Blind Over-the-Air Computation and Data Fusion via Provable Wirtinger Flow

Jialin Dong, *Student Member, IEEE*, Yuanming Shi, *Member, IEEE*, and Zhi Ding, *Fellow, IEEE*

***Abstract*—Over-the-air computation (AirComp) shows great promise to support fast data fusion in Internet-of-Things (IoT) networks. AirComp typically computes desired functions of distributed sensing data by exploiting superposed data transmission in multiple access channels. To overcome its reliance on channel state information (CSI), this work proposes a novel *blind over-the-air computation* (BlairComp) without requiring CSI access, particularly for low complexity and low latency IoT networks. To solve the resulting non-convex optimization problem without the initialization dependency exhibited by the solutions of a number of recently proposed efficient algorithms, we develop a Wirtinger flow solution to the BlairComp problem based on *random initialization*. We establish the global convergence guarantee of Wirtinger flow with random initialization for BlairComp problem, which enjoys a model-agnostic and natural initialization implementation for practitioners with theoretical guarantees. Specifically, in the first stage of the algorithm, the iteration of randomly initialized Wirtinger flow given sufficient data samples can enter a local region that enjoys strong convexity and strong smoothness within a few iterations. We also prove the estimation error of BlairComp in the local region to be sufficiently small. We show that, at the second stage of the algorithm, its estimation error decays exponentially at a linear convergence rate.**

***Index Terms*—Over-the-air computation, data fusion, bilinear measurements, Wirtinger flow, regularization-free, random initialization.**

I. Introduction

THE broad range of Internet-of-Things (IoT) applications continues to contribute substantially to the economic development and the improvement of our lives [1]. In particular, the wirelessly networked sensors are growing at an unprecedented rate, making data aggregation highly critical for IoT services [2]. For large scale wireless networking of sensor nodes, orthogonal multiple access protocols are highly impractical because of their low spectrum utilization efficiency for IoT and the excessive network latency [3]. In response, the concept of *over-the-air computation* (AirComp) has recently been considered for computing a class of nomographic functions, such as arithmetic mean, weighted sum, geometric mean and Euclidean norm of distributed sensor data via concurrent, instead of the sequential, node transmissions [4]. AirComp exploits the natural superposition of co-channel transmissions from multiple data source nodes [5].

There have already been a number of published works related to AirComp. Among them, one research thread takes on the information theoretic view and focuses on achievable computation rate under structured coding schemes. Specifically, in the seminal work of [6], linear source coding was designed to reliably compute a function of distributed sensor data transmitted over the multiple-access channels (MACs). Lattice codes were adopted in [6], [7] to compute the sum of source signals over MACs efficiently. Leveraging lattice coding, a compute-and-forward relaying scheme [5] was proposed for relay assisted networks. On the other hand, a different line of studies [8], [9] investigates the error of distributed estimation in wireless sensor networks. In particular, linear decentralized estimation was investigated in [8] for coherent multiple access channels. Power control was investigated in [9] to optimize the estimation distortion. It was shown in [10] that pre- and post-processing functions enable the optimization of computation performance by harnessing the interference for function computations. Another more recent line of studies focused on designing transmitter and receiver matrices in order to minimize the distortion error when computing desired functions. Among others, MIMO-AirComp equalization and channel feedback techniques for spatially multiplexing multi-function computation have been proposed [3]. Another work developed a novel transmitter design at the multiple antennas IoT devices with zero-forcing beamforming [11].

Most recently, the paper [12] integrated wireless power transfer into MIMO AirComp, thereby supporting self-sustainable AirComp for low-power devices. In addition, [13] proposed an intelligent reflecting surface (IRS) for AirComp to generate controllable wireless environments in order to improve received signal power. The AirComp has also play a vital role in various applications of wireless sensor networks, such as enabling low-latency global model aggregation to support large-scale distributed machine learning for edge AI in 6G [14]. Recently, the papers [15], [16] exploited AirComp in a federated learning system to support fast global model aggregation for locally updated model on each device. [17] further investigated the power control problem for AirComp over fading channels to adaptively control the devices' transmit power to combat channel distortion, thereby improving magnitude alignment of simultaneous signals.

Manuscript received August 2, 2019; revised December 12, 2019 and January 23, 2020; accepted January 24, 2020. Date of publication January 29, 2020; date of current version February 14, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. N. Gillis. This material is based upon the work of J. Dong and Y. Shi supported by by National Nature Science Foundation of China under Grant 61601290, and the work of Z. Ding supported by the National Science Foundation under Grant NSF 1824553. *(Corresponding author: Yuanming Shi.)*

J. Dong is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, and with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: dongjl@shanghaitech.edu.cn).

Y. Shi is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: shiym@shanghaitech.edu.cn).

Z. Ding is with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA 95616 USA (e-mail: zding@ucdavis.edu).

Digital Object Identifier 10.1109/TSP.2020.2970338

Similar to the key idea of AirComp, federated learning (FL) or the collaborative machine learning usually operates at a wireless edge network instead of in a data center [18]–[20]. The paper [19] studied wireless collaborative machine learning (ML), where mobile edge devices operate distributed stochastic gradient descent (DSGD) over-the-air with the help of a wireless access server. The channel state information (CSI) is available only at this wireless access server. Additionally, the paper [20] investigated machine learning at the wireless edge which contains power and bandwidth-limited devices (workers). The authors first introduced a digital DSGD (D-DSGD) scheme and an analog scheme, called A-DSGD to solve collaborative machine learning problem. The work [21] has recently considered a distributed learning problem over multiple access channel (MAC) where the objective function is a sum of the nodes' local loss functions. A novel Gradient-Based Multiple Access (GBMA) algorithm is developed to solve this distributed learning problem over MAC. Furthermore, to address the communication issue induced by the edge devices under the federated edge learning scheme, the paper [22] proposed a novel digital version of broadband over-the-air aggregation, called one-bit broadband digital aggregation (OBDA).

However, the main limitation of current AirComp is the dependence on channel-state-information (CSI), which leads to high latency and significant overhead in the massive Internet-of-Things networks with a large number of devices. Although in the works [19], [23] the channel state information is unknown to the transmitters, the receivers still need to obtain the channel state information. Recently, blind demixing has become a powerful tool to exclude channel-state-information (i.e., without channel estimation at both transmitters and receivers) thereby enabling low-overhead communications [24]–[26]. Specifically, in blind demixing, a sequence of source signals can be recovered from the sum of bilinear measurements without the knowledge of channel information [27]. Inspired by the recent progress of blind demixing, in this paper, we shall propose a novel *blind over-the-air computation* (BlairComp) scheme for low-overhead data aggregation, thereby computing the desired function (e.g., arithmetic mean) of sensing data vectors without the prior knowledge of channel information. The advantage of BlairComp is that without the extra cost of obtaining the CSI, BlairComp can achieve sufficiently low estimation error, which is illustrated in Fig. 1. It can also support low-overhead communications by excluding CSI from data packet transmission. However, the BlairComp problem turns out to be a highly intractable nonconvex optimization problem due to the bilinear signal model.

There is a growing body of recent works to tame the nonconvexity in solving the high-dimensional bilinear systems. Specifically, semidefinite programming was developed in [25] 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 [24], was further developed to optimize the variables in the natural vector space. Nevertheless, the theoretical guarantees for the regularized gradient [24] provide a pessimistic convergence rate and require carefully-designed initialization. The Riemannian trust-region optimization algorithm without regularization was further proposed in [26] to improve the convergence rate. However, the second-order algorithm brings unique challenges in providing statistical guarantees. Recently, theoretical

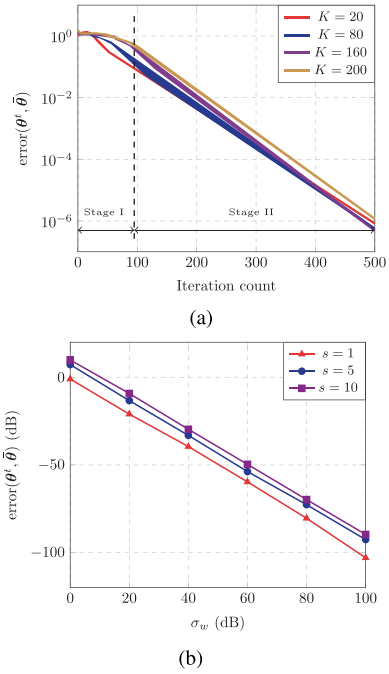


Fig. 1. (a) Linear convergence rate of randomly initialized Wirtinger flow, plotted semi-logarithmically. (b) Relative error $\mathrm{error}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})$ vs. σ_w (dB).

guarantees concerning regularization-free Wirtinger flow with spectral initialization for blind demixing was provided in [27]. However, this regularization-free method still calls for spectral initialization. To find a natural implementation for the practitioners that works equally well as spectral initialization, in this paper, we shall propose to solve the BlairComp problem via *randomly initialized Wirtinger flow* with provable optimality guarantees.

Based on the random initialization strategy, a line of research studies the benign global landscapes for the high-dimensional nonconvex estimation problems, followed by designing generic saddle-point escaping algorithms, e.g., noisy stochastic gradient descent [28], trust-region method [29], perturbed gradient descent [30]. With sufficient samples, these algorithms are guaranteed to converge globally for phase retrieval [29], matrix recovery [31], matrix sensing [32], robust PCA [32] and shallow neural networks [33], where all local minima are provably as good as global and all the saddle points are strict. However, the theoretical results developed in [28]–[33] 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 [34].

In this paper, our main contribution is to establish the global convergence guarantee of Wirtinger flow with random initialization for the BlairComp problem, which enjoys a model-agnostic and natural initialization implementation for practitioners with theoretical guarantees. It turns out that, for BlairComp, the procedure of Wirtinger flow with random initialization can be separated into two stages:

- Stage I: the estimation error is nearly stable, which takes only a few iterations,

- Stage II: the estimation error decays exponentially at a linear convergence rate.

In addition, we identify the exponential growth of the magnitude ratios of the signals to perpendicular components, which explains why Stage I lasts only for a few iterations. Compared with the theoretical analysis on the phase retrieval problem [34], the theoretical analysis on the BlairComp problem is much more complex and challenging. The primary challenge arises since the "incoherence" between multiple sources in BlairComp leads to distortion in the statistical property. Moreover, unlike the Gaussian designed vector $\boldsymbol{a}_j$ concerned in the phase retrieval problem, the designed vector $\boldsymbol{b}_j$ in the BlairComp problem is deterministic, not random. We will clarify the technical details exploited to address above issues in our paper in the sequel.

Notations: Throughout this paper, $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ denotes that there exists a constant $c > 0$ such that $|f(n)| \leq c|g(n)|$ whereas $f(n) \gtrsim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \geq c|g(n)|$. $f(n) \gg g(n)$ denotes that there exists some sufficiently large constant $c > 0$ such that $|f(n)| \geq c|g(n)|$. In addition, the notation $f(n) \asymp g(n)$ means that there exists constants $c_1, c_2 > 0$ such that $c_1|g(n)| \leq |f(n)| \leq c_2|g(n)|$. Let superscripts $(\cdot)^\top$ and $(\cdot)^\mathsf{H}$ denote the transpose and conjugate transpose of a matrix/vector, respectively. Let the superscript $(\cdot)^*$ denote the conjugate transpose of a complex number.

II. Problem Formulation

Blind over-the-air computation (BlairComp) aims to facilitate low-overhead data aggregation in IoT networks without a priori knowledge of CSI. This is achieved by computing the desired functions of the distributed sensing data based on the natural signal superposition of transmission over multi-access channels.

A. Blind Over-the-Air Computation

We consider a wireless sensor network consisting of s active sensor nodes and a single fusion center. Let $\boldsymbol{d}_i = [d_{i1} \ \ldots, d_{i,N}]^\top \in \mathbb{C}^N$ denote the sensor data vector collected at the i-th node. The fusion center, through AirComp, aims to compute nomographic functions of distributed data that can be decomposed as [10] $\mathcal{H}_\ell(d_{1\ell}, \ldots, d_{s\ell}) = \mathcal{F}_\ell(\sum_{i=1}^s \mathcal{G}_{i\ell}(d_{i\ell})), \ell = 1, \ldots, N$. Function $\mathcal{G}_{i\ell}(\cdot) : \mathbb{C} \to \mathbb{C}$ denotes the pre-processing function by the sensor nodes and $\mathcal{F}_\ell(\cdot) : \mathbb{C} \to \mathbb{C}$ denotes the post-processing function at the fusion center. Typical nomographic functions by AirComp include the arithmetic mean, weighted sum, geometric mean, polynomial, Euclidean norm [10].

In this work, we focus on a specific nomographic function

$$\bar{\boldsymbol{\theta}} = \sum_{i=1}^s \bar{\boldsymbol{x}}_i \tag{1}$$

where $\bar{\boldsymbol{x}}_i = [\mathcal{G}_{i1}(d_{i1}), \ldots, \mathcal{G}_{iN}(d_{iN})]^\top \in \mathbb{C}^N$ is the preprocessed data vector transmitted by the i-th node. The transmitted signals over m time slots from the i-th node are represented as

$$\boldsymbol{f}_i = \boldsymbol{C}_i \boldsymbol{x}_i, \tag{2}$$

where $\boldsymbol{C}_i \in \mathbb{C}^{m \times N}$ with $m > N$ is the encoding matrix and is known at the fusion center. The signals $\boldsymbol{f}_i$'s are transmitted through individual time-invariant channels denoted by their respective CSI vectors $\boldsymbol{h}_i$'s where a maximum delay of at most K samples is contained in $\boldsymbol{h}_i \in \mathbb{C}^K$. The zero-padded channel vector $\boldsymbol{g}_i \in \mathbb{C}^m$ is given as

$$\boldsymbol{g}_i = [\boldsymbol{h}_i^\top, 0, \ldots, 0]^\top. \tag{3}$$

Hence, based on the cyclic convolution operation, the received signal is given as

$$\boldsymbol{p} = \sum_{i=1}^s \boldsymbol{f}_i \circledast \boldsymbol{g}_i + \boldsymbol{n}, \tag{4}$$

where $\boldsymbol{n}$ is the additive white complex Gaussian noise. For ease of algorithm design and theoretical analysis, the blind demixing model based on cyclic convolution is presented in the Fourier domain. This is achieved by left multiplying the signals in the time domain with the unitary discrete Fourier transform (DFT) matrix, and converting the time domain convolution into component-wise production operation in the Fourier domain

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{p} = \sum_i (\boldsymbol{F}\boldsymbol{C}_i\boldsymbol{x}_i) \odot \boldsymbol{B}\boldsymbol{h}_i + \boldsymbol{F}\boldsymbol{n}, \tag{5}$$

where the operation $\odot$ is the component-wise product. Here, the first K columns of the unitary discrete Fourier transform (DFT) matrix $\boldsymbol{F} \in \mathbb{C}^{m \times m}$ satisfying property $\boldsymbol{F}\boldsymbol{F}^\mathsf{H} = \boldsymbol{I}_m$ form the known matrix

$$\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]^\mathsf{H} \in \mathbb{C}^{m \times K} \tag{6}$$

with $\boldsymbol{b}_j \in \mathbb{C}^K$ for $1 \leq j \leq m$. Hence, over m channel access opportunities (e.g., time slots), the received signals at fusion center in the frequency domain can be written as [19], [21]

$$y_j = \sum_{i=1}^s \boldsymbol{b}_j^\mathsf{H} \bar{\boldsymbol{h}}_i \bar{\boldsymbol{x}}_i^\mathsf{H} \boldsymbol{a}_{ij} + e_j, \ 1 \leq j \leq m, \tag{7}$$

where $\boldsymbol{b}_j \in \mathbb{C}^K$ for each $1 \leq j \leq m$ is an access vector, which means that the vector is accessible to the fusion center. Additionally, $\boldsymbol{a}_{ij} \in \mathbb{C}^N$ denotes the j-th column of $(\boldsymbol{F}\boldsymbol{C}_i)^\mathsf{H}$, $\bar{\boldsymbol{h}}_i \in \mathbb{C}^K$ is the CSI vector that contains channel gains, and e_j is an independent circularly symmetric complex Gaussian measurement noise.

To compute the desired functions via BlairComp without knowledge of $\{\bar{\boldsymbol{h}}_i\}$, we can consider a precoding scheme with randomly selected known vectors $\boldsymbol{a}_{ij} \in \mathbb{C}^N$ follows i.i.d. circularly symmetric complex normal distribution $\mathcal{N}(\boldsymbol{0}, 0.5\boldsymbol{I}_N) + \mathrm{i}\mathcal{N}(\boldsymbol{0}, 0.5\boldsymbol{I}_N)$ for $1 \leq i \leq s, 1 \leq j \leq m$. The target of BlairComp is to compute the desired function vector $\bar{\boldsymbol{\theta}}$ via concurrent transmissions without channel information, thereby providing low-overhead data aggregation in the IoT networks.

B. Multi-Dimensional Nonconvex Estimation

For each $1 \leq i \leq s$, $\bar{\boldsymbol{h}}_i$ and $\boldsymbol{h}_i$ denote the ground-truth CSI vector and the corresponding estimate, respectively. $\bar{\boldsymbol{x}}_i$ and $\boldsymbol{x}_i$ denote the ground-truth data vector generated by a node in the sensor network and the corresponding estimate, respectively.

