\nabla_{\boldsymbol{h}_i} f(\boldsymbol{z})$ and $\nabla_{\boldsymbol{x}_i} f(\boldsymbol{z})$ denote the Wirtinger gradient of $f(\boldsymbol{z})$ with respect to $\boldsymbol{h}_i$ and $\boldsymbol{x}_i$ respectively:

$$\nabla_{\boldsymbol{h}_i} f(\boldsymbol{z}) = \sum_{j=1}^{m} \left(\sum_{k=1}^{s} \boldsymbol{b}_j^* \boldsymbol{h}_k \boldsymbol{x}_k^* \boldsymbol{a}_{kj} - y_j \right) \boldsymbol{b}_j \boldsymbol{a}_{ij}^* \boldsymbol{x}_i, \quad (5a)$$

$$\nabla_{\boldsymbol{x}_i} f(\boldsymbol{z}) = \sum_{j=1}^{m} \left(\sum_{k=1}^{s} \boldsymbol{h}_k^* \boldsymbol{b}_j \boldsymbol{a}_{kj}^* \boldsymbol{x}_k - \overline{y}_j \right) \boldsymbol{a}_{ij} \boldsymbol{b}_j^* \boldsymbol{h}_i. \quad (5b)$$

In light of the Wirtinger gradient (5), the update rule of Wirtinger flow is given by

$$\begin{bmatrix} \boldsymbol{h}_i^{t+1} \\ \boldsymbol{x}_i^{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}_i^{t} \\ \boldsymbol{x}_i^{t} \end{bmatrix} - \eta \begin{bmatrix} \frac{1}{\|\boldsymbol{x}_i^t\|_2^2} \nabla_{\boldsymbol{h}_i} f(\boldsymbol{z}^t) \\ \frac{1}{\|\boldsymbol{h}_i^t\|_2^2} \nabla_{\boldsymbol{x}_i} f(\boldsymbol{z}^t) \end{bmatrix}, i = 1, \cdots, s, \quad (6)$$

where $\eta > 0$ is the step size.

1.3 Numerical Results

The power of randomly initialized WF for solving problem $\mathscr{P}$ (2) will be illustrated by numerical example. The ground truth signals and initial points are randomly generated as

$$\boldsymbol{h}_i^\natural \sim \mathcal{N}(\boldsymbol{0}, K^{-1}\boldsymbol{I}_K), \ \boldsymbol{x}_i^\natural \sim \mathcal{N}(\boldsymbol{0}, N^{-1}\boldsymbol{I}_N), \quad (7)$$

$$\boldsymbol{h}_i^0 \sim \mathcal{N}(\boldsymbol{0}, K^{-1}\boldsymbol{I}_K), \ \boldsymbol{x}_i^0 \sim \mathcal{N}(\boldsymbol{0}, N^{-1}\boldsymbol{I}_N), \quad (8)$$

for $i = 1, \cdots, s$. In all simulations, we set $K = N$ and normalize $\|\boldsymbol{h}_i^\natural\|_2 = \|\boldsymbol{x}_i^\natural\|_2 = 1$ for $i = 1, \cdots, s$. Specifically, for each $K \in \{10, 20, 40, 80, 100\}$, $s = 10$ and $m = 50K$, the design vectors $\boldsymbol{a}_{ij}$'s follow $\boldsymbol{a}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_N) + i\mathcal{N}(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_N)$. With the chosen step size $\eta = 0.1$ in all settings, Fig. 1(a) shows the relative error, i.e., $\sum_{i=1}^{s} \|\boldsymbol{h}_i^t \boldsymbol{x}_i^{t*} - \boldsymbol{h}_i^\natural \boldsymbol{x}_i^{\natural *}\|_F / \sum_{i=1}^{s} \|\boldsymbol{h}_i^\natural \boldsymbol{x}_i^{\natural *}\|_F$, versus the iteration count, where $\|\boldsymbol{X}\|_F$ denotes the Frobenius norm of the matrix $\boldsymbol{X}$. We observe that, the iterations of randomly initialized Wirtinger flow can be separately into two stages: Stage I: within first tens of iterations, the relative error maintains nearly flat, Stage II: the relative error enjoys linear convergence rate which rarely changes as the problem size varies.

We further illustrate that the performance and convergence rate of the Wirtinger flow with random initialization depend on the condition number, i.e., $\kappa := \frac{\max_i \|\boldsymbol{x}_i^\natural\|_2}{\min_i \|\boldsymbol{x}_i^\natural\|_2}$. In this experiment, let $K = 50$, $m = 800$, $s = 2$, the step size be $\eta = 0.2$. We then set for the first component $\|\boldsymbol{h}_1^\natural\|_2 = \|\boldsymbol{x}_1^\natural\|_2 = 1$ and for the second one $\|\boldsymbol{h}_2^\natural\|_2 = \|\boldsymbol{x}_2^\natural\|_2 = \kappa$ with $\kappa \in \{1, 2, 3\}$. The random initialization (8) is utilized. Fig. 1(b) shows the relative error versus the iteration count. As we can see, a larger κ yields slower convergence rate.

2 MAIN RESULTS

In this section, we shall verify the preceding numerical results with theoretical guarantees. Through theoretical analysis, we assume the design vectors $\boldsymbol{a}_{ij}$'s as $\boldsymbol{a}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_N) + i\mathcal{N}(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_N)$. To present the main theorem, we first introduce several fundamental notions and definitions.

Throughout this paper, $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ denotes that there exists a constant $c > 0$ such that $|f(n)| \le c|g(n)|$, whereas $f(n) \gtrsim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \ge c|g(n)|$. $f(n) \gg g(n)$ denotes that there exists some sufficiently large constant $c > 0$ such that $|f(n)| \ge c|g(n)|$. In addition, the notation $f(n) \asymp g(n)$ means that there exists constants $c_1, c_2 > 0$ such that $c_1|g(n)| \le |f(n)| \le c_2|g(n)|$.

Furthermore, the incoherence parameter [9], which characterizes the incoherence between $\boldsymbol{b}_j$ and $\boldsymbol{h}_i$ for $1 \le i \le s, 1 \le j \le m$.

Definition 1 (Incoherence for blind demixing)**.** *Let the incoherence parameter μ be the smallest number such that*

$$\max_{1\le i\le s, 1\le j\le m} \frac{|\boldsymbol{b}_j^* \boldsymbol{h}_i^\natural|}{\|\boldsymbol{h}_i^\natural\|_2} \le \frac{\mu}{\sqrt{m}}. \tag{9}$$

Let $\widetilde{\boldsymbol{h}}_i^t$ and $\widetilde{\boldsymbol{x}}_i^t$ be

$$\widetilde{\boldsymbol{h}}_i^t = \frac{1}{\overline{\omega_i^t}} \boldsymbol{h}_i^t \quad \text{and} \quad \widetilde{\boldsymbol{x}}_i^t = \omega_i^t \boldsymbol{x}_i^t, \tag{10}$$

for $i = 1, \cdots, s$, respectively, where ω_i's are alignment parameters. We further define the norm of signal component and the perpendicular component with respect to $\boldsymbol{h}_i$ for $i = 1, \cdots, s$, as

$$\alpha_{\boldsymbol{h}_i} := \langle \boldsymbol{h}_i^\natural, \widetilde{\boldsymbol{h}}_i^t \rangle / \|\boldsymbol{h}_i^\natural\|_2, \tag{11}$$

$$\beta_{\boldsymbol{h}_i} := \left\| \widetilde{\boldsymbol{h}}_i^t - \frac{\langle \boldsymbol{h}_i^\natural, \widetilde{\boldsymbol{h}}_i^t \rangle}{\|\boldsymbol{h}_i^\natural\|_2^2} \boldsymbol{h}_i^\natural \right\|_2, \tag{12}$$

respectively. Here, ω_i's are the alignment parameters. Similarly, the norms of the signal component and the perpendicular component with respect to $\boldsymbol{x}_i$ for $i = 1, \cdots, s$, can be represented as

$$\alpha_{\boldsymbol{x}_i} := \langle \boldsymbol{x}_i^\natural, \widetilde{\boldsymbol{x}}_i^t \rangle / \|\boldsymbol{x}_i^\natural\|_2, \tag{13}$$

$$\beta_{\boldsymbol{x}_i} := \left\| \widetilde{\boldsymbol{x}}_i^t - \frac{\langle \boldsymbol{x}_i^\natural, \widetilde{\boldsymbol{x}}_i^t \rangle}{\|\boldsymbol{x}_i^\natural\|_2^2} \boldsymbol{x}_i^\natural \right\|_2, \tag{14}$$

respectively.

Theorem 1. *Assume that the initial points obey (8) for $i = 1, \cdots, s$ and the stepsize $\eta > 0$ satisfies $\eta \asymp s^{-1}$. Suppose that the sample size satisfies $m \ge C\mu^2 s^2 \kappa^4 \max\{K, N\} \log^{12} m$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - c_1 m^{-\nu} - c_1 m e^{-c_2 N}$ for some constants $\nu, c_1, c_2 > 0$, there exists a sufficiently small constant $0 \le \gamma \le 1$ and $T_\gamma \lesssim s \log(\max\{K, N\})$ such that*

1. *The randomly initialized WF linearly converges to $\boldsymbol{z}^\natural$, i.e.,*

$$\operatorname{dist}(\boldsymbol{z}^t, \boldsymbol{z}^\natural) \le \gamma \left(1 - \frac{\eta}{16\kappa}\right)^{t - T_\gamma} \left\|\boldsymbol{z}^\natural\right\|_2, \ t \ge T_\gamma,$$

2. *The magnitude ratios of the signal component to the perpendicular component with respect to $\boldsymbol{h}_i^t$ and $\boldsymbol{x}_i^t$ obey*

$$\max_{1\le i\le s} \frac{\alpha_{\boldsymbol{h}_i^t}}{\beta_{\boldsymbol{h}_i^t}} \gtrsim \frac{1}{\sqrt{K \log K}} (1 + c_3\eta)^t, \tag{15a}$$

$$\max_{1\le i\le s} \frac{\alpha_{\boldsymbol{x}_i^t}}{\beta_{\boldsymbol{x}_i^t}} \gtrsim \frac{1}{\sqrt{N \log N}} (1 + c_4\eta)^t, \tag{15b}$$

respectively, where $t = 0, 1, \cdots$ for some constant $c_3, c_4 > 0$.

Theorem 1 provides precise statistical analysis on the computational efficiency of WF with random initialization. Specifically, in Stage I, it takes $T_\gamma = \mathcal{O}(s \log(\max\{K, N\}))$ iterations for randomly initialized WF to a local region near the ground truth that enjoys strong convexity and strong smoothness. The short duration of Stage I is own to the exponential growth of the magnitude ratio of the signal component to the perpendicular components (15). Moreover, in Stage II, it takes $\mathcal{O}(s \log(1/\varepsilon))$ iterations to reach ε-accurate solution at a linear convergence rate. Thus, the randomly initialized WF is guaranteed to converge to the ground truth with the iteration complexity $\mathcal{O}(s \log(\max\{K, N\}) + s \log(1/\varepsilon))$ given the sample size $m \gtrsim s^2 \max\{K, N\} \operatorname{poly}\log(m)$.

To further illustrate the relationship between the signal component $\alpha_{\boldsymbol{h}_i}$ (resp. $\alpha_{\boldsymbol{x}_i}$) and the perpendicular component $\beta_{\boldsymbol{h}_i}$ (resp. $\beta_{\boldsymbol{x}_i}$) for $i = 1, \cdots, s$, we provide the simulation results under the setting of $K = N = 10$, $m = 50K$, $s = 4$ and $\eta = 0.1$ with $\|\boldsymbol{h}_i^\natural\|_2 = \|\boldsymbol{x}_i^\natural\|_2 = 1$ for $1 \le i \le s$. In particular, $\alpha_{\boldsymbol{h}_i}$, $\beta_{\boldsymbol{h}_i}$ versus iteration count (resp. $\alpha_{\boldsymbol{h}_i}$, $\beta_{\boldsymbol{h}_i}$ versus iteration count) for $i = 1, \cdots, s$ is demonstrated in Fig. 1(c) (resp. Fig. 1(d)). Consider Fig. 1(a), Fig. 1(c) and Fig. 1(d) collectively, it shows that despite the rare decline of the relative error during Stage I, the sizes of the signal components, i.e., $\alpha_{\boldsymbol{h}_i}$ and $\alpha_{\boldsymbol{x}_i}$ for each $i = 1, \cdots, s$, exponentially increase and the signal component becomes dominant component at the end of Stage I. Furthermore, the exponential growth of the ratio $\alpha_{\boldsymbol{h}_i}/\beta_{\boldsymbol{h}_i}$ (resp. $\alpha_{\boldsymbol{x}_i}/\beta_{\boldsymbol{x}_i}$) for each $i = 1, \cdots, s$ is illustrated in Fig. 1(e) (resp. Fig. 1(f)).