BlairComp facilitates low-latency data aggregation in the IoT network, which aims to compute the desired function vector $\bar{\boldsymbol{\theta}} = \sum_{i=1}^s \bar{\boldsymbol{x}}_i$ via concurrent transmissions. Instead of concerning the sum of individual relative error of data vectors, i.e., $\sum_{i=1}^s \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i\|_2 / \|\bar{\boldsymbol{x}}_i\|$, the computational performance of BlairComp is characterized by the estimation error of the nomographic function $\bar{\boldsymbol{\theta}}$. To estimate the vector $\bar{\boldsymbol{\theta}}$ from the

received signal $\boldsymbol{y}$, we need to minimize the relative error between $\bar{\boldsymbol{\theta}}$ and the estimated vector $\boldsymbol{\theta} = \sum_{i=1}^{s} \omega_i \boldsymbol{x}_i$ which is denoted by

$$\mathrm{error}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}) = \frac{\|\sum_{i=1}^{s} \omega_i \boldsymbol{x}_i - \sum_{i=1}^{s} \bar{\boldsymbol{x}}_i\|_2}{\|\sum_{i=1}^{s} \bar{\boldsymbol{x}}_i\|_2}, \tag{8}$$

where $\omega_i \in \mathbb{C}$ alignment parameters that align the estimated vectors to the ground truth. The alignment parameters can be estimated via

$$\omega_i = \underset{\bar{\omega}_i \in \mathbb{C}}{\arg\min} \left(\|(\bar{\omega}_i^*)^{-1} \boldsymbol{h}_i - \bar{\boldsymbol{h}}_i\|_2^2 + \|\bar{\omega}_i \boldsymbol{x}_i - \bar{\boldsymbol{x}}_i\|_2^2 \right). \tag{9}$$

To estimate ω_i, one reference symbol in $\boldsymbol{x}_i$ is needed. One way to address this problem is to develop a bilinear estimation approach [26]:

$$\mathscr{P}: \underset{\{\boldsymbol{h}_i\},\{\boldsymbol{x}_i\}}{\text{minimize}} \; f(\boldsymbol{h}, \boldsymbol{x}) := \sum_{j=1}^{m} \left| \sum_{i=1}^{s} \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}_i \boldsymbol{x}_i^{\mathsf{H}} \boldsymbol{a}_{ij} - y_j \right|^2, \tag{10}$$

which estimates $\{\boldsymbol{h}_i\}$ and $\{\boldsymbol{x}_i\}$ from the sum of bilinear measurements $\boldsymbol{y}$. Even though problem $\mathscr{P}$ is nonconvex, some algorithms, e.g., Wirtinger flow with spectral initialization, can solve it with low statistical and computational guarantees [27].

In this paper, to find a model-agnostic and natural implementation for practitioners that works equally well as spectral initialization, we shall propose to solve BlairComp problem $\mathscr{P}$ via Wirtinger flow with *random initialization*. Our main contribution is to provide the statistical optimality and convergence guarantee for the randomly initialized Wirtinger flow algorithm by exploiting the benign geometry of the high-dimensional BlairComp problem.

III. Main Approach

In this section, we first propose an algorithm based on randomly initialized Wirtinger flow to solve the BlairComp problem $\mathscr{P}$. We shall present a statistical analysis to demonstrate the optimality of this algorithm for solving the high-dimensional nonconvex estimation problem.

A. Randomly Initialized Wirtinger Flow Algorithm

Wirtinger flow with random initialization is an iterative algorithm with a simple gradient descent update procedure without regularization. Specifically, the gradient step of Wirtinger flow is represented by the notion of Wirtinger derivatives [35], i.e., the derivatives of real valued functions over complex variables.

To simplify the notations, we denote $f(\boldsymbol{z}) := f(\boldsymbol{h}, \boldsymbol{x})$, where

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \cdots \\ \boldsymbol{z}_s \end{bmatrix} \in \mathbb{C}^{s(N+K)} \text{ with } \boldsymbol{z}_i = \begin{bmatrix} \boldsymbol{h}_i \\ \boldsymbol{x}_i \end{bmatrix} \in \mathbb{C}^{N+K}. \tag{11}$$

For each $i = 1, \ldots, s$, $\nabla_{\boldsymbol{h}_i} f(\boldsymbol{z})$ 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 as:

$$\nabla_{\boldsymbol{h}_i} f(\boldsymbol{z}) = \sum_{j=1}^{m} \left(\sum_{k=1}^{s} \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}_k \boldsymbol{x}_k^{\mathsf{H}} \boldsymbol{a}_{kj} - y_j \right) \boldsymbol{b}_j \boldsymbol{a}_{ij}^{\mathsf{H}} \boldsymbol{x}_i, \tag{12a}$$

$$\nabla_{\boldsymbol{x}_i} f(\boldsymbol{z}) = \sum_{j=1}^{m} \left(\sum_{k=1}^{s} \boldsymbol{h}_k^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{a}_{kj}^{\mathsf{H}} \boldsymbol{x}_k - y_j^* \right) \boldsymbol{a}_{ij} \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}_i. \tag{12b}$$

In light of the Wirtinger gradient (12), the update rule of Wirtinger flow uses a stepsize $\eta > 0$ via

$$\begin{bmatrix} \boldsymbol{h}_i^{t+1} \\ \boldsymbol{x}_i^{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}_i^{t} \\ \boldsymbol{x}_i^{t} \end{bmatrix} - \eta \begin{bmatrix} \frac{1}{\|\boldsymbol{x}_i^t\|_2^2} \nabla_{\boldsymbol{h}_i} f(\boldsymbol{z}^t) \\ \frac{1}{\|\boldsymbol{h}_i^t\|_2^2} \nabla_{\boldsymbol{x}_i} f(\boldsymbol{z}^t) \end{bmatrix}, \; i = 1, \ldots, s. \tag{13}$$

Compared with the paper [27] that solves the blind demixing problem via Wirtinger flow with spectral initialization, we solve the BlairComp via Wirtinger flow by utilizing random initialization. Random initialization is a model-agnostic and natural implementation for practitioners and works equally well as the spectral initialization strategy. Moreover, different from the sum of error, i.e., $\sum_{i=1}^{s} \|\omega_i \boldsymbol{x}_i - \bar{\boldsymbol{x}}_i\|_2 / \sum_{i=1}^{s} \|\bar{\boldsymbol{x}}_i\|_2$ considered in blind demixing, this work focuses on the relative error (8) as the performance metric. This performance metric (8) can be computed via exploring the superposition property of a wireless multiple-access channel. Since computing the relative error (8) of BlairComp does not require to transmit individual data information to the fusion center, it can address the issue of communication bandwidth limitation and support fast wireless data aggregation [16].

Before proceed to theoretical analysis, we first present an example to illustrate the practical efficiency of Wirtinger flow with random initialization for solving problem $\mathscr{P}$ (10). The ground truth values $\{\bar{\boldsymbol{h}}_i, \bar{\boldsymbol{x}}_i\}$ and initial points $\{\boldsymbol{h}_i^0, \boldsymbol{x}_i^0\}$ are randomly generated according to

$$\bar{\boldsymbol{h}}_i \sim \mathcal{N}(\boldsymbol{0}, K^{-1}\boldsymbol{I}_K), \; \bar{\boldsymbol{x}}_i \sim \mathcal{N}(\boldsymbol{0}, N^{-1}\boldsymbol{I}_N), \tag{14}$$

$$\boldsymbol{h}_i^0 \sim \mathcal{N}(\boldsymbol{0}, K^{-1}\boldsymbol{I}_K), \; \boldsymbol{x}_i^0 \sim \mathcal{N}(\boldsymbol{0}, N^{-1}\boldsymbol{I}_N), \tag{15}$$

for $i = 1, \ldots, s$. In all simulations, we set $K = N$. For each value of $K \in \{20, 80, 160, 200\}$, $s = 10$ and $m = 50\,K$, the design vectors $\boldsymbol{a}_{ij}$'s and $\boldsymbol{b}_j$'s for each $1 \le i \le s, 1 \le j \le m$, are generated according to the descriptions in Section II. With the chosen step size $\eta = 0.1$ in all settings, Fig. 1(a) shows the relative error, i.e., $\mathrm{error}(\boldsymbol{\theta}^t, \bar{\boldsymbol{\theta}})$ (8), versus the iteration count. We observe the convergence of Wirtinger flow with random initialization exhibits two stages: Stage I: within dozens of iterations, the relative error remains nearly flat, Stage II: the relative error shows exponential decay despite the different problem size.

In practical scenario, the estimation error of ambiguity alignment parameters would have influences on the relative error, i.e., $\mathrm{error}(\boldsymbol{\theta}^t, \bar{\boldsymbol{\theta}})$ (8). Hence, we illustrate the relationship between the estimation error of ambiguity alignment parameters and the relative error via the following experience. Let $K = 10$, $m = 100$, the step size be $\eta = 0.1$ and the number of users $s \in \{1, 5, 10\}$. In each iteration, for $i = 1, \ldots, i$, the estimated ambiguity alignment parameter $\hat{w}_i$ is represent by $\hat{w}_i = w_i + e_{w_i}$, where w_i is given by (9) and $e_{w_j} \sim \mathcal{N}(0, 0.5\sigma_w^{-1}) + \mathrm{i}\mathcal{N}(0, 0.5\sigma_w^{-1})$. In the experiment, the parameter σ_w varies from 1 to 10^5. Fig. 1(b) shows the relative error $\mathrm{error}(\boldsymbol{\theta}^t, \bar{\boldsymbol{\theta}})$ versus the parameter σ_w. Both the relative error and the parameter σ_w are shown in the dB scale. As we can see, the relative error scales linearly with the parameter σ_w.

We further study the relative error $\mathrm{error}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})$ in noisy scenario and explore the robustness of the Wirtinger flow with random initialization. We assume that the additive noise in (7) follows $\boldsymbol{e} = \varsigma \cdot \|\boldsymbol{y}\|_2 \cdot \frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|_2}$, where $\boldsymbol{e} \in \mathbb{C}^m$ and $\boldsymbol{\omega} \in \mathbb{C}^m$ is a standard complex Gaussian vector. Here, the constant ς equals

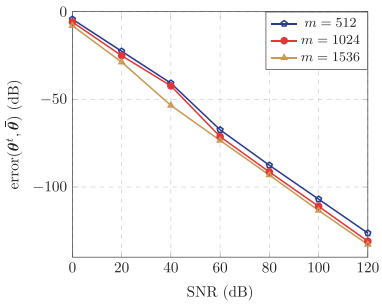


Fig. 2. Relative error $\mathrm{error}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})$ vs. SNR (dB).

the signal to noise ratio (SNR). Consider the realistic applications in wireless communication, we further explore the robustness of Wirtinger flow with random initialization in the setting of Hadamard-type encoding matrices with $s = 5, K = N = 10$ and different sample sizes $m = 512, 1024, 1536$. Here, the encoding matrix in (2) is a Hadamard-type matrix. Specifically, for $1 \le i \le s$, the Hadamard-type matrix is given by [26]

$$\boldsymbol{C}_i = \boldsymbol{F}\boldsymbol{D}_i\boldsymbol{H}, \tag{16}$$

where $\boldsymbol{F} \in \mathbb{C}^{m\times m}$ is the DFT matrix, $\boldsymbol{D}_i$'s are diagonal matrices with independent binary ± 1 entries, and $\boldsymbol{H} \in \mathbb{C}^{m\times N}$ is a fixed partial deterministic Hadamard matrix. For each setting, 100 independent trails are performed and the algorithm stops when the relative error $\mathrm{error}(\boldsymbol{\theta}^t, \bar{\boldsymbol{\theta}}) < 10^{-15}$ or the iterations $t > 500$. The relative error $\mathrm{error}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})$ in dB against the signal to noise ratio (SNR) in the settings of Hadamard-type encoding matrices is illustrated in Fig. 2. It depicts that the relative error $\mathrm{error}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})$ of WF with random initialization scales linearly with SNR.

B. Theoretical Analysis

To present the main theorem, we first introduce several fundamental definitions. Specifically, the incoherence parameter [24], 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 BlairComp): Let the incoherence parameter μ be the smallest number such that $\max_{1\le i\le s, 1\le j\le m} \frac{|\boldsymbol{b}_j^{\mathsf{H}}\bar{\boldsymbol{h}}_i|}{\|\bar{\boldsymbol{h}}_i\|_2} \le \frac{\mu}{\sqrt{m}}$.

The incoherence between $\boldsymbol{b}_j$ and $\boldsymbol{h}_i$ for $1 \le i \le s, 1 \le j \le m$ specifies the smoothness of the loss function (10). It is the smoothness along with the strong convexity of the loss function in the local region that guarantees Wirtinger flow to linearly converge to the global optimal, which plays a vital role in the theoretical analysis in Stage II. Let $\widetilde{\boldsymbol{h}}_i^t$ and $\widetilde{\boldsymbol{x}}_i^t$, respectively, denote

$$\widetilde{\boldsymbol{h}}_i^t = (\omega_i^{t*})^{-1}\boldsymbol{h}_i^t \quad \text{and} \quad \widetilde{\boldsymbol{x}}_i^t = \omega_i^t\boldsymbol{x}_i^t,\ i = 1, \ldots, s, \tag{17}$$

where ω_i^t's are alignment parameters. We further define the norm of the signal component and the perpendicular component with respect to $\boldsymbol{h}_i^t$ for $i = 1, \ldots, s$, as

$$\alpha_{\boldsymbol{h}_i^t} := \langle \bar{\boldsymbol{h}}_i, \widetilde{\boldsymbol{h}}_i^t\rangle / \|\bar{\boldsymbol{h}}_i\|_2, \tag{18}$$

$$\beta_{\boldsymbol{h}_i^t} := \left\| \widetilde{\boldsymbol{h}}_i^t - \frac{\langle \bar{\boldsymbol{h}}_i, \widetilde{\boldsymbol{h}}_i^t\rangle}{\|\bar{\boldsymbol{h}}_i\|_2^2}\bar{\boldsymbol{h}}_i \right\|_2, \tag{19}$$

respectively. Here, ω_i's are the alignment parameters. Similarly, the norms of the signal component and the perpendicular component with respect to $\boldsymbol{x}_i^t$ for $i = 1, \ldots, s$, can be represented as

$$\alpha_{\boldsymbol{x}_i^t} := \langle \bar{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_i^t\rangle / \|\bar{\boldsymbol{x}}_i\|_2, \tag{20}$$

$$\beta_{\boldsymbol{x}_i^t} := \left\| \widetilde{\boldsymbol{x}}_i^t - \frac{\langle \bar{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_i^t\rangle}{\|\bar{\boldsymbol{x}}_i\|_2^2}\bar{\boldsymbol{x}}_i \right\|_2, \tag{21}$$

respectively.

Without loss of generality, we assume $\|\bar{\boldsymbol{h}}_i\|_2 = \|\bar{\boldsymbol{x}}_i\|_2 = q_i$ $(0 < q_i \le 1)$ for $i = 1, \ldots, s$ and $\alpha_{\boldsymbol{h}_i^0}, \alpha_{\boldsymbol{x}_i^0} > 0$ for $i = 1, \ldots, s$. Define the condition number $\kappa := \frac{\max_i \|\bar{\boldsymbol{x}}_i\|_2}{\min_i \|\bar{\boldsymbol{x}}_i\|_2} \ge 1$ with $\max_i \|\bar{\boldsymbol{x}}_i\|_2 = 1$. Then the main theorem is presented in the following.

Theorem 1: Assume that the initial points obey (15) for $i = 1, \ldots, s$ 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 Wirtinger flow leads to exponentially decaying estimation error, i.e., $\mathrm{error}(\boldsymbol{\theta}^t, \bar{\boldsymbol{\theta}}) \le \gamma(1 - \frac{\eta}{16\kappa})^{t - T_\gamma}$, $t \ge T_\gamma$,
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

$$\max_{1\le i\le s} \frac{\alpha_{\boldsymbol{h}_i^t}}{\beta_{\boldsymbol{h}_i^t}} \gtrsim \frac{1}{\sqrt{K\log K}}(1 + c_3\eta)^t, \tag{22a}$$

$$\max_{1\le i\le s} \frac{\alpha_{\boldsymbol{x}_i^t}}{\beta_{\boldsymbol{x}_i^t}} \gtrsim \frac{1}{\sqrt{N\log N}}(1 + c_4\eta)^t, \tag{22b}$$

respectively, where $t = 0, 1, \ldots$ for some constants $c_3, c_4 > 0$.
3) The normalized root mean square error $\mathrm{RMSE}(\boldsymbol{x}_i^t, \bar{\boldsymbol{x}}_i) = \frac{\beta_{\boldsymbol{x}_i^t}}{\|\boldsymbol{x}_i^t\|_2}$ for $i = 1, \ldots, s$ obey

$$\mathrm{RMSE}(\boldsymbol{x}_i^t, \bar{\boldsymbol{x}}_i) \lesssim \sqrt{N\log N}(1 + c_4\eta)^{-t}, \tag{23}$$

for some constants $c_4 > 0$.