3 DYNAMICS ANALYSIS

In this section, we shall briefly summarize the proof of the main theorem which is based on investigating the

dynamics of the iterates of WF with random initialization. The steps of proving Theorem 1 are summarized as follows.

1. **Stage I**:
   - **Dynamics of population-level state evolution.** Provide the population-level state evolution of $\alpha_{\boldsymbol{x}_i}$ (20a) and $\beta_{\boldsymbol{x}_i}$ (20b), $\alpha_{\boldsymbol{h}_i}$ (21a), $\beta_{\boldsymbol{h}_i}$ (21b) respectively, where the sample size approaches infinity. We then develop the approximate state evolution (23), which are remarkably close to the population-level state evolution, in the finite-sample regime. See details in Section 3.1.
   - **Dynamics of approximate state evolution.** Show that there exists some $T_\gamma = \mathcal{O}(s\log(\max\{K,N\}))$ such that $\mathrm{dist}(\boldsymbol{z}^{T_\gamma}, \boldsymbol{z}^\natural) \le \gamma$, if $\alpha_{\boldsymbol{h}_i}$ (11), $\beta_{\boldsymbol{h}_i}$ (12), $\alpha_{\boldsymbol{x}_i}$ (13) and $\beta_{\boldsymbol{x}_i}$ (14) satisfy the approximate state evolution (23). The exponential growth of the ratio $\alpha_{\boldsymbol{h}_i}/\beta_{\boldsymbol{h}_i}$ and $\alpha_{\boldsymbol{x}_i}/\beta_{\boldsymbol{x}_i}$ are further demonstrated under the same assumption. Please refer to Section 3.2.
   - **Leave-one-out arguments.** Prove that with high probability $\alpha_{\boldsymbol{h}_i}$, $\beta_{\boldsymbol{h}_i}$, $\alpha_{\boldsymbol{x}_i}$ and $\beta_{\boldsymbol{x}_i}$ satisfy the approximate state evolution (23) if the iterates $\{\boldsymbol{z}_i\}$ are independent with $\{\boldsymbol{a}_{ij}\}$. See details in Section A in the supplemental material. To achieve this, the "near-independence" between $\{\boldsymbol{z}_i\}$ and $\{\boldsymbol{a}_{ij}\}$ is established via exploiting leave-one-out arguments and some variants of the arguments. Specifically, the leave-one-out sequences and random-sign sequences are constructed in Section 3.3. The concentrations between the original and these auxiliary sequences are provided in the supplemental material.
2. **Stage II**: **Local geometry in the region of incoherence and contraction.** We invoke the prior theory provided in [2] to show local convergence of the random initialized WF in Stage II. Claim 15 in Stage II are further proven. Please refer to Section A.3 in the supplemental material.

3.1 Dynamics of Population-level State Evolution

In this subsection, we investigate the dynamics of population-level (where we have infinite samples) state evolution of $\alpha_{\boldsymbol{h}_i}$ (11), $\beta_{\boldsymbol{h}_i}$ (12), $\alpha_{\boldsymbol{x}_i}$ (13) and $\beta_{\boldsymbol{x}_i}$ (14).

Without loss the generality, we assume that $\boldsymbol{x}_i^\natural = q_i\boldsymbol{e}_1$ for $i = 1,\cdots,s$, where $0 < q_i \le 1, i = 1,\cdots,s$ are some constants and $\kappa = \frac{\max_i q_i}{\min_i q_i}$, and $\boldsymbol{e}_1$ denotes the first standard basis vector. This assumption is based on the rotational invariance of Gaussian distributions. Since the deterministic nature of $\{\boldsymbol{b}_j\}$, the ground truth signals $\{\boldsymbol{h}_i^\natural\}$ (channel vectors) cannot be transferred to a simple form, which yields more tedious analysis procedure. For simplification, for $i = 1,\cdots,s$, we denote

$$x_{i1}^t \qquad \text{and} \qquad \boldsymbol{x}_{i\perp}^t := [x_{ij}^t]_{2\le j\le N}, \tag{16}$$

as the first entry and the second through the N^{th} entries of $\boldsymbol{x}_i^t$, respectively. Based on the assumption that $\boldsymbol{x}_i^\natural = q_i\boldsymbol{e}_1$ for $i = 1,\cdots,s$, (13) and (14) can be reformulated as

$$\alpha_{\boldsymbol{x}_i} := \widetilde{x}_{i1}^t \quad \text{and} \quad \beta_{\boldsymbol{x}_i} := \left\|\widetilde{\boldsymbol{x}}_{i\perp}^t\right\|_2. \tag{17}$$

To study the population-level state evolution, we start by considering the case where the sequences $\{\boldsymbol{z}_i^t\}$ (refer to (3)) are established via the population gradient, i.e., for $i = 1,\cdots,s$,

$$\begin{bmatrix}\boldsymbol{h}_i^{t+1}\\ \boldsymbol{x}_i^{t+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{h}_i^{t}\\ \boldsymbol{x}_i^{t}\end{bmatrix} - \eta\begin{bmatrix}\frac{1}{\|\boldsymbol{x}_i^t\|_2^2}\nabla_{\boldsymbol{h}_i}F(\boldsymbol{z}^t)\\ \frac{1}{\|\boldsymbol{h}_i^t\|_2^2}\nabla_{\boldsymbol{x}_i}F(\boldsymbol{z}^t)\end{bmatrix}, \tag{18}$$

where

$$\nabla_{\boldsymbol{h}_i}F(\boldsymbol{z}) := \mathbb{E}[\nabla_{\boldsymbol{h}_i}f(\boldsymbol{h},\boldsymbol{x})] = \|\boldsymbol{x}_i\|_2^2\,\boldsymbol{h}_i - (\boldsymbol{x}_i^{\natural *}\boldsymbol{x}_i)\boldsymbol{h}_i^\natural,$$

$$\nabla_{\boldsymbol{x}_i}F(\boldsymbol{z}) := \mathbb{E}[\nabla_{\boldsymbol{x}_i}f(\boldsymbol{h},\boldsymbol{x})] = \|\boldsymbol{h}_i\|_2^2\,\boldsymbol{x}_i - (\boldsymbol{h}_i^{\natural *}\boldsymbol{h}_i)\boldsymbol{x}_i^\natural.$$

Here, the population gradients are computed based on the assumption that $\{\boldsymbol{x}_i\}$ (resp. $\{\boldsymbol{h}_i\}$) and $\{\boldsymbol{a}_{ij}\}$ (resp. $\{\boldsymbol{b}_j\}$) are independent with each other. With simple calculations, the dynamics for both the signal and the perpendicular components with respect to $\boldsymbol{x}_i^t$, $i = 1,\cdots,s$ are given as:

$$\widetilde{x}_{i1}^{t+1} = (1-\eta)\,\widetilde{x}_{i1}^t + \eta\frac{q_i^2}{\|\widetilde{\boldsymbol{h}}_i^t\|_2^2}\boldsymbol{h}_i^{\natural *}\widetilde{\boldsymbol{h}}_i^t, \tag{19a}$$

$$\widetilde{\boldsymbol{x}}_{i\perp}^{t+1} = (1-\eta)\,\widetilde{\boldsymbol{x}}_{i\perp}^t. \tag{19b}$$

Assuming that $\eta > 0$ is sufficiently small and $\|\boldsymbol{h}_i^\natural\|_2 = \|\boldsymbol{x}_i^\natural\|_2 = q_i$ $(0 < q_i \le 1)$ for $i = 1,\cdots,s$ and recognizing that $\|\widetilde{\boldsymbol{h}}_i^t\|_2^2 = \alpha_{\boldsymbol{h}_i^t}^2 + \beta_{\boldsymbol{h}_i^t}^2$, we arrive at the following population-level state evolution for both $\alpha_{\boldsymbol{x}_i^t}$ and $\beta_{\boldsymbol{x}_i^t}$:

$$\alpha_{\boldsymbol{x}_i^{t+1}} = (1-\eta)\alpha_{\boldsymbol{x}_i^t} + \eta\frac{q_i\alpha_{\boldsymbol{h}_i^t}}{\alpha_{\boldsymbol{h}_i^t}^2 + \beta_{\boldsymbol{h}_i^t}^2}, \tag{20a}$$

$$\beta_{\boldsymbol{x}_i^{t+1}} = (1-\eta)\beta_{\boldsymbol{x}_i^t}. \tag{20b}$$

Likewise, the population-level state evolution for both $\alpha_{\boldsymbol{h}_i^t}$ and $\beta_{\boldsymbol{h}_i^t}$:

$$\alpha_{\boldsymbol{h}_i^{t+1}} = (1-\eta)\alpha_{\boldsymbol{h}_i^t} + \eta\frac{q_i\alpha_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{x}_i^t}^2 + \beta_{\boldsymbol{x}_i^t}^2}, \tag{21a}$$

$$\beta_{\boldsymbol{h}_i^{t+1}} = (1-\eta)\beta_{\boldsymbol{h}_i^t}. \tag{21b}$$

In finite-sample case, the dynamics of the randomly initialized WF iterates can be represented as

$$\boldsymbol{z}_i^{t+1} = \begin{bmatrix} \boldsymbol{h}_i^{t+1} \\ \boldsymbol{x}_i^{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}_i^t - \eta/\|\boldsymbol{x}_i^t\|_2^2 \cdot \nabla_{\boldsymbol{h}_i} F(\boldsymbol{z}) \\ \boldsymbol{x}_i^t - \eta/\|\boldsymbol{x}_i^t\|_2^2 \cdot \nabla_{\boldsymbol{x}_i} F(\boldsymbol{z}) \end{bmatrix} - \\ - \begin{bmatrix} \eta/\|\boldsymbol{x}_i^t\|_2^2 \cdot (\nabla_{\boldsymbol{h}_i} f(\boldsymbol{z}) - \nabla_{\boldsymbol{h}_i} F(\boldsymbol{z})) \\ \eta/\|\boldsymbol{h}_i^t\|_2^2 \cdot (\nabla_{\boldsymbol{x}_i} f(\boldsymbol{z}) - \nabla_{\boldsymbol{x}_i} F(\boldsymbol{z})) \end{bmatrix}. \tag{22}$$

Under the assumption that the last term in (22) is well-controlled, which will be justified in Section D in the supplemental material, we arrive at the approximate state evolution:

$$\alpha_{\boldsymbol{h}_i}^{t+1} = (1-\eta+\frac{\eta q_i \psi_{\boldsymbol{h}_i^t}}{\alpha_{\boldsymbol{x}_i^t}^2+\beta_{\boldsymbol{x}_i^t}^2})\alpha_{\boldsymbol{h}_i^t} + \eta(1-\rho_{\boldsymbol{h}_i^t})\frac{q_i\alpha_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{x}_i^t}^2+\beta_{\boldsymbol{x}_i^t}^2}, \tag{23a}$$

$$\beta_{\boldsymbol{h}_i}^{t+1} = (1-\eta+\frac{\eta q_i \varphi_{\boldsymbol{h}_i^t}}{\alpha_{\boldsymbol{x}_i^t}^2+\beta_{\boldsymbol{x}_i^t}^2})\beta_{\boldsymbol{h}_i^t}, \tag{23b}$$

$$\alpha_{\boldsymbol{x}_i}^{t+1} = (1-\eta+\frac{\eta q_i \psi_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{h}_i^t}^2+\beta_{\boldsymbol{h}_i^t}^2})\alpha_{\boldsymbol{x}_i^t} + \eta(1-\rho_{\boldsymbol{x}_i^t})\frac{q_i\alpha_{\boldsymbol{h}_i^t}}{\alpha_{\boldsymbol{h}_i^t}^2+\beta_{\boldsymbol{h}_i^t}^2}, \tag{23c}$$

$$\beta_{\boldsymbol{x}_i}^{t+1} = (1-\eta+\frac{\eta q_i \varphi_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{h}_i^t}^2+\beta_{\boldsymbol{h}_i^t}^2})\beta_{\boldsymbol{x}_i^t}, \tag{23d}$$

where $\{\psi_{\boldsymbol{h}_i^t}\}, \{\psi_{\boldsymbol{x}_i^t}\}, \{\varphi_{\boldsymbol{h}_i^t}\}, \{\varphi_{\boldsymbol{x}_i^t}\}, \{\rho_{\boldsymbol{h}_i^t}\}$ and $\{\rho_{\boldsymbol{x}_i^t}\}$ represent the perturbation terms.