Theorem 1 provides a precisely statistical analysis on the computational efficiency of Wirtinger flow with random initialization. In summary, for the BlairComp problem, $\boldsymbol{\theta}^t$ updated by the Wirtinger flow with random initialization can linearly converge to the optimum solution, i.e., $\bar{\boldsymbol{\theta}}$. The computation performance is demonstrated in Fig. 1(a) which shows that the estimation error declines linearly after a few iterations. $\bar{\boldsymbol{\theta}}$, as long as the sample size is sufficiently large. The computation performance is demonstrated in Fig. 1 which shows that the estimation error declines linearly after a few iterations. Specifically, in Stage I, it takes $T_\gamma = \mathcal{O}(s\log(\max\{K, N\}))$ iterations for randomly initialized Wirtinger flow to reach sufficient small relative error, i.e., $\mathrm{error}(\boldsymbol{\theta}^{T_\gamma}, \bar{\boldsymbol{\theta}}) \le \gamma$ where $\gamma > 0$ is some sufficiently small constant. The short duration of Stage I is own to the exponential growth of the magnitude ratio of the signal component to the perpendicular components. 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 iteration complexity of randomly initialized WF is guaranteed to be $\mathcal{O}(s\log(\max\{K, N\}) + s\log(1/\varepsilon))$ as long as the sample size exceeds $m \gtrsim s^2\max\{K, N\}\mathrm{poly}\log(m)$. Compared with

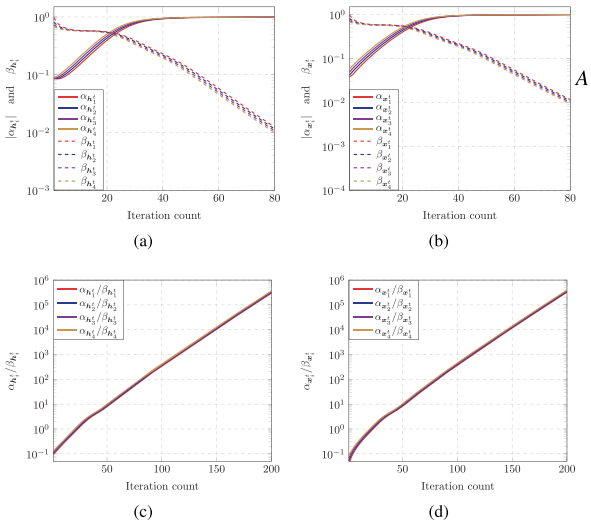


Fig. 3. Numerical example of signal versus perpendicular components.

Wirtinger flow with spectral initialization [27], Wirtinger flow with random initialization is a model-agnostic and natural for practitioners to implement. Moreover, we have demonstrated in Theorem 1 that random initialization works equally well as spectral initialization from the perspective of both computational complexity and statistical complexity.

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 $i = 1, \ldots, s$, we provide the simulation results under the setting of $K = N = 10$, $m = 50\,K$, $s = 4$ and $\eta = 0.1$ with $\|\bar{\boldsymbol{h}}_i\|_2 = \|\bar{\boldsymbol{x}}_i\|_2 = 1$ 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 $i = 1, \ldots, s$ is demonstrated in Fig. 3(a) (resp. Fig. 3(b)). Consider Fig. 1(a), Fig. 3(a) and Fig. 3(b) collectively, it shows that despite the rare decline of the estimation error, i.e., $\mathrm{error}(\boldsymbol{\theta}^t, \bar{\boldsymbol{\theta}})$, during Stage I, the size of the signal component, i.e., $\alpha_{\boldsymbol{h}_i}$ and $\alpha_{\boldsymbol{x}_i}$ for each $i = 1, \ldots, s$, 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 $i = 1, \ldots, s$ is illustrated in Fig. 3(c) (resp. Fig. 3(d)).

IV. Proof of Theorem 1

In this section, we prove the main theorem by investigating the dynamics of the iterates of Wirtinger flow 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 $\bar{\alpha}_{\boldsymbol{x}_i}$ (28a) and $\bar{\beta}_{\boldsymbol{x}_i}$ (28b), $\bar{\alpha}_{\boldsymbol{h}_i}$ (29a), $\bar{\beta}_{\boldsymbol{h}_i}$ (29b) respectively, where the sample size approaches infinity. We then develop the approximate state evolution (31), which are remarkably close to the population-level state evolution, in the finite-sample regime. See details in Section IV-A.
   - **Dynamics of approximate state evolution.** Show that there exists some $T_\gamma = \mathcal{O}(s \log(\max\{K, N\}))$ such that $\mathrm{error}(\boldsymbol{x}^{T_\gamma}, \bar{\boldsymbol{x}}) \le \gamma$, if $\alpha_{\boldsymbol{h}_i}$ (18), $\beta_{\boldsymbol{h}_i}$ (19), $\alpha_{\boldsymbol{x}_i}$ (20) and $\beta_{\boldsymbol{x}_i}$ (21) satisfy the approximate state evolution (31). 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 Lemma 1.
   - **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 (31) if the iterates $\{\boldsymbol{z}_i\}$ are independent with $\{\boldsymbol{a}_{ij}\}$. Please refer to Lemma 2. 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 IV-C. The concentrations between the original and these auxiliary sequences are then provided in Lemma 4-Lemma 9.
2) **Stage II**: **Local geometry in the region of incoherence and contraction.** We invoke the prior theory provided in [27] to show local convergence of the random initialized Wirtinger flow in Stage II.

Claims (22) and (23) are further proven in Section IV-F.

A. Dynamics of Population-Level State Evolution

In this subsection, we investigate the dynamics of population-level (where we have infinite samples) state evolution of $\alpha_{\boldsymbol{h}_i}$ (18), $\beta_{\boldsymbol{h}_i}$ (19), $\alpha_{\boldsymbol{x}_i}$ (20) and $\beta_{\boldsymbol{x}_i}$ (21). Then, we derive the approximate state evolution in the finite-sample case from the population-level state evolution. The procedure of derivation is based on the assumption that the difference between the approximate state evolution and the population-level state evolution is sufficiently small. This assumption is identified in Appendix B.

Without loss the generality, we assume that $\bar{\boldsymbol{x}}_i = q_i \boldsymbol{e}_1$ for $i = 1, \ldots, s$, where $0 < q_i \le 1, i = 1, \ldots, s$ 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 $\{\bar{\boldsymbol{h}}_i\}$ (channel vectors) cannot be transferred to a simple form, which yields more tedious analysis procedure. For simplification, for $i = 1, \ldots, s$, we denote

$$x_{i1}^t \quad \text{and} \quad \boldsymbol{x}_{i\perp}^t := [x_{ij}^t]_{2 \le j \le N} \tag{24}$$

as the first entry and the second through the N-th entries of $\boldsymbol{x}_i^t$, respectively. Based on the assumption that $\bar{\boldsymbol{x}}_i = q_i \boldsymbol{e}_1$ for $i = 1, \ldots, s$, (20) and (21) can be reformulated as

$$\alpha_{\boldsymbol{x}_i^t} := \widetilde{x}_{i1}^t \quad \text{and} \quad \beta_{\boldsymbol{x}_i^t} := \left\|\widetilde{\boldsymbol{x}}_{i\perp}^t\right\|_2. \tag{25}$$

To study the population-level state evolution, we start with considering the case where the sequences $\{\boldsymbol{z}_i^t\}$ (refer to (11)) are established via the population gradient, i.e., for $i = 1, \ldots, s$,

$$\begin{bmatrix} \boldsymbol{h}_i^{t+1} \\ \boldsymbol{x}_i^{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}_i^t \\ \boldsymbol{x}_i^t \end{bmatrix} - \eta \begin{bmatrix} \frac{1}{\|\boldsymbol{x}_i^t\|_2^2} \nabla_{\boldsymbol{h}_i} F(\boldsymbol{z}^t) \\ \frac{1}{\|\boldsymbol{h}_i^t\|_2^2} \nabla_{\boldsymbol{x}_i} F(\boldsymbol{z}^t) \end{bmatrix}, \tag{26}$$

where $\nabla_{\boldsymbol{h}_i} F(\boldsymbol{z}) := \mathbb{E}[\nabla_{\boldsymbol{h}_i} f(\boldsymbol{h}, \boldsymbol{x})] = \|\boldsymbol{x}_i\|_2^2 \boldsymbol{h}_i - (\bar{\boldsymbol{x}}_i^{\mathsf{H}} \boldsymbol{x}_i) \bar{\boldsymbol{h}}_i$, $\nabla_{\boldsymbol{x}_i} F(\boldsymbol{z}) := \mathbb{E}[\nabla_{\boldsymbol{x}_i} f(\boldsymbol{h}, \boldsymbol{x})] = \|\boldsymbol{h}_i\|_2^2 \boldsymbol{x}_i - (\bar{\boldsymbol{h}}_i^{\mathsf{H}} \boldsymbol{h}_i) \bar{\boldsymbol{x}}_i$. 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 $\boldsymbol{x}_i^t$, $i = 1, \ldots, s$ are given as

$$\widetilde{x}_{i1}^{t+1} = (1 - \eta)\, \widetilde{x}_{i1}^t + \eta \frac{q_i^2}{\|\widetilde{\boldsymbol{h}}_i^t\|_2^2} \bar{\boldsymbol{h}}_i^{\mathsf{H}} \widetilde{\boldsymbol{h}}_i^t, \tag{27a}$$

$$\widetilde{\boldsymbol{x}}_{i\perp}^{t+1} = (1-\eta)\,\widetilde{\boldsymbol{x}}_{i\perp}^{t}. \tag{27b}$$

Assuming that $\eta > 0$ is sufficiently small and $\|\bar{\boldsymbol{h}}_i\|_2 = \|\bar{\boldsymbol{x}}_i\|_2 = q_i$ $(0 < q_i \le 1)$ for $i = 1, \ldots, s$ 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 $\bar{\alpha}_{\boldsymbol{x}_i^t}$ and $\bar{\beta}_{\boldsymbol{x}_i^t}$:

$$\bar{\alpha}_{\boldsymbol{x}_i^{t+1}} = (1-\eta)\bar{\alpha}_{\boldsymbol{x}_i^t} + \eta\frac{q_i\bar{\alpha}_{\boldsymbol{h}_i^t}}{\bar{\alpha}_{\boldsymbol{h}_i^t}^2 + \beta_{\boldsymbol{h}_i^t}^2}, \tag{28a}$$

$$\bar{\beta}_{\boldsymbol{x}_i^{t+1}} = (1-\eta)\bar{\beta}_{\boldsymbol{x}_i^t}. \tag{28b}$$

The population-level state evolution for both $\bar{\alpha}_{\boldsymbol{h}_i^t}$ and $\bar{\beta}_{\boldsymbol{h}_i^t}$:

$$\bar{\alpha}_{\boldsymbol{h}_i^{t+1}} = (1-\eta)\bar{\alpha}_{\boldsymbol{h}_i^t} + \eta\frac{q_i\bar{\alpha}_{\boldsymbol{x}_i^t}}{\bar{\alpha}_{\boldsymbol{x}_i^t}^2 + \beta_{\boldsymbol{x}_i^t}^2}, \tag{29a}$$

$$\bar{\beta}_{\boldsymbol{h}_i^{t+1}} = (1-\eta)\bar{\beta}_{\boldsymbol{h}_i^t}. \tag{29b}$$

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

$$\boldsymbol{z}_i^{t+1} = \begin{bmatrix}\boldsymbol{h}_i^{t+1}\\ \boldsymbol{x}_i^{t+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{h}_i^{t} - \eta/\|\boldsymbol{x}_i^t\|_2^2\cdot\nabla_{\boldsymbol{h}_i}F(\boldsymbol{z})\\ \boldsymbol{x}_i^{t} - \eta/\|\boldsymbol{x}_i^t\|_2^2\cdot\nabla_{\boldsymbol{x}_i}F(\boldsymbol{z})\end{bmatrix} - \begin{bmatrix}\eta/\|\boldsymbol{x}_i^t\|_2^2\cdot(\nabla_{\boldsymbol{h}_i}f(\boldsymbol{z}) - \nabla_{\boldsymbol{h}_i}F(\boldsymbol{z}))\\ \eta/\|\boldsymbol{h}_i^t\|_2^2\cdot(\nabla_{\boldsymbol{x}_i}f(\boldsymbol{z}) - \nabla_{\boldsymbol{x}_i}F(\boldsymbol{z}))\end{bmatrix}. \tag{30}$$

We would derive the state evolution in the finite-sample case based on the update rule (30). The finite-sample state evolution is similar to the population-level state evolution (28) and (29) except for the perturbation terms which come from the last term in (30). Specifically, under the assumption that the last term in (30) is well-controlled, which will be justified in Appendix B, we arrive at the approximate state evolution:

$$\alpha_{\boldsymbol{h}_i^{t+1}} = \left(1-\eta+\frac{\eta q_i\psi_{\boldsymbol{h}_i^t}}{\alpha_{\boldsymbol{x}_i^t}^2+\beta_{\boldsymbol{x}_i^t}^2}\right)\alpha_{\boldsymbol{h}_i^t} + \eta(1-\rho_{\boldsymbol{h}_i^t})\frac{q_i\alpha_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{x}_i^t}^2+\beta_{\boldsymbol{x}_i^t}^2}, \tag{31a}$$

$$\beta_{\boldsymbol{h}_i^{t+1}} = \left(1-\eta+\frac{\eta q_i\varphi_{\boldsymbol{h}_i^t}}{\alpha_{\boldsymbol{x}_i^t}^2+\beta_{\boldsymbol{x}_i^t}^2}\right)\beta_{\boldsymbol{h}_i^t}, \tag{31b}$$

$$\alpha_{\boldsymbol{x}_i^{t+1}} = \left(1-\eta+\frac{\eta q_i\psi_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{h}_i^t}^2+\beta_{\boldsymbol{h}_i^t}^2}\right)\alpha_{\boldsymbol{x}_i^t} + \eta(1-\rho_{\boldsymbol{x}_i^t})\frac{q_i\alpha_{\boldsymbol{h}_i^t}}{\alpha_{\boldsymbol{h}_i^t}^2+\beta_{\boldsymbol{h}_i^t}^2}, \tag{31c}$$

$$\beta_{\boldsymbol{x}_i^{t+1}} = \left(1-\eta+\frac{\eta q_i\varphi_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{h}_i^t}^2+\beta_{\boldsymbol{h}_i^t}^2}\right)\beta_{\boldsymbol{x}_i^t}, \tag{31d}$$

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.

B. Dynamics of Approximate State Evolution

To begin with, we define the discrepancy between the estimate $\boldsymbol{z}$ and the ground truth $\bar{\boldsymbol{z}}$ as the distance function, given as

$$\mathrm{dist}(\boldsymbol{z}, \bar{\boldsymbol{z}}) = \left(\sum_{i=1}^{s}\mathrm{dist}^2(\boldsymbol{z}_i, \bar{\boldsymbol{z}}_i)\right)^{1/2}, \tag{32}$$

where $\mathrm{dist}^2(\boldsymbol{z}_i, \bar{\boldsymbol{z}}_i) = \min_{\alpha_i\in\mathbb{C}}(\|\frac{1}{\alpha_i^*}\boldsymbol{h}_i - \bar{\boldsymbol{h}}_i\|_2^2 + \|\alpha_i\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i\|_2^2)/d_i$ for $i = 1, \ldots, s$. Here, $d_i = \|\bar{\boldsymbol{h}}_i\|_2^2 + \|\bar{\boldsymbol{x}}_i\|_2^2$ and each α_i is the alignment parameter. It is easily seen that if $\alpha_{\boldsymbol{h}_i^t}$ (18), $\beta_{\boldsymbol{h}_i^t}$ (19), $\alpha_{\boldsymbol{x}_i^t}$ (20) and $\beta_{\boldsymbol{x}_i^t}$ (21) obey

$$|\alpha_{\boldsymbol{h}_i^t} - q_i| \le \frac{\gamma}{2\kappa\sqrt{s}} \quad\text{and}\quad \beta_{\boldsymbol{h}_i^t} \le \frac{\gamma}{2\kappa\sqrt{s}} \quad\text{and}\quad |\alpha_{\boldsymbol{x}_i^t} - q_i| \le \frac{\gamma}{2\kappa\sqrt{s}} \quad\text{and}\quad \beta_{\boldsymbol{x}_i^t} \le \frac{\gamma}{2\kappa\sqrt{s}}, \tag{33}$$

for $i = 1, \ldots, s$, then $\mathrm{dist}(\boldsymbol{z}, \bar{\boldsymbol{z}}) \le \gamma$. Moreover, based triangle inequality, there is $\mathrm{error}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}) \le \mathrm{dist}(\boldsymbol{z}, \bar{\boldsymbol{z}}) \le \gamma$.