3.2 Dynamics of Approximate State Evolution

It is easily seen that if $\alpha_{\boldsymbol{h}_i^t}$ (11), $\beta_{\boldsymbol{h}_i^t}$ (12), $\alpha_{\boldsymbol{x}_i^t}$ (13) and $\beta_{\boldsymbol{x}_i^t}$ (14) obey

$$\begin{aligned} &|\alpha_{\boldsymbol{h}_i^t} - q_i| \le \frac{\sqrt{2}\gamma}{4\kappa\sqrt{s}} \quad \text{and} \quad \beta_{\boldsymbol{h}_i^t} \le \frac{\sqrt{2}\gamma}{4\kappa\sqrt{s}} \quad \text{and} \\ &|\alpha_{\boldsymbol{x}_i^t} - q_i| \le \frac{\sqrt{2}\gamma}{4\kappa\sqrt{s}} \quad \text{and} \quad \beta_{\boldsymbol{x}_i^t} \le \frac{\sqrt{2}\gamma}{4\kappa\sqrt{s}}, \end{aligned} \tag{24}$$

for $i = 1, \cdots, s$, then

$$\operatorname{dist}(\boldsymbol{z}, \boldsymbol{z}^\natural) \le \left[\frac{s\kappa^2}{2}\left(|\alpha_{\boldsymbol{h}_i^t} - q_i| + |\beta_{\boldsymbol{h}_i^t}|\right)^2 + \frac{s\kappa^2}{2}\left(|\alpha_{\boldsymbol{x}_i^t} - q_i| + |\beta_{\boldsymbol{x}_i^t}|\right)^2\right]^{1/2} \le \gamma. \tag{25}$$

In this subsection, we shall show that as long as the approximate state evolution (23) holds, there exists some constant $T_\gamma = \mathcal{O}(s\log\max\{K,N\})$ satisfying condition (24). This is demonstrated in the following Lemma. Prior to that, we first list several conditions and definitions that contribute to the lemma.

- The initial points obey

$$\alpha_{\boldsymbol{h}_i^0} \ge \frac{q_i}{K\log K} \quad \text{and} \quad \alpha_{\boldsymbol{x}_i^0} \ge \frac{q_i}{N\log N}, \tag{26a}$$

$$\sqrt{\alpha_{\boldsymbol{h}_i^0}^2 + \beta_{\boldsymbol{h}_i^0}^2} \in \left[1 - \frac{1}{\log K}, 1 + \frac{1}{\log K}\right] q_i, \tag{26b}$$

$$\sqrt{\alpha_{\boldsymbol{x}_i^0}^2 + \beta_{\boldsymbol{x}_i^0}^2} \in \left[1 - \frac{1}{\log N}, 1 + \frac{1}{\log N}\right] q_i, \tag{26c}$$

  for $i = 1, \cdots, s$.

- Define

$$T_\gamma := \min\left\{t : \text{satifes (24)}\right\}, \tag{27}$$

  where $\gamma > 0$ is some sufficiently small constant.

- Define

$$T_1 := \min\Big\{t : \min_i \frac{\alpha_{\boldsymbol{h}_i^t}}{q_i} \ge \frac{c_7}{\log^5 m}, \ \min_i \frac{\alpha_{\boldsymbol{x}_i^t}}{q_i} \ge \frac{c_7'}{\log^5 m}\Big\}, \tag{28}$$

$$T_2 := \min\left\{t : \min_i \frac{\alpha_{\boldsymbol{h}_i^t}}{q_i} > c_8, \ \min_i \frac{\alpha_{\boldsymbol{x}_i^t}}{q_i} > c_8'\right\}, \tag{29}$$

  for some small absolute positive constants $c_7, c_7', c_8, c_8' > 0$.

- For $0 \le t \le T_\gamma$, it has

$$\begin{aligned} &\frac{1}{2\sqrt{K\log K}} \le \frac{\alpha_{\boldsymbol{h}_i^t}}{q_i} \le 2, \ c_5 \le \frac{\beta_{\boldsymbol{h}_i^t}}{q_i} \le 1.5 \quad \text{and} \\ &\frac{\alpha_{\boldsymbol{h}_i^{t+1}}/\alpha_{\boldsymbol{h}_i^t}}{\beta_{\boldsymbol{h}_i^{t+1}}/\beta_{\boldsymbol{h}_i^t}} \ge 1 + c_5\eta, \ i = 1, \cdots, s, \end{aligned} \tag{30}$$

$$\begin{aligned} &\frac{1}{2\sqrt{N\log N}} \le \frac{\alpha_{\boldsymbol{x}_i^t}}{q_i} \le 2, \ c_6 \le \frac{\beta_{\boldsymbol{x}_i^t}}{q_i} \le 1.5 \quad \text{and} \\ &\frac{\alpha_{\boldsymbol{x}_i^{t+1}}/\alpha_{\boldsymbol{x}_i^t}}{\beta_{\boldsymbol{x}_i^{t+1}}/\beta_{\boldsymbol{x}_i^t}} \ge 1 + c_6\eta, \ i = 1, \cdots, s, \end{aligned} \tag{31}$$

  for some constants $c_5, c_6 > 0$.

Lemma 1. *Assume that the initial points obey condition (26) and the perturbation terms in the approximate state evolution (23) obey* $\max\left\{|\psi_{\boldsymbol{h}_i^t}|, |\psi_{\boldsymbol{x}_i^t}|, |\varphi_{\boldsymbol{h}_i^t}|, |\varphi_{\boldsymbol{x}_i^t}|, |\rho_{\boldsymbol{x}_i^t}|\right\} \le \frac{c}{\log m}$, *for* $i = 1, \cdots, s$, $t = 0, 1, \cdots$ *and some sufficiently small constant* $c > 0$.

1. *Then for any sufficiently large* K, N *and the step-size* $\eta > 0$ *that obeys* $\eta \asymp s^{-1}$, *it follows*

$$T_\gamma \lesssim s\log(\max\{K,N\}), \tag{32}$$

   and (30), (31).

2. *Then with the stepsize $\eta > 0$ following $\eta \asymp s^{-1}$, one has that $T_1 \le T_2 \le T_\gamma \lesssim s \log \max\{K, N\}$, $T_2 - T_1 \lesssim s \log\log m$, $T_\gamma - T_2 \lesssim s$.*

Proof. Please refer to Appendix C in the supplemental material. □

The random initialization (8) satisfies the condition (26) with probability at least $1 - \mathcal{O}(1/\sqrt{\log \min\{K, N\}})$ [16]. According to this fact, Lemma 1 ensures that under both random initialization (8) and approximate state evolution (23) with the stepsize $\eta \asymp s^{-1}$, Stage I only lasts a few iterations, i.e., $T_\gamma = \mathcal{O}(s \log \max\{K, N\})$. In addition, Lemma 1 demonstrates the exponential growth of the ratios, i.e., $\alpha_{\boldsymbol{h}_i^{t+1}}/\alpha_{\boldsymbol{h}_i^t}, \beta_{\boldsymbol{h}_i^{t+1}}/\beta_{\boldsymbol{h}_i^t}$, which contributes to the short duration of Stage I.

Moreover, Lemma 1 defines the midpoints T_1 when the sizes of the signal component, i.e., $\alpha_{\boldsymbol{h}_i^t}$ and $\alpha_{\boldsymbol{x}_i^t}$, $i = 1, \cdots, s$, become sufficiently large, which is crucial to the following analysis. In particular, when establishing the approximate state evolution (23) in Stage I, we analyze two subphases of Stage I individually:

- Phase 1: consider the iterations in $0 \le t \le T_1$,
- Phase 2: consider the iterations in $T_1 < t \le T_\gamma$,

where T_1 is defined in (28).

where T_1 is defined in (28).

3.3 Leave-One-Out Approach

According to Section 3.1 and Lemma 1, the unique challenge for establishing the approximate state evolution (23) is to bound the perturbation terms to certain order, i.e., $|\psi_{\boldsymbol{h}_i^t}|, |\psi_{\boldsymbol{x}_i^t}|, |\varphi_{\boldsymbol{h}_i^t}|, |\varphi_{\boldsymbol{x}_i^t}|, |\rho_{\boldsymbol{h}_i^t}|, |\rho_{\boldsymbol{x}_i^t}| \ll 1/\log m$ for $i = 1, \cdots, s$. To achieve this goal, we exploit some variants of leave-one-out sequences [16] to establish the "near-independence" between $\{\boldsymbol{z}_i^t\}$ and $\{\boldsymbol{a}_i\}$. Hence, some terms can be approximated by a sum of independent variables with well-controlled weights, thereby be controlled via central limit theorem.

In the following, we define three sets of auxiliary sequences $\{\boldsymbol{z}^{t,(l)}\}$, $\{\boldsymbol{z}^{t,\mathrm{sgn}}\}$ and $\{\boldsymbol{z}^{t,\mathrm{sgn},(l)}\}$, respectively.

- *Leave-one-out sequences* $\{\boldsymbol{z}^{t,(l)}\}_{t\ge 0}$. For each $1 \le l \le m$, the auxiliary sequences $\{\boldsymbol{z}^{t,(l)}\}$ are established by dropping the l^{th} sample and runs randomly initialized WF with objective function

$$f^{(l)}(\boldsymbol{z}) = \sum_{j:j\neq l} \Big| \sum_{i=1}^{s} \boldsymbol{b}_j^* \boldsymbol{h}_i \boldsymbol{x}_i^* \boldsymbol{a}_{ij} - y_j \Big|^2. \tag{33}$$

  Thus, the sequences $\{\boldsymbol{z}_i^{t,(l)}\}$ (recall the definition of $\boldsymbol{z}_i$ (3)) are statistically independent of $\{\boldsymbol{a}_{il}\}$.

- *Random-sign sequences* $\{\boldsymbol{z}^{t,\mathrm{sgn}}\}_{t\ge 0}$. Define the auxiliary design vectors $\{\boldsymbol{a}_{ij}^{\mathrm{sgn}}\}$ as

$$\boldsymbol{a}_{ij}^{\mathrm{sgn}} := \begin{bmatrix} \xi_{ij} a_{ij,1} \\ \boldsymbol{a}_{ij,\perp} \end{bmatrix}, \tag{34}$$

  where $\{\xi_{ij}\}$ is a set of standard complex uniform random variables independent of $\{\boldsymbol{a}_{ij}\}$, i.e.,

$$\xi_{ij} \overset{\text{i.i.d.}}{=} u/|u|, \tag{35}$$

  where $u \sim \mathcal{N}(0, \frac{1}{2}) + i\mathcal{N}(0, \frac{1}{2})$. Moreover, with the corresponding ξ_{ij}, the auxiliary design vector $\{\boldsymbol{b}_j^{\mathrm{sgn}}\}$ is defined as $\boldsymbol{b}_j^{\mathrm{sgn}} = \xi_{ij}\boldsymbol{b}_j$. With these auxiliary design vectors, the sequences $\{\boldsymbol{z}^{t,\mathrm{sgn}}\}$ are generated by running randomly initialized WF with respect to the loss function $f^{\mathrm{sgn}}(\boldsymbol{z})$

$$\sum_{j=1}^{m} \Big| \sum_{i=1}^{s} \boldsymbol{b}_j^{\mathrm{sgn}*} \boldsymbol{h}_i \boldsymbol{x}_i^* \boldsymbol{a}_{ij}^{\mathrm{sgn}} - \boldsymbol{b}_j^{\mathrm{sgn}*} \boldsymbol{h}_i^\natural \boldsymbol{x}_i^{\natural *} \boldsymbol{a}_{ij}^{\mathrm{sgn}} \Big|^2. \tag{36}$$

  Note that these auxiliary design vectors, i.e., $\{\boldsymbol{a}_{ij}^{\mathrm{sgn}}\}, \{\boldsymbol{b}_j^{\mathrm{sgn}}\}$ produce the same measurements as $\{\boldsymbol{a}_{ij}\}, \{\boldsymbol{b}_j\}$, i.e., $\boldsymbol{b}_j^{\mathrm{sgn}*} \boldsymbol{h}_i^\natural \boldsymbol{x}_i^{\natural *} \boldsymbol{a}_{ij}^{\mathrm{sgn}} = \boldsymbol{b}_j^* \boldsymbol{h}_i^\natural \boldsymbol{x}_i^{\natural *} \boldsymbol{a}_{ij} = q_i a_{ij,1} \boldsymbol{b}_j^* \boldsymbol{h}_i^\natural$ for $1 \le i \le s, 1 \le j \le m$.

Note that all the auxiliary sequences are assumed to have the same initial point, namely, for $1 \le l \le m$, $\{\boldsymbol{z}^0\} = \{\boldsymbol{z}^{0,(l)}\} = \{\boldsymbol{z}^{0,\mathrm{sgn}}\} = \{\boldsymbol{z}^{0,\mathrm{sgn},(l)}\}$.