In this subsection, we shall show that as long as the approximate state evolution (31) holds, there exists some constant $T_\gamma = \mathcal{O}(s\log\max\{K, N\})$ satisfying condition (33). 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

$$\alpha_{\boldsymbol{h}_i^0} \ge \frac{q_i}{K\log K} \quad\text{and}\quad \alpha_{\boldsymbol{x}_i^0} \ge \frac{q_i}{N\log N}, \tag{34a}$$

$$\sqrt{\alpha_{\boldsymbol{h}_i^0}^2 + \beta_{\boldsymbol{h}_i^0}^2} \in \left[1-\frac{1}{\log K}, 1+\frac{1}{\log K}\right]q_i, \tag{34b}$$

$$\sqrt{\alpha_{\boldsymbol{x}_i^0}^2 + \beta_{\boldsymbol{x}_i^0}^2} \in \left[1-\frac{1}{\log N}, 1+\frac{1}{\log N}\right]q_i, \tag{34c}$$

  for $i = 1, \ldots, s$.
- Define

$$T_\gamma := \min\{t : \text{satifes (33)}\}, \tag{35}$$

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

$$T_1 := \min\left\{t : \min_i\frac{\alpha_{\boldsymbol{h}_i^t}}{q_i} \ge \frac{c_7}{\log^5 m}, \min_i\frac{\alpha_{\boldsymbol{x}_i^t}}{q_i} \ge \frac{c_7'}{\log^5 m}\right\}, \tag{36}$$

$$T_2 := \min\left\{t : \min_i\frac{\alpha_{\boldsymbol{h}_i^t}}{q_i} > c_8, \min_i\frac{\alpha_{\boldsymbol{x}_i^t}}{q_i} > c_8'\right\}, \tag{37}$$

  for some small absolute positive constants $c_7, c_7', c_8, c_8' > 0$.
- For $0 \le t \le T_\gamma$, it has

$$\frac{1}{2\sqrt{K\log K}} \le \frac{\alpha_{\boldsymbol{h}_i^t}}{q_i} \le 2,\ c_5 \le \frac{\beta_{\boldsymbol{h}_i^t}}{q_i} \le 1.5 \quad\text{and}\quad \frac{\alpha_{\boldsymbol{h}_i^{t+1}}/\alpha_{\boldsymbol{h}_i^t}}{\beta_{\boldsymbol{h}_i^{t+1}}/\beta_{\boldsymbol{h}_i^t}} \ge 1+c_5\eta,\ i = 1, \ldots, s, \tag{38}$$

$$\frac{1}{2\sqrt{N\log N}} \le \frac{\alpha_{\boldsymbol{x}_i^t}}{q_i} \le 2,\ c_6 \le \frac{\beta_{\boldsymbol{x}_i^t}}{q_i} \le 1.5 \quad\text{and}\quad \frac{\alpha_{\boldsymbol{x}_i^{t+1}}/\alpha_{\boldsymbol{x}_i^t}}{\beta_{\boldsymbol{x}_i^{t+1}}/\beta_{\boldsymbol{x}_i^t}} \ge 1+c_6\eta,\ i = 1, \ldots, s, \tag{39}$$

  for some constants $c_5, c_6 > 0$.

Lemma 1: Assume that the initial points obey condition (34) and the perturbation terms in the approximate state evolution (31) obey $\max\{|\psi_{\boldsymbol{h}_i^t}|, |\psi_{\boldsymbol{x}_i^t}|, |\varphi_{\boldsymbol{h}_i^t}|, |\varphi_{\boldsymbol{x}_i^t}|, |\rho_{\boldsymbol{x}_i^t}|\} \le \frac{c}{\log m}$, for $i = 1, \ldots, s$, $t = 0, 1, \ldots$ and some sufficiently small constant $c > 0$.

1) Then for any sufficiently large K, N and the stepsize $\eta > 0$ that obeys $\eta \asymp s^{-1}$, it follows $T_\gamma \lesssim s\log(\max\{K,N\})$ and (38), (39).
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: The proof of Lemma 1 is similar to the proof of Lemma 1 in [34]. ■

Remark 1: The key point of proving this lemma is to deal with complicated approximate state evolution (31) which involves the relationship between $\alpha_{\boldsymbol{x}_i^t}$ and $\alpha_{\boldsymbol{h}_i^t}$. To address this issue, we approximate $\alpha_{\boldsymbol{x}_i^t}$ in (31a) with $\alpha_{\boldsymbol{h}_i^t}$ by multiplying proper constant and approximate $\alpha_{\boldsymbol{h}_i^t}$ in (31c) with $\alpha_{\boldsymbol{x}_i^t}$ by multiplying proper constant. The proper constants are derived by computing the relationship between $\alpha_{\boldsymbol{h}_i^t}$ and $\alpha_{\boldsymbol{h}_i^t}$ based on (31a) and (31c).

The random initialization (15) satisfies the condition (34) with probability at least $1 - \mathcal{O}(1/\sqrt{\log\min\{K,N\}})$ [34]. According to this fact, Lemma 1 ensures that under both random initialization (15) and approximate state evolution (31) 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 $\alpha_{\boldsymbol{x}_i^t}, i = 1,\ldots,s$, become sufficiently large, which is crucial to the following analysis. In particular, when establishing the approximate state evolution (31) 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 (36).

C. Leave-One-Out Approach

According to Section IV-A and Lemma 1, the unique challenge in establishing the approximate state evolution (31) 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 $i = 1,\ldots,s$. To achieve this goal, we exploit some variants of leave-one-out sequences [27], [34] 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 weight, 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* $\{\boldsymbol{z}^{t,(l)}\}_{t\ge0}$*:* For each $1 \le l \le m$, the auxiliary sequence $\{\boldsymbol{z}^{t,(l)}\}$ is established by dropping the l-th sample and runs randomly initialized Wirtinger flow with objective function

$$f^{(l)}(\boldsymbol{z}) = \sum_{j:j\ne l}\left|\sum_{i=1}^{s}\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i\boldsymbol{x}_i^{\mathsf{H}}\boldsymbol{a}_{ij} - y_j\right|^2. \tag{40}$$

  Thus, the sequences $\{\boldsymbol{z}_i^{t,(l)}\}$ (recall the definition of $\boldsymbol{z}_i$ (11)) are statistically independent of $\{\boldsymbol{a}_{il}\}$.
- *Random-sign sequences* $\{\boldsymbol{z}^{t,\mathrm{sgn}}\}_{t\ge0}$*:* Define the auxiliary design vectors $\{\boldsymbol{a}_{ij}^{\mathrm{sgn}}\}$ as

$$\boldsymbol{a}_{ij}^{\mathrm{sgn}} := \begin{bmatrix}\xi_{ij}a_{ij,1}\\ \boldsymbol{a}_{ij,\perp}\end{bmatrix}, \tag{41}$$

  where $\{\xi_{ij}\}$ is a set of standard complex uniform random variables independent of $\{\boldsymbol{a}_{ij}\}$, i.e., $\xi_{ij} \overset{\text{i.i.d.}}{=} u/|u|$, where $u \sim \mathcal{N}(0,\frac{1}{2}) + \mathrm{i}\mathcal{N}(0,\frac{1}{2})$. 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 Wirtinger flow with respect to the loss function

$$f^{\mathrm{sgn}}(\boldsymbol{z}) = \sum_{j=1}^{m}\left|\sum_{i=1}^{s}\boldsymbol{b}_j^{\mathrm{sgn}\,H}\boldsymbol{h}_i\boldsymbol{x}_i^{\mathsf{H}}\boldsymbol{a}_{ij}^{\mathrm{sgn}} - \boldsymbol{b}_j^{\mathrm{sgn}\,H}\bar{\boldsymbol{h}}_i\bar{\boldsymbol{x}}_i^{\mathsf{H}}\boldsymbol{a}_{ij}^{\mathrm{sgn}}\right|^2. \tag{42}$$

  Note that these auxiliary design vectors, i.e., $\{\boldsymbol{a}_{ij}^{\mathrm{sgn}}\}$, $\{\boldsymbol{b}_j^{\mathrm{sgn}}\}$ produce the same measurements as $\{\boldsymbol{a}_{ij}\},\{\boldsymbol{b}_j\}$: $\boldsymbol{b}_j^{\mathrm{sgn}\mathsf{H}}\bar{\boldsymbol{h}}_i\bar{\boldsymbol{x}}_i^{\mathsf{H}}\boldsymbol{a}_{ij}^{\mathrm{sgn}} = \boldsymbol{b}_j^{\mathsf{H}}\bar{\boldsymbol{h}}_i\bar{\boldsymbol{x}}_i^{\mathsf{H}}\boldsymbol{a}_{ij} = q_i a_{ij,1}\boldsymbol{b}_j^{\mathsf{H}}\bar{\boldsymbol{h}}_i$, 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 $1 \le l \le m$,

$$\{\boldsymbol{z}^0\} = \{\boldsymbol{z}^{0,(l)}\} = \{\boldsymbol{z}^{0,\mathrm{sgn}}\} = \{\boldsymbol{z}^{0,\mathrm{sgn},(l)}\}. \tag{43}$$

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

$$\begin{aligned}\omega_{i,\mathrm{mutual}}^{t,(l)} := \arg\min_{\omega\in\mathbb{C}}&\left\|\frac{1}{\omega^*}\boldsymbol{h}_i^{t,(l)} - \frac{1}{\omega_i^{t*}}\boldsymbol{h}_i^t\right\|_2^2\\ &+\left\|\omega\boldsymbol{x}_i^{t,(l)} - \omega_i^t\boldsymbol{x}_i^t\right\|_2^2,\end{aligned} \tag{44}$$

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

$$\widehat{\boldsymbol{h}}_i^{t,(l)} := \frac{1}{(\omega_{i,\mathrm{mutual}}^{t,(l)})^*}\boldsymbol{h}_i^{t,(l)} \quad\text{and}\quad \widehat{\boldsymbol{x}}_i^{t,(l)} := \omega_{i,\mathrm{mutual}}^{t,(l)}\boldsymbol{x}_i^{t,(l)}. \tag{45}$$

Define the alignment parameter between $\boldsymbol{z}_i^{t,\mathrm{sgn}} = [\boldsymbol{h}_i^{t,\mathrm{sgn}\top}\ \boldsymbol{x}_i^{t,\mathrm{sgn}\top}]^\top$ and $\boldsymbol{z}_i^t = [\boldsymbol{h}_i^{t\top}\ \boldsymbol{x}_i^{t\top}]^\top$ as

$$\begin{aligned}\omega_{i,\mathrm{sgn}}^t := \arg\min_{\omega\in\mathbb{C}}&\left\|\frac{1}{\omega^*}\boldsymbol{h}_i^{t,\mathrm{sgn}} - \frac{1}{\omega_i^{t*}}\boldsymbol{h}_i^t\right\|_2^2\\ &+\left\|\omega\boldsymbol{x}_i^{t,\mathrm{sgn}} - \omega_i^t\boldsymbol{x}_i^t\right\|_2^2,\end{aligned} \tag{46}$$

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

$$\check{\boldsymbol{h}}_i^{t,\mathrm{sgn}} := \frac{1}{(\omega_{i,\mathrm{sgn}}^t)^*}\boldsymbol{h}_i^{t,\mathrm{sgn}} \quad\text{and}\quad \check{\boldsymbol{x}}_i^{t,\mathrm{sgn}} := \omega_{i,\mathrm{sgn}}^t\boldsymbol{x}_i^{t,\mathrm{sgn}}. \tag{47}$$

D. Establishing Approximate State Evolution for Phase 1 of Stage I

In this subsection, we will justify that the approximate state evolution (31) for both the size of the signal component and the

size of the perpendicular component is satisfied during Phase I. In particular, we establish a collection of induction hypotheses which are crucial to the justification of approximate state evolution (31), and then identify these hypotheses via inductive argument.

To begin with, we list all the induction hypotheses: for $1 \le i \le s$,

$$\max_{1\le l\le m} \operatorname{dist}\left(\boldsymbol{z}_i^{t,(l)}, \widetilde{\boldsymbol{z}}_i^t\right) \le (\beta_{\boldsymbol{h}_i^t}+\beta_{\boldsymbol{x}_i^t})\left(1+\frac{1}{s\log m}\right)^t C_1\frac{s\mu^2\kappa\sqrt{\max\{K,N\}\log^8 m}}{m} \tag{48a}$$

$$\max_{1\le l\le m} \operatorname{dist}\left(\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{h}_i^{t,(l)}, \bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t\right)\cdot\|\bar{\boldsymbol{h}}_i\|_2^{-1} \le \alpha_{\boldsymbol{h}_i^t}\left(1+\frac{1}{s\log m}\right)^t C_2\frac{s\mu^2\kappa\sqrt{K\log^{13} m}}{m} \tag{48b}$$

$$\max_{1\le l\le m} \operatorname{dist}\left(x_{i1}^{t,(l)}, \widetilde{x}_{i1}^t\right) \le \alpha_{\boldsymbol{x}_i^t}\left(1+\frac{1}{s\log m}\right)^t C_2\frac{s\mu^2\kappa\sqrt{N\log^{13} m}}{m} \tag{48c}$$

$$\max_{1\le i\le s} \operatorname{dist}\left(\boldsymbol{h}_i^{t,\mathrm{sgn}}, \widetilde{\boldsymbol{h}}_i^t\right) \le \alpha_{\boldsymbol{h}_i^t}\left(1+\frac{1}{s\log m}\right)^t C_3\sqrt{\frac{s\mu^2\kappa^2 K\log^8 m}{m}} \tag{48d}$$

$$\max_{1\le i\le s} \operatorname{dist}\left(\boldsymbol{x}_i^{t,\mathrm{sgn}}, \widetilde{\boldsymbol{x}}_i^t\right) \le \alpha_{\boldsymbol{x}_i^t}\left(1+\frac{1}{s\log m}\right)^t C_3\sqrt{\frac{s\mu^2\kappa^2 N\log^8 m}{m}} \tag{48e}$$

$$\max_{1\le l\le m}\left\|\widetilde{\boldsymbol{h}}_i^t-\widehat{\boldsymbol{h}}_i^{t,(l)}-\widetilde{\boldsymbol{h}}_i^{t,\mathrm{sgn}}+\widehat{\boldsymbol{h}}_i^{t,\mathrm{sgn},(l)}\right\|_2 \le \alpha_{\boldsymbol{h}_i^t}\left(1+\frac{1}{s\log m}\right)^t C_4\frac{s\mu^2\sqrt{K\log^{16} m}}{m}, \tag{48f}$$

$$\max_{1\le l\le m}\left\|\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}-\widetilde{\boldsymbol{x}}_i^{t,\mathrm{sgn}}+\widehat{\boldsymbol{x}}_i^{t,\mathrm{sgn},(l)}\right\|_2 \le \alpha_{\boldsymbol{x}_i^t}\left(1+\frac{1}{s\log m}\right)^t C_4\frac{s\mu^2\sqrt{N\log^{16} m}}{m}, \tag{48g}$$

$$c_5 \le \left\|\boldsymbol{h}_i^t\right\|_2, \left\|\boldsymbol{x}_i^t\right\|_2 \le C_5, \tag{48h}$$

$$\left\|\boldsymbol{h}_i^t\right\|_2 \le 5\alpha_{\boldsymbol{h}_i^t}\sqrt{\log^5 m}, \tag{48i}$$

$$\left\|\boldsymbol{x}_i^t\right\|_2 \le 5\alpha_{\boldsymbol{x}_i^t}\sqrt{\log^5 m}, \tag{48j}$$

where $C_1,\ldots,C_5$ 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 IV-C.

Specifically, (48a), (48c), (48d) and (48e) 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 (48f) and (48g), $\widetilde{\boldsymbol{h}}_i^t-\widetilde{\boldsymbol{h}}_i^{t,\mathrm{sgn}}$ (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 (48h) illustrates that the norm of the iterates $\{\boldsymbol{h}_i^t\}$ (resp. $\{\boldsymbol{x}_i^t\}$) is well-controlled in Phase 1. Moreover, (48i) (resp. (48j)) indicates that $\alpha_{\boldsymbol{h}_i^t}$ (resp. $\alpha_{\boldsymbol{x}_i^t}$) is comparable to $\|\boldsymbol{h}_i^t\|_2$ (resp. $\|\boldsymbol{x}_i^t\|_2$).

We are moving to prove that if the induction hypotheses (48) hold for the t-th iteration, then $\alpha_{\boldsymbol{h}_i}$, $\beta_{\boldsymbol{h}_i}$, $\alpha_{\boldsymbol{x}_i}$ and $\beta_{\boldsymbol{x}_i}$ obey the approximate state evolution (31). This is demonstrated in Lemma 2.

Lemma 2: Suppose $m \ge Cs^2\mu^2\max\{K,N\}\log^{10} m$ for some sufficiently large constant $C>0$. For any $0\le t\le T_1$ (36), if the t-th iterate satisfies the induction hypotheses (48), then for $i=1,\ldots,s$, with probability at least $1-c_1 m^{-\nu}-c_1 m e^{-c_2 N}$ for some constants $\nu, c_1, c_2>0$, the approximate evolution state (31) holds for some $|\psi_{\boldsymbol{h}_i^t}|, |\psi_{\boldsymbol{x}_i^t}|, |\varphi_{\boldsymbol{h}_i^t}|, |\varphi_{\boldsymbol{x}_i^t}|, |\rho_{\boldsymbol{h}_i^t}|, |\rho_{\boldsymbol{x}_i^t}| \ll 1/\log m$, $i=1,\ldots,s$.