In view of the ambiguities, i.e., $\boldsymbol{h}_i^\natural \boldsymbol{x}_i^\natural = \frac{1}{\omega} \boldsymbol{h}_i^\natural (\omega \boldsymbol{x}_i^\natural)^*$, several alignment parameters are further defined for the sequel analysis. Specifically, the alignment parameter between $\boldsymbol{z}_i^{t,(l)} = [\boldsymbol{h}_i^{t,(l)*} \quad \boldsymbol{x}_i^{t,(l)*}]^*$ and $\widetilde{\boldsymbol{z}}_i^t = [\widetilde{\boldsymbol{h}}_i^{t*} \quad \widetilde{\boldsymbol{x}}_i^{t*}]^*$, where $\widetilde{\boldsymbol{h}}_i^t = \frac{1}{\omega_i^t} \boldsymbol{h}_i^t$ and $\widetilde{\boldsymbol{x}}_i^t = \omega_i^t \boldsymbol{x}_i^t$, is represented as

$$\begin{aligned} \omega_{i,\mathrm{mutual}}^{t,(l)} := \arg\min_{\omega\in\mathbb{C}} & \left\| \frac{1}{\overline{\omega}} \boldsymbol{h}_i^{t,(l)} - \frac{1}{\overline{\omega_i^t}} \boldsymbol{h}_i^t \right\|_2^2 \\ & + \left\| \omega \boldsymbol{x}_i^{t,(l)} - \omega_i^t \boldsymbol{x}_i^t \right\|_2^2, \end{aligned} \tag{37}$$

for $i = 1, \cdots, s$. In addition, we denote $\widehat{\boldsymbol{z}}_i^{t,(l)} = [\widehat{\boldsymbol{h}}_i^{t,(l)*} \ \widehat{\boldsymbol{x}}_i^{t,(l)*}]^*$ where

$$\widehat{\boldsymbol{h}}_i^{t,(l)} := \frac{1}{\overline{\omega_{i,\mathrm{mutual}}^{t,(l)}}} \boldsymbol{h}_i^{t,(l)} \quad \text{and} \quad \widehat{\boldsymbol{x}}_i^{t,(l)} := \omega_{i,\mathrm{mutual}}^{t,(l)} \boldsymbol{x}_i^{t,(l)}. \tag{38}$$

Define the alignment parameter between $\boldsymbol{z}_i^{t,\mathrm{sgn}} =$

$[\boldsymbol{h}_i^{t,\mathrm{sgn}*} \ \ \boldsymbol{x}_i^{t,\mathrm{sgn}*}]^*$ and $\boldsymbol{z}_i^t = [\boldsymbol{h}_i^{t*} \ \ \boldsymbol{x}_i^{t*}]^*$ as

$$\omega_{i,\mathrm{sgn}}^t := \arg\min_{\omega\in\mathbb{C}} \left\| \frac{1}{\overline{\omega}}\boldsymbol{h}_i^{t,\mathrm{sgn}} - \frac{1}{\overline{\omega_i^t}}\boldsymbol{h}_i^t \right\|_2^2 + \left\| \omega \boldsymbol{x}_i^{t,\mathrm{sgn}} - \omega_i^t \boldsymbol{x}_i^t \right\|_2^2, \tag{39}$$

for $i = 1, \cdots, s$. In addition, we denote $\check{\boldsymbol{z}}_i^{t,\mathrm{sgn}} = [\check{\boldsymbol{h}}_i^{t,\mathrm{sgn}*} \ \check{\boldsymbol{x}}_i^{t,\mathrm{sgn}*}]^*$ where

$$\check{\boldsymbol{h}}_i^{t,\mathrm{sgn}} := \frac{1}{\overline{\omega_{i,\mathrm{sgn}}^t}}\boldsymbol{h}_i^{t,\mathrm{sgn}} \quad \text{and} \quad \check{\boldsymbol{x}}_i^{t,\mathrm{sgn}} := \omega_{i,\mathrm{sgn}}^t \boldsymbol{x}_i^{t,\mathrm{sgn}}. \tag{40}$$

3.4 Induction Hypotheses

In this subsection, we shall establish a collection of induction hypotheses which are crucial to the justification of approximate state evolution (23). We list all the induction hypotheses: for $1 \le i \le s$,

$$\max_{1\le l\le m} \mathrm{dist}\left(\boldsymbol{z}_i^{t,(l)}, \widetilde{\boldsymbol{z}}_i^t\right) \le (\beta_{\boldsymbol{h}_i^t} + \beta_{\boldsymbol{x}_i^t}) \left(1 + \frac{1}{s\log m}\right)^t C_1 \frac{s\mu^2\kappa\sqrt{\max\{K,N\}\log^8 m}}{m} \tag{41a}$$

$$\max_{1\le l\le m} \mathrm{dist}\left(\boldsymbol{h}_i^{\natural *}\boldsymbol{h}_i^{t,(l)}, \boldsymbol{h}_i^{\natural *}\widetilde{\boldsymbol{h}}_i^t\right) \cdot \|\boldsymbol{h}_i^\natural\|_2^{-1} \le \alpha_{\boldsymbol{h}_i^t} \left(1 + \frac{1}{s\log m}\right)^t C_2 \frac{s\mu^2\kappa\sqrt{K\log^{13} m}}{m} \tag{41b}$$

$$\max_{1\le l\le m} \mathrm{dist}\left(x_{i1}^{t,(l)}, \widetilde{x}_{i1}^t\right) \le \alpha_{\boldsymbol{x}_i^t} \left(1 + \frac{1}{s\log m}\right)^t C_2 \frac{s\mu^2\kappa\sqrt{N\log^{13} m}}{m} \tag{41c}$$

$$\max_{1\le i\le s} \mathrm{dist}\left(\boldsymbol{h}_i^{t,\mathrm{sgn}}, \widetilde{\boldsymbol{h}}_i^t\right) \le \alpha_{\boldsymbol{h}_i^t} \left(1 + \frac{1}{s\log m}\right)^t C_3 \sqrt{\frac{s\mu^2\kappa^2 K\log^8 m}{m}} \tag{41d}$$

$$\max_{1\le i\le s} \mathrm{dist}\left(\boldsymbol{x}_i^{t,\mathrm{sgn}}, \widetilde{\boldsymbol{x}}_i^t\right) \le \alpha_{\boldsymbol{x}_i^t} \left(1 + \frac{1}{s\log m}\right)^t C_3 \sqrt{\frac{s\mu^2\kappa^2 N\log^8 m}{m}} \tag{41e}$$

$$\max_{1\le l\le m} \left\| \widetilde{\boldsymbol{h}}_i^t - \widehat{\boldsymbol{h}}_i^{t,(l)} - \widetilde{\boldsymbol{h}}_i^{t,\mathrm{sgn}} + \widehat{\boldsymbol{h}}_i^{t,\mathrm{sgn},(l)} \right\|_2 \le \alpha_{\boldsymbol{h}_i^t} \left(1 + \frac{1}{s\log m}\right)^t C_4 \frac{s\mu^2\sqrt{K\log^{16} m}}{m}, \tag{41f}$$