Proof: Please refer to Appendix B for details. ■

Remark 2: Due to the "incoherence" between multiple signals, extra technical arguments are required to be developed. Take the term J_{i1} in (68) as an example, we present $J_{i1}=\bar{\boldsymbol{h}}_k^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t a_{ij,1}{}^* q_k a_{ij,1}+\sum_{j\ne i}^m\sum_{k=1}^s \bar{\boldsymbol{h}}_k^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t a_{kj,1}{}^* q_k a_{ij,1}$. Therein, the i.i.d. random variable a_{ij} ensures the statistical property $\mathbb{E}(\boldsymbol{a}_{ij}\boldsymbol{a}_{kj}^*)=0$ for $k\ne i$ that facilitates the proof. This technique is also exploited in the following lemma, which is the cornerstone of the theoretical analysis in the BlairComp problem.

In the sequel, we will prove the hypotheses (48) hold for Phase 1 of Stage I via inductive arguments. Before moving forward, we first investigate the incoherence between $\{\boldsymbol{x}_i^t\}$, $\{\boldsymbol{x}_i^{t,\mathrm{sgn}}\}$ (resp. $\{\boldsymbol{h}_i^t\}$, $\{\boldsymbol{h}_i^{t,\mathrm{sgn}}\}$) and $\{\boldsymbol{a}_{ij}\}$, $\{\boldsymbol{a}_{ij}^{\mathrm{sgn}}\}$ (resp. $\{\boldsymbol{b}_j\}$, $\{\boldsymbol{b}_j^{\mathrm{sgn}}\}$).

Lemma 3: Suppose that $m \ge Cs^2\mu^2\max\{K,N\}\log^8 m$ for some sufficiently large constant $C>0$ and the t-th iterate satisfies the induction hypotheses (48) for $t\le T_0$ (36), then with probability at least $1-c_1 m^{-\nu}-c_1 m e^{-c_2 N}$ for some constants $\nu, c_1, c_2>0$,

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{a}_{il}^{\mathsf{H}}\widetilde{\boldsymbol{x}}_i^t\right|\cdot\|\widetilde{\boldsymbol{x}}_i^t\|_2^{-1}\lesssim\sqrt{\log m},| \tag{49a}$$

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{a}_{il,\perp}^{\mathsf{H}}\widetilde{\boldsymbol{x}}_{i\perp}^t\right|\cdot\|\widetilde{\boldsymbol{x}}_{i\perp}^t\|_2^{-1}\lesssim\sqrt{\log m}, \tag{49b}$$

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{a}_{il}^{\mathsf{H}}\check{\boldsymbol{x}}_i^{t,\mathrm{sgn}}\right|\cdot\|\check{\boldsymbol{x}}_i^{t,\mathrm{sgn}}\|_2^{-1}\lesssim\sqrt{\log m}, \tag{49c}$$

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{a}_{il,\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right|\cdot\|\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\|_2^{-1}\lesssim\sqrt{\log m}, \tag{49d}$$

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{a}_{il}^{\mathrm{sgn}\,\mathsf{H}}\check{\boldsymbol{x}}_i^{t,\mathrm{sgn}}\right|\cdot\|\check{\boldsymbol{x}}_i^{t,\mathrm{sgn}}\|_2^{-1}\lesssim\sqrt{\log m}, \tag{49e}$$

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t\right|\cdot\|\widetilde{\boldsymbol{h}}_i^t\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m, \tag{50a}$$

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}_i^{t,\mathrm{sgn}}\right|\cdot\|\check{\boldsymbol{h}}_i^{t,\mathrm{sgn}}\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m, \tag{50b}$$

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{b}_l^{\mathrm{sgn}\,\mathsf{H}}\check{\boldsymbol{h}}_i^{t,\mathrm{sgn}}\right|\cdot\|\check{\boldsymbol{h}}_i^{t,\mathrm{sgn}}\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m. \tag{50c}$$

Proof: Based on the induction hypotheses (48), we can prove the claim (49) in Lemma 3 by invoking the triangle inequality, Cauchy-Schwarz inequality and standard

Gaussian concentration. Furthermore, based on the induction hypotheses (48), the claim (50) can be identified according to the definition of the incoherence parameter in Definition 1 and the fact $\|\boldsymbol{b}_j\|_2 = \sqrt{K/m}$. Moreover, to address the deterministic property of $\boldsymbol{b}_\ell$ and facilitate the proof of (50), we divide $\{\boldsymbol{b}_j\}_{1\le j\le m}$ into consecutive bins in order to exploit the random property of $\boldsymbol{a}_{i\ell}$ within each bin. Here, we assume each bin contains $\Delta \asymp \mathrm{poly}\log m$ contiguous vectors. For instance, for $l = 1,\ldots,m, i = 1,\ldots,s$, to bound the term $|\boldsymbol{b}_\ell \sum_{j=1}^{m}\sum_{k=1}^{s} \boldsymbol{b}_j\boldsymbol{b}_j^*\widetilde{\boldsymbol{h}}_k^t(\boldsymbol{x}_k^\natural{}^*\boldsymbol{a}_{kj}\boldsymbol{a}_{ij}^*\boldsymbol{x}_i^\natural - \|\boldsymbol{x}_k^\natural\|_2^2)|$, we consider the consecutive vector in individual bins, given by $|\boldsymbol{b}_l \sum_{j=1}^{\Delta}\sum_{k=1}^{s} \boldsymbol{b}_{\iota+j}\boldsymbol{b}_{\iota+j}^*\widetilde{\boldsymbol{h}}_k^t(\boldsymbol{x}_k^\natural{}^*\boldsymbol{a}_{k,\iota+j}\boldsymbol{a}_{i,\iota+j}^*\boldsymbol{x}_i^\natural - \|\boldsymbol{x}_k^\natural\|_2^2)|$, for each $0\le\iota\le m-\Delta$, which enables to exploit the randomness within each bin. ■

Now we are ready to specify that the hypotheses (48) hold for $0\le t\le T_1$ (36). We aim to demonstrate that if the hypotheses (48) hold up to the t-th iteration for some $0\le t\le T_1$, then they hold for the $(t+1)$-th iteration. Since the case for $t=0$ can be easily justified due to the equivalent initial points (43), we mainly focus the inductive step.

Lemma 4: Suppose the induction hypotheses (48) hold true up to the t-th iteration for some $t\le T_1$ (36), then for $i=1,\ldots,s$, with probability at least $1-c_1\,m^{-\nu}-c_1me^{-c_2\,N}$ for some constants $\nu, c_1, c_2>0$, there is $\max_{1\le l\le m}\mathrm{dist}(\boldsymbol{z}_i^{t+1,(l)},\widetilde{\boldsymbol{z}}_i^{t+1})\le(\beta_{\boldsymbol{h}_i^{t+1}}+\beta_{\boldsymbol{x}_i^{t+1}})\cdot(1+\frac{1}{s\log m})^{t+1}C_1\cdot\frac{s\mu^2\kappa\sqrt{\max\{K,N\}\log^8 m}}{m}$ holds $m\ge Cs\mu^2\kappa\sqrt{\max\{K,N\}\log^8 m}$ with some sufficiently large constant $C>0$ as long as the stepsize $\eta>0$ obeys $\eta\asymp s^{-1}$ and $C_1>0$ is sufficiently large.

In terms of the difference between $\boldsymbol{x}^t$ and $\boldsymbol{x}_i^{t,(l)}$ (resp. $\boldsymbol{h}_i^t$ and $\boldsymbol{h}_i^{t,(l)}$) along with the signal direction, i.e., (48b) and (48c), we reach the following lemma.

Lemma 5: Suppose the induction hypotheses (48) hold true up to the t-th iteration for some $t\le T_1$ (36), then with probability at least $1-c_1\,m^{-\nu}-c_1me^{-c_2\,N}$ for some constants $\nu, c_1, c_2>0$,

$$\max_{1\le l\le m}\mathrm{dist}\left(\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{h}_i^{t+1,(l)},\bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^{t+1}\right)\cdot\|\bar{\boldsymbol{h}}_i\|_2^{-1} \le\alpha_{\boldsymbol{h}_i^{t+1}}\left(1+\frac{1}{s\log m}\right)^{t+1}C_2\frac{s\mu^2\kappa\sqrt{K\log^{13}m}}{m} \tag{51}$$

$$\max_{1\le l\le m}\mathrm{dist}\left(x_{i1}^{t+1,(l)},\widetilde{x}_{i1}^{t+1}\right) \le\alpha_{\boldsymbol{x}_i^{t+1}}\left(1+\frac{1}{s\log m}\right)^{t+1}C_2\frac{s\mu^2\kappa\sqrt{N\log^{13}m}}{m} \tag{52}$$

holds for some sufficiently large $C_2>0$ with $C_2\gg C_4$, provided that $m\ge Cs\mu^2\kappa\max\{K,N\}\log^{12}m$ for some sufficiently large constant $C>0$ and the stepsize $\eta>0$ obeys $\eta\asymp s^{-1}$.

Proof: Please refer to Appendix C for details. ■

The next lemma concerns the relation between $\boldsymbol{h}_i^t$ and $\boldsymbol{h}_i^{t,\mathrm{sgn}}$, i.e., (48d), and the relation between $\boldsymbol{x}_i^t$ and $\boldsymbol{x}_i^{t,\mathrm{sgn}}$, i.e., (48e).

Lemma 6: Suppose the induction hypotheses (48) hold true up to the t-th iteration for some $t\le T_1$ (36), then with probability at least $1-c_1\,m^{-\nu}-c_1me^{-c_2\,N}$ for some constants $\nu, c_1, c_2>0$,

$$\max_{1\le i\le s}\mathrm{dist}\left(\boldsymbol{h}_i^{t+1,\mathrm{sgn}},\widetilde{\boldsymbol{h}}_i^{t+1}\right) \le\alpha_{\boldsymbol{h}_i^{t+1}}\left(1+\frac{1}{s\log m}\right)^{t+1}C_3\sqrt{\frac{s\mu^2\kappa^2K\log^8 m}{m}} \tag{53a}$$

$$\max_{1\le i\le s}\mathrm{dist}\left(\boldsymbol{x}_i^{t+1,\mathrm{sgn}},\widetilde{\boldsymbol{x}}_i^{t+1}\right) \le\alpha_{\boldsymbol{x}_i^{t+1}}\left(1+\frac{1}{s\log m}\right)^{t+1}C_3\sqrt{\frac{s\mu^2\kappa^2N\log^8 m}{m}} \tag{53b}$$

holds for some sufficiently large $C_3>0$, provided that $m\ge Cs\mu^2\kappa^2\max\{K,N\}\log^8 m$ for some sufficiently large constant $C>0$ and the stepsize $\eta>0$ obeys $\eta\asymp s^{-1}$.

We still need to characterize the difference $\widetilde{\boldsymbol{h}}_i^t-\widehat{\boldsymbol{h}}_i^{t,(l)}-\widetilde{\boldsymbol{h}}_i^{t,\mathrm{sgn}}+\widehat{\boldsymbol{h}}_i^{t,,\mathrm{sgn},(l)}$, i.e., (48f), and the difference $\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}-\widetilde{\boldsymbol{x}}_i^{t,\mathrm{sgn}}+\widehat{\boldsymbol{x}}_i^{t,\mathrm{sgn},(l)}$, i.e., (48g), in the following lemma.

Lemma 7: Suppose the induction hypotheses (48) hold true up to the t-th iteration for some $t\le T_1$ (36), then with probability at least $1-c_1\,m^{-\nu}-c_1me^{-c_2\,N}$ for some constants $\nu, c_1, c_2>0$,

$$\max_{1\le l\le m}\left\|\widetilde{\boldsymbol{h}}_i^{t+1}-\widehat{\boldsymbol{h}}_i^{t+1,(l)}-\widetilde{\boldsymbol{h}}_i^{t+1,\mathrm{sgn}}+\widehat{\boldsymbol{h}}_i^{t+1,\mathrm{sgn},(l)}\right\|_2 \le\alpha_{\boldsymbol{h}_i^{t+1}}\left(1+\frac{1}{s\log m}\right)^{t+1}C_4\frac{s\mu^2\sqrt{K\log^{16}m}}{m} \tag{54a}$$

$$\max_{1\le l\le m}\left\|\widetilde{\boldsymbol{x}}_i^{t+1,}-\widehat{\boldsymbol{x}}_i^{t+1,,(l)}-\widetilde{\boldsymbol{x}}_i^{t+1,\mathrm{sgn}}+\widehat{\boldsymbol{x}}_i^{t+1,\mathrm{sgn},(l)}\right\|_2 \le\alpha_{\boldsymbol{x}_i^{t+1}}\left(1+\frac{1}{s\log m}\right)^{t+1}C_4\frac{s\mu^2\sqrt{N\log^{16}m}}{m} \tag{54b}$$

holds for some sufficiently large $C_4>0$, provided that $m\ge Cs\mu^2\max\{K,N\}\log^8 m$ for some sufficiently large constant $C>0$ and the stepsize $\eta>0$ obeys $\eta\asymp s^{-1}$.

Remark 3: The arguments applied to prove Lemma 4-Lemma 7 are similar to each other. We thus mainly focus on the proof of (52) in Lemma 5 in Appendix C.

E. Establishing Approximate State Evolution for Phase 2 of Stage I

In this subsection, we move to prove that the approximate state evolution (31) holds for $T_1<t\le T_\gamma$ (T_γ and T_1 are defined in (35) and (36) respectively) via inductive argument. Different from the analysis in Phase 1, only $\{\boldsymbol{z}^{t,(l)}\}$ is sufficient to establish the "near-independence" between iterates and design vectors when the sizes of the signal component follow $\alpha_{\boldsymbol{h}_i^t}$, $\alpha_{\boldsymbol{x}_i}\gtrsim 1/\log m$ in Phase 2 (according to the definition of T_1). As in Phase 1, we begin with specifying the induction hypotheses: for $1\le i\le s$,

$$\max_{1\le l\le m}\mathrm{dist}\left(\boldsymbol{z}_i^{t,(l)},\widetilde{\boldsymbol{z}}_i^t\right)\le(\beta_{\boldsymbol{h}_i^t}+\beta_{\boldsymbol{x}_i^t})\left(1+\frac{1}{s\log m}\right)^t \times C_6\frac{s\mu^2\kappa\sqrt{\max\{K,N\}\log^{18}m}}{m} \tag{55a}$$

$$c_5\le\left\|\boldsymbol{h}_i^t\right\|_2,\left\|\boldsymbol{x}_i^t\right\|_2\le C_5, \tag{55b}$$

From (55), we can conclude that one has

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{a}_{il}^{\mathsf{H}}\widetilde{\boldsymbol{x}}_i^t\right|\cdot\|\widetilde{\boldsymbol{x}}_i^t\|_2^{-1}\lesssim\sqrt{\log m}, \tag{56}$$

$$\max_{1\le i\le s,1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t\right|\cdot\|\widetilde{\boldsymbol{h}}_i^t\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m, \tag{57}$$

with probability at least $1-c_1 m^{-\nu}-c_1 m e^{-c_2 N}$ for some constants $\nu, c_1, c_2>0$ during $T_1<t\le T_\gamma$ as long as $m\gg Cs\mu^2\kappa K\log^8 m$.

We then move to prove that if the induction hypotheses (48) hold for the t-th iteration, then $\alpha_{\boldsymbol{h}_i}$, $\beta_{\boldsymbol{h}_i}$, $\alpha_{\boldsymbol{x}_i}$ and $\beta_{\boldsymbol{x}_i}$ obey the approximate state evolution (31). This is demonstrated in Lemma 8.

Lemma 8: Suppose $m\ge Cs^2\mu^2\kappa^4\max\{K,N\}\log^{12} m$ for some sufficiently large constant $C>0$. For any $T_1\le t\le T_\gamma$ (T_1 and T_γ are defined in (35) and (36) respectively), if the t-th iterate satisfies the induction hypotheses (48), then for $i=1,\ldots,s$, with probability at least $1-c_1 m^{-\nu}-c_1 m e^{-c_2 N}$ for some constants $\nu, c_1, c_2>0$, the approximate evolution state (31) hold for some $|\psi_{\boldsymbol{h}_i^t}|, |\psi_{\boldsymbol{x}_i^t}|, |\varphi_{\boldsymbol{h}_i^t}|, |\varphi_{\boldsymbol{x}_i^t}|, |\rho_{\boldsymbol{h}_i^t}|, |\rho_{\boldsymbol{x}_i^t}|\ll 1/\log m, i=1,\ldots,s$.

It remains to proof the induction step on the difference between leave-one-out sequences $\{\boldsymbol{z}^{t,(l)}\}$ and the original sequences $\{\boldsymbol{z}^t\}$, which is demonstrated in the following lemma.

Lemma 9: Suppose the induction hypotheses (48) are valid during Phase 1 and the induction hypotheses (55) hold true from T_1-th to the t-th for some $t\le T_\gamma$ (35), then for $i=1,\ldots,s$, with probability at least $1-c_1 m^{-\nu}-c_1 m e^{-c_2 N}$ for some constants $\nu, c_1, c_2>0$,

$$\max_{1\le l\le m}\operatorname{dist}\left(\boldsymbol{z}_i^{t,(l)},\widetilde{\boldsymbol{z}}_i^t\right)\le\left(\beta_{\boldsymbol{h}_i^{t+1}}+\beta_{\boldsymbol{x}_i^{t+1}}\right)\left(1+\frac{1}{s\log m}\right)^{t+1}\times C_6\frac{s\mu^2\kappa\sqrt{K\log^{18} m}}{m} \tag{58}$$

holds $m\ge Cs\mu^2\kappa K\log^8 m$ with some sufficiently large constant $C>0$ as long as the stepsize $\eta>0$ obeys $\eta\asymp s^{-1}$ and $C_6>0$ is sufficiently large.