$$\max_{1\le l\le m} \left\| \widetilde{\boldsymbol{x}}_i^t - \widehat{\boldsymbol{x}}_i^{t,(l)} - \widetilde{\boldsymbol{x}}_i^{t,\mathrm{sgn}} + \widehat{\boldsymbol{x}}_i^{t,\mathrm{sgn},(l)} \right\|_2 \le \alpha_{\boldsymbol{x}_i^t} \left(1 + \frac{1}{s\log m}\right)^t C_4 \frac{s\mu^2\sqrt{N\log^{16} m}}{m}, \tag{41g}$$

$$c_5 \le \left\|\boldsymbol{h}_i^t\right\|_2, \left\|\boldsymbol{x}_i^t\right\|_2 \le C_5, \tag{41h}$$

$$\left\|\boldsymbol{h}_i^t\right\|_2 \le 5\alpha_{\boldsymbol{h}_i^t}\sqrt{\log^5 m}, \tag{41i}$$

$$\left\|\boldsymbol{x}_i^t\right\|_2 \le 5\alpha_{\boldsymbol{x}_i^t}\sqrt{\log^5 m}, \tag{41j}$$

where $C_1, \cdots, C_5$ and c_5 are some absolute positive constants and $\widehat{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_i, \widehat{\boldsymbol{h}}_i, \widetilde{\boldsymbol{h}}_i$ are defined in Section 3.3.

Specifically, (41a), (41c), (41d) and (41e) identify that the auxiliary sequences $\{\boldsymbol{z}^{t,(l)}\}$ and $\{\boldsymbol{z}^{t,\mathrm{sgn}}\}$ are extremely close to the original sequences $\{\boldsymbol{z}^t\}$. In addition, as claimed in (41f) and (41g), $\widetilde{\boldsymbol{h}}_i^t - \widetilde{\boldsymbol{h}}_i^{t,\mathrm{sgn}}$ (resp. $\widetilde{\boldsymbol{x}}_i^t - \widetilde{\boldsymbol{x}}_i^{t,\mathrm{sgn}}$) and $\widehat{\boldsymbol{h}}_i^{t,(l)} - \widehat{\boldsymbol{h}}_i^{t,\mathrm{sgn},(l)}$ (resp. $\widehat{\boldsymbol{x}}_i^{t,(l)} - \widehat{\boldsymbol{x}}_i^{t,\mathrm{sgn},(l)}$) are also exceedingly close to each other. The hypotheses (41h) illustrates that the norm of the iterates $\{\boldsymbol{h}_i^t\}$ (resp. $\{\boldsymbol{x}_i^t\}$) is well-controlled in Phase 1. Moreover, (41i) (resp. (41j)) indicates that $\alpha_{\boldsymbol{h}_i^t}$ (resp. $\alpha_{\boldsymbol{x}_i^t}$) is comparable to $\|\boldsymbol{h}_i^t\|_2$ (resp. $\|\boldsymbol{x}_i^t\|_2$).

Lemma 4-Lemma 9 in the supplemental material demonstrate that if induction hypotheses (41) for $1 \le i \le s$ hold true up to the t^{th} iteration, then they are still satisfied in the $(t+1)^{\mathrm{th}}$ iteration under the condition of sufficient samples. Thus auxiliary sequences are close sufficient to the original sequences, which contribute to bound the last term in (22) thereby validating the approximate evolution (23).

4 CONCLUSION

In this paper, we develop a least square estimation approach to solve blind demixing problem. To address the limitations of state-of-the-art algorithms, e.g., high computational complexity, low convergence rate and requirement of careful initialization, we developed the randomly initialized Wirtinger flow to solve the blind demixing problem. The global convergence guarantee of this algorithm is further provided in this paper. It shows that iterates of the randomly initialized Wirtinger flow can enter the local region that enjoys strong convexity and strong smoothness within a few iterations, followed by linearly converging to the ground truth.

Acknowledgements

This work was supported in part by the National Nature Science Foundation of China under Grant 61601290 and the Shanghai Sailing Program under Grant 16YF1407700.

References

[1] S. Ling and T. Strohmer, "Blind deconvolution meets blind demixing: Algorithms and performance bounds," *IEEE Trans. Inf. Theory*, vol. 63, pp. 4497–4520, Jul. 2017.

[2] J. Dong and Y. Shi, "Nonconvex demixing from bilinear measurements," *IEEE Trans. Signal Process.*, vol. 66, pp. 5152–5166, Oct. 2018.

[3] Y. Zhang, Y. Lau, H.-w. Kuo, S. Cheung, A. Pasupathy, and J. Wright, "On the global geometry of sphere-constrained sparse blind deconvolution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4894–4902, Jun. 2017.

[4] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang, *et al.*, "Simultaneous denoising, deconvolution, and demixing of calcium imaging data," *Neuron*, vol. 89, pp. 285–299, Jan. 2016.

[5] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 391–398, 2013.

[6] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2528–2535, Jun. 2010.

[7] J. Dong and Y. Shi, "Blind demixing for low-latency communication," *IEEE Trans. Wireless Commun.*, Dec. 2018.

[8] X. Wang and H. V. Poor, "Blind equalization and multiuser detection in dispersive CDMA channels," *IEEE Trans. Commun.*, vol. 46, pp. 91–103, Jan. 1998.

[9] S. Ling and T. Strohmer, "Regularized gradient descent: A nonconvex recipe for fast joint blind deconvolution and demixing," *Inf. Inference: J. IMA*, Mar. 2018.

[10] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points-online stochastic gradient for tensor decomposition.," in *28th Annu. Proc. Conf. Learn. Theory (COLT)*, pp. 797–842, Jul. 2015.

[11] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Found. Comput. Math*, vol. 18, pp. 1131–1198, Oct. 2018.

[12] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017.

[13] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Adv. Neural. Inf. Process. Syst. (NIPS)*, pp. 3873–3881, 2016.

[14] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1233–1242, 2017.

[15] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Trans. Inf. Theory*, Jul. 2018.

[16] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Math. Program.*, to appear, 2018.

[17] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, pp. 1985–2007, Apr. 2015.

[18] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 3345–3354, Jul. 2018.

[19] K. Pothoven, *Real and functional analysis*. Springer Science & Business Media, 2013.