Remark 4: The proof of Lemma 8 and Lemma 9 is inspired by the arguments used in Section H and Section I in [34].

F. Proof for Claims (22) and (23)

Combining the analyses in Phase 1 and Phase 2, we complete the proof for claims (22) with $0\le t\le T_\gamma$ (35). Consider the definition of T_γ (35) and the incoherence between iterates and design vectors given in (56) and (57), we arrive at

$$\left\|\widetilde{\boldsymbol{x}}_i^{T_\gamma}-\bar{\boldsymbol{x}}_i\right\|_2\le\frac{\gamma}{\sqrt{2}\,s} \tag{59}$$

$$\operatorname{dist}(\boldsymbol{z}^{T_\gamma},\bar{\boldsymbol{z}})\le\gamma \tag{60}$$

$$\operatorname{error}(\boldsymbol{\theta}^{T_\gamma},\bar{\boldsymbol{\theta}})\le\gamma \tag{61}$$

$$\max_{1\le i\le s,1\le j\le m}\left|\boldsymbol{a}_{ij}^{\mathsf{H}}\widetilde{\boldsymbol{x}}_i^{T_\gamma}\right|\cdot\|\widetilde{\boldsymbol{x}}_i^{T_\gamma}\|_2^{-1}\lesssim\sqrt{\log m}, \tag{62}$$

$$\max_{1\le i\le s,1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^{T_\gamma}\right|\cdot\|\widetilde{\boldsymbol{h}}_i^{T_\gamma}\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m, \tag{63}$$

which implies that $\max_{1\le i\le s,1\le j\le m}|\boldsymbol{a}_{ij}^{\mathsf{H}}(\widetilde{\boldsymbol{x}}_i^{T_\gamma}-\bar{\boldsymbol{x}}_i)|\lesssim\frac{\gamma\sqrt{\log m}}{\sqrt{2}\,s}$, based on the inductive hypothesis (55a). Based on these properties, we can exploit the techniques applied in [36, Section IV] and the triangle inequality to prove that for $t\ge T_\gamma+1$,

$$\operatorname{error}(\boldsymbol{\theta}^t,\bar{\boldsymbol{\theta}})\le\operatorname{dist}\left(\boldsymbol{x}^t,\bar{\boldsymbol{x}}\right)\le\gamma\left(1-\frac{\eta}{16\kappa}\right)^{t-T_\gamma}, \tag{64}$$

where the stepsize $\eta>0$ obeys $\eta\asymp s^{-1}$ as long as $m\gg s^2\mu^2\kappa^4\max\{K,N\}\log^8 m$. It remains to prove the claim (22) for Stage II. Since we have already demonstrate that the ratio $\alpha_{\boldsymbol{h}_i^t}/\beta_{\boldsymbol{h}_i^t}$ increases exponentially fast in Stage I, there is $\frac{\alpha_{\boldsymbol{h}_i^{T_1}}}{\beta_{\boldsymbol{h}_i^{T_1}}}\ge\frac{1}{\sqrt{2K\log K}}(1+c_3\eta)^{T_1}$. By the definition of T_1 (see (36)) and Lemma 1, one has $\alpha_{\boldsymbol{h}_i^{T_1}}\asymp\beta_{\boldsymbol{h}_i^{T_1}}\asymp 1$ and thus

$$\frac{\alpha_{\boldsymbol{h}_i^{T_1}}}{\beta_{\boldsymbol{h}_i^{T_1}}}\asymp 1. \tag{65}$$

When it comes to $t>T_\gamma$, based on (64), we have

$$\frac{\alpha_{\boldsymbol{h}_i^t}}{\beta_{\boldsymbol{h}_i^t}}\ge\frac{1-\operatorname{dist}(\boldsymbol{z}^t,\bar{\boldsymbol{z}})}{\operatorname{dist}(\boldsymbol{z}^t,\bar{\boldsymbol{z}})}\gtrsim\frac{1}{\sqrt{K\log K}}(1+c_3\eta)^t,$$

where (i) is derived from (65) and the fact that γ is a constant, (ii) arises from $T_\gamma-T_1\asymp s^{-1}$ based on Lemma 1, and the last inequality is satisfied as long as $c_3>0$ and $\eta\asymp s^{-1}$. Likewise, we can apply the same arguments to the ratio $\alpha_{\boldsymbol{x}_i^t}/\beta_{\boldsymbol{x}_i^t}$, thereby concluding that $\frac{\alpha_{\boldsymbol{x}_i^t}}{\beta_{\boldsymbol{x}_i^t}}\gtrsim\frac{1}{\sqrt{N\log N}}(1+c_4\eta)^t$. Claim (23) can be further derived via the equation $\mathrm{RMSE}(\boldsymbol{x}_i^t,\bar{\boldsymbol{x}}_i)=\frac{\beta_{\boldsymbol{x}_i^t}}{\sqrt{\alpha_{\boldsymbol{x}_i^t}^2+\beta_{\boldsymbol{x}_i^t}^2}}<\frac{\beta_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{x}_i^t}}$.

V. Conclusion

In this paper, we proposed a blind over-the-air computation scheme to compute the desired function of distributed sensing data without the prior knowledge of the channel information, thereby providing low-overhead data aggregation in IoT networks. To harness the benefits of computational efficiency, fast convergence guarantee, regularization-free and careful *initialization-free*, the BlairComp problem was solved by randomly initialized Wirtinger flow with provable guarantees. Specifically, the statistical guarantee and fast global convergence guarantee concerning randomly initialized Wirtinger flow for solving the BlairComp problem were provided. It demonstrated that with sufficient samples, in the first tens iterations, the randomly initialized Wirtinger flow enables the iterates to enter a local region that enjoys strong convexity and strong smoothness, where the estimation error is sufficiently small. At the second stage of this algorithm, the estimated error experiences exponential decay.

Appendix A
Preliminaries

For $\boldsymbol{a}_{ij}\in\mathbb{C}^N$, the standard concentration inequality gives that, for $i=1,\ldots,s$,

$$\max_{1\le j\le m}|a_{ij,1}|=\max_{1\le j\le m}\left|\boldsymbol{a}_{ij}^{\mathsf{H}}\bar{\boldsymbol{x}}\right|\le 5\sqrt{\log m} \tag{66}$$

with probability $1-\mathcal{O}(m^{-10})$ [36]. In addition, by applying the standard concentration inequality, we arrive at, for $i=1,\ldots,s$,

$$\max_{1\le j\le m}\|\boldsymbol{a}_{ij}\|_2\le 3\sqrt{N} \tag{67}$$

with probability $1-C'\exp(me^{-cK})$ for some constants, $c,C'>0$ [36].

Lemma 10: Fix any constant $c_0>1$. Define the population matrix $\nabla^2_{\boldsymbol{z}_i}F(\boldsymbol{z})$ as

$$\begin{bmatrix} \|\boldsymbol{x}_i\|_2^2\boldsymbol{I}_K & \boldsymbol{h}_i\boldsymbol{x}_i^{\mathsf{H}}-\bar{\boldsymbol{h}}_i\bar{\boldsymbol{x}}_i^{\mathsf{H}} & \boldsymbol{0} & \bar{\boldsymbol{h}}_i\bar{\boldsymbol{x}}_i^{\top} \\ \boldsymbol{x}_i\boldsymbol{h}_i^{\mathsf{H}}-\bar{\boldsymbol{x}}_i\bar{\boldsymbol{h}}_i^{\mathsf{H}} & \|\boldsymbol{h}_i\|_2^2\boldsymbol{I}_K & \bar{\boldsymbol{x}}_i\bar{\boldsymbol{h}}_i^{\top} & \boldsymbol{0} \\ \boldsymbol{0} & (\bar{\boldsymbol{x}}_i\bar{\boldsymbol{h}}_i^{\top})^{\mathsf{H}} & \|\boldsymbol{x}_i\|_2^2\boldsymbol{I}_K & (\boldsymbol{h}_i\boldsymbol{x}_i^{\mathsf{H}}-\bar{\boldsymbol{h}}_i\bar{\boldsymbol{x}}_i^{\mathsf{H}})\mathsf{H} \\ (\bar{\boldsymbol{h}}_i\bar{\boldsymbol{x}}_i^{\top})^{\mathsf{H}} & \boldsymbol{0} & (\boldsymbol{x}_i\boldsymbol{h}_i^{\mathsf{H}}-\bar{\boldsymbol{x}}_i\bar{\boldsymbol{h}}_i^{\mathsf{H}})\mathsf{H} & \|\boldsymbol{h}_i\|_2^2\boldsymbol{I}_K \end{bmatrix}$$

Suppose that $m>c_1s^2\mu^2K\log^3 m$ for some sufficiently large constant $c_1>0$. Then with probability exceeding $1-\mathcal{O}(m^{-10})$,

$$\begin{aligned}&\left\|\left(\boldsymbol{I}_{4K}-\eta\nabla^2 f(\boldsymbol{z})\right)-\left(\boldsymbol{I}_{4K}-\eta\nabla^2 F(\boldsymbol{z})\right)\right\| \\ &\lesssim\sqrt{\frac{s^2\mu^2K\log m}{m}}\max\left\{\|\boldsymbol{z}\|_2^2,1\right\} \\ \text{and}\quad&\left\|\nabla^2 f(\boldsymbol{z})\right\|\le 5\|\boldsymbol{z}\|_2^2+2\end{aligned}$$

hold simultaneously for all $\boldsymbol{z}$ obeying $\max_{1\le i\le s,1\le l\le m}|\boldsymbol{a}_{il}^{\mathsf{H}}\boldsymbol{x}_i|\cdot\left\|\boldsymbol{x}_i\right\|_2^{-1}\lesssim\sqrt{\log m}$ and $\max_{1\le i\le s,1\le l\le m}|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}_i|\cdot\left\|\boldsymbol{h}_i\right\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m$, provided that $0<\eta<\frac{c_2}{\max\{\|\boldsymbol{z}\|_2^2,1\}}$ for some sufficiently small constant $c_2>0$.

Appendix B
Proof of Lemma 2

According to the Wirtinger flow gradient update rule (12b), and the expression $\boldsymbol{a}_{kj}^{\mathsf{H}}\boldsymbol{x}_k^t=x_{k\|}^t a_{kj,1}^*+\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\boldsymbol{x}_{k\perp}^t$ and reformulate terms, we arrive at

$$\widetilde{x}_{i1}^{t+1}=\widetilde{x}_{i1}^t+\eta' J_{i1}-\eta' J_{i2}-\eta' J_{i3}, \tag{68}$$

where

$$\begin{aligned}J_{i1}&=\sum_{j=1}^m\sum_{k=1}^s\bar{\boldsymbol{h}}_k^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t a_{kj,1}{}^* q_k a_{ij,1},\\ J_{i2}&=\sum_{j=1}^m\sum_{k=1}^s\widetilde{\boldsymbol{h}}_k^{t\,H}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t a_{kj,1}{}^*\widetilde{x}_{k\|}^t a_{ij,1},\\ J_{i3}&=\sum_{j=1}^m\sum_{k=1}^s\widetilde{\boldsymbol{h}}_k^{t\,H}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\boldsymbol{x}_{i\perp}^t a_{ij,1},\\ \eta'&=\eta/\|\widetilde{\boldsymbol{h}}_i^t\|_2^2.\end{aligned}$$

We will control the above three terms J_{i1}, J_{i2} and J_{i3} separately in the following.

- With regard to the first term J_{i1}, it has $\sum_{j=1}^m\sum_{k=1}^s q_k\bar{\boldsymbol{h}}_k^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t a_{kj,1}{}^* a_{ij,1}=\sum_{k=1}^s q_k\bar{\boldsymbol{h}}_k^{\mathsf{H}}\cdot\left(\sum_{j=1}^m a_{kj,1}{}^* a_{ij,1}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right)\widetilde{\boldsymbol{h}}_i^t$. According to Lemma 11 and Lemma 12, there is $J_{i1}=q_i\bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t+r_1$, where the size of the remaining term r_1 satisfies $|r_1|\lesssim\sum_{k=1}^s q_k\bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t\sqrt{\frac{K}{m}\log m}\lesssim\sqrt{\frac{s^2K}{m}\log m}\cdot\bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t$, based on the fact that $\|\bar{\boldsymbol{h}}_k\|^2\lesssim 1$ and $\|\widetilde{\boldsymbol{h}}_k^t\|^2\lesssim 1$ for $k=1,\ldots,s$.
- Similar to the first term, the term J_{i2} can be represented as $J_{i2}=\|\widetilde{\boldsymbol{h}}_i^t\|_2^2\widetilde{x}_{i1}^t+r_2$, where the term r_{i2} obeys

$$|r_2|\lesssim|\widetilde{x}_{i1}^t|\sum_{k=1}^s\widetilde{\boldsymbol{h}}_k^{t\,H}\widetilde{\boldsymbol{h}}_i^t\sqrt{\frac{K}{m}\log m}\lesssim\sqrt{\frac{s^2K}{m}\log m}|\widetilde{x}_{i1}^t|. \tag{69}$$

- For the last term J_{i3}, it follows that

$$\begin{aligned}&\sum_{j=1}^m\sum_{k=1}^s\widetilde{\boldsymbol{h}}_k^{t\,H}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^t\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\widetilde{\boldsymbol{x}}_{i\perp}^t a_{ij,1}\\ &=\sum_{k=1}^s\widetilde{\boldsymbol{h}}_k^{t\,H}\left(\sum_{j=1}^m a_{ij,1}\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\boldsymbol{x}_{i\perp}^t\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right)\widetilde{\boldsymbol{h}}_i^t.\end{aligned} \tag{70}$$

By exploiting the random-sign sequence $\{\boldsymbol{x}_i^{t,\mathrm{sgn}}\}$, one can decompose

$$\begin{aligned}&\sum_{j=1}^m a_{ij,1}\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\widetilde{\boldsymbol{x}}_{i\perp}^t\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}=\sum_{j=1}^m a_{ij,1}\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}+\\ &\sum_{j=1}^m a_{ij,1}\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}_{i\perp}^t-\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right)\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}.\end{aligned} \tag{71}$$

Note that $a_{ij,1}\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}$ in (71) is statistically independent of ξ_{ij} (41) and $\boldsymbol{b}_j^{\mathrm{sgn}}\boldsymbol{b}_j^{\mathrm{sgn}\,H}=\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}$. Hence we can consider $\sum_{j=1}^m a_{ij,1}\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}$ as a weighted sum of the ξ_{ij}'s and exploit the Bernstein inequality to derive that

$$\begin{aligned}&\left\|\sum_{j=1}^m\xi_{ij}\left(a_{ij,1}\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right)\right\|\\ &\lesssim\sqrt{V_1\log m}+B_1\log m\end{aligned} \tag{72}$$

with probability exceeding $1-\mathcal{O}(m^{-10})$, where $V_1:=\sum_{j=1}^m|a_{ij,1}|^2|\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}|^2|\,\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}|^2$, $B_1:=\max_{1\le j\le m}|a_{ij,1}||\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}||\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}|$. In view of Lemma 17 and the incoherence condition (49d) to deduce that with probability at least $1-\mathcal{O}(m^{-10})$,

$$\begin{aligned}V_1&\lesssim\left\|\sum_{j=1}^m|a_{i,1}|^2\left|\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right|^2\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right\|\\ \|\boldsymbol{b}_j\|_2^2&\lesssim\frac{K}{m}\left\|\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right\|_2^2\end{aligned}$$

with the proviso that $m\gg\max\{K,N\}\log^3 m$. Furthermore, the incoherence condition (49d) together with the fact (66) implies that $B_1\lesssim\frac{K}{m}\log m\|\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\|_2$. Substitute

the bounds on V_1 and B_1 back to (72) to obtain

$$\left\|\sum_{j=1}^{m} a_{ij,1}\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right\| \lesssim \sqrt{\frac{K\log m}{m}}\left\|\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right\|_2 \tag{73}$$

as long as $m \gtrsim K\log^3 m$. In addition, we move to the second term on the right-hand side of (71). Let $\boldsymbol{u} = \sum_{j=1}^{m} a_{ij,1}\boldsymbol{a}_{kj}^{\mathsf{H}}\boldsymbol{z}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}$, where $\boldsymbol{z} \in \mathbb{C}^{N-1}$ is independent with $\{\boldsymbol{a}_{kj}\}$ and $\|\boldsymbol{z}\|_2 = 1$. Hence, we have

$$\begin{aligned}&\left\|\sum_{j=1}^{m} a_{ij,1}\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}_{i\perp}^{t}-\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right)\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right\| \\ &\le \|\boldsymbol{u}\|_2\left\|\widetilde{\boldsymbol{x}}_{i\perp}^{t}-\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right\|_2 \lesssim \sqrt{\frac{K\log m}{m}}\left\|\widetilde{\boldsymbol{x}}_{i\perp}^{t}-\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right\|_2,\end{aligned} \tag{74}$$

with probability exceeding $1-\mathcal{O}(m^{-10})$, as long as that $m \gg K\log^3 m$. Here, the last inequality of (74) comes from Lemma 13. Substituting the above two bounds (73) and (74) into (71), it yields

$$\begin{aligned}\left\|\sum_{j=1}^{m} a_{ij,1}\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\widetilde{\boldsymbol{x}}_{i\perp}^{t}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right\| &\lesssim \sqrt{\frac{K\log m}{m}}\left\|\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right\|_2 \\ &+\sqrt{\frac{K\log m}{m}}\left\|\widetilde{\boldsymbol{x}}_{i\perp}^{t}-\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right\|_2.\end{aligned} \tag{75}$$

Combining (70) and (75), we arrive at

$$\begin{aligned}|J_{i3}| &\lesssim \sqrt{\frac{s^2K\log m}{m}}\left\|\widetilde{\boldsymbol{x}}_{i\perp}^{t}\right\|_2 \\ &+\sqrt{\frac{s^2K\log m}{m}}\left\|\widetilde{\boldsymbol{x}}_{i\perp}^{t}-\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right\|_2,\end{aligned} \tag{76}$$

by exploiting the fact that $\|\widetilde{\boldsymbol{h}}_k^t\|^2 \lesssim 1$ for $k = 1,\ldots,s$ and the triangle inequality $\|\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\|_2 \le \|\widetilde{\boldsymbol{x}}_{i\perp}^{t}\|_2 + \|\widetilde{\boldsymbol{x}}_{i\perp}^{t}-\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\|_2$.

- Collecting the bounds for J_{i1}, J_{i2} and J_{i3}, we arrive at

$$\begin{aligned}\widetilde{x}_{i1}^{t+1} &= \widetilde{x}_{i1}^{t}+\eta' J_{i1}-\eta' J_{i2}-\eta' J_{i3} \\ &= (1-\eta)\,x_{i1}^{t}+\eta q_i\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{h}_i^{t}/\|\widetilde{\boldsymbol{h}}_i^{t}\|_2^2+R,\end{aligned} \tag{77}$$

where the residual term R follows that

$$\begin{aligned}|R| \lesssim \frac{\eta}{\|\widetilde{\boldsymbol{h}}_i^{t}\|_2^2}\sqrt{\frac{s^2K}{m}\log m}\Big(&\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{h}_i^{t}+\left|\widetilde{x}_{i1}^{t}\right|+\left\|\widetilde{\boldsymbol{x}}_{i\perp}^{t}\right\|_2 \\ &+\left\|\widetilde{\boldsymbol{x}}_{i\perp}^{t}-\check{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}\right\|_2\Big).\end{aligned} \tag{78}$$

Substituting the hypotheses (48) into (77) and in view of the fact $\alpha_{\boldsymbol{x}_i^t} = \langle\boldsymbol{x}^t,\bar{\boldsymbol{x}}\rangle/\|\bar{\boldsymbol{x}}_i\|_2$ and the assumption that $\|\bar{\boldsymbol{h}}_i\|_2 = \|\bar{\boldsymbol{x}}_i\|_2 = q_i$ for $i = 1,\ldots,s$, one has

$$\begin{aligned}&\alpha_{\boldsymbol{x}_i^{t+1}} \\ &= (1-\eta)\,\alpha_{\boldsymbol{x}_i^t}+\eta'' q_i\bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^{t}+\mathcal{O}\left(\eta''\sqrt{\frac{s^2K}{m}\log m}\alpha_{\boldsymbol{x}_i^t}\right) \\ &+\mathcal{O}\left(\eta''\sqrt{\frac{s^2K}{m}\log m}\beta_{\boldsymbol{x}_i^t}\right)+\mathcal{O}\left(\eta''\sqrt{\frac{s^2K}{m}\log m}\cdot\alpha_{\boldsymbol{h}_i^t}\right) \\ &+\mathcal{O}\left(\eta''\alpha_{\boldsymbol{x}_i^t}\left(1+\frac{1}{s\log m}\right)^t C_3\sqrt{\frac{s\mu^2N\log^8 m}{m}}\right) \\ &= \left(1-\eta+\frac{\eta q_i\psi_{\boldsymbol{x}_i^t}}{\alpha_{\boldsymbol{x}_i^t}^2+\beta_{\boldsymbol{x}_i^t}^2}\right)\alpha_{\boldsymbol{x}_i^t}+\eta(1-\rho_{\boldsymbol{x}_i^t})\frac{q_i\alpha_{\boldsymbol{h}_i^t}}{\alpha_{\boldsymbol{h}_i^t}^2+\beta_{\boldsymbol{h}_i^t}^2},\end{aligned} \tag{79}$$

where $\eta'' = \eta/(q_i\|\boldsymbol{h}_i^t\|_2^2)$, for some $|\psi_{\boldsymbol{x}_i^t}|, |\rho_{\boldsymbol{x}_i^t}| \ll \frac{1}{\log m}$, provided that

$$\sqrt{\frac{s^2K\log m}{q_i^2\,m}} \ll \frac{q_i}{\log m}, \tag{80a}$$

$$\sqrt{\frac{s^2K\log m}{q_i^2\,m}}\beta_{\boldsymbol{x}_i^t} \ll \frac{q_i}{\log m}\alpha_{\boldsymbol{x}_i^t}, \tag{80b}$$

$$\left(1+\frac{1}{s\log m}\right)^t C_3\sqrt{\frac{s\mu^2N\log^8 m}{q_i^2\,m}} \ll \frac{q_i}{\log m}, \tag{80c}$$

where the parameter q_i is assumed to be $0 < q_i \le 1$. Therein, the first condition (80a) naturally holds as long as $m \gg s^2K\log^3 m$. In addition, the second condition (80b) holds true since $\beta_{\boldsymbol{x}_i^t} \le \|\boldsymbol{x}_i^t\|_2 \lesssim \alpha_{\boldsymbol{x}_i^t}\sqrt{\log^5 m}$ (based on (48j)) and $m \gg s^2K\log^8 m$. For the last condition (80c), we have for $t \le T_1 = \mathcal{O}(s\log\max\{K,N\})$, $(1+\frac{1}{s\log m})^t = \mathcal{O}(1)$, which further implies $(1+\frac{1}{s\log m})^t C_3\sqrt{\frac{s\mu^2N\log^8 m}{q_i^2\,m}} \lesssim C_3\sqrt{\frac{s\mu^2N\log^8 m}{q_i^2\,m}} \ll \frac{q_i}{\log m}$ as long as the number of samples obeys $m \gg s\mu^2N\log^{10} m$. This concludes the proof.

Despite it turns to be more tedious when proving (31a), similar arguments used above can be applied to the proof of (31a). Specifically, according to the Wirtinger flow gradient update rule (12a), the signal component $\langle\bar{\boldsymbol{h}}_i,\widetilde{\boldsymbol{h}}_i^t\rangle$ can be represented as follows

$$\begin{aligned}\bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^{t+1} &= \bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^{t}-\frac{\eta}{\|\widetilde{\boldsymbol{x}}_i^t\|_2^2} \\ &\times\sum_{j=1}^{m}\left(\sum_{k=1}^{s}\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_k^{t}\widetilde{\boldsymbol{x}}_k^{t\,H}\boldsymbol{a}_{kj}-y_j\right)\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{a}_{ij}^{\mathsf{H}}\widetilde{\boldsymbol{x}}_i^{t}.\end{aligned}$$

Expanding this expression using $\boldsymbol{a}_{kj}^{\mathsf{H}}\boldsymbol{x}_k^t = x_{k\|}^t a_{kj,1}^* + \boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\boldsymbol{x}_{k\perp}^t$ and rearranging terms, we are left with

$$\bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^{t+1} = \bar{\boldsymbol{h}}_i^{\mathsf{H}}\widetilde{\boldsymbol{h}}_i^{t}-\eta_i' L_{i1}+\eta_i' L_{i2}+\eta_i' L_{i3}, \tag{81}$$

where

$$L_{i1} = \sum_{j=1}^{m}\sum_{k=1}^{s}\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}_k^{t}\widetilde{\boldsymbol{x}}_k^{t\,H}\boldsymbol{a}_{kj}\boldsymbol{a}_{ij}^{\mathsf{H}}\boldsymbol{x}_i,$$

$$L_{i2} = \sum_{j=1}^{m}\sum_{k=1}^{s}\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\bar{\boldsymbol{h}}_k a_{kj,1}q_k a_{ij,1} * t\,\widetilde{x}_{i1}^{t},$$

$$L_{i3}=\sum_{j=1}^{m}\sum_{k=1}^{s}\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\bar{\boldsymbol{h}}_k\boldsymbol{a}_{ij,\perp}^{\mathsf{H}}\boldsymbol{x}_{i\perp}^{t}a_{kj,1}q_k,$$

$$\eta_i'=\eta/\|\widetilde{\boldsymbol{x}}_i^t\|_2^2.$$

Here, L_{i1}, L_{i2} and L_{i3} can be controlled via the strategies exploited to control J_{i1}, J_{i2} and J_{i3}. The proof of (31d) can be derived based on the same argument.

APPENDIX C
PROOF OF (52) IN LEMMA 5

By applying the arguments in [27, Appendix F], it yields that

$$\begin{aligned}&\operatorname{dist}\left(\boldsymbol{x}_i^{t+1,(l)},\widetilde{\boldsymbol{x}}_i^{t+1}\right)\\&\le\kappa\sqrt{\sum\nolimits_{k=1}^{s}\max\left\{\left|\frac{\omega_i^{t+1}}{\omega_i^t}\right|,\left|\frac{\omega_i^t}{\omega_i^{t+1}}\right|\right\}^2\|\boldsymbol{J}_k\|^2},\end{aligned}\tag{82}$$

where ω_i^t is the alignment parameter and $\boldsymbol{J}_k=\omega_k^t\boldsymbol{x}_k^{t+1}-\omega_{k,\text{mutual}}^{t,(l)}\boldsymbol{x}_k^{t+1,(l)}$ where $\omega_{k,\text{mutual}}^{t,(l)}$ is defined in (45). According to (17) and (45), we arrive at

$$\begin{aligned}&\omega_i^t x_{i1}^{t+1}-\omega_{i,\text{mutual}}^{t,(l)}x_{i1}^{t+1,(l)}\\&=\widetilde{x}_{i1}^t-\widehat{x}_{i1}^{t,(l)}-\eta'\boldsymbol{e}_1^\top\left(\nabla_{\boldsymbol{x}_i}f\left(\widetilde{\boldsymbol{z}}^t\right)-\nabla_{\boldsymbol{x}_i}f^{(l)}\left(\widehat{\boldsymbol{z}}_i^{t,(l)}\right)\right)\\&\quad-\eta'\left(\sum_{k=1}^{s}\widehat{\boldsymbol{h}}_i^{t,(l)H}\boldsymbol{b}_l\boldsymbol{a}_{kl}^{\mathsf{H}}\widehat{\boldsymbol{x}}_i^{t,(l)}-\bar{\boldsymbol{h}}_k^{\mathsf{H}}\boldsymbol{b}_l\boldsymbol{a}_{kl}^{\mathsf{H}}\bar{\boldsymbol{x}}_k\right)\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}_i^{t,(l)}a_{il,1},\end{aligned}$$

where the stepsize $\eta'=\eta/\|\widetilde{\boldsymbol{h}}_i^t\|_2^2$. It follows from the fundamental theorem of calculus [37, Theorem 4.2] that

$$\begin{aligned}&\widetilde{x}_{i1}^{t+1}-\widehat{x}_{i1}^{t+1,(l)}=\\&\left\{\widetilde{x}_{i1}^t-\widehat{x}_{i1}^{t,(l)}-\eta'\left(\int_0^1\boldsymbol{e}_1^\top\nabla_{\boldsymbol{x}_i}^2f\left(\boldsymbol{z}\left(\tau\right)\right)\mathrm{d}\tau\right)\begin{bmatrix}\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\\\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\end{bmatrix}\right\}\\&-\eta'\left[\left(\sum_{k=1}^{s}\widehat{\boldsymbol{h}}_i^{t,(l)H}\boldsymbol{b}_l\boldsymbol{a}_{kl}^{\mathsf{H}}\widehat{\boldsymbol{x}}_i^{t,(l)}-\bar{\boldsymbol{h}}_k^{\mathsf{H}}\boldsymbol{b}_l\boldsymbol{a}_{kl}^{\mathsf{H}}\bar{\boldsymbol{x}}_k\right)\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}_i^{t,(l)}a_{il,1}\right],\end{aligned}\tag{83}$$

where $\boldsymbol{z}(\tau)=\widetilde{\boldsymbol{z}}^t+\tau(\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t)$ with $0\le\tau\le1$ and the Wirtinger Hessian with respect to $\boldsymbol{x}_i$ is

$$\nabla_{\boldsymbol{x}_i}^2f\left(\boldsymbol{z}\right)=\begin{bmatrix}\boldsymbol{D}&\boldsymbol{E}\\\boldsymbol{E}^{\mathsf{H}}&\left(\boldsymbol{D}^{\mathsf{H}}\right)^\top\end{bmatrix},\tag{84}$$

with $\boldsymbol{D}=\sum_{j=1}^m|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i|^2\boldsymbol{a}_{ij}\boldsymbol{a}_{ij}^{\mathsf{H}}$, $\boldsymbol{E}=\sum_{j=1}^m\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i\cdot\left(\boldsymbol{a}_{ij}\boldsymbol{a}_{ij}^{\mathsf{H}}\boldsymbol{x}_i\right)^\top$.

- We begin by controlling the second term of (83). Based on (50a) and the hypothesis (48a), we obtain $\max_{1\le i\le s,1\le l\le m}|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}_i^{t,(l)}|\cdot\|\widehat{\boldsymbol{h}}_i^{t,(l)}\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m$. Along with the standard concentration results $|\boldsymbol{a}_{il}^{\mathsf{H}}\boldsymbol{x}_i^{t,(l)}|\lesssim\sqrt{\log m}\|\boldsymbol{x}_i^{t,(l)}\|_2$, one has

$$\begin{aligned}&\left|\left(\sum_{k=1}^{s}\widehat{\boldsymbol{h}}_i^{t,(l)H}\boldsymbol{b}_l\boldsymbol{a}_{kl}^{\mathsf{H}}\widehat{\boldsymbol{x}}_i^{t,(l)}-\bar{\boldsymbol{h}}_k^{\mathsf{H}}\boldsymbol{b}_l\boldsymbol{a}_{kl}^{\mathsf{H}}\bar{\boldsymbol{x}}_k\right)\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}_i^{t,(l)}a_{il,1}\right|\\&\lesssim\frac{s\mu^2\log^5 m}{m}\left\|\widehat{\boldsymbol{x}}_i^{t,(l)}\right\|_2.\end{aligned}\tag{85}$$

- It remains to bound the first term in (83). To achieve this, we first utilize the decomposition $\boldsymbol{a}_{ij}^{\mathsf{H}}(\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)})=a_{ij,1}{}^*(\widetilde{x}_{i1}^t-\widehat{x}_{i1}^{t,(l)})+\boldsymbol{a}_{ij,\perp}^{\mathsf{H}}(\widetilde{\boldsymbol{x}}_{i\perp}^t-\widehat{\boldsymbol{x}}_{i\perp}^{t,(l)})$ to obtain that

$$\begin{aligned}&\boldsymbol{e}_1^\top\left(\nabla_{\boldsymbol{x}_i}^2f\left(\boldsymbol{z}\left(\tau\right)\right)\mathrm{d}\tau\right)\begin{bmatrix}\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\\\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\end{bmatrix}\\&=\omega_1\left(\tau\right)+\omega_2\left(\tau\right)+\omega_3\left(\tau\right),\end{aligned}$$

where

$$\begin{aligned}\omega_1\left(\tau\right)&=\sum_{j=1}^{m}|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i(\tau)|^2a_{ij,1}a_{ij,1}^*\left(\widetilde{x}_{i1}^t-\widehat{x}_{i1}^{t,(l)}\right),\\\omega_2\left(\tau\right)&=\sum_{j=1}^{m}|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i(\tau)|^2a_{ij,1}\boldsymbol{a}_{ij,\perp}^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}_{i\perp}^t-\widehat{\boldsymbol{x}}_{i\perp}^{t,(l)}\right),\\\omega_3\left(\tau\right)&=\sum_{j=1}^{m}\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i(\tau)\boldsymbol{a}_{ij}^{\mathsf{H}}\boldsymbol{x}_i(\tau)b_{j,1}\boldsymbol{a}_{ij}^\top\left(\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\right).\end{aligned}$$

Based on Lemma 10, Lemma 14 and the fact $\|\boldsymbol{b}_j\|_2=\sqrt{K/m}$, by exploiting the techniques in Appendix B, $\omega_1(\tau)$, $\omega_2(\tau)$ and $\omega_3(\tau)$ can be bounded as follows:

$$\begin{aligned}\omega_1\left(\tau\right)&=\|\boldsymbol{h}_i(\tau)\|_2^2\left(\widetilde{x}_{i1}^t-\widehat{x}_{i1}^{t,(l)}\right)\\&\quad+\mathcal{O}\left(\sqrt{\frac{s^2\mu^2K\log m}{m}}\left(\widetilde{x}_{i1}^t-\widehat{x}_{i1}^{t,(l)}\right)\right)\end{aligned}\tag{86}$$

$$\begin{aligned}|\omega_2(\tau)|&\lesssim\sqrt{\frac{K\log^2 m}{m}}\left(\left\|\widetilde{\boldsymbol{x}}_{i\perp}^t-\widehat{\boldsymbol{x}}_{i\perp}^{t,(l)}\right\|_2\right.\\&\quad\left.+\left\|\widetilde{\boldsymbol{x}}_{i\perp}^t-\widehat{\boldsymbol{x}}_{i\perp}^{t,(l)}-\widetilde{\boldsymbol{x}}_{i\perp}^{t,\text{sgn}}-\widehat{\boldsymbol{x}}_{i\perp}^{t,\text{sgn},(l)}\right\|_2\right)\end{aligned}\tag{87}$$

$$\begin{aligned}\omega_3\left(\tau\right)&=|h_{i1}(\tau)|\left(\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\right)^{\mathsf{H}}\boldsymbol{x}_i(\tau)\\&\quad+\mathcal{O}\left(\frac{1}{\log^5 m}\left\|\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\right\|_2\right)\end{aligned}\tag{88}$$

with probability at least $1-\mathcal{O}(m^{-10})$, provided that $m\gg\mu^2K\log^{13}m$.

- Combining the bounds (85) (86), (87) and (88), one has

$$\begin{aligned}&\widetilde{x}_{i1}^{t+1}-\widehat{x}_{i1}^{t+1,(l)}\\&=\left(1-\eta\frac{\int_0^1\|\boldsymbol{h}_i(\tau)\|_2^2\mathrm{d}\tau}{\|\widetilde{\boldsymbol{h}}_i^t\|_2^2}+\mathcal{O}\left(\eta'\sqrt{\frac{s^2\mu^2K\log m}{m}}\right)\right)\cdot\\&\left(\widetilde{x}_{i1}^t-\widehat{x}_{i1}^{t,(l)}\right)+\mathcal{O}\left(\eta'\frac{s\mu^2\log^5 m}{m}\left\|\widehat{\boldsymbol{x}}_i^{t,(l)}\right\|_2\right)\\&+\mathcal{O}\left(\eta'\sqrt{\frac{K\log^2 m}{m}}\left(\left\|\widetilde{\boldsymbol{x}}_{i\perp}^t-\widehat{\boldsymbol{x}}_{i\perp}^{t,(l)}\right\|_2\right.\right.\end{aligned}$$

$$+\left\|\widetilde{\boldsymbol{x}}_{i\perp}^{t}-\widehat{\boldsymbol{x}}_{i\perp}^{t,(l)}-\widetilde{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn}}-\widehat{\boldsymbol{x}}_{i\perp}^{t,\mathrm{sgn},(l)}\right\|_2\Bigg)\Bigg)$$

$$+\mathcal{O}\left(\eta'\frac{1}{\log^5 m}\left\|\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\right\|_2\right)$$

$$+\eta'\sup_{0\le\tau\le 1}|h_{i1}(\tau)|\left(\widetilde{\boldsymbol{x}}_i^t-\widehat{\boldsymbol{x}}_i^{t,(l)}\right)^{\mathsf{H}}\boldsymbol{x}_i(\tau).$$

By exploiting similar arguments in Appendix E in [34], we can arrive at

$$\begin{aligned}&\mathrm{dist}\left(x_{i1}^{t+1,(l)},\widetilde{x}_{i1}^{t+1}\right)\\&\le(1-\eta+\varrho_2)\,\alpha_{\boldsymbol{x}_i^t}\left(1+\frac{1}{s\log m}\right)^t C_2\frac{s\mu^2\kappa\sqrt{N\log^{13}m}}{m}\\&\le\alpha_{\boldsymbol{x}_i^{t+1}}\left(1+\frac{1}{s\log m}\right)^{t+1}C_2\frac{s\mu^2\kappa\sqrt{N\log^{13}m}}{m}\end{aligned}$$

for some $|\varrho_2|\ll\frac{1}{\log m}$ provided that $m\ge Cs\mu^2\kappa N\log^{12}m$ for some sufficiently large constant $C>0$.

APPENDIX D
TECHNICAL LEMMAS

Lemma 11: Suppose $m\gg K\log^3 m$. With probability exceeding $1-\mathcal{O}(m^{-10})$, we have $\|\sum_{j=1}^m a_{ij,1}^*a_{ij,1}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}-\boldsymbol{I}_K\|\lesssim\sqrt{\frac{K}{m}\log m}$.

Lemma 12: Suppose $m\gg K\log^3 m$. For $k\ne i$, we have $\|\sum_{j=1}^m a_{kj,1}^*a_{ij,1}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\|\lesssim\sqrt{\frac{K}{m}\log m}$, $\|\sum_{j=1}^m|a_{kj,1}||a_{ij,1}|\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\|\lesssim\sqrt{\frac{K}{m}\log m}$, with probability exceeding $1-\mathcal{O}(m^{-10})$.

Lemma 13: Suppose $m\gg K\log^3 m$ and $\boldsymbol{z}\in\mathbb{C}^{N-1}$ with $\|\boldsymbol{z}\|_2=1$ is independent with $\{\boldsymbol{a}_{kj}\}$. With probability exceeding $1-\mathcal{O}(m^{-10})$, we have $\|\sum_{j=1}^m a_{ij,1}\boldsymbol{a}_{kj,\perp}^{\mathsf{H}}\boldsymbol{z}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\|\lesssim\sqrt{\frac{K}{m}\log m}$.

Remark 5: Lemma 12, Lemma 13 and Lemma 11 can be proven by applying the arguments in [36, Section D.3.3].

Lemma 14: Suppose $m\gg(\mu^2/\delta^2)N\log^5 m$. With probability exceeding $1-\mathcal{O}(m^{-10})$, we have $\|\sum_{j=1}^m|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i|^2\boldsymbol{a}_{ij,\perp}\boldsymbol{a}_{ij,\perp}^{\mathsf{H}}-\|\boldsymbol{h}_i\|_2^2\boldsymbol{I}_{N-1}\|\lesssim\delta\|\boldsymbol{h}_i\|_2^2$, obeying $\max_{1\le l\le m}|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}_i|\cdot\left\|\boldsymbol{h}_i\right\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m$. Furthermore, there is $\|\sum_{j=1}^m\sum_{k=1}^s b_{j,1}\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i\boldsymbol{a}_{ij}\boldsymbol{a}_{kj}^{\mathsf{H}}-h_{i1}\boldsymbol{I}_N\|\lesssim\delta\|\boldsymbol{h}_i\|_2$, with probability exceeding $1-\mathcal{O}(m^{-10})$, provided $m\gg(\mu/\delta^2)s^2N\log^3 m$.

Proof: Please refer to Lemma 11 and Lemma 12 in [27]. ■

Lemma 15: Suppose the sampling size $m\gg s\mu^2\sqrt{N\log^9 m}$, then with probability exceeding $1-\mathcal{O}(m^{-10})$, we have $\|\sum_{j=1}^m\sum_{k=1}^s\boldsymbol{h}_k^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i\boldsymbol{a}_{ij}\boldsymbol{a}_{kj}^{\mathsf{H}}-\|\boldsymbol{h}_i\|_2^2\boldsymbol{I}_N\|\lesssim\frac{s\mu^2\sqrt{K\log^9 m}}{m}\|\boldsymbol{h}_i\|_2^2$, obeying $\max_{1\le i\le s,1\le j\le m}|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i|\cdot\left\|\boldsymbol{h}_i\right\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m$.

Lemma 16: Suppose the sampling size follows that $m\gg s\mu^2\sqrt{N\log^5 m}$. With probability exceeding $1-\mathcal{O}(m^{-10})$, we have $\|\sum_{j=1}^m\sum_{k=1}^s\bar{\boldsymbol{h}}_k^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_i\boldsymbol{a}_{ij}\boldsymbol{a}_{kj}^{\mathsf{H}}-(\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{h}_i)\boldsymbol{I}_N\|\lesssim\frac{s\mu^2\sqrt{K\log^5 m}}{m}|\bar{\boldsymbol{h}}_i^{\mathsf{H}}\boldsymbol{h}_i|$, obeying $\max_{1\le l\le m}|\boldsymbol{b}_l^{\mathsf{H}}\bar{\boldsymbol{h}}_i|\cdot\left\|\bar{\boldsymbol{h}}_i\right\|_2^{-1}\le\frac{\mu}{\sqrt{m}}$ and $\max_{1\le l\le m}|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}_i|\cdot\left\|\boldsymbol{h}_i\right\|_2^{-1}\lesssim\frac{\mu}{\sqrt{m}}\log^2 m$.

Remark 6: The proof of Lemma 15 and 16 exploits the same strategy as [34, Section K] does.

Lemma 17: Suppose that $\boldsymbol{a}_{ij}$ and $\boldsymbol{b}_j$ follows the definition in Section II. $1\le i\le s,1\le j\le m$. Consider any $\epsilon>3/n$ where $n=\max\{K,N\}$. Let $\mathcal{S}:=\{\boldsymbol{z}\in\mathbb{C}^{N-1}|\max_{1\le j\le m}|\boldsymbol{a}_{ij,\perp}^{\mathsf{H}}\boldsymbol{z}|\le\beta\|\boldsymbol{z}\|_2\}$, where β is any value obeying $\beta\ge c_1\sqrt{\log m}$ for some sufficiently large constant $c_1>0$. Then with probability exceeding $1-\mathcal{O}(m^{-10})$, one has

1) $|\sum_{j=1}^m|a_{ij,1}|^2|\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\boldsymbol{z}|^2\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}-\|\boldsymbol{z}\|_2\boldsymbol{I}_K|\le\epsilon\|\boldsymbol{z}\|_2$ for all $\boldsymbol{z}\in\mathcal{S}$, provided that $m\ge c_0\max\{\frac{1}{\epsilon^2}n\log n,\ \frac{1}{\epsilon}\beta^2 n\log^2 m\}$.
2) $|\sum_{j=1}^m|a_{ij,1}||\boldsymbol{a}_{kj\perp}^{\mathsf{H}}\boldsymbol{z}||\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}|\le\epsilon\|\boldsymbol{z}\|_2$ for all $\boldsymbol{z}\in\mathcal{S}$, provided that $m\ge c_0\max\{\frac{1}{\epsilon^2}n\log n,\ \frac{1}{\epsilon}\beta n\log^{\frac{1}{2}}m\}$.

Here, $c_0>0$ is some sufficiently large constant.

Proof: Please refer to Lemma 12 in [34]. ■

REFERENCES

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, Jun. 2015.

[2] J. G. Andrews *et al.*, "What will 5G be?," *IEEE J. Sel. Area. Comm.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] G. Zhu and K. Huang, "MIMO over-the-air computation for high-mobility multimodal sensing," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6089–6103, Aug. 2019.

[4] M. Goldenbaum, H. Boche, and S. Stanczak, "Nomographic functions: Efficient computation in clustered gaussian sensor networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2093–2105, Apr. 2015.

[5] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[6] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, Sep. 2007.

[7] R. Soundararajan and S. Vishwanath, "Communicating linear functions of correlated Gaussian sources over a MAC," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1853–1860, Mar. 2012.

[8] J. J. Xiao, S. Cui, Z. Q. Luo, and A. J. Goldsmith, "Linear coherent decentralized estimation," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 757–770, Feb. 2008.

[9] C. H. Wang, A. S. Leong, and S. Dey, "Distortion outage minimization and diversity order analysis for coherent multiaccess," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6144–6159, Dec. 2011.

[10] M. Goldenbaum, H. Boche, and S. Stańczak, "Harnessing interference for analog function computation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 4893–4906, Oct. 2013.

[11] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Over-the-air computation for IoT networks: Computing multiple functions with antenna arrays," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5296–5306, Dec. 2018.

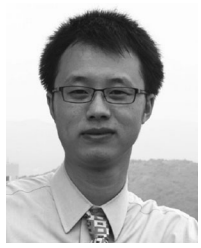
[12] X. Li, G. Zhu, Y. Gong, and K. Huang, "Wirelessly powered data aggregation for iot via over-the-air function computation: Beamforming and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3437–3452, Jul. 2019.

[13] T. Jiang and Y. Shi, "Over-the-air computation via intelligent reflecting surfaces," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Waikoloa, Hawaii, USA, Dec. 2019.

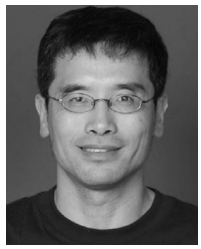
[14] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: Ai empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[15] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.
[16] K. Yang, Y. Shi, J. Zhang, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, Jan. 2019, doi 10.1109/TWC.2019.2961673.
[17] D. Wen, G. Zhu, and K. Huang, "Reduced-dimension design of MIMO over-the-air computing for data aggregation in clustered IoT networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5255–5268, Nov. 2019.
[18] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, Jan. 2020.
[19] M. M. Amiri, T. M. Duman, and D. Gunduz, "Collaborative machine learning at the wireless edge with blind transmitters," in *Proc. IEEE Global Conf. Signal and Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.
[20] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, 2019, pp. 1432–1436.
[21] T. Sery and K. Cohen, "A sequential gradient-based multiple access for distributed learning over fading channels," in *Proc. IEEE 53 rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, 2019, pp. 303–307.
[22] G. Zhu, Y. Du, D. Gunduz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," 2020, *arXiv: 2001.05713*.
[23] M. Goldenbaum and S. Stanczak, "On the channel estimation effort for analog computation over wireless multiple-access channels," *IEEE Wireless Commun. Lett.*, vol. 3, no. 3, pp. 261–264, Jun. 2014.
[24] S. Ling and T. Strohmer, "Regularized gradient descent: A nonconvex recipe for fast joint blind deconvolution and demixing," *Inf. Inference: J. IMA*, vol. 8, no. 1, pp. 1–49, Mar. 2018.
[25] S. Ling and T. Strohmer, "Blind deconvolution meets blind demixing: Algorithms and performance bounds," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4497–4520, Jul. 2017.
[26] J. Dong, K. Yang, and Y. Shi, "Blind demixing for low-latency communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 897–911, Feb. 2019.
[27] J. Dong and Y. Shi, "Nonconvex demixing from bilinear measurements," *IEEE Trans. Signal Process.*, vol. 66, no. 19, pp. 5152–5166, Oct. 2018.
[28] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points-online stochastic gradient for tensor decomposition," in *Proc. Conf. Learn. Theory*, 2015, pp. 797–842.
[29] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Found. Comput. Math*, vol. 18, no. 5, pp. 1131–1198, Oct. 2018.
[30] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1724–1732.
[31] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Proc. Adv. Neural. Inf. Process. Syst.*, 2016, pp. 3873–3881.
[32] 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.*, 2017, pp. 1233–1242.
[33] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 742–769, Feb. 2019.
[34] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Math. Program.*, vol. 176, no. 1/2, pp. 5–37, Jul. 2019.
[35] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.
[36] 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.*, 2018, pp. 3345–3354.
[37] K. Pothoven, *Real and Functional Analysis*. Berlin, Germany: Springer Science & Business Media, 2013.

Jialin Dong (Student Member, IEEE) received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. She is currently working toward the graduate degree with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. Her research interests include mathematical optimization and high-dimensional probability.



Yuanming Shi (Member, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011 and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2015. Since September 2015, he has been with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, where he is currently a tenured Associate Professor. He visited University of California, Berkeley, CA, USA, from October 2016 to February 2017. He was the recipient of the 2016 IEEE Marconi Prize Paper Award in Wireless Communications, and the 2016 Young Author Best Paper Award by the IEEE Signal Processing Society. He is an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. His research interests include optimization, statistics, machine learning, signal processing, and their applications to 6 G, IoT, AI, and FinTech.



Zhi Ding (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, USA, in 1990. He is a Professor of electrical and computer engineering with the University of California, Davis, CA, USA. From 1990 to 2000, he was a Faculty Member with Auburn University and later, University of Iowa. He has held visiting positions with Australian National University, Hong Kong University of Science and Technology, NASA Lewis Research Center, and USAF Wright Laboratory. He has active collaboration with researchers from Australia, Canada, China, Finland, Hong Kong, Japan, Korea, Singapore, and Taiwan.

He has serving on technical programs of several workshops and conferences. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1994 to 1997 and 2001 to 2004, and an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS 2002–2005. He was a member of technical committee on Statistical Signal and Array Processing and a member of technical committee on Signal Processing for Communications (1994–2003). He was the General Chair of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing and the Technical Program Chair of the 2006 IEEE Globecom. He was also an IEEE Distinguished Lecturer (Circuits and Systems Society, 2004–2006, Communications Society, 2008–2009). He served on as IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS Steering Committee Member (2007–2009) and its Chair (2009–2010). He is a Co-Author of the text: *Modern Digital and Analog Communication Systems*, 5th ed., (Oxford University Press, 2019